

Classificação da Dívida Ativa do Estado de Sergipe Classification of Active Debt of Tax From the Sergipe State

Yúri Faro Dantas de Sant’Anna², Luiz Felipe da Conceição Souza¹,
Antônio José Alves Neto¹, Max Castor Rodrigues Junior³,
André Britto de Carvalho¹, Renê Pereira de Gusmão¹

¹ Departamento de Computação – Universidade Federal de Sergipe (UFS)
São Cristóvão – SE – Brazil

² Centro de Informática – Universidade Federal de Pernambuco
Recife, PE.

³ Departamento de Sistemas e Computação
Universidade Estácio – Aracaju, SE – Brazil,

{luizfcs, antonio.neto, andre, rene}@dcomp.ufs.br,

yfds@cin.ufpe.br, max.carojun@professores.estacio.br

Abstract. *As a way of ensuring the efficiency of the services provided, the public sector depends on revenue from tax collection. In the State of Sergipe, the Goods and Services Circulation Tax (ICMS) is responsible for more than 90% of state revenue. Therefore, the effective collection of this resource is essential for state administration. Active debt is the name for debts that are judicialized and whose payment is still pending. Many taxpayers, in active debt, are unable to honor their debts or have debts considered insoluble. The costs generated by charges directed to these elements, in addition to the non-generation of revenue for the State through these processes, result in a series of additional expenses. This work proposed the evaluation and selection of a classification algorithm that, based on the taxpayer’s behavior and the characteristics of the debt, classifies the possibility of full, partial, or non-payment of the debt. After the proposed tests, the Random Forest algorithm presented the best performance among all the options explored, obtaining an f1-score of 81.84%. In this context, the classifier provides an efficient and viable solution for the real scenario.*

Resumo. *Como forma de assegurar a eficiência dos serviços providos, o setor público depende da receita proveniente da coleta de impostos. No Estado de Sergipe, o Imposto de Circulação de Bens e Serviços (ICMS) é responsável por mais de 90% da receita estadual. Portanto, a coleta efetiva deste recurso é essencial para a administração estadual. A dívida ativa é a denominação para débitos que são judicializados e cujo pagamento ainda é pendente. Muitos contribuintes, em processo de dívida ativa, são incapazes de honrar suas dívidas ou tem débitos considerados insolúveis. Os custos gerados pelas cobranças direcionadas a esses elementos, em adição a não geração de receita para o Estado por estes processos, resultam em uma série de gastos adicionais. Este trabalho propôs a avaliação e seleção de um algoritmo de classificação que, baseando-se no comportamento do contribuinte e nas características da dívida, classifique*

a possibilidade de pagamento total, parcial ou não pagamento da dívida. Após os testes propostos, o algoritmo Random Forest apresentou o melhor desempenho dentre todas as opções exploradas, obtendo um f1-score de 81,84%. Neste contexto, o classificador fornece uma solução eficiente e viável para o cenário real.

1. Introdução

Para atender aos interesses da população, realizar investimentos e honrar os seus compromissos financeiros regulares, o Estado arrecada os seus recursos através de tributos e taxas que são cobrados do conjunto de contribuintes adequados para cada imposto e sua forma de arrecadação. Os tributos são fundamentais para o oferecimento de serviços essenciais como infraestrutura, saúde, educação e segurança. Portanto, é responsabilidade de todas as pessoas físicas ou jurídicas, que se enquadrem como contribuintes pela Lei Complementar nº 87/1996 [Brasil 1996], efetuarem o pagamento dos seus respectivos tributos, para assegurar o cumprimento das obrigações do Estado para com a sociedade [Corvalão et al. 2012].

A inadimplência de parte dos contribuintes gera dificuldades na arrecadação de algumas das fontes de recolhimento do Estado, resultando em pendências financeiras e creditárias denominadas dívida ativa. De acordo com [Soares and Oliveira 2016], “*A dívida ativa, classificada como receita corrente, é um crédito de receitas públicas provenientes do não pagamento aos cofres públicos, representando um direito a receber asseguração do Estado*”. A dívida ativa é constituída de valores oriundos de créditos tributários ou não tributários e encontra suporte na Lei Federal nº 6.830/1980 [Brasil 1980], a qual “*Dispõe sobre a cobrança judicial da Dívida Ativa da Fazenda Pública*”.

A Secretaria de Estado da Fazenda de Sergipe (SEFAZ-SE), órgão responsável pela cobrança de todos os tributos estaduais e conseqüentemente na fiscalização da dívida ativa, possui formas de cobrar e negociar os débitos administrativamente e, quando necessário, encaminhar os contribuintes em débito para a Procuradoria Geral do Estado (PGE), onde poderá ser efetuada uma cobrança por vias jurídicas. No entanto, nem sempre essas ações trazem retorno aos cofres públicos, gerando frustração de arrecadação ao Estado e custos processuais evitáveis. Esse déficit, por sua vez, pode afetar negativamente o oferecimento de serviços públicos, como saúde, educação e infraestrutura [Portella and Teixeira 2016].

O processo da dívida ativa na SEFAZ-SE, partindo da sua inscrição e considerando as possibilidades de parcelamento e execução judicial é realizado da seguinte forma:

1. Ao extrapolar o prazo de pagamento do tributo, o contribuinte é contatado pela SEFAZ-SE, que realiza a cobrança administrativa.
2. Após a cobrança administrativa, caso o contribuinte ainda não tenha regularizado sua situação, este é inscrito na dívida ativa.
3. Uma vez inscrito na dívida ativa, o débito é negociado pela SEFAZ-SE, podendo ser parcelada.
4. Caso persista a irregularidade, a dívida pode ser acrescida em multa e juros de mora.
5. A SEFAZ-SE pode executar o contribuinte judicialmente, caso em que a dívida passará a ser cobrada pela PGE, a cobrança jurídica.

6. Há a possibilidade de o contribuinte solicitar o pagamento parcelado da dívida. Nesse caso, o saldo devedor é virtualmente zerado e não pode ser judicializado enquanto o parcelamento estiver sendo devidamente honrado.
7. Caso a dívida seja parcelada, mas as parcelas não sejam efetivamente pagas, o débito restante volta para o registro da dívida ativa.

Conforme [SEFAZ-SE 2021], em 2021, houve um crescimento de 28,53% no montante da dívida ativa em relação ao ano anterior, acumulando mais de 10 bilhões de reais, dos quais apenas 20 milhões foram recuperados no mesmo ano. Destaca-se, ainda, a concentração desses fundos em apenas 0,71% dos contribuintes inscritos do Estado [SEFAZ-SE 2021], o que salienta a relevância da recuperação desses valores para a arrecadação, dada a magnitude do deficit.

Dentre as fontes de origem da dívida ativa, o montante gerado pelo tributo de Imposto Sobre Circulação de Mercadorias e Serviços (ICMS) acumula a quantia mais relevante a ser recuperada. Em 2021, o valor de ICMS correspondeu a 82,68% do total arrecadado da dívida ativa [SEFAZ-SE 2021], mostrando-se um componente de suma importância para recuperação de passivos desta receita em débito.

Para oferecer suporte ao processo de tomada de decisão, a SEFAZ-SE se vale de um aparato de *Business Intelligence* (BI), que permite o armazenamento, processamento e análise de dados da dívida ativa e de outras formas de arrecadação do Estado. Dessa forma, torna-se possível a utilização de técnicas de mineração de dados (*data mining*).

Compreendendo a importância da classificação de dívidas para auxiliar a tomada estratégica de decisão, este trabalho tem como objetivo principal selecionar, dentre um *pool* de classificadores, um modelo de detecção de padrões para realizar a categorização de registros de dívidas inscritas, tornando possível priorizar, com maior confiabilidade, a cobrança administrativa ou jurídica de dívidas com maiores chances de retorno financeiro.

O restante deste trabalho está organizado da seguinte maneira: A seção 2 apresenta trabalhos relacionados. A Seção 3 descreve a metodologia utilizada na condução do presente trabalho. A Seção 4 é dedicada à descrição do tratamento de dados e treinamento do classificador para a dívida ativa. Na Seção 5 são discutidos os resultados. Na Seção 6 é apresentada a conclusão e são propostos os trabalhos futuros. Por fim, na Seção 7 são dados os agradecimentos.

2. Trabalhos Relacionados

Nesta seção, são discutidos trabalhos relacionados à aplicação de técnicas de aprendizagem de máquina na previsão ou classificação da dívida ativa em outros Estados Federativos e contextos, como sonegação fiscal.

No trabalho de [Corvalão et al. 2012], os autores desenvolveram um modelo de regressão logística para classificar empresas contribuintes do ICMS do Estado de Santa Catarina. Baseando-se em dados das atividades econômicas, como faturamento e receita bruta, dados financeiros e registros de auditorias anteriores, foi realizada a classificação binária dos contribuintes como propensos à sonegação ou não propensos, obtendo um resultado de 71% no conjunto de validação.

Similarmente, [Rocha et al. 2017] utilizaram dados da SEFAZ-GO, referentes à contribuição de ICMS de médias e grandes empresas do setor atacadista, como: dados

cadastrais do contribuinte atacadista, registro de apuração de ICMS, registro de controle de crédito de ICMS, registro de inventário e resultado da auditoria do ICMS. Partindo desses dados, foi treinado um modelo preditivo de classificação baseado em Árvores de Decisão. O modelo tem como objetivo indicar quais empresas poderiam estar cometendo sonegação do tributo em questão. Os autores realizaram três experimentos com diferentes conjuntos de atributos, aplicando o algoritmo J48 do *software* WEKA, obtendo assim um modelo com 84% de acurácia.

Buscando aperfeiçoar o método de seleção de contribuintes do ICMS da SEFAZ-BA, [Oliveira 2012] aplicou uma rede neural artificial que busca indicar se o contribuinte tem um risco baixo, moderado ou alto de sonegar o tributo. Para isso, utilizou dados referentes ao tipo de estabelecimento do contribuinte, tipo de inscrição estadual, natureza jurídica, porte da empresa, segmento econômico, capital social e valor arrecadado. O modelo apresenta uma taxa de acurácia de 71%, demonstrando ser uma ferramenta útil para a seleção de contribuintes para fiscalização.

No estudo de [Sisnando and de Sousa Freitas 2006], os autores levantam a hipótese de que um modelo de rede neural pode detectar padrões de desempenho arrecadatório dos contribuintes do ICMS da SEFAZ do estado do Ceará e até superar o modelo estatístico em vigor na Secretaria. Para tal, utilizaram-se de dados referentes à identificação do contribuinte, localização, objeto social, tipo de sociedade, natureza jurídica, desempenho econômico e desempenho fiscal, para treinar um modelo de rede neural multicamada. De fato, o modelo proposto se mostra superior ao método em vigor utilizado nesta instituição.

Visando melhorar e reduzir custos da auditoria do Imposto de Valor Agregado (do inglês, *Value Added Tax*), [Gupta and Nagadevara 2007] utilizam técnicas de *data mining* e algoritmos de aprendizagem supervisionada para identificar contribuintes que possuem maior probabilidade de cometer evasão fiscal. Os dados utilizados se referem ao perfil do contribuinte, tendência de retorno, valor pago em imposto e variação no pagamento ao longo do tempo. Os autores treinaram 8 modelos, dos quais um modelo de *Random Forest* se destacou com uma acurácia de 86%.

Buscando detectar fraudadores do Imposto de Valor Agregado [Castellón González and Velásquez 2013] fazem uso de dados históricos dos contribuintes, dados de agentes, parceiros e representantes legais, além de características pessoais, como atividades econômicas, idade do contribuinte, tempo de atuação da companhia e patrimônio. Os contribuintes foram separados em dois grupos: “micro e pequenos” e “médios e grandes”. Em ambos os segmentos, os melhores resultados foram provenientes do modelo de rede neural, com uma acurácia de 92,6% no grupo de micro e pequenos contribuintes e 88,8% no grupo de médios e grandes.

É possível notar que a maioria dos trabalhos ligados a avaliação, classificação e previsão de tributos está relacionada à tentativa de prever o não pagamentos de impostos e reduzir, desta forma, uma possível inadimplência. Entretanto, a cobrança da dívida ativa diz respeito a contribuintes que não mostraram capacidade, ou interesse, em pagar os seus tributos devidos nem de maneira espontânea e nem coercitiva, podendo até estar, inclusive, em situação de insolvência, tornando a previsão da sua capacidade de pagamento ainda mais complexa, considerando ainda que os cadastros de contribuintes podem ser efetuados para pessoas físicas, quanto pessoas jurídicas.

Além disto, diferentemente dos estudos citados que levantam a hipótese de padrões nas características dos contribuintes, este estudo supõe a existência de padrões em atributos da própria dívida, podendo ser detectados e classificados por modelos de detecção de padrões. Além de propor uma abordagem genérica que seja aplicável à contribuintes de todos os espectros, desde empresas à pessoas físicas. Tornando a classificação, possivelmente, mais complexa.

3. Metodologia

O método utilizado neste trabalho envolveu o treinamento dos cinco algoritmos de AM citados: *k-Nearest Neighbors* (kNN), *Support Vector Machines* (SVM), *Bernoulli Naive Bayes*, *Decision Tree* e *Random Forest*. Os algoritmos foram selecionados por se tratarem de técnicas de baixo custo computacional [Duda et al. 2001], largamente utilizadas na indústria e academia para a resolução de problemas complexos sem a necessidade de hardwares específicos ou elevados tempos de treinamento [Kubat 2017]. Como evidenciam as referências expostas na seção 2 e trabalhos que visam o desenvolvimento de técnicas portáteis [Martín et al. 2013] ou aplicáveis em cenários de constante reprocessamento e recalibragem [de Sant’Anna et al. 2024].

Uma vez treinados, a seleção do modelo final realizou-se por meio da comparação de seus respectivos desempenhos para a classificação da dívida ativa, através das métricas de acurácia e *f1-score*.

Os dados utilizados foram extraídos da base de dados da SEFAZ-SE, sendo compostos por atributos numéricos e categóricos da dívida ativa, além de dados referentes aos contribuintes. Ambos os conjuntos de dados foram unidos, resultando em 132.252 amostras, contendo 67.066 dívidas não executadas judicialmente, 65.186 executadas, 90.832 dívidas não parceladas e 41.420 em parcelamento.

Antes da execução dos passos subsequentes, tornou-se necessária a remoção dos registros das dívidas em parcelamento, de forma a evitar que essas amostras, cujo valor do saldo foi zerado (não considerando-se mais o contribuinte como um devedor) em consequência da negociação, causassem algum viés durante o tratamento dos dados. Os demais passos de modelagem e tratamento são descritos nos parágrafos seguintes.

Dentre os valores da dívida ativa, é possível encontrar débitos que chegam aos milhões de reais, enquanto outros atingem apenas centenas ou milhares. Além de contribuintes que tendem a pagar suas dívidas de forma parcial, o que é, ainda que não ideal, considera um retorno financeiro válido para os cofres públicos. Devido a essa variância, notou-se a necessidade de criar uma nova variável que melhor acompanhasse as vicissitudes das características da dívida ativa, para além do binarismo “quitada” e “não quitada”. Desta forma, foram criadas quatro possíveis *status* para a quitação da dívida: “Não quitada”, “Parcialmente quitada inferior”, “Parcialmente quitada superior” e “Quitada”, estas serão melhor definidas na subseção 4.1.

O conjunto de dados final, obtido após a remoção de atributos desinteressantes para o treinamento dos algoritmos, contou com um conjunto de 25 atributos, os quais podem ser conferidos na Tabela 1.

A primeira etapa deste projeto foi o "embaralhamento" dos dados coletados, visto que estes encontram-se naturalmente armazenados de acordo com o critério temporal da

Tabela 1. Atributos presentes no conjunto de dados final

Descrição do Atributo	Tipo do Atributo
Código da Pessoa Física ou Jurídica	Discreto
Data da Primeira Inscrição na Dívida	Data
Data da Última Inscrição na Dívida	Data
Valor do Primeiro Saldo da Dívida	Contínuo
Valor do Saldo do Imposto	Contínuo
Valor do Saldo da Multa	Contínuo
Valor do Saldo da Procuradoria Geral do Estado	Contínuo
Valor da Multa na Primeira Inscrição	Contínuo
Valor do Imposto na Primeira Inscrição	Contínuo
Valor Total na Primeira Inscrição	Contínuo
Executada Judicialmente	Textual
Data da Execução	Data
Indicativo de Protesto	Textual
Valor do Protesto	Contínuo
Comarca	Textual
Vara Judicial	Textual
Data da Ação judicial	Contínuo
Valor Original do Imposto	Contínuo
Valor Original da Multa	Contínuo
Percentual Pago	Contínuo
Código da Situação do Contribuinte	Discreto
Valor do Capital Social	Contínuo
Valor da Previsão da Receita Bruta	Contínuo
Valor da Aquisição Acumulada no Ano	Contínuo
<i>Status</i> de Quitação da Dívida	Textual

dívida. Dados financeiros são suscetíveis a variações no comportamento em momentos distintos [Guralnik and Srivastava 1999]. Desta forma, tentou-se evitar que alguma das bases de dados a serem montadas posteriormente concentrasse dados de uma mesma sazonalidade específica, algo que poderia enviesar o estudo.

Visto que os algoritmos de aprendizagem de máquina não lidam com dados textuais ou em formatos específicos com datas, foi realizada, na fase de pré-processamento, a codificação desses dados. Para lidar com a alta variância das informações, os atributos numéricos contínuos, por sua vez, foram normalizados. Essa decisão visou evitar que a escala dos dados causasse viés no modelo e impactasse negativamente na generalização e previsão [Ahsan et al. 2021].

É possível notar que existe grande desbalanceamento de dados, fato que poderia acarretar tendência dos classificadores em apontar as classes com o maior número de amostras. Como forma de tratar o desbalanceamento dos dados, utilizou-se a técnica de re-amostragem. Sua aplicação consistiu na redução do número de amostras das classes maioritárias para coincidir com o número de amostras da segunda menor classe. Os detalhes são discutidos na seção 5. O processo de aumento do registros de dívidas para

diversificar e balancear os dados da menor classe do conjunto de treinamento é conhecido como *data augmentation*. Neste trabalho o processo não foi aplicado ao conjunto de teste, visto que a ideia é mantê-lo original e o classificador, comparável à outras metodologias e a um cenário de execução mais próximo da realidade.

Por fim, o desempenho de cada um dos modelos foi medido através da acurácia e *f1-score*. Na seções seguintes, serão detalhados os conceitos por trás das técnicas utilizadas e descrita a condução do trabalho.

4. Desenvolvimento do Classificador para a Dívida Ativa

Nesta seção, é descrito o processo de exploração e tratamento dos dados e o treinamento do classificador da dívida ativa. Em todas as fases, a linguagem de programação *Python* foi utilizada em sua versão 3.7. A linguagem foi escolhida pela sua diversidade de bibliotecas voltadas para *data mining* e detecção de padrões, além da abrangente documentação. Foi utilizada a biblioteca *scikit-learn* [Pedregosa et al. 2011], em sua versão 1.0.2.

O conjunto de dados original consistiu na junção entre a base de dados com atributos da dívida ativa e a base de dados com atributos de cada contribuinte. Essas duas bases foram unidas, a partir do atributo referente ao código do contribuinte, resultando em 132.252 amostras.

4.1. Base de Dados

Levando em consideração a normativa legal que torna explícita a relação entre o *status* de parcelamento e o valor zerado virtualmente do saldo da dívida, tornou-se necessária a remoção dos registros "*em parcelamento*", uma vez que os valores zerados virtualmente poderiam causar ruído em relação às dívidas quitadas de fato.

Conforme descrito na seção de metodologia, uma nova variável objetivo foi criada para uma melhor discriminação dos débitos. Essa variável foi composta por quatro possíveis *status* para a quitação da dívida: "Não quitada", quando o valor pago é R\$ 0,00 (zero reais), "Parcialmente quitada inferior", quando o valor pago é menor que o valor do saldo devedor, "Parcialmente quitada superior", quando o valor pago é maior ou igual ao valor do saldo devedor e "Quitada", quando o valor do saldo devedor é R\$ 0,00 (zero reais). Finalmente, dados referentes à quantidade de parcelamento, valor do saldo, valor pago e código de identificação dos contribuintes foram removidos.

Em seguida, procedeu-se o embaralhamento do conjunto de dados e sua divisão em dois subconjuntos. Seguindo a convenção, alocou-se 80% do conjunto de dados original para treinamento do algoritmo, enquanto que os 20% restantes foram destinados ao conjunto de teste. A Tabela 1 exibe os atributos componentes do conjunto de dados final. Na mesma tabela, é possível observar a quantidade de amostras por classe nesse conjunto de dados, além do número de amostras por classe no conjunto de teste.

É possível notar o desbalanceamento das classes. Para contornar esse problema, foi realizada uma re-amostragem, reduzindo o número de amostras das classes maioritárias para coincidir com o número de amostras da segunda menor classe, a classe "Quitada", conforme pode ser observado na Tabela 2. Em testes anteriores, o balanceamento de acordo a menor classe se mostrou pouco representativo, com apenas 1.167 registros de cada uma das classes. Realizado esse passo, foi feito o *oversampling* para

Tabela 2. Balanceamento das Classes

Classes	Nº de Amostras			
	Antes da Divisão em Treinamento e Teste	Após a Divisão em Treinamento e Teste	Classe na Base de Treinamento e Teste	Classe após o balanceamento e Data Augmentation
Não quitada	53.688	11.265	45.049	9.064
Parc. quitada inferior	21.688	4.423	17.385	9.064
Parc. quitada superior	5.835	4.668	1.167	9.064
Quitada	11.318	2.280	9.064	9.064

a classe “Parcialmente quitada inferior”, através do algoritmo *Gaussian Noise*, técnica de *data augmentation* largamente utilizada para dados de tipos heterogêneos e diversos [Arslan et al. 2019], igualando, desta forma, a quantidade de amostras de cada uma das classes.

O algoritmo *Gaussian Noise* gera ruídos estatísticos nos dados originais seguindo uma distribuição normal, de forma gerar novos registros diferentes dos originais. Como parâmetros dessa função, a media da distribuição foi configurada como 0 (zero), enquanto que o desvio padrão foi dado como 0,1.

4.2. Pré-Processamento dos Dados e Treinamento dos Algoritmos

Na fase de pré-processamento dos dados, atributos de *flag*, como “executado” e “parce-lado” tiveram seus valores substituídos por 0 ou 1. Os atributos textuais, como “comarca” e “vara”, assim como os atributos de data, foram codificados de forma que cada categoria fosse representada por um número inteiro. Os valores numéricos contínuos, por sua vez, foram normalizados para uma escala entre 0 e 1. A Figura 1 compara a escala original dos dados com a escala após a normalização.

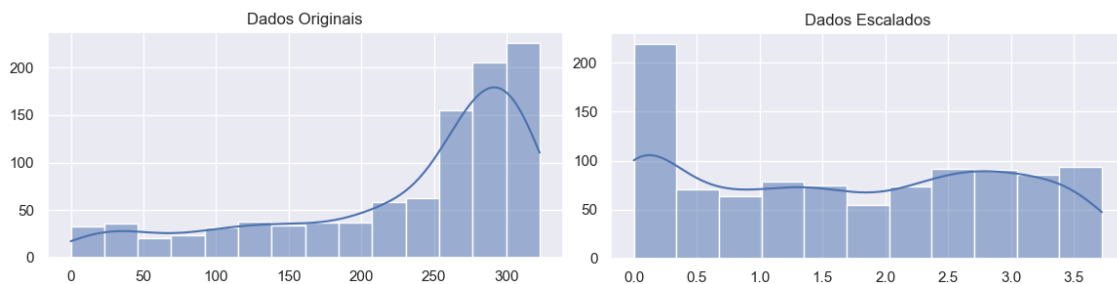


Figura 1. Comparação da Escala dos Dados Contínuos

Todos os algoritmos foram utilizados com seus parâmetros padrões, exceto o kNN, que teve seu parâmetro referente ao número de vizinhos configurado com o valor 2 (dois), que se mostrou um número satisfatório de acordo com os testes executados. Como variável objetivo, foi selecionada a variável criada com as classes: “Não quitada”, “Quitada parcialmente superior”, “Quitada parcialmente inferior” e “Quitada”.

4.3. Testes de desempenho

Finalizada a montagem e estruturação das bases de dados necessárias para este projeto, os algoritmos selecionados foram treinados utilizando o conjunto de dados discriminado para este fim. Para verificar o desempenho preditivo dos modelos treinados, foram feita as

classificações dos registros contidos no conjunto de teste. Em posse das classes reais de cada registro foram calculadas métricas de avaliação que permitiram comparar de maneira objetiva o desempenho destes modelos, sendo medidas a acurácia e *f1-score* (métrica indicada para bases desbalanceadas) em cada um dos 5 modelos. Os resultados podem ser conferidos na seção 5.

5. Resultados

A análise dos resultados da avaliação dos modelos foi realizada através da métrica de acurácia e do *f1-score*.

Os resultados da acurácia para os modelos *kNN*, *SVC*, *Bernoulli Naive Bayes*, *Decision Tree* e *Random Forest* foram de 72,75%, 71,94%, 48,56%, 72,28% e 81,14%, respectivamente. Isso demonstra que o *Random Forest* obteve o melhor desempenho nesta métrica. Para o *f1-score*, os resultados foram de 71,19% para o *kNN*, 74,52% para o *SVC*, 45,12% para o *Bernoulli Naive Bayes*, 72,54% para o *Decision Tree* e 81,84% para o *Random Forest*. Novamente, o *Random Forest* apresentou o melhor desempenho. A Tabela 3 explicita o resultado dos algoritmos e possibilita a comparação entre eles baseado nas métricas encontradas.

Tabela 3. Desempenho dos Modelos

Modelo	Acurácia	F1-Score
kNN	72,75%	71,19%
SVC	71,94%	74,52%
Bernoulli Naive Bayes	48,56%	45,12%
Decision Tree	72,28%	72,54%
Random Forest	81,14%	81,84%

Os resultados apresentados demonstram que o modelo *Random Forest* foi o que apresentou o melhor desempenho, tanto na métrica de acurácia quanto na métrica de *f1-score*. Tal performance se mostra satisfatória para a previsão da quitação ou não quitação de dívidas, tornando o modelo útil para o auxílio da tomada de decisão na priorização da cobrança.

6. Conclusão

O objetivo deste trabalho foi desenvolver um modelo de classificação da dívida ativa, tornando possível prever a possibilidade de quitação ou não quitação de uma nova dívida inscrita. Com esse intuito, foram comparados cinco algoritmos de classificação: *kNN*, *SVC*, *Bernoulli Naive Bayes*, *Decision Tree* e *Random Forest*, utilizando a acurácia e o *f1-score* como métricas. Foram analisadas as matrizes de confusão de cada algoritmo, importantes insumos para compreender como os modelos estão classificando as amostras.

Os resultados apresentados neste estudo indicam que o modelo *Random Forest* possui desempenho satisfatório, concluindo, assim, que é uma opção promissora para a classificação da dívida ativa do Estado de Sergipe. O modelo pode ser utilizado para auxiliar na tomada de decisão estratégica para a cobrança de débitos inscritos na dívida ativa, evitando custos administrativos e judiciários gerados por cobranças infrutíferas. O possível retorno financeiro, ao priorizar a cobrança de dívidas, pode ser convertido no

oferecimento de serviços públicos, trazendo benefícios à sociedade. Portanto, o trabalho realizado se mostra uma aplicação relevante de conceitos de Sistemas de Informação ao domínio governamental.

6.1. Trabalhos Futuros

Como trabalhos futuros, evidencia-se a necessidade de, junto à PGE seguir com o desenvolvimento de uma ferramenta final funcional e que atenda a área de negócio na seleção dos processos mais promissores para judicialização, integrando o modelo selecionado. Essa ferramenta deverá ter uma interface intuitiva para o usuário final, sendo capaz de indicar, com confiabilidade, se uma nova dívida inscrita pode ser quitada, não quitada ou parcialmente quitada. Deverá, também, ser capaz de indicar se alguma dívida poderá ter sua classificação alterada mediante mudanças em características da dívida. Os trabalhos futuros incluem, para implantação e análise dos resultados desta ferramenta:

- **Estudo de viabilidade da ferramenta**
Fase de validação da ferramenta desenvolvida. Será medida a efetividade da ferramenta no auxílio à seleção dos processos;
- **Análise de características mais importantes**
Estudos para se saber quais dois atributos da dívida mais influenciam em sua quitação ou não quitação;
- **Ajuste de parâmetros do algoritmo**
Configurações podem ser feitas antes do treinamento do algoritmo, para ajustar o modelo aos dados em questão, melhorando seu desempenho.

7. Agradecimentos

À SEFAZ-SE por ceder a sua infraestrutura e dados sem os quais este projeto não seria possível, além de todos os revisores que proporcionaram uma melhora no desenvolvimento deste trabalho.

Referências

- Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., and Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9(3):52.
- Arslan, M., Guzel, M., Demirci, M., and Ozdemir, S. (2019). Smote and gaussian noise based sensor data augmentation. In *2019 4th International Conference on Computer Science and Engineering (UBMK)*.
- Brasil (1980). Lei nº 6.830, de 22 de setembro de 1980. dispõe sobre a cobrança judicial da dívida ativa da fazenda pública, e dá outras providências. *Diário Oficial [da] República Federativa do Brasil*.
- Brasil (1996). Lei complementar nº 87, de 13 de setembro de 1996. dispõe sobre o imposto dos estados e do distrito federal sobre operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual e intermunicipal e de comunicação, e dá outras providências. (lei kandir). *Diário Oficial [da] República Federativa do Brasil*.

- Castellón González, P. and Velásquez, J. D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, 40(5):1427–1436.
- Corvalão, E. D. et al. (2012). Classificação de contribuintes: um modelo em duas fases.
- de Sant’Anna, Y. F. D., de Farias, M. L., Júnior, M. C., Dantas, D., and Junior, M. R. (2024). Fuel classification in electronic tax documents. In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM*, pages 337–343. INSTICC, SciTePress.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, New York, 2 edition.
- Gupta, M. and Nagadevara, V. (2007). Audit selection strategy for improving tax compliance: application of data mining techniques. In *Foundations of Risk-Based Audits. Proceedings of the eleventh International Conference on e-Governance, Hyderabad, India, December*, pages 28–30. Citeseer.
- Guralnik, V. and Srivastava, J. (1999). Event detection from time series data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–42.
- Kubat, M. (2017). *An Introduction to Machine Learning*. Springer International Publishing, Gewerbestrasse 11, 6330 Cham, Switzerland, 2 edition.
- Martín, H., Bernardos, A. M., Iglesias, J., and Casar, J. R. (2013). Activity logging using lightweight classification techniques in mobile devices. 17(4):675–695.
- Oliveira, F. N. d. (2012). Estratégias para aperfeiçoar o processo de recuperação de receitas tributárias no estado da bahia: um modelo para o icms baseado em redes neurais artificiais.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Portella, A. A. and Teixeira, R. P. (2016). Federalismo fiscal e efetividade da dignidade da pessoa humana: análise da posição do município na estrutura do financiamento público brasileiro e a escassez de recursos para as ações de saúde/fiscal federalism and effectiveness of human dignity... *Revista de Direito da Cidade*, 8(2):631–679.
- Rocha, S. M. et al. (2017). Mineração de dados aplicada à classificação dos contribuintes de icms da sefaz-go.
- SEFAZ-SE (2021). Contas anuais do estado 2021. <https://www.sefaz.se.gov.br/transparencia/Responsabilidade%20Fiscal/CONTAS%20ANUAIS%20DO%20ESTADO/2021/Contas%20Anuais%202021.pdf>. Acesso em: 23 de junho de 2023.
- Sisnando, S. R. A. and de Sousa Freitas, M. A. (2006). Previsão e avaliação do desempenho dos contribuintes do icms do estado do ceará utilizando as redes neurais artificiais. *Revista econômica do Nordeste*, 37(1):131–149.

Soares, L. E. and Oliveira, M. F. d. (2016). O protesto extrajudicial de certidão de dívida ativa como meio alternativo eficaz de recuperação do crédito público.