

Artigo

Identificando evasão fiscal em empresas de fachada e em créditos ilegais de ICMS

Gunther Siqueira Lemos Gomes ¹Remis Balaniuk ²¹ Secretaria de Estado de Economia do Distrito Federal, Brasília / DF – Brasil² Universidade Católica de Brasília (UCB), Brasília / DF – Brasil

Empresas que emitem documentos fiscais para fraudar o fisco com a transferência de crédito do ICMS sem a circulação de mercadorias causam prejuízo ao erário público e, por conseguinte, à sociedade. Diversas iniciativas de combate a fraudes fiscais têm utilizado, com sucesso, técnicas de análise de dados e aprendizagem de máquina. Este trabalho buscou investigar o uso dessas técnicas na identificação de uma prática específica de fraude fiscal realizada por empresas popularmente conhecidas como “empresas noteiras”, que formadas exclusivamente para emitir créditos não devidos de ICMS, imposto sobre operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual, intermunicipal e de comunicação. Com base na análise documental e em consulta com auditores e especialistas, foram identificadas tipologias e variáveis relevantes na determinação de eventos de sonegação fiscal realizados pelas empresas noteiras. Em torno dessas variáveis, procedeu-se à coleta e à preparação de dados provenientes da Secretaria de Fazenda do Distrito Federal. Com esses dados, foi possível explorar o uso de modelos preditivos baseados em aprendizagem de máquina capazes de apontar comportamentos potencialmente fraudulentos. Os bons resultados obtidos por esses modelos demonstram seu potencial como parte de uma sistemática de monitoramento e auditorias fiscais realizadas pelos órgãos fazendários.

Palavras-chave: ICMS; noteiras; aprendizagem de máquina; empresas de fachada; sonegação fiscal.

Identificación de evasión tributaria en empresas fachada y créditos ilegales del ICMS

Las empresas que emiten documentos tributarios para defraudar al fisco con la transferencia de crédito del ICMS (impuesto a las operaciones relacionadas con la circulación de bienes y de prestación de servicios interestatales, interurbanos y de comunicaciones) sin movimiento de mercancías causan daños al erario público y, por ende, a la sociedad en su conjunto. Varias iniciativas para combatir el fraude fiscal han utilizado con éxito técnicas de análisis de datos y aprendizaje automático. Este trabajo buscó investigar el uso de estas técnicas en la identificación de una práctica específica de fraude fiscal, practicada por empresas conocidas popularmente como ‘empresas factureras’, constituidas exclusivamente para emitir créditos no vencidos del ICMS. A partir del análisis documental y la consulta a auditores y especialistas, se identificaron tipologías y variables relevantes para la identificación de eventos de evasión fiscal realizados por empresas factureras. En torno a estas variables se recolectaron y prepararon datos desde la Secretaría de Hacienda del Distrito Federal. Con estos datos fue posible explorar el uso de modelos

DOI: <https://doi.org/10.1590/0034-761220230303>ISSN: 1982-3134 

Artigo submetido em 20 de setembro de 2023 e aceito para publicação em 07 de junho de 2024.

Editora-chefe:Alketa Peci (Fundação Getúlio Vargas, Rio de Janeiro / RJ – Brasil) **Editor adjunto:**Sandro Cabral (Insper Instituto de Ensino e Pesquisa, São Paulo / SP – Brasil) **Pareceristas:**João Paulo de Oliveira Louzano (Universidade Federal dos Vales do Jequitinhonha e Mucuri, Teófilo Otoni / MG – Brasil) Ana Carolina Giuberti (Universidade Federal do Espírito Santo, Vitória / ES – Brasil) **Relatório de revisão por pares:** o relatório de revisão por pares está disponível neste [link](#).

predictivos basados en *machine learning* capaces de señalar comportamientos potencialmente fraudulentos. Los buenos resultados obtenidos por estos modelos demuestran su potencial como parte de un seguimiento sistemático y auditorías fiscales por parte de las autoridades tributarias.

Palabras clave: ICMS; factureras; aprendizaje automático; empresas de fachada; evasión de impuestos.

Identifying tax evasion in shell companies and fraudulent credits of Brazil's state value-added tax (ICMS)

Companies that issue tax documents to defraud the tax authorities with the transfer of credits of Brazil's state value-added tax (ICMS) without the movement of goods cause financial losses to the government and, therefore, to society as a whole. Several initiatives to combat tax fraud have successfully used data analysis and Machine Learning techniques. This work sought to investigate the use of these techniques in identifying a specific practice of tax fraud, practiced by shell companies, formed exclusively to issue non-due ICMS credits, the tax on operations related to the circulation of goods, and the provision of interstate, intercity, and communication services. Based on document analysis and consultation with auditors and specialists, typologies and variables relevant to identifying tax evasion events carried out by shell companies were identified. Around these variables, data from the Finance Department of the Federal District were collected and prepared. With this data, it was possible to explore the use of predictive models based on Machine Learning capable of pointing out potentially fraudulent behavior. The good results obtained by these models demonstrate their potential as part of systematic monitoring and fiscal audits by tax authorities.

Keywords: ICMS; machine learning; shell companies; tax evasion.

1. INTRODUÇÃO

No cenário econômico e político brasileiro, as demandas da sociedade por melhores serviços e infraestrutura, aliadas à percepção de uma carga tributária já por demais elevada, permanecem como temas centrais de discussão e pressão. Em um contexto de crescente conscientização cívica e participação social, os cidadãos brasileiros têm cobrado, de forma cada vez mais enfática, soluções para problemas crônicos, como a falta de investimentos em saúde, educação, transporte e segurança, enquanto também clamam por alívio nos encargos fiscais que impactam suas vidas e o ambiente de negócios. A expectativa por políticas públicas eficientes e transparentes que promovam o desenvolvimento sustentável e a inclusão social, ao mesmo tempo que busquem aliviar o peso dos tributos sobre a população e as empresas, tornou-se evidente, exigindo respostas ágeis e inovadoras por parte das autoridades governamentais.

Na contramão dessas demandas e expectativas, num cenário de elevada carga tributária, a administração pública vê perdas de receitas por causa de fraudes fiscais ganharem espaço. Observa-se que quanto maior o ganho em não pagar imposto, maior também é o ânimo do contribuinte em sonegar.

No centro da questão fiscal e tributária estão os impostos indiretos, que consistem em taxas sobre o consumo inseridas nos preços de toda e qualquer mercadoria e serviço. Em 2015, esses impostos representavam 49,4% da carga tributária bruta da União, um patamar distante do praticado nos países desenvolvidos (Souza, 2018). Exemplos de impostos indiretos são o Imposto sobre Produtos Industrializados (IPI) e o Imposto sobre Circulação de Mercadorias e Serviços (ICMS). Cerca de 82% da arrecadação de impostos dos estados e do Distrito Federal, em 2020, foram baseados no ICMS, tendo sido este, juntamente com o Imposto de Renda, os impostos mais sonegados naquele ano (Santos et al., 2022). O ICMS é também de particular relevância para os pequenos municípios que dependem da transferência de recursos por meio da repartição constitucional de tributos (Azevedo et al., 2015).

Por causa de sua amplitude e materialidade, esquemas fraudulentos na coleta do ICMS costumam acarretar perdas bilionárias aos cofres públicos. Em apenas uma operação realizada pela fiscalização

tributária do estado do Paraná, em 2022, foram identificadas 844 empresas de fachada criadas exclusivamente para sonegar impostos. Essas empresas emitiram notas fiscais de operações que somaram R\$ 4,8 bilhões, acarretando a sonegação de R\$ 542,8 milhões de ICMS (Ortiz, 2022).

Esta pesquisa buscou identificar um tipo específico de fraude associada ao ICMS, caracterizada pelo uso de operações simuladas de venda de mercadorias praticadas pelas chamadas “empresas noteiras”: empresas criadas com o objetivo de operar fraude fiscal, principalmente por meio da emissão de notas fiscais falsas (Carvalho, 2018). Trata-se da emissão de documento fiscal sem que o vendedor – emitente da nota – transacione a mercadoria que está comercializando, o que indica que não há a circulação física nem jurídica (troca de propriedade) dela, e sem o pagamento do imposto devido, ou seja, trata-se apenas de operação fictícia. São as chamadas “notas frias”. Para Ferreira (2019), as empresas noteiras também são conhecidas como “empresas de fachada”, já que são criadas apenas para fornecer um aspecto legal a uma transação fraudulenta.

O mecanismo de comércio de créditos contidos em notas frias se explica, pois o ICMS tem a característica da não cumulatividade, situação em que um comerciante, ao adquirir uma mercadoria para revenda, possui o direito de aproveitar o ICMS pago pelo seu fornecedor como crédito, pagando ao estado em que está cadastrado apenas a diferença ou o valor adicionado do imposto na operação seguinte de venda da mesma mercadoria. Nesse esquema, a empresa de fachada emite a nota fria, sem, obviamente, pagar o ICMS devido no seu estado, e vende a nota por uma fração do valor do imposto devido a uma empresa destinatária – tipicamente de outra unidade da federação – que deseja legalizar as próprias vendas sem o pagamento do imposto no seu estado.

A principal dificuldade em identificar tal fraude é que as informações fornecidas nos documentos fiscais emitidos são, aparentemente, de operações normais e, para dificultar ainda mais a ação do fisco, os sonegadores, cientes da dificuldade de comunicação entre os entes federados, simulam vendas interestaduais, em que o emitente e o destinatário da nota são de estados distintos.

Como a identificação dessas empresas e de sua prática fraudulenta requer tempo e esforço, as noteiras acabam se beneficiando desse atraso, gerando valores expressivos de créditos de ICMS, aproveitados pelos seus receptores, até serem descobertas.

Allingham e Sandmo (1972) ressaltam que as empresas trabalham sob limite de risco ao sonegar impostos, e que aplicar multas é uma das principais formas de aumentar a sensação de risco pelo contribuinte e evitar a sonegação.

O experimento realizado por Kleven et al. (2011), na Dinamarca, que corrobora os estudos apresentados por Allingham e Sandmo, revelou que o contribuinte que sofre uma auditoria tributária tende a responder melhor à arrecadação espontânea. Entretanto, Lima (2007) destaca que os órgãos fiscalizadores não possuem capacidade de fiscalizar todos os contribuintes. Nesse contexto, verifica-se que, para imprimir uma sensação de risco mais efetiva num contexto de elevado número de contribuintes e, por conseguinte, de operações a serem fiscalizadas, se faz necessário utilizar métodos automatizados para a detecção de fraudes em larga escala, de modo a tornar tal processo mais abrangente, ágil e com menor custo.

Esses métodos automatizados precisam, obviamente, utilizar plataformas computacionais robustas, grandes bases de dados e modelos analíticos capazes de processar as enormes quantidades de transações diárias e identificar comportamentos atípicos com indicativos de fraude fiscal. Trata-se, portanto, de uma aplicação típica para a ciência de dados, uma área multidisciplinar que utiliza técnicas computacionais, estatísticas e matemáticas para resolver problemas complexos e que vem

sendo utilizada nos mais diversos domínios para extrair informações significativas para os negócios e as tomadas de decisão. Na fiscalização tributária, o uso de técnicas derivadas da ciência de dados vem evoluindo expressivamente nas últimas décadas e vem produzindo resultados em diversas frentes e países (Abrantes & Ferraz, 2016).

Nesse contexto, o objetivo do presente trabalho foi explorar o potencial dos métodos da ciência de dados e da aprendizagem de máquina como base para uma sistemática de seleção de indícios de simulação na emissão de documentos fiscais, buscando burlar a tributação pelo ICMS por empresas noteiras.

Como diferencial com relação aos trabalhos existentes na literatura sobre o tema da fraude fiscal, esse estudo busca contribuir com uma estratégia de análise adequada à realidade tributária brasileira, às particularidades do ICMS e às práticas fraudulentas praticadas nas unidades da federação na atualidade. Além disso, a pesquisa se baseia no conhecimento de negócio dos próprios auditores fiscais e explora práticas já adotadas pelas secretarias de fazenda, como a denegação automática de notas fiscais como insumo para a preparação dos modelos preditivos.

A metodologia utilizada foi o CRISP-DM (Chapman, 2000), referência em projetos de mineração de dados. Toda a compreensão do negócio, definição de premissas e escolha dos dados utilizados na análise foi realizada com a participação de auditores fiscais. Para a concepção dos modelos preditivos foram utilizadas somente bases de dados tipicamente disponíveis nas secretarias de fazenda estaduais, a saber: documentos fiscais eletrônicos, especificamente a Nota Fiscal Eletrônica (NFe), dados cadastrais e societários das empresas e eventos da NFe ligados à denegação automática de notas. De maneira acessória, foram usados dados de Conhecimento de Transporte Eletrônico (CTe), Nota Fiscal do Consumidor Eletrônica (NFCe), Escrituração Fiscal Digital (EFD) e Declaração de Meios de Pagamento (Dimp). Os dados utilizados compreendem o período entre 2014 e 2023 obtidos na Secretaria de Fazenda do Distrito Federal.

Este artigo é composto por cinco seções. Na primeira, trouxemos essa introdução; na segunda, apresentaremos uma revisão da literatura sobre o tema, além de alguns conceitos necessários para a compreensão do contexto no qual a pesquisa se desenrolou; na terceira, detalharemos a metodologia da pesquisa; na quarta seção, apresentaremos os resultados das etapas da pesquisa; finalmente, na quinta e última seção, traremos algumas considerações finais.

2. REVISÃO DA LITERATURA

Castellón-González e Velásquez (2013 como citado por De Roux et al., 2018) descrevem como a escolha de empresas para fiscalização tributária se baseia, tradicionalmente, em regras e no uso da intuição pelo agente fiscal. Esse cenário é similar ao descrito por Zumaya et al. (2021), que relatam que as técnicas tradicionais de seleção manual de contribuintes para a fiscalização, sem o apoio de métodos computacionais, dão eficácia restrita, uma vez que várias empresas não aparecem nos filtros convencionais e não são fiscalizadas.

Com o avanço da digitalização de documentos fiscais, de informações cadastrais de contribuintes, quadros societários de empresas e do registro de transações comerciais e financeiras, enormes bases de dados oficiais passaram a ser alimentadas cotidianamente e estão à disposição dos órgãos fiscalizadores. Avanços dramáticos no cenário tecnológico abriram novas possibilidades de uso desses dados e trouxeram a ciência de dados para dentro da realidade da fiscalização tributária. Nas últimas

décadas, diversas iniciativas, que propuseram a aplicação de uma diversidade significativa de técnicas e produziram resultados marcantes, demonstraram que a automação da fiscalização tributária é um caminho sem volta (Abrantes & Ferraz, 2016).

A literatura sobre o tema é farta, mas alguns trabalhos com enfoque aplicado e baseados em casos reais, como é o caso do estudo aqui apresentado, merecem destaque.

O Tax Administration Diagnostic Assessment Tool (Tadat, 2019), patrocinado pela União Europeia e o Fundo Monetário Internacional, entre outras instituições, aponta o uso da informação em larga escala com cruzamento massivo de informações como uma das boas práticas no combate à sonegação, melhorando a capacidade do fisco de identificar desvios e aumentar a eficiência da fiscalização. Um dos exemplos citados é o uso das denominadas malhas fiscais, em que o fisco aponta indícios de irregularidades fiscais para que o contribuinte os sane de modo espontâneo. Portanto, apesar de a prática não ser voltada à auditoria, demonstra a importância da ciência de dados na fiscalização tributária.

Matos (2019) levanta a importância de se levarem em consideração outras informações durante a escolha de alvos de auditoria, como as características cadastrais dos contribuintes (features), mas destaca a difícil interpretação desses dados sem a ajuda de inteligência de máquina, pois elas possuem baixas correlações lineares, e as fraudes utilizam estratégias complexas que, muitas vezes, não são identificáveis diretamente nos dados disponíveis. O autor sugere a técnica de grafos como forma de interligar tais características e apresenta um método de seleção de features baseado em regras de associação e lógica proposicional.

Já Bittencourt (2018) trabalhou com dados fiscais da base da Secretaria de Fazenda do Distrito Federal, a mesma base primária desta dissertação. Seu objetivo foi o estudo de outliers para identificar períodos fiscais mensais em que os contribuintes possuíam um comportamento de sonegação. O autor sugere o uso do método Redes Neurais com fins preditivos para determinar condutas fundamentadas em dados fiscais cadastrais, o que é proposto neste trabalho para a fraude específica que envolve empresas noteiras.

Ippolito e Lozano (2020), por sua vez, demonstram a aplicabilidade de uma diversidade de técnicas, como Random Forests (Florestas Aleatórias), Naive Bayes, Árvore de Decisão, Regressão Logística, Ensemble Learning e Redes Neurais, na identificação de crimes contra o fisco na cidade de São Paulo.

Andrade et al. (2021) utilizaram dados da Secretaria de Fazenda do Estado do Espírito Santo para a detecção de fraude e também diferentes técnicas para o mesmo fim. Para os testes, foram utilizados os algoritmos K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM) e Neural Network (NN).

Castellón-González e Velásquez (2013) realizaram um estudo para a Administração Tributária do Chile, no qual dividem os contribuintes em clusters (agrupamentos) pelo porte, utilizando algoritmos como Self Organizing Maps (SOM), Neural Gas (NG) e Árvores de Decisão e, depois, trabalham com um algoritmo supervisionado para a solução do problema de classificação de fraudes.

Xavier et al. (2022) aplicaram Random Forests, Redes Neurais e Grafos para identificar o perfil de potenciais sonegadores, utilizando apenas dados abertos e públicos disponibilizados pela Receita Federal do Brasil, o Conselho Administrativo de Recursos Fiscais do Estado de Goiás e outras fontes públicas.

Por fim, Zumaya et al. (2021) atacaram a fraude fiscal no México e propuseram a identificação de fraudes utilizando Redes Neurais e Random Forest. Especificamente, em relação à Rede Neural,

Zumaya et al. (2021) utilizaram Dynamic Recurrent Neuronal Network (DRNN) para considerar os documentos fiscais emitidos de maneira distinta, por meio de séries temporais.

Conclui-se que há uma diversidade expressiva de técnicas de ciência de dados e aprendizagem de máquina aplicadas na fiscalização tributária, especialmente na detecção de fraudes. Diversas dessas técnicas foram testadas e algumas, efetivamente, aproveitadas para estruturar o método escolhido por este estudo. Entretanto, o trabalho aqui proposto se distingue dos demais já publicados por atacar especificamente o esquema das empresas noteiras e as notas frias na fraude contra o ICMS. Além disso, nossa técnica utiliza conhecimento de negócio dos auditores fiscais brasileiros e dados de denegação automática de notas fiscais na preparação dos modelos preditivos.

3. CONCEITOS ESTRUTURANTES SOBRE OS DADOS UTILIZADOS NO ESTUDO

O Brasil adotou o Sistema Público de Escrituração Digital em 2007. Esse sistema é composto de diversos subprojetos, com destaque para o cadastro fiscal, a escrituração fiscal e contábil e os documentos fiscais eletrônico, como a Nota Fiscal Eletrônica (NFe), a Nota Fiscal ao Consumidor Eletrônica (NFCe) e a Nota Fiscal de Serviço Eletrônica (NFSe).

Para o presente estudo, é de grande relevância o Cadastro Fiscal e a Nota Fiscal Eletrônica. O Cadastro Fiscal contém informações cadastrais prestadas pelas pessoas jurídicas aos fiscos estaduais, como endereço, responsável pela contabilidade, capital social, classificação em atividades econômicas de maneira padronizada (CNAE-Fiscal) e composição societária.

Cada documento fiscal possui um fim específico, e a Nota Fiscal Eletrônica (NFe) é o documento hábil para registrar uma operação interestadual de venda de mercadoria ou uma operação entre contribuintes do imposto. Assim, quando a operação envolve crédito do ICMS, o documento para registrar essa transação é uma NFe, sendo vedada a emissão de outros documentos. Produtores rurais constituem uma exceção a essa regra, pois podem emitir a Nota Fiscal do Produtor Rural, que ainda é enviada por meio de formulários e não possui versão eletrônica.

O Imposto sobre Circulação de Mercadorias (ICMS) é estadual e possui uma característica constitucional de ser não cumulativo, ou seja, compensa o valor já pago. As noteiras se apoiam nesse aspecto do ICMS para oferecer vantagem ao beneficiário do esquema. O adquirente de uma nota fiscal fria, emitida por uma noteira, faz uso desse imposto a compensar e, com isso, reduz o imposto devido. A noteira, por ser uma empresa de fachada, não recolhe o tributo a nenhum fisco estadual, diminuindo a tributação efetiva dos produtos.

Um evento relevante para esse estudo, que serviu como ponto de partida para a anotação de dados e a criação de bases para treinamento e teste dos modelos preditivos, é a denegação de NFes, praticada por boa parte das secretarias de fazenda estaduais. A NFe denegada ocorre quando uma nota fiscal eletrônica é emitida, mas a Secretaria de Fazenda (Sefaz) identifica algum problema por parte do emitente ou do destinatário. Dizer que a nota fiscal eletrônica foi denegada significa que a Sefaz identificou alguma suspeita de irregularidade do emissor ou do destinatário da NFe, e ela não pode ser faturada/autorizada enquanto não for auditada, e essa irregularidade estar associada à prática das empresas noteiras ou simplesmente a um problema cadastral. Por meio da denegação automática, as secretarias de fazenda conseguem estancar a ação de empresas suspeitas antes mesmo de confirmadas as ações fraudulentas. As empresas cujas notas fiscais foram denegadas são auditadas e podem recorrer da denegação. No final, o auditor fiscal registra o motivo padronizado do bloqueio

da NFe e se as suspeitas de irregularidades foram confirmadas. Dessa forma, o registro das denegações fornece uma base rica de exemplos de empresas auditadas, das quais algumas foram identificadas como empresas fantasmas ou noteiras e outras foram confirmadas como empresas legítimas e não fraudulentas. As bases rotuladas dessa maneira facilitam muito a implementação de modelos analíticos preditivos fundamentados em aprendizagem supervisionada.

4. MÉTODO

Para atacar o problema específico da evasão fiscal operada pelas empresas noteiras, recorreu-se a uma metodologia aplicada e quantitativa, baseada em dados fiscais típicos da prática fiscal das unidades da federação. Assim, optou-se por uma pesquisa cuja finalidade é exploratória, considerando que o fenômeno, embora conhecido pelos auditores fiscais, carece de uma sistematização e automação na hora de identificar o contribuinte fraudador.

A pesquisa baseou-se em dados fiscais documentais em poder da Secretaria de Fazenda do Distrito Federal, mas pode, a princípio, ser aplicada em outras unidades da federação, já que dados similares são utilizados por todas as secretarias de fazenda estaduais. Neste trabalho, utilizamos técnicas e métodos das ciências estatísticas e de dados.

As tarefas realizadas durante o estudo foram:

- Mapeamento das fragilidades da sistemática de tributação e de coleta do ICMS exploradas pelas empresas noteiras nos principais esquemas de evasão fiscal;
- Identificação das variáveis relevantes para a evidenciação de fraude fiscal fundamentada nos dados digitais de documentos fiscais eletrônicos contidos na base de dados da Secretaria de Fazenda do DF;
- Escolha das variáveis estatisticamente relevantes a serem utilizadas nos modelos preditivos;
- Implementação de processo automatizado de carga e transformação dos atributos brutos provenientes dos bancos de dados para derivação das variáveis analíticas selecionadas;
- Experimentação com modelos preditivos baseados em aprendizagem de máquina, buscando aqueles capazes de, em larga escala e de forma rápida, identificar indícios de atuação de empresas noteiras na emissão fraudulenta de documentos fiscais com o intuito de burlar a tributação pelo ICMS;
- Avaliação da adequabilidade e potencial de aplicação dos métodos propostos na atuação cotidiana da Secretaria de Fazenda do DF.

Para guiar o processo de análise de dados e a execução das tarefas, foi utilizado o Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). O CRISP-DM é uma metodologia dividida em etapas e baseada em ciclos.

FIGURA 1 VISÃO GERAL DO CRISP-DM



Fonte: Chapman et al. (2000).

4.1. Detalhando as Fases do Processo CRISP-M

4.1.1 Entendimento do negócio

A primeira etapa do “entendimento do negócio” consiste em conhecer o domínio do negócio, definir o problema a ser tratado e os objetivos segundo a perspectiva do negócio. Essa etapa demanda várias interações com os detentores do conhecimento da atividade ou do processo em estudo. No caso presente, em que se busca apoiar o processo de auditoria das secretarias de fazenda, os especialistas são, obviamente, os auditores fiscais. Nessa fase, são também identificadas as fontes de dados potencialmente úteis à análise do problema. Para isso, faz-se necessária a participação de especialistas da área de tecnologia desses órgãos, conhecedores das bases de dados existentes.

4.1.2 Entendimento dos dados

A fase de “entendimento dos dados” consiste na fase mais longa do projeto, em que se busca identificar os problemas de qualidade e seus subconjuntos interessantes à análise, como também formar hipóteses com base na sua visualização.

Chapman et al. (2000) descrevem a fase de entendimento de dados como a fase de seleção de tabelas, registros, atributos e transformação e limpeza daqueles considerados não adequados à análise.

Dados de diferentes fontes são integrados, sumarizados e preparados para as fases subsequentes de análise.

4.1.3. Preparação dos dados

Uma vez identificados os atributos presentes nas bases de dados relacionadas com o contexto da pesquisa e que sejam relevantes para a identificação dos esquemas típicos de fraude fiscal operados pelas empresas noteiras, são necessários a definição de variáveis a serem utilizadas pelos modelos preditivos e o processo de transformação dos atributos brutos provenientes dos bancos de dados nessas variáveis analíticas. Essa fase é conhecida como “engenharia de atributos”.

O correto relacionamento entre tabelas é um passo relevante para dar sentido ao dado. Em alguns casos, os dados brutos devem ser agrupados, como, por exemplo, o conjunto de notas de entrada ou de saída de um contribuinte.

Chapman et al. (2000) destacam a relevância de se trabalhar com estatística descritiva e também a utilização de painéis de Inteligência de Negócio para a melhor compreensão dos dados, inclusive com a verificação de correlação e de dependência entre variáveis.

4.1.4. Modelagem

Nessa fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ótimos. Em regra, existem diversos métodos para o mesmo tipo de problema de mineração de dados. Alguns têm requisitos específicos na forma dos dados de entrada e saída. Portanto, com frequência, é necessário voltar à fase de preparação de dados e reformular as bases de treinamento e teste. Cada modelo testado é avaliado segundo as métricas de acurácia e o potencial preditivo. O modelo uma vez treinado e avaliado pode, então, ser utilizado para realizar previsões em dados novos.

4.1.5. Avaliação

Na avaliação, o mais importante é verificar se o modelo estudado atende aos objetivos do negócio, ou seja, se vai ser capaz de resolver a questão proposta. Para isso são utilizados métodos, como validação cruzada e métricas de desempenho, para avaliar o quão bem os modelos se saem em dados não vistos. Com base nessa avaliação, é possível ajustar e aprimorar os modelos, se necessário.

4.1.6. Implantação

Caso a avaliação indique a adequação dos modelos criados, o resultado deve ser colocado em produção, de modo a agregar valor para o negócio. A forma como isso é feito varia muito, e deve se integrar aos processos de trabalho da organização. Idealmente, os modelos utilizados devem ser continuamente reavaliados, por meio da comparação dos resultados de predições com a realidade prática do negócio, para retroalimentar o processo de concepção de modelos e sua evolução.

5. DESENVOLVIMENTO

5.1 Entendimento do negócio

Com o objetivo de mapear as fragilidades da sistemática atual de tributação e de coleta do ICMS utilizada pelas unidades da federação, foi solicitado aos auditores fiscais estaduais envolvidos no combate a noteiras a resposta a um questionário (ver Anexo I). O questionário foi enviado a todas

as secretarias de fazenda estaduais. Ao todo, 16 auditores fiscais estaduais de dez UFs diferentes responderam. Participaram representantes dos estados de Alagoas, do Distrito Federal, do Espírito Santo, de Goiás, do Maranhão, de Minas Gerais, de Pernambuco, do Piauí, do Rio Grande do Norte e de Tocantins. Os entrevistados têm entre seis e 28 anos de serviço dedicados à fiscalização tributária.

Das respostas coletadas por meio do questionário foram extraídas informações relevantes ao estudo, entre as quais destacamos:

- a) Até o momento, nenhuma UF faz uso de técnicas analíticas baseadas em inteligência artificial para a detecção de noteiras, o que reforça o pioneirismo deste trabalho no tocante à fiscalização tributária entre entes federados brasileiros;
- b) Seis estados declaram usar Business Intelligence (BI) na análise e visualização de dados relacionados com a questão tributária;
- c) Como citado, uma importante ação no combate a noteiras é a Denegação Automática da Nota Fiscal. Dos nove estados entrevistados, apenas três não usam esse procedimento no cotidiano. Todas as seis unidades federadas que usam a denegação o fazem depois de um despacho fundamentado para garantir a segurança jurídica do procedimento. Um estado declarou o uso de Indício de Grave Irregularidade (IGI), por ele proposto, como forma de reduzir a discricionariedade no processo.

Quanto à natureza e integração de diferentes órgãos de fiscalização no combate a noteiras, restou clara a opção majoritária de que o problema deve ser tratado com a integração do Poder Judiciário, do Ministério Público, da Polícia Civil e de fiscalização tributária. Apenas dois auditores fiscais acreditam que a questão se resolve somente no âmbito da fiscalização tributária.

Também foi questionado sobre qual ação efetivamente os estados adotam quando identificam que uma empresa é noteira, permitindo respostas múltiplas:

- a) Cinco estados autuam diretamente a empresa noteira, mas destes apenas um não arrola outros responsáveis solidários, justificando que empresas noteiras não possuem patrimônio e não pagam autos de infrações lavrados. Nesse mesmo sentido, um estado não arrola os sócios do contrato social e os outros arrolam todos os possíveis;
- b) Oito estados informam aos demais que a empresa tem comportamento de noteira, o que faz com que o estado de destino da nota possa dar tratamento ao crédito fiscal das entradas, glosando-os;
- c) Dois estados não publicam no diário oficial do estado o ato de inidoneidade, o que dificulta o trabalho de fiscalização tributária do estado de destino e, ainda, um estado que faz a publicação não dá ciência aos demais, o que dificulta o estado de destino a atuar juridicamente;
- d) O real beneficiário da fraude é o destinatário da nota fiscal, que aproveita um crédito indevido do imposto. Apenas quatro estados afirmam estornar esses créditos, o que faz com que a replicação da fraude seja vantajosa para os seus autores;
- e) Quatro estados oferecem denúncias fiscais ao Ministério Público e fazem chegar ao conhecimento do Poder Judiciário as fraudes praticadas. Isso demonstra que, para parte dos estados estudados, tal prática não possui uma natureza criminal.

Os entrevistados foram estimulados a indicar condutas típicas das empresas noteiras, classificando-as entre 1 (Pouco relevante) e 5 (Muito relevante), no que se refere a um indicativo da prática de evasão fiscal. Entre as condutas aferidas como “Muito relevantes”, pode-se observar a tendência de um sócio ou contador de uma empresa noteira replicar essa fraude em outras empresas do seu grupo econômico. Outro fato recorrente é uma empresa transacionar com noteiras e, *a posteriori*, ela mesma adotar tal postura.

Os auditores fiscais consultados não veem como indicativo de fraude o fato de a empresa não estar em dia com suas obrigações acessórias de escriturar o imposto. Segundo Tadat (2019), existe uma distinção entre o fraudador e o não conforme. Em regra, o “não conforme” pode não estar em dia com a obrigação acessória, e uma ação de fiscalização ou um estímulo, como a malha fiscal, pode trazê-lo à conformidade. Já o fraudador, pela experiência de fiscalização, pode simular registros na entrega da Escrituração Fiscal Digital (EFD) para parecer conforme. O fato de a empresa ter sócios em outros estados também não foi considerado indicativo relevante da prática de noteira.

Além das propostas, os auditores fiscais foram estimulados a sugerir outros indicadores de possível fraude, e os resultados foram:

- Verificar se o sócio é beneficiário de um programa social do Governo;
- Analisar as formas jurídicas utilizadas para a constituição da empresa, basicamente se a empresa é individual ou é uma sociedade;
- Destacar empresas localizadas em apartamentos e salas comerciais no lugar de galpões, principalmente para atacadistas de grãos;
- Validar os telefones e endereços constantes do Cadastro Fiscal e se existe contador cadastrado;
- Suspeitar de empresas no mesmo local onde foi identificada anteriormente uma empresa como sendo noteira.

Ao entrevistado foi, ainda, solicitada sua avaliação com relação à relevância de algumas variáveis quantitativas sugeridas, que podem ser estimadas em dados típicos em poder das secretarias de fazenda, classificando-os entre 1 (Pouco relevante) e 5 (Muito relevante) no que se refere à detecção da prática de noteira. Para a ampla maioria dos auditores consultados, as seguintes variáveis possuem alto poder preditivo nesse sentido:

- Relação entre o capital social da empresa e seu total de saídas: procura verificar se a empresa tem saídas (vendas) em montante muito superior ao seu capital inicial;
- Relação entre o ICMS a recolher, que é uma medida que avalia a subtração do ICMS destacado na saída deduzido do ICMS creditado na entrada, e o total de saídas da empresa;
- Quantidade de documentos emitidos;
- Relação entre o valor contábil da entrada e o valor contábil da saída, pois, numa relação normal, o custo de aquisição das mercadorias, adicionado de uma margem de valor agregado, deve ser próximo ao valor das saídas.

Não houve consenso com relação à relevância das seguintes variáveis:

- Faturamento líquido, que é uma medida que desconta os custos do total de vendas (faturamento bruto);
- Total do gasto com salários;
- Total de clientes, medido numa contagem distinta de destinatários;
- Razão entre a quantidade de empregados e o total de saídas;
- Total de fornecedores, medido numa contagem distinta de emitentes.

Os auditores foram convidados a sugerir outras variáveis quantitativas que podem indicar a prática de noteira. Entre as respostas, destacam-se:

- Quantidade de dias entre a liberação do cadastro fiscal e a primeira nota emitida;
- Concentração do movimento econômico em poucos dias do mês;
- Coerência da atividade econômica prevista (CNAE) e os produtos de vendas;
- Coerência entre produtos na entrada e produtos na saída;
- Verificação de inscrição estadual reativada, ou seja, uma empresa que havia perdido o direito a emitir notas fiscais que pede que sua inscrição seja reativada.

Por fim, a pesquisa perguntou aos auditores fiscais se tinham outros apontamentos relevantes que pudessem contribuir com a pesquisa. Nesse sentido, houve respostas referentes a permitir maior integração entre as unidades da federação e também entre o Poder Judiciário, a Polícia Civil e o Ministério Público e a fiscalização tributária no combate a noteiras.

É importante ressaltar que as respostas obtidas, por não representarem nem 50% das secretarias de fazenda estaduais, não podem ser utilizadas como um retrato completo da questão da fraude fiscal perpetrada pelas empresas noteiras no país, nem do seu combate pelas unidades da federação. Mas nota-se também uma percepção comum, entre os respondentes, da gravidade, amplitude e complexidade do problema, além de uma convergência na seleção de indicadores e variáveis que podem auxiliar na identificação dessas fraudes. Essa convergência se mostrou suficientemente robusta para que esses indicadores e variáveis selecionados fossem, então, utilizados como base para as etapas subsequentes desta pesquisa.

Mas, obviamente, este estudo, realizado no contexto de uma única secretaria de fazenda, não tem a pretensão de propor uma solução, ou um método, de apoio à auditoria das noteiras, generalizável a todas as UFs, tratando-se somente de um estudo exploratório que tentou, na medida do possível, utilizar dados e entendimentos comuns a um número maior de unidades de fiscalização.

5.2 Entendimento dos dados

As bases utilizadas neste estudo se limitam àquelas que um fisco estadual típico tem acesso, de forma a propor uma solução que seja replicável num maior número de UFs. Foi implementado um processo de Extração, Transformação e Carga (ETL) que retirou dados brutos do *data lake* da Secretaria de Fazenda do DF em formato XML, que foram transformados em tabelas em um banco de dados relacional. Essas tabelas foram, então, processadas por meio da linguagem Structured Query Language (SQL), que gerou coleções de dados que foram carregadas e processadas em equipamentos da própria secretaria pela linguagem de programação Python.

A seleção das bases de dados para extração foi feita fundamentada em destaques e sugestões de variáveis e indicadores de risco feitos pelos próprios auditores fiscais nas suas respostas ao questionário. Essas bases contêm dados de documentos fiscais eletrônicos, especificamente as NFes e seus eventos, além de dados cadastrais e societários. De maneira acessória, foram usados dados de Conhecimento de Transporte Eletrônico (CT-e), Nota Fiscal do Consumidor Eletrônica (NFCe), Escrita Fiscal Digital (EFD) e Declaração de Meios de Pagamento (DIMP).

Os eventos das NFes indicam se houve denegação automática e qual o tipo de ocorrência que levou à denegação imediata da emissão de notas fiscais. O auditor fiscal registra, inicialmente, um motivo padronizado para o bloqueio da NFe, que pode ser consequência de uma irregularidade cadastral ou um volume de operações incompatível com o porte, a situação cadastral ou o capital social da empresa. Muitas vezes, os contribuintes que sofrem uma ação de denegação automática por parte da fiscalização tributária, mas que são contribuintes idôneos, buscam o fisco que, por sua vez, reverte essa situação impeditiva da emissão do documento fiscal. Uma investigação subsequente feita pelos auditores da Secretaria de Fazenda do Distrito Federal pode identificar alguma fraude fiscal, classificando a empresa como noteira, se for o caso.

Dos cerca de 655 mil contribuintes cadastrados no Distrito Federal, apenas 174 mil emitiram notas fiscais de janeiro de 2014 a junho de 2023. Para este estudo, foi selecionado, inicialmente, um conjunto de 2.801 empresas indicadas como suspeitas de alguma irregularidade e que, posteriormente, foram classificadas como noteiras ou não pelos auditores fiscais do DF, por meio de validações em campo. Essa classificação foi, então, utilizada como rótulo para definir a base de treinamento dos modelos preditivos.

Assim, o indicativo de denegação isoladamente não é suficiente para determinar se houve fraude fiscal, pois ela pode ocorrer por outros fatores, mas a confirmação por meio de auditoria dá robustez à seleção de casos de fraude. Da mesma forma, não se pode afirmar que empresas que não foram analisadas pelos auditores fiscais e não possuem evento de denegação automática não são noteiras, o que dificultaria a seleção de empresas “não noteiras” simplesmente por amostragem entre aquelas sem denegação automática.

O passo seguinte foi unir a base de dados de noteiras e não noteiras identificadas e validadas pelos auditores fiscais, com os demais documentos fiscais eletrônicos emitidos por elas.

Para o estudo, foram extraídos dados de Nota Fiscal Eletrônica (NFe) de janeiro de 2014 a junho de 2023 e de Nota Fiscal Eletrônica ao Consumidor (NFCe) de janeiro de 2016 (a NFCe é um documento com adesão mais recente pelos contribuintes) a junho de 2023.

O percentual de empresas sem registro de entrada ou saída por Documento Fiscal Eletrônico (DFE) foi de 4,1%, e das empresas sem documentos de saída ou em que estes se encontravam zerados, 9,7% da amostra nos dados, como pode ser observado na Tabela 1. Para essas empresas, os únicos dados disponíveis são os cadastrais e societários, o que prejudica sua comparação às demais empresas contidas na base de dados. Por esse motivo, este estudo prosseguiu apenas com as empresas que possuem documentos fiscais eletrônicos de entrada ou saída com valores de ICMS.

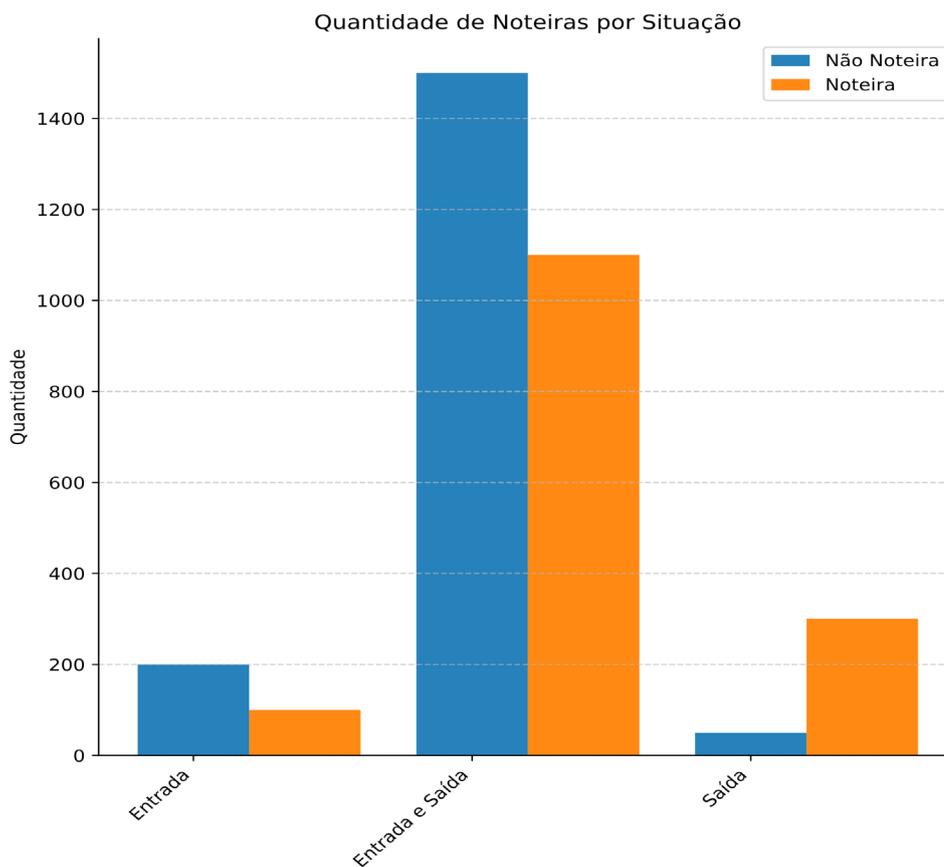
TABELA 1 EXISTÊNCIA OU NÃO DE DOCUMENTO FISCAL ELETRÔNICO (DFE)

	Contagem	Sem documento fiscal de saída ou com valor do zerado	Sem registro de entrada ou saída
Noteira	1.369	144	42
Não noteira	1.432	173	92
Total	2.801	317	134

Fonte: Elaborada pelos autores.

Na Figura 2, temos a frequência de contribuintes da base selecionada que promovem apenas a entrada de mercadorias, apenas a saída ou ambos.

FIGURA 2 ENTRADAS VS. SAÍDAS

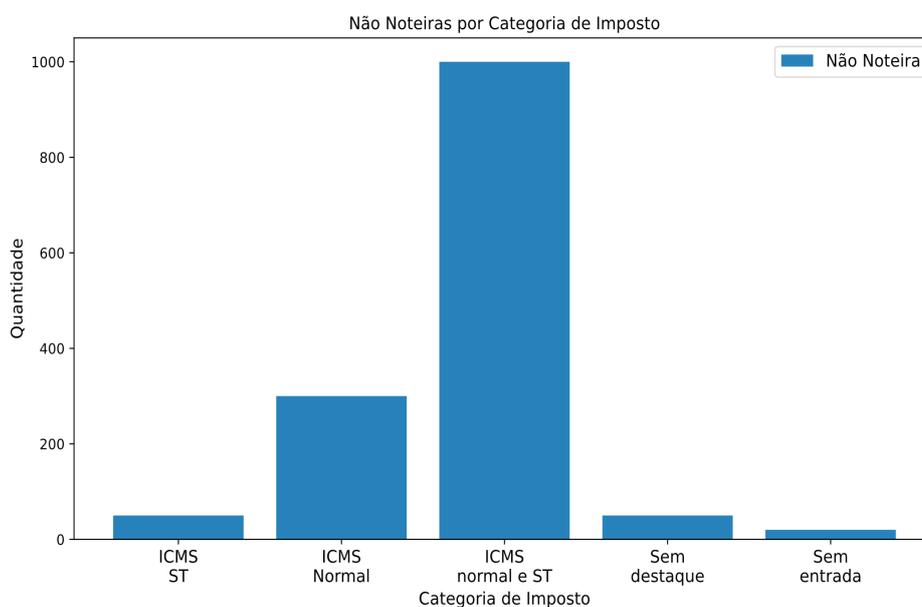


Fonte: Elaborada pelos autores.

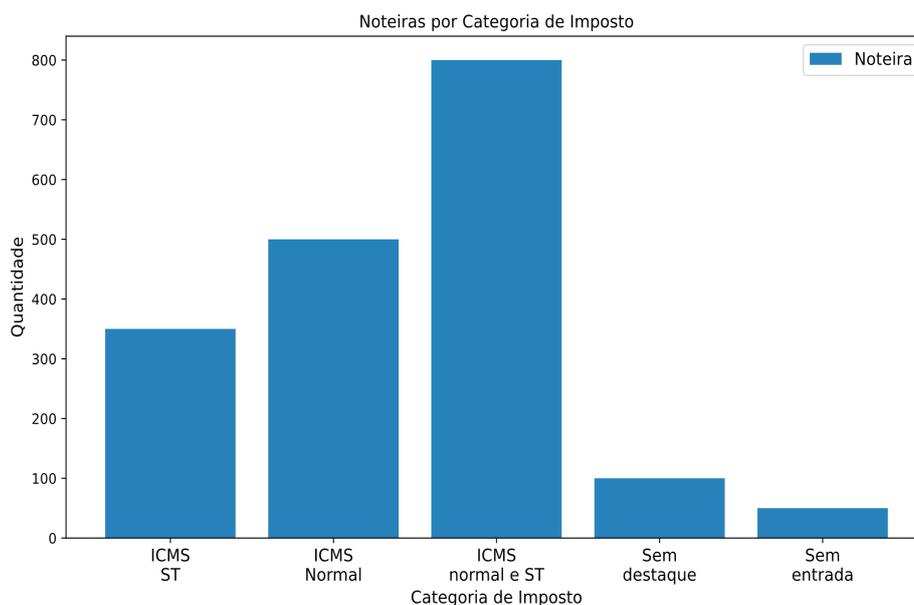
Observa-se, então, que as empresas que têm apenas a entrada de mercadorias são menos frequentes e elas há um equilíbrio entre noteiras e não noteiras. As noteiras, entretanto, possuem a tendência marcante de vender o que não compraram, como se nota no terceiro histograma da Figura 2, que mostra a frequência de empresas para as quais há somente saída de mercadorias. A maioria dos casos é de empresas que compram e vendem (entrada e saída), e, nesse caso, não é tão evidente diferenciar todo o universo de noteiras.

Outra característica relevante, ilustrada na Figura 3, trata do imposto que é destacado em nota, se é o ICMS Normal, se é o ICMS Substituição Tributária (ICMS-ST¹), que antecipa o recolhimento do ICMS em uma etapa anterior, se não tem imposto destacado ou se tem ambos. Novamente, a característica que se destaca para as noteiras é o recebimento de notas com destaque de ICMS e ICMS-ST, seguida do destaque apenas do ICMS e, por fim, daquelas que não têm notas de entrada (compras). Já a tendência a comprar sem destaque de ICMS é mais pronunciada entre as não noteiras.

FIGURA 3 DESTAQUE DE ICMS OU SUBSTITUIÇÃO TRIBUTÁRIA NA ENTRADA



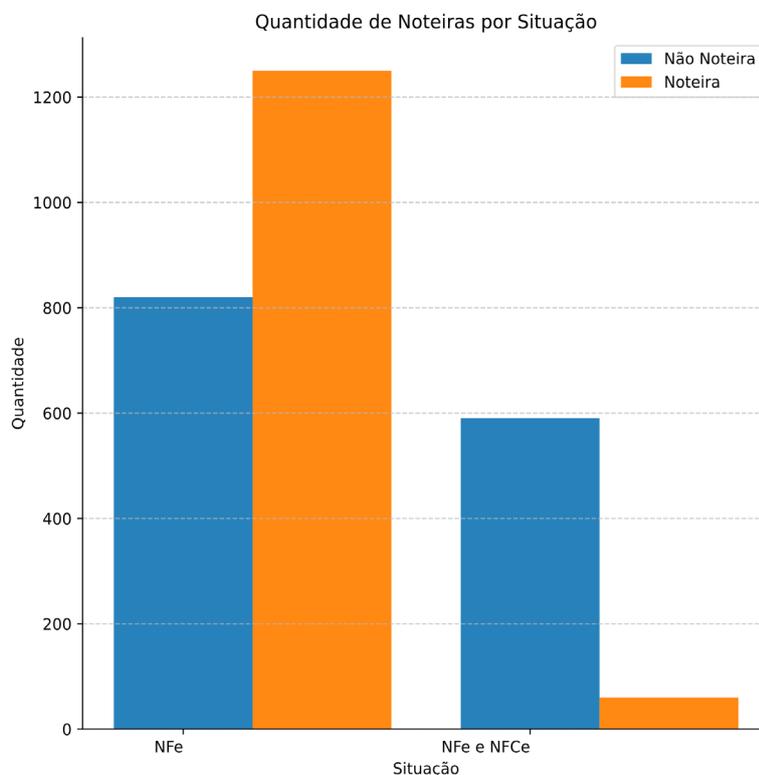
¹A Substituição Tributária (ST) do ICMS é o regime pelo qual a responsabilidade pelo imposto devido em relação às operações ou prestações de serviços é atribuída a outro contribuinte.



Fonte: Elaborada pelos autores.

Os dados confirmam um conhecimento empírico de que empresas noteiras trabalham apenas com o ICMS Normal, de modo que não utilizam mercadorias sujeitas somente ao regime de ICMS Substituição Tributária. Nas empresas não noteiras, existe uma divisão maior entre o ICMS Normal e a Substituição Tributária. A noteira inscrita no regime Simples Nacional destaca o ICMS na saída, algo previsto apenas para as empresas maiores desse regime em situações bem específicas. A não noteira do Simples Nacional tem um comportamento oposto, de não destacar o ICMS, o que é esperado segundo o regramento tributário dessas empresas.

É possível observar também que existe uma distinção relevante entre a emissão de NFe e NFCe por noteiras e não noteiras, como ilustrado na Figura 4. O volume de transações apenas com NFe é muito maior entre as noteiras do que entre as não noteiras, pois essas últimas guardam uma relação mais próxima de vendas com o consumidor, logo, com a NFCe. Entretanto, o conjunto de empresas classificadas como noteiras e não noteiras pelos auditores fiscais do DF não é suficiente para afirmarmos que noteiras não emitem NFCe. Da mesma forma, não é possível afirmar que todas as empresas não noteiras emitem ambos os documentos, uma vez que, por causa de sua natureza de comércio, é possível que transacionem apenas com pessoas jurídicas utilizando somente NFes.

FIGURA 4 USO EXCLUSIVO DE NFE OU EM CONJUNTO COM NFCE

Fonte: Elaborado pelos autores.

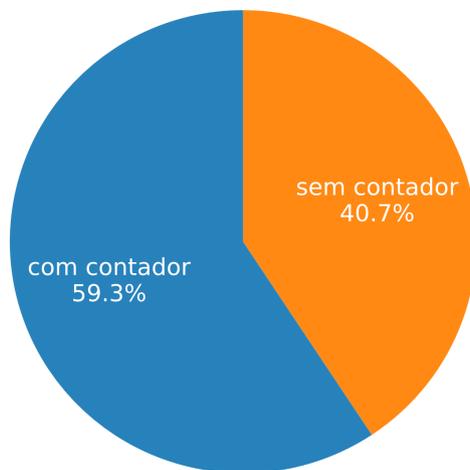
Empresas que se beneficiam do esquema das noteiras são aquelas que compram notas para poder se creditar no ICMS e pagar menos imposto. Espera-se que as empresas destinatárias de uma dada noteira comprem de várias noteiras. Com o objetivo de desenvolver o estudo, foram extraídos, do banco de NFes, os emitentes e destinatários distintos em todas as notas. Procedeu-se a uma extração com o seguinte algoritmo:

1. Para as empresas sabidamente noteiras, foram identificados seus destinatários;
2. Dos destinatários do passo 1, foram extraídas as empresas que venderam (emitentes) para elas, diferente da empresa do passo 1;
3. Verifica-se a classificação das empresas do passo 2.

Verifica-se que as empresas destinatárias de uma dada noteira tendem a comprar de outras noteiras em uma proporção atípica: quase metade de seus fornecedores (46,3%) são noteiras. Isso corrobora o identificado no questionário. É uma informação relevante, mas que, isolada, não possui poder preditivo.

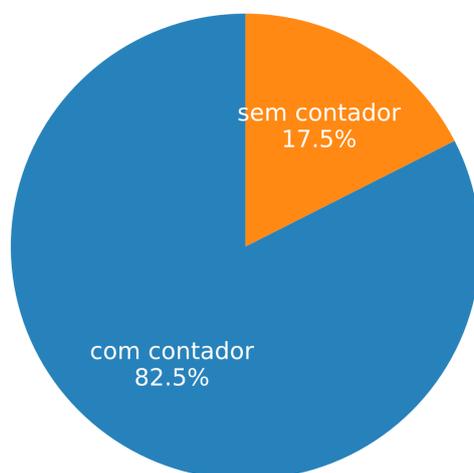
A relação contador-cliente pode indicar várias práticas fiscais e contábeis e o comportamento da empresa. A ausência de um contador cadastrado não quer dizer, necessariamente, que a empresa não possua a orientação de um contador, mas, eventualmente, que esse profissional não foi informado à Secretaria de Fazenda do Distrito Federal.

FIGURA 5 CONTADOR EM NÃO NOTEIRA



Fonte: Elaborada pelos autores.

FIGURA 6 CONTADOR EM NOTEIRA

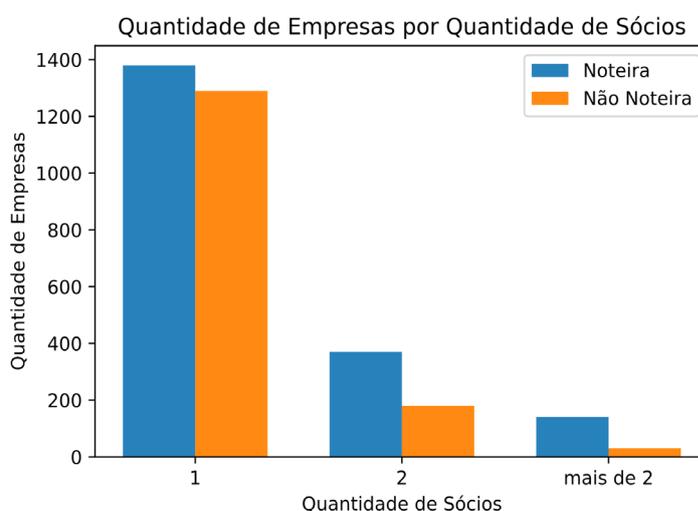


Fonte: Elaborada pelos autores.

Ao se comparar os gráficos das Figuras 5 e 6, fica evidente que o percentual de empresas com contador é significativamente menor em empresas noteiras. Como geralmente se trata de uma empresa de fachada e com possível indiciamento por crime contra a ordem tributária, fica marcada a tendência do contador oculto para, no caso de o esquema ser descoberto, não lhe atribuir crime. Entretanto, alguns contadores procuram prestar serviço no sentido de dar um certo ar de legalidade à empresa.

A análise de composição societária é outro ponto relevante dentro do estudo. Microempreendedor Individual e Empresa Individual de Responsabilidade Limitada (Eireli) são empresas de apenas um titular (ou sócio único). As demais empresas são constituídas na forma de uma sociedade, em que dois ou mais sócios se unem para realizar os objetivos sociais da empresa. A Figura 7 mostra a quantidade de empresas por quantidade de sócios nas empresas da base.

FIGURA 7 QUANTIDADE DE SÓCIOS POR EMPRESA

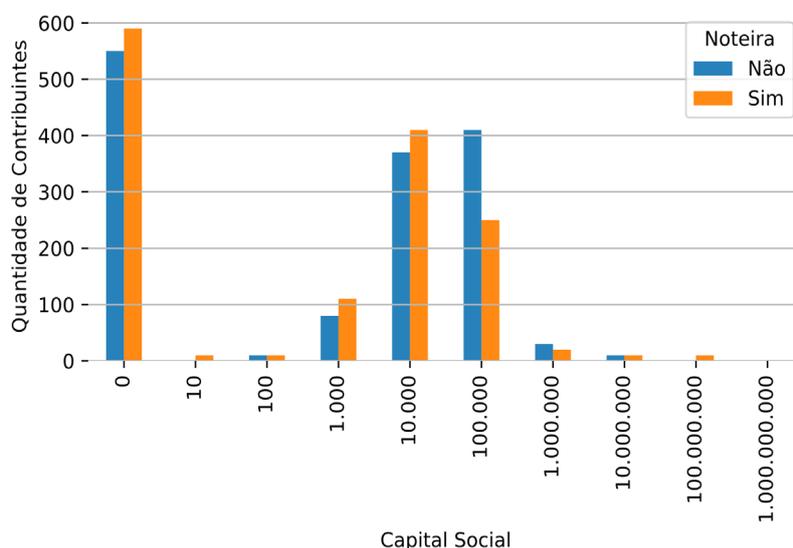


Fonte: Elaborada pelos autores.

Observa-se que as empresas identificadas como noteiras têm a tendência de ter um único sócio, além de a ocorrência de mais de dois sócios ser visivelmente baixa nessa categoria.

Quanto ao capital social da empresa, observou-se que existe uma pequena tendência na base das não noteiras possuírem um capital social maior que o das noteiras, como pode ser visto na Figura 8. Inclusive, entre aquelas empresas que não declaram capital social, a presença de noteiras é bem substancial.

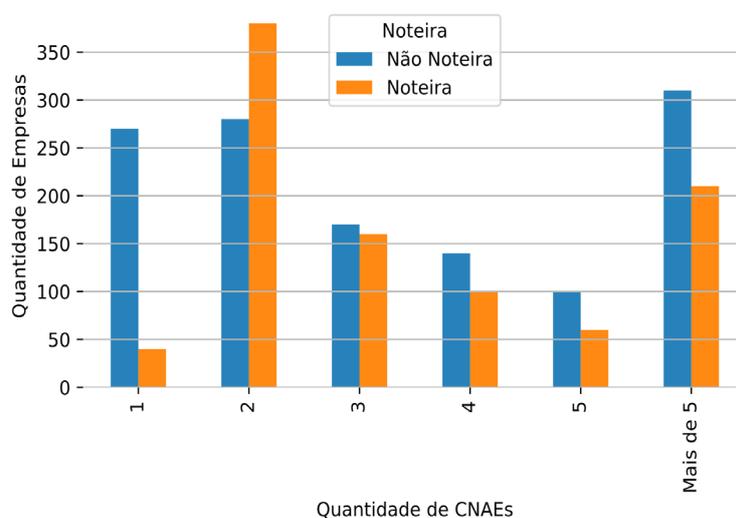
FIGURA 8 CAPITAL SOCIAL DAS EMPRESAS



Fonte: Elaborada pelos autores.

A Classificação Nacional de Atividades Econômicas (CNAE) também revelou que existe uma sutil tendência de a noteira declarar menos CNAEs que as não noteiras, como pode ser verificado na Figura 9.

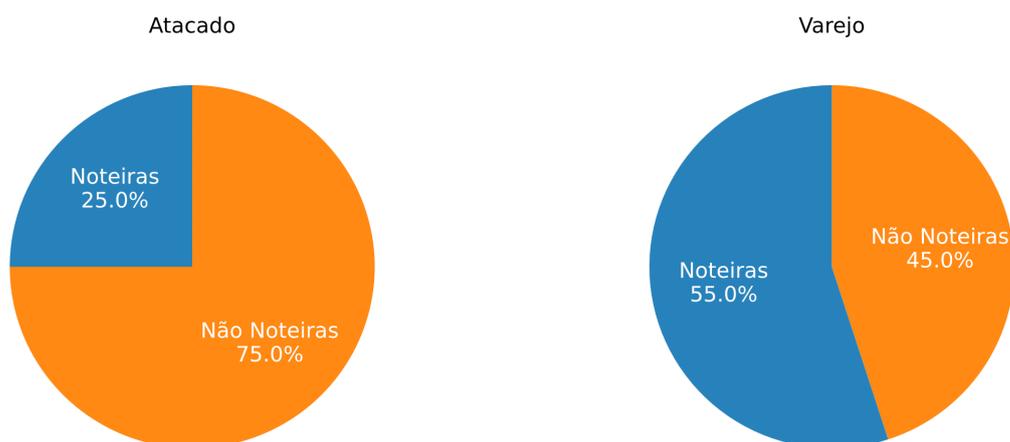
FIGURA 9 QUANTIDADE DE CNAES POR TIPO DE EMPRESA



Fonte: Elaborada pelos autores.

Ainda com relação à atividade econômica, as noteiras atuam, principalmente, no comércio e, em geral, preferem o varejo, como mostra a Figura 10.

FIGURA 10 COMPARAÇÃO ENTRE ATACADO E VAREJO



Fonte: Elaborada pelos autores.

5.3 Preparação dos dados

O objetivo dessa fase é preparar uma coleção de dados que será utilizada na fase de modelagem. Essa coleção contém dados sumarizados por empresa, ou seja, cada empresa possui um registro único com seus dados agregados do período em que permaneceu em atividade. Para o treinamento dos modelos preditivos, essa coleção vem adicionada da indicação da empresa sendo ou não uma noteira. Uma vez treinados, os modelos preditivos são capazes de sugerir essa classificação segundo dados de entrada.

5.3.1 Definição dos atributos (features)

As empresas selecionadas para a base de dados de treinamento e teste emitiram, aproximadamente, 7 milhões de notas entre 2014 e 2023. Como forma de criar um perfil da atuação de cada empresa, foi feita uma agregação, em que cada uma tem seus valores sumarizados em uma única tupla.

A sumarização da atividade de cada empresa levou em consideração os seguintes critérios:

- Atividade econômica
 - Atua no varejo e/ou no atacado?
- Perfil da clientela
 - Os clientes compram de empresas já identificadas como noteiras?
- Contador
 - A empresa tem contador?
 - O contador atua ou atuou para empresas já identificadas como noteiras?

- Quadro societário
 - Quantidade de sócios.
 - Algum sócio da empresa já esteve ligado a empresa identificada como noteira?
- Dados financeiros
 - Valores contábeis das operações da empresa:
 - Discriminados entre entradas e saídas de mercadorias;
 - Discrimina ICMS destacado.
 - Contagem de notas emitidas e recebidas.
 - Emite nota para consumidor final (NFCE)?
 - Compra de empresa identificada como noteira e de quantas?
 - Quantidade de fornecedores da empresa alvo que também vende para empresa noteira.
 - Quantidade de meses em que recebeu e vendeu mercadorias:
 - Diferença entre o primeiro e o último mês com emissão fiscal.
 - Capital social da empresa.

De forma a evitar diferenças muito amplas de valores entre empresas de diferentes portes e volume de transações, os atributos foram normalizados sempre que possível. Para isso, buscou-se adotar, ao máximo, critérios que possuem justificativas do ponto de vista do negócio para manter a semântica de cada atributo.

De forma a definir regras adequadas a cada atributo, eles foram separados em três categorias:

- De origem financeira;
- De contagem distinta de pessoa (sócio, cliente, fornecedor);
- De contagem distinta de documentos fiscais eletrônicos emitidos.

Para os dados financeiros, por exemplo, um valor interessante para se tomar a base da normalização é a soma do valor contábil (VNF) de todas as notas de entrada e saída. No volume de saída, são consideradas, inclusive, as vendas ao consumidor final com o documento Nota Fiscal Eletrônica ao Consumidor (NFCe). Com base nesse valor, são calculadas proporções de entrada e de saída. Dessa forma, um contribuinte que apenas dá saída (venda) sem dar entrada (compra), ao ter suas entradas normalizadas pela VNF, obterá um valor próximo de zero e suas saídas normalizadas terão um valor próximo de um. Já um contribuinte típico, que compra mercadorias para comercialização, terá seus valores de entrada e saída normalizados com valores em torno de 0,5.

A agregação descrita anteriormente procurou refletir, de novo, os elementos indicados pelos auditores fiscais estaduais como relevantes para a análise da fraude fiscal do esquema das noteiras, descritos na seção 5.1.

Para a seleção final dos atributos com maior valor preditivo, foi feita uma análise de correlação, que utilizou o coeficiente de Pearson, entre as variáveis de entrada, que descrevem a empresa, e entre cada variável de entrada e a de saída, que indica a classe da empresa (noteira ou não noteira).

Shimakura (2006) define as variáveis com coeficientes maiores que 0,90 como altamente correlacionadas. Variáveis de entrada fortemente correlacionadas entre si significam duplicidade de informação, o que indica que a eliminação de uma delas não causa prejuízo ao modelo preditivo. Já variáveis de entrada fracamente correlacionadas às de saída indicam atributos com fraco poder

preditivo. Dessa forma, para cada par de variáveis de entrada cujo coeficiente de Pearson foi superior a 0,90, uma delas foi descartada, e variáveis de entrada cuja correlação com a variável-alvo apresentaram coeficiente de Pearson inferior a 0,90, também o foram. Depois da análise do valor informativo de cada atributo, foram selecionados 11 numéricos e oito binários para compor a base final, descritos da Quadro 1.

QUADRO 1 DESCRIÇÃO DOS ATRIBUTOS SELECIONADOS

Descrição	Tipo
Se atua no atacado	Binária
Se atua no varejo	Binária
Se clientes já compraram de noteiras	Binária
Se contador atua em noteiras	Binária
Se tem contador	Binária
Se emite NFCe	Binária
Se tem sócio que é sócio de outras empresas	Binária
Se tem sócio que é ou foi sócio de noteira	Binária
Valor contábil das vendas com a soma de NFe e NFCe	Numérica
Valor total do ICMS na saída	Numérica
Valor contábil das operações de entrada de mercadoria (compra)	Numérica
Valor do ICMS Substituição Tributária (ST) destacado	Numérica
Valor do ICMS destacado para compra de mercadoria	Numérica
Valor do capital social registrado no contrato social da empresa	Numérica
Número de meses entre o primeiro e o último mês com emissão fiscal	Numérica
Quantidade de sócios da empresa	Numérica
Quantidade de notas fiscais emitidas para entrada de mercadorias	Numérica
Razão entre a quantidade de documentos de saída e o total de documentos emitidos	Numérica
Razão entre o valor das vendas e o total das transações de compra e venda	Numérica

Fonte: Elaborada pelos autores.

5.3.2 Preparação da base de treinamento e do teste

Os dados foram separados em treinamento e teste de maneira aleatória, como é praxe em experimentos de aprendizagem de máquina supervisionados, e foram tomadas também todas as precauções para se evitarem vieses por causa do vazamento de dados (*data leakage*) entre a base de treinamento e a de teste. Foram utilizados 75% dos dados para treinamento e 25% para testes, de forma que, para treinamento, havia 966 exemplos de noteiras e 981 de não noteiras, enquanto para teste, havia 341 empresas noteiras e 309 não noteiras, constituindo uma base razoavelmente balanceada.

De forma a corrigir os dados faltantes (*missing value*), foi realizada sua substituição por medidas de tendência central adequadas a cada tipo de dado.

5.3.3 Verificação de viés nos dados

O estudo se baseia numa base classificada por auditores fiscais do Distrito Federal. As empresas não noteiras do estudo são empresas que tiveram suas notas preventivamente denegadas e que, depois de análise das autoridades fiscais, foram liberadas à emissão. Não há na base, portanto, empresas sobre as quais pairou dúvida se são ou não noteiras no final da análise. Mas, ao mesmo tempo, sobre todas as empresas da base classificadas, em algum momento, pairaram dúvidas a respeito de seu comportamento, mesmo que por razões diversas. Por conta disso, restou a dúvida de se a base rotulada como não noteira é representativa do universo de empresas contribuintes, já que a prática da fraude é uma exceção, ou se já contém um viés por ter sido selecionada previamente pelos auditores.

Para examinar a possibilidade de haver um viés embutido na base de não noteiras, foi criada uma nova amostra aleatória de empresas no restante das empresas não analisadas. Foi feita, então, uma análise estatística, por meio de um teste de hipótese, que buscou verificar se o subconjunto de empresas não noteiras do conjunto de denegações apresenta as mesmas distribuições de frequência da nova amostra aleatória. O teste confirmou a hipótese nula de que ambas as amostras têm o mesmo comportamento estatístico.

5.4 Modelagem

5.4.1 Seleção de modelos para treinamento

O experimento aqui realizado é considerado, na literatura de aprendizagem de máquina, uma aprendizagem supervisionada, baseada num conjunto de exemplos rotulados que buscam encontrar um modelo preditivo que, uma vez treinado, seja capaz de classificar se o comportamento de uma empresa, por meio de suas transações de compra e venda, dados cadastrais e sócios, fornecedores e clientes, é o de uma empresa noteira.

Os algoritmos usados no treinamento são todos da biblioteca Sklearn² para Python, que foi escolhida por causa da facilidade de uso, farta documentação, disponibilidade gratuita e diversidade de métodos ali contidos.

Em projetos que envolvam aprendizagem de máquina é comum explorar uma ampla gama de modelos preditivos e, no final, selecionar aqueles que melhor se comportam segundo métricas e critérios preestabelecidos. Algumas dessas métricas avaliam a acurácia das predições, outras, sua interpretabilidade ou, ainda, o perfil dos erros que o modelo normalmente comete.

Em Carvalho et al. (2019), é possível obter um completo tutorial sobre métricas e estratégias de avaliação de modelos preditivos. Nosso trabalho explorou os algoritmos de classificação mais citados e utilizados na literatura, com especial atenção àqueles aplicados na detecção de fraudes, incluindo os utilizados nos trabalhos destacados na seção 2.1. Depois de diversas rodadas de testes e ajustes de parâmetros, os algoritmos selecionados como os mais adequados ao problema tratado são estes, apresentados na Quadro 2. Uma visão geral de como esses algoritmos funcionam pode ser obtida em Mahesh (2020).

² <https://scikit-learn.org/stable/>

QUADRO 2 MODELOS DE TREINAMENTO

Função Sklearn	Algoritmo
RandomForestClassifier	Árvores de decisão aleatórias
GradientBoostingClassifier	Descida de gradiente
MLPClassifier	Rede neural
AdaBoostClassifier	Ada Boost
KNeighborsClassifier	Vizinhos mais próximos
DecisionTreeClassifier	Árvore de decisão
SVC	Máquina de vetores de suporte

Fonte: Elaborada pelos autores.

5.4.2 Treinamento e teste

Para obter a melhor configuração de hiper-parâmetros dos modelos, optou-se por usar o recurso *pipeline* estendido para processar uma validação cruzada (*cross validation*) de teste com diversas configurações dos algoritmos (Claesen & Moor, 2015). Para tanto, foi utilizado o algoritmo GridSearchCV, que realiza uma pesquisa exaustiva sobre os valores de parâmetros especificados para cada classificador.

Os melhores resultados foram obtidos com os seguintes parâmetros:

- AdaBoostClassifier
 - learning_rate (taxa de aprendizado do modelo durante o treinamento): 0,15
 - n_estimators (número de estimadores no modelo): 200
- GradientBoostingClassifier
 - learning_rate: 0,05
 - n_estimators: 200
- KNeighborsClassifier
 - n_neighbors (número de vizinhos mais próximos que serão considerados): 9
- MLPClassifier:
 - activation (função de ativação usada nas camadas ocultas da rede neural): tanh
 - hidden_layer_sizes (o número e o tamanho das camadas ocultas da rede neural): 150
 - solver (algoritmo de otimização usado para ajustar os pesos da rede neural): adam
- RandomForestClassifier
 - criterion (medida de qualidade usada para avaliar a divisão de cada nó da árvore de decisão): entropy
 - n_estimators: 40
- SVC
 - kernel (método de estimativa de fronteiras): rbf

Feito o treinamento dos modelos com os hiperparâmetros anteriores, realizou-se o teste no conjunto de dados não utilizados no treinamento. Cada instância é apresentada a cada um dos modelos e gera um dos seguintes resultados:

- Verdadeiro positivo – noteiras, identificadas como noteiras;
- Verdadeiro negativo – não noteiras, identificadas como não noteiras;
- Falso positivo – não noteiras que são identificadas como noteiras;
- Falso negativo – noteiras que são identificadas como não noteiras.

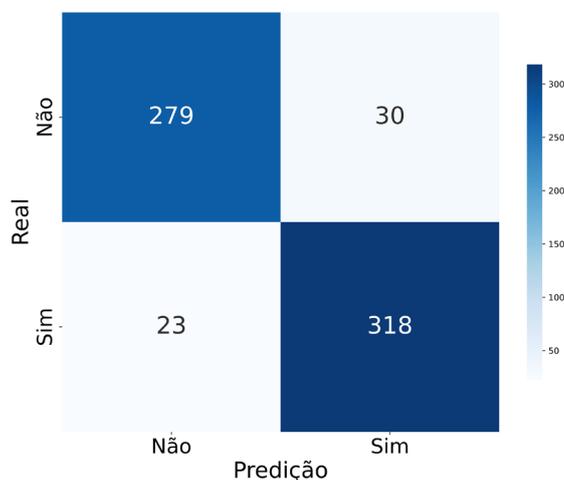
A Tabela 2 sumariza os resultados com base nas métricas convencionais utilizadas na aprendizagem de máquina.

TABELA 2 RESULTADOS DE AVALIAÇÃO DOS MODELOS

Algoritmo	Acurácia ¹	Precisão ²	Recall ³	Especificidade ⁴	F1Score ⁵
MLPC	0,918	0,933	0,914	0,924	0,923
Random Forest	0,915	0,903	0,933	0,897	0,918
Gradient Boosting	0,911	0,912	0,917	0,904	0,915
SVC	0,908	0,912	0,912	0,903	0,912
Ada Boost	0,895	0,9	0,9	0,89	0,9
K Neighbors	0,889	0,909	0,883	0,896	0,896
Decision Tree	0,715	0,745	0,695	0,738	0,719

Fonte: Elaborada pelos autores.

Como indicado na Tabela 4, o modelo que teve o melhor desempenho foi a rede neural Multi-layer Perceptron Classifier com a matriz de confusão da Figura 11.

FIGURA 11 MATRIZ DE CONFUSÃO DO TESTE COM O MULTI-LAYER PERCEPTRON

Fonte: Elaborada pelos autores.

6. CONSIDERAÇÕES FINAIS

Este trabalho buscou investigar o uso de técnicas preditivas baseadas em aprendizagem de máquina na identificação de uma prática específica de fraude fiscal, praticada por empresas popularmente conhecidas como “empresas noteiras”, formadas, exclusivamente, para emitir créditos não devidos de ICMS.

Seguindo uma metodologia consolidada de análise de dados, o estudo começou pela compreensão do contexto negocial que envolvia a questão tributária e sua fiscalização pelas secretarias de fazenda estaduais.

Por meio de uma enquete com os auditores fiscais de diversas unidades da federação, foi possível mapear as fragilidades da sistemática de tributação e coleta do ICMS exploradas pelas empresas noteiras e as estratégias utilizadas pelos auditores para combatê-las.

Em seguida, um trabalho extenso de obtenção, tratamento, carga e compreensão dos dados tipicamente disponíveis para os órgãos fiscalizadores permitiu a materialização e mensuração da fraude fiscal em suas diversas dimensões. Depois da seleção das variáveis mais relevantes derivadas desses dados, fundamentada no conhecimento negocial dos auditores fiscais e no uso de técnicas estatísticas adequadas, foi preparada uma base de dados para o treinamento e teste dos algoritmos de aprendizagem de máquina. A aplicação de um amplo espectro de algoritmos, incluindo os mais referenciados na literatura, demonstrou a viabilidade de identificação das empresas com comportamento típico de noteiras.

Os métodos ora propostos de predição baseados em ciência de dados têm o potencial para melhorar a seleção dos contribuintes que serão incluídos na programação fiscal para auditoria baseada em indícios analíticos e vistoria *in loco*, em razão do critério adotado pela fiscalização de comprovar os indícios.

O modelo apresentado neste trabalho está sendo aplicado, na prática, na Secretaria de Fazenda do Distrito Federal (Sefaz-DF) na busca por indícios de fraude fiscal, com o intuito de aumentar a efetividade no combate a noteiras.

Como trabalho futuro, a coleta de resultados durante um período estatisticamente relevante permitirá comparar as estimativas de acurácia dos modelos com a realidade em campo, assim como a proposta de melhorias na sistemática proposta. A metodologia utilizada é transparente e condizente com a prática da maioria dos órgãos fiscalizadores estaduais e, da mesma forma, as fontes de dados necessárias para sua aplicação estão, a princípio, disponíveis também na maioria das unidades da federação, de forma que o trabalho realizado pode ser útil ou servir de ponto de partida para outras iniciativas similares.

As principais fragilidades do presente estudo estão centradas em fatores externos, como a inviabilidade do uso de bases não acessíveis pela fiscalização tributária, entre elas pode-se citar o Cadastro Geral de Empregados e Desempregados (Caged), que poderia trazer melhor adequação ao cadastro de sócios.

REFERÊNCIAS

- Abrantes, P. C., & Ferraz, F. (2016). Big data applied to tax evasion detection: a systematic review. In *International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA.
- Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: a theoretical analysis. *Journal of Public Economics*, 1(3-4), 323-38. [http://www.sciencedirect.com/science/article/pii/0047-2727\(72\)90010-2](http://www.sciencedirect.com/science/article/pii/0047-2727(72)90010-2)
- Andrade, J. P. A., Paulucio, L. S., Paixão, T. M., Berriel, R. F., Carneiro, T. C. J., Carneiro, R. V., Souza, A. F., Badue, C., & Oliveira-Santos, T. (2021). A machine learning-based system for financial fraud detection. In *Anais do 18º Encontro Nacional de Inteligência Artificial e Computacional*, Porto Alegre, RS, Brasil. <https://doi.org/10.5753/eniac.2021.18250>
- Azevedo, R. R., Silva, J. M., & Gatsios, R. C. (2015). Comparação de modelos de previsão de série temporal com base no ICMS estadual. Contabilidade e controladoria no século XXI. In *Anais do 4º Congresso Controladoria e Contabilidade*, São Paulo, SP, Brasil. <https://congressosp.fipecafi.org/anais/artigos152015/35.pdf>
- Bittencourt, S. A. P., Neto. (2018). *Análise de “outliers” para o controle do risco de evasão tributária do ICMS* (Dissertação de Mestrado). Universidade de Brasília, Brasília, DF, Brasil.
- Carvalho, L. (2018, 07 de dezembro). Mato Grosso integra operação nacional de combate às empresas noteiras. *Sefaz Notícias*. <http://www5.sefaz.mt.gov.br/-/10942780-mato-grosso-integra-operacao-nacional-de-combate-as-empresas-noteiras>
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: a survey on methods and metrics. *Electronics*, 8(8), 832. <https://doi.org/10.3390/electronics8080832>
- Castellón-González, P., & Velásquez, J. D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, 40(5), 1427-36. <https://doi.org/10.1016/j.eswa.2012.08.051>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: step-by-step data mining guide*. SPSS Inc.
- Claesen, M., & Moor, B. (2015). *Hyperparameter search in machine learning*. <https://doi.org/10.48550/arXiv.1502.02127>
- Ferreira, R. P. (2018). *Reconhecimento de cenários baseado nas localizações dos fornecedores do governo federal* (Dissertação de Mestrado). Universidade de Brasília, Brasília, DF, Brasil.
- Ippolito, A., & Lozano, A. (2020). Tax crime prediction with machine learning: a case study in the municipality of São Paulo. In *Proceeding of the 22º International Conference on Enterprise Information Systems*. <https://www.scitepress.org/Papers/2020/95647/95647.pdf>
- Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica*, 79(3), 651-92. <https://eml.berkeley.edu/~saez/kleven-knudsen-kreiner-pedersen-saezEMA11taxaudit.pdf>
- Lima, S. L. M. (2007). O acompanhamento tributário - um novo paradigma em fiscalização para a Receita Federal do Brasil. In Ministério da Fazenda, & Secretaria da Receita Federal (Orgs.), *Administração pública: prêmio de criatividade e inovação auditor fiscal da Receita Federal José Antônio Schöntag: 6º prêmio Schöntag: monografias premiadas* (pp. 877-917). SRF. <https://repositorio.enap.gov.br/handle/1/4575>
- Mahesh, B. (2020). Machine learning algorithms: a review. *International Journal of Science and Research*, 9, 381-386. <https://www.ijsr.net/archive/v9i1/ART20203995.pdf>
- Matos, R. T. B. R. (2019). *Feature selection with low correlated binary features for potential tax fraudsters classification* (Tese de Doutorado). Universidade Federal do Ceará, Fortaleza, CE, Brasil. <https://repositorio.ufc.br/handle/riufc/43348>
- Ortiz, J. (2022, 01 de fevereiro). Fazenda: Receita estadual identificou 844 empresas falsas nos últimos cinco anos. *Agência Estadual de Notícias*. <https://rrmais.com.br/noticia/noticias/parana/fazenda-receita-estadual-identificou-844-empresas-falsas-nos-ultimos-cinco-anos>
- Redação DM Anápolis. (2021, 11 de novembro). *Operação desarticula esquema de notas fiscais frias*. <https://www.dmanapolis.com.br/noticia/16207/>

operacao-desarticula-esquema-de-notas-fiscais-frias

Ruzgas, T., Kižauskienė, L., Lukauskas, M., Sinkevičius, E., Frolovaitė, M., & Arnastauskaitė, J. (2023). Tax fraud reduction using analytics in an East European country. *Axioms*, 12(3), 288. <https://doi.org/10.3390/axioms12030288>

Santos, A. H. S., Rocha, K. de L., Toldo, L. de A., & Fabel, V. H. B. (2022). *Estimativa da carga tributária bruta do governo geral*. Tesouro Nacional. <https://www.anfip.org.br/wp-content/uploads/2023/03/RT-CARGA-TRIBUTARIA-ANUAL-2021-1.pdf>

Shimakura, S. E. (2006, 30 de agosto). *Interpretação do coeficiente de correlação, CE003 – Estatística II*. Universidade Federal do Paraná. <http://leg.ufpr.br/~silvia/CE003/node74.html>

Souza, J. M. (2018). Tributos sobre consumo: novo modelo para um Brasil mais justo. In Fagnani, E.

(Org.), *A reforma tributária necessária: diagnóstico e premissas* (804 p). Anfip, Fenafisco e Plataforma Política Social.

Tadat. (2019, 01 de novembro). *Tadat Subnational Field Guide*. <https://www.tadat.org/assets/files/TADAT%20Subnational%20Field%20Guide%20-%20November%202019.pdf>

Xavier, O., Pires, S., Marques, T., & Soares, A. (2022). Identificação de evasão fiscal utilizando dados abertos e inteligência artificial. *Revista de Administração Pública*, 56(3), 426-40. <https://doi.org/10.1590/0034-761220210256>

Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G., & Pineda, C. (2021). Identifying tax evasion in Mexico with tools from network science and machine learning. In O. M. Granados, & J. R. Nicolás-Carlock (Eds.), *Corruption networks: understanding complex systems*. Springer. https://doi.org/10.1007/978-3-030-81484-7_6

Gunther Siqueira Lemos Gomes

Mestre em Governança, Tecnologia e Inovação pela Universidade Católica de Brasília (UCB); Auditor Fiscal da Receita do Distrito Federal. E-mail: gunther.gomes@economia.df.gov.br

Remis Balaniuk

Doutor em Ciência da Computação pelo Institut National Polytechnique de Grenoble; Professor e pesquisador da Universidade Católica de Brasília (UCB). E-mail: remis@p.ucb.br

CONTRIBUIÇÃO DOS AUTORES

Gunther Siqueira Lemos Gomes: Conceituação (Suporte); Curadoria de dados (Liderança); Análise formal (Igual); Investigação (Igual); Metodologia (Igual); Administração de projeto (Suporte); Recursos (Liderança); Software (Liderança); Validação (Igual); Visualização (Liderança); Escrita - rascunho original (Liderança); Escrita - revisão e edição (Igual).

Remis Balaniuk: Conceituação (Liderança); Administração de projeto (Liderança); Supervisão (Liderança); Validação (Igual); Visualização (Suporte); Escrita - rascunho original (Suporte); Escrita - revisão e edição (Igual).

DISPONIBILIDADE DE DADOS

O conjunto de dados que dá suporte aos resultados deste estudo não está disponível publicamente por conter informações sensíveis e protegidas por sigilo fiscal.

FINANCIAMENTO

Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF).

Anexo

QUESTIONÁRIO PARA AUDITORES FISCAIS ESTADUAIS

Prezado colega do fisco estadual, muito obrigado por sua ajuda!

Este questionário é parte de uma pesquisa sobre o comportamento de empresas noteiras e a compreensão de quais seriam as variáveis preditoras com maior eficácia para identificar esse comportamento no universo das empresas do cadastro de cada estado. Por sua característica nacional, pedimos aos colegas auditores fiscais que trabalham diretamente com o combate à prática de noteiras nas secretarias de fazenda ou em órgãos equivalentes para responder às questões abaixo.

Um primeiro conjunto de questões visa verificar a situação do combate a noteiras em cada estado. Em seguida, apresentaremos temas que buscam identificar as variáveis indicadoras de comportamentos associados a fraudes fiscais cometidas por empresas noteiras. Cada sugestão de variável será descrita, e pedimos que o colega a avalie numa escala que vai de 0 (Irrelevante) a 5 (Muito relevante) para a detecção de noteiras.

Em campos livres, os colegas podem propor novas variáveis preditoras, que serão analisadas no estudo e podem vir a ser testadas contra a bases da Secretaria de Fazenda do DF.

IDENTIFICAÇÃO DO COLEGA AUDITOR FISCAL

- a. Nome:
- b. E-mail:
- c. Cargo efetivo:
- d. Tempo de serviço no cargo:

FISCO ESTADUAL

- e. Unidade federada:
- f. Possui COI ou órgão equivalente para combater as noteiras?
- g. Possui algum procedimento amparado em software para a detecção de noteiras?
 - i. Sim – amparado em inteligência artificial.
 - ii. Sim – amparado em painel de BI.
 - iii. Não.
- h. Usa denegação da emissão de documentos fiscais eletrônicos como forma de impedir a fraude com noteiras?

DO COMBATE A NOTEIRAS NO SEU ESTADO

- i. Empresas noteiras são um problema que atinge:
 - i. Apenas o fisco do emissor;
 - ii. Apenas o fisco do destinatário;
 - iii. Apenas o fisco do emissor e do destinatário;
 - iv. Toda a fiscalização tributária.

- j. Noteiras são um problema que necessita:
- i. De operações de natureza tributária;
 - ii. De operações de natureza policial;
 - iii. De oferecimento de denúncia por crime contra a ordem tributária;
 - iv. Nenhuma das alternativas anteriores;
 - v. Todas as alternativas anteriores.
- k. Qual é a importância da participação da Polícia Civil no combate a noteiras?
- i. De 0 a 5 (Não é importante – Muito importante).
- l. A Polícia Civil tem papel efetivo no combate a noteiras no seu estado?
- i. Sim ou não.
- m. Qual é a importância da participação do Ministério Público no combate a noteiras?
- i. De 0 a 5 (Não é importante – Muito importante).
- n. O Ministério Público tem papel efetivo no combate a noteiras no seu estado?
- i. Sim ou não.
- o. Qual é a importância da participação do Poder Judiciário no combate a noteiras?
- i. De 0 a 5 (Não é importante – Muito importante).
- p. O Poder Judiciário tem papel efetivo no combate a noteiras no seu estado?
- i. Sim ou não.

VARIÁVEIS PREDITORAS DO TIPO BINÁRIO (SIM OU NÃO)

Avalie cada sugestão de variável segundo sua experiência no combate a noteiras. Variáveis que você achar que não influenciam a detecção de noteiras devem ser marcadas com zero, e variáveis que têm alta influência, 5, sendo os valores intermediários indicativos de importância relativa.

1. A empresa comprou de uma empresa já classificada como noteira?
Nesse caso, a empresa em estudo, sendo ou não noteira, comprou de uma empresa sabidamente noteira.
2. A empresa já vendeu para empresas que compram de empresas já classificadas como noteiras?
Nesse caso, a empresa em estudo, sendo ou não noteira, vendeu para uma empresa que sabidamente compra de noteira.
3. O sócio/contador da empresa é sócio/contador de empresa classificada como noteira?
Nesse caso, a empresa em estudo, sendo ou não noteira, tem sócio ou contador que tem relação de responsabilidade contábil ou sociedade com empresa que já foi classificada como noteira.
4. A empresa já comprou de empresas que têm sócio/contador que tem empresas que já foram classificadas como noteiras?
Nesse caso, a empresa em estudo, sendo ou não noteira, tem sócio ou contador que comprou de empresa que tem sócio ou contador que tem empresas classificadas como noteiras, mas que a empresa de quem ela comprou não foi classificada como noteira.
5. O sócio tem trabalho com carteira assinada fora da unidade federada para qual tem uma empresa?
Esse caso avalia a possibilidade de o sócio ser laranja pelo fato de possuir trabalho formal em outro estado.

6. O sócio tem trabalho com carteira assinada com salário de até 2 mil reais?

Esse caso avalia a possibilidade de o sócio ser laranja pelo fato de possuir trabalho formal com salário baixo.

7. A empresa é do regime Simei, mas emitiu nota com destaque de ICMS?

8. O NCM preponderante na entrada é o mesmo NCM preponderante na saída?

9. Entregou EFD (se normal) ou PGDAS (se Simples Nacional)?

VARIÁVEIS PREDITORAS DO TIPO NUMÉRICO

1. Faturamento líquido.
2. Total de entradas/total de saídas.
3. Capital social/total de saídas.
4. Quantidade de documentos fiscais de saída.
5. Quantidade de clientes.
6. Quantidade de fornecedores.
7. Folha de pagamento/total de saídas.
8. Quantidade de empregados/total de saídas.
9. ICMS a recolher/total de saídas.

SUGESTÕES DE VARIÁVEIS

- Do tipo binário (campo texto – livre).
- Do tipo numérico (campo texto – livre).
- Outras sugestões (campo texto – livre).