Predição da Inadimplência do IPVA no Estado do Ceará Utilizando Aprendizado de Máquina

Maria Inês V. Silva[1,2], **Francisco Victor da S. Pinheiro**[1]
César Lincoln C. Mattos[1], **José Maria da S. Monteiro Filho**[1]
Rossana M. C. Andrade [1]

[1]Mestrado e Doutorado em Ciência da Computação
Universidade Federal do Ceará (UFC)
Fortaleza-CE

[2]Secretaria da Fazenda do Estado do Ceará (SEFAZ-CE)

ines.vale@sefaz.ce.gov.br, victor.pinheiro.ce@alu.ufc.br

cesarlincoln@dc.ufc.br, monteiro@dc.ufc.br, rossana@ufc.br

Abstract. *For Brazilian states and municipalities, the Motor Vehicle Ownership Tax (IPVA) is the second most important source of revenue. Compliance with the collection of IPVA depends on different factors, such as the country's economy, the market value and residence of the vehicles, among other factors. Predicting whether taxpayers will be compliant or not in relation to the payment of IPVA can provide subsidies that help governments to develop public policies, planning tax actions and directing campaigns to encourage timely tax payment. In this work, we conducted a series of experiments aiming to build an efficient solution to the problem of classifying taxpayers in terms of their compliance with the IPVA payment. Real data referring to the IPVA of the State of Ceará was used from 2019 to 2023. In total, four classification algorithms were explored to classify taxpayers into two groups: compliant and non-compliant. The best results achieved an F1 score of 0.86 proving the viability of the proposed solution.*

Resumo. *Para os estados e municípios brasileiros, o Imposto sobre a Propriedade de Veículos Automotores (IPVA) é a segunda fonte mais importante de receita. A adimplência na arrecadação do IPVA depende de diferentes fatores, tais como a economia do país, o valor venal e a residência dos veículos, dentre outros fatores. Prever se os contribuintes serão adimplentes ou não em relação ao pagamento do IPVA pode fornecer subsídios que auxiliem os governos a elaborar políticas públicas, planejando ações fiscais e direcionando as campanhas de incentivo ao pagamento tempestivo do imposto. Neste trabalho, foi realizada uma série de experimentos buscando construir uma solução para o problema de classificação de contribuintes quanto à adimplência em relação ao pagamento do IPVA. Foram utilizados dados reais referentes ao IPVA do Estado do Ceará no período de 2019 a 2023. Ao todo, quatro algoritmos de classificação foram explorados para classificar os contribuintes em dois grupos: adimplentes e inadimplentes. Os melhores resultados alcançaram uma pontuação F1 de 0,86 comprovando a viabilidade da solução proposta.*

1. Introdução

Nos últimos anos, a capacidade computacional e a quantidade de dados gerados e disponíveis em todas as áreas de conhecimento têm aumentado muito rapidamente, impulsionando o desenvolvimento de técnicas e algoritmos capazes de apoiar o direcionamento de políticas públicas [Amarasinghe et al. 2023]. A área de aprendizado de máquina tem se destacado como uma ferramenta promissora para a análise de dados complexos, proporcionando suporte à decisão em diversas aplicações práticas.

O Imposto sobre a Propriedade de Veículos Automotores (IPVA) é a segunda fonte mais importante de receita para os estados e municípios brasileiros, ficando atrás apenas do Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS). Para o cálculo do IPVA, a Secretária da Fazenda (SEFAZ) toma como base o levantamento de preços anual realizado pela Fundação Instituto de Pesquisas Econômica (Fipe)[1], além dos dados do Sindicato dos Revendedores de Veículos Automotores do Ceará (Sindivel)[2]. As alíquotas utilizadas no cálculo do tributo variam entre 1% e 3,5% sobre o valor da venda dos veículos. A arrecadação do IPVA é direcionada ao próprio estado (50%) e aos seus municípios (50%). E, ao contrário do que se pensa, de acordo com o artigo 167 da Constituição Federal [Brasil 1988], o imposto estadual tem como finalidade não apenas a melhoria das vias, podendo ser utilizado também em áreas como saúde, educação, cultura e segurança,ou seja, é um tributo desvinculado.

A adimplência na arrecadação do IPVA depende de diversos fatores como a própria economia do país, o valor venal e a residência dos veículos, dentre outros fatores [Governo do Estado do Ceará 2023]. Prever se os contribuintes serão adimplentes ou não em relação ao pagamento do IPVA pode fornecer subsídios que auxiliem os governos a elaborar políticas públicas, planejando ações fiscais e campanhas específicas para aumentar a arrecadação desse importante imposto

Este trabalho é o relato de uma série de experimentos que podem auxiliar na classificação dos contribuintes quanto à adimplência em relação ao pagamento do IPVA. Foram utilizados dados reais referentes ao IPVA do Estado do Ceará no período de 2019 a 2023. Ao todo, quatro algoritmos clássicos para tarefas de classificação, abrangendo diferentes categorias, foram explorados para classificar os contribuintes em dois grupos: adimplentes e inadimplentes. Os algoritmos avaliados foram: *Random Forest - RF*, *Logistic Regression - LR*, *Cat Boost* e *Multilayer Perceptron - MLP*. Os melhores resultados alcançaram uma pontuação F1 de 0,86, comprovando a viabilidade da solução proposta. A classificação dos contribuintes nas classes "adimplente" e "inadimplente" permite ainda que os auditores da SEFAZ possam não apenas compreender padrões presentes nos dados históricos, como também prever antecipadamente a inadimplência do pagamento de IPVA.

O restante deste artigo está organizado da seguinte forma. A Seção 2 discute os trabalhos relacionados. Na Seção 3, a metodologia utilizada no desenvolvimento deste trabalho é apresentada em detalhes, já mostrando os resultados da análise exploratória dos dados e descrevendo a avaliação experimental realizada. A Seção 4 discute os resul-

[1] `https://veiculos.fipe.org.br/`
[2] `https://www.sindivel.com.br/`

tados dos experimentos relacionados à classificação dos contribuintes. Por fim, a Seção 5 apresenta as conclusões desta pesquisa e aponta direções para trabalhos futuros.

2. Trabalhos Relacionados

Na literatura, diversos estudos exploram a avaliação de modelos de aprendizagem de máquina com o objetivo de solucionar problemas do setor público, utilizando dados abertos. A seguir, algumas dessas pesquisas, voltadas para questões fiscais e tributárias, são relatadas.

O trabalho de [Lima and Delen 2020] explora vários potenciais preditores para Índices de Percepção de Corrupção em 132 países, a partir da perspectiva da análise preditiva, utilizando técnicas de aprendizado de máquina. Em contraste com estudos anteriores baseados em correlações e análises estatísticas, este estudo emprega modelos não lineares para alcançar alta precisão preditiva. Entre os algoritmos testados, o *Random Forest* apresentou maior precisão na predição/classificação, seguido por SVM e Redes Neurais Artificiais.

O trabalho de [Silva et al. 2022] aborda a tributação do comércio de bens e serviços, este trabalho descreve a aplicação de algoritmos clássicos de aprendizado supervisionado para identificar padrões de comportamento circular no comércio envolvendo contribuintes do Estado de Goiás, por meio da análise de suas operações de compra e venda de bens e serviços. Os experimentos obtiveram resultados indicaram que o algoritmo KNN [Faceli 2011] foi a técnica mais precisa, considerando as características específicas do contexto brasileiro.

Em [Akinrinola et al. 2024], os autores realizam uma revisão abrangente da evolução dos modelos de predição de arrecadação fiscal, destacando a importância de previsões fiscais precisas no planejamento econômico e realizando uma análise crítica dos estudos anteriores, estabelecendo assim uma base sólida para a compreensão do estado atual e das tendências na utilização do aprendizado de máquina na área de tributação. As principais conclusões revelam que, embora os modelos de aprendizado de máquina apresentem resultados animadores na predição de arrecadação, ainda existem desafios relacionados à complexidade dos dados, à interpretabilidade dos modelos, bem como a considerações éticas.

Em comparação aos trabalhos relacionados, a proposta deste artigo é descrever uma solução que viabilize a classificação de contribuintes quanto à adimplência em relação ao pagamento do IPVA. Para isso, quatro diferentes algoritmos de aprendizagem de máquina foram explorados. Adicionalmente, este estudo visa analisar padrões e tendências nos dados. Os resultados obtidos irão contribuir para aprimorar a gestão fiscal, direcionando ações a serem tomadas para o incentivo do pagamento dentro do prazo previsto.

3. Metodologia Utilizada

Nesta seção, é apresentada a metodologia utilizada neste trabalho, a qual é estruturada em seis etapas, conforme ilustra a Figura 1, já sendo detalhados os resultados de cada uma delas.

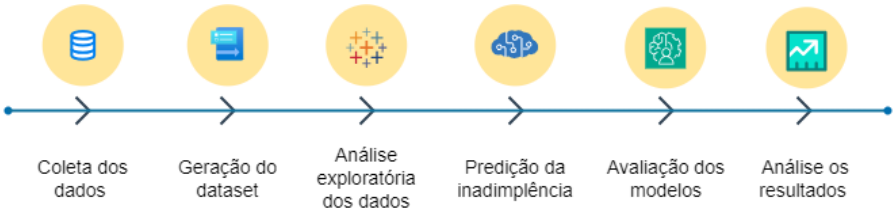

Figura 1. Fluxo metodológico para a execução deste trabalho.

3.1. Coleta dos Dados e Geração do *Dataset*

A extração de dados foi realizada a partir da base de dados *Oracle* do sistema de IPVA, considerando os anos de 2019 a 2023. As linhas foram extraídas em arquivos no formato csv, incluindo informações como identificador único do veículo, CPF do proprietário, ano de fabricação, tipo e descrição do veículo, valor do IPVA, indicador de inadimplência considerando o último dia do ano do imposto, indicador se o IPVA foi pago parcelado e o município do veículo.

Adicionalmente, foram agregadas informações que podem influenciar na adimplência, como um indicador de sociedade de empresa do proprietário e dados municipais de PIB, renda per capita e região dentro do estado. A partir desses dados, foi criado um atributo indicando se o veículo pertence à região metropolitana de Fortaleza. Como os dados de PIB e renda per capita para 2022 e 2023 não estavam disponíveis no IBGE, esses valores foram estimados usando um modelo de regressão linear, tendo como base histórica os anos de 2012 a 2021. Além disso, os valores do imposto, de 2019 a 2022, foram corrigidos pelo Índice Nacional de Preços ao Consumidor Amplo - IPCA antes das próximas etapas.

3.2. Análise Exploratória dos Dados

Após a extração inicial, que considerou todos os tipos de veículos e gerou um total de 11.246.726 linhas, foi realizada uma análise preliminar dos dados. Esta etapa envolve o exame inicial dos dados coletados com o objetivo de identificar padrões, tendências e anomalias. Esse exame foi conduzido por meio de estatísticas descritivas, que ajudam na compreensão básica da distribuição e correlação entre diferentes variáveis. Essa análise foi fundamental para orientar as etapas subsequentes de modelagem e análise, assegurando que os dados sejam compreendidos em maior profundidade antes da aplicação de técnicas preditivas.

Inicialmente, foram carregados todos os tipos de veículos. Porém, os tipos "Micro-ônibus", "Caminhão Trator", "Ônibus", "Triciclo", "Motor Home", "Trator de rodas", "Ciclomotor" e "Semirreboque" foram removidos da amostra por representarem menos de 1% da arrecadação prevista total, que é próxima de 2 bilhões de reais ao ano. Desse modo, os tipos de veículos considerados no estudo foram apenas: "Automóvel", "Caminhão", "Caminhonete", "Camioneta", "Motocicleta", "Motoneta" e "Utilitário" e, a partir deste momento, os totais referenciados sempre serão relativos apenas a esses últimos tipos citados. Após a remoção dos veículos que pouco contribuem para a arrecadação total, o conjunto de dados ficou com as quantidades de veículos apresentadas na Tabela 1, valores utilizados para a análise dos dados brutos a seguir.

Tabela 1. Quantidade de veículos com pagamento previsto de IPVA por ano.

Ano	Qtd. de Veículos com imposto taxado	Aumento/Redução com relação ao ano anterior
2019	2.174.057	-
2020	2.230.326	+2,9%
2021	2.257.461	+1,22%
2022	2.259.469	+0,09%
2023	2.249.282	-0,45%

Desdobrando essas quantidades totais de veículos por ano e por tipo de veículo, é possível visualizar na Figura 2 a evolução das quantidades normalizadas de diferentes tipos de veículos ao longo dos anos. A normalização temporal foi realizada utilizando o ano de 2019 como base para analisar qual tipo de veículo cresceu ou decresceu em vendas durante os cinco anos observados. O comportamento das curvas, especialmente no período mais crítico da pandemia de COVID-19 no Brasil e logo após, onde aparece o acréscimo na aquisição de utilitários, caminhonetes e motocicletas e um decréscimo na compra de automóveis, pode significar um aumento nos serviços de entregas.

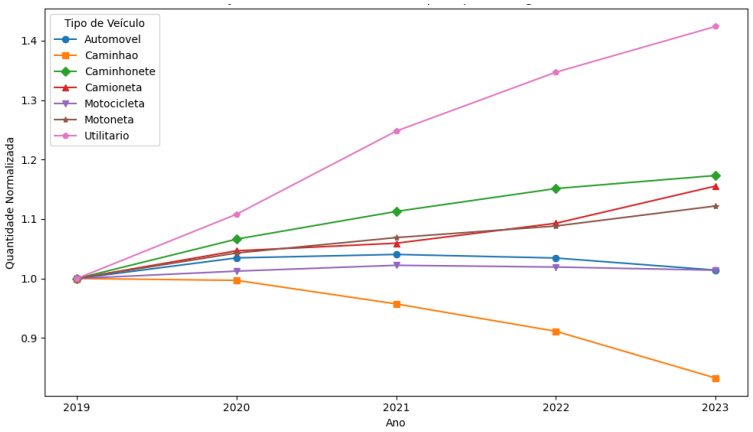

Figura 2. Observação do acréscimo/decréscimo da aquisição de veículos por tipo e ano.

Falando sobre a distribuição dos veículos por região do estado, foi observado que a região metropolitana de Fortaleza reteve, em 2023, cerca de 43% dos veículos do estado, detendo 62% dos automóveis e apenas 13% do total de motocicletas e motonetas. As motocicletas sempre representam o maior número de veículos em todas as regiões do interior do estado.

A Figura 3 apresenta os valores calculados de IPVA, atualizados pelo IPCA acumulado até 2023, para cada ano, além da quantidade de veículos e o percentual de inadimplência observado ao final de cada ano listado. De acordo com esta tabela, os automóveis sempre representam, em valores reais, o maior percentual esperado da arrecadação. Com relação à inadimplência, as motocicletas e motonetas sempre saem na frente. Até o final de cada ano, a média de inadimplência, considerando todos os tipos de veículos citados, é cerca de 25%.

Os próximos passos antes da execução dos modelos incluíram a escolha das variáveis independentes, o pivotamento de valores anuais para cada veículo, a separação do

Ano	Tipo Veículo	Valor IPVA Total (R$)	% Total IPVA no Ano	Qtde. Veículos	% Inadimplência no Ano
2019	Automóvel	678.497.583,77	55,2	724.829	19,83
2019	Motocicleta	190.376.398,15	15,49	1.129.243	44,48
2019	Caminhonete	176.707.940,22	14,38	91.376	20,52
2019	Utilitário	90.570.404,30	7,37	25.669	15,97
2019	Camioneta	69.118.299,57	5,62	37.582	21,14
2019	Motoneta	12.458.855,43	1,01	153.020	43,07
2019	Caminhão	11.516.185,96	0,94	12.338	20,8
	Soma	**1.229.245.667,40**		**2.174.057**	
2020	Automóvel	697.908.934,39	54,81	749.966	20,28
2020	Motocicleta	195.578.701,16	15,36	1.143.265	44,16
2020	Caminhonete	185.960.511,34	14,61	97.428	21,03
2020	Utilitário	98.021.150,50	7,7	28.444	17,79
2020	Camioneta	71.170.665,75	5,59	39.334	22,14
2020	Motoneta	13.308.216,57	1,05	159.589	43,69
2020	Caminhão	11.301.214,16	0,89	12.300	19,36
	Soma	**1.273.249.393,87**		**2.230.326**	
2021	Automóvel	697.276.196,64	53,84	754.171	19,96
2021	Motocicleta	199.467.141,53	15,4	1.154.392	39,92
2021	Caminhonete	195.417.720,61	15,09	101.685	18,96
2021	Utilitário	108.866.265,85	8,41	32.033	15,57
2021	Camioneta	69.635.014,45	5,38	39.815	20,85
2021	Motoneta	13.889.686,06	1,07	163.554	37,93
2021	Caminhão	10.610.378,29	0,82	11.811	17,84
	Soma	**1.295.162.403,43**		**2.257.461**	

Ano	Tipo Veículo	Valor IPVA Total (R$)	% Total IPVA no Ano	Qtde. Veículos	% Inadimplência no Ano
2022	Automóvel	823.047.482,82	52,17	749.816	21,49
2022	Caminhonete	249.951.043,41	15,84	105.202	20,15
2022	Motocicleta	245.615.364,35	15,57	1.151.024	42,1
2022	Utilitário	143.359.855,46	9,09	34.575	17,67
2022	Camioneta	85.435.083,92	5,42	41.074	22,06
2022	Motoneta	17.775.613,94	1,13	166.535	40,16
2022	Caminhão	12.409.354,13	0,79	11.243	19,36
	Soma	**1.577.593.798,03**		**2.259.469**	
2023	Automóvel	898.245.165,27	50,9	735.068	20,68
2023	Motocicleta	282.287.025,08	16	1.145.078	39,84
2023	Caminhonete	281.671.083,42	15,96	107.205	18,95
2023	Utilitário	166.918.713,87	9,46	36.549	17,14
2023	Camioneta	101.536.174,99	5,75	43.427	21,21
2023	Motoneta	21.074.890,69	1,19	171.682	36,97
2023	Caminhão	12.917.626,11	0,73	10.273	17,01
	Soma	**1.764.650.679,43**		**2.249.282**	

Figura 3. Valores agregados por ano e tipo de veículo

conjunto em dados de treino e teste, o balanceamento dos dados de treino, a normalização dos valores e a execução dos algoritmos de classificação.

3.3. Predição da Inadimplência

Devido ao grande volume de dados, onde um veículo poderia aparecer em cinco linhas diferentes, uma para cada ano, a execução dos modelos de aprendizado de máquina ficou inviável com os recursos de máquina disponíveis. Desse modo, algumas estratégias foram utilizadas para reduzir o conjunto de treino. Inicialmente, cada veículo foi pivotado (agrupado), gerando uma única linha, contendo informações de 2019 a 2023, tais como o próprio valor cobrado de IPVA a cada ano e o valor do PIB anual do município do veículo. Esta abordagem, ao contrário dos métodos clássicos de séries temporais, tem a vantagem de incorporar naturalmente múltiplas variáveis contextuais e a sua relação entre si durante o treino. Dados pivotados simplificam a estrutura dos dados, permitindo que os modelos de aprendizado de máquina se concentrem em variáveis mais relevantes e diretas, em vez de tentar discernir padrões em um conjunto de dados redundantes e, possivelmente, ruidosos.

De toda forma, para que as informações fossem utilizadas em sua disposição original, onde uma linha contém os atributos de um veículo para cada ano de cobrança, além do número de linhas ser significativamente maior, seria necessária uma intervenção específica para séries com valores históricos, onde deveriam ser avaliados os componentes de tendência, sazonalidade e aleatoriedade, além de serem tratados os *outliers* e ruídos do *dataset*.

Ao final do pivotamento, o conjunto utilizado como entrada para os modelos ficou com 1.730.079 linhas, uma massa de dados ainda considerada muito grande para a execução dos modelos. Nesse conjunto, foram retirados os veículos que não tinham IPVA cobrado em algum dos anos. Mesmo assim, cerca de 75% dos veículos apareceram nos cinco anos de amostra e foi possível obter bons resultados ao final. A Figura 4 apresenta os atributos utilizados como entrada, além da variável dependente.

Na separação dos dados em conjuntos de treino e teste, a distribuição na proporção 80 e 20% resultou, respectivamente, em 1.384.063 e 346.016 linhas. Para viabilizar a

Atributo	Qtde. Valores para cada Veículo (ou linha)	Tipo
TipoVeiculo	1	categórica multivalorada
eCapital	1	categórica binária
ESocioAno	5 (de 2019 a 2023)	categórica binária
Parcelado	5 (de 2019 a 2023)	categórica binária
Pib	5 (de 2019 a 2023)	float64
IdadeVeiculo	5 (de 2019 a 2023)	int64
Populacao	5 (de 2019 a 2023)	int64
PerCapita	5 (de 2019 a 2023)	float64
ValorIPVA	5 (de 2019 a 2023)	float64
InadimplenteAno (alvo)	5 (de 2019 a 2023)	categórica binária

Figura 4. Atributos utilizados nos algoritmos.

execução dos modelos na fase de treino, foi realizada uma redução em 75% dos dados de treino, com outra subamostragem em seguida, utilizando a técnica *Random Undersampling*, resultando em 242.702 linhas de treino, mantendo-se as linhas de teste intactas.

Antes do balanceamento, com relação à variável alvo, ”InadimplenteAno", a classe que representa os adimplentes representava 65% do conjunto de treino. Depois do balanceamento, as classes ficaram com a mesma representatividade. Depois disso, os dados de treino e teste foram normalizados, com o estimador *Column Transformer* da biblioteca *scikit-learn*[3]. As colunas categóricas multivaloradas foram transformadas antes com *One Hot Encoding* e a padronização dos valores numéricos utilizou a classe *Standard Scaler*. Os atributos binários não precisam ser normalizados. É importante pontuar que o atributo multivalorado TipoVeiculo é desdobrado em sete atributos após a aplicação do *One Hot Encoding*, o que resulta em 47 (quarenta e sete) atributos de entrada para os experimentos.

Ao todo, quatro algoritmos de aprendizado de máquina foram escolhidos para treinamento e teste com os dados históricos de pagamento de IPVA, devido à sua eficácia comprovada em problemas de classificação, especialmente envolvendo atributos categóricos [Hastie et al. 2009]. Os modelos mencionados na seção de Introdução foram implementados a partir da biblioteca Python *scikit-learn* (classes *RandomForestClassifier*, *LogisticRegression*) e *MLPClassifier*), além da biblioteca catboost da Yandex (classe *CatBoostClassifier*).

Para cada experimento realizado, foi utilizada a técnica de validação cruzada, com 10 *Folds*, combinada com o *Grid Search* para encontrar os hiperparâmetros ótimos de cada modelo. Foi realizada uma tentativa de configurar a estratégia de divisão de validação cruzada como *StratifiedKFold*, o que dispensa a necessidade do balanceamento preliminar do conjunto de dados, porém não foi observada melhoria nos resultados. A estratificação é especialmente importante para problemas de classificação com classes desbalanceadas, pois mantém a proporção de classes em cada dobra de validação, mas não surtiu o efeito esperado.

Após os melhores hiperparâmetros terem sido encontrados, os modelos foram executados sobre o conjunto completo dos dados de treino (sem balanceamento) e, em seguida, essa nova instância treinada foi aplicada nos dados de teste, antes já normaliza-

[3]`https://scikit-learn.org/stable/index.html`

dos. A partir dessa última execução é que as métricas de desempenho, tais como acurácia, precisão, revocação e medida-F1, foram calculadas. Para a avaliação dos algoritmos e a análise dos resultados, detalhadas na próxima seção, os desempenhos foram comparados utilizando-se as métricas citadas, visando identificar o modelo mais assertivo na predição da inadimplência do IPVA. Também foram calculadas as matrizes de confusão para a execução do melhor modelo com os dados de teste, para que se entenda melhor o impacto das taxas de falsos positivos e falsos negativos.

4. Avaliação dos Modelos e Análise dos Resultados

Nesta seção, serão apresentadas as métricas de desempenho obtidas após a execução dos quatro algoritmos avaliados, considerando os melhores hiper-parâmetros obtidos na execução do *Grid Search* com a validação cruzada de 10 iterações.

Na avaliação dos modelos para o problema de classificação de inadimplência, é muito importante analisar os Falsos Positivos (FP) e Falsos Negativos (FN) encontrados, uma vez que podem ser tomadas ações indevidas pelo fisco. Por exemplo, a partir de falsos positivos, podem ser disparadas notificações ou cobranças injustas, trazendo desconforto ou insatisfação para alguns contribuintes. Já um falso negativo representa um contribuinte inadimplente que foi considerado adimplente. Isso pode levar as autoridades fiscais a pecarem por falta de ações para recuperar os valores devidos aos cofres do estado. No caso, a métrica considerada mais adequada para executar o *Grid Search* foi a revocação, uma vez que é o mais importante para o estado seria reduzir os falsos negativos. Uma revocação alta para o problema posto significa que o modelo está identificando a maioria dos inadimplentes corretamente.

Os resultados dos modelos testados com os dados de pagamento de IPVA de 2019 a 2023 foram muito semelhantes, com a média da medida-F1 superior a 82% e precisão média maior que 84% para todos os algoritmos executados, sendo o *Cat Boost* o algoritmo campeão, reafirmando que é particularmente eficaz quando aplicado a dados categóricos. Esta alta precisão geral demonstra a eficácia dos modelos em prever bons resultados para o conjunto de dados apresentado.

Para o algoritmo *Random Forest - RF*, os melhores valores para os hiperparâmetros, encontrados após 160 execuções, foram: *bootstrap: True*, *max_depth: 30*, *max_features: sqrt*, *min_samples_leaf: 4*, *min_samples_split: 10* e *n_estimators: 300*. A Figura 5 mostra os resultados obtidos pelo algoritmo *Random Forest - RF*, após a seleção dos melhores hiperparâmetros.

Classe	Precisão	Revocação	Medida F1	Suporte
0	0,87	0,92	0,89	224.422
1	0,83	0,75	0,79	121.594
Acurácia			0,86	346.016
Média macro	0,85	0,83	0,84	346.016
Média ponderada	0,86	0,86	0,86	346.016

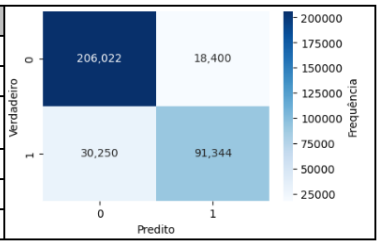

Figura 5. Métricas e matriz de confusão obtidas na execução do algoritmo *Random Forest - RF*.

Já para o algoritmo *Logistic Regression - LR*, os melhores valores para os hiperpa-

râmetros, encontrados após 550 execuções, foram: C: 0.01, *l1_ratio: 0.5*, *max_iter: 500*, *penalty: elasticnet* e *solver: saga*. A Figura 6 ilustra os resultados obtidos pelo algoritmo *Logistic Regression - LR*, após a seleção dos melhores hiperparâmetros.

Classe	Precisão	Revocação	Medida F1	Suporte
0	0,87	0,91	0,89	224.422
1	0,81	0,75	0,78	121.594
Acurácia			0,85	346.016
Média macro	0,84	0,83	0,83	346.016
Média ponderada	0,85	0,85	0,85	346.016

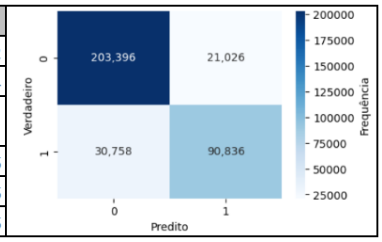

Figura 6. Métricas e matriz de confusão obtidas na execução do algoritmo *Logistic Regression - LR*.

No experimento realizado com o algoritmo *Cat Boost*, os melhores valores para os hiperparâmetros, encontrados após 120 execuções, foram: *border_count: 128*, *depth: 4*, *iterations: 100*, *l2_leaf_reg: 3* e *learning_rate: 0.2*. A Figura 7 ilustra os resultados obtidos pelo algoritmo *Cat Boost*, após a seleção dos melhores hiperparâmetros.

Classe	Precisão	Revocação	Medida F1	Suporte
0	0,87	0,92	0,90	224.422
1	0,83	0,75	0,79	121.594
Acurácia			0,86	346.016
Média macro	0,85	0,84	0,84	346.016
Média ponderada	0,86	0,86	0,86	346.016

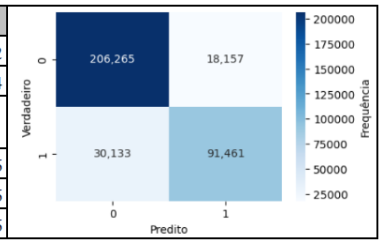

Figura 7. Métricas e matriz de confusão obtidas na execução do *Cat Boost*.

Para a rede neural *Multilayer Perceptron - MLP*, os melhores valores para os hiperparâmetros, encontrados após 320 execuções, foram: *activation: relu*, *alpha: 0.0001*, *hidden_layer_sizes: (50)*, *learning_rate: constant* e *solver: adam*. A Figura 8 ilustra os resultados obtidos pelo algoritmo *Multilayer Perceptron - MLP*, após a seleção dos melhores hiperparâmetros.

Classe	Precisão	Revocação	Medida F1	Suporte
0	0,87	0,92	0,89	224.422
1	0,83	0,75	0,79	121.594
Acurácia			0,86	346.016
Média macro	0,85	0,84	0,84	346.016
Média ponderada	0,86	0,86	0,86	346.016

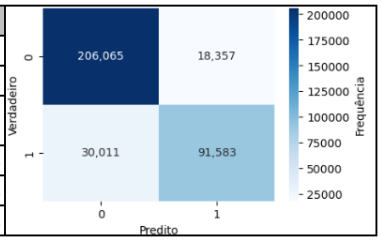

Figura 8. Métricas e matriz de confusão obtidas pela rede neural *Multilayer Perceptron - MLP*.

De acordo com as matrizes de confusão geradas, pode-se afirmar que o percentual de falsos positivos, ou seja, os contribuintes adimplentes que foram considerados inadimplentes, foi pequeno (cerca de 8%) para todos os algoritmos avaliados, o que não geraria grandes transtornos para os contribuintes. Já o o percentual do falsos negativos, para a

maioria dos algoritmos investigados, foi mais elevada, ficando em torno de 25%, mesmo utilizando a revocação como métrica do *Grid Search*, o que pode indicar a necessidade de avaliar diferentes estratégias para a seleção de hiperparâmetros ou a inclusão de outros atributos dependentes.

5. Conclusão e Trabalhos Futuros

Neste trabalho, foi realizada uma série de experimentos buscando construir uma solução para o problema de classificação de contribuintes quanto à adimplência em relação ao pagamento do IPVA. Para isso, foi construído um conjunto de dados reais contendo informações referentes ao IPVA do Estado do Ceará no período de 2019 a 2023. Ao todo, quatro algoritmos clássicos para tarefas de classificação foram explorados para classificar os contribuintes em dois grupos: adimplentes e inadimplentes. Os melhores resultados alcançaram uma pontuação F1 de 0,86, comprovando a viabilidade da solução proposta.

Como trabalhos futuros, podem ser introduzidas variáveis socioeconômicas ou que permitam a análise segmentada por tipo de veículo. O Índice de Desenvolvimento Humano (IDH), a renda da pessoa física e informações sobre outras dívidas do proprietário podem melhorar significativamente a precisão dos modelos, proporcionando uma visão mais detalhada do perfil dos contribuintes. Analisar a inadimplência por tipo de veículo pode ajudar a identificar padrões específicos e desenvolver políticas segmentadas. Por exemplo, se motocicletas têm uma taxa de inadimplência mais alta, estímulos específicos, como um maior parcelamento, poderiam ser implementados para esse grupo.

Referências

Akinrinola, O., Addy, W. A., Ajayi-Nifise, A. O., Odeyemi, O., and Falaiye, T. (2024). Application of machine learning in tax prediction: A review with practical approaches. *Global Journal of Engineering and Technology Advances*, page 102–117.

Amarasinghe, K., Rodolfa, K. T., Lamba, H., and Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data 38; Policy*, 5:e5.

Brasil (1988). Constituição da República Federativa do Brasil. Acesso em: 25 de maio de 2024.

Faceli, K. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen - LTC.

Governo do Estado do Ceará (2023). Sefaz divulga tabela do IPVA 2024, que apresenta redução média de 4,59 Acesso em: 28 de maio de 2024.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Lima, M. S. M. and Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*, 37(1):101407.

Silva, D., Carvalho, S. T., and Silva, N. (2022). Comparative analysis of classification algorithms applied to circular trading prediction scenarios. In Kő, A., Francesconi, E., Kotsis, G., Tjoa, A. M., and Khalil, I., editors, *Electronic Government and the Information Systems Perspective*, pages 95–109, Cham. Springer International Publishing.