REVISTA Estratégias & Soluções

Artigo oriundo de Trabalho de Conclusão de Curso de Especialização em Data Science e Analytics do MBA USP/Esalq
Submetido: 12/09/2024
Aceito: 28/08/2025
E&S 2025, 5: e2024082
DOI: 10.22167/2675-6528-2024082

Otimização das ações de auditoria fiscal através do ranqueamento de empresas utilizando aprendizado de máquina

Giovana Amorim Zanato[1]*

Cláudio Tucci Junior[2]
https://orcid.org/0000-0002-1361-0639

*autora correspondente: giovanazanato@gmail.com

[1]Especialista em Data Science e Analytics. Auditora Fiscal da Receita Estadual. Secretaria de Econômia do Estado de Goiás. Avenida Vereador José Monteiro, 2233, Setor Vila Nova, 74653-900, Goiânia, Goias, Brasil.

[2]Doutor em Ciências Sociais. Professor orientador. Avenida Paulista, 1159, Conjuntos 612 e 613, Cerqueira César, São Paulo, SP, Brasil.

A auditoria fiscal é elemento central da arrecadação tributária e da conformidade no setor público, garante a aplicação correta das normas e exerce papel estratégico ao viabilizar políticas públicas por meio da receita arrecadada[1]. Com a transição do Imposto sobre Circulação de Mercadorias e Prestação de Serviços (ICMS) para o Imposto sobre Bens e Serviços (IBS)[2], prevista para 2026, cresce a relevância de métodos de ciência de dados e aprendizado de máquina como ferramentas para identificar riscos de evasão fiscal, proporcionam redução de custos, maior transparência e detecção de *outliers* ao comparar o comportamento de um contribuinte com o de sua população de referência[3].

Estudos anteriores sobre o uso de aprendizado de máquina em auditoria fiscal demonstraram sua eficácia na detecção de fraudes tributárias. Ruzgas e Kizauskiene[4] desenvolveram um estudo com vários modelos para detecção de crimes fiscais e mostraram que a técnica de mineração de dados identifica a evasão fiscal e extrai conhecimento oculto aplicável à redução das perdas de receita decorrentes da evasão.

Silva e Rigitano[5] realizaram um estudo de caso de fraude financeira com redes bayesianas em dados de contribuintes de São Paulo, Brasil. Wahab e Bakar[6] aplicaram um conjunto de algoritmos de aprendizagem automática para classificar o imposto sobre o rendimento no *Inland Revenue Board* da Malásia e verificaram a eficácia de diferentes algoritmos de machine learning na seleção de auditorias. Shi e Dong[7] estudaram um modelo de rede neural de gráfico heterogêneo para apoiar a detecção de evasão fiscal. Esses trabalhos

evidenciam o potencial do aprendizado de máquina para aprimorar a precisão e a eficiência das auditorias fiscais.

O relatório da Organização para Cooperação e Desenvolvimento Econômico (OCDE) (*Tax audits in a changing environment*) enfatiza que auditorias fiscais devem ser concebidas como instrumentos de compliance tributário e de monitoramento baseado em risco**,** capazes de identificar potenciais inconformidades de forma preventiva, reforçar o cumprimento voluntário das obrigações tributárias e fortalecer a legitimidade e a transparência do sistema tributário [8].

Nesse contexto, este estudo utiliza dados de propriedade da Secretaria da Economia do Estado de Goiás, com registros históricos de auditorias e informações declaradas pelos contribuintes para comparar diferentes técnicas de aprendizado de máquina com o objetivo de classificar empresas com maior risco de não conformidade fiscal em relação ao ICMS.

Portanto, este estudo tem como objetivo aplicar e comparar modelos de aprendizado de máquina para ranquear empresas, de modo a direcionar de forma mais eficaz as ações de auditoria fiscal, reduzir custos, aumenta a transparência e fortalecer o combate à concorrência desleal.

Esta pesquisa adota uma abordagem quantitativa, baseada em dados numéricos de declarações fiscais e registros de auditoria, submetidos a técnicas estatísticas e de aprendizado de máquina, e possui caráter descritivo ao analisar padrões de comportamento dos contribuintes e comparar o desempenho de diferentes modelos preditivos para priorizar as ações de auditoria fiscal[9].

Segundo Alpaydin[10], aprendizado de máquina é a área da inteligência artificial que desenvolve métodos para que computadores identifiquem padrões a partir de dados e realizem previsões sem programação explícita. Um modelo de aprendizado de máquina é a representação matemática construída a partir de dados, capaz de identificar padrões e gerar previsões ou classificações em novas situações[11].

A variável alvo, que representa o resultado que o modelo busca prever, é quantitativa e indica o valor financeiro, em reais, calculado a partir de auditorias fiscais realizadas pelos auditores fiscais da Secretaria de Economia do Estado de Goiás, entre 2019 e 2023, para 448 empresas varejistas de médio e grande porte. Essa variável, denominada TRIBUTAVEL_AUTO na base de dados utilizada, representa o valor de ICMS sonegado pelo contribuinte no exercício fiscal analisado (Quadro 1) que representa o valor que o modelo em questão pretende prever para aqueles contribuintes ainda não auditados.

Quadro 1. Dados do auto de infração[1]

Nome da variável	Descrição
Período do fato gerador autuado	Mês e ano referentes ao fato gerador autuado
Tributavel_auto	Valor nominal do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) autuado

Fonte: Dados originais da pesquisa.
Nota. [1]O auto de infração é o documento formal que instrui o processo administrativo tributário.

As variáveis explicativas servem de parâmetro para estimar o provável valor de ICMS sonegado. A seleção das variáveis explicativas considerou critérios teóricos e de relevância prática para a apuração do ICMS, privilegiando informações diretamente relacionadas ao comportamento fiscal dos contribuintes e às especificidades setoriais. Evitou-se incluir variáveis redundantes ou altamente correlacionadas, como o "ICMS recolhido", para reduzir multicolinearidade e preservar a interpretabilidade do modelo. Reconhece-se, contudo, a possibilidade de viés de seleção decorrente da própria base de dados, que reflete apenas contribuintes com EFDs completas e consistentes. Para mitigar esse risco, foram adotadas verificações de integridade e coerência das informações utilizadas.

A apuração do contribuinte, representada no registro E110 da Escrituração Fiscal Digital (EFD)[12,13], consolida toda a movimentação fiscal mensal do contribuinte, e conforme detalhado no Quadro 2, foram selecionadas variáveis constantes neste registro nos períodos auditados.

Quadro 2. Apuração mensal do Contribuinte informada na Escrituração Fiscal Digital (EFD)

Nome da variável	Campo na EFD	Descrição
Debito	VL_TOT_DEBITOS	Valor total dos débitos por saídas e prestações com débito do imposto
Ajuste_debito_doc	VL_AJ_DEBITOS	Valor total dos ajustes a débito decorrentes de documento fiscal
Ajuste_debito	VL_TOT_AJ_DEBITOS	Valor total dos ajustes a débito
Estorno_credito	VL_ESTORNO_CREDITOS	Valor total de estornos de créditos
Credito	VL_TOT_CREDITOS	Valor total dos créditos por entradas e aquisições com crédito do imposto
Ajuste_credito_doc	VL_TOT_AJ_CREDITOS	Valor total dos ajustes a crédito decorrentes de documento fiscal
Ajuste_credito	VL_AJ_CREDITOS	Valor total dos ajustes a crédito
Estorno_debito	VL_ESTORNO_DEBITO	Valor total dos estornos de débitos
Saldo_credor_ant	VL_SLD_CREDOR_ANT	Saldo credor do período anterior transferido ao período de referência
Saldo_apurado	VL_SLD_APURADO	Saldo devedor apurado
Deducao	VL_TOT_DED	Valor total das deduções
Icms_recolher	VL_ICMS_RECOLHER	Valor total do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) a ser pago aos cofres públicos
Saldo_credor_t	VL_SLD_CREDOR_TRANSPORTAR	Valor total do saldo credor que, quando existente, será transportado ao período seguinte

Extra_apur	DEB_ESP	Valores recolhidos ou a recolher, extra-apuração

Fonte: Dados originais da pesquisa.
Nota. Dados do registro E110, do bloco E, da Escrituração Fiscal Digital (EFD)[13].

A carga tributária do setor em que a empresa atua (Quadro 3), também compõe o rol de variáveis explicativas, uma vez que traz particularidades fiscais de cada setor varejista impactando de forma relevante o resultado do modelo.

Quadro 3. Carga tributária do setor

Nome da variável	Descrição
Ct_setor	Carga tributária do setor, calculada pela seleção das notas fiscais de operações de saída por tipo de atividade comercial e pela divisão do somatório do valor do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) pelo somatório dos valores nominais dos produtos

Fonte: Dados originais da pesquisa.
Nota: Foram utilizados dados de totalizadores de operações de saídas constantes no registro C190 da Escrituração Fiscal Digital (EFD)[13].

Todos os dados necessários para o trabalho estão armazenados nos bancos de dados da Secretaria de Economia do Estado de Goiás. A extração, a seleção e o saneamento dos dados foram realizados com o software SAS *Enterprise Guide* v.9.2 (*Statistical Analysis Sistema, Cary, NC, USA*), por meio da criação de fluxos para geração dos arquivos de saída em formato CSV. O desenvolvimento dos algoritmos ocorreu no programa R, com as bibliotecas *tidyverse, rgl, jtools, Rmisc, caret, neuralnet* aplicadas no processamento e análise dos dados.

Os dados trabalhados não apresentam identificação do contribuinte e os valores estão totalizados para impossibilitar sua identificação, em consonância com as normas específicas de sigilo fiscal. No caso em estudo, não foram incluídas empresas em conformidade fiscal devido à impossibilidade de determinação, uma vez que uma empresa distribuída para auditoria apresenta algum indício de irregularidade. Em contrapartida, não é possível afirmar que as empresas não distribuídas para auditoria estejam em plena regularidade fiscal quanto à apuração do ICMS.

Dentre as técnicas de aprendizado de máquina disponíveis, foram selecionados para estudo comparativo os modelos de regressão linear multivariada, *random forest* e redes neurais artificiais. O modelo que apresentar o melhor desempenho, conforme as métricas específicas de avaliação, será aplicado sobre a base dos demais contribuintes, com o intuito de estimar o valor potencial de ICMS omitido a partir das variáveis fornecidas e apoiar a priorização das ações de fiscalização.

O estudo foi iniciado com o desenvolvimento da técnica de regressão linear multivariada a fim de compreender o comportamento do fenômeno em questão a partir das variáveis explicativas. Com os dados históricos de resultados de auditorias fiscais, esperava-se que o

modelo fornecesse a estimativa de ICMS autuado, ou seja, o valor de ICMS omitido e não recolhido aos cofres públicos estaduais, conforme a apuração mensal e a carga tributária do setor do contribuinte analisado.

A normalização das variáveis e a execução inicial do modelo, por meio do comando lm (Figura 1), revelaram a presença de parâmetros sem significância estatística, o que indicou a necessidade de refinamento. Para incluir no modelo final apenas variáveis com influência estatisticamente significativa sobre o resultado, foi aplicado o procedimento *stepwise*[14], que avalia e remove gradualmente as variáveis sem contribuição relevante.

```
modelo_score <- lm(formula = TRIBUTAVEL_AUTO ~
                     CT_SETOR +
                     DEBITO +
                     AJUSTE_DEBITO +
                     ESTORNO_CREDITO +
                     CREDITO +
                     AJUSTE_CREDITO +
                     AJUSTE_CREDITO_DOC +
                     ESTORNO_DEBITO +
                     SALDO_CREDOR_ANT +
                     SALDO_APURADO +
                     DEDUCAO +
                     SALDO_CREDOR_T +
                     ICMS_RECOLHER +
                     EXTRA_APUR,
                   data = EMPRESAS_GEAV)
```

Figura 1. Algoritmo inicial de regressão linear multivariada, antes do procedimento de Box-Cox
Fonte: Dados originais da pesquisa.

Em seguida, aplicou-se a transformação de Box-Cox[15], técnica matemática que ajusta as variáveis para aproximar sua distribuição da normalidade, o que fortalece a validade e a confiabilidade do modelo. Os resultados são descritos no Quadro 4.

Quadro 4. Resultado da estimação do modelo após transformação de Box-Cox e procedimento stepwise

Nome da Variável	Estimado	Padrão	Erro	t-value	Pr(>|t|)
(Intercept)	0,05591	0,01297	4,312	2e-05	***
CT_SETOR	-0,05005	0,02213	-2,262	0,02419	*
DEBITO	-4,18415	1,86388	-2,245	0,02528	*
AJUSTE_DEBITO	-0,19378	0,08748	-2,215	0,02727	*
ESTORNO_CREDITO	-0,13765	0,05635	-2,443	0,01498	*
CREDITO	2,65803	1,13202	2,348	0,01932	*
AJUSTE_CREDITO	0,79261	0,29350	2,701	0,00719	**
ESTORNO_DEBITO	1,02801	0,07873	13,057	< 2e-16	***
SALDO_APURADO	0,94823	0,45529	2,083	0,03787	*
DEDUCAO	0,12780	0,05925	2,157	0,03155	*
EXTRA_APUR	-0,21454	0,06758	-3,175	0,00161	**

Fonte: Dados originais da pesquisa.
Nota. Erro residual padrão: 0,07244 em 434 graus de liberdade; R-quadrado: 0,437; R-quadrado ajustado: 0,424; estatística-F: 33,69 em 10 e 434 DF; p-value: < 2,2e-16; Pr(>t): níveis de significância: ***: entre 0 e 0,001; **: entre 0,001 e 0,01; *: entre 0,01 e 0,05.

Após gerar o modelo, foi aplicado o teste de Shapiro-Francia, teste simples e robusto para verificação de resíduos à normalidade. Conforme apresentado no Quadro 5, o modelo em estudo apresentou P-value = 2,2e-16, o que evidencia incompatibilidade com uma distribuição normal, e o modelo não se mostrou adequado para o estudo em questão[14].

Quadro 5. Teste de normalidade de Shapiro-Francia

dado: step_modelo_bc_SCORE$residuals
W = 0,51394, p-value < 2,2e-16

Fonte: Dados originais da pesquisa.

Após a análise inicial, constatou-se que o modelo de regressão linear não era adequado para o estudo, pois os dados apresentavam características de não-linearidade. Diante disso, foi selecionado o *random forest*, algoritmo de aprendizado de máquina reconhecido por sua robustez. Essa técnica combina o poder de múltiplas árvores de decisão e apresenta eficácia no tratamento de relações complexas e não-lineares nos dados. Além de aumentar a precisão do modelo, a abordagem minimiza o risco de *overfitting* (ajuste excessivo aos dados) e garante que as conclusões sejam mais confiáveis e generalizáveis para outros cenários[16].

A primeira etapa da aplicação da técnica, após o carregamento dos dados, é a separação da base em treino, validação e teste. A base de treino é usada para ajustar os parâmetros do modelo e para que ele aprenda padrões a partir dos dados. Em relação à base de validação, ela é utilizada durante o treinamento para definir hiperparâmetros, comparar modelos e evitar *overfiting*. E a base de teste é aplicada somente após o treinamento e a validação, fornece uma forma independente de medir o desempenho do modelo em dados inéditos, o que permite avaliar sua capacidade de generalização. No modelo em estudo, a base de treino compreendeu 60% dos dados, a base de validação 20% e a base de teste os 20% restantes.

Foram testadas várias combinações de variáveis e configurações de parâmetros. Os melhores resultados ocorreram quando a variável AJ_DEBITO_DOC foi desconsiderada, pois revelou ser irrelevante para o modelo. A configuração de parâmetro que apresentou maior R^2 (proporção da variabilidade da variável alvo explicada pelo modelo) na base de teste utilizou a base de validação incorporada à base de treino, com o propósito de ampliar o conjunto de treinamento.

O modelo foi treinado com 50 árvores (ntree = 50), conforme algoritmo representado na Figura 2.

```
rf <- randomForest::randomForest(
  TRIBUTAVEL_AUTO ~
    CT_SETOR +
    DEBITO +
    AJUSTE_DEBITO +
    ESTORNO_CREDITO +
    CREDITO +
    AJUSTE_CREDITO_DOC +
    SALDO_CREDOR_ANT +
    SALDO_APURADO +
    DEDUCAO +
    ICMS_RECOLHER +
    SALDO_CREDOR_T +
    EXTRA_APUR +
    AJUSTE_CREDITO +
    ESTORNO_DEBITO,
  data = treino_combinado,
  ntree = 50
)
```

Figura 2. Algoritmo de *random forest*
Fonte: Dados originais da pesquisa.

O R^2 obtido na base de teste foi de 61%, com erro quadrático médio (*Mean of Squared Error* - MSE) = 2,496516e+14 e percentual de variabilidade explicada de 15,12%, conforme apresentado nos Quadros 6 e 7.

Quadro 6. Avaliação do modelo com a utilização da técnica de *random forest*

Base	MSE[1]	R^{2}[2]
Treino	5,816345e+13	0,8022467
Teste	2,239706e+14	0,6133123

Fonte: Dados originais da pesquisa.
Nota: [1]MSE: *Mean Squared Error* - erro quadrático médio. [2]R^2: coeficiente de determinação.

Quadro 7. Características da variável criada com algoritmo de *random forest*

Tipo de *random forest*	**Regressão**
Número de árvores	50
Número de variáveis testadas a cada divisão	4
Média de resíduos quadrados	2,496516e+14
Percentual de explicabilidade da variável (%)	15,12

Fonte: Dados originais da pesquisa.

A análise das variáveis na base de teste, à luz do contexto específico do problema, indica que o R^2 de 61% pode ser considerado moderadamente bom, ou seja, 61% da variabilidade da variável TRIBUTAVEL_AUTO é explicada pelo modelo. Porém, ao observar-se um percentual de explicabilidade das variáveis de 15, 12 e um MSE de 2,239706e+14, conclui-se que o modelo não apresenta capacidade preditiva adequada e pode não constituir solução satisfatória para o problema. A partir dessa avaliação, identificou-se a necessidade de analisar o problema com a técnica de redes neurais artificiais.

As redes neurais artificiais foram inspiradas no funcionamento do cérebro humano. Esses sistemas computacionais são compostos por unidades interconectadas, denominadas neurônios artificiais, que processam informações e aprendem padrões complexos a partir dos dados. A estrutura das redes neurais consiste em camadas de neurônios, nas quais cada neurônio efetua operações matemáticas para transformar os dados de entrada em saídas úteis. A principal característica das redes neurais é a habilidade de aprendizagem, com o aprimoramento constante do desempenho do modelo, cujo treinamento ocorre por meio de um processo iterativo de ajustes dos pesos entre os nós, com técnicas como a descida do gradiente. [17]

O primeiro passo na aplicação do modelo de redes neurais no caso em estudo consistiu na normalização das variáveis, seguida da divisão da base em 70% para treino e 30% para teste. O próximo passo foi a execução do algoritmo de redes neurais, demonstrado na Figura 3. Foram testadas várias combinações de variáveis explicativas e hiperparâmetros. Assim como no modelo de *random forest*, os melhores resultados ocorreram quando a variável AJ_DEBITO_DOC foi desconsiderada, pois mostrou ser irrelevante para o modelo. A combinação de hiperparâmetros que apresentou melhor desempenho inclui quatro camadas intermediárias: a primeira com cinco neurônios, a segunda com quatro, a terceira com três e a quarta com dois, conforme arquitetura do modelo apresentada na Figura 4, e a utilização de uma função de ativação na saída.

```
nn <- neuralnet(
  TRIBUTAVEL_AUTO ~
    CT_SETOR +
    DEBITO +
    AJUSTE_DEBITO +
    ESTORNO_CREDITO +
    CREDITO +
    AJUSTE_CREDITO_DOC +
    AJUSTE_CREDITO +
    ESTORNO_DEBITO +
    SALDO_CREDOR_ANT +
    SALDO_APURADO +
    DEDUCAO +
    ICMS_RECOLHER +
    SALDO_CREDOR_T +
    EXTRA_APUR,
  data = treino,
  hidden=c(5,4,3,2),
  linear.output=T
)
```

Figura 3. Algoritmo de redes neurais
Fonte: Dados originais da pesquisa.

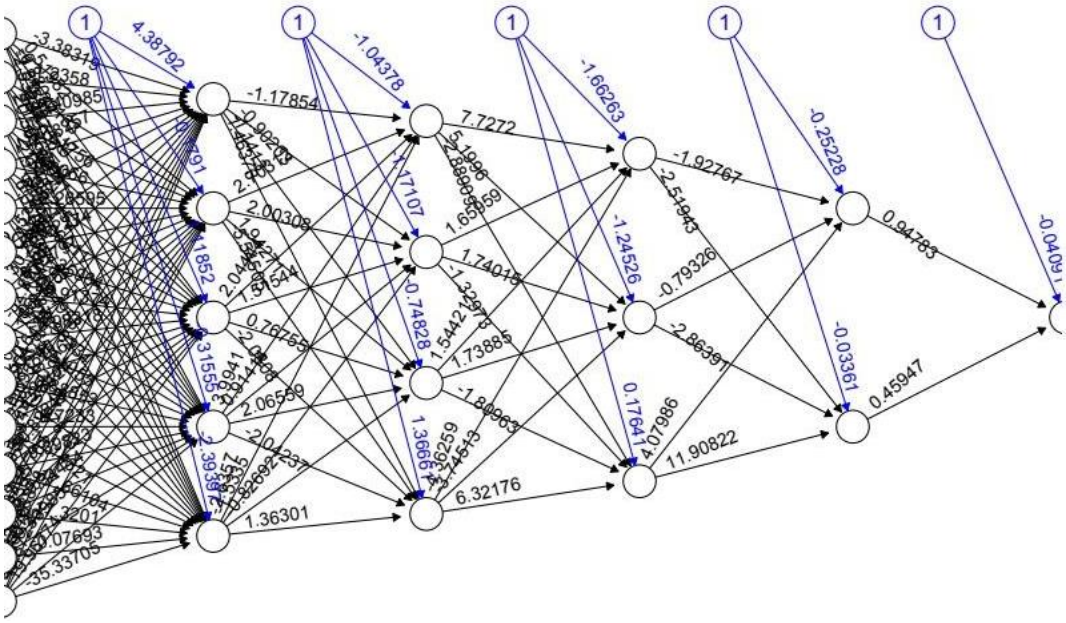

Figura 4. Representação gráfica da arquitetura do modelo de rede neural para o caso em estudo com as camadas, os nós e os respectivos pesos
Fonte: Dados originais da pesquisa.

O modelo apresentou um MSE de 0,007816 na base de teste, conforme apresentado no Quadro 8. Ele fornece uma medida de precisão geral do modelo de redes neurais em relação aos dados de teste. Quanto menor for o valor do MSE, mais próximas às previsões estão dos valores reais[11]. Como o MSE foi calculado sobre dados de teste independentes, o modelo demonstra boa capacidade de generalização, o que permite sua aplicação em outras bases de contribuintes varejistas e a identificação daqueles com potenciais indícios de evasão fiscal.

Quadro 8. Dados de avaliação do modelo com a utilização da técnica de redes neurais

Base	MSE[1]	R^2
Treino	0,002264842	0,704415
Teste	0,007816353	0,3610342

Fonte: Dados originais da pesquisa.
Nota: Nota: [1]MSE: *Mean Squared Error* - erro quadrático médio. R^2: coeficiente de determinação.

O gráfico apresentado na Figura 5 foi gerado a partir do modelo de redes neurais artificiais e mostra, em vermelho, os valores reais de omissão de ICMS, ou seja, os valores de ICMS autuado em ações fiscais, e, em azul, os valores de ICMS omissos previstos pelo modelo na base de teste.

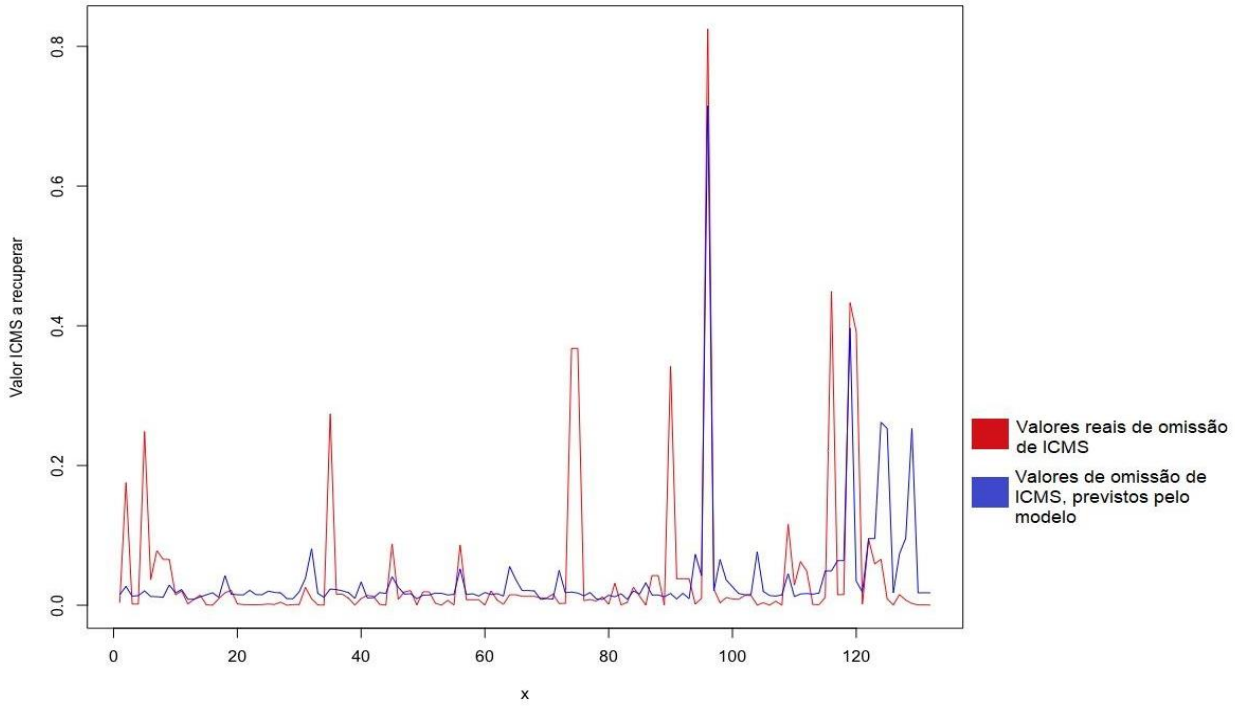

Figura 5. Representação gráfica do modelo de rede neural para o caso em estudo
Fonte: Dados originais da pesquisa.
Nota. Gráfico gerado no software R a partir do modelo gerado no mesmo software.

Os resultados deste estudo corroboram os achados de pesquisas anteriores, que indicam que modelos de redes neurais são eficazes na detecção de padrões complexos e não-lineares em dados fiscais[18]. Embora o modelo de *random forest* tenha apresentado um R^2 superior, o baixo MSE da rede neural destaca sua capacidade de fornecer previsões mais precisas.

A abordagem metodológica deste estudo — que combinou regressão linear multivariada, *random forest* e redes neurais — contribuiu não apenas para avaliar diferentes níveis de complexidade estatística, mas também para gerar interpretações práticas relevantes à gestão fiscal. A compreensão das relações entre as variáveis permitiu identificar padrões associados à probabilidade de sonegação, fornecendo subsídios para o direcionamento mais eficiente das auditorias e para o uso mais racional de recursos de fiscalização.

A comparação entre os modelos também evidenciou o potencial das técnicas de aprendizado de máquina em aumentar a precisão das previsões, o que pode resultar em redução de custos operacionais e impacto positivo na arrecadação, ao concentrar esforços em contribuintes com maior risco fiscal. Dessa forma, os resultados obtidos extrapolam o campo metodológico e oferecem ferramentas concretas de apoio à decisão na administração tributária.

A análise aprofundada do ICMS e a identificação de padrões de evasão fiscal oferecem subsídios importantes para a futura gestão do IBS, uma vez que ambos compartilham

características estruturais: incidem, em essência, sobre a circulação de mercadorias e serviços, possuem natureza de tributos indiretos, cujo ônus é transferido ao consumidor final, e adotam a não cumulatividade, com creditamento do imposto pago nas etapas anteriores da cadeia. O IBS apresenta, entretanto, uma vantagem significativa em relação à amplitude da base de incidência, que abrangerá não apenas as operações sujeitas atualmente ao ICMS, mas também outros serviços, locações e atividades financeiras[2].

Ademais, sua regulamentação será mais uniforme, diferentemente do ICMS, cuja disciplina normativa é fragmentada entre os estados. Nesse cenário, os conjuntos de dados disponíveis para análises fiscais tendem a ser substancialmente maiores e mais consistentes. Esse fator favorece a aplicação de modelos de aprendizado de máquina com menor suscetibilidade a ruídos e maior capacidade de identificação robusta de relações complexas .

O estudo cumpriu seu objetivo de investigar como modelos de inteligência artificial, especificamente redes neurais, podem ser aplicados para identificar padrões de evasão fiscal e fornecer subsídios à gestão tributária. Apesar de algumas limitações, como a restrição a 15 variáveis específicas e a dependência de dados de setores selecionados, que no âmbito estadual, especificamente no Estado de Goiás, constituem uma população restrita para o estudo e que pode inclusive ser um limitador à generalização dos resultados, foi possível desenvolver um modelo com alto nível de previsibilidade para os valores observados.

Ao identificar padrões complexos de evasão, o modelo proposto pode ser adaptado ao novo sistema tributário brasileiro, de modo a auxiliar no desenvolvimento de mecanismos mais robustos de auditoria e conformidade. Pesquisas futuras podem explorar bases de dados mais amplas, incluir variáveis adicionais e avaliar a aplicação do modelo em diferentes setores, com o intuito de aprimorar a precisão e a eficácia das estratégias de fiscalização no contexto da reforma tributária.

COMO CITAR

Zanato, G.A.; Tucci Jr, C. Otimização das ações de auditoria fiscal através do ranqueamento de empresas utilizando aprendizado de máquina. Revista E&S. 2025; 6: e2024082.

REFERÊNCIAS

[1] Santos, C. 2015. Auditoria Fiscal e Tributária. 3ed. São Paulo, SP: IOB.

[2] Brasil. 1988. Constituição da República Federativa do Brasil. Brasília, DF: emenda constitucional n. 132. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm . Acesso em: 15 set. 2025

[3] Alexopoulos, A.; Kotsogiannis, C.; Dellaportas, P.; Olhede, S. C.; Gyoshev, S.; Pavkov, T. 2025. A network approach to detect Value Added Tax fraud. Department of Economics, AUEB, Greece. Disponível em: https://arxiv.org/pdf/2106.14005 . Acesso em: 15 set. 2025.

[4] Ruzgas, T.; Kižauskienė, L.; Lukauskas, M.; Sinkevičius, E.; Frolovaitė, M.; Arnastauskaite, J. 2023. Tax Fraud Reduction Using Analytics in an East European Country. MDPI (Multidisciplinary Digital Publishing Institute). Disponível em: https://www.mdpi.com/2075-1680/12/3/288 . Acesso em: 15 set. 2025.

[5] Silva, L. S.; Rigitano, H. C.; Carvalho, R. N.; Souza, J. C. F. 2016. Bayesian Networks on Income Tax Audit Selection - A Case Study of Brazilian Tax Administration. Brasília, DF. Disponível em: https://ceur-ws.org/Vol-1663/bmaw2016_paper_3.pdf . Acesso em: 15 set. 2025.

[6] Wahab, R. A. S. R.; Bakar, A. A. 2021. Digital Economy Tax Compliance Model in Malaysia using Machine Learning Approach. Disponível em: https://www.ukm.my/jsm/pdf_files/SM-PDF-50-7-2021/20.pdf . Acesso em: 15 set. 2025.

[7] Shi, B.; Dong, B. 2023. An edge feature aware heterogeneous graph neural network model to tax evasion detection. Artigo. Expert System with Applications: An International Journal. Disponível em: https://dl.acm.org/doi/10.1016/j.eswa.2022.118903 . Acesso em: 15 set. 2025.

[8] Organização para a Cooperação e Desenvolvimento Econômico (OECD). 2017.. Tax Changing Tax Compliance Enviroment and the Role of Audit. 2017 OCDE – Organização para a Cooperação e Desenvolvimento Econômico. Paris. Disponível em: https://www.oecd.org/en/publications/the-changing-tax-compliance-environment-and-the-role-of-audit_9789264282186-en.html . Acesso em: 15 set. 2025.

[9] Creswell, J. W. 2022 Research design: qualitative, quantitative, and mixed methods approaches. 6ed. Thousand Oaks, CA, EUA: Sage.

[10] Alpaydin, E. 2020. Introduction to Machine Learning. 4ed. Cambridge, MA, USA: The MIT Press.

[11] Hastie, T.; Tibshirani, R.; Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2ed. New York: Springer.

[12] Brasil. 2007. Decreto n. 6.022, de 22 de janeiro de 2007. Institui o Sistema Público de Escrituração Digital - Sped. Receita Federal do Brasil, Brasília, DF. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/Decreto/D6022.htm. Acesso em: 15 set. 2025.

[13] Sistema Público de Escrituração Digital.2023. Guia prático da escrituração fiscal digital – EFD ICMS/IPI. v.3.1.5. Disponível em: http://sped.rfb.gov.br/arquivo/show/7273. Acesso em: 15 set. 2025.

[14] Fávero, L. P.; Belfiore, P. 2022. Análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata. Rio de Janeiro: LTC.

[15] Box, G. E. P.; Cox, D. R. 1964. An Analysis of Transformations. Journal of the Royal Statistical Society. Series B. 26(2). 211-252.

[16] Angshuman, P.; Dipti, P. M.; Prasun, D.; Abhinandan, G.; Appa, R. C.; Saurabh, K. 2018. *Random Forest* aprimorado para classificação. IEEE Transactions on Image Processing, 27. Disponível em: https://ieeexplore.ieee.org/document/8357563 . Acesso em: 15 set. 2025.

[17] Hardesty, L. 2017. Explained: Neural networks: Ballyhooed artificial-inteligence technique known as “deep learming” revives 70-years-old idea. MIT News. Disponível em: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414. Acesso em: 15 set. 2025.

[18] Pérez López, C.; Delgado Rodríguez, M.J.; Lucas Santos, S. 2019. Tax Fraud Detection through Neural Networks: An Application Using a Sample of Personal Income Taxpayers. Future Internet. 11(4), p. 86. doi:10.3390/fi11040086. Disponível em: https://www.mdpi.com/1999-5903/11/4/86 . Acesso em: 15 set. 2025.