Prediction of active debt in the State of Pernambuco, Brazil

Prediction of active debt in the State of Pernambuco, Brazil

Álvaro Farias Pinheiro [1] orcid.org/0000-0002-6254-7293
João Alberto da Silva Amaral [1] orcid.org/0000-0002-8141-4787
Geraldo Torres Galindo Neto [1] orcid.org/0000-0001-7244-8822
José Nilo Martins Sampaio [1] orcid.org/0000-0002-1752-9926
Wedson Lino Soares [1] orcid.org/0000-0002-0078-3944

[1] Polytechnic School of Pernambuco, University of Pernambuco, Recife, Brazil.

Lead author's email: Álvaro Pinheiro afp@ecomp.poli.br

Abstract

Application of data mining (DM) techniques to optimize the process of collection of Active Debt (AD) of the State of Pernambuco, Brazil. We apply the following data mining techniques: Decision Tree (DT), Logistic regression (LR), Naive Bayes (NB), Support vector machine (SVM), also applied to the Random Forest technique which is considered an assemble method. We observed that the RF technique obtained better results than all the techniques of classification, reaching higher values in all metrics analyzed. We note that the creation of a data mining model to choose which debts can succeed in the collection process can bring benefits to the Pernambuco government. With the application of RF technique, we obtained indexes above 85% in the evaluation of the metrics.

Keywords: Collection; Active Debt; Artificial Neural Network; Prediction.

1 INTRODUCTION

In recent years, amount of data is growing in information systems, it is estimated volume of data present in organizations is double every 20 months **[1]**. This large volume of data can create difficulties the decision-making process of organizations. In this context, the data mining process seeks to fill this gap, second Witten *et al*. **[1]**: "Data mining is about solving problems by analyzing data already present in databases", and "Data mining is defined as the process of discovering patterns in data". Recently data mining is an approach used is several fields including economy, **[2]** healthy **[3]**, industrial process **[4]**, education **[5]**, software engineering **[6]**, social media **[7]**, and agriculture **[8]**. Currently, the data mining process has also been used to analyze the granting of customer credit **[9]**, as well as optimize debt collection **[10]**. Second Hunt **[11]** the process of collecting debts has recently been completely restructured with the use of Information Technology (IT). In the process of debt collection, government agencies have an important regulatory role for three reasons: "(1) there is excessive racing in collections by unsecured creditors, (2) creditors cannot easily distinguish between those who can't pay and those who simply won't pay, and finally (3) consumers are either unwilling or unable to file for bankruptcy". Government agencies face two challenges in the debt collection process. First the large volume of data stored [1], and second **[11]**: "To be effective, collectors must be able to distinguish consumers who can't pay from those who won't pay even though they have the resources to do so". In this paper we aim to use data mining techniques to optimize the debt collection process in the state government of Pernambuco, Brazil. In section II we present the motivations for conducting this work.

DOI: 1025286/repa.v5i1.1299

2 MOTIVATION

In this section, we presented two motivations for conduct this study. Motivation 1: Evolution of active debt over the years. The Pernambuco State Attorney General's Office (PGE) is the government agency of the state of Pernambuco responsible for collecting the state's debts. However, despite all the efforts made by the agency, the amount of debt has been increasing annually (See Figure 1).

Motivation 2: Increase in the number of debtors. In addition to the large amount of debts that the state of Pernambuco accumulates, the number of debtors has also increased exponentially year after year (See Figure 2).

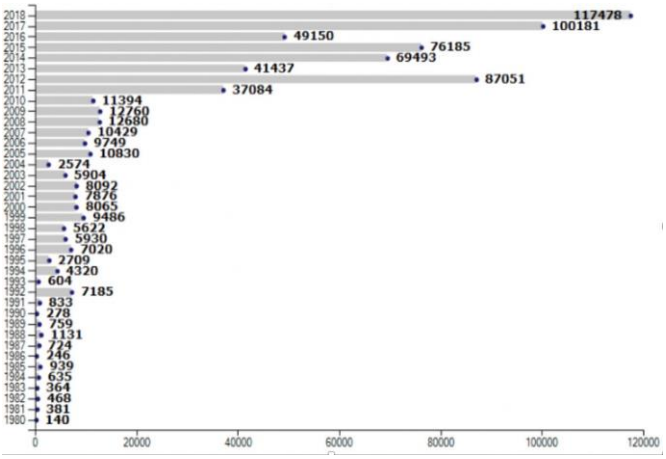

Figure 1: Number of active debts expressed in R$ in each year (1980-2018).
Fonte: The Author (2018).

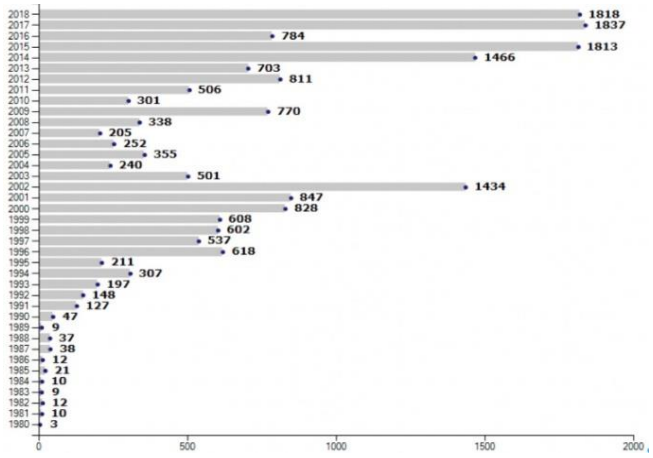

Figure 2: Number of debtors in each year (1980-2018).
Fonte: The Author (2018).

3 BACKGROUND

In this section, we first present the methodology Cross Industry Standard Process for Data Mining (CRISP-DM). Second, we present an overview of the data mining techniques used in this work. We finish the section presenting the metrics used for analysis the techniques for data mining classification.

3.1 CRISP-DM

This methodology is composed of six phases: (1) Business understanding; (2) Data understanding; (3) Data preparation; (4) Modeling; (5) Evaluation; and (6) Deployment **[12]** (See Figure 3). The phases are flexible, with a set of actions defined for each of them, triggered cyclically **[12]**.

1) ***Business understanding***: Second Chapman et al. [12] at this stage the data analyst should seek to understand what the business objectives are; and how the use of data mining can bring some benefit to the organization. According to Chapman *et al*. **[12]** this step is extremely important, **[12]** argues that.

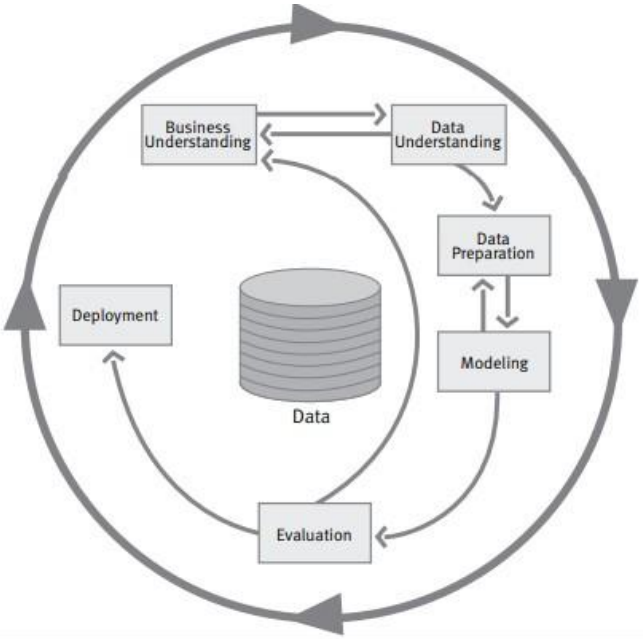

Figure 3: CRIP-DM Methodology.
Fonte: Chapman *et al.* (2000).

"A possible consequence of neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions".

2) ***Data understanding***: This phase consists of four steps **[12]**: (1) Collect initial data: access data and, for example, collect data from multiple databases; (2) Describe data: perform an initial analysis of data, for example, create or modify an

existing data dictionary; (3) Explore data: after the initial analysis of the data in the previous step, observe, for example, the relationship between the data and which data may be most relevant to the solution of the problem; and (4) Check the data quality: observe what data is missing or incomplete, and verify that the data is represented correctly.

3) ***Data preparation***: This phase consists of five steps **[12]**: (1) Select data: The data that will be used for application of the data mining techniques are chosen, inclusion or exclusion of rows and columns of the database must be justified **[12]**; (2) Clean data: in the previous phase of the CRISPDM a verification of data quality is realized, in this step a deeper analysis is performed, for example categorical data can be transformed into numerical data; (3) Construct dada: new data may emerge, for example data from two columns may be derived in a new attribute, as exemplified by Chapman *et al*. **[12]**: "Example: area = length * width"; (4) Integrate data: Multiple-table data can be linked, Chapman *et al.* **[12]** gives an example: "converting a table of customer purchases where there is one record for each purchase into a new table where there is one record for each customer, with fields such as number of purchases, average purchase amount, percent of orders charged to credit card, percent of items under promotion, etc."; (5) Format Data: syntactic changes are performed on the data, depending on the problem addressed and what tools will be used. For example: "Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict."

4) ***Modeling***: At this stage the data mining technique is chosen and applied in the database, in addition an initial data model is generated.

5) ***Evaluation***: An initial version of the data model generated in the previous step is tested in this step.

6) ***Deployment***: The data model tested in the previous step is applied in the organization.

3.2 Classification techniques

In this section we will briefly present the data mining techniques that will be used in this study.

1) ***Decision tree*** (DT): are methods of learning nonparametric supervised machines, widely used in classification and regression tasks. In its construction is used a training set formed by inputs and outputs where the model is trained to extract the characteristics of the existing classes in the data and then be able to be intended classify new data submitted to the model **[13]**. These models are among the most popular inference algorithms and has been applied in several areas such as medical diagnosis and credit risk, and from them one can extract" if then" which are easily understood **[13]**.

2) ***Naive Bayes*** (NB): is a classification algorithm, which uses historical data to predict the classification of a new data. The basic principle of the operation of this model is the calculation of the probability of an event occurring since another event has already occurred, so it is called "naive" because it disregards the correlation between the variables **[13]**.

3) ***Logistic regression*** (LR): is a model that allows us to estimate the probability associated to the occurrence of a given event in the face of a set of explanatory variables. it is a statistical technique that aims to model, from a set of observations, the" logistic" relationship between a variable dichotomous response and a series of numerical or categorical explanatory variables **[13]**.

4) ***Support Vector Machine*** (SVM): is a representation of examples as points in space, mapped so that the examples in each category are divided by a clear space that is as broad as possible. The new examples are then mapped in the same space and predicted as belonging to a category based on which side of space they are placed. Therefore, what an SVM does is find a line of separation (a hyperplane) between data of two classes. That is, the hyperplane seeks to maximize the distance between the closest points in relation to each of the classes **[13]**.

5) ***Random Forest*** (RF): this algorithm is an ensemble classifier, second **[14]** can be defined as: "Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest".

Each tree trains an initial sample of the original training data and searches only for a randomly selected subset of the input variables to determine a split **[15]**. Each tree in the random forest casts a unit vote for the most popular class at the entrance and the output of the classifier is determined by the vote of most trees **[15]**.

3.3 Evaluation Measures

In our experiments we used four types of metrics to evaluate the data mining techniques presented in Section III-B. The metrics used were: Accuracy, Precision, Recall, and F-Measure. All these metrics use the data present in the confusion matrix (See Table 1).

DOI: 1025286/repa.v5i1.1299

Table 1: Confusion Matrix.

	Detected	
	Positive	**Negative**
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

According **[16]** "Accuracy is the proportion of the total number of predictions where correctly calculated."

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

Precision is: "the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases."

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Second Powers **[16]** Recall metric is: "is the proportion of Real Positive cases that are correctly Predicted Positive".

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Also used was the F-Measure metric **[16]**: "combine the recall and precision which is considered a good indicator of the relationship between them".

$$F - Measure = 2\frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Finally, the ROC (Receiver Operating Characteristic) metric, ROC is a metric to compare systems performance, represented by the area under the ROC curve. The ROC curve is plotted as a diagram of the true positive values (TP) as a function of the false positive (PF) ratio. The more the value of this metric is closer to 1 the better the performance of the classifier **[16]**.

4 METHOD

In this section we present the application of CRISP-DM in the database obtained from the Judicial Automation system, with data belonging to the Attorney General's Office State of Pernambuco PGE and referring to the register of active debt. Below we present the steps of CRISP-DM and what was performed at each stage.

1) ***Understanding of the business***: The "PGE" is the organ of the government of the state responsible for the judicial representation of Pernambuco, with its activities legal consultancy to the executive branch and The collection of the active debt, in addition to the exercise of the other attributions assigned by law, being divided into specialized which are: litigation, advisory, support to the Governor and Ministry of Finance, the latter being responsible for the collection process, which is subdivided into areas such as the Core Processes Priority, Core of Tax Intelligence, Nucleus of Successes and Donations and Core of Active Debt. All work together to better identify and promote debt collection actions.

The proposal of this research is to provide the techniques of intelligence artificial to assist the activity of the recovery of the State's financial health, grouping to the debts entered, identifying the rules that characterize the actions taken by the prosecutors and indicating which of the best debts to be charged priority. The collection process that is intended to be set up is to determine which debts should be considered for sending letters, emails and phone calls that will inform the need to pay the debt, and if the debtor refuses to san it, an legal action should be recommended for the redemption of this debt, which should be used in the latter case, since is an expensive resource, so other processes can and should be used in alternative, such as SERASA and protest in notary's office.

For this control, the "PGE" has an active debt database that's available in two public agencies, the Financial Secretary of the State of Pernambuco "SEFAZ" and the "PGE", the first is the database of the efisco System and the second is the database of the System of Automation of Justice (SAJ). For this work we used the data available on the second, but they receive some fields from the first.

2) ***Understanding of the data***: to apply the techniques it was necessary to follow some steps, how to determine which model best fits the problem you want to attack, and to this requires standardized data, but the existence of many data, makes it known which ones should be worked on, and here comes the importance of understanding the business. For this problem, which aims to predict which debt should be electronically assisted or not, according to the values obtained from THE SAJ database, the data displayed in list were used to be treated.

Table 2: The SAJ database.

Variable	Description
NUCDA and CDPROCESSO	Composite key for the PGE to identify the record.
CDTIPOCDA	Binary identifier that indicates whether it is a charge of ICMS (1) or another tribute (0).
CDPESSOA	Person Table Code.
TPPESSOA	Binary identifier that indicates whether you are a legal person (1) or a physical person (0).
CDGRUPOECONOMICO	The first two digits of the CNAE code that indicates the economic activity area.
VLCDA	Initial debt value.
VLCDAATUALIZADO	Updated debt value.
NUPARCELAMENTO	Binary identifier that indicates whether the debt was parceled (1) or not (0).
DTAJUIZAMENTO	Field with the date that the process was aided, if it has not gone to justice is empty.
CDORGAOCOMARCA	code of the district where the debt will be treated.
CDSITUACAOPROC	indicates whether the company responsible for the debt is an active.

3) ***Data preparation***: in this phase the activities of selection, cleaning, creation, integration and formatting of the data were performed. For this, the preparation of the variables cited in Table I was performed. We identify some independent variables: CDTIPOCDA; TPPESSOA; CDGROUPECONOMICO; NUPARCELAMENTO, and other that we could use to calculate the dependent variable: VLCDATUALIZADO; VLCDA and DTAJUIZAMENTO. The following variables were excluded: NUCDA and CDPROCESS by composing the composite key so that the PGE identifies the record. Thus, we verified that these fields do not add value to the prediction and were eliminated during the cleanup process. CDPESSOA was not perceived duplicity in the codes, we decided not to include this field in the analysis. The variable CDGRUPOECONOMICO was divided in many binaries' fields according to CNAE code treated to indicate the area of economic activity (See Table 2).

Table 3: Economy Activities.

Variable	Description
01..03	AGRICULTURE, LIVESTOCK, FORESTRY PRODUCTION, FISHERIES AND AQUACULTURE.
05..09	EXTRACTIVE INDUSTRIES.
10..33	PROCESSING INDUSTRIES.
35	ELECTRICITY AND GAS.
36..39	WATER, SEWAGE, WASTE MANAGEMENT AND DECONTAMINATION ACTIVITIES.
41..43	CONSTRUCTION.
45..47	TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES.
49..53	TRANSPORT, STORAGE AND MAIL.
55..56	ACCOMMODATION AND FOOD.
58..63	INFORMATION AND COMMUNICATION.
64..66	FINANCIAL, INSURANCE AND RELATED SERVICES ACTIVITIES.
68	REAL ESTATE ACTIVITIES.
69..75	PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES.
77..82	ADMINISTRATIVE ACTIVITIES AND COMPLEMENTARY SERVICES.
84	PUBLIC ADMINISTRATION, DEFENSE AND SOCIAL SECURITY.
85	EDUCATION.
86..88	HUMAN HEALTH AND SOCIAL SERVICES.
90..93	ARTS, CULTURE, SPORT AND RECREATION.
94..96	OTHER SERVICE ACTIVITIES.
97	DOMESTIC SERVICES.
99	INTERNATIONAL BODIES AND OTHER EXTRATERRITORIAL INSTITUTIONS.

These independent variables were used as data entry in the artificial neural network, the independent variables were normalized to balance the different or very heterogeneous values and the dates fields were converted to days. Also, to normalize and get the Y of X, the rule y = (x-min)/(max-min) was applied.

In relation to the dependent variable, it is identified which debt should be aided or not by an identifier, keeping the debtor's identification in absolute secrecy, due to the confidential nature of the taxpayer's data and the legal requirements of the Brazilian General Law of Data Protection. Consequently, the data of this sample are real data, but do not identify the debtor, since the fields used to allow the identification of the debtors were extracted, so guaranteeing their anonymity.

It is important to emphasize that the objective of this research was to verify whether the method applied satisfactorily meets time and processing capacity the use of a prediction function to indicate which form of debt collection should be applied to according to her characteristics.

In this survey, x entries were created as the most significant variables in the set of available fields in the database, the target being the debt value, a single hidden layer consisting of processing neurons with

DOI: 1025286/repa.v5i1.1299

weight was created. A synaptic called a w, and a single output neuron, which is the variable that indicates whether it is a debt that is judgable or not. Thus, the network will make possible the modeling of indication of the form of collection in the case of outputs with debt value that covers the costs of the process and would also serve to evaluate what other forms of collection for the debts are not feasible.

An advantage of using initial variables is the decrease in the effects of atypical values, moreover, the confidentiality of data is increased since it is very difficult to identify individuals from the values. And as another advantage is the normality that is induced in the variables of the model due to the behavior of the main components to be symptomatically normal.

In the selection process the following techniques were applied: Vertical selection, for the selection of records was applied the criterion of bringing only processes that have already been completed, because there is no way to calculate the dependent variable of processes in progress the That would prevent supervised training. As a result, 57,098 lines were made available.

From the calculation of the dependent variable, as there is no variable on the basis that indicates whether the judicial collection was timely or not, the suggestion was to consider sending to justice when the process has date of filing (A) and in the case of ICMS (I) the amount paid (P) (initial value at is greater than R$ 26,000.00 or in the case of other taxes (O) the amount paid (P) is greater than R$ 13,000.00 which can be presented as the following logical proposition: Y = A ∩ ((I ∩ (P > 26, 000)) ∪ (O ∩ (P > 13, 000)).

From the cleaning and transformation of the fields, after receiving the base in CSV file we proceeded to the cleaning and transformation of the data to better adapt the classification technique with supervised training chosen artificial neural networks. The following is a list of the procedures used to prepare the base:

• When wen analyze the field VLCDA, we found some negatives values, after validated the situation with the government managers responsible for the database it was perceived that it was an inconsistency in the database, they recommended that we use zeros instead the negative values;

• Based on the DTAJUISHING field that has null dates and values, a calculated field has been created that can assume the binary value ”0” for unaided and ”1” for the processes that have been helped;

• Created a VLPG (paid amount) field that is the result of the following subtraction: VLCDA-VLCDA;

• The CDGROUPECONOMICO field that represents the first two digits of the CNAE code, for being a categorical data was splatted into several binary fields for each category, as represented below: CNAEAgriculture, CNAE-Extractive, CNAE-Transformation, CNAE-Energy, CNAE-Water, CNAE-Construction, CNAE-Trade, CNAE-Transport, CNAE-Food, CNAECommunication, CNAE-Financial, CNAE-RealEstate, CNAE-Professional, CNAE-Adm, CNAE-Public, CNAEEducation, CNAE-Cheers, CNAE-Arts, CNAE-Others, CNAE-Domestic, Cnae-International;

• Created the column to determine whether the process is in the minimum limit value established by law to be forwarded to the LIMITEVALOR justice. Which can be calculated by the following logical proposition: LIMIT EV ALOR = (I ∩ (P > 26, 000)) ∪ (O ∩ (P > 13, 000) Where: I = ICMS, p = Paid amount, O = Other taxes;

• Calculated the dependent variable following the formula: Y = A ∩ ((I ∩ (P > 26, 000)) ∪ (The ∩ (P > 13, 000));

• Renamed the input fields to facilitate understanding, being at the end represented by the following attributes: ICMS, PJ, normalized value, installment, CNAEAgriculture, CNAE-Extractive, CNAE-Transformation, CNAE-Energy, CNAE-Water, CNAE-Construction, CNAE-Trade, CNAE-Transport, CNAE-Food, CNAECommunication, CNAE-Financial, CNAE-RealEstate, CNAE-Professional, CNAE-Adm, CNAE-Public, CNAEEducation, CNAE-Cheers, CNAE-Arts, CNAE-Others, CNAE-Domestic, Cnae-International, LIMITEVALOR;

• Due to the publication of a recent normative that determines minimum values to forward a process to justice and the fact that the system data possess many records that do not follow this normative, we had to made a severe restriction on the records for that training does not learn with cases that should no longer be followed. We used the field LIMITEVALOR = 1.0 for the filter criterion, which reduced the number of records from 57,098 to 10,210 records.

4) ***Data modeling***: The modeling step is shown to be the intermediate step of the CRISP-DM process and which has a very high importance due to the fact that from the techniques applied here it will be possible to obtain the results to be evaluated in the next step. In this work the classification techniques applied to the data were: tree, sum, random forest, naive Bayes and logistic regression. The techniques were applied based

on preliminary tests and comparisons performed with the Orange Canvas tool.

5) ***Evaluation***: In this step we observe how the results obtained with the data mining techniques meet the specified problem. It is an important step because it is necessary to carefully check the characteristics and qualities of the models before they are implanted. In this study, the evaluation was performed through techniques that seek to influence a better decision-making, through the training of some classification techniques by means of some metrics, such as accuracy, precision, recall, F-measure and the area under the ROC curve. The results obtained in the evaluation stage can be checked in the results chapter

6) ***Deployment***: In this step all the knowledge obtained through the mining work becomes a subsidy for the development of strategies related to the problem that is in focus and that can be used by the client after all the steps of CRISP-DM. This work does not cover this stage, so it remains open for possible future work.

5 RESULTS

In this section we present the results of the application of data mining techniques. See (Figure 4).

Regarding the analysis of the metrics, we observed that the RL technique obtained the highest index of accuracy with 86.3, precision with 85.2, and recall with 86.3

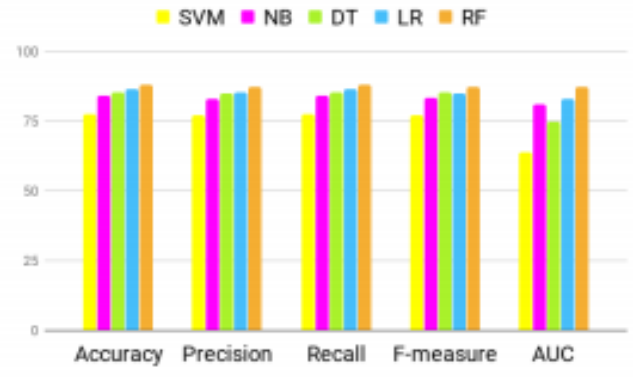

Figure 4: Classification Method Results Using Ensemble Methods.
Fonte: The Author (2018).

The DT technique obtained better performance by observing the F-measure metric with 85.2. Finally, the AUC technique obtained greater indices with the RL technique reaching 83.1.

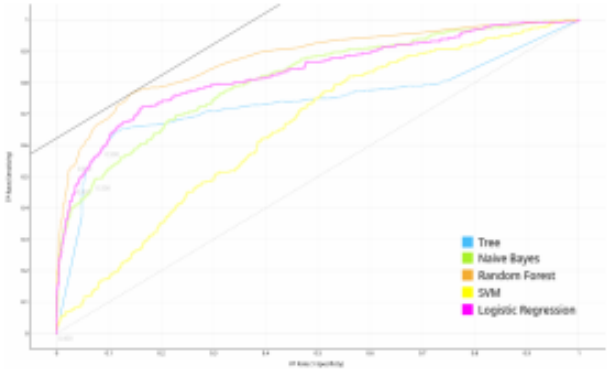

Figure 4: ROC curve.
Fonte: The Author (2018).

Table 4: Classification method results using ensemble methods.

Evaluation measure	Classification technique				Ensemble method
	SVM	LR	NB	DT	RF
Accuracy	77,6	86,3	84,0	85,3	88,1
Precision	77,1	85,2	82,8	85,1	87,4
Recall	77,5	86,3	84,0	85,3	88,1
F-measure	77,3	84,9	83,2	85,2	87,3
ROC	64	83,1	81,0	74,6	87,3

6 CONCLUSION

In this paper, we seek to analyze how data mining techniques can optimize the process of debt collection, we use a database of debtors from the state of Pernambuco, Brazil. The classification methods used in this study achieved good assertiveness results at 88% and 89%, we believe that the proposed model can be considered reliable since all the research metrics have achieved a satisfactory result. As future work, it is expected that this model will be applied in the PGE, and that its results without being analyzed by the users of the model. In addition, in this study we used only the database containing data for the year 2018, a study that synthesizes more data can offer as contribution a much more robust model.

REFERENCES

[1] WITTEN, I. H., *et al.* **Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2016.

[2] ARSLAN, A. K.; COLAK, C.; SARIHAN M. E. **Different medical data mining approaches-based prediction of ischemic stroke**.

DOI: 1025286/repa.v5i1.1299

Computer methods and programs in biomedicine, vol. 130, pp. 87–92, 2016.

[3] GENG, R.; BOSE, I.; CHEN, X. **Prediction of financial distress: An empirical study of listed chinese companies using data** mining. European Journal of Operational Research, vol. 241, no. 1, pp. 236–247, 2015.

[4] GE, Z.; *et al.* **Data mining and analytics in the process industry: The role of machine learning**. IEEE Access, vol. 5, pp. 20590–20616, 2017.

[5] KAUR, P.; SINGH, M.; JOSAN, G. S. **Classification and prediction-based data mining algorithms to predict slow learners in education sector**. Procedia Computer Science, vol. 57, pp. 500–508, 2015.

[6] GOUSIOS, G.; SPINELLIS, D. **Mining software engineering data from github.** IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C). IEEE, pp. 501–502, 2017.

[7] INJADAT, M.; SALO, F.; NASSIF, A. B. **Data mining techniques in social media: A survey**. Neurocomputing, vol. 214, pp. 654–670, 2016.

[8] GANDHI, N.; ARMSTRONG, L. J. **A review of the application of data mining techniques for decision making in agriculture**. 2nd International Conference on Contemporary Computing and Informatics (IC3I). IEEE, pp. 1–6, 2016.

[9] KOUTANAEI, F. N.; SAJEDI, H.; KHANBABAEI, M. **A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring**. Journal of Retailing and Consumer Services, vol. 27, pp. 11–23, 2015.

[10] van de GEER R.; WANG, Q.; BHULAI, S. **Data-driven consumer debt collection via machine learning and approximate dynamic programming.** Available at SSRN 3250755, 2018.

[11] HUNT, R. M. **Collecting consumer debt in America**. Available at SSRN 993249, 2007.

[12] CHAPMAN P., *et al.* **Crisp-dm 1.0 step-by-step data mining guide**. 2000.

[13] BREIMAN L. **Random forests**. Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[14] GISLASON, P. O.; BENEDIKTSSON, J. A.; SVEINSSON, J. R. **Random forests for land cover classification**. Pattern Recognition Letters, vol. 27, no. 4, pp. 294–300, 2006.

[15] AMRIEH, E. A.; HAMTINI, T.; ALJARAH, I. **Mining educational data to predict student's academic performance using ensemble methods**. International Journal, vol.9 No8, pp119-136, 2016.

[16] POWERS, D. M. **Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation**, Technical Report SIE-07-001, School of Informatics and Engineering, Australia, 2007.