Volume 3 – Número 3 – Agosto 2018

Revista de Engenharia e Pesquisa Aplicada

Edição Especial - Ciência de Dados e Analytics

http://dx.doi.org/10.25286/repa.v3i3

ISSN: 2525-4251

Qualis CAPES B5

Engenharias

Coordenação Setorial de Pós-Graduação e Pesquisa

UPE

POLI ESCOLA POLITÉCNICA DE PERNAMBUCO

Revista de

Engenharia e Pesquisa Aplicada

Edição Especial Ciência de Dados e Analytics

Volume 3 - Número 3 – Agosto 2018

ISSN: 2525-4251 (versão on line)

Revista de Engenharia e Pesquisa Aplicada

Volume 3 - **Número 3 –Agosto 2018**

Foco e Escopo
A Revista de Engenharia e Pesquisa Aplicada é uma publicação da Universidade de Pernambuco que tem como objetivo ser um canal de divulgação de trabalhos nas áreas de engenharia, computação e áreas tecnológicas convergentes.

Processo de Avaliação
O processo de avaliação da Revista de Engenharia e Pesquisa Aplicada será realizado por pares acadêmicos com expertise na área.

Periodicidade
Trimestral.

Política de Acesso Livre
Esta revista oferece acesso livre imediato ao seu conteúdo, seguindo o princípio de que disponibilizar gratuitamente o conhecimento científico ao público proporciona maior democratização mundial do conhecimento.

Universidade de Pernambuco
Reitor: Pedro Henrique de Barros Falcão
Vice-Reitor: Maria do Socorro Cavalcanti

Escola Politécnica de Pernambuco
Diretor: José Roberto Cavalcanti
Vice-Diretor: Alexandre Duarte Gusmão

Endereço
Rua Benfica, 455 – Madalena
Recife/PE - CEP: 50/720-001
Telefone: 55 81 3184-7513
Email: repa@poli.br

CIP Catalogação-na-Publicação
Universidade de Pernambuco Escola Politécnica de Pernambuco
Biblioteca Central

Revista de Engenharia e Pesquisa Aplicada / Universidade de Pernambuco, Escola Politécnica de Pernambuco - Vol.3, no. 3 (2018) - Recife: UPE, 2018.
Trimestral
ISSN 2525-4251 (versão online)
Título abreviado: Rev. Eng. Pesquisa Aplicada.
[1] ENGENHARIA - Periódicos

DOI: http://dx.doi.org/10.25286/repa.v3i3

Revista de Engenharia e Pesquisa Aplicada

Volume 3 – Número 3 – Agosto 2018

Edição Especial Ciência de Dados e Analytics

Iniciativas de Ciência de Dados e Analytics da Universidade de Pernambuco

Uma edição especial voltada para os trabalhos acadêmicos desenvolvidos na UPE

Alexandre Magno Andrade Maciel [1] orcid.org/0000-0003-4348-9291
Rodrigo Lins Rodrigues[2] orcid.org/0000-0002-3598-5204
Mailson Melo dos Santos Filho [3] orcid.org/0000-0002-1711-5301

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Departamento de Educação, Universidade Federal Rural Pernambuco, Recife, Pernambuco, Brasil,
[3] Fábrica de Negócios – *Analytics & Data Minin*g, Recife, Pernambuco, Brasil.

E-mail do autor principal: amam@ecomp.poli.br

Em fevereiro de 2010 a revista The Economist publicou uma edição especial intitulada *Data, Data Everywhere* na qual foi realizado uma profunda análise do cenário mundial sobre o que eles chamaram de uma quantidade inimaginável de informação digital que está ficando cada vez mais vasta, mais rapidamente. Naquele momento, eles avaliaram que esse grande volume de dados poderia proporcionar grandes avanços para a economia e para a sociedade tais como: tendências de mercado, prevenção de doenças, combate à criminalidade entre outros. Por outro lado, estas benesses viriam acompanhadas de uma série de novos problemas tais como: dificuldades de armazenamento de dados oriundos de diversas fontes (*web*, sensores, dispositivos móveis), segurança dos dados e proteção da privacidade dos usuários **[1]**.

No Brasil, este cenário tem demorado alguns anos para ser replicado, como mostra a pesquisa da *Computer World* que mostrou que o mercado nacional representou em 2016 46,8% do mercado na América Latina, gerando receita de US$ 1,16 bilhão. Esta é uma fatia ínfima do mercado visto que a América Latina como um todo corresponde a apenas por 5,1% do mercado global **[2]**. Este crescente interesse do mercado tem demandado também a formação de um conjunto de novos profissionais especializados em Ciência de Dados e Analytics. Cientista de Dados, Arquiteto de Dados, Engenheiro de Dados, Analista de Negócios e Desenvolvedor de Visualização de Dados são alguns dos cargos que tem ganhado cada vez mais relevância dentro das organizações **[3].**

Diante deste contexto global, em maio de 2016 foi proposto o Curso de Especialização em Ciência de Dados e *Analytics* na Escola Politécnica (POLI) da Universidade de Pernambuco (UPE). A proposta, essencialmente inovadora, foi construída a partir de uma colaboração entre as quatro universidades públicas de Pernambuco: UPE, Universidade Federal Rural de Pernambuco (UFRPE), a Universidade Federal de Pernambuco (UFPE), a Universidade de Vale do São Francisco (UNIVASF) e com a participação da empresa Fábrica de Negócios – *Analytics & Data Mining*. O curso está fundamentado num tripé de conhecimento baseado em disciplinas da área de inteligência computacional, de infraestrutura e banco de dados e de negócios e tem como público alvo profissionais da área de computação, engenharias, estatística e administração. O curso encontra-se na sua terceira turma e esta edição especial da Revista de Engenharia e Pesquisa Aplicada (REPA) é aberta com dez artigos que representam o trabalho de conclusão de curso referente a turma de 2016.

Outra ação acadêmica importante da Universidade de Pernambuco que vem ganhando atenção do mercado e do governo local é uma iniciativa chamada Sala de Aula Aberta. Esta ação iniciou no segundo semestre de 2016 com as disciplinas de inteligência artificial e mineração de dados do Curso de Graduação e de Mestrado em Engenharia da Computação, e tem como objetivo, aplicar uma metodologia de Aprendizagem Baseada em Problemas, trazendo para sala de aula

 http://dx.doi.org/10.25286/repa.v3i3.976

demandas reais de problemas complexos de empresas e das unidades gestoras do governo do estado. Durante a disciplina os alunos se debruçam sobre o entendimento das problemáticas, análise e pré-processamento dos dados e, a partir do estudo das diversas abordagens de inteligência artificial, propõem soluções, em nível de prova de conceito, que são entregues aos demandantes e tem potencial para tornar-se soluções tecnológicas de apoio à decisão. Os últimos oito artigos desta edição apresentam os resultados obtidos a partir do Sala de Aula Aberta.

É notória a importância que a área de Ciência de Dados e *Analytics* vem ganhando no Recife, e, isto se comprova na busca por estas duas ações da UPE, além de outras ações que vêm surgindo na cidade como novos cursos de extensão e especialização, *hackthons* promovidos a partir de análise de dados de diversas empresas, bem como o surgimento de diversas startups focadas nesta área. Isto posto, é com enorme satisfação que entregamos esta edição especial da REPA, resultado de um esforço visionário da POLI/UPE, como apoio dos diversos parceiros acadêmicos e da indústria, para nos colocarmos na vanguarda da formação tecnológica do estado, e consolidarmos um grupo de professores/pesquisadores a atuarem de maneira decisiva no fortalecimento do ecossistema de inovação do país.

Referências

[1] Data, data everywehre. **The Economist**, 25 Feb.2010. Disponível em: < https://www.economist.com/special-report/2010/02/25/data-data-everywhere>

[2] Mercado brasileiro de big data e analytic fatura US$ 1,16 BI e já representa 50% AL. **Computerworld**, 21 mar. 2017. Disponível em: < http://computerworld.com.br/2017/03/21/mercado-brasileiro-de-big-data-e-analytics-fatura-us-116-bi-e-ja-representa-quase-50-da-al/>

[3] 10 Carreiras em Big Data e Data Science. **Data Science Academy**, 25 out. 2016. Disponível em: <http://datascienceacademy.com.br/blog/10-carreiras-em-big-data-e-data-science/>

Business Intelligence for Detection of Anomalies in Records of Fueling

A case study on automobiles used for illegal fuel storage in Bolivia

Vanessa Adriana Gironda Aquize[1] orcid.org/0000-0001-9792-0396
Mailson Melo dos Santos Filho[2] orcid.org/0000-0001-5727-2427

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil.
[2] Fábrica de Negócios – Analytics & Data Mining, Recife, Pernambuco, Brasil.

E-mail do autor principal: vanessa.gironda@gmail.com

Abstract

At present, in any organization it is necessary to make decisions, very strategic meetings to achieve a satisfactory development. It is the case of the National Hydrocarbon Agency of Bolivia (ANH), that due the smuggling of fuel, implemented the RFID technology in order to register the records of fueling of all fleet vehicular. From this, a model of Anomaly Detection in records of fueling was proposed trough machine learning techniques. Nevertheless, the huge volume of information about anomalies scores in a local and global level needs to analyzed and take decisions (e.g. strict control in some risk zones according variables analyzed). The collection and analysis of this information, given its heterogeneous character and its volume, usually become a problem for the government institution and this is where Business Intelligence (BI) intervenes, through the Systems of Support of "human decision-making". Today there are many BI solutions, being the Open Source Pentaho Business Intelligence platform one of the most used currently. This management platform covers data analysis and reporting operations, making this a flexible solution to cover our study case: "Anomaly Detection in Records of Fueling in automobiles used for illegal fuel storage in Bolivia". So, in this paper, this technological platform will be applied making some adjustments in the anomaly scores context. The principal contribution is design and development a BI solution responsible of analyze in large amount of records of anomalies in Bolivia and in this way to allow to make better decisions of control of fuel smuggling, having the right information in the right place at the right time.

Key-words:*Pentaho BI;Machine Learning;Anomaly Detection;Decision Making;Data Warehouse.*

 http://dx.doi.org/10.25286/repa.v3i3.925

1 Introduction

Bolivia subsidizes more than 50 % of the fuel costs currently giving rise to the existence of unscrupulous people that are dedicated to the smuggling of this re-source. To address this problem, the institution in charge of regulating, controlling and supervising all the activities of the hydrocarbon in (National Agency of Hydrocarbons, ANH for its acronym in Spanish), implemented the called B-SISA system in which it records the fuel supply of each vehicle through a Radio Frequency Identification (RFID) technology as a control action.

In this context, Buarque et al. **[1]** develop a computational approach for Anomaly Detection based in records of fueling in order to determinate scores of anomalies and in this way identify possible cases of illegal storage via profiling and unsupervised clustering algorithms. However, the massive volumes of stored records after processing the model need to be analyzed and from this to make decisions. The collection and analysis of this information, given its heterogeneous character and its volume, usually become a problem for organizations and this is where Business Intelligence (BI) intervenes, through the Systems of Sup-port of "human decision-making". Today there are many BI solutions, being the Open Source Pentaho Business Intelligence platform one of the most used currently. The platform proposed covers data analysis and reporting operations, making this a flexible solution to cover our study case.

On the one hand, the strategic objective of the ANH is based on: Intelligently manage resources from the model proposed in **[1]** also its implementation in a framework open source flexible to needs of the institution, and in this way achieve maximum operational efficiency, to improve the control of fuel smuggling in the all country.

On the other hand, it is clear that the strategic role that the ANH must assume is to have an updated anomaly in records of fueling and also fuel sales control, which can be guided by the anomalies that vehicles re-fuel most.

Therefore, this paper faced with the challenge of efficiently accomplishing strategic activities, is opportunely evaluated on the use of information technologies for the extraction, transformation, loading and exploitation of the data stored in the company databases in order to adopt the best strategic improvement decisions.

2 Theoretical background

2.1 Business Intelligence (BI)

BI systems ensure obtaining of useful, correct and in-time information, usually taken from disparate data sources. They close the gap between the huge amount of data available to the decision factor, and the report analysis presented in a suggestive way that should sup-port the decision-making process **[3]**. BI offers sophisticate information analysis and information discovery technologies such as Data Warehouse, On-line Analytical Processing (OLAP), Data Mining, etc. BI solutions arrived to the third generation BI, providing access to information, advanced graphical and web-based OLAP, in-formation mining tools and prepackaged applications that exploit the power of those tools **[4]**.

A BI system has four major components: a data warehouse (with its data source), business analytics (a collection of tools for manipulating, mining and analyzing the data from the warehouse), business performance management (for monitoring and analyzing performance) and a user interface (connecting to the system via a browser) **[5]**.

A data warehouse is the core component of a BI infrastructure. The dimensional model of a data warehouse consists in numeric measures, dimensions and fact tables. Related measures are collected into fact tables. The measures can be looked upon in different ways, those ways being called dimensions. A dimension is a particular area of interest such as time, geographic position, category and so on **[4]**.

An OLAP instrument is a combination of analytical processing procedures and graphic presentations **[5]**. OLAP uses the word cube to describe what in the relational world would be the integration of the fact table with dimension tables **[4]**. It generally includes a calculation engine for adding complex analytical logic to the cube, and a query language. Because the standard relational query language (SQL) is not well suited to work with cubes, Multidimensional Expression (MDX), an OLAP-specific query language, has been developed.

Data mining is a technology that uses complex algorithms for data analyzing and discovering valuable in-formation for the decision maker **[5]**.

The emphasis is on data's quality to be valid, previously unknown, comprehensible and actionable.

When designing the data scheme of the warehouse, the following types of schemes may be used: star, snow-flake or constellation **[6]**.

This paper presents a practical solution implemented in a suite of open source Business Intelligence products called Pentaho Business Analytics, providing data integration, OLAP services, reporting, dashboarding, data mining and ETL capabilities.

2.2 Pentaho

Pentaho, founded in 2004, has an open source heritage and provides commercial professional and enterprise editions of its Pentaho Business Analytics technologies through a subscription model as well as open source versions. Along with reporting, interactive data discovery, and predictive analytics capabilities, Pentaho also provides data access and integration. Built to meet demand for ad hoc discovery, visualization, and exploration of large and diverse big data source **[7]**.

The collection of analysis components in Pentaho Business Analytics enables visualizations of data trends by creating static reports from an analysis data source, traversing an analysis cube, showing how data points compare by using charts, and monitoring the status of certain trends and thresholds with dashboards.

The process starts by using any client tools, consolidating data from disparate sources into one canonical source and optimizing it for the metrics wanted to be analyzed; creating an analysis schema to describe the data; iteratively improve that schema so that it meets the users' needs; and create aggregation tables for frequently computed views **[7]**. The architecture of an Open Source BI solution is depicted in Figure 1 **[8]**.

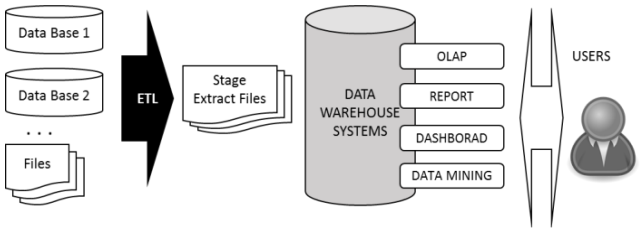


Figure1 - Layer of Information Business Information.
The Pentaho User Console, includes:

(i) Interactive Reporting for quick and easy data-driven reports;

(ii) Pentaho Analyzer an interactive analysis tool that provides a rich Web-based, drag-and-drop user interface;

(iii) Pentaho Dashboard Designer, a layout template, theme, and the content are design.

3 Preliminaries

3.1 About the Anomalies

For clarity, we now introduce some definitions and assumptions to refer the previous model evaluated in **[1]**. This is also used as axiom for our proposed anomaly detection method.

For smuggling fuel, fraudsters accumulate this re-source by making several high purchases in short periods of time (e.g. fifteen times per day). This activity is considerate as irregular fuel supplies since those purchases are above the average consumption referring his own history consumption or historical of vehicles with similar characteristics.

Figure 2 shows an example of a sequence of records that are considered as anomaly due to increased fueling amount in a short period of time, compared to its own historic. (e.g. 350 – 400 liters of fueling in just two days is not normal in reference to its historical behavior). We define these anomaly as "Local Anomaly": vehicles that have irregular fueling records according to its historical data.

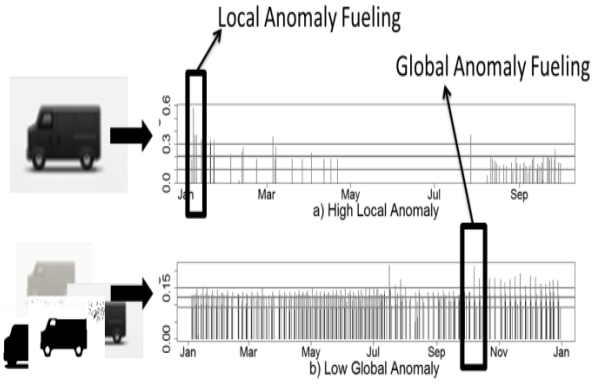


Figure 2 - Fuel supply with consumption approximation and illegal fuel storage.

http://dx.doi.org/10.25286/repa.v3i3.925

On the other hand, the anomaly explained before could be considerate as a normal in reference to others vehicles from similar characteristics like: (1) Type of vehicle (e.g. track, van, motorcycle, etc.), (2) Monthly consumption behavior (e.g. constant, variable, high or low), (3) Area of fuel consumption (e.g. rural, urban), (4) Type of fuel (e.g. diesel, gasoline, vehicular natural gas or combinations).

We define these characteristics such as "context variables" that can influence in the anomaly determination for a set of vehicles. Depending of the context variable the anomaly can be different in reference with Local Anomaly. Thus, we define as "Global Anomaly" to the anomalies of vehicles corresponding to the context variables.

3.2 About the Data

The data provided by the government institution refer to 190.456 records from vehicles samples in a random way. The vehicles correspond to locations in all Bolivia and after cleaning process, we have the following attributes:

Frame 1 - Atributtes of vehicles.

No	Original Attribute	Type	Values
1	Id	Categorical	1000 different id
2	Type of vehicle	Categorical	10 different type
3	Brand of vehicle	Categorical	82 different brands
4	Location of fueling	Categorical	9 Locations
5	Fuel Type	Categorical	4 Fuel type
6	Services Station	Categorical	638 different GS
7	Amount of fueling	Numeric	From 0 to 4480
8	Time of fueling	data	During 2015–2016

3.3 About the Anomaly Detection in Records of Fueling Model

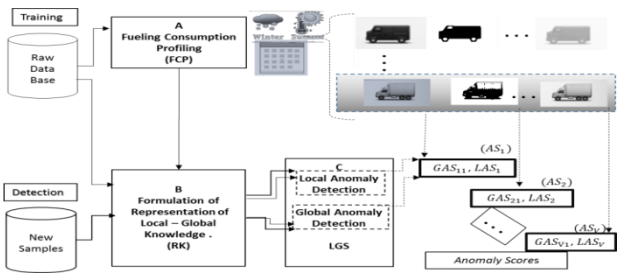


Figure 3 - Anomaly Detection in Records of Fueling Model.

According with Buarque et al**. [1]**, the Anomaly Detection in Records of Fueling (ADRF) uses Local and Global information grouped in four-step process, A-D as is showed in Figure 3. The flowchart starts in step A, in which "Fuel Consumption Profiling" (FCP) computes fuel consumption profiles of a set of vehicles that are expected to be influenced similarly by the same context variables such as location, season, weeks, etc. Thus, the FCP refer to clustering techniques based in contextual information that serves as profiling algorithm. With the profiles predetermined, the step B performs a particular and novel "Representation of Knowledge" (RK) based in historic of fueling records, in order to extracts features using a sliding overlapping window from the historic of fueling per: (1) each vehicle and (2) group profile. The RK is responsible for compute the principal features of what irregular patterns of consumption would mean by the case study.

After the processing performed by the RK, the step C called "Local and Global Scores" (LGS) is activated in order to score each window corresponding to the knowledge that had already been acquired both locally and globally on the last step. This step is responsible for quantifying the level of the anomaly for each fueling record in reference to the historic in fueling records per each vehicle and each profile. The algorithm that score the level of anomalies employ technologies of clustering in order to associate similar patterns corresponding to the historic fueling records. So, it has the objective of favoring the recognition of irregular patterns, besides promoting an analysis of divergences and similarities between patterns. During this step, the memorized values act as an organized repository of anomaly knowledge, storing prototypes that represent the behavior of each vehicle and each profile. This one provides a base set of information that allows the LGS to retrieve irregular patterns making it possible to explain how the final score computed was built by the model.

So, the model compute two important scores by each record of fueling for each vehicle: the first one is the Local Anomaly Score (LAS) and the second one Global Anomaly Score (GAS).

In a proof of use of the model, the results of the accuracy are 82.75 % for Local Anomalies and 99.7 % for Global Anomalies. These results represent a good performance of the proposal due

that 4.314 of 4.433 record of fueling were labeled correctly for the Global Case and 3.583 of 4.433were labeled correctly for the Local Case.

4 Methodology Proposed

Both the B-SISA system and the implementation of ADFR model work with information, data that is enriched by online transactions, data that needs to be purified, to incorporate a visualization personalized that allows analysis in order to help the institution to manage the decision-making process. In this context, we propose a methodology based on the architecture of Business Intelligence (i.e. transactional systems, management system technologies and management systems interfaces). Note that we used Pentaho as the framework for the BI solution.

To achieve this goal, we start from a point of origin that are all transactional systems, next we go integrating, debugging and visualizing. For this, the following steps are described, which are also shown in Figure 4:

•Transactions originated by the B-SISA system (fuel supplies of 1000400 vehicles in 638 service stations) depends on other applications dispersed in different systems (i.e. DB1, DB2, DB3).So, it is extracted, transformed and loaded in a large data warehouse to normalize and create the Operational Data Store of the B-SISA (ODS-B-SISA) through a tool called ETL (implemented in the Pentaho Data Integration specifically in the Spoon application of the Pentaho platform).

•The model that refers to the ADRF is implemented in order to compute the global and local anomalies scores corresponding to each of the B-SISA transaction records. This implementation requires a variety of statistical and machine learning tools available in the R environment. So, we integrate R with Pentaho Data Integration (PDI).The results are deposited in a database that will allow per-forming the following ETL.

•An ETL2 is built which after a dimensional modeling is responsible for extracting, loading and trans-forming available metadata to generate datamarts in order to analyze anomalies.

•Next, we design the Schema that refer to the data warehouses as large structured data responsible for constructing datamarts. In this step the Online Analytical Processing (OLAP) analyzing large quantities of data in real-time (deals with data in bulk) andemploying a technique called Multidi-mensional. We explain the details of this analyze in the experiments section.

•Finally, a customized dashboard is developed after of: (1) connect the Pentaho Server with database created, (2) create of some queries according with specific needs of the specialist (3) verifying these queries trough a plugin Sayku. The dashboard refers to the databases, PDI, OLAP created. We used the IvyDashboad Components as Plugin to create more user-friendly interfaces.

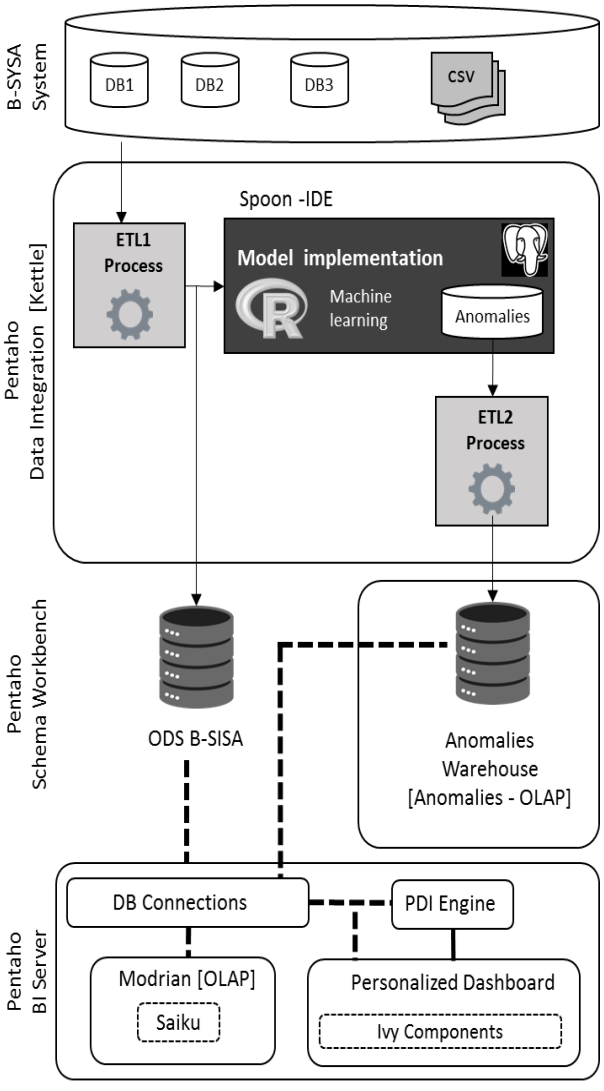


Figure 4 - Methodology Proposed.

http://dx.doi.org/10.25286/repa.v3i3.925

5 Experiments and Results

In a general way, this section describes different experiments during the implementation of the methodology proposed that will allow to have a support system of decision for the ANH's specialists in order to resolve the problems to analyze large amounts of anomalies scores and in this way to have control of fuel consumption in the Bolivian state.

To perform the tests of our approach, we describe steps implemented from de methodology in the next subsections.

5.1 Multidimensional Modeling

According with Ralph Kimballto in **[3]** and in order to perform the Multidimensional modeling, we described the formulation of the steps of the business to analyze (i.e our study case) through the flow chart showed in Figure 5. In it 4 steps are described. (A) The ADRF refer to the implementation of the model proposed by **[1]**, The supervisors of each location want to analyze which vehicles are extremely abnormal (locally and globally), at what gas stations or locations. (B)Level of detail define the granularity that have to able in the dimensional model,(C) Dimensions refer the answer to the following questions: Which vehicle has the most local and global anomalies scores? Which department has more anomalies scores? in which gas stations have more anomalies scores?In this context this refer to the time, vehicle type, locations and the vehicle type (these dimensions are re-quired by the ANH institution),(D) Indicators refer two principal measures: sum and maximum of anomaly scores (both of them considering the Local and Global scores).

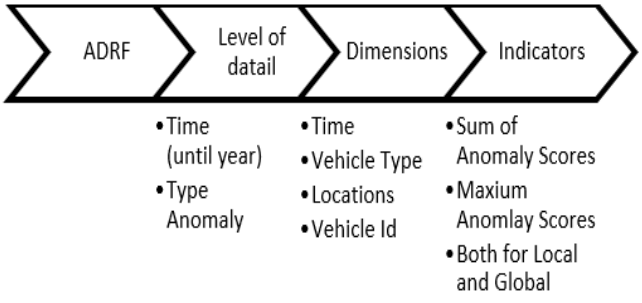


Figure 5 - Flow Chart of the process of business in the Anomalies context.

5.2 ETL 1

As the methodology refer, the ETL1 integrate three data sources from B-SISA system. They were extracted from the Vehicle Registration Application (i.e. " Motorized Kardex "),Sales Record Application (i.e. "Consumption Details") and Geospatial Information of Services Stations according to the ANH Raw Data. This tree data sources are integrated in order to implement the ADRF model.

The Figure 6 show the design and test of ETL1 trough the graphical tool Spoon (part of Pentaho Data Integration solution). In it, we show the different transformations and jobs contemplated by the model (i.e. De-termination of Profiling Groups, Formulations of knowledge and Local and Global Detection). Note that each icon in the Figure refer to an step from the ADRF and could represent as well as just an execution ofR script (e.g. A_DeterminationProfilingGroup) or a conjunct of transformations, called Jobs in the Spoon tool (e.g. ANH_DetectionStep_Job and ANH_TrainingStep_Job).

The R scripts refer the application of machine learning techniques (e.g. Unsupervised learning from SOM net-works and hierarchal clustering) in order to compute the scores of anomalies based in **[1]**.

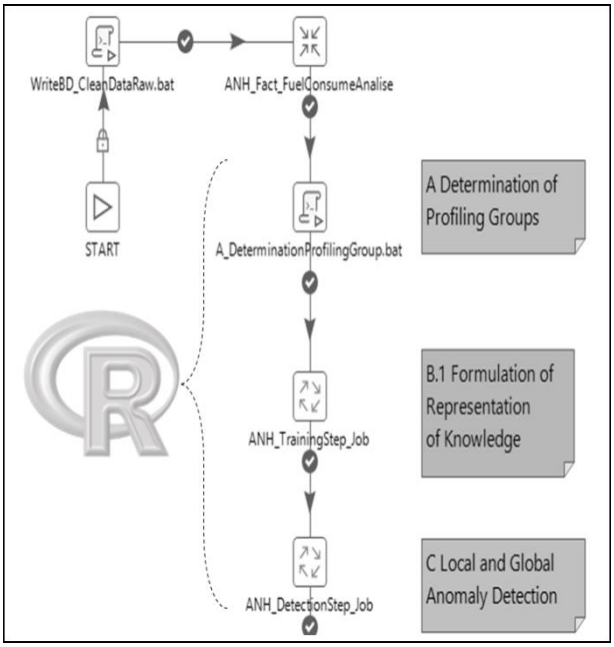


Figure 6 - ETL of ADFR model.

5.3 ETL 2

After a multidimensional modeling, the ETL2 describe refer to the extraction, transformation and loading of each dimension established according with the needs of our study case. The Figure 7 shows the design and test of ETL2. Each number refer to the metadata generated for the different dimensions of the cube.

5.4 Anomalies Data Warehouse

The Anomalies Data Warehouse is constructed after the ADFR implemented through R environment and its integration with Pentaho Data Integration (PDI). The data is stored over multiple dimension tables as the star schema is showed in the Figure 8. Note that these dimensions were created according with the needs of the ANH institution. They want to analyze the vehicles with more and maximum anomalies scores during a specific time, modifying the vehicle type, the location and also visualizing the current anomalies of each vehicle in a map referring the Geo Position of the service station.

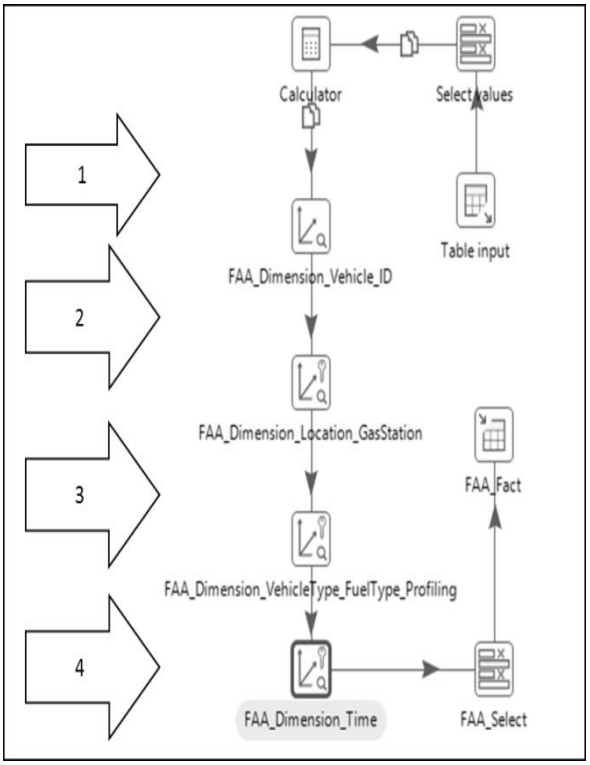


Figure 7- ETL of Anomalies Fact.

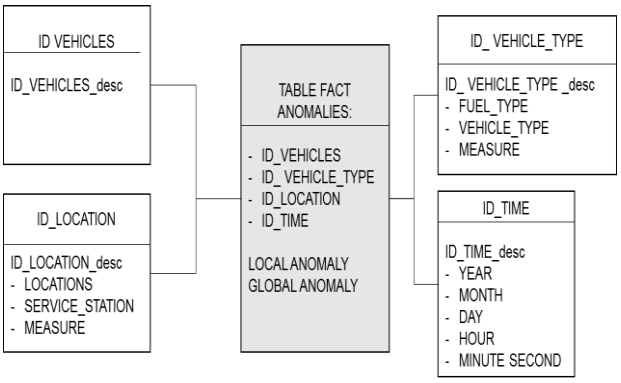


Figure 8 - Star Schema design for Anomalies Fact.

5.5 OLAP

The OLAP designed employs the Multidimensional modeling explained before. It consists of four dimensions (i.e. Time, Vehicle Type, Locations, Vehicle Id). For better understanding, Figure 9 shows and example of a cube with three of the four dimensions. Dimension of "Time" in the x-axis, dimension of "Vehicle Type" in the y-axis, dimension of "location" in the z-axis.

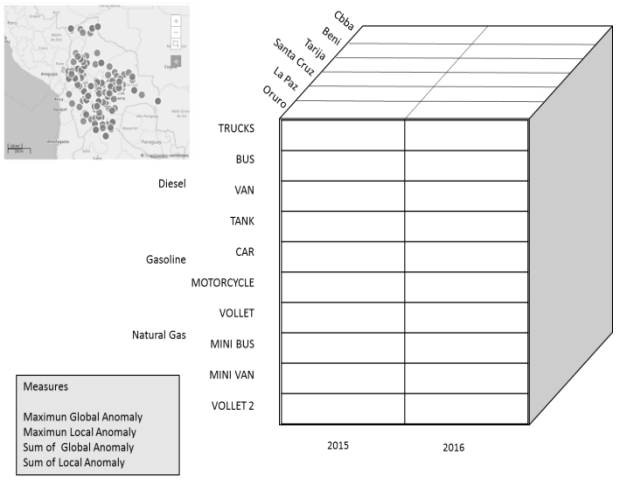


Figure 9 - Star Schema design for Anomalies Fact.

5.6 Dashboard Customized

Dashboards increase the analytical power of the visualization by allowing multiple perspectives on the dataset in the same location. For our approach we used the Community Dashboard Editor (CDE) and the plugin Ivy in order to have the graphical interface more friendly for the end users (i.e. specialist of the

http://dx.doi.org/10.25286/repa.v3i3.925

ANH institution). We create the layout, components and data source panels based in Charts, Data Tables, and Files created before using the Analysis or Report features.

When creating a dashboard, the Data Table content type allows a tabular representation of a database query in a dashboard. It also allows the manipulation of the data, directly from the dashboard.

In the Figure 10, the final dashboard to analyze anomalies in Boliviais showed.

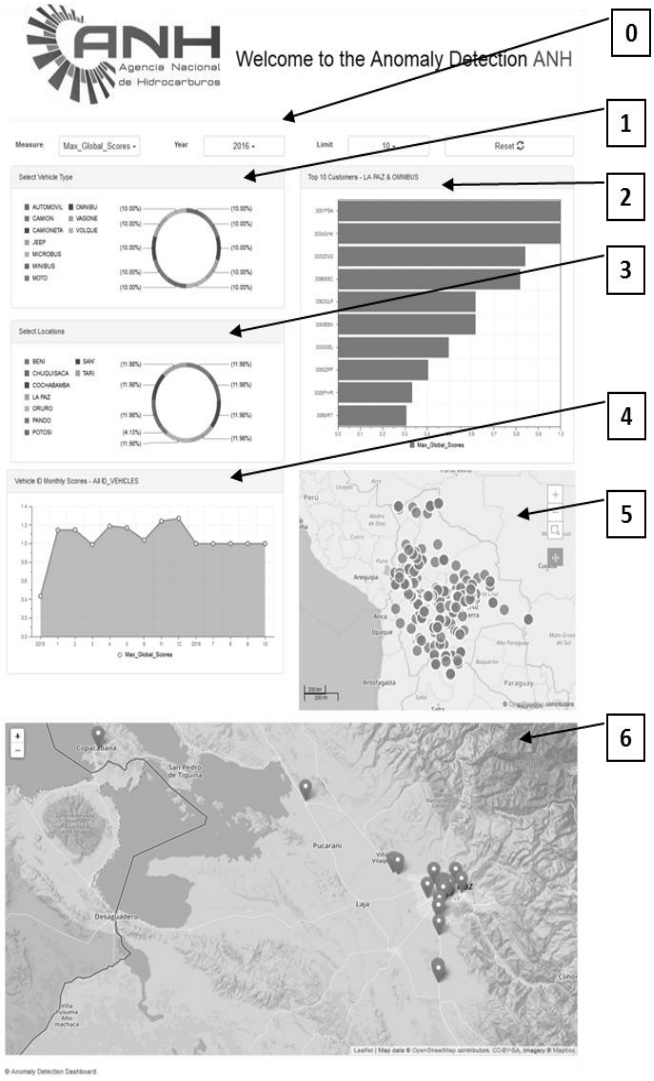


Figure 10 - Last Dashboard to analyze anomalies in Bolivia.

Next, seven macro components were designed, numbered from 0 to 6:

0- Initial Settings. At the first time, it is able for the user to give initial information: (a) corresponds to the measure (i.e. Sum of Anomaly Scores and Maximum Anomaly Scores for the Local and Global levels), (b) corresponds the year of analyze and (c) is the limit to visualize the top "n" vehicles with high measures

1 - Select Vehicle Type. The specialist can visualize the percent of amount anomalies (maximum or sum of them) per vehicle type.

2 - Select Locations. The specialist can visualize the percent of amount anomalies (maximum or sum of them) per location

3 - Top Consumers with high Anomalies. Also the user will visualize the top 10 vehicles respect to the vehicle type and the location selected.

4 - Vehicle Monthly Scores. Also the data is updated if the specialist chooses one of the top 10 vehicles in order to know the fuel consumption by month.

5 - Map of Bolivia geo referring to all Service Station. As reference the total service station this map show all of them.

6 - Map of Bolivia geo referring just the Service Station with the anomalies and also it is showed the record of real fueling by each vehicle with respect to those Service Station. This is the tool more significant by the specialist because he can know the regions of lot of anomalies in order to make strict control in the sales from some service stations. To have reference the real fueling of the anomaly cases permit to verify the anomalies scores.

With whatever change of visualization by the user, with just a click in whatever macro-component, all the data is updated.

5.7 Analyze of Anomalies from Dashboard

With the sampled data which was implemented this paper shows the following analysis:

- The Maximum departmental anomalies represent LA PAZ corresponding to the vehicles ID: 3001-FSA, 3034-GHK and other eight.
- The maximum anomalies per service station are the ones that are located in the border with Peru.

The maximum consumption per month is in the December.

6 Conclusions

It was possible to design and develop the consultations for the necessary dashboard and thus having a better structure of information in the anomaly detection context.

Also it was possible to analyze, design and construct the technologies from an BI solution (i.e. ETL, OLAP, Cubes, and customize and Dashboard), thus achieving weaknesses in the ANH.

In general, a BI architecture provided the concepts in order to implement a open source BI solution based in the compute of anomalies scores trough a machine learning techniques.

References

[1] AQUIZE, Vanessa Gironda; EMERY, Eduardo; DE LIMA NETO, Fernando Buarque. Self-organizing maps for anomaly detection in fuel consumption. Case study: Illegal fuel storage in Bolivia. In: IEEE LATIN AMERICAN CONFERENCE ON COMPUTATIONAL INTELLIGENCE LA-CCI, 4., 2017, Peru. **Proceedings**... Peru: IEEE, 2017. p.1-6. Available in: <https://ieeexplore.ieee.org/abstract/document/8285697/>

[2] STODDER, David. **Data visualization and discovery for better business decisions**. [E-book] TDWI Research,2013. p.30-31. Available in: <http://solutiondesignteam.com/wp-content/uploads/data-visualization-discovery-better-business-decisions-106672.pdf>

[3] TARNAVEANU, Diana; MUNTEAN, Mihaela. Free Business Intelligence – An Easy and Reliable Alternative. **Mathematical Models & Methods in Applied Sciences**, WSEAS Press, p. 158-164, 9 Set. 2012.
Available in: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2143945>

[4] MUTEAN, Mihaela; BRANDAS, Claudio. Business Intelligence Support Systems and Infrastructures. **Economy Informatics**, n. 7, p. 100-104, 2007.

[5] BUTUZA, Antoanela et al. Increasing the Business Performance using Business Intelligence**. Analele Universităţii Eftimie Murgu Reşiţa, Fascilua de Ingenerie**, v. 18., n.3 , p. 67-72, 2011.

[6] MIRCEA, Marinela et. Al. Agile Development for Service Oriented Business Intelligence Solutions. **Database Systems Journal**, v. 2., n.1, p. 43-56, 2011.

[7] PENTAHO. Big Data Integration and Analytics. Disponível em: <http://www.pentaho.com/>

[8] GOLFARELLI, Matteo. Open Source BI Platforms: a Functional and Architectural Comparison. In: INTERNATIONAL CONFERENCE ON DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, 11., 2009, Berlin. **Proceedings**... Berlim: Springer, 2009. p. 287 – 297.

http://dx.doi.org/10.25286/repa.v3i3.925

Solução IoT de Monitoramento de Poços para Gerenciamento de Recursos Hídricos

Estudo de caso para aplicado na região metropolitana do Recife

Victor Mendonça de Azevedo [1] orcid.org/0000-0003-2943-4622
Alexandre Magno Andrade Maciel [1] orcid.org/0000-0003-4348-9291
Kiev Santos da Gama [2] orcid.org/0000-0003-1508-6196

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Centro de Informática, Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil,

E-mail do autor principal: vma@ecomp.poli.br

Resumo

O objetivo principal das cidades inteligentes é alcançar o uso sustentável dos recursos. Para fazer o uso correto dos recursos, é necessário um monitoramento e gerenciamento precisos. Em alguns lugares, como os aquíferos subterrâneos, o acesso para monitoramento pode ser difícil, portanto, o uso de sensores pode ser uma boa solução. Os aquíferos subterrâneos são um importante recurso hídrico. O correto monitoramento, ajuda a prevenir seu esgotamento e a evitar impactos ambientais. Este artigo tem como objetivo apresentar um esforço que vem sendo realizado pela Agência Pernambucana de Águas e Clima – APAC, no monitoramento de poços subterrâneos, com o uso de uma solução IoT para um monitoramento continuo e em tempo real, com a análise dos dados obtidos e geração de histórico, possibilitando o estudo do comportamento dos aquíferos subterrâneos da Região Metropolitana de Recife. Apresentamos como essa solução é composta e como os dados gerados podem ser analisados.

Palavras-Chave: Aquíferos; Condutividade; IoT; Análise de dados; Poços;

Abstract

The main goal of smart cities is to achieve the sustainable use of resources. Accurate monitoring and management is required to make the correct use of resources. In some places, such as underground aquifers, access for monitoring can be difficult, so the use of sensors can be a good solution. Underground aquifers are an important water resource. Correct monitoring helps prevent depletion and avoid environmental impacts. This paper aims to present an effort that has been carried out by the Pernambuco Water and Climate Agency (APAC), in the monitoring of underground wells, with the use of an IoT solution for a continuous and real time monitoring, with the analysis of the obtained data and generation of history, it could enable the study of the behavior of subterranean aquifers of the Metropolitan Region of Recife. We present how this solution is composed and how the generated data can be analyzed.

Key-words: *Aquifers; Conductivity; IoT; Data Analysis; Wells*

1 Introdução

O surgimento de novas tendências sempre vem acompanhado de tecnologias de base existentes, que fornecem os meios para sua popularização. Este também é o caso da IoT (*Internet of Things*, ou Internet das Coisas) e sua relação com o que é chamado de *Big Data*. Enquanto a IoT conecta diversos dispositivos e sistemas por meio de sensores (relação máquina-máquina), tornando cada vez menor a necessidade de interação humana nestes processos, o *Big Data* toma conta da imensa quantidade de dados gerados.

Em uma pesquisa realizada pela Gartner em 2014, analistas previram que em 2020 a quantidade de dispositivos conectados passará dos atuais 5 bilhões para 25 bilhões **[1]**.

A IoT vem para facilitar ainda mais, tornando os dispositivos móveis conectados à internet verdadeiros chamarizes de informação e produtores de dados. Só que essa grande quantidade de dados deve ir para algum lugar para não se perder, e deve ainda ser "minerada" para ter valor agregado e ser passível de utilização. Neste cenário entra a *Big Data*.

O termo *Big Data* é dado ao armazenamento e processamento da mistura de dados estruturados (números) e não-estruturados (imagem, mídias sociais, *Twitter*, etc.). Trabalha com grande volume de dados em uma velocidade de aquisição que pode até ser em tempo real, ao mesmo tempo em que recebe diferentes tipos de entradas.

Através do uso de técnicas de análise estatística automatizada, ou seja, o *Data Mining*, as empresas estão descobrindo as tendências e padrões de comportamento que antes passava despercebidos. Uma vez descoberta essa inteligência vital, ela pode ser usada de forma preditiva para uma variedade de coisas.

O primeiro passo de para construir um programa de Mineração de Dados é a Coleta de Dados. A maioria das empresas já realizam essas tarefas de coleta de dados, contudo, nem todos esses dados coletados são valiosos, sendo necessária uma seleção.

A chave aqui é para localizar os dados críticos para o negócio, refiná-lo e prepará-lo para o processo de Mineração de Dados. Dados úteis não são mais chamados de dados, mas sim de informação. Dessa maneira, informações já possuem aspectos úteis e valor agregado.

A mineração de dados busca encontrar padrões (classificar e criar modelos de predição), segmentar informações (isolando variáveis) ou buscar correlações entre dados existentes. É aí que reside a grande vantagem da *Big Data*. Estes métodos de coleta, armazenamento e pós-processamento que são utilizados em conjunto com a IoT.

Os sistemas de sensores e monitoramentos estão fornecendo atualmente às empresas de serviços de água grandes quantidades de fluxos de dados em tempo real. No entanto, esse fluxo de dados ainda não é chamado de *Big Data*. *Big Data* abrange muito mais. Para exemplificar esse fato, considere os dados derivados de sensores instalados em bombas de água, que são vistos em monitores em qualquer momento, mas os dados desses sistemas de supervisão e aquisição de dados (SCADA) se concentram no que está acontecendo agora ou no que aconteceu no passado. Esta situação seria como olhar em um espelho retrovisor de um carro, não se torna muito eficaz se o caminho tiver muitas curvas a frente. Felizmente, esses sistemas SCADA poderiam potencialmente dizer muito mais, poderiam melhorar a eficiência, o uso de energia e as condições da rede que podem afetar seus requisitos de confiabilidade e manutenção, especialmente quando combinados com dados de outros locais da própria infraestrutura ou de fora do sistema de água. (Como dados meteorológicos).

Sistemas de análise de dados poderiam fazer exatamente isso. Às vezes, simplesmente armazenando e visualizando os dados de maneira amigável ao usuário, por exemplo, em um mapa, podem ser obtidas grandes informações. A análise de dados provenientes de ambientes de *Big Data*, poderia adicionar modelos preditivos para finalmente olhar para o futuro, possibilitando a manutenção no momento certo e local de forma a otimizar ainda mais o equilíbrio entre desempenho e confiabilidade.

Estas análises também podem evitar desastres causados pelo homem, como quedas súbitas na qualidade da água ou o surto de uma doença contagiosa **[2]**.

http://dx.doi.org/10.25286/repa.v3i3.937

Neste artigo, propomos algumas analises e ações que já podem ser tomadas, mesmo com poucos dados disponíveis, não nos limitando apenas a visualização dos dados gerados.

2 Gestão de Águas Subterrâneas

2.1 Importância da Gestão dos Poços

Um dos parâmetros mais importantes para monitorar nestas águas é a condutividade, pois os altos níveis de condutividade indicam a salinização das águas subterrâneas e a impossibilidade para consumo humano.

A contaminação das águas subterrâneas geralmente está ligada ao escoamento e infiltração de águas superficiais das áreas urbanas e agrícolas. O problema da salinização está relacionado às zonas costeiras onde o aquífero chega perto do mar. Às vezes, é muito difícil evitar misturar água doce com água do mar. Isso ocorre porque a água doce emerge em algumas áreas marinhas, mas se muita água é extraída do aquífero, então, a água do mar pode infiltrar-se para o aquífero. Se isso acontecer, a água subterrânea não será potável por um longo período de tempo. Além disso, as mudanças climáticas também influenciam o agravamento desse efeito.

O processo de urbanização também afeta o equilíbrio natural devido à impermeabilização do solo que reduz a infiltração local. De acordo com trabalhos publicados por Padowski et al. **[3]** dentre as 70 maiores cidades do mundo, 28 delas (40%) usam apenas fontes de água subterrânea. Destas 28 cidades, apenas seis deles usam um reservatório não ameaçado, 20 dependem de um reservatório ameaçado e dois em um vulnerável. No entanto, as simulações para 2.040 mostram que apenas três dessas cidades terão seus reservatórios ainda não ameaçados e cinco cidades terão reservatórios vulneráveis. Os aquíferos estão em perigo e esses aquíferos são a fonte de água de muitas pessoas.

Nas últimas décadas, a população está aumentando e o efeito da migração é que essa população se concentra nas áreas costeiras, de modo que as cidades das áreas costeiras sofreram um grande crescimento em poucos anos. Usando dados das cidades com mais de 1 milhão de habitantes **[4]**, foi calculado a porcentagem de cidades que podem ser consideradas cidades costeiras (50 km ou mais perto do mar). A distribuição das cidades com mais de 1 milhão de habitantes é mostrada na Figura 1 e essas cidades são representadas em diferentes cores, dependendo da proximidade com o mar. Pode-se ver pela Figura 1 que a maioria das grandes cidades estão localizadas próximas a costa.

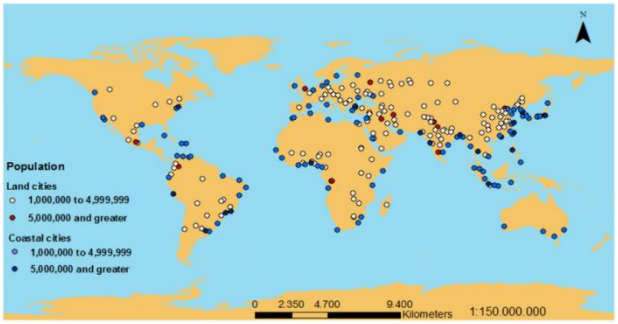


Figura 1 - Distribuição de cidades com mais de 1 milhão de habitantes **[4]**

As águas subterrâneas são o abastecimento de água para uma grande parte da população mundial. A maior parte da população mundial vive em zonas costeiras. Nessas áreas, a água subterrânea é suscetível a intrusão salina, o que pode tornar a água inalterável. O processo de urbanização, as mudanças climáticas e o bombeamento das águas subterrâneas podem alterar o equilíbrio do aquífero costeiro entre o sal e a água doce. Todos esses fatos levam a uma situação alarmante para a sustentabilidade das cidades e alguns pesquisadores já mencionaram e racionalizaram a necessidade do monitoramento das águas subterrâneas **[5,6]**, enquanto outros estão fazendo algumas medidas e simulações.

No entanto, estas simulações não são tão precisas devido ao alto número de variáveis que afetam cada aquífero. Cada cidade deve ser considerada como um caso específico que requer monitoramento e diagnóstico específicos **[6]**. Atualmente para monitorar os parâmetros das águas subterrâneas e, especificamente, a salinidade, os esforços são baseados na amostragem manual de poços com periodicidade temporal diferente. Esta metodologia acarreta um desperdício de energia, dinheiro e esforços

humanos em comparação com a possibilidade de usar dispositivos IoT. O uso destes dispositivos para monitoramento de aquíferos possibilita a coleta de dados em tempo real e é uma boa opção para o monitoramento de longo prazo.

2.1 Problema de Água Subterrânea em Recife

O grave problema no fornecimento de água em Pernambuco tem obrigado a população a recorrer à perfuração de poços. A Região Metropolitana do Recife possui mais de 5.000 poços reconhecidos atualmente, que precisam ser monitorados. Um exemplo clássico é a exploração de águas subterrâneas que aconteceu no bairro de Boa Viagem a partir da estiagem rigorosa entre 1997 e 1998. Na ocasião, houve um aumento de novas perfurações ocasionando o rebaixamento do nível piezométrico, que ocorre quando se retira mais do que o aquífero é capaz de renovar e suas recargas naturais se tornam insuficientes, e a salinização local constatada, que ocorre com poços mal perfurados no litoral, quando a água do mar encontra a água doce, por causa do desequilíbrio na pressão do subsolo, em algumas unidades de captação **[7]**.

Poços artesianos, desde que planejados e executados corretamente, podem resolver parte do problema, mas precisam de controle rigoroso do volume de água extraído através de hidrômetros. Desde então, os novos poços recebem uma outorga que lhes permite extrair um volume diário limitado de água do solo.

O problema, portanto, está na fiscalização dos poços, dos volumes extraídos e do nível de condutividade de cada um. As respostas das questões abaixo ajudam no monitoramento e na tomada de decisões importantes para esse controle e fiscalização:

- Quais os volumes de água extraídos dos poços?
- Quais os picos de consumo por período?
- Quais os níveis de condutividade elétrica dos reservatórios aquíferos por bairros da RMR?

As respostas para essas perguntas podem ser obtidas através da análise de dados. Com os dados fornecidos pelos hidrômetros instalados e equipamentos que possam medir o nível de condutividade é possível relacionar com as informações de horários e locais onde estão instalados e extrair informações importantes para o controle de consumo por poço ou por região e verificar se um aquífero está próximo do seu limite de consumo, possibilitando aos órgãos reguladores a tomada de ações com o objetivo de evitar o desgaste destes reservatórios e possíveis problemas ambientais.

3 Solução IoT para Gestão dos Recursos de Água Subterrânea

3.1 Trabalhos Relacionados

E*m 2007 foi lançado, no condado de* Rocky View, em Alberta, Canadá, um programa chamado *Rocky View Well Watch* **[8]**. Um dos principais objetivos deste projeto foi formar uma rede comunitária de monitoramento de águas subterrâneas usando informações voluntárias dos donos dos poços. Os dados utilizados no projeto eram do nível da agua e a coleta era feita manualmente e de forma voluntária pelos proprietários dos poços.

A fim de fazer previsões sobre como as mudanças nas práticas de uso da terra e clima afetaram o balanço hídrico no condado de *Rocky View*, os pesquisadores utilizaram os dados obtidos para criar modelos numéricos para simular possíveis resultados. Variáveis que afetam a elevação do nível da água foram utilizadas nestes modelos como, clima, recarga de água subterrânea, geologia, e bombeamento de águas subterrâneas. No site do projeto é possível verificar um mapa com os poços e os níveis de agua de cada um deles.

Muitas pesquisas ainda estão sendo feitas na área do uso de IoT para gestão de águas. Alguns sistemas que já fazem uso dessa tecnologia têm provado que uma redução da intervenção humana pode alcançar um aumento significativo de eficiência.

Monitoramento e gerenciamento contínuo do meio ambiente através da integração do IoT e sistemas que utilizem sensores é um tópico ativo para pesquisadores, engenheiros e até

http://dx.doi.org/10.25286/repa.v3i3.937

administradores públicos. Como trabalhos notáveis nessa área de sistemas de gestão de água, podemos citar o *Sustainable Water Distribution Strategy with Smart Water Grid* **[9]**, que explica como preencher a lacuna entre conectar vários recursos hídricos e otimizar a gestão do sistema com novas soluções de tecnologia de informação. Água da chuva, recursos externos, e outros tipos alternativos de recursos hídricos são integrados nesse sistema para fornecer um ambiente sustentável.

O artigo intitulado *Internet of Things for Smart cities* **[10]** fornece uma visão sobre a realização de uma rede IoT junto com a rede e os serviços de *back-end* necessários. O artigo também descreve os protocolos utilizados nesta arquitetura.

O artigo Micro *Water distribution resources* **[11]** descreve sobre a construção sustentável de um sistema de distribuição de água em vilas na Índia, considerando fatores locais como trabalho, necessidades da comunidade, clima e tempo de implementação. O artigo dá uma ideia sobre os vários desafios que tiveram que enfrentar ao construir esse sistema na Índia.

3.2 Projeto de Sustentabilidade Hídrica do Governo do Estado de Pernambuco

Atualmente o controle do volume de consumo por poço é feito manualmente, onde cada proprietário do poço anota a informação de consumo e envia mensalmente a Agência Pernambucana de Águas e Clima – APAC. Esse fluxo de informação demanda bastante tempo para ser concluído, o que acarreta em uma demora na análise destes dados e de um completo mapeamento dos índices de consumo de toda a região, impactando ainda na tomada de possíveis ações de controle. As informações sobre os índices de condutividade são inexistentes.

Pensando na situação atual e em uma possível solução, o Governo do Estado desenvolveu um projeto chamado de Projeto de Sustentabilidade Hídrica de Pernambuco – PSHPE, que visa a instalação e monitoramento, através da Agência Pernambucana de Águas e Clima – APAC, de 356 poços artesianos na Região Metropolitana do Recife.

Esta solução para gestão, otimização, monitoramento e controle dos sistemas de água subterrânea, consiste na implantação de uma rede de monitoramento dos poços através de sistema de telemetria (IoT), em poços da Região Metropolitana do Recife. Nesta rede, serão instalados hidrômetros, módulos de comunicação remota, que permitem a comunicação com um sistema informatizado, onde se torna possível conferir as leituras (vazões) dos hidrômetros, e sondas que permitem o monitoramento do nível e da salinidade da água destes poços. Esta rede de monitoramento requer uma continuidade na coleta de dados, tendo como objetivo a obtenção de um monitoramento adequado para os aquíferos.

Este projeto prevê a instalação dos seguintes itens, de acordo a Figura 2:

(a) Módulos de comunicação remota: Equipamento responsável pela promoção da conexão lógica dos parâmetros monitorados nos poços com o servidor de banco de dados remoto.

(b) Hidrômetros de telemetria: Equipamento responsável pelo registro do consumo de agua com conexão do tipo Reed switch.

(c) Sistema de monitoramento: Software para análise e controle do sistema de monitoramento instalado.

(d) Sondas para monitoramento de nível d'água e de salinidade da água: Os Sensores de medição de nível d'água são do tipo cerâmico-capacitivo para envio em tempo real de informações sobre o nível de coluna de água acima da posição em que o sensor estiver instalado no poço, suportam operação a uma profundidade de até 100 m de coluna de água e possui acessórios que permitam sua instalação a até 200m do nível do solo, operam na faixa de 1 a 1.000 µS (micro-siemens).

Figura 2 - Equipamentos que compõem a solução IoT

4 Análises Propostas com os Dados Obtidos

4.1 Análise Preliminar

A solução proposta pelo Projeto de Sustentabilidade Hídrica de Pernambuco – PSHPE otimiza e automatiza o envio das informações para o órgão regulador, APAC. Com os dados em tempo real, é possível visualizar o que acontece e o que pode ser feito.

Com a aplicação de técnicas de análise de dados, podemos gerar gráficos de consumo, verificar os níveis de condutividade e identificar quais poços e regiões já possuem um esgotamento no aquífero, e com o passar do tempo, de posse de dados históricos gerados por este sistema, prever quais aquíferos serão impactados, quais regiões vão ter aumento de consumo e desta forma, tomar ações preditivas.

De posse dos dados provenientes de 20 módulos de comunicação, com hidrômetro de telemetria e sonda de condutividade, podemos iniciar algumas análises.

Os módulos estão instalados pela região metropolitana do Recife, de acordo a Figura 3, e enviam informações de consumo, condutividade e temperatura dos poços a cada cinco minutos.

Os dados obtidos correspondem ao período de 15 de julho de 2017 a 6 de dezembro de 2017.

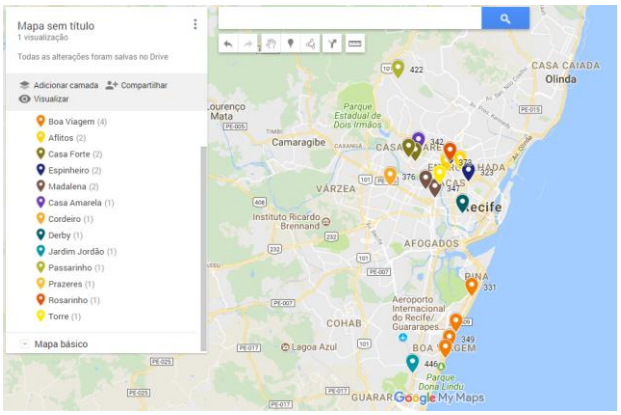


Figura 3 - Módulos instalados pela Região Metropolitana de Recife

Como no início nem todos os poços analisados estavam conectados ao sistema, então, ao longo do período fomos obtendo mais informações de mais poços.

A Figura 4, mostra um exemplo de como são as colunas dos dados como foram obtidos:

	A	B	C	D	E	F	G	H	I
1	cod	cod_modulo	volume	dh_leitura	contad	icm	nivel_poco	temperatura	analog1
2	7541776	323	0.000	04/07/2017 17:45	0	N;N	N;N	N;N	N;4.183"
3	7566757	342	0.000	12/07/2017 12:59	0	N;N	N;N	N;N	N;4.146"
4	7566771	342	0.000	12/07/2017 12:54	1	N;N	N;N	N;N	N;4.151"
5	7566785	342	0.000	12/07/2017 12:59	2	N;N	N;N	N;N	N;4.151"
6	7566798	342	0.000	12/07/2017 13:04	3	N;N	N;N	N;N	N;4.151"
7	7566814	342	0.000	12/07/2017 13:09	4	N;N	N;N	N;N	N;4.150"
8	7566827	342	0.000	12/07/2017 13:14	5	N;N	N;N	N;N	N;4.151"
9	7566841	342	0.000	12/07/2017 13:19	6	N;N	N;N	N;N	N;4.150"
10	7566855	342	0.000	12/07/2017 13:24	7	N;N	N;N	N;N	N;4.150"
11	7566869	342	0.000	12/07/2017 13:29	8	N;N	N;N	N;N	N;4.149"
12	7566883	342	0.000	12/07/2017 13:34	9	N;N	N;N	N;N	N;4.151"
13	7566897	342	0.000	12/07/2017 13:39	10	N;N	N;N	N;N	N;4.150"

Figura 4 - Exemplo da coluna da tabela dos dados obtidos

4.2 Pré-Processamento

Os dados estavam no formato bruto e foram necessários procedimentos de limpeza e organização dos dados.

Foi incluída uma coluna período, cujos valores correspondem aos turnos Madrugada, Manhã, Tarde e Noite, onde:

http://dx.doi.org/10.25286/repa.v3i3.937

Horário das 5:00 as 11:59, corresponde ao turno Manhã;

Horário das 12:00 as 17:59, corresponde ao turno Tarde;

Horário da 18:00 as 23:59, corresponde ao turno Noite;

Horário das 00:00 as 4:59, corresponde ao turno Madrugada;

Foi incluída também uma coluna com a informação do Bairro onde o modulo está instalado.

Com os dados limpos e organizados, foi possível a geração de gráficos, de acordo a Figura 5, com o intuito de obter algumas informações preliminares.

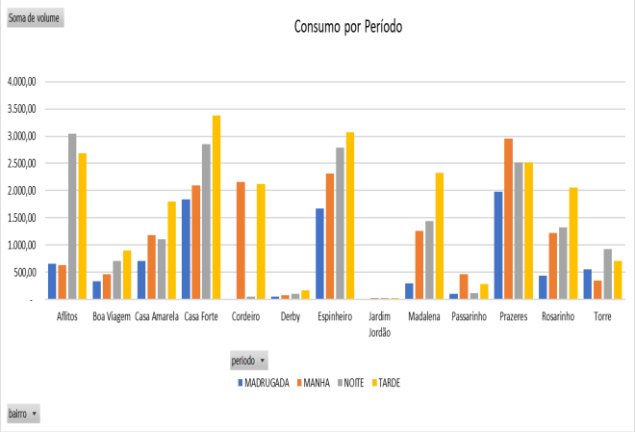


Figura 5 - Gráfico de Consumo por Período

Tabela 1 - Consumo por Bairro e Período

Volume por Período	Rótulos de Colun				
Rótulos de Linha	MADRUGADA	MANHA	NOITE	TARDE	Total Geral
Aflitos	657,30	635,10	3.039,50	2.688,00	7.019,90
Boa Viagem	333,20	462,40	705,80	903,90	2.405,30
Casa Amarela	714,40	1.189,70	1.100,10	1.794,50	4.798,70
Casa Forte	1.843,30	2.100,80	2.851,30	3.378,60	10.174,00
Cordeiro		2.161,40	52,60	2.118,60	4.332,60
Derby	53,30	76,00	106,80	173,70	409,80
Espinheiro	1.675,60	2.313,10	2.782,80	3.070,00	9.841,50
Jardim Jordão	15,70	28,30	29,90	23,30	97,20
Madalena	297,70	1.255,70	1.435,10	2.325,10	5.313,60
Passarinho	101,30	459,20	117,40	290,10	968,00
Prazeres	1.981,10	2.950,10	2.525,30	2.515,40	9.971,90
Rosarinho	438,40	1.218,80	1.320,30	2.061,80	5.039,30
Torre	557,90	346,50	928,30	711,00	2.543,70
Total Geral	8.669,20	15.197,10	16.995,20	22.054,00	62.915,50

Pela informação do gráfico, na Figura 5, e da Tabela 1, podemos concluir que o período de maior consumo de água dos poços artesianos, em metros cúbicos, é no período da tarde, seguido do período da noite e manha, sendo o menor consumo no período da madrugada.

Pode-se constatar também que o bairro que tem o maior consumo pelo período da manhã é Prazeres, pelo período da tarde é Espinheiro e pelo período da noite é Aflitos.

4.3 Análise de Condutividade

A condutividade elétrica da água representa a facilidade ou dificuldade de passagem da eletricidade na água. A Tabela 2 apresenta a condutividade por bairro e por módulo.

Tabela 2 - Condutividade Por Bairro e Módulo.

Bairros	Média de icm	DesvPad de icm
Aflitos		
373	0,37	0,02
386	0,12	0,00
Boa Viagem		
331	1.739,59	613,39
349	0,29	0,06
388	6.517,29	2.297,34
395	0,25	0,01
Casa Amarela		
342	0,13	0,01
Casa Forte		
346	0,40	0,03
378	0,61	0,00
Cordeiro		
376	0,75	0,01
Derby		
387	2.486,47	1.275,72
Espinheiro		
323	0,35	0,01
377	0,19	0,02
Jardim Jordão		
446	0,14	0,01
Madalena		
347	0,64	0,02
365	164,73	382,67
Passarinho		
422	0,07	0,00
Prazeres		
366	1.246,33	486,19
Rosarinho		
411	0,17	0,01
Torre		
394	0,88	0,02
Total Geral	**253,63**	**658,57**

De acordo a Tabela 2, é possível verificar que alguns módulos instalados no Bairro de Boa Viagem, possuem altos índices de condutividade,

que indicam que a água é salina. Em alguns pontos como Prazeres e Derby, a água está como a de um Rio marginal e salobra. Para o caso estudado, os bairros que obtiveram altas médias de índices de condutividade, são as que possuem altos valores de desvio padrão, o que indica que a condutividade nesses locais varia bastante, podendo estar relacionadas a fatores como aumento no consumo, ou variação nos índices de chuva no local. Nas Figuras 6, 7 e 8, são listadas as localizações em que estes módulos.

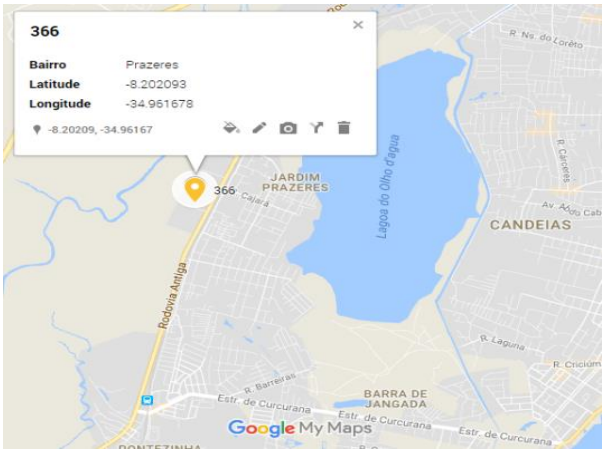


Figura 6 - Módulo instalado no bairro de Prazeres, próximo a lagoa do Olho d'agua

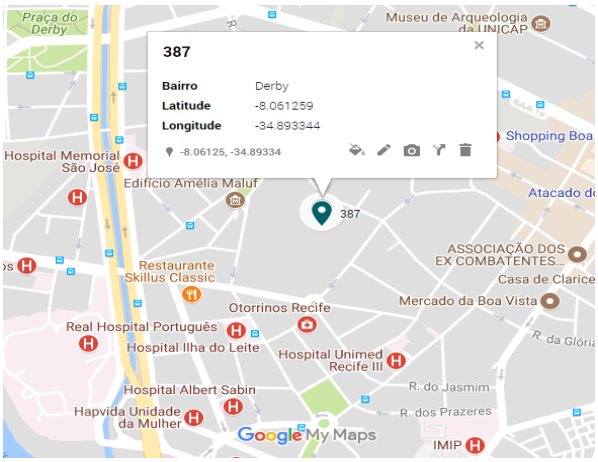


Figura 7 - Módulo instalado no bairro do Derby, próximo ao canal

Figura 8 - Módulos instalados no bairro de Boa Viagem, próximos ao mar.

Podemos constatar que os módulos estão em poços localizados próximo a um lago, ao canal, e próximo a beira mar, o que pode ser um indicio do alto valor da condutividade encontrado nos poços destes locais.

Com essas análises preliminares podemos destacar os seguintes pontos:

Na Região Metropolitana de Recife, o consumo desenfreado que aconteceu, a partir da estiagem rigorosa que aconteceu entre 1997 e 1998, no bairro de Boa Viagem, ocasionou um rebaixamento do nível piezométrico, ocasionando um aumento na salinidade da água e consequentemente, aumento da condutividade. Fato este que podemos comprovar através das análises preliminares obtidas, onde verificamos que em dois módulos instalados no bairro de Boa Viagem, os índices de condutividades mostram que a água é salina.

4.4 Correlação entre Índice de Chuva, Maré e Condutividade

É possível encontrar alguns trabalhos que tentam correlacionar a influência dos índices de chuvas com o nível dos poços **[12]**. Como ainda não possuímos dados concretos sobre os níveis de poços da região, vamos verificar se existe alguma correlação entre os índices de maré e chuva com a condutividade dos poços. Para isso, foi obtido no site da APAC **[13]** os dados históricos de chuva de todo o estado de Pernambuco. No site Tábua

http://dx.doi.org/10.25286/repa.v3i3.937

de Marés **[14]** também foi possível obter o nível das marés do estado.

De posse desses dados, selecionados o período que compreende entre os meses de setembro a novembro de 2017 e os poços que possuem dados completos desse período. Realizamos a limpeza dos dados e tentamos correlacionar com os dados dos poços disponíveis. Distribuímos na Tabela 4, os dados de condutividade, índices de chuva e maré, separados por período e por poço.

Tabela 4 - Valores de condutividade, nível mare e índice de chuvas por poço, dia e período.

Rótulos de Linha	Média de icm	Média de nivel_mare	Média de indice_chuvas
323	0,35	1,32	0,78
set	0,35	1,31	1,61
01/set			
MADRUGADA	0,35	0,50	-
MANHA	0,35	2,10	-
NOITE	0,35	1,90	-
TARDE	0,35	0,60	-
02/set			
MADRUGADA	0,35	0,60	-

Como estamos tentando correlacionar duas variáveis de escala métrica, o coeficiente de correlação de Pearson, ou r de Pearson, é o mais indicado. O r de Pearson pode ser obtido através da fórmula (1) abaixo:

$$r = \frac{\sum (xi - \underline{x})(yi - \underline{y})}{\sqrt{(\sum (xi - \underline{x})2)(\sum (yi - \underline{y})2)}} \quad (1)$$

Onde x_i e y_i são os valores das variáveis que vamos correlacionar, $\underline{x}$ e $\underline{y}$ são respectivamente as médias dos valores de x_i e y_i.

De posse da Tabela 4, aplicamos a formula de correlação r de Pearson para verificar a correlação existente entre esses dados, para cada poço, como mostrado na Tabela 5.

Tabela 5 - Correlação entre condutividade, nível mare e índice de chuvas por poço.

Poço	icm	nivel_mare	indice_chuvas	Correlação icm x nivel_mare	Correlação icm x indice_chuvas
323	0,35	1,32	0,78	0,05	- 0,16
342	0,13	1,37	0,82	- 0,02	0,07
346	0,39	1,38	0,86	0,05	0,23
349	0,29	1,24	1,19	0,00	0,04
365	172,35	1,32	0,89	- 0,11	- 0,11
366	1.218,00	1,29	0,94	- 0,03	0,14
373	0,37	1,23	1,15	0,05	0,12
376	0,75	1,35	0,88	- 0,01	- 0,32
378	0,61	1,33	0,94	0,02	- 0,20
386	0,12	1,33	0,93	- 0,05	- 0,04

Pela Tabela 5, os índices de correlação, para todos os poços, estão próximos de zero, o que indica que não há relação entre as variáveis de condutividade, nível de maré e índice de chuvas. Geramos também o gráfico de dispersão para cada poço.

Nas Figuras 9 e 10 é mostrado apenas, como exemplo, o gráfico de dispersão correspondente ao poço 323.

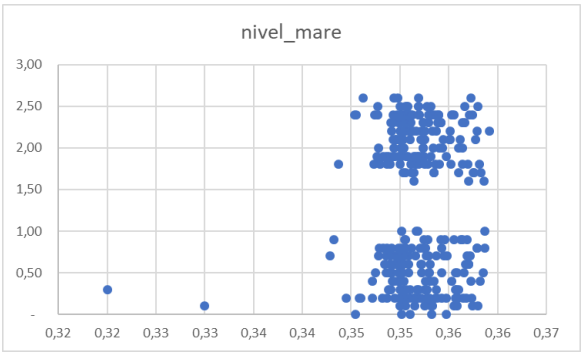


Figura 9 - Gráfico de dispersão do nível mare com a condutividade do poço 323.

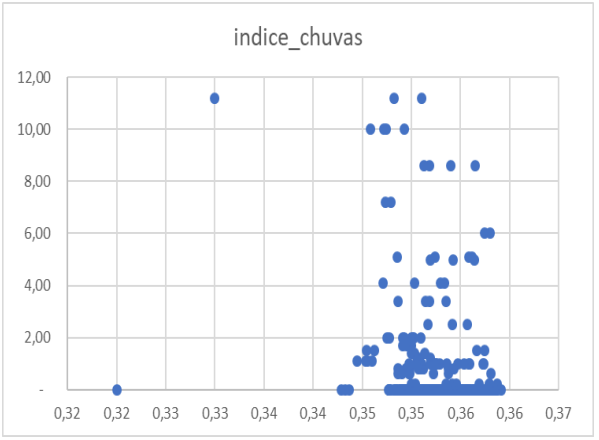


Figura 10 - Gráfico de dispersão do índice de chuvas com a condutividade do poço 323

Pelas Figuras 9 e 10, é possível observar que não existe uma relação entre as variáveis condutividade, índice de chuva e nível maré. Os demais poços possuem gráficos semelhantes.

5. Análises e Discussões

Importante destacar que para se obter um melhor cenário, mais dados devem ser obtidos.

Estamos apenas avaliando um período de amostragem de 3 meses e de apenas 20 poços que já possuem o sistema de telemetria instalado.

Com relação as análises de índices de consumo por mês, mais dados são necessários, assim como um maior período de amostragem para que possamos elaborar uma análise mais detalhada.

Com as análises preliminares podemos destacar alguns pontos:

No bairro de Boa Viagem, na Região Metropolitana de Recife, é possível comprovar através das análises preliminares obtidas, que em dois módulos instalados os índices de condutividades mostram que a água é salina, fato este que pode constatar um rebaixamento no nível piezométrico do aquífero na região.

O consumo por período nos permite gerar um perfil de consumo dos poços. Com a análise feita, constatamos que o período da tarde apresenta o maior consumo, fato este que pode estar relacionado ao período de retorno do trabalho e/ou escola.

6. Conclusão

Apesar de abundantes, as águas subterrâneas não são inesgotáveis, e o seu uso sem uma gestão consciente pode trazer problemas a curto e a médio prazo.

A extração de água subterrânea de forma exagerada, que ultrapassa os limites de produção das reservas reguladoras ou ativas do aquífero, ocasiona um processo de rebaixamento do nível do aquífero e irá provocar danos ao meio ambiente ou para o próprio recurso. Portanto, a água subterrânea pode ser retirada de forma permanente e em volumes constantes, por muitos anos, desde que esteja condicionada a estudos prévios do volume armazenado no subsolo e das condições climáticas e geológicas de reposição**[15]**.

Para a correta gestão do uso de águas subterrâneas é necessário um monitoramento continuo dos parâmetros de consumo e condutividade. Os índices de condutividade são um parâmetro crucial para determinar se um aquífero está sendo prejudicado. A dificuldade de um monitoramento continuo pode ser resolvida com o uso de soluções IoT que possa proporcionar, em tempo real, a informação desses parâmetros **[16]**.

Destacamos neste artigo a importância das águas subterrâneas e de sua gestão, assim como a implantação de uma solução do Governo do Estado para controle no uso dessas águas.

A Solução IoT, ainda em fase de implantação pelo Governo do Estado, se mostra eficaz no envio de informações em tempo real do que acontece em cada poço, porém não possibilita uma análise completa de todo o cenário.

Propomos neste artigo, uma análise dos dados gerados, em conjunto, de todos os poços, com o uso dos dados provenientes da Solução IoT. Geramos gráficos de análise do consumo por bairros e por período, o que nos permite visualizar qual o período do dia que o consumo é maior. Geramos também uma análise dos níveis de condutividade da Região.

Com mais dados disponíveis, melhores análises poderão ser efetuadas, e com o uso de dados históricos e aplicação de técnicas de aprendizagem de máquina, com a seleção adequada de atributos, como foi efetuado neste artigo para as analises, podemos gerar previsões sobre quais bairros terão um aumento de consumo e de índice de condutividade, possibilitando a tomada de ações preditivas afim de evitar o esgotamento dos aquíferos e impactos ambientais significativos.

6.1 Trabalhos Futuros

O monitoramento em tempo real pode fornecer *insights* ao vivo sobre o estado corrente da rede de monitoramento de águas subterrâneas. Com a solução IoT implantada, os dados de consumo, nível e condutividade da água podem ser obtidos em tempo real. Isso possibilita detectar estados críticos e acionar uma ação de resposta. No sistema IoT do projeto de sustentabilidade hídrica, cada poço possui sensores de temperatura, nível, condutividade e consumo. Os dados são enviados a cada 5 minutos. Quando o projeto estiver finalizado, teremos dados de 100 poços, enviando 4 medidas a cada 5 minutos, num total de 4.800 medições por hora.

http://dx.doi.org/10.25286/repa.v3i3.937

A análise dos dados de consumo de água, nível, condutividade e temperatura é caracterizada por atualizações incrementais ao longo do tempo. Para essa configuração, o processamento de fluxo e a tecnologia CEP podem ser uma solução. Para deduzir o estado atual de um sistema, é necessário apenas cálculos sobre as últimas informações do sensor. Portanto, o estado que é necessário para processamento é relativamente pequeno. O processamento CEP, mantêm o estado na memória e, assim, permitem um alto rendimento. O Uso do mecanismo CEP de código aberto poderia executar as análises necessárias para os 100 poços em paralelo em uma única máquina. O desempenho vai depender dos detalhes de implementação e das análises solicitadas.

7. Referências

[1] STAMFORD. Gartner Says 6.4 Billion Connected "Things" Will Be in Use in 2016, Up 30 Percent From 2015. **Newsroom** ,10 Nov. 2015. Disponível em: <https://www.gartner.com/newsroom/id/3165317>. Acesso em: 2 mar. 2018.

[2] CHEUNG, Cy; NUIJTEN, Martijn. Big Data and the Future of Water Management. **Rijksdienst voor Ondernemend Nederland**. Disponível em: <https://www.rvo.nl/sites/default/files/2014/05/Big%20Data%20and%20the%20Future%20of%20Water%20Management.pdf>. Acesso em: 18 out. 2017.

[3] PADOWSKI, Julie C.; GORELICK, Steven M. Global analysis of urban surface water supply vulnerability. **Environmental Research Letters**, v. 9, n. 10, p. 104004, 2014.

[4] Intl ESRI Data. Disponível em: <http://www.baruch.cuny.edu/geoportal/data/esri/esri_intl.htm#world> Acesso em: 16 out. 2017.

[5] HAYASHI, T. et al. Effects of human activities and urbanization on groundwater environments: An example from the aquifer system of Tokyo and the surrounding area. **Science of the total environment**, v. 407, n. 9, p. 3165-3172, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18823643> Acesso em: 16 out. 2017.

[6] FOSTER, S. S. D. The interdependence of groundwater and urbanisation in rapidly developing cities. **Urban water**, v. 3, n. 3, p. 185-192, 2001.

[7] SANTOS, Ivaneide de Oliveira et al. Aqüífero Boa Viagem, uma Discussão de seus Usos versus suas Potencialidades, Recife–PE. **Revista Brasileira de Geografia Física**, v. 4, n. 4, p. 848-856, 2012.

[8] UNIVERSITY OF CALGARY. Research Connection. **Groundwater Connections connect, connecting research, community & education**, 2013. Disponível em: <http://groundwaterconnections.weebly.com/research-connections.html> Acesso em: 17 out. 2017.

[9] BYEON, Seongjoon et al. Sustainable water distribution strategy with smart water grid. **Sustainability**, v. 7, n. 4, p. 4240-4259, 2015.Disponível em: <http://www.mdpi.com/2071-1050/7/4/4240/htm> Acesso em: 2 fev. 2018.

[10] ZANELLA, A et al. Internet of things for smart cities. **IEE Internet of Things Journal**, v. 1, n. 1, p.22-32, Feb. 2014. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6740844> Acesso em: 2 fev. 2018.

[11] RAMESH, Maneesha Vinodini et al. Micro water distribution networks: A participatory method of sustainable water distribution in rural communities. In: GLOBAL HUMANITARIAN TECHNOLOGY CONFERCE (GHTC), 2016, Seattle. **Proceedings.**.. Seattle: IEEE, 2016. p. 797-804.

[12] SILVA, Iara Lina de S.; MOTA, Elison José; OLIVEIRA, Leidiane Leão de. Relação da água da chuva com os poços de abastecimento público do Urumari em Santarém Pará, Brasil. In: CONGRESSO BRASILEIRO DE ÁGUAS SUBTERRÂNEAS, 13., 2014, Belo Horizonte. **Proceedings**... Belo Horizonte: ABAS, 2014..

[13] Turismo no Recife. Disponível em: <http://www.turismonorecife.com.br/pt-br/informacoes-importantes/tabua-de-mares> Acesso em: 15 mar. 2018.

[14] Agência Pernambucana de Águas e Climas – APAC. Disponível em: <http://www.apac.pe.gov.br/meteorologia/> Acesso em: 15 mar. 2018.

[15] Associação Brasileira de Águas Subterrâneas, ABAS. Disponível em: <http://www.abas.org/educacao.php> Acesso em 2 mar. 2018.

[16] PERLMAN, Howard. Electrical Conductivity and Water. **The USGS Water Science School**, 2014. Disponível em: <http://water.usgs.gov/edu/electrical-conductivity.html> Acesso em: 1 mar. 2018.

http://dx.doi.org/10.25286/repa.v3i3.937

Utilização de Dataflow para previsão de aceitação de respostas no fórum StackOverflow.com

Using Dataflow to predict answers' acceptance in the StackOverflow.com forum

Talita Albuquerque de Araújo[1] orcid.org/0000-0003-4002-8238
Jairson Barbosa Rodrigues[2] orcid.org/0000-0003-1176-3903

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, PE, Brasil
[2] Colegiado de Engenharia da Computação, Universidade do Vale do São Francisco, Juazeiro, BA, Brasil

E-mail do autor principal: taa2@ecomp.poli.br

Resumo

Nos últimos anos processar dados em larga escala tem sido um grande desafio, sendo, para isso, necessária a utilização de sistemas de alto desempenho para esse processamento. Este trabalho tem como objetivo apresentar um framework que permita que seja desempenhada essa função de forma rápida e simples, tirando proveito da estrutura do *DataFlow* para processamento de Big Data. A análise realizada é do tipo preditiva, em uma base disponibilizada on-line. A partir dela, será mostrado o uso do framework e se procurará verificar se o modelo gerado teve sucesso ou não. Os indicadores usados para essa comprovação serão a acurácia, a curva ROC, a especificidade e a sensitividade. Como resultado, espera-se extrair conhecimento sobre a aplicação do framework *DataFlow* para análise de grandes quantidades de dados e mostrar algumas vantagens no seu uso prático.

Palavras-chave: Big Data; Aprendizado de Máquina; *Dataflow*; Apache Beam;

Abstract

In recent years large-scale data processing has been a major challenge, requiring the use of high-performance systems for this processing. This work aims to present a framework that allows this function to be performed quickly and easily, taking advantage of the DataFlow structure for Big Data processing. The analysis performed is of the predictive type, in a database made available online. From this, it will be shown the use of the framework and will try to verify if the model generated was successful or not. The indicators used for this verification will be the accuracy, the ROC curve, the specificity and the sensitivity. As a result, it is expected to extract knowledge about the application of the DataFlow framework for analyzing large amounts of data and to show some advantages in its practical use.

***Key-words**: Big Data; machine learning; Dataflow; Apache Beam;*

 http://dx.doi.org/10.25286/repa.v3i3.942

1 Introdução

O presente artigo realiza, como estudo de caso, uma análise preditiva na base de dados do site *StackOverflow* **[1]**— disponibilizada na *BigQuery*, um *Data Warehouse* do *Google Cloud Plataform* **[2]** —, empregando o *framework Dataflow*, usado para processamento de *Big Data*.

É esperado que mais dados sejam gerados nos próximos 5 anos do que nos últimos 5.000 anos **[3]**. Tal previsão reforça a necessidade de ferramentas de análise de dados em escala *Big Data*. O *framework Hadoop* apresentou-se, então, como uma evolução natural nesse cenário; todavia, há casos como os de análises em tempo real e análises com Aprendizagem de Máquina nos quais o *Hadoop* não é plenamente satisfatório **[4]**.

E como isso apresenta-se o *framework Dataflow* que pode ser apresentado como um modelo de programação unificado e um serviço gerenciado que desenvolve e executa uma grande variedade de padrões de processamento de dados, essa análise será realizada sobre os dados do fórum *StackOverflow*, no qual desenvolvedores postam perguntas referentes a tecnologia em busca de ajuda ou explicações para as suas dúvidas. Por ser um fórum, existe muita interação entre os usuários, através de *posts* com respostas, comentários ou até novas dúvidas. O usuário que postou a pergunta inicial deve definir uma das respostas ou comentários como aquele (a) que resolveu o seu problema, e essa interação é marcada como a resposta válida à pergunta realizada no fórum.

Na seção 4 do presente artigo, serão expostos alguns trabalhos nos quais houve uso do *framework* Dataflow. Já os conceitos importantes para a compreensão do tema serão esclarecidos na seção 5, que corresponde ao Referencial Teórico. Daí, parte-se para a explicação do caminho percorrido na elaboração do trabalho, quando, na seção 6, será descrito como alguns dos conceitos especificados foram utilizados para a realização da análise na base de dados escolhida. Os resultados verificados e as conclusões a que se chegou são apresentados e interpretados na seção 7 e, ao final, algumas ideias para trabalhos futuros se encontram na seção 8.

2 Contexto Atual

Sabe-se que houve avanços quanto às ferramentas utilizadas no processamento de *Big Data*, e a evolução do modelo monoestágio do *MapReduce* para o modelo de processamento multiestágio do *framework Spark* consiste em um dos casos mais conhecidos **[4]**, **[5]**, **[6]**. Já trabalhos comparativos entre *Spark* e *Dataflow* podem ser encontrados em **[7]**, **[8]** e **[9]**. Destas referências, as duas primeiras foram criadas pela empresa Google, que desenvolveu o *Dataflow* como um *framework* integrado com outras ferramentas na sua plataforma *Google Cloud*. Ao se gerar um código usando o *framework* para processar dados em paralelo (em fluxo contínuo) não se faz necessário gerar um novo código para se trabalhar com dados em batch: o mesmo código serve para os dois tipos de processamento, facilitando a análise de grandes quantidades de dados.

> [...] nenhum outro modelo de programação paralela de dados em grande escala fornece a enorme capacidade e a facilidade de uso do Dataflow/Beam **[8]**.

Pelo fato de o *Dataflow* ser apresentado como um sucessor do *MapReduce* **[10]** e trazer soluções diferentes das propostas pelo *Spark*, e ainda por ser disponibilizado de maneira simples pelo Google Cloud, foi decidido elaborar este artigo com foco no processamento de dados com *Dataflow*, pois ele oferece um *SDK* (Kit de Desenvolvimento de Software) de rápida curva de aprendizado que permite a construção de *Jobs* em vários motores de tempos de execução (*runtime engine*), conforme é descrito na proposta de Onofre e Jagielski **[10]**.

3 Objetivos

- Criar um modelo preditivo usando *DataFlow* para definir se um *post* no fórum *StackOveflow* terá uma resposta marcada como aceita; e
- Demonstrar o potencial de aplicação de novas ferramentas e novos paradigmas de

Aprendizado de Máquina em grandes volumes de dados.

4 Trabalhos Relacionados

A análise de dados através de *DataFlow* é demonsytada em **[11]**, que explica o *framework* em profundidade, descrevendo seu modelo de programação e o funcionamento do SDK, não sedo feita a analise de uma base, que mostraria o uso de forma pratica do *DataFlow*. Já em [12], encontra-se um estudo comparativo das funcionalidades do *framework Dataflow* com *Spark* e *Flink*, analisando-se o processamento de cada um, o mecanismo de tolerância a falhas e suas forças e fraquezas, mostrando, a partir de outros *frameworks* utilizados no mercado, os atributos do *Dataflow*.

5 Referencial Teórico

Nesta seção, será apresentada uma introdução acerca dos conceitos e tecnologias utilizadas neste trabalho — especificamente: *Big Data*, *Hadoop*, *MapReduce*, *Spark,* além do *DataFlow* e das bibliotecas *TensorFlow* e *Scikit-learn*.

5.1 Big Data

Apesar de ser um termo bastante usado na atualidade, o conceito de *Big Data* pode variar de autor para autor ou quando se têm objetivos diferentes. Doug Laney **[13]** define *Big Data* com foco nas três dimensões de um conjunto dos dados. Segundo o autor, uma massa de dados pode ser caracterizada como *Big Data* quando supre os requisitos de: volume (ou seja, a quantidade de dados); velocidade (em que as informações são geradas); e variedade (em relação às fontes e tipos de dados — de diversos formatos, estruturados ou não estruturados).

Com o passar do tempo, muitas outras dimensões foram adicionadas a esse conceito. A título de exemplo: veracidade **[14]** e valor **[15]**, dentre outros **[16]**. Muitos desses termos se referem ao resultado do emprego de técnicas de ciência de dados, não refletindo essencialmente a atribuição de novas características. Sob uma ótica, essas classificações adicionais podem ser percebidas como um esforço para se entender melhor o conceito *Big Data*; por outra perspectiva, pode-se incorrer em falta de precisão ou mesmo publicidade extravagante sobre o tema.

Dados em escala *Big Data* não podem ser manipulados através de ferramentas e métodos tradicionais. Dessa forma, ferramentas e técnicas específicas têm sido desenvolvidas para seu tratamento adequado perante os desafios impostos na atualidade.

5.2 Hadoop

O *Hadoop* é um *framework* usado para armazenamento distribuído e para processamento de grandes conjuntos de dados **[17]**. Dentre seus componentes principais, encontram-se o HDFS (*Hadoop Distributed File System*) e a máquina de execução MapReduce.

O *HDFS* trata do armazenamento distribuído de dados em *hardware* de forma escalável e confiável. A sua arquitetura faz com que a informação seja armazenada em blocos, que são distribuídos em diversas máquinas ao longo de um *cluster*. Na arquitetura *HDFS*, existe uma máquina que controla as demais, chamada *namenode*; os dados em si são armazenados nos *datanodes,* e em cada nó do *cluster* pode existir um ou mais *datanodes*. **[18]**. Essa arquitetura foi desenhada para ser executada em *hardware* de baixo custo.

O *MapReduce*, por sua vez, é um modelo de programação desenvolvido pela Google usado no processamento de vastas informações em *clusters* **[19]**. O seu modelo de processamento é paralelo e escalável, o que gera um grande conjunto de dados organizados em forma de blocos, que são distribuídos em todos os nós do *cluster* utilizando duas funções: *map* e *reduce* **[5]**.

Esse modelo funciona bem quando submetido a processamento de dados em lote, lidos de forma sequencial, e por isso a velocidade não é essencial, já que o *MapReduce* tem que buscar os dados armazenados pelo *HDFS*, ler esses dados e realizar a operação que foi designada. Após isso, o resultado deve ser salvo no *armazenamento* HDFS. O procedimento é repetido até que todo o trabalho seja finalizado. Tais operações tornam o processamento lento devido ao excessivo número de operações de entrada e saída. Em cenários de tratamento de dados iterativos ou em um fluxo

http://dx.doi.org/10.25286/repa.v3i3.942

contínuo de dados, fazem-se necessárias abordagens que diminuam o gargalo do *Hadoop*.

Dessa necessidade, surgiu o *Spark*, um novo *framework* que trabalha como um motor de computação distribuída em memória **[20]** resultando em velocidade de processamento maior.

5.3 Spark

Em se tratando de processamento de tarefas iterativas e distribuídas (tais como Aprendizagem de Máquina), o *framework Apache Spark* é uma solução largamente adotada, como se pode ver pela lista, disponibilizada em sua página oficial, de empresas que atualmente utilizam o *Spark* **[21]**. Com esta plataforma, é possível manipular dados de vários tipos — texto, grafos e imagens — e que vêm de diferentes origens — em lote ou fluxo contínuo de dados em tempo real. Entre as suas vantagens, encontra-se o fato do *Spark* ser escalável, rápido e possuir curva de aprendizado simples, permitindo codificação de tarefas em *Python, Java, R* ou *Scala* **[22]**.

O armazenamento em memória é possível graças a uma estrutura de dados paralela e tolerante a falhas chamada RDD — Conjunto de Dados Distribuídos Resilientes **[23]**, que salva os resultados intermediários na memória do nó correspondente; assim, podem ser executados algoritmos iterativos de Aprendizagem de Máquina com maior eficiência.

Esse conjunto de dados não é salvo em armazenamentos físicos, e sim na memória volátil dos nós do *cluster*; isso torna sua recuperação e escrita muito rápidas e, assim, pode-se reutilizar esses resultados parciais para uso em operações paralelas do tipo *MapReduce* **[24]**.

O programador pode criar diversos *pipelines*, que contêm em si várias etapas de diferentes complexidades, e os dados que se encontram nesses *pipelines* podem ser compartilhados, permitindo que os *Jobs* do *Spark* possam trabalhar com os mesmos dados. Os resultados do processamento dos dados são retidos em memória, minimizando a escrita no HDFS ou outro armazenamento que possa ser escolhido.

Assim, para aplicações que exigem processamento rápido e análises em tempo real, o *Apache Spark* supera o seu predecessor, o *Hadoop*.

5.3 Apach Beam e Processamento *Big Data*

O *Dataflow*, disponibilizado pela Google na plataforma *Google Cloud,* permite que sejam desenvolvidos modelos de análise de *Big Data*, que possa ser integrada com diversos produtos disponibilizados pela plataforma para o trabalho com *Big Data*, como: o *Big Query*, para trabalhos com bases *Big Data*; o *Storage*, para armazenamento de grandes quantidades de dados; o *Dataproc*, um serviço de gerenciamento do *Spark* e do *Hadoop*; dentre outros. Enquanto no *Google Cloud* é disponibilizada a versão 1.0 do *Dataflow*, existe uma versão mais atual, 2.x, que é chamada de *Apache Beam*.

O *Apache Beam* é mantido pela Apache e tem distribuição gratuita do *SDK* utilizado pelos programadores para gerar os códigos pertinentes.

O *SDK* permite a criação de *pipelines* de processamento, que poderão ser executados em diversas plataformas, inclusive na nuvem. A arquitetura suporta linguagens como *Java* e *Python*.

Para fins de compreensão, ressalta-se que, ao se usar o termo *Dataflow* neste artigo, estará sendo mencionado o *framework* disponibilizado pela Google na sua plataforma *Google Cloud* — cuja infraestrutura foi usada para maior rapidez no processamento — e, quando for mencionado *Apache Beam*, o que estará sendo citado é o *SDK*, que foi usado para o desenvolvimento do código.

O *Apache Beam* — que contém parte do modelo de programação/*SDK* do *Google Cloud Dataflow*, o qual foi criado a partir de uma evolução do *MapReduce* e incubado pela Apache — tem seu nome originado da junção de duas palavras, que nomeiam os dados "Batch" e o processamento "strEAM". Disso, infere-se que o *Apache Beam* pode ser usado para dados em *batch* — que são dados separados em lotes, sendo processado um lote de cada vez —, para um *stream* contínuo de dados e para vários tipos de processamento de dados, incluindo ETL. Tudo isso pode ocorrer de forma simplificada em motores variados, como *Spark Runner* ou *Google Cloud Dataflow*, e em qualquer escala **[10]**.

Além do *SDK*, outro grande diferencial do Apache Beam é o seu modelo de programação unificado. Em relação ao modelo, existem quatro conceitos principais de nível superior que devem ser pensados ao se construir o *Job* de

processamento, que são: *Pipeline*, *PCollection*, *PTransform* e *Pipeline Runner* **[10]**.

O *Pipeline* representa um *Job* que recebe dados de entrada, realiza alguma computação e armazena a saída, seja um resultado ou os próprios dados tratados na saída. O *PCollection* representa o conjunto de dados dentro do *Pipeline*; esse conjunto possui tamanho praticamente ilimitado, podendo ser trabalhado com dados em lote ou cuja origem tenha atualização contínua, criando-se, assim, um *Job* para um *stream* de dados. Vários tipos de *Jobs* podem ser feitos dentro do *Pipeline*, como construção de histogramas e criação de modelos de Aprendizagem de Máquina, e é através do *PCollection* que *Jobs* paralelos podem acontecer **[25]**.

O *PTransform* é a computação em si, que transforma dados de entrada em dados de saída **[26]**. Ele pode executar transformações em elementos, agregar vários elementos em conjunto ou ser uma combinação composta de outros *PTransforms* **[25]**. Observe a Figura 1.

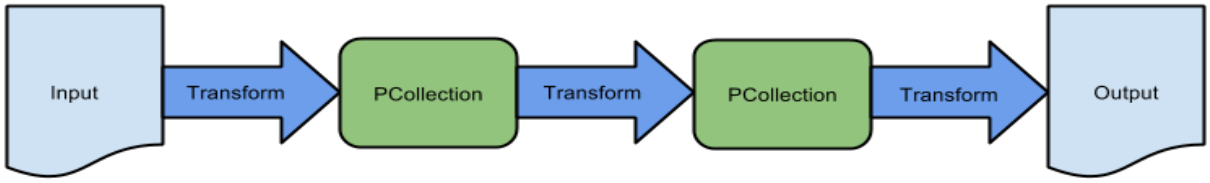


Figura 1: Funcionamento de Jobs do Spark. Fonte:**[45]**

Ao se criar programas com o *SDK* do *Apache Beam*, um *Job* de processamento é gerado e será executado por um dos serviços de *backend* de processamento distribuído e múltiplo, através do que é chamado de executor de *Pipelines*: os *PipelineRunners* **[27]**. Existem várias opções; algumas são: *DirectRunner*, que é executado diretamente em uma máquina local, após a instalação do *Beam* na máquina; *DataflowRunner*, no qual o *Pipeline* a ser executado é submetido no *Google Cloud Data Flow* (executor utilizado neste artigo) — esta opção provê uma integração com outros sistemas da plataforma; e *SparkRunner*, que executa o *Pipeline* em um *Apache Spark Cluster* **[28]**. Este fornece um *framework* de computação cluster *open-source*, além de poder ser executado localmente ou na nuvem.

5.5 Bibliotecas de Aprendizagem de Máquina

O treinamento e teste do modelo preditivo foi desenvolvido usando o *Tensor Flow* [29], uma biblioteca para Aprendizagem de Máquina de código aberto. As demais atividades do projeto foram desenvolvidas usando a biblioteca *Scikit-learn* [30].

O *Tensor Flow* disponibiliza ferramentas para definir modelos novos, não oferecendo soluções prontas de Aprendizagem de Máquina. Assim, o desenvolvedor com conhecimento técnico pode criar modelos flexíveis com um conjunto extenso de funções e classes e realizar os cálculos a partir de chamadas, podendo o *Tensor Flow* também ser usado para execução de códigos matemáticos complexos [29].

Além disso, o *Tensor Flow* permite implantar aplicativos em clusters distribuídos, em estações de trabalho locais, em dispositivos móveis e em aceleradores customizados. Pode-se trabalhar com o *Tensor Flow* da maneira que mais se adequa ao caso do projeto **[31]**.

Uma computação *Tensor Flow* é descrita por um gráfico de fluxo de dados, que é composto por um conjunto de nós. Esse gráfico representa todos os cálculos feitos, as operações matemáticas, os parâmetros e suas regras e o pré-processamento de entrada. Vale a pena salientar que o modelo suporta várias execuções simultâneas sobre o processamento geral. Cada nó possui zero ou mais entradas e saídas e representa a instanciação de uma operação **[32]**.

A biblioteca *Scikit-learn* **[33]** possui código aberto e é específica para a linguagem *Python*. Ela contém um conjunto bastante amplo de algoritmos para modelos estatísticos e implementa muitos

http://dx.doi.org/10.25286/repa.v3i3.942

algoritmos de Aprendizagem de Máquina, o que visa a facilitar a vida do desenvolvedor **[34]**. A biblioteca é focada na modelagem dos dados, sendo usada em conjunto com outras bibliotecas *Python* para carregamento, manipulação e sumarização dos dados.

5.6 Métricas no modelo preditivo

A base de dados do fórum *StackOverflow* se encontra armazenada no *Big Query* **[35]** na área denominada pela plataforma como *bigquery-public-data* **[36]** e contém as interações entre os usuários e outras informações dessa comunidade on-line. A base sempre reflete o conjunto atual de dados compartilhados da comunidade.

O *Big Query* permite a análise e a consulta dos dados através de *scripts* SQL; com isso, podem ser realizados tratamentos na base e, assim, utilizá-la para responder diversas perguntas — no caso deste estudo, analisando as variáveis dos *posts*, determinar se as perguntas cadastradas no fórum *StackOverflow* terão uma resposta aceita ou não pelo usuário que fez a pergunta. Com a base escolhida, foi realizada uma análise preditiva para tentar responder à questão.

A análise preditiva é diferente de outros tipos de análise de dados (como a descritiva ou a prescritiva) por usar padrões do passado — como apontado por **[37]** — para determinar eventos ou respostas futuras, não se utilizando de conhecimento empírico baseado somente na experiência prévia e pessoal do analista. A análise preditiva tem, assim, o objetivo de usar dados estatísticos e históricos para decidir as melhores ações ou mesmo saber como uma determinada situação irá ocorrer. Em uma abordagem de negócios, por exemplo, a análise preditiva pode ser utilizada para várias atividades, como: identificar tendências; prever comportamentos; entender as reais necessidades de clientes; determinar se um paciente possui determinada doença; ou mesmo se um voo irá atrasar ou não.

O modelo de Aprendizagem de Máquina escolhido foi a regressão logística. Uma das razões dessa escolha foi sua capacidade para trabalhar com modelos binários — no caso da base analisada, *true* ou *false*. A regressão logística calcula a relação entre a variável categórica e todas as outras variáveis dependentes. Um bom modelo deve avaliar quais dessas características realmente fazem a diferença na contagem de probabilidade. A regressão logística funciona fazendo suposições sobre os dados analisados para o aprendizado sobre a base e, dessa forma, prevê probabilidades, e não somente as categorias possíveis **[39]**.

Para desenvolvimento do modelo, os dados devem ser divididos em grupos de treinamento e de teste. Para tal, podem ser usados vários métodos; no caso concreto, foi usado o *Cross Validation.* Os conjuntos de dados são usados, em um momento, para treinamento e, em outro, para teste, e assim não há o risco de o modelo aprender apenas sobre uma categoria específica **[38]**.

As métricas descritas a seguir foram utilizadas para definir se o modelo foi bem treinado.

A matriz de confusão mostra o número de classificações corretas em comparação com as classificações que foram previstas para as classes existentes na base. Bases com categorias binárias, do tipo *true* ou *false*, geram uma matriz de confusão menor e de melhor compreensão do que se tivessem mais categorias. Esse tipo de visualização (apresentado na Tabela 2) é mais simples e foi gerado para a matriz de confusão da análise da base do *StackOverflow*.

Com os valores definidos na matriz, algumas métricas foram calculadas, como a acurácia, que consiste na proporção de predições corretas, observando-se o acerto total. Por utilizar todos os valores de desempenho gerados pela matriz de confusão, ela é considerada um classificador geral.

Outra métrica utilizada foi a curva ROC. No campo da Ciência de Dados, essa curva é útil para se trabalhar com domínios cujas classes estão desbalanceadas, pelo fato de ela ser baseada nas taxas médias de VP e FP, que não dependem da distribuição das classes.

A curva ROC é construída ordenando-se as linhas, validando-se os valores (se positivos ou negativos) e construindo-se a curva de baixo para cima. Quando o valor é positivo, a curva sobe; quando negativo, ela inclina-se para a direita.

A curva ROC é baseada nas métricas VP e FP, mas não nos valores absolutos, e sim nas suas taxas, podendo estas serem chamadas de Taxa VP/Taxa FP (*VP rate/FP rate*) ou identificadas por seus nomes específicos, que são: sensitividade, ou *recall*, e especificidade, respectivamente. Essas métricas são obtidas das seguintes formas:

- Sensitividade: porcentagem de amostras classificadas como positivas e corretas, sobre o total de amostras positivas.

- Especificidade: porcentagem de amostras negativas identificadas corretamente, sobre o total de amostras negativas.

6 Metodologia

Os dados do *StackOverflow* utilizados para o estudo de caso totalizam um tamanho de 174 GB, com 786.588 amostras, contendo variáveis diversas, descritas na Tabela 1, tais como: quais foram as perguntas feitas, quem são os usuários, quais foram as interações com os *posts*, dentre outros dados.

A base total do fórum *StackOverflow* não está totalmente disponível no *Big Query*, por isso não foram usadas na pesquisa todas as informações existentes no fórum, apenas as já se encontravam na base disponibilizada no *Big Query*. Além disso, deve-se observar que, pela natureza do fórum, a base tende a crescer de maneira constante, exigindo do programador um maior conhecimento em análise de *stream* contínuo. No entanto, esse não foi o foco do trabalho e sim demonstrar o potencial da ferramenta e a aplicabilidade do modelo, mesmo que não manipulando dados em volumes que se encaixem plenamente no conceito *Big Data* em relação ao seu tamanho.

A base possui desbalanceamento de 21% para perguntas respondidas e 79% para posts que não tiveram respostas aceitas. Apesar disso, não houve tratamento de balanceamento, pois, uma vez que esses 21% não incorrem em eventos raros, como foi demonstrado em [39], o resultado não seria alterado, havendo impacto apenas na métrica acurácia que foi complementada com outras métricas — curva ROC, sensitividade e especificidade. A curva ROC é particularmente útil em domínios em que existe uma grande desproporção entre as classes **[40]**.

Durante o pré-processamento da base foram constatados dados quantitativos — numéricos, e variáveis textuais de identificação dos *posts* — como o título e o próprio texto relacionado. As variáveis numéricas podem ser classificadas como discretas ou contínuas **[41]**: as primeiras são baseadas em contagens — no caso da base em questão, número de views de um *post* ou número de vezes em que ele foi marcado como favorito; já as segundas consistem de variáveis numéricas que podem assumir qualquer valor em um determinado intervalo, podendo ser data/hora, por exemplo (no caso da base escolhida, a data em que foi cadastrada a pergunta e a data da última interação com o *post* eram variáveis contínuas, mais difíceis de se trabalhar, então foi criada uma variável que registra o tempo que o *post* permaneceu ativo no fórum, nomeada *days_it_has_been_active*).

http://dx.doi.org/10.25286/repa.v3i3.942

Tabela 1: Esquema da base de dados analisada.

Nome da Coluna	Tipo	Descrição
Answer_count	Int	Número de respostas sugeridas
Comment_count	Int	Número de comentários feitos no *post*
Favorite_count	Int	Número de vezes em que o *post* foi marcado como favorito
Score	Int	Pontuação dada ao *post* pela plataforma
View_count	Int	Número de vezes em que o *post* foi visualizado
Days_it_has_been_active	Int	Número de dias corridos entre a data da criação do *post* e a última data ativa
Accepted_answe_id	Int	ID do *post* aceito como resposta; caso não possua, o valor é zero
Has_accepted_answer	Bool	Variável criada a partir da anterior; caso tenha resposta, o valor é true (1); caso não possua, é false (0)

Fonte: o autor.

As variáveis textuais foram retiradas da base usada pelo modelo, pois não forneciam informações de classificação (funcionando apenas como um Identificador Único do objeto cadastrado na linha da base); com isso, foram utilizadas apenas as variáveis numéricas da tabela já existente, sem uso de categorias textuais.

Não foram observadas variáveis categóricas que pudessem ser extraídas a partir das outras existentes, ficando a base com as variáveis discretas e com a variável alvo que identifica se um *post* possui uma resposta aceita ou não (*Accepted Answer ID*). No entanto, como não se pôde trabalhar com IDs específicos, passou-se a utilizar a variável criada has_accepted_answer, como mostrado na Tabela 1

Essa variável define se existe ou não uma resposta ao *post*; esse tipo de variável lida com situações em que o valor final é categorizado em um número finito de possibilidades mutuamente exclusivas **[42]**. Por ser uma variável binária, o 0 representa *posts* sem resposta aceita, e o 1, os que possuem resposta aceita, como mostrado na Tabela 1.

No fim, as variáveis utilizadas foram: quantidade de respostas sugeridas; número de comentários feitos no *post*; quantas vezes a pergunta foi marcada como favorita; pontuação do *post* dado pela plataforma; número de vezes em que a postagem foi visualizada; ID da postagem aceita como resposta; e as duas novas criadas — dias em que a postagem passou ativa e uma variável booleana, que informa se houve resposta aceita. Tais variáveis se encontram relacionadas na Tabela 1.

Com a criação da variável *days_it_has_been_active*, esperava-se poder auxiliar o modelo a identificar melhor se um *post* teria uma resposta válida ou não, mas foi observado que o tempo de vida de um *post* não influenciou a análise probabilística, havendo apenas uma pequena variação das métricas utilizadas pelo artigo, não sendo realmente considerada como uma melhora no modelo.

Com as alterações na base descritas antes e com a escolha do modelo preditivo, procurou-se prever o resultado quanto à aceitação de uma resposta ao *post* — dadas as informações da Tabela 1, sem as variáveis *has_accepted_answer* e *accepted_answer*_id — e, assim, responder à pergunta proposta no artigo.

Para executar o modelo produzido, o *SDK* ajudou na criação do *Pipeline*, que busca a base hospedada no Storage com as novas variáveis já geradas. Além do *SDK*, também foram usadas as bibliotecas de Aprendizagem de Máquina citadas anteriormente — *Tensor Flow* e *Scikit-learn* — especializadas e disponibilizadas em *Pyhton*. Por ser uma base grande e a máquina local utilizada não possuir um grande poder de processamento, foi utilizado o executor do *Google Data Flow* e, assim, a computação foi feita na nuvem, tendo sido geradas as métricas explicadas no Referencial Teórico para analisar se o modelo foi treinado com sucesso.

7 Resultados e Conclusões

A partir do modelo produzido e da análise feita partindo das variáveis obtidas, foi possível criar um modelo com resultados altos para as métricas consideradas e, assim, realizar a predição desejada.

A aplicação do SDK Apache Beam demonstrou facilidade de uso e praticidade quanto à criação de Jobs no executor de Pipeline utilizado para a avaliação descrita neste estudo. O fato de as ferramentas serem distribuídas de maneira gratuita e separadas umas das outras — como é o caso do *SDK* ou do *Runner* — possibilita a geração da computação desacoplada de qualquer serviço, ficando o desenvolvedor livre para escolher como montar uma arquitetura de processamento que seja ideal para o seu contexto específico. Por exemplo, o usuário pode escolher se usará o *SDK* com o Hadoop ou com todo o conjunto do Apache *Beam*, além de poder utilizar o *Pipeline Runner* local ou no Apache Spark Cluster.

A fase de pré-processamento depende muito da capacidade do analista para identificar o que mais se adequa ao estudo **[43]** e para verificar quais variáveis farão com que o modelo seja bem-sucedido.

Ao fim dessa fase e após o modelo gerado ser rodado obteve-se a Tabela 2, a matriz de confusão, da análise feita.

Tabela 2: Matriz de confusão gerada pela análise

		Valor Previsto	
		0	1
Valor Verdadeiro	0	484.462	103.234
	1	72.530	126.362

Fonte: o autor.

A partir da matriz, é possível obter o valor da acurácia, que, para o modelo testado, foi de 0,85385. Porém, a acurácia não é uma métrica confiável para o caso de bases desbalanceadas **[44]**. Isso acontece porque o classificador acaba se mostrando tendencioso para o lado com mais classes; por essas serem as que mais existem na base, a probabilidade tende a ser maior.

Contudo, esse tipo de tendência não influencia a curva ROC, porque ela é invariante quanto à proporção de exemplos entre as classes — desde que seja assumida a distribuição de uma classe como característica do domínio **[40]** — , e, por isso, a curva ROC é a métrica mais importante na análise e ajuda a complementar o valor informado pela acurácia. O valor da curva foi de 0,92 e pode ser observado na Figura 2. Pode-se observar também, nessa imagem, a linha randômica, que mostra quando o modelo não consegue prever a categoria, atuando como um modelo de adivinhação.

A área abaixo da curva ROC (AUC) está associada ao poder discriminante da análise de avaliar o modelo, pois quanto maior a área, maior a assertividade do modelo.

Os valores obtidos para a sensitividade e para a especificidade foram, respectivamente: 0,9181 e 0,8249. Esses valores altos demonstram o nível de acerto do modelo, já que ambos analisam a proporção das vezes que o modelo acertou a variável categórica analisada (has_accepted_answer) — seja negativamente ou positivamente — sobre o número dos valores reais positivos e negativos encontrados na base.

http://dx.doi.org/10.25286/repa.v3i3.942

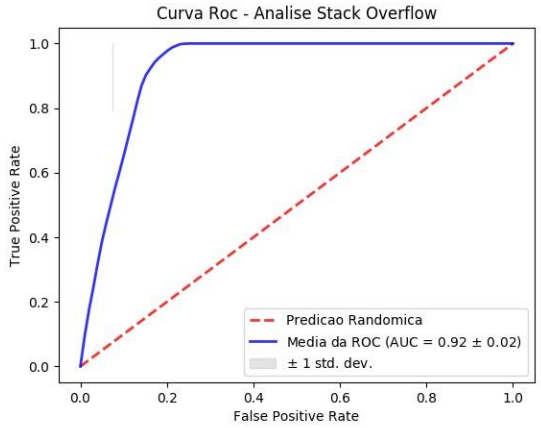


Figura 2: Curva ROC da análise preditiva
Fonte: o autor.

O uso da arquitetura montada executando um modelo de Regressão Logística, a geração do código aplicando o *Beam* e a escolha do *DataflowRunner* para o processamento se mostraram aceitáveis para o desenvolvimento de um modelo preditivo que pode realizar a validação probabilística.

Com base nos resultados obtidos, pode-se confirmar que o modelo gerado para realizar a análise preditiva nos dados coletados teve sucesso, e, com isso, foi possível prever se uma pergunta no fórum teria ou não uma resposta marcada como.

8 Projetos Futuros

A análise descrita no artigo poderia se beneficiar de um balanceamento da base para a verificação da existência de uma melhora nas métricas, e poderia ser feita uma análise com rede neural. Gerando um novo modelo que se beneficiaria da estrutura paralela que este tipo de análise apresenta e da sua habilidade de aprender e generalizar.

Esse modelo pode ser empregado para buscar o mesmo tipo de resposta que foi encontrada para a análise da base do *StackOverflow* em outros fóruns de colaboração. Esse tipo de análise pode dar mais conhecimento aos usuários do fórum uma vez que saberão se podem esperar ou não por uma resposta com base nas interações no *post* da sua pergunta.

Assim, seria possível obter um uso prático para o modelo, além de uma aplicação em uma variedade de cenários.

Referências

[1] STACK OVERFLOW - Where Developers Learn, Share, & Build Careers. Disponível em: < https://stackoverflow.com/>. Acesso em: 18 abr. 2018.

[2] Computação em nuvem, serviços de hospedagem e APIs do Google. Disponível em: <https://cloud.google.com/>. Acesso em: 18 abr. 2018.

[3] PRAMANA, Setia et al. Big data for government policy: Potential implementations of bigdata for official statistics in Indonesia. In: INTERNATIONAL WORKSHOP ON BIG DATA AND INFORMATION SECURITY, 2017, Jakarta. **Proceedings**...Jakarta: IEEE, 2017. p. 17-21.

[4] AGNEESWARAN, Vijay S. **Big Data Analytics Beyond Hadoop**. New Jersey: Perason, 2014.

[5] ZHAO, Disheng. Performance comparison between Hadoop and HAMR under laboratory environment. **Procedia Computer Science**, v. 111, p. 223–229, 2017.

[6] AGNEESWARAN, Vijay S. **Big Data Analytics Beyond Hadoop**: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives. FT Press, 2014.

[7] GONZALEZ, Jose Ugia; KRISHNAN, S. P. T. **Building Your Next Big Thing with Google Cloud Platform: A Guide for Developers and Enterprise Architects**. New York: Apress, 2015.

[8] AKIDAU, Tyler; PERRY, Frances. **DATAFLOW/Beam e Spark**: uma comparação de modelo de programação. Apach Beam Commiters, 2016. Disponível em: <https://cloud.google.com/dataflow/blog/dataflow-beam-and-spark-comparison?hl=pt-br>. Acesso em: 9 mar. 2018.

[9] MORGAN, Timothy Prickett. **Google Pits Dataflow Against Spark**. The Next Plataform, 2016. Disponível em: <https://www.nextplatform.com/2016/05/03/google-pits-dataflow-spark/>. Acesso em: 13 mar. 2018.

[10] ONOFRE, J. B.; JAGIELSKI, J. **BeamProposal - Incubator Wiki**. Disponível em: <https://wiki.apache.org/incubator/BeamProposal>. Acesso em: 26 dez. 2017.

[11] AKIDAU, Tyler et al. The Dataflow Model: a Practical Approach to Balancing Correctness, Latency and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing, **Proceedings of the VLDB Endowmen**, v. 8, n. 12, p. 1792-1803, 2015.

[12] SALEM, Farouk. **Comparative Analysis of Big Data Stream Processing Systems**. Aalto University, 2016.

[13] LANEY, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety Application Delivery Strategies. **META group research note**, v. 6, n. 70, p. 1, 2001.

[14] KEPNER, Jeremy et al. Computing on masked data: A high performance method for improving big data veracity. In: HIGH PERFORMACE EXTREME COMPUTING CONFERENCEi, 8., 2014, Massachusetts. **Proceedings...** Massachuttes: IEE**E**,2014. p. 1–6.

[15] LIU, S. M. et al. Big Data A Survey. **Mobile Networks and Applications,** New York 2014.

[16] GRADY, Nancy; CHANG, Wo (ed.) N**IST Big Data Interoperability Framework**: Volume 1, Definitions. Gaithersburg: NIST, 2015. p. 32.

[17] UZUNKAYA, C.; ENSARI, T.; KAVURUCU, Y. Hadoop Ecosystem and Its Analysis on Tweets. **Procedia - Social and Behavioral Sciences**, s.l., v. 195, p. 1890–1897, 2015.

[18] APACHE Hadoop 3.0.0 – HDFS Architecture. Disponível em: <http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html#NameNode_and_DataNodes>. Acesso em: 22 fev. 2018.

[19] MINER, Donald; SHOOK, Adam. **MapReduce design patterns**. Sebastopol: O'Reilly, 2012.

[20] BENGFORT, B.; KIM, J. **Analítica de dados com Hadoop**. São Paulo: Novatec, 2016.
[21] APACHE Spark. Disponível em: <http://spark.apache.org/powered-by.html>. Acesso em: 8 mar. 2018.

[21] KANE, F. **Frank Kane's Taming Big Data with Apache Spark and Python**. Birmingham: Packt Publishing, 2017.

[22] ZAHARIA, M.; CHOWDHURY, M.; DAS, T.; DAVE, A. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing, **Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation**. USENIX Association, 2012.

[23] ZAHARIA, Matei. et al. **Spark**: Cluster Computing with Working Sets. In: Conferece on Hot topics in Cloud Computer, 2., 2010, Boston. **Proceedings**... Boston: USENIX, 2010.

[24] THE WORLD beyond batch: Streaming 102 - O'Reilly Media. Disponível em: <https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-102>. Acesso em: 15 fev 2018.

[25] MODELO de programação do Dataflow. Disponível em: <https://cloud.google.com/dataflow/model/programming-model>. Acesso em: 15 fev. 2018.

[26] PIPELINERUNNER (Google Cloud Dataflow SDK 1.9.1 API). Disponível em: <https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/PipelineRunner>. Acesso em: 14 mar. 2018.

http://dx.doi.org/10.25286/repa.v3i3.942

[27] APACHE Beam. Disponível em: <https://beam.apache.org/>. Acesso em: 15 fev 2018.

[28] TENSORFLOW. Disponível em: <https://www.tensorflow.org/>. Acesso em: 10 mar. 2018.

[29] SCIKIT-LEARN: machine learning in Python. Disponível em: <http://scikit-learn.org/stable/>. Acesso em: 14 mar. 2018.

[30] ABADI, M. et al. TensorFlow: A System for Large-Scale Machine Learning. In: USENIX CONFERENCE ON OPERATING SYSTEMS DESIGN AND IMPLEMENTATION, 12., 2016, Geórgia. **Proceedings**... Geórgia: Berkeley, 2016. p. 265–284.

[31] MARTIN, A. et al. **TensorFlow**: Large-scale machine learning on heterogeneous systems. Disponível em: < https://www.tensorflow.org/>. Acesso: 14 mar. 2018.

[32] GÉRON, A. **Hands-On Machine Learning with Scikit-Learn and TensorFlow**. Sebastopol: O'Reilly Media, 2017.

[33] ABRAHAM, A. et al. Machine Learning for Neuroimaging with Scikit-Learn. F**rontiers in neuroinformatics**, v.8, p.14, 2014.

[34] GOOGLE BigQuery. Disponível em: <https://bigquery.cloud.google.com/dataset/bigquery-public-data:stackoverflow>. Acesso em: 23 fev. 2018.

[35] CONJUNTOS de dados públicos do Google BigQuery. Disponível em: <https://cloud.google.com/bigquery/public-data/>. Acesso: 14 mar. 2018.

[36] PYNE, S.; PRAKASA; B.L.S.; RAO, S. B. **Big Data Analytics**: Methods and Applications. New Delhi: Springer, 2016.

[37] TRAIN/TEST Split and Cross Validation in Python – Towards Data Science. Disponível em: <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>. Acesso em: 15 fev. 2018.

[38] KING, Gray; ZENG, Langche. Logistic Regression in Rare Events Data. **Political Analysis**, v. 9, n. 2, p. 137–163, 2001.

[39] PRATI, R. C.; BATISTA, GEAPA; MONARD, M. C. Curvas ROC para avaliação de classificadores. **IEEE Lat. Am. Trans.**, s.l., v. 6, n. 2, p. 215–222, 2008.

[40] LE BLANC, D. C. **Statistics**: Concepts and Applications for Science. Sudbury: Jones and Bartlett, 2004. v. 2.

[41] HAIDER, M. **Getting Started with Data Science: Making sense of Data with Analytics**, IBM Press, Indianápolis, 2016.

[42] ENRIQUE, G.; PRADO, D. A.; BATISTA, A. Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. **Instituto de Ciências Matemáticas e de Computação**, São Carlos, 2003.

[43] GRIGOREV, A. **Mastering Java for data science** : building data science applications in Java. Birmingham: Packt Publishing, 2017.

[44] GOOGLE Dataflow And Apache Beam (I). Disponível: <http://www.datio.com/development/google-dataflow-and-apache-beam-i/>. Acesso em: 14 mar. 2018.

Análise de relação entre variáveis de ocorrências de crimes da cidade do Recife

Estudo de caso por meio da avaliação de dados da Secretaria de Defesa Social

Carolina Lima Gomes de Melo[1] orcid.org/0000-0001-7826-6392
Rodrigo Lins Rodrigues[2] orcid.org/0000-0002-3598-5204

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Departamento de Educação, Universidade Federal Rural de Pernambuco, Recife, Brasil

E-mail do autor principal: clgm@ecomp.poli.br

Resumo

O aumento da criminalidade nos últimos anos na cidade do Recife tem deixado em alerta a população e a polícia. A sociedade tem modificado seus hábitos com receio da violência, enquanto a polícia tenta prevenir e combater novos crimes de acordo com as ocorrências relatadas às delegacias da cidade. Este estudo tem por objetivo encontrar e analisar possíveis relações existentes entre variáveis representativas de ocorrências de crimes das delegacias da cidade do Recife, para auxiliar na análise e na criação de estratégias para melhor prevenção de crimes. Inicialmente foi realizado um pré-processamento das bases por meio das ferramentas Pentaho Kettle e R, em seguida foram realizadas estatísticas descritivas e por fim foram aplicados os testes de hipótese de Qui-quadrado e o de Fisher.

Palavras-Chave:*Ocorrências de crimes em Recife; Qui-quadrado; Fisher; R; Kettle;*

Abstract

The increase in crime in recent years in the city of Recife has left the population and the police alert. Society has modified its habits for fear of violence, while police try to prevent and fight new crimes according to the incidents reported to city police stations. This study aims to find and analyze possible relationships between representative variables of crime occurrences in the city of Recife, to assist in the analysis and creation of strategies for better crime prevention. Initially, a pre-processing of the bases was carried out using the Pentaho Kettle and R tools, followed by descriptive statistics and finally the Chi-square and Fisher hypothesis tests were applied.

Key-words:*Occurrences of crimes in Recife Chi-square; Fisher; R; Kettle;*

http://dx.doi.org/10.25286/repa.v3i3.943

1 Introdução

Segundo George Kelling **[1]**, um dos autores do artigo "Teoria da Janela Quebrada" e do livro "Consertando janelas quebradas", se um prédio tem uma janela quebrada, é melhor consertá-la logo, antes que um vândalo quebre todas as outras. Nesta perspectiva acredita-se que crimes pequenos podem atrair crimes mais graves. Essa ideia tem orientado programas nos Estados Unidos, onde o controle das pequenas infrações tem ajudado a prevenir crimes mais graves**[1]**.

O crescente aumento da violência no estado de Pernambuco aponta para uma urgência na criação de soluções estratégicas, desafiadoras para a gestão da segurança pública. Algumas cidades pelo mundo têm adotado medidas proativas de controle e prevenção a crimes com a ajuda da gestão de informações registradas em sistema de informação, apresentando significativo impacto na redução da taxa de criminalidade. Um exemplo disso é o CompStat, utilizado pela prefeitura de Nova Iorque, que pode ser definido como uma técnica de gestão de processos orientada por metas, baseada na utilização de tecnologia computacional, estratégia operacional e responsabilidade gerencial para estruturar o modo como o departamento de polícia fornece serviços voltados para o controle da criminalidade **[2]**. Logo, o Compstat não é um método de mapeamento do crime por si só, mas um instrumento de gestão baseado em informação gerada pela tecnologia da informação. Sua utilização teve impactos significativos na experiência de Nova Iorque nos anos 90 e foi expandida para outros departamentos de polícia, dentro e fora dos Estados Unidos. O sistema Compstat caracteriza-se pela utilização de dados criminais com o mapeamento geográfico das áreas a serem policiadas. Esse instrumento está associado a técnicas como o mapeamento de "zonas quentes" (*hotspot mapping*), que são definidas a partir da utilização das estatísticas criminais **[3]**.

Exemplos como este nos encorajou a desenvolver este trabalho, onde buscamos analisar a relação existente entre as variáveis de ocorrência de crimes fomentando, assim, possíveis orientações à Segurança Pública no estado de Pernambuco, através da promoção de estratégias adequadas para prevenção e combate a crimes.

2 Trabalhos relacionados

Vários trabalhos vêm sendo desenvolvidos nesta temática nos últimos anos, um estudo para identificar o *modus operandi* do crime de roubo a transeuntes em Belém do Pará foi realizado em 2015 **[4]** e, nele, a partir de dados coletados nos registros dos boletins de ocorrência, foi possível realizar uma análise descritiva, identificando o modo como os assaltantes agem e o meio empregado para a locomoção no espaço geográfico por ocasião do cometimento do crime. Além disso, pôde-se analisar a variável temporal (horário e dia da semana) e o número de autores envolvidos.

No estudo **[4]**, foi feito um levantamento dos dados de 2011, 2012 e 2013 para observar se houve alguma alteração ao longo dos anos no comportamento do cometimento do crime de roubo, em termos da faixa de hora predominante. Em 2011, o delito ocorria sobretudo das 12h00 às 17h59; já em 2012 e 2013, a tônica se dava entre 18h00 e 23h59; e o ano de 2013 caracterizou-se pelo distanciamento entre as duas faixas de hora, fugindo do padrão dos anos anteriores. Este conjunto de informações revela a relação entre a luz do dia e a escuridão na preferência do delinquente no cometimento do crime de roubo.

O mesmo estudo também mostrou uma nítida distribuição da ocorrência do roubo ao longo dos dias da semana: em 2011 foi o sábado; em 2012, a sexta-feira; e em 2013, a quarta-feira **[4]**.

Um outro estudo semelhante **[5]**, realizado com variáveis de criminalidade em Teresina, Piauí, apontou os locais e horários onde há o maior índice de criminalidade, quais os principais crimes cometidos, causas e instrumentos utilizados. Também traça um perfil das vítimas e criminosos. Nesse estudo foi possível verificar a diferença entre os crimes cometidos contra homens e mulheres. Enquanto os homens são as principais vítimas de homicídios dolosos, as mulheres são as principais atingidas por lesões dolosas **[5]**.

3 Referencial Teórico

Os métodos utilizados para analisar dados em busca de relações existentes entre as variáveis diferem de acordo com o tipo de dado coletado. Neste trabalho, as variáveis analisadas são

categóricas, e, para descobrir possíveis dependências entre elas, foram utilizados dois tipos de testes de hipótese, o teste do *qui-quadrado* e o teste exato de Fisher.

Um teste de hipótese é um procedimento estatístico que permite tomar uma decisão (aceitar ou rejeitar a hipótese nula H_0) entre duas ou mais hipóteses (hipótese nula H_0 ou hipótese alternativa H_1), utilizando os dados observados de um determinado experimento **[6]**. Uma hipótese nula geralmente afirma que não existe relação entre dois fenômenos medidos, enquanto que a hipótese alternativa é contraditória à primeira.

3.1 Teste de Qui quadrado

O teste do *qui quadrado* pode não ser tão preciso quando os valores das amostras são muito pequenos, sendo, nesse caso, recomendado utilizar testes exatos. Por esse motivo, duas abordagens foram utilizadas neste trabalho, o teste de aproximação de *quiquadrado* e o teste exato de Fisher.

O teste do *qui quadrado* compara as frequências observadas com as frequências esperadas da amostra. As frequências observadas são obtidas diretamente dos dados das amostras, enquanto que as frequências esperadas são calculadas matematicamente a partir destas.

É preciso obter as duas estatísticas: o X^2 calculado e o X^2 tabelado. O X^2 calculado é obtido através da equação **(1)**. Já o χ^2 tabelado depende do número de graus de liberdade e do nível de significância adotado.

O teste é feito sobre uma tabela de contingência formada por (m) linhas e (n) colunas e o grau de liberdade é dado pelo produto de (m-1) x (n-1). O nível de significância representado por α é definido previamente, mas costuma-se adotar, por convenção, o valor de $\alpha = 0{,}05$. O nível de significância representa a máxima probabilidade de erro (valor-p) que se tem ao rejeitar uma hipótese. Em outras palavras, isso significa que os resultados experimentais que atingem esse nível de significância têm, no máximo, 5% de chance de ser resultado do mero acaso.

O teste calcula a relação: quadrado da diferença entre as frequências obtida (O_i) e esperada (O_e) em cada casa da tabela de contingência, dividido pela frequência esperada, e soma esses quadrados. O resultado dessa soma é o valor do X^2 encontrado **(1)**.

$$X^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \qquad \textbf{(1)}$$

Considerando o grau de liberdade calculado e o nível de significância pré-definido, consultando-se a tabela da distribuição de X^2, encontramos o valor crítico de X^2 e o valor p. Uma vez que o valor de X^2 encontrado é maior ou igual ao valor crítico, e, consequentemente, o valor-p é menor do que o nível de significância adotado, rejeita-se a hipótese nula. Ou seja, quando o valor obtido no teste estatístico excede o correspondente do valor crítico estabelecido, a hipótese nula deve ser rejeitada e aceita-se a associação entre as variáveis. Caso contrário, a hipótese nula não pode ser descartada.

3.2 Teste Exato de Fisher

O teste exato de Fisher é mais preciso do que o teste do *qui quadrado* quando os números esperados são pequenos. John H. McDonald recomenda que seja utilizado o teste exato de Fisher quando o tamanho total da amostra for inferior a 1000, e o *qui quadrado* seja utilizado para tamanhos de amostra maiores **[7]**.

O Teste Exato de Fisher se caracteriza por fornecer diretamente o valor-p, sem o uso de uma estatística intermediária.

O teste também pode ser realizado no caso de amostras grandes, porém, como envolve cálculos maiores e mais complexos que utilizam fatoriais, o que pode conduzir a números excessivamente elevados, necessita de maiores recursos do computador.

O teste exato de Fisher é baseado na distribuição hipergeométrica **(2)**:

$$P = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{N!\,A!\,B!\,C!\,D!} \qquad \textbf{(2)}$$

A probabilidade calculada será igual ao produto dos fatoriais dos totais marginais pelo fatorial do total geral multiplicado pelo inverso do produto dos fatoriais dos valores observados em cada classe.

http://dx.doi.org/10.25286/repa.v3i3.943

4 Metodologia

A análise realizada neste artigo utilizou cinco bases de dados da Secretaria de Defesa Social do Estado de Pernambuco[1]. Essas bases contêm ocorrências que foram registradas nas delegacias do estado no período de 01 de janeiro de 2017 a 30 de setembro de 2017, isto é, durante um período de nove meses, resultando em um total de 52.873 instâncias. São elas: CVLI (Crimes Violentos Letais Intencionais), CVP (Crimes Violentos contra o Patrimônio), Furto, Furto de Veículo e Roubo de Veículo.

O processo de análise envolveu:

- Verificar se existe relação significativa entre o dia da semana ser dia útil ou final de semana e a natureza do crime.
- Verificar se existe relação significativa entre o período do dia e a natureza do crime.
- Verificar se existe relação entre o trimestre e a natureza do crime.
- Verificar se existe relação significativa entre a área onde ocorreu o crime e a natureza do mesmo.
- Verificar se existe relação significativa entre a faixa etária da vítima e a natureza do crime para os casos de crimes violentos letais intencionais.
- Verificar se existe relação significativa entre o sexo da vítima e a natureza do crime para os casos de crimes violentos letais intencionais.

4.1 Pré-processamento

Para dar início à pesquisa, foi necessário realizar um trabalho de extração, transformação e carga (ETL – *Extract, Transform and Load*) nas bases para que a integração entre elas e suas análises fossem possíveis.

O processo de ETL envolveu a extração dos dados das bases obtidas, a transformação dos mesmos, incluindo higienização, padronização e categorização de variáveis, e, por fim, a integração das bases em uma base única. Para esta última etapa do ETL, foram selecionadas as variáveis mais significativas e em comum entre as bases.

As bases foram extraídas dentro da ferramenta Kettle, também conhecida como PDI (*Pentaho Data Integration*[2]). Nela, cada base passou individualmente por uma transformação. Uma variável representando um contador para os registros foi criada, além de mais outras duas variáveis categóricas, ambas derivadas da variável data. Uma delas representando o mês, e a outra, o dia da semana. As mesmas transformações foram reproduzidas para cada base.

Em seguida, as saídas das bases tratadas no Kettle continuaram o pré-processamento com mais algumas categorias sendo criadas. Dessa vez, utilizando o R[3], um ambiente de software livre para computação estatística e gráficos. As categorias trimestre e semana – indicando quando se trata de dia útil ou final de semana, mais especificamente – foram criadas, derivadas das variáveis recém-criadas mês e dia da semana, respectivamente. Também foi criada – apenas para a base CVLI – a categoria faixa etária, derivada da variável idade. A base CVLI era a única que continha as variáveis idade e gênero.

O pré-processamento no R gerou novas saídas, que foram utilizadas novamente no Kettle para realizar a integração.

A integração envolveu a junção de todas as cinco bases utilizando apenas as suas variáveis em comum, resultando em uma base única, conforme Figura 1.

[1] http://www.sds.pe.gov.br/
[2] http://www.pentaho.com/product/data-integration
[3] https://www.r-project.org/

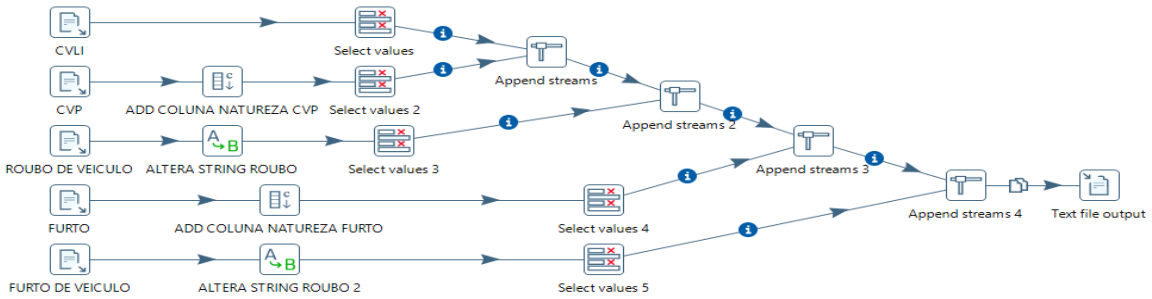


Figura 1 – Pré-processamento e integração das bases no PDI

4.2 Dicionário de dados

A base integrada, resultante da fase de ETL, contém nove variáveis, como mostra o Quadro 1:

Quadro 1 – Dicionário de Dados das bases integradas

AREA	Área onde ocorreu o crime, representada pela AIS – Área Integrada de Segurança, divisão territorial feita em Pernambuco para acompanhamento de ações e resultados. Existem, ao todo, 26 AIS em Pernambuco, mas as bases só envolvem as da cidade do Recife: AIS 1 – Santo Amaro, Boa Vista, Ilha Joana Bezerra e São José. AIS 2 – Espinheiro, Iputinga, Cordeiro, Madalena, Água Fria e Campo Grande. AIS 3 – Boa Viagem, Pina, Imbiribeira, Ibura, Brasília Teimosa AIS 4 – Várzea, Curado, Jardim São Paulo, Torrões e Afogados. AIS 5 – Apipucos, Guabiraba, Brejo da Guabiraba, Passarinho, Dois Unidos, Vasco da Gama e Alto do Mandú.
DATA	Informa a data e a hora em que ocorreu o crime.
DIA_SEMANA	Variável categórica derivada da variável DATA. Informa o dia da semana quando ocorreu o crime.
DIA_ÚTIL	Variável categórica derivada da variável DIA_SEMANA. Informa se o dia em que ocorreu o crime era dia útil ou final de semana.
MÊS	Variável categórica derivada da variável DATA. Informa o mês em que ocorreu o crime.
NATUREZA	Informa a natureza do crime (homicídio, roubo, latrocínio, etc.).
PERÍODO	Variável categórica derivada da variável DATA. Informa o período do dia em que o crime ocorreu (manhã, tarde, noite ou madrugada).
TRIMESTRE	Variável categórica derivada da variável MÊS. Informa o trimestre em que o crime ocorreu.
TOTAL	Contador de registros.

A base CVLI, além das variáveis mostradas na tabela acima, também contém três variáveis exclusivas, conforme o Quadro 2:

Quadro 2 – Dicionário de Dados da base CVLI

SEXO	Variável categórica que representa o sexo da vítima de crimes violentos letais intencionais.
IDADE	Informa a idade da vítima.
FAIXA_ETARIA	Variável categórica derivada da variável IDADE que representa a faixa etária da vítima, onde "Adolescente" representa a faixa de 12 a 18 anos, "Jovem Adulto" represente a faixa de 19 a 30 anos, "Adulto" representa a faixa de 31 a 49 anos, e "Idoso", a partir de 60 anos.

4.3 Análises Descritivas

4.3.1 Tabelas de Frequência

Inicialmente, utilizando a ferramenta R, foram construídas individualmente tabelas de frequência para todas as nove variáveis da base integrada e para as duas variáveis categóricas exclusivas da base CVLI.

A tabela de frequência auxilia na orientação da pesquisa, mostrando a distribuição de frequência das variáveis. A frequência de uma variável é o número de ocorrências ou repetições dessa variável.

Utilizando as funções *table()* e *prop.table()*, uma tabela de frequência foi construída para cada variável, assim como uma tabela de frequência relativa.

A tabela de frequência da variável que representa o sexo da vítima de crimes violentos letais intencionais, por exemplo, apontou para um

http://dx.doi.org/10.25286/repa.v3i3.943

percentual maior de vítimas do sexo masculino, com 92,19% de incidência.

Ainda na base de crimes violentos letais intencionais, a tabela de frequência construída em cima da variável que representa a faixa etária da vítima mostrou que as maiores vítimas são jovens adultos, com uma incidência de 55,41%, seguidos de adultos e adolescentes, com 30,45% e 13,48% respectivamente.

Todas as distribuições citadas até aqui foram construídas em cima da base CVLI. A seguir, serão apresentadas as tabelas de distribuição de frequências em cima da integração de todas as cinco bases (CVLI, CVP, ROUBO DE VEÍCULO, FURTO E FURTO DE VEÍCULO). A Tabela1, referente à natureza do crime, mostra que há um índice maior de casos de crimes violentos ao patrimônio, e o segundo lugar ficando com crimes de furto.

Tabela 1 - Distribuição de frequência relativa referente à natureza do crime

Natureza	Frequência
Cvp	56,05%
Furto	34,72%
Roubo de veículo	5,65%
Furto de veículo	2,45%
Homicídio	1,11%
Latrocínio	0,03%
Lesões corporais seguida de morte	0,00%

Como o período da base vai de janeiro a setembro do mesmo ano, a variável trimestre, sozinha, não indica algo muito significativo, já a tabela de frequência relativa da variável referente ao mês da ocorrência apresentou uma distribuição praticamente uniforme para todos os meses, de forma que, sozinha, na base integrada, ela não tem grande relevância.

Tabela 2 - Distribuição de Frequência relativa referente ao dia da semana

Dia da semana	Frequência
Sexta-feira	15,65%
Quarta-feira	15,05%
Terça-feira	14,74%
Quinta-feira	14,70%
Segunda-feira	14,47%
Sábado	14,04%
Domingo	11,34%

Pode-se observar na Tabela2 que ocorreram mais crimes nas sextas-feiras e nas quartas-feiras, com o domingo apresentando o menor percentual de ocorrências, coincidindo com o estudo realizado em Belém do Pará **[4]**, em que as sextas-feiras e as quartas-feiras apresentaram mais incidentes de crimes de roubo.

Consequentemente, a tabela de frequências da variável que categorizou o dia da semana em dia útil ou final de semana ratificou o resultado acima, revelando que os crimes ocorreram com mais frequência nos dias úteis do que nos finais de semana.

O estudo também identificou que, assim como em Teresina **[5]**, crimes de violência contra o patrimônio e homicídio são mais recorrentes no período da noite. Neste estudo, encontramos um resultado de 40,45% de incidência de crimes de CVP durante a noite e 8,81% no período da madrugada. Para homicídio, houve uma incidência de 42,56% no período da noite e 17,42% durante a madrugada, enquanto que o estudo realizado em Teresina apresentou uma incidência de 49% para crimes de furto e 55% para crimes de roubo - ambos são categorizados como crimes de violência ao patrimônio (CVP) - e uma incidência de 76% para crimes de homicídio, ambos no período de 18h às 03h (noite e madrugada).

Tabela 3– Distribuição de frequência relativa referente ao período do dia.

Período	Frequência
Noite	34,26%
Tarde	28,12%
Manhã	23,81%
Madrugada	13,81%

A distribuição de frequências relativas por período, conforme apresentado na Tabela 5, mostra que houve uma maior incidência de crimes gerais durante a noite.

Tabela 4 – Distribuição de frequência relativa referente à área da ocorrência.

Área	Frequência
1	24,29
3	23,25
4	20,23
2	19,05
5	12,20
99	0,99

Por fim, a distribuição por área (Tabela 4) indicou que a AIS 1 e a AIS3 apresentaram um percentual

maior de crimes. A AIS 1 é composta dos seguintes bairros: Santo Amaro, Boa Vista, Ilha Joana Bezerra e São José, e a AIS 3 é formada pelos bairros Boa Viagem, Pina, Imbiribeira, Ibura e Brasília Teimosa. A área 99 corresponde aos valores ausentes.

Este resultado pode incentivar um outro estudo que envolva o levantamento das características urbanas dos bairros do estado para detectar se esses bairros mais violentos possuem características que indiquem degradação urbana, negligência ou desmazelo. As características socioeconômicas dos bairros também devem ser levadas em consideração, para identificar se existe um alto índice de desemprego, pobreza, baixa escolaridade e outras características que podem ser reveladas potenciais para desencadear a violência.

5 Testes de relação

Analisando o Gráfico 1, é possível ratificar que existe uma incidência maior de crimes de violência ao patrimônio durante o período da noite.

Aplicando o teste de *qui quadrado* na tabela de contingência construída a partir dessas variáveis, "natureza" e "período", utilizando a função *chisq.test()* do R, obtivemos o valor do $X^2 = 3348.9$, com o grau de liberdade = 12 e o valor p *<0,0000000000000022*, isto é, muito abaixo do nível de significância *0,05*. Portanto, há um forte indício de que podemos rejeitar a hipótese nula e concluir que existe uma relação entre a natureza do crime e o período do dia em que ele ocorre.

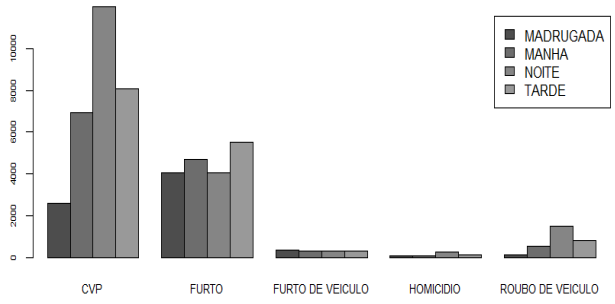


Gráfico 1 – Incidência de crime por natureza e por período

Para identificar se a possível relação que parece existir entre as variáveis "natureza" e "dia_util" no Gráfico 2, também aplicamos o teste de *qui-quadrado* na tabela de contingência das duas variáveis, obtendo um resultado de *233.51* para o X^2, com grau de liberdade = *4* e valor p ≈ *0,000*, indicando que podemos rejeitar a hipótese nula e constatar que existe relação entre essas duas variáveis.

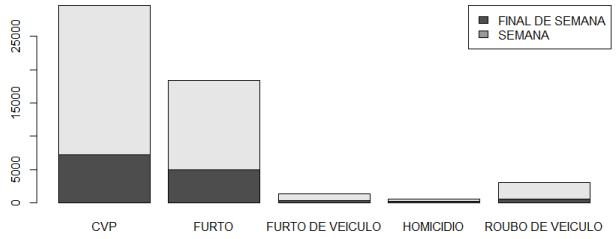


Gráfico 2 – Incidência de crime por natureza e por dia de semana (dia útil ou final de semana)

Ainda utilizando o teste de *qui quadrado*, vimos que a hipótese nula, que afirma não existir relação entre as variáveis que representam a natureza do crime e sexo da vítima não pode ser descartada, visto que o resultado do X^2 foi de *0.23795*, com um valor *p = 0.9935*, acima do nível de significância estabelecido, de 0,05. Portanto, não podemos afirmar que essas variáveis são dependentes.

Por meio do teste exato de *Fisher*, pôde-se concluir que a distribuição do número de homicídios difere (*p = 0,01212*) entre as faixas etárias das vítimas e o sexo, sendo maior para jovens adultos e adultos do sexo masculino, com 53% e 26,36% respectivamente. O Gráfico 3 ilustra bem essa diferença. Este resultado converge com o observado no estudo realizado em Teresina, Piauí.

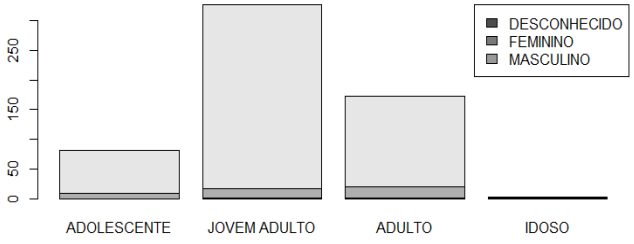


Gráfico 3 – Incidência de homicídios por faixa etária e por sexo

Analisando também a relação entre o período do dia e o dia da semana das ocorrências de crimes de homicídio isoladamente, vimos que, assim como o estudo em Teresina também apontou, eles seguem

http://dx.doi.org/10.25286/repa.v3i3.943

um padrão temporal, com uma maior incidência à noite e nos finais de semana, com o sábado à noite apresentando o maior número de ocorrências, 13,67%. Aplicando o teste de *qui quadrado*, pudemos constatar que existe uma relação entre essas duas variáveis (período do dia e dia da semana). O resultado do teste apresentou um $X^2 = 39.876$ e um valor $p = 0.00217$.

O Gráfico 4 mostra uma distribuição dos crimes por área e natureza, indicando que crimes contra o patrimônio (roubos e furtos) ocorrem mais na AIS 1, enquanto que crimes relacionados a veículos (roubos e furtos) ocorrem mais na AIS 4, formada pelos bairros Várzea, Curado, Jardim São Paulo, Torrões e Afogados, e AIS 3, formada pelos bairros Boa Viagem, Pina, Imbiribeira, Ibura e Brasilia Teimosa. Isso foi ratificado através do teste do *qui-quadrado*, que resultou em um p valor menor do que *0,05*.

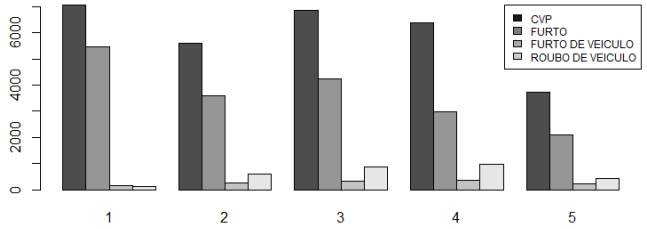


Gráfico4 – Incidência de crimes de CVP, furto ou roubo de veículo por AIS.

Tabela 5 – Distribuição de Frequência Natureza X AIS

	CVP	FURTO	FURTO DE VEICULO	ROUBO DE VEICULO
1	13,35%	10,35%	0,28%	0,19%
2	10,62%	6,77%	0,48%	1,14%
3	12,99%	8,01%	0,61%	1,64%
4	12,05%	5,64%	0,68%	1,85%
5	7,03%	3,94%	0,40%	0,82%

A Tabela 6 destrincha os crimes relacionados apenas a veículos (roubos e furtos) por período do dia e dia da semana. Nela pode-se observar que existe maior incidência nas noites de sexta e quarta-feira. Aplicando o teste de *qui quadrado*, obtivemos o resultado de 0.4866 para o valor p.

Tabela 6 – Distribuição de Frequência de roubos e furtos de veículos por período e dia da semana

	Madrugada	Manhã	Noite	Tarde
DOM	1,40%	1,87%	3,60%	2,27%
SEG	1,78%	2,69%	5,65%	3,79%
TER	1,82%	3,18%	6,61%	4,30%
QUA	1,89%	2,80%	7,31%	4,58%
QUI	1,73%	3,39%	6,82%	3,74%
SEX	1,85%	3,41%	7,34%	4,18%
SAB	1,61%	2,29%	4,77%	3,34%

A Tabela 7 representa uma síntese de todas as variáveis e suas respectivas significâncias encontradas através dos testes de relação.

Tabela 7 – Resultado do valor p nos testes de relação entre as variáveis

	natureza	periodo	dia_semana	area	sexo	faixa etaria	dia da semana (veículo)
natureza		2,2E-15		< 2,2e-16			
periodo	< 2,2e-16		0,00217				
dia_semana	0,9935	0,00217					
area	< 2,2e-16						
sexo						0,01212	
faixa etaria					0,01212		
Período (veículo)							0,4866

6 Análise e Discussões

Vimos nos resultados das análises que o período de maior ocorrência de crimes em geral é o da noite. Isso acontece possivelmente pela visibilidade ser menor com a pouca iluminação que existe à noite e por ter menos movimento nas ruas. Esse tipo de informação merece ser identificado para que haja uma melhor e mais estratégica distribuição do policiamento por horário. Outra possível solução também seria a readequação da iluminação da cidade de acordo com o tipo de movimentação de cada bairro.

Vimos também nos resultados das análises que alguns bairros são mais suscetíveis a crimes de roubos e furtos de veículos. Com poucos dados suficientes para inferir a causa dessa estatística, podemos apenas conjecturar que esses números podem estar relacionados ao fato de se tratarem de bairros residenciais compostos, em sua maioria, de casas, implicando em muitos carros estacionados nas ruas e moradores que precisam descer dos carros para abrir os portões de suas residências, ficando mais vulnerável e se tornando um alvo fácil para os criminosos.

A preferência por alvos de determinada faixa etária por parte dos criminosos poderia ter sua causa descoberta, caso existissem algumas outras variáveis, tais como condição social da vítima, a localização exata do crime e o sexo do criminoso. Uma base de dados mais completa e mais integrada poderia auxiliar nessa descoberta.

Vimos também que o tipo de ocorrência de crimes varia de acordo com a área. A base utilizada nesse estudo tem limitações de quantidade de variáveis a um período de tempo pequeno. Porém, as relações encontradas aqui podem ser utilizadas para alimentar estudos mais detalhados e mais aprofundados, e também para um planejamento estratégico da segurança segmentado por AIS, onde cada AIS teria um policiamento adequado.

Ainda em relação à distribuição de ocorrências por AIS, este estudo nos possibilitou observar que a AIS 1 possui maior ocorrência em todos os tipos de crime, e talvez isso comprove a teoria das janelas quebradas, sugerindo que crimes menores atraiam crimes mais graves. A partir da prevenção de pequenos crimes pode ser possível evitar que a área vá se transformando em um espaço cada vez mais violento. Para isso é necessário identificar os pontos e quais os crimes ali praticados, como o mostrado nesse estudo.

7 Considerações Finais

Com esses levantamentos, é possível elaborar um diagnóstico melhor do problema da violência no estado e começar a ter uma visão estratégica para solucioná-lo. A mudança do tipo de policiamento em determinadas áreas, ou o aumento do efetivo policial em determinados dias e horários, de acordo com os resultados dessas estatísticas, pode ser uma das soluções, juntamente com a instalação de câmeras de segurança em pontos estratégicos.

8 Contribuições Futuras

Cabe fazer algumas considerações relevantes para que haja uma melhor gestão das informações sobre os dados das ocorrências de crimes e delitos no Estado de Pernambuco.

Este estudo pode fomentar outros estudos que devem ser realizados envolvendo mais variáveis além das mostradas aqui, mas que não foram fornecidas pela Secretaria de Defesa Social por razões de sigilo da informação, como variáveis geográficas e variáveis relacionadas a dados urbanos, buscando possíveis relações entre tipos de crimes e características urbanas do local de ocorrência.

Como trabalhos futuros serão utilizadas técnicas de mineração de dados, com o objetivo de encontrar padrões nos crimes relatados, além da criação de uma ferramenta web para melhor visualização e análise dos resultados.

Este estudo e possíveis extensões do mesmo podem orientar a Segurança Pública na promoção de estratégias adequadas para a redução da criminalidade no Estado. Também podem fornecer subsídios para serem feitos posteriormente monitoramentos para identificar se houve real redução na criminalidade após a implementação de novos controles.

http://dx.doi.org/10.25286/repa.v3i3.943

Agradecimentos

À Secretaria de Defesa Social de Pernambuco, pelo fornecimento das bases de dados utilizadas neste trabalho.

Referências

[1] WILSON, James Q.; KELLING, George L. BROKEN WINDOWS: The police and neighborhood safety. **The Atlantic**, Mar. 1982. Disponível em: <https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/>. Acesso em: 30 set. 2017.

[2] WALSH, William F. Compstat: An analysis of an emerging police managerial paradigm. Policing: An International. **Journal of Police Strategies & Management**, v. 24, n.3, p. 347, 2001.

[3] RATCLIFFE, J.erry H. Crime mapping and the training needs of law enforcement**. European Journal on Criminal Policy and Research**, v. 10, n.1, p.10-65, 2004. Disponível em: <https://doi.org/10.1023/B:CRIM.0000037550.40559.1c>. Acesso em: 1 fev. 2018.

[4] CAVALCANTE, Lucidéia S.; ALMEIDA, Silvia dos S.; ARAÚJO, Adrilayne dos R. O *Modus operandi* do crime de roubo a transeuntes em Belém. **Planejamento e Políticas Públicas**, n. 47, jul/dez. 2016. Disponível em: <http://www.ipea.gov.br/ppp/index.php/PPP/article/view/614>. Acesso em: 30 jan. 2018

[5] SANTOS, Laura Castro de Carvalho dos. Violência e criminalidade: Um estudo dos dados existentes em Teresina - PI. **Âmbito Jurídico**, Rio Grande, v. 15, n. 99, abr 2012. Disponível em: <http://www.ambito-juridico.com.br/site/?n_link=revista_artigos_leitura&artigo_id=11448>. Acesso em: mar 2018.

[6] DÁVILA, Víctor Hugo L. **Teste de Hipóteses**. Instituto de Matemática, Estatística e Computação, UNICAMP. Disponível em: <https://www.ime.unicamp.br/~hlachos/Inferencia_Hipo1.pdf> p. 3. Acesso em: 15 mar. 2018.

[7] MCDONALD, John H. **Handbook of Biological Statistics**. 3rd ed. Baltimore: Sparky House Publishing, 2014. Disponível em: <http://www.biostathandbook.com/HandbookBioStatThird.pdf>. Acesso em: mar. 2018.

Business Intelligence para uma análise da qualidade da entrega dos objetos postais: um estudo de caso nos Correios de Alagoas

Jean Barros Teixeira[1] orcid.org/0000-0001-9485-6149
Mailson Melo dos Santos Filho[2] orcid.org/0000-0002-1711-5301
Carlos André Duarte Costa[3] orcid.org/0000-0002-7729-1120

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, PE, Brasil.
[2] Fábrica de Negócios – Analytics & Data Mining, Recife, PE, Brasil.
[3] Empresa Brasileira de Correios e Telegráfos (GERAE), Alagoas, Brasil.

E-mail do autor principal: jbt@ecomp.poli.br

Resumo

Atualmente as organizações necessitam de sistemas que forneçam suporte a tomada de decisão através de análises do grande volume de dados oriundos de seus sistemas transacionais. É o caso dos Correios que carece de uma solução que possa trazer vantagem competitiva no mercado em que atua. Dessa forma o objetivo desse trabalho é propor e implementar uma solução de Business Intelligence que irá auxiliar os gestores na análise da qualidade dos objetos postais e em uma tomada de decisão mais assertiva. Para isso foi utilizada a ferramenta *Open Source Pentaho* e técnicas que permitiram a criação de um *Data Mart*, realização de consultas analíticas e o desenvolvimento de um *dashboard*.

Palavras-Chave*:* *Pentaho*; *Business Intelligence*; *Data Warehouse*; *Data Mart*; Tomada de decisão;

Abstract

Currently, organizations need systems that support decision making through analysis of the large volume of data coming from their transactional systems. This is the case of the Post Office which offers a solution to create a competitive competitor in the market in which it operates. In this way the objective of this work is to propose and implement a Business Intelligence solution that is compatible with the managers in the analysis of the quality of the objects in a more assertive decision making. Pentaho and techniques that allows the creation of a Data Mart, conducting analytical consultations and development of a control panel.

Key-words: *Pentaho; Business Intelligence; Data Warehouse; Data Mart; Decision make;*

http://dx.doi.org/10.25286/repa.v3i3.94

Introdução

No contexto atual as organizações, públicas ou privadas, são pressionadas pelo mercado diariamente, devido a essa competitividade no mundo dos negócios o ecossistema organizacional sofre constante mudança. Com a finalidade de melhorar continuamente os produtos e serviços prestados aos seus clientes, o aprimoramento dos processos internos, e o aperfeiçoamento do conhecimento de negócio as organizações devem responder de uma forma célere às condições de mudança impostas pelo mercado, isto se torna um fator crítico de sucesso. Porém, as tomadas de decisões necessitam de grandes quantidades de dados, informação e conhecimento.

Os dados após serem transformados em informação torna-se um recurso primordial para auxiliar os gestores em uma melhor tomada de decisão. "A informação é considerada como um ingrediente básico do qual dependem os processos de decisão" **[1]**. Para isso a informação precisa ser disponibilizada com uma boa qualidade, confiável, na quantidade e no momento certo.

Assim como as demais organizações, os Correios também sofrem constantemente com a pressão exercida pelos concorrentes, mais precisamente no mercado de logística e encomendas já que o mesmo detém apenas o monopólio postal (entrega de cartas). Para melhorar sua competitividade, a empresa necessita acompanhar as tendências tecnológicas do mercado, utilizando sistemas capazes de coletar, organizar e publicar as informações necessárias para que os gestores possam tomar decisões condizentes com a realidade e mais assertivas.

Diante dessa necessidade de um sistema que possa auxiliar as organizações a lidar com os seus dados, diversos softwares e técnicas são desenvolvidos no mercado, como por exemplo, uma técnica chamada *Business Intelligence*. Uma solução de *Business Intelligence* (BI) proporciona aos gestores uma visão real da organização através de indicadores alinhados aos objetivos estratégicos do negócio. O BI permite coletar diversos dados isolados de várias fontes de dados da empresa, organizá-los e consolidá-los de forma que se possa extrair diversas informações úteis para o negócio.

Sendo assim o objetivo deste trabalho é a implementação de uma solução de Business Intelligence para auxiliar os gestores na realização de análises sobre a qualidade da entrega dos objetos postais através de um Dashboard de gestão operacional.

Revisão Bibliográfica

2.1 *Business Intelligence* (BI)

O BI ou em sua tradução, inteligência de negócio, é um processo que envolve diversos conceitos, metodologias, arquiteturas, tecnologias e infraestrutura, e tem como objetivo auxiliar as organizações em tomadas de decisões mais inteligentes e que agreguem valor ao negócio. Isso é possível pois através do BI pode-se coletar dados de diversas fontes, organizá-los, analisá-los e compartilhá-los com todos os tomadores de decisões.

De forma mais ampla, pode ser entendido como a utilização de variadas fontes de informação para definir estratégias de competitividade nos negócios da empresa. Podem ser incluídos nessa definição os conceitos de estruturas de dados, representadas pelos bancos de dados tradicionais, *data warehouse*, e *data marts*, criados objetivando o tratamento relacional e dimensional de informações, bem como as técnicas de data mining aplicadas sobre elas, buscando correlações e fatos "escondidos" **[2]**.

2.2 *Data Warehouse* e *Data Mart*

Data warehouse (DW) ou Armazém de dados é a nomenclatura utilizada para definir um repositório de dados históricos, relacional ou multidimensional, que serve aos interesses de todos os departamentos da organização **[3]**. Assim, o DW é projetado para armazenar de forma consolidada o grande volume de dados de uma organização e através da análise dos mesmos, extrair informações que respondam às necessidades do negócio.

Por sua vez, um *Data Mart* pode receber o mesmo conceito de um DW, a maior diferença entre eles é que o *Data Mart* é constituído por dados de uma mesma área da organização, por exemplo um *Data Mart* de Vendas. Um conjunto de *Data Marts*, ou seja, diversos conjuntos de dados específicos de cada área da organização formam o DW.

2.3 Online Analytical Processing (OLAP)

OLAP é considerado como um processo ou arquitetura que possibilita ao usuário uma análise profunda dos dados em diversos ângulos, geralmente através de interfaces gráficas que facilitam o manuseio do usuário. Têm o objetivo de trabalhar dados existentes, buscando consolidações em vários níveis, trabalhando fatos em dimensões variadas **[1]**. As funções básicas do OLAP são:

- visualização multidimensional dos dados;
- exploração;
- rotação;
- vários modos de visualização.

O OLAP e o DW trabalham de forma integrada, já que os dados armazenados de forma eficiente no DW são recuperados pela ferramenta OLAP com a mesma eficiência e rapidez. Desde a idealização do DW deve-se levar em consideração o que se deseja apresentar na ferramenta OLAP.

2.4 *Pentaho*

O *Pentaho* é uma suíte composta por diversos softwares voltados para a criação de soluções de BI de ponta a ponta. Existem soluções para diversas áreas desde a integração dos dados (ETL - *Extract, Transform e Load*), relatórios pré-formatados e ad hoc, análises *online* (OLAP - *Online Analytical Processing*), ferramentas para criação de dashboard, mineração de dados etc. Vale ressaltar que é uma ferramenta *open source*, dessa forma o cliente pode customizá-la de acordo com as necessidades do seu negócio.

A plataforma se divide em duas partes **[4]**.

O *Solution Engine* e seus componentes são responsáveis pela execução e controle das soluções. A base de seu funcionamento é uma máquina de *workflow* interna, que sequencia as chamadas de cada componente para o resultado desejado.

O Portal, a porção do *Pentaho* visível ao cliente final. Através dele o cliente navega entre as soluções e aciona a execução de qualquer recurso, como um relatório ou *dashboard*.

3 Preliminares

3.1 Metodologia Atual

Atualmente os Correios possuem um *Data Warehouse* com dados de diversos sistemas que a empresa possui, de todas as unidades espalhadas pelo Brasil. A ferramenta utilizada para acessar esses dados é o *MicroStrategy* em sua versão *web*. Geralmente o acesso a essa ferramenta é liberado para gestores de nível estratégico e para alguns gestores operacionais caso desejem.

Conforme a Figura 1, os dados são organizados por pastas para facilitar o seu acesso por parte dos usuários. Geralmente as pastas contêm dados de alguns sistemas ou de indicadores específicos da empresa.

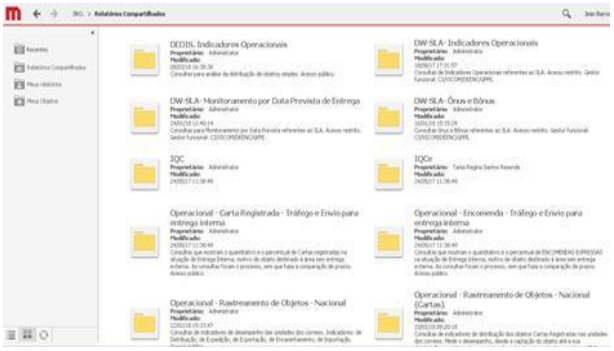


Figura 1 - Pastas de trabalho no *MicroStrategy*..

O sistema possui uma interface gráfica amigável que permite o usuário a realizar consultas de forma simples, escolhendo alguns parâmetros pré-determinados. Um dos problemas é o tempo de duração para se realizar uma consulta, pois devido a grande quantidade de dados de todo o Brasil algumas consultas

http://dx.doi.org/10.25286/repa.v3i3.94

demoram bastante. Na Figura 2 podemos ver como é gerado um relatório.

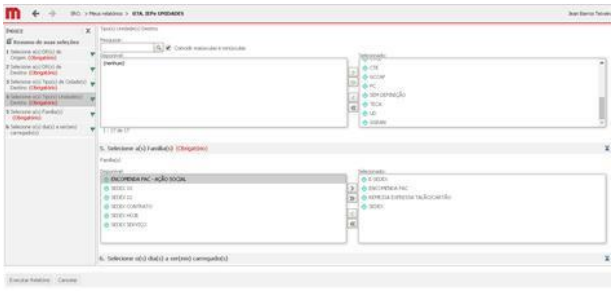


Figura 2 - Tela de geração do relatório.

Na Figura 3 podemos ver o resultado de um relatório criado na ferramenta. O usuário tem algumas opções, como adicionar mais alguns atributos disponíveis para agregar mais dados ao relatório, realizar exportação assim como gerar um gráfico.

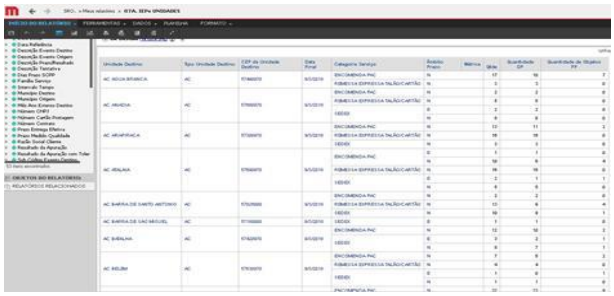


Figura 3 - Relatório gerado.

A Figura 4 apresenta um gráfico gerado pela ferramenta, um gráfico simples para uma análise rápida e algumas opções de customização do gráfico.

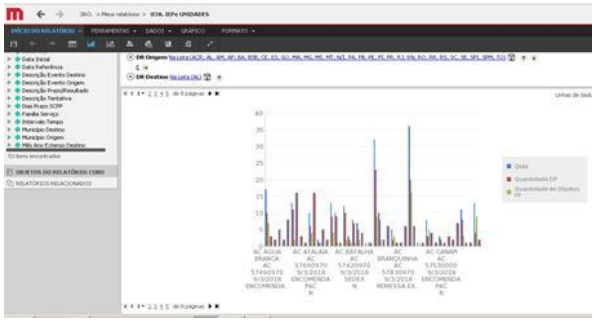


Figura 4 - Gráfico da consulta realizada.

Em resumo a ferramenta utilizada atualmente nos Correios ajuda os gestores a encontrarem os dados dos diversos sistemas em um único local, de forma simples, mas para se realizar análises bem mais profundas, cruzar dados ou até mesmo gerar *dashboards* que auxiliem na tomada de decisão o usuário precisa recorrer a outras ferramentas, exportar os dados e utilizá-los em planilhas eletrônicas, exigindo um pouco mais de habilidades técnicas. Além disso vale ressaltar que o tempo de duração das consultas costuma ser consideravelmente alto dependendo da sua complexidade.

3.2 Sobre os dados

Os dados foram obtidos do DW dos Correios através da ferramenta *MicroStrategy Web*, o relatório escolhido foi referente ao Índice de Entrega no Prazo - IEP e possui 60.825 registros de novembro/2017 a Março/2018. Tabela 1 mostra os atributos:

Tabela 1: Tabela com os principais atributos.

N	Atributo	Tipo	Valores
1	Unidade	Categórico	112
2	Tipo de Unidade	Categórico	3
3	Data final	Data	Nov/2017 a Mar/2018
4	Categoria Seviço	Categórico	5
5	Âmbito	Categórico	2
6	Quantidade de objetos entregues	Numérico	De 1 a 2280
7	Quantidade de objetos fora do prazo	Numérico	De 0 a 1402
8	Quantidade de objetos fora do prazo unidade	Numérico	De 0 a 1500

4 Metodologia proposta

Nesta seção será apresentada a proposta de uma solução de *Business Intelligence Open Source*, foi escolhida a suite *Pentaho* pois, além de ser uma ferramenta livre a mesma possui um conjunto de ferramentas robustas. A Figura 5 mostra a metodologia proposta.

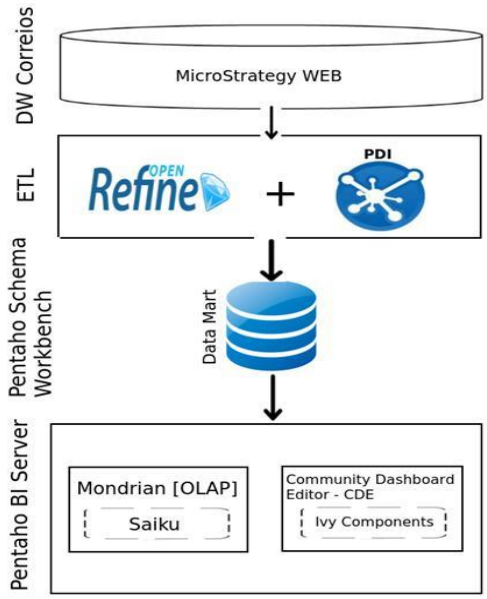


Figura 5 - Metodologia proposta

Na primeira camada, o DW dos Correios será acessado através do *MicroStrategy Web*, já que não se conseguiu acesso direto ao mesmo, e dele exportado um arquivo .CSV que será utilizado na etapa de ETL.

Na segunda camada, de ETL, utilizou-se duas ferramentas, a primeira foi o *Open Refine* pois apesar dos dados serem oriundos de um DW foi necessário alguns refinamentos nos mesmos. A outra ferramenta utilizada foi o *Pentaho Data Integration - PDI, que faz parte da suite* pentaho e foi utilizada para dar carga dos dados no *Data Mart*.

Na terceira camada através da ferramenta *Pentaho Schema Workbench* foi desenvolvido um esquema baseado na técnica de modelagem multidimensional para a criação do cubo OLAP utilizado posteriormente para consultas em tempo real e auxiliando no desenvolvimento do *Dashboard*.

Na quarta e última camada foi utilizado o *Pentaho BI Server* com algumas de suas ferramentas. Para realizar consultas no cubo de acordo com as necessidades apontadas pelo especialista, foi utilizada uma ferramenta OLAP denominada *Saiku*. Essas consultas serviram de base para o desenvolvimento de um *Dashboard*, através do *Community Dashboard Editor* - CDE e utilizando um *plugin* chamado *IvyDashboard* que possui alguns componentes que permitem uma melhor experiência do usuário.

5 Desenvolvimento e Resultados

Modelagem multidimensional

Para facilitar a modelagem multidimensional foram utilizadas as quatro etapas propostas por *Ralph Kimball* para analisar um processo de negócio, conforme ilustrado na Figura 7. No estudo de caso em questão foi definido como processo de negócio a qualidade da entrega (etapa 1), em seguida foi definido que o grão ou nível de detalhamento seria o dia (etapa 2), as dimensões (etapa 3) foram escolhidas para atender os questionamentos dos gestores: Quais unidades ofensoras por tipo de serviço? Qual o âmbito com mais perda de prazo? Quais tipos de unidades são as mais ofensoras? E por último foram definidos as métricas ou indicadores (etapa 4).

5.1 ETL

Esta fase é referente à extração, transformação e carga dos dados no *Data Mart* como mostra a Figura 6. O ETL foi feito com a ferramenta PDI, como entrada de dados temos um arquivo no formato csv que passou por um refinamento anteriormente através da ferramenta *OpenRefine*, foi realizado também uma higienização dos nomes dos atributos além de outras transformações e por fim temos a carga em todas as dimensões e na tabela fato.

http://dx.doi.org/10.25286/repa.v3i3.94

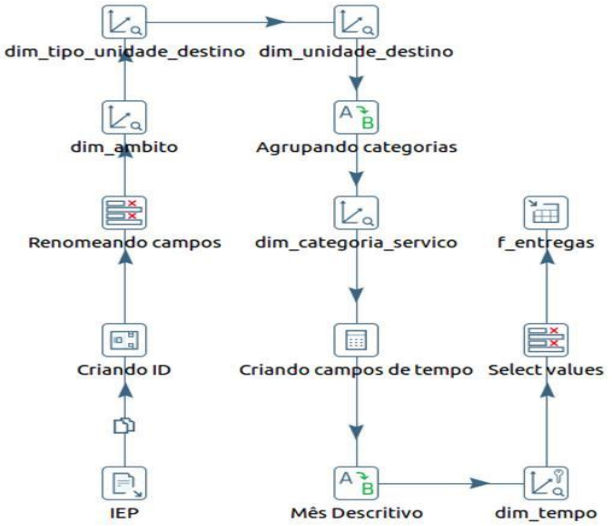


Figura 6 - ETL para criação do *data mart*.

5.4 Data Mart

O *Data Mart* foi desenvolvido baseando-se na modelagem multidimensional mencionada anteriormente, dessa forma os dados estão armazenados em várias dimensões (tabelas) que seguem um esquema denominado de estrela como ilustrado na Figura 7. O fluxo de processo do negócio realizado na primeira etapa de desenvolvimento facilitou na implementação do *Data Mart*.

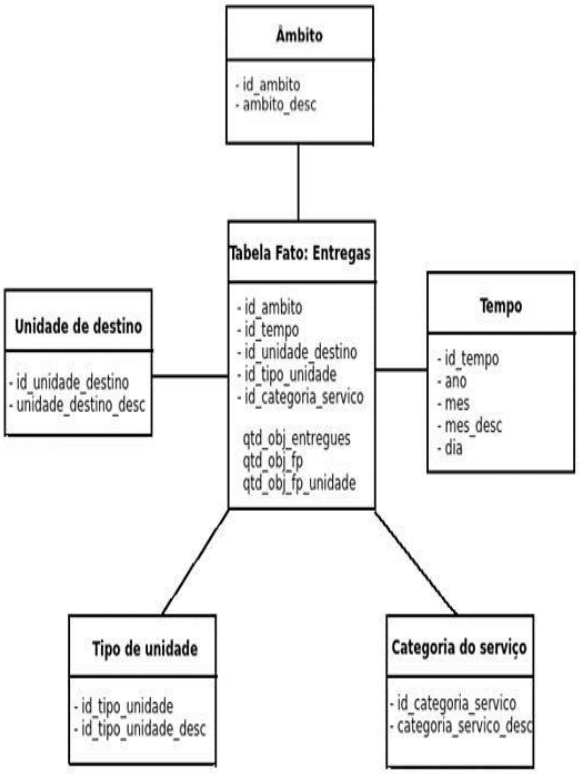


Figura 7 - Modelagem multidimensional na arquitetura em estrela.

5.4 OLAP

Após a criação do *Data Mart* foi possível realizar algumas consultas no cubo através da ferramenta Saiku que possui uma interface bem amigável permitindo manipular as dimensões e métricas de forma fácil e prática. A Figura 8 mostra o resultado de uma consulta simples criada a partir da dimensão tipo de unidade e da medida quantidade de objetos fora do prazo.

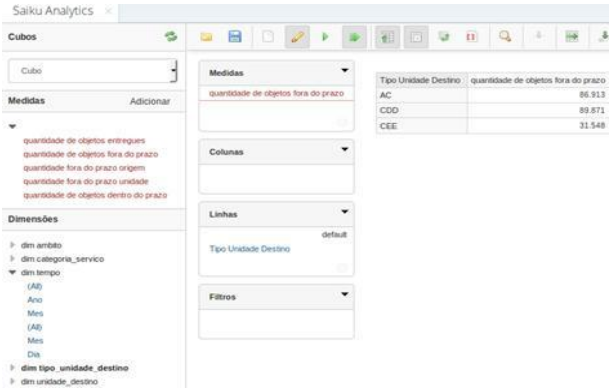


Figura 8 - Consulta OLAP utilizando o Saiku.

5.5 *Dashboard*

Um *dashboard* ou painel de bordo expõe uma visão ampla das informações mais importantes de uma organização, geralmente em uma única tela e permitindo uma interatividade por parte do usuário.

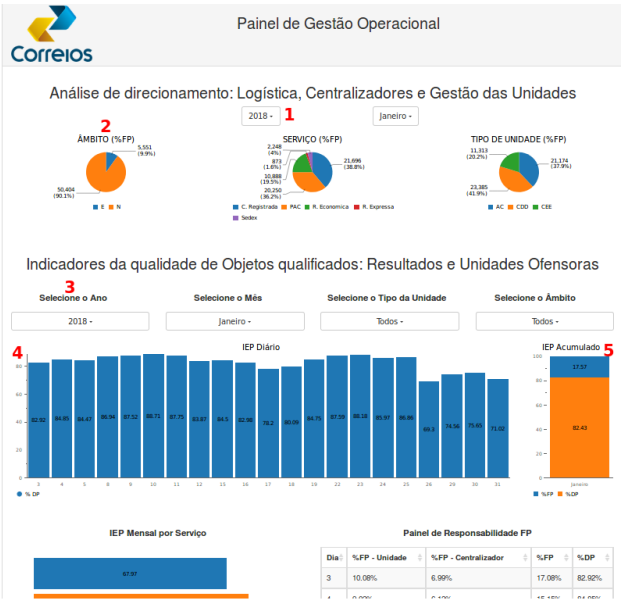


Figura 9 - *Dashboard* de gestão operacional.

Para desenvolver o *dashboard* proposto utilizamos o *Community Dashboard Editor* (CDE), uma ferramenta integrante da suíte *Pentaho* voltada para o desenvolvimento de *dashboards*. Também foi utilizado o *plugin Ivy,* pois ele possui alguns componentes que permitem tornar a interface gráfica mais atrativa e de fácil interação.

As figuras 9 e 10 apresentam o *dashboard* criado com o apoio de um especialista em estatística dos Correios que apontou quais informações eram relevantes para os usuários (Diretor, gestores operacionais, coordenadores etc). Cada componente do *dashboard* será descrito a seguir de acordo com seu respectivo número nas figuras.

Filtros (1): Existem dois filtros vinculados apenas aos gráficos da análise de direcionamento onde o usuário pode escolher o ano e o mês;

Gráficos direcionadores (2): Esses três gráficos de pizza tem o objetivo de alertar e direcionar a gestores do nível estratégico qual o âmbito (nacional ou estadual), serviço (PAC, Sedex etc) e tipo de unidade (AC, CDD ou CEE) são os ofensores em um respectivo mês do ano. Como por exemplo, na figura 11, um gestor estratégico poderia identificar que em Alagoas no mês de Janeiro a carta registrada (serviço) nacional (âmbito) nos CDD's (tipo de unidade) foi um ponto crítico e precisa ser verificada.

Filtros (3): Esses são o segundo conjunto de filtros vinculados a segunda parte do painel, o usuário além de poder escolher ano e mês, também tem a opção de escolher o tipo da unidade e o âmbito;

Índice de Entrega no Prazo – IEP diário (4): Este gráfico mostra o resultado diário do IEP, calculado pela razão entre os objetos entregues no prazo e o somatório de objetos entregues (dentro e fora do prazo).

IEP acumulado (5): Este gráfico foi adicionado, pois os resultados diários apresentam oscilações que muitas vezes não estão ligadas a um problema crônico. Pode ter ocorrido por um fator isolado (ausência elevada, quebra de veículo etc). Já resultado do mês (acumulado) mostra a real situação do indicador. O desvio causado por esse fator isolado é diluído na média do mês e não é uma variável significativa.

IEP mensal por serviço (6): Esse gráfico mostra o resultado acumulado mensal por tipo de serviço. Através dele pode-se verificar os problemas em um nível mais elevado já que podem ser verificados particularidades como logística, origem da carga, prazos, prioridades na entrega entre outros.

Painel de responsabilidade – FP (7): Nesta tabela são apontados os percentuais de objetos fora do prazo, com caracterização de responsabilidade, ou seja, se a perda do prazo foi da unidade ou do centralizador;

IEP das maiores unidades em quatidade de objetos (8): As 15 maiores unidades em quantidade de objetos qualificados (Sedex, PAC etc) representam 70% da carga do estado de Alagoas. Considerando a "lei dos poucos vitais" (80/20), 80% das consequências (quantidade de

http://dx.doi.org/10.25286/repa.v3i3.94

objetos total, dentro do prazo, fora do prazo etc) decorrem de 20% das causas (unidades).

Unidades ofensoras (9): Nesse gráfico são apontadas as unidades ofensoras, ordenadas pela quantidade de objetos fora do prazo. Assim como a tabela do item 8, o gráfico ajuda aos gestores a identificar unidades que causam grande impacto negativo nos resultados do estado.

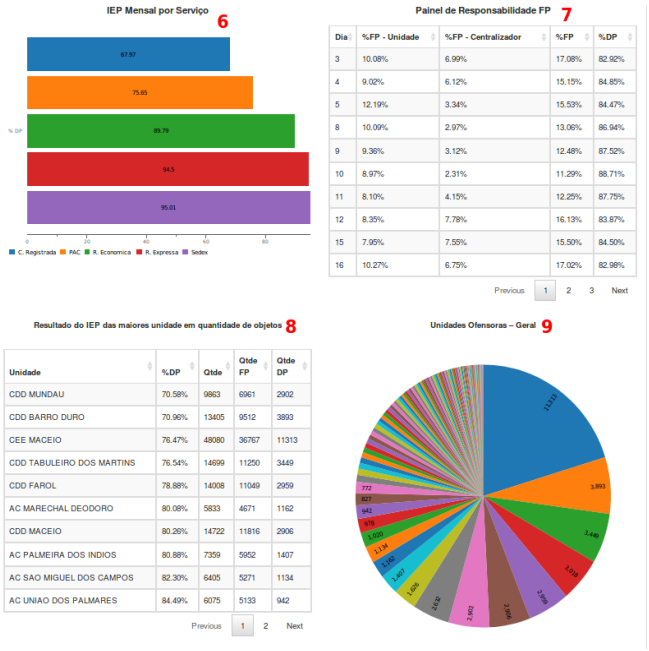


Figura 10 - *Dashboard* de gestão operacional.

6 Conclusão

Diante da necessidade de se obter vantagem competitiva e um suporte a tomada de decisão, utilizar um software que ajude nessa tarefa é primordial em uma organização. Dessa forma a solução de Business Intelligence proposta e implementada nesse trabalho conseguiu suprir as necessidades dos gestores e responder diversos questionamentos sobre o processo de negócio estudado.

Anteriormente era necessário utilizar-se de outros meios para realizar algumas análises com os dados provenientes do DW dos Correios. Além disso, o tempo de duração para gerar os relatórios é um pouco elevado. Com a nova solução, o usuário pode acessar um *dashboard* que contém diversas informações relevantes em um único local e com uma interface fácil de utilizar.

Conclui-se que o objetivo do trabalho foi alcançado visto que os gestores após análise do *dashboard*, conseguem a partir de suas conclusões, gerar planos de ações e tomadas de decisão mais assertivas para melhorar a qualidade da entrega dos objetos postais.

Agradecimentos

Agradeço a Empresa Brasileira de Correios e Telegráfos pelo custeio de 80% do curso.

Referências

[1] BARBIERI, Carlos. **BI2 - Business Intelligence**: modelagem e qualidade. Rio de Janeiro: Elsevier, 2011. p.392. P.

[2] BATISTA, Cleisson Fabrício Leite et al. Proposta de *data mart* para análise de faturamento de empresa de varejo utilizando *software* livre. **Revista Brasileira de Administração Científica**, v.3, n. 2, p.163–180, 2012.

[3] CAUTELA, A. L.; POLIONI, F. G. F. S**istemas de informação**. São Paulo: Livros Científicos e Técnicos, 1982.

[4] DIAS, Jorge Luis Ferreira. Pentaho - BI: conhecendo a plataforma, arquitetura e infraestrutura. **Devmidea**, 2014. Disponível em <https://www.devmedia.com.br/pentaho-bi-conhecendo-a-plataforma-arquitetura-e-infraestrutura/31502> Acesso em:10 mar. 2018.

[5] Arquitetura genérica de Business Intelligence. Disponível em: <www.it4biz.com.br> Acesso em: 12 mar. 2018.

[6] Conceitos sobre Pentaho. Disponível em: <http://www.pentaho.com/> Acesso em: 20 mar. 2018.

Aplicação de Algoritmos de Clusterização em uma Base de Dados de Reservas de Hotéis

Application of Clustering Algorithms in Hotels Reservation Datasets

Pedro Alexandre de Araújo Aguiar[1] orcid.org/0000-0002-9973-763X
Clodomir Joaquim de Santana Junior[1] orcid.org/0000-0001-7869-7184
Carmelo José Albanez Bastos Filho[1] orcid.org/0000-0002-0924-5341

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Pernambuco, Brasil.

E-mail do autor principal: paaa@ecomp.poli.br

Resumo

Este artigo faz uma análise da aplicação dos algoritmos de clusterização K-Means e Fuzzy C-Means. O estudo de caso visa identificar perfis de clientes de uma agência de viagens online, com o objetivo de melhorar a eficácia do envio de ofertas através de e-mail marketing, possibilitando o envio de anúncios personalizados para cada perfil. O processo de clusterização foi feito baseado na similaridade entre os usuários, levando em conta 13 característicasextraídas das vendas dos clientes.O resultado mostra que, apesar de chegaram a grupos parecidos, o K-Means teve desempenho levemente superior ao Fuzzy C-Means, no que diz respeito a avaliação através da métrica de estatística Gap.

Palavras-Chave: Clusterização;K-Means; Fuzzy C-Means;

Abstract

This paperanalyzes the application of K-Means and Fuzzy C-Means clustering algorithms.The case study aims to identify customer profiles of an online travel agency, with the objective of improving the effectiveness ofemail marketing campaigns, allowing to send personalized advertisements for each profile.The clustering process was based on similarity among users, considering 13 characteristics extracted from customer sales.The result shows that although they obtained similar groups, the K-Means performed slightly better than Fuzzy C-Means, considering the evaluation through the metric of Gap statistics.

Key-words:*Clustering;K-Means; Fuzzy C-Means;*

http://dx.doi.org/10.25286/repa.v3i3.945

1 Introdução

Uma das técnicas utilizadas durante a mineração de dados é a chamada clusterização (ou agrupamento). O objetivo da clusterização é segmentar determinado conjunto de dados em subgrupos de acordo com similaridades encontradas dentro da base de dados **[1]**. Geralmente essa tarefa é executada de forma não-supervisionada, isto é, sem interferência humana, já que a classificação das amostras é desconhecida.

Um dos algoritmos de clusterização mais tradicionais é o *K-Means*. Ele é bastante difundido devido à sua eficiência e simplicidade, entretanto, o mesmo apresenta alguns problemas quando aplicado em bases de dados mais complexas, sendo os principais: a tendência em convergir para os ótimos locais e o fato de que a escolha dos centroides iniciais interfere bastante na qualidade do agrupamento **[2]**. Tentando corrigir esses problemas, diversos algoritmos alternativos foram sugeridos, dentre eles o *Fuzzy C-Means* (FCM). O FCM faz uso da lógica de agrupamento *fuzzy,* isto é, o conceito de que determinada amostra pode não pertencer a somente um grupo (como no *K-Means*), mas sim a diversos grupos, cada qual com seu grau de pertinência **[3]**.

Este artigo aborda os dois algoritmos. Será feito umestudo de caso utilizando uma base de dados de 2.959 reservas de hotéis de uma agência de viagens online do Brasil durante os anos de 2016 e 2017. A ideia é agrupar os clientes dessas reservas de acordo com similaridades de seus hábitos de compras, utilizando as duas técnicas (*K-Means* e FCM), fazendo um comparativo entre elas, além de realizar uma análise nos agrupamentos de melhor qualidade.

O artigo foi organizado da seguinte maneira: a Seção 2 irá abordar os tipos de clusterização dos algoritmos aqui utilizados,além de apontar alguns trabalhos relacionados. Já na Seção 3 serão detalhados os algoritmos *K-Means*e*Fuzzy C-Means*. A Seção 4 apresenta os detalhes da base de dados utilizada como estudo de caso e a Seção 5 apresenta os resultados. Para finalizar, a Seção 6 apresenta as conclusões sobre estudo de caso e possíveis aplicações futuras.

2 Fundamentação Teórica

2.1 Tipos de Clusterização

A literatura aborda algumas técnicas de clusterização **[4]**, neste artigo vamos apresentar os dois tipos que se aplicam aos algoritmos utilizados no estudo de caso, são eles: particional e *fuzzy*.

O agrupamento particional (do qual o *K-Means* faz parte) tem por objetivo dividir as amostras em grupos (*clusters*, em inglês) que tenham alto grau de similaridade entre seus elementos, e alto grau de separação entre elementos de *clusters* diferentes. Além disso, num algoritmo do tipo particional, cada instância só pode estar atribuída a um *cluster*.

No agrupamento do tipo*fuzzy*, o algoritmo busca dividir as amostras em grupos que podem se sobrepor, isto é, determinada amostra pode pertencer a mais de um cluster. Dada essa natureza, cada amostra apresenta um grau de pertinência em relação a determinado grupo. Caso desejado, um algoritmo *fuzzy* pode ser utilizado para gerar um agrupamento particionalatribuindo determinada amostra ao grupo em que a mesma apresentar o maior grau de pertinência.

2.2 Trabalhos Relacionados

A tarefa de tentar conhecer melhor o comportamento dos passageiros para melhorar a qualidade das indicações de hotéis e pacotes de viagem foi abordada por outros autores. A seguir serão destacados alguns trabalhos que têm essa linha de estudo.

Trabalhos já foram desenvolvidos com o intuito de sugerir pacotes de viagem aos usuários utilizando diversos agentes queinteragem entre si, para buscar as melhores opções de voo, estadia e atrações **[5] [6]**.

http://dx.doi.org/10.25286/repa.v3i3.945

Outros autores avaliaram a utilização de inteligência artificial na área de turismo **[7]**, propondo novos modelos para sistemas de recomendação **[8]** e também elencando ferramentas que já existem com esse propósito.

Também foram encontrados trabalhos que avaliaram técnicas de mineração de dados (filtragem colaborativa e regras de associação) aplicadas ao setor de turismo **[9]**.

Apesar de nenhum dos trabalhos mencionados abordar exclusivamente o tema específico de análise de algoritmos de clusterização numa base de dados de turismo, todos trazem tópicos relevantes no sentido do uso da inteligência artificial para melhorar o relacionamento com os clientes.

3 Algoritmos de clusterização

3.1 *K-Means*

O K-Means é um algoritmo particional proposto porMacQueen em 1967 **[10]**, que é bastante popular por sua simplicidade e eficiência. A ideia do algoritmo é dividir a base de dados em K grupos que tenham instâncias semelhantes, considerando uma medida de similaridade. A similaridade entre duas amostras é medida utilizando uma função de distância, que geralmente é a distância euclidiana. A função de distância euclidiana entre duasamostras x_i e x_j, ambas de dimensão *d (*a dimensão é a quantidade de características de uma amostra), é dada por:

$$d(x_i, x_j) \sqrt{\sum_{p=1}^{d} (x_{i,p} - x_{j,p})^2} \quad \textbf{(1)}$$

Um dos passos mais importantes do *K-Means* é a atualização dos centroides ao fim de cada iteração, através do cálculo das novas médias de cada característica do mesmo. Esse cálculo é feito utilizando a média de cada característica para cada elemento de um determinado. O cálculo é definido através da fórmula:

$$C_k \frac{1}{n_k} . \sum_{i=1}^{n_k} x_i^k \quad \textbf{(2)}$$

Onde C_k representa o centroide do grupo K, en_ko total de amostras presentes no cluster K. Os passos se repetem até que determinado critério de parada seja atingido (por exemplo, quando não houver mais mudança em nenhum dos grupos, ou caso uma quantidade máxima de iterações seja atingida). O Algoritmo 1apresenta o pseudo-algoritmo do *K-Means*.

Algoritmo 1: *K-Means*
Entrada: base de dados com *i* instâncias e *d* dimensões, *K* grupos desejados
Saída: Base de dados dividida em *K* grupos

```
1 início
2 inicializa os K centroides com valores aleatórios;
3 enquanto critério de parada não atingido;
4   para cada amostra x_i faça
5     Adicione x_i ao grupo do centroide C_k de menor distância, de acordo com a equação (1);
6   fim
7   Atualiza os centroides C_k de acordo com a equação (2);
8 fim
```

3.2 *Fuzzy C-Means* (FCM)

O FCM foi introduzido em 1984 por Bezdek **[11]**, como uma extensão do C-Meansparticional. O objetivo do algoritmo é encontrar clusters *fuzzy* para determinado conjunto de dados. A lógica *fuzzy* nesse algoritmo diz respeito ao fato de que, para determinado elemento da base de dados, o algoritmo irá encontrar graus de pertinência para cada cluster, ou seja, um elemento pode pertencer a mais de um cluster.

Para atingir seu objetivo, o FCM segue alguns passos: inicialmente, definimos a quantidade cde clusters desejados, sendo2 ≤ c ≤ n, sendo n a quantidade de amostras.Definimos também o valor do coeficiente de "fuzzificação" m, e iniciamos a matriz de pertencimento $U^{(0)}$ com valores de pertencimento aleatórios. Depois disso, seguimos os seguintes passos:

1. Calculamos os centróides de cada grupo, de acordo com a equação:

http://dx.doi.org/10.25286/repa.v3i3.945

$$v_{ij} = \frac{\sum_{k=1}^{n}(\mu_{ij})^m x_i}{\sum_{k=1}^{n}(\mu_{ij})^m} \quad \textbf{(3)}$$

2. Calcula a distância euclidiana D_{ij}, utilizando a equação (1) de cada ponto ipara cada centroidej.
3. Atualiza os valores μ_{ij} da matriz de pertencimento U, seguindo a equação:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}(\frac{D_{ij}}{Dkj})^{\frac{2}{m-1}}} \quad \textbf{(4)}$$

O algoritmo executa esses passos até que o módulo da diferença entre duas matrizes de pertencimento U^k e U^{k-1} seja menos que o coeficiente de erro ε definido pelo usuário. Essa condição é definida na equação:

$$\|U^k - U^{k-1}\| < \varepsilon \quad \textbf{(5)}$$

Outros critérios de paradas também podem ser adotados, como por exemplo: atingir um número máximo de interações. Sendo assim, poderíamos definir o pseudo-algoritmo do *Fuzzy C-Means* da seguinte maneira:

Algoritmo 2: *Fuzzy C-Means (FCM)*

Entrada: coeficiente de "fuzzificação" m, *c* grupos desejados, base de dados com i instâncias, ε erro aceitável.

Saída: Matriz de pertencimento M com c colunas e i linhas, onde o elemento M_{ic} define o grau de pertencimento do elemento i no conjunto c.

1. **início**
2. inicializa a matriz de pertencimento M com valores aleatórios entre 0 e 1;
3. **repita**;
4. cacula os centroides de acordo com a equação (3), para cada cluster c.
5. cacula a distância euclidiana de cada instância, para cada centroide de acordo com a equação (1)
6. Atualiza a matriz de pertinência de acordo com a equação (4);
7. **até que** equação (5) seja verdadeira;
8. **fim**

4 Base de Dados de Reservas de Hotéis

Uma agência de viagens online, possui um *site* de reservas de hotéis, onde oferece hospedagens para hotéis em todo o Brasil.Os clientes, em sua grande maioria, são atingidos através de disparos de e-mail marketing com ofertas determinadas de maneira arbitrária pelo departamento de marketing da agência, sem nenhum tipo de filtro sobre os destinatários da oferta. Essa abordagem acarreta um grande custo com os envios do e-mail marketing, uma vez que o custo é determinado pela quantidade de e-mails destinatários.

Visando melhorar a eficiência do uso do dinheiro investido em e-mail marketing, foi pensando em realizar uma classificação dos usuários do *site*, com base em seu histórico de compras, com o objetivo de identificar perfis de clientes e, dessa maneira, enviar oferta de maneira mais eficaz, ou seja, apenas aos usuários com mais probabilidade de se interessar pelo anuncio.

Para realizar o agrupamento dos usuários, foram selecionadas 2.959 vendas realizadas pelo site da agência. Essas vendas foram escolhidas devido ao fato de ser o total de vendas aprovadas durante os anos de 2016 e 2017, já os foram

http://dx.doi.org/10.25286/repa.v3i3.945

escolhidos pois a empresa focou nas vendas online de viagens a partir de 2016.

As amostras são compostaspor 13 características, descritas na Tabela 1. Todos os valores das amostras foram normalizados para uso nos algoritmos.

Tabela 1 - Características de cada elemento da amostra

Característica	Descrição
Uf	Estado onde o cliente reside
Idade	Idade do clientes
Total Comprado	Soma total de todas as compras do cliente
Ticket Médio	Valor médio de compra do cliente
Quantidade de Compras	Quantidade de compras do cliente
Tipo de Pagamento	Forma que o cliente pagou a compra (Cartão de crédito, boleto)
Quantidade de Parcelas	Quantidade de parcelas que o cliente escolheu dividir
Cidade	Cidade do cliente
Valor da venda	Valor da venda específica
Hotel	Hotel comprado pelo cliente
Destino	Cidade de destino da viagem
Mês de estadia	Mês que o cliente escolheu se hospedar
Quantidade de diárias	Total de dias que o cliente irá ficar hospedado

Tabela 2 - Resultado da simulação dos algoritmos utilizando 30 simulações e condição de parada 1.000 iterações.

Algoritmo-K	Estatística GAP	Distância Inter-cluster	Erro Quantizado	Distância Intra-Cluster
K-Means-2	0.14(0.05)	0.58(0.03)	0.53(0.008)	1595.98(8.13)
K-Means-3	**0.25(0.07)**	2.14(0.08)	0.50(0.014)	1480.49(28.91)
K-Means-4	0.31(0.07)	4.77(0.32)	0.47(0.010)	1407.52(10.05)
K-Means-5	0.37(0.06)	8.49(0.51)	0.46(0.013)	1354.52(10.16)
K-Means-6	0.44(0.07)	13.81(0.66)	0.45(0.011)	1306.17(15.31)
K-Means-7	0.48(0.07)	19.76(0.98)	0.44(0.006)	1264.06(14.54)
K-Means-8	0.49(0.08)	27.84(0.90)	0.43(0.006)	1233.99(18.03)
K-Means-9	0.54(0.09)	36.94(1.67)	0.42(0.005)	1205.38(12.19)
K-Means-10	0.56(0.07)	**47.34(2.19)**	**0.41(0.004)**	**1183.41(6.33)**
FCM-2	0.12(0.04)	0.05(0.014)	0.569(0.005)	1690.90(21.83)
FCM-3	**0.21(0.05)**	0.24(0.002)	**0.564(0.006)**	1645.34(2.00)
FCM-4	0.19(0.04)	0.55(0.003)	0.575(0.017)	1624.03(1.51)
FCM-5	0.20(0.06)	0.91(0.003)	0.581(0.030)	1615.03(1.33)
FCM-6	0.17(0.06)	1.30(0.004)	0.563(0.018)	1611.44(1.75)
FCM-7	0.16(0.06)	1.72(0.004)	0.564(0.021)	**1609.54(1.65)**
FCM-8	0.17(0.07)	2.17(0.004)	0.597(0.020)	1609.94(2.10)
FCM-9	0.13(0.07)	2.65(0.004)	0.585(0.036)	1610.99(2.36)
FCM-10	0.12(0.07)	**3.17(0.005)**	0.593(0.030)	1612.85(3.36)

http://dx.doi.org/10.25286/repa.v3i3.945

4.1 Pré-processamento dos Dados

Com o intuito de melhorar a qualidade dos dados trabalhados, alguns pré-processamentos foram feitos em algumas características. Todos os tratamentos foram feitos utilizando a ferramenta de códigoaberto *Open* Refine **[12]**. As características que sofreram algum tipo de manipulação foram:

1) Idade: característica extraída através da data de nascimento do cliente.
2) Cidade: como no site da agência essa característica é um campo aberto para digitação do usuário, foi necessário padronizar o nome das cidades para evitar que registros da mesma cidade aparecessem com valores distintos.
   Exemplo: São Paulo, São Paulo e Sampa foram unificados para São Paulo, e assim por diante.
3) Destino: alguns hotéis de um mesmo destino constavam como destinos diferentes. Exemplo:

   dois hotéis A e B ficam em Porto de Galinhas/PE, mas o hotel A tinha como destino Ipojuca/PE. Como Porto de Galinhas é uma praia de Ipojuca, foram unificados para Porto de Galinhas (sempre unificamos para o nome que o destino é mais conhecido).
   Mês de estadia: característica foi derivada com base no dia da entrada do cliente no hotel, isto é, se a estadia de um cliente começou em 01/04/2018, o mês de estadia foi abril, mesmo que a estadia tenha se estendido por meses distintos.

5 Experimento e Resultados

O estudo de caso foi realizado utilizando a seguinte estratégia: como ambos os algoritmos têm como entrada uma quantidade K de clusters desejados, foram realizadas 30 execuções do algoritmo, para cada quantidade K de clusters, com o K variando de 2 a 10 grupos. Posteriormente, foi realizada uma média de 4 métricas para cada valor de K relativas às 20 execuções, de maneira que avaliando essas métricas, fosse possível definir qual o melhor K a ser escolhido para cada algoritmo. O resultado consolidado das execuções está presente na Tabela 2.

As métricas utilizadas para avaliar os grupos foram as seguintes: Estatística GAP, Distâncias Intra e Inter-cluster e o Erro quantizado. A seguir explicamos cada uma dessas métricas e, posteriormente, iremos avaliar os grupos que tiveram a melhor estatística GAP para ambos os algoritmos, pois essa métrica se mostra eficiente na decisão de escolher a melhor quantidade de grupos **[13]**.

5.1 Descrição das métricas

Estatística GAP: é uma métrica proposta por Tibshirani **[13]**, em 2001, que tem por objetivo encontrar o número ideal de clusters K. Para escolha do K ideal, são avaliados os diversos valores da métrica (que faz uso do logaritmo da distância intra-cluster) para os diversos valores de K. Essa mesma análise também é feita para um conjunto de dados aleatórios. A estatística GAP representa justamente a diferença entreo valor encontrado para a amostra real em relação à amostra aleatória. Por isso o valor dessa métrica deve ser maximizado, mostrando que o agrupamento escolhido se difere de um agrupamento aleatório. Para o cálculo dessa métrica, utilizam-se as seguintes equações:

$$D_k = \sum_{\forall xi,xj \in Ck} d(x_i, x_j) \quad \textbf{(6)}$$

Onde d(x_i,x_j) é a distância euclidiana entre os pontos. Posteriormente, utilizamos a equação (6) para encontrar a dispersão entre valores crescentes de K, através da equação:

http://dx.doi.org/10.25286/repa.v3i3.945

$$W_k = \sum_{i=1}^{k} \frac{1}{2n_r} D_i \quad \textbf{(7)}$$

Uma vez que temos a dispersão interna dos grupos, utiliza-se a versão amortizada logW_k com o valor da mesma métrica para uma amostra de dados aleatória. Sendo assim, a estatística GAP tem por objetivo maximizar o valor da equação:

$$Gap_n(\mathrm{k}) = E_n^*(\log W_k) - \log W_k \quad \textbf{(8)}$$

Onde E_n^* (log W_k)representa o valor da métrica esperado para uma amostra aleatória. Distância Intra-Cluster: é utilizada para validar a distância entre dois elementosx_i e x_j,pertencentes ao mesmo grupoC_k. Sendo assim, um bom valor para essa métrica é um valor baixo, indicando proximidade entre os elementos do cluster. Ela é medida conforme a equação:

$$D_{intra} = \sum_k \frac{1}{2N_k} \sum_{xi,xj \in Ck} d(x_i, x_j) \quad \textbf{(9)}$$

Distância Inter-Cluster: é utilizada para validar a distância entre dois centroides C_k e $C_{K'}$. Sendo assim, um bom valor para essa métrica é um valor alto, indicando que os clusters estão bem separados. É calculada através da equação:

$$D_{inter} = \sum_{\forall k,k' | k \neq k'} d(c_k, c_{k'}) \quad \textbf{(10)}$$

Erro Quantizado:é utilizada para medir a eficiência do algoritmo para valores crescentes de K. Leva em consideração a distância euclidiana de cada ponto x_i ao seu centroide c_k em relação ao total de amostras pertencentes ao cluster C_k , referenciado na fórmula como $|C_k|$. Sendo assim, o valor ideal para essa métrica são valores baixos. Segue a equação:

$$J_e = \frac{\sum_k \sum_{\forall x_i \in C_k} d(x_i, c_k)/|C_k|}{N_k} \quad \textbf{(11)}$$

5.2 Resultados *K-Means*

Os resultados da execução do *K-Means*, estão exibidos na Tabela 2. Através dela, podemos perceber que o algoritmo conseguiu cumprir o que se esperava dele, ao maximizar a distância inter-cluster e minimizar a distância intra-cluster, conforme o numero de K aumenta. Também podemos perceber que o algoritmo também conseguiu diminuir a métrica do erro quantizado, que também leva em consideração a qualidade interna dos grupos. Analisando os valores da estatística Gap, também percebemos sucesso ao maximizar o valor da métrica. Nas figuras 1, 2, 3 e 4 podemos ver a variação das métricas das distâncias intra e inter cluster, o erro quantizado e a estatística gap, respectivamente.

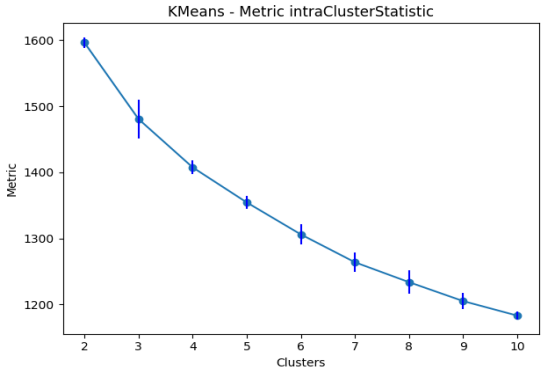


Figura 1 - Variação da distância intra-cluster para valores crescentes de K, utilizando K-Means.

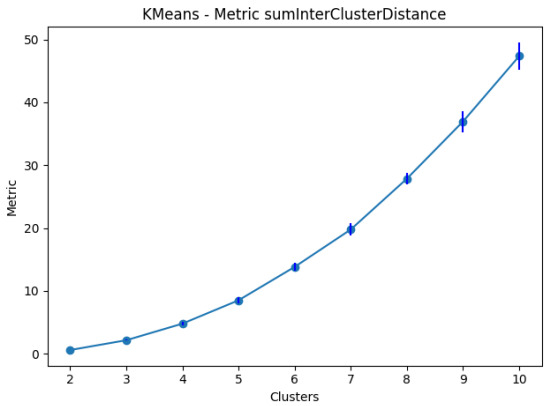


Figura 2 - Variação da distância inter-cluster para valores crescentes de K, utilizando *K-Means*.

http://dx.doi.org/10.25286/repa.v3i3.945

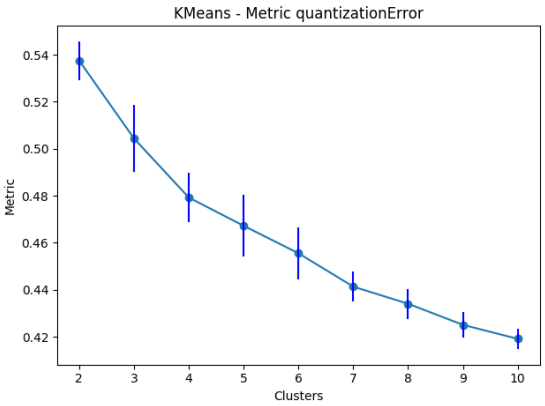


Figura 3 - Variação do erro quantizado para valores crescentes de K, utilizando *K-Means*.

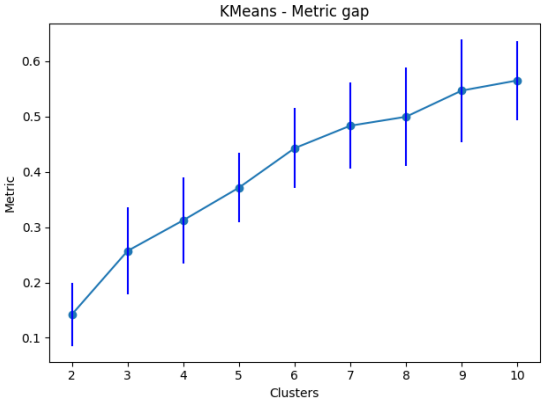


Figura 4 - Variação da estatística Gap para valores crescentes de K, utilizando *K-Means*.

Com relação ao número de clusters ideal de acordo com cada métrica, podemos analisar que as métricas das distâncias intra e inter-cluster e a do erro quantizado, apontaram o número de grupos 10 como ideal. Já a estatística gap, apontou como quantidade ideal 3. Apesar de três métricas apontarem 10 como quantidade ideal, podemos indicar 3 como quantidade de grupos a ser escolhida, uma vez que as distâncias intra e inter-cluster têm a tendência de se otimizarem com a quantidade de K aumentando. A quantidade 3 também foi apontada como ideal pela equipe de negócio do site de viagens, uma vez que uma granularização de 10 clusters não seria tão necessária para o direcionamento das ações de e-mail marketing.

Posteriormente, com o intuito de identificar o perfil de usuários criados pelo *K-Means*, com 3 grupos, foi analisada uma das amostras agrupadas a fim de identificar quais características pesaram na escolha dos clusters:foi percebido que o *K-Means* criou os clusters principalmente considerando a variável "Mês de Estadia", sendo um cluster de usuários que se hospedam mais nos primeiros meses do ano, outro grupo de usuários que se hospedam mais próximo ao meio do ano e o último mais próximo do final do ano.

5.3 Resultados *Fuzzy C-Means* (FCM)

Os resultados da execução do *Fuzzy C-Means*, estão exibidos na tabela 2. Podemos perceber que ao contrário do que ocorreu no *K-Means*, as métricas de erro quantizado e a distância intra-cluster, não mantiveram a minimização conforme a quantidade de grupos aumentou.

Podemos perceber que para a distância intra-cluster, a minimização funcionou bem até o k igual 7, depois disso o índice começou a subir. A dificuldade de encontrar grupos coesos pode ser percebida também na oscilação da métrica do erro quantizado, que ficou oscilando para cada número de k.Em relação à distância intra-cluster, o algoritmo demonstrou o comportamento esperado, maximizando a distância entre os grupos, conforme aumenta a quantidade de clusters. Já a estatística Gap, não manteve um aumento durante os testes com número crescente de K, ela atingiu o valor ótimo com k igual a 3, e depois teve uma pequena oscilação até diminuir no k igual a 10, indicando perda de qualidade nos agrupamentos.Nas figuras 5, 6, 7 e 8, podemos ver a variação das métricas das distâncias intra e inter cluster, o erro quantizado e a estatística gap, respectivamente.

http://dx.doi.org/10.25286/repa.v3i3.945

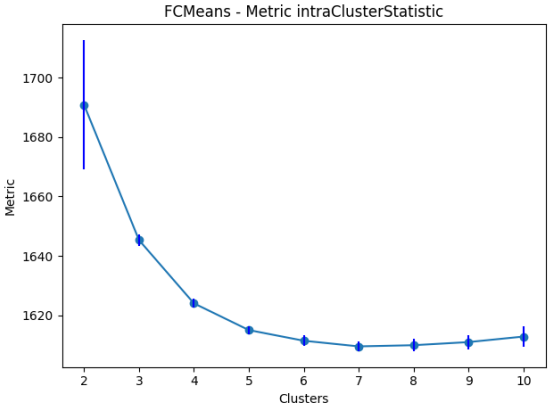


Figura 5 - Variação da distância intra-cluster para valores crescentes de K, utilizando FCM.

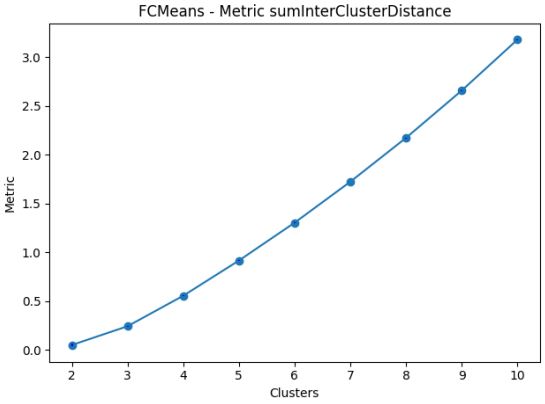


Figura 6 - Variação da distância inter-cluster para valores crescentes de K, utilizando FCM

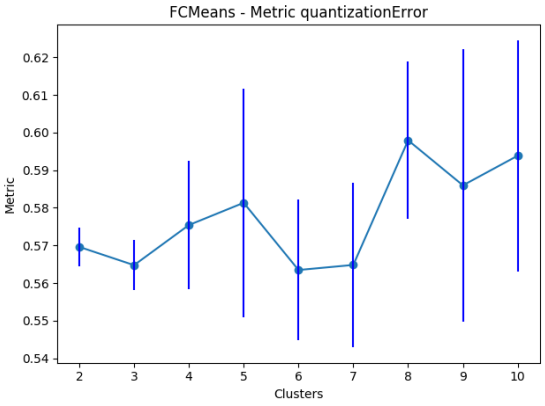


Figura 7- Variação do erro quantizado para valores crescentes de K, utilizando FCM.

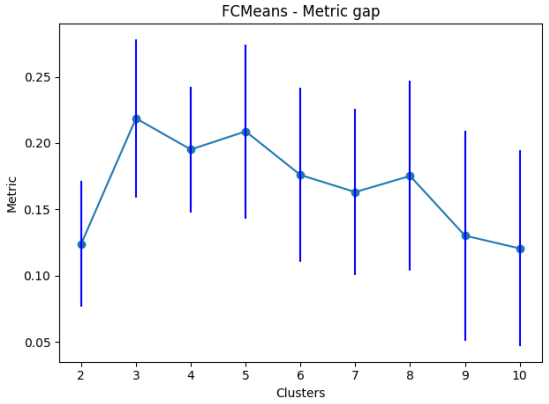


Figura 8- Variação da estatística Gap para valores crescentes de K, utilizando FCM.

Com relação à quantidade de grupos ideal, podemos perceber que as métricas do erro quantizado e estatística gap, apontaram a quantidade de grupos igual a 3 como melhor escolha (apesar do erro quantizado do k igual 7 possui o valor da métrica igual ao 3, o desvio padrão é maior, fazendo com que predominasse o k igual a 3 para essa métrica). As distâncias intra e inter-cluster, apontaram o k igual a 7 e 10, respectivamente.

Para identificação do perfil dos usuários criados pelo FCM com 3 grupos, também foi analisada uma amostra agrupada com o intuito de identificar características que se destacam entre os perfis: também foi identificado que a característica "Mês da estadia" puxou a criação dos grupos, de maneira igual ao K-Means, com usuários que se hospedam no início, meio e fim do ano.

5.4 Análise e Discussão

Durante a análise e comparação dos resultados dos algoritmos, pudemos perceber que existiram duas características que também poderiam definir os perfis dos usuários, foram elas o Destino e o Hotel. Quando essas variáveis foram transformadas de categóricas para numéricas, o processo utilizado foi o de ordená-las em ordem alfabética e atribuir números sequênciais, iniciando em 1. Exemplificando: supondo que tivéssemos 3 cidades: Atibaia, Bauru e Campo Grande.Seguindo a lógica utilizada, elas seriam transformadas em 1, 2 e 3, respectivamente.

http://dx.doi.org/10.25286/repa.v3i3.945

Devido à forma como essas características foram transformadas, o ganho de informação não foi relevante, isto porque os grupos em que os usuários se dividiram seguiram a ordem alfabética das características, o que apontou, por exemplo, grupos em que determinado usuário compra mais hotéis/destinos com as letras do começo do alfabeto.

Como uma forma de melhorar esse cenário, poderíamos trabalhar essas duas características para gerar uma terceira, que seria o tipo de hotel. A lógica seria, tendo como base o Hotel e o Destino, apontar qual o perfil em que o hotel se encaixaria melhor (Praia, Campo, entre outros). Com essa terceira característica, o modelo poderia conseguir uma melhor generalização, apontando perfis de usuário por tipo de hotel.

6 Conclusões

Este artigo fez uma análise da aplicação de dois algoritmos de agrupamento (*K-Means* e *Fuzzy C-Means*) em uma base de dados de reservas de hotéis de uma agência de viagens online do Brasil. As características do conjunto de dados eram detalhes sobre as vendas efetuadas no site da agência durante os anos de 2016 e 2017. O objetivo foi identificar perfis de clientes com o intuito de melhorara eficácia do envio de ofertas através de e-mail marketing, possibilitando o envio de anúncios personalizados para cada perfil.

Foram escolhidos o *K-Means* e o *Fuzzy C-Means* por serem alguns dos algoritmos mais populares na aplicação de agrupamento em bases de dados, além do fato, de cada um utilizar abordagens de clusterização diferentes (*K-Means* é particionista e o FCM utiliza lógica *fuzzy*).

Os resultados mostraram que ambos os algoritmos chegaram a um número de K ideal igual a 3, considerando a métrica de avaliação de clusters Estatística Gap. Considerando os dados agrupados, os dois algoritmos também tiveram resultados semelhantes, dando ênfase à característica "Mês de Estadia" para criar os grupos. Apesar da similaridade nos resultados, podemos apontar que o *K-Means* teve desempenho levemente superior, uma vez que o valor da sua estatística Gap foi maior que a do FCM, demonstrando grupos mais consistentes.

Como trabalhos futuros, podemos agregar outros algoritmos de clusterização para serem analisados. Também podemos aplicar a clusterização resultada desse estudo de caso, para ser validada na prática, avaliando métricas de envio de campanhas por e-mail como a taxa de abertura das campanhas segmentadas em relação com as das campanhas sem segmentação.

Referências

[1] PROVOST, Foster; FAMCETT, Tom. Data science and its relationship to big data and data-driven decision making. **Big Data**, v. 1, n.1 , p. 51-59, 2013. Disponível em: <https://www.liebertpub.com/doi/abs/10.1089/big.2013.1508>.

[2] MUMTAZ, Karam; DURAISWAMY Karthig. A Novel Density based improved k-means Clustering Algorithm – Dbkmeans. **International Journal on Computer Science and Engineering**, v. 2, n.2, p. 213-218, 2010. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.5204&rep=rep1&type=pdf>

[3] GHOSH, Soume; DUBEY, Sanajy Kumar. Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. **International Journal of Advanced Computer Science and Applications**, v. 4, n.4, p. 35-39, 2013. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.683.5131&rep=rep1&type=pdf#page=46>.

[4] JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data Clustering: A Review. **ACM Computing Surveys**, v. 31, n.3, p. 264-323, 1999. Disponível em: <https://dl.acm.org/citation.cfm?id=331504>.

http://dx.doi.org/10.25286/repa.v3i3.945

[5] LORENZIi, Fabiana et al. Enhancing the Quality of Recommendations Through Expert and Trusted Agents. In: IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE, 23., 2011, Florida. **Proceedings**... Florida: IEEE, 2011. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6103346/>

[6] Lorenzi, Fabiana; LOH, Stanley; ABEL, Mara. PersonalTour: a recommender system for travel packages. In: IEEE/WIC/ACM INTERNACIONAL CONFERENCE ON WEB INTELLIGENCE AND INTELLIGENT AGENT TECNOLOGY, 4., 2011, Lyon. **Proceedings**... Lyon: IEEE, 2011. p. 333-336. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6040800/>.

[7] COELHO, Bruno; MARTINS, Constantino; ALMEIDA, Ana. Web Intelligence in Tourism: user modeling and recommender system. In IIEEE/WIC/ACM INTERNACIONAL CONFERENCE ON WEB INTELLIGENCE AND INTELLIGENT AGENT TECNOLOGY, 3., 2010, Califórnia. **Proceedings**... Califórnia: IEEE, 2010. p. 619-622. Disponível em: <https://ieeexplore.ieee.org/abstract/document/5616446/>.

[8] SANTOS, Filipe et al. Tourism Recommendation System based in user's profile and functionality levels. In: INTERNATIONAL C* CONFERCENCE ON COMPUTER SCIENCE & SOFTWARE ENGINEERING, 9., 2016, Porto. **Proceedings**... Porto: ACM, 2016. p. 93-97. Disponível em: <https://dl.acm.org/citation.cfm?id=2948995>.

[9] ZHAO, Xuesong; JI, Kkaifan. Tourism E-Commerce Recomender System Based on Web Data Mining. In: The International Conference on Computer Science & Education, 8., 2013, Sri Lanka. **Proceedings**... Sri Lanka: IEEE, 2013. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6554161/>.

[10] MacQueen, James. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, Califórnia, v. 1, n.14, p. 281-297, 1967. p. 281-297.

[11] BEZDEK, James C.; EHRLICH, Robert; FULL, Willam. FCM: The Fuzzy C-Means Clustering Algorithm. **Computers & Geosciences**, v. 10, n.2-3, p. 191-203, 1984.

[12] Open Refine. Disponível em: <http://openrefine.org>. Acessoem: 23 abr. 2018.

[13] TIBSHIRANI, Roberto; GUENTHER, Walther; HASTIE, Trevor. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society**: Series B, v. 63, Parte 2, p. 411-423, 2001.

http://dx.doi.org/10.25286/repa.v3i3.945

Análise de Regressão Aplicada a Previsão de Reprovação de Alunos em Plataforma de Ensino a Distância

Analysis of Regression Applied to the Reproducibility Forecast of Students in Distance Learning Platform

Francisco de Assis de Araújo [1] orcid.org/0000-0002-6052-5330
Rodrigo Lins Rodrigues [2] orcid.org/0000-0002-4521-3806

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Pernambuco, Brasil.
[2] Departamento de Educação, Universidade Federal Rural de Pernambuco, Recife, Pernambuco, Brasil.

E-mail do autor principal: faa@ecomp.poli.br

Resumo

Um dos principais problemas enfrentados no Ensino a Distância são os riscos de reprovação e evasão de alunos. Com o objetivo de auxiliar Professores e gestores nessa modalidade de ensino, este trabalho demonstra resultados das aplicações práticas de técnicas estatísticas e mineração de dados para previsão de reprovação de Alunos através da Análise de Regressão Logística que demonstrou sua eficácia através de excelentes índices de desempenho em três modelos de dados utilizados, índices estes que foram considerados estatisticamente iguais através da Análise de Variância (ANOVA) aplicada ao comparar os índices de desempenho dos modelos de Regressão gerados. Através dos índices de significância das variáveis selecionadas em cada modelo é possível identificar os meios de interação que mais contribuem com o desempenho do aluno, auxiliando no combate a reprovação.

Palavras-Chave: Previsão de Reprovação; Análise de Regressão; EAD;

Abstract

One of the main problems faced in Distance Learning is the risks of student disapproval and avoidance. With the objective of assisting teachers and managers in this teaching modality, this paper demonstrates the results of the practical applications of statistical techniques and data mining to predict student disapproval through Logistic Regression Analysis that demonstrated its effectiveness through excellent performance indices in three data models used, which were considered statistically equal by the Analysis of Variance (ANOVA) applied when comparing the performance indices of the Regression models generated. Through the indices of significance of the variables selected in each model, it is possible to identify the means of interaction that contribute most to the student's performance, helping to combat failure.

Key-words: *Forecast of Reprobation; Regression Analysis; EAD;*

 http://dx.doi.org/10.25286/repa.v3i3.954

1 Introdução

A Educação a Distância (EAD) no Brasil tem se consolidado com diversos estudantes optando por essa modalidade de ensino para ampliar suas formações e realização profissional. Um dos principais diferenciais desta modalidade de ensino é a grande quantidade de dados gerada pelas interações nas plataformas de suporte online, conhecidas como AVA ou Ambientes Virtuais de Aprendizagem que abre novas possibilidades para pesquisas buscando compreender os processos de aprendizagem por meio das interações de alunos e professores. De acordo com Cechinel et al, algumas áreas de pesquisas surgiram nos últimos anos com intuito de auxiliar em questões como essas **[1]**.

De acordo com Cristobal et al, a Mineração de Dados Educacionais (do inglês Educacional Data Mining - EDM) é uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados originados no contexto educacional **[2]**.

Baker et al., aponta a possibilidade de identificação de estudantes com alto risco de evasão e reprovação a partir de modelos automáticos como um dos potenciais problemas a serem atacados pela comunidade brasileira que atua na área de mineração de dados educacionais **[7]**.

Conforme CECHINEL et al, a eficácia e a eficiência de estudantes têm frequentemente sido associadas a diferentes medidas de suas interações dentro dos Ambientes Virtuais de Aprendizagem-AVA, medidas estas que normalmente possuem uma alta correlação com o sucesso do aprendizado dos alunos **[1]**.

As interações dos alunos e professores com os ambientes virtuais de aprendizagem (AVA) proveem os dados que alimentam as pesquisas nessas áreas e possibilitam a descoberta de novos conhecimentos.

Este trabalho tem como objetivo estimar um modelo de relação existente entre a reprovação do aluno e as interações quantificadas através dos logs de interações com o ambiente de ensino, tendo como resultado um Modelo Regressão Logística capaz de prever o risco de reprovação do aluno.

Este trabalho está organizado da seguinte maneira: A Seção 2 apresenta alguns trabalhos relacionados que demonstram semelhanças com o trabalho atual, a Seção 3 apresenta um breve histórico sobre Análise de Regressão e o Modelo de Regressão Logístico, a Seção 4 apresenta o objetivo do trabalho e os modelos de dados utilizados no experimento de Análise de Regressão, a Seção 5 a base de dados através de algumas análises descritivas das variáveis, a Seção 6 o pré-processamento e limpeza dos dados, a Seção 7 apresenta o desenvolvimento da modelagem através da Análise de Regressão logística, na Seção 8 os Modelos Logísticos gerados são avaliados através de seus índices de desempenho e na Seção 9 são apresentadas as conclusões.

2 Trabalhos Relacionados

Em diversos trabalhos relacionados à EAD, observa-se que o sucessos dos alunos em ambiente de AVA está diretamente relacionado às diferentes medidas de interações com o ambiente. Por exemplo, Murray et al., observou que estudantes que apresentaram as mais altas taxas de acesso aos conteúdos no AVA obtiveram desempenhos mais satisfatórios nas avaliações **[4]**.

Dickson et al**.**, descobriu que o número total de cliques dados por estudantes é fortemente correlacionado com as suas notas finais em um curso **[5]**. Manhães et al**.**, utilizou técnicas de mineração de dados para prever a evasão de estudantes em cursos presenciais da Escola Politécnica da Universidade do Rio de Janeiro **[6]**.

Cechinel et al, descreve resultados da aplicação de técnicas de aprendizado de máquina para demonstrar a viabilidade de utilizar apenas a quantidade de interações dos alunos para gerar predições razoavelmente precisas e que a introdução de atributos derivados das contagens (e.g. médias) é útil para previsões mais precisas quando a quantidade de dados é esparsa **[1]**.

Esses estudos assemelham-se ao presente trabalho através da utilização de atributos de contagens de interações para treinamento do modelo de predição, contudo diferente de alguns trabalhos, este busca desenvolver uma modelagem capaz de prevê a reprovação do aluno no intuito de colaborar com os Professores no combate a essa reprovação e consequentemente a evasão no ambiente de ensino.

3 Análise de Regressão

Segundo Batista, a expressão "regressão" em estatística significa a dependência funcional entre duas ou mais variáveis aleatórias, correspondendo em termos matemáticos à obtenção de uma função que melhor represente a dependência entre aquelas variáveis. O Modelo *logit* foi inserido na terminologia estatística médica por Berkson, em 1944, que batizou o referido termo, por analogia com o modelo desenvolvido por Bliss em 1934, designado por *probit*. O Modelo de Regressão Logístico, também denominado por Modelo *logit,* é especialmente adaptável aos casos em que existe uma variável dependente binária ou dicotômica **[8].**

Através da Regressão Logística, avaliamos o relacionamento entre uma variável dependente e diversas outras variáveis intituladas de independentes, sendo a variável dependente ou resposta Y_j binária, essa variável binária pode assumir os valões $Y_j=0$ ou $Y_j=1$ que em nosso caso significa "aprovado" e "reprovado", respectivamente. Neste caso, "sucesso" é o evento de interesse e significa prever se o aluno será reprovado conforme as suas interações com o ambiente educacional.

De acordo com Baker et al, Uma forma de realizar essa previsão é aplicar o modelo de Regressão Logística, muito popular em EDM para predições binárias **[7]**. Segundo Batista **[8]**, o Modelo Logístico não tem muitas exigências de pressupostos e é usado para prever a probabilidade de eventos binários ocorrerem em que: se a probabilidade for > 0,5 a previsão é de que o resultado do evento seja 1 (em nosso caso significa reprovado) e <= 0,5 a previsão é de que o resultado do evento seja 0 (em nosso caso significa aprovado). Um modelo de regressão Logística segue a equação **(1)**.

$$\boldsymbol{logit}(\boldsymbol{p_j}) = \boldsymbol{ln}\left(\frac{\boldsymbol{p_j}}{\boldsymbol{1-p_j}}\right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{x}_1 + \cdots + \boldsymbol{\beta}_n\boldsymbol{x}_n \quad \textbf{(1)}$$

Em que $\mathbf{p}_j$ indica a probabilidade de ocorrência, $x_1 \ldots x_n$ representa o vetor de variáveis explicativas (ou independentes) e β_o e β_x indicam os coeficientes do modelo **[8]**.

4 Experimento Realizado

O experimento deste trabalho se dera através de uma base de dados referente às interações de Alunos em uma plataforma de Ensino a Distância (EAD) com os diversos artefatos educacionais disponibilizados durante os cursos de Pedagogia, Administração e Biologia através do primeiro ao oitavo períodos letivos, com o objetivo de descobrir um modelo de dados ideal para previsão de reprovação de alunos através de Regressão Logística. Neste experimento foram utilizados os modelos de dados Base Geral, Base recortada por Cursos e Base recortada por períodos, as variáveis participantes foram selecionadas através do método de Regressão *Stepwise*. Para testar se existe diferença entre os desempenhos dos três modelos gerados, os índices de desempenho dos modelos de predição foram comparados entre si através da Análise de Variância (*ANOVA*), assim como a comparação visual através da Curva ROC de cada modelo gerado.

5 Análise Descritiva dos Dados

A Base de dados disponibilizada no formato original xls, refere-se às interações de 1738 Alunos em uma plataforma de Ensino a Distância (EAD) com os diversos artefatos educacionais disponibilizados durante os cursos de Administração, Biologia, Letras e Pedagogia. Esta base contém 30217 linhas de dados e 39 variáveis, sendo 6 variáveis ordinais, 32 variáveis discretas, 1 variável continua.

Também fazem parte da base de dados as variáveis ordinais Curso, Período, Semestre, Id do Aluno, Id da Disciplina, Nome da disciplina e a variável contínua DESEMPENHO além das variáveis candidatas a preditoras descritas a seguir no quadro 1.

http://dx.doi.org/10.25286/repa.v3i3.954

Quadro 1 - Descrição das variáveis

Variável	Descrição sobre as variáveis
VAR01	Quantidade de diferentes locais (IP´s) a partir dos quais a(o) aluna(o) acessou o ambiente.
VAR02	Quantidade de mensagens enviadas por aluna(o) às(os) Professoras(es) pelo ambiente.
VAR03	Quantidade de mensagens enviadas por aluna(o) às(os) Tutor(es) pelo ambiente.
VAR04	Quantidade geral de mensagens enviadas pela(o) aluna(o) dentro do ambiente.
VAR05	Quantidade geral de mensagens recebidas pela(o) aluna(o) dentro do ambiente.
VAR06	Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo "tira-dúvidas".
VAR07	Quantidade de postagens no "Fórum tira dúvidas";
VAR08	Quant. de postagens de um(a) aluno(a) em fóruns que foram respondidas por outros(as) alunos(as).
VAR09	Quantidade de postagens de um(a) aluno(a) em fóruns que foram respondidas pelo(a) professor(a) ou tutor(a).
VAR10	Quantidade de colegas diferentes para quem o(a) aluno(a) enviou mensagens dentro do ambiente.
VAR12	Quantidade de visualizações da aba "Conteúdo" do curso, onde constam os arquivos com o conteúdo programático do curso
VAR13	Horário que mais realizou atividades;
VAR14	Turno do dia em que realizou mais atividades.
VAR16	Quantidade de atividades entregues por um(a) aluno(a) fora do prazo, por disciplina;
VAR17	Tempo médio entre a abertura da atividade e sua submissão;
VAR18	Quantidade de leituras feitas ao fórum (*pageviews*);
VAR20	Quantidade de respostas ao tópico principal (refazer opinião em fórum);
VAR21	Quantidade de pageviews ao quadro de notas;
VAR22	Quantidade de vezes que o aluno visualiza o (*Cheklist* Atividades)
VAR23	Quantidade de visualizações de notas por atividade;
VAR24	Média semanal da quantidade de acessos de um(a) aluno(a) ao ambiente.
VAR25	Tempo médio entre a criação de um tópico no fórum temático e a primeira postagem do aluno;
VAR31	Quantidade de acessos do(a) aluno(a) ao ambiente.
VAR31b	Quantidade de dias distintos que o aluno entrou na disciplina
VAR31c	Quantidade de dias distintos que o aluno entrou na plataforma
VAR32a	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Manhã).
VAR32b	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Tarde).
VAR32c	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Noite).
VAR32d	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Madrugada).
VAR33	Quantidade de atividades entregues por um(a) aluno(a) no prazo, por disciplina;
VAR34	Quantidade geral de postagens de um(a) aluno(a) em fóruns.
VAR35	Quantidade de respostas de um(a) professor(a) para as dúvidas de alunos(as) em fóruns.

Para um melhor entendimento, fez-se necessário a realização de algumas análises descritivas, para isto o gráfico de barras apresentado na figura 10 mostra as médias de acessos por aluno em cada meio através do qual o aluno interagiu com o ambiente de ensino, portanto a média do total de acessos por aluno (Var31) foi de 1470, o meio através do qual o Aluno menos interagiu com o ambiente foi em "Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo tira-dúvidas" (VAR06), "Quantidade de postagens no Fórum tira dúvidas" (VAR07), "Quantidade de respostas ao tópico principal (refazer opinião em fórum)" (VAR20) e "Quantidade de respostas de um(a) professor(a) para as dúvidas de alunos(as) em fóruns" (VAR35) o meio através do qual os alunos mais interagiram com o ambiente foi em "quantidade de mensagens recebidas dentro do ambiente" (Var05) com uma média de 667 mensagens por aluno, porém a média de mensagens enviadas pelos alunos (Var03) foi de 284 mensagens, o turno em que o aluno mais acessou o ambiente foi o turno da noite (Var32c) com uma média de 544 acessos e o turno que menos o aluno acessou foi o turno da madrugada (Var32d) com uma média de 32 acessos, o meio onde houve menos interações foi em "Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo tira-dúvidas" (Var06) com uma média de 0,48 por aluno, foi realizado uma análise da média de acessos para o DESEMPENHO e constatado que os DESEMPENHOS >4 tiveram uma média de 1315 acessos e os DESEMPENHOS <=4 tiveram uma média de 489 acessos. Através do gráfico de barras apresentado na figura 1, visualiza-se um comparativo ente as médias de acessos dos alunos em cada ambiente de interação através das variáveis candidatas a preditoras.

6 Pré-Processamento dos Dados

Esta etapa tem como objetivo melhorar a qualidade dos dados e a eficiência do processo de

mineração através da remoção de dados ruidosos, valores faltantes e dados inconsistentes oriundos da coleta dos mesmos. Na base de dados original no formato .xls, foi efetuado a mudança do padrão numérico Brasileiro para o padrão americano, substituindo a virgula (separador decimal) por ponto. A variável "Quantidades de Time Out" (VAR28), mesmo constando no dicionário de dados, foi excluída das análises devido inexistência na base de dados original. Conforme quadro 2, não sendo constatado interações no Semestre 2009.2, foi considerado como dados faltantes ou inexistentes e o referido semestre foi excluído da base de dados.

Conforme quadro 3, referente ao sumário da variável DESEMPENHO, constatando-se o limite superior igual a 11, considerado como inconsistência o referido DESEMPENHO foi reduzido para 10.

Através da variável DESEMPENHO, foi criada a variável DESEMPENHO_BINÁRIO, através do seguinte critério: A variável criada recebeu "0" para aprovado quando DESEMPENHO >= 4 e "1" para reprovados quando DESEMPENHO < 4.

Quadro 2 - Totais de Alunos e interações por Semestre

Semestre	Alunos	Interações
2009.2	50	0
2010.1	170	54349
2010.2	259	211757
2011.1	276	183265
2011.2	260	172672
2012.1	195	100904
2012.2	24	3069
2013.2	543	583463
2014.1	530	241758
2014.2	1295	524079
2016.1	1294	480342

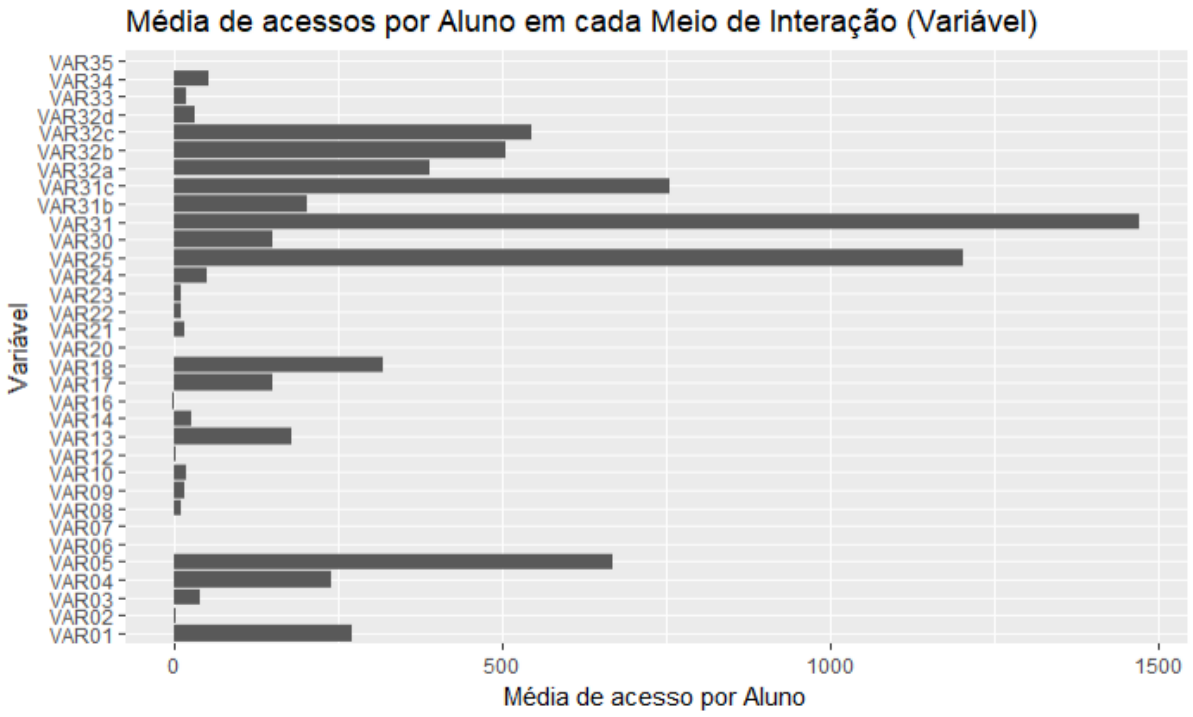


Figura 1: Média de acessos por aluno ao meio de interação

http://dx.doi.org/10.25286/repa.v3i3.954

Quadro 3 - Sumarização da variável DESEMPENHO

Min.	0.000
1st Qu.	1.000
Median	5.000
Mean	4.298
3rd Qu.	7.000
Max.	11.000

7 Desenvolvimento da Modelagem

Neste trabalho, estamos interessados em encontrar o melhor modelo de dados para o desenvolvimento de um modelo de Regressão Logístico capaz de realizar previsão de reprovação dos Estudantes com excelentes índices de precisão, para isto buscou-se descobrir o modelo de dados com melhor potencial de eficácia para o Modelo de Regressão. Os modelos de dados utilizados foram a Base Geral sem recortes e os recortes realizados através das variáveis Curso e Período. O objetivo deste recorte foi identificar as melhores configurações para a construção do modelo: (1) Modelo Genérico com todos os alunos, (2) Modelo por Período e (3) Modelo por Curso.

Em cada um dos 13 modelos de dados analisados, 1 modelo através da Base Geral sem recortes, 4 através dos recortes por Curso e 8 através dos recortes por Período, foi aplicado o método de Regressão *stepwise* para adicionar sistematicamente a variável mais significativa ou remove a variável menos significativa e determinar um melhor subconjunto de variáveis preditoras para o modelo de dados. O quadro 5 apresenta as variáveis selecionadas a partir do modelo de dados Base Geral sem recortes, e seus índices de significância no modelo.

Conforme os modelos de dados, as análises realizadas geraram 13 Modelos de Regressão Logística, a partir desses modelos foram geradas as matrizes de confusão e extraídos os índices de desempenho do modelo Base Geral conforme quadro 6, as médias dos índices de precisão dos modelos por Período conforme quadro 7 e as médias dos índices precisão dos modelos por Curso conforme quadro 8. Conforme Kuhn M (2013), o quadro 4 descreve os índices de desempenho usados na avaliação dos modelos. A seguir serão apresentados três cenários nos quais foram analisados os três modelos de dados através da Análise de Regressão Logística e definidos os modelos de previsão a serem comparados entre si através de seus índices de desempenho, assim como a Curva ROC que, segundo Batista, baseada na taxa de verdadeiros positivos TPR e na taxa de falsos positivos FPR, descreve graficamente o desempenho do sistema classificador binário **[8].**

Quadro 4 - Índices de desempenho usados na avaliação dos modelos

Accuracy	Descreve com que frequência o classificador está correto
Kappa	Diferença entre a precisão e a taxa de erro nulo
Sensitivity	Razão entre *True* Positivos para o verdadeiro sim geral
Specificity	Razão de positivos verdadeiros para o total real não
Precision	Razão de positivos verdadeiros para o previsto global sim.
Prevalence	Relação entre o sim real e o número total de instâncias.
Balanced_Accuracy	Média de sensibilidade e especificidade

7.1 Cenário "Modelo Geral"

Neste modelo de dados foi utilizado a base sem recortes, as variáveis Selecionadas e seus índices de significância estão descritos conforme quadro 4 e aplicando o Modelo de Classificação Logístico obteve-se, através da matriz de confusão, os índices de precisão apresentados no quadro 6 assim como a curva ROC do modelo produzido e apresentada na figura 2.

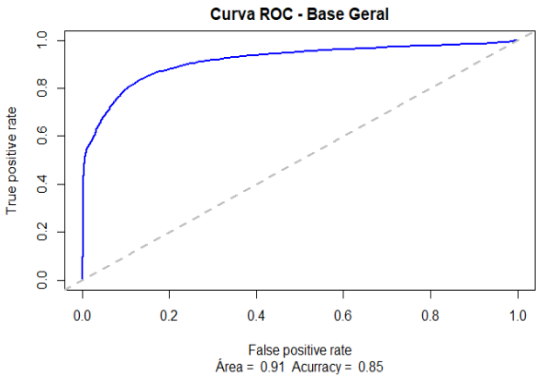


Figura 2 - Curva ROC do modelo gerado a partir da base geral

Quadro 5 - Variáveis e seus índices de significância, selecionadas através da Regressão *Stepwise* aplicada a Base Geral

	Estimate	Std.Error	t-value	Pr(>|t|)
Intercept	0.0129876	0.0129876	78.122	< 2e-16 ***
VAR01	0.0011975	0.0003955	3.028	0.002487 **
VAR06	0.0484049	0.0306022	1.582	0.113841
VAR08	0.0227512	0.0062065	3.666	0.000252 ***
VAR09	0.0148059	0.0047020	3.149	0.001660 **
VAR12	-0.0255182	0.0127252	-2.005	0.045042 *
VAR14	0.0614615	0.0073499	8.362	< 2e-16 ***
VAR16	-0.1459795	0.0209859	-6.956	4.51e-12 ***
VAR17	0.0029227	0.0005390	5.422	6.48e-08 ***
VAR18	0.0021749	0.0005241	4.150	3.45e-05 ***
VAR21	0.0286768	0.0054393	5.272	1.47e-07 ***
VAR23	-0.0135823	0.0058131	-2.336	0.019549 *
VAR25	-0.0007355	0.0001780	-4.132	3.72e-05 ***
VAR31	-0.0029649	0.0006821	-4.347	1.44e-05 ***
VAR32	-0.0010347	0.0003187	-3.247	0.001182 **
VAR33	-0.4312691	0.0137166	-31.441	< 2e-16 ***
VAR34	-0.0374658	0.0051717	-7.244	5.85e-13 ***
VAR35	0.0522678	0.0260286	2.008	0.044747 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Quadro 6 - Índices de precisão do modelo Logístico obtido através da base geral

Índices	Base Geral
Accuracy	0.85
Kappa	0.70
Sensitivity	0.83
Specificity	0.84
Precision	0.87
Prevalence	0.50
Balanced_Accuracy	0.85

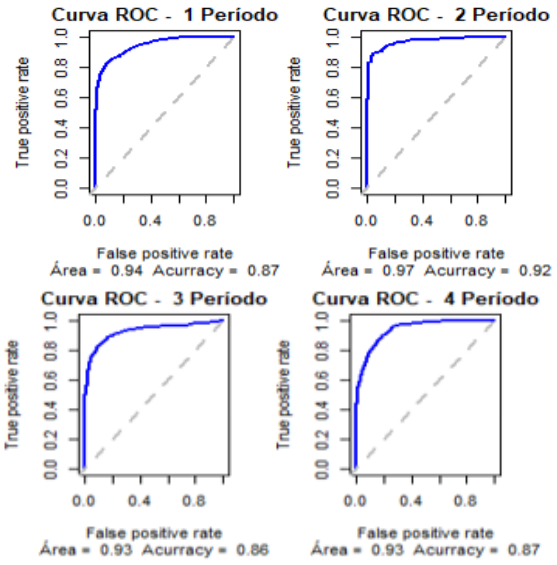
7.2 Cenário "Modelo por Período"

Neste modelo de dados foi utilizado a base geral recortada por período sem discriminação de Curso, em cada um dos oito recortes gerados foi aplicado o modelo de classificação Logístico e através da matriz de confusão foram obtidos os índices de precisão descritos no quadro 7 e a curva ROC para cada Período com as respectivas áreas abaixo da curva e *accuracy* do modelo, conforme apresentado na figura 3.

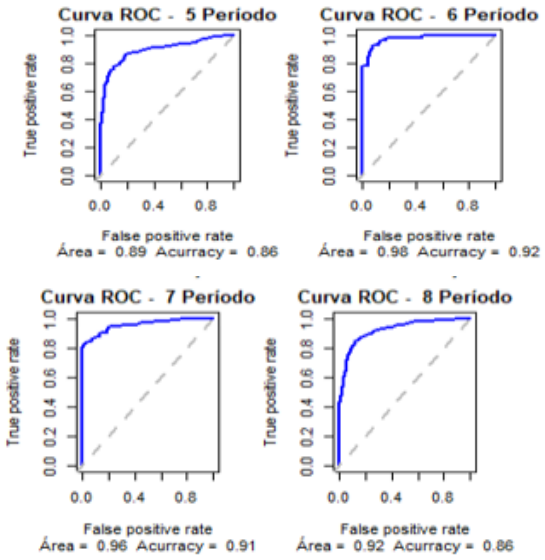
Quadro 7 - Índices de precisão do modelo Logístico obtido através dos recortes por período

Índices	Períodos							
	1	2	3	4	5	6	7	8
Accuracy	0.87	0.92	0.86	0.87	0.86	0.92	0.91	0.86
Kappa	0.74	0.84	0.72	0.71	0.69	0.83	0.81	0.71
Sensitivity	0.79	0.87	0.84	0.93	0.72	0.88	0.84	0.86
Specificity	0.85	0.89	0.87	0.85	0.86	0.91	0.87	0.89
Precision	0.91	0.96	0.86	0.87	0.87	0.93	0.96	0.82
Prevalence	0.45	0.47	0.46	0.64	0.36	0.45	0.48	0.43
Balanced_ Accuracy	0.86	0.92	0.86	0.84	0.83	0.91	0.91	0.86

Figura 3: Curvas ROC dos modelos gerados a partir dos recortes por Período



http://dx.doi.org/10.25286/repa.v3i3.954



7.3 Cenário "Modelo por Curso"

Neste modelo de dados foi utilizado a base geral recortada por Curso sem discriminação de Período, em cada um dos quatro recortes gerados foi aplicado o modelo de Regressão Logístico e através da matriz de confusão foram obtidos os índices de precisão descritos no quadro 8 e a curva ROC para cada Período com as respectivas áreas abaixo da curva e *accuracy* do modelo, conforme apresentado na figura 4.

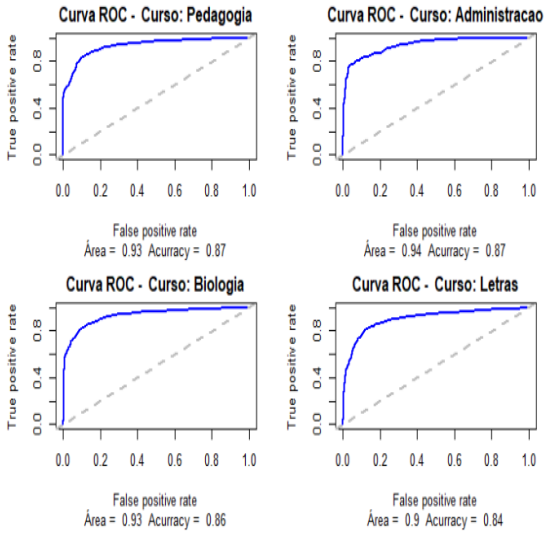


Figura 4 - Curvas ROC dos modelos gerados a partir dos recortes por Curso

Quadro 8: Índices de precisão dos modelos gerados a partir dos recortes por Curso

Índices	Pedag	Administ	Biologia	Letras
Accuracy	0.87	0.87	0.86	0.84
Kappa	0.73	0.73	0.72	0.69
Sensitivity	0.86	0.81	0.86	0.83
Specificity	0.86	0.86	0.86	0.83
Precision	0.87	0.88	0.87	0.86
Prevalence	0.50	0.44	0.50	0.52
Balanced_ Accuracy	0.87	0.86	0.86	0.84

8 Avaliação dos Modelos

Neste trabalho foi aplicado análise de Regressão Logística na geração de um modelo para previsão de reprovação de estudantes através dos números de interações dos mesmos com o ambiente virtual de ensino.

Através do modelo de dados Base Geral, sem recortes, o Modelo de Regressão Logística gerado apresentou um índice de *accuracy* de 85%, em busca de melhorar a acurácia deste modelo a base de dados foi recortada por Curso e Período, aplicada a Regressão Logística em cada recorte e calculada a média dos índices de desempenho dos modelos gerados através dos referidos recortes conforme quadro 9, gerou-se dois vetores com as medias dos índices de desempenho em cada recorte e outro vetor com os índices do modelo de dados sem cortes, nestes três vetores foi aplicada análise de variância (ANOVA) que resultou em um valor p associado a estatística t igual a 0,955 conforme demonstrado no quadro 10, portanto pode-se afirmar que não há diferença entre as médias dos índices de desempenho dos modelos aplicados a base geral e os recortes, sendo assim pressupõe-se que os modelos de regressão Logística obtidos a partir da base de dados geral e seus recortes tem índices de desempenho semelhantes o que podemos constatar visualmente na curva ROC dos referidos modelos.

Quadro 9 - Índices de *accuracy* do modelo gerados a partir da Base geral e média dos índices gerados a partir dos recortes por Período e Curso

Índice	Base geral	Índice médio por recorte	
		Curso	Período
Accuracy	0.85	0.86	0.88
Kappa	0.70	0.72	0.76
Sensitivity	0.83	0.84	0.84
Specificity	0.84	0.85	0.87
Precision	0.87	0.87	0.90
Prevalence	0.50	0.49	0.47
Balanced_ Accuracy	0.85	0.86	0.88

Quadro 10 – A nova aplicada aos índices dos modelo apresentados na Quadro 2.

	Df	SumSq	Mean Sq	Fvalue	Pr(>F)
Bases	2	0,0016	0.000817	0.046	0.955
Residuals	21	0.3707	0.017651		

De acordo com as análises realizadas, o Modelo de Classificação Logístico para previsão de reprovação do aluno pode ser estimado através do modelo de dados generalizado através da base geral sem recortes com índice de desempenho mediano de 0.835, através dos modelos de dados gerados a partir dos recortes por Curso os quais apresentaram índice de desempenho mediano de 0.845 ou através dos modelos de dados gerados a partir dos recortes por Período que apresentaram índice de desempenho mediano de 0,855 conforme o quadro 10.

Quadro 10 - Sumário dos índices de desempenho por modelo e dados.

Índice	Base Geral	Curso	Período
Min.	0.5000	0.4900	0.4700
1st Qu.	0.7975	0.8100	0.8200
Median	0.8350	0.8450	0.8550
Mean	0.7837	0.7913	0.8037
3rd Qu.	0.8500	0.8600	0.8725
Max.	0.8700	0.8700	0.9000

Com o *BoxPlot* comparativo apresentado na figura 5 podemos verificar que os índices de desempenho possuem distribuição assimétrica, variabilidades semelhantes e que os modelos gerados através dos recortes por Curso apresentam índices de desempenho mediano mais próximo do modelo gerado a partir da Base Geral e que apesar do modelo gerado a partir dos recortes por Período apresentar índices de precisão levemente superior aos demais modelos, estatisticamente a ANOVA comprova a igual capacidade preditiva dos três modelos.

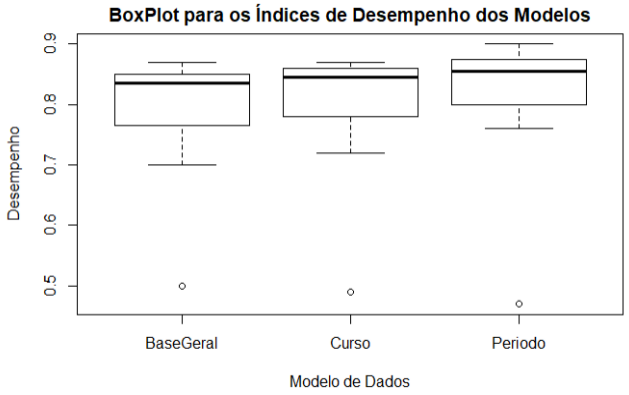


Figura 5 - Gráfico box-plot, índices desempenho dos modelos de classificação

9 Conclusões

A consistência dos dados através de forma padronizada e estruturada é fundamental para o sucesso das pesquisas e implementação de modelos de classificação e previsão.

Neste trabalho, através de algumas análises descritivas como visto em outros trabalhos, constatou-se que o desempenho do aluno está diretamente associado ao total de interações pois os alunos que tiveram desempenho superior também realizaram uma número superior de acessos ao ambiente de ensino.

Como foi demonstrado, os modelos de previsão obtidos através da Regressão Logística realizada através dos três modelos de dados obtiveram excelentes índices de desempenho possibilitando a aplicação de qualquer um dos modelos na previsão de reprovação do aluno, portanto tomando como referência a Base Geral sem recorte, conforme variáveis selecionadas e apresentadas no quadro 4, com nível de confiança de 95% e significância de 5% os meios de interação que mais contribuem com a previsão de

http://dx.doi.org/10.25286/repa.v3i3.954

reprovação do aluno são representados através da variáveis VAR12, VAR23 e VAR35, com nível de confiança de 99% e nível de significância de 1% os meios de interação que mais contribuem são representados através das variáveis VAR01, VAR09 e VAR32. Assim como na figura 4, seria possível verificar as variáveis que mais contribuem com a previsão de reprovação nos demais modelos de dados.

Apesar das análises realizadas, não podemos desconsiderar que o modelo de previsão pode cometer o grave erro de apresentar elevada taxa de falso positivo em que o algoritmo classifica um aluno no grupo de aprovação enquanto o mesmo encaminha-se para a reprovação ou um erro menos grave, falso negativo, em que o aluno erroneamente é classificado no grupo de reprovados, erros esses que podem comprometer o modelo de classificação.

Agradecimentos

Os autores agradecem a colaboração do Núcleo de Educação a Distância - NEAD/UPE pelo fornecimento da base de dados. Os autores também agradecem ao Grupo de Pesquisa em Ciência de Dados Educacionais (CiDE/UFRPE).

Referências

[1] CECHINEL, Cristian; ARAUJO, Ricardo Matsumura; DETONI, Douglas. Modelling and Prediction of Distance Learning Students Failure by using the Count of Interactions. **Brazilian Journal of Computers in Education**, v. 23, n. 03, p. 1, 2015.

[2] ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining: a review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 40, n. 6, p. 601-618, 2010.

[3] BROWN, Malcoln. Learning Analytics: Moving from Concept to Practice, **EDUCAUSE Learning Initiative Brief**, v. 7, 2012.

[4] MURRAY, Meg et al. Student interaction with content in online and hybrid courses: Leading horses to the proverbial water. In: INFORMING SCIENCE AND INFORMATION TECHNOLOGY EDUCATION CONFERENCE, 2013, Porto. **Proceedings**…Porto: Informing Science Institute, 2013. p. 99-115.

[5] DICKSON, W. Patrick. Toward a deeper understanding of student performance in virtual high school courses: Using quantitative analyses and data visualization to inform decision making. **A synthesis of new research in K–12 online learning**, p. 21-23, 2005.

[6] MANHÃES, Laci Mary Barbosa et a. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: *WORKSHOP* DE INFORMÁTICA NA ESCOLA, SIMPÓSIO BRASILEIRO DE INFORMATICA NA EDUCAÇÃO, 22., 2011, Aracajú. Anais..., Aracaju, 2011.

[7] BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de dados educacionais: Oportunidades para o brasil. **Brazilian Journal of Computers in Education**, v. 19, n. 02, p. 03, 2011.

[8] BATISTA, Antonio Sarmento. **Regressão Logística**: uma introdução ao modelo estatístico. Porto: Vida Econômica, 2015.

Uma Proposta de um *Framework* para Gerir o Dado como Ativo de Valor nas Empresas de Trânsito

A Proposal fora Framework to Manage the Data as a Valuable Asset in Transit Companies

Luciene Maria dos Santos[1] orcid.org/0000-0003-0606-0293
Andrêza Leite de Alencar[2] orcid.org/0000-0002-7083-0646

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Pernambuco, Brasil.
[3]Bacharelado em Ciência da Computação, Universidade Federal Rural de Pernambuco, Recife, Pernambuco, Brasil.

E-mail do autor principal: lms2@ecomp.poli.br

Resumo

Um dos grandes desafios das organizações é saber mensurar o valor monetário dos dados. Nos últimos tempos com o surgimento dos grandes e massivos volumes de dados, "Big Data" esta tarefa é cada vez mais desafiadora. Tratar os dados como um ativo de valor auxiliará a organização nas tomadas de decisão de forma mais eficaz e eficiente do contrário a organização poderá está perdendo vantagens e benefícios por não conhecer o potencial verdadeiro do valor monetário dos seus dados. Este trabalho apresenta uma proposta de um framework que auxiliará as empresas a gerir os dados como ativo de valor. Como resultado foi possível monetizar e rentabilizar o valor dos dados e planejar a estimativa de crescimento em uma perspectiva de 5 anos, ponto fundamental que possibilitou que a Instituição não apenas conhecesse o quanto vale o valor de sua informação bem como aprovasse todo planejamento de orçamento financeiro para o ano de 2018 necessário para proteger seu ativo de dados.

Palavras-Chave*:* *Framework*; Dados; Ativo; Valor;*Big Data*;

Abstract

One of the great challenges of organizations is knowing how to measure the monetary value of data, in recent times with the emergence of large and massive data volumes, "Big Data" this task is increasingly challenging. Treating data as a value asset will help the organization in decision making more effectively and efficiently otherwise the organization may be losing its advantages and benefits by not knowing the true potential of the monetary value of its data. This work shows a proposal for a framework that will help companies to manage data as a valuable asset. As a result, it was possible to monetize and improve the value of the data, planning the growth estimate from a 5 years perspective, as a fundamental point that enabled the Institution, not only to know the value of its information, as well as approving all financial budget planning for the year 2018, required to protect your asset data.

Key-words: *Framework; Data; Asset; Value; Big Data;*

 http://dx.doi.org/10.25286/repa.v3i3.962

1 Introdução

Nos últimos tempos um termo constante quando se refere a dados é o "*Big Data*" que são extensos conjuntos de dados, primariamente com as características de volume, variedade, velocidade e/ou variabilidade **[2].** Com o amadurecimento do *Big Data* foram encontradas outras características das quais algumas são: Valor, Veracidade e Visualização.

Com a explosão do *Big Data* é cada vez mais comum extrair conhecimento dessa quantidade massiva de dados e muito mais desafiador gerar valor de negócio através destes. Outra tarefa desafiadora é saber mensurar o valor monetário do dado tratando-o como ativo intangível, que segundo as normas contábeis um ativo intangível são as contas representativas das aplicações de recursos em bens incorpóreos que contribuirão para a formação do resultado de exercício(s) futuro(s)**[10]**.

Outra maneira para rentabilizar os dados é utilizando-se de práticas de governança de dados adequadas para este fim **[14]**. Um fator importante a considerar, se a organização não estiver quantificando o valor da informação, é que provavelmente não esteja gerando ou demonstrando seu valor de forma eficaz e suficiente, nem esteja colhendo qualquer outro benefício potencial advindo desta monetização **[6]**.

Diante dessa necessidade surgiu a "*infonomics*" (disciplina que gerencia e contabiliza o valor da informação com mesmo rigor e formalidade como qualquer outro ativo tradicional que as empresas possuem tais como: ativo financeiro, capital humano, físico (bens e imóveis) e etc.). Cada vez mais cabe às organizações se comportarem como se fossem otimizar a capacidade da informação de gerar valor comercial **[4]**.

É importante mencionar que há um processo natural de maturidade para melhor adotar políticas de governança de dados de acordo com a área de negócio da organização e as atuais circunstâncias desta em relação ao mercado competitivo **[14]**. Baseado neste processo este trabalho propõe o desenvolvimento de um *framework* que auxilie a quantificar o valor da informação, para justificar custos e investimentos em infraestrutura de TI, capital humano além de outros custos que proteja e mantenha o seu ativo de dados.

2 Fundamentação Teórica

O Dicionário Longman de inglês contemporâneo fornece uma descrição relacionada de um *Framework:* "um conjunto de fatos, ideias, etc... a partir da qual são desenvolvidas mais ideias ou sobre as quais são baseados" **[9]**.

Frameworks facilitam o entendimento e comunicação da estrutura e do relacionamento dentro de um sistema, para um propósito definido, são utilizados como uma forma de traduzir temas complexos em formas que possam ser estudadas e analisadas**[17]**.Suportam o processo de tomada de decisão e de resolução de problemas, fornecendo as categorias e representações normalmente em uma linguagem de símbolos**[11]**.

As descrições acima não são mutuamente exclusivas, e parecem combinar com a maioria dos *frameworks* de gerenciamento existentes. No entanto, uma definição padrão não existe. Os *frameworks* são cada vez mais utilizados na disciplina de gestão como traduzir questões complexas em um formato simples e analisável **[17]**.

2.1 O dado como ativo de valor

Apesar do burburinho em torno do "*big data"*apenas alguns *CIOS* e outros executivos de TI informaram que suas organizações estavam conseguindo gerar valor comercial significativo a partir de seus dados **[1]**. Qual o valor dos seus dados? muitas empresas ainda não sabem a resposta para essa pergunta, mas no futuro as empresas precisarão desenvolver maior expertise na avaliação dos seus ativos de dados.

Ao invés de estudar o valor dos dados no resumo, foram analisados os eventos que provocaram a necessidade de tal avaliação e que pode ser comparado entre as empresas tais eventos. Por tanto foi definido o valor dos dados

como um composto de três fontes de valor: 1- O ativo, 2- O valor da atividade, 3- O valor esperado ou futuro **[17]**.

Infonomics, expressa o crescente comportamento e a importância da informação como um ativo econômico. Monetiza, gerência e mensura informações como um ativo para vantagem competitiva **[7]**. O termo *Infonomics* foi cunhado para transmitir o valor subjacente da informação em termos de produção, demanda de mercado e impacto econômico.

É também a criação do valor moderno para os serviços de informação e aborda a questão de saber se a informação se tornou ou não uma *commodity* e também mudou a forma de como as empresas devem criar e comercializar suas informações **[13]**.

3 Objetivo

Criar um *framework* de uso comum para auxiliar empresas e organizações de trânsito no suporte à gestão, a estimar uma margem monetária e comercial dos dados da organização, utilizando práticas básicas de fácil acesso e de simples implementação. Este *framework* a princípio foi aplicado e validado numa instituição pública de trânsito para trabalhos futuros, será empregado em empresas privadas. Este *Framework* foi desenvolvido inspirado a princípio no trabalho *Corporate Governance of Big Data***[19]**e uma compilação das ideias do demais autores citados nos trabalhos relacionados.

4 Metodologia

Este trabalho caracteriza-se como uma pesquisa exploratória baseada na Governança de Dados para a elaboração e aplicação de procedimentos para auxiliar a monetizar ou rentabilizar os dados da organização. A pesquisa para embasamento teórico contou com dados qualitativos por meio de pesquisa a artigos científicos.

Como estratégia desta pesquisa foi utilizado o estudo de caso para validar o processo de implantação deste *framework*. No contexto desta pesquisa é utilizada a estratégia de estudo de caso único. Baseado nesta estratégia os estudos foram realizados em uma organização apenas.

A Tabela 1 mostra o Fluxo da Metodologia utilizada nesta pesquisa.

Tabela 1 - Fluxo da Metodologia.

Tarefa	Descrição
1-	Consulta exploratória a referências bibliográficas, artigos científicos, para o embasamento teórico;
2-	Definição dos procedimentos técnicos para aplicação do objeto deste estudo. Neste caso foi utilizado um estudo de caso para validação do framework;
3-	Estudo de Caso para implantação e validação do *framework*;
4-	Implantação *Framework*;
5-	Validação do *Framework*.

5 *Framework* para Gerir o Dado como Ativo de Valor na Organização

Este *framework* é composto por seis etapas que são executadas sequencialmente conforme demonstra a Figura 1.Cada etapa é essencial para se chegar ao valor do dado. Em cada etapa é explicado "como fazer". Nós tópicos seguintes será mostrado o detalhamento de cada etapa do *framework*.

http://dx.doi.org/10.25286/repa.v3i3.962

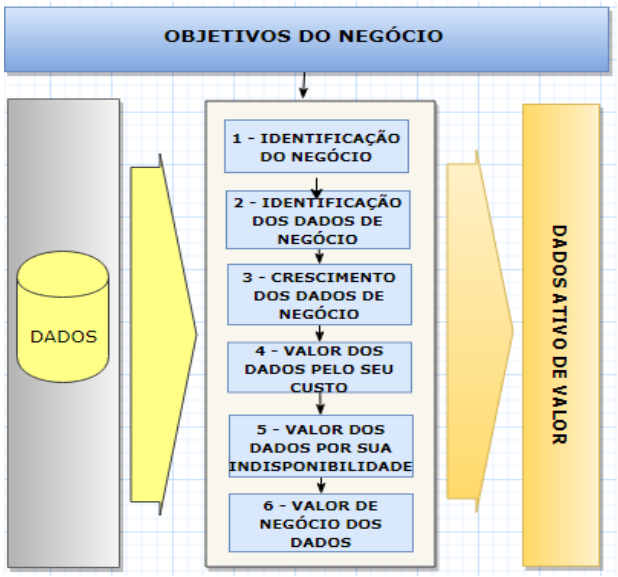


Figura 1 - Representação em Imagem do *Framework* o Dado como Ativo de Valor - DAV

ETAPA 1 - A Identificação do Negócio

Esta é a primeira e principal fase para encontrar o valor dos dados da organização, identificar e definir bem o negócio é muito importante. Para isso entrevistas com a alta gestão da empresa e analistas de negócio deixará claro a missão, atividade e objetivo da mesma, a entrada na Figura 1 desta fase é na área de Objetivos do Negócio.

Para empresas que não possuem um setor com especialistas de negócio é importante se deixar claro a atividade da empresa, suas metas e objetivos, identificar bem o que gera receita e o que pode gerar dentro do escopo de atividade da organização, é primordial.

Por exemplo num comércio eletrônico a principal atividade ou negócio da empresa é a venda de produtos por meios eletrônicos que pode ser a clientes finais ou a outras empresas.

ETAPA 2 - A Identificação dosDados de Negócio

Nesta etapa será feita a identificação e seleção dos dados relevantes para o negócio, ou seja, os dados que geram receita ou que podem ocasionar a perda dela. Aqui nessa fase a área de Negócio juntamente com o DBA ou alguém responsável pelo repositório de dados vai facilitar o acesso e a identificação destas.

Esta etapa é a sequência da Etapa 1 a entrada desta fase se dáà consultas ao repositório de dados. Por exemplo, em um comércio eletrônico pode-se identificar como tabelas de negócio a tabela de CLIENTE, VENDAS, PRODUTO.

ETAPA 3–Crescimento dos Dados de Negócio

É importante, independente do porte da empresa, seja ela pública ou privada, ter um controle efetivo de seus dados de negócio. Esse conhecimento será utilizado como base para os cálculos do valor do dado de negócio e da importância da integridade e planejamento dos *storages* para mantê-los intactos e disponíveis. Para saber a alocação dos dados a ajuda de um DBA ou de uma pessoa com conhecimento técnico em Banco de Dados se faz necessário.

ETAPA 4 – O Valor do Dado pelo seu Custo

Neste estudo iremos considerar o valor do custo do dado, o custo total do departamento de Tecnologia da Informação representado na Tabela 2, pois este responsável pela aquisição, manutenção e disponibilidade dos dados. Na verdade, o termo "custo" é utilizado, porém agrega valor à informação pois preserva e protege o valor da mesma.

O que compõe o custo do setor de Tecnologia da Informação compreende Custos Fixos (mão de obra + infraestrutura) e as licenças de uso de software e equipamentos se for o caso. A Equação (1) representa o Custo Médio Anual:

$$CustoMedioAnual = CustoMensal/12 \quad (1)$$

Tabela 2 - Custo Mensal do Departamento de Tecnologia da Informação.

Custos	IndicadorCusto
Custos Fixos e Variáveis (Folha de Pagamento + Treinamentos + Compra de Equipamentos para o departamento de Tecnologia da Informação	A
Licenças de *Software*	B
Licenças de *Hardware* caso existam	C
Total	Soma(A+B+C)

Custos Mensais do Departamento de Tecnologia da Informação + Compra de

equipamentos para o departamento de Tecnologia da Informação.

A = Folha de Pagamento + Treinamentos ou algum custo eventual com o capital humano

B = Licenças de *Software* tais como licenças do(s) banco(s) de dados e suporte + licenças de *software* de desenvolvimento e etc.

C = Licenças de *Hardware* caso existam tais como licenças com equipamentos físicos aluguel de computadores, *blader*, *Storages* e etc.

ETAPA 5 – Valor do Dado baseado em sua Indisponibilidade

É muito comum as empresas perderem receita e/ou responderem a processos jurídicos devido à ausência, indisponibilidade e/ou erro na aquisição e manutenção do seu ativo de dados.

Existe outros tipos de prejuízos nesta etapa, normalmente a grande maioria das empresas já sofreram algum prejuízo por negligenciar uma política de controle, segurança e disponibilidade da informação. É comum as grandes empresas perderem valor na bolsa de valores ou valor de mercado.

ETAPA 6 – Valor de Negócio do Dado

Valor de Negócio do Dado ou Valor da Informação será utilizado nesta fase as tabelas de negócio já identificadas na etapa 2 as que geram receita ou que geram perda de receita. Nesta fase também iremos utilizar a identificação do negócio que foi feita na Etapa 1.

Podemos citar um exemplo de um comércio eletrônico onde a identificação do negócio que no nosso *framework* fica na Etapa 1 é Comércio Virtual ou Comércio não presencial as tabelas de negócio que no nosso *framework* fica na etapa 2 são normalmente identificadas como tabela de Cliente, Vendas ou Faturamento.

A empresa também pode além do comércio virtual ter uma outra atividade de negócio como por exemplo a empresa presta serviços de consultoria isso outro negócio então poderia se calcular o valor de negócio do comércio virtual e da prestação de serviços pois são faturamentos distintos. É preciso considerar um tempo médio de receita pois, apenas um ano não reflete muitas vezes a realidade do quanto a empresa de fato fatura.

6 Estudo de Caso e suas Principais Características

O primeiro passo para iniciar uma pesquisa adotando a estratégia de estudo de caso é a definição do objetivo e da abordagem que normalmente são: qualitativa, quantitativa ou uma combinação das duas. O estudo de caso tem como propósito reunir informações detalhadas e sistemáticas sobre um fenômeno **[12]**.

O estudo de caso é uma investigação empírica que investiga um fenômeno contemporâneo dentro de seu contexto da vida real **[21]** adequado quando as circunstâncias são complexas e podem mudar, quando as condições que dizem respeito não foram encontradas antes, quando as situações são altamente politizadas e onde existem muitos interessados **[8].**

Para aplicar e validar este trabalho foi utilizada a técnica de estudo de caso exploratório em uma única empresa. Seguindo as diretrizes propostas por Runeson et al. **[15].** Este estudo possui uma abordagem qualitativa com estudos exploratórios, estudos exploratórios são todos aqueles que buscam descobrir ideias e soluções **[16],** abordando cada evento e situação de cada etapa nele proposto.

Este *framework* baseia-se nas características e objetivos dos *frameworks* de gerenciamento e gestão com a particularidade de envolver a área de negócio e ter como escopo empresas e organizações de trânsito.

6.1 Dados e Características da Organização deste Trabalho

A organização selecionada para este estudo é em uma organização pública de trânsito do estado de Pernambuco e envolverá as áreas de negócio e de banco de dados. Num primeiro momento será identificado o negócio ou atividade principal da organização.

Serão utilizados como amostragem deste estudo de caso para análise, os dados contidos nas tabelas de negócio do banco de dados da instituição coletadas via consultas *SQL*. Neste estudo a atividades principais da organização é de Serviços e Fiscalização de Trânsito para a população.

http://dx.doi.org/10.25286/repa.v3i3.962

A arrecadação é feita através destas duas atividades. Estas informações foram levantadas com as áreas de negócio e em consultas a missão e objetivo do órgão.

Também foi identificado uma divisão para os serviços pois possuem taxas de cobrança diferentes e as regras de negócio para o funcionamento delas também é diferente. Portanto estão divididas as áreas em Habilitação e Veículo, diferente de um comércio eletrônico ou um comércio tradicional pois basicamente o negócio é baseado no valor do produto.

As tabelas de dados de negócio neste caso que são identificadas na Etapa 2 do *framework*, as mais relevantes são: Tabela de USUARIO, TAXASSERVICOS, VEICULO, SERVICO, na grande maioria das empresas a tabela de USUARIO é a tabela de CLIENTE e a tabela de VENDAS, PRODUTOS são as tabelas de negócio.

No capítulo a seguir será descrito de forma detalhada todas as etapas do *framework* definido.

6.2 *Framework* (DAV) aplicado e validado no DETRAN-PE

A seguir a aplicação deste *framework* na Instituição Pública de Trânsito DETRAN-PE.

ETAPA 1 - **A Identificação do Negócio**

Nesta etapa foi identificado o negócio ou atividades de negócio do DETRAN-PE com entrevistasaos gestores das áreas de negócio. Foi identificado dois grandes pilares que norteia a área de negócio desta instituição que são elas: Área de Habilitação que arrecada com serviços de habilitação de trânsito para a população. Área de Veículo que arrecada com serviços de veículo, Licenciamento de Veículo e Infrações de Trânsito.

ETAPA 2 - **A Identificação dosDados de Negócio**

Os dados de negócio nesta etapa foram identificados com consultas às áreas de negócio com suporte do DBA. Para área de Habilitação foi identificado a tabela de **USUARIO** e na área de Veículo a tabela de **VEICULO**. Estas são as tabelas chave onde nasce toda arrecadação.

ETAPA 3 – **Crescimento dos Dados de Negócio**

O Código de Trânsito Brasileiro não impõe uma limitação de idade máxima para possuir uma habilitação que é revista com exames médicos em média entre 2 a 5 anos, portanto neste estudo consideramos os usuários habilitados até 70 anos de idade o que representa 94% da tabela de USUARIO os outros 6% consideramos *outliers* visto que o acesso dos mesmos aos serviços de condutores são inexpressivos, estas considerações foram feitas para área de Habilitação.

Para área de Veículo consideramos os veículos ativos, ou seja, que a situação esteja normal para abertura de serviços e cobranças de taxas e licenciamentos. Os veículos ativos representam 78% da base da tabela de VEICULO os 22% restantes são considerados veículos com situação que não geram serviços.

O Quadro 1 representa o Percentual Médio de Crescimento Anual dos Dados das Tabelas de Negócio.

Quadro 1 - Percentual Médio de Crescimento Anual dos Dados das Tabelas de Negócio

Percentual Médio de Crescimento Anual dos Dados das Tabelas de Negócio	
Área de Habilitação (USUÁRIO)	7,26%
Área de Veículo (VEÍCULO)	6,27%

ETAPA 4 – **O Valor do Dado pelo seu Custo**

Abaixo na Tabela 3 segue os custos do Departamento de Tecnologia da Informação do DETRAN-PE estimados para o ano de 2018.

Tabela 3 - Custo Mensal do Departamento de Tecnologia da Informação para o ano de 2018.

Custos	**IndicadorCusto**
Custos Fixos e Variáveis (Contratos de Pessoal + Treinamentos + Compra de Equipamentos para o departamento de Tecnologia da Informação	R$ 25.452.613,33 Ano R$ 2.121.051,11 Mês

Licenças de *Software*	R$ 9.696.588,81 Ano R$ 808.049,07 Mês
Licenças de *Hardware* caso existam	R$ 13.463.504,42 Ano R$ 1.121.958,70 Mês
Total	R$ 48.612.706,56

ETAPA 5 – Valor do Dado baseado em sua Indisponibilidade

Neste estudo de caso um exemplo de perda de receita por falta ou indisponibilidade da informação se deu da seguinte forma:

Por consultas à base de dados de Infração, a infração mais cometida nos últimos cinco anos foi: Transitar em velocidade superior à máxima permitida para o local em até 20% (vinte por cento), considerada uma infração Média que acarreta 4 pontos na Carteira Nacional de Habilitação o valor desta multa é de R$ 130,16.

A média mensal dos últimos cinco anos desta infração dá um valor de: 35.468,38 x R$ 130,16 = R$ 4.616.564,77 em arrecadação mês só da infração mais cometida, este valor é deixado de arrecadar no mês caso a organização não consiga coletar os dados das multas dos sensores e agentes de trânsito ou simplesmente não conseguir entregá-las aos destinatários.

Cálculo da Média Mensal

A média aritmética da quantidade de infração mais cometida nos últimos 5 anos. Isto foi feito através de consulta a base de dados.

Quadro 2 - Quantidade da Infração mais cometida por ano.

AnoQuantidade de Infrações	
2013	165.566
2014	305.315
2015	490.680
2016	638.597
2017	527.945

Na Equação (2) têm-se a Média dos últimos 5 anos da infração mais cometida.

Média Anual = (Soma dos Anos)/5

$$= \frac{(165.566 + 305.315 + 490.680 + 638.597 + 527.945)}{5} \quad \textbf{(2)}$$

Média Anual = 425.620,60

Na Equação (3) têm-se a Média Mensal da Infração mais cometida nos últimos 5 anos.

Média Mensal = Média Anual/12

Média Mensal = 425.620,60/12

Média Mensal = 35.468,38 **(3)**

ETAPA 6 – Valor deNegócio do Dado

A seguir o cálculo do valor do negócio baseado em nosso estudo de caso que possui mais de um tipo de negócio.

Negócio Área Habilitação

Arrecadação com Serviços

Tabela 3 - Demonstrativo da Média de Arrecadação com Serviços dos Últimos 5 Anos.

Média Arrecadação	Indicador Média
Média de Arrecadação com Serviços últimos 5 anos	A
Média Estimada Arrecadação com Serviços últimos 5 anos	B

Na Equação (4) temos a Média de Arrecadação com Serviços últimos 5 anos:

A = Média de Arrecadação com Serviços dos Últimos 5 anos = (Σ das Arrecadações dos últimos 5 anos)/5 **(4)**

Na Equação (5) temos a Média Estimada de Arrecadação com Serviços últimos 5 anos:

B = Média Estimada de Arrecadação com Serviços últimos 5 anos = Média de Arrecadação dos Últimos 5 anos + (Média de Arrecadação dos Últimos 5 anos * Percentual Médio de Crescimento Anual da tabela de Negócio).

http://dx.doi.org/10.25286/repa.v3i3.962

Média Estimada de Arrecadação com Serviços últimos 5 anos = A + (A * 7,26%)

(5)

Negócio Área Veículo

Arrecadação com Serviços, com Licenciamento e Infrações

Tabela 4 - Demonstrativo da Média de Arrecadação com Serviços, com Licenciamento e Infrações dos Últimos 5 Anos.

Média Arrecadação	Indicador Média
Média de Arrecadação com Serviços, Licenciamento, Infrações últimos 5 anos	W
Média Estimada Arrecadação com Serviços, Licenciamento e Infrações últimos 5 anos	X
Média de Arrecadação com Serviços, Licenciamento e Infrações últimos 5 anos	Y
Média Estimada Arrecadação com Serviços, Licenciamento e Infrações últimos 5 anos	Z

Na Equação (6) é demonstrada a Média de Arrecadação com Serviços dos Últimos 5 anos:

W = Média de Arrecadação com Serviços dos Últimos 5 anos = (Σ das Arrecadações dos últimos 5 anos)/5

(6)

Na Equação (7) é demonstrada a Média Estimada de Arrecadação com Serviços últimos 5 anos:

X = Média Estimada de Arrecadação com Serviços últimos 5 anos = Média de Arrecadação dos Últimos 5 anos + (Média de Arrecadação com Serviços dos Últimos 5 anos * Percentual Médio de Crescimento Anual da tabela de Negócio)

Média Estimada de Arrecadação com Serviços últimos 5 anos = W + (W * 6,27%)

(7)

Na Equação (8) é demonstrada a Média de Arrecadação com Licenciamento e Infrações dos Últimos 5 anos:

Y = Média de Arrecadação com Licenciamento e Infrações dos Últimos 5 anos = (Σ das Arrecadações dos últimos 5 anos)/5

(8)

Na Equação (9) é demonstrada a Média Estimada de Arrecadação com Licenciamento e Infrações últimos 5 anos:

Z = Média Estimada de Arrecadação com Licenciamento e Infrações últimos 5 anos = Média de Arrecadação dos Últimos 5 anos + (Média de Arrecadação com Licenciamento e Infrações dos Últimos 5 anos * Percentual Médio de Crescimento Anual da tabela de Negócio)

Média Estimada de Arrecadação com Licenciamento e Infrações últimos 5 anos = Y + (Y * 6,27%)

(9)

Na Equação (10) é demonstrado o Valor Total dos Dados da Instituição:

Valor Total dos Dados da Instituição = Σ (A,W,Y)

(10)

7 Trabalhos Relacionados

No trabalho de Pauline Glikman e Nicolas Glady**[5]** um Processo de 6 etapas é proposto para calcular o valor dos dados da empresa onde o dado é avaliado sob quatro perspectivas: pelos acionistas, ou seja, o quanto os acionistas de mercado estão dispostos a pagar pelos dados de clientes de uma empresa que toma decisões através dos dados, pela empresa considerando as receitas geradas pelos seus clientes, por avaliações do próprio cliente individualmente provocando questões do tipo até que ponto valorizamos nossos próprios dados, pela internet um mercado mundial de dados onde leva-se em questão a tendência do mercado de monetizar os dados oriundos da mesma.

O presente estudo difere do estudo supracitado porque é a proposta de um *framework* para auxiliar a calcular o valor dos dados. É também aplicado em um estudo de caso e mostrado os resultados obtidos.

Infonomia e o valor da informação na economia digital **[3]**observa que o valor da informação atende as normas de contabilidade, porém este ativo ainda não está sendo declarado em nenhum balanço contábil empresarial. Os

resultados deste estudo expandem a noção da infonomia e que esta, tem como objetivo mostrar os métodos de contabilidade bem como os fatores que afetam os ativos da informação e seu valor econômico.

O presente estudo tem por objetivo uma proposta de um *framework* para auxiliar a calcular o valor dos dados o que difere do estudo anterior "Infonomia e o valor da informação na economia digital" que tem por principal objetivo mostrar que o valor dos dados atende as normas de contabilidade.

O trabalho de Todd Tramba **[20]** apresenta um processo de 6 etapas para calcular o valor dos dados da empresa que são: 1-O valor do negócio, 2-O valor dos dados do cliente, 3-Definir o valor da empresa sem os dados, 4- Custo para reposição dos dados de negócio, 5- Tempo é dinheiro, 6- Tempo empregado para repor dados perdidos.

Apesar do estudo anterior relacionar seis etapas para se calcular o valor dos dados da empresa o que difere deste estudo que tem como proposta para calcular o valor dos dados, um *framework* que auxilie neste objetivo e a validação deste num estudo de caso.

O trabalho de Garifova L.F **[3]** apresenta um método que auxilia a contabilidade da informação. O autor relaciona como etapas deste método: O valor da informação (VI), O valor da informação para o negócio (VIB), Perda do valor da informação (LIV), Valor da produtividade da informação (VIP), Valor econômico da informação (EVI), Valor de mercado da informação (MVI) e cada etapa acima relacionada é calculada através de fórmulas e variáveis.

O que difere este estudo do estudo supracitado é que além do *framework* que auxilia a chegar no valor dos dados da organização é aplicado para validá-lo num estudo de caso e cada etapa para se obter o valor dos dados é bem distinta de um trabalho para o outro.

Obs.: Uma característica comum a todos os trabalhos relacionados é que nenhum deles cita a sua aplicação na prática nem mesmo utilizando um estudo de caso para tal experimento.

7.1 Resumos Esquemáticos dos Trabalhos Relacionados

A Tabela 5 relaciona um resumo esquemático dos trabalhos relacionados.

Tabela 5 - Resumo Esquemático dos Trabalhos Relacionados (continua)

Trabalho	Propósito	Aplicabilidade do Estudo para sua Avaliação	
Pauline Glikman, Nicolas Glady **[5]**	Avaliação do valor dos dados por quatro perspectivas (pelos acionistas, pela própria empresa, pelos próprios usuários, pela Internet um mercado mundial de dados)	Os Autores não citam aplicação do estudo.	Os autores não apresentam conclusão e resultados do estudo.

Tabela 5 - Resumo Esquemático dos Trabalhos Relacionados (continua)

Trabalho	Propósito	Aplicabilidade do Estudo para sua Avaliação	
Garifova L.F **[3]**	O autor observou que o valor da informação atende as normas de contabilidade e chegou à conclusão este ativo ainda não está sendo declarado em nenhum balanço contábil empresarial.	O autor não cita aplicação do estudo.	O autor sugere que sejam feitos mais estudos por organizações de vários setores distintos.
Todd Tramba**[20]**	Um processo de 6 passos para calcular o valor dos dados da empresa.	O autor não cita aplicação do estudo.	O autor não apresenta conclusão e resultados do estudo.
Laney, Gartner **[6]**	Um método que auxilia a	O autor não cita aplicação	O autor não cita aplicação

http://dx.doi.org/10.25286/repa.v3i3.962

	contabilidade da informação.	do estudo, mas sugere que sejam feitos mais estudos por organizações de vários setores distintos.	do estudo além do que se propõe o próprio método desenvolvido.
Framework o Dado como Ativo de Valor –DAV	Um framework para gerir os dados como ativo de valor..	O estudo foi validado mediante estudo de caso único em instituição pública de trânsito.	Resultado comparativo entre o valor do dado e o valor do hardware.

8 Conclusões e Resultados

Este presente estudo possibilitou o desenvolvimento de um *framework* para auxiliar a monetizar o valor dos dados, além disso, permitiu sua aplicabilidade em uma instituição pública de trânsito. Em seu desenvolvimento foi possível conhecer de fato a importância dos dados de negócio para a organização, permitiu que a organização conhecesse o valor dos seus dados.

Depois de concluído, o seu resultado foi útil para aprovação de orçamentos necessários para a manutenção e proteção dos ativos de dados, finalmente a organização teve conhecimento de que o orçamento para 2018 para proteção de sua informação representa apenas **1%** em relação ao valor monetário total dos dados, encontrado após aplicação deste *framework*.

A Tabela 6 demonstra a estimativa de crescimento em 5 anos do Valor dos Dados por Área de Negócio - Habilitação. Este percentual 6,27% foi encontrado na Etapa 3.

Tabela 6 - Estimativa de crescimento do Valor dos Dados em 5 anos – Negócio de Habilitação.

Ano	Valor dos Dados
2018	R$ 93.967.743,67
2019	R$ 99.859.521,20
2020	R$ 106.120.713,18
2021	R$ 112.774.481,89
2022	R$ 119.845.441,91

Cálculo da Estimativa de crescimento da Área de Negócio de Habilitação expressa na Equação (11):

Ano Seguinte = Valor dos Dados * 6,27% + Ano **(11)**

Exemplo: 93.967.743,67 * 6,27% + 93.967.743,67

A Tabela 7 demonstra a estimativa de crescimento em 5 anos do Valor dos Dados por Área de Negócio– Veículo. Este percentual 7,26% foi encontrado na Etapa 3.

Tabela 7 - Estimativa de crescimento do Valor dos Dados em 5 anos – Negócio de Veículo.

Ano	Valor dos Dados
2018	R$ 2.901.319.483,51
2019	R$ 3.111.955.278,01
2020	R$ 3.337.883.231,19
2021	R$ 3.580.213.553,78
2022	R$ 3.840.137.057,78

Cálculo da Estimativa de crescimento da Área de Negócio de Veículo expressa na Equação (12):

Ano Seguinte = Valor dos Dados * 7,26% + Ano **(12)**

Exemplo: 2.901.319.483,51 * 7,26% + 2.901.319.483,51

A Figura 2 e a Figura 3 representam o Valor da Informação aplicando o percentual médio de crescimento dos dados denegócio, ou seja, representa as médias estimadas de arrecadação numa perspectiva em cinco anos que corresponde aos valores contidos nas Tabelas 6 e 7, esta informação foi utilizada para o planejamento dos próximos orçamentos e foi fundamental para aprovação de todos os recursos financeiros pendentes de aprovação.

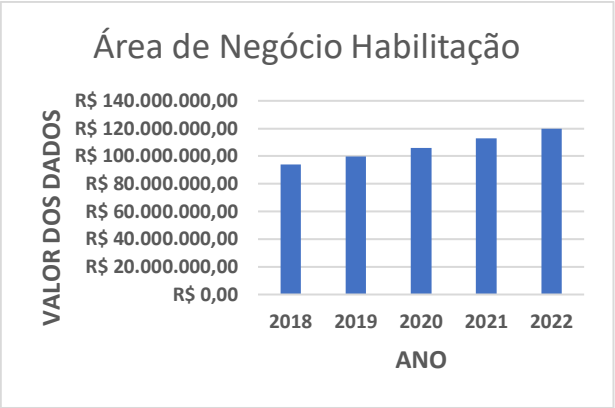


Figura 2: Crescimento Médio Anual Área de Negócio de Habilitação perspectiva em 5 anos

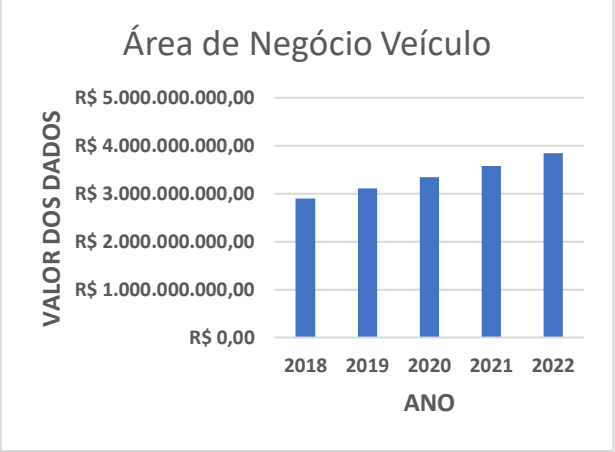


Figura 3: Crescimento Médio Anual Área de Negócio de Veículo perspectiva em 5 anos

A Tabela 8 mostra o Valor Total dos Dados encontrado na última etapa deste Framework e o Plano de Orçamento Anual (despesas para manter e proteger o Ativo de Dados) para o Ano de 2018 o custo para manter e proteger os dados representa 1% do seu valor.

Tabela 8 - Valor Total dos Dado X Orçamento Anual para o Ano de 2018.

Descrição	Valor
Valor Total dos Dados	R$ 4.087.304.268,13
Plano de Orçamento Anual 2018	R$ 48.206.901,23
Plano de Orçamento Ano 2018 em relação ao Valor dos Dados	**0,011794302 ≈ 1%**

A Figura 4 mostra o Valor Total dos Dados aplicando o fator de crescimento dos dados de negócio numa perspectiva em 5 anos.

Figura 4 - Valor Total dos Dados uma perspectiva em 5 anos.

Referências

[1] BEATH, Cynthia et al. Finding value in the information explosion.**MIT Sloan Management Review**, v. 53, n. 4, p. 18, 2012.

[2] GRECY, Nance; CHANG, Wo. **Big Data Interoperability Framework**: volume 1, definitions. Massachusetts: NIST, 2018. Disponível em: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf>. Acesso em: 10 abr. 2018 as 09h15min.

[3] GARIFOVA, L. F. Infonomics and the Value of Information in the Digital Economy. **Procedia economics and finance**, v. 23, p. 738-743, 2015. Disponível em:<https://www.sciencedirect.com/science/article/pii/S2212567115004232>. Acesso em: 13 de abril 2018 as 20h02min.

[4] GARTNER. **IT Glossary**: Infonomics. Disponível em:<https://www.gartner.com/it-glossary/infonomics>. Acesso em: 13 abr. 2018 as 10h45min.

[5] GILKMAN, Pauline; GLADY, Nicolas. What's the Value of your Data, **Techcrunch**, 2015.

http://dx.doi.org/10.25286/repa.v3i3.962

Disponível em: <https://techcrunch.com/2015/10/13/whats-the-value-of-your-data/>. Acesso em: 13 abr. 2018 15h37min.

[6] LANEY, Doug. Infonomics: the economics of information and principles of information asset management. In**:** INFORMATION QUALITY INDUSTRY SYMPOSIUM, 5., 2011, Massachusetts. **Proceedings...**Massachusetts: MIT, 2011. p. 2. Disponível em: <http://mitiq.mit.edu/IQIS/Documents/CDOIQS_201177/Papers/05_01_7A-1_Laney.pdf>. Acesso em: 14 abr. 2018 as 14h16min.

[7] LANEY, Douglas. **Infonomics**: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage. Reino Unido: Routledge, 2017.

[8] LLEWELLYN, Sue; NORTHCOTT, Deryl. The "singular view" in management case studies. **Qualitative Research in Organizations and Management: An International Journal**, v. 2, n. 3, p. 194-207, 2007.

[9] LONGMAN, Dictionary of America English (HardCover). 5 ed. Lognman Dictionaries, 2014.

[10] LUDÍCIBUS, S. et al. **Contabilidade Introdutória**. 11 ed. São Paulo: Editora Atlas, 2010.

[11] ODEH, Mohammed; KAMM, Richard. Bridging the gap between business models and system models.**Information and Software Technology**, v. 45, n. 15, p. 1053-1060, 2003.

[12] PATTOM, M. G. **Qualitative Research and Evaluation Methods**. 3 ed. Thousand Oaks, CA: Sage, 2002.

[13] REGAZZI, John J**. Infonomics and Values Creation in the New Business of Free**. Hershey: IGI Global, 2013.

[14] ROMER, Rafael. Empresas Brasileiras têm prejuízo de US$ 26 bi com perda de dados. **Canaltec**, 4 dez. 2014. Disponível em: <https://canaltech.com.br/seguranca/Empresas-brasileiras-tem-prejuizo-de-US-26-bi-com-perda-de-dados-aponta-estudo/>. Acesso em: 13 abr. 2018 as12h04min.

[15] RUNESON, Per et al. **Case Study Research in Software Engineering**: Guidelines and Examples. New Jersey: John Wiley & Sons, 2012.

[16] SELLTIZ, C.; JAHODA, M.; DEUTSCH, M. **Métodos de Pesquisa nas Relações Sociais**. São Paulo: EDUSP, 1974.

[17] SHEHABUDDEEN, N., PROBERT, D., PHAAL, R. Representing and approaching complex management issues: part 1 – role and definition. Working Paper. **Institute for Manufacturing, University of Cambridge**, 2000.

[18] SHORT, James; TODD, Steve. What's Your Data worth? **MIT Sloan Management Review**, v. 58, n. 3, p. 17, 2017. Disponível em: <https://sloanreview.mit.edu/article/whats-your-data-worth/> Acesso em: 5 jul. 2018 as 11h55min.

[19] TALLON, Paul P. Corporate governance of big data: Perspectives on value, risk, and cost.**Computer**, v. 46, n. 6, p. 32-38, 2013.

[20] TRAMBA, Todd, 2014. A 6 Step process to Calculate the Value of Company Date**. Teclink**, July 2014. Disponível em: <http://www.tccohio.com/blog/a-six-step-process-to-calculate-the-value-of-company-data>. Acesso em: 14 abr. 2018 as 11h19min.

[21] YIN, Robert K. **Estudo de caso**: planejamento e métodos. 3 ed.Porto Alegre: Bookman, 2005.

Uso de Técnicas de Clusterização em uma Base de Dados Financeira

Use of Clustering Techniques in a Financial Database

Armando Pereira Pontes Júnior[1] orcid.org/0000-0002-8212-4589
Clodomir Joaquim de Santana Junior[1] orcid.org/0000-0001-7869-7184
Carmelo José Albanez Bastos-Filho[1] orcid.org/0000-0002-0924-5341

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Pernambuco, Brasil.

E-mail do autor principal: appj@ecomp.poli.br

Resumo

O artigo tem como foco o uso de duas importantes técnicas computacionais para problemas de clusterização. Os algoritmos utilizados foram o *K-Means* e o *Fuzzy C-Means* (FCM), que aplicados em uma base de dados financeira de concessão de crédito pessoal podem auxiliar o tomador de decisão a identificar as principais características dos mutuários que se encontravam adimplentes e mutuários que estavam inadimplentes. O processo de clusterização investigou, através de 15 características (divididas entre características pessoais, condições de emprego e renda e condições da operação de crédito), similaridades que pudessem ajudar na formação de k grupos distintos. O resultado demonstra que as técnicas de agrupamentos aplicadas podem ser eficientes como ferramentas complementares para auxiliar o gestor financeiro nas suas atividades de classificação de risco, tomada de decisão e gerenciamento do crédito.

Palavras-Chave*:* Clusterização; Adimplência, Inadimplência; *K-Means*, *Fuzzy C-Means*.

Abstract

The article focuses on the use of two important computational techniques for clustering problems. The algorithms used were the K-Means and the Fuzzy C-Means (FCM), which have a financial database of credit granting, which help the decision maker to identify the main borrowing factors that were in arrears and borrowers that were defaulting. The clustering process investigated, through 15 characteristics (personal comparisons, income subsidies and operating conditions), similarities that can help in the formation of work groups. The following evidence that will be stored in companies in the activities of the date will be such as classifications to backup the risk and risk.

Key-words: *Clustering; Payment, Default; K-Means, Fuzzy C-Means.*

http://dx.doi.org/10.25286/repa.v3i3.976

1 Introdução

1.1 Intermediação Financeira

O nível de atividade econômica de um país é essencial para o crescimento, para a geração de emprego e renda e para a melhoria das condições socioeconômica de sua população. A atividade produtiva é altamente dependente, por um lado, dos investimentos realizados pelas empresas na produção de bens e serviços, e por outro lado, do consumo, que tem sua maior fatia realizada pelas famílias. Dados do IBGE - Instituto Brasileiro de Geografia e Estatística revelam que no ano de 2017 o consumo das famílias foi responsável por 63,43% do PIB, quando consideramos o cálculo do PIB pela renda **[1]**.

A ausência de capital, parcial ou total, dos agentes econômicos na demanda real ou latente, por consumo dos bens e serviços ofertados pelo mercado é facilmente verificado. É neste contexto de desequilíbrio entre os agentes superavitários e os agentes deficitários que surge a importante figura do crédito.

As instituições financeiras atuam como agentes de intermediação financeira no mercado, captando recursos juntos aos investidores pessoas físicas, empresas e Governos que possuem fundos excedentes e canalizam àqueles que necessitam de recursos para financiar seu déficit orçamentário **[2]**. A intermediação financeira é uma atividade que requer algumas condições básicas, tais como a existência de moeda, a consolidação de uma base legal e institucional e a existência de agentes econômicos superavitários e deficitários. Atendidas as duas primeiras condições, as instituições financeiras prestam o papel de intermediação entre os agentes econômicos, o primeiro, que possuem recursos financeiros em abundâncias e estão dispostos a emprestar, e o segundo, que necessitam de aportes extras para equilibrar suas finanças ou fazer frente a novos investimentos.

1.2 Risco de Crédito

Por definição, crédito é todo ato de vontade de alguém (pessoa física, jurídica ou Governo) em ceder, temporariamente, parte de seu patrimônio a um outro (pessoa física, jurídica ou Governo) com a expectativa de que essa parcela do patrimônio volte a sua posse de forma integral, acrescida de remuneração, e que seja feito no tempo aprazado **[3]**. De forma mais simples, crédito é o ato de entregar um certo valor mediante promessa de pagamento futuro de um montante maior ao que foi emprestado.

Entretanto, em toda operação de crédito é inerente a figura do risco, algo a ser administrado, mas nunca poderá ser completamente eliminado. A melhor forma de gerenciar a concessão de crédito é através da elaboração de mecanismos de identificação, mensuração e classificação de riscos, de uma forma que o tomador de decisão possa usufruir de ferramentas que o ajude a minimizar o risco de crédito.

Assim, este artigo investigará duas importantes técnicas de agrupamentos que podem servir de ferramentas complementares na tarefa de administração do risco de crédito.

Este artigo encontra-se dividido em seis seções. Além da introdução, a segunda seção trata de forma simples duas importantes técnicas de *clusterizaçã*o e detalha os dois algoritmos utilizados no artigo, mostrando a lógica de suas execuções. A terceira seção descreve em detalhes a base de dados financeira usada para o estudo de caso e como foi feito o pré-processamento. Na quarta seção são explicadas as métricas de validação usadas na execução dos algoritmos e também são apresentados os resultados numéricos. Na quinta seção é apresentada uma análise comparativa dos resultados. E por fim, a sexta e última seção discorre sobre as conclusões e as possíveis contribuições futuras.

1 Clusterização

2.1 Teoria

Os aumentos consideráveis na geração de dados demandam cada vez mais o uso técnicas que são capazes de realizar a extração do conhecimento de forma eficiente e automática.

Desta forma, podemos definir *clusterização* como processos computacionais muito utilizados em *Data Mining, e* que são bastante úteis na resolução de problemas de classificações e agrupamentos de conjunto de dados **[4]**.

A análise de clusters envolve a organização de um conjunto de padrões, comumente

representado na forma de vetores de atributos ou pontos em um espaço multidimensional, em grupos de acordo com uma medida de similaridade.

Existem diversas medidas de similaridade, a depender da natureza do problema que estamos a tratar. Assim, a formação dos grupos dependerá exclusivamente dos critérios de similaridade pré-definidos e da escolha das técnicas a serem utilizadas.

Neste artigo serão utilizadas duas técnicas bastante difundidas, *K-Means* e o *Fuzzy C-Means* (FCM). Depois serão comparadas as respectivas respostas de cada algoritmo para o problema de partição de uma base financeira que apresentam dados de bons (adimplentes) e maus (inadimplentes) pagadores.

2.2 *K-Means*

É um algoritmo do tipo não supervisionado proposto por MacQueen em 1967 **[5]**. Ele é muito eficiente e ao mesmo tempo de simples execução, o que o faz ser bastante utilizado para resolver o bem conhecido problema de clusterização. Em 2009, o *K-Means* foi considerado um dos dez algoritmos mais importante no campo da mineração de dados, considerando, como já mencionado, sua simplicidade mais também a sua escalabilidade. O *K-Means* possui complexidade ***O*** (t. N. K), onde t é a quantidade de iterações, N é a quantidade de objetos e K é a quantidades de grupos. Portanto, a complexidade é linear para qualquer variável do problema. Todavia, o algoritmo apresenta algumas restrições quando aplicado em bases mais complexas. Um dos problemas é fato da escolha inicial dos centroides (a lógica do algoritmo será melhor detalhada no próximo tópico) poder interferir nas soluções apresentadas. Isso o leva muitas vezes a obter soluções convergidas para ótimos locais.

2.3 Lógica do *K-Means*

O algoritmo se baseia na subdivisão de um conjunto de dados em k subgrupos, onde cada observação pertencerá apenas e, somente apenas, a um único subgrupo. Assim, dado um conjunto de observações $(x_1, x_2, \ldots, x_n)$ em que cada x_i tem d-dimensões (a dimensão representa a quantidade de características da observação x_i), o algoritmo *K-Means* particionará as n observações em k subgrupos ($2 \leq k \leq n$), que guardem o maior grau de semelhança entre suas observações.

Este artigo utilizará a distância euclidiana como a função de similaridade. Assim, dadas duas observações quaisquer, a distância euclidiana entre elas é calculada de acordo com a equação:

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{d} (x_{i,p} - x_{j,p})^2} \quad \textbf{(1)}$$

O passo seguinte é associar cada observação ao centroide mais próximo. Quando nenhuma observação estiver pendente, a primeira iteração estará concluída. O passo seguinte é recalcular a posição dos centroides tornando-os baricentros dos subgrupos definidos na etapa anterior. O cálculo leva em consideração as médias das distâncias euclidianas, novamente de acordo com a equação (1), das observações dentro de cada grupo. O recalculo da posição dos centroides é dado por:

$$C_k = \frac{1}{n_k} . \sum_{i=1}^{n_k} x_i^k \quad \textbf{(2)}$$

Onde C_k representa o baricentro do grupo K e n_k é o total de observações associadas ao cluster K. A partir daí o algoritmo entra na sua regra de *loop*, até que um dos critérios de parada ocorra. Os critérios mais comuns para interromper as iterações são: a) não haja mais mudança na posição dos centroides e b) se atinja o número máximo de iterações predefinidas pelo usuário.

O Algoritmo 1 abaixo representa, de forma simplista, os passos para implementação do *K-Means* clássico, onde a base de dados com *i* observações, *d* dimensões e *K* centroides são os parâmetros de entrada do algoritmo.

http://dx.doi.org/10.25286/repa.v3i3.976

Algoritimo 1:	K-means Clássico.
Entrada:	Conjundo de daddos, K centroides.
Saída:	Base de dados dividida em K grupos.
1.	Escolher K centroide aleatoriamente.
2.	Calcular a distância de cada objeto aos centroides, conforme equação (1).
3.	Atribuir cada objeto ao seu centroide mais próximo.
4.	Atualiza a posição dos centroides, conforme equação (2), para as médias das instâncias de cada grupo.
5.	Repete passos 2, 3 e 4 até que nenhum centroide mude de posição ou se atinja o número máximo de iterações.
6.	FIM.

2.4 *Fuzzy C-Means* (FCM)

Alguns problemas envolvem grupos mais delineados que não podem ser separados de uma maneira *hard*, como é feito no *K-Means.* Em outras palavras, há situações em que as categorias se sobrepõem umas às outras e em diferentes níveis. Nesses casos pode-se recorrer a lógica *fuzzy*, ou seja, em alguns agrupamentos as observações pertencem a todos grupos, com diferentes graus ou níveis de pertinência, assumindo valores contínuos de pertinência (o que contrapõe a lógica binária do *K-Means*). Assim, tratando da possibilidade de partição com sobreposição (*overlapping,* em inglês) diversos algoritmos foram sendo apresentados, dentro os quais o *Fuzzy C-Means* (FCM) que foi introduzido em 1984 por Bezdek **[6]**.

2.5 Lógica do *Fuzzy C-Means* (FCM)

Quando um algoritmo fuzzy é aplicado a um conjunto de dados, o resultado é uma matriz *fuzzy* de modo que:

$$\begin{cases} P = [p_{i,j}] \\ p_{i,j} \in [0,1] \end{cases} \quad \textbf{(3)}$$

Onde a P é uma matriz de dimensão K x N, sendo que o *K* representa a quantidade de grupos e o *N* a quantidade de objetos. O valor de cada $p_{i,j}$ é o grau de associação do *j-ésimo* objeto ao *i-ésimo* grupo *fuzzy*. Assim, todos os objetos possuem algum grau de pertinência com todos os grupos, inclusive pertinência de valor nulo. Para se executar o algoritmo deve-se seguir as seguintes etapas:

1. Defina-se o número *c* de grupos *fuzzy*, com $2 \leq c \leq n$;
2. Defina-se o valor do coeficiente de "*fuzzificação" m*;
3. Inicializa-se a matriz de pertencimento $P^{(0)}$ com valores aleatórios;
4. Calcula-se os centroides de cada grupo *c*, com a seguinte equação:

$$c_{i,j} = \frac{\sum_{c=1}^{n}(p_{i,j})^m x_i}{\sum_{c=1}^{n}(p_{i,j})^m} \quad \textbf{(4)}$$

5. Calcula-se a distância euclidiana $D_{i,j}$, como mostrado na equação (1), entre cada ponto *i* para cada centroide *j*;
6. Atualiza-se os valores $p_{i,j}$ da matriz de pertencimento P, conforme a seguinte equação:

$$p_{i,j} = \left[\sum_{c=1}^{c}\left(\frac{D_{i,j}}{D_{c,j}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad \textbf{(5)}$$

7. Volta-se a executar as etapas de 4 a 6 até que o módulo da diferença entre as duas matrizes de pertencimento, a atual e a da iteração anterior, seja menor que um coeficiente de erro εdefinido pelo usuário. Formalmente:
8.

$$\| P^t - P^{t-1} \| < \varepsilon \quad \textbf{(6)}$$

Onde o (*t*) representa a iteração atual, (*t-1*) a iteração imediatamente anterior.

3 Base de Dados

A base de dados apresentada em sua forma original contava com 40.320 registros de operações financeiras de crédito pessoal. Cada registro continha 21 características conforme apresentado na Tabela 1.

Tabela 1. Dicionário de dados (original)

ID	ID do mutuário
Setor de Atividade	Atividade que mutuário exerce
Data de Nascimento	Data de nascimento, dd/mm/aaaa
Data da proposta	Data da primeira porposta
Estado Civil	Estado civil
Sexo	Sexo
Grau de Instrução	Grau de Instrução
CEP	CEP residencial
UF_Endereço	UF residencial
DDD Residencial	Código de DDD residencial
Tipo Tel. Residencial	Tipo de telefone que possui na residência
Tipo de Residência	Situação e tipo da sua residência
Classe Profissional	Categoria profissional
Renda	Renda salarial do mutuário
Outras Rendas	Demais renda, exceto salarial
Dependentes	Número de dependentes
Renda do Cônjuge	Renda (diversas) do cônjuge
Produto Comercializado	Identificação do produto comercializado: cartão de crédito, CDC, cheque especial
Idade	Idade do mutuário (em anos)
Tempo de Emprego	Tempo no emprego (em anos)
Alvo	Identficação da situação do contrato: adimplência/inadimplência

3.2 Pré-processamento

Com objetivo de tratar, limpar, organizar e melhorar a apresentação dos dados foi feito um pré-processamento visando a remoção das observações onde verificamos total ausência de informação ou valores incoerentes com atributo. À título de exemplo, foram retiradas as observações que possuíam valores negativos para o atributo Renda ou valores acima de 120 (anos) para o atributo Idade.

Posteriormente foram feitas as remoções dos atributos ID, Setor de Atividade, Data de Nascimento, CEP, UF_Endereço, DDD Residencial e Tipo Tel. Residencial. O primeiro atributo removido foi o ID, que apenas identificava a quantidade de registros existente na base (40.320). Os atributos Setor de Atividade e CEP apresentavam uma gama diversa de informações que por si só não agregavam muito valor a base (particularmente em relação ao atributo Setor de Atividade, verificou-se que a mesma informação estava registrada com escritas diversas ou com uso de abreviações). Já o atributo Data de Nascimento foi removido uma vez que guardava redundância com o atributo Idade, sendo este último mais fácil de manusear. E de forma empírica, por entender que esses atributos não trariam ganhos de informação à pesquisa, foram feitas as remoções dos atributos CEP, UF_Endereço, DDD Residencial e Tipo Tel. Residencial.

O próximo passo foi categorizar cada atributo que se apresentava como *string*, dando as possíveis classificações do atributo um valor discreto. À título de exemplo, o atributo Sexo que podia ser classificado em "masculino" ou "feminino" foi reclassificado com os valores 1, para feminino e 2, para masculino. E ainda foi criada um atributo chamado de Idade2 que categorizava o valor discreto de cada idade dos mutuários em faixas etárias (até 25 anos, de 25 a 35 anos, de 35 a 45 anos, de 45 a 65 anos, acima de 65 anos).

Por fim, após todos os atributos já se apresentarem na forma numérica, foram feitas as normalizações para o intervalo de [0,1].

Ao concluir a fase de pré-processamento, a base de dados que foi submetida a execução dos algoritmos contava agora com 28.700 registros e 15 características.

4 Resultados

4.1 Métricas - Definições

Os resultados preliminares foram submetidos ao crivo das quatro métricas, no intuito de verificar a qualidade dos resultados e também o valor ideal de k. As observações foram feitas em cada uma das categorias do atributo Alvo (adimplentes e inadimplentes) e, também, por algoritmo executado. As métricas utilizadas foram as seguintes:

Estatística GAP: tem por objetivo encontrar um número ideal de *cluster*. Seu cálculo é feito pela diferença do logaritmo da distância *intra-cluster* do grupo analisado e de um conjunto de dados aleatório. Trata-se, portanto, de maximizar a diferença entre as distâncias do agrupamento

http://dx.doi.org/10.25286/repa.v3i3.976

escolhido e de um agrupamento aleatório. O objetivo é mostrar que o agrupamento escolhido é diferente de um aleatório.

Distância *Intra-Cluster*: é utilizada para calcular a distância de duas observações pertencentes ao mesmo cluster. Sua otimização se dar quando os valores são baixos, que indica proximidade das observações dentro do cluster.

Distância *Inter-Cluster*: é a métrica utilizada para calcular a distância entre dois centroides. Seu valor de otimização se dar quando os valores crescem, que demonstram que os centroides são realmente díspares.

Erro Quantizado: uma forma de avaliar a quantização do espaço obtido mediante a aplicação de um algoritmo de agrupamento é a lógica desta métrica. Ela está baseada no cálculo da média das distâncias entre os dados e o vetor que representa a região onde eles estão localizados. É uma métrica que avalia a eficiência do algoritmo para valores crescentes de K. Ela é otimizada quando se têm valores baixos.

4.2 Resultados das Métricas

As Tabelas 2 e 3 apresentam os resultados das simulações dos dois algoritmos (*K-Means* e FCM) para os dois grupos pesquisados (adimplentes e inadimplentes). Para que os resultados apresentassem consistência estatística, foram feitas 30 execuções para cada quantidade de *K* desejado nos dois algoritmos e por cada categoria do atributo Alvo (adimplentes e inadimplentes). Os valores escolhidos para K partiram de 2 até 10 *clusters*.

A escolha pelo número ideal de K dos grupos será feita observando primordialmente a estatística GAP, visto que essa métrica é bastante eficiente para escolha da melhor quantidade de grupos **[7]**.

Tabela 2. Resultados dos algoritmos utilizando 30 simulações e condição de parada 1.000 iterações . Resultados para o grupo de adimplentes

Algoritmo-K	GAP	Distância Intra-Cluster	Distância Inter-Custer	Erro Quantizado
K-Means - 2	**0,185 (0,063)**	15281,714 (146,79)	2,256(0,125)	0,873(0,009)
K-Means - 3	0,317(0,145)	13779,275 (486,18)	7,33(0,570)	0,77(0,011)
K-Means - 4	0,456(0,109)	12848,772 (439,14)	15,513(1,210)	0,72(0,023)
K-Means - 5	0,516(0,112)	12025,928(197,20)	27,809(1,764)	0,677(0,015)
K-Means - 6	0,560(0,132)	11617,746(265,07)	42,896(1,977)	0,658(0,013)
K-Means - 7	0,586(0,112)	11136,261(209,23)	60,838(2,879)	0,632(0,014)
K-Means - 8	0,634(0,130)	10710,921(144,57)	82,663(3,257)	0,612(0,011)
K-Means - 9	0,643(0,120)	10459,493(153,54)	109,637(3,555)	0,596(0,014)
K-Means - 10	0,690(0,107)	**10269,051(180,47)**	**138,55(6,133)**	**0,584(0,012)**
FCM - 2	0,182 (0,056)	16109,797(1,854)	0,298(0,001)	0,920(0,000)
FCM - 3	**0,409(0,061)**	14290,551(1,338)	2,308(0,306)	0,825(0,000)
FCM - 4	0,381(0,075)	14263,469(15,246)	4,017(0,039)	0,839(0,000)
FCM - 5	0,416(0,081)	14214,267(8,488)	6,891(0,471)	0,810(0,000)
FCM - 6	0,383(0,091)	14190,755(38,057)	9,327(0,341)	0,845(0,009)
FCM - 7	0,476(0,085)	13902,936(435,158)	14,793(2,408)	0,815(0,12)
FCM - 8	0,532(0,014)	13566,182(370,003)	21,394(2,876)	0,786(0,048)
FCM - 9	0,534(0,112)	13506,843(216,494)	26,372(3,343)	**0,775(0,028)**
FCM - 10	0,519(0,095)	**13453,251(180,276)**	**32,675(0,400)**	0,807(0,027)

Tabela 3. Resultados dos algoritmos utilizando 30 simulações e condição de parada 1.000 iterações . Resultados para o grupo de inadimplentes

Algoritmo-K	GAP	Distância Intra-Cluster	Distância Inter-Custer	Erro Quantizado
K-Means - 2	0,224(0,120)	8376,685(475,019)	1,978(0,160)	0,752(0,045)
K-Means - 3	**0,338(0,107)**	7661,666(89,506)	5,893(0,793)	0,695(0,034)
K-Means - 4	0,353(0,141)	7328,630(120,082)	12,012(1,039)	0,653(0,027)
K-Means - 5	0,499(0,105)	6907,897(117,411)	22,839(2,198)	0,632(0,019)
K-Means - 6	0,517(0,139)	6636,410(97,191)	35,001(3,300)	0,613(0,018)
K-Means - 7	0,591(0,131)	6385,989(77,171)	50,334(4,014)	0,594(0,016)
K-Means - 8	0,635(0,128)	6167,879(75,412)	70,889(4,865)	0,578(0,020)
K-Means - 9	0,708(0,116)	6002,152(107,063)	94,405(6,328)	0,567(0,014)
K-Means - 10	0,687(0,129)	**5842,265(90,158)**	**120,511(8,965)**	**0,549(0,013)**
FCM - 2	0,277(0,053)	8376,470(0,574)	0,952(0,001)	0,752(0,000)
FCM - 3	**0,346(0,057)**	8178,752(49,754)	1,955(0,066)	0,750(0,026)
FCM - 4	0,413(0,099)	7484,355(1,717)	5,730(0,009)	**0,673(0,000)**
FCM - 5	0,435(0,087)	7454,696(29,490)	8,488(0,156)	0,721(0,014)
FCM - 6	0,457(0,083)	7364,961(35,215)	12,593(0,903)	0,709(0,041)
FCM - 7	0,446(0,089)	7326,424(27,724)	17,193(0,117)	0,720(0,016)
FCM - 8	0,439(0,098)	7280,361(41,205)	22,866(0,754)	0,725(0,015)
FCM - 9	0,439(0,113)	7282,343(49,161)	28,452(0,169)	0,742(0,016)
FCM - 10	0,391(0,097)	**7249,279(52,416)**	**34,981(1,074)**	0,755(0,019)

5. Análise dos Resultados

5.1 Resultados do *K-Means* e do FCM Para o Grupo dos Adimplentes

Os resultados para o grupo de bons pagadores estão descritos na Tabela 2. Os valores se mostram consistentes com as métricas estabelecidas. Houve uma maximização da distância *inter-cluster* e uma minimização da distância *intra-cluster* nos dois algoritmos, à medida que o número de K aumentou. O erro quantizado também foi atendido, uma vez que houve diminuição desta métrica a medida que o número de cluster aumentava. Porém, houve uma pequena diferença, enquanto que no *K-Means* este valor se mostrou ótimo para K = 10, no FCM foi quando K atingiu o valor 9.

Com relação ao número de cluster ideal apontado principalmente pela métrica estatística GAP, o *K-Means*, maximizou no k=2 enquanto que o FCM foi no K=3.

Analisando uma amostra agrupada foi possível investigar de forma mais pormenorizada o perfil dos mutuários bons pagadores na criação, pelo K-Means, de dois *cluster*. Verificou-se que as características mais relevantes para formação do grupo foram "tempo de emprego", "prazo da proposta" e "produto contratado".

Já em relação ao FCM, como já descrito, o número ideal de cluster foi quando o K atingiu o valor 3. Verificando-se numa amostra agrupada que as características mais importantes para formação dos cluster foram "sexo", "produto contratado" e "prazo da proposta".

5.2 Resultados *K-Means* e do FCM Para o Grupo dos Inadimplentes

Os resultados para o grupo dos inadimplentes estão exibidos na Tabela 3. As métricas da distância *intra-cluster* e da distância *inter-cluster*

http://dx.doi.org/10.25286/repa.v3i3.976

obtiveram os resultados esperados, ou seja, minimizaram o valor da distância da primeira métrica e maximizaram o valor da distância da segunda. As duas métricas indicaram o k ideal como sendo de valor 10.

Também para a estatística GAP houve uma convergência dos dois algoritmos para o valor ideal de K = 3.

Agora analisando uma amostra agrupada do cluster k = 3 do *K-Means*, verificou que a principal característica para formação do cluster foi o "sexo" e o "tipo de residência".

Por fim, já em relação ao FCM, na análise mais apurada do agrupamento para K=3, verifica-se que as principais características para formação dos grupos foram "Tempo no Emprego", "Idade2" e "Estado Civil".

5.3 Análise Comparativa das Características dos Grupos Gerados Pelos Algoritmos

Adotou-se também uma outra abordagem comparativa para entender como cada algoritmo classificou os subgrupos do ponto de vista dos atributos. Assim, foram feitas as comparações de cada atributo estabelecendo sua importância para criação do subgrupo comparativamente com que foi estabelecido no outro algoritmo.

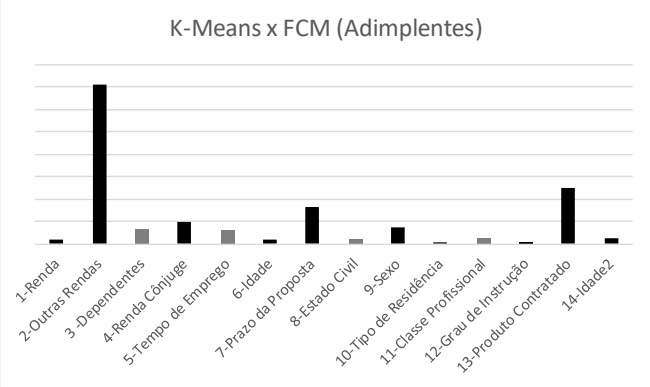


Figura 1: Comparativo das classificações.

A Figura 1 mostra como os algoritmos diferenciaram os 14 atributos da base de dados. Cada barra representa a diferença dada na importância dos atributos quando comparados *vis a vis* nos dois algoritmos. As barras na cor cinza, identificam que o FCM estabeleceu uma importância maior que o *K-Means* para um determinado atributo. As barras na cor preta dizem o contrário.

Esta mesma análise comparativa foi feita para o grupo de inadimplentes, como se pode observar na figura abaixo:

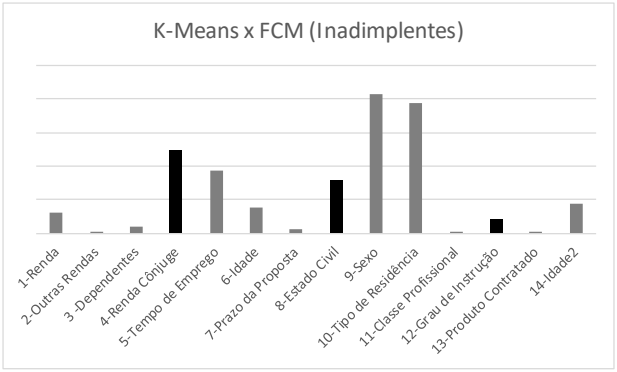


Figura 2: Comparativo das classificações.

Como se pode verificar, no grupo de adimplentes, os atributos que mais contribuíram para diferenciar o resultado do FCM do resultado do K-Means foram "Outras Rendas" e "Produto Comercializado". Enquanto que no grupo dos inadimplentes, os atributos "Sexo" e "Tipo de Residência" se mostraram como os mais relevantes para diferenciar os resultados obtidos em cada um dos algoritmos.

6 Conclusões

Este artigo analisou a aplicação de técnicas de *clusterização*, com a execução de dois importantes algoritmos: *K-Means* e *Fuzzy C-Means*. Foi utilizada uma base de dados financeira com informações sobre operações de crédito, dados pessoais e informações relativas à ocupação do mutuário. A base estava dividida em dois grupos, os adimplentes e os inadimplentes.

Assim, o objetivo era identificar perfis similares de mutuários adimplentes para que um gestor financeiro tivesse como estreitar a relação e potencializar os negócios. Bem como a importância de identificar perfis similares de mutuários inadimplentes para que o gestor fizesse uma administração mais próxima e cautelosa para com esse perfil de mutuário.

Do ponto de vista dos resultados apresentados pelos algoritmos, especificamente para o grupo das pessoas adimplentes, não foi possível compara-los. Enquanto que o *K-Means* retornou 2 como número ideal partição do conjunto de dados, o FCM fixou o número de cluster como

sendo 3. Contudo, ambos concluíram que as características "prazo da proposta" e "produto contratado" são bons rótulo para se fazer um agrupamento.

Já em relação ao grupo de inadimplentes, ambos os algoritmos chegaram ao número ideal de partição em 3 grupos. Porém, divergiram nas características que esses grupos têm que ser particionados. Podemos considerar que FCM mostrou uma performance levemente superior ao *K-Means*uma vez que primeiro apresentou um valor de estatística gap maior que a do segundo.

Como trabalhos futuros, poderíamos agregar esse trabalho na modelagem de um sistema de *credit score* como etapa de pré-processamento. Sistematizando que perfis de mutuários inadimplentes como encontrados neste estudo de caso pontuariam menos no sistema de *credit score*. Já os perfis de mutuários adimplentes teriam uma pontuação maior no sistema, uma vez que apresentam um risco menor de crédito.

Referências

[1] INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Quadros Completos: PIB 2017**. – IBGE. IBGE, 10 ABR. 2018. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-detalhe-de-midia.html?view=mediaibge&catid=2102&id=1800>. Acesso em 19.05.2018, 20:56:15.

[2] GITMAN, Lawrence J. **Princípios de Administração Financeira**. 7 ed. São Paulo: Habra, 2002.

[3] SCHICKEL, Wolfgang K. A**nálise de Crédito Concessão e gerência de empréstimos**. 5 ed. São Paulo: Atlas, 2000.

[4] ALAM, Shafiq et al. Research on particle swarm optimization based clustering: a systematic review of literature and techniques. **Swarm and Evolutionary Computation**, v. 17, p. 1-13, 2014

[5] MACQUEEN, James et al. Some methods for classification and analysis of multivariate observations. In: **Berkeley symposium on mathematical statistics and probability, 5**., 1967, Berkely. **Proceedings...** Berkely: University of California Press, 1967. p. 281-297.

[6] BEZDEK, James C.; EHRLICH, Robert; FULL, William. FCM: The fuzzy c-means clustering algorithm. **Computers & Geosciences**, v. 10, n. 2-3, p. 191-203, 1984.

[7] TIBSHIRANI, Robert; WALTHER, Guenther; HASTIE, Trevor. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 63, n. 2, p. 411-423, 2001.

http://dx.doi.org/10.25286/repa.v3i3.976

Especificação de um Repositório de Soluções Inovadoras para o Laboratório de Inteligência Governamental (LiGOV)

Specification of a repository of innovative solutions for the Governmental Intelligence Laboratory (LiGOV)

Eronita M. L. Van Leijden[1,2] orcid.org/0000-0002-8434-7954
Alexandre M. A. Maciel[1] orcid.org/0000-0003-4348-9291

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Agência Estadual de Tecnologia da Informação, Recife, Brasil,

E-mail do autor principal: emlvl@ecomp.poli.br

Resumo

Em busca de reduzir gastos de custeio, racionalizar despesas e aumentar a eficiência das ações nas políticas públicas, governos empreendem esforços no sentido de conectar as demandas da sociedade aos serviços prestados aos cidadãos. Em paralelo, a Universidade Estadual de Pernambuco (UPE) vem trabalhando numa metodologia de ensino que incentiva o desenvolvimento de trabalhos acadêmicos com problemas reais do Setor Público. Para maximizar uma integração e colaboração entre Universidade e Governo que se firmou um convênio na forma de um Laboratório de Inteligência Governamental (LiGOV). Neste contexto, várias soluções foram desenvolvidas e, por ainda não existir um ambiente físico adequado para armazenamento das soluções resultantes dos trabalhos acadêmicos, professores ficavam responsáveis pela guarda e compartilhamento dessas soluções. Sendo assim, este artigo objetiva especificar um repositório de conteúdo apropriado para o armazenamento de código-fonte das soluções, dos arquivos de documentação e de apresentação, além de viabilizar condições adequadas para o desenvolvimento dos trabalhos de pesquisa.

Palavras-Chave: Repositório; Laboratório de Governo; Gestão de Conhecimento;

Abstract

In order to reduce costs, rationalize expense and increase the efficiency of public policy will, governments are making efforts to connect society's demands with services provided to citizens. Concomitant, the University of Pernambuco State (UPE) has been working on a teaching methodology to encourage the development of academic work with real public sector problems. To maximize an integration and collaboration between University and Government an agreement was signed in the form of a Government Intelligence Laboratory (LiGOV). In this context, several solutions have been developed and, because there is not yet a suitable physical environment for storing the solutions resulting from academic work, teachers were responsible for the custody and sharing of these solutions. Thus, this article aims to specify an appropriate content repository for the solution source store, documentation and presentation files, and enables appropriate conditions for the development of work of research.

Key-words:*Repository; Government Laboratory; Knowledge Management;*

 http://dx.doi.org/10.25286/repa.v3i3.973

1 Introdução

Diante de cenários econômicos adversos é fundamental que governos de todas as esferas reduzam gastos de custeio, racionalizem despesas e aumentem a eficiência e a eficácia de suas ações nas políticas públicas. Situações deste tipo requerem dos gestores, além da realização de ajustes organizacionais, fortes investimentos em tecnologia da informação e comunicação **[1]**.

Em 2016, o tema geral do Fórum Econômico Mundial, realizado em Davos, foi "*Mastering the Fourth Industrial Revolution*" (i.e. dominando a quarta revolução industrial). Uma importante avaliação do momento tecnológico atual que remete a uma coleção de tecnologias empoderadoras da sociedade, e.g. Internet das Coisas, Big-Data, Computação em Nuvens, Inteligência Artificial (IA), dentre outras **[2]**.

Nesse sentido, cada vez mais, torna-se necessário que os governos empreendam esforços no sentido de conectar as demandas da sociedade aos serviços prestados aos cidadãos. Observa-se, no Brasil, que alguns governos têm investido fortemente em tecnologia de Big-Data e IA com objetivo de otimizar seus recursos humanos e de capital disponível **[3][4]**, bem como ofertar melhores serviços **[5]**.

Em Pernambuco, no âmbito da Administração Pública Estadual (APE) também tem se verificado essa tendência. Órgãos e entidades públicas já possuem um conjunto rico de dados e informações que são consumidos por relatórios gerenciais e ferramentas de análise de dados **[6]**. Além disso, outras iniciativas vêm empregando soluções de mineração de dados, e técnicas de inteligência artificial para apoiar o processo de tomada de decisão **[7]**.

Desde 2016, uma ação do curso de Engenharia da Computação da Universidade de Pernambuco (UPE) vem implantando uma metodologia de ensino-aprendizagem baseada em problemas, denominada Sala de Aula Aberta. Esta ação tem como objetivo além de melhorar a preparação dos jovens universitários para o mercado de trabalho, bem como desenvolver soluções a problemas reais tanto do mercado quanto do governo **[8]**.

Professores de duas disciplinas, essencialmente inovadoras, IA e Mineração de Dados (MD), em processo experimentativo, têm convidado gestores de órgãos públicos para participarem desta nova dinâmica de aula. Os gestores interessados têm a oportunidade de ampliar suas competências e habilidades, e contribuem no processo metodológico, fornecendo problemas complexos da administração pública. Ao final das disciplinas são disponibilizados aos gestores soluções inovadoras que se tornam ativos aplicáveis no órgão de origem.

Para maximizar os resultados desta metodologia, foi firmado um convênio entre a UPE e Agência Estadual de Tecnologia da Informação (ATI), cuja competência institucional é propor e prover soluções integradoras de meios, métodos e competências de Tecnologia da Informação para o Estado **[9]**. Essa parceria tem como finalidade estabelecer a política de integração e cooperação entre Governo e Universidade, na forma de um Laboratório de Inteligência Governamental (LiGOV).

Ao longo do ano de 2017 houveram quatro turmas na sala de aula aberta, com participação de nove gestores, de cinco órgãos distintos, que resultaram em 18 soluções de uso específico desses órgãos. As soluções entregues antes do LiGOV foram salvaguardadas pelos próprios professores de cada disciplina, que de forma particular, armazenou e, quando requisitado, compartilhou o resultado dos estudos entregues com outros órgãos do Estado que tivessem interesse.

Estando previstos mais 15 soluções em 2018, com perspectiva de inclusão de outras disciplinas neste novo modelo de metodologia de ensino, observou-se a necessidade de um ambiente de armazenamento apropriado que viabilize condições adequadas para o desenvolvimento dos trabalhos de forma organizada e colaborativa.

Este trabalho objetiva especificar um repositório de soluções inovadoras para o LiGOV. Para isto, faz-se necessária a realização de uma fundamentação teórica necessárias ao desenvolvimento do trabalho (seção 2), a proposição do repositório aderente aos requisitos técnicos e operacionais do LiGOV (seção 3) e por fim, a realização de uma avaliação dos resultados obtidos (seção 4). A seção 5 apresenta as conclusões e trabalhos futuros.

2 Fundamentação Teórica

Esta seção apresenta conceitos necessários para ao desenvolvimento deste trabalho. Na Seção 2.1 é apresentado o embasamento sobre o sistema de gestão de conhecimento, na seção 2.2 trata especificamente da gestão de conteúdo, e na seção 2.3 é feita uma contextualização a respeito do uso de repositórios aplicados à gestão do conhecimento.

2.1 Gestão de Conhecimento

A Gestão do Conhecimento (GC) refere-se ao conjunto de processos desenvolvidos em uma organização para criar, armazenar, transferir e aplicar conhecimento [9].

Das ações voltadas à GC dentro da organização, as mais utilizadas são:

- Benchmarking;
- Fóruns de discussão;
- Gestão de conteúdo;
- Gestão eletrônica de documentos;
- Mapeamento de conhecimentos, competências e processos.

Dalkir (2005) classifica o ciclo da GC em três etapas:

1. Captura e/ou criação do conhecimento;
2. Compartilhamento e disseminação do conhecimento;
3. Aquisição e aplicação do conhecimento.

Para cada etapa, existem vários tipos de ferramentas e técnicas que são usadas na Gestão do Conhecimento. Muitas destas são emprestadas de outras disciplinas e outras são específicas da GC **[11]**. A Tabela 1 apresenta os tipos de ferramentas classificadas de acordo com as etapas da GC.

Tabela 1 - Tipos de Ferramentas por etapas da GC.

Etapas da GC	Tipos de Ferramentas
Captura e/ou criação do conhecimento	Ferramenta de Criação de Conteúdo; Mineração de Dados e Descoberta de Conhecimento; *Blogs;* Ferramenta de Gerenciamento de Conteúdo; Repositório de Conteúdo
Compartilhamento e disseminação do conhecimento	*Groupware*e Colaboração *Wiki* Tecnologia de Rede
Aquisição e aplicação do conhecimento	Ferramenta inteligentes de filtragem Tecnologia adaptativa

Fonte: Dalkir**[11]**

Todos eles precisam ser combinados da maneira apropriada para atender a todas as necessidades da disciplina de GC, e a escolha de ferramentas a serem incluídas no kit de ferramentas de GC deve ser consistente com a estratégia geral de negócios da organização **[11]**.

2.2 Gestão de Conteúdo

A gestão de conteúdo é uma combinação de tecnologia e processos organizacionais: a tecnologia facilita a criação, o armazenamento e a disponibilidade do conteúdo, e os processos organizacionais são a essência para o sucesso da implantação **[12]**.

As ferramentas de criação e gestão de conteúdo mais usadas variam do geral (por exemplo, processamento de texto) ao mais especializado (por exemplo, gerenciamento de código fonte) **[11]**. Questões como qual conteúdo será armazenado e como as informações serão recuperadas são avaliados para especificação desse ambiente.

Características observadas nas ferramentas de gestão de conteúdo, como; gestão integrada, gestão do ciclo de vida, classificação automática, flexibilidade nas possibilidades de apresentação, controle de versão, equilíbrio entre centralização e descentralização e segurança e monitoramento, são aspectos analisados para se escolher a plataforma que apoie a gestão de conteúdo em uma organização **[13]**.

http://dx.doi.org/10.25286/repa.v3i3.973

2.3 Repositórios

As tecnologias destinadas à gestão de conteúdo consistem em: tipos de redes (intranets e extranets), repositórios de conhecimento, portais de conhecimento e espaços de trabalho compartilhados baseados na web.

Repositórios de conhecimento são como um repositório on-line de conhecimento, experiências e documentação sobre um determinado domínio de especialização baseado em computador. Tais repositórios são por vezes referidos como bases de experiência ou memórias corporativas **[11]**.

Na obra de Dalkir **[11]** encontram-se classificações de repositório de conhecimento de acordo com três aspectos; um quanto a pessoa que faz a inserção do conteúdo, outro quanto ao propósito do repositório e outro quanto aos elementos do conteúdo. A Tabela 2 apresenta as classificações de acordo com esses aspectos.

Tabela 2 - Classificação de Repositório.

Aspectos	Classificações
Pessoa que faz a inserção no repositório	- **Coleção passiva** -> onde os próprios trabalhadores reconhecem que conhecimento tem valor suficiente para ser armazenado no repositório; - **Coleção ativa** -> onde algumas pessoas da organização estão analisando os processos de comunicação para detectar conhecimento.
Propósito do repositório	- **Repositórios externos** (como inteligência competitiva); - **Repositórios interno estruturados** (como relatórios de pesquisa e material de mercado orientado para o produto); - **Repositórios internos informais** (como "lições aprendidas").
Elemento do conteúdo	- **Conhecimento declarativo** (por exemplo, conceitos, categorias, definições, suposições - conhecimento do que); - **Conhecimento processual** (por exemplo, processos, eventos, atividades, ações, manuais - conhecimento de como ou know-how); - **Conhecimento causal** (por exemplo, raciocínio para decisões, para decisões rejeitadas - conhecimento do porquê); - **Contexto** (por exemplo, circunstâncias de decisões, conhecimento informal, o que é e o que não é feito, aceito, etc. - conhecimento de cuidado - por quê).

Fonte: Dalkir **[11]**

Um repositório de conhecimento difere de um data warehouse e de um repositório de informações principalmente na natureza do conteúdo armazenado. O conteúdo do conhecimento consistirá tipicamente em conteúdo contextual, subjetivo e bastante pragmático. O conteúdo em repositórios de conhecimento tende a ser desestruturado (por exemplo, trabalhos em andamento, relatórios de rascunho, apresentações). Os repositórios de conhecimento também tendem a ser mais dinâmicos do que outros tipos de arquiteturas, porque o conteúdo do conhecimento é continuamente atualizado e fragmentado em perspectivas variadas para atender a uma ampla variedade de diferentes usuários e contextos de usuários. Para esse fim, os repositórios normalmente acabam sendo uma série de mini-portais vinculados distribuídos por uma organização **[11]**.

Dependendo das atividades da organização, ou da sua área de atuação, podem ser necessários mais de uma tecnologia. Em engenharia de software, por exemplo, o gerenciamento de documentos e a gestão de competências são atividades que precisam de sistemas de apoio à conteúdos distintos **[14]**.

Para Rus, Lindvall e Sinha **[15]**, os artefatos resultantes de tarefas de desenvolvimento de software representam os ativos de conhecimento explícitos da organização e devem ser gerenciados com eficiência. Eles não apenas servem ao projeto para o qual foram produzidos, mas também podem ser analisados para gerar novos conhecimentos à organização **[15]**.

3 Repositório Proposto

3.1 Necessidades Tecnológicas e Requisitos do Ambiente

Considerando as disciplinas inicialmente participantes do Programa Sala de Aula Aberta, IA e MD, foram identificados três formatos de conteúdo comuns: código-fonte, material acadêmico entregue (artigos e apresentações), e os dados disponibilizados pelos gestores.

Para cada um dos formatos identificados, foram identificados os requisitos junto aos professores das disciplinas para a implementação

do repositório do LiGOV. A Tabela 3 mostra esses requisitos.

Tabela 3 - Requisitos para implementação do LiGov.

Tipo	Requisitos
Conteúdo (Código-fonte, documentação e dados)	- ser um ambiente colaborativo (internet); - permitir um desenvolvimento distribuído; -não restringir a linguagem de programação em que as soluções são desenvolvidas; - Não deve restringir o formato ou tamanho do material; -permitir gerenciamento de permissões de compartilhamento das soluções; - deve haver controle de acesso em níveis específicos de usuários; - permitir a rastreabilidade de código se houver evolução da solução em fase posterior da sala de aula aberta, com possibilidade de fazer ramificações de código (desenvolvimento não linear).
Gestão (pesquisa, relatórios, estatísticas)	- apresentar uma opção para busca por características das soluções; - apresentar estatísticas descritivas das soluções implementadas; - inventariar as soluções desenvolvidas.

Para atender a esses requisitos, a implementação do Repositório do LiGOV foi proposta a utilização de três ferramentas de propósito específico: I) Controle de Versão, II) Catalogação de Soluções e III) Painel de Gestão.

3.2 Controle de Versão

Considerando os requisitos necessários para o controle de conteúdo, e sabendo que o resultado das disciplinas avaliadas para este artigo (IA e MD) são soluções de software, avaliou-se as ferramentas de gestão de conteúdo para controle de versão em código-fonte.

Subvertion (SVN), GitHub e o Git foram selecionadas para o estudo por serem ferramentas disponíveis na ATI, e com possibilidade de uso imediato pelo LiGOV. A Tabela 4 apresenta a análise comparativa entre as ferramentas mencionadas.

Tabela 4 - Análise Comparativa com Ferramentas

Critérios de Avaliação	Subversion	GitHub	Git
Permite Colaboração com disponibilidade de ferramenta de internet web na ATI	Não	Sim	Sim
Desenvolvimento Distribuído	Não	Sim	Sim
Edições de qualquer linguagem de programação	Sim	Sim	Sim
Edições de qualquer tipo de arquivo e tamanho	Sim	Sim	Sim
Gerenciamento de permissões de compartilhamento das soluções	Sim	Não	Sim
Controle de Acesso	Sim	Não	Sim
Desenvolvimento não linear	Não	Sim	Sim

Por atender a maioria dos critérios avaliados, demonstrado na Tabela 4, optou-se por utilizar a ferramenta Git.

Git é um sistema de controle de versões distribuídos, usado principalmente no desenvolvimento colaborativo de software, podendo armazenar e controlar vários artefatos resultantes dessa área, mas pode também ser usado para registrar o histórico de edições de qualquer tipo de arquivo **[16]**.

A instância do Git hospedada e gerenciada na ATI foi denominada Git-PE.

No início de cada semestre, os professores que estiverem com gestores de órgão público contribuindo nas aulas com problemas reais do seu órgão, encaminhará para o responsável do repositório na ATI as informações descritivas sobre os casos de estudo fornecidos pelo(s) gestor(es); informando uma breve descrição do problema, a classe (busca, otimização, classificação, agrupamento ou previsão), a técnica ou modelo que será trabalhado na sala e a relação dos alunos e gestores responsáveis por cada caso de estudo.

http://dx.doi.org/10.25286/repa.v3i3.973

De posse dessas informações, o responsável pelo repositório cadastra os alunos e gestores no Git-PE, cria o diretório de trabalho e configura as concessões de acesso para cada diretório de trabalho. Na Tabela 5 é apresentado as padronizações estabelecidas.

Tabela 5 - Padronizações estabelecidos no Git-PE.

Objetos	Padronizações
Nome do Diretório de trabalho (Nome do Projeto)	[Sigla Disciplina][Período Letivo]- [Ident da Equipe]-(se IA)[Classe do Problema]] (se MD)[Nomenclatura do caso de estudo] Ex.: IA20171-8-Otimização MD20172-1-AnaliseChamadosTelematicaSEE
Descrição do caso de Estudo (Descrição do Projeto)	[Faz uma breve descrição sobre o caso de estudo em 1 ou no máximo 3 linhas]. Órgão: [Nome do Órgão] (Se a disciplina for IA inserir também). Técnica: [técnica de IA aplicado] Ex: Sistemas de fronteira. Órgão: Secretaria da Fazenda - SEFAZ. Técnica: Árvore de Decisão Melhorar a eficiência nos atendimentos da Rede de Telemática da SEE. Órgão: Agência Estadual de Tecnologia da Informação de PE
Permissão de acesso de usuário	Alunos = "Developer" Gestor Público = "Master"
Arquivos	Na Raiz ter o README; Criar as estrutura de pastas: Código, Dados e Documentação
Permissão de acesso na linha de desenvolvimento	"Master" é protegido
Nível de Visibilidade do Conteúdo	Durante o semestre letivo, o nível de visibilidade é Privado. Ao finalizar a disciplina, o nível de visibilidade é alterado para Público ou Internal

No fim do semestre, assim que o professor sinalizar à ATI a conclusão da disciplina, tendo validado o material produzido na sala de aula aberta, a ATI realiza o *merge* do *branch* trabalhado pelos alunos.

Por fim, existe a alteração no nível de visibilidade do diretório de trabalho, conforme a padronização definida. A Figura 1 mostra o processo de inserção de conteúdo no Git-PE.

3.3 Catálogo de Soluções

O Catálogo de Soluções deve, além de viabilizar a estruturação das informações descritivas sobre os problemas trabalhados em sala de aula e suas respectivas soluções inventariadas, ser um instrumento de comunicação para envio dessas informações entre a POLI e ATI que viabilizam a criação inicial do diretório de trabalho no Git-PE.

Para inventariar as soluções ficou decidido que este catálogo deveria conter informações sobre: órgão que forneceu o caso de estudo, identificação do representante do órgão, classe do problema, técnica(s) utilizada na solução, relação dos alunos que trabalharam em cada solução, uma avaliação descritiva do representante do órgão quanto ao material entregue e as especificações técnicas do código. A Figura 2 mostra o diagrama de entidade e relacionamento.

No primeiro momento, por razões estratégicas para validação do modelo, decidiu-se utilizar planilhas eletrônicas para implementar este diagrama.

Para cada sala de aula, é criado um arquivo cuja nomenclatura foi padronizada como "Inventário_[sigla disciplina][período letivo]". Por exemplo, Inventario_MD20172 para guardar as informações descritivas dos problemas trabalhados na disciplina de Mineração de Dados do segundo semestre de 2017. A Figura 3 mostra a Tela Inicial do Formulário de Inventário.

Figura 3 - Tela Inicial do Formulário de Inventário das Soluções.
Fonte: O autor.

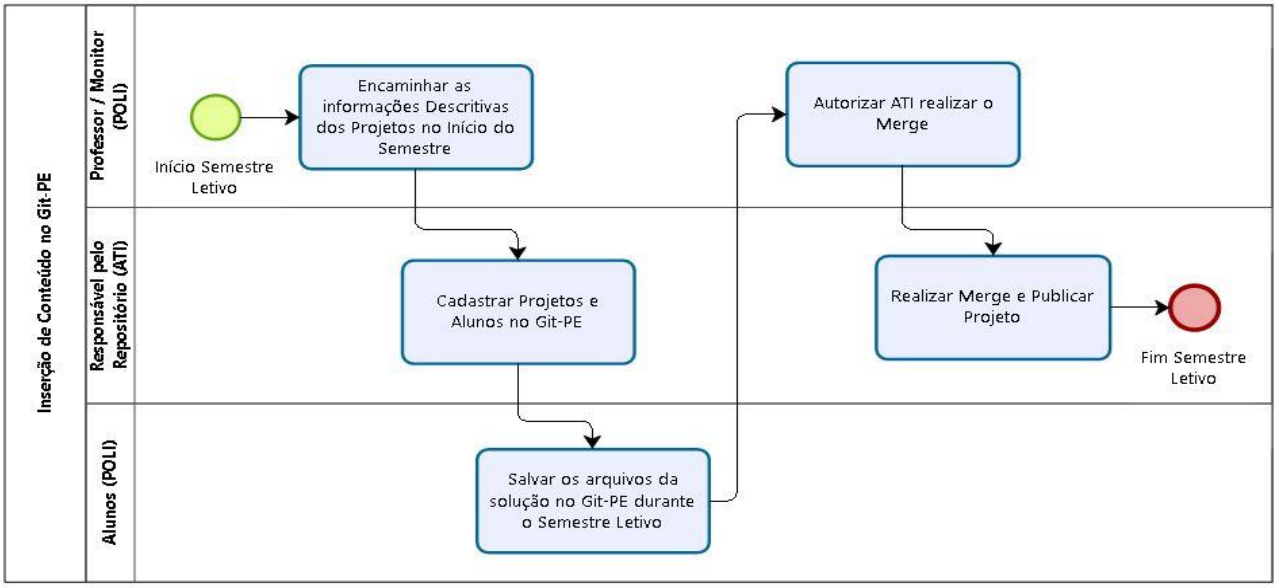


Figura 1 - Processo de inserção de conteúdo no Git-PE
Fonte:O autor.

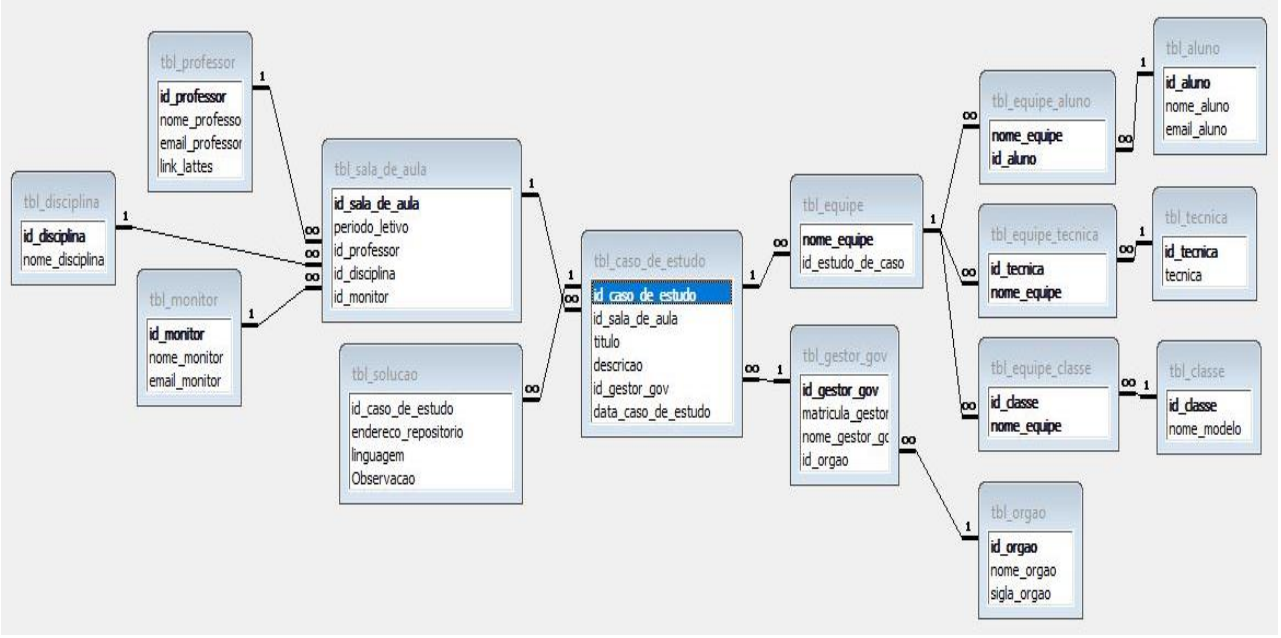


Figura 2 - Diagrama de Entidade e Relacionamento do Catálogo de Soluções
Fonte:O autor.

2.4 Painel de Gestão

Visando atender aos requisitos de busca e estatísticas descritivas, decidiu-se utilizar uma ferramenta de *Business Intelligence* (BI).

Segundo DRUZDZEL e FLYNN (2002), Sistemas de Apoio à Decisão, como é o caso de ferramentas de BI, fornecem armazenamento e recuperação de dados, em modelagem específica, que aprimora a forma de acesso aos dados e a recuperação das informações **[17]**. Este tipo de ferramenta é caracterizado ainda por possibilitar visões integradas de várias fontes de informação, de forma agregada e única.

No intuito de aproveitar as licenças disponíveis na ATI, optou-se por utilizar a ferramenta

http://dx.doi.org/10.25286/repa.v3i3.973

Qlikview, já que esta atende às necessidades de formular consultas e extrair relatórios, preparar painéis de monitoramento e controle (*Dashboard*) e até fazer auditoria nos dados.

Qlikview é uma ferramenta que se propõe, em um único aplicativo, a atender todo o processo de BI de forma integrada; do desenvolvimento do processo ETL ao desenvolvimento do layout para visualização e análise dos dados **[18]**.

O Painel Gestor faz a integração da tecnologia escolhida para armazenar os conteúdos resultantes da Sala de Aula Aberta (Git-PE) e a solução escolhida para inventariar os casos de estudos trabalhados em sala de aula (Planilhas para Catalogação).

Foi desenvolvido, no próprio Qlikview, um script de extração que faz uma leitura sistemática de todas as planilhas preenchidas com a catalogação das soluções. A engrenagem de processamento do Qlikview realiza automaticamente a integração dos dados e disponibiliza em memória para as análises no painel. A Figura 4 mostra uma ilustração do processo de ETL.

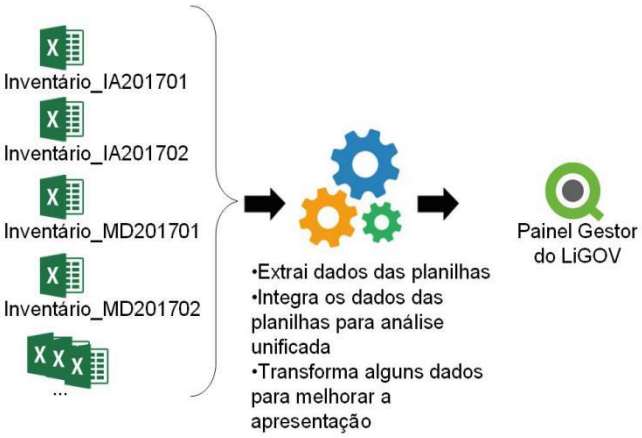


Figura 4 - Processo ETL
Fonte**:** O autor.

4 Resultados Obtidos

Para efeito de teste do repositório proposto, foi dividido o processo de validação em três fases. No primeiro momento validou-se a etapa de captura e criação de conteúdo, que no final do primeiro semestre de 2017, individualmente inclui-se todas as informações e conteúdo no ambiente implantado.

Posteriormente foi testado o potencial de compartilhamento e colaboração do repositório. Para tanto, durante o segundo semestre de 2017, seguiu-se o processo apresentado na Figura 1, cujas informações e conteúdos foram inseridos pelos responsáveis de cada etapa do processo.

Por fim, foi avaliado o resultado do repositório proposto quanto aos requisitos de gestão.

Do total, foram inseridos no repositório 16 casos de estudo que resultaram em 18 soluções aplicáveis ao Governo. Dos quais, 11 soluções foram de 2017.1 e 7 de 2017.2. O resultado da avaliação do Git-PE é apresentado na Tabela 6.

Tabela 6 - Avaliação do Git-PE.

Critérios de Avaliação	Resultado
Permite Colaboração com disponibilidade de ferramenta de internet web na ATI	Sim. Disponível em: https://www.git.pe.gov.br/groups/LIGOV
DesenvolvimentoDistribuído	Sim
Edições de qualquer linguagem de programação	Sim. Das 18 soluções inseridas no Git-PE tiveram 10 soluções em Pyton, 06 em Java, 01 em R e 01 em Weka.
Edições de qualquer tipo de arquivo e tamanho	Parcial. Material acadêmico -atendeu, pois houveram arquivos com extensão PPT, DOC e DOCx. Dados -não atendeu,pois a performance de armazenar grandes arquivos foi baixa. Além disso, identificou ser desnecessário o armazenamento para esse tipo de conteúdo.
Gerenciamento de permissões de compartilhamento das soluções	Sim
Controle de Acesso	Sim
Desenvolvimentonão linear	Parcial. Não houve a oportunidade de testar a evolução de alguma solução armazenada, mas foi realizado uma simulação dessa situação.

Na validação da aderência do Git-PE ao processo do LiGOV observou-se uma boa aceitação. O ambiente disponibilizado para os alunos de uso opcional teve, das sete soluções resultantes de 2017.2, 4 equipes utilizando o ambiente definido colaborativamente conforme a expectativa, 2 desenvolveram no Github e no final do projeto o merge foi realizado para o Git-PE e apenas uma equipe fez a entrega do conteúdo em mídia para inserção via upload. A Figura 5 mostra a tela do Git-PE com a relação das soluções cadastradas com visibilidade pública.

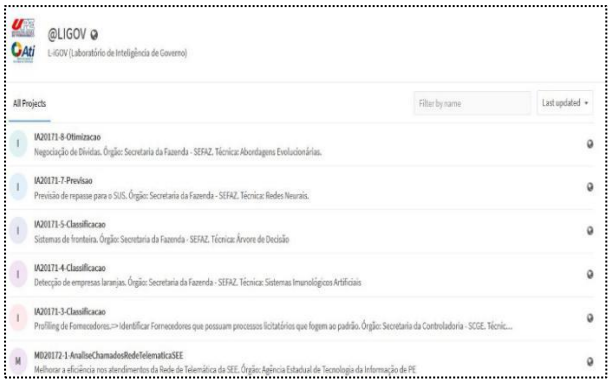


Figura 5 - Tela Git-PE com Lista das Soluções
Fonte**:**O autor.

Para realizar buscas por características das soluções são disponibilizadas duas opções. A Figura 6 mostra a opção no Git-PE, cujo o filtro limita-se a informações contidas no nome do diretório de trabalho e na descrição do caso de estudo informados.

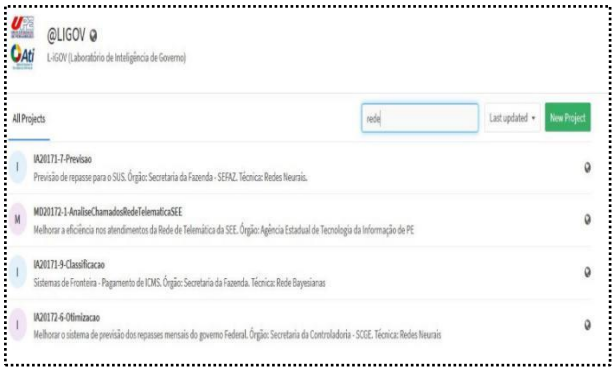


Figura 6 - Exemplo de Filtro no Git-PE
Fonte**:** O autor.

A outra opção de busca com maior autonomia e qualidade é realizada no Painel Gestor. Neste é possível pesquisar informações por qualquer característica inserida no formulário de inventário. Na Figura 7 mostra tela de busca no Qlikview.

Figura 7 - Exemplo de Filtro no Painel Gestor
Fonte**:**O autor.

O Painel Gestor cumpre a necessidade de se ter as estatísticas descritivas dos dados, sendo possível realizar várias análises, como por exemplo, num gráfico de pizza que visualmente apresenta a produção das soluções por período letivo, por disciplina, por órgão público, por classe de problema ou por técnica de resolução utilizada. A Figura 8 exemplifica essa possibilidade.

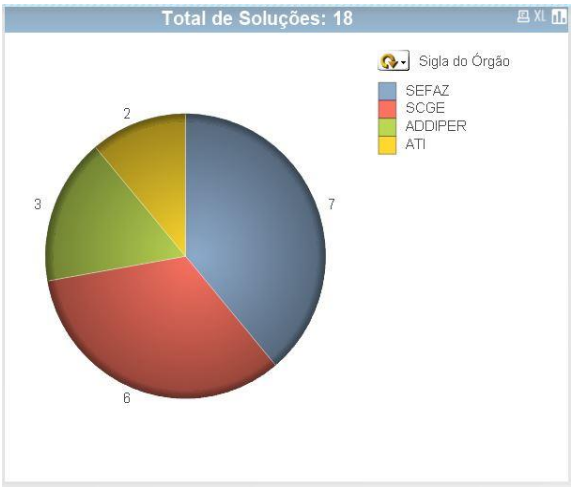


Figura 8 - Número de soluções entregues por órgão
Fonte: O autor

Outro exemplo é cruzar a classe dos problemas versus as técnicas utilizadas, como mostrado na Figura 9.

http://dx.doi.org/10.25286/repa.v3i3.973

Total de Soluções: 18

Técnica	Classe de Problema: Classificação e Otimização	Classificação	Associação e Classificação	Busca	Otimização	Previsão
Total	**1**	**10**	**1**	**3**	**2**	**1**
Swarm Intelligence	-	1	-	1	-	-
Árvore de Decisão	-	2	-	-	-	-
Rede Bayesianas	-	1	-	-	-	-
Sistemas Especialistas e...	-	1	-	-	-	-
Abordagens Evolucionárias	-	-	-	-	1	-
Abordagens Probabilísticas	-	1	-	-	-	-
Algoritmo Genético	1	-	-	-	-	-
Apriori e Regressão logís...	-	-	1	-	-	-
Busca Local	-	-	-	1	-	-
Cross-validation	-	1	-	-	-	-
Rede neural MLP e Rand...	-	1	-	-	-	-
Redes Neurais	-	-	-	-	1	-
Redes neurais	-	-	-	-	-	1
Sistema multi agente	-	-	-	1	-	-
Sistemas de busca local	-	-	-	1	-	-
Sistemas Imunológicos A...	-	2	-	-	-	-

Figura 9 - Número de Soluções entregues por Classe de Problema versus Técnica
Fonte: O autor

5. Conclusões

Neste artigo pode-se conhecer a proposta de Repositório que atende a necessidade de armazenamento, compartilhamento e gestão das soluções produzidas no Laboratório de Inteligência Governamental (LiGOV).

Considerando algumas necessidades do LiGOV, baseado nas teorias da Gestão de Conhecimento, Gestão de Conteúdo e Repositório de Conteúdo, foi especificado um ambiente composto por três tecnologias; uma ferramenta para armazenar os conteúdos resultantes da Sala de Aula Aberta (Git-PE), outra para inventariar os casos de estudos trabalhados em sala de aula (Formulário de Inventário) e a terceira para integrar os dados do Git-PE e do Formulário de Inventário, além de proporcionar buscas por características das soluções e apresentar suas estatísticas descritivas.

Ao longo de 2017, foi possível validar a especificação proposta. Como resultado foi possível validar a inclusão de 18 soluções trabalhadas na Sala de Aula Aberta no repositório; com os códigos-fonte, material acadêmico e características das soluções catalogadas.

Dessas 18 soluções, em quatro foram também testadas a utilização do ambiente colaborativo.

Assim, podemos considerar que o repositório proposto atingiu os objetivos. Passou a oferecer o armazenamento de soluções inovadoras alinhada com as necessidades do LiGOV; ambiente de desenvolvimento colaborativo e possibilidade de compartilhamento de soluções para reutilização.

5.1 Trabalhos Futuros

Para trabalhos futuros, pretende-se aprimorar a biblioteca de busca textual, a exemplo do Apache Lucene para aperfeiçoar as buscas pelas soluções no Git-PE, tendo em vista a grande escala que se pode alcançar este Repositório.

Almeja-se ainda criar um mecanismo de catalogação automática utilizando a busca textual no Git-PE.

Por fim, pretende-se criar um mecanismo (*analytics*) para avaliação das métricas de uso dos recursos disponibilizados pelo Git-PE, objetivando mensurar o valor, enquanto ativo, para o Estado.

Referências

[1] DINIZ, Eduardo Henrique et al. O governo eletrônico no Brasil: perspectiva histórica a partir de um modelo estruturado de análise, **Rev. Adm Pública**, Rio de Janeiro, v. 43, n. 1, p.23-48, 2009.

[2] SCHWAB, Klaus. **A Quarta Revolução Industrial**. 2 ed. São Paulo: Edipro, 2017.

[3] BLUMENSCHEIN, Fernando **Estratégias de Compras Governamentais no Brasil:** Teoria dos Leilões e "Big data". Rio de Janeiro: FGV Projetos, 2014.

[4] SILVA, Luís Andre. A evolução do Controle na era digital. **Revista do TCU**, n. 137, 2016.

[5] FERNANDES, Janderson Gabriel L. et al. Inteligência Artificial: Uma Visão Geral. **Revista Eletrônica Engenharia e Debate**, v. 1, 2018.

[6] Pernambuco. **Novo Portal da Transparência oferece melhor compreensão da aplicação dos recursos públicos**. Diário Oficial de PE. Ano XCV Nº 27, P.1 de 08/02/2018.

[7] D'ALMEIDA, B. J. Instrução Normativa da Coordenação da Administração Tributária (CAT) Nº 006, de 9.3.2017. **Diário Oficial de PE**, ano 94, n.61, p. 22, 31 mar. 2017.

[8] PINHEIRO, Cauanne L. **Melhoria de Processo de Negócio da Colaboração entre Sala de Aula Aberta e Agência de Tecnologia da Informação**. Monografia de Graduação, Universidade de Pernambuco, Recife, 2017.

[9] PERNAMBUCO. L**ei Complementar Nº 049, de 31/01/2003**: dispõe sobre as áreas de atuação, a estrutura e o funcionamento do Poder Executivo. Assembléia Legislativa do Estado, Recife, 31 jan. 2003.

[10] FIGUEIREDO, Saulo P. **Gestão do Conhecimento - Estratégia Competitiva para a Criação e Mobilização do Conhecimento na Empresa**. Rio de Janeiro: Qualitymarky 2005.

[11] DALKIR, Kimiz. **Knowledge Management in Theory and Practice**. Massashusetts, Oxford: Elsevier, 2005.

[12] PPARREIRAS, F. S., BAX, M. P. A gestão de conteúdos no apoio a engenharia de software. In: KMBrasil, 2003, São Paulo. **Anais**... São Paulo: SBGC - Sociedade Brasileira de Gestão do Conhecimento. 2003. CD-ROM. Disponível em: <http://www.fernando.parreiras.nom.br/publicacoes/pgct142.pdf>

[13] SANTOS, Marcelo; FRANCO, Carlos; TERRA, José. **Gestão de Conteúdo 360° Integrando Negócio, Design e Tecnologia**. São Paulo: Saraiva, 2009.

[14] RUS, Ioana; LINDVALL, Mikael. Knowledge Management in Software Engineering. **IEEE Software**, v. 19, n. 3, 2002.

[15] RUS, Ioana; LINDVALL, Mikael; SUM, Sachin. **A State of the Art Report: Knowledge Management in Software Engineering**. Fraunhofer Center for Experimental, Software Engineering Maryland, The University of Maryland. Park, Maryland: The University of Maryland, 2001.

[16] GIT. **Git fast version control**. Copyright 2014 by Tutorials Point (I) Pvt. Ltd.

[17] DRUZDZEL, Marek J. ; FLYNN, Roger R. **Decision Support Systems**. Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Program University of Pittsburgh. Pittsburgh: University of Pittsburgh, 2002.

[18] QLIKVIEW. **Tutorial Qlikview**. Copyright 1994-2005 QlikTech International AB, Sweden. 7º Edição.

http://dx.doi.org/10.25286/repa.v3i3.973

Extração de Informação e Mineração de Dados no Diário Oficial de Pernambuco

Information Extraction and Data Mining in the Official Gazette of Pernambuco

Ricardo Batista das Neves Junior[1] orcid.org/0000-0001-9538-6505
Weverton Fernandes de Medeiros Melo[1] orcid.org/0000-0003-3429-2892
Roberta Andrade de Araújo Fagundes[1] orcid.org/0000-0000-7172-4183
Alexandre Magno Andrade Maciel[1] orcid.org/0000-0003-4348-9291

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: rbnj@ecomp.poli.br

Resumo

O uso de técnicas de mineração de dados tem sido amplamente utilizado para o processamento de uma grande quantidade de dados documentados. No entanto, atualmente, poucos aplicativos mostraram-se efetivos para extrair e minerar dados em diários oficiais. Este trabalho tem como objetivo apresentar um método para construção de uma aplicação que usa um algoritmo para indexar conteúdo da base do Diário Oficial do Estado de Pernambuco, transformando as informações anteriormente disponíveis no texto para o formato estruturado, para aplicar uma Mineração de Dados. Para o desenvolvimento do método, a linguagem Java foi utilizada, com a possibilidade do aplicativo web. O estudo de caso baseou-se em documentos publicados no Diário Oficial de janeiro de 2007 a abril de 2017. Os resultados mostram que é possível indexar e estruturar esses dados, mas ainda há necessidade de uma melhor padronização dos dados.

Palavras-Chave: Mineração de Dados; Diário Oficial; Árvore de Decisão.

Abstract

The use of Data Mining techniques has been widely applied for processing a high amount of documented data. However, to date, there are very few effective applications for extracting and mining data in official journals. This work aims to present a method for the construction of an application that uses an algorithm to index contents of the base of the Official Gazette of the state of Pernambuco, transforming the information previously available in the text to structured format, to apply a Mining of Data. For the development of the method, the Java language was used, with the possibility of the web application. The case study was based on documents published in the Official Gazette from January 2007 to April 2017. The results show that it is possible to index this data and give meaning to it, but there is still a need for a better standardization of the data.

Key-words: *Data Mining; Official Diary; Decision Tree.*

http://dx.doi.org/10.25286/repa.v3i3.892

1 Introdução

O Diário Oficial de Pernambuco (DOE) foi criado no ano de 1924, ano da criação da Companhia Editora de Pernambuco (CEPE) **[1]**. A disponibilização do DOE tem como objetivo manter público aos cidadãos informações pertinentes aos poderes Executivo, Legislativo e Judiciário. Graças ao DOE, todos os atos administrativos do estado tornam-se públicos e alcançáveis por qualquer cidadão aumentando a transparência entre o governo e os indivíduos. Atualmente o DOE é disponibilizado gratuitamente no site da CEPE (*http://cepe.com.br*). As informações disponibilizadas pelo Diário Oficial do Estado podem ser consideradas como dados não estruturados, pois trata-se de um arquivo no formato PDF, com um texto corrido, imagens agregadas ao texto e difícil leitura computacional no contexto da divisão das sessões.

Atualmente, a Controladoria do Estado de Pernambuco, têm a obrigação de utilizar o DOE para obter algumas informações referentes à Inquéritos Administrativos, Processos Administrativos, Sindicâncias Administrativas e entre outras. Para a obtenção destas informações, é necessário efetuar o *download* dos Diários Oficiais, abrir estes documentos, utilizar o atalho *Control Find* (Ctrl + F), pesquisar pelo termo desejado e retirar a informação. Pode-se notar que este é um processo que gastar um tempo que poderia ser investido em outras atividades.

Este trabalho propõe uma solução de extração de informações associada a um mecanismo de mineração de dados para predição dos resultados das sindicâncias, com o objetivo de automatizar e otimizar este processo de acompanhamento. Para isto será desenvolvido um algoritmo capaz de ler a base de dados (Diários Oficiais), pesquisar pelos termos necessários, extrair as informações pertinentes ao conteúdo, estruturar os dados e implementado um motor de inferência baseado em árvore de decisão para predição de resultado das sindicâncias.

Este trabalho está organizado da seguinte maneira: A sessão 2 fala sobre a fundamentação teórica, onde pode-se entender o que é mineração de dados e observar alguns trabalhos que aplicam técnicas de mineração. A sessão 3 elucida sobre materiais e métodos, tais como preparação/transformação dos dados e técnica escolhida. A sessão 4 fala sobre os experimentos realizados bem como os resultados obtidos. A sessão 5 mostra as conclusões e considerações finais sobre o trabalho.

2 Fundamentação Teórica

2.1 Mineração de Dados

A mineração de dados vem atraindo a atenção na indústria da informação e na sociedade como um todo nos últimos anos, devido à grande disponibilidade de enormes quantidades de dados e a iminente necessidade de transformar esses dados em informações e conhecimentos úteis. As informações e os conhecimentos adquiridos podem ser utilizados em aplicações que vão desde análise, detecção de fraude e fidelização de clientes, controle de produção e exploração. A mineração de dados pode ser vista como resultado da evolução natural da informação tecnologia.

A mineração de dados é o processo de extração de conhecimento em grandes quantidades de dados. Ela está inserida em um processo maior denominado Descoberta de conhecimento (KDD – *Knowledge Discovery in Database*) **[3,8]**.

A descoberta de conhecimento como processo consiste numa sequência iterativa de algumas etapas tais como: (*i*) Limpeza de dados - para remover ruídos e dados inconsistentes; (*ii*) Integração de dados - onde várias fontes de dados podem ser combinadas; (*iii*) Seleção de dados - onde os dados relevantes para a tarefa de análise são recuperados da base de dados; (*iv*) Transformação de dados - onde os dados são transformados ou consolidados em formulários para mineração executando operações de resumo ou agregação, por exemplo; (*v*) Mineração de dados - um processo essencial onde são aplicados métodos inteligentes para extrair padrões de dados; (*vi*) Avaliação de padrões - para identificar os padrões verdadeiramente interessantes que representam o conhecimento com base em algumas medidas de interesse e (*vii*) Apresentação do conhecimento - onde técnicas de visualização e de representação do conhecimento são usados para apresentar o conhecimento minado ao usuário **[3]**.

Os passos *i* a *iv* são formas diferentes de pré-processamento de dados, onde os dados são preparados para a mineração. O passo de mineração de dados pode interagir com o usuário ou uma base de conhecimento. Os padrões interessantes são apresentados ao usuário e

podem ser armazenados como novos conhecimentos na base de conhecimento.

Concordamos que a mineração de dados é um passo no processo de descoberta de conhecimento. No entanto, na indústria, na mídia e no ambiente de pesquisa de banco de dados, o termo mineração de dados está se tornando mais popular do que o longo prazo de descoberta de conhecimento a partir de dados **[3]**.

Em princípio, a mineração de dados deve ser aplicável a qualquer tipo de repositório de dados, bem como a dados transitórios, como fluxos de dados. Os sistemas de banco de dados avançados incluem bancos de dados objeto-relacionais e bancos de dados específicos orientados a aplicativos, como bancos de dados espaciais, bancos de dados de séries temporais, bancos de dados de texto e bancos de dados multimídia. Os desafios e técnicas de mineração podem diferir para cada um dos sistemas de repositório **[3]**.

As funcionalidades de mineração de dados são usadas para especificar o tipo de padrões a serem encontrados nas tarefas de mineração de dados. Em geral, as tarefas de mineração de dados podem ser classificadas em duas categorias: descritiva e preditiva. As tarefas de mineração descritivas caracterizam as propriedades gerais dos dados no banco de dados. As tarefas de mineração preditivas realizam inferência nos dados atuais para fazer previsões.

Em alguns casos, os usuários podem não ter ideia sobre quais tipos de padrões em seus dados podem ser interessantes e, portanto, podem gostar de procurar vários tipos diferentes de padrões em paralelo. Assim, é importante ter um sistema de mineração de dados que pode explorar vários tipos de padrões para acomodar diferentes expectativas ou aplicações de usuários. Além disso, os sistemas de mineração de dados devem ser capazes de descobrir padrões em várias granularidades (isto é, diferentes níveis de abstração). Os sistemas de mineração de dados também devem permitir que os usuários especifiquem dicas para orientar ou focalizar a busca por padrões interessantes. Como alguns padrões podem não ser válidos para todos os dados do banco de dados, uma medida de certeza ou "confiabilidade" é geralmente associada a cada padrão descoberto **[3]**.

2.2 Árvore de Decisão

A Árvore de Decisão é uma técnica de classificação de dados dentro da Mineração de Dados (*Data Mining*). Podem ser usadas em conjunto com outras tecnologias de regras, mas são as únicas a apresentar os resultados hierarquicamente (com priorização). Nela, o atributo mais importante é apresentado na árvore como o primeiro nó, e os atributos menos relevantes são mostradas nos nós subsequente. A vantagem principal das Árvores de Decisão é a tomada de decisões levando em consideração os atributos mais relevantes, além de compreensíveis para a maioria das pessoas. Ao escolher e apresentar os atributos em ordem de importância, as Árvores de Decisão permitem aos usuários conhecer quais fatores mais influenciam os seus trabalhos.

Está técnica é uma representação simples das informações e um caminho eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados **[9]**. Uma Árvore de Decisão utiliza a estratégia chamada dividir-para-conquistar, ou seja, um problema complexo é decomposto em subproblemas mais simples. Repetidamente, a mesma estratégia é aplicada a cada subproblema **[10]**. A capacidade de discriminação de uma Árvore de Decisão vem das características de divisão do espaço definido pelos atributos em subespaços e da associação de uma classe a cada subespaço.

Segundo **[9]**, as Árvores de Decisão são compostas de: nós, que representam os atributos, e de ramos, oriundos desses nós e que recebem os valores possíveis para esses atributos (cada ramo descendente corresponde a um possível valor desse atributo). Nas árvores existem nós folha, que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está associada a uma classe. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

3 Materiais e Métodos

3.1 Preparação dos Dados

Durante a preparação dos dados para a realização da mineração de dados foi considerado

http://dx.doi.org/10.25286/repa.v3i3.892

algumas etapas descritas em **[2]**. Inicialmente, foram utilizados Diários Oficiais do Estado referentes ao Poder Executivo publicados no período de janeiro de 2007 a abril de 2017. Essa base de dados foi lida e pré-processada. A fase de pré-processamento dos dados foi dividida nas seguintes etapas: Tokenização, Remoção de stop *word* e *Stemming*.

No processo de tokenização foi definida algumas palavras-chave tais como: "Sindicância Administrativa", "Sindicância Investigativa", "Sindicância Disciplinar" e "Sindicância Administrativa Disciplinar". O objetivo do algoritmo é buscar por essas palavras-chave e entregar informações referentes aos termos solicitados. Dentre o conteúdo retornado pelo algoritmo estão inclusos os termos: "data", "palavra-chave", "informação", "número portaria", "tipo comissão", "secretaria", "lei número", "vencimento" e "resultado". Onde "data" é a data de publicação do DOE, "palavra-chave" é qual palavra-chave é correspondente à informação retornada, "informação" é todo o parágrafo o qual a palavra chave está inserida, "número portaria" é o número da portaria envolvida na sindicância, "tipo comissão" é o tipo da comissão da sindicância que pode ser comissão específica ou comissão permanente, "secretaria" é qual secretaria do Estado está envolvida na sindicância retornada, "lei número" é qual o número da lei é referente à sindicância, "vencimento" é a data de vencimento da sindicância e "resultado" é em que resultou a sindicância (i.e. processo arquivado, funcionário repreendido ou abertura de um inquérito administrativo).

Não houve dificuldades para remover as palavras desprezíveis, dado que é necessário extrair uma baixa quantidade de informações, em comparação com à quantidade de informação existente no DOE. Então, o algoritmo automaticamente despreza tudo que não há relação com as palavras-chave.

O processo de *Stemming* consiste em reduzir ao radical algumas palavras que se deseja buscar com o objetivo de identificar a palavra independente do tempo verbal ou se a palavra está no gerúndio, infinitivo ou particípio. Ao encontrar uma palavra-chave no DOE, o algoritmo precisa verificar se a sindicância encontrada tem um resultado conclusivo (i.e. funcionário repreendido, processo arquivado ou abertura de um inquérito administrativo), se houver as palavras-chave podem aparecer no algoritmo de formas diferentes (e.g. "arquivado", "arquivou-se", "arquivando", "repreendido", "repreensão" e etc.), então, para encontrar o resultado em qualquer terminação da palavra, o resultado foi buscado por "arquiv", "repreen", "inquérito". Na busca pelos resultados, além de utilizar a técnica *Stemming*, foi realizada algumas variações nas palavras com todas as letras minúsculas, a primeira letra maiúscula e todas as letras maiúsculas.

3.2 Transformação dos Dados

O algoritmo de extração de informação, através das regras estabelecidas, conseguiu formar uma base de dados com 339 registros. Infelizmente, a falta de consistência e padrão nas publicações do Diário Oficial afetou na formação da base e gerou alguns registros com campos vazios, necessitando realizar transformações nos dados para prepara-los para a mineração.

Primeiramente, foram filtrados os resultados, selecionando apenas os que tinham sindicância fechada (inquérito administrativo, arquivado e repreensão). Em seguida, foram eliminados os registros que estavam com o campo "secretaria" vazio. Por último, os atributos que não seriam utilizadas para a classificação foram retirados, restando apenas os campos "Secretaria", "Tipo de Comissão" e "Resultado".

Após a transformação, a base contou com 40 registros, onde o campo "Tipo de Comissão" foi alterado para a forma binária (comissão permanente = 0 e comissão especifica = 1) e "Secretaria" em forma numérica (ADAGRO = 1, DETRAN = 2, HSE = 3, IRH = 4, JUCEPE = 5, SAD = 6, SASSEPE = 7, SCGE = 8, SDS = 9, SES = 10, SUAPE = 11 e UPE = 12).

3.3 Técnica Utilizada

A base de dados utilizada foram os Diários Oficiais de Pernambuco, após realizar a extração de informação e estruturação dos dados, restou uma base de dados com apenas 339 registros, destes registros, a minoria dos campos estavam preenchidos valor do "resultado" (que informa o resultado final da sindicância). Como o objetivo do trabalho é prever qual será o resultado de uma sindicância, existiam apenas 40 registros disponíveis para executar a previsão.

Diante deste contexto, a base de dados foi preparada com intuito de relacionar as secretarias e comissões com os resultados das sindicâncias, para que fosse possível encontrar uma relação com essa classificação. Para isso, foi aplicado a técnica de Árvore de Decisão utilizando o software Weka, que já possui uma biblioteca para a mesma e é muito simples para implementar. Os tipos de árvore utilizados foram a *J48* e *RandomTree*. O *J48* é uma implementação do algoritmo C4.5 dentro do programa Weka, que gera uma árvore de classificação, onde, a cada nó, o algoritmo escolhe um atributo que irá subdividir de forma mais eficiente o conjunto de dados em subconjuntos homogêneos e qualificados por sua classe. Já o *RandomTree* constrói a árvore de classificação escolhendo K atributos de forma aleatória em cada nó, não realizando a poda, o que acaba plotando uma árvore grande.

4 Experimentos

A técnica aplicada foi capaz de relacionar as variáveis e classificar as possibilidades, onde na Árvore *J48* o tipo de comissão é o atributo mais importante, enquanto na *RandomTree* a secretaria é o atributo mais importante. Como a base de dados foi transformada, transformando campos texto em valores numéricos, as árvores fazem a verificação do valorcontido no campo, por exemplo: $Secretaria >= 8$ (SCGE = 8, SDS = 9, SES = 10, SUAPE = 11 e UPE = 12) e $Tipo\ Comissao > 0$ (comissão especifica = 1).

Dada o tipo de comissão e a secretaria, as Árvores de Decisão realizam a previsão do resultado de uma sindicância, essa informação pode ser muito valiosa para o tomador de decisão

Da Controladoria do Estado de Pernambuco, pois, de uma determinada secretaria sempre arquivar as sindicâncias, ou qual tipo de comissão é mais propensa a arquivar um processo, abrir inquérito administrativo ou repreender um funcionário.

4.1 Árvore J48

A árvore J48 é uma técnica simples, logo, apresenta como vantagem um menor custo computacional. Entretanto, em seu resultado, além de alcançar uma menor taxa de acerto em relação à árvore *RandomTree*, apresenta também um menor índice Kappa, que é o índice que reflete a confiabilidade do modelo, ou seja, quanto maior o índice Kappa, mais confiável é o modelo. A Figura 1 mostra a árvore gerada pelo algoritmo J48, pode-se notar a baixa complexidade da árvore gerada obtendo uma taxa de acerto de 70% com um índice Kappa de 0.4217. O valor do Erro médio absoluto alcançado foi de 0.1759 e o Erro médio quadrático igual a 0.2966.

4.2 Árvore *RandomTree*

A árvore *RandomTree* traz como desvantagem um maior custo computacional em relação à J48, mas, a desvantagem é compensada pelo seu resultado significativamente superior em termos de Taxa de acerto e índice Kappa. Na Figura 4 pode-se visualizar que o algoritmo *RandomTree* gerou uma árvore maior, mais complexa e com maior riqueza de detalhes em relação à J48. A aplicação de um grande conjunto de dados nessa técnica pode resultar em um alto custo computacional. O resultado obtido é superior, com uma taxa de acerto de 80% com um índice Kappa de 0.624. O valor do Erro médio absoluto foi de 0.09 e o valor do Erro médio quadrático alcançado igual a 0.2121.

http://dx.doi.org/10.25286/repa.v3i3.892

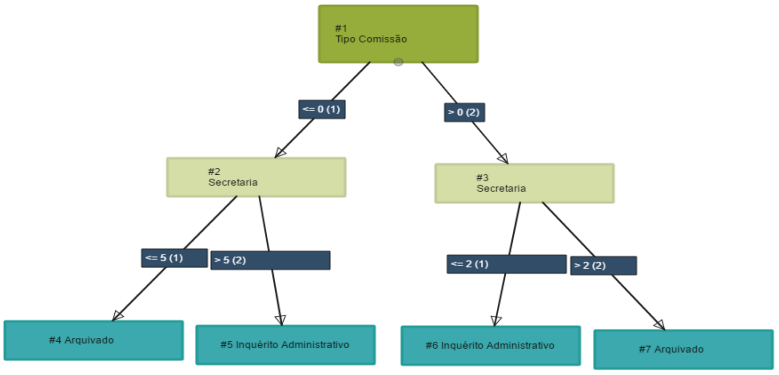


Figura 1:Árvore de Decisão - J48.

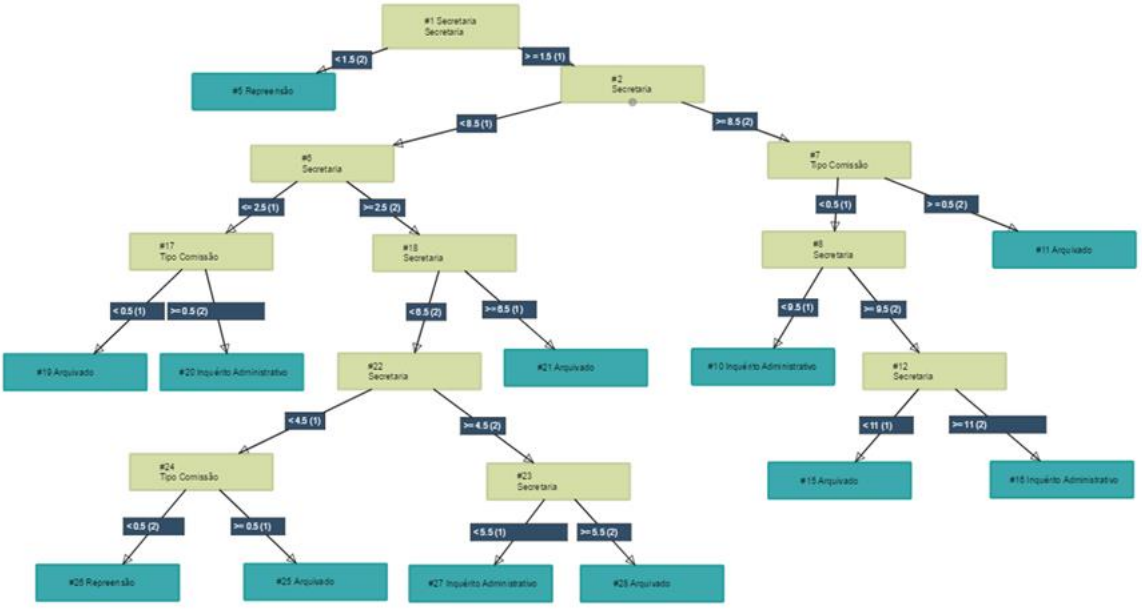


Figura 2:Árvore de Decisão – *RandomTree.*

5 Conclusões

Este trabalho entrega um projeto piloto que poderá ser expandido para Diários Oficiais de outros estados. Desenvolveu-se um motor de inferência para extração de informações de texto corrido e aplicou-se a técnica de mineração de dados conhecida como Árvore de Decisão a fim de facilitar a visualização, organizar os documentos publicados no Diário Oficial do Estado de Pernambuco e retirar informações relevantes a

partir dos dados. Através das etapas de pré-processamento e extração propostas pelo trabalho, pôde-se recuperar informação, antes apresentada em formato de linguagem natural para posteriormente transforma-la em uma base de dados possível de ser persistida. A etapa de mineração de dados demonstrou que é possível classificar e relacionar os dados, gerando informações que podem ser de extrema relevância para os tomadores de decisão.

Caso os órgãos desejem um algoritmo com melhores resultados, é necessário realizar uma padronização mínima dos Diário Oficiais para que a extração seja mais eficiente. Além disso, existe uma necessidade de um maior estudo das regras de padrões nos termos exibidos no Diário Oficial.

Referências

[1] COMPANHIA EDITORA DE PERNAMBUCO. A Cepe.**CEPE**. Disponível em: <https://www.cepe.com.br/index.php/cepe.html>. Acesso em: 24 abr. 2017.

[2] PATEL, Falguni N.; SONI, Neha R. Text mining: A Brief survey. **International Journal of Advanced Computer Research**, v. 2, n. 4, p. 243-248, 2012.

[3]JIAWEI, Han;PEI, Jian; KAMBER, Micheline.**Data mining: concepts and techniques**. Amsterdã: Elsevier, 2011.

[4]IBM Knowledge Center. Disponível em: <https://www.ibm.com/support/knowledgecenter/pt-br/SS3RA7-17.1.0 /modeler-crispdm-ddita/clementine/crisp-help/crisp-overview.html>. Acesso em: 24 abr. 2017.

[5]BHARANIPRIYA, V.; PRASAD, V. Kamakshi. Web content mining tools: a comparative study. **International Journal of Information Technology and Knowledge Management**, v. 4, n. 1, p. 211-215, 2011.
.

[6] Revista de Ciências Exatas e Tecnologia, v. 3 n. 3, 2008.

[7]MORAIS, Edison Andrade M.; AMBRÓSIO, Ana Paula L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

[8]FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory;SMYTH, Padhraic. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: International Conference on Knowledge Discovery and Data Mining, 2., 1996, Portland. **Proceedings**... Portland: KDD, 1996. p. 82-88.

[9]GARCIA, Simone C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. In: SEMANA ACADÊMICA, 2000,Rio Grande do Sul. **Anais**.... Rio Grande do Sul: UFRGS, 2000.

[10]Gama, J. **Árvores de decisão**. 2000. Disponível em:<http://www.liacc.up.pt/~jgama/Mestrado/ECD1/Arvores.html>. Acesso em: 14 ago. 2002.

[11]QUINLAN, J.Ross. **C4.5**: programs for machine learning. Sydney, Australia: Morgan Kaufmann Publishers, 1993. p. 302.

http://dx.doi.org/10.25286/repa.v3i3.892

Um modelo de inferência para a classificação de resultados processuais a partir de causas jurídicas oriundas da Justiça Estadual

An inference model for the classification of procedural results from legal causes originating from State Courts

Manoel Alves de Almeida Neto[1] orcid.org/0000-0003-4941-6376
Vinícius Malloni Moura[2] orcid.org/0000-0002-6547-4448
Jonathan da Silva Bandeira[1] orcid.org/0000-0003-1693-2091
Pedro Rudá Cavalcanti Gomes de Freitas[1] orcid.org/0000-0002-8151-1114
Roberta Andrade de Araújo Fagundes[1] orcid.org/0000-0002-7172-4183

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Kurier Tecnologia em Informação, Recife, Pernambuco, Brasil.

E-mail do autor principal: Manoel Alves de A. Neto maan@poli.br

Resumo

Este artigo tem como objetivo apresentar um modelo de inferência para classificar processos jurídicos relacionados à Justiça Estadual utilizando dados documentados de jurisprudência, tais como: comentários dos juízes realizados durante os veredictos, classes jurídicas do processo e UF. Os dados foram extraídos de websites de cortes judiciais, como o Portal do Tribunal de Justiça do Estado de Minas Gerais e o Poder Judiciário do Estado de Alagoas. Em toda a base de dados, foi realizada uma seleção no campo textual da descrição da sentença para extrair as leis que foram consideradas nos veredictos. Para tal seleção, o atributo de publicação e a quantidade de ocorrências da lei na base de dados foram considerados. As técnicas utilizadas para realizar a mineração de dados e classificar os processos como procedentes ou improcedentes foram a árvore de decisão e as redes neurais artificiais. Os testes realizados mostraram resultados satisfatórios e superiores ao valor comum para classificação de dados de jurisprudência, de normalmente 60%.

Palavras-Chave: Mineração de dados; Redes neurais; C4.5; PART; Jurisprudência;

Abstract

This paper aims to develop an inference model for classify the judiciary processes related to State Courts using documental data, such as comments made of the judges in the verdict moment, states and processual juridical class. The data were extracted from Justice courts websites, such as Poder Judiciário do Estados de Alagoas and Portal do Tribunal de Justiça do Estado de Minas Gerais. To extract the laws used in veredicts from database, its amount of ocurrence and the Publication attribute were considered. The techniques used were decision tree and artificial neural network. The tests showed a good shape of the laws and articles to classify appropriate and unfounded cases, which help the decision maker to take a decision. The results obtained were higher than the common rate of 60%.

Key-words: *Data mining; Neural networks; C4.5; PART; Law.*

1. Introdução

O acesso a informações sobre dados Jurídicos é imprescindível para o operador do direito no exercício de suas funções. Contudo, grande parte dessas informações tem fontes distintas e numerosas, como os processos disponibilizados através do sistema Processo Judicial Eletrônico – Pje **[1]**, tornando difícil a correlação destes dados para um processamento estatístico.

Em uma pesquisa de mercado realizada pelo CONJUR **[2]**, é possível chegar à conclusão de que não há uma solução amplamente utilizada no mercado que possua um banco de dados com informações jurídicas relacionadas entre si, de modo a permitir uma abordagem estatística que possibilite entender a relação entre os processos e as variáveis que os levam a classificação de suas sentenças.

Devido à dificuldade de acesso a estas informações, torna-se evidente a importância de um estudo aplicado à área jurídica. Este trabalho tem como objetivo desenvolver um modelo para classificação de resultados processuais a partir de informações extraídas de causas jurídicas, como leis, artigos, fórum e classe processual pertencente ao poder judiciário da Justiça Estadual.

2 Fundamentação Teórica

Para amparar teoricamente a aplicação de conhecimento deste trabalho, se fez necessária a observação e análise de alguns dos fundamentos relacionados à área das ciências jurídicas e da mineração de dados.

2.1 Área de Conhecimento Jurídico

A ciência jurídica ou ciência do direito estuda o fenômeno jurídico em todas as suas manifestações e momentos, tendo como objeto de estudo o conhecimento do direito **[3]** e como meio de expressão a chamada linguagem forense **[4]**.

No Brasil, o direito é dividido em duas principais ramificações: o direito público, que rege os interesses públicos e as relações do estado e o direito privado, que rege os interesses individuais de cada um e as suas respectivas relações particulares **[5]**.

Um dos principais objetivos do meio jurídico é gerar e documentar os conhecimentos obtidos na área, sejam leis e normativas, informações processuais, etc. Todo dado extraído neste meio é denominado de informação jurídica. Segundo Alonso **[6]**, a informação jurídica pode ser conceituada como qualquer dado ou fato extraído de toda e qualquer forma de conhecimento da área jurídica, obtido por todo e qualquer meio disponibilizado e que pode ser usado, transferido ou comunicado sem a preocupação de estar integrado a um contexto.

A informação jurídica pode ser classificada de três maneiras: legislação, jurisprudência e doutrina. Segundo Passos **[7]**, a informação jurídica também pode ser gerada, registrada e recuperada em três formas distintas: a normativa (legislação), a interpretativa (jurisprudência) e a descritiva (doutrina).

Segundo Passos **[7]** e Barros **[8]**, a atual migração das documentações e extrações de dados jurídicos do meio físico para o digital trouxeram uma série de vantagens, tais como: variedade e quantidade de material disponível, maior acessibilidade e redução de custos. Em contrapartida, há uma série de dificuldades na recuperação de informações jurídicas no meio digital, tais como: a obtenção de toda legislação sobre um determinado assunto, realização de pesquisas de jurisprudência e presença de possíveis deficiências nas bases de dados.

Por fim, quando se fala de bases de dados jurídicas, também é importante saber o que caracteriza uma informação jurídica e se reflete em uma base de dados desta natureza, que segundo Martinho **[9]**, se dá pelos seguintes itens: grandes volumes de informação e rapidez na sua desatualização face a um constante crescimento e criação de novas fontes; público-alvo exigente e diversificado; Necessidade de grande rigor e precisão da sua conservação e rapidez na transmissão de seus dados para assegurar sua correta utilização e aplicação.

2.2 Mineração de Dados

Mineração de dados é uma etapa de um processo maior que envolve várias áreas de

conhecimento diferentes. É comum ao processo o uso de técnicas estatísticas, modelos matemáticos e/ou inteligência artificial para reconhecimento de padrões, proficiência em trabalhar com grandes quantidades de dados, saber lidar com a captura de dados em sistemas de sensores ou sistemas embarcados, conhecimento do problema a ser abordado e capacidade de organizar e apresentar resultados.

Dado que a mineração dos dados é apenas uma etapa, é necessário identificar como todo o processo se desenvolve. O *Knowledge Discovery from Data* (KDD), segundo Fayyad **[10]**, é um processo não trivial, interativo e iterativo, para identificação de padrões que sejam compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. O procedimento de KDD se divide nas fases de agrupamento das informações, pré-processamento, seleção, limpeza, transformação e mineração de dados, além de avaliação do conhecimento extraído durante o procedimento para tomadas de decisões.

3 Materiais e Métodos

Aqui estão descritas as metodologias utilizadas nas etapas de pré-processamento e transformação dos dados.

3.1 Pré-Processamento

Todos os dados utilizados foram extraídos dos portais de cada Tribunal de Justiça, tais como Portal de Serviços de São Paulo **[11]** e Portal do Tribunal de justiça do estado de Minas Gerais **[12]**. Devido cada tribunal possuir sua própria fonte de dados e forma de apresentação, as informações são extraídas de acordo com regras específicas implementadas para cada fonte de dados utilizando recursos como expressões regulares e dicionários de palavras para homogeneizar os termos com a finalidade de agregar valor estatístico.

Na homogeneização dos dados, destacam-se as seguintes transformações: espaços duplos entre palavras para espaços únicos; convenção de todos os caracteres para maiúsculo; remoção de todos os diacríticos, espaços no início e fim de cada frase; e, remoção dos caracteres especiais.

Indicador Êxito é um campo que representa a classe de saída para os algoritmos de classificação, podendo ser Parcialmente Procedente, Procedente ou Improcedente. Para determinar qual indicador de êxito deve ser atribuída a cada processo jurídico, utilizou-se a aplicação de um dicionário de palavras-chave no campo publicação (i.e. todos os comentários dos juízes), a fim de identificar o pensamento dos juízes sobre cada processo julgado.

Após ter feito a atribuição das classes à cada processo jurídico, o próximo passo é realizar a extração das leis e artigos do campo publicação. Para isso, a seguinte expressão regular foi aplicada no campo Publicação:

```
(LEI|ART|ARTIGO)(\W)?(\s+)?(\d{1,10})(\W)?(\d*)(\/)?(\d{1,10})?
```

3.2 Transformação dos Dados

As transformações nos dados foram para utilizar nos algoritmos e classificação em duas etapas.

Na primeira etapa, utilizou-se os seguintes campos:

- Dispositivos (leis e artigos) como entradas;
- Indicador de êxito como saída.

A Tabela I mostra um exemplo da base de dados utilizada nos primeiros experimentos após a aplicação das transformações.

Neste exemplo, o processo 01 foi considerado procedente utilizando a lei 6.830/80 e o artigo 33. O processo 04 teve seu indicador de êxito classificado como parcialmente procedente com a utilização do artigo 33. Por fim, o processo 07 foi classificado como improcedente através do artigo 535.Já segunda etapa, foram utilizados os seguintes dados:

Tabela I – Primeira transformação dos dados

N. Proc.	Ind. de Êxito	LEI. 6.830/86	ART. 535
1	Procedente	1	0
4	Parcialmente Procedente	0	0
7	Improcedente	0	1

Após o resultado dos primeiros experimentos utilizando as transformações citadas anteriormente, os seguintes processamentos foram adicionados com a finalidade de agregar mais informações à base de dados final:

http://dx.doi.org/10.25286/repa.v3i3.907

- Adição das colunas Classe, UF e Fórum ao conjunto de dados;
- Transformação dos valores destas colunas para binário, de forma a representar a existência do valor deste atributo no processo determinado.

A Tabela II mostra um exemplo da base de dados após estas transformações aplicada nesta segunda abordagem. Neste exemplo, o processo 02 foi classificado como procedente para os dados de entrada: PE, Fórum Joana Bezerra, Lei 6.830/80 e artigo 535. O processo 08 foi classificado como procedente tendo como dados de entrada Acre, Trabalhista e o artigo 535. Por último, o processo 10 foi classificado como improcedente para o conjunto de entrada PE, Fórum Joana Bezerra e lei 6.858/80.

Tabela II – Segunda transformação dos dados

N. Proc.	Ind. de Êxito	PE	AC	Joana Bezerra	Trabalhista	LEI. 6.830/80
2	Procedente	1	0	1	0	1
8	Procedente	0	1	0	1	0
10	Improcedente	1	0	1	0	1

4 Parametrização das Técnicas

Para este trabalho, as técnicas utilizadas foram árvore de decisão e Rede Neural Artificial. A utilização da árvore de decisão teve como objetivo mostrar, de forma estrutura e sequência, os dados de entrada, UF, Fórum, Dispositivos (leis e artigos) e Classe Processual para explicar os resultados. Já o propósito da utilização da Rede Neural Artificial foi realizar um estudo comparativo com a árvore de decisão para aferir, por meio de experimentações, qual das duas técnicas oferece maior precisão em seus resultados, levando em consideração o volume e desbalanceamento das bases de dados jurídicas analisadas.

4.1 Árvores de Decisão

Os algoritmos de árvores de decisão utilizados foram, C4.5 e *Recursive Partitioning and Regression Trees* (PART). O algoritmo C4.5 utiliza o cálculo da entropia, equação (1), para montar a estrutura da árvore. Já o algoritmo PART é um método estatístico para análises de multi-variáveis, e utiliza o particionamento recursivo para montar a árvore de decisão com apenas as variáveis de entrada que contém maiores índices de correlação entre si. Para ambas técnicas, a abordagem pós-poda foi adotada para recalcular os nós que possuem baixa relevância na árvore.

$$H(T) = -\sum_{j=1}^{k} \frac{freq(c_j, T)}{|T|} x log_2 \frac{freq(c_j, T)}{|T|} \quad \textbf{(1)}$$

Onde:

- freq(cj , T) quantidade de registros da classe cj em T;
- |T| número total de registros do conjunto T;
- k número de classes distintas que ocorrem em registros de T.

4.2 Redes Neurais

A rede neural utilizada foi a MLP (*Multi-Layer Perceptron*) com o algoritmo de *backpropagation*, funções de ativação não lineares (sigmóde logística), uma camada de entrada com 2900 neurônios, uma camada intermediária com 30 neurônios e uma camada de saída com um neurônio. O algoritmo de treinamento é clássico e tem duas fases: *Feedfoward* e *Backward* ou *Backpropagation*.

Primeiramente, os pesos de cada neurônio foram inicializados com valores pequenos e randômicos e foram definidas as condições de parada da execução do algoritmo (atingir um número máximo de ciclos ou repetir vinte ciclos sem ganhos significativos nos resultados e sem mudanças nos pesos dos neurônios). Em seguida, na fase *Foward*, os padrões de treinamento foram passados pelas unidades da camada de entrada, intermediária e saída. As unidades da camada de entrada receberam os dados e os dissiparam para as unidades da camada seguinte (intermediária). As unidades da camada intermediária ponderaram os sinais recebidos por meio dos seus pesos, aplicaram sua função de ativação para computar suas saídas e as enviaram para as unidades da camada de saída. As unidades da camada de saída

ponderaram seus sinais de entrada e aplicaram sua função de ativação para computarem seus resultados.

Na fase seguinte, a *Backward*, os erros foram calculados na camada de saída e retropropagados para as camadas anteriores que com base nestes, também calcularam seus respectivos erros. Após essas duas fases do algoritmo, os pesos e bias foram ajustados conforme a necessidade.

5 Experimentos Realizados

Os experimentos foram realizados em duas etapas para testar as técnicas de árvore de decisão C4.5 e *Recursive Partitioning and Regression Trees* (PART), e a Rede Neural Artificial. A primeira etapa utilizou-se apenas as leis e artigos citados nas sentenças dos processos (i.e campo publicação - tabela I). Na segunda etapa foram utilizados os campos UF, Fórum, Classe Processual, leis e artigos. O intuito da primeira abordagem foi analisar quais são as influências que as leis e artigos têm no resultado de cada causa jurídica. A intenção da segunda abordagem foi verificar o comportamento das leis e artigos para cada UF, Fórum e Classe Processual.

Foram realizadas 30 execuções para cada técnica. O método *cross validation*, *K-folds* foi adotado para realizar o processo de treinamento, teste e validação, e o valor de k foi igual a 3. Todas as configurações utilizadas para cada técnica serão descritas a seguir.

Para a primeira abordagem, a base de dados utilizada continha 4.157 instâncias com 140 colunas. A distribuição dos dados estava estruturada em 139 colunas representando artigos e leis, e uma coluna representando a saída da classificação, podendo ser avaliada como Procedente, Parcialmente Procedente ou Improcedente.

- C4.5:
  - Limitação de até 7 nós;
  - Pós-poda;
- PART:
  - Aceitação das folhas que contêm pelo menos 25% de relacionamento com os dados;
  - Pós-poda;
- RNA:
  - Máximo de ciclos de treinamento = 600 ciclos;
  - Neurônios na camada de saída = 40 neurônios;
  - Funções de ativação (Na camada escondida e na camada de saída): Sigmóide logística;

Para a segunda abordagem, a base de dados utilizada continha 8.050 instâncias com 2.900 colunas. O mapeamento dos dados estava estruturado nas seguintes colunas: 1.798 leis e artigos, 5 classes, 21 UFs e 1.076 Fóruns. A saída única apresentou os resultados de classificação, sendo avaliada como Procedente ou Improcedente.

- C4.5:
  - Limitação de até 7 nós;
  - Pós-poda;
- PART:
  - Aceitação das folhas que contém acima de 35% de relacionamento com os dados;
  - Pós-poda;
- RNA:
  - Alpha (taxa de aprendizado) = 0,85 e beta (momentum) = 0,25;
  - 100 ciclos máximo de treinamento;
  - 30 Neurônios na camada de saída;
  - Funções de ativação (Na camada escondida e na camada de saída): *Sigmóide* logística.

6 Resultados

Os resultados obtidos através dos experimentos foram organizados em formato visual utilizando gráficos no formato boxplot e tabelas contendo as informações pertinentes à cada técnica.

Todos os resultados obtidos na primeira abordagem estão descritos na Tabela III. Como é possível notar, ao aplicar apenas as leis e artigos como entradas e utilizar três possíveis valores de saída: Parcialmente Procedente, Procedente ou Improcedente, os resultados obtidos das três técnicas não foram muito favoráveis, pois a média das taxas de acerto foram de 65.23%, 65.92% e 59.65% para as técnicas C4.5, PART e Rede Neural.

http://dx.doi.org/10.25286/repa.v3i3.907

O valor do índice Kappa foi 17.68% para C4.5, 19.5% para a técnica PART, e 16.40% para a RNA. Os valores da média do erro absoluto de todas as execuções foram, 35.27%, 34.65% e 42.87% para as técnicas C4.5, PART e Rede Neural. Nesta primeira abordagem, a técnica que obteve o melhor resultado foi a *Recursive Partitioning and Regression Trees* (PART).

Figura 1 – boxplot da primeira abordagem

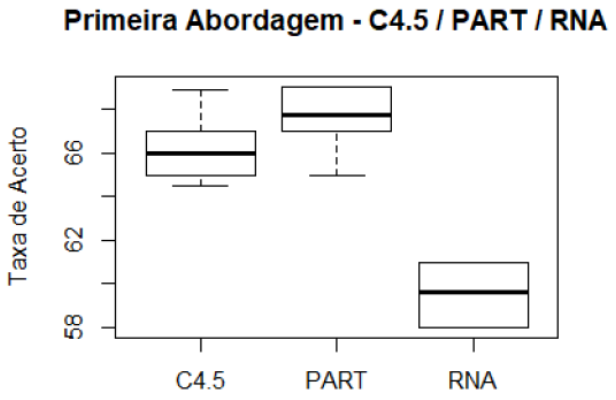


Tabela III – Segunda transformação dos dados

Métrica	C4.5	PART	Rede Neural
Média da taxa de acerto	66.20%	67.59%	59.65%
Índice Kappa	17.68%	19.57%	16.48%
Média do erro absoluto	35.27%	34.65%	42.87%

A Figura 3 mostra o gráfico boxplot dos resultados obtidos na primeira abordagem. Na Tabela IV estão os valores das análises descritiva.

Tabela IV – Resultados estatísticos da primeira abordagem

Métrica	C4.5	PART	RNA
Mínimo	64.50%	65%	58%
Máximo	68.90%	69.04%	61%
Médiana	65.90%	67.75%	59.65%
Média	66.20%	67.59%	59.55%
Desvio Padrão	1.64%	1.56%	1.40%

Com relação a segunda abordagem, os resultados obtidos estão descritos na Tabela V. Nessa segunda abordagem, ao adicionar os demais dados, UF, Fórum e Classe Processual junto com as leis e artigos, e considerando apenas duas possíveis respostas no resultado dos algoritmos, procedente ou improcedente, fica evidente que os resultados obtidos foram muito superiores em comparação com a primeira abordagem. Nessa segunda abordagem, a técnica que obteve maior ganho e melhores resultados em comparação com as demais, foi a Rede Neural Artificial. Na primeira abordagem, a RNA obteve 59.65% para a média da taxa de acerto, porém na segunda abordagem, esse resultado passou para 76.06%. Ou seja, um ganho de 16.41% na média geral da taxa de acerto. Já para as técnicas de árvores de decisão, o ganho na média da taxa de acerto não foi tão significativo, 8.05% para C4.5 e 6.72% para PART.

Tabela V – Resultados obtidos na segunda abordagem

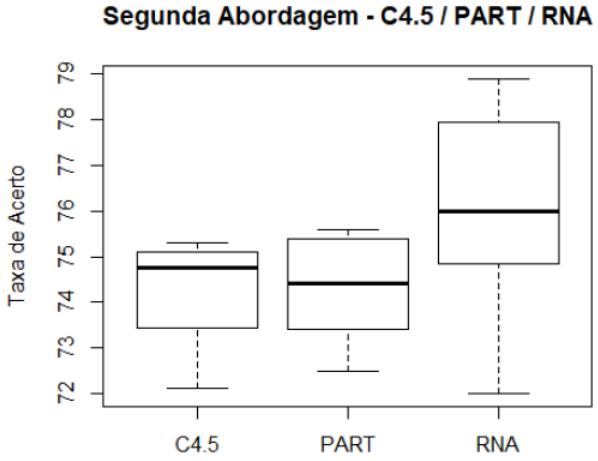
Métrica	C4.5	PART	Rede Neural
Média da taxa de acerto	74.25%	74.31%	76.06%
Índice Kappa	49.14%	50.15%	53.00%
Média do erro absoluto	28.01%	27.43%	27.72%

A Figura 2 mostra o gráfico boxplot dos resultados obtidos na segunda abordagem.

Tabela VI – Resultados estatísticos da primeira abordagem

Métrica	C4.5	PART	RNA
Mínimo	72.10%	72.50%	72%
Máximo	75.29%	75.60%	78.90%
Médiana	74.75%	74.40%	76.10%
Média	74.25%	74.31%	76.60%
Desvio Padrão	1.16%	1.17%	2.28%

Figura 2 – boxplot da segunda abordagem



7 Conclusões e Trabalhos futuros

Com o término deste trabalho, pôde-se observar que a maior dificuldade esteve em lidar com o tratamento de uma base de dados de natureza jurídica, de modo que os resultados finais pudessem convergir para algo dentro do confiável e aceitável aos padrões técnicos e necessidades corporativas e acadêmicas. Além disto, a subjetividade do tipo de informação jurídica (jurisprudência, que gera dados interpretativos)

reduz consideravelmente a eficácia da aplicação de técnicas de aprendizado de máquina, que realizam classificações objetivas.

Como trabalhos derivados e relacionados, visando o aprimoramento dos resultados obtidos e modelos desenvolvidos neste trabalho, pode-se destacar a importância da realização de:

- Análises preditivas utilizando modelos estatísticos e/ou computacionais para estimar percentuais de confiabilidade na utilização de determinadas leis e artigos para dados processos judiciais, tendo como base os resultados já obtidos e modelos desenvolvidos neste trabalho;
- Com o objetivo de auxiliar os juristas, desenvolver aplicação para realização de classificação de documentação normativa segundo conteúdo do assunto. Isto colaboraria para otimizar pesquisas referentes a processos de determinados casos que se encontrem vinculados aos assuntos em questão. A organização desse tipo de informação já proporcionaria aos profissionais da área e pesquisadores que desejam realizar aplicações na mesma, bases específicas para busca e extração do conhecimento desejado;
- Avaliação de envolvimento de súmulas unificadas nas análises de jurisprudência, pois essas informações fornecem maior confiabilidade quanto ao comportamento dos magistrados com relação às análises dos processos;
- Aplicar técnicas de mineração textual afim de analisar todo o contexto geral da publicação, não apenas às leis e artigos.

Referências

[1] Portal CNJ - Processo Judicial Eletrônico (PJe). Disponível em <http://www.cnj.jus.br/tecnologia-da-informacao/processo-judicialeletronico-pje>. Acesso em: 30 de junho 2017.

[2] Portal Conjur. Disponível em <http://www.conjur.com.br/2015-jun 24/conheca-softwares-juridicos-usados advogados>. Acesso em: 30 de junho 2017.

[3] REALE, Miguel. **Lições preliminares de direito**. São Paulo: Saraiva, 2003, p. 321.

[4] REOLON, Suzana Minuzzi. A linguagem jurídica e a comunicação entre o advogado e seu cliente na atualidade. Direito & Justiça, v. 36, n. 2, p. 180-191, jul./dez. 2010. Disponível em: <http://revistaseletronicas.pucrs.br/ojs/index.php/fadir/article/viewFile/9101/6347>.

[5] SILVA, Andréia Gonçalves. **Fontes de informação jurídica: conceitos e técnicas da leitura para o profissional da informação**. Rio de Janeiro: Interciência, 2010.

[6] ALONSO, Cecília Andreotti Atienza. A informação jurídica face às comunidades da área do direito e a dos fornecedores da informação jurídica. In: CIBERNÉTIA, SIMPÓSIO INTERNACIONAL DE PROPRIEDADE INTELECTUAL, INFORMÁTICA E ÉTICA, 1998, Florianópolis. **Anais**... Florianópolis, 1998.

[7] PASSOS, Edilenice (Org.). **Informação Jurídica: teoria e prática**. Brasília, DF: Thesaurus, 2004.

[8] BARROS, Lucivaldo Vasconcelos. Avaliação de Fontes de informação para busca de documentos jurídicos na Internet: uma reflexão à luz das cinco leis de Ranganathan e dos critérios de acessibilidade. In: SEMINÁRIO NACIONAL DE DOCUMENTAÇÃO E INFORMAÇÃO JURÍDICAS, 2., 2010, Brasília. **Anais**... Brasília:

[9] MARTINHO, A. M. O bibliotecário jurídico: identidade e competências profissionais. In: ENCONTRO NACIONAL DE BIBLIOTECAS JURÍDICAS, 1., 2004, Lisboa. **Anais**... Lisboa: Faculdade de Direito da Universidade de Lisboa, 2006.

[10] FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

[11] Portal de Serviços de São Paulo. Disponível em <https://esaj.tjsp.jus.br/cpopg/open.do>. Acesso em: 28 abr. 2017.

http://dx.doi.org/10.25286/repa.v3i3.907

[12] Portal do Tribunal de Justiça do Estado de Minas Gerais. Disponível em <http://www.tjmg.jus.br/portal/>. Acesso em: 28 abr. 2017.

Mineração de Dados na Identificação de Empresas Irregulares Quanto ao Pagamento de Impostos

Data Mining in the Identification of Irregular Companies Regarding the Payment of Taxes

Rafaella Leandra Souza do Nascimento[1] orcid.org/0000-0001-9548-5079
Pedro José Buarque Lins dos Santos[1] orcid.org/0000-0001-8151-9127
Jorge Felipe Lessa Santiago[1] orcid.org/0000-0001-7828-1226
Bettina Cavalcanti Araújo[1] orcid.org/0000-0002-9821-1812
Fernando Baptistella de Lima[1] orcid.org/0000-0002-1021-7321
Alexandre Magno de Andrade Maciel[1] orcid.org/0000-0003-4348-9291

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: rlsn@ecomp.poli.br

Resumo

Este artigo descreve o processo de descoberta de conhecimento utilizando base de dados da Secretaria da Fazenda de Pernambuco. As atividades desempenhadas consistem no pré-processamento dos dados, limpeza, mineração e avaliação dos resultados obtidos. O órgão governamental possui a necessidade de classificar e identificar perfis de empresas com maior potencial de se comportarem de maneira irregular em relação a legislação dos impostos estaduais. Portanto, o objetivo deste trabalho consistiu em aplicar algoritmos de Mineração de Dados, através das tarefas de classificação e clusterização. Os resultados apontam para uma maior taxa de acerto com o classificador Random Forests e identificou níveis de empresas nocivas na base de dados através dos algoritmos de clusterização.

Palavras-Chave: Mineração de Dados; Classificação; Clusterização; Empresas Irregulares; Impostos;

Abstract

This article describes the process of knowledge discovery using the database of the Pernambuco Department of Finance. The activities performed consist of data pre-processing, cleaning, mining and evaluation of the results obtained. The government agency has the need to classify and identify profiles of companies with greater potential to behave in an irregular manner in relation to the state taxes legislation. Therefore, the objective of this work was to apply Data Mining algorithms, through the tasks of classification and clustering. The results point to a higher hit rate with the Random Forests classifier and identified levels of noxious companies in the database through clustering algorithms.

Key-words: *Data Mining; Classification; Clustering; Irregular Companies; Taxes.*

 http://dx.doi.org/10.25286/repa.v3i3.932

1 Introdução

ICMS é a sigla para Imposto sobre a Circulação de Mercadorias e Serviços e é o principal imposto arrecadado pelo Governo de Pernambuco, tendo recebido, de acordo com os dados da Secretaria da Fazenda de Pernambuco (SEFAZ), mais de dois bilhões e quinhentos milhões no ano de 2015 **[1]**. Esse imposto é regulamentado pelo artigo 155, II e § 2ª da Constituição Federal de 1988 e é obrigatório às empresas, produtores rurais e prestadoras de serviços **[2]**. Todas elas devem possuir uma inscrição estadual junto a Secretaria da Fazenda.

A Secretaria da Fazenda é um órgão administrativo do Poder Executivo e tem por finalidade desenvolver e executar a política de tributos do Estado. Ela é responsável pela tributação, arrecadamento e fiscalização desses tributos. É estimado pela SEFAZ do Estado de Pernambuco que em 2015 aproximadamente 105 milhões de reais foram sonegados **[1]**. Para o órgão há a necessidade de melhor monitorar as empresas com o perfil nocivo, ou seja, empresas que possuem potencial de se comportarem de maneira irregular em relação a legislação dos impostos, como o ICMS.

Para o auxílio nesta tarefa, a SEFAZ possui atualmente um sistema de classificação de empresas como nocivas e não nocivas. Para isto, é utilizada como entrada para um classificador de redes neurais artificiais, uma base de dados com informações cadastrais sobre as empresas. No entanto, este modelo não realizou nenhuma tarefa anterior de limpeza, pré-processamento e tratamento dos dados, nem foi realizado nenhuma medição de eficácia do modelo de classificação desenvolvido.

Sendo assim, o objetivo deste trabalho consiste em aperfeiçoar a identificação do perfil de empresas nocivas, ou não, com novos experimentos utilizando a técnica de redes neurais artificiais, comparando os resultados com a técnica Random Forest, SVM e método *Ensemble*. É realizado um trabalho de tratamento dos dados e ajuste do modelo a fim de obter melhores índices de acerto sob a classificação realizada. Também, criar grupos com perfis de nocividade semelhantes, e para isto, será utilizada a técnica de clusterização.

Este trabalho está dividido da seguinte forma: o item 2 apresenta a Fundamentação Teórica, onde encontram-se trabalhos relacionados ao tema deste artigo, assim como definições sobre as técnicas utilizadas para resolução do problema; no item 3 é desenvolvido os Materiais e Métodos utilizados, onde são apresentadas as informações referentes à base de dados utilizada e as técnicas utilizadas no desenvolvimento deste trabalho; o 4 mostra os Experimentos Realizados; e o item 5 compõe Conclusões e Trabalhos Futuros, desenvolvidos após às análises dos resultados obtidos ao final dos experimentos realizados.

2 Trabalhos Relacionados

Com o aumento das irregularidades na contabilidade financeira evidenciado no atual cenário econômico, o tema de detecção de fraude tornou-se de grande importância para o setor acadêmico, de pesquisas, político e industrial. As falhas presentes nos sistemas de auditoria e controle internos criaram a necessidade de as organizações usarem procedimentos mais especializados para detectar a fraude financeira, seja ela de qualquer natureza.

Desta forma, técnicas de mineração de dados estão fornecendo grande ajuda na detecção destas irregularidades, uma vez que lidar com a complexidade de grandes volumes de dados financeiros são grandes desafios para quem administra as organizações.

Tendo como objetivo analisar o processo gerencial e de análise e suporte dos dados, Power e Power **[3]** faz um levantamento sobre fraudes em empresas de seguros. Expõem que para este cenário o problema é multimilionário, e este pode ser detectado e impedido se os dados forem coletados corretamente, bem analisados e compartilhados entre companhias de seguro, para assim serem aplicadas técnicas de suporte e análise apropriadas.

Power e Power **[3]** ainda desenvolvem que para criar estas capacidades de apoio à decisão deve haver um envolvimento de questões gerenciais, tecnológicas e de propriedade de

dados. Portanto, artigo desenvolvido examina tais questões no contexto do uso de novas fontes de dados e análise preditiva para reduzir a fraude de seguros e melhorar o serviço ao cliente. Um modelo de processo é desenvolvido para incentivar a discussão e a inovação na detecção e redução de fraudes.

Já o trabalho de Junqué de Fortuny *et al.* **[4]** aborda o tema de fraude de residência corporativa, onde há uma limitação de pesquisas por causa da disponibilidade dos dados e da alta sensibilidade destes. Esta pesquisa contou com a colaboração do governo Belga, o qual se propôs a abordar o tema cooperando com outras instituições (como a academia), sendo o objetivo final ter um sistema de tributação justo e eficiente. No trabalho é descrito os problemas envolvidos na construção de tal sistema de detecção de fraude, que são principalmente relacionados com dados (por exemplo, assimetria de dados, qualidade, volume, variedade) e relacionados com a implementação (por exemplo, a necessidade de explicações das previsões feitas).

Ainda, para de Moura, de Lavor Lopes e Faria **[5]** é levantado o tema de sonegação fiscal no Brasil. O governo federal vem elaborando métodos informatizados de obrigações a serem entregues a fim de se evitar a prática de sonegação, obrigações das quais tem o objetivo de mostrar a situação da empresa em um todo, pois através destes pode-se ter clareza nas operações de entradas e saídas de uma empresa. Com a entrega dessas obrigações as empresas passam por uma auditoria externa mensal. Através de pesquisas bibliográficas, neste artigo evidenciou-se a história da auditoria externa de forma geral, assim, podendo identificar métodos de auditoria que facilitam ajudar a detectar uma empresa que está praticando a sonegação de impostos.

Segundo Silva **[6]**, algumas formas de informatização das informações existentes consistem em notas fiscais eletrônicas, o Sistema Público de Escrituração Digital (SPED), as diversas declarações existentes, entre outros. Além disto, cita que no processo de auditoria por parte dos fiscais, ainda há grande dificuldade em detectar a fraude, principalmente por muitas vezes algumas informações estarem ausentes. Esta informatização consiste em um passo para terem sistemas de fiscalização mais confiáveis.

No que se trata de técnicas existentes para processar dados e para apoio à decisão, o trabalho de Sharma e Panigrahi **[7]** realiza uma revisão abrangente da literatura sobre a aplicação de técnicas de mineração de dados para a detecção de fraudes de contabilidade financeira. Esta revisão sistemática e abrangente da literatura das técnicas de mineração de dados aplicáveis à detecção de fraude tem o intuito de fornecer uma base para pesquisas neste campo. Como resultados foi exposto que técnicas de mineração de dados como redes neurais, modelos logísticos, redes bayesianas e árvores de decisão foram aplicadas mais extensamente para fornecer soluções para os problemas inerentes à detecção e classificação de dados fraudulentos.

Tendo em vista estes trabalhos, evidencia-se que o tema de detecção de fraude, nos mais diferentes tipos de organizações, é de bastante interesse. Abrangendo desde os processos gerenciais, de controle e tratamento dos dados, até o apoio a decisão com base em técnicas de extração de informação. Sendo assim, a realização deste trabalho mostra-se de bastante contribuição para esta área de interesse.

3 Materiais e Métodos

Este capítulo consiste em fazer uma descrição das bases de dados selecionadas para estudo, assim como descrever o desenvolvimento com base no processo KDD, desenvolvido por Fayyad et. al**. [8].**

3.1 Descrição da Base de Dados

A base fornecida pela Secretaria da Fazenda de Pernambuco, caracteriza o cenário onde empresas ou entidades estaduais que tem seu nível de licitação medido através de uma série de 46 atributos categóricos e numéricos, sendo inicialmente estes atributos utilizados para fazer uma classificação de empresas ou entidades estaduais em lícitas ou ilícitas. Ela possui uma coleção de 662.942 instâncias do problema mencionado, estando relativamente desbalanceada contendo 613.658 instâncias classificadas como lícitas e apenas 49.284 instâncias classificadas como ilícitas.

Ainda sobre a base fornecida pela Secretaria da Fazenda de Pernambuco, algumas

http://dx.doi.org/10.25286/repa.v3i3.932

considerações iniciais valem ser feitas a respeito dos atributos coletados. Dentre os 46 atributos, majoritariamente 39 são variáveis categóricas (podendo ser nominal ou ordinal) e apenas 7 são atributos contendo valores numéricos.

3.2 Análise Descritiva dos Dados

A análise descritiva dos dados é uma importante etapa para o processo de descoberta do conhecimento, pois, antes de executar as técnicas de mineração, pode-se utilizar recursos capazes de explicar de uma forma prévia os dados. Estes podem ser organizados usando distribuição de frequência, por exemplo, representados visualmente por mapas, gráficos, diagramas. Desta forma, alguns pontos da base de dados foram definidos como importantes para esta análise inicial.

A Figura 1 mostra a distribuição dos segmentos econômicos por região de desenvolvimento. Pode ser observado que, há uma elevada concentração de empresas na região metropolitana do Recife, com pico para o segmento econômico de valor 17 (Fabricação de Celulose, Papel e Produtos de Papel). As demais regiões não possuem valores elevados, mas, de certa forma mantém-se constantes entre si.

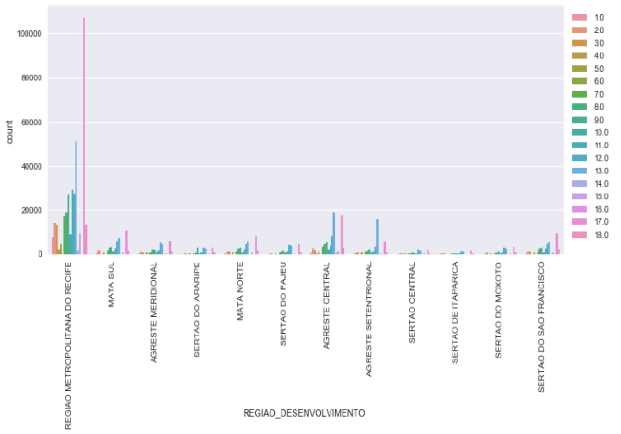


Figura 1 - Distribuições do segmento econômico por região de desenvolvimento.
Fonte**:** Autores.

As análises preliminares também indicam que o maior nível de arrecadação no período de 12 meses do estado de Pernambuco está concentrado na região metropolitana do Recife, como mostrado na Figura 2, com pico para a classe "A", ou seja, indica que a maioria das empresas concentram o menor nível de arrecadação.

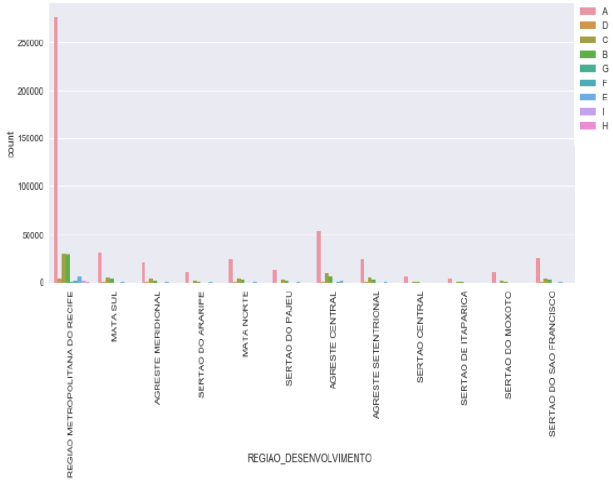


Figura 2 - Distribuições da arrecadação do período de 12 meses por região de desenvolvimento.
Fonte**:** Autores.

O gráfico da Figura 3 mostra a distribuição das classes por segmento econômico, onde a classe "0" representa as empresas não nocivas e a classe "1" representa as empresas nocivas. É notório que devido ao desbalanceamento da base para as classes, a cor azul é mais frequente (classe "0"). No entanto, podemos analisar que a cor verde (classe "1") possui destaque para segmento econômico representado, por exemplo, pelas classes 17, 13 e 11, cujos segmentos são: Fabricação de Celulose, Papel e Produtos de Papel; Fabricação de Produtos Têxteis; e Fabricação de Bebidas, respectivamente. Isto quer dizer que há uma concentração maior de empresas nocivas nesses segmentos.

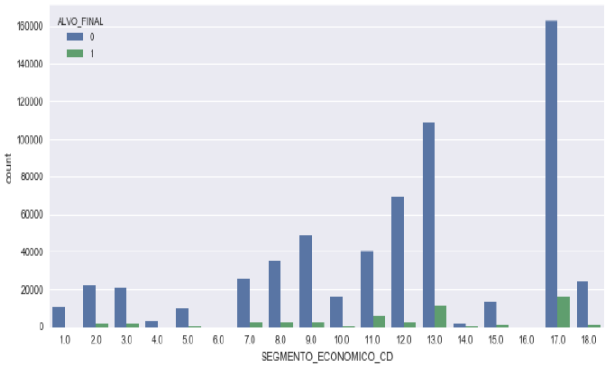


Figura 3 - Distribuições do grau de nocividade por segmento econômico (valor 0 significa lícita e valor 1 ilícita).
Fonte**:** Autores.

3.3 O Processo KDD

Atualmente a mineração de dados é aplicada comumente para fazer correlações, encontrar padrões, oferecendo maior clareza para análise da base de dados. Contudo, atualmente a mineração pode ser considerada como uma parte do processo de descoberta de conhecimento em banco de dados (KDD – *Knowledge Discovery in Databases*), desenvolvido por Fayyad et. al. **[8]**.

A descoberta de conhecimento em bancos de dados (KDD) é um processo amplo consistindo em algumas etapas. Seu uso tem como finalidade melhorar a qualidade dos dados que serão processados, refletindo consequentemente nos resultados obtidos. De acordo com as fases do processo, a inicial é a seleção e o pré-processamento, na qual o foco da seleção consiste nas escolhas dos possíveis dados e registro da análise da massa de dados a ser minerada, podendo ser um conjunto de dados ou um subconjunto de variáveis onde a extração será realizada. Já o pré-processamento visa assegurar a qualidade dos dados, eliminando os possíveis ruídos e dados discrepantes do conjunto.

A fase seguinte consiste na transformação, em que os dados serão transformados utilizando o padrão ideal para aplicação de algoritmos de mineração. Na fase de mineração de dados são aplicadas algumas técnicas inteligentes para obter padrões de interesse de determinadas variáveis dos registros. A mineração de dados possui classificações das determinadas tarefas **[9]**, as mais utilizadas são classificação, análise de agrupamento e associação.

Por fim, a etapa de interpretação e avaliação visa encontrar padrões interessantes de acordo com algum critério estabelecido na análise, sendo assim utilizado técnicas de representação de conhecimento. Consequentemente, após a extração de conhecimento é realizada a tomada de decisão, que visa otimizar processos, podendo definir estratégias mais adequadas para se aplicar a determinados cenários.

3.3.1 Seleção dos dados

Na análise inicial foi constatado que apenas um grupo reduzido de 18 atributos se mostraram candidatos de relevância para o problema em questão. Chegou-se a estes atributos com ajuda de um especialista da área do problema.

3.3.2 Pré-Processamento e Transformação dos dados

Foi possível observar inconsistências na base fornecida, como atributos ausentes para algumas das instâncias do problema, bem como má representação dos dados para a posterior aplicação de técnicas numéricas para detecção de padrões. Para poder superar estes problemas, se fez uma análise estatística dos dados bem como transformações dos valores representados por "0" e/ou "-1" para não existentes (NaN).

Para ter uma melhor visualização quanto aos dados faltantes, é feito um mapeamento dos dados, como mostra a Figura 4. Pode-se notar que, para as colunas CONTADOR_PESSOA_CD, PADRAO_A1 até PADRAO_A5 a quantidade de valores ausentes é grande (cor amarela), então foi realizada a eliminação vertical, ou seja, estas colunas foram excluídas, uma vez que a perda de informação é elevada e a importância das informações para o estudo não é indispensável.

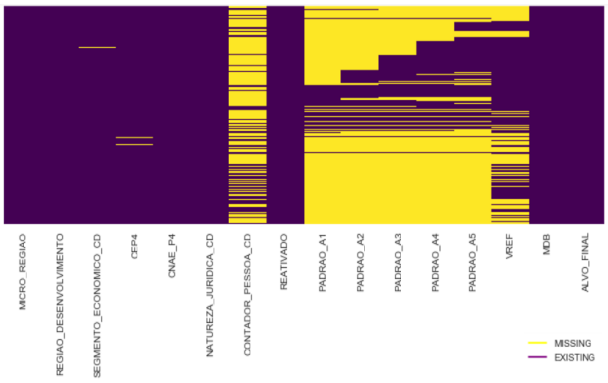


Figura 4 - Mapeamento dos dados faltantes.
Fonte**:** Autores.

A Figura 6 mostra o mapeamento dos dados após a eliminação vertical realizada. Pode-se perceber que a coluna VREF não foi excluída, isto se dá pela importância do seu significado para o estudo em questão. Para resolver os *missing values* de VREF é usada a técnica de interpolação, na qual cada valor nulo é substituído pela média do VREF para cada região de desenvolvimento (REGIAO_DESENVOLVIMENTO).

Ainda, como pode ser observado na Figura 5 encontram-se dados faltantes nas colunas SEGMENTO_ECONOMICO_CD e CEP4, mas como se apresentam em menor quantidade, é realizada

http://dx.doi.org/10.25286/repa.v3i3.932

a eliminação horizontal, ou seja, os registros (linhas) que não possuem valores são excluídos. A Tabela 1 contém as colunas finais para aplicar as técnicas de mineração, totalizando 10 colunas.

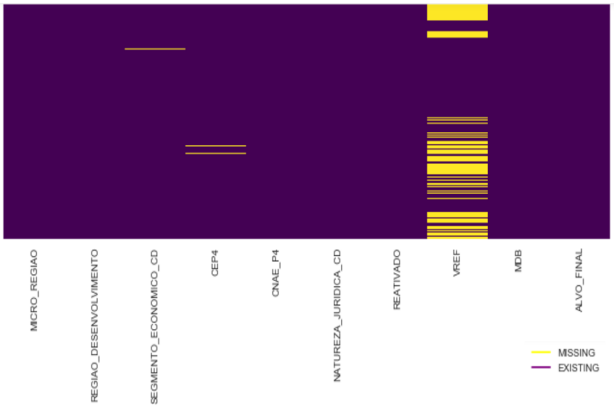


Figura 5 - Mapeamento dos dados após eliminação vertical.
Fonte**:** Autores.

Com exceção da coluna VREF, na qual se aplicou a técnica de interpolação mencionada anteriormente, e da coluna ALVO_FINAL, na qual se apresentam as classes de saída, todas as outras colunas residuais são categórico-nominais. Tendo isso em mente, aplicou-se a técnica de transformação para variáveis categórico-nominais, e após essa transformação obtiveram-se 77 colunas onde foram efetuados os primeiros testes de classificação utilizando Redes Neurais MLP e Random Forests. Ao final de todas as transformações e eliminações, a base final constituiu-se de 77 colunas e de um total de 646.846 instâncias.

Tabela 1 - Dicionário de dados após pré-processamento e tratamento dos dados.

Num	Nome da Variável	Num	Nome da Variável
1	MICRO_REGIAO	**6**	NATUREZA_JURIDICA
2	REGIAO_DESENVOLVIMENTO	**7**	REATIVADO
3	SEGMENTO_ECONOMICO_CD	**8**	VREF
4	CEP4	**9**	MDB
5	CNAE_P4	**10**	ALVO_FINAL

Fonte**:** Autores.

3.3.3 Mineração de Dados

Para a implementação das técnicas de mineração, foi utilizada a biblioteca *Scikit-learn* (sklearn), que é *open source*, desenvolvida em *Python* e interage com outras bibliotecas como *Numpy/Scipy* e *Matplotlib*. Ela inclui vários algoritmos de classificação, regressão, agrupamento, como SVM, redes neurais, Random Forest, *K-means*, entre outros. Neste trabalho, a implementação se deu pelos classificadores de rede neural MLP e Random Forest.

Os experimentos se deram em dois processos, particionando os dados em treino e teste, sendo 70% dos dados e 30%, respectivamente. As configurações da rede neural MLP e Random Forest utilizadas são mostradas na Tabela 2. Tais configurações foram obtidas utilizando uma combinação das metodologias de busca *Grid Search* para busca de parâmetros ótimos juntamente com *Stratified K-Fold Cross Validation*. Esta metodologia faz combinações entre os parâmetros da técnica de aprendizado de máquina a fim de encontrar as melhores configurações.

4 Resultados dos Experimentos

Nesta seção são apresentados os experimentos realizados para os algoritmos de classificação, através da base de dados balanceada e não balanceada, e para os algoritmos de clusterização.

4.1 Algoritmos de Classificação

Os resultados dos experimentos resultam na classificação da base pela rede neural MLP definida e pelo Random Forest. A Figura 6 apresenta o *output* da execução, por meio da matriz de confusão e a Tabela 2 mostra os comparativos entre as métricas de desempenho das técnicas de mineração.

Como pode ser observado na Tabela 2, a rede neural MLP e a Random Forest classificam muito bem a classe "0", mas a performance de ambas é terrível para classificar instâncias da classe "1".

Isso se dá devido ao altíssimo nível de desbalanceamento entre essas duas classes. Para a classe "0" existem um total de 599.854 instâncias e para a classe "1" um total de 46.992 instâncias.

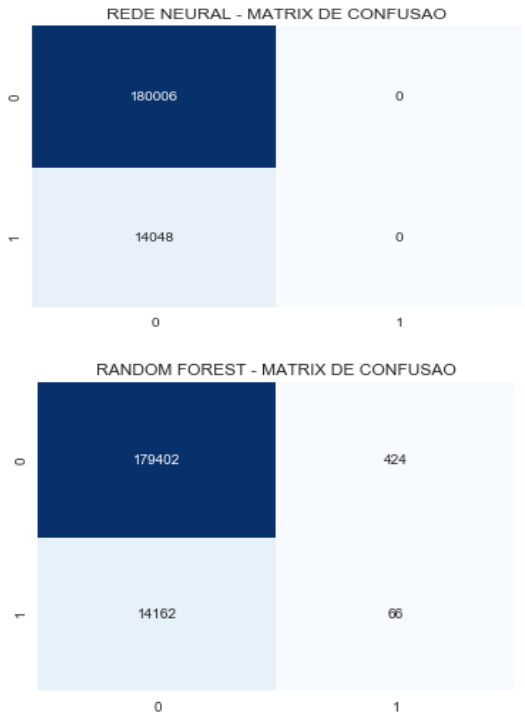


Figura 6 - Matriz de confusão para a rede neural MLP acima e matriz de confusão para Random Forest abaixo. Fonte**:** Autores.

Tabela 2 - Comparativo entre métricas de desempenho dos algoritmos. Legenda: C – classe; P – precision; R – recall; F-S – f-score; S – support.

Tec.	C	P	R	F-S	S	Score
MLP	0	0,93	1,00	0,96	180006	**0,927**
	1	0,00	0,00	0,00	14048	
	Avg/ total	0,86	0,93	0,89	194054	
Random Forest	0	0,93	1,00	0,96	179826	**0,924**
	1	0,13	0,00	0,01	14228	
	Avg/ total	0,87	0,92	0,89	194054	

Fonte: Autores.

Para resolver esse problema foi feito um balanceamento entre classes onde tentou-se igualar a quantidade de instâncias para ambas por meio de uma extração aleatória. Após isso, é realizada a seleção de variáveis utilizando-se a própria técnica Random Forest, que pode ser utilizada para medir a importância de cada coluna.

Com este resultado, selecionou-se então as 10 *features* mais relevantes de acordo com a técnica Random Forest e se construiu uma nova base de dados, dessa vez balanceada e com as colunas de maior importância selecionadas. Após isto, foi feito um novo treinamento para a rede MLP e a Random Forest. Os resultados obtidos são mostrados na Figura 7 e na Tabela 3.

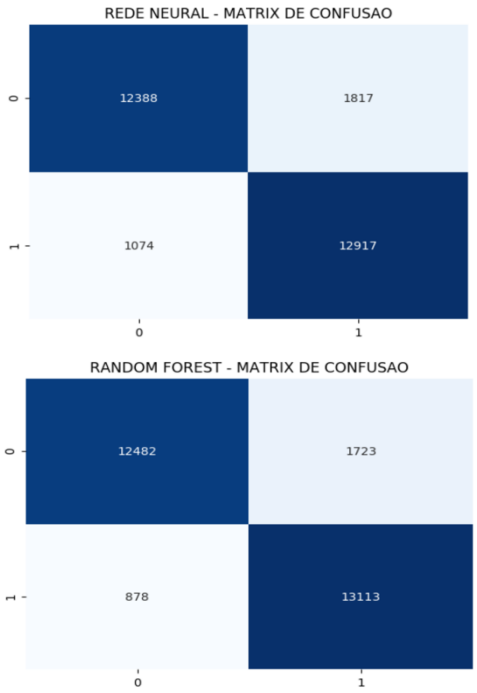


Figura 7 - Matriz de confusão para a rede neural MLP a esquerda e matriz de confusão para Random Forest a direita após o balanceamento e seleção de *features.* Fonte**:** Autores.

Tabela 3 - Comparativo entre métricas de desempenho dos algoritmos após o balanceamento e seleção de *features.* Legenda: C – classe; P – precision; R – recall; F-S – f-score; S – support.

Tec.	C	P	R	F-S	S	Score
MLP	0	0,92	0,87	0,90	14205	**0,897**
	1	0,88	0,92	0,90	13991	
	Avg/ total	0,90	0,90	0,90	28196	
Random Forest	0	0,93	0,88	0,91	14205	**0,907**
	1	0,88	0,94	0,91	13991	
	Avg/ total	0,91	0,91	0,91	28196	

Fonte: Autores.

Após a execução para a base balanceada e com as colunas de maior importância selecionadas, duas novas técnicas são incluídas, pois antes estas se mostraram inviáveis por exigirem grande esforço computacional. A primeira delas foi o SVM (Support Vector

http://dx.doi.org/10.25286/repa.v3i3.932

Machines), com configuração de kernel rbf, C de 10 e gamma de 0.001; e a outra foi uma técnica *ensemble* das três técnicas já utilizadas. Uma técnica *ensemble* nada mais é do que a execução combinada das três técnicas juntas e o resultado final foi obtido através do voto majoritário. Os resultados obtidos são mostrados na Figura 8 e Tabela 4.

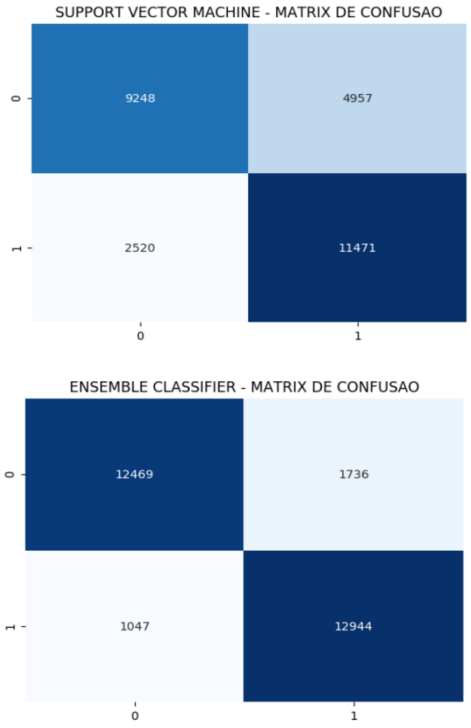


Figura 8 - Matriz de confusão para o SVM a esquerda e matriz de confusão para Ensemble Classifier a direita após o balanceamento e seleção de *features.*
Fonte**:** Autores.

Tabela 4 - Comparativo entre métricas de desempenho dos novos algoritmos após o balanceamento e seleção de *features.* Legenda: C – classe; P – precision; R – recall; F-S – f-score; S – support.

Tec.	C	P	R	F-S	S	Score
SVM	0	0,79	0,65	0,71	14205	**0,734**
	1	0,70	0,82	0,75	13991	
	Avg/ total	0,74	0,73	0,73	28196	
Ensemble	0	0,92	0,88	0,90	14205	**0,901**
	1	0,88	0,93	0,90	13991	
	Avg/ total	0,90	0,90	0,90	28196	

Fonte: Autores.

Ao final da execução de todas as técnicas de classificação foi construído a curva ROC (*Receiver Operator Characteristic*), como mostra a Figura 9, que mede a eficiência do classificador permitindo-se obter visualmente uma análise dos classificadores a respeito da taxa de falsos positivos e verdadeiros positivos. O melhor classificador em termos de taxa falsos positivos e verdadeiros negativos seria a curva que mais se aproxima ao eixo da esquerda e ao eixo superior do gráfico.

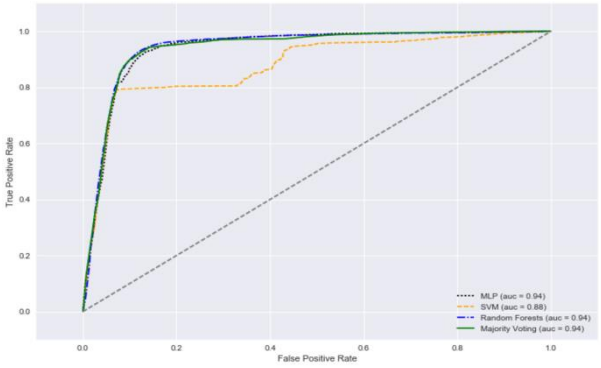


Figura 9 - Curva ROC das técnicas utilizadas no processo: MLP, Random Forest, SVM e Ensemble com Voto Majoritário.
Fonte: Autores.

4.2 Algoritmos de Clusterização

Após todos os resultados dos experimentos para os algoritmos utilizados na tarefa de classificação, é iniciado a tarefa de clusterização. O objetivo desta tarefa consiste em identificar os níveis de nocividade para empresas presentes na base de dados. Para isto, é levado em consideração apenas os registros classificados como nocivos na base, correspondente a classe "1". Inicialmente, existem 49.284 instâncias pertencentes a classe "1", e devido a limitações no processo de experimentação, foi utilizado uma amostra aleatória com 10.000 instâncias nos experimentos para o K-means, Fuzzy C-Means e PSO.

Os resultados são mostrados na Figura 10, Figura 11 e Figura 12 para o K-means, Fuzzy C-Means e PSO, respectivamente. Foi possível mostrar os gráficos da relação entre a variância intra-*cluster* e o número de grupos, e procurar assim por um ponto de estagnação no processo

de minimização dessa métrica, o que indica o número ideal de grupos. Esta é uma boa forma de estimar o número de *clusters*, já que indica que o conjunto de *cluster* é bom para um certo K.

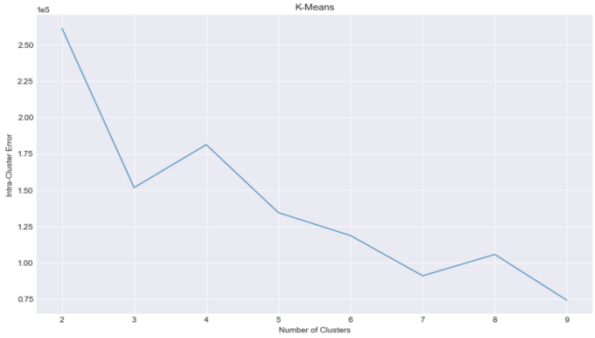


Figura 10 - Gráfico da relação entre o erro intra-*cluster* e o número de grupos para o K-Means.
Fonte**:** Autores.

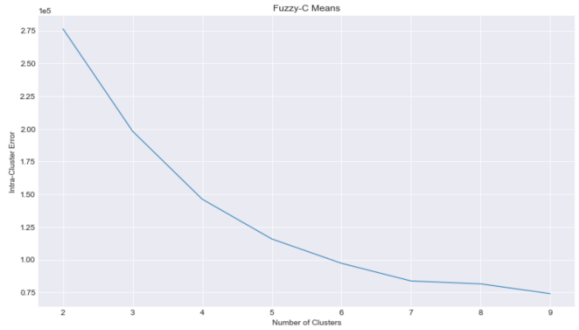


Figura 11: Gráfico da relação entre o erro intra-*cluster* e o número de grupos para o Fuzzy C-Means.
Fonte**:** Autores.

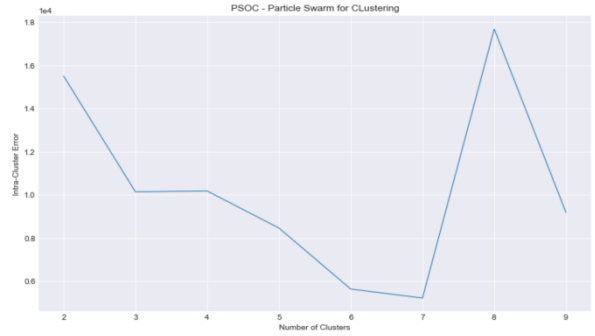


Figura 12: Gráfico da relação entre o erro intra-*cluster* e o número de grupos para o PSO.
Fonte**:** Autores.

5 Análises e Discussões

Esta seção consiste na discussão e avaliação das técnicas utilizadas com seus respectivos resultados obtidos referentes ao capítulo 4. Comparando o modelo antes e após o balanceamento e com a seleção dos *features* selecionadas.

5.1 Estimação da classificação de empresas nocivas

Como pode ser observado na Tabela 2, a rede neural MLP é ótima para classificar elementos da classe "0", mas obtém resultados ruins para elementos de classe "1" na base desbalanceada. O Random Forest apresenta resultados um pouco melhores para a classificação de elementos da classe "1". Apesar de obterem a taxa de acerto parecido, 92,7% para MLP e 92,4% para Random Forest, analisando as métricas de desempenho, percebe-se que, apesar de pequena, a precisão para o Random Forest é maior.

Para o experimento após o processo de balanceamento e seleção de 10 *features* mais relevantes de acordo com a técnica Random Forest, verifica-se que houve uma melhoria bastante significativa quanto ao resultado da classificação. Como pode ser observado na Tabela 3, a rede neural MLP passa a ter um desempenho muito superior para a classe "1". A técnica Random Forest também apresentou resultados muito melhores para a classificação de elementos da classe "1", e em comparação a MLP e os resultados da Random Forest mostraram-se, em média, superiores.

Foram feitas a adição de duas novas técnicas para efeitos comparativos uma vez que a base foi balanceada, como mostra Tabela 4. O SVM se mostrou pouco eficiente para esse trabalho de classificação tendo uma performance bastante inferior a rede neural MLP e a Random Forest, com a taxa de acerto parecido de 73,4%. Entretanto, a técnica *ensemble* com voto majoritário teve uma performance muito próxima da Random Forest, isto é, com a taxa de acerto de 90,1%**,** porém com relativo maior esforço computacional. Com base na figura 9, a curva ROC, percebemos que a Random Forest é que melhor técnica com taxa verdadeiro positivo/falso positivo e também a que possuiu as melhores taxas nas métricas computadas, e por ser menos custosa do que a técnica *ensemble* aplicada, pode-se então concluir que seria a mais apropriada para ser aplicada nesse problema.

5.2 Caracterização do perfil de empresas nocivas

http://dx.doi.org/10.25286/repa.v3i3.932

Como visto na Figura 10, Figura 11 e Figura 12, são mostrados os gráficos que indicam o número de grupos que minimizam a variância intra-cluster. Para valores de K de 2 até 10, são executados experimentos para as diferentes técnicas de clusterização.

Para o K-Means o melhore valor de K é 3. Como pode ser visto no gráfico da Figura 10, a depressão maior (e que minimiza o erro) está presente para esse valor, tendo em vista que o algoritmo apresentou dificuldades na minimização da métrica intra-cluster para valores subsequentes de K. Já para o Fuzzy C-Means não se pode fazer uma inferência do valor ideal de K, tendo em vista que o algoritmo apresentou o resultado esperado. O gráfico da Figura 11 mostra que a curva que minimiza o erro é decrescente em relação ao valor de K, o que é um comportamento esperado para essa métrica. Por fim, para o PSO, o K que minimiza o erro é para K igual a 7, pois como pode ser observado no gráfico da Figura 12, percebe-se que esse é o menor valor de K antes de um comportamento de maximização indesejado.

Analisando os resultados, e com base na literatura e nas técnicas de agrupamento para minimizar o erro intra-*cluster*, podemos inferir que podem existir 3 ou 7 perfis de empresas nocivas na base de dados. Para garantir uma inferência mais assertiva, seria interessante realizar mais simulações e utilizar diferentes métricas de desempenho para validar onde os algoritmos convergem em opinião.

6 Conclusões

Com a análise dos resultados obtidos neste trabalho pode-se concluir que os objetivos foram alcançados, uma vez que se obteve resultados satisfatórios nas duas tarefas de descoberta de conhecimento. Primeiramente, a classificação das empresas quanto à nocividade, a qual era a maior necessidade da Secretaria da Fazenda de Pernambuco, mostrou métricas de desempenhos para os algoritmos utilizados bastante significativas. Inicialmente, pretendia-se apenas utilizar a técnica de redes neurais, no entanto, as demais técnicas utilizadas foram bastante importantes para assim confrontar os resultados entre os classificadores, determinando assim, os resultados do Random Forest como os melhores (90,7% de acerto).

Posteriormente, com a clusterização buscou-se determinar níveis de nocividade entre as empresas presentes na base de dados. Com os resultados desta análise a organização em questão pode criar estratégias diferenciadas para lidar com as empresas de acordo com seu nível ilícito. Os resultados das técnicas de clusterização mostraram que existem 3 ou 7 níveis de nocividade entre as empresas pernambucanas.

Os fatores importantes para determinar os bons resultados consistiram nas diferentes análises dos dados e estratégias de pré-processamento adotadas ao longo do processo de descoberta de conhecimento, as quais muitas vezes precisaram ser refeitas e lapidadas para um maior acerto no resultado final. Estas estratégias foram importantes, inclusive para escapar de limitações no processamento das técnicas, uma vez que a base de dados utilizada possui uma grande quantidade de informação e desbalanceamento entre classes. Tudo isto mostra que a necessidade de conhecer bem o problema e os dados em questão são decisivas para o bom andamento do processo, principalmente quando estes possuem grande volume e complexidade.

Referências

[1] SECRETARIA DA FAZENDA DE PERNAMBUCO. SEFAZ. Disponível em: <http://www.sefaz.pe.gov.br/RPM/Scripts/TransfConstitucionalRelatorio2.asp>. Acesso em: 27 abr. 2017.

[2] BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília: Senado Federal, 1988.

[3] POWER, Daniel J.; POWER, Mark L. Sharing and Analyzing Data to Reduce Insurance Fraud. In: ANNUAL MWAIS CONFERENCE, 10., 2015, Pittsburg. **Proceddings**... Pittsburg, 2015.

[4] JUNQUÉ DE FORTUNY, Enric et al. Corporate residence fraud detection. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 20., 2014**,**.New York. **Proceedings**... New York: ACM, 2014. p. 1650-1659.

[5] MOURA, Renan Gomes de; LAVOR LOPES, Paloma de Lavor; FARIA, Sandi Siqueira L. A. O papel da auditoria externa no combate à sonegação. **Cadernos UniFOA**, v. 11, n. 31, p. 75-86, 2016.

[6] SILVA, Jéssica Bonomo. **Sonegação fiscal**: percepções de fiscalizações tributárias nos órgãos federais, estaduais e municipais. Monografia. Bacharelado em Ciências Contábeis, Universidade Caxias do Sul. Rio Grande do Sul, 2017.

[7] SHARMA, Anuj; PANIGRAHI, Prabin K. A review of financial accounting fraud detection based on data mining techniques. **International Journal of Computer Applications**, v. 39, n. 1, 2012.

[8] FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases.**AI magazine**, v. 17, n. 3, p. 37, 1996.

.

[9] LAROSE, Daniel T. **Discovering knowledge in data: an introduction to data mining**. New Jersey: John Wiley & Sons, 2014.

 http://dx.doi.org/10.25286/repa.v3i3.932

Projeto 500 Cities: Detecção de Comunidades utilizando Algoritmos de Clusterização

500 Cities Project: Detection of Communities Using Clustering Algorithms

Anderson Vinícius Alves Ferreira[1] orcid.org/0000-0001-8598-6574
Lizandra Raflesia Monteiro de Lira[1] orcid.org/0000-0002-2379-5868
Thiago José da Silva[1] orcid.org/0000-0002-1710-2148
Carmelo José Albanez Bastos Filho[1] orcid.org/0000-0002-0924-5341

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: avaf@ecomp.poli.br

Resumo

A saúde pública é uma extensa área com problemas complexos e que constantemente necessita de bastantes investimentos. Frequentemente os órgãos governamentais enfrentam desafios para entender como oferecer melhores serviços de saúde e prevenir epidemias futuras. Métodos preventivos tem sido até o momento a melhor opção para controlar doenças e epidemias ou mesmo extingui-las. Este trabalho utilizou dados epidemiológicos provenientes do projeto 500 Cities e técnicas de agrupamento (clusterização) de dados para identificar comunidades com características relevantes para dar suporte na prevenção de epidemias e doenças.

Palavras-Chave: Saúde Pública; Agrupamento; Clusterização; Comunidades;

Abstract

Public health is a large area with complex problems and constantly needs huge investments. Frequently, government agencies face challenges in understanding how to deliver effective and targeted health services and prevent future epidemics. Prevention has so far been the best option to control diseases and epidemics or even extinguish them. This work used epidemiological data provided by the 500 Cities project along with data clustering techniques to identify communities with relevant characteristics to support epidemics and diseases prevention.

Key-words:*Public health; Clustering; Communities;*

 http://dx.doi.org/10.25286/repa.v3i3.966

1 Introdução

A saúde pública é uma extensa área com problemas complexos e que constantemente necessita de bastantes investimentos. Frequentemente, os órgãos governamentais enfrentam desafios para entender como oferecer melhores serviços de saúde e prevenir epidemias futuras. Como controlar as epidemias? Como identificar as epidemias? Como o governo pode interferir na prevenção de epidemias e doenças? Existem tendências regionais de saúde? Existem regiões com comportamento incomum de piora na saúde? Alguma região se destaca por altos níveis de bem-estar? Será que é possível predizer as condições de saúde de uma cidade baseado na situação de cidades vizinhas?

A fim de criar soluções e intervenções plausíveis e efetivas para as atuais necessidades, os métodos preventivos têm sido até o momento a melhor opção, fazendo com que em um futuro próximo algumas doenças sejam controladas ou mesmo extintas. Torna-se importante, portanto, saber quais os hábitos de saúde das populações das cidades e quais impactos tais hábitos poderiam ter sobre as condições gerais de saúde das populações.

Em 2015, nos Estados Unidos, a Fundação Robert Wood Johnson lançou o projeto 500 Cities **[1]** (em tradução livre, 500 Cidades), que tem como objetivo reportar dados epidemiológicos das 500 maiores cidades americanas. O projeto prevê que os dados sejam utilizados por órgãos governamentais para ajudar a desenvolver e implementar atividades de prevenção efetivas e direcionadas, identificar problemas de saúde pública emergentes, e estabelecer e monitorar objetivos relevantes para a saúde pública.

O conjunto de medidas disponibilizado pelo projeto *500 Cities* baseia-se em doenças crônicas prioritárias e de maior impacto na saúde pública. As medidas incluem os comportamentos de maior risco que causam doenças, sofrimento e morte precoce relacionados a doenças e condições crônicas, bem como as condições e doenças mais comuns, dispendiosas e previníveis entre todos os problemas de saúde. O total de 27 medidas inclui 5 comportamentos não saudáveis, 13 indicadores de saúde e 9 práticas de prevenção.

O projeto *500 Cities* representa o primeiro projeto a fornecer informações em larga escala para cidades e pequenas áreas dentro das cidades. E, apesar dos dados reportados serem provenientes de questionários respondidos pelos residentes das cidades, pode ser possível obter análises interessantes a respeito das condições de saúde de uma determinada população.

Este trabalho visa a identificar características relevantes para dar suporte na prevenção de epidemias e doenças e agrupar as diferentes cidades de acordo com seus hábitos e condições de saúde.

2 Fundamentação teórica

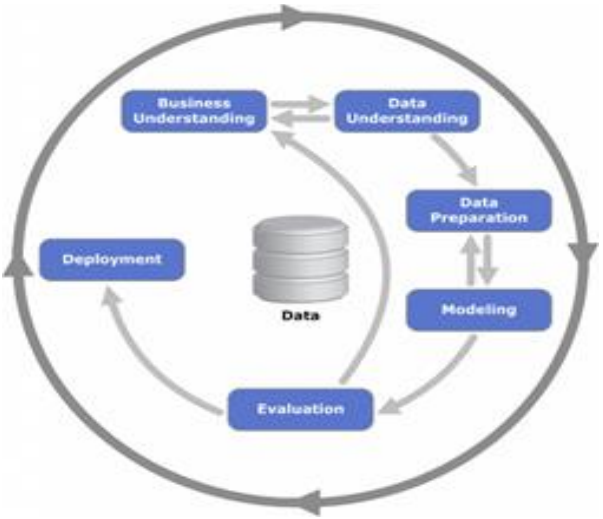
2.1 Área de conhecimento

A quantidade de dados que são gerados e armazenados cresce exponencialmente a cada ano. Dessa forma, é fundamental utilizar essas informações e verificar se existe alguma informação ou padrão dentro delas. Mineração de dados, ou *Data Mining* em inglês, refere-se a extração de informação implícita, previamente desconhecida e potencialmente útil em grandes conjuntos de dados.

O processo de mineração de dados comumente utilizado é chamado de CRISP-DM.

2.2 Mineração de Dados

A técnica de trabalho CRISP-DM é dividida em 6 principais etapas:



2.2.1 Entendimento do Negócio

Esta é a etapa de compreender de forma adequada o problema que necessita ser resolvido. É preciso buscar detalhes sobre como a questão afeta a organização e quais são os principais objetivos e expectativas em relação ao trabalho como um todo.

2.2.2 Compreensão dos dados

Após a primeira etapa, o objetivo torna-se inspecionar, organizar e descrever todos os dados disponíveis. É fundamental a avaliação em busca de quais dados podem ser relevantes para decifrar o problema.

2.2.3 Preparação dos dados

Definidos, organizados e bem inspecionados, nesta etapa é preciso preparar todas as databases, definir o formato que será necessário para a análise e ajustar demais questões técnicas.

2.2.4 Modelagem

Neste quarto momento, são selecionadas e aplicadas as técnicas de mineração de dados mais apropriadas, dependendo dos objetivos identificados na primeira fase.

2.2.5 Avaliação

Considerada uma etapa de *after-work*, mas ainda assim extremamente importante para a vitalidade do ciclo, a quinta fase pede o acompanhamento dos resultados objetivos e a avaliação da aplicabilidade confiável dos insights e conhecimentos obtidos.

2.2.6 Desenvolvimento

Todo o conhecimento que for obtido por meio do trabalho de mineração e modelagem agora poderá ser aplicado de forma prática.

2.3 Agrupamento (Clusterização)

Agrupamento (ou clusterização) é a tarefa de agrupar um conjunto de objetos de forma que os objetos do mesmo grupo (chamados de cluster) sejam mais semelhantes (em algum sentido) uns aos outros do que àqueles de outros grupos (clusters). É uma tarefa principal de mineração exploratória de dados e uma técnica comum para análise de dados estatísticos, usada em muitos campos, incluindo aprendizado de máquina, reconhecimento de padrões, análise de imagens, recuperação de informações, bioinformática, compactação de dados e computação gráfica.

Em Clusterização, não existe o conceito de um algoritmo "correto", mas sim o algoritmo mais apropriado para um problema em particular. Normalmente, a escolha do algoritmo precisa ser feita experimentalmente, a menos que haja uma razão matemática para preferir um modelo de cluster a outro. A visão geral a seguir listará apenas os exemplos mais proeminentes de algoritmos de agrupamento.

2.3.1 *K-Means*

O algoritmo *K-Means* **[2]** agrupa dados tentando separar amostras em *n* grupos de igual variância, minimizando um critério conhecido como inércia ou soma de quadrados intra-*cluster*. Esse algoritmo requer que o número de *clusters* seja especificado. Ele se adapta bem a um grande número de amostras e foi usado em uma grande variedade de áreas de aplicação em muitos campos diferentes.

O algoritmo *K-Means* divide um conjunto de **N** amostras **X** em **K** clusters disjuntos **C**, cada um descrito pela média μj das amostras no cluster. As médias são comumente chamadas de "centróides" do *cluster*; note que eles não são, em geral, pontos de **X**, apesar de estarem no mesmo espaço. O algoritmo *K-Means* visa escolher centróides que minimizem a inércia, ou a soma dos quadrados *intra-cluster*. A inércia pode ser considerada como uma medida de quão internamente coerentes são os clusters.

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_j - \mu_i||^2) \quad \textbf{(1)}$$

http://dx.doi.org/10.25286/repa.v3i3.966

2.3.2 Fuzzy C-Means

O Fuzzy C-Means (FCM) **[3]** é uma técnica de agrupamento em que cada objeto pode pertencer a mais de um *cluster*.

No algoritmos não-fuzzy, os dados são divididos em *clusters* distintos, em que cada objeto pode pertencer apenas a um *cluster*. No FCM, os objetos podem pertencer a vários *clusters*.

Os graus de pertinência são atribuídos a cada um dos objetos. A pertinência indica o grau em que os objetos pertencem a cada *cluster*. Assim, os pontos na borda de um *cluster*, com graus de pertinência inferiores, podem estar no *cluster* em um grau menor que os pontos no centro do *cluster*.

O FCM visa minimizar a seguinte função objetivo:

$$\arg\min_{C} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \|\mathbf{x}_i - \mathbf{c}_j\|^2, \tag{2}$$

onde:

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_j\|}{\|\mathbf{x}_i - \mathbf{c}_k\|} \right)^{\frac{2}{m-1}}}. \tag{3}$$

2.3.3 *Affinity Propagation*

O algoritmo *Affinity Propagation* **[4]** cria *clusters* enviando mensagens entre pares de amostras até a convergência. Um conjunto de dados é então descrito usando um pequeno número de exemplares, que são identificados como os mais representativos de outras amostras. As mensagens enviadas entre pares representam a adequação para uma amostra ser o exemplar da outra, que é atualizada em resposta aos valores de outros pares. Essa atualização ocorre iterativamente até a convergência, momento em que os exemplares finais são escolhidos e, portanto, o agrupamento final é dado.

As mensagens enviadas entre pontos pertencem a uma das duas categorias. A primeira é a responsabilidade $r(i,k)$, que é a evidência acumulada de que a amostra *'k'* deve ser o exemplar da amostra *'i'*. A segunda é a disponibilidade $a(i,k)$, que é a evidência acumulada de que a amostra *'i'* deve escolher a amostra *'k'* para ser seu exemplar, e considera os valores para todas as outras amostras que *'k'* devem ser exemplares. Deste modo, os exemplares são escolhidos por amostras se forem semelhantes o suficiente para muitas amostras e escolhidas por muitas amostras para serem representativas de si mesmas.

Mais formalmente, a responsabilidade de uma amostra *'k'* ser o exemplar da amostra *'i'* é dada por:

$$r(i,k) \leftarrow s(i,k) - max[a(i,k') + s(i,k') \forall k' \neq k] \tag{4}$$

Onde $s(i,k)$ é a semelhança entre amostras *'i'* e *'k'*. A disponibilidade de amostra *'k'* para ser o exemplar da amostra *'i'* é dada por:

$$a(i,k) \leftarrow min[0, r(k,k) + \sum_{i' \ s.t. \ i' \notin \{i,k\}} r(i',k)] \tag{5}$$

2.3.4 *Mean Shift*

O algoritmo *Mean Shift* **[5]** visa descobrir 'bolhas' em uma densidade suave de amostras. É um algoritmo baseado em centróide, que funciona através da atualização de candidatos para centróides para ser a média dos pontos dentro de uma determinada região. Esses candidatos são então filtrados em um estágio de pós-processamento para eliminar quasi-duplicatas para formar o conjunto final de centróides.

Dado um centróide candidato, a atualização do centróide é realizada de acordo com a seguinte equação:

$$x_i^{t+1} = x_i^t + m(x_i^t) \tag{6}$$

Onde:

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)} \tag{7}$$

$N(x_i)$ é a vizinhança de amostras dentro de uma determinada distância ao redor de x_i e m é o vetor de deslocamento médio que é calculado para cada centróide e que aponta para uma região do aumento máximo na densidade de pontos.

O algoritmo Mean Shift define automaticamente o número de clusters, em vez de depender de um parâmetro.

2.3.5 *Spectral Clustering*

O algoritmo *Spectral Clustering* **[6]** faz uma incorporação de baixa dimensão da matriz de afinidade entre as amostras, seguida por um K-Means no espaço de baixa dimensão. O *Spectral Clustering* requer o número de clusters a serem especificados. Ele funciona bem para um pequeno número de clusters, mas não é aconselhável ao usar muitos clusters.

Para dois clusters, ele resolve um relaxamento convexo do problema de cortes normalizados no grafo de similaridade: cortar o gráfico em dois, de modo que o peso do corte das bordas seja pequeno em comparação com os pesos das bordas dentro de cada cluster.

2.3.6 *Agglomerative Clustering*

O algoritmo *Agglomerative Clustering* **[7]** executa um agrupamento hierárquico usando uma abordagem de baixo para cima: cada objeto inicia em seu próprio cluster e os clusters são mesclados sucessivamente. Os critérios de ligação determinam a métrica usada para a estratégia de mesclagem:

- 'Ward' minimiza a soma das diferenças quadradas dentro de todos os clusters. É uma abordagem de minimização de variações e, nesse sentido, é semelhante à função objetivo do K-Means, mas com uma abordagem hierárquica aglomerativa;
- A ligação máxima ou completa minimiza a distância máxima entre observações de pares de clusters;
- A ligação média minimiza a média das distâncias entre todas as observações de pares de clusters.

2.3.7 *Birch*

O algoritmo *Birch* **[8]** constrói uma árvore chamada *Characteristic Feature Tree* (CFT) para os dados fornecidos. Os dados são essencialmente compactados para um conjunto de nós *Characteristic Feature* (CF Nodes). Os nós CF têm um número de *subclusters* chamados *subclusters Characteristic Feature* (CF Subclusters) e estes *subclusters* CF localizados nos nós CF-não-terminais podem ter CF Nodes como filhos.

Os *subclusters* CF armazenam as informações necessárias para o agrupamento em cluster, o que elimina a necessidade de manter todos os dados de entrada na memória.

O algoritmo *Birch* possui dois parâmetros, o limiar e o fator de ramificação. O fator de ramificação limita o número de *subclusters* em um nó e o limiar limita a distância entre a amostra inserida e os *subclusters* existentes.

2.3.8 *Particle Swarm Optimization for Clustering* (PSOC)

A otimização de enxame de partículas (PSO) é um processo de busca estocástica, modelado a partir do comportamento social de um bando de aves **[9]**. O algoritmo mantém uma população de partículas, onde cada partícula representa uma potencial solução para um problema de otimização. No contexto do PSO, um enxame refere-se a um número de soluções potenciais para o problema de otimização, onde cada solução potencial é referida como uma partícula. O objetivo do PSO é encontrar a posição das partículas que resulta na melhor avaliação de uma determinada função de aptidão (objetivo).

No contexto do algoritmo PSOC **[10]**, uma única partícula representa os vetores dos centróides dos *clusters*. Ou seja, cada partícula é construída do seguinte modo:

$$\mathbf{x}_i = \left(\mathbf{m}_{i1}, \cdots, \mathbf{m}_{ij}, \cdots, \mathbf{m}_{iN_c}\right) \quad \textbf{(8)}$$

onde mij se refere ao vetor centróide do j-ésimo cluster da i-ésima partícula no cluster Cij. Portanto, um enxame representa um número de agrupamentos candidatos para os vetores de dados atuais. A aptidão das partículas é medida como o erro quantização:

$$J_e = \frac{\sum_{j=1}^{N_c} [\sum_{\forall \mathbf{z}_p \in C_{ij}} d(\mathbf{z}_p, \mathbf{m}_j)/|C_{ij}|]}{N_c}$$

http://dx.doi.org/10.25286/repa.v3i3.966

(9)

onde d é a distância euclidiana e |Cij| é o número de objetos pertencentes ao cluster Cij, ou seja, a freqüência desse cluster.

2.3.9 *Particle Swarm Clustering* (PSC)

O algoritmo *Particle Swarm* for *Clustering* (PSC) **[11]** apresenta uma codificação diferente das partículas em relação ao PSOC para o problema de agrupamento. Nessa abordagem, cada partícula pi representa um único centróideck, que navega pelo espaço de buscas até encontrar a posição ótima correspondente aos centróides das regiões de alta densidade na base de dados. Como cada partícula é um único centróide, a solução para o problema, neste caso, é todo o enxame.

2.3.10 Métricas para Avaliação de *Clusters*

A avaliação de desempenho de um algoritmo de clusterização não é trivial quanto contar o número de erros ou a precisão e a recuperação de um algoritmo de classificação supervisionado. Em particular, qualquer métrica de avaliação não deve levar em consideração os valores absolutos dos rótulos dos clusters, mas sim se esse agrupamento define uma separação dos dados semelhante a algum conjunto verdadeiro preestabelecido de classes ou então se satisfaz alguma suposição de que os membros que pertencem à mesma classe são mais semelhantes que membros de classes diferentes de acordo com alguma métrica de similaridade.

2.3.10.1 Coeficiente de Silhueta

O Coeficiente de Silhueta **[12]** é calculado usando a distância intra-cluster média (a) e a distância média do *cluster* mais próximo (b) para cada amostra. O coeficiente de silhueta para uma amostra é (b - a) / max (a, b). Para esclarecer, b é a distância entre uma amostra e o *cluster* mais próximo do qual a amostra não faz parte.
O melhor valor do Coeficiente de Silhueta é 1 e o pior valor é -1. Valores próximos a 0 indicam *clusters* sobrepostos. Valores negativos geralmente indicam que uma amostra foi atribuída ao *cluster* errado, pois um *cluster* diferente é mais semelhante.

2.3.10.2 Índice de Calinski-Harabasz

Se os rótulos para as classes dos dados não forem conhecidos a priori, o índice de Calinski-Harabasz **[13]** pode ser usado para avaliar o modelo, onde uma pontuação mais elevada de Calinski-Harabasz se relaciona a um modelo com clusters melhor definidos.

Para clusters, o índice de Calinski-Harabasz é dado como a razão entre a média da dispersão entre os clusters e a dispersão dentro do cluster:

$$s(k) = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1}$$

(10)

onde B_K é a matriz de dispersão entre grupos e W_K é a matriz de dispersão dentro do cluster definida por:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$
$$B_k = \sum_{q} n_q (c_q - c)(c_q - c)^T$$

(11)

com Nigual ao número de objetos, C_q o conjunto de objetos no cluster q, c_q o centro do cluster q, c o centro de E, n_q o número de objetos no cluster q.

3 Materiais e Métodos

3.1 Descrição da base de dados

Os dados utilizados neste trabalho são provenientes de uma base pública fornecida pelo projeto 500 Cities. O projeto 500 Cities reporta dados de cidades e dados censitários, obtidos usando métodos de estimação em pequenas áreas, para 27 medidas de doenças crônicas nas 500 maiores cidades americanas. O total de 27 medidas inclui 5 comportamentos não-saudáveis, 13 indicadores de saúde e 9 práticas de prevenção.

Indicadores de Saúde	Práticas de Prevenção	Comportamentos Não-Saudáveis
Artrite	Triagem de colesterol	Beber compulsivamente
Câncer (excluindo câncer de pele)	Falta de seguro de saúde entre adultos de 18 a 64 anos	Tabagismo
Doença renal crônica	Tomar remédio para o controle da pressão alta entre aqueles com pressão alta	Nenhuma atividade física de lazer
Doença de obstrução pulmonar crônica	Visitas ao médico para check-up de rotina no último ano	Dormir menos de 7 horas
Doença cardíaca coronariana	Visitas ao dentista ou clínica odontológica	Obesidade
Asma	Uso de mamografia entre mulheres de 50 a 74 anos	

Indicadores de Saúde	Práticas de Prevenção	Comportamentos Não-Saudáveis
Diabetes	Papanicolaou entre mulheres adultas com idade entre 21-65 anos	
Pressão alta	Exame de sangue oculto nas fezes, sigmoidoscopia ou colonoscopia entre adultos de 50 a 75 anos	
Colesterol alto entre adultos com idade ≥ 18 anos que foram rastreados nos últimos 5 anos	Idosos ≥ 65 anos que estejam atualizados em um conjunto de serviços clínicos preventivos (Homens: vacina contra Polissacarídeos Pneumocócicos - PPV, Rastreio do câncer colorretal; Mulheres: Igual ao anterior e Mamografia nos últimos 2 anosanos)	
Saúde mental instável por ≥14 dias		
Saúde física instável por ≥14 dias		
Acidente vascular encefálico		
Todos os dentes perdidos entre adultos com idade ≥65 anos		

O número de cidades por estado varia de 1 a 121. E as cidades variam de 42.417 pessoas em Burlington (Vermont) a 8.175.133 em Nova York (Nova York). Entre a 500 cidades, existem aproximadamente 28.000 setores censitários, para os quais foram fornecidos dados. Os intervalos para os setores variam em população de menos de 50 a 28.960, e em tamanho de menos de 1 milha quadrada a mais de 642 milhas quadradas. O número de setores por cidade varia de 8 a 2.140.

O projeto 500 Cities inclui uma população total de 103.020.808, o que representa 33,4% da população total dos Estados Unidos de 308.745.538.

A base de dados tem um total de 117 dimensões para um total de 500 entradas (cidades).

Existem 3 categorias principais:

- Indicadores de Saúde;
- Prevenção;
- Comportamentos não saudáveis.

http://dx.doi.org/10.25286/repa.v3i3.966

Cada uma dessas categorias é dividida em subcategorias para as quais são fornecidas 4 medições:

- Dados brutos;
- Dados ajustados por idade;
- Nível de confiança de 95% bruto;
- Nível de confiança de 95% ajustado por idade.

A maioria das porcentagens é definida como a proporção de entrevistados com idade superior ou igual a 18 anos que responderam "sim" a uma determinada pergunta de um questionário, sobre entrevistados com idade superior ou igual a 18 anos que relataram "sim" ou "não" (excluindo aqueles que se recusaram a responder, ou que responderam "não sei / não tenho certeza").

Neste trabalho, serão considerados apenas os dados ajustados por idade.

3.1.1 Dicionário de Dados

Como especificado no tópico anterior (3.1), a base possui 117 atributos, dos quais 112 estão relacionados às categorias e suas subcategorias. Para cada atributo são fornecidas 4 medições, que apresentam a seguinte configuração:

- Dados brutos: seu tipo é Numeric, com tamanho máximo igual a 4. Aceita o seguinte formato de valores: 00.0 ;
- Dados ajustados por idade: segue o mesmo padrão de Dados brutos;
- Nível de confiança de 95% bruto: seu tipo é String, com tamanho máximo igual a 12. Aceita o seguinte formato de valores: "(ponto mais inferior, ponto mais superior)";
- Nível de confiança de 95% ajustado por idade: segue o mesmo padrão do nível de confiança bruto

Os demais 5 atributos estão relacionados a:

- Estado: do tipo String, relacionado à abreviação do Estado, com apenas 2 caracteres;
- Nome da cidade: do tipo String do lugar ou subdivisão do município;
- Código FIPS: do tipo Numeric, com no máximo 7 caracteres. Relacionado Código FIPS do estado + código FIPS de subdivisão do local ou do condado;
- População: do tipo Numeric, com no máximo 8 caracteres. Dados retirados do Censo 2010 dos EUA;
- Geolocalização: do tipo String, informando latitude e longitude do local.

O Dicionário de Dados completo pode ser visualizado em Anexos.

3.1.2 Distribuição de Frequência

Abaixo seguem as tabela que contém um resumo dos dados obtido em uma amostra. A distribuição é organizada em formato de tabela, e cada entrada da tabela contém a frequência dos dados em um determinado intervalo, ou em um grupo.

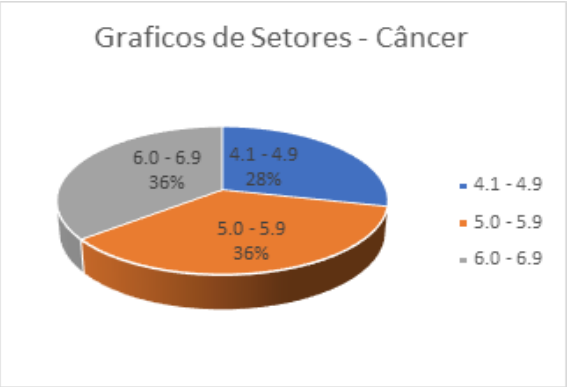
Câncer (excluindo câncer de pele) entre adultos com idade> = 18 anos - 2015					
Classe	Limite Inferior	Ponto Médio	Limite Superior	Frequência Absoluta	Amplitude de classe
1	4,1	4,5	4,9	8	0,8
2	5	5,45	5,9	10	0,9
3	6	6,45	6,9	10	0,9

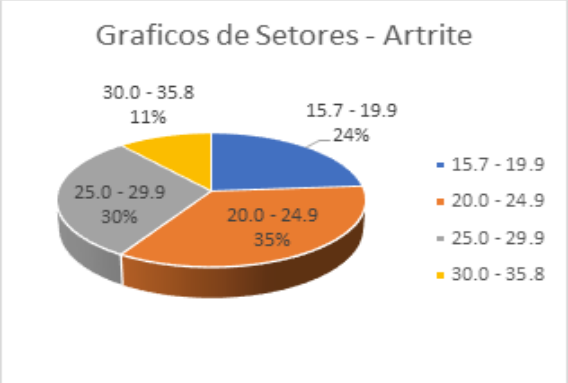
Artrite entre adultos com idade> = 18 anos - 2015					
Classe	Limite Inferior	Ponto Médio	Limite Superior	Frequência Absoluta	Amplitude de classe
1	15,7	17,8	19,9	34	4,2
2	20	22,45	24,9	50	4,9
3	25	27,45	29,9	43	4,9
4	30	32,9	35,8	16	5,8

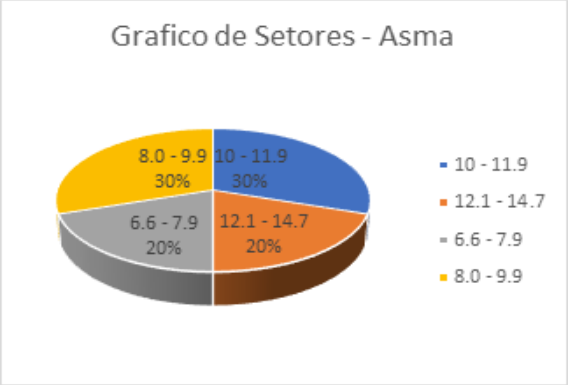
Asma atual entre adultos com idade> = 18 anos - 2015					
Classe	Limite Inferior	Ponto Médio	Limite Superior	Frequência Absoluta	Amplitude de classe
1	6,6	7,25	7,9	13	1,3
2	8	8,95	9,9	20	1,9
3	10	10,95	11,9	20	1,9
4	12,1	13,4	14,7	13	2,6

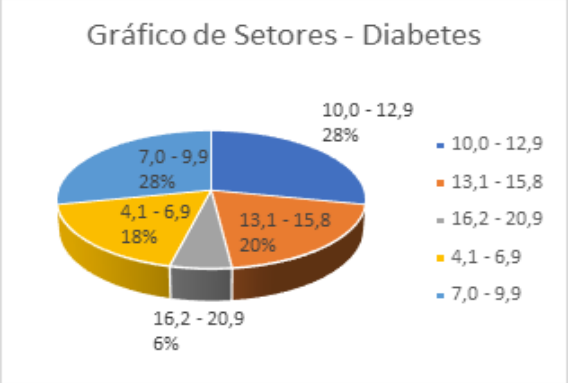
Diabetes diagnosticado entre adultos com idade> = 18 anos - 2015					
Classe	Limite Inferior	Ponto Médio	Limite Superior	Frequência Absoluta	Amplitude de classe
1	4,1	5,5	6,9	19	2,8
2	7	8,45	9,9	30	2,9
3	10	11,45	12,9	30	2,9
4	13,1	14,45	15,8	21	2,7
5	16,2	18,55	20,9	6	4,7

3.1.3 Visualização dos dados









3.1.4 Medidas de Resumo

Porcentagem - Diabetes	
Moda	8,9
Mediana	10,3
Média	10,5

Porcentagem - Artrite	
Moda	22,4
Mediana	23,65
Média	23,9

Porcentagem - Câncer	
Moda	6
Mediana	5.6
Média	5,55

Porcentagem - Asma	
Moda	9
Mediana	14,65
Média	10,1

3.2 Parametrização das técnicas

1. K-Means:
a. K variando de 2 a 13
b. Inicialização aleatória dos centróides
2. *Fuzzy C-Means*:
a. K variando de 2 a 13
b. *Fuzzificador* m igual a 2
3. Affinity Propagation
a. *Preference* igual a -50
4. Agglomerative Clustering
a. K variando de 2 a 13
b. Tipo de ligação: média
5. Birch
a. K variando de 2 a 13
6. Mean Shift
a. *Bandwidth* igual a 0,81
7. Spectral Clustering
a. K variando de 2 a 13
b. Eigen solver: arpack
c. Afinidade: *nearest_neighbors*
8. PSC
a. Quantidade de partículas (K) variando de 2 a 13
b. Número de iterações igual a 1000
c. Inércia igual a 0,95
d. c1=c2=2,05
e. c3=c4=1,0
f. Velocidade máxima das partículas: 0,001
9. PSOC
a. K variando de 2 a 13
b. Quantidade de partículas igual a 30
c. Número de iterações igual a 1000
d. Inércia igual a 0,72,
e. c1=c2=1,49

4 Experimentos realizados

Cada algoritmo de clusterização foi executado 30 vezes para cada conjunto de parâmetros para observação das métricas de Silhueta e Calinski-Harabasz. As duas métricas foram utilizadas para a escolha do melhor número de clusters para o problema. A Tabela 1 mostra as estatísticas para os algoritmos com o melhor número de clusters baseado nas duas métricas.

Tabela 1 - Estatísticas das métricas

Algoritmo	Silhueta	Calinski-

http://dx.doi.org/10.25286/repa.v3i3.966

(n_clusters)	(desvio-padrão)	Harabasz(desvio-padrão)
Fuzzy C-Means **(2)**	0.27399 **(4.53246e-17)**	247.8647 **(3.476e-17)**
PSOC **(2)**	0.27398 **(7.3687e-18)**	247.845 **(0.007212)**
PSC **(2)**	0.27228 **(0.00046)**	229.572 **(14.57811)**
K-Means **(2)**	0.27106 **(2.67750e-05)**	247.85433 **(0.00572)**
Spectral **(2)**	0.25755 **(5.55111e-17)**	231.66074 **(8.52651e-14)**
Birch **(2)**	0.25136 **(0.0)**	196.50644 **(0.0)**
Affinity **(3)**	0.16760 **(5.55111e-17)**	163.81091 **(2.84217e-14)**
Agglomerative **(3)**	0.27787 **(0.0)**	43.16529 **(0.0)**
Mean Shift **(5)**	0.14607 **(2.77555e-17)**	45.52593 **(1.42108e-14)**

Os algoritmos com os melhores valores para as métricas distribuíram os dados em dois clusters. O *Affinity Propagation* e o *Agglomerative Clustering* indicaram a divisão em três clusters, porém com valores menores para as métricas. O *Agglomerative Clustering* teve um resultado bom comparando a Silhueta, mas um resultado ruim quando analisando o índice de *Calinski-Harabasz*. O *Mean Shift* por sua vez em seu melhor resultado agrupou os dados em 5 clusters e teve o pior resultado dentre os algoritmos testados nessa base de dados, considerando as duas métricas utilizadas nesse trabalho.
A visualização geográfica dos clusters para os resultados do K-Means e do AffinityPropagation pode ser observada nas figuras a seguir.

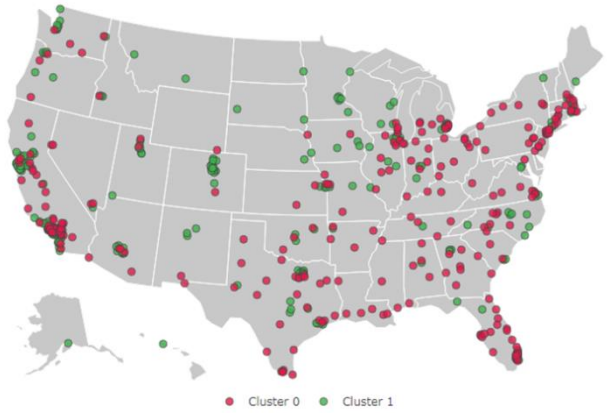


Figura - K-Means

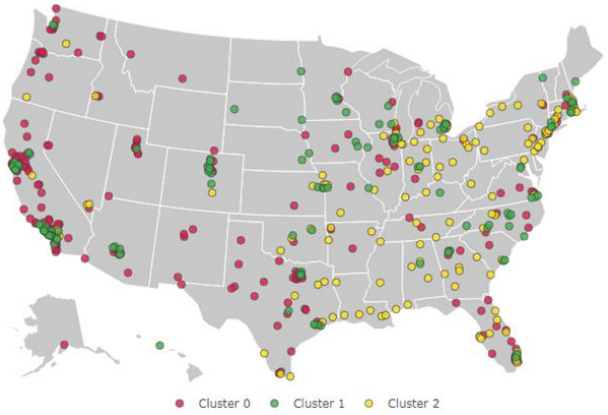


Figura – *Affinity Propagation*

Para analisar mais profundamente o que cada cluster significa a matriz de correlação para cada cluster pode ser útil para inferir as relações entre as diferentes dimensões do problema. Por exemplo, para o K-Means, temos as seguintes matrizes de correlação:

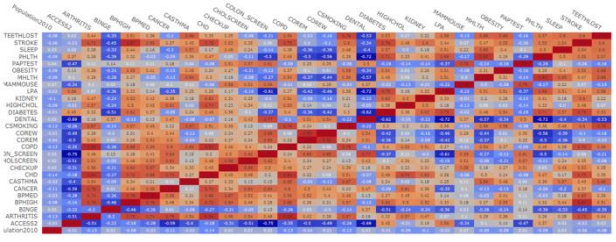


Figura - Matriz de correlação Cluster 0 (K-Means)

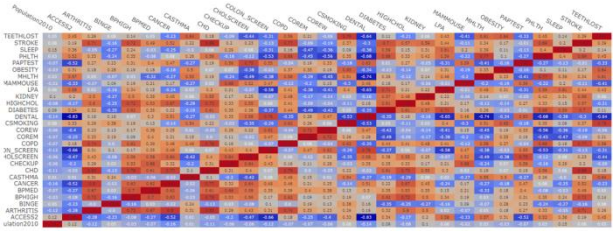


Figura - Matriz de correlação Cluster 1 (K-Means)

Juntamente com as matrizes de correlação, os centróides de cada cluster indicam as principais características dos clusters.

O K-Means, por exemplo, agrupou os clusters em cidades saudáveis (Cluster 1) e não-saudáveis (Cluster 0). As principais variáveis para considerar uma cidade saudável foram:

- Menores índices de doenças (artrite, asma, doenças cardíacas, doenças pulmonares, diabetes, colesterol alto e doenças renais);
- Menores índices de maus hábitos (alcoolismo, tabagismo, sedentarismo);
- Melhores índices de prevenção (controle de pressão alta, colonoscopia, exames de rotina para homens e mulheres, visitas ao dentista).

Dentre as 10 cidades mais populosas presentes na base de dados, temos o seguinte agrupamento:

- Cidades saudáveis: San Diego, Honolulu;
- Cidades não-saudáveis: Nova York, Los Angeles, Chicago, Houston, Philadelphia, Phoenix, San Antonio e Dallas.

As principais características das cidades de San Diego e Honolulu em relação às outras é o alto índice de acesso a planos de saúde, visitas periódicas ao dentista e a prática de atividades físicas de lazer regulares.

Por sua vez, o *Affinity Propagation* agrupou os dados em três clusters: cidades saudáveis (Cluster 1), cidades não-saudáveis (Cluster 0) e cidades críticas (Cluster 2). As diferenças principais entre as cidades não-saudáveis e as cidades críticas são que as cidades não-saudáveis tem menores índices de prevenção, porém as cidades críticas tem índices elevados na maioria das doenças (exceto câncer) e em todos os maus hábitos (alcoolismo, tabagismo, sedentarismo, obesidade e poucas horas diárias de sono).

Dentre as 10 cidades mais populosas presentes na base de dados, temos o seguinte agrupamento:

- Cidades saudáveis: Honolulu
- Cidades não-saudáveis: Nova York, Los Angeles, Chicago, Houston, Phoenix, San Antonio, Dallas e San Diego
- Cidades críticas: Philadelphia

Em relação ao resultado do K-Means, a cidade de Philadelphia passou a ser classificada como crítica e a cidade de San Diego passou a ser considerada como não saudável. Entretanto, ao realizar uma análise de similaridade entre as cidades de Honolulu e San Diego, as duas são mais similares do que San Diego e Nova York, por exemplo. Isso pode indicar que San Diego foi erroneamente agrupada pelo AffinityPropagation. Tendo em vista que o AffinityPropagation teve resultados de métrica piores que o K-Means, este pode ser um bom indício.

5 Conclusões e trabalhos futuros

A saúde pública é uma área com problemas bastante complexos e que os órgãos governamentais frequentemente enfrentam desafios para entender como oferecer melhores serviços de saúde e prevenir epidemias futuras. Os métodos preventivos normalmente são a melhor para controlar ou extinguir doenças, tornado importante saber quais os hábitos de saúde das populações das cidades e quais impactos tais hábitos poderiam ter sobre as condições gerais de saúde das populações.

O projeto 500 Cities reporta dados epidemiológicos das 500 maiores cidades americanas e baseia-se em doenças crônicas prioritárias e de maior impacto na saúde pública. As medidas reportadas incluem os índices de doenças, comportamentos de maior risco que causam doenças e práticas de prevenção.

Este trabalho procurou identificar características relevantes para dar suporte na prevenção de epidemias e doenças e agrupar as diferentes cidades de acordo com seus hábitos e condições de saúde. Essa detecção de comunidades foi feita utilizando vários algoritmos de clusterização para identificar os principais agrupamentos e suas principais características.

A maioria dos algoritmos agrupou as cidades em dois clusters, sendo um com bons indicadores

http://dx.doi.org/10.25286/repa.v3i3.966

de saúde e outro com indicando cidades não-saudáveis. Outros algoritmos subdividiram o grupo de cidades não-saudáveis para incluir o grupo de cidades críticas (com péssimo indicadores). Porém, essa divisão em mais agrupamentos levou a uma diminuição da avaliação dos clusters, o que pode indicar um aumento de cidades sendo erroneamente agrupadas.

Trabalhos futuros que possam agregar informações relevantes ao presente trabalho incluem:

- Estudar como utilizar efetivamente a informação de geolocalização no processo de clusterização. Métricas euclidianas tendem a não ser efetivas com dados de latitude e longitude;
- Analisar o comportamento de outras métricas de avaliação de agrupamentos (principalmente a estatística Gap, mas possivelmente outras como índice de Xu, índice de Hartigan);
- Utilizar outros algoritmos de agrupamento que utilizem outras abordagens para cálculo de similaridade, como os algoritmos que utilizam conceitos de Teoria da Informação e que não dependam da distância euclidiana.

Referências

[1] CENTERS FOR DISEASE CONTROL AND PREVENTION. **500 Cities**: local data for better health. National Center for Chronic Disease Prevention and Health, Promotion, Division of Population Health, 2016. Disponível em: <https://www.cdc.gov/500cities> Acesso em: 23 out. 2017.

[2] MACQUEEN, James B. et al. Some Methods for classification and Analysis of Multivariate Observations. In: Berkeley Symposium on Mathematical Statistics and Probability, 5., 1967, California. **Proceedings**…Berkeley, CA: University of California, Press. p. 281–297.

[3] BEZDEK, James C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. New York, London: Plenum Press, 1981.

[4] FREY, Brendan J.; DUECK, Delbert. Clustering by passing messages between data points, **Science**, v. 315, n. 5814, p. 972-976, 2007.

[5]COMANICIU, Dorin; MEAN, Peter. Shift: A robust approach toward feature space analysis. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n.5, p. 603-619, 2002.

[6] NG, Andrew Y.; JORDAN, Michael I.; WEISS, Yair. On spectral clustering: Analysis and an algorithm. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 15., 2002. **Proceedings**... Press,2002. p. 849-856.

[7] ZHANG, et al. Graph degree linkage: Agglomerative clustering on a directed graph. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 12., 2012, Florence. **Proceedings**… Florence, IT, 2012.

[8] ZHANG, Tian; RAMAKRISHNAN, Raghu; LIVNY, Miron. BIRCH: an efficient data clustering method for very large databases. In: **ACM Sigmod Record**. ACM, 1996. p. 103-114.

[9] KENNEDY, James. Particle swarm optimization. In: Encyclopedia of machine learning, p. 760-766. Springer, 2011.

[10] MERWE, D. W. van der; ENGELBRCHT, A. P. Data clustering using particle swarm optimization. In: Congress on Evolutionary Computation, 2003. **Proceedings**… 2003.

[11] COHEN, Sandra C. M.; CASTRO Leandro N. de. Data Clustering with Particle Swarms. In: IEEE Congress on Evolutionary Computations. 2006.

[12] ROUSSEEUW Peter J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. **Computational and Applied Mathematics**, v. 20, p. 53-65, 1987.

[13] CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, v. 3, p. 1-27, 1974.

http://dx.doi.org/10.25286/repa.v3i3.966

Análise de Crédito Utilizando uma Abordagem de Mineração de Dados

Credit Analysis Utilizing a Data Mining Approach

Joyce Maria do Carmo de Sá[1] orcid.org/0000-0001-8224-1323
Iago Richard Rodrigues Silva[1] orcid.org/0000-0002-8242-9059
Raniel Gomes da Silva[1] orcid.org/0000-0003-4874-3447
Luís Gustavo Arcoverde Souto[1] orcid.org/0000-0002-0410-0151
Paloma Gabriela Santos Silva[1] orcid.org/0000-0003-1477-9986

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: Joyce de Sá jmcs@ecomp.poli.br

Resumo

O crédito é um instrumento aplicado para incrementar e facilitar a realização de vendas de bens e serviços. Ele é o responsável por grande parte dos resultados auferidos nas empresas e pelo desenvolvimento e crescimento da economia do país. No entanto, faz-se necessário uma rígida avaliação para onde este crédito deve ir, uma vez que, sendo aplicado para empresas ou pessoas erradas, o credenciado pode acumular prejuízos. Desta forma, este trabalho propõe uma abordagem utilizando Mineração de Dados para análise de crédito através da aplicação de algoritmos de Inteligência Computacional, proporcionando uma tomada de decisão mais assertiva no momento de concessão do crédito.

Palavras-Chave: Análise de Crédito; Mineração de Dados; Inteligência Computacional;

Abstract

Creditis an instrument used to increase and facilitate sales of goods and services. He is responsible for a great part of the results obtained in the companies and for the development and growth of the economy of the country. However, a rigid assessment is necessary to where this credit should go, since, being applied to companies or wrong people, the credentialed can accumulate losses. In this way, this work proposes an approach using Data Mining for credit analysis through the application of Computational Intelligence algorithms, providing a more assertive decision making at the moment of credit granting.

Key-words: *Credit Analysis; Data Mining; Computational Intelligence;*

 http://dx.doi.org/10.25286/repa.v3i3.967

1 Introdução

Segundo Ross, Westerfield e Jordan (2002), a concessão de crédito é motivada pela necessidade de estimular vendas, mas isso acarreta para empresa concessora custos de imobilização do capital, bem como o risco do cliente não pagar, por isso é necessário definir como conceder e como cobrar, ou seja, uma política de crédito **[1]**. Entretanto, é necessário saber que política de crédito possui melhores resultados. Para isso, é essencial uma análise minuciosa das possíveis variáveis que venham a influenciar o bom do ruim credenciado.

Recentemente as necessidades dos clientes e a economia nacional têm sofrido diversas alterações. O processo de mudança de atitude, no que tange o crédito para pessoas físicas, dos tempos da inflação elevada para o momento de estabilidade, gerou uma desorientação para as pessoas e instituições financeiras, acarretando um aumento considerável na inadimplência **[2]**. Tal fato direcionou a necessidade de se ter, a cada dia, critérios mais precisos para a análise e concessão de crédito. Com os avanços tecnológicos foi possível verificar de diversas maneiras a análise do crédito. De acordo com Schrickel **[3]**, a análise de crédito envolve a habilidade de fazer uma decisão de crédito dentro de um cenário de incertezas e constantes mutações e transformações incompletas. Esta habilidade depende da capacidade de analisar logicamente situações complexas, e chegar a uma conclusão clara, prática e factível, de ser implementada.

Com a finalidade de obter uma análise mais rigorosa a pesquisa será guiada por metodologias de mineração de dados (*Data Mining*), como o CRISP-DM (*Cross Industry Standard Process for Data Mining*), o qual defende um modelo de processo que fornece uma estrutura para realização de projetos de mineração de dados que são independentes da indústria e da tecnologia utilizada, focando o descobrimento de padrões e regras significativos **[4]**. Pode-se descrever Mineração de Dados como parte do processo de descoberta de conhecimento em CRISP-DM, que tem por objetivo selecionar técnicas que serão utilizadas para localização de padrões nos dados, gerando por fim uma busca dos referidos padrões relacionados a um dado interesse **[5]**. Suas etapas podem apresentar-se de forma cognitiva, interativa e exploratória, compreendendo nos seguintes passos: entendimento do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e desenvolvimento **[6]**.

O presente trabalho tem por finalidade apresentar uma abordagem de análise de crédito através de Mineração de Dados, utilizando métodos de classificação para aprovação da concessão de crédito. O restante do trabalho está organizado da forma que segue. A seção 2 traz um embasamento teórico sobre análise de crédito e mineração de dados; a seção 3 apresenta e descreve a base, dicionário e alterações de dados juntamente com a parametrização das técnicas, além de descrever com detalhes os experimentos realizados durante a pesquisa; a seção 4 analisa os experimentos e apresenta conclusões.

2 Referencial Teórico

Nesta seção serão apresentados os principais temas que formam a base teórica para realização deste trabalho.

2.1 Análise de Crédito

Crédito é um conceito que está presente no cotidiano das pessoas e empresas, com o passar do tempo é perceptível uma maior necessidade da utilização desse conceito, que para Schrickel **[3]** crédito significa, "todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente, parte do seu patrimônio a um terceiro, com a expectativa de que esta parcela volte a sua posse integralmente, após decorrido o tempo estipulado". Silva **[7]** defende que a função do crédito consiste em avaliar a capacidade de pagamento do tomador, visando assegurar a reputação e a solidez do emprestador.

Com o aumento da procura por crédito, ocasionou também aumentou do índice de inadimplência, tornando-se necessário que as empresas buscassem ferramentas para auxiliar nas decisões de riscos, como a análise de crédito. Para que essas instituições possam mensurar o risco da concessão de crédito, faz-se necessário o uso de inteligência computacional (IC), que prover redução de custos, aumento de

produtividade, precisão e flexibilidade na operacionalização de mudanças na estratégia de concessão de crédito **[8]**, conseguindo análises mais precisas através das suas abordagens e técnicas de mineração de dados que podem extrair informações importantes de um conjunto de dados.

2.2 Mineração de Dados

Com a exacerbada quantidade de dados crescendo diariamente, responder uma questão tornou-se necessário **[9]**: O que fazer com os dados armazenados? As técnicas tradicionais de exploração de dados não são mais adequadas para tratar a grande maioria dos repositórios. Com a finalidade de responder a esta questão, foi proposta, no final da década de 80, a Mineração de Dados, do inglês *Data Mining*.

Para Fayyad et al. **[10]**. A extração de conhecimento de base de dados (mineração de dados) é o processo de identificação de padrões válidos, novos potencialmente úteis e compreensíveis embutidos nos dados. Portanto, mineração de dados nada mais é do que a procura de respostas para perguntas que ainda não existem em um grande volume de dados, extração de conhecimento, sabedoria.

2.2.1 CRISP-DM

Atualmente diversos processos definem e padronizam as fases e atividades da Mineração de Dados. Apesar das particularidades, todos em geral contém a mesma estrutura. Neste trabalho, escolhemos o CRISP-DM (*Cross-Industry Standard Processof Data Mining*) como modelo, devido à vasta literatura disponível e por atualmente ser considerado o padrão de maior aceitação.

O processo CRISP-DM consiste de seis fases organizadas de maneira cíclica, conforme mostra a figura abaixo. Além disto, apesar de ser composto por fases, o fluxo não é unidirecional, podendo ir e voltar entre as fases.

As fases do processo CRISP-DM são:

I. **Entendimento do Negócio:** Nessa etapa, o foco é entender qual o objetivo que se deseja atingir com a mineração de dados. O entendimento do negócio irá ajudar nas próximas etapas.

II. **Entendimento dos Dados:** as fontes fornecedoras dos dados podem vir de diversos locais e possuírem diversos formatos. Após definir os objetivos, é necessário conhecer os dados visando:

   a. Descrever de forma clara o problema;

   b. Identificar os dados relevantes para o problema em questão;

   c. Certificar-se de que as variáveis relevantes para o projeto não são interdependentes

   Normalmente as técnicas de agrupamento e de exploração visual também são utilizadas nesta etapa.

III. **Preparação dos Dados:** devido às diversas origens possíveis, é comum que os dados não estejam preparados para que os métodos de Mineração de Dados sejam aplicados diretamente. Dependendo da qualidade desses dados, algumas ações podem ser necessárias. Este processo de limpeza dos dados geralmente envolve filtrar, combinar e preencher valores vazios.

IV. **Modelagem:** é nesta fase que as técnicas (algoritmos) de mineração serão aplicadas. A escolha da(s) técnica(s) depende dos objetivos desejados.

V. **Avaliação:** considerada uma fase crítica do processo de mineração, nesta etapa é necessária a participação de especialistas nos dados, conhecedores do negócio e tomadores de decisão. Diversas ferramentas gráficas são utilizadas para a visualização e análise dos resultados (modelos). Testes e validações, visando obter a confiabilidade nos modelos, devem ser executados (*crossvalidation*, *suppliedtest set*, use *training set*, *percentage Split*).

VI. **Desenvolvimento:** Após executado o modelo com os dados reais e completos é necessário que os envolvidos conheçam os resultados.

http://dx.doi.org/10.25286/repa.v3i3.967

2.2.2 Classificação

A etapa de classificação pode ajudar no planejamento e na tomada de decisão, consiste em preparar os dados utilizados para treinamento, onde também será analisado o critério de parada que pode influenciar na qualidade final da previsão e testes. A classificação é aplicada na etapa da modelagem do CRISP-DM.

Uma abordagem geral para o aprendizado deste modelo consiste, primeiramente, em fornecer dados de treinamento, cujos resultados são conhecidos. Os dados de treinamento são então usados para gerar o modelo de classificação, que é posteriormente aplicado aos dados de teste, cujos resultados são desconhecidos. O objetivo é criar um modelo capaz de categorizar corretamente tanto os dados utilizados em seu treinamento, como dados nunca vistos antes, ou seja, um modelo com boa capacidade de generalização **[6]**.

As próximas subseções apresentam os algoritmos para classificação dos dados utilizados neste trabalho.

2.2.2.1 *NaiveBayes*

O *NaiveBayes* mostra ser uma ótima alternativa devido à sua utilidade para grandes volumes de dados e rapidez na execução quando comparados com outros algoritmos de classificação.

O *NaiveBayes* é uma técnica de aprendizado probabilístico supervisionado baseado no teorema de *Bayes* com uma suposição de independência entre os preditores. Basicamente, um classificador *NaiveBayes* assume que a presença de uma característica particular em uma classe não está relacionada com a presença de qualquer outro recurso.

2.2.2.2 *K-NearestNeighbors*

Os K vizinhos mais próximos ou (K-NN) é um algoritmo simples e é um classificador onde o aprendizado é baseado na analogia. O conjunto de treinamento é formado por vetores n-dimensionais e cada elemento deste conjunto representa um ponto no espaço n-dimensional. A ideia principal deste algoritmo é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento. Para determinar a classe de um elemento que não pertença ao conjunto de treinamento, o classificador K-NN procura K elementos do conjunto de treinamento que estejam mais próximos deste elemento desconhecido, ou seja, que tenham a menor distância. Estes K elementos são chamados de K-vizinhos mais próximos. Verifica-se quais são as classes desses K vizinhos e a classe mais frequente será atribuída à classe do elemento desconhecido. Dois pontos chaves que devem ser determinados para aplicação do K-NN são: a forma como se calcula a distância e o valor do K.

2.2.2.3 Regressão Logística

A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em detrimento de um conjunto de variáveis categóricas. Busca estimar a probabilidade da variável dependente dela assumir um determinado valor em função dos conhecidos de outras variáveis. Os resultados da análise ficam contidos no intervalo de zero a um.

O modelo de regressão logística (RL) é um modelo linear generalizado, sendo um tipo de análise de regressão muito utilizado para realizar predições ou explicar a ocorrência de um evento específico quando a variável dependente (variável resposta) é de natureza binária. Quanto às variáveis independentes, estas podem ser tanto quantitativas quanto qualitativas.

Por se tratar de um modelo linear generalizado, a RL apresenta três componentes: uma componente aleatória, que consiste em uma combinação das variáveis independentes (preditoras); um componente sistemático, que relaciona as variáveis independentes com os parâmetros do modelo correspondente à variável resposta que se quer modelar; e uma função de ligação. Como a regressão logística funciona com modelos preditivos binários, ela pode ser utilizada em classificação de dados cuja saída sejam binárias.

2.2.2.4 *Deep Learning*

Uma Rede Neural Artificial (RNA) consiste de vários neurônios simples conectados, cada um

produzindo uma sequência de valores de ativação. O aprendizado, ou a tarefa que a RNA se propõe a ser utilizada depende dos pesos encontrados que fazem a rede ter o comportamento desejado. Dependendo do problema e como os neurônios estão conectados tais comportamentos podem precisar de grandes blocos computacionais com vários neurônios interligados, onde cada bloco realiza transformações, normalmente de forma não linear. O *Deep Learning* (DL) se propõe a, de forma precisa, encontrar os pesos para cada bloco ou camada de neurônios **[16]**.

Podem se caracterizar por DL redes neurais que: usam uma cascata de diversas camadas, com unidades de processamento, normalmente, não-linear para a extração e transformação de características; cada camada sucessiva usa a saída da camada anterior como entrada; são baseados na aprendizagem (supervisionada) de vários níveis de características ou representações dos dados; realizam parte de uma área da aprendizagem de máquina mais ampla que é a aprendizagem de representações de dados; aprendem vários níveis de representações que correspondem a diferentes níveis de abstração; os níveis formam uma hierarquia de conceitos **[17]**.

Durante o treinamento da rede neural pode ocorrer dois problemas distintos ligados a fase de treinamento, são eles o *overfitting* e *underfitting*. O *overfitting* é o treinamento excessivo, fazendo com que a rede memorize padrões da base de dados atual e perca sua capacidade de generalização com entradas de dados novas a rede. O *underfitting* é o treinamento insuficiente, fazendo com que a rede não aprenda os padrões e comportamentos e assim não possa generalizar para novos dados de entrada.

Sendo assim é necessário determinar um critério de parada, e um dos mais usados é a validação cruzada. Esta é a verificação da diferença entre a saída encontrada e a saída desejada, onde os pesos são inéditos a cada ciclo de validação. Enquanto o erro de validação estiver diminuindo, a rede continua treinando, isto é, no momento em que o erro da validação começar a aumentar e o de treinamento continuar a diminuir, a rede está começando a memorizar padrões, sendo este o ponto de parada para o treinamento.

2.3 Teste de *Kolmogorov-Smirnov*

O teste estatístico de *Kolmogorov-Smirnov* (KS) foi proposto pelos soviéticos A.N. Kolmogorov e N.V. Smirnov **[11]**, é uma técnica não paramétrica, usada para testar se duas amostras podem ser provenientes de uma mesma função de distribuição **[13]**. A estatística KS é definida como a máxima diferença entre as distribuições acumuladas dos scores dos "bons" e "maus" pagadores **[11]**. Pode ser definida pela equação 1:

$$KS = \max_{s} \left\{ \left| F_M(s) - F_B(s) \right| \right\} \quad (1)$$

O KS mede a máxima separação entre a frequência relativa acumulada de maus pagadores, Fm(s) e a frequência relativa acumulada de bons pagadores, Fb(s). Sob a hipótese que as distribuições sejam iguais, o p-valor indica se esta hipótese é rejeitada o não a um nível de significância.

3 Materiais e Métodos

A metodologia utilizada neste trabalho foi a CRISP-DM. Suas etapas serão descritas nas próximas subseções.

3.1 Entendimento do Negócio

Essa é a primeira etapa para buscar compreensão adequada do problema, onde se definem os objetivos do projeto de mineração de dados. Esta etapa também estabelece os critérios para definição e interpretação dos resultados obtidos do processo de mineração de dados.

A base de dados a ser utilizada contém informações sobre clientes que podem determinar na análise de crédito, identificando se o cliente é um bom ou mau pagador. Para esta abordagem foi escolhido o método de classificação como técnica de mineração de dados.

O problema proposto refere-se a avaliar se o cliente é um bom ou mau pagador para a concessão de crédito através da base de dados proposta. O objetivo deste trabalho é de aplicar técnicas de mineração de dados com diversos

http://dx.doi.org/10.25286/repa.v3i3.967

algoritmos de classificação a fim de encontrar o melhor desempenho.

3.2 Entendimento dos Dados

A base de dados foi obtida com a empresa Neurotech, a qual se dispôs a oferecer uma parte de sua real base de dados. A base de dados possui 176 atributos e 500000 instâncias e nela encontram-se informações de clientes que podem ou não serem bons para receberem o crédito. Os tipos de dados encontram-se da seguinte forma:

A. **Cadastrais:** dados relacionados ao local onde o indivíduo vive como classe social, vizinhança, entre outros.
B. **Demográficos:** comparações com a vizinhança como renda, por exemplo.
C. **Financeiros:** atividade do indivíduo como consumidor.
D. **Geográficos:** exposição do indivíduo em locais considerados importantes.
E. **Partidos:** possíveis filiações a partidos políticos.
F. **Programas Sociais:** o indivíduo faz parte de ONGs, bolsa família, prouni, etc.
G. **Riscos:** verifica a exposição do indivíduo a alguns fatores considerados de risco.
H. **Servidor:** verifica se o indivíduo faz parte de algum serviço militar ou público.
I. **Web:** analisa a exposição do indivíduo em sites na internet com temas pré-selecionados.

Informações como classe social do consumidor, renda da vizinhança, atividade do consumidor no mercado financeiro, exposição a endereço de hotéis, filiação política, *flag* bolsa família, exposição risco web, *flag* servidor militar, *flag* servidor civil, etc. podem ser levadas em consideração no momento de escolha para aplicação de crédito, inclusive o seu montante.

3.2.1 Análise Estatística Bivariada

O mercado está cada vez mais competitivo e não existe mais espaço para interpretações errôneas e/ou incompletas. Mais do que nunca é preciso a obtenção de informações sólidas para a tomada de decisões. Por este motivo a análise de dados torna-se essencial para qualquer negócio que almeja ter sucesso.

Após a realização da PCA verificamos que a variável **RENDA** tem uma grande correlação (aproximadamente 0.6) com o alvo e a partir disso foi realizada a sua análise. Ela é do tipo categórica**,** suas categorias são: "ATE 2 SM" , "2 A 4 SM","4 A 10 SM", "10 A 20 SM", "ACIMA DE 20 SM" ou "NULL" quando não há informações de renda para aquele indivíduo. Foram realizados os seguintes procedimentos:

A. **Agrupamento:** O agrupamento foi realizado para determinar a quantidade de 'indivíduos' por categoria de renda e como eles estão relacionados com o alvo (**INDICE_BOM_CLI**).
B. **Porcentagem Total:** É a porcentagem da quantidade de indivíduos por categoria.
C. **Porcentagem da Taxa de Maus:** Foi realizada para calcular a 'taxa de maus', variável que indica quantos indivíduos por categoria pertencem ao alvo = '1'.

A partir dessas informações foi possível gerar o gráfico apresentado na Figura 1. No eixo Y estão as variáveis que representam a PORCENTAGEM TOTAL, no eixo X estão as categorias, e a curva indica a taxa de maus em porcentagem, em relação à porcentagem total.

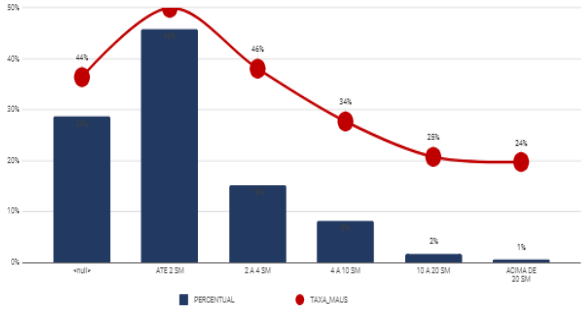


Figura 1 - Gráfico referente à análise bivariada da renda em relação ao alvo**.**

Após as transformações, formação de gráficos e análise, foi possível obter as seguintes conclusões:

- Quanto maior a renda do indivíduo, menor o 'risco' ou a probabilidade de ele ser um mal pagador (PERTENCER AO ALVO '1');
- A quantidade expressiva de nulos (143421) pode atrapalhar no desempenho do modelo final, uma forma de contornar essa situação é considerar que os

indivíduos dessa categoria se enquadram na categoria "ATÉ DOIS SM", realizar o cálculo da média entre essas variáveis.

Como os resultados obtidos após os processamentos não foram suficientemente relevantes, foi necessário o reajuste e utilização de outras técnicas de preenchimento de dados faltosos e pré-processamento. Para certificar que os dados estavam coerentes e obter um melhor entendimento do seu comportamento, após gerar o KS de cada uma foi realizada a análise bivariada. Com ela, foi possível visualizar outras formas de preencher os dados faltosos e como cada variável se comporta em relação ao alvo (**INDICE_BOM_CLI)**.

3.2.2 Análise de Correlação Entre as Variáveis

Para critério de análise dos atributos mais significativos, foi aplicada a técnica de correlação de Spearman. A técnica de spearman realiza a correlação entre duas variáveis para determinar as *features*mais relevantes da base de dados **[18]**. A execução de um algoritmo *featureselection*é de suma importância no processo de pré-processamento, pois é possível remover atributos que não trarão bons resultados na acurácia, além de diminuir o custo computacional para os algoritmos de classificação de dados.

Junto com a técnica de *Spearman*, é possível aliar recursos de visualização de dados, como o *heatmap*para que o cientista de dados consiga visualizar de maneira geral, a relevância dos atributos na base de dados. A Figura 2 apresenta o resultado da correlação de *Spearman* aplicado ao *heatmap*.

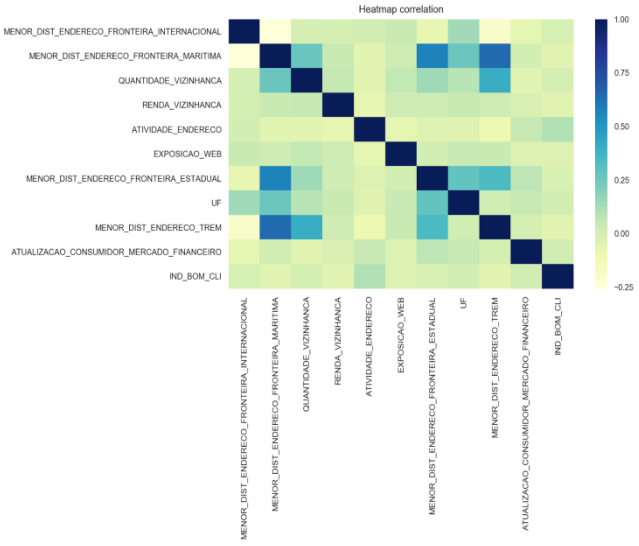


Figura 2 - *Heatmap* para as 10 primeiras variáveis mais relevantes.

3.2 Preparação da Base de Dados

Ao realizar a análise da base de dados utilizada no projeto, foram encontrados alguns problemas como dados faltosos, atributos contendo strings, alta amplitude dos valores e alta quantidade de atributos. Para a preparação dos dados, foram utilizados diversos softwares e plataformas, tais quais: Python, Excel e R, cada um utilizado para uma finalidade específica.

A resolução dos problemas citados serão discutidos nos pontos a seguir:

A. Tratamento de Dados Faltosos: Em 168 atributos faltaram dados a serem preenchidos, em todas as instâncias da base de dados esse problema foi detectado. Para resolução deste problema, foram analisados previamente os tipos de dados que compõem os atributos para posterior aplicação de um método de estimação de tendência central da base de dados, como mediana e moda. Nos espaços vazios correspondentes a dados categóricos que não poderiam conter dados faltosos foram aplicados a moda (16 atributos), nos dados categóricos que poderiam assumir dados faltosos foram preenchidos com zero (2 atributos), e nos dados contínuos foram aplicados à mediana (150). Apenas em

http://dx.doi.org/10.25286/repa.v3i3.967

oito atributos não foi feito o tratamento de dados faltosos.

B. Transformação de Dados: Alguns atributos contendo valores do tipo String estão presentes na base de dados, como por exemplo, Estados Brasileiros, Bancos, Classe Social, etc. Estes valores foram convertidos em dados numérico-categóricos para posterior execução dos algoritmos de classificação de dados, visto que estes utilizam dados numéricos (contínuos ou não) em suas execuções. Por exemplo, no atributo **UF**, que corresponde ao Estado do indivíduo analisado, os valores possíveis são as siglas correspondentes a eles e para transformação cada sigla recebeu um identificador único, de 1 a 27 (Unidades Federativas existentes no Brasil) e assim foram substituídos na base de dados. Processo igual a este foi realizados nos outros atributos, alterando apenas a quantidade de identificadores únicos.

Além disso, os dados da base não se encontram normalizados, existindo um desbalanceamento na amplitude dos valores dos atributos. Por exemplo, um dos atributos do tipo demográfico, que geralmente são nomeados com o prefixo **EXPOSICAO_ENDERECO** possuem valor mínimo igual a 0 e valor máximo de 2850 em uma de suas variáveis, enquanto o atributo **MENOR_DIST_BANCO** possui o mesmo valor mínimo, entretanto valor máximo igual a 998886. Isso ocorre na maioria dos atributos da base de dados, sendo necessário a aplicação de uma função de normalização em todos os atributos para que seus *ranges* tornem-se entre 0 e 1, esta função é descrita na equação 2.

$$F(i) = \frac{Xi - min(x)}{max(x) - min(x)} \quad \textbf{(2)}$$

Onde:F(i) = Valor a ser setado na cédula atual; Xi = Valor cédula atual; min(x) = menor valor do atributo; max(x) = maior valor do atributo.

Desta forma, com os dados normalizados, é possível executar os algoritmos de classificação sem a interferência de variáveis menos relevantes (pelo fato de ter uma amplitude maior) sobre as variáveis mais relevantes, proporcionando uma classificação estatisticamente mais confiável.

C. Seleção de atributos: A quantidade de atributos é de 176. A análise estatística de correlação com a aplicação do algoritmo PCA (*Principal Components Analysis*) torna-se essencial para redução de redundância e informações desnecessárias contidas na base de dados. Uma das vantagens que a redução de características pode proporcionar é uma mais rápida classificação e taxa de acurácia quase equivalente a que se obtém rodando o classificador na base original. Antes da aplicação do PCA, foi observado que o primeiro atributo da base de dados é um identificador (ID) assumindo-se chave primária da base de dados, o que é irrelevante para o processo de classificação, sendo esta excluída da execução dos experimentos subsequentes. Outra variável removida foi a variável demográfica**CAD_DEMOGRAFICO_VAR_35_1** porque a mesma estava duplicada.

Foi utilizada a técnica de *Kolmogorov-Smirnov*, a qual forneceu uma visão geral da base de dados e propôs uma melhor e menor seleção de variáveis contendo um número de 43 colunas. Para implementar esse teste estatístico, foi utilizado a linguagem python. O algoritmo foi aplicado na base de dados, gerando um csv com todas as variáveis e resultados do KS, sendo possível montar um ranking com as variáveis que possuem melhor valor.

D. Balanceamento das Classes: Em relação à quantidade de instâncias da classe 0 (aptos para receber o crédito) e da classe 1 (não aptos), pode-se afirmar que a base de dados encontra-se balanceada, não sendo necessária a aplicação de algum algoritmo para o balanceamento das classes. A classe 0 (aprovados na análise de crédito) possui 253804 instâncias, enquanto a classe 1 (não aprovados) possui 246196 instâncias.

3.4 Modelagem

Nesta seção serão apresentados os algoritmos utilizados para o processo de modelagem. A base de dados foi gerada através do teste KS, resultando em um ranking com as variáveis de maior correlação, sendo 4 bases de dados com 10, 20, 30 e 40 variáveis de maior correlação respectivamente. O conjunto de dados foi separado em 80% para treinamento e 20% para teste. Cada algoritmo foi executado 30 vezes para obtenção de um resultado estatisticamente mais confiável.

3.4.1 *Naive Bayes*

O algoritmo *Naive Bayes* foi utilizado com o auxílio da biblioteca *scikit-learn* para Python, utilizando suas configurações padrão. Para criação do modelo de classificação, foi utilizada a função Gaussian NB.

3.4.2 K-NN

Para execução do treinamento com KNN, foi utilizado com o auxílio da biblioteca *scikit-learn* para Python. O K-NN também possui a necessidade de inicialização de alguns parâmetros como a distância a ser utilizada e o valor K. A configuração do K para os experimentos foi de K=5, devido o número de classes ser par foi definido um número ímpar e maior que 3 para uma maior capacidade de consideração de vizinhos com características similares. Como o vetor de dimensões foi diminuído, foi utilizada a distância euclidiana para cálculo da distância de todos os pontos entre si.

3.4.3 Regressão Logística

A Regressão Logística foi aplicada utilizando a biblioteca *scikit-learn*para Python, utilizando suas configurações padrão, sendo que esta não oferece opções para personalização de parâmetros.

3.4.4 *Deep Learning*

Para o treinamento da *Deep Learning* (DL) é necessário definir os dados de entrada da rede neural, a quantidade de neurônios na camada de entrada, a quantidade de neurônios nas camadas intermediárias e na camada de saída, definir a taxa de conectividade, o número de ciclos do warmup, a função de ativação na camada intermediária e a equação para o cálculo do erro. A Tabela 1 apresenta os valores utilizados para os parâmetros citados.

Tabela 1 - Valores dos parâmetros utilizados na Deep Learning.

Parâmetro	Valor
Quantidade de Neurônios na Camada de Entrada	18
Quantidade deNeurônios nas Camadas Escondidas	Qtdneurônios entrada * 1.5 (progressivamente)
Quantidade de Neurônios na Camada de Saída	136,6875
Função de Ativaçãona Camada Intermediária	sigmoide

4 Resultados

Esta seção descreve os resultados para cada algoritmo utilizado para as bases de dados.

4.1 *Naive Bayes*

O Quadro 1 apresenta os resultados obtidos com a aplicação do *Naive Bayes.*

Quadro 1 - Resumo dos resultados obtidos com a aplicação do Naive Bayes.

BASE	ACURÁCIA (MÉDIA)	DESVIO PADRÃO
10	0,554776	0,001
20	0,56265	0,001
30	0,554799333	0,001
40	0,559786	0,01

Com a execução do *Naive Bayes*, foi possível observar que o número de atributos em relação às bases em estudo não possui nenhuma relevância para o modelo, onde há uma oscilação para mais ou menos conforme o número de atributos é aumentado. Este é outro modelo

http://dx.doi.org/10.25286/repa.v3i3.967

preciso, pois o desvio padrão do mesmo foi relativamente baixa, em contrapartida o modelo proporcionou também uma baixa acurácia.

4.2 K-NN

O Quadro 2 apresenta os resultados obtidos com a aplicação do NaiveBayes.

Quadro 2 - Resumo dos resultados obtidos com a aplicação do Naive Bayes.

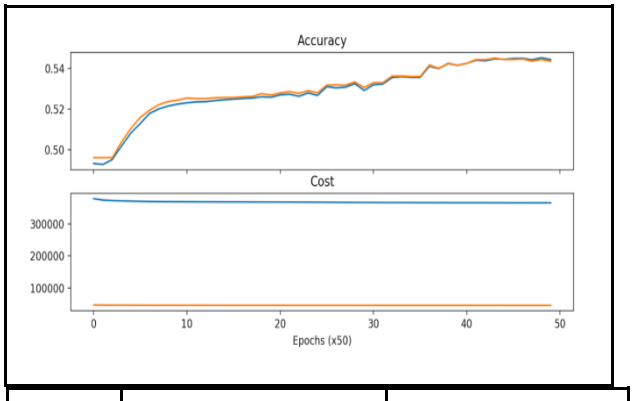
BASE	ACURÁCIA (MÉDIA)	DESVIO PADRÃO
10	0,5488	0,00112015393
20	0,550019	0,00114078119
30	0,55390	0,00116039911
40	0,55621	0,00116178119

O algoritmo K-NN seguiu o mesmo padrão em relação a acurácia dos outros modelos, permanecendo na faixa de 54% a 55%. O mesmo cenário de avaliação da acurácia dos outros modelos é também válida para este modelo. Semelhantemente a regressão logística, este algoritmo pode apresentar resultados melhores de acordo quando o número de atributos é aumentado.

4.3 Regressão Logística

O Quadro 3 apresenta os resultados obtidos com a aplicação da regressão Regressão Logística.

Quadro 1 - Resumo dos resultados obtidos com a aplicação da Regressão Logística

BASE	ACURÁCIA (MÉDIA)	DESVIO PADRÃO
		
10	0,551208666	0,001553385032
20	0,556811666	0,001319963645
30	0,562282	0,001521241326
40	0,576364	0,001419870319

Foi observado que o número de amostras proporcionou pouco impacto na acurácia do modelo, obtendo apenas um ganho de 2% na base de dados de 40 atributos, em relação à execução na base de 10 atributos. Foi observado que o desvio padrão foi considerado baixo, e o modelo apesar da baixa acurácia é preciso. De todos os modelos testados neste trabalho, este modelo proporcionou a maior acurácia, tendo aproximadamente 57,64% nesta métrica utilizando na base de dados com 40 atributos.

4.4 *Deep Learning*

As Figuras 3, 4, 5, 6 e 7 apresentam os resultados (em forma de gráficos) obtidos com a aplicação da *Deep Learning*.

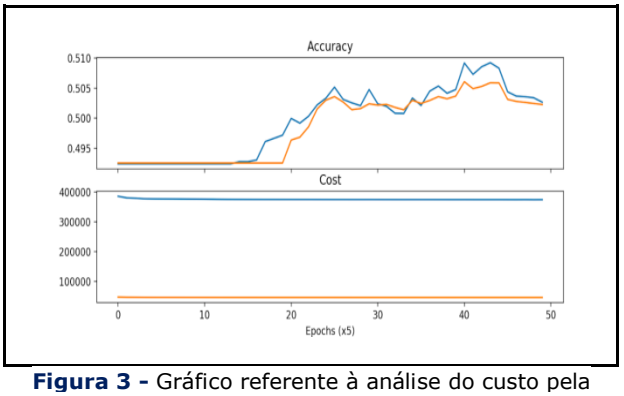


Figura 3 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando apenas 10 variáveis.

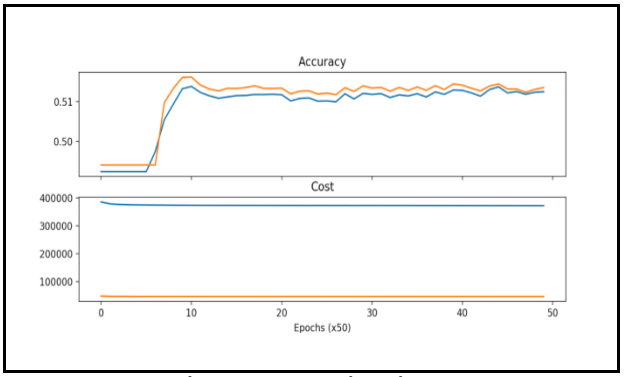


Figura 4 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando apenas 20 variáveis.

Figura 5 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando apenas 30 variáveis.

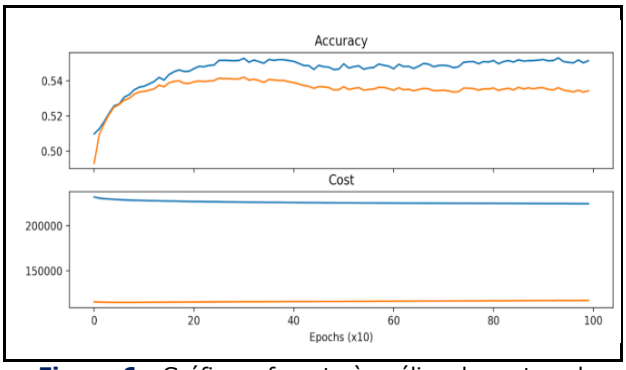


Figura 6 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando apenas 40 variáveis.

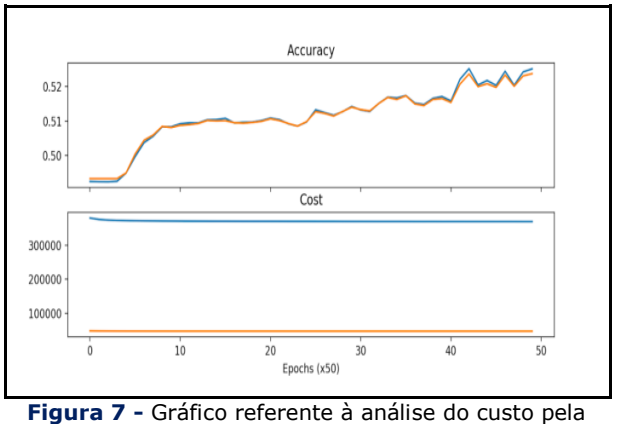


Figura 7 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando todo o resto da base.

Portanto, o melhor resultado foi obtido utilizando a base de dados completa, como pode ser visto na Figura 5, pois além de obter o melhor resultado de acurácia foi o que obteve a melhor relação entre o custo com a acurácia.

5 Conclusão

Este trabalho apresentou uma abordagem utilizando aprendizado de máquina para extração de conhecimento em uma base de dados, com a finalidade de classificar indivíduos aptos ou não para receber crédito, sendo este um sistema para análise de crédito. Esta abordagem seguiu a metodologia CRISP-DM, onde foram testadas diversas abordagens para preenchimento de dados faltosos, seleção de atributos e modelagem dos experimentos.

Foram utilizados diversos algoritmos clássicos para criação de modelos de predição para análise de crédito, tais quais: NaiveBayes, Regressão Logística, Deep Learning e K-NN, sendo estes analisados observando a métrica de acurácia. Cada algoritmo foi executado 30 vezes, tendo sido coletados em cada execução o dado correspondente a acurácia, sendo este coletado e para análise foi calculado sua média e desvio padrão. De acordo com os resultados obtidos, não foi possível obter uma alta acurácia de classificação dos dados separados para teste (20% da base).

A baixa acurácia avaliada neste projeto, ocorreu pela baixa correlação das variáveis. A *feature*mais conceituada foi a de ATIVIDADE_EMAIL, cujo *ranking*, de acordo com Spearman, foi de 0.11. Para um algoritmo de classificação, esse valor é considerado extremamente baixo e com isso, torna-se inviável o reconhecimento de padrões **[19]**.

Apesar das inconsistências identificadas nessa base de dados, é possível obter resultados mais significativos através do Deep Learning, mas será necessário aplicar alguns experimentos, como monitorar a acurácia a partir do acréscimo e remoção das *hiddenlayers* e neurônios, adicionar mais dados para que os neurônios consigam aprender mais rapidamente e avaliar outros algoritmos do *Framework Tensorflow*.

Referências

[1] ROSS, Stephen A. et al. **Administração financeira**. São Paulo: Editora Atlas, 1995.

[2] ALMEIDA, Hamilton. Políticas econômicas serão iguais até 95. **Zero Hora**, Porto Alegre, 24 mai 1992.

[3] SCHRICKEL, W. K. **Análise de crédito**: Concessão e Gerência de Empréstimos, São Paulo: Atlas, 1994.

[4] BERRY, Michael J. A.; LINOFF, **Gordon. Data mining techniques**. USA : Wiley Computer Publishing, 1997.

http://dx.doi.org/10.25286/repa.v3i3.967

[5] NARENDRAN, C. R. Data Mining-Classification Algorithm–Evaluation. May 8th, 2009.

[6] WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: International conference on the practical applications of knowledge discovery and data mining, 4., 2000, Citeseer. **Proceedings**... Citesser, 2000. p. 29-39.

[7] SILVA, José Pereira da. **Gestão e análise de risco de crédito**. 6 ed. São Paulo: Atlas, 2008.

[8] YU, L.; WANG, S.; LAI, K. K.; ZHOU, L. **Bio-Inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machines**. Berlim, Heidelberg: Springer-Verlag, 2008.

[9] LAROSE, Daniel T. **Discovering Knowledge in Data**: An Introduction to Data Mining. New Jersey: John Wiley & Sons, Inc, 2005.

[10] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge Discovery and Data Mining: towards a unifying framework. In: FELIX, Priscila et al. **Proceedings of second international conference on electrical system , tecnology and informacion.** Lecture Notes in Electrical Engineering, v.365. Berlim: Springer, 2015. p. 82-88.

[11] ANDERSON, R. **The Credit Scoring Toolkit**: Theory and Practice for Retail Credit Risk Management and Decision Automation. New York: Oxford University Press, 2007.

[12] ARSENAULT, MARC-OLIVER. Kolmogorov-Smirnov tes: a needed tool in your data science toolbox. **Towards data science**. 21 Nov. 2017. Disponível em: <https://towardsdatascience.com/kolmogorov-smirnov-test-84c92fb4158d>. Acesso em: 5 jun. 2018.

[13] CONOVER, W. J. **Practical Nonparametric Statistics**. 3. ed. New York: John Wiley and Sons, 1999.

[14] FRIEDMAN, Nir; GEIGER, Dan; GOLDSZMIDT, Moises. Bayesian network classifiers. **Machine learning**, v. 29,, 1997.

[15] FRIEDMAN, N.; GOLDSZMIDT,M. Building Classifiers Using Bayesian Networks. In: National Conference on Artificial Intelligence (AAAI96), 30., 1996, Portland. **Proceedings**... Prontland, AAAI Press,1996. v.2, p.1277-1284.

[16] SCHMIDHUBER, Jürgen. Deep learning in neural networks: An overview. **Neural networks**, v. 61, p. 85-117, 2015

[17] DENG, Li et al. Deep learning: methods and applications. **Foundations and Trends in Signal Processing**, v. 7, n. 3–4, p. 197-387, 2014.

[18] Correlation (Pearson, Kendall, Spearman). Disponivelem: <http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/>. Acesso em :8 djul. 2018.

[19] BROWNLEE, Jason. How to Use Correlation to Understand the Relationship Between Variables. Statistical Methods, **Machine Learning Mastery**, 27 April 2018. Disponivel em: <https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/>. Acesso em: 8 jul. 2018.

Aplicação de Regras de Associação em Dados da Criminalidade da Cidade do Recife

Application of association rules between criminality data from the city of Recife

Bettina Cavalcanti Araújo[1] orcid.org/0000-0002-9821-1812
Alexandre Magno Andrade Maciel[1] orcid.org/0000-0003-4348-9291

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: bca@ecomp.poli.br

Resumo

Os crescentes números da criminalidade na cidade do Recife fez com que o Governo do Estado aumentasse a frota de policiais nas ruas no primeiro trimestre de 2018 esperando uma queda nos crimes. Porém, combater a violência apenas de forma reativa é uma estratégia insuficiente para uma redução significativa, por conta da complexidade do perfil criminal de cada localidade, dificultando a organização tática dos policiais, sendo necessária a criação de ações preventivas em paralelo. Com a aplicação dos algoritmos Apriori e FP-Growth, foram extraídas Regras de Associação que geraram uma análise preditiva de dados de ocorrências de crimes relatados nas delegacias do Recife, como o fato de um crime ocorrer no período da noite implica que sua natureza seja homicídio. Espera-se assim auxiliar a polícia nas investigações e estratégias para um combate mais eficaz à criminalidade.

Palavras-Chave: Criminalidade na cidade do Recife; Delegacias do Recife; Mineração de Dados; Regras de Associação; Apriori; FP-Growth.

Abstract

The rising numbers of crime in the city of Recife have prompted the State Government to increase the fleet of police officers on the streets in the first quarter of 2018, expecting a decrease in crime. However, combating violence only in a reactive way is an insufficient strategy for a significant reduction, because to the complexity of the criminal profile of each locality, making difficult the tactical organization of the police, and it is necessary to create preventive actions in parallel. With the application of the algorithms Apriori and FP-Growth, Association Rules were extracted that generated a predictive analysis of data of occurrences of crimes reported in the police stations of Recife, as the fact that a crime occurs in the night period implies that its nature is homicide . It's hoped to assist the police in investigations and strategies for a more effective fight against crime.

Key-words: *Criminality in the city of Recife; Police stations in Recife; Data Mining; Association Rules; Apriori; FP-Growth.*

 http://dx.doi.org/10.25286/repa.v3i3.974

1 Introdução

Em Pernambuco, temos visto a criminalidade crescer nos últimos anos, num ritmo três vezes maior do que o Rio de Janeiro, cidade que possui a maior fama de ser o epicentro da violência no país, apresentando de 2014 a 2017 um avanço de 37,8%, comparado aos 12,6% do Rio **[1]**. Em paralelo, Recife atingiu em 2017, a 22ª posição no ranking das cidades mais violentas do mundo, pela ONG Mexicana Segurança, Justiça e Paz, com uma média de 54,43 homicídios a cada 100 mil habitantes **[2]**. Analisando os números de 2006 a 2016, percebe-se que o Estado está voltando ao patamar de dez anos atrás, sendo alguns desses anos de sucesso na redução de crimes com a implantação do programa Pacto Pela Vida **[3]**.

O Governo de Pernambuco aumentou a frota de policiais em 1.214 novos contratos realizados nos primeiros meses de 2018, visando reprimir a violência com a reativação de delegacias que estavam fechadas por falta de delegados e mais profissionais focados na investigação do tráfico de drogas, grande problema instaurado na cidade do Recife **[4]**. Embora essas medidas reativas não sejam suficientes para obter uma queda significativa nos números de crimes, pois assim como foi feito em Medellín, é preciso investir em inteligência de dados para maior eficácia no combate à criminalidade. Segundo Weriqui Bezerra, "quando você parar de olhar para o problema, talvez encontrará a solução que tanto procura. Não foque nos sintomas apenas, mas na raiz do problema".

A cidade colombiana Medellín, é um exemplo de sucesso na redução da violência por meio de investimento em segurança pública, inteligência e equipamentos para polícia. Nos anos 90, Medellín foi considerada a metrópole mais violenta do mundo, com média de 380 homicídios por 100 mil habitantes. Após aplicação de tecnologia e maior capacitação da polícia, a cidade possui atualmente a taxa de 21 homicídios por 100 mil, tornando-se uma cidade-modelo em um país subdesenvolvido **[5]**.

Foi realizado um estudo utilizando uma abordagem estatística para analisar possíveis relações entre variáveis de ocorrências na cidade do Recife **[6]**. Os resultados descritivos foram obtidos através de testes de relações conduzidos manualmente pelo usuário, o que traz algumas limitações para alcançar um nível maior de abrangência e detalhe nesses resultados. Com o objetivo de realizar uma análise preditiva dos dados de ocorrências criminais da capital Pernambucana, através da utilização da mineração de dados para encontrar possíveis padrões frequentes de crimes. Com o presente estudo espera-se auxiliar os órgãos competentes à segurança pública a compreender possíveis relações associativas entre as variáveis cadastradas nos boletins de ocorrências nas delegacias da cidade, para contribuir com a promoção de estratégias preventivas de combate ao crime.

2 Fundamentação teórica

2.1 Trabalhos relacionados

Em busca de auxiliar a segurança pública a traçar estratégias mais assertivas no combate ao crime, no Brasil já existem alguns trabalhos com objetivos semelhantes tendo a criminalidade como centro do estudo. Na Universidade Federal Fluminense, foi construída uma ferramenta web chamada SiAPP (Sistema de Apoio ao Policiamento Preditivo) utilizando dados referentes a ocorrências da cidade de Niterói, em busca da predição de crimes a partir do aprendizado de máquina. Foram usados 781 registros que geraram 281 pontos de importância na cidade. Por meio da geolocalização e uso de algoritmos, foram extraídas regras de associação visualizadas em um mapa da região, onde cada regra foi evidenciada pelo bairro a qual pertence. Após a visualização de resultados e análise, foram encontradas indicações do comportamento da criminalidade em determinados pontos, como por exemplo, que houve um aumento na ocorrência de furtos entorno de escolas em dias comerciais e aumento de roubos de carteira na região do centro a noite **[7]**.

Outro estudo foi realizado na Universidade Federal de Viçosa, com o objetivo de realizar estudos em relação a segurança pública em cidades de pequeno porte, com a implementação

de um sistema SIG (Sistemas de Informações Geográficas) para auxiliar no cadastro de atividades e mapeamento de crimes, já que cidades pequenas geralmente não possuem recursos tecnológicos avançados. Foi utilizada uma base disponibilizada pela Polícia Militar da cidade de Rio Pomba, com dados de 2009,2010 e 2011. Com isso, foram utilizados os algoritmos de redes neurais para classificar o quanto a polícia deve atuar nas ações de combate e prevenção para alcançar a diminuição de crimes, e Apriori para extrair padrões criminais relevantes para tornar as ações policiais mais eficazes contra a violência **[8]**.

Ao analisar os resultados obtidos nesses estudos, pode-se compreender o quanto as regras de associação extraídas dos algoritmos aplicados às bases de dados, podem ser relevantes para pesquisa e análise em relação ao combate a crimes. Os resultados gerados dificilmente seriam encontrados sem mineração de dados, por conta do tamanho e diversidade dos registros de ocorrências criminais. A expectativa é que assim como foi feito em Niterói e Viçosa, sejam encontrados padrões frequentes de crimes da cidade do Recife.

2.2 Regras de Associação

Por sua grande capacidade de aplicações, é umas das técnicas de mineração de dados mais utilizadas atualmente em campanhas de marketing, logística no comércio, atividades que atuam com criação de estratégias e tomada de decisão **[9]**. De origem descritiva, estas aplicações têm como princípio descobrir possíveis associações, da presença de um item com outro em uma mesma transação, ou seja, na mesma operação de consulta, no conjunto de dados, encontrando padrões ou tendências frequentes relevantes para análise, em forma de regras de associação.

O problema clássico que mostra o poder das regras de associação surgiu com a observação da cesta de compras dos clientes de uma rede de supermercados dos Estados Unidos, para descobrir relações entre os itens comprados juntos frequentemente. Descobriu-se uma regra curiosa, que indicava {cerveja} -> {fraldas}, informação essa que após análise, concluiu que homens que iam comprar fraldas para os filhos, aproveitavam para levar cerveja. Seguindo o resultado do estudo, os produtos foram colocados um ao lado do outro, e as vendas aumentaram 30%.

Uma regra de associação é uma implicação da forma A ⇒ B, onde A e B são conjuntos de itens, também chamados de *itemsets*, e A ∩ B = ∅. A regra A ⇒ B vale no conjunto de transações T com suporte s, onde s é a porcentagem de transações em T que contém A ∪ B, ou seja, pode ser entendida como a relevância estatística de uma regra **(1)**. A regra A ⇒ B tem confiança c no conjunto de transações T, onde c é a porcentagem de transações em T contendo A que também contém B, sendo considerada a certeza da ocorrência de uma regra **(2)**. A determinação do suporte e da confiança é de extrema importância, pois serve para eliminar regras que sejam pouco significativas por apresentarem valores muito baixos. Outro fator para verificação de regras consideráveis, é o lift, capaz de indicar o quanto B é frequente, quando A aparece **(3)**. A convicção é uma medida responsável por indicar a força de uma implicação, tentando exprimir até que ponto o antecedente de determinada regra e a negação do consequente dessa regra são independentes **(4)**. Além disso, dois parâmetros devem ser informados pelo usuário ao algoritmo utilizado: o *minSup* e *minConf*, que são respectivamente, um valor mínimo para o suporte e um valor mínimo para a confiança. Uma regra é dita como frequente, se atende a um suporte mínimo. E caso atenda um suporte e confiança mínima, é dita como uma regra forte.

$$Sup(A \Rightarrow B) = \frac{P(A\ U\ B)}{|T|} \quad \textbf{(1)}$$

$$Conf(A \Rightarrow B) = \frac{P(A \mid B)}{|T|} \quad \textbf{(2)}$$

$$Lift(A \Rightarrow B) = \frac{Conf(A \Rightarrow B)}{Sup(B)} \quad \textbf{(3)}$$

$$Conv(A \Rightarrow B) = \frac{Sup(A)\ x\ Sup(\neg B)}{Sup(AU\ \neg B)} \quad \textbf{(4)}$$

2.2.1 Apriori

O algoritmo foi proposto em 1994, por Agrawal e Srikant, foi o pioneiro, sendo atualmente o mais famoso e utilizado quando o assunto é regras de

http://dx.doi.org/10.25286/repa.v3i3.974

associação por ele ser eficaz em encontrar *itemsets* frequentes em grandes bancos de dados, gerando posteriormente regras fortes de associação. O nome Apriori, é baseado no fato do algoritmo usar conhecimento *a priori* para estudo das características dos *itemsets* frequentes **[10]**. O funcionamento do Apriori, é dividido em duas partes, geração do conjunto de itens frequentes e geração das regras, ocasionando uma economia de custo computacional.

Primeiramente ocorre uma geração de um conjunto de *k-itemsets* candidatos, logo após ele percorre toda base de dados, verificando se os candidatos são mesmo frequentes, ou seja, aqueles que possuem o suporte com valor maior do que o *minSup* determinado, gerando um conjunto de apenas itens frequentes.

De posse de um conjunto de *k-itemsets* frequentes, com k ≥ 2, é gerado as regras de associação, de forma que caso os itens AB e ABCD sejam frequentes, podemos avaliar a regra AB ⇒ CD, aplicando o cálculo da confiança, sendo conf(AB ⇒ CD) = sup(ABCD)/sup(AB). Caso o valor da confiança seja maior ou igual ao *minConf* determinado, a regra é considerada válida.

O princípio dos *Itemsets* Frequentes do algoritmo, diz que se um itemset é frequente, todos os seus subconjuntos não vazios também são frequentes, logo todo *itemset* que não é frequente, da mesma forma seus subconjuntos também não são frequentes. Este preceito só é válido de acordo com a propriedade anti-monotônica do suporte, que garante que o suporte de um conjunto frequente nunca exceda o suporte de seus subconjuntos **(5)**.

$$\forall A, B: (A \subseteq B) \Rightarrow s(A) \geq s(B) \quad \textbf{(5)}$$

A validade do princípio permite que ao identificar um conjunto de itens não frequentes, o subgrafo que o contém pode ser podado, evitando que todos os nós sejamos visitados desnecessariamente, consequentemente economizando tempo de execução.

2.2.2 FP-Growth

Com o uso do Apriori e de outros algoritmos que utilizam abordagem semelhante, foram encontradas algumas dificuldades, como a execução de muitos acessos ao banco de dados e no tratamento de uma grande quantidade de conjuntos de itens candidatos, ocasionados por um grande número de itens frequentes ou caso o valor do *minSup* seja muito baixo. Para solucionar essas e outras limitações, foi desenvolvido por Han, Pey e Yin, em 2000, o algoritmo FP-Growth (*Frequent Pattern Growth*) **[11]**.

Baseado em uma estrutura de dados de árvore de prefixos para padrões frequentes, usada para extração dos conjuntos de itens constantes na própria estrutura, capaz de armazenar essas informações de forma compactada, permitindo uma mineração de dados bastante eficaz pois não necessita de vários acessos a base de dados, sendo apenas duas vezes, uma para encontrar e ordenar os *itemsets* frequentes e outra para construção da árvore, chamada de FP-Tree (*Frequent Pattern Tree*).

A solução do algoritmo é dada através de três pilares, a compactação do banco de dados é realizada dando lugar a uma estrutura geralmente bastante menor em uma árvore FP-Tree, em seguida é usado o algoritmo para minerar a árvore em busca de evitar uma grande geração de conjuntos de itens candidatos, e por fim, as tarefas de mineração são decompostas em tarefas menores usando o método particional. O processo de construção da FP-Tree acontece primeiramente após a seleção do valor do *minSup*, com a varredura da base de dados e o armazenamento e ordenação decrescente dos conjuntos de itens frequentes encontrados.

3 Metodologia

A metodologia empregada neste trabalho está baseada em um dos mais conhecidos processos para realização de um projeto de mineração de dados chamado CRISP-DM (*Cross Industry Standard Process for Data Mining*). Criado em 1996, para ser aplicado em diversas áreas de negócios, sem a dependência de uma ferramenta, o modelo é composto por seis etapas não rigorosas que podem prosseguir ou retornar em diferentes fases, possuindo o caráter cíclico da mineração de dados, como mostra a Figura 1.

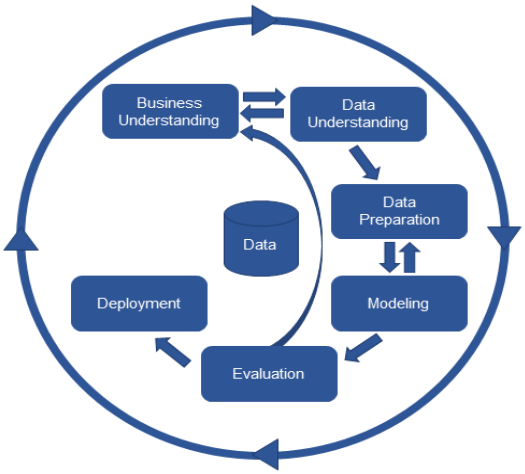


Figura 1 – Ciclo de vida do modelo CRISP-DM

3.1 Entendimento do negócio

Para realização deste estudo, foram fornecidas pela Secretaria de Defesa Social de Pernambuco, cinco bases de dados, referentes a ocorrências criminais registradas nas delegacias do Estado ou pela *internet* no período de 1 de Janeiro à 30 de Setembro de 2017. Cada base é caracterizada pela separação por natureza do crime, onde tem seus registros relacionados a apenas essa natureza identificada no cadastro do boletim de ocorrência, sendo elas: CVLI (Crimes Violentos Letais Intencionais), CVP (Crimes Violentos contra o Patrimônio), Furto, Furto de Veículo e Roubo de Veículo. De posse do banco de dados de ocorrências de crimes, é possível analisar, preparar e minerar os dados para encontrar regras de associação entre os registros, ou seja, relacionamentos frequentes entre determinados atributos.

3.2 Entendimento dos dados

A base CVLI contém 9 variáveis: ÁREA, DATA, MÊS, PERÍODO DO DIA, IDADE, SEXO, NATUREZA JURÍDICA, LOCAL GENÉRICO, TOTAL. A base CVP contém 5 variáveis: ÁREA, LOCAL, DATA, PERÍODO DO DIA, TOTAL. As base de Roubo e Furto de Veículo contém as mesmas 5 variáveis: ÁREA, DATA, NATUREZA, PERÍODO, TOTAL. E por último a base de Furto contém 5 variáveis: ÁREA, DATA, LOCAL, PERÍODO, TOTAL.

Todas as bases possuem o atributo ÁREA como informação geográfica da localidade onde ocorreu o crime, sendo os registros desse campo, dados por AIS (Área Integrada de Segurança), divisão territorial feita em Pernambuco para melhor acompanhamento de ações e resultados, gerando 26 AIS para todo o Estado, porém nas bases de dados fornecida, encontram-se apenas registros das AIS referentes a capital Pernambucana, sendo elas:

- AIS 1 - Santo Amaro (compreende bairros como Boa Vista, São José, Ilha Joana Bezerra).
- AIS 2 - Espinheiro (compreende bairros como Cordeiro, Água Fria, Campo Grande).
- AIS 3 - Boa Viagem (compreende bairros como Ipsep, Jordão, COHAB).
- AIS 4 - Várzea (compreende bairros como Jardim São Paulo, Afogados, Mustardinha).
- AIS 5 - Apipucos (compreende bairros como Casa Amarela, Vasco da Gama, Macaxeira).

Para melhor entendimento dos atributos e seus registros presentes nas seis bases de dados, será descrito informações e significados no quadro 1 referente ao dicionário de dados:

Quadro 1 – Dicionário de Dados dos campos presentes nas bases utilizadas (continua)

CAMPO	DESCRIÇÃO
ÁREA	Atributo nominal relacionado a área onde ocorreu o crime (AIS1, AIS2, AIS3, AIS4 e AIS5).
MÊS	Atributo nominal derivado da variável DATA, que informa o mês da ocorrência do crime.
PERÍODO	Atributo nominal que informa o período do dia em que o crime aconteceu (manhã, tarde, noite ou madrugada).
NATUREZA	Atributo nominal que informa a natureza do crime (CVP, furto, homicídio, roubo de veículo, etc.).

Quadro 1 – Dicionário de Dados dos campos presentes nas bases utilizadas (continuação)

CAMPO	DESCRIÇÃO
DIA_SEMANA	Atributo nominal que informa o dia da semana que ocorreu o crime.

http://dx.doi.org/10.25286/repa.v3i3.974

SEXO	Atributo nominal que informa o gênero da vítima (feminino, masculino ou desconhecido).
LOCAL	Atributo nominal que informa o local que o crime ocorreu.

3.3 Preparação dos dados

Após um estudo das bases de dados fornecidas e dos tipos de entradas necessárias para aplicação dos algoritmos de regras de associação, Apriori e FP-Growth, foi realizado um pré-processamento dos dados para obtenção de melhores resultados.

O primeiro algoritmo utilizado foi o Apriori, e ele admite apenas variáveis nominais, pensando nisso, primeiramente foi criada uma nova base chamada de INTEGRADA, com a junção de variáveis em comum e relevantes entre as cinco bases, considerando campos com possibilidade de transformação das variáveis de quantitativas para categóricas, foram selecionados cinco atributos: ÁREA, MÊS, PERÍODO, NATUREZA, DIA, SEMANA.

Cada base foi analisada individualmente, e também sofreram transformações. A base CVLI, que tinha 9 variáveis, ficou com 5, todas nominais: ÁREA, PERÍODO, SEXO, NATUREZA, LOCAL. As bases CVP e Furto, dos 5 atributos, ficaram 4: ÁREA, MÊS, LOCAL, PERÍODO. E as bases Furto de Veículo e Roubo de Veículo, que tinham 5 variáveis, ficaram com 3: ÁREA, MÊS, PERÍODO.

3.1 Modelagem

Com as bases de dados devidamente tratadas, foi escolhido para minerar os dados o *software Weka* na versão pois possui alguns algoritmos já implementados agregados a ferramenta, e uma das técnicas presentes são regras de associação, com os algoritmos Apriori e FP-Growth, os quais podem ser aplicados nas bases fornecidas, resultando em padrões frequentes nos registros. Para o algoritmo Apriori foi usado o banco de dados transformado, com todas as variáveis categóricas. Para aplicação do algoritmo FP-Growth na ferramenta *Weka*, foi realizada uma transformação nos atributos, tornando todos binários.

3.4 Avaliação

Durante a aplicação dos algoritmos na ferramenta *Weka*, foram experimentados vários valores de *minSup* e *minConf* para extrair as melhores regras. Foi percebido que valores muito altos, conseguiram extrair poucas ou nenhuma regra por conta da grande variedade de registros nas bases, e insistir nesses valores poderia ocasionar a perda de alguns conjuntos de itens raros e relevantes. Foi preciso realizar diversos testes até encontrar os valores ideais de *minSup* e *minConf* como parâmetros para serem aplicados em cada base de dados.

3.5 Desenvolvimento

Para esta pesquisa não foi realizada a etapa de desenvolvimento de um sistema, mas sim a entrega da análise dos resultados gerados pela aplicação de regras de associação por meio deste artigo.

4 Experimentos e Resultados

Inicialmente foi aplicado o algoritmo Apriori, nas seis bases individualmente, gerando diferentes regras relevantes para análise. Para a base Integrada, que possui 52.873 registros foi escolhido um *minSup* de 0,05 e um *minConf* de 0,5 que geraram os *itemsets* candidatos na tabela 1.

Tabela 1 – Conjunto de itens candidatos da base Integrada

k	Quantidade
1	28
2	49
3	1

Dos itens candidatos encontrados, podemos destacar que as áreas de maior ocorrência são as AIS 1 e AIS 3, a natureza dos crimes mais cometidos é a CVP, sendo a maior parte dos crimes ocorridos à noite, os meses de mais registros foram Agosto e Janeiro, e o dia da

semana de maior incidência foi a sexta-feira. Logo após a geração dos candidatos, foram geradas as melhores regras, onde algumas delas e os valores de sua Confiança estão descritas respectivamente na tabela 2.

Tabela 2 - Regras de associação geradas para a base Integrada

Regra	Confiança
{PERÍODO=Noite}⇒ {NATUREZA=CVP}	0,66
{ÁREA=AIS3, PERÍODO=Noite} ⇒ {NATUREZA=CVP}	0,65
{ÁREA=AIS4} ⇒ {NATUREZA=CVP}	0,60
{MÊS=Janeiro}⇒{NATUREZA=CVP}	0,59
{DIA_SEMANA=Segunda} ⇒ {NATUREZA=CVP}	0,58

Para a base CVLI, que possui 601 registros foi escolhido um *minSup* de 0,05 e um *minConf* de 0,5 que geraram os *itemsets* candidatos na Tabela 3.

Tabela 3 – Conjunto de itens candidatos da base CVLI

k	Quantidade
1	8
2	11
3	6
4	4

Dos itens candidatos encontrados, podemos analisar que a área de maior incidência de casos tem seus dados não informados, sendo caracterizada pela sigla NI(Não Informada), a natureza dos crimes mais cometidos foi Homicídio, sendo a maior parte dos crimes ocorridos no período da noite, a grande maioria das vítimas é do sexo masculino, o local de maior ocorrência de crimes acontece em logradouros público, e o mês com mais registros de crimes CVLI foi Setembro. Logo após a geração dos candidatos, foram geradas as melhores regras, onde algumas delas e os valores de sua Confiança estão descritas respectivamente na tabela 4.

Tabela 4 – Regras de associação geradas para a base CVLI (continua)

Regra	Confiança
{MÊS=Setembro} ⇒ {NATUREZA=Homicídio}	0,98
{PERÍODO=Noite} ⇒ {NATUREZA=Homicídio}	0,98
{SEXO=Masculino, LOCAL= Logradouro público} ⇒ {NATUREZA=Homicídio}	0,97
{PERÍODO=Noite, LOCAL=Logradouro público, NATUREZA=Homicídio} ⇒ {SEXO=Masculino}	0,97
{PERÍODO=Manhã, LOCAL=Logradouro público, NATUREZA=Homicídio} ⇒ {SEXO=Masculino}	0,95

Para a base CVP, que possui 18.026 registros foi escolhido um minSup de 0,01 e um minConf de 0,3 que geraram os itemsets candidatos na tabela 5.

Tabela 5 - Conjunto de itens candidatos da base CVP

k	Quantidade
1	36
2	161
3	25

Dos itens candidatos encontrados, podemos destacar que a área com mais registros de Crimes Violentos contra o Patrimônio foi a AIS1, ocorridos principalmente em Vias públicas, sendo a maioria dos delitos cometidos no período da noite, e o mês de maior incidência desse tipo de crime foi em Janeiro. Logo após a geração dos candidatos, foram geradas as melhores regras, onde algumas delas e os valores de sua Confiança estão descritas respectivamente na tabela 6.

Tabela 6 – Regras de associação geradas para a base CVP

Regra	Confiança
{ÁREA=AIS4, MÊS=Janeiro} ⇒ {PERÍODO=Noite}	0,44
{LOCAL=Praça pública} ⇒ {ÁREA=AIS1}	0,42
{LOCAL=Ônibus} ⇒ {PERÍODO=Noite}	0,40

http://dx.doi.org/10.25286/repa.v3i3.974

{ÁREA=AIS3, LOCAL=Calçada} ⇒ {PERÍODO=Noite}	0,38
{ÁREA=AIS5, PERÍODO=Manhã} ⇒ {LOCAL=Via pública}	0,35

Para a base Furto, que possui 14.154 registros, foi escolhido um *minSup* de 0,01 e um *minConf* de 0,3, que geraram os itemsets candidatos na tabela 7.

Tabela 7 – Conjunto de itens candidatos da base Furto

k	Quantidade
1	40
2	156
3	9

Dos itens candidatos encontrados, podemos destacar que a área com maior incidência de furtos foi a AIS 1, sendo a maioria deles ocorridos em Vias públicas, e principalmente acontecidos no período da tarde, com Agosto como mês de maior ocorrências. Logo após a geração dos candidatos, foram geradas as melhores regras, onde algumas delas e os valores de sua Confiança estão descritas respectivamente na tabela 8.

Tabela 8 - Regras de associação geradas para a base Furto

Regra	Confiança
{LOCAL=Parada de ônibus} ⇒ {ÁREA=AIS1}	0,57
{LOCAL=Shopping Center} ⇒ {ÁREA=AIS3}	0,54
{LOCAL=Ônibus, PERÍODO=Tarde} ⇒ {ÁREA=AIS1}	0,44
{LOCAL=Calçada} ⇒ {PERÍODO=Tarde}	0,37
{LOCAL=Residência térrea} ⇒ {PERÍODO=Madrugada}	0,35

Para a base Furto de Veículo, que possui 1.098 registros, foi escolhido um minSup de 0,01 e um *minConf* de 0,3, que geraram os itemsets candidatos tabela 9.

Tabela 9 - Conjunto de itens candidatos da base Furto de Veículo

k	Quantidade
1	18
2	100
3	20

Dos itens candidatos encontrados, podemos destacar que a área com mais registros de furto de veículos é a AIS4, sendo a maioria dos crimes ocorridos no período da madrugada, e o mês de maior incidência foi Maio. Logo após a geração dos candidatos, foram geradas as melhores regras, onde algumas delas e os valores de sua Confiança estão descritas respectivamente na tabela 10.

Tabela 10 – Regras de associação geradas para a base Furto de Veículo

Regra	Confiança
{ÁREA=AIS2, MÊS=Janeiro} ⇒ {PERÍODO=Madrugada}	0,52
{PERÍODO=Tarde, MÊS=Janeiro} ⇒ {ÁREA=AIS3}	0,52
{ÁREA=AIS5, MÊS=Março} ⇒ {PERÍODO=Madrugada}	0,50
{ÁREA=AIS2, MÊS=Julho} ⇒ {PERÍODO=Tarde}	0,44
{PERÍODO=Madrugada, MÊS=Fevereiro} ⇒ {ÁREA=AIS4}	0,43

E por fim, para a base Roubo de Veículo, que possui 1.884 registros, foi escolhido um *minSup* de 0,01 e um *minConf* de 0,3, que geraram os itemsets candidatos tabela 11.

Tabela 11 – Conjunto de itens candidatos da base Roubo de Veículo

k	Quantidade
1	18
2	82
3	35

Dos itens candidatos encontrados, podemos destacar que a área com a maioria das ocorrências foi a AIS4, onde a maior parte dos registros no período da noite, e Março foi o mês com mais ocorrências de roubo de veículos. Logo após a geração dos candidatos, foram geradas as melhores regras, onde algumas delas e os valores de sua Confiança estão descritas na tabela 12.

Tabela 12 – Regras de associação geradas para a base Roubo de Veículo

Regra	Confiança
{ÁREA=AIS5, MÊS=Janeiro} ⇒ {PERÍODO=Noite}	0,69
{ÁREA=AIS2, MÊS=Setembro} ⇒ {PERÍODO=Noite}	0,51
{ÁREA=AIS2, MÊS=Março} ⇒ {PERÍODO=Noite}	0,48
{ÁREA=AIS3, MÊS=Maio} ⇒ {PERÍODO=Noite}	0,45
{ÁREA=AIS4, MÊS=Março} ⇒ {PERÍODO=Noite}	0,42

Em seguida foi realizada uma transformação nos dados na própria ferramenta *Weka*, para converter os atributos de nominais para binários, para aplicar o algoritmo FP-Growth. Foram utilizados os mesmos parâmetros de *minSup* e *minConf* para cada base usados no algoritmo Apriori. Após a aplicação do FP-Growth, foi observado que as regras extraídas e o valor de sua confiança, foram as mesmas do Apriori, corroborando a relevância das regras geradas.

5. Análise e discussão

Em comparativo com o estudo estatístico **[6]**, percebe-se que há uma confirmação entre os resultados encontrados entre as pesquisas, porém foram encontradas novas associações entre as variáveis existentes no banco de dados que oferecem mais detalhes em relação ao comportamento criminal. Informações como o dia da semana com mais registros, sendo a sexta-feira, a AIS 1 liderando as ocorrências referentes a Crimes Violentos contra o Patrimônio, além da AIS 4 com maior ocorrência de furtos e roubos de veículos, foram confirmadas pela geração de itens candidatos na aplicação do algoritmo Apriori.

Uma observação pode ser realizada considerando a natureza de crimes para exemplificar e comprovar a efetividade dos resultados dos estudos feitos com o mesmo banco de dados, onde a maior quantidade de registros é de Crimes Violentos ao Patrimônio, de forma que no estudo anterior foi identificado que 40,45% dos crimes dessa natureza acontecem à noite, e no presente estudo encontramos uma regra que mostra que {PERÍODO=Noite}⇒ {NATUREZA=CVP} com 0,66 de Confiança, fortalecendo a veracidade e importância dos fatos para análise.

Por possuir um nível alto de granularidade dos dados das ocorrências de crimes, os resultados encontrados no estudo de Viçosa sugerem poucas informações sobre os crimes, como na regra que indica que {DATA=Julho, SETOR=S8, TURNO=Noite}⇒ {GRUPO=B}, onde o atributo SETOR refere-se às subdivisões da cidade e o atributo GRUPO refere-se ao tipo de crime,onde B equivale a crimes contra pessoas. No estudo realizado em Niterói, como possui um nível mais baixo de granularidade dos dados das ocorrências de crimes, os resultados encontrados foram mais objetivos pela presença da localização geográfica do crime, como por exemplo, na regra que indica que {BAIRRO=Icaraí,PERÍODO=Manhã,ENTORNO=Banco}⇒{OCORRÊNCIA=Furto}, enquanto possuímos como localização apenas as AIS.

6. Considerações Finais

De posse dos resultados em forma de regras de associação, temos a representação de padrões de relacionamento frequentes entre itens das bases de dados estudadas, que permitem um aprofundamento em pesquisas e análises em relação a essas regras. Os resultados mostraram que as predições geradas tiveram um valor de Confiança médio maior que 57%, confirmando a relevância das informações, sendo possível através delas, que a Polícia atue de forma mais

http://dx.doi.org/10.25286/repa.v3i3.974

efetiva no combate e prevenção de crimes, pois eles saberão por exemplo quais dias ou períodos precisarão intensificar a frota de policiais nas ruas, assim como determinada AIS que apresenta maior ocorrência de crimes precisa de maior prioridade nas investigações.

Um fator importante para estudos posteriores, seria o fornecimento de novos registros referentes à todo Estado de Pernambuco, e não apenas da capital, permitindo um estudo geral da Segurança Pública do Estado. Para que regras mais robustas e com maiores valores de confiança sejam geradas, é necessário a inclusão de mais variáveis às bases de dados fornecidas pela Secretaria de Defesa Social de Pernambuco, que possam mostrar mais detalhes sobre as ocorrências registradas, possibilitando uma análise urbanística, geográfica e comportamental dos crimes ocorridos em Recife.

Com este estudo damos um passo para pesquisas maiores, para que com o uso da mineração de dados, possa haver uma contribuição no auxílio à tomada de decisões para um melhor planejamento das rondas policiais de acordo com a AIS e a natureza dos crimes, foco em investigações para combate a determinados crimes frequentes, e com a criação de novas estratégias de medidas preventivas. Com o uso desta pesquisa, espera-se cooperar com as autoridades e órgãos responsáveis pela Segurança Pública, para alcançarmos uma redução significativa na violência da cidade do Recife.

Agradecimentos

À Secretaria de Defesa Social de Pernambuco, pelo fornecimento das bases de dados utilizadas neste trabalho.

Referências

[1] FALCÃO, Marina. Violência em Pernambuco é maior e homicídios crescem mais que no Rio. **Valor Econômico**. Disponível em: <http://www.valor.com.br/brasil/5138178/violencia-em-pernambuco-e-maior-e-homicidios-crescem-mais-que-no-rio> Acesso: 21 abr. 2018.

[2] JC ONLINE. Recife é a 22ª cidade mais violenta do mundo, segundo ONG Mexicana. **Jornal do Comercio**, Recife, 7 mar. 2018.Disponível em: <http://jconline.ne10.uol.com.br/canal/cidades/policia/noticia/2018/03/07/recife-e-a-22-cidade-mais-violenta-do-mundo-segundo-ong-mexicana-330506.php> Acesso: 22 abr. 2018.

[3] MONTESANI, Beatriz. Por que a violência sobe de forma preocupante em Pernambuco, segundo especialista. **Nexo Jornal**, São Paulo, 18 abr. 2017. Disponível em: <https://www.nexojornal.com.br/entrevista/2017/04/18/Por-que-a-viol%C3%AAncia-sobe-de-forma-preocupante-em-Pernambuco-segundo-este-especialista> Acesso: 10 mai. 2018.

[4] CAVALCANTE, Diogo. Com novos policiais, Governo de PE espera reforçar delegacias. **Folha de Pernambuco**, Recife, 5 fev.2018. Disponível em: <https://www.folhape.com.br/noticias/noticias/cotidiano/2018/02/05/NWS,57935,70,449,NOTICIAS,2190-COM-NOVOS-POLICIAIS-GOVERNO-ESPERA-REFORCAR-DELEGACIAS.aspx> Acesso em: 21 abr. 2018.

[5] SANT'ANNA, Lourival. Como Medellín virou a cidade-modelo que está vencendo o crime. **Revista Exame**, 5 out. 2017. Disponível em: <https://exame.abril.com.br/revista-exame/menos-violenta-e-mais-prospera/> Acesso em: 10 mai. 2018.

[6] MELO, Carolina; RODRIGUES, Rodrigo Lins; CAVALCANTI, Bettina. **Análise de relações entre variáveis de ocorrências de crimes da cidade do Recife**. Universidade de Pernambuco, 2018.

[7] LOURENÇO, Vitor; MANN, Paulo; PAES, Aline; OLIVEIRA, Daniel de. SiAPP: Um sistema para análise de ocorrências de crimes baseado em aprendizado lógico-relacional. In: BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS, 12., 2016, Florianópolis. **Anais**... Florianópolis:SBSI, 2016. p.168-175.

[8] LAMAS, João Paulo Campolina. **Predição de crimes e otimização de ações de segurança pública para cidades de pequeno porte utilizando geotecnologias**. Tese de Doutorado. Programa de Pós-Graduação em Engenharia Civil, Universidade Federal de Viçosa. Viçosa, 2013. Disponível em: <http://www.locus.ufv.br/bitstream/handle/123456789/840/texto%20completo.pdf?sequence=1>

[9] BARANAUSKAS, José Augusto. **Regras de Associação**. Departamento de Física e Matemática, Universidade de São Paulo. Disponível em: <http://dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-Regras-Associacao.pdf>

[10] AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. Fast algorithms for mining association rules. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994, Santiago. **Proceedings**... Santiago, 1994. p. 487–499. Disponível em: <http://www.vldb.org/conf/1994/P487.PDF>

[11] HAN, Jiawei.; PEI, Jian; YIN, Yiwen. Mining Frequent Patterns Without Candidate Generation In: INTERNATIONAL CONFERENCE 2000 ON MANAGEMENT OF DATA, 2002, Dallas. **Proceedings**... Dallas: ACM SIGMOD, 2002. p. 1-12.

http://dx.doi.org/10.25286/repa.v3i3.974

Desenvolvimento de um Sistema de Apoio a Decisão para priorização de Pedidos de Desembolso no Estado de Pernambuco

Development of a decision system for prioritization of Disbursement Requests in the State of Pernambuco

Itallo Henrique de Santana Santos[1] orcid.org/0000-0003-1124-4150
Alexandre Magno Andrade Maciel[1] orcid.org/0000-0003-4348-9291

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: ihss@ecomp.poli.br

Resumo

A Secretaria da Controladoria Geral do Estado (SCGE) analisa mensalmente despesas geradas pelos diferentes órgãos do Estado de Pernambuco com o objetivo de garantir os pagamentos daquelas que são mais sensíveis. Mais de 1 bilhão de reais em despesas ficaram pendentes no exercício de 2016, demonstrando importância de priorização dos pagamentos. Nesse contexto o artigo apresenta o processo de desenvolvimento de um Sistema de Apoio a Decisão (SAD) para a Secretaria da Controladoria Geral do Estado de Pernambuco. O sistema proposto tem a capacidade de classificar as despesas públicas por meio de árvore de decisão auxiliando o trabalho de análise dos gestores responsáveis na priorização de pagamentos. No artigo é detalhado a caracterização do problema, a fundamentação teórica usada no trabalho, a aplicação da metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) e o sistema. A utilização de árvore de decisão na classificação das despesas teve como resultado uma acurácia de 99%, mostrando que o uso desse tipo de modelo atendeu satisfatoriamente na solução do problema encontrado.

Palavras-Chave: Sistemas de Apoio à Decisão; Sistemas de informações; Árvore de Decisão; Mineração de Dados;

Abstract

The Secretaria da ControladoriaGeral do Estado (SCGE) analyzes the monthly expenses generated by the different organs of the State of Pernambuco with the objective of pay sensitive expenses. Over one billion reais in expenses were outstanding in 2016, showing the importance of prioritizing payments . In this context the article presents the process of developing a Decision Support System (DSS) for the SCGE. The proposed system has the ability to classify public expenditures through a decision tree, assisting the analysis work of the responsible managers in the prioritization of payments. In the article is detailed the characterization of the problem, the theoretical basis used in the work, the application of CRISP-DM (Cross Industry Standard Process for Data Mining) and the system. The utilization decision tree in the classification of expenditures resulted in an accuracy of 99%, showing that the use of this type of model satisfactorily attended to the problem.

Key-words: Decision Support System*; Information Systems; Decision trees; Data Mining.*

 http://dx.doi.org/10.25286/repa.v3i3.975

1 Introdução

Nos últimos anos houve um grande crescimento do uso de tecnologias na area de gestão de organizações públicas ou privadas, elas hoje buscam cada vez mais garantir eficácia e eficiência nas suas atividades por meio de sistemas de informação. Laudon **[1]** define sistema de informação *(SI)* como um conjunto de componentes inter-relacionados que trabalham juntos na coleta, recuperação, processamento armazenamento e distribuição de informações com a finalidade de facilitar o planejamento, o controle, a coordenação, a análise e o processo decisório em organizações. Os SI transformam as informações tornando mais fácil o acesso e entendimento pelos seus usuários.

Nesse contexto, Góes **[2]** comenta:

> Governadores e prefeitos fazem inúmeras tentativas de 'revolucionar' na gestão dos seus Estados e Municípios, respectivamente, na ânsia de fazer de seu mandato um exemplo de gestão eficiente e com o maior atendimento possível ao interesse público. A 'nova' Administração Pública está em toda parte. Ela preocupa-se, essencialmente, com a gestão dos recursos financeiros e com o marketing público, cujos objetivos são a qualidade e a racionalidade, entendidas como sinônimos de economia e meios de redução de custos **[2]**.

Os pagamentos das Despesas Públicas realizadas pelo Estado de Pernambuco seguem, normalmente, uma organização baseada em Pedidos de Desembolsos (PDs) inseridos no sistema eletrônico de gestão financeira do Estado. Pedidos de Desembolsos, conforme o nome intuitivo, são pedidos ou solicitações de pagamentos das despesas realizadas pelas unidades gestoras do Estado de Pernambuco. Diante do cenário de crise econômica, sobretudo a crise na arrecadação dos impostos estaduais, nem todos os PDs foram pagas dentro do prazo previsto. Tal situação demanda uma preocupação maior no controle dos gastos por parte do órgão de controle interno, no caso a Controladoria Geral do Estado de Pernambuco.

Em 2016, o governo do Estado de Pernambuco movimentou mais de 28 bilhões de reais com PDs objetivando o pagamento das próprias despesas, conforme informações extraídas da base de dados dos PDs fornecida pela Secretaria da Fazenda.

Apesar do esforço no cumprimento das obrigações financeiras, ainda restou mais de 1 bilhão de reais de PDs com competências referentes ao exercício de 2016, e que, antes, passaram a integrar como novos PDs do exercício de 2017. Diante de tal situação, houve a necessidade de priorizar o controle do pagamento de certos temas de PDs mais urgentes, como por exemplo, PDs de medicamentos, alimentação, programas sociais do estado e entre outros. E ainda priorizar o fomento do controle interno de PDs em unidades gestoras consideradas sensíveis, como por exemplo, o FEAS (Fundo Estadual de Assistência Social).

A Secretaria da Fazenda de Pernambuco (SEFAZ) detém o controle da base dos Pedidos de Desembolsos. Portanto, a SEFAZ busca melhorar o processo de priorização de PDs, inclusive com categorização e a criação de alternativas de Business Intelligence. Em reconhecimento e apoio a SEFAZ, a Secretaria da Controladoria Geral do Estado de Pernambuco (SCGE) inicia um estudo de análise das PDs, tendo em vista as atribuições de Controle Interno deste órgão segundo o artigo 70 da Constituição Federal e Lei Complementar nº 119 de 26/06/2008 do Estado de Pernambuco.

Atualmente, a SCGE utiliza a ferramenta Qlikview para gerar dados referentes aos PDs mensais no formato .xls, fornecida pela SEFAZ, para extração das informações onde é considerado os aspectos de unidades gestoras, temas dos gastos e valores considerados sensíveis ou fora da normalidade, de forma que no final sejam gerados alertas, com base nos aspectos citados, para que sejam encaminhados às autoridades competentes do executivo estadual assim tomando as medidas corretivas necessárias. Entretanto, o processo para geração dos alertas é considerado um problema por ser considerado ainda muito manual.

Nesse contexto, o artigo tem como objetivo desenvolver um sistema de apoio a decisão, descrevendo conceitos, técnicas e ferramentas usadas em seu desenvolvimento. O sistema será capaz de classificar Pedidos de Desembolso utilizando modelos de árvore de decisão.

2 Fundamentação Teórica

2.1 Sistemas de Apoio à Decisão

Os sistemas de apoio a decisão SAD são uma classificação dos SI. Esses sistemas tem como foco o suporte às decisões específicas de um processo, tem como objetivo melhorar a capacidade de compreensão da informação por meio de modelos e simulações permitindo o gestor ampliar conhecimento e esclarecimento de um problema. Vale salientar que os SADs estão limitados ao suporte e não na automatização das decisões.

Na literatura encontramos várias definições para os SADs. Sprague e Watson **[3]** definem SAD como sistemas computacionais que ajudam os responsáveis pela tomada de decisões a enfrentar problemas não-estruturais através da interação direta com modelos de dados e análises. Segundo Bidgoli **[4]**, SAD é um sistema de informação baseado em computador, que consiste de *hardware* e *software* e elemento humano, para assistir qualquer decisão em qualquer nível, e que enfatiza tarefas não-estruturadas ou semi-estruturadas. Para Courtney (2001), os SAD são sistemas de gerenciamento de decisões interativos, baseados em computador, que ajudam os decisores a utilizar dados e modelos para resolver problemas não-estruturados.

Barbosa **[5]** complementa que conceito de SAD ainda não está completamente livre de divergências entre os pesquisadores, usuários e desenvolvedores, não havendo, portanto, nenhuma definição exata. Entretanto, podem ser destacados alguns aspectos em comum, tais como: conseguem juntar o pensamento humano e a informação automatizada; abrangem todas as fases do processo de decisão; objetivam ajudar e não substituir o decisor; utilizam modelos para análise de situações de tomada de decisão, possibilitar experiências com diferentes estratégias sob diferentes configurações facilitadas pela capacidade de modelagem permitem, e apresentam uma interface amigável **[5]**.

2.2 Tipos de Sistema de Apoio à Decisão

Alter **[6]**, com referência de dezenas de sistemas de informação, propôs um conjunto de operações genéricas que podem ser executadas por SAD, as quais se classificam em sistemas orientados a dados e sistemas orientados a modelos.

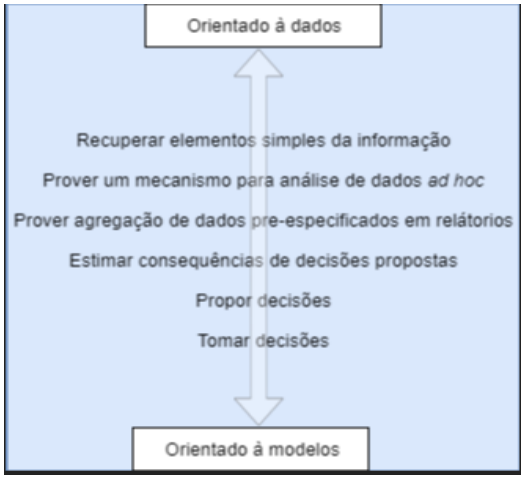


Figura 1 - Conjunto de operações genéricas
Fonte: adaptado de **[6]**.

Além disso, Alter **[6]** com base nas operações genéricas classificou sete tipos diferentes de SAD. Barbosa **[5]** resume essas operações como:

- Sistemas de desenho de arquivos: representam basicamente uma versão automatizada dos sistemas de arquivamento manual, proporcionam maior segurança e rapidez na recuperação de informações.
- Sistemas de análise de dados: facilitam a análise a partir de arquivos com dados atuais ou históricos, geram indicadores.
- Sistemas de análise de informações: provêm acesso a uma série de dados orientados a decisão e pequenos modelos para prover informação gerencial, possibilitando a análise atavés do uso de dados internos.
- Modelos de contas: calculam a consequência de ações planejadas sobre a base de definições de contas e estimam variações das entradas nas fórmulas das contas.

http://dx.doi.org/10.25286/repa.v3i3.975

- Modelos de representação: incluem modelos de simulação, estimam consequência de ações sobre a base de modelos, tais como probrabilidades de ocorrências.
- Modelos de otimização: oferecem linhas de ação para uma solução ótima, consideram restrições; usados para decisões repetitivas que podem ser descritas matematicamente.
- Modelos de sugestão: consideram uma sugestão específica para uma decisão estruturada e repetitiva, substituindo procedimento menos eficientes.

2.3 Árvores de Decisão

Árvores de Decisão em Mineração de Dados (Data Mining) faz parte de métodos que envolvem classificação dos dados. Podem ser usadas como tecnologia de indução de regras **[7]**. Crepaldi et al. **[7]** complementa que:

> A vantagem principal das Árvores de Decisão é a tomada de decisões levando em consideração os atributos mais relevantes, além de compreensíveis para a maioria das pessoas. Ao escolher e apresentar os atributos em ordem de importância, as Árvores de Decisão permitem aos usuários conhecer quais fatores mais influenciam os seus trabalhos.

Segundo Garcia **[8]**, as Árvores de Decisão consistem de: nodos (nós), que representam os atributos, e de arcos (ramos), provenientes desses nodos e que recebem os valores possíveis para esses atributos (cada ramo descendente corresponde a um possível valor desse atributo). Nas árvores existem nodos folha (folha da árvore), que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está associada a uma classe **[8]**.

2.4 Representação de uma Árvore de Decisão

A Figura 2 apresenta um exemplo de árvore de decisão.

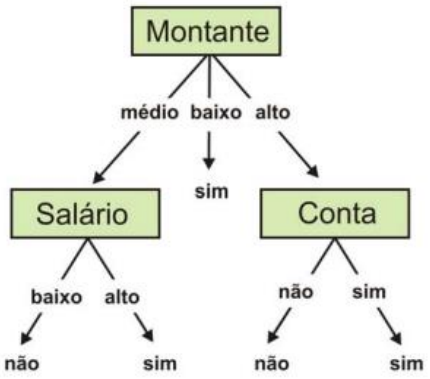


Figura 2 - Representação de uma árvore de decisão. Fonte: adaptado de **[9]**.

Na Figura 2, são analisados objetos que descrevem o recebimento de um empréstimo por uma pessoa. É considerado o valor do empréstimo (Montante), se ele é médio, baixo ou alto. Alguns objetos são exemplos positivos de uma classe sim, ou seja, os requisitos exigidos a uma pessoa, por um banco, são satisfatórios à concessão de um empréstimo, e outros são negativos, onde os requisitos exigidos não são satisfatórios à concessão de um empréstimo **[9]**.

Com a árvore de decisão em mãos é possível gerar regras. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação **[10]**. Dois exemplos de regras são definidos a seguir baseado na figura (árvore).

Se montante = médio e salário = baixo
então classe = não
Se montante = médio e salário = alto
então classe = sim

2.5 Técnicas de Construção de Árvores de Decisão

A construção de uma Árvore de Decisão é dada pelo particionamento dos dados, onde a verificação de um atributo divide o conjunto de dados, criando um ramo para cada valor deste atributo. Para cada ramo é criado um nodo e um novo atributo é analisado para fazer o particionamento do subconjunto, esse processo é repetido sucessivamente, e tem como finalidade, separar os dados em classes onde diferentes classes tendem a fazer parte de um subconjunto diferente dos dados **[10]**.

A Figura 3 apresenta algoritmo genérico para construir Árvores de Decisão, em que S

representa o conjunto de exemplos aplicado a árvores de decisão, primeiramente *S* contém todo conjunto de dados treino.

Se todos os exemplos no atual conjunto de exemplos S satisfazem um critério de parada;

Então
Cria um nodo folha com algum nome da classe e para;

Senão

Seleciona um atributo A para ser utilizado como um atributo de particionamento e cria um
nodo com o nome do atributo de particionamento.

Escolhe um teste sobre os valores de A, com resultados mutualmente exclusivos e coletivamente exaustivos $R_1, \dots R_k$, *e cria um ramo, a partir do nodo recentemente criado, para cada teste;*

Particiona S nos subconjuntos $S_1,\dots S_k$, *tal que cada* S_i, $i = 1..,k$, *contenha todos os exemplos em S com resultado* R_i, *do teste escolhido;*

Aplica esse algoritmo recursivamente para cada subconjunto S_i, $i=1,\dots k$;

Fim_senão

fim_se

Figura 3 - Algoritmo genérico para gerar árvores de decisão. Fonte: adaptado de **[10]**.

2.6 Técnica de Seleção de Atributos

A seleção de atributos na geração de uma árvore de decisão consiste em determinar qual o atributo melhor se enquadra no nodo em análise, observando o aspecto de como esse atributo define o conjunto de dados. Logo é possível determinar a melhor partição a ser realizada **[10]**. A utilização do índice de Gini é amplamente utilizado nesse tipo de problema.

2.7 Índice de Gini

O índice de Gini, criado pelo matemático italiano Conrado Gini, mede o grau de heterogeneidade dos dados.

O índice de Gini em um determinado nó é:

$$Indice\ Gini = 1 - \sum_{i=1}^{c} p_i^2$$

Onde p_i é a frequência relativa de cada classe em cada nó e c é o número de classes. Se o índice de Gini for igual a zero, significa que o nó é puro, ao contrário, quando se aproxima de um o nó é considerado impuro, logo é possível particionar os dados de uma melhor forma com essa informação.

3 Metodologia

A metodologia utilizada para o desenvolvimento desse projeto é o CRISP-DM (Cross Industry Standard Process for Data Mining). O CRISP-DM tem com objetivo a padronização de conceitos e técnicas na busca de informações para a tomada de decisões(SILVA, 2002). O CRISP-DM é dividido em seis fases de maneira cíclica como mostra a Figura 4.

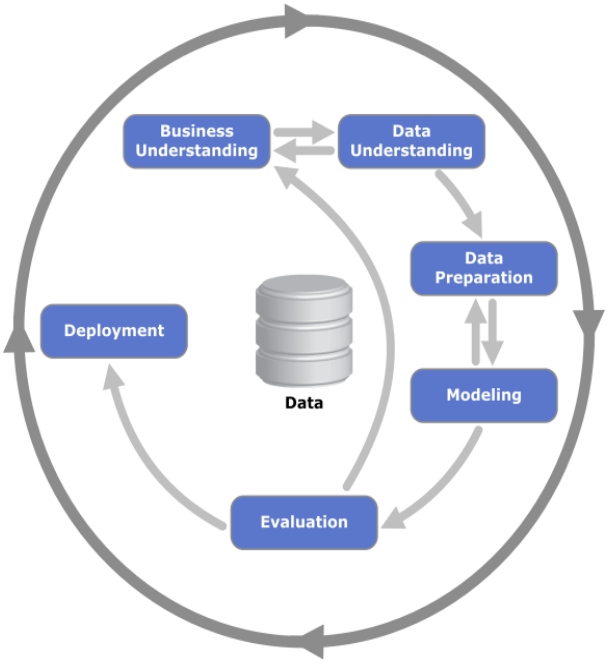


Figura 4 - Fluxo do CRISP-DM

3.1 Entendimento do Negócio

Nesta fase adquiriu-se entendimento de todo processo de análise dos Pedidos de Desembolso feito pela Secretaria da Controladoria Geral do Estado de Pernambuco. A Figura 5 o fluxo do processo de análise dos Pedidos.

http://dx.doi.org/10.25286/repa.v3i3.975

Figura 5 - Processo de classificação de Pedidos de Desembolso feito pela SCGE.

ASecretaria da Controladoria da Geral do Estado mensalmente extrai os dados de Pedidos de Desembolso através da plataforma de BI Qlik View, onde um arquivo .xls é gerado, cada linha desse arquivo representa um Pedido de Desembolso que precisará ser analisado e classificado.

Na etapa de transformações alguns campos do Pedido de Desembolso são verificados e/ou processados com o objetivo de transformar os dados de forma que caiba a análise de acordo com as regras de priorização definidas pela gestão. Através da ferramenta Excel é feito os processos a seguir.

Verificação da Unidade Gestora: Para cada Pedido é feita a verificação se a Unidade Gestora daquele Pedido de Desembolso é prioridade.

Verificação do Tema: Para cada Pedido é feita a verificação se o Tema daquele Pedido de Desembolso é prioridade.

Verificação do Valor: Para cada Pedido é feita a verificação se o valor em R$ ultrapassa uma quantia específica. Pedido de Desembolso com valor muito alto são priorizados.

Extração da Data de Competência: Cada Pedido de Desembolso contém um campo textual que o descreve, nessa descrição pode está contida a data que a despesa ocorreu. Essa data é importante, já que é possível saber quantos dias já se passaram desde que ocorreu a despesa. Se não for possível extrair essa data, a data de competência deverá ser a data que o Pedido de Desembolso foi adicionado ao sistema.

Inclusão da data Prazo Limite: Para cada Pedido de Desembolso é adicionado uma data que defina o prazo limite que o Pedido de Desembolso poderá atingir antes que ele precise ser priorizado.

Inclusão dadataGeração:A data de Geração é a data que o Pedido de Desembolso está sendo analisado, a partir dessa data é possível saber se o Pedido de Desembolso passou do Prazo Limite. Pedidos de Desembolso com data de Geração depois da data Prazo Limite são priorizados.

Na etapa de classificação é feita a análise das regras de priorização definida pela gestão. Para este projeto os Pedidos de Desembolso foram classificados pela Secretaria da Controladoria Geral do Estado em:

Alerta Tempo: Pedidos de Desembolso acima do prazo limite.

Alerta Unidade Gestora: Pedidos de Desembolso acima do prazo limite e de Unidades Gestoras consideradas prioritárias.

Alerta Tema: Pedidos de Desembolso acima do prazo limite e relacionados a Temas considerados prioritários.

Alerta Valor: Pedidos de Desembolso acima do prazo limite e com valor acima de uma quantia espécifica.

Aguardando Ordem Bancária: Pedidos de Desembolso que na análise em questão não são priorizados.

Com todo processo em mãos houve o entendimento que um método de classificação utilizando treinamento supervisionado com árvores de decisão tornaria o processo de classificação mais eficiente, pois reduziria a interferência humana no processo de classificação utilizando o excel que é uma ferramenta não ideal para fazer esse tipo de tarefa, além disso, padronizaria a classificação por meio de modelos fixos, é de se mencionar que com o modelo gerado o processo de classificação é feita de forma algorítmica.

3.2 Entendimento dos dados

3.2.1 Base de dados

A base de dados utilizada pelo SAD proposto é a base de dados dos Pedidos de Desembolsos fornecida pela Secretaria da Fazenda. Tal base contém as informações de todos os Pedidos de Desembolsos das unidades gestoras do Estado. A base existente possui 29333 registros com 12 atributos relativos mês de agosto do ano de 2016.

3.2.2Dicionário dos dados

A Tabela 1 apresenta o dicionário dos dados.

Tabela 1 - Dicionário de dados.

Campo	Descrição	Tipo
UO SIGLA	Refere-se à unidade gestora queexecuta a despesa.	Char
Fonte Sintética	Refere-se à fonte de recurso público que será utilizada para pagamento da despesa.	Num
NU_PD_ EXP	Data de inclusão do Pedido deDesembolso (PD) no sistema da Secretaria da fazenda	Char
CGS	Referem-se aos temas de gasto dadespesa	Char
DT_INCLUSAO_ DOCUMENTO_E XP	Data da inclusão do Pedido de Desembolso no sistema	DATE
PD RazãoSocial	Nome da empresa fornecedora do estado	CHAR
OBS_EXP	Campo aberto de observação. Contém detalhes sobre a despesa, além de poder incluir a data de competência.	CHAR
NU ORDEM BANCARIA EXP	Número da Ordem Bancária Expedida. Fato posterior ao Pedido de Desembolso que autoriza o pagamento da despesa pelo banco.	CHAR
DH_COMPETÊN CIA_AJUST	Data extraída do campo OBS_EXP, é a data de competência ajustada	DATE
Valor PD	Custo da despesa	NUM
SITUACAO	Classificação atribuída ao Pedido de Desembolso pelos analistas de acordo com as regras	CHAR

3.3 Preparação dos Dados

3.3.1Pentaho Data Integration

É uma ferramenta de ETL open source que permite manipular dados. Utilizada amplamente em aplicações de *Analitycs*.

Através do PDI (*Pentaho Data Integration*) foram realizados os processos descritos a seguir:

Extração:

- Extração da base de treino
- Extração de dados de Pedidos de Desembolso que serão classificados

Transformação:

- Seleção dos dados que serão utilizados pelo modelo.
- Criação de novas colunas que serão utilizadas pelo modelo.
- Processo de limpeza (conversão de tipos).

Carga:

- Base de treino transformada para classificador dos Pedidos de Desembolso
- Dados de Pedidos de Desembolso transformadas para serem classificados
- Carga da base de treino para classificador dos Pedidos de Desembolso.
- Transformações de dados na base de treino com o objetivo de alimentar o algoritmo que gera árvore de decisão.

Nas figuras 6 e 7 encontram-se exemplos de*ETL* usado no sistema.

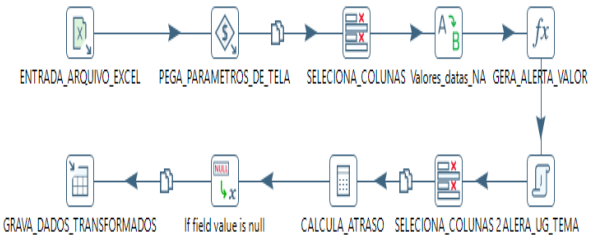


Figura 6 - *Transformation* que carrega dados a serem utilizados no treinamento do modelo.

Figura 7 - *Job*resposável por carregar dados transformados e gerar modelo de classificação.

3.3.2 Modelagem

Nessa fase, foi utilizado a biblioteca rpart (*Recursive partitioning for classification, regression trees*) **[11]**, disponível no ambiente de desenvolvimento integrado R **[12]**. O rpart implementa ideias encontradas no CART

http://dx.doi.org/10.25286/repa.v3i3.975

(*Classification and Regression Trees*) de Reiman, Friedman, Olshen e Stone **[13]**.

O rpart dá a possibilidade de criar modelos de classificação e regressão utilizando árvores binárias. O rpartutiliza a técnica TDIDT (*Top-Down Induction of Decision Tree*) **[14]**. O critério de seleção de atributos de particionamento escolhido foi o índice de gini, disponibilizado implementado pelo rpart.

O processo de ETL para etapa do treinamento e validação do modelo preparou os dados da seguinte forma:

Seleção da variável alvo: A coluna *SITUACAO* é o alvo do modelo, essa coluna descreve qual a situação do Pedido de Desembolso já classificado pela Secretaria da Controladoria da Geral do Estado de Pernambuco.

Criação e seleção de colunas preditoras: Para alimentação do modelo foram criadas colunas baseada nas regras de classificação de Pedidos de Desembolso são elas:

- UG PRIORIDADE: Essa coluna informa se a unidade gestora de um Pedido de Desembolso em questão é prioritária.
- TEMA PRIORIDADE: Essa coluna informa se o tema de um Pedido de Desembolso em questão é prioritário.
- VALOR PRIORIDADE: Essa coluna informa se o valor do Pedido de Desembolso em questão está acima de uma certa quantia.
- ATRASO: Essa coluna informa quantos dias o Pedido de Desembolso está atrasado.

Unidade Gestora Prioridade	Tema Prioridade	Valor Prioridade	Atraso	Situação
SIM	NÃO	NÃO	60	AGUARDANDO OB
SIM	NÃO	NÃO	100	ALERTA UNIDADE GESTORA
NÃO	NÃO	SIM	130	ALERTA VALOR

Figura 8 - Exemplo de variáveis de entrada para treino do modelo.

3.4 Avaliação

Para o processo de validação foi selecionado a base para treino e validação do modelo de agosto de 2016 com 27 mil onde cada registroestá classificado com uma situação.

Foram selecionados de forma aleatória 80% dos dados para treinamento, e 20% para validação. O balanceamento da base de treino por meio do método*oversampling* não aumentou significadamente o grau de informação das classes minoritárias. A árvore gerada pelo rpart é mostrada na Figura 9. Para a avaliação do resultado foi utilizado a matriz de confusão, conforme mostra a Figura 10.

Como resultado tivemos a classificação correta 99% dos dados de validação. A especificidade que representa a proporção de verdadeiros negativos e a sensibilidade que representa a proporção de verdadeiros positivos para cada classe obtiveram um resultado acima dos 99%.

O resultado mostrou que para as regras de priorização de Pedidos de Desembolso utilizada na base de agosto de 2016 a árvore binária conseguiu se adequar quase perfeitamente para cada tipo de situação.

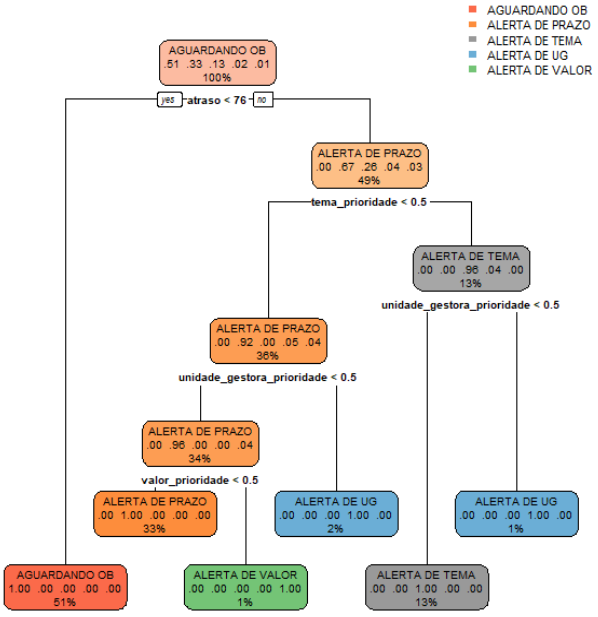


Figura 9 - Árvore de decisão gerada para classificação de Pedidos de Desembolso baseado na regra utilizada em Agosto de 2016.

	AGUARDANDO OB	ALERTA DE PRAZO	ALERTA DE TEMA	ALERTA DE UG	ALERTA DE VALOR
AGUARDANDO OB	3016	0	0	0	0
ALERTA DE PRAZO	0	1900	0	0	0
ALERTA DE TEMA	0	0	763	0	0
ALERTA DE UG	0	0	0	122	0
ALERTA DE VALOR	0	2	1	0	83

Figura 10 - Matriz de confusão.

3.5 Implantação

Para o desenvolvimento da aplicação foi utilizado a plataforma *Pentaho* que é uma plataforma de BI (*Business Intelligence*), o Pentaho dá a capacidade de desenvolver *dashboards* totalmente integráveis aos servidores Linux e Windows, facilitando o *deploy* da aplicação. Além disso ele prover segurança ao uso da aplicação, ideal para o sistema proposto.

O *pentaho* possibilita o desenvolvimento de dashboards através do Community Dashboard Editor (CDE). O CDE permite o rápido desenvolvimento de dashboards através da simplificação do processo de criação, design e integração de dados.

4 Ferramenta

Os objetivos do sistema de apoio a decisão proposto são descritos a seguir:

- Disponibilidade à usuários simultâneos.
- Capacidade de segurança de acesso.
- Automatização do processo de priorização.
- Adaptabilidade a novas regras de priorização.

4.1 Telas

O sistema tem duas telas, a primeira tela chamada de Tela Treino, é responsável, por gerar modelos de classificação. O usuário definirá a base de treino para gerar um modelo, assim como também irá definir quais são as regras de priorização que foram utilizadas para essa base de treino. Um nome para o modelo será definido. Este modelo será salvo no sistema para ser usado em uma classificação. A Figura 11 mostra a tela de treino.

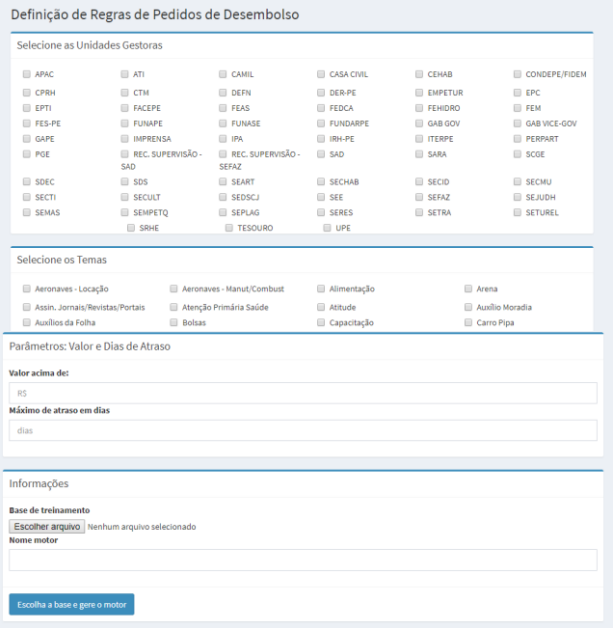


Figura 11 - Tela Treino.

A segunda tela chamada de Tela de Classificação será responsável por utilizar um modelo já salvo no sistema para classificar Pedidos de Desembolso. O usuário selecionará um modelo, assim como um arquivo de Pedidos de Desembolso para ser classificado em situações de alerta. A Figura 12 mostra a tela de classificação.

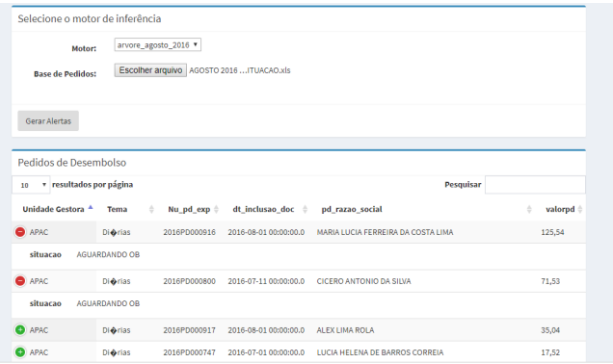


Figura 12 - Tela Classificação.

5 Conclusões

O seguinte trabalho alcançou o objetivo de criar um sistema que auxilie o processo decisório relacionado a um assunto importante que impacta diretamente todos na sociedadeque são as despesas públicas. A aplicação de métodos, técnicas e ferramentas proporcionou de forma satisfátoria a automatização do processo de

http://dx.doi.org/10.25286/repa.v3i3.975

classificação dos Pedidos de Desembolso do Estado de Pernambuco por meio de árvores de decisão.

Para o problema proposto de classificaçãoo uso de árvores de decisão se adequou bem, com um resultado satisfatório baseado na acurácia apresentada de 99%. O sistema desenvolvido dá a possibilidade da aplicação real de classificação dos Pedidos de Desembolso podendo ajudar ao processo de tomada de decisão e análise de dados pela Secretaria da Fazenda do Estado de Pernambuco (SEFAZ) e pela Secretaria Controladoria da Geral do Estado de Pernambuco (SCGE).

Referências

[1] LAUDON, Kenneth C.; LAUDON, Jane Price. **Sistemas de informação com internet**. Rio de Janeiro: LTC, 1999.

[2] GÓES, B. C. Administração Pública sob o Príncipio da Eficiência. Artigo, **Escola da Magistratura do Estado do Rio de Janeiro**, Rio de Janeiro, 2014.

[3] SPRAGUE, Jr.; WATSON, H. **Decision support systems: putting theory into practice**. USA: Prentice-Hall, 1989.

[4] BIDGOLI, H. **Decision Support System - Principles and Practice**. New York: West Publishing Company, 1989.

[5] BARBOSA, G. R. **Sistemas de Apoio A Decisão sob enfoque dos profissionais de tecnologia da informação e decisores**. Dissertação de Mestrado. Universidade Federal de Pernambuco, Recife, 2003.

[6] ALTER, S. **Decision Support Systems: current practice and continuing challenges**. California: Addison-Wesley Publishing Company, 1980.

[7] CREPALDI G. Paola; Avila R. N. P; Oliveira J. P. M.; Rodrigues P. R.; Martins L. R. Um estudo sobre a árvore de decisão e sua importância nahabilidade de aprendizado. Artigo, Faculdade Integrada Inesul Londrina, Londrina.

[8] GARCIA, S.C.. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. In: SEMANA ACADÊMICA, 2000, Rio Grande do Sul. **Anais**... Rio Grande do Sul:UFRGS, 2000.

[9] PETERMANN, R.J. **Modelo de Mineração de Dados para Classificação de Clientes em Telecomunicações**. Dissertação de Mestrado. Pontíficia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2006.

[10] STAIR, R. **Princípios de sistemas de informação**: uma abordagem gerencial, 2. Ed. Rio de Janeiro: LTC, 1998.

[11] THERNEAU, T.; ATKINSON, B.; RIPLEY, B. Rpart: Recursive Partitioning and Regression Trees, 2012. R package. Disponível em: <http://CRAN.R-project.org/package=rpart>.

[12] R Core Team, R: A Language and Environment for Statiscal Computing Vienna, Austria, 2012. Disponível em: <http://R-project.org/>.

[13] SILVA, E. M. **Descoberta de Conhecimento com o uso de Text Mining**: Cruzando o Abismo de Moore. Dissertação de Mestrado. Universidade Católica de Brasília, Brasília - DF. 2002.

[14] BREIMAN, L. et al. **Classification and Regression Trees**. Chapman and Hall, 1984.

[15] FRIZARINI. C. **Algoritmo para Indução de Árvores de Classificação para Dados Desbalenceados**. Dissertação de Mestrado. Universidade São Paulo, São Paulo, 2013.

Desenvolvimento de um Framework Integrador de Mineração de Dados Educacionais

Development of an Integrated Framework for Educational Data Mining

Italo Yoshito Fujisawa[1] orcid.org/0000-0002-5026-1722
Alexandre Magno de Andrade Maciel[1] orcid.org/0000-0003-4348-9291

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: iyf@ecomp.poli.br

Resumo

Com o avanço da tecnologia nos últimos tempos, a área de educação a distância (EAD) vem ganhando espaço na sociedade moderna. Professores e alunos não precisam, necessariamente, estar fisicamente próximos no processo de aprendizagem, *utilizando sistemas gerenciadores de cursos que são categorizados como "Ambiente Virtual de Aprendizagem" (AVA) para realizar a comunicação* que, também geram grandes volumes de dados. Os dados gerados não são muito exploradas devido a grande dificuldade existente em realizá-lo pela sua diversidade de informações e seus formatos. Em meio a esse problema, o presente trabalho tem como objetivo desenvolver um *framework* integrador de mineração de dados, inicialmente testado com o AVA chamado Moodle, para permitirem que desenvolvedores construam com facilidade aplicações para as pessoas realizarem análises de maneira facilitada utilizando técnicas de mineração de dados no intuito de obter informações mais aprimoradas. Para Atingir tais objetivos foram construídas: uma aplicação *Web* para demonstração do *framework* integrador utilizando o *framework* chamado React criada pela Facebook, e a proposta deste trabalho foi construída em forma de *WebService* utilizando a linguagem *python*, juntamente com a biblioteca *Flask.*

Palavras-Chave*:* Ambiente Virtual de Aprendizagem. Mineração de Dados; *Framework*; Web Service. *Moodle*;

Abstract

With the advancement of technology in recent times, the field of distance learning has been gaining space in modern society. Teachers and students do not necessarily need to be physically close in the learning process, using course management systems that are categorized as "Virtual Learning Environment" (AVA) to perform communication that also generate large volumes of data. The data generated is not much explored due to the great difficulty in performing it for its diversity of information and formats. In the midst of this problem, the present work aims to develop a data mining integrator framework, initially tested with AVA called Moodle, to allow developers to easily build applications for people to perform analyzes in an easy way using datamining techniques. To achieve these objectives, a Web application for demonstration of the integrator framework was created using the framework called React created by Facebook, and the proposal of this work was built as a Web Service using the python language, together with the Flask library.

Key-words:*Virtual Learning Enviroment; Data Mining; Framework; Web Service; Moodle; React; Python; Flask;*

 http://dx.doi.org/10.25286/repa.v3i3.976

1 Introdução

No processo de ensino da sociedade moderna observa se que cada vez mais a presença da utilização de ferramentas eletrônicas de apoio ao ensino. Isso se dá pelo motivo das informações estarem presentes eletronicamente e de fácil acesso para todos e de maneira instantânea através de dispositivos eletrônicos conectados à *internet*, o mesmo sendo considerado como principal fonte de informação e estudo da atualidade. **[1]** E com a grande disseminação da informação pela *internet*, fez com que ocorresse grandes mudanças no setor da educação; o aprimoramento e formalização da educação a distância (EAD) **[2]**.

Com o surgimento dessa modalidade de ensino e com o avanço tecnológico, foram desenvolvidas sistemas categorizados como "Ambientes Virtuais de Aprendizagem" (AVA) para distribuição e gerenciamento de disciplinas ministradas pelas instituições de ensino, com o objetivo de aproveitar a facilidade que o meio eletrônico e a *internet* permite de ter acesso aos dados em diversos tipos de dispositivos em diversos formatos **[3]**.

O Moodle é um exemplo de uma ferramenta dessa categoria. Ela consiste em ser uma plataforma de ensino *open-source*, ou seja, é disponibilizado gratuitamente sendo de fácil customização **[4]**. Além disso, a plataforma possui uma grande maturidade, ampla comunidade de desenvolvedores, grande quantidade de documentação.

O Brasil, segundo o próprio Moodle **[15]**, é o quarto país que possui mais sites registrados que estão em funcionamento no momento, usados principalmente para o gerenciamento de cursos e disciplinas em universidades.

Esses tipos de sistemas estão gerando grandes volumes de dados que são poucos analisados e explorados pela sua grande dificuldade existente para realizar esse processo. O motivo disso está no fato por dados possuírem diversos formatos sem nenhuma padronização. Assim permitindo que possíveis informações úteis, que poderiam ser usados para o melhoramento dos cursos prestados pela instituição educacional, sejam despercebidas **[6]**.

Uma forma de contornar esse problema seria utilizar ferramentas que permitissem a automação do processo de análise que resume ou agrega os dados de maneira a extrair informações relevantes de forma visualmente notável para auxiliar no acompanhamento ou até na tomada de decisão do instrutor. Decisões estas que poderiam vir a ser utilizada na melhoria da qualidade dos cursos.

Gonçalves **[10]** propôs uma arquitetura de *framework* de mineração de dados integrado ao Moodle em forma de aplicação Web, com o objetivo de permitir que desenvolvedores construam ferramentas de visualização e análise de dados que usem técnicas de mineração de dados de maneira facilitada.

A estrutura proposta no trabalho de Gonçalves utiliza-se da linguagem R em conjunto com a biblioteca *Shiny*, aproveitando-se da facilidade de criação e utilização de técnicas de mineração de dados em que a tecnologia permite ter. Cada visualização ou análise consiste em ser uma aplicação Web construída utilizando o *Shiny* que é integrado ao Moodle em forma de extensão, que muitas vezes são popularmente chamados de *plugins*. Além disso, a arquitetura apresentou ser pouco flexível por ser especificamente integrado ao Moodle, não permitindo integrar em outros tipos de sistemas.

Dito isto, este trabalho tem como objetivo propor um *framework* de mineração de dados no contexto educacional que seja flexível e que possa permitir ser usado em diversos outros sistemas, independente da tecnologia que ela esteja usando. Permitindo que desenvolvedores usem em seus sistemas AVA, para analisar e explorar os dados gerados por elas sem dificuldades.

2 Fundamentação Teórica

2.1 Mineração de Dados Educacional

O uso de sistemas na forma de aplicações Web vem ganhando muita popularidade nos últimos tempos que como consequência causou diversas mudanças em diversos setores que vem utilizando esse tipo de tecnologia. como por exemplo: Nos setores da Educação, Saúde, Negócios e etc. Na educação, é muito comum observar que, está cada vez mais presente, instituições que se utilizam de plataformas de ambientes virtuais de

aprendizagem (AVA) para ministrar cursos, presenciais ou semipresenciais **[15]**.

Esses sistemas geram diversos tipos de dados oriundas das interações entre os elementos envolvidos que são alunos e professores. Informações como: quantidades de vezes que o aluno entrou, atividades submetidas por elas, desempenho de alunos em disciplinas, trocas de mensagens entre alunos e professores e etc, são todos armazenados no banco de dados do sistema.

Devido a sua grande quantidade de dados, surge novas possibilidades e ambientes que poderiam ser analisados por intermédio do uso de técnicas de mineração de dados no contexto educacional que é denominada Mineração de dados Educacional **[6]**.

Ela consiste em ser um sub-ramo da mineração de dados que está preocupada em aplicar suas técnicas para detectar padrões em grandes conjuntos de dados educacionais, que de outra forma seriam difíceis ou impossíveis de ser percebidas ou analisadas devido ao enorme volume de dados e variedade de formatos.

2.2 Framework

De acordo com Sommerville **[8]**, a engenharia de software trata-se de um ramo da ciênciada computação ou engenharia da computação que é voltada à especificação, desenvolvimento e manutenção de sistemas de software em relação a todos os aspectos de produção.

Atualmente, é essencial a utilização de técnicas e conceitos dessa área no desenvolvimento de qualquer projeto de software para seguir boas práticas no intuito de evitar uma possível inviabilidade na manutenção, consequentemente, na descontinuidade de um projeto. A reutilização de componentes de software é uma das ideias propostas pela área que ajuda a mitigar esse tipo de problema.

Na literatura, existem diversas definições de *framework* e que podem ser diferentes dependendo do contexto em que esteja sendo aplicada:

- Segundo Fayad **[12]**, *framework* consiste em um conjunto de classes abstratas para a solução de uma família de problemas.
- Em Mattson **[13]**, é uma estrutura pré-definida com o objetivo de atingir a máxima reutilização de componentes de software.
- Para Johnson **[11]**, framework consiste em um conjunto de objetos que colaboram com o objetivo de atender a um conjunto de responsabilidades para uma aplicação específica ou um domínio de aplicação.

As definições de Fayad e Mattson estão no contexto de desenvolvimento de software. A definição de Johnson está bem genérica, podendo ser aplicada tanto no sentido de desenvolvimento como na área de gestão de projetos, onde cada componente do *framework* consiste em atividades ou ações bem definidas para atingir um determinado objetivo etapa a etapa.

Para o presente artigo, a definição de Mattson se aproximou mais do posicionamento do trabalho, portanto consideramos ela como definição do framework.

2.3 Web Service

Segundo o *Open soft* **[7]**, *Web Service* trata-se de uma solução ou aplicação que oferecem serviços ou informações que são consumidos por outras. As informações são solicitadas por requisições como POST, GET, UPDATE, DELETE e etc. Ela pode ser implementada usando diversas linguagens de programação existentes, no entanto há uma necessidade de padronização no formato de transmissão de informações para que a comunicação seja efetiva com diversos sistemas.

Então foram criadas protocolos para fazer esse intermédio entre o WS e as aplicações que irão consumir o serviço, como por exemplo: REST e SOAP. Apesar de existir outros protocolos, esses dois formatos são os mais conhecidos hoje na literatura.

O protocolo REST padroniza o formato dos dados enviados em JSON, que é uma estrutura bem similar a objetos da linguagem *javascript* ou dicionários em outras linguagens. No protocolo SOAP, as informações são passadas no formato de XML.

http://dx.doi.org/10.25286/repa.v3i3.976

3. *Framework* Proposto

No presente trabalho, foram construídas duas aplicações distintas: uma em forma de *Web Service* (WS), que consiste em ser uma aplicação Web que provê recursos a diversas outras aplicações através de requisições HTTP, não possuindo uma interface como em outras aplicações Web. A outra aplicação foi construída como uma aplicação Web comum que será consumidora dos recursos do WS.

Na aplicação WS, é onde está presente toda a estrutura do framework proposto, ela foi construída nesse tipo de tecnologia pelo motivo de permitir que ela seja usada por vários tipos de AVA's, independente da tecnologia usada na sua construção, que é a principal ideia por trás do WS.

A aplicação Web comum, foi construída somente como objetivo de demonstração do uso do *framework*. Ambas aplicações são descritas com detalhe nos tópicos seguintes.

3.1 Ferramentas Utilizadas

As tecnologias utilizadas para a elaboração do *framework* Python e React.

Python como linguagem no *back-end* por ser uma linguagem referência quando se trata de ciência dos dados, juntamente com a linguagem R **[16]**. Em conjunto, foi utilizada o *Scikit-learn* que consiste em ser uma biblioteca de *Machine Learning* que possui diversos classificadores já implementados e prontos para uso. Também foi usada a biblioteca *Flask* que é usada na criação do *Web Services*, fácil de ser utilizada, exigindo uma curva de aprendizado relativamente pequena. **[17]**

Para a confecção da aplicação Web que consumirá o WS, foi decidido utilizar o *framework front-end* chamado React, desenvolvido pelo Facebook. Essa ferramenta trouxe um novo paradigma de desenvolvimento *Web* que consiste em modelar a aplicação separados em componentes. Componentes estes que possuem "estados" armazenados, e com base nesses estados o componente pode se comportar e ser exibido de maneira diferente, dependendo de como foi programado para lidar com tal situação.

Essa forma de desenvolver uma aplicação *Web* faz com que o desenvolvedor deixe toda a aplicação separadas em peças independentes que podem ser facilmente aproveitados em outros projetos, ou seja, deixando altamente escalável permitindo adicionar, com facilidade, novas funcionalidades e recursos na aplicação. Além dessa vantagem, ela foi escolhida para desenvolver uma aplicação *single-page* ou de página única. Isso garante um bom desempenho porque a página é carregada apenas uma única vez.

3.2 Arquitetura

A arquitetura da ferramenta, em uma visão mais ampla, está ilustrada na Figura 1. Como foi mencionado anteriormente, foram construídas duas aplicações. A primeira sendo uma aplicação React fazendo a comunicação com o segundo que é um *Web Service*, seguindo o padrão REST. A segundo aplicação se comunica diretamente com a base de dados do Moodle extraindo informações e dados via conexão direta com o banco.

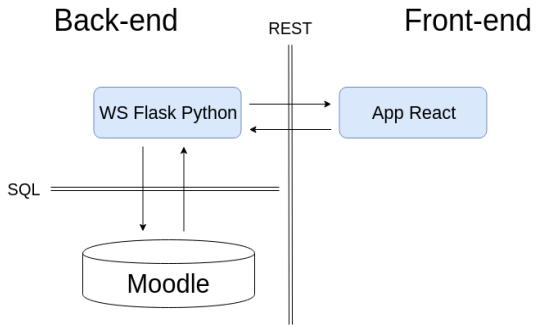


Figura 1 - Visão Geral do Projeto.
Fonte: Autor (2018).

3.2.1 Aplicação React

No desenvolvimento de uma aplicação React é comumente utilizado um script que automatiza a geração de uma aplicação básica com requisitos mínimos para começar o seu desenvolvimento, chamada create-react-app. Ela foi criada pela própria equipe de desenvolvimento do Facebook.

A aplicação proposta possui a seguinte estrutura de pastas, além das criadas pelo *scriptcreate-react-app*, sendo uma delas a pasta "*src*", e interna a elas foram criadas outras pastas:

- **Components:** Pasta com finalidade de guardar todos os componentes da interface do usuário.
- **Endpoints:** onde é guardada um arquivo javascript com variáveis constantes especificando todas as URLs de endpoints do Web Service.
- **Store:** Pasta que contém 2 outras pastas: actions e reducers que são relativos a utilização do Redux, um framework utilizado internamente ao React com o objetivo de permitir a centralização de estados dos componentes da aplicação em um local, assim fazendo com que a manipulação e comunicação entre os estados da aplicação se torne mais fácil.
- **Utils:** Local onde é armazenada arquivos javascript com funções auxiliares e de utilidades com diversas finalidades. Como por exemplo, funções de manipulação de listas, dados temporais, *Strings* e etc.
- **Views:** Arquivos javascript possuindo macro componentes React que, juntamente com os componentes localizados na pasta "components", forma uma página Web por completo.
- **Visualization:** Pasta onde está localizada os componentes referentes a cada visualização criada para cada Classe de treinamento criado pelos usuários. Não possuindo nenhum padrão, a forma de visualização dos dados está livre para ser definida pelo usuário do framework proposto neste artigo, são arquivos que vão consumir resultados vindos do WS.

Resultado da estruturação da aplicação mostrada na Figura 2:

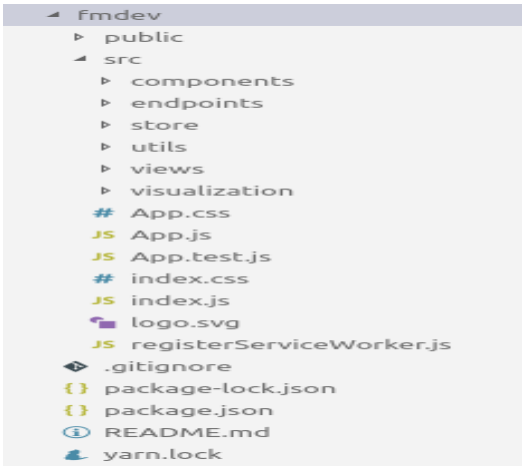


Figura 2 - Estrutura da aplicação React.
Fonte: Autor (2018).

3.2.2 *Web Service Flask*

A aplicação de *Web Service* está estruturada como mostra a Figura 3:

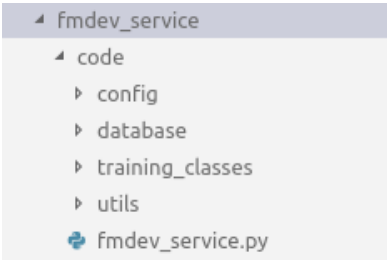


Figura 3 - Estrutura do *Web Service*.
Fonte: Autor (2018).

O arquivo **fmdev_service.py** é onde está localizado o código de inicialização do serviço, assim como as definições dos recursos e rotas oferecidos pelo mesmo.

A pasta **config** possui um arquivo de configuração onde são especificadas informações necessárias para abrir uma conexão com o banco, como por exemplo: tipo da base, tendo duas possibilidades, PostgreSQL ou MySQL, nome da base, usuário, senha, *host* e porta onde o serviço está disponível.

A pasta ***database*** possui dois arquivos *python*, um para PostgreSQL e outro para MySQL, classes estas responsáveis por iniciar uma conexão com a base de dados configurada e realizar consultas definidas por seus métodos, cada um com uma finalidade específica.

O ***training_classes*** é onde é criada arquivos com classes que são implementadas de acordo com a classe abstrata, localizada no mesmo diretório, chamada **ModelClass.py** que obriga o usuário do *framework* a fazer implementação de dois métodos, uma de tratamento de dados e outra de treinamento.

3.3 Utilização do *Framework*

Para adicionar uma nova classe de treinamento, o desenvolvedor deve levar os seguintes passos em consideração:

1. Criar uma Classe python na pasta **training_classes** no **fmdev_service** estendendo a classe abstrata **ModelClass** e

http://dx.doi.org/10.25286/repa.v3i3.976

implementando os dois métodos exigidos por ela que são: **tratar_dados** e **treinar**.

O primeiro método deve receber como parâmetro as seguintes informações:

- os dados em variáveis separadas contendo atributos e com os dados da coluna a ser classificada pelo classificador implementado.
- variável contendo o JSON com as opções de configuração recebida dos *inputs* de formulários do *front-end*, também contendo dados de configuração. vindas da aplicação que está utilizando o WS.

No segundo método é importante definir o formato de retorno dos resultados obtidos no treinamento que será usado no componente de visualização pela aplicação que está utilizando o WS. A Figura 4 ilustra a criação descrita anteriormente:

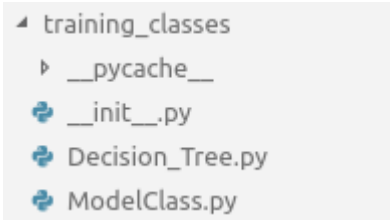


Figura 4 - Passo 1: criação do arquivo (exemplo usando árvore de decisão).
Fonte: Autor.

```
1  from abc import ABCMeta, abstractmethod
2
3  # Classe Modelo para ser seguido por outras classes
4  # Implementação Obigratória de alguns métodos da Classe
5
6  class ModelClass(object):
7
8      __metaclass__ = ABCMeta
9
10     def __init__(self, processing_op):
11         self.processing_op = processing_op
12
13     @abstractmethod
14     def tratar_dados(self): pass
15
16     @abstractmethod
17     def treinar(self): pass
```

Figura 5: Passo 1 - Classe Modelo que deve ser implementada.
Fonte: Autor.

```
13  def tratar_dados(self, X, Y, config):
14      # Conversão Numérica Categórica
15      Y['nota'] = np.where(Y['nota']>=50, 'aprovado', 'reprovado')
16
17      # Particionamento de Dados
18      part = config['part']['teste']
19      part = round(part/100, 2)
20      X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=part )
21
22      # Atribuição do resultado do tratamento como Atributos da Classe
23      self.X_train = X_train
24      self.X_test = X_test
25      self.Y_train = Y_train
26      self.Y_test = Y_test
```

Figura 6 - Passo 1: Exemplo de implementação de treinamento (Árvore de Decisão).
Fonte: Autor.

2. Após a implementação da nova Classe de treinamento, precisamos adicionar o código de instanciação em uma área específica no arquivo fmdev_service.py. Local onde existe um trecho de código condicional, que verifica o campo com o nome da classe que a aplicação consumidora deseja executar na variável contendo informações de configuração. Caso esse arquivo exista na para training_classes, o código cria uma nova instância atribuindo a variável training_class.

```
177     # Área de inserção de Classes Implementadas
178     # Inserir condição aqui para instanciar a classe solicitada
179     if (nome_arquivo == "Decision_Tree"):
180         training_class = Decision_Tree()
181
182     # Executa o tratamento dos dados implementados na Classe Selecionada
183     if (training_class != None):
184         training_class.tratar_dados(X, Y, config)
185         resultado_treino = training_class.treinar()
186         return {'result': resultado_treino}
187     else:
188         return {'result': "Class does not exist !"}
```

Figura 6 - Passo 2 :Inserir código de instanciação na condicional.
Fonte: Autor.

Os passos descritos anteriormente são suficientes para que outras aplicações externas comecem a utilizar o recurso criado pelo usuário do *framework*.

Os passos descritos logo em seguida são referentes aos passos que devem ser seguidas, especificamente, na aplicação que está utilizando o WS, que são específicas a aplicação React que foi criada para demonstrar o consumo do WS.

Esses passos podem variar dependendo de como o desenvolvedor da aplicação deseja consumir os resultados dos classificadores.

Criar novo componente de visualização para consumir os resultados do treinamento obtido. Para isso basta criar um novo componente dentro da pasta **visualization** na aplicação React. A estrutura interna do componente está livre para ser definida pelo usuário do *framework*, ou seja, pode adotar qualquer tipo de bibliotecas de visualizações de dados a ser exibido pelo componente criado.

Esta etapa é opcional, caso o usuário deseje acrescentar um novo tipo de tratamento de dados e configuração, é necessário criar um componente de formulários para o usuário inserir que tipo de tratamento deseja usar nos dados que está enviando para o WS. E implementar códigos no WS, que lide com a configuração desejada.

4 Testes e Resultados

O teste da ferramenta foi realizado em cima da base do Moodle do NEAD (Núcleo em Educação a Distância) da Universidade de Pernambuco, que foi disponibilizada por Maciel **[1]**. Os atributos que foram usados para os testes foram criados por Santana **[7]**, que consiste em consultas SQL que extraem as seguintes informações do banco de dados do Moodle do NEAD:

- Número de Postagens Realizados do Aluno.
- Número de Postagens de outros Alunos lidas pelo Aluno.
- Número de logins do Aluno.
- Número de Revisão de Postagens Anteriores.

Com base nestes atributos tentamos classificar se o aluno foi aprovado ou reprovado utilizando o algoritmo de árvore de decisão. O método que retorna o resultado foi implementado de maneira que ela apenas retorne como a estrutura da árvore. Não foi realizada nenhuma validação do resultado retornado pelo motivo de não estar no escopo do projeto, ficando apenas focado na construção das aplicações. Os resultados das construções são descritos detalhadamente no tópico seguinte.

A aplicação React trata-se de uma ferramenta onde o usuário pode realizar processos relacionados à mineração de dados sequencialmente: seleção de atributos, pré-processamento, treinamento, resultado. As imagens referentes às telas encontram-se anexo a este artigo e são descritas logo abaixo:

Tela de Autenticação: Onde o usuário precisa informar uma credencial válida cadastrada no Moodle. Após a autenticação o botão do próximo é habilitado para ir a próxima página. (anexo - tela 01).

Tela de seleção dos atributos: Nesta tela, o usuário pode selecionar atributos para passar pelo pré-processamento e posteriormente para a fase de treinamento. (anexo - tela 02).

Tela de edição, criação e pré-visualização: criar, editar consultas SQL para extrair os atributos do banco de dados. É, também, o local onde o pode executar SQL e pré-visualizar o resultado. (Anexo - tela 03)

Tela de Seleção e tratamento de Dados: Nesta etapa são listadas as colunas resultantes das consultas selecionadas na etapa anterior. É onde o usuário pode verificar o tipo dos dados, selecionar colunas que vão servir como entrada para o algoritmo de classificação, a coluna Classe, a coluna ID que vai servir para realizar o INNER JOIN das consultas selecionadas a partir da consulta com a coluna ID. É importante que na etapa anterior, as consultas possuírem colunas em comum para realizar esse processo, caso contrário serão desconsideradas para o envio na próxima etapa. (anexo - tela 04)

Tela de seleção de algoritmos de classificação: Tela de escolha do algoritmo a ser executado. As opções são disponibilizadas conforme o desenvolvedor adicione classes na pasta *training classes*. (anexo - tela 05)

Tela de Resultados: Após a escolha do algoritmo e sua execução, o resultado é apresentado conforme definido no componente de visualização criada pelo desenvolvedor. (anexo - tela 06).

5 Conclusões e trabalhos futuros

O *framework* proposto apresentou ser funcional e permitiu que novos classificadores sejam adicionados com poucos passos de

http://dx.doi.org/10.25286/repa.v3i3.976

codificação. Embora exista vários pontos que ainda devem ser analisados:

- Conexões diretas com a base do Moodle não é recomendada segundo a própria Moodle. Existem meios mais seguros para tal finalidade que consiste em realizar tais operações de consulta de informações através da API do próprio Moodle.
- Realizar experimentos mais elaborados com bases mais populosas. A base do NEAD que foi disponibilizada apresentou poucas informações referentes aos atributos selecionados para teste, apresentando 1500 alunos. No entanto, desses 1500 alunos, somente 97 apresentou ter tais atributos, e entre os 97 somente 8 apresentaram ser aprovadas, deixando o algoritmo com *Overfitting* de alunos aprovados.
- Mapear parâmetros dos classificadores. Existem diversos classificadores na biblioteca utilizada (*Scikit-learn*), cada um com suas propriedades e parâmetros específicos. Para melhor desempenho dos algoritmos classificadores, existe a necessidade de considerar tais características como parâmetro de configuração para serem processadas pelo *framework* criado.

6. Referências

[1] MACHADO, Liliana Dias; MACHADO, Elian de Castro. O papel da tutoria em ambientes de EAD. In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA , 11., 2004, Salvador. **Anais**... Salvador: ABED, 2014. p. 1-3.

[2] LOPES, Maria Cristina et al. O processo histórico da educação à distância e suas implicações: desafios e possibilidades. IN: JORNADA DO HISTEDBR, 7., 2007, Campo Grande. **Anais**... Campo Grande, MS: UCDB. 2007.

[3] NUNES, E. R.; AQUINO, G. A. M.; FURTADO, A. M. A importância dos ambientes Virtuais de Aprendizagem na Busca de novos domínios do EAD. In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA , 13., 2004, Salvador. **Anais**... Curitiba: ABED, 2007. p. 1-3.

[4] FRANCISCATO, Fábio Teixeira et al. Avaliação dos Ambientes Virtuais de Aprendizagem Moodle, TelEduc e Tidia-ae: um estudo comparativo. **Novas Tecnologias na Educação (RENOTE)**, v. 6, n. 2, 2008.

[5] HERMIDA, Jorge Fernando; BONFIM, Claudia Ramos de Souza. A educação à distância: história, concepções e perspectivas. **Revista HISTEDBR On-line**, Campinas, n. especial, p. 166-181, 2006.

[6] COSTA, Evandro et al. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. In: Jornada de Atualização em Informática na Educação, 2., 2013, Campinas. **Anais**... Campinas:Sociedade Brasileira de Computação, 2013.

[7] SANTANA, L. C. D. Integração de um Mecanismo de Mineração de Dados educacionais ao Moodle. 2015.

[8] SOMMERVILLE, Ian. **Engenharia de software.** 6 ed. São Paulo: Addison Wesley, 2003.

[9] MACIEL, A. M. A.; RODRIGUES, R. L.; CARVALHO, E. C. B. Desenvolvimento de um Assistente Virtual Integrado ao Moodle para Suporte à Aprendizagem Online. In: Simpósio Brasileiro de Educação a Distância, 2., 2014, São Carlos. **Anais**... São Carlos, SP: SEAD, 2014.

[10] GONÇALVES A. F. D. Desenvolvimento de uma arquitetura integrada para a visualização de dados em ambientes de ensino a distância. 2017.

[11] JOHNSON, Ralph E. Components, frameworks, patterns. In: GEORGANS, Jhoan. **ACM SIGSOFT Software Engineering Notes**. v. 43, n.2. New York: ACM, 1997. p. 10-17.

[12] FAYAD M., SCHMIDT, D. JOHNSON, R. **Building Application Frameworks: Object-Oriented Foundations of Framework Design**. John Wiley & Sons, 1999.

[13] MATTSSON, Micheal. **Object-oriented Frameworks**: A survey of methodological issues. Licentiate thesis. Department of Computer Science, Lund University. Swend: Lund UNiverty, 1996.

[14] OPENSOFT. **Web Service**: o que é, como funciona e para que serve, 2016. Disponível em: <https://www.opensoft.pt/web-service/> Acesso em: 12 jul. 2018.

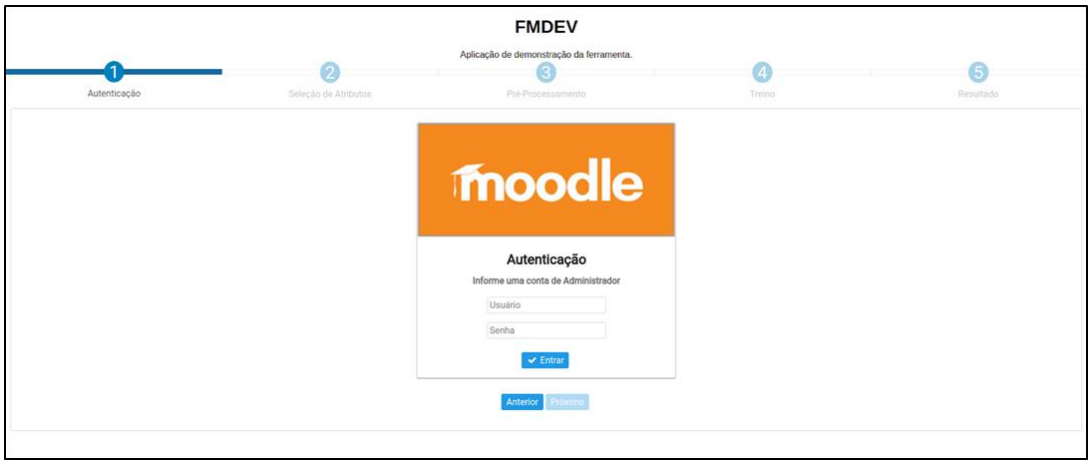
[15] MOODLE, **Moodle Statistics**, 2018, Disponível em: <https://moodle.net/stats/>. Acesso em: 12 jul.2018.

[16] CANTO, Carlos Eurico. As linguagens populares na ciência dos dados. **Propus Data Science**, 9 maio 2017. Disponível em: :<http://propus.science/as-linguagens-mais-populares-em-ciencia-de-dados/>. Acesso em: 08 ago. 2018.

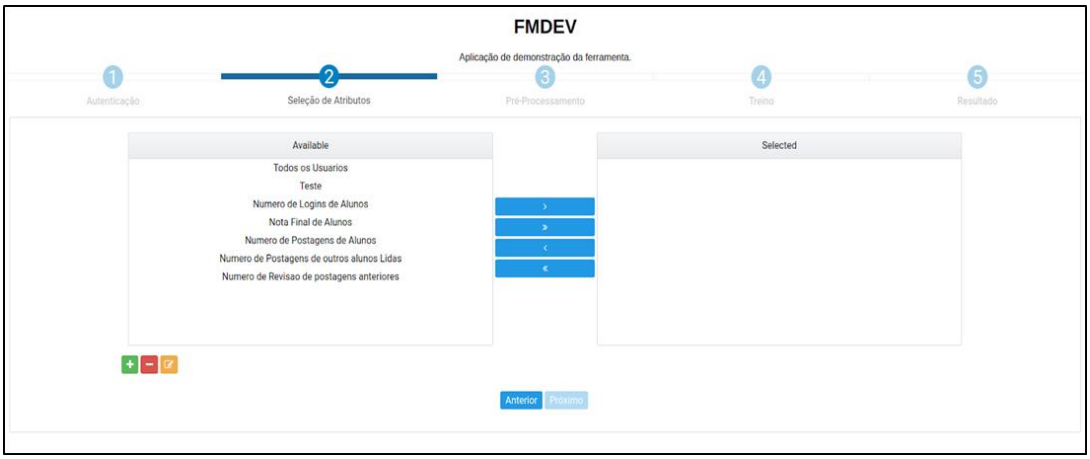
[17] FATIMA, H. Flask x Django: como escolher o framework concreto para seu aplicativo Web. **iMasters**, 10 abr. 2018. Disponível em: <https://imasters.com.br/back-end/flask-x-django-como-escolher-o-framework-correto-para-seu-aplicativo-web>. Acesso em: 08 ago. 2018.

http://dx.doi.org/10.25286/repa.v3i3.976

Anexos

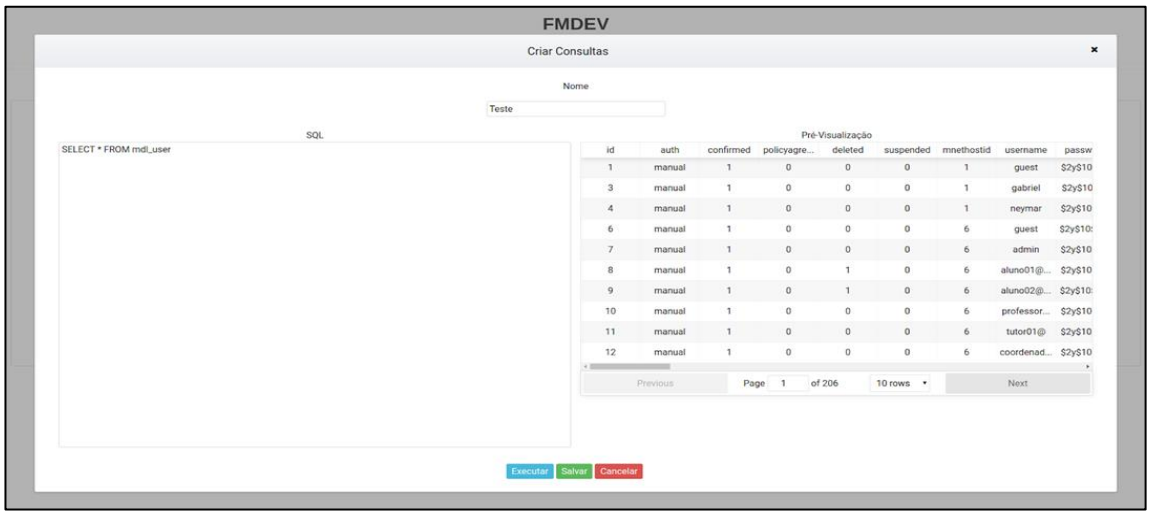


Tela 1: Tela de autenticação. Fonte: Autor.



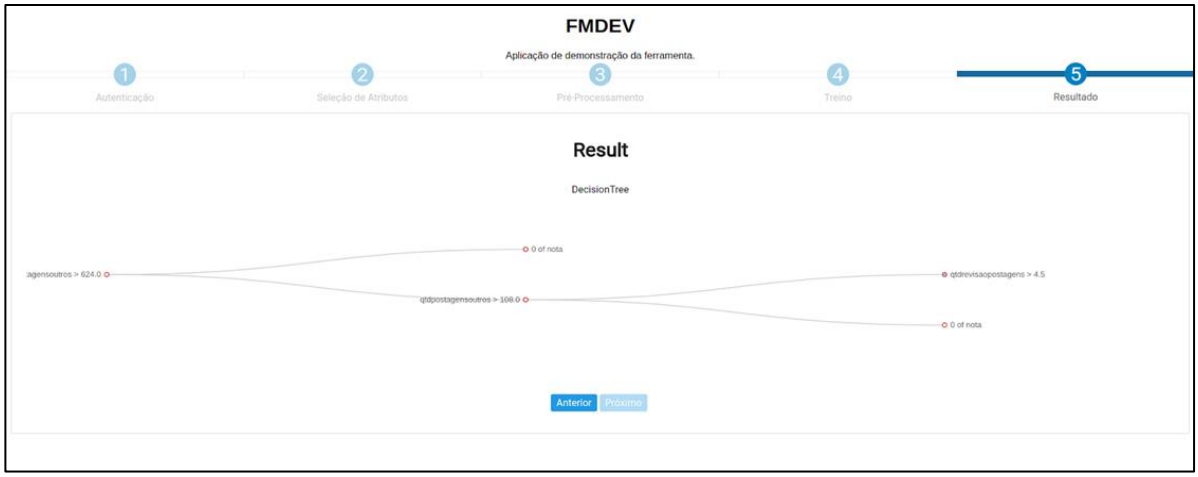
Tela 2: Tela de seleção de atributos. Fonte: Autor.

Tela 3: Tela de criação de atributos. Fonte: Autor.



Tela 4: Tela de pré-processamento. Fonte: Autor.

Tela 5: Tela de seleção de classificadores Fonte: Autor.



Tela 6: Tela de resultados Fonte: Autor.

http://dx.doi.org/10.25286/repa.v3i3.976


UPE
UNIVERSIDADE
DE PERNAMBUCO
POLI
ESCOLA POLITÉCNICA
DE PERNAMBUCO