



EIXO 4 – INTELIGÊNCIA ARTIFICIAL E TRANSFORMAÇÃO DIGITAL NA AUDITORIA PÚBLICA

APLICAÇÃO DE MACHINE LEARNING NA AUDITORIA FISCAL: UM MODELO PREDITIVO PARA DETECÇÃO DE EMPRESAS DE FACHADA EM UM CICLO DE AUDITORIA CONTÍNUA

Remis Balaniuk

Pós-doutor em Realidade Virtual pela Stanford University - EUA (2002), Doutor em Informática pelo Institut National Polytechnique de Grenoble - França (1996), Mestre em Ciências da Computação pela Universidade Federal do Rio Grande do Sul (1989), graduado em Processamento de Dados pela Universidade de Brasília (1986). Já atuou como pesquisador associado da Universidade de Stanford - EUA, pesquisador visitante da Universidade de Oxford, no Reino Unido e do Institut National de Recherche en Informatique et en Automatique, na França, e consultor do Programa das Nações Unidas para o Desenvolvimento - PNUD. Atualmente é professor e pesquisador da Universidade Católica de Brasília e do Instituto Serzedello Corrêa do Tribunal de Contas da União. Atuou por mais de vinte e cinco anos como cientista de dados na detecção de fraudes e na avaliação de políticas públicas utilizando técnicas de mineração de dados, Inteligência Artificial, Geotecnologias e Business Intelligence no Tribunal de Contas da União, onde foi Auditor Federal de Controle Externo. Na pesquisa atua principalmente na área de Ciência da Computação, com ênfase em métodos de aprendizagem de máquina, Inteligência Artificial Generativa, Redes Neurais Profundas, Sistemas Multiagentes e métodos de geoprocessamento.

Thiago Segati Silva

Mestrando em Governança, Tecnologia e Inovação pela Universidade Católica de Brasília (2025), Pós-graduação *latu sensu* em Gestão de Riscos e Cibersegurança pela Faculdade Focus (2022), graduado em Ciências Contábeis pela Universidade Estácio de Sá (2021), graduado em Ciência da Computação pela Universidade Católica de Goiás (2005). Ocupo o cargo de Auditor Fiscal da Receita do Distrito Federal há 4 anos na Secretaria de Economia do DF. Auto na Gerência de Documentos Fiscais Eletrônicos, diretamente na área de recepção, validação, armazenamento e distribuição desses documentos para público interno e externo. Atuei na Secretaria de Saúde do Estado de Goiás, como executor administrativo na área de orçamento e finanças por 9 anos.

RESUMO



Fraudes fiscais, especialmente as praticadas por “empresas noteiras” que emitem créditos indevidos de ICMS, representam um desafio crítico para a arrecadação tributária, a justiça fiscal e a concorrência leal no Brasil. Métodos tradicionais de fiscalização são predominantemente reativos, baseados em regras estáticas de Business Intelligence (BI), e demonstram baixa eficácia, com taxas de confirmação de fraude raramente superando 10-20% das empresas selecionadas, resultando em um alto custo de oportunidade na alocação de recursos públicos. Este estudo propõe e valida um protocolo inovador e adaptativo que integra Auditoria Contínua (AC) e Aprendizado de Máquina (Machine Learning – ML) para a detecção de empresas noteiras, aplicado no contexto da Secretaria de Economia do Distrito Federal (SEEC/DF). A metodologia abrange o mapeamento de processos, inventário e engenharia de atributos, seleção de variáveis via LASSO, e o treinamento de cinco modelos ensemble (Random Forest, Gradient Boosting, XGBoost, LightGBM e CatBoost) integrados por votação suave. Implementado em um ciclo contínuo com retroalimentação a partir de auditorias presenciais, o modelo alcançou uma precisão média de 77,1% nos ciclos iniciais de auditoria, superando amplamente o método tradicional. Os resultados demonstram ganhos expressivos de eficiência e assertividade na alocação de recursos, além de um elevado potencial de replicabilidade em outras administrações tributárias. A principal contribuição reside na proposição de um framework adaptativo que alia inteligência artificial e validação empírica, fortalecendo a governança fiscal e a capacidade de resposta contra fraudes estruturadas.

Palavras-chave: ICMS; empresas de fachada; auditoria contínua; aprendizado de máquina; administração tributária.

1 INTRODUÇÃO

A necessidade de financiar políticas públicas eficazes sem aumentar a carga tributária impõe ao Estado o imperativo de combater a sonegação fiscal. As fraudes fiscais envolvendo a emissão de documentos inidôneos por “empresas noteiras” representam um desafio significativo para as administrações tributárias, ao permitirem a transferência de direitos de crédito de ICMS indevidos e prejudicarem as receitas



públicas. Tais empresas, que frequentemente simulam operações comerciais sem estoque físico ou jurídico, são criadas para facilitar esquemas de sonegação e podem acobertar diversos ilícitos, incluindo lavagem de dinheiro, falsa exportação e ocultação de cargas roubadas.

Os métodos de auditoria fiscal convencionais são predominantemente periódicos e retrospectivos, e sua rigidez dificulta a identificação ágil e precisa desses esquemas fraudulentos, especialmente dado o volume e a velocidade das informações fiscais contemporâneas. Isso resulta em perdas substanciais de arrecadação e limita as ações preventivas.

O cenário internacional e a literatura em ascensão apontam para a integração de Auditoria Contínua (AC) e Aprendizado de Máquina (Machine Learning – ML) como um caminho promissor para superar essas limitações. A AC, ao permitir o monitoramento em tempo real e a análise automatizada de dados fiscais, possibilita a detecção rápida de anomalias e padrões suspeitos. O ML, por sua vez, oferece capacidade adaptativa para identificar padrões complexos e mutáveis de fraude, com modelos que se ajustam continuamente a novas tendências. Embora existam pesquisas relevantes nessa área, há uma lacuna de estudos que apliquem essas técnicas de forma sistemática e operacional em administrações tributárias brasileiras, especialmente com retroalimentação baseada em resultados reais de campo.

Este estudo se propõe a preencher essa lacuna, desenvolvendo e avaliando um protocolo inovador que integra AC e ML no contexto da Secretaria de Economia do Distrito Federal (SEEC/DF). O objetivo geral é aprimorar a capacidade dos órgãos de fiscalização tributária na detecção e resposta rápida a essas práticas ilícitas, aumentando a eficácia das medidas de controle e proteção dos recursos públicos.

2 METODOLOGIA

Este trabalho caracteriza-se como uma pesquisa aplicada, de abordagem quantitativa e natureza mista, combinando análise exploratória e desenvolvimento tecnológico com validação empírica. Foi conduzido como um estudo de caso na Secretaria de Estado de Economia do Distrito Federal (SEEC/DF), em parceria com o Núcleo de Gestão e Controle de Operações Especiais (NGCOE). O desenho metodológico foi



concebido para ser replicável e escalável, utilizando exclusivamente bases de dados efetivamente disponíveis à administração tributária.

A metodologia foi estruturada em seis etapas principais:

- **Mapeamento de Processos e Identificação de Oportunidades:** Realizado por meio de entrevistas estruturadas com auditores fiscais e análise de procedimentos internos da SEEC/DF. O fluxo tradicional de fiscalização foi descrito em quatro fases (planejamento, análise, execução, encerramento), predominantemente reativas e dependentes de regras estáticas. Identificaram-se cinco pontos estratégicos para a inserção da AC-ML, incluindo a seleção de contribuintes, análise cadastral, preparação para auditoria de campo, avaliação dos resultados das vistorias e acompanhamento pós-auditoria.
- **Inventário e Estruturação das Fontes de Dados:** Mapeamento detalhado das principais bases de dados acessadas pela fiscalização tributária. As fontes incluíram Cadastro Fiscal (CFDF), Declarações de Apuração (DAR/DF), Escrituração Fiscal Digital (EFD), Documentos Fiscais Eletrônicos (NF-e/NFC-e), Declaração de Meios de Pagamento (DIMP), NF3e (Nota Fiscal de Energia elétrica) e Dados Federais Compartilhados (RFB). Uma base consolidada de 1.218 empresas foi construída, com 21,76% classificadas como "noteiras" (classe "SIM") e 78,24% como "legítimas" (classe "NÃO"), evidenciando um desbalanceamento de classes.
- **Engenharia de Atributos:** Partindo de 94 atributos originais, foram criadas 38 novas variáveis (totalizando 132 atributos candidatos). Essas variáveis derivadas foram baseadas em conhecimento contábil e fiscal, categorizadas em agregações temporais (sufixos Atual e Antigo), monetárias (Mon), quantitativas (Int), razões (Rel), subtrações (Sub) e booleanas (Bol). O pré-processamento incluiu imputação de zeros para valores ausentes (semanticamente coerente para dados fiscais), *One-hot encoding* para variáveis categóricas e padronização com *StandardScaler*.
- **Seleção de Variáveis:** Implementou-se uma metodologia híbrida e multifásica. Primeiramente, a base foi dividida em treino (80%) e teste (20%) com um *random_state* fixo (25) para garantir estabilidade. A técnica SMOTE (*Synthetic*



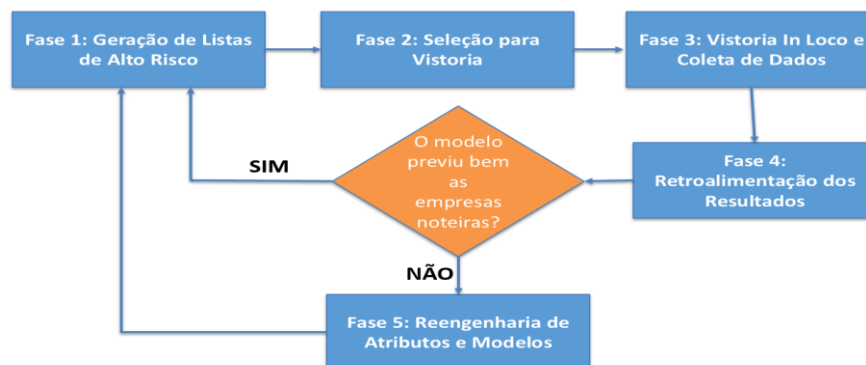
Minority Over-sampling Technique) foi aplicada exclusivamente ao conjunto de treino para balancear as classes. Em seguida, a Regressão LASSO foi utilizada como filtro inicial, reduzindo os 132 atributos para 39 variáveis com maior poder preditivo linear. A seleção final de características foi implicitamente realizada pelos modelos ensemble, que capturam interações não lineares complexas.

- **Treinamento e Avaliação de Modelos de ML:** Foi conduzido um experimento comparativo com um portfólio diversificado de algoritmos, incluindo modelos de linha de base (Regressão Logística, Árvore de Decisão) e modelos *ensemble* (*Random Forest*, *Gradient Boosting*, *XGBoost*, *LightGBM*, *CatBoost*). Todos foram submetidos a um protocolo rigoroso de validação cruzada estratificada k-fold (k=5) com otimização de hiperparâmetros via *RandomizedSearchCV*, utilizando o F1-score como métrica principal devido ao desbalanceamento de classes. O modelo final foi construído como um ensemble de votação suave (*soft voting*), combinando as previsões probabilísticas dos cinco modelos de melhor desempenho.

- **Integração em Ciclo Contínuo:** O elemento mais inovador é o protocolo de Auditoria Contínua Preditiva, concebido como um ciclo dinâmico de detecção, validação e aprimoramento, indicado na Figura 1. Este ciclo inclui as fases de: Geração de listas de alto risco (pelo modelo ML); Seleção para vistoria (refinada por auditores); Vistoria in loco e coleta de dados (com protocolo padronizado); Retroalimentação dos resultados (classificação final das empresas); e um Bloco Condicional que, se o desempenho do modelo se degradar, ativa um subprocesso de Reengenharia de atributos e modelos para retreinar e otimizar o sistema.



Figura 1 - O Ciclo Operacional da Auditoria Contínua Preditiva



Fonte: Elaborada pelos autores.

3 RESULTADOS E DISCUSSÃO

A fase inicial de implementação do protocolo foi avaliada em três ciclos operacionais consecutivos, totalizando 48 empresas auditadas presencialmente pelo NGCOE da SEEC/DF. Devido a um viés de verificação (apenas casos de alta probabilidade de fraude foram auditados), a métrica principal de avaliação foi a Precisão (Precision). Empresas com probabilidade superior a 0.60 de serem "noteiras" foram consideradas de alto risco para auditoria.

Os resultados de precisão constam na Tabela 1:

Tabela 1: Resultados por ciclo

Ciclo	Empresas auditadas	Empresas confirmadas (noteiras)	Precisão (%)	Falsos Positivos (%)
1	7	6	85,7	14,3
2	33	25	75,8	24,2
3	8	6	75	25
Total	48	37	77,1	22,9

Fonte: elaborada pelos autores.

A precisão consolidada para todo o período foi de 77,1% (37 acertos em 48 auditorias). Isso significa que, das empresas selecionadas pelo modelo de ML e auditadas,



aproximadamente 8 em cada 10 foram de fato confirmadas como fraudulentas. Esta taxa de acerto representa um ganho relativo superior a 280% em comparação com o método tradicional baseado em painel de BI, que confirmava apenas entre 10% e 20% das empresas indicadas como suspeitas. Esse avanço expressivo demonstra uma alocação consideravelmente mais eficiente dos recursos de auditoria.

É notável que, nesta fase inicial, a retroalimentação e o retreinamento do modelo não resultaram em um aumento progressivo da precisão ao longo dos três ciclos (de 85,7% para 75,0%). Isso pode ser atribuído ao volume relativamente pequeno de empresas auditadas por ciclo, à força dos sinais de fraude já capturados pelo modelo inicial, e à necessidade de mais dados rotulados ou de introdução de novos padrões para melhorias incrementais. No entanto, o desempenho se manteve consistentemente alto e significativamente superior ao método anterior, validando a eficácia da abordagem baseada em ML.

4 CONCLUSÃO

O presente estudo demonstrou a viabilidade e a eficácia de um protocolo inovador de auditoria contínua baseado em Machine Learning para a detecção de empresas de fachada que emitem créditos indevidos de ICMS. Os resultados iniciais atestam ganhos substanciais em eficiência e precisão, com uma taxa média de acerto superior a 77%, um aumento significativo em relação aos 10-20% do método tradicional.

A principal contribuição deste trabalho transcende a criação de um algoritmo de alta precisão. O que se propõe é um *framework* de governança de dados e auditoria que articula a inteligência artificial com a validação empírica dos auditores em um ciclo de aprendizagem contínua. Essa sistemática, ao conectar a geração de listas de risco, a vistoria em campo e a retroalimentação estruturada, representa uma inovação de processo para a administração tributária, transformando a fiscalização de uma atividade reativa para um ecossistema proativo e adaptativo.

Apesar dos avanços, o estudo identificou desafios centrais para o amadurecimento do protocolo, incluindo a necessidade de:



- Ampliar a equipe de auditores para verificar um volume maior de casos, incluindo amostras de negativos.
- Desenvolver mecanismos de Interpretabilidade do Modelo (XAI) para fornecer aos auditores justificativas claras sobre a classificação de risco, tornando a vistoria mais assertiva.
- Testar a automação da ação de denegação de documentos fiscais para previsões de altíssimo risco (ex: probabilidade > 90%), visando um ganho de escala na contenção de fraudes.

Como recomendações para trabalhos futuros, sugere-se: implementar a auditoria de uma amostra aleatória de empresas de baixo risco para obter métricas de avaliação mais completas (*recall*, *acurácia*, *F1-score*); avançar em ferramentas móveis para coleta de dados em campo e painéis de controle automatizados e promover a expansão institucional do protocolo para outras unidades da federação.

Em síntese, a experiência prática na SEEC/DF comprova que a auditoria contínua preditiva é tecnicamente viável e operacionalmente aplicável. A continuidade deste trabalho, com a incorporação das recomendações propostas, tem o potencial de fortalecer ainda mais o combate à fraude fiscal e a promoção da justiça fiscal.

REFERÊNCIAS BIBLIOGRÁFICAS

COSTA, T. W. F. ICMS e os procedimentos de combate às “empresas noteiras” no Distrito Federal nos anos de 2022 e 2023. 2024. Dissertação (Mestrado em Direito), UCB, 2024.

GOMES, G. S. L. Identificação de práticas de evasão fiscal utilizando Aprendizagem de Máquina: O caso das empresas de fachada e os créditos ilegais de ICMS. 2023. Dissertação (Mestrado em Governança, Tecnologia e Inovação) – UCB, 2023.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York: Springer, 2017.

HE, H.; GARCIA, E. A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, v. 21, n. 9, p. 1263–1284, 2009.

OKUSI, O.; IKEMEFUNA, C. D.; IWUH, A. C.; YUSUF, S. Adaptive fraud detection systems: Using machine learning to identify and respond to evolving financial threats. International Research Journal of Modernization in Engineering Technology and Science, v. 6, n. 9, 2024.