UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Algoritmos de Classificação e Aprendizado de Máquina para Identificação de Empresas de Fachada

Ingrid Alves Novais dos Santos

Monografia - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

Ingrid Alves Novais dos Santos

Algoritmos de Classificação e Aprendizado de Máquina para Identificação de Empresas de Fachada

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador : Prof. Dr. Vicente Garibay Cancho

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A474a

Alves Novais dos Santos, Ingrid
Algoritmos de Classificação e Aprendizado de Máquina para Identificação de Empresas de Fachada / Ingrid Alves Novais dos Santos; orientador Vicente Garibay Cancho. -- São Carlos, 2025.
96 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Aprendizado de Máquina. 2. Classificação. 3. Empresas de Fachada. 4. Administração Tributária. 5. Fiscalização. I. Garibay Cancho, Vicente, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Ingrid Alves Novais dos Santos

Classification Algorithms and Machine Learning to Identify Shell Companies

Monograph presented to the Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Data Science.

Concentration area: Data Science

Advisor: Prof. Dr. Vicente Garibay Cancho

Original version

São Carlos

2025

Dedico esse trabalho à minha família (Darcio, Bruno, Clara e Eunice),
que tanto amo e apoia em todos os desafios que me proponho
e nos que aparecem pelo caminho.

AGRADECIMENTOS

Agradeço a Deus em quem acredito estar toda a sabedoria e conhecimento.

Agradeço ao meu marido e filhos, que suportaram minha ausência e atuaram colaborativamente para que essa tarefa fosse cumprida. Vocês são a familha ideal para mim.

Agradeço à minha mãe, que me deu o primeiro impulso para chegar até aqui e que mesmo com sua saúde comprometida, soube esperar e me apoiar.

Agradeço aos colegas de trabalho que me acompanharam nesse curso, em especial Sylvia e Marcelo. Sem a ajuda de vocês não seria possível concluir no prazo proposto.

Finalmente, agradeço às instituições SEFAZ-SP e USP-São Carlos que promovem com essa parceria a difusão do conhecimetno para enriquecimento do serviço público.

"...Da determinação, que tens tomada,
Não tornes por detrás, pois é fraqueza
Desisitir-se da cousa começada."
Luiz Vaz de Camões. Os Lusíadas (1572)

RESUMO

Santos, I. A. N. dos Santos **Algoritmos de Classificação e Aprendizado de Máquina para Identificação de Empresas de Fachada**. 2025. 94 p. Monografia (MBA em Ciências de Dados) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Este trabalho teve como objetivo a obtenção de um modelo de classificação baseado em aprendizado de máquina para identificar empresas de fachada no Estado de São Paulo. Ele se baseia no fato de que esse comportamento é disseminado e causa enorme prejuízo ao interesse público. Foram testados diversos modelos utilizando variáveis cadastrais e socioeconômicas, respeitando o sigilo fiscal e a legislação de proteção de dados. A variável alvo é a situação cadastral da empresa, definida como ativa (1) ou inapta (0). Os algoritmos testados foram K-Nearest Neighbors, Florestas Aleatórias, Gradient Boosting, Regressão Logística e Naive Bayes. O modelo de Florestas Aleatórias sem nenhum tratamento apresentou as melhores métricas, tanto no treinamento quanto na classificação de dados inéditos, destacando-se como uma ferramenta simples e eficaz para a detecção de empresas fraudulentas, destacando-se como uma ferramenta valiosa para a fiscalização. Este modelo apresenta baixo custo computacional, boa precisão e é pouco propenso a overfitting, podendo ser combinado com outros algoritmos no futuro. Em um grupo de 452 empresas, o modelo identificou corretamente 134 das 143 empresas inaptas, alcançando uma precisão de quase 94%. No entanto, é importante se observar que houve uma quantidade não desprezível de falsos negativos que deverão ser tratados, eventualmente. A importância das variáveis socioeconômicas e do CEP atual foi destacada, sugerindo que a inclusão de mais variáveis desse tipo pode melhorar ainda mais o modelo. Embora os modelos testados tenham mostrado bons resultados, exceto o Naive Bayes, a escolha inicial das variáveis foi coerente e influente no desempenho. Sugere-se aprofundar o modelo de Florestas Aleatórias com estratégias adicionais, como histórico de conformidade e padrões de documentos fiscais, além de testar novos algoritmos e ajustar as variáveis. O trabalho conclui que o modelo pode ser continuamente aprimorado com a aplicação prática e a ajuste fino da fiscalização, tornando-o cada vez mais eficaz..

Palavras-chave: Classificação. Aprendizado de Máquina. Empresas de Fachada. Documentos Fiscais. Estado de São Paulo. Variáveis Cadastrais. Variáveis Socioeconômicas. Sigilo Fiscal. Proteção de Dados. Algoritmos. Precisão. Robustez. Falsos Negativos. Compliance. Aprimoramento Contínuo.

ABSTRACT

Santos, I. A. N. dos Santos **Classification Algorithms and Machine Learning to Identify Shell Companies**. 2025. 94 p. Monograph (MBA in Data Sciences) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

This work aimed to develop a classification model based on machine learning to identify shell companies in the State of São Paulo. It is based on the fact that this behavior is widespread and causes enormous damage to public interest. Various models were tested using cadastral and socioeconomic variables, respecting fiscal confidentiality and data protection legislation. The target variable is the company's registration status, defined as active (1) or inactive (0). The algorithms tested were K-Nearest Neighbors, Random Forests, Gradient Boosting, Logistic Regression, and Naive Bayes. The untreated Random Forest model presented the best metrics, both in training and in classifying new data, standing out as a simple and effective tool for detecting fraudulent companies, and proving to be a valuable tool for inspection. This model has low computational cost, good accuracy, and is not prone to *overfitting*, and it can be combined with other algorithms in the future. In a group of 452 companies, the model correctly identified 134 of the 143 inactive companies, achieving an accuracy of almost 94%. However, it is important to note that there was a non-negligible amount of false negatives that will need to be addressed eventually. The importance of socioeconomic variables and the current postal code was highlighted, suggesting that the inclusion of more variables of this type could further improve the model. Although the tested models showed good results, except for Naive Bayes, the initial choice of variables was coherent and influential in performance. It is suggested to deepen the Random Forest model with additional strategies, such as compliance history and fiscal document patterns, as well as testing new algorithms and adjusting the variables. The work concludes that the model can be continuously improved with practical application and fine-tuning of inspection, making it increasingly effective.

Keywords: Classification. Machine Learning. Shell Companies. Fiscal Documents. State of São Paulo. Cadastral Variables. Socioeconomic Variables. Fiscal Confidentiality. Data Protection. Algorithms. Accuracy. Robustness. False Negatives. Compliance. Continuous Improvement.

LISTA DE FIGURAS

Figura 1 – Matriz de Confusão 46
Figura 2 – Fluxograma do Método CRISP-DM 57
Figura 3 – Distribuição - CEP's para cada grupo de empresas. 58
Figura 4 – Distribuição - CNAE's para cada grupo de empresas. 60
Figura 5 – Distribuição - Sócios que já pertenceram ao quadro da empresa. 60
Figura 6 – Distribuição - Contadores que já prestaram serviço à empresa. 61
Figura 7 – Distribuição - Quantidades de Clientes. 62
Figura 8 – Distribuição - Razão - Quantidade de Clientes/ Total de Clientes. 62
Figura 9 – Distribuição - Fornecedores. 63
Figura 10 – Distribuição - Razão - Quantidade de Fornecedores/ Total de Fornecedores. 64
Figura 11 – Distribuição - Razão - Quantidade de Fornecedores/ Quantidade de Clientes. 64
Figura 12 – Distribuição - Capital Social, submetidos a transformação logarítmica. . . . 65
Figura 13 – Distribuição - Valores totais das Notas Fiscais Eletrônicas Emitidas, submetidos a transformação logarítmica. 66
Figura 14 – Distribuição - ICMS de Notas Fiscais Emitidas. 66
Figura 15 – Distribuição - Valores dos CT-e's submetidos a transformação logarítmica . 67
Figura 16 – Distribuição - Quantidade de CT-e's. 68
Figura 17 – Distribuição - Valores de Notas Fiscais Recebidas, submetidos a transformação logarítmica. 69
Figura 18 – Comparação da Razão de Valores de Notas Fiscais Recebidas/ Valores Notas Fiscais Emitidas. 70
Figura 19 – Comparação da Razão do ICMS destacado nas Notas Fiscais Recebidas/ ICMS das Notas Fiscais Emitidas. 70
Figura 20 – Mapa de Calor das Variáveis do Grupo de Ativas 72
Figura 21 – Mapa de Calor das Variáveis do Grupo de Inaptas 72
Figura 22 – Mapa de Calor das Variáveis do Grupo de Mistas 72
Figura 23 – Matriz de Confusão - K-Nearest Neighbors 75
Figura 24 – Importância das Variáveis - KNN 76
Figura 25 – Matriz de Confusão - Florestas Aleatórias 77
Figura 26 – Importância das Variáveis - RF 77
Figura 27 – Matriz de Confusão - Gradient Boosting 78
Figura 28 – Importância das Variáveis - GB 79
Figura 29 – Matriz de Confusão - Regressão Logística 80
Figura 30 – Importância das Variáveis - RL 81
Figura 31 – Matriz de Confusão - Naive Bayes 81
Figura 32 – Importância das Variáveis - NB 82

Figura 33 – Matriz de Confusão - K-Nearest Neighbors (Dados Novos) 84
Figura 34 – Matriz de Confusão - Florestas Aleatórias (Dados Novos) 85
Figura 35 – Matriz de Confusão - Gradient Boosting 86
Figura 36 – Matriz de Confusão - Regressão Logística (Dados Novos) 87
Figura 37 – Matriz de Confusão - Naive Bayes 87
Figura 38 – Créditos de ICMS transferidos por empresas inidôneas 89

LISTA DE TABELAS

Tabela 1 – Distribuição percentual de CEP's para cada grupo de empresas. 59
Tabela 2 – Distribuição percentual de quant_socios para as empresas. 60
Tabela 3 – Distribuição percentual de quant_contadores para as empresas 61
Tabela 4 – Parâmetros Estatísticos do Capital Social das Amostras 65
Tabela 5 – Parâmetros Estatísticos dos Valores de NF-e's emitidas (total_emit) 66
Tabela 6 – Parâmetros Estatísticos dos Valores de NF-e's recebidas (total_dest). . . . 69
Tabela 7 – Métricas do Modelo K-Nearest Neighbors para os dados em treinamento . . 75
Tabela 8 – Métricas do Modelo Florestas Aleatórias para os dados em treinamento . . 76
Tabela 9 – Métricas do Modelo Gradient Boosting para os dados em treinamento . . . 78
Tabela 10 – Métricas do Modelo Regressão Logística para os dados em treinamento . . . 79
Tabela 11 – Métricas do Naive Bayes para os dados em treinamento 81
Tabela 12 – Métricas do Modelo K-Nearest Neighbors para dados inéditos 83
Tabela 13 – Métricas do Modelo Florestas Aleatórias para dados inéditos 84
Tabela 14 – Métricas do Modelo Gradient Boosting para dados inéditos 85
Tabela 15 – Métricas do Modelo Regressão Logística para dados inéditos 86
Tabela 16 – Métricas do Modelo Naive Bayes para dados inéditos 87
Tabela 17 – Melhores Métricas Obtidas nas Classificações 90

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

CRISP-DM	Processo Padrão Intersetorial para Mineração de Dados (Cross-industry standard process for data mining).
CEP	Código de Endereçamento Postal.
CNAE	Classificação Nacional de Atividade Econômica.
CT-e	Conhecimento de Transporte Eletrônico.
ICM	Imposto Sobre Circulação de Mercadorias.
ICMS	Imposto Sobre Circulação de Mercadorias e Serviços de Comunicação e Transporte Interestadual.
KNN	K-Vizinhos mais próximos (K-Nearest Neighbors).
RF	Florestas Aleatórias (RF).
GB	Gradient Boosting.
RL	Regressão Logística.
NB	Naive Bayes.
NF-e	Nota Fiscal Eletrônica.
IBGE	Instituto Brasileiro de Geografia e Estatística.
SEFAZ-SP	Secretaria da Fazenda e Planejamento do Estado de São Paulo.
RFB	Receita Federal do Brasil.
SF	Secretaria da Fazenda
NCM	Nomenclatura Comum do Mercosul.
IA	Inteligência Artificial.
CAT	Coordenadoria da Administração Tributária.
CNPJ	Cadastro Nacional da Pessoa Jurídica.
USP	Universidade de São Paulo.
USPSC	Campus USP de São Carlos.

SUMÁRIO

1 INTRODUÇÃO 27
1.1 História da Tributação de Mercadorias no Brasil 27
1.2 O ICMS 28
1.3 Complexidade Tributária e Evasão Fiscal 29
1.4 Empresas Fictícias ou Empresas de Fachada 29
1.5 Estratégias de Fiscalização 31
1.6 Uso de Machine Learning 31
1.7 Modelos de Classificação 32

2 REVISÃO TEÓRICA 35
2.1 Modelos de Classificação Utilizados 35
2.1.1 K-Nearest Neighbors (KNN) 35
2.1.1.1 Processo de Treinamento do KNN 35
2.1.1.2 Vantagens do KNN: 35
2.1.2 Florestas Aleatórias - Random Forests (RF) 37
2.1.2.1 Árvores de Decisão 37
2.1.2.2 Processo de Treinamento das Florestas Aleatórias 39
2.1.2.3 Vantagens do Modelo de Florestas Aleatórias 39
2.1.2.4 Desvantagens do Modelo 40
2.1.3 Gradient Boosting (GB) 40
2.1.3.1 Processo de Treinamento do Gradient Boosting 40
2.1.3.2 Vantagens do Gradient Boosting 41
2.1.3.3 Desvantagens 41
2.1.4 Regressão Logística (RL) 42
2.1.4.1 Vantagens da Regressão Logística 43
2.1.4.2 Desvantagens da Regressão Logística 43
2.1.5 Naive Bayes (NB) 44
2.1.5.1 Vantagens do Modelo Naive Bayes 45
2.1.5.2 Desvantagens 45
2.2 Métricas de Avaliação 45
2.2.1 Acurácia 46
2.2.2 Precisão 46
2.2.3 Revogação 46
2.2.4 F1-Score 47
2.3 Refinamento 47
2.4 Trabalhos Relacionados 48

2.5 Legislação Aplicada . . . 53

3 OBJETIVO DO TRABALHO . . . 55

4 METODOLOGIA . . . 57
4.1 Entendimento do Negócio . . . 57
4.2 Entendimento dos Dados . . . 57
4.2.1 Dados Cadastrais . . . 58
4.2.2 Dados Econômico-Fiscais . . . 61
4.2.3 Correlação entre as variáveis . . . 71
4.2.4 Preparação do Dados . . . 73
4.3 Treinamento dos Algoritmos . . . 74
4.4 Avaliação dos Modelos . . . 74
4.4.1 KNN - K-NearestNeighbors . . . 74
4.4.2 Florestas Aleatórias (RF) . . . 76
4.4.3 Gradient Boosting (GB) . . . 78
4.4.4 Regressão Logística (RL) . . . 79
4.4.5 Naive Bayes (NB) . . . 81
4.5 Deployment - Implantação da solução . . . 82
4.5.1 KNN - K-NearestNeighbors . . . 83
4.5.2 Florestas Aleatórias (RF) . . . 84
4.5.3 Gradient Boosting (GB) . . . 85
4.5.4 Regressão Logística (RL) . . . 86
4.5.5 Naive Bayes (NB) . . . 87

5 CONCLUSÃO . . . 89

REFERÊNCIAS . . . 93

1 INTRODUÇÃO

A tributação no Brasil e no mundo é um sistema em constante transformação e sua evolução tem grande relevância para a compreensão das dinâmicas econômicas e sociais ao longo dos séculos.

Desde os primeiros tributos na Antiguidade, passando pelo desenvolvimento dos sistemas tributários na Idade Média e Moderna, até os dias atuais, a tributação tem sido um instrumento crucial para o financiamento do Estado e a promoção do bem-estar social.(Júnior, 2012)

No Brasil, a trajetória tributária reflete as transformações políticas e econômicas do país, ao longo dos anos, culminando na recente reforma tributária que busca simplificar e tornar mais justa a arrecadação de impostos.

A análise de grandes volumes de dados permite identificar padrões e anomalias que seriam de difícil detecção com os métodos tradicionais.

Algoritmos de Machine Learning podem contribuir para a prevenção de comportamentos de contribuintes, sugerir ações preventivas e aumentar a eficiência e a eficácia das políticas tributárias.(Nembe J. K., 2024)

Essas tecnologias não apenas modernizam a gestão tributária, mas também representam um avanço significativo no processo histórico da tributação. Já agora e, principalmente, no futuro, a utilização de dados e inteligência artificial promete transformar ainda mais o cenário tributário, possibilitando uma administração mais transparente, equitativa e adaptada às necessidades da sociedade. (Júnior, 2012)

A reforma tributária, em curso no Brasil, não tem opção senão incorporar essas inovações, com o potencial de criar um sistema mais eficiente e justo, aproveitando ao máximo os dados disponíveis para a tomada de decisões informadas e estratégicas.

Com a ciência de dados e o Machine Learning, as administrações tributárias têm à disposição ferramentas poderosas para otimizar a arrecadação e combater, sonegação fiscal, harmonicamente com a indispensável modernização. (Nembe J. K., 2024)

1.1 História da Tributação de Mercadorias no Brasil

O início da história da tributação no Brasil se confunde com a de mercadorias e origina-se do período colonial sincronizada com a evolução econômica e política do país. (Júnior, 2012)

Como primícias dessa atividade a Coroa Portuguesa impôs tributos sobre produtos importados, por meio de uma Carta Régia em 1603. (Júnior, 2012)

Após a sua independência, em 1822, o Brasil começou a desenvolver um sistema tributário próprio. A Proclamação da República em 1889 adotou uma estrutura federativa em que estados e municípios ganharam maior autonomia na arrecadação de tributos. (Júnior, 2012)

Já, a Constituição de 1891 estabeleceu base para a criação de impostos estaduais e municipais. (Júnior, 2012)

A promulgação da Constituição de 1988 foi um marco na reorganização do sistema tributário nacional. Entre as várias inovações, destacou-se a criação do Imposto sobre Circulação de Mercadorias e Serviços (ICMS), que substituiu o Imposto sobre Circulação de Mercadorias (ICM) da Constituição anterior. (Brasil, 1988)

1.2 O ICMS

ICMS é a sigla definida para o imposto sobre operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual, intermunicipal, de comunicação. Destaca-se que nessa seara incluem-se energia elétrica e combustíveis, como mercadorias. É um tributo de competência dos Estados e do Distrito Federal e sua regulamentação está prevista na Lei Complementar 87/1996. (Brasil, 1996)

Como tributo de competência estadual se estabelece como a principal fonte de receita dos estados brasileiros e de municípios de menor porte, neste último caso, decorrente das transferências estabelecidas na Constituição Federal. (Campanelle, 2012)

O imposto é conhecido por sua complexidade, devido às diferentes alíquotas aplicadas por cada estado e às várias isenções e incentivos fiscais existentes decorrentes de uma guerra fiscal constante entre os entes federativos. Apesar de sua considerável importância na arrecadação, sempre foi alvo de críticas devido a essa complexidade e seu correspondente impacto sobre o consumo e a produção. A guerra fiscal entre os estados, que concedem diferentes benefícios para atrair investimentos de empresas para si, cria um ambiente de incerteza jurídica e aumenta os custos de conformidade para as empresas. (Teixeira, 2024)

Além disso, o ICMS por ser um imposto cumulativo é frequentemente apontado como um fator que contribui para a alta carga tributária no Brasil, afetando a competitividade das mercadorias brasileiras no mercado interno e externo. (Campanelle, 2012)

Todos esses pontos ressaltam que a complexidade e impacto econômico do ICMS contribuem para a evasão fiscal, de forma que o trabalho de orientação e fiscalização é fundamental para assegurar a arrecadação tão necessária para o desenvolvimento econômico e social do país. (Marcos Vinícius dos Santos, Hugo Leonardo Menezes de Carvalho, Lucimar Antonio de Cabral de Ávila, 2019)

1.3 Complexidade Tributária e Evasão Fiscal

Cada estado brasileiro, desde que atendendo os dispositivos do artigo 155 da Constituição Federal e seus incisos, têm a autonomia para definir suas alíquotas de ICMS. (Brasil, 1988) Autonomia essa que resulta em uma diversidade de tributos aplicados sobre produtos e serviços.

Essas variações criam distorções no mercado, incentivando práticas de planejamento tributário agressivas e a busca por jurisdições com alíquotas mais favoráveis. (Mendonça, 2021) A diferença nas alíquotas pode levar empresas a mudarem suas operações para estados com tributação mais baixa, a tal prática conhecida como "guerra fiscal."

As políticas praticadas pelos estados, já mencionadas, de isenções e incentivos fiscais complicam ainda mais o cenário tributário. Essas, embora, visem promover o desenvolvimento econômico local, muitas vezes resultam em perda de receita e aumento da complexidade tributária. (Teixeira, 2024). A falta de harmonização na concessão de benefícios fiscais entre estados torna o cumprimento das obrigações tributárias um desafio ainda maior para as empresas.

Essa guerra não só dificulta a conformidade tributária, como também cria oportunidades para a evasão e sonegação fiscal. A evasão fiscal refere-se às práticas ilegais usadas para evitar o pagamento de tributos, enquanto a sonegação fiscal envolve a omissão deliberada de informações para reduzir a carga tributária. A complexidade do ICMS favorece esses comportamentos, uma vez que as empresas podem explorar brechas na legislação ou utilizar documentos falsos para reduzir sua carga tributária. (Mendonça, 2021)

A falta de clareza e a constante mudança nas normas do ICMS proporcionam um terreno fértil e empresas mal intencionadas exploram brechas na legislação tributária para minimizar sua carga fiscal. (Marcos Vinícius dos Santos, Hugo Leonardo Menezes de Carvalho, Lucimar Antonio de Cabral de Ávila, 2019) Um exemplo e a declaração a menor do valor das mercadorias ou realizar operações fictícias entre filiais de diferentes estados para aproveitar alíquotas mais baixas.

Levando em conta os problemas causados pela complexidade e o comportamento irregular, torna-se oportuno o uso do que há de moderno na tecnologia da informação para sua solução.

1.4 Empresas Fictícias ou Empresas de Fachada

A prática que se quer explorar nesse trabalho é a falsificação de documentos fiscais, as chamadas notas fiscais frias. A utilização de documentos falsos é uma tática recorrente na sonegação do ICMS, permitindo que as empresas reduzam artificialmente o valor do tributo devido. (Mendonça, 2021)

A emissão de notas fiscais frias é forma corriqueira e bastante encontrada em estudos. Ela ocorre através de empresas simuladas, fictícias ou empresas “de fachada”. São elas entidades criadas com o propósito de simular operações comerciais legítimas, mas que na realidade não possuem atividade econômica real.

A criação de empresas de fachada para a emissão de notas fiscais frias é uma das mais impactantes práticas ilegais para se obter vantagens financeiras de uma forma geral.

Em particular, emitindo notas fiscais frias, as empresas podem declarar despesas fictícias, reduzindo assim a base de cálculo dos tributos devidos. (Gomes, 2023)

Além disso, essas empresas são frequentemente utilizadas para lavagem de dinheiro. Ao emitir notas fiscais frias, os criminosos podem justificar a origem de recursos ilegais, integrando-os ao sistema financeiro formal. (Gomes, 2023)

No entanto, a vantagem indevida mais comum, e que mais nos interessa, é a primeira. Pois, através dessas emissões geram-se créditos fiscais indevidos para as empresas. Elas criam notas fiscais frias e obtêm créditos de impostos, como o ICMS. (Gomes, 2023)

Esse esquema funciona com a criação de documentos fiscais falsos que registram compras ou serviços que nunca ocorreram. Quando uma empresa emite uma nota fiscal fria, ela declara ao fisco que realizou uma compra ou contratou um serviço, gerando um crédito de ICMS correspondente ao valor, suposta e falsamente pago anteriormente, para essa transação fictícia. Esses créditos podem ser usados para abater o próprio ICMS em operações futuras, reduzindo assim o montante de impostos a ser pago. Em alguns casos, as empresas podem até solicitar a restituição desses créditos ao fisco, o chamado Crédito Acumulado, obtendo um retorno financeiro direto.

A evasão e sonegação do ICMS têm impactos significativos tanto econômicos quanto sociais. A perda de arrecadação compromete a capacidade dos estados de investir em serviços públicos essenciais, como saúde, educação e infraestrutura. A sonegação fiscal do ICMS representa uma perda significativa de receita, que poderia ser utilizada para melhorar a qualidade de vida da população.

Também contribui para a desigualdade e a injustiça fiscal. A evasão do ICMS beneficia empresas que adotam práticas ilegais em detrimento daquelas que cumprem suas obrigações tributárias, criando um ambiente de concorrência desleal, prejudicando ainda mais o bem-estar social, uma vez que, em decorrência disso, a carga tributária acaba sendo transferida para os consumidores finais, que pagam preços mais altos pelos produtos e serviços. (Mendonça, 2021)

Esse último ponto implica que além de reduzir a carga tributária, essa prática pode distorcer a concorrência no mercado. Empresas que operam legalmente e pagam todos os seus impostos acabam em desvantagem competitiva em relação àquelas que utilizam créditos fiscais indevidos para reduzir seus custos operacionais. Aí está a concorrência

desleal e empresas que seguem as regras são prejudicadas. (Mendonça, 2021)

Vale ressaltar que a prática de obtenção de créditos fiscais indevidos também pode ter implicações legais severas. Empresas envolvidas nesse tipo de fraude podem enfrentar multas significativas, processos judiciais e até mesmo a prisão de seus responsáveis. Além disso, a reputação da empresa pode ser gravemente afetada, resultando em perda de confiança por parte de clientes, fornecedores e investidores.

Por fim, a emissão de notas fiscais frias pode ser utilizada para manipular demonstrações financeiras, apresentando uma situação econômica mais favorável do que a real. Isso pode atrair investidores ou facilitar a obtenção de financiamentos de forma a aumentar a concorrência disleal e incentivar ainda mais práticas ilegais.

1.5 Estratégias de Fiscalização

A utilização de operações fictícias e a criação de empresas de fachada para minimizar a carga tributária representam desafios significativos para as administrações tributárias. Sendo assim, os orgãos fiscais têm adotado uma série de métodos para fiscalizar e combater a evasão fiscal por meio de empresas simuladas. Entre as práticas adotas encontram-se auditorias detalhadas, análise de risco baseada em big data e até mesmo a cooperação internacional para troca de informações fiscais. (Nembe J. K., 2024)

As auditorias detalhadas continuam sendo uma ferramenta tradicional e essencial na batalha contra a evasão fiscal. Essas auditorias envolvem a revisão minuciosa de registros contábeis, documentos fiscais e transações financeiras para identificar possíveis fraudes e inconsistências.

Cada vez mais eficaz e necessária é a utilização de tecnologias avançadas, como big data e inteligência artificial para a identificação de padrões suspeitos e na análise de risco. Essas tecnologias permitem às administrações tributárias analisarem grandes volumes de dados em busca de inconsistências e transações atípicas, facilitando a identificação das empresas que possam estar envolvidas nessas práticas irregulares. (Nembe J. K., 2024)

1.6 Uso de Machine Learning

O aprendizado de máquina, conforme descrito no livro “Pattern Recognition and Machine Learning” de Christopher M. Bishop (2006), é um campo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem aos computadores aprender a partir de dados. O objetivo principal é criar sistemas que possam melhorar seu desempenho em tarefas específicas através da experiência. (Bishop, 2006)

Segundo esse autor, o aprendizado de máquina envolve a construção de modelos matemáticos que capturam padrões nos dados. Esses modelos são treinados usando conjuntos de dados de treinamento, onde o algoritmo ajusta seus parâmetros para minimizar o erro

de previsão. Existem diferentes tipos de aprendizado de máquina, incluindo aprendizado supervisionado, onde o modelo é treinado com pares de entrada e saída conhecidos, e aprendizado não supervisionado, onde o modelo tenta encontrar padrões ou estruturas ocultas nos dados sem rótulos explícitos. (Bishop, 2006)

O aprendizado de máquina é, além de tudo, uma área interdisciplinar que combina estatística, matemática e ciência da computação. Diante disso, mostra-se adequado o uso dessa tecnologia para enfrentar os desafios já mencionados de evasão fiscal para a identificação dessas empresas fraudulentas.

Por fim, como exemplo concreto e demonstração de que este é um caminho sem volta, o uso de tecnologias da informação para detecção de fraudes fiscais já acontece e não está restrito ao meio acadêmico. A Receita Federal do Brasil tem desenvolvido tecnologias inovadoras que utilizam inteligência artificial e análise de redes complexas para detectar fraudes fiscais. Além disso, a modernização da fiscalização com o uso de Big Data e machine learning tem sido destacada como uma ferramenta essencial para melhorar a eficiência na detecção de fraudes e na arrecadação tributária. (RFB, 2024)

As aplicações destacadas pelo trabalho do fisco federal foram: Irregularidades Tributárias em Grupos Econômicos e Importações; Fraudes com Criptomoedas; e Identificação de Fraudes em Pedidos de Ressarcimento. (RFB, 2024)

1.7 Modelos de Classificação

Os modelos de classificação em machine learning são ferramentas poderosas que permitem a categorização de dados em classes predefinidas, desempenhando um papel crucial em diversas aplicações, desde diagnósticos médicos até a detecção de fraudes financeiras. (Bishop, 2006)

A escolha do algoritmo e das métricas apropriadas para cada trabalho depende do tipo de dados e do problema específico a ser resolvido.

O avanço contínuo em machine learning tem permitido a criação de modelos cada vez mais precisos e eficientes, contribuindo significativamente para a inovação tecnológica e a solução de problemas complexos.

O processo de classificação começa com a coleta e preparação dos dados, que são divididos em conjuntos de treinamento e teste. O conjunto de treinamento é utilizado para ensinar o algoritmo, enquanto o conjunto de teste avalia a performance do modelo. Durante o treinamento, o algoritmo ajusta seus parâmetros internos para minimizar a diferença entre as previsões e as classificações reais dos dados de treinamento. (Bishop, 2006)

Os modelos de classificação em machine learning podem ser divididos em duas categorias principais: supervisionados e não supervisionados. Cada abordagem tem suas

próprias características e aplicações, sendo escolhida de acordo com a natureza dos dados e os objetivos do problema a ser resolvido.

Nos modelos supervisionados, o algoritmo é treinado utilizando um conjunto de dados rotulados, ou seja, dados que já possuem uma classificação conhecida. Durante o treinamento, o modelo aprende a associar as características dos dados às suas respectivas classes. Depois do aprendizado e teste da classificação, pode comparar os resultados aos resultados reais. Essa abordagem é amplamente utilizada em tarefas como reconhecimento de padrões, diagnóstico médico e detecção de fraudes (Bishop, 2006).

Por outro lado, os modelos não supervisionados trabalham com dados não rotulados, ou seja, dados que não possuem uma classificação prévia. O objetivo desses modelos é identificar padrões ou agrupamentos naturais nos dados. Essa abordagem é útil em situações onde a rotulagem dos dados é impraticável ou muito custosa. (Lindholm, 2022)

Esses modelos são frequentemente utilizados em tarefas de segmentação de mercado, análise de redes sociais e compressão de dados.

A escolha entre modelos supervisionados e não supervisionados depende de vários fatores, incluindo a disponibilidade de dados rotulados, o custo de rotulagem e a natureza do problema. Em muitos casos, uma combinação de ambas as abordagens pode ser utilizada para obter melhores resultados. Por exemplo, técnicas de aprendizado semi-supervisionado combinam um pequeno conjunto de dados rotulados com um grande conjunto de dados não rotulados para melhorar a performance do modelo.

Esse trabalho utilizará um modelo supervisionado tendo como parâmetro a experiência já existente na SEFAZ-SP, se propondo a utilizar um modelo treinado por machine learning para encontrar empresas inidôneas.

A seguir será apresentado o desenvolvimento do trabalho contendo uma revisão teórica de trabalhos relacionados a esse, legislação aplicada e explanação sobre os modelos utilizados.

Na sequência a metodologia utilizada e , finalment, as conclusões de acordo com os resultados.

2 REVISÃO TEÓRICA

2.1 Modelos de Classificação Utilizados

2.1.1 K-Nearest Neighbors (KNN)

O modelo de classificação K-Nearest Neighbors (KNN) é um dos métodos mais simples e intuitivos em machine learning. Ele é baseado na ideia de que objetos semelhantes estão próximos uns dos outros no espaço de características. Este método é amplamente utilizado devido à sua simplicidade e eficácia em diversos problemas de classificação e regressão. (Lindholm, 2022)

O KNN é um algoritmo de aprendizado supervisionado que classifica uma nova amostra com base nas classes das (K) amostras de treinamento mais próximas. A proximidade é geralmente medida por uma métrica de distância. A mais comum é a Euclidiana, que basicamente é que a menor distância em linha reta entre dois pontos. O valor de (K) é um hiperparâmetro que deve ser determinado com cuidado, pois influencia diretamente o desempenho do modelo. (Lindholm, 2022)

Quando o K é pequeno, a classificação fica mais sensível a regiões bem próximas, submetendo-se ao risco de sobreajuste, somente fazendo uma boa classificação para os dados treinados.

Com K grande, a classificação fica menos sujeita a ruídos pode ser considerada mais robusta. Porém se K for grande demais, pode ser que haja o problema de underfiting, que é um subajuste, ou seja, um treinamento pobre. O modelo poderia, neste caso, não capturar bem as diferenças entre as classes, gerando previsões genéricas e imprecisas.

2.1.1.1 Processo de Treinamento do KNN

(Lindholm, 2022):

1. Cálculo da Distância: Para uma nova amostra, calcula-se a distância entre esta amostra e todas as amostras do conjunto de treinamento.
2. Seleção dos Vizinhos: Identificam-se os (K) vizinhos mais próximos com base nas distâncias calculadas.
3. Votação: A classe da nova amostra é determinada pela classe majoritária entre os (K) vizinhos mais próximos.

2.1.1.2 Vantagens do KNN:

- Simplicidade: O KNN é fácil de entender e implementar.

- Flexibilidade: Pode ser usado tanto para problemas de classificação quanto de regressão.
- Menor Tempo de configuração: Como o KNN não requer uma fase de treinamento, ele pode ser configurado rapidamente. Isso é especialmente útil quando se precisa de um modelo rápido para uma tarefa específica ou para uma análise exploratória inicial dos dados. Em vez de gastar tempo ajustando parâmetros e treinando o modelo, você pode começar a fazer previsões imediatamente após definir o valor de (K) e escolher a métrica de distância adequada.

Desvantagens do KNN (Lindholm, 2022):

- Custo Computacional: O cálculo das distâncias para todas as amostras de treinamento pode ser computacionalmente intensivo, especialmente para grandes conjuntos de dados.
- Sensibilidade a Ruídos: O desempenho do KNN pode ser mais afetado por outliers e ruídos nos dados.
- Escolha de (K): A escolha do valor de (K) é crucial e pode exigir validação cruzada para otimização.
- Pouco ganho ao se Reutilizar o Modelo: Uma das principais desvantagens do KNN é que ele não éficiente para "aprender" um modelo que possa ser reutilizado para novos dados. Cada vez que uma nova amostra precisa ser classificada, o KNN deve recalcular as distâncias entre essa amostra e todas as amostras do conjunto de treinamento. Isso pode ser computacionalmente caro, especialmente para grandes conjuntos de dados.
- Desempenho em Tempo Real: Devido à necessidade de calcular distâncias para cada nova previsão, o KNN pode ser ineficiente em aplicações de tempo real ou em sistemas com recursos computacionais limitados. Em contraste, modelos que passam por uma fase de treinamento, como redes neurais ou máquinas de vetor de suporte (SVM), podem fazer previsões muito mais rapidamente após o treinamento inicial.

Embora a ausência de uma fase de treinamento explícita no KNN ofereça a vantagem de uma configuração rápida e simples, ela também traz desafios significativos em termos de eficiência computacional e reutilização do modelo. A escolha de usar KNN deve considerar o balanço entre essas vantagens e desvantagens, bem como as especificidades do problema em questão.

Algumas técnicas para melhorar o desempenho do KNN podem ser aplicadas (Lindholm, 2022):

- Normalização dos Dados: Normalizar as características para que todas tenham a mesma escala.
- Ponderação dos Vizinhos: Atribuir pesos aos vizinhos com base na distância, dando mais importância aos vizinhos mais próximos.
- Redução de Dimensionalidade: Utilizar técnicas como PCA (Análise de Componentes Principais) para reduzir a dimensionalidade dos dados e, assim, diminuir o custo computacional.
- Indexação de Dados: Utilizar estruturas de dados como KD-Trees ou Ball Trees para acelerar a busca dos vizinhos mais próximos.
- Amostragem: Em casos de conjuntos de dados muito grandes, pode-se usar uma amostra representativa dos dados de treinamento para reduzir o tempo de cálculo.

O KNN é utilizado em diversas áreas, como reconhecimento de padrões, detecção de anomalias, e sistemas de recomendação. Sua simplicidade e eficácia o tornam uma escolha popular para problemas onde a interpretabilidade e a facilidade de implementação são importantes.

2.1.2 Florestas Aleatórias - Random Forests (RF)

(Lindholm, 2022)

O modelo de classificação de Florestas Aleatórias (Random Forests) é uma técnica de aprendizado de máquina que combina múltiplas árvores de decisão para melhorar a precisão e a robustez das previsões.

Antes de apresentar as Florestas Aleatórias é importante apresentar outro modelo de classificação que são as Árvores de Decisão.

2.1.2.1 Árvores de Decisão

(Lindholm, 2022)

O modelo de classificação de árvores de decisão é uma técnica amplamente utilizada para classificação e regressão. Este modelo é baseado na divisão iterativa do espaço de características em subconjuntos mais homogêneos, utilizando uma estrutura de árvore onde cada nó interno (ponto em que se toma a decisão) representa uma decisão baseada em uma característica específica, e cada folha representa uma classe ou valor de saída.

A construção de uma árvore de decisão começa com o nó raiz, que contém todo o conjunto de dados. Em cada nó, o algoritmo seleciona a característica que melhor separa os dados em termos de pureza, utilizando métricas como a entropia ou o índice de Gini. O índice de Gini é o mais utilizado. A característica escolhida divide os dados em dois ou

mais subconjuntos, que se tornam nós filhos. Este processo é repetido recursivamente até que os nós sejam puros (ou seja, contenham dados de uma única classe) ou até que um critério de parada seja atingido, como a profundidade máxima da árvore que, eventulmente, se estabeleça.

Vantagens das Árvores de Decisão: (Lindholm, 2022):

- Simplicidade: fáceis de interpretar e visualizar, o que facilita a compreensão dos critérios de decisão.
- Flexibilidade: Podem ser usadas tanto para problemas de classificação quanto de regressão.
- Pré-processamento simples: Requerem pouco pré-processamento dos dados, não necessitando de normalização ou escalonamento.

Desvantagens das Árvores de Decisão: (Lindholm, 2022)

- Possibilidade de Overfitting: Árvores de decisão podem facilmente se ajustar demais aos dados de treinamento, especialmente se não forem podadas.
- Instabilidade: Pequenas variações nos dados podem resultar em árvores completamente diferentes.
- Bias (viés): Podem ser enviesadas se uma classe dominar o conjunto de dados.

Para mitigar as desvantagens, técnicas como poda (pruning), agregação (bagging) e florestas aleatórias (random forests) podem ser utilizadas.

A poda remove ramos que têm pouca importância, reduzindo o overfitting (sobreajuste), que é quando o modelo fica muito treinado para aqueles dados, não funcionando bem se mudarem os dados e suas características. O bagging combina várias árvores de decisão treinadas em diferentes subconjuntos dos dados para melhorar a robustez e a precisão. As **florestas aleatórias**, uma extensão do bagging, introduzem ainda mais aleatoriedade ao selecionar subconjuntos aleatórios de características para cada árvore.

As árvores de decisão são aplicadas em diversas áreas, como diagnóstico médico, análise de crédito, e reconhecimento de padrões. Sua capacidade de lidar com dados categóricos e contínuos, além de sua interpretabilidade, as tornam uma ferramenta valiosa em aprendizado de máquina.

Para utilizar um modelo mais aperfeiçoado que parte das Árvores de Decisão, optou-se pelo uso das **Florestas Aleatórias**. Este método, introduzido por Leo Breiman

em 2001, é uma extensão do método de bagging (bootstrap aggregating) e é amplamente utilizado devido à sua eficácia e simplicidade.

Uma Floresta Aleatória é composta por um conjunto de árvores de decisão, cada uma treinada em um subconjunto aleatório dos dados de treinamento. Além disso, durante a construção de cada árvore, um subconjunto aleatório de características é considerado para a divisão em cada nó, o que introduz ainda mais variabilidade e reduz a correlação entre as árvores.

2.1.2.2 Processo de Treinamento das Florestas Aleatórias

(Lindholm, 2022):

1. Amostragem Bootstrap: Para cada árvore, um subconjunto dos dados de treinamento é selecionado com reposição. O que significa que, ao selecionar um subconjunto dos dados de treinamento, cada amostra é escolhida aleatoriamente e pode ser selecionada mais de uma vez. Em outras palavras, após selecionar uma amostra, ela é "reposta" no conjunto original, permitindo que possa ser escolhida novamente.

2. Seleção Aleatória de Características: Em cada nó, um subconjunto aleatório de características é escolhido, e a melhor divisão é determinada apenas entre essas características.

3. Construção da Árvore: Cada árvore é construída até a profundidade máxima ou até que os nós sejam puros, sem poda.

4. Agregação das Previsões: As previsões finais do modelo de Florestas Aleatórias são obtidas agregando as previsões de todas as árvores. Para o caso específico da classificação, a classe final é determinada pela votação majoritária das árvores, quer dizer, quais resultados se repetiram mais vezes em cada árvore.

2.1.2.3 Vantagens do Modelo de Florestas Aleatórias

- Redução do Overfitting: Ao combinar múltiplas árvores, o modelo reduz a variância e o risco de sobreajuste do modelo aos dados de treinamento.

- Robustez: É menos sensível a outliers e ruídos nos dados, pois a classificação é trabalhada de forma mais intensiva em várias árvores.

- Importância das Características: Fornece uma estimativa da importância de cada característica na previsão.

2.1.2.4 Desvantagens do Modelo

- Complexidade: pode ser computacionalmente intensivo e requer mais memória.
- Interpretação: Embora cada árvore individual seja interpretável, a combinação de muitas árvores pode ser difícil de interpretar.

As Florestas Aleatórias são usadas em diversas áreas, como bioinformática, para a classificação de genes; finanças, para a detecção de fraudes; e marketing, para a segmentação de clientes. Sua capacidade de lidar com grandes conjuntos de dados e características torna este método uma escolha popular para muitos problemas de classificação. (Lindholm, 2022)

2.1.3 Gradient Boosting (GB)

(Lindholm, 2022) e (Bishop, 2006)

O modelo Gradient Boosting é uma técnica poderosa e popular em *machine learning*, utilizada principalmente para problemas de classificação e regressão. Este método combina a simplicidade dos modelos de árvores de decisão com a robustez de técnicas de ensemble, resultando em um modelo altamente eficaz e preciso.

Os modelos de árvores de decisão são conhecidos por sua simplicidade e facilidade de interpretação. Eles funcionam dividindo os dados em subconjuntos baseados em características específicas, criando uma estrutura de árvore onde cada nó representa uma decisão. Apesar da simplicidade que as árvores de decisão têm e da facilidade que apresentam para se entender e implementar, a técnica de ensemble (ou associação de modelos) fornece robustez a um modelo que pode se demonstrar muito sensível a ruídos.

O uso das técnicas de ensemble, como o Gradient Boosting, tem como premissa a combinação de múltiplos modelos para melhorar a precisão e a robustez das previsões. Agregar resultados de vários modelos, em vez de depender de um só, ajuda a reduzir a variância e o risco de overfitting. Isso é feito através de métodos como bagging (bootstrap aggregating) e boosting.

O bagging é basicamente uma alternância de amostagens dentro da amostra e o boosting é uma forma iterativa de combinar métodos simples como árvores de decisão sequencialmente. Cada nova árvore é treinada para corrigir os erros cometidos pelas árvores anteriores. O objetivo é minimizar uma função de perda, ajustando o modelo de forma a melhorar a precisão das previsões.

2.1.3.1 Processo de Treinamento do Gradient Boosting

(Lindholm, 2022):

1. Inicialização: O processo começa com uma previsão inicial simples. No caso da Regressão, por exemplo seria a média dos valores de saída.
2. Cálculo dos Resíduos: Em cada iteração, calcula-se o resíduo, que é a diferença entre a previsão atual e o valor real.
3. Novo Treinamento: O modelo é retreinado tentando-se obter melhora nos resíduos.
4. Atualização do Modelo: As previsões do modelo são atualizadas somando-se uma fração das previsões da nova árvore, controlada por um parâmetro de taxa de aprendizado (learning rate).
5. Repetição: Este processo é repetido até que um número predefinido de árvores seja adicionado ou até que a melhoria no erro de previsão se estabilize.

2.1.3.2 Vantagens do Gradient Boosting

(Lindholm, 2022):

- Alta Precisão: geralmente oferece alta precisão devido à sua capacidade de corrigir erros iterativamente.
- Flexibilidade: pode ser aplicado a uma ampla gama de problemas de classificação e regressão.
- Controle de Overfitting: parâmetros como a profundidade das árvores e a taxa de aprendizado ajudam a controlar o overfitting.

2.1.3.3 Desvantagens

(Lindholm, 2022) e (Bishop, 2006)

- Custo Computacional: O treinamento pode ser lento e exigir muitos recursos computacionais, especialmente para grandes conjuntos de dados.
- Sensibilidade a Ruídos: Pode ser sensível a outliers e ruídos nos dados, o que pode levar a overfitting se não for bem regulado.
- Complexidade: A interpretação do modelo final pode ser complexa devido à combinação de muitas árvores.

Uma técnica para melhorar o desempenho do Gradient Boosting pode ser a regularização que se constitui em técnicas que adicionam penalidades ao modelo para evitar o overfitting, que é quando o modelo se ajusta demais aos dados de treinamento e perde a capacidade de generalizar para novos dados.

Uma forma comum de regularização é limitar a complexidade das árvores de decisão individuais. Isso pode ser feito restringindo a profundidade máxima das árvores, o número mínimo de amostras por folha ou o número mínimo de amostras necessárias para dividir um nó. Árvores mais simples são menos propensas a capturar ruídos nos dados de treinamento. (Lindholm, 2022)

Outra estratégia é se controlar a taxa de aprendizado (Learning Rate), ou seja, o quanto cada nova árvore contribui para o modelo final. Uma taxa de aprendizado menor significa que cada árvore tem um impacto menor, o que pode ajudar a evitar overfitting, pois o modelo é ajustado de forma mais gradual. No entanto, isso geralmente requer um número maior de árvores para alcançar uma boa performance. (Lindholm, 2022)

Ainda, considerando mais uma forma de regularização, algumas implementações, como o XGBoost, permitem adicionar penalidades diretamente na função de perda. Isso ajuda a manter os valores dos parâmetros sob controle e reduz a complexidade do modelo.

Por último é importante comentar que pode ser utilizada a Stochastic Gradient Boosting que introduz aleatoriedade ao Gradient Boosting. Essa aleatoriedade pode ser aplicada de duas maneiras principais: na seleção das amostras de treinamento e na seleção das características (features) usadas para dividir os nós das árvores. (Lindholm, 2022)

O Gradient Boosting é amplamente utilizado para previsão de vendas, detecção de fraudes, e análise de risco de crédito. Sua capacidade de lidar com dados heterogêneos e fornecer previsões precisas o torna uma ferramenta valiosa em diversas áreas. (Lindholm, 2022)

2.1.4 Regressão Logística (RL)

(Lindholm, 2022) e (Bishop, 2006)

A Regressão Logística é um modelo bastante utilizado para tarefas de classificação. De certa forma, é um aperfeiçoamento da regressão linear.

A regressão linear é um método utilizado para prever valores contínuos. Por exemplo, pode ser usada para prever o preço de uma casa com base em suas características, como tamanho e localização.. Ela trabalha com a ideia principal de que há uma relação linear entre as variáveis de entrada (características) e a variável de saída (valor contínuo). A regressão busca minimizar a diferença entre os valores previstos e os valores reais, utilizando uma função de custo como o erro quadrático médio.

Por outro lado, a regressão logística é utilizada para prever categorias discretas. Em vez de prever um valor contínuo, a regressão logística estima a probabilidade de uma amostra pertencer a uma classe específica. Isso é feito através de uma função de ativação, também chamada de função sigmoide, que transforma a combinação linear das características de entrada em uma probabilidade. Essa probabilidade é então usada para

determinar a classe da amostra.

Essa função sigmoide mapeia qualquer valor real em um intervalo entre 0 e 1. Esse mapeamento é crucial para interpretar a saída do modelo como uma probabilidade. A função de custo, geralmente a entropia cruzada, é utilizada para otimizar os parâmetros do modelo, ajustando-os para melhorar a precisão das previsões.

A regressão logística, assim como a linear, é muito utilizada pela medicina, para prever doenças, por profissionais de marketing para pesquisas de consumidor. Mas, também cita seu uso na detecção de fraudes, como é o caso deste trabalho.

2.1.4.1 Vantagens da Regressão Logística

(Lindholm, 2022) e (Bishop, 2006):

- Interpretabilidade: Uma das principais vantagens. Os coeficientes do modelo podem ser facilmente interpretados, permitindo que os usuários compreendam como cada característica de entrada influencia a probabilidade de uma amostra pertencer a uma classe específica.
- Simplicidade: é relativamente simples de entender e implementar. Por isso ela pode ser utilizada como um ponto de partida para problemas de classificação. A simplicidade do modelo também facilita a sua implementação em diversas plataformas e linguagens de programação.
- Razoável eficiência computacional: Essa característica torna o modelo adequado para grandes conjuntos de dados. A otimização dos parâmetros do modelo pode ser realizada de forma rápida utilizando métodos de descida de gradiente (gradient descent), permitindo que o modelo seja treinado e aplicado de forma mais rápida que outros.
- Baixa influência de ruídos: é um modelo que não é tão influenciado por outliers. Especialmente, quando comparado a outros modelos de classificação. Isso se deve ao fato de que a função de custo utilizada na Regressão Logística penaliza mais erros grandes, tornando o modelo menos sensível a valores extremos nos dados de treinamento.
- Flexibilidade: torna fácil sua utilização, não só na definição de classes binárias, mas também em problemas de classificação multiclasse pelo uso da regressão logística multinomial.

2.1.4.2 Desvantagens da Regressão Logística

(Lindholm, 2022) e (Bishop, 2006)

- Requer certa linearidade: Segundo (Lindholm, 2022), a Regressão Logística assume uma relação linear entre as características de entrada e o logaritmo das probabilidades. Isso pode ser uma limitação significativa quando os dados possuem relações não lineares complexas. Em tais casos, o modelo pode não capturar adequadamente os padrões nos dados, resultando em baixo desempenho.

- Não está imune a Outliers: Embora tenha relevante robustez, em situações específicas, ela pode ser sensível a outliers. Eles podem influenciar os coeficientes do modelo, especialmente, se não forem tratados adequadamente.

- Demanda grandes conjuntos de dados para que as previsões sejam precisas e confiáveis: Em situações onde os dados são escassos, o modelo pode não ter informações suficientes para aprender os padrões, resultando em desempenho insatisfatório.

- Sensível a multicolinearidade: se duas ou mais características de entrada são altamente correlacionadas, pode ser um problema para a Regressão Logística. A multicolinearidade pode dificultar a estimativa precisa dos coeficientes do modelo, levando a interpretações enganosas e a um modelo menos estável.

- Funciona melhor em classificações binárias. Embora a regressão logística Multinomial possa lidar com problemas de classificação multiclasse, ela pode ser menos eficiente e mais complexa de implementar do que outros modelos, como árvores de decisão ou redes neurais, que são naturalmente adequados para múltiplas classes.

2.1.5 Naive Bayes (NB)

(Bishop, 2006)

O modelo Naive Bayes é baseado no Teorema de Bayes, que é uma maneira de calcular a probabilidade de uma classe, dado um conjunto de características. Essencialmente, ele ajuda a atualizar as crenças sobre a probabilidade de uma classe com base em novas evidências.

A característica "ingênua", como a tradução do nome diz, é a suposição de que todas as características são independentes umas das outras, dado o valor da classe. Isso significa que o modelo assume que a presença ou ausência de uma característica não afeta a presença ou ausência de outra característica, o que simplifica muito os cálculos.

As principais aplicações do modelo Naive Bayes são em áreas como filtragem de spam, classificação de documentos, análise de sentimentos e sistemas de recomendação de produtos, serviços ou informações.

Existem alguns tipos de modelos Naive Bayes (Lindholm, 2022) e (Bishop, 2006):

- Gaussian Naive Bayes : é o mais amplamente utilizado e serve para características que são contínuas e seguem uma distribuição normal (ou Gaussiana).
- Multinomial Naive Bayes: é adequado para dados que representam contagens ou frequências, como a contagem de palavras em documentos. É amplamente utilizado em tarefas de classificação de texto, como filtragem de spam e categorização de notícias.
- Bernoulli Naive Bayes: é utilizado para características binárias, ou seja, características que podem estar presentes ou ausentes. É também utilizado para detecção de spam.

2.1.5.1 Vantagens do Modelo Naive Bayes

(Lindholm, 2022) e (Bishop, 2006)

- Simplicidade: simples de implementar e computacionalmente eficiente. Ele requer uma quantidade relativamente pequena de dados de treinamento para funcionar bem.
- Escalabilidade: o número de parâmetros que precisa estimar cresce linearmente com o número de características. Isso significa que ele pode lidar com grandes conjuntos de dados de maneira eficaz.

2.1.5.2 Desvantagens

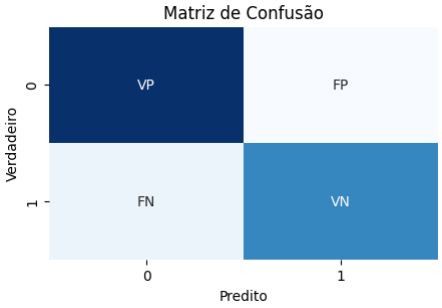
(Bishop, 2006)

- A principal limitação do Naive Bayes é que ele supõe independência entre as características, que raramente é verdadeira. No entanto, mesmo com essa suposição simplificada, o modelo frequentemente apresenta bom desempenho em muitas aplicações reais.
- Embora o Naive Bayes seja um bom classificador, ele pode não ser um bom estimador de probabilidades. As probabilidades preditas pelo modelo podem não ser confiáveis, especialmente em casos de classes desbalanceadas.

2.2 Métricas de Avaliação

As métricas utilizadas neste trabalho são a Acurácia, Precisão, Revogação e o F1-Score. Elas são melhor entendidas após a apresentação do Gráfico denominado Matriz de Confusão:

Para melhor entender, o verdadeiro positivo é o que se espera de um modelo de classificação, a identificação correta como verdadeira para o valor que é verdadeiro, o falso



Fonte: Elaborada pela Autora

Figura 1 – Matriz de Confusão

positivo, seria falar que um valor falso é classificado como a correto para a variável alvo. O falso negativo é o correto, classificado como falso. Por último, o verdadeiro negativo, também é uma resposta esperada, porém é dizer que aquele valor classificado como falso é, realmente, falso.

Segue o detalhamento de cada métrica e como ela pode ser melhor aplicada.

2.2.1 Acurácia

A acurácia nos dá a proporção de previsões corretas (VP + VN), ou seja, ela faz a soma de todas as previsões corretas que se obtém na classificação. Não importa se a empresa era Inapta (identificada pelo algarismo 0) ou Ativa (identificada pelo algarismo 1), se a previsão foi correta, ela soma tudo e divide pelo total de previsões (VP + FP + FN + VN).

A acurácia é bastante utilizada e uma excelente métrica. Ela é adequada ao modelo que estamos trabalhando, uma vez que, especialmente, para os dados de treinamento, há um balanceamento entre as amostras e para a classificação binária.

2.2.2 Precisão

A precisão irá fazer a relação entre as previsões positivas corretas (VP), no caso são as empresas Ativas (1), sobre todas que foram classificadas como positivas (VP + FP).

A precisão é mais apropriada para classes desbalanceadas e são a métrica correta quando se há um custo alto de falsos positivos. No nosso caso, para dados inéditos, pode ser uma boa alternativa, pois, há um maior desbalanceamento dos dados e encontrar com segurança as empresas ativas (valores verdadeiros) é bastante desejado.

2.2.3 Revogação

A revogação, também chamada pelo inglês, recall, dá a proporção de previsões positivas corretas (VP) sobre todas que deveriam ter sido acertadas como corretas (VP + FP).

A revogação é essencial quando há um olhar mais crítico para os falsos negativos (detecção de doenças, por exemplo). Ela funciona melhor também em dados desbalanceados, o que se mostra também muito interessante para o trabalho.

2.2.4 F1-Score

Finalmente, essa métrica relaciona as duas anteriores. Ela é a média harmônica da precisão e da revogação, dada pela fórmula F1 = [2x (precisão x revogação)]/(precisão + revogação).

Ela é um equilíbrio entre as duas últimas métricas detalhadas. Bastante, utilizada para amenizar o efeito de um possível desbalanceamento entre as amostras sem deixar de fazer uma boa avaliação. Como também é uma métrica que funciona bem com dados desbalanceados, foi escolhida como a principal na eleição do modelo mais robusto deste estudo.

É importante informar que o cálculo das métricas foi feito em python. Para as variáveis Precisão, Revogação e F1-Score foi utilizado um método ponderado que leva em consideração os valores verdadeiros para as empresas inaptas e ativas, como forma de minimizar um possível desbalanceamento, que ocorre, nos testes com os novos dados.

2.3 Refinamento

O refinamento nada mais é que o ajuste de hiperparâmetros. Esse processo é crucial no desenvolvimento dos modelos de classificação.

O GridSearchCV é uma técnica amplamente utilizada para essa finalidade. A escolha adequada dos hiperparâmetros pode melhorar significativamente a performance do modelo. Ele atua em uma busca exaustiva em um espaço pré-definido de hiperparâmetros, testando todas as combinações possíveis para encontrar a configuração que resulta no melhor desempenho do modelo. Esse processo é particularmente útil em algoritmos de classificação, onde a precisão e a capacidade de generalização são essenciais. (Bishop, 2006)

Ao aplicar o GridSearchCV, é importante definir qual é o conjunto de hiperparâmetros que se deseja otimizar. Por exemplo, em uma Random Forest, pode-se variar o número de árvores, a profundidade máxima das árvores e o número mínimo de amostras por folha. O GridSearchCV então treina e avalia o modelo para cada combinação desses valores, utilizando algum tipo de validação cruzada para garantir que os resultados sejam robustos e não enviesados. É um método computacionalmente intensivo, mas garante que todas as possíveis configurações sejam consideradas, aumentando as chances de encontrar a melhor combinação. (Bishop, 2006)

Importante considerar que o GridSearchCV deve ser usado com cuidado, especialmente em conjuntos de dados grandes, devido ao seu alto custo computacional. Nesse caso,

vale a pena considerar utilizar uma busca aleatória em vez da exaustiva, para reduzir o tempo de computação. Mesmo assim, o GridSearchCV continua sendo uma ferramenta poderosa quando se dispõe de recursos computacionais suficientes. (Lindholm, 2022)

Nesse trabalho, decidiu-se por associar o GridSearchCV com a validação cruzada de Monte Carlo que é uma abordagem poderosa para otimizar modelos de classificação que não tenha um número extenso de dados, como é o caso.

A validação cruzada de Monte Carlo é uma técnica de validação que envolve a repetição de processos de treinamento e teste várias vezes com diferentes divisões aleatórias dos dados, proporcionando uma estimativa mais robusta da performance do modelo. A técnica envolve a repetição de processos de treinamento e teste várias vezes com diferentes divisões aleatórias dos dados, proporcionando uma estimativa mais confiável do desempenho do modelo. (Bishop, 2006)

O processo começa com a divisão aleatória do conjunto de dados em dois subconjuntos: um para treinamento e outro para teste. Em cada iteração, uma nova divisão aleatória é realizada, e o modelo é treinado no subconjunto de treinamento e avaliado no subconjunto de teste. Esse processo é repetido várias vezes (geralmente dezenas ou centenas), e os resultados são então agregados para fornecer uma estimativa média do desempenho do modelo. (Bishop, 2006)

A principal vantagem da validação cruzada de Monte Carlo é que ela reduz a variabilidade associada a uma única divisão dos dados. Ao realizar múltiplas divisões aleatórias, a técnica captura uma ampla gama de possíveis variações nos dados, proporcionando uma estimativa mais robusta e confiável da capacidade de generalização do modelo. Essa ferramenta é particularmente útil em conjuntos de dados pequenos ou quando há uma alta variabilidade nos dados como acontece com os dados utilizados no modelo e que se verá mais à frente.

Dessa forma, a validação cruzada de Monte Carlo ajuda a mitigar o risco de overfitting, garantindo que o modelo não se ajuste excessivamente a uma única divisão dos dados. Além disso, ao utilizar o GridSearchCV, é possível otimiza a escolha dos parâmetros podendo proporcionar um modelo mais próximo do melhor possível.

2.4 Trabalhos Relacionados

A ciência de dados torna-se cada vez mais relevante para qualquer área de conhecimento e desenvolvimento de trabalhos. Para a gestão tributária não é diferente. Ao longo do tempo diversos estudos podem ser encontrados para que se entenda a importância de se desenvolver técnicas para a detecção de fraudes fiscais e, de forma particular, o uso da tecnologia da informação com esse objetivo. Nesta seção, pode-se verificar quão rico é o desenvolvimento de trabalhos nessa seara e alguns trabalhos acadêmicos desenvolvidos na

área de fraude fiscal, combinados ou não, com aprendizado de máquina se destacam.

Veja que o artigo "A corrupção no ambiente de negócios: survey com as micro e pequenas empresas da cidade de São Paulo", escrito por Felipe Mendes Borini e Fernando Correa Grisi, investiga a presença e os impactos da corrupção no ambiente empresarial, ainda que seja focado em micro e pequenas empresas. Infelizmente, pode-se aferir que é um comportamento disseminado em vários setores econômicos e não faz distinção de capacidade financeira. A pesquisa revela que a corrupção é uma prática comum tanto nas relações entre empresas quanto nas interações com o setor público, que é onde estamos interessados a atuar. (Borini, 2009)

O estudo utilizou uma métrica desenvolvida especificamente para avaliar a corrupção no ambiente de negócios dessas empresas, aplicando um survey com 206 respostas de micro e pequenas empresas do centro comercial de São Paulo. Os resultados indicam a presença de diferentes tipos de corrupção, classificados como corrupção preta, cinza e branca. A pesquisa conclui que a corrupção no ambiente de negócios é um fenômeno complexo e multifacetado. (Borini, 2009)

Andréa de Mendonça e outros também publicaram algo relacionado na revista Cadernos EBAPE da FGV, o artigo "Misbehavior empreendedor associado às práticas de sonegação fiscal", sobre a investigação de como o comportamento inadequado de empreendedores está relacionado às práticas de sonegação fiscal em micro e pequenas empresas no Brasil. A pesquisa adota a perspectiva do "jeitinho brasileiro", um mecanismo social onde indivíduos contornam regras e normas para obter vantagens. Através de entrevistas semiestruturadas com micro e pequenos empreendedores, o estudo revela que a sonegação fiscal é frequentemente vista como uma escolha racional, devido à percepção de baixa fiscalização e penalidades pouco relevantes. (Mendonça, 2021)

Os resultados mostram que muitos empresários justificam a sonegação como uma estratégia necessária para a sobrevivência e competitividade de seus negócios. A pesquisa destaca que a formação cultural do país, que reforça a desconfiança no sistema político e econômico, contribui para a aceitação dessas práticas. A análise de conteúdo das entrevistas identifica categorias de comportamento que incluem a violação intencional de regras e normas sociais, caracterizando o misbehavior organizacional. (Mendonça, 2021)

O conhecimento desse comportamento é altamente relevante para trabalhos de investigação e fraudes usando ciência de dados. As variáveis e padrões de comportamento identificados podem ser incorporados em modelos preditivos de aprendizado de máquina para detectar práticas de sonegação fiscal. Além disso, a compreensão dos fatores culturais e comportamentais que incentivam a sonegação pode ajudar a desenvolver algoritmos mais precisos e contextualizados. A integração de dados abertos e técnicas de inteligência artificial pode potencializar a identificação de fraudes, permitindo uma fiscalização mais eficaz e proativa. (Mendonça, 2021)

No artigo “Ciência de Dados Aplicada à Auditoria Interna” de Gustavo Fleury Soares há diversas menções de como a ciência de dados pode transformar as atividades de auditoria. No caso desse artigo, a menção é de forma especial às auditorias internas, apresentando novos métodos e desafios. Contudo, é fácil a extrapolação para auditorias feitas pela administração público-tributária. (Soares, 2020)

O trabalho começa definindo e interligando diversas especialidades da ciência de dados, como inteligência artificial, mineração de dados e big data. Em seguida, revisa a literatura acadêmica contemporânea, destacando os principais métodos, benefícios e desafios para cada etapa da auditoria interna.(Soares, 2020)

A ciência de dados oferece técnicas avançadas para extrair informações e insights dos dados, um objetivo crucial na auditoria interna. Entre os métodos discutidos estão a análise preditiva, a detecção de anomalias e a mineração de textos, que podem identificar padrões suspeitos e comportamentos atípicos. A aplicação dessas técnicas pode melhorar a conformidade, apoiar a tomada de decisões gerenciais e detectar fraudes de maneira mais eficaz.(Soares, 2020)

Modelos preditivos e algoritmos de detecção de anomalias podem ser treinados para identificar transações fraudulentas, enquanto a mineração de textos pode ser usada para analisar documentos e comunicações em busca de sinais de fraude. O trabalho considera integração de big data permite o processamento de grandes volumes de dados em tempo real, aumentando a precisão e a eficiência das auditorias, servindo de paradigma para este trabalho.(Soares, 2020)

Otávio Calaça Xavier e outros publicaram na Revista de Administração Pública da FGV o artigo: "Identificação de Evasão Fiscal Utilizando Dados Abertos e Inteligência Artificial"que explora como a combinação de dados abertos e técnicas de inteligência artificial podem ser utilizados para identificar práticas de evasão fiscal. Com a mesma proposta que rege o atual trabalho qual seja a de desenvolver soluções inteligentes que auxiliem a administração tributária e os auditores na detecção de sonegadores, utilizando dados disponíveis. A linha deles é utilizar dados públicos, diferentemente, do que fizemos, mas, a estratégia é muito semelhante. (Xavier, 2022)

Otávio e seus pares apresentam três modelos preditivos baseados em algoritmos de Random Forest, Redes Neurais e Grafos. Esses modelos foram treinados para identificar padrões e perfis de contribuintes com alto risco de evasão fiscal. A validação dos modelos mostrou uma acurácia superior a 98% na predição de perfis inadimplentes, destacando a eficácia dessas técnicas na detecção de fraudes fiscais. (Xavier, 2022)

Além disso, o artigo discute a implementação de uma solução de software visual para uso pelos auditores fiscais, permitindo uma análise mais eficiente e precisa dos dados. A pesquisa conclui que o uso de dados abertos, combinado com inteligência artificial,

pode revolucionar a fiscalização tributária, tornando-a mais proativa e menos dependente de dados sensíveis dos contribuintes. Essa proposta pode ser muito interessante como complemento ao trabalho atual e prospecções futuras. (Xavier, 2022)

O trabalho desenvolvido por Joseph Nembe e outros "The Role of Artificial Intelligence in Enhancing Tax Compliance and Financial Regulation" explora como a Inteligência Artificial (IA) pode transformar a conformidade tributária e a regulação financeira. O trabalho se propôs a analisar como soluções baseadas em IA podem melhorar a administração tributária, detectar não conformidades e reduzir a evasão fiscal. Isso porque, algoritmos de aprendizado de máquina são capazes de analisar grandes volumes de dados financeiros com rapidez e precisão, identificando padrões que indicam fraudes ou evasões fiscais. (Nembe J. K., 2024)

De forma similar ao que se pretende o trabalho atual, o de Joseph se baseou em que a análise preditiva com algoritmos de *machine learning* possam permitir que as autoridades fiscais antecipem comportamentos dos contribuintes e aloque recursos de forma mais eficaz para fins de fiscalização. (Nembe J. K., 2024)

Foi também explorado o potencial no campo da regulação financeira, a IA desempenha um papel crucial na monitoração e cumprimento das normas regulatórias complexas. Com o crescimento exponencial das transações financeiras e a sofisticação dos instrumentos financeiros, os mecanismos tradicionais de regulação não conseguirão acompanhar. Sistemas de IA, equipados com capacidades de processamento de linguagem natural, podem examinar grandes volumes de documentos regulatórios e dados financeiros para identificar possíveis violações e avaliar riscos sistêmicos. Modelos de avaliação de risco baseados em IA permitem que os reguladores identifiquem proativamente ameaças emergentes à estabilidade financeira, facilitando intervenções oportunas para prevenir crises. (Nembe J. K., 2024)

No entanto, o estudo de Joseph deixa claro que embora a IA ofereça grandes promessas para melhorar a conformidade tributária e a regulação financeira, sua implementação deve ser acompanhada por estruturas de governança robustas e diretrizes éticas. A dependência de decisões algorítmicas levanta preocupações sobre transparência, responsabilidade e mitigação de vieses. Nessa mesma linha, o presente trabalho teve preocupação em trabalhar em conformidade com a legislação de proteção de dados evitando-se a utilização de dados sensíveis. (Nembe J. K., 2024) (São Paulo, 2006)

Mais explicitamente trabalhado o aprendizado de máquina para encontrar empresas de fachada, tem-se o trabalho de Mestrado de Gunther Siqueira Lemes Gomes: "Identificação de Práticas de Evasão Fiscal Utilizando Aprendizagem de Máquina: O caso das empresas de fachada e os créditos ilegais de ICMS." (Gomes, 2023)

Aquele trabalho procurou demonstrar como técnicas de aprendizado de máquina

podem ser aplicadas para identificar fraudes fiscais específicas, especialmente aquelas realizadas por empresas de fachada que emitem créditos ilegais de ICMS. A proposta, muito alinhada ao presente trabalho, buscou analisar a eficácia dessas técnicas na detecção de fraudes, contribuindo para a melhoria das práticas de fiscalização tributária. (Gomes, 2023)

Foram identificadas variáveis relevantes para a detecção de fraudes: padrões de emissão de notas fiscais, comportamentos atípicos nas transações comerciais, histórico de conformidade fiscal, relações entre empresas, dados financeiros e informações cadastrais. Utilizando esses dados, foram desenvolvidos modelos preditivos de aprendizado de máquina para identificar empresas com risco de práticas fraudulentas. (Gomes, 2023)

Os resultados se mostraram eficazes e foi ressaltada, mais uma vez, a necessidade de um olhar ético e transparente para uso dessas tecnologias. (Gomes, 2023)

Também fundamental se entende o reconhecimento de trabalhos anteriormente realizados por participantes do MBA em Ciência de Dados que deram enfoque à detecção de fraudes fiscais com aprendizado de máquina.

O trabalho do colega Marcelo Fernandez, "Proposta de modelo de predição de operações comerciais interestaduais simuladas: análise de operações com a Zona Franca de Manaus / Área de Livre Comércio"foi de grande inspiração para o corrente, pois focando nas transações envolvendo a Zona Franca de Manaus (ZFM) e Áreas de Livre Comércio (ALC) e utilizando técnicas de classificação em *Machine Learning* buscou identificar padrões e anomalias nas operações comerciais, com o objetivo de detectar possíveis fraudes tributárias. Ele utilizou um grante volume de dados comerciais para identificar padrões de comportamento normais e anômalos, enquanto o aprendizado de máquina implementa algoritmos para melhorar a precisão na detecção de fraudes.(Fernandez, 2024)

Bem como nos padrões que estamos querendo encontrar, as simulações de operações comerciais envolvendo a Zona Franca de Manaus ajudaram a prever possíveis cenários adicionais de fraude, sendo especialmente relevante para essas regiões com incentivos fiscais, onde a fraude pode ser mais prevalente. (Fernandez, 2024)

O trabalho do colega reforça a importância destes modelos para a detecção de fraudes tributárias, pois aumenta a eficiência, melhora a precisão na identificação desse tipo de operação, permite uma atuação anterior à sua ocorrência, economizando recursos e tempo. Além disso, promove maior transparência nas operações comerciais, fortalecendo a confiança no sistema tributário. (Fernandez, 2024)

Destacam-se tambem, três trabalhos bastante correlacionados desenvolvidos pelos colegas Walter Bentivegna, Pedro Henriques e Rodrigo Viti. Eles trabalharam com o aprendizado de máquina usando classificação com variáveis como: valor das transações, frequência de emissão de notas fiscais, histórico de irregularidades, setor de atuação e

relações comerciais, localização geográfica das empresas e o tempo de operação no mercado (Bentivegna, 2022) (Henriques, 2024) (Viti, 2024). Os modelos de classificação utilizados por Bentivegna foram Regressão Logística, Florestas Aleatórias, Máquinas de Vetores de Suporte (SVM), K-Nearest Neighbors (KNN), Redes Neurais Artificiais, árvores de decisão e o Gradient Boosting. (Bentivegna, 2022) (Henriques, 2024) (Viti, 2024)

Os resultados foram se aprimorando com as inovações em cada trabalho e apresentaram resultados favoráveis na identificação de “noteiras”.

Isso estimula que técnicas como as desenvolvidas no trabalho dos colegas e o presente sejam integradas aos sistemas de monitoramento e auditoria fiscal, melhorando a capacidade de detecção e prevenção de fraudes.

Por último, também apresentado na conclusão do Curso de MBA da USP-São Carlos, o trabalho de análise de sobrevivência: "Utilização de análise de sobrevivência para inidôneos: uma análise dos tempos de inaptidão e cassação dos casos de 2015 a 2023", pelo colega Jônatas Dell Lucas, mostra que a Secretaria da Fazenda e Planejamento do Estado de São Paulo tem desempenhado um bom trabalho na detecção e cessação de fraudes. Porém, a sua conclusão, sugere que os prazos poderiam ser diminuídos. Daí, surge a oportunidade de que aplicação de algoritmos de classificação poderia complementar essa abordagem, ajudando a identificar fraudes de forma mais rápida e eficiente. Algoritmos de *machine learning* poderiam ser treinados para automatizar a detecção e reduzir o tempo de resposta. A integração dessas técnicas com a análise de sobrevivência poderia criar um sistema robusto para a detecção e gestão de fraudes, beneficiando a SEFAZ-SP com uma fiscalização mais ágil e precisa. (Ducas, 2024)

2.5 Legislação Aplicada

A Portaria CAT 95/2006, emitida pela Secretaria da Fazenda do Estado de São Paulo, estabelece procedimentos para a suspensão, cassação e nulidade da inscrição no Cadastro de Contribuintes do ICMS. Mais especificamente, sobre a declaração de nulidade de empresas inidôneas, a portaria define que a inscrição no cadastro pode ser considerada nula quando constatada a participação do contribuinte em fraudes fiscais estruturadas. (São Paulo, 2006)

A nulidade da inscrição ocorre em situações em que há evidências de que a empresa foi constituída para práticas fraudulentas, como a emissão de notas fiscais sem a correspondente circulação de mercadorias ou serviços. O procedimento administrativo de constatação de nulidade envolve a apuração de fatos que comprovem a inidoneidade da empresa, podendo resultar na anulação da inscrição no cadastro ou sua cassação a partir de um determinado momento. Destafeita, a portaria detalha os critérios e processos para a suspensão e cassação da eficácia da inscrição, visando garantir que apenas empresas em

conformidade com as normas fiscais permaneçam ativas no cadastro.(São Paulo, 2006)

A Receita Federal também tem sua metodologia para declaração de inidoneidade de empresas feita na Instrução Normativa RFB nº 1.863/2018 que estabelece um conjunto de critérios detalhados para a identificação desses contribuintes e a subsequente declaração de inaptidão de suas inscrições no Cadastro Nacional da Pessoa Jurídica (CNPJ). (RFB, 2018)

Um dos principais critérios é a omissão na entrega de declarações fiscais obrigatórias por um período mínimo de 90 dias. Empresas que não cumprem essa exigência são consideradas para inaptidão, pois a falta de declarações pode indicar tentativas de ocultar informações financeiras relevantes. Além disso, a Receita Federal presta atenção especial às inconsistências em documentos fiscais. A emissão de notas fiscais frias, que são notas fiscais emitidas sem a correspondente circulação de mercadorias ou prestação de serviços, é um sinal claro de práticas fraudulentas. A autenticidade e a consistência das informações contidas nesses documentos são rigorosamente verificadas. (RFB, 2018)

Empresas que simulam transações para gerar créditos fiscais indevidos ou para ocultar a real movimentação financeira são identificadas como inidôneas. A Receita Federal utiliza de cruzamento de dados e análise de padrões para detectar essas simulações. Irregularidades cadastrais também são um fator importante. Discrepâncias nos dados cadastrais, como endereços falsos ou inexistentes, e a utilização de "laranjas" (pessoas que emprestam seus nomes para a constituição de empresas fraudulentas) são indicadores de inidoneidade. (RFB, 2018)

Aquelas com maior histórico de não conformidade, incluindo reincidência em práticas fraudulentas, são propensas a serem declaradas inaptas. A identificação dos beneficiários finais, ou seja, as pessoas naturais que, em última instância, possuem ou controlam a empresa, também é levado em consideração pela Receita que verifica se esses beneficiários estão envolvidos em práticas fraudulentas.

Após essa compilação de material acadêmico e legislativo vamos dar sequência ao desenvolvimento propriamente dito do trabalho proposto. Apresenta-se na sequência o objetivo do trabalho, a metodologia, resultado e conclusão.

3 OBJETIVO DO TRABALHO

Partindo do conhecimento de que empresas são constituídas de forma simulada para fraudar o fisco e o mercado, abalando a estabilidade deste último e prejudicando o desenvolvimento da sociedade e das políticas públicas, há a necessidade de se encontrar as melhores estratégias para combater esse comportamento.

Nesse caminho, este trabalho visa a construção de um modelo classificação que possa colaborar na identificação de tais empresas. Para isso, se utilizará de informações cadastrais e economico-fiscais obtidas nos bancos de dados da Secretaria de Fazenda do Estado de São Paulo para um modelo de classificação, utilizando as ferramentas de aprendizado de máquina.

Foi escolhido para o trabalho um modelo supervisionado, tomando como alvos empresas que foram investigadas e classificadas como **inaptas**, de acordo com um conjunto de evidências e provas, definidas na legislação ((São Paulo, 2006)), por não existirem de fato, ou ainda, por não existirem da forma como foram cadastradas nos sistemas fazendários (simulação de quadro societário, entre outros). Por outro lado, temos empresas que já foram fiscalizadas, que estão **ativas**, pois atendem, ainda que sob fiscalização e imposição de penalidades, as demandas oriundas do fisco de forma objetiva e concreta de forma que permitiram concluir como existentes de fato e direito.

O termo inidônea aqui, bastante utilizado, quer se referir a empresas que simulam sua existência de fato, existindo apenas no "mundo"cadastral, ou simulam seu quadro societário, ou ainda, outras informações, com o finco de obter vantagens, quais sejam, reduzir a carga tributárias para elas ou para outrem.

Tendo-se criado um modelo e obtendo-se uma avaliação, significativamente, positiva, poder-se-á aplicá-lo a empresas que ainda não foram fiscalizadas, com o passar do tempo, se poderá utilizar até mesmo para empresas no início de seu funcionamento.

Uma ferramenta dessas em mãos, proporcionará uma maior eficiência e ganho de escala nas ações do fisco.

4 METODOLOGIA

Considerando-se a importância de um projeto de Ciência de Dados como esse que pode trazer resultados com grande impacto positivo para a fiscalização estadual, torna-se interessante a utilização do Método CRISP-DM (Processo Padrão da Indústria Transversal para Mineração de Dados) que envolve as seguintes etapas que devem suceder-se de forma suave e coerente. A Figura 2 traz a visualização do processo:



Fonte: Bruno Lima Santos, enviado pelo aplicativo WhatsApp

Figura 2 – Fluxograma do Método CRISP-DM

4.1 Entendimento do Negócio

O primeiro passo da Metodologia CRISP-DM é o **Entendimento do Negócio**, precisamos saber qual é a área de conhecimento e qual é o problema a ser resolvido. No caso em voga, trata-se de um problema da área da fiscalização de tributos, mais especificamente, o Imposto Sobre Circulação de Mercadorias e Serviços de Telecomunicação (ICMS) do Estado de São Paulo. Procura-se evitar fraudes que podem levar à evasão fiscal, por emissão de documentos fiscais inábeis, as chamadas notas frias feitas por empresas inidôneas.

No capítulo de 1 procurou-se detalhar o negócio e as questões que justificaram o desenvolvimento desse estudo.

4.2 Entendimento dos Dados

O próximo passo é **Entendimento dos Dados**, quer dizer, saber quais são os dados disponíveis ou necessários, como vão ser obtidos, estão eles identificados ou estão mesclados a outros dados. Também precisa-se saber como serão trabalhados os dados em termos de limpeza, transformação, balanceamento etc.

Os nossos dados foram foram extraídos do banco de dados da SEFAZ-SP, disponibilizados no aplicativo Rich Client, observando Lei Geral de Proteção de Dados Pessoais (LGPD) - Lei 13.709/ 2018 – e a Resolução SF 20/2012. (Brasil, 2018) e (São Paulo, 2012)

São dados individualizados, que foram fornecidos pelos próprios contribuintes, quando da inscrição no Estado de São Paulo ou no momento em que emitem os documentos fiscais que acompanharam as operações supostamente realizadas.

Outras variáveis poderiam ser utilizadas, pensando-se em aumentar a efetividade da modelagem, no entanto, para garantir a observância da legislação supracitada, preferiu-se a modéstia na não utilização de variáveis que pudessem conter dados sensíveis e sigiloso.

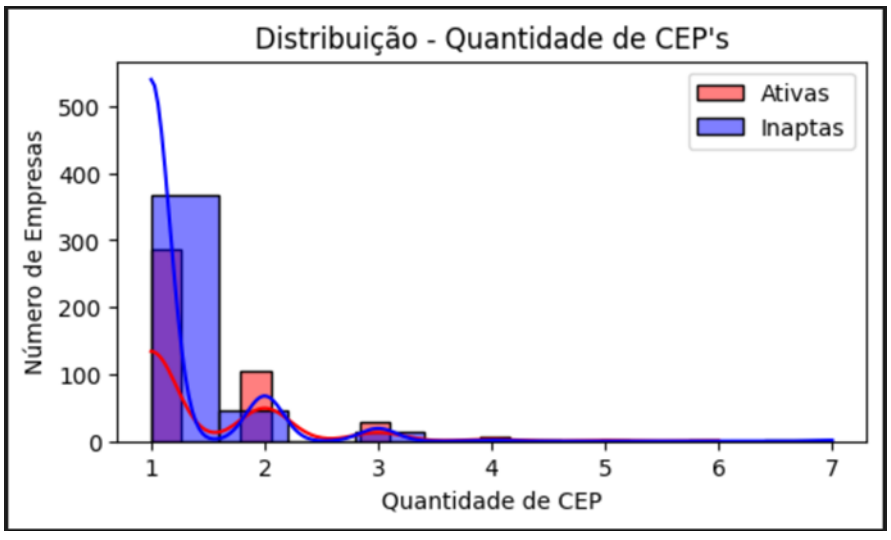
Poderíamos, por exemplo, utilizar os números de documentos dos indivíduos, quais sejam, pessoas físicas ou jurídicas envolvidas nas empresas. Ou mesmo avançar com variáveis categóricas, tornando o modelo mais complexo.

Essas possibilidades não estão descartadas para um estudo futuro, porém, para permitir uma maior liberdade de publicidade para o presente trabalho, escolheu-se restringir dados, nesse momento.

No próximo item, iremos fazer uma análise exploratória mais aprofundada, explicando cada variável e, quando adequado, faz-se uma apresentação gráfica ou apresentação de parâmetros estatísticos.

4.2.1 Dados Cadastrais

1. Quantidade de CEP do Estabelecimento (quant_cep): é uma variável numérica inteira, escolhida por representar o número de vezes que houve alteração de endereço de uma empresa.



Fonte: Elaborado pela autora.

Figura 3 – Distribuição - CEP's para cada grupo de empresas.

A percepção que temos pela Figura 3 é que as Empresas Ativas têm um padrão mais instável de endereço. Enquanto as Inaptas estão bastante concentradas com um só CEP, ou seja, uma tendência a não ter uma variação de endereço tão grande para a maioria delas. Mais de 85% (noventa e cinco por cento) das empresas se mantiveram no endereço desde sua abertura após o ano de 2019. Isso pode ser decorrência do fato que uma empresa constituída apenas com fins ilícitos, quer seja para ludibriar o fisco, quer seja para enganar o mercado, em geral, oferece um endereço residencial, ou mesmo um endereço inexistente, pelo simples fato que é uma exigência o fornecimento do endereço pela Junta Comercial. Inclusive, nessas muitas vezes, quando encontrado o endereço, percebe-se que não há condições de desenvolvimento da atividade declarada pela empresa.

Abaixo, na tabela 1, pode-se ver de forma mais clara a distribuição das quantidades dos ceps de acordo com o tipo de empresa.

Quantidade de CEP's	% de Ativas	% de Inaptas
1	66,82	87,75
2	24,53	10,75
3	5,54	3,04
4	1,40	0,00
5	0,47	0,00
6	0,24	0,24

Tabela 1 – Distribuição percentual de CEP's para cada grupo de empresas.

2. CEP atual (cep_atual): variável numérica inteira. É o atual código postal fornecido, que remete à suposta localização da empresa.

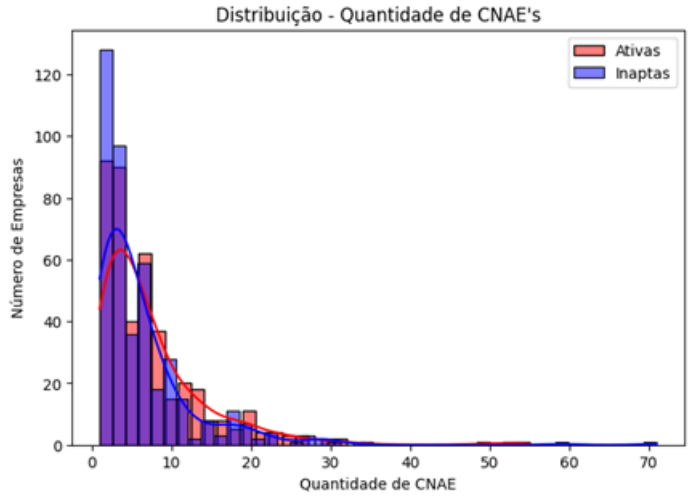
3. Quantidade de CNAE's do Estabelecimento (quant_cnaes): CNAE é a sigla para classificação nacional de atividades econômicas e consiste em um código de sete números utilizado para padronizar e categorizar as atividades econômicas no país.

Para obter um número de CNPJ é necessário escolher a CNAE correta, permitindo um correto enquadramento tributário. A CNAE é direcionada pelo IBGE (Instituto Brasileiro de Geografia e Estatística) e coordenado pela Secretaria da Receita Federal.

A variável aqui apresentada é numérica inteira e poderia permitir a existência de um padrão próprio para as empresas inaptas, por ser o próprio responsável pela empresa que faz o enquadramento das atividades econômicas elegendo dentre os definidos o que mais considera adequado.

Em que pese essa expectativa, apenas pela obervação da Figura 4, a distribuição da quantidade CNAE's entre os dois grupos de empresas não parece ter uma grande diferença.

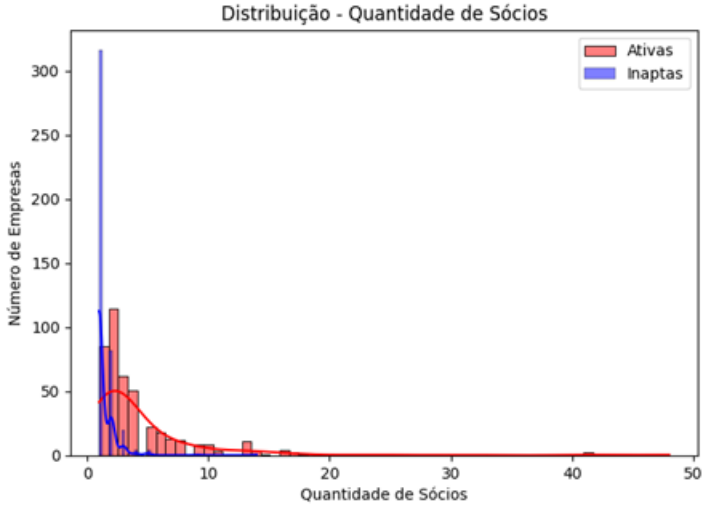
Apenas, timidamente, as empresas inaptas tendem a ter um número menor CNAE's cadastrados, evidente com a concentração em valores menores.



Fonte: Elaborado pela autora.

Figura 4 – Distribuição - CNAE's para cada grupo de empresas.

4. Quantidade Sócios (quant_socios): é uma variável numérica inteira que indica o número total de indivíduos que já fizeram parte do quadro societário da empresa.



Fonte: Elaborado pela autora.

Figura 5 – Distribuição - Sócios que já pertenceram ao quadro da empresa.

Quantidade de Sócios	% de Ativas	% de Inaptas
1	19,86	73,83
2	26,64	19,16
3	14,49	4,67
4	11,92	0,93
5	5,14	0,92
6 a 48	21,96	0,47

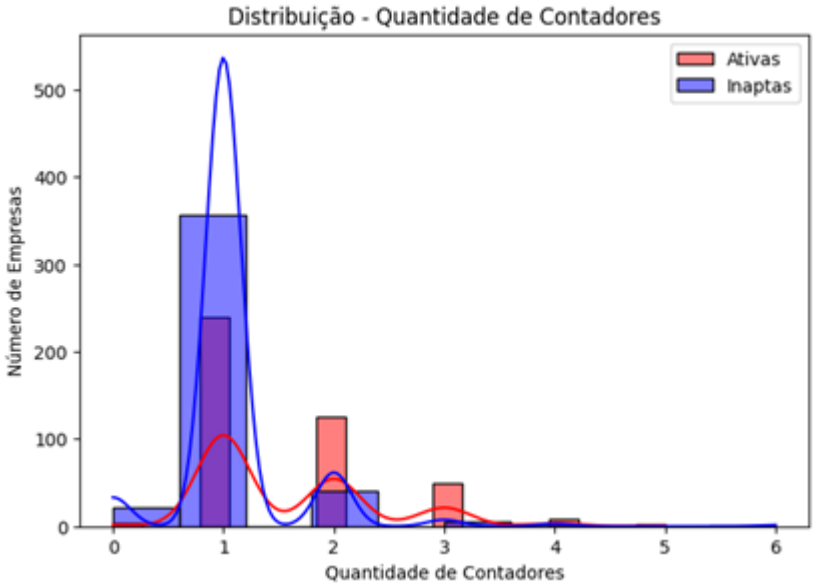
Tabela 2 – Distribuição percentual de quant_socios para as empresas.

Na Figura 5, percebe-se uma concentração dessa quantidade para as empresas inaptas, não existentes, em uma ou duas pessoas. Vê-se que, aproximadamente, 94% têm

essa característica, enquanto há uma maior distribuição de valores para empresa ativas com constatação de existência. A tabela 2 reforça abaixo essa informação.

5. Quantidade de Contadores do Estabelecimento (quant_contadores): é uma variável numérica inteira que indica o número total de contadores que prestaram serviço à empresa.

O mesmo fenômeno dos sócios, ainda que com menos discrepância, acontece. As empesas inaptas ficaram, em sua maioria, no máximo com dois contadores. Interessante perceber que algumas empresas não cadastraram contador, sendo que, as inaptas têm esse caso em maior número.



Fonte: Elaborado pela autora.

Figura 6 – Distribuição - Contadores que já prestaram serviço à empresa.

De forma mais evidente, através dos números na Tabela 3:

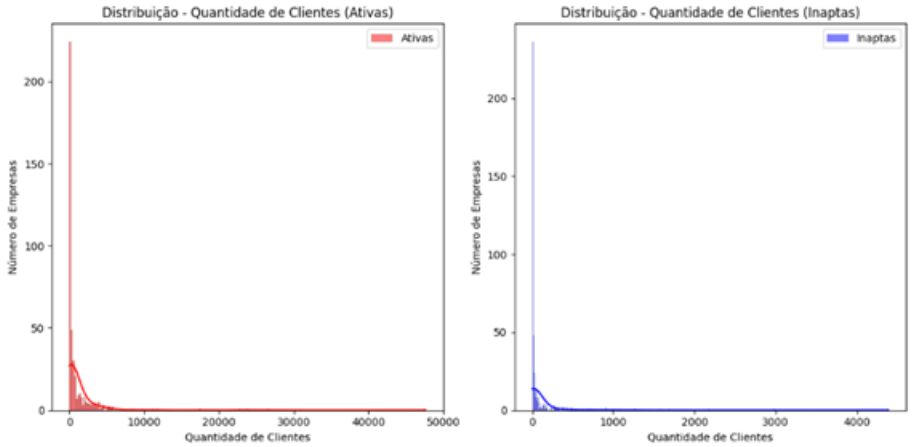
Quantidade de Contadores	% de Ativas	% de Inaptas
0	0,93	5,14
1	56,07	83,41
2	29,21	9,58
3	11,45	1,17
4	1,87	0,47
5	0,47	0,00
6	0,00	0,23

Tabela 3 – Distribuição percentual de quant_contadores para as empresas

4.2.2 Dados Econômico-Fiscais

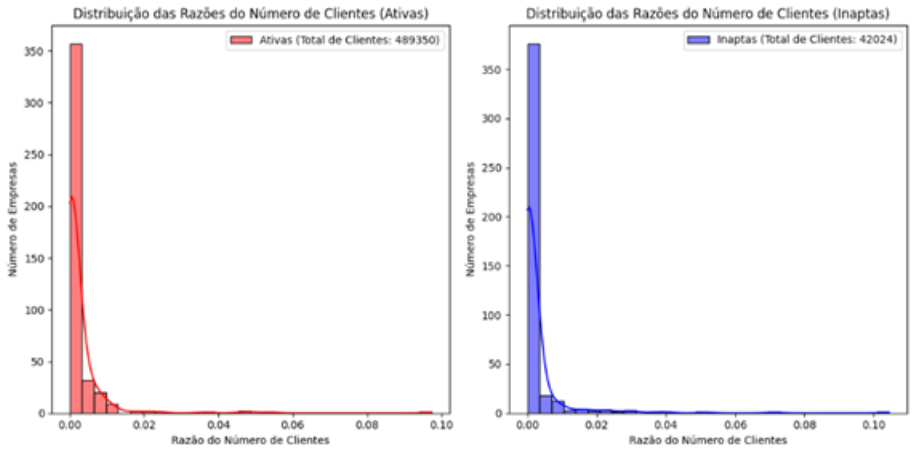
6. Quantidade de Clientes (quant_client): é uma variável numérica inteira que se refere, na verdade, ao número de empresas para as quais foram emitidas Notas Fiscais Ele-

trônicas pela empresa amostrada. Não necessariamente se refere a um cliente propriamente dito, mas, utilizou-se essa nomenclatura para simplificação e melhor entendimento.



Fonte: Elaborado pela autora.

Figura 7 – Distribuição - Quantidades de Clientes.



Fonte: Elaborado pela autora.

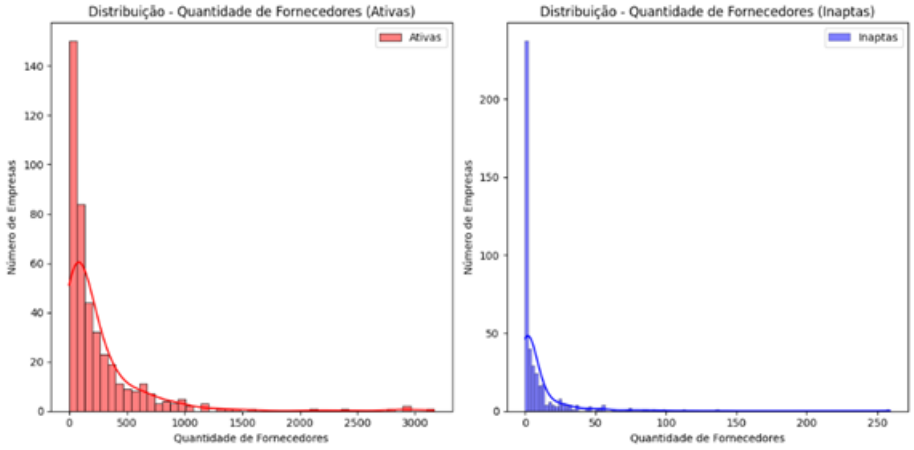
Figura 8 – Distribuição - Razão - Quantidade de Clientes/ Total de Clientes.

Uma característica que se observou ao verificar o número de Notas Fiscais Eletrônicas e os valores envolvidos de cada grupo é que as empresas inaptas apresentam um número bem menor de clientes. O número de destinatários de NF-e's para algumas empresas ativas ultrapassa 5 (cinco) mil empresas, enquanto para as inaptas apenas um percentual reduzido ultrapassa o milhar, nenhuma delas chegando aos 5(cinco) mil. Portanto, a visualização, tal como está na Figura 7, fica um pouco prejudicada.

Preferiu-se, então, apresentar o gráfico da razão de clientes de cada empresa pela soma de todos os clientes de cada grupo. Seja dizer o número de clientes de uma empresa inapta amostrada, sobre o somatório dos clientes das empresas inaptas. Esse procedimento resultou na Figura 8.

Percebe-se que o padrão de clientes é bem semelhante para os dois grupos, sendo que a maioria das empresas ficaram concentradas à esquerda do gráfico, indicando que, para os dois grupos, uma grande porcentagem das empresas tem 20% (vinte por cento) do número de clientes e os demais 80% (oitenta por cento), aproximadamente, seguem uma distribuição regular do número de clientes entre as empresas amostradas.

7. Quantidade de Fornecedores (quant_forn): é uma variável numérica inteira que se refere, na verdade, ao número de empresas que emitiram Notas Fiscais Eletrônicas para a empresa amostrada. Da mesma forma que os clientes, o número não necessariamente se refere a um fornecedor propriamente dito, mas, utilizou-se essa nomenclatura para simplificação e melhor entendimento.



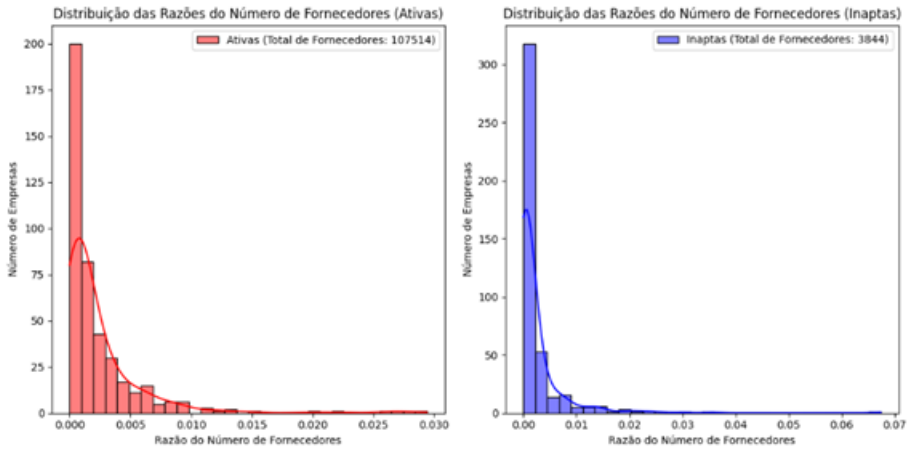
Fonte: Elaborado pela autora.

Figura 9 – Distribuição - Fornecedores.

Conforme se nota na Figura 9, como para ocorreu para os clientes, o número de NF-e's recebidas é relativamente menor para as empresas inaptas. Porém,algo particular para as inaptas é que há uma porcentagem considerável, cerca de 33% (trinta e três por cento), que não recebeu nenhuma nota fiscal, mesmo as empresas tendo emitido tais documentos, seja dizer, apesar de que tinham 0 (zero) fornecedores, o número de clientes foi diferente de 0 (zero).

A razão entre fornecedores entre as empresas, visualizada na Figura 10, é um pouco diferente entre os dois grupos. A distribuição dessa quantidade é mais gradual para as ativas. Os valores das notas recebidas estão mais pulverizados e há uma concentração de valores próximos ao zero para esta relação.

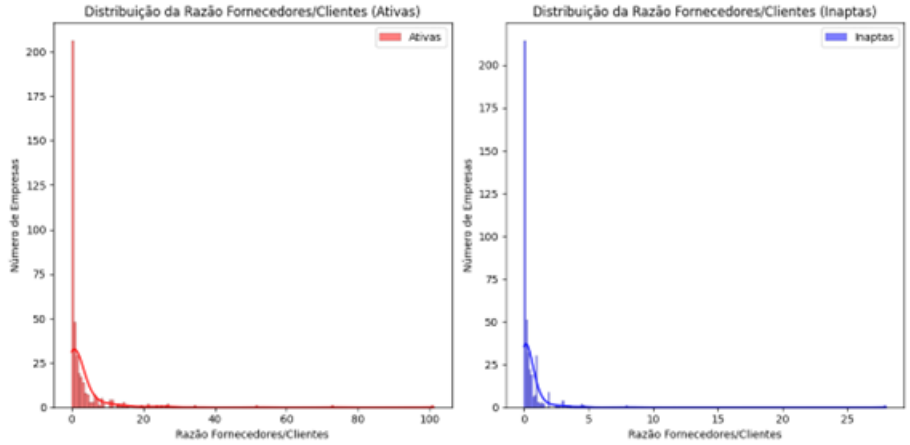
Isso se alinha ao fato de que há um padrão diferente de emissão de notas fiscais emitidas e recebidas pelas empresas inaptas em relação às ativas. Veja a Figura 11, a proporção dessa razão que, inicialmente, se aproxima de 0 (zero) para ambas, na sequência, para as ativas, parece quatro vezes maior, para cerca da metade das demais empresas. Reflete um comportamento típico das empresas inaptas que reforça sua razão de existirem



Fonte: Elaborado pela autora.

Figura 10 – Distribuição - Razão - Quantidade de Fornecedores/ Total de Fornecedores.

para emissão de documentos fiscais frios para saídas de mercadorias que não tiveram entradas correspondentes.

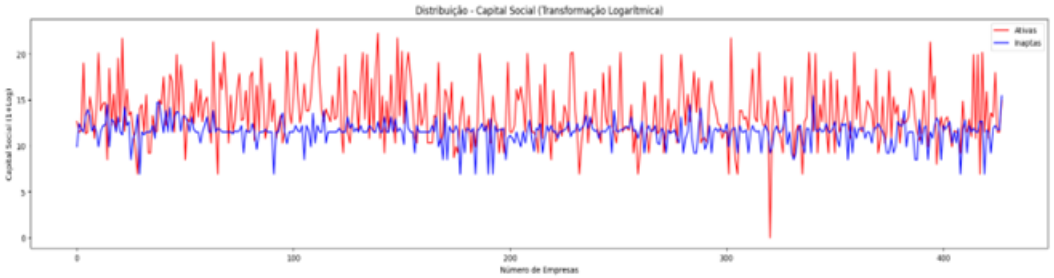


Fonte: Elaborado pela autora.

Figura 11 – Distribuição - Razão - Quantidade de Fornecedores/ Quantidade de Clientes.

8. Capital Social (capital_social): é uma variável numérica flutuante (float) que indica o valor declarado pela empresa como o valor aplicado para dar início ao funcionamento da empresa, ou seu crescimento. É uma variável que se enquadra tanto como dados econômico-fiscais, quanto cadastral, uma vez que é declarada pela empresa quando de sua constituição e cadastro na Junta Comercial do Estado.

Para o gráfico da Figura 12, utilizou-se a transformação logarítmica (1+Log) dos valores do Capital Social das empresas para melhor visualização. Tais valores, para as empresas ativas, variavam mais e em maior magnitude, não sendo fácil a visualização no gráfico de barras. Fica, após a transformação, mais favoráveo a observação desse comportamento. As empresas ativas possuem maiores valores e mais variação. As empresas



Fonte: Elaborado pela autora.

Figura 12 – Distribuição - Capital Social, submetidos a transformação logarítmica.

inaptas tem valores mais modestos, chegando ao máximo de milhões, enquanto às ativas têm exemplos com trilhões de reais.

Vale a pena apresentar os parâmetros estatísticos que favorecem a compreensão das impressões obtidas na Figura 12. Foi elaborada a Tabela 4 com eles:

Parâmetros Amostrais do Capital Social	Ativas	Inaptas
Média	R$ 94.438.100,17	R$ 214.048,83
Mediana	R$ 400.000,00	R$ 100.000,00
Desvio Padrão	488.077.681,34	456.362,51
Variância	2,38 e+17	2,08 e+11

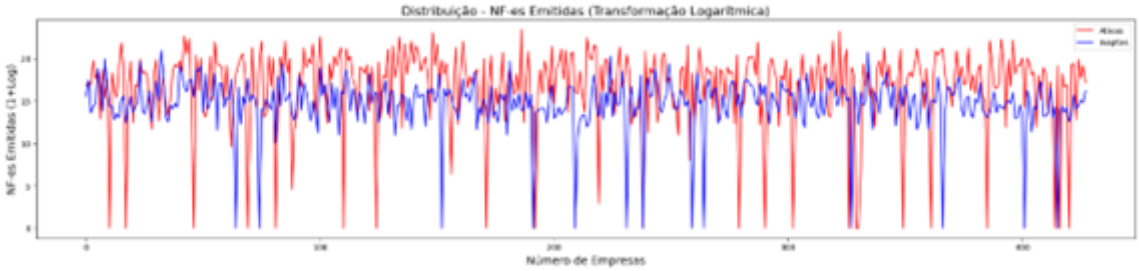
Tabela 4 – Parâmetros Estatísticos do Capital Social das Amostras

Para as empresas ativas obteve-se uma média significativamente maior, com uma mediana também maior, mas que não acompanhou o padrão da média, mostrando que a maior parte das ativas estão bastante abaixo da média dos valores para essa variável.

Já os parâmetros desvio padrão e variância que apresentam valores altos para as empresas ativas, demonstram que há uma variabilidade bem alta para essas empresas. Relativamente, essa variação é bem menor para as empresas inaptas. Isso vai de encontro ao fato de que os valores de capital social dessas últimas são de fato menores e tendem a gerar essa maior concentração.

9. Total de Notas Fiscais - Emitidas (total_emit): é uma variável numérica flutuante (float) que indica o valor total agregado das operações com mercadorias constantes das notas fiscais eletrônicas emitidas pela empresa, durante o período de Janeiro de 2019 a Agosto de 2024, ou até que a empresa tenha tido sua situação cadastral declarada como Inapta.

Diferentemente do Capital Social, para as notas fiscais emitidas, como pode ser percebido na Figura 13, há uma distância menor entre os valores tanto para as ativas quanto para as inaptas, ainda que essa diferença não seja imperceptível. Percebe-se também que ambos os grupos têm uma variabilidade grande nos valores de notas fiscais emitidas.



Fonte: Elaborado pela autora.

Figura 13 – Distribuição - Valores totais das Notas Fiscais Eletrônicas Emitidas, submetidos a transformação logarítmica.

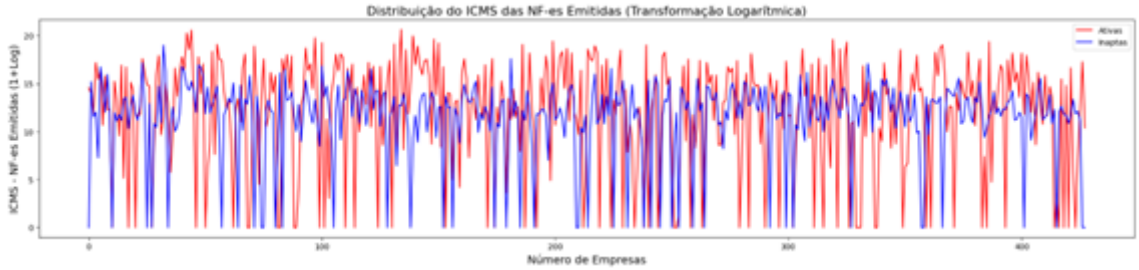
Os valores das empresas inaptas são menores, porém com um comportamento muito similar quando se olha para os dois grupos.

Para essa variável também foi elaborada a Tabela 5 com os parâmetros estatísticos.

Parâmetros Amostrais das NF-e's emitidas	Ativas	Inaptas
Média	R$ 502.070.530,05	R$ 23.350.918,43
Mediana	R$ 64.204.011,89	R$ 3.434.226,24
Desvio Padrão	1.402.952.537,20	94.997.060,00
Variância	1.97 e+18	9,02 e+15

Tabela 5 – Parâmetros Estatísticos dos Valores de NF-e's emitidas (total_emit)

10. Total de ICMS-Próprio - Emitidas (total imp emit): é uma variável numérica flutuante (float) que indica o valor total agregado do ICMS que deveria ter sido recolhido e está destacado nas notas fiscais eletrônicas emitidas pela empresa, durante o período de janeiro de 2019 a agosto de 2024, ou até que a empresa tenha tido sua situação cadastral declarada como inapta.



Fonte: Elaborado pela autora.

Figura 14 – Distribuição - ICMS de Notas Fiscais Emitidas.

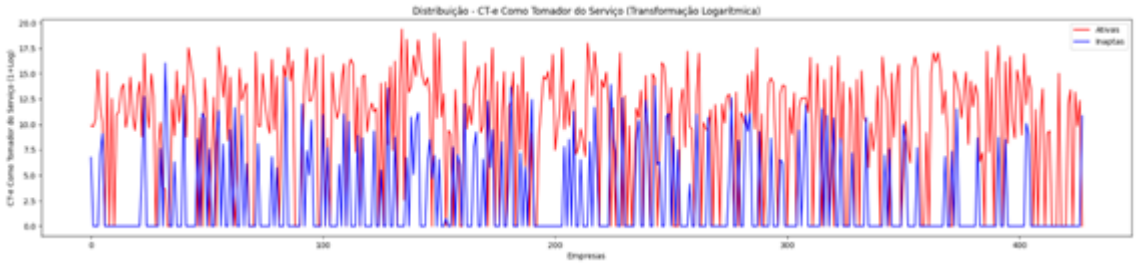
Visualmente a performance gráfica para o imposto destacado nas notas emitidas foi muito similar ao dos valores totais das notas fiscais. Grande variabilidade e valores menores para as empresas inaptas, em uma proporção semelhante.

11. Total de ICMS-ST - Emitidas (total_ st_emit): é uma variável numérica flutuante (float) que indica o valor total agregado do ICMS que deveria ter sido recolhido

como substituto tributário e está destacado nas notas fiscais eletrônicas emitidas pela empresa, durante o período de janeiro de 2019 a agosto de 2024, ou até que a empresa tenha tido sua situação cadastral declarada como inapta.

12. Quantidade de Notas Fiscais - Emitidas – (quant_nf_emit): é uma variável numérica inteira que indica o número total de notas fiscais eletrônicas emitidas pela empresa, durante o período de janeiro de 2019 a agosto de 2024, ou até que a empresa tenha tido sua situação cadastral declarada como inapta.

13. Base de Cálculo do Imposto dos Conhecimentos de Transporte (CTes) como Tomador do Serviço (total_bc_cte): é uma variável numérica flutuante (float) que indica o valor total tributável nos documentos fiscais de transporte de suas mercadorias, denominados Conhecimentos de Transporte. Esses documentos são emitidos pela empresa transportadora, mas, em geral, é a empresa proprietária das supostas mercadorias que consta nesses documentos como tomadora do serviço de transporte e é responsável pelo pagamento do ICMS. Os valores agregados englobam o período de janeiro de 2019 a agosto de 2024, ou até que a empresa tenha tido sua situação cadastral declarada como inapta.



Fonte: Elaborado pela autora.

Figura 15 – Distribuição - Valores dos CT-e's submetidos a transformação logarítmica

Como nos demais casos, o gráfico mostra a distribuição do total de Conhecimentos de Transporte Eletrônico (CT-e) como tomador do serviço para empresas ativas e inaptas, utilizando uma transformação logarítmica. Utilizou-se a transformação logarítmica com o objetivo de reduzir a escala dos dados e facilitar a visualização de diferenças entre os grupos.

Esse gráfico se apresentou bem interessante, pois é possível notar que a diferença de comportamento entre os grupos é bem grande. Percebe-se que há muitas empresas inaptas que não emitiram nenhum conhecimento de transporte, ou seja, no processo de simulação de operações não se deram ao trabalho de também simular o transporte das mercadorias.

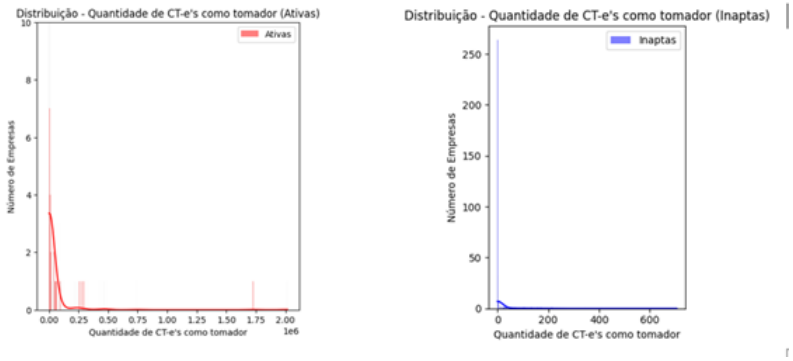
Isso gera uma variabilidade maior para os valores das inaptas. Ressalta-se que entre as 428 (quatrocentos e vinte e oito) empresas inaptas amostradas, 264 (duzentos e sessenta e quatro) não emitiram nenhum conhecimento de transporte e olhando os dados, mais de 85% (oitenta e cinco por cento) emitiram menos de 10 (dez) desses documentos.

Já para as ativas, os casos zerados são poucos mais de 10% (dez por cento) e estão relacionados, na maioria expressiva, a casos de não emissões de notas fiscais também. Isso indica que não emitiram conhecimentos de transporte por não terem registrado nenhuma saída de mercadorias, o que faz sentido, contrariamente ao caso anterior.

12. ICMS de Conhecimentos de Transporte (CTes) como Tomador do Serviço (total_imp_cte): é uma variável numérica flutuante (float) que indica o valor total agregado do ICMS que deveria ter sido recolhido e está destacado nos Conhecimentos de Transporte para a empresa como tomador do serviço, durante o período de janeiro de 2019 a agosto de 2024, ou até que tenha tido sua situação cadastral declarada como inapta.

13. Total de ST retido nos CTes como Tomador do Serviço (total_st_cte): é uma variável numérica flutuante (float) que indica o valor total agregado do ICMS que deveria ter sido recolhido pela empresa como substituto tributário e está destacado nos Conhecimentos de Transporte para a empresa como tomador do serviço, durante o período de janeiro de 2019 a agosto de 2024, ou até que tenha tido sua situação cadastral declarada como inapta.

14. Quantidade de CTes como Tomador do Serviço (quant_cte): é uma variável numérica inteira que indica o número total de Conhecimentos de Transporte para a empresa como tomador do serviço, durante o período de janeiro de 2019 a agosto de 2024, ou até que tenha tido sua situação cadastral declarada como inapta.



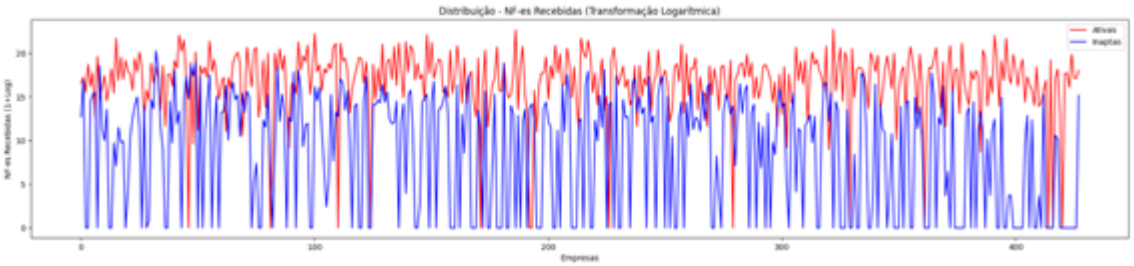
Fonte: Elaborado pela autora.

Figura 16 – Distribuição - Quantidade de CT-e's.

A Figura 16 confirma o que foi verificado no gráfico "Distribuição - Valores dos CT-e's Como Tomador do Serviço (Transformação Logarítmica)" – Figura 15, em que se tem maiores valores para os CT-e's para as ativas, com uma boa distribuição. Essas ativas chegam com valores de até milhões de CT-e's. Sendo que temos um número considerável de inaptas sem algum CT-e ou com muito poucas emissões e os valores apresentados na Figura 15 são baixos. Além disso, o número máximo de emissões por uma empresa inapta não chega a 800 (oitocentos) na Figura 16.

15. Total de Notas Fiscais - Destinatário – (total_dest): é uma variável numérica

flutuante (float) que indica o valor total agregado das operações com mercadorias constantes das notas fiscais eletrônicas emitidas para a empresa, durante o período de Janeiro de 2019 a Agosto de 2024, ou até que tenha tido sua situação cadastral declarada como Inapta.



Fonte: Elaborado pela autora.

Figura 17 – Distribuição - Valores de Notas Fiscais Recebidas, submetidos a transformação logarítmica.

Parâmetros Amostrais da Varíavel total_dest	Ativas	Inaptas
Média	R$ 264.905.052,29	R$ 6.670.460,98
Mediana	R$ 37.702.813,18	R$ 92.834,20
Desvio Padrão	721.524.958,52	34.799.224,36
Variância	5,21 e+17	1,21 e+15

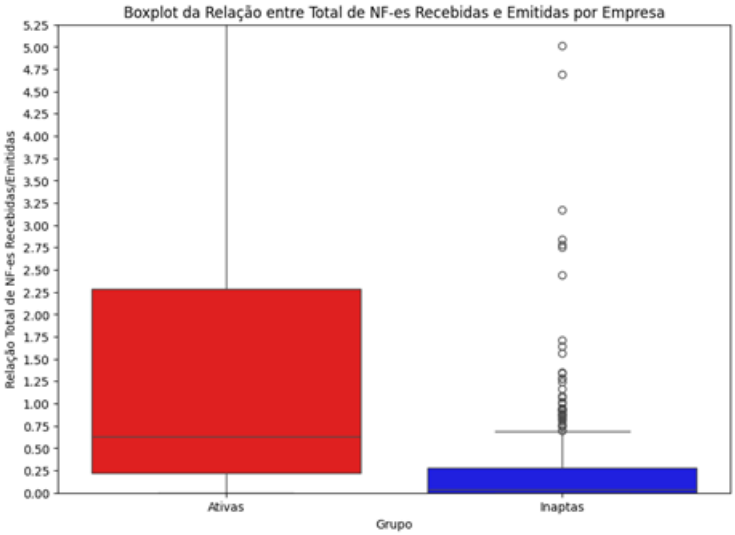
Tabela 6 – Parâmetros Estatísticos dos Valores de NF-e's recebidas (total_dest).

A observação desse gráfico (Figura 17) e dos parâmetros estatísticos (Tabela 6 dos valores totais das Notas Fiscais recebidas nos permite constatar que a variabilidade de valores dos dois grupos é grande sendo maior para as ativas, uma vez que têm um espectro maior. No entanto, se confrontarmos a média e a mediana, podemos observar que há uma quantidade muito maior de valores abaixo da média para as inaptas. Fica mais uma vez demonstrado que uma grande quantidade de empresas inaptas que emitem Notas Fiscais de saída, sem terem recebido nenhuma nota de entrada (o que se observa nos valores zerados no gráfico).

Ainda, quando se volta o olhar para os dados propriamente ditos, descobre-se que 199 (cento e noventa e nove) das 428 (quatrocentos e vinte e oito) empresas não receberam notas fiscais. Mesmo que quase todas essas, em um total de 190 (cento e noventa), tenham dado saída a mercadorias ou serviços.

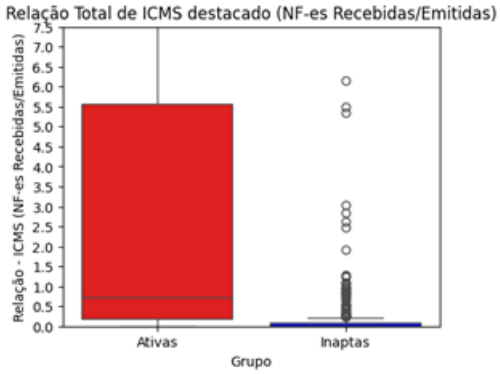
Os boxplot's que seguem (Figura 18 e 19) nos permitem visualizar de forma bem clara a questão de Notas Fiscais emitidas sem prévio recebimento por parte das inaptas.

Importante é notar que as empresas ativas (que foram já fiscalizadas pela Secretaria da Fazenda – SP) apresentaram uma tendência a ter mais registros de notas recebidas (que lhes dão crédito) do que notas emitidas. Isso se pode dizer pois, as relações para as ativas apresenta valores maiores que 1 (um), tanto das notas fiscais como do imposto, na



Fonte: Elaborado pela autora.

Figura 18 – Comparação da Razão de Valores de Notas Fiscais Recebidas/ Valores Notas Fiscais Emitidas.



Fonte: Elaborado pela autora.

Figura 19 – Comparação da Razão do ICMS destacado nas Notas Fiscais Recebidas/ ICMS das Notas Fiscais Emitidas.

Figura 19. No entanto, as Inaptas que tiveram uma quantidade menor de notas fiscais, apresentaram as relações em sua maior parte abaixo de 1 (um).

Ambos os casos acima, são situações atípicas que chamam a atenção do fisco, comprovando coerência na escolha de seus alvos para fiscalização.

16. Total de ICMS-próprio - Destinatário – (total_imp_dest): é uma variável numérica flutuante (float) que indica o valor total agregado do ICMS que deveria ter sido recolhido e está destacado nas notas fiscais eletrônicas emitidas para a empresa, durante o período de janeiro de 2019 a agosto de 2024, ou até que tenha tido sua situação cadastral declarada como inapta.

17. Total de ICMS-ST - Destinatário (total_st_dest): é uma variável numérica flutuante (float) que indica o valor total agregado do ICMS que deveria ter sido recolhido

como substituto tributário e está destacado nas notas fiscais eletrônicas emitidas para a empresa, durante o período de janeiro de 2019 a agosto de 2024, ou até que tenha tido sua situação cadastral declarada como inapta.

18. Quantidade de Notas Fiscais - Destinatário (quant_nf dest): é uma variável numérica inteira que indica o número total notas fiscais eletrônicas emitidas para a empresa, durante o período de janeiro de 2019 a agosto de 2024, ou até que tenha tido sua situação cadastral declarada como inapta.

Por último, porém de grande importância é a variável alvo, também denominada *label* no ramo de Ciência de Dados.

19. Atividade (atividade): Refere-se à situação cadastral do contribuine e consiste em uma variável numérica booleana que representará ou o grupo empresa declarada como inapta (não está em atividade) pela SEFAZ-SP ou o grupo empresa ativa (existe e está em atividade) cuja existência foi constatada pela SEFAZ-SP, por meio de fiscalização.

A empresa ativa recebeu o número 1 (um) como rótulo e a empresa inapta recebeu o numeral 0 (zero).

Para ambos os grupos foram consideradas apenas empresas que tiveram abertura a partir do primeiro dia de janeiro de dois mil e dezenove (01/01/2019) até trinta e um de agosto de dois mil e vinte e quatro (31/08/2024).

4.2.3 Correlação entre as variáveis

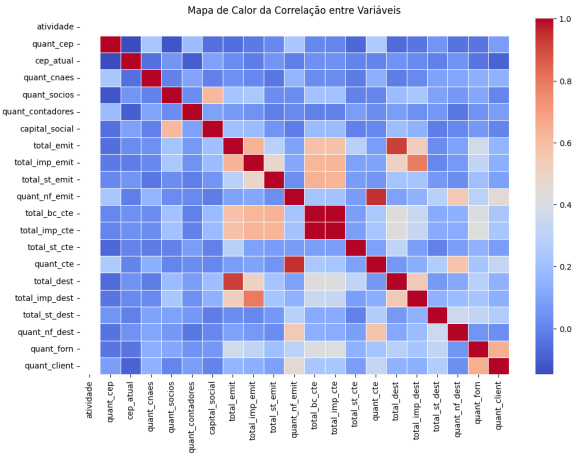
Apresenta-se aqui a correlação entre as variáveis, tanto para o grupo de treinamento, como para os dados inéditos.

A Figura 20 traz a correlação dos dados de empresas ativas, a Figura 21, das empesas inaptas, seguidas pela correlação das empresas mistas, na Figura 22.

As variáveis variaram em termos de correlação. Há algumas com forte correlação, algumas com média correlação e sem correlação.

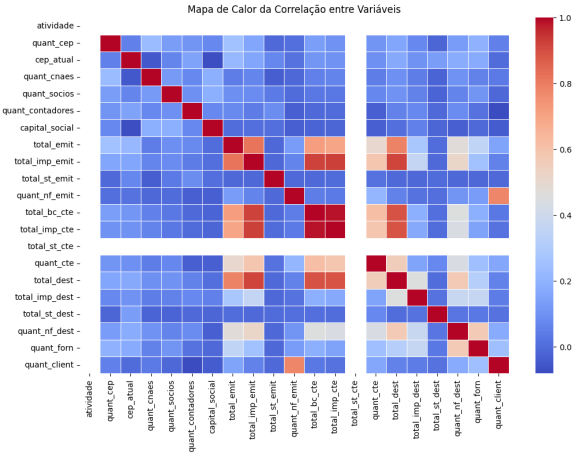
No entanto, é importante avaliar os três mapas em conjunto e verificar as correlações próximas não se repetem nos três grupos.

Exceto para a variável total_bc_cte e total_imp_cte que houve forte correlação nos três grupos. Porém, considerando que na análise gráfica dos CT-e's, percebeu-se a importância desses documentos, optou-se por manter todas as variáveis a eles relacionadas.



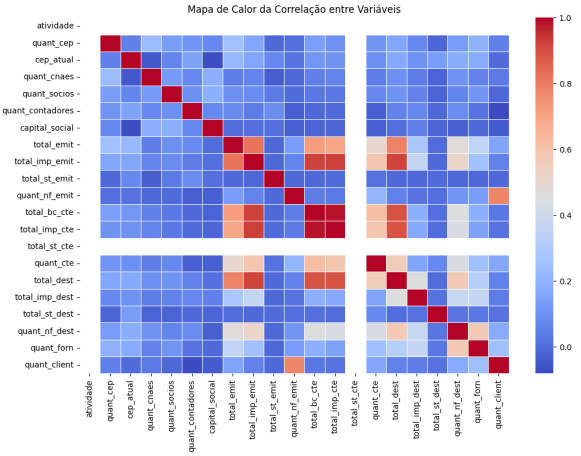
Fonte: Elaborado pela autora.

Figura 20 – Mapa de Calor das Variáveis do Grupo de Ativas



Fonte: Elaborado pela autora.

Figura 21 – Mapa de Calor das Variáveis do Grupo de Inaptas



Fonte: Elaborado pela autora.

Figura 22 – Mapa de Calor das Variáveis do Grupo de Mistas

4.2.4 Preparação do Dados

Após o entendimento, passa-se à **Preparação dos Dados**, ou seja, olhar para as variáveis de forma a compreender de que forma elas atenderão à solução do problema. É preciso verificar se há dados faltantes para alguma das variáveis e é preciso fazer os ajustes necessários que foram identificados no passo anterior.

No caso em estudo, foi feita a distribuição aleatória com o uso de python dos dados descritos abaixo.

Foram extraídas do banco de dados da Secretaria (1308) um mil e trezentos e oito observações no total, conforme se descreverá abaixo.

Foram encontradas 571 (quinhentos e setenta e uma) empresas que foram declaradas inaptas pela SEFAZ-SP de acordo com a Portaria CAT 95/2006 (São Paulo, 2006), todas foram consideradas para o estudo.

Foram encontradas 737 empresas fiscalizadas, para as quais constatou-se a atividade e que se mantiveram ativas até a coleta dos dados. Dessa forma, presume-se com razoável nível de certeza que a empresa existe e está em funcionamento, nos servindo como boa fonte para empresas ativas.

Para a criação de um modelo os grupos serão formados da seguinte maneira:

1. Primeira metade do Grupo de Modelagem (Inaptas)

- Composição: 428 empresas de fachada.
- Uso: Exclusivamente para treinar o modelo com empresas de fachada.

2. Segunda metade do Grupo de Modelagem (Empresas Ativas)

- Composição: 428 empresas ativas.
- Uso: Exclusivamente para treinar o modelo com empresas consideradas corretas.

3. Grupo de Validação (Mistas) o Composição: 143 empresas inaptas e 309 empresas ativas). o Uso: Para avaliar o desempenho do modelo em dados não vistos.

Os dados de modelagem apresentaram um balanceamento perfeito (428 Inaptas) para 428 (Ativas), enquanto os dados inéditos (Mistas), tiveram um desbalanceamento que será considerado para as análises de avaliação. Foram aproximadamente 30% (trinta por cento) de empresas inaptas, para 70% (setenta por cento) de empresas ativas.

4.3 Treinamento dos Algoritmos

O treinamento dos algoritmos, que é o passo seguinte e o mais técnico, consistiu, basicamente, na divisão dos dados em treinamento e teste, no caso, utilizou-se oitenta por cento (80%) da mesclagem dos Grupos de Treinamento, Inaptas e Ativas, para o treinamento e vinte por cento (20%) para teste.

Os algoritimos utilizados foram os já, mais detalhada mente, apresentados na Seção 2.1, quais sejam: K-Nearest Neighbor (KNN), Florestas Aleatórias (Random Forest), Gradient Boosting, Naive Bayes e Regressão Logísitica.

Para tanto foram utilizadas bibliotecas scikit-learn da linguagem de programação python para tal treinamento. (learn, 2007-2024)

Os modelos foram treinados da seguinte forma:

- Inicialmente, fez-se uma modelagem pura e simples partindo-se de uma baseline com a biblioteca correspondente a cada Modelo e obteve-se para cada uma delas as métricas que foram conservadas para a próxima etapa que é a avaliação.

- Para cada modelo foi testada a possibilidade de normalização de dados que também será avaliada oportunamente.

- Finalmente, no aprimoramento do estudo, foi feito o refinamento de cada modelo utilizando-se o GridSearchCV com a validação cruzada de Monte Carlo, para completar as informações a serem avaliadas.

A próxima seção traz as avaliações detalhadas.

4.4 Avaliação dos Modelos

Após o treinamento, conforme a Seção 4.3, foi feita a coleta das métricas para cada algoritmo e o respectivo modelo testado.

Nesses modelos os dados utilizados são balanceados (428 (quatrocentos e vinte e oito) inaptas/ 428 ativas), dessa forma vamos fazer as comparações nessa fase a métrica acurácia, que é a melhor para dados balanceados.

4.4.1 KNN - K-NearestNeighbors

Tem-se abaixo a Tabela 7 para as métricas obtidas para o grupo teste após o treinamento.

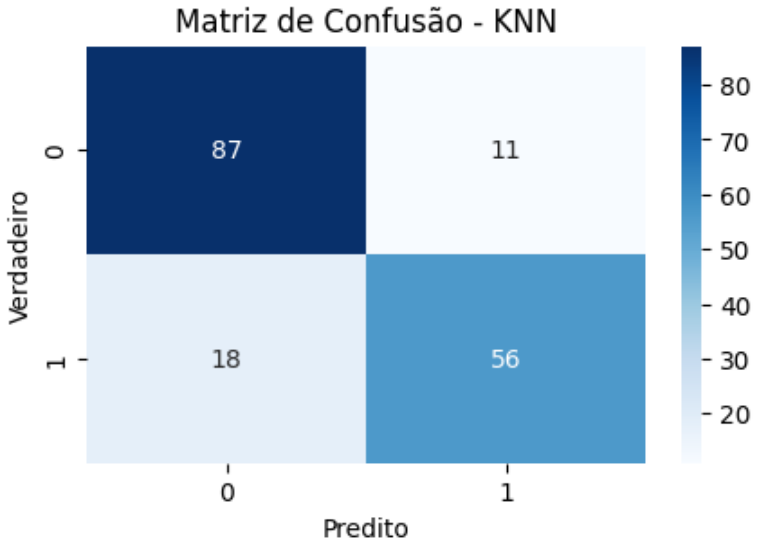
Com métricas coerentes entre si em termos de variação entre os modelos de KNN, o melhor resultado para o grupo de teste foi o KNN-Simples, com acurácia de 0.8314 ou 83.14%.

Modelo	Acurácia	Precisão	Revogação	F1-Score
KNN Simples	0.837209	0.838060	0.837209	0.835750
KNN Simples Normalizado	0.808140	0.807604	0.808140	0.807172
KNN com Monte Carlo	0.831395	0.831690	0.831395	0.830117

Tabela 7 – Métricas do Modelo K-Nearest Neighbors para os dados em treinamento

A normalização não melhorou o treinamento do modelo. Apesar da normalização em geral ser benéfica para os modelos, não é uma regra absoluta. Alguns fatores podem não corresponder idealmente à normalização. Por exemplo, a distribuição dos dados pode já estar adequada, a interação entre as variáveis que vimos não é absolutamente independente, ou mesmo a versão do algoritmo pode ser favorecida ou não pela normalização. Nesse caso, não houve ganho.

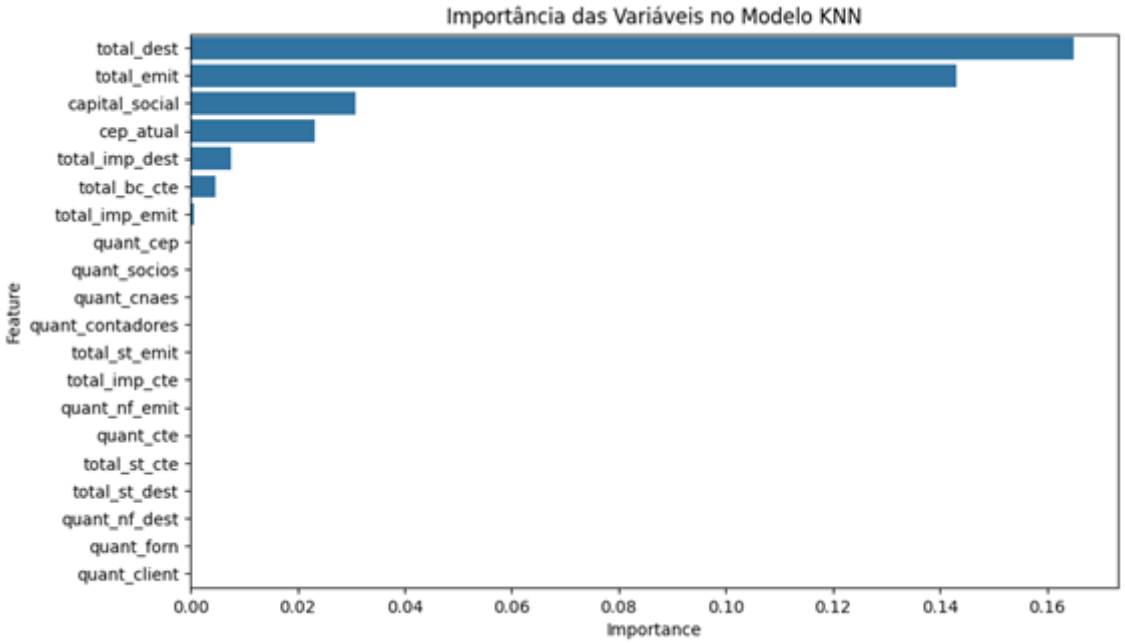
A matriz de confusão na Figura 23 favorece o entendimento do resultado da classificação, com a identificação dos grupos de empresas. Percebe-se que o algoritmo classificou corretamente 87 (oitenta e sete) empresas ativas e 56 (cinquenta e seis) empresas inaptas do total de 172 (cento e setenta duas) empresas usadas para o grupo de teste. Sendo ainda, que 11 (onze) inaptas foram classificadas como ativas e 18 (dezoito) ativas erroneamente como inaptas. Em resumo, esse modelo permitiu a detecção de 75% das empesas inaptas do grupo de teste. Não é um resultado, propriamente, ruim, mas, espera-se mais. Além disso, tivemos quase 20% (vinte por cento) de falsos negativos que, nesse caso, é ruim. No uso deses resultados para ações fiscais, empresas idôneas podem vir a sofrer penalidades indevidamente.



Fonte: Elaborado pela autora.

Figura 23 – Matriz de Confusão - K-Nearest Neighbors

Em relação à importância das variáveis mostrada na Figura 24 para o modelo



Fonte: Elaborado pela autora.

Figura 24 – Importância das Variáveis - KNN

KNN, percebeu-se pouca influência dos dados cadastrais e um peso significativo dos sócio-econômicos. Fica sugerido um novo modelo mais enxuto nas *features* ou com a inserção de mais variáveis nesse último enfoque.

4.4.2 Florestas Aleatórias (RF)

Apresenta-se, adiante, a Tabela 8 para as métricas obtidas para o grupo teste após o treinamento.

As métricas foram batante coerentes entre si. O melhor resultado para o grupo de teste do modelo foi o RF-Simples, o modelo clássico do algoritmo de Florestas Aleatórias da biblioteca Scikit_learn em python, com a acurácia de 0.9167 ou 91.67%. Repare que é o mesmo resultado obtido para o modelo com dados normalizados (para Acurácia e para Precisão também). Sendo assim, a opção é pela não normalização pois oferece um ótimo modelo, mais simples e com menos custo computacional.

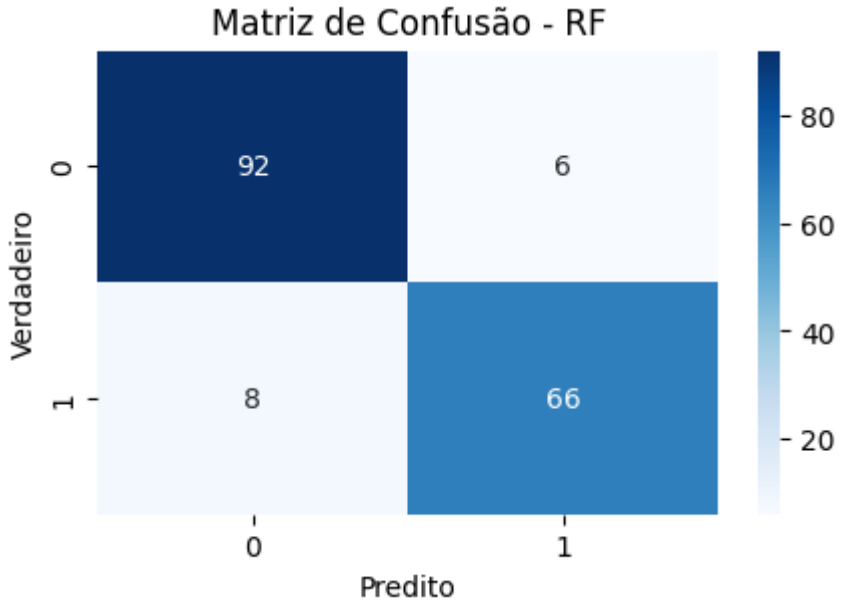
Modelo	**Acurácia**	**Precisão**	**Revogação**	**F1-Score**
RF Simples	0.918605	0.916667	0.891892	0.904110
RF Simples Normalizado	0.918605	0.916667	0.891892	0.904110
RF com Monte Carlo	0.912791	0.912871	0.912791	0.912546

Tabela 8 – Métricas do Modelo Florestas Aleatórias para os dados em treinamento

O refinamento na mesma direção, além dos resultados menos favoráveis, gastou mais tempo para treinar. O treinamento do algoritmo simples foi imediato, mesmo contando o

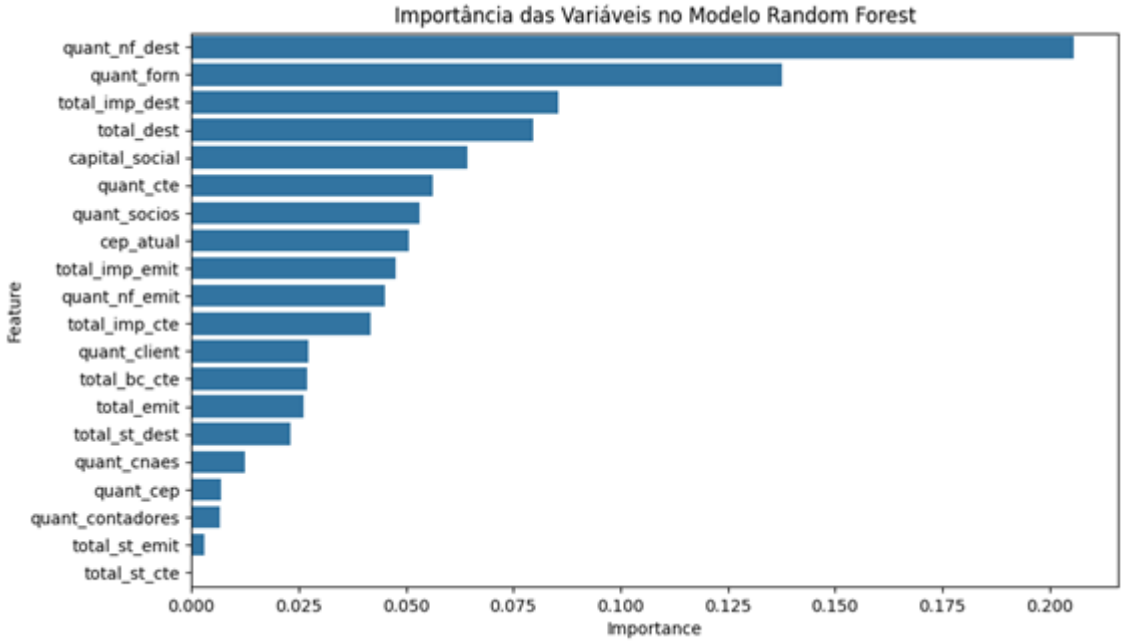
tempo de normalização, se fosse necessária. Enquanto isso, com o refinamento houve o decurso de 8.9s (oito segundos e 9 décimos).

Apesar de a classificação dos novos dados ter sido imediata, teme-se que com dados mais extensos, que o modelo refinado seja mais desfavorável.



Fonte: Elaborado pela autora.

Figura 25 – Matriz de Confusão - Florestas Aleatórias



Fonte: Elaborado pela autora.

Figura 26 – Importância das Variáveis - RF

A matriz de confusão está na Figura 25. No caso, o algoritmo classificou corretamente 92 (noventa e duas) empresas ativas e 66 (sessenta e seis) empresas inaptas do total de 172 (cento e setenta duas) empresas usadas para o grupo de teste. Sendo ainda, que 6 (seis) inaptas foram classificadas como ativas e 8 (oito) ativas erroneamente como inaptas. A porcentagem de acerto para as inaptas nesse grupo de teste foi considerada ótima, superando 90% (noventa por cento).

Quase todas as variáveis disponíveis foram aproveitadas pelo algoritmo no treinamento (Figura 26). A importância parece diminuir em algumas das variáveis disponíveis, que poderiam ter avaliadas sua relevância em um novo estudo.

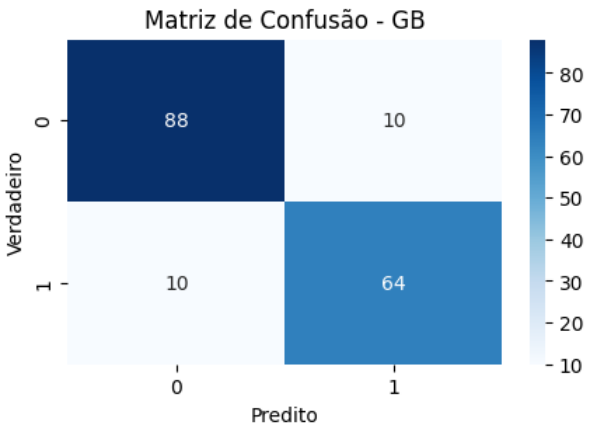
4.4.3 Gradient Boosting (GB)

Tem-se abaixo a Tabela 9 para as métricas obtidas para o grupo teste após o treinamento.

Modelo	Acurácia	Precisão	Revogação	F1-Score
GB Simples	0.906977	0.907263	0.906977	0.906615
GB Simples Normalizado	0.906977	0.914286	0.864865	0.888889
GB com Monte Carlo	0.883721	0.883721	0.883721	0.883721

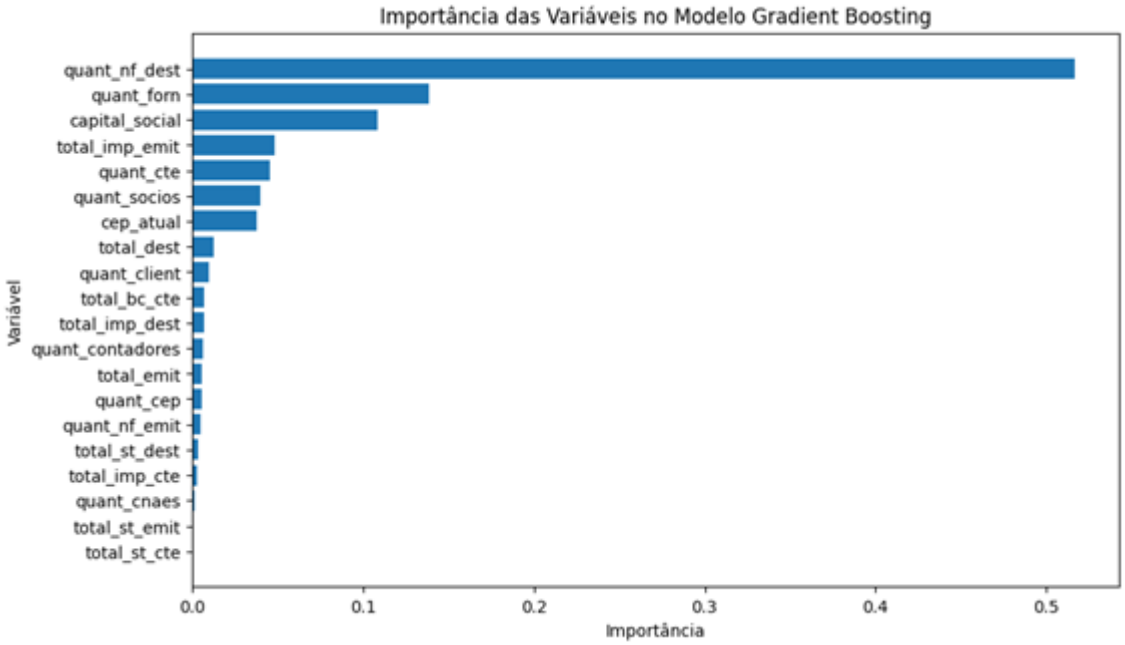
Tabela 9 – Métricas do Modelo Gradient Boosting para os dados em treinamento

O melhor resultado para o grupo de teste do modelo foi o GB-Simples, o modelo clássico do algoritmo de Gradient Boosting da biblioteca Scikit_learn em python, com a acurácia de 0.9070 ou 90.70%. Repare na Tabela 9 que é o mesmo resultado obtido para o modelo com dados normalizados. A precisão foi maior nos dados normalizados. Ainda assim, definimos a acurácia como referência, pela sua afinidade a dados balanceados e mantemos como preferido aqui o modelo simples. Além disso, as outras métricas oferecem oscilação.



Fonte: Elaborado pela autora.

Figura 27 – Matriz de Confusão - Gradient Boosting



Fonte: Elaborado pela autora.

Figura 28 – Importância das Variáveis - GB

A matriz de confusão para esse modelo está na Figura 27. No caso, o algoritmo classificou corretamente 88 (oitenta e oito) empresas ativas e 64 (sessenta e quatro) empresas inaptas do total de 172 (cento e setenta duas) empresas usadas para o grupo de teste. Sendo ainda, que 10 (dez) inaptas foram classificadas como ativas e 10 (dez) ativas erroneamente como inaptas. Oitenta e seis por cento (86%) das inaptas foram detectadas, resultando muito boa a classificação.

Para o Gradient Boosting uma variável, a Quantidade de Notas Fiscais Recebidas, foi destaque. Algumas outras independentes entre si tiveram uma importância média, com algumas outras com pouca ou nenhuma importância.

4.4.4 Regressão Logística (RL)

Tem-se abaixo a tabela 10 para as métricas obtidas para o grupo teste após o treinamento.

Modelo	**Acurácia**	**Precisão**	**Revogação**	**F1-Score**
RL Simples	0.848837	0.848603	0.848837	0.848249
RL Simples Normalizado	0.866279	0.859155	0.824324	0.841379
RL com Monte Carlo	0.883721	0.883799	0.883721	0.883269

Tabela 10 – Métricas do Modelo Regressão Logística para os dados em treinamento

O melhor resultado para o grupo de teste do modelo foi o RL com Monte Carlo, com a acurácia de 0.8838 ou 88.38%, sem muita discussão, afinal todas as métricas foram

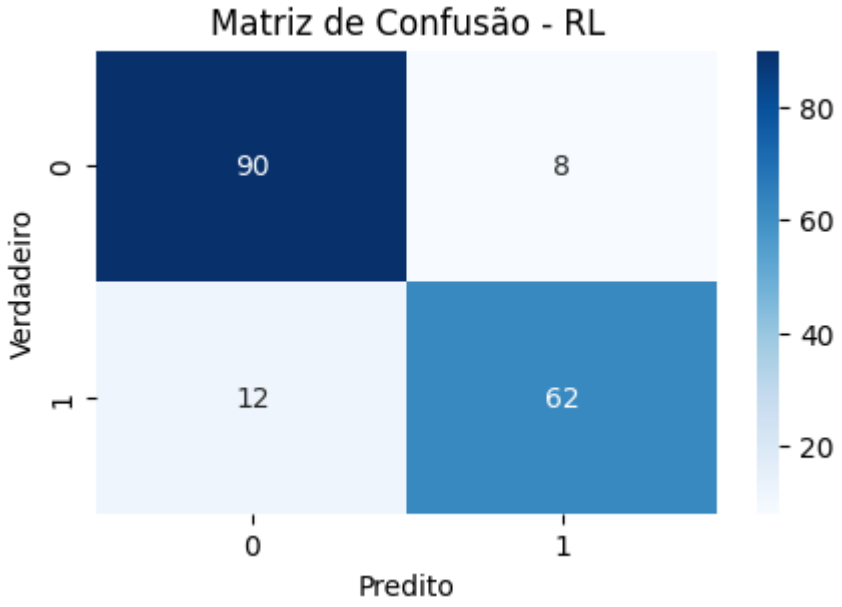
melhores para o modelo otimizado. Isso demonstra que aumentar a variabilidade das amostras teve um impacto significativo para a Regressão Logística.

A matriz de confusão para esse modelo está na Figura 29. No caso, o algoritmo classificou corretamente 90 (noventa) empresas ativas e 62 (sessenta e dois) empresas inaptas do total de 172 (cento e setenta duas) empresas usadas para o grupo de teste. Sendo ainda, que 8 (oito) inaptas foram classificadas como ativas e 12 (doze) ativas erroneamente como inaptas. Mais um resultado muito bom, quase 89% de acertos para as inaptas, na classificação do grupo teste.

A gráfico da importância das variáveis, Figura 30, tem uma visualização interessante para o modelo de Regressão Logística. Nesse modelo as variáveis que afetam negativamente a probabilidade de se classificar a classe verdadeira aparecem. Mas, sua influência, que é negativa para as ativas, é importante, se pensarmos que ela é positiva para detecção das empresas inaptas.

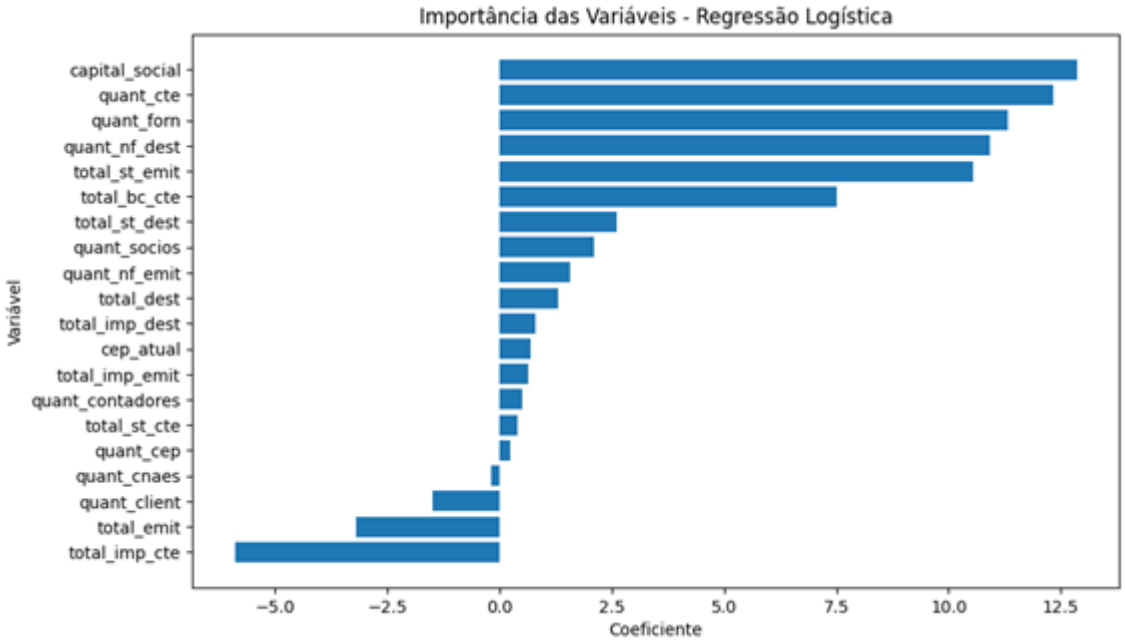
Outra observação importante é que as variáveis para as quais já se havia observado diferenças e variabilidade tiveram grande importância na classificação com esse modelo. Essas informações são interessantes, quando se pensa em insistir em um modelo de Regressão Logística.

É natural e intuitivo, mas principalmente para esse modelo a variabilidade pesa bastante e ficou evidenciado a importância, de encontrar variáveis com essas características nas empresas e nas suas operações.



Fonte: Elaborado pela autora.

Figura 29 – Matriz de Confusão - Regressão Logística



Fonte: Elaborado pela autora.

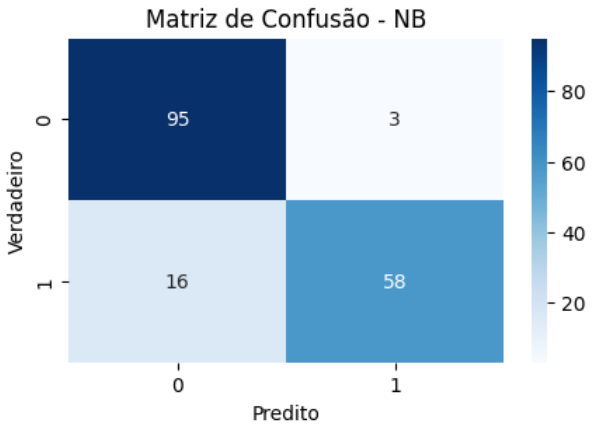
Figura 30 – Importância das Variáveis - RL

4.4.5 Naive Bayes (NB)

Tem-se a tabela 11 para as métricas obtidas para o grupo teste após o treinamento.

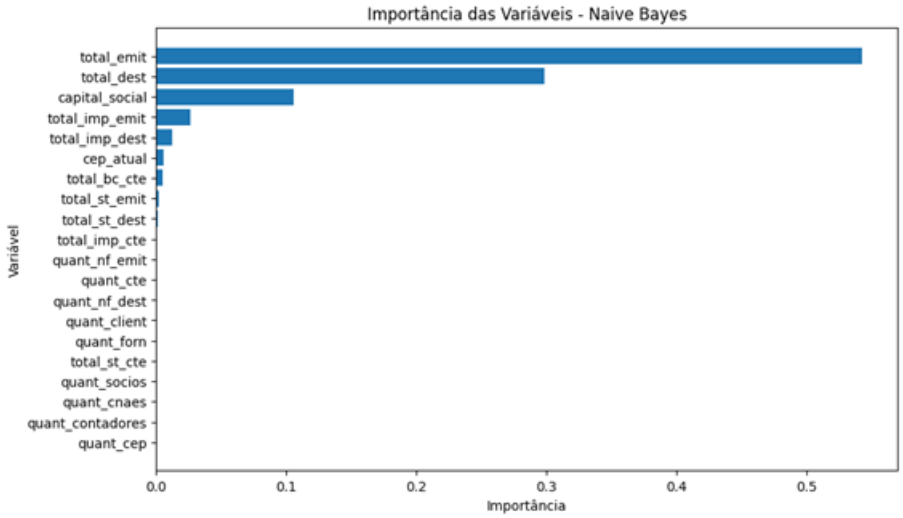
Modelo	Acurácia	Precisão	Revogação	F1-Score
NB Simples	0.802326	0.831760	0.802326	0.791618
NB Simples Normalizado	0.889535	0.896712	0.889535	0.887652
NB com Monte Carlo	0.889535	0.896712	0.889535	0.887652

Tabela 11 – Métricas do Naive Bayes para os dados em treinamento



Fonte: Elaborado pela autora.

Figura 31 – Matriz de Confusão - Naive Bayes



Fonte: Elaborado pela autora.

Figura 32 – Importância das Variáveis - NB

O melhor resultado para o grupo de teste do modelo, com a acurácia de 0.8895 ou 88.95%, foi o algoritmo clássico de Naive Bayes da biblioteca Scikit_learn em python, com dados normalizados. O modelo treinado com otimização de parâmetros e validação cruzada de Monte Carlo teve um resultado igual ao escolhido, porém com um tempo maior de execução.

A matriz de confusão para esse modelo está na Figura 31. No caso, o algoritmo classificou corretamente 95 (noventa e cinco) empresas ativas e 58 (cinquenta e oito) empresas inaptas do total de 172 (cento e setenta duas) empresas usadas para o grupo de teste. Sendo ainda, que 3 (três) inaptas foram classificadas como ativas e 16 (dezesseis) ativas erroneamente como inaptas. Foi um acerto de 95% das inaptas, um resultado bom, nesse sentido. Porém, os falsos negativos aqui chegaram a quase 15%.

A importância das variáveis atestadas na Figura 32 é principalmente relativa aos valores totais dos documentos fiscais, o imposto destacado nelas e das variáveis cadastrais apenas o Capital Social se destacou, com uma pequena influência do CEP atual.

A próxima etapa e final para o trabalho é aplicação do modelo a dados inéditos.

4.5 Deployment - Implantação da solução

Foram aplicados os modelos treinados e testados previamente nos dados novos. Optou-se por testar todos os modelos e avaliar para se ter uma percepção ampla do comportamento dos modelos diante dos novos dados.

A ideia é compreender se os dados utilizados no treino favoreceram ou não um sub ou sobreajuste dos modelos.

Outra avaliação que se quer fazer é que considerando o balanceamento nos dados

de treinamento, como o desbalanceamento dados novos poderia impactar nos resultados.

É óbvio que no mundo real há uma quantidade maior de empresas que existem de fato contra aquelas fraudulentas, de forma que se deseja que o modelo funcione bem ainda que os dados estejam desbalanceados.

Em realidade, o desbalanceamento não é agressivo, até mesmo por conta do número instâncias. O grupo de empresas, ao qual se chamou de mistas, ficou com todas as empresas que não foram utilizadas no treinamento. Conforme já descrito na Seção4.2.4, o número de empresas ativas inclusas no grupo de mistas é de 309 (trezentos e nove) e o de inaptas 142 (cento e quarenta e duas), resultando em uma proporção de pouco mais de 30% de de inativas para aproximadamente 70% de ativas. Considerando que tal desbalanceamento não é, realmente, grande, especialmente pela quantidade total de amostras, optou-se por não utilizar métodos como Undersampling ou Oversampling (redução ou aumento da amostragem), por meio de algoritmos.

Caso o resultado obtido nas classificações ficassem muito aquém do esperado, poder-se-ia pensar nessas opções. A questão é que como não temos muitos casos que nos dessem uma garantia precisa da atividade ou inaptidão das empresas, não se desejou desperdiçar as amostras obtidas. Caso tivesse sido feito o balanceamento, provavelmente, seria o Oversampling.

Na busca de uma avaliação que pudesse reduzir o efeito do desbalanceamento optou-se pela F1-Score como referência para escolha do melhor modelo de classificação.

Em seguida percorre-se cada modelo e sua performance nos novos dados.

4.5.1 KNN - K-NearestNeighbors

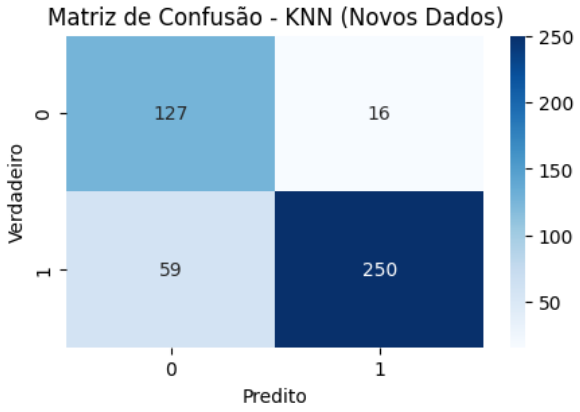
Veja abaixo a tabela 12 com as métricas obtidas para a classificação dos dados inéditos (Grupo de Empresas Mistas).

Modelo	**Acurácia**	**Precisão**	**Revogação**	**F1-Score**
KNN Simples (Novos)	0.818584	0.851397	0.818584	0.824206
KNN com Monte Carlo(Novos)	0.834071	0.858525	0.834071	0.838710

Tabela 12 – Métricas do Modelo K-Nearest Neighbors para dados inéditos

O melhor resultado para os dados inéditos do KNN, com o F1-Score de 0.8387 ou 83.87%, modelo em que se utilizou o GridSearchCV com a validação cruzada de Monte Carlo.

A matriz de confusão para esse modelo está na Figura 33. No caso, o algoritmo classificou corretamente 127 (cento e vinte e sete) empresas ativas e 250 (duzentos e cinquenta) empresas inaptas do total de 452 (quatrocentos e cinquenta e duas) dados não vistos anteriormente. Sendo que 16 (dezesseis) inaptas foram classificadas como ativas e



Fonte: Elaborado pela autora.

Figura 33 – Matriz de Confusão - K-Nearest Neighbors (Dados Novos)

59 (cinquenta e nove) ativas, erroneamente, como inaptas. Foi cerca de 89% das inaptas detectadas.

É importante observar que para o KNN os resultados no treinamento do modelo com dados normalizados não foram os melhores. Dessa forma, se utilizou o modelo para dados não normalizados nos dados inéditos e o mesmo procedimento para o modelo otimizado.

O modelo KNN treinado teve um desempenho melhor nos dados novos quando comparado ao grupo teste. Pode ser que isso aconteça pelo fato de que, apesar da pouca capacidade de aprendizado do modelo, por outro lado, ele parece permitir que a nova classificação capte ("aprenda") informações dos dados novos.

Da mesma forma que ocorreu no treinamento, uma quantidade indesejável que deve ser considerada de empresas ativas foram classificadas como inaptas.

4.5.2 Florestas Aleatórias (RF)

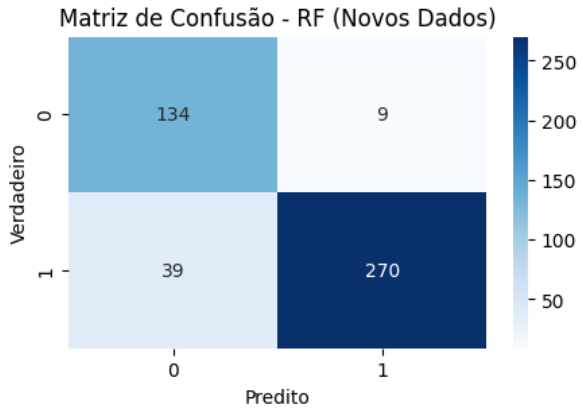
Tem-se abaixo a tabela 13 com as métricas obtidas para a classificação dos dados inéditos (Grupo de Empresas Mistas).

Modelo	**Acurácia**	**Precisão**	**Revogação**	**F1-Score**
RF Simples (Novos)	0.89381	0.96774	0.87379	0.91837
RF Simples (Novos-Normalizados)	0.89159	0.96429	0.87379	0.91681
RF com Monte Carlo(Novos)	0.89159	0.90262	0.89159	0.89379

Tabela 13 – Métricas do Modelo Florestas Aleatórias para dados inéditos

O melhor resultado para os dados inéditos da Florestas Aleatórias, com o F1-Score de 0.8938 ou 89.38%, modelo clássico do algoritmo de Florestas Aleatórias da biblioteca Scikit_learn em python. O resultado, novamente considerado ótimo, de classificação confirmou o que já tinha sido demonstrado no treinamento: o modelo mais simples de

Florestas Aleatórias é suficiente para fazer uma das melhores classificações dos dados existentes.



Fonte: Elaborado pela autora.

Figura 34 – Matriz de Confusão - Florestas Aleatórias (Dados Novos)

A matriz de confusão para esse modelo está na Figura 34. No caso, o algoritmo classificou corretamente 134 (cento e trinta e quatro) empresas ativas e 270 (duzentos e setenta) empresas inaptas do total de 452 (quatrocentos e cinquenta e duas) dados não vistos anteriormente. Sendo que 9 (nove) inaptas foram classificadas como ativas e 39 (trinta e nove) ativas erroneamente como inaptas. Resultou em um acerto de quase 97% (noventa e sete por cento) das inaptas. A quantidade de falsos negativos foi cerca de 3% (três por cento).

4.5.3 Gradient Boosting (GB)

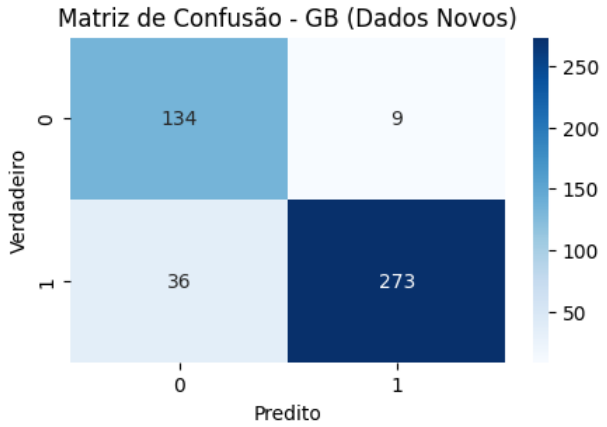
Tem-se abaixo a tabela 14 com as métricas obtidas para a classificação dos dados inéditos (Grupo de Empresas Mistas).

Modelo	Acurácia	Precisão	Revogação	F1-Score
GB Simples (Novos)	0.89823	0.90845	0.89823	0.90024
GB Simples (Novos-Normalizados)	0.90044	0.91119	0.90044	0.90246
GB com Monte Carlo(Novos)	0.88274	0.89529	0.88274	0.88525

Tabela 14 – Métricas do Modelo Gradient Boosting para dados inéditos

O melhor resultado para os dados inéditos do Gradient Boosting, com o F1-Score de 0.9025 ou 90.25%, modelo clássico do algoritmo de Gradient Boosting com dados normalizados. Uma pequena diferença aconteceu em relação aos dados do treinamento que funcionaram melhor com os dados não normalizados e, dessa vez, as métricas foram coordenadas entre si, sem alternâncias, dentro do mesmo modelo.

A matriz de confusão para esse modelo está na Figura 35. No caso, o algoritmo classificou corretamente 134 (cento e trinta e quatro) empresas ativas e 273 (duzentos e



Fonte: Elaborado pela autora.

Figura 35 – Matriz de Confusão - Gradient Boosting

setenta e três) empresas inaptas do total de 452 (quatrocentos e cinquenta e duas) dados não vistos anteriormente. Sendo que 9 (nove) inaptas foram classificadas como ativas e 36 (trinta e seis) ativas erroneamente como inaptas. Foi mais de 97% (noventa e quatro por cento) de acerto das inaptas desse grupo de empresas. Ainda para esse modelo que foi o melhor em termos de métricas, a quantidade de falsos negativos poderiam ser mitigados.

4.5.4 Regressão Logística (RL)

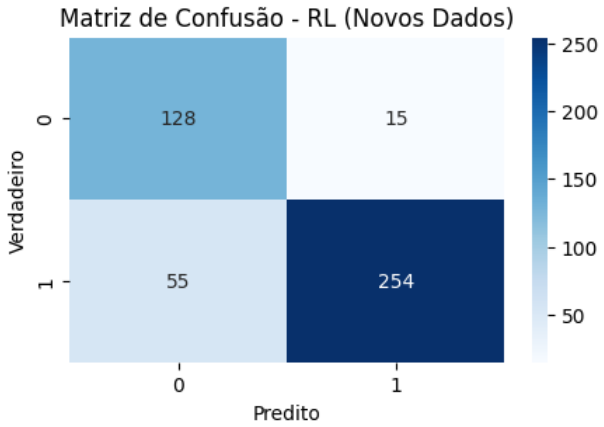
Tem-se abaixo a tabela 15 com as métricas obtidas para a classificação dos dados inéditos (Grupo de Empresas Mistas).

Modelo	Acurácia	Precisão	Revogação	F1-Score
RL Simples (Novos)	0.82743	0.85636	0.82743	0.83257
RL Simples (Novos-Normalizados)	0.84513	0.94424	0.82201	0.87889
RL com Monte Carlo(Novos)	0.85620	0.88327	0.85620	0.86043

Tabela 15 – Métricas do Modelo Regressão Logística para dados inéditos

O melhor resultado para os dados inéditos do Regressão Logística, com o F1-Score de 0.8789 ou 87.89%, modelo clássico do algoritmo de Regressão Logística com dados normalizados. Para a Regressão Logística, o melhor restultado para os dados inéditos não necessitou de refinamento, ao contrário dos dados teste do treinamento.

A matriz de confusão para esse modelo está na Figura 36. No caso, o algoritmo classificou corretamente 128 (cento e vinte e oito) empresas ativas e 254 (duzentos e cinquenta e quatro) empresas inaptas do total de 452 (quatrocentos e cinquenta e dois) dados não vistos anteriormente. Sendo que 15 (nove) inaptas foram classificadas como ativas e 55 (trinta e seis) ativas erroneamente como inaptas. Para essa classificação, cerca de 95% de inaptas foram identificadas, e um dos piores, senão o pior resultado de falsos negativos que excedeu a 20% (vinte por cento).



Fonte: Elaborado pela autora.

Figura 36 – Matriz de Confusão - Regressão Logística (Dados Novos)

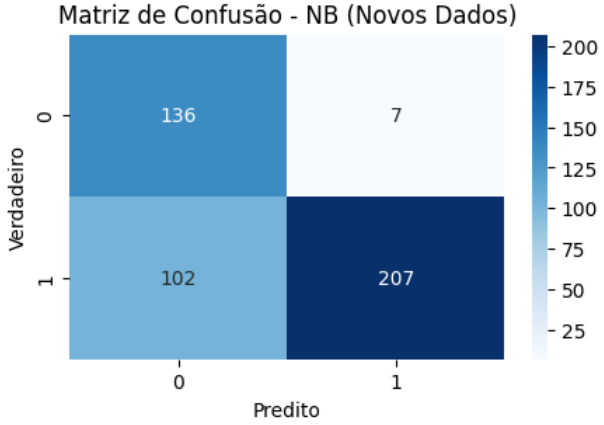
4.5.5 Naive Bayes (NB)

Tem-se a tabela 16 com as métricas obtidas para a classificação dos dados inéditos (Grupo de Empresas Mistas).

Modelo	Acurácia	Precisão	Revogação	F1-Score
NB Simples (Novos)	0.64823	0.81589	0.64823	0.65213
NB Simples (Novos-Normalizados)	0.75885	0.84205	0.75885	0.767012
NB com Monte Carlo(Novos)	0.75885	0.84205	0.75885	0.76701

Tabela 16 – Métricas do Modelo Naive Bayes para dados inéditos

O melhor resultado para os dados inéditos do Naive Bayes, com o F1-Score de 0.7670 ou 76.70%, modelo clássico do algoritmo Naive Bayes com dados normalizados. O modelo treinado com otimização de parâmetros e validação cruzada de Monte Carlo teve um resultado igual ao escolhido, porém com um tempo maior de execução.



Fonte: Elaborado pela autora.

Figura 37 – Matriz de Confusão - Naive Bayes

A matriz de confusão para esse modelo está na Figura 37. No caso, o algoritmo classificou corretamente 136 (cento e trinta e seis) empresas ativas e 207 (duzentos e sete) empresas inaptas do total de 452 (quatrocentos e cinquenta e duas) dados não vistos anteriormente. Sendo que 7 (sete) inaptas foram classificadas como ativas e 102 (cento e duas) ativas erroneamente como inaptas.

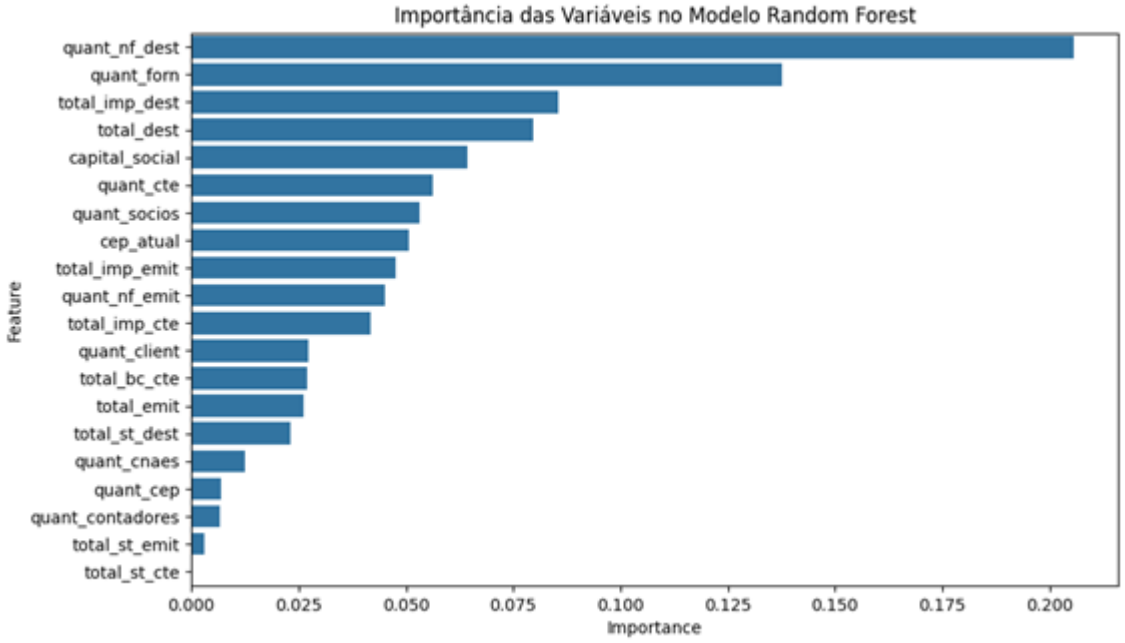
Para essa classificação a detecção das inaptas foram de quase 97% (noventa e sete por cento), o problema nessa classificação é que 102 (cento e duas) empresas classificadas como inaptas não o são. Sendo o pior resultado de falsos negativos, resultaria que a classificação equivocada se tornaria problemática. Pois, poderia desencadear processos de penalização em um número grande empresas idôneas, causando um custo razoável para o fisco reverter essa situação inadequada.

5 CONCLUSÃO

Tendo em vista a importância de combater o comportamento de indivíduos que constituem empresas de fachada, surge a grande importância de se estabelecer mecanismos de identificação e bloqueamento de tais empresas.

Com esse objetivo, a proposta desse trabalho é contribuir para a fiscalização realizada pelo Estado de São Paulo, oferecendo um Modelo de Classificação treinado por Aprendizado de Máquina, capaz de encontrar em um universo de empresas cadastradas no fisco estadual, quais delas são empresas de fato e quais são empresas fraudulentas, que simulam existência para obter de forma ilícita benefícios financeiros para seus idealizadores ou para outrem.

Para se ter um ideia, conforme levantado por (Oliveira, 2024), demonstrado na Figura 38, entre 2019 e 2024, foi computada a transferência de aproximadamente dois bilhões de reais em créditos espúrios de ICMS por empresas inidôneas.



Fonte: (Oliveira, 2024)

Figura 38 – Créditos de ICMS transferidos por empresas inidôneas

Um modelo como o proposto pode tornar possível uma detecção mais ampla dessas empresas e o desenvolvimento de estratégias, para que de forma mais rápida e eficiente se possa impedir sua continuidade.

Foram, portanto, selecionadas variáveis (features) preditoras tanto de origem cadastral, como socioeconômicas de empresas que fornecessem uma variável alvo com

a maior precisão possível da existência ou não das empresas e foram treinados diversos modelos de classificação, de algoritmos já conhecidos por sua performance adequada para os tipos de variáveis encontradas.

Para a escolha dessas variáveis foi levado em consideração o sigilo fiscal constitucional e a legislação de proteção de dados, não sendo possível a identificação das empresas pela simples observação dos códigos.

A variável alvo foi definida pela situação cadastral da empresa que denominamos de **atividade**. Se a situação era ativa, o registro para a determinada empresa foi 1(um) e para a situação inapta, o registro foi 0 (zero). As empresas ativas foram aquelas todas que, tendo sido criadas a partir do ano de 2019 (dois mil e dezenove), tivessem sido fiscalizadas e interagido com o fisco, de forma que ao final de sua fiscalização continuaram com essa situação. Já as empresas inaptas foram as que tendo como parâmetro o regramento aposto na Portaria CAT 95/2006 (São Paulo, 2006), editada pela Secretaria da Fazenda e Planejamento do Estado de São Paulo, foram declaradas, também após fiscalização, como inaptas desde sua criação, que também ocorreu depois de janeiro de 2019.

Para o desenvolvimento desse projeto utilizou-se a Metodologia CRISP-DS. Os algoritmos foram testados inicialmente em uma linha de base (baseline) e depois em novos ciclos de aperfeiçoamento foram percorridos, buscando aprimoramento e mais segurança para uma escolha mais acertada.

Utilizamos os algoritmos: K-Nearest Neighbors, Florestas Aleatórias, Gradient Booster, Regressão Logística e Naive Bayes. Todos foram testados quanto à sensibilidade ou não à normalização dos dados, foram avaliados por refinamento com GridSearch, usado com a validação cruzada de Monte Carlo.

Após se verificar o resultados de cada modelo verificou-se que o Modelo de Florestas Aleatórias puro e simples foi aquele que apresentou as melhores métricas, tanto durante o treinamento, como nos dados inéditos. Na Tabela 17, trouxemos as melhores métricas para cada algoritmo:

Modelo	KNN	RF	GB	RL	NB
Treinamento (Acurácia)	0.8314	0.9186	0.9070	0.8721	0.8895
Inéditos (F1-Score)	0.8387	0.9184	0.9025	0.8789	0.7670

Tabela 17 – Melhores Métricas Obtidas nas Classificações

Observa-se, e pode ser visto com mais detalhes na Seção 4.5.2, o modelo de Florestas Aleatórias com a Normalização dos Dados foi o que demonstrou os melhores resultados de classificação e robustez quando diante de novos dados.

Esse resultado fornece um modelo que pode agregar muito ao trabalho de fiscalização. É um modelo que tem um custo computacional, relativamente, baixo, apresentou uma precisão boa, e um aumento no volume de instâncias pode não prejudicar muito sua performance. Outras vantagens das Florestas Aleatórias é são pouco propensas a overfitting e podem ser associadas a outros algoritmos, permitindo que não se descarte essa possibilidade no futuro.

No grupo de empresas mistas com 452 (quatrocentos e cinquenta e duas) instâncias, 143 (cento e quarenta e três) empresas eram inaptas e o modelo conseguiu identificar 134 (cento e trinta e quatro). Isso significa quase de 94% (noventa e sete por cento), um valor ótimo.

No entanto, quando se olha a Matriz de Confusão (Figura 34) e se encontra as 39 (trinta e nove) empresas que foram classificadas como inaptas, mas são ativas, deve-se ter a consciência de que um tratamento humano deverá ser feito com o resultado final. Será importante verificar se há outras informações disponíveis que permitam identificar uma empresa ativa ao ser selecionada para a fiscalização. Dessa forma, evitar-se-á que uma empresa receba uma penalidade indevidamente. Apesar de não ser uma porcentagem grande de falsos positivos, não deve ser ignorada.

Outra observação que se faz importante é feita a partir da Figura 26. A importância de cada variável para esse modelo. Como se percebeu uma importância bem concentrada nas variáveis socio econômicas mais a variável CEP atual. Pode ser que uma inclusão de mais variáveis desse tipo possam ajudar a obter um sucesso ainda maior.

De qualquer maneira o resultado é muito interessante e é um grande impulso para se aprofundar ainda mais no modelo.

Além disso, é importante notar que nenhum dos modelos se mostrou inadequado. Exceto, pelo Naive Bayes que demonstrou necessitar de mais esforço para sua melhoria, os demais tiveram bons resultados.

A escolha inicial das variáveis se mostrou muito coerente. A maioria delas teve uma boa influência para o desempenho de algum dos modelos. Reforçando que as socioeconômicas foram as que mais se destacaram. A depender do modelo vale a pena se excluir algumas e tentar encontrar outras que correspondam à mesma linha e que possam enriquecer o aprendizado.

Para complementar, fica a sugestão de novos trabalhos como por exemplo, aprofundar o modelo de Florestas Aleatórias associando com éstratégias já usadas outros trabalhos, como por exemplo histórico de conformidade, relações com outras empresas, padrões de documentos fiscais etc. Poder-se ia agregar variáveis alegóricas e tornar o modelo mais complexo.

Há ainda outras variáveis que se poderiam agregar, como utilizar o CNAE principal

das empresas, a principal mercadoria comercializada, por meio da Nomenclatura Comum do Mercosul (NCM) ou mesmo número de documentos de envolvidos, sócios e contadores, que poderiam oferecer um padrão de comportamento. Neste último caso, há que se avaliar o cuidado com o sigilo inerente.

Além disso, outros algoritmos poderiam ser testados, ou mesmo os utilizados ter o retrabalho de separar as *features* pouco importantes e agregar alguma outra ou outras mais.

Referente ao presente trabalho, se conclui aqui a monografia, mas, não é um trabalho que será encerrado. A partir do resultado aqui encontrado, pode ser utilizado o modelo e à medida de sua aplicação pode ser feito um aperfeiçoamento, à medida que se ampliar a abrangência da fiscalização novas empresas serão rotuladas de forma mais precisa e poderão ser agregadas ao modelo, tornando-o cada vez mais eficaz.

REFERÊNCIAS

BENTIVEGNA, W. Identificando Emitentes de Nota Fiscal Inidônea por Meio de Modelos Baseados Em Algoritmos de Classificação. **2º Workshop de Defesas de Trabalhos de Conclusão de Curso (MBA em Ciência de Dados)**, 2022.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. 1°. ed. [*S.l.: s.n.*]: Springer, 2006.

BORINI, F. C. G. F. M. A corrupção no ambiente de negócios: survey com as micro e pequenas empresas da cidade de são paulo. **Revista de Administração - RAUSP**, 2009.

BRASIL. **Constituição da República Federativa do Brasil**. 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/ConstituicaoCompilado.htm. Acesso em: 21 jul. 2024.

BRASIL. **Lei Complementar n°123 de 14 de dezembro de 2006**. 1996. Disponível em: https://www.planalto.gov.br/ccivil_03/Leis/LCP/Lcp87.htm. Acesso em: 27 set.2024.

BRASIL. **Lei Geral de Proteção de Dados Pessoais (LGPD). Lei 13.709 de 2018.** 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

CAMPANELLE, D. M. **Sistema Tributário Brasileiro: uma análise crítica**. 2012. Disponível em: https://pantheon.ufrj.br/bitstream/11422/2062/1/DMCampanelle.pdf. Acesso em: 30 out. 2024.

DUCAS, J. D. Utilização de análise de sobrevivência para inidôneos: uma análise dos tempos de inaptidão e cassação dos casos de 2015 a 2023. **4º Workshop de Defesas de Trabalhos de Conclusão de Curso (MBA em Ciência de Dados)**, 2024.

FERNANDEZ, M. L. A. Proposta de modelo de predição de operações comerciais interestaduais simuladas: análise de operações com a Zona Franca de Manaus / Área de Livre Comércio. **4º Workshop de Defesas de Trabalhos de Conclusão de Curso (MBA em Ciência de Dados)**, 2024.

GOMES, G. S. L. **Identificação de Práticas de Evasão Fiscal Utilizando Aprendizado de Máquina: O Caso das Empresas de Fachada e os Créditos Ilegais de ICMS**. 2023. Dissertação (Mestrado) — Universidade Católica de Brasília - UCB, 2023. Disponível em: https://bdtd.ucb.br:8443/jspui/handle/tede/3311.

HENRIQUES, P. M. Detecção de fraudes no ICMS na Grande São Paulo usando aprendizado de máquina supervisionado. **4º Workshop de Defesas de Trabalhos de Conclusão de Curso (MBA em Ciência de Dados)**, 2024.

JúNIOR, C. P. **História Econômica do Brasil**. 2012. Disponível em: https://drive.google.com/file/d/0B8_gvWjrwU3ZQjFITVBwYWZUbjg/view?resourcekey=0-6jJjbtIZ1SrUO9hcA-lwkA. Acesso em: 27 set.2024.

LEARN, s. scikit. 2007–2024. Disponível em: https://scikit-learn.org/stable/index.html. Acesso em: 22 jul. 2024.

LINDHOLM, e. a. A. **Machine learning: A first course for engineers and scientists.** Cambridge University Press (Virtual Publishing)., 2022. Disponível em: https://smlbook.org/book/sml-book-draft-latest.pdf. Acesso em: 20 jul. 2024.

Marcos Vinícius dos Santos, Hugo Leonardo Menezes de Carvalho, Lucimar Antonio de Cabral de Ávila. Complexidade Tributária e Evasão Fiscal no Brasil. **XVI Congresso USP de Iniciação Científica em Contabilidade**, 2019. Disponível em: https://congressousp.fipecafi.org/anais/19UspInternational/ArtigosDownload/1682.pdf.

MENDONçA, M. E. B. T. C. N. A. T. B. B. D. E. Misbehavior empreendedor associado às práticas de sonegação fiscal. **Cadernos EBAPE BR**, 2021.

NEMBE J. K., e. a. The role of artificcial intelligence in enhancing tax compliance and financial regulation. **Finance & Accounting Research Journal**, 2024.

OLIVEIRA, M. S. de. Identificação de Empresas Inidôneas no Cadastro de Contribuintes de ICMS: Uma Abordagem Baseada em Algoritmos de Classificação. **4º Workshop de Defesas de Trabalhos de Conclusão de Curso (MBA em Ciência de Dados)**, 2024.

RFB. **Instrução Normativa RFB nº 1863**. 2018. Disponível em: http://normas.receita.fazenda.gov.br/sijut2consulta/link.action?idAto=97729.

RFB. **Receita Federal desenvolve tecnologia inovadora capaz de ampliar a detecção de fraudes e ilegalidades tributárias e aduaneiras**. 2024. Disponível em: https://www.gov.br/receitafederal/pt-br/assuntos/noticias/2024/setembro. Acesso em: 21 dez. 2024.

SOARES, G. F. Ciência de dados aplicada à auditoria interna. **Revista da CGU**, 2020.

São Paulo. **Portaria CAT 95/2006**. 2006. Disponível em: https://legislacao.fazenda.sp.gov.br/Paginas/pcat952006.aspx. Acesso em: 31 out. 2024.

São Paulo. **Resolução SF Nº 20 de 2012**. 2012. Disponível em: https://legislacao.fazenda.sp.gov.br/Paginas/resf202012.aspx. Acesso em: 22 jul. 2024.

TEIXEIRA, J. de A. Complexidade do Atual Código Tributário Nacional e as Principais Mudanças trazidas pela PEC 45/2019. **Revista Governança Tributária**, 2024. Disponível em: https://educacao.ibpt.com.br/revista.

VITI, R. Algoritmos de agrupamento e classificação para a identificação de empresas emissoras de notas fiscais inidôneas. **4º Workshop de Defesas de Trabalhos de Conclusão de Curso (MBA em Ciência de Dados)**, 2024.

XAVIER, e. a. O. C. Identificação de evasão fiscal utilizando dados abertos e inteligência artificial. **Revista de Administração Pública - FGV EBAPE**, 2022.