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.

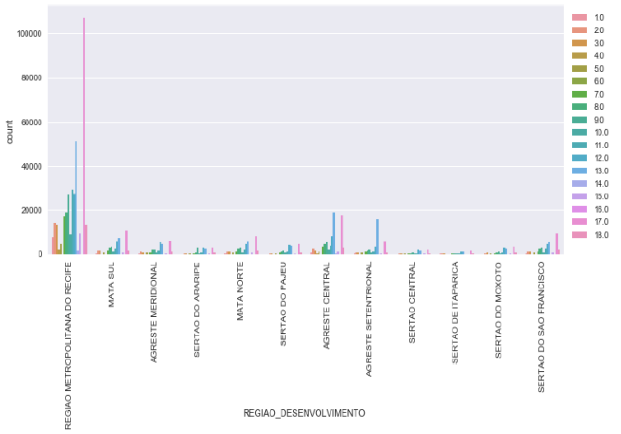

Figura 1 - Distribuições do segmento econômico por região de desenvolvimento.
Fonte**:** Autores.

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.

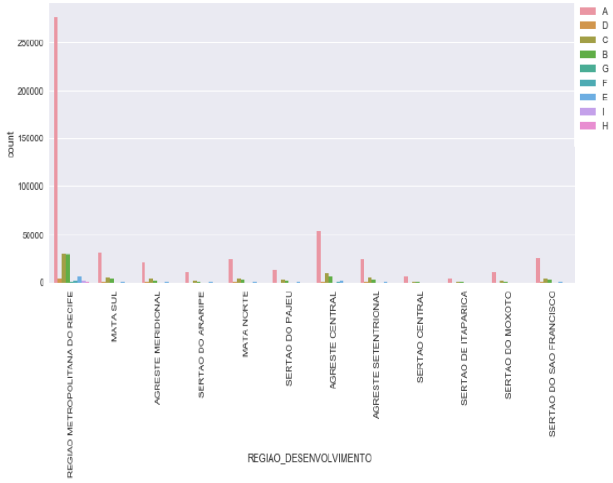

Figura 2 - Distribuições da arrecadação do período de 12 meses por região de desenvolvimento.
Fonte**:** Autores.

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.

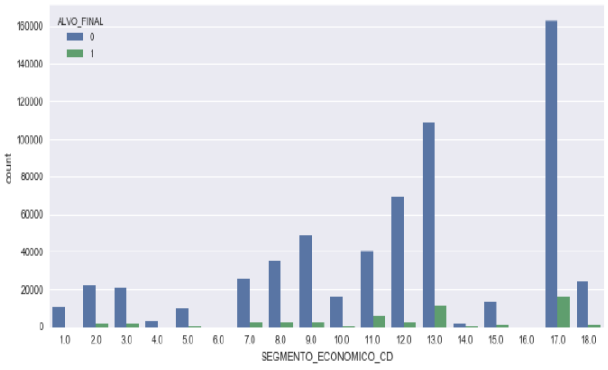

Figura 3 - Distribuições do grau de nocividade por segmento econômico (valor 0 significa lícita e valor 1 ilícita).
Fonte**:** Autores.

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.

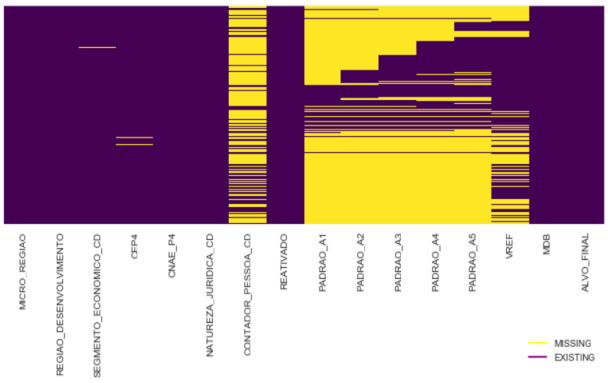

Figura 4 - Mapeamento dos dados faltantes.
Fonte**:** Autores.

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.

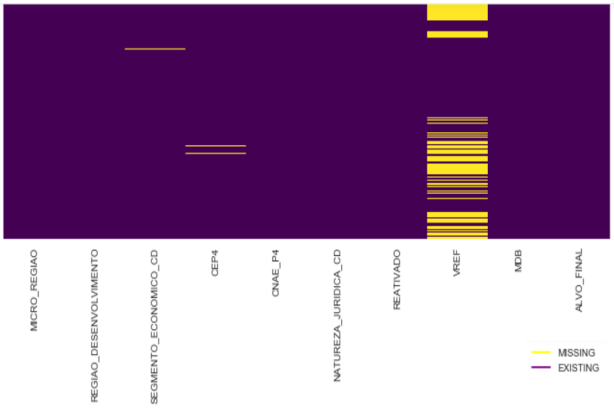

Figura 5 - Mapeamento dos dados após eliminação vertical.
Fonte**:** Autores.

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.

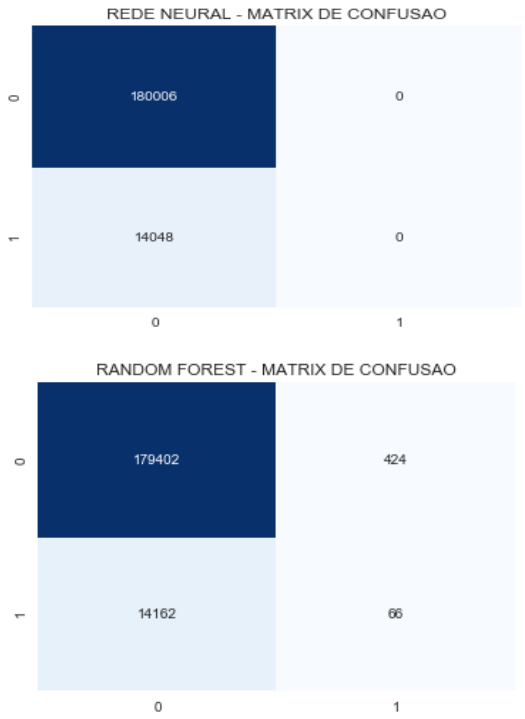

Figura 6 - Matriz de confusão para a rede neural MLP acima e matriz de confusão para Random Forest abaixo. Fonte**:** Autores.

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.

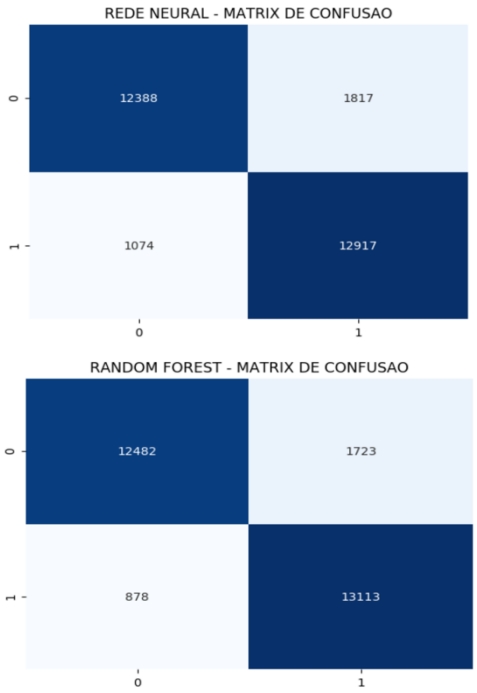

Figura 7 - Matriz de confusão para a rede neural MLP a esquerda e matriz de confusão para Random Forest a direita após o balanceamento e seleção de *features.* Fonte**:** Autores.

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.

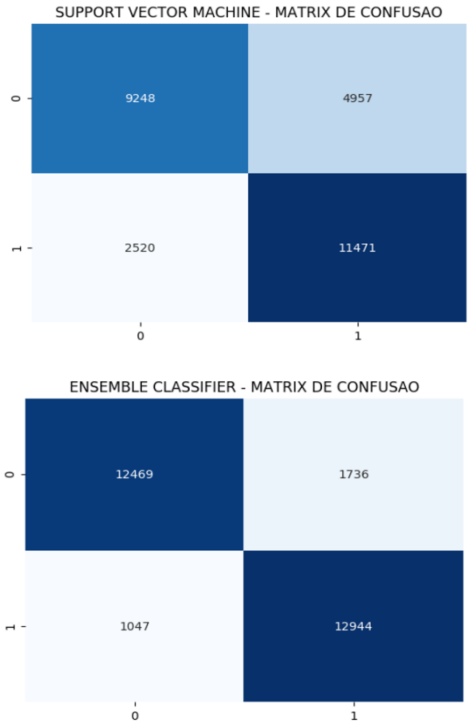

Figura 8 - Matriz de confusão para o SVM a esquerda e matriz de confusão para Ensemble Classifier a direita após o balanceamento e seleção de *features.*
Fonte**:** Autores.

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.

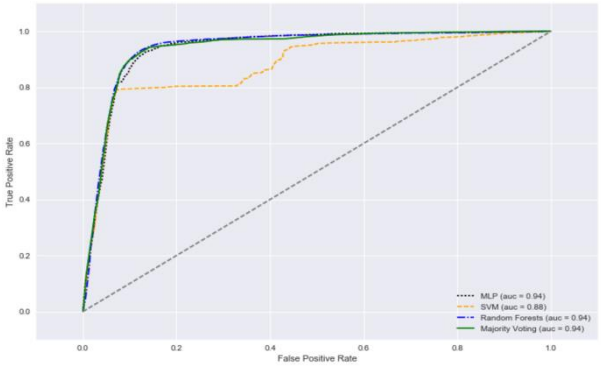

Figura 9 - Curva ROC das técnicas utilizadas no processo: MLP, Random Forest, SVM e Ensemble com Voto Majoritário.
Fonte: Autores.

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.

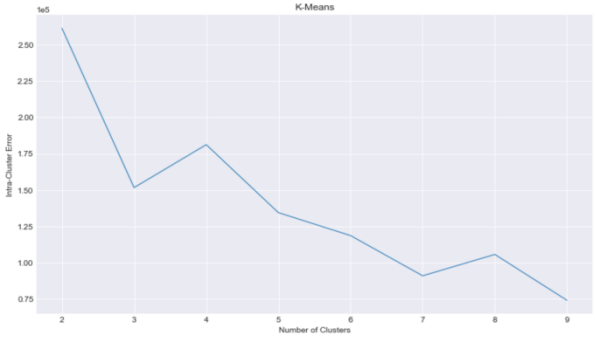

Figura 10 - Gráfico da relação entre o erro intra-*cluster* e o número de grupos para o K-Means.
Fonte**:** Autores.

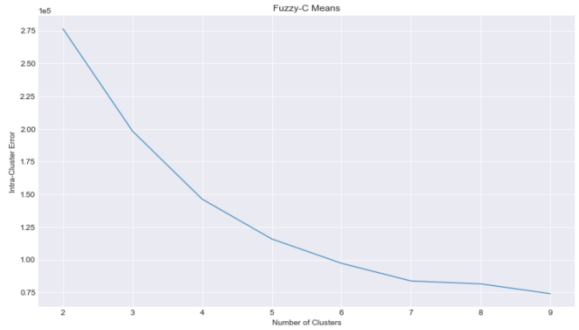

Figura 11: Gráfico da relação entre o erro intra-*cluster* e o número de grupos para o Fuzzy C-Means.
Fonte**:** Autores.

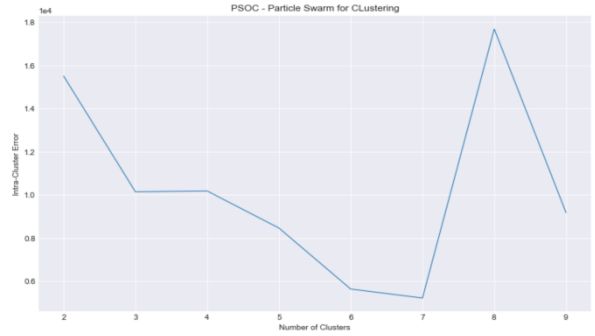

Figura 12: Gráfico da relação entre o erro intra-*cluster* e o número de grupos para o PSO.
Fonte**:** Autores.

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