FGV EBAPE

RAP
REVISTA DE ADMINISTRAÇÃO PÚBLICA

Fórum: Perspectivas Práticas

Identificação de evasão fiscal utilizando dados abertos e inteligência artificial

Otávio Calaça Xavier [1]
Sandrerley Ramos Pires [2]
Thyago Carvalho Marques [2]
Anderson da Silva Soares [3]

[1] Instituto Federal de Educação, Ciência e Tecnologia de Goiás/ Departamento de Áreas Acadêmicas IV, Goiânia / GO – Brasil
[2] Universidade Federal de Goiás / Escola de Engenharia Elétrica, Mecânica e de Computação, Goiânia / GO – Brasil
[3] Universidade Federal de Goiás / Instituto de Informática, Goiânia / GO – Brasil

A evasão fiscal é a consequência da prática da sonegação. Apenas no Brasil, estima-se que ela corresponda a 8% do PIB. Com isso, os governos necessitam de sistemas inteligentes para apoiar os auditores fiscais na identificação de sonegadores. Tais sistemas dependem de dados sensíveis dos contribuintes para o reconhecimento dos padrões, que são protegidos por lei. Com isso, o presente trabalho apresenta uma solução inteligente, capaz de identificar os perfis de potenciais sonegadores com o uso apenas de dados abertos, públicos, disponibilizados pela Receita Federal e pelo Conselho Administrativo Tributário do Estado de Goiás, entre outros cadastros públicos. Foram gerados três modelos que utilizaram os recursos Random Forest, Redes Neurais e Grafos. Em validação depois de melhorias finas, foi possível obter acurácia superior a 98% na predição do perfil inadimplente. Por fim, criou-se uma solução de software visual para uso e validação pelos auditores fiscais do estado de Goiás.
Palavras-chave: evasão fiscal; redes neurais; inteligência artificial; dados abertos; auditoria fiscal.

Identificación de la evasión fiscal mediante datos abiertos e inteligencia artificial

La evasión fiscal es la consecuencia de la práctica de la defraudación tributaria. En Brasil, se estima que corresponde al 8% del PIB. Por lo tanto, los gobiernos necesitan y utilizan sistemas inteligentes para ayudar a los agentes de hacienda a identificar a los defraudadores fiscales. Dichos sistemas se basan en datos confidenciales de los contribuyentes para el reconocimiento de patrones, que están protegidos por ley. Este trabajo presenta una solución inteligente, capaz de identificar perfiles de potenciales defraudadores fiscales, utilizando únicamente datos públicos abiertos, puestos a disposición por la Hacienda Federal y por el Consejo Administrativo Tributario del Estado de Goiás, entre otros registros públicos. Se generaron tres modelos utilizando *random forest* y *neural networks*. En la validación después de finas mejoras, fue posible obtener una precisión superior al 98% en la predicción del perfil moroso. Finalmente, se creó una solución de software visual para uso y validación por parte de los auditores fiscales del estado de Goiás.
Palabras clave: evasión de impuestos; redes neuronales; inteligencia artificial; datos abiertos; auditoría de impuestos.

Tax evasion identification using open data and artificial intelligence

Tax evasion is the practice of the non-payment of taxes. In Brazil alone, it is estimated as 8% of GDP. Thus, governments must use intelligent systems to support tax auditors to identify tax evaders. Such systems seek to recognize patterns and rely on sensitive taxpayer data that is protected by law and difficult to access. This research presents a smart solution, capable of identifying the profile of potential tax evaders, using only open and public data, made available by the Brazilian internal revenue service, the administrative council of tax appeals of the State of Goiás, and other public sources. Three models were generated using Random Forest, Neural Networks, and Graphs. The validation after fine improvements offered an accuracy greater than 98% in predicting tax evading companies. Finally, a web-based solution was created to be used and validated by tax auditors of the State of Goiás.
Keywords: tax evasion; neural networks; artificial intelligence; open data; tax auditing.

DOI: http://dx.doi.org/10.1590/0034-761220210256
Artigo recebido em 17 jul. 2021 e aceito em 06 abr. 2022.
ISSN: 1982-3134

AGRADECIMENTOS

Agradecimentos a Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) e ao Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG) por financiarem esta pesquisa.

1. INTRODUÇÃO

A evasão fiscal é um tópico importante para os governos de todo o mundo (Bethencourt & Kunze, 2019). Segundo o Sindicato Nacional dos Procuradores da Fazenda Nacional (Sinprofaz, 2019), no Brasil, estima-se que a arrecadação tributária poderia ser 23,1% maior sem as evasões, quase 8% do PIB do país. Se não houvesse evasão, a carga tributária poderia ser reduzida em quase 30%, sem queda de arrecadação. Em Goiás, Brasil, a situação não é diferente. Diversos trabalhos como Vieira (2015) e Esteves (2013) mostram uma situação semelhante, tanto na evasão fiscal quanto na recuperação da receita.

Ferramentas computacionais inteligentes podem aumentar a produtividade dos órgãos fiscalizadores tanto na recuperação de receita quanto na atuação preventiva, antecipando-se às ações de sonegação (Nasution, Salim, & Budhiarti, 2020).

Um problema da área de ciência de dados é a obtenção de dados para a criação de modelo de aprendizado de máquina (Nasution et al., 2020). Muitos dados usados para a identificação da evasão fiscal não são facilmente cedidos. São informações sigilosas das empresas e de seus sócios, protegidas pela lei brasileira (Matos, Macedo, & Monteiro, 2015). Uma alternativa ainda pouco explorada é a utilização de dados abertos, ou seja, aqueles que devem estar disponíveis gratuitamente para todos usarem e republicar como desejarem (Auer et al., 2007). A Receita Federal, por exemplo, divulga as informações cadastrais das empresas do Brasil na forma de dados abertos.

Este estudo tem como objetivo a utilização de dados abertos de várias fontes, para a construção de modelos de aprendizagem de máquina para a detecção de evasão fiscal. Nenhum dado não público foi utilizado neste texto, o que torna a abordagem genérica e aplicável a qualquer unidade da federação.

Uma das principais contribuições deste estudo é a utilização de dados abertos no problema da identificação da evasão fiscal, uma vez que a maioria dos trabalhos necessita de dados sensíveis. Outra contribuição é a utilização das redes neurais para grafos, ou GNNs (Zhang, Song, Huang, Swami, & Chawla, 2019), no problema da evasão fiscal em comparação com abordagens clássicas de aprendizagem de máquina, como o algoritmo Random Forest (Ho, 1995) e Redes Neurais Multicamadas (Haykin, 2007).

O trabalho focou na separação de empresas com perfil de potencial sonegadora das que não apresentam esse perfil. Como rótulo do viés "sonegador", utilizou-se a presença da empresa na dívida ativa do estado. Para a rotulação do viés "não sonegador", foram criados filtros com base na experiência dos auditores fiscais.

O artigo é organizado em introdução; seção 2, que apresenta um referencial teórico e alguns trabalhos correlatos; seção 3, que mostra a abordagem proposta; seção 4, que apresenta os resultados obtidos e uma análise deles; e, finalmente, seção 5, em que consta a conclusão do trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta alguns conceitos importantes para melhor leitura do trabalho, bem como alguns trabalhos correlatos.

2.1 Evasão fiscal e inadimplência tributária

Para Carrazza (2020), a evasão fiscal é o ato praticado por aquele que adota uma conduta ilegal com a intenção de não pagar ou reduzir tributos devidos ou, ainda, adiar seu recolhimento. Ela está definida legalmente no Brasil pela lei nº 4.729, de 14 de julho de 1965.

Inadimplência tributária é a falta de pagamento de taxas, impostos e/ou contribuições. Ela por si só não caracteriza a sonegação, uma vez que a inadimplência por não conhecimento das regras tributárias não visa fraudar a fiscalização tributária (Projeto de Lei 6520/2019). Todavia, o Supremo Tribunal Federal (STF) julgou como crime a inadimplência tributária (para o ICMS), conforme Informativo STF nº 963/2019 (Supremo Tribunal Federal, 2019).

Como a diferenciação entre inadimplência e sonegação ainda está em fase de amadurecimento no Brasil, o trabalho foca a identificação de empresas inadimplentes com perfil de má pagadora de impostos, conceito que já está bem definido.

Várias técnicas para a detecção de fraude financeira, sonegação e evasão fiscal já foram propostas na literatura. Existem pesquisas abrangentes (*surveys*) que descrevem várias delas (Barman et al., 2016; West & Bhattacharya, 2016). As pesquisas nessa área podem ser divididas em métodos tradicionais de auditoria fiscal, métodos de auditoria baseados em aprendizagem de máquina e métodos baseados em grafos (Ruan, Yan, Dong, Zheng, & Qiana, 2019).

Seleção manual de casos, seleção baseada em denúncias e seleções utilizando ferramentas computacionais são três métodos tradicionais frequentemente utilizados. A seleção manual de casos ou baseada em denúncias é demorada e requer experiência especializada do auditor fiscal (Ruan et al., 2019). Segundo o Sinprofaz (2019), a mão de obra é cara e inferior ao necessário.

Matos et al. (2015) e Wu, Ou, Lin, Chang, e Yen (2012) utilizaram regras de associação para a triagem de declarações fiscais e identificação de padrões de fraudes frequentes. Já Assylbekov et al. (2016) e Liu, Pan, e Chen (2010) adotaram a clusterização para inspeção fiscal e identificação de anomalias que podem levar a fraude fiscal. Noguera, Quesada, Tapia, e Llàcer (2014) abordaram um modelo com base em agentes para uma simulação de cumprimento das obrigações fiscais que combinava mecanismos de influência social com escolhas racionais.

Há também pesquisas que utilizaram grafos para a identificação da evasão fiscal, como fizeram Beutel, Akoglu, e Faloutsos (2015) e Dreżewski, Sepielak, e Filipkowski (2015), que usaram dados de redes sociais e bancários para a identificação de fraudes realizadas de maneira direta e indireta. Por fim, Ruan et al. (2019) e Zha et al. (2019) propõem a utilização de redes neurais para grafos para cenários mais complexos.

Todos os trabalhos mencionados utilizam dados sigilosos. O presente estudo se assemelha aos trabalhos mais recentes de Ruan et al. (2019) e Zha et al. (2019), tendo como diferencial a utilização de dados abertos.

2.2 Dados abertos

De acordo com Auer et al. (2007), dados abertos devem estar disponíveis gratuitamente para todos usarem e republicarem, sem restrições de direitos autorais, patentes ou outros mecanismos de controle. Há ainda a definição dada pelo Open Definition[1], que diz que um conjunto de dados é dito aberto quando pode ser acessado, utilizado, compartilhado e replicado por qualquer pessoa.

Muitos governos utilizam o conceito e a tecnologia dos dados abertos para disponibilizar dados de acesso público. No Brasil, apenas o Portal dados.gov.br oferece mais de 10 mil tipos de dados abertos. Há ainda portais de dados abertos de estados e municípios.

Os dados abertos são um instrumento para o exercício da cidadania (Ribeiro & Almeida, 2011), bem como fonte para o desenvolvimento dos mais diversos propósitos de pesquisa, notadamente na área de ciência de dados, como apresentado por Prado (2020).

Quando se tem muitos dados e busca-se a obtenção de padrões não triviais, mecanismos computacionais de aprendizagem de máquina podem ser utilizados.

2.3 Aprendizagem de máquina

Aprendizagem de máquina, considerada parte da inteligência artificial, é o estudo e a aplicação de algoritmos computacionais que se aprimoram automaticamente por meio da experiência e pelo uso de dados (Mitchell, 1997). Os algoritmos de aprendizagem de máquina constroem modelos inteligentes baseados em amostras de dados para realizar tarefas de predição ou apoio a decisão em novos dados.

Este estudo utiliza a abordagem de aprendizagem de máquina denominada Aprendizado Supervisionado, que aprende com um conjunto de pares de entrada-saída desejada. A entrada é composta por um conjunto de informações a respeito de uma entidade e a saída é a classificação desejada para ela. O treinamento consiste em mapear essa relação entrada-saída, permitindo que o modelo preveja a saída de novas entidades que não compunham o conjunto de treinamento.

Trabalhos como de Ruan et al. (2019) e Zha et al. (2019) utilizam o Aprendizado Supervisionado no processo de geração de modelos de predição.

Técnicas como Random Forest (Ho, 1995) e Redes Neurais Multicamadas (Haykin, 2007), utilizadas neste estudo, são comumente aplicadas em aprendizagem supervisionada com bons resultados. Por outro lado, regras de associação e clusterização, citadas na seção 2.1, são abordagens não supervisionadas.

O trabalho modela a identificação da evasão fiscal como um problema de aprendizagem supervisionada, do tipo classificação, uma vez que se pretende classificar empresas entre idôneas e inadimplentes.

3. ABORDAGEM PROPOSTA

A abordagem proposta foi modelada como uma classificação binária entre os potenciais perfis "inadimplente" e "idôneo". As atividades de construção dos *datasets* e a análise dos dados seguiram o modelo de processo para mineração de dados chamado CRISP-DM (Wirth & Hipp, 2000). As etapas do processo são apresentadas na Figura 1. As seções 1, 2 e 3 deste artigo correspondem à primeira fase (Compreensão do Negócio).

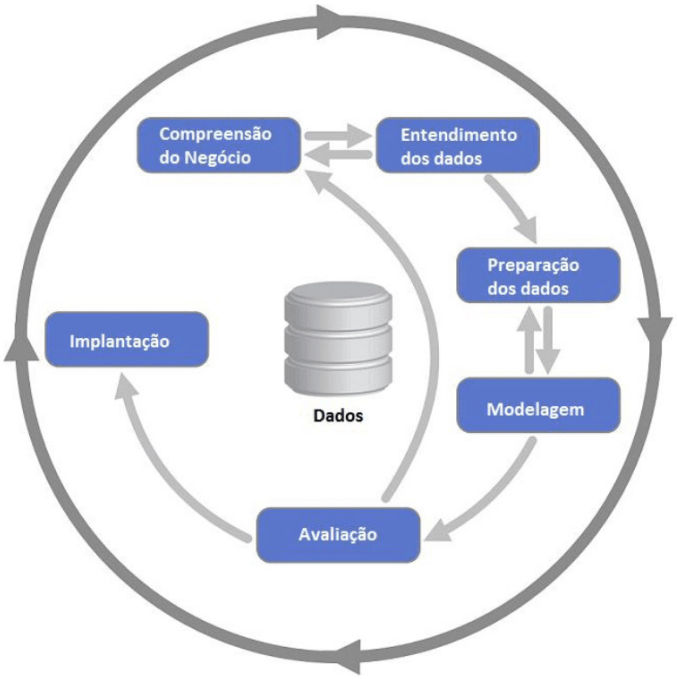
[1] Open Definition: organização cujo objetivo é promover definições do termo Open. Recuperado de http://opendefinition.org/

3.1 Entendimento dos dados

Os dados originam-se dos seguintes cadastros públicos:

- **Dados públicos CNPJ**[2] – relação de empresas da Receita Federal que contém os dados cadastrais de empresas e a identificação de seus sócios.

FIGURA 1 FASES DO MODELO DE PROCESSO CRISP-DM



Fonte: Adaptada de CRISP-DM (Wirth & Hipp, 2000).

- **Sintegra**[3] – Sistema Integrado de Informações sobre Operações Interestaduais com Mercadorias e Serviços. Possui dados das empresas que praticam operações em outros estados da federação.
- **Cadastro do Conselho Administrativo Tributário do Estado de Goiás**[4] – contém os processos das empresas que foram autuadas e julgadas na esfera administrativa do estado de Goiás.

Ao todo, as massas de dados utilizados contêm 1,6 milhão de empresas do estado de Goiás. Para a composição do perfil "inadimplente", foram consideradas as empresas inscritas na Dívida Ativa do Estado de Goiás, em um total de 193.987 empresas. Para a composição do perfil "idôneo", foram criados filtros com base na experiência dos auditores fiscais e foi possível extrair 617.622 empresas. Nota-se que as classes estão desbalanceadas. A classe idônea é 3,2 vezes maior que a inadimplente.

[2] Dados públicos CNPJ. Recuperado de https://cutt.ly/Mcn6RM0
[3] Sintegra. Recuperado de http://www.sintegra.gov.br/
[4] CAT Goiás. Recuperado de https://cutt.ly/Vcn617O

3.1.1 Balanceamento dos dados

Para igualar a quantidade de empresas em cada um dos perfis, foi extraída uma amostra aleatória do perfil "idôneo", que contém a mesma quantidade de empresas do perfil "inadimplente". Observou-se uma heterogeneidade entre os tipos de empresa existentes no cadastro, assim, o balanceamento dos dados deve considerar uma segmentação de grupos específicos de empresas. Neste trabalho foram utilizados a data de abertura e a categoria do CNPJ (se está inscrito como MEI).

3.2 Preparação dos dados

As seguintes atividades foram realizadas nesta etapa:

- Eliminação das variáveis alfabéticas não categóricas, como nome fantasia, razão social, etc.
- Formatação e simplificação de dados categóricos. Dados como se é optante pelo Simples Nacional ou MEI possuem mais de um valor com o mesmo significado.
- Binarização de variáveis categóricas, para a não correlação de grandeza em dados categóricos.
- Remoção dos campos nulos. Amostras com mais que 50% de nulos foram removidas.

Nessa etapa, foi observado que vínculos societários podem indicar o perfil sonegador. Uma empresa pode ter como sócio outra empresa com perfil sonegador ou uma pessoa física que é figura no quadro societário de uma empresa sonegadora. Há ainda o vínculo entre a empresa matriz e suas filiais, que pode ser obtido pelos números de CNPJ das empresas.

Para separação da massa de validação, foi utilizada uma data limiar, o dia 1º de janeiro de 2020. Assim, para treino e teste, foram consideradas apenas empresas abertas até essa data, totalizando 375.062 empresas, divididas igualmente entre os dois perfis. Foram utilizados 80% (300.050) dos dados para treinamento e 20% (75.012) para teste.

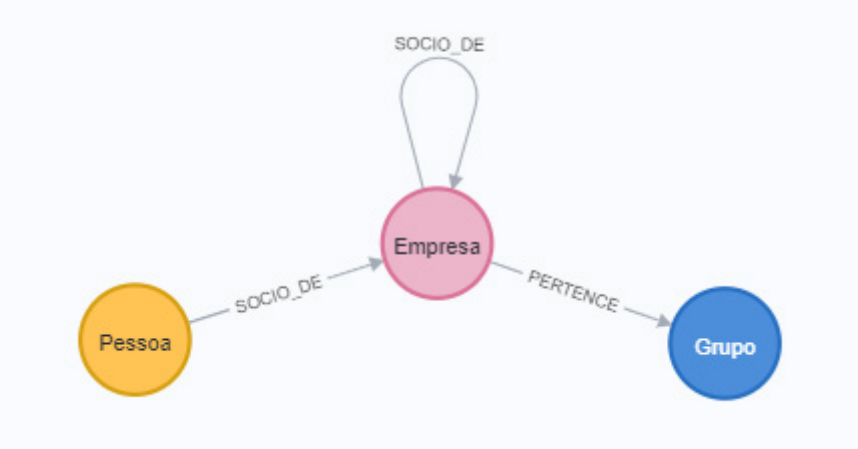
3.3 Modelagem

Os diversos relacionamentos que uma empresa pode ter com diferentes empresas e pessoas (quadro societário) introduzem complexidade na representação tabular dos dados. Por exemplo, dois CNPJs de empresas de um mesmo grupo não são numericamente próximos. Uma solução para esse tipo de problema é a utilização de um modelo de redes neurais baseadas em grafos, as GNNs, como descrito por Kipt et al. (2016). Nelas os dados de entrada são representados por grafos, os quais não pertencem a um espaço euclidiano. Neste trabalho, foi adotado um tipo específico de GNN chamado R-GCN (Zhang et al., 2019).

Para comparação, também foram construídos modelos baseados em Redes Neurais Multicamadas, que possuem como ideia central a representação matemática de neurônios interligados em várias camadas, com o objetivo de simular uma rede neural natural (Haykin, 2007), e Random Forest, que é um método de aprendizagem que opera por meio da construção de várias árvores de decisão no momento do treinamento. Árvores de decisão são técnicas que usam um modelo de decisão em formato de árvore, em que a montagem da estrutura da árvore mapeia a capacidade de o algoritmo classificar eventos (Ho, 1995).

A Figura 2 mostra a estrutura do grafo criado para a realização do experimento. Cada nó na figura representa um tipo de nó no grafo de entrada da R-GCN. Os nós do tipo Empresa são os únicos que contêm dados detalhados, e os outros tipos de nó se limitam ao identificador ou ao nome do item. Ainda assim, o grafo consegue representar, mediante os relacionamentos, informações que não são facilmente representáveis por intermédio de dados tabulares.

FIGURA 2 GRAFO HETEROGÊNEO DE ENTRADA PARA R-GCN



Fonte: Elaborada pelos autores.

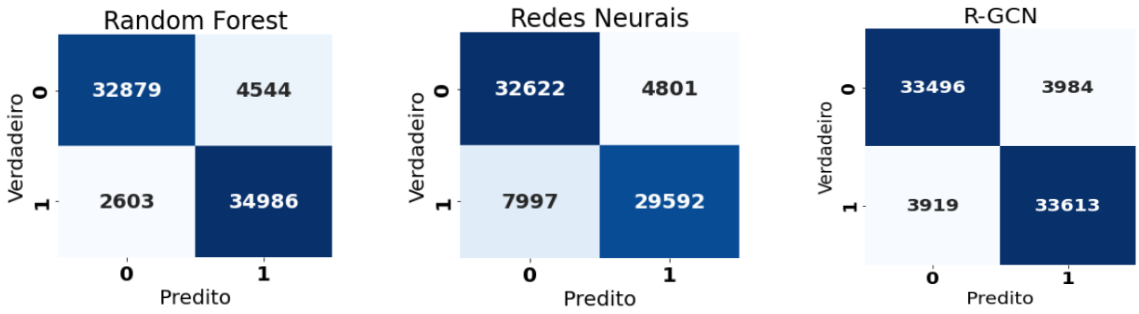
3.4 Avaliação dos resultados

A avaliação foi realizada com matrizes de confusão e métricas de acurácia, que consistem na divisão da quantidade de amostras classificadas corretamente pelo total de amostras e a área sob a curva ROC (AUC).

Para classificações binárias (sim/não), a matriz de confusão possui dimensão 2 × 2, na qual as linhas representam as classes previstas e as colunas, as classes reais. Quanto mais forte for a diagonal principal da matriz, melhor é o processo de classificação. Já a curva ROC mostra o comportamento entre a taxa de verdadeiro-positivo com a taxa de falso-positivo. AUC é a área abaixo da curva ROC e permite obter um valor numérico para a comparação da eficiência do classificador.

A Figura 3 mostra as matrizes de confusão com os resultados para cada modelo e a Tabela 1 as demais métricas.

FIGURA 3 MATRIZES DE CONFUSÃO COM RESULTADOS DO TESTE PARA OS TRÊS MODELOS PROPOSTOS



Random Forest Rede neural de múltiplas camadas R-GCN

Nota: 0: classe das empresas potencialmente idôneas; 1: classe das potencialmente inadimplentes.
Fonte: Elaborada pelos autores.

TABELA 1 RESULTADOS OBTIDO COM OS MODELOS CLÁSSICOS E R-GCN

Modelo		Treino	Teste
Random Forest	Acurácia AUC	**99,99%** **99,99%**	**90,47%** **96,85%**
Redes Neurais	Acurácia AUC	82,73% 82,75%	82,93% 82,96%
R-GCN	Acurácia AUC	89,10% 95,67%	89,13% 95,64%

Fonte: Elaborada pelos autores.

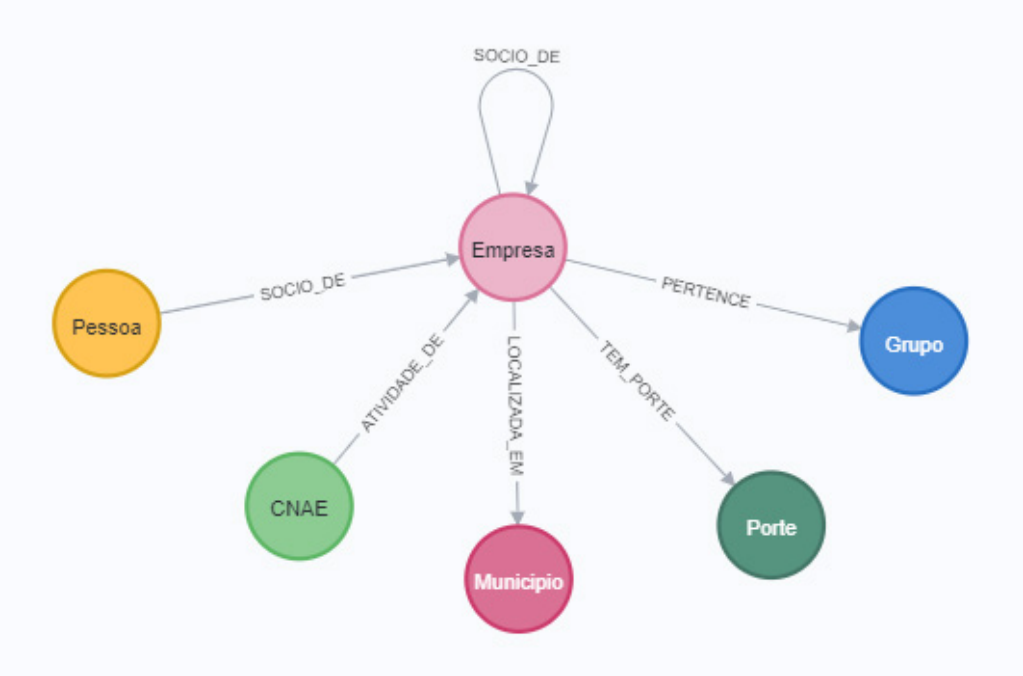
Nota-se, na coluna "Teste" da tabela, que os modelos com Random Forest e R-GCN obtiveram resultados semelhantes, com acurácia um pouco melhor com o modelo baseado em Random Forest. Todavia, a R-GCN obteve um resultado mais uniforme, com quantidades similares de falso-positivos e falso-negativos. Também é possível observar que o modelo com R-GCN teve valores de métricas de treino semelhantes aos de teste e validação, um indicativo de que o modelo é mais generalista.

3.4.1 Melhorias finas nos modelos

Durante as análises dos resultados, observou-se que a R-GCN dá mais importância para os relacionamentos (arestas) entre os nós do grafo que para os atributos de cada nó. Com isso, levantou-se a hipótese de que a transformação de atributos dos nós (como Porte da Empresa, Natureza Jurídica, Município etc.) em nós separados poderia evidenciar tais atributos para a R-GCN. Logo, um novo grafo foi criado com esse propósito, apresentado na Figura 4.

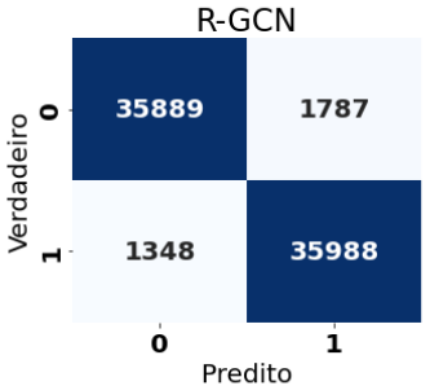
A Figura 5 mostra a matriz de confusão e a Tabela 2, as demais métricas para esse novo experimento com a R-GCN. Nota-se significativa melhora com as mudanças realizadas nesse novo experimento. Uma hipótese inédita foi levantada: como as características não euclidianas são poucas para essa massa de dados (apenas os vínculos entre empresas, sócios e CNAEs), pode ser possível melhorar os modelos clássicos transformando os dados, tanto quanto possível, em tabulares.

FIGURA 4 **GRAFO HETEROGÊNEO MODIFICADO**



Fonte: Elaborada pelos autores.

FIGURA 5 **MATRIZ DE CONFUSÃO DA R-GCN MODIFICADA**



Fonte: Elaborada pelos autores.

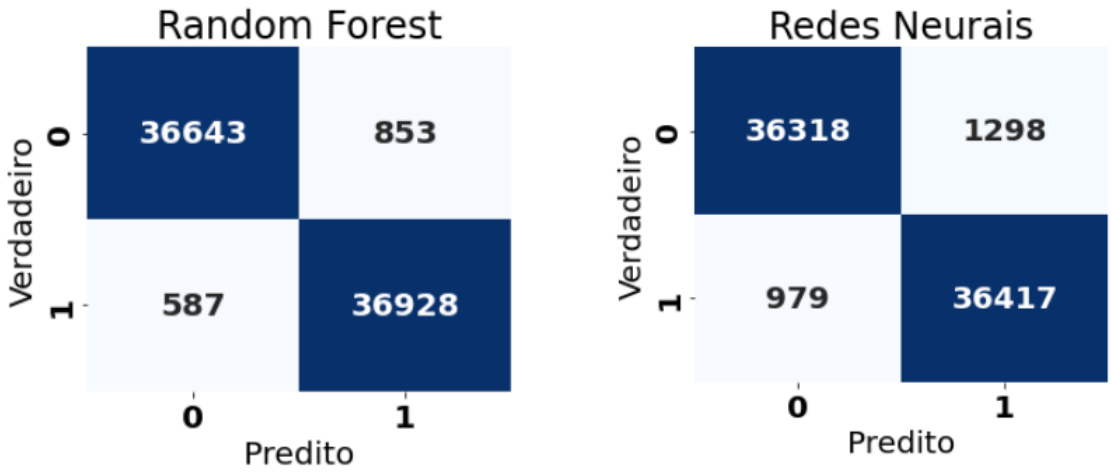
O uso da entrada de dados de forma tabular contradiz a hipótese inicial deste trabalho, pois, dessa forma, os métodos clássicos podem vir a ter uma acurácia semelhante à GNN ou até melhor. Depois de tal transformação, um novo experimento foi realizado, e os resultados podem ser vistos na Figura 6 e na Tabela 2. Nota-se que os modelos Random Forest e as redes neurais profundas obtiveram uma significativa melhora, com resultados ainda mais positivos que aqueles apresentados pelo segundo modelo com R-GCN.

TABELA 2 RESULTADOS OBTIDOS APÓS REFINAMENTO

Modelo		Treino	Teste
Random Forest	Acurácia AUC	**99,99%** **99,99%**	**98,08%** **99,65%**
Redes Neurais	Acurácia AUC	98,15% 99,73%	96,96% 99,22%
R-GCN	Acurácia AUC	95,96% 99,17%	95,82% 99,15%

Fonte: Elaborada pelos autores.

FIGURA 6 MATRIZ DE CONFUSÃO RESULTANTE DO APRIMORAMENTO DO RANDOM FOREST E DA REDE NEURAL



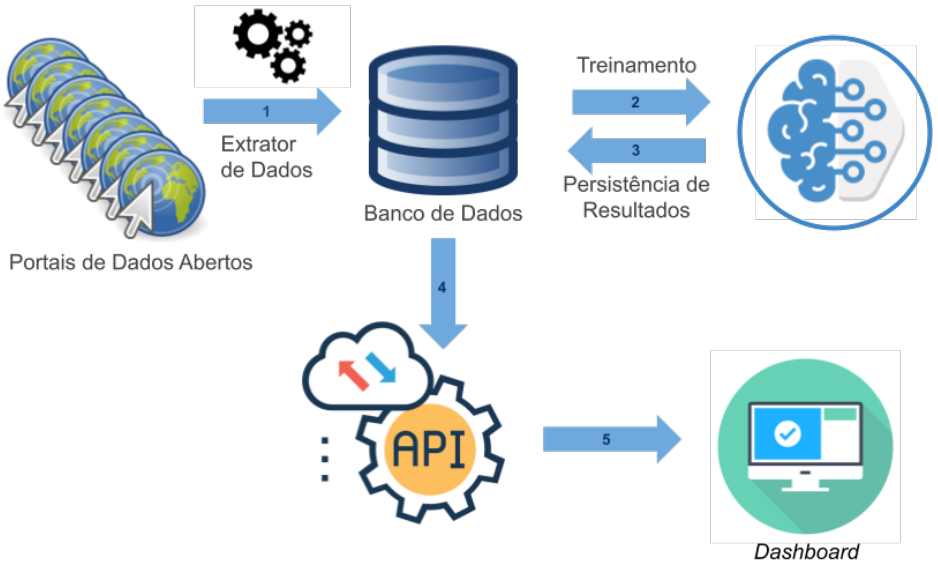
(a) *Random Forest* Rede neural de múltiplas camadas

Fonte: Elaborada pelos autores.

3.5 Implantação

A última etapa do modelo CRISP-DM é a implantação. A Figura 7 apresenta a arquitetura do sistema proposto. A extração dos dados abertos em portais é automatizada, com atualizações periódicas que visam à melhoria dos modelos inteligentes. Os modelos treinados, por sua vez, são executados para todas as empresas do estado de Goiás e disponibilizam uma probabilidade de pertencimento à classe das inadimplentes. Tais chances estão à disposição de outros sistemas, por meio de uma Application Program Interface (API). Para tanto, foi viabilizado um *dashboard web* (Figura 8), com filtros interativos que permitem a visualização das predições pelos usuários.

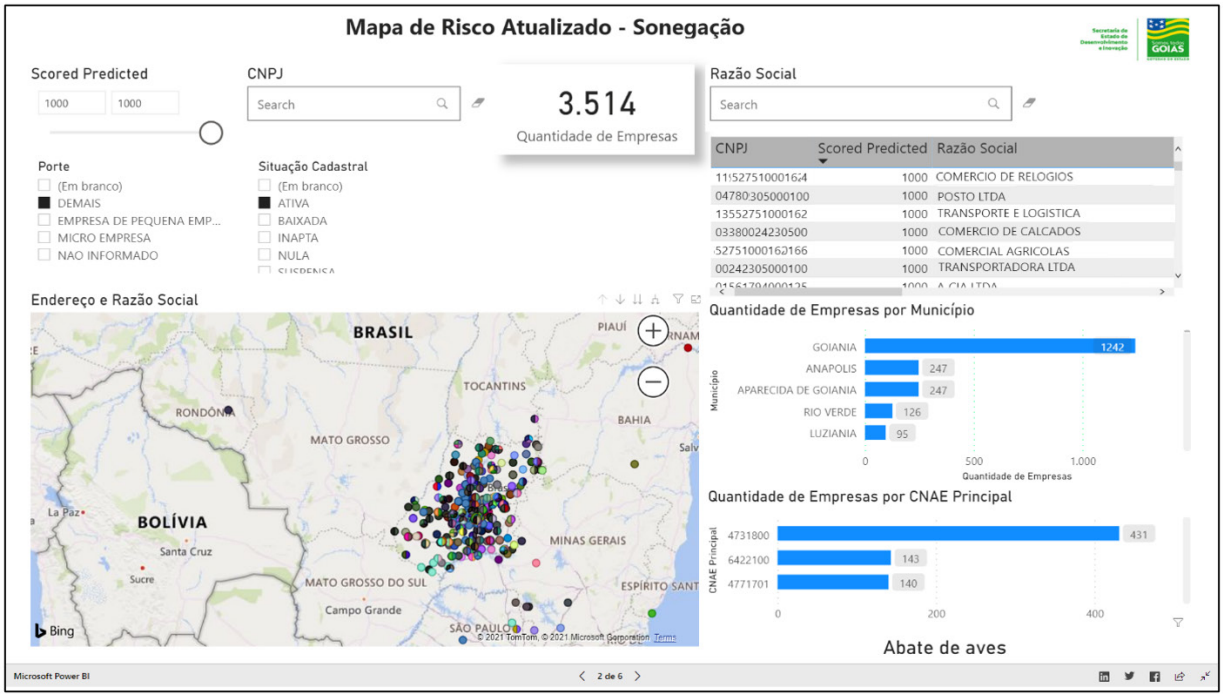
FIGURA 7 ARQUITETURA DA SOLUÇÃO PROPOSTA



Fonte: Elaborada pelos autores.

No *dashboard* pode-se filtrar pelo índice de perfil inadimplente, o qual varia de zero a mil. A funcionalidade "Mapa" apresenta a localização geográfica das empresas, facilitando a análise.

FIGURA 8 PAINEL DO MAPA DE RISCO DE SONEGAÇÃO



Fonte: Elaborada pelos autores.

4. CONCLUSÃO

Produziram-se modelos inteligentes com dados abertos para a identificação de empresas que praticam evasão fiscal. Foram comparados três modelos e realizadas melhorias, de forma a obter acurácia acima de 98%, o que reforça a hipótese inicial do trabalho, de que é possível avaliar o perfil das empresas com base em dados abertos.

Com isso, concluiu-se que a utilização de dados relacionais, representados em grafos, é equivalente aos dados tabulares utilizados em classificações. Os experimentos ainda permitiram melhor entendimento dos dados, sendo possível, assim, representar as características relacionais (antes apresentadas apenas nos grafos) também por meio de dados tabulares. Com isso, o modelo clássico Random Forest obteve melhora de quase 8% de acurácia, sendo o escolhido para a construção da solução final.

A contribuição científica deste trabalho é mostrar a viabilidade da utilização de dados públicos para o tratamento do problema da evasão fiscal, proposta não observada em nenhum outro trabalho. A utilização de Redes Neurais para Grafos Heterogêneas e Relacionais, mesmo não obtendo os melhores resultados, contribuiu para o aprimoramento dos dados utilizados nas outras técnicas.

Não é de conhecimento dos autores, depois de vasta revisão bibliográfica, que exista outro texto que utilizou tal técnica para o problema da evasão fiscal.

O resultado da pesquisa realizada neste trabalho superou as expectativas dos autores e dos auditores fiscais. Com isso, o sistema já está em uso pelos auditores e delegados fiscais.

REFERÊNCIAS

Assylbekov, Z., Melnykov, I., Bekishev, R., Baltabayeva, A., Bissengaliyeva, D., & Mamlin, E. (2016). Detecting value-added tax evasion by business entities of Kazakhstan. In *Proceedings of the 8º International Conference on Intelligent Decision Technologies*, Tenerife, Spain.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In K. Aberer, K. S. Choi, N. Noy, D. Allemang, K. I. Lee, L. Nixon, … P. Cudré-Mauroux (Eds.), *The Semantic Web* (Lecture notes in computer science, Vol., 4825, pp. 722-35). Springer, Berlin, Heidelberg.

Barman, S., Pal, U., Sarfaraj, M. A., Biswas, B., Mahata, A., & Mandal, P. (2016). A complete literature review on financial fraud detection applying data mining techniques. *International Journal of Trust Management in Computing and Communications*, *3*(4), 336-359. Recuperado de https://doi.org/10.1504/IJTMCC.2016.084561

Bethencourt, C., & Kunze, L., (2019, abril). Tax evasion, social norms, and economic growth. *Journal of Public Economic Theory*, *21*(2), 332-46. Recuperado de https://doi.org/10.1111/jpet.12346

Beutel, A., Akoglu, L., & Faloutsos, C. (2015). Fraud detection through graph-based user behavior modeling. In *Proceedings of the 22º ACM SIGSAC Conference on Computer and Communications Security*, Denver, CO.

Carrazza, R. A. (2020). *ICMS*. Salvador, BA: Editora Juspodivm.

Dreżewski, R., Sepielak, J., & Filipkowski, W. (2015, fevereiro). The application of social network analysis algorithms in a system supporting money laundering detection. *Information Sciences*, *295*, 18-32. Recuperado de https://doi.org/10.1016/j.ins.2014.10.015

Esteves, R. E. S. (2013). *Pesquisas em contabilidade tributária e planejamento tributário: uma análise bibliométrica* (Trabalho de conclusão). Universidade Federal de Goiás, Goiânia, GO.

Haykin, S. (2007). *Redes neurais: princípios e prática*. Porto Alegre, RS: Bookman Editora.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3º International Conference on Document Analysis and Recognition*, Montreal, Canada.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5º International Conference on Learning Representations*, Toulon, France.

Lei nº 4.729, de 14 de julho de 1965. (1965). Define o crime de sonegação fiscal e dá outras providências. Brasília, DF. Recuperado de http://www.planalto.gov.br/ccivil_03/leis/1950-1969/l4729.htm

Liu, X., Pan, D., & Chen, S. (2010). Application of hierarchical clustering in tax inspection case-selecting. In *Proceedings of the 2010 International Conference on Computational Intelligence and Software Engineering*, Wuhan, China.

Matos, T., Macedo, J. A. F., & Monteiro, J. M. (2015). An empirical method for discovering tax fraudsters: a real case study of Brazilian fiscal evasion. In *Proceedings of the 19º International Database Engineering & Applications Symposium*, Yokohama, Japan.

Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw-Hill

Nasution, M. K. M., Salim, S. O., & Budhiarti, N. E. (2020). Data science. *Journal of Physics: Conference Series*, *1566*(1), 20-27. Recuperado de https://doi.org/10.1088/1742-6596/1566/1/012034

Noguera, J. A., Quesada, F. J. M., Tapia, E., & Llàcer, T. (2014). Tax compliance, rational choice, and social influence: An agent-based model. *Revue française de sociologie*, *55*(4), 765-804.

Prado, K. H. J. (2020). *Data science aplicada à análise criminal baseada nos dados abertos governamentais do Brasil* (Dissertação de Mestrado). Universidade Federal de Sergipe, Laranjeiras, SE.

Projeto de Lei 6520/2019. (2019). Altera a Lei nº 8.137, de 27 de dezembro de 1990, para esclarecer que a conduta tipificada em seu art. 2º, inciso II, abarca somente as relações de responsabilidade tributária e não abrange as hipóteses em que o sujeito passivo deixa de recolher valor de tributo descontado ou cobrado caso ele tenha declarado o tributo na forma da legislação aplicável. Brasília, DF. Recuperado de https://www.camara.leg.br/

proposicoesWeb/fichadetramitacao?idProposicao=2234636

Ribeiro, C. J. S., & Almeida, R. F. (2011). Dados abertos governamentais (open government data): instrumento para exercício de cidadania pela sociedade. In *Anais do 12º Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação,* Brasília, DF.

Ruan, J., Yan, Z., Dong, B., Zheng, Q., & Qiana, B. (2019, março). Identifying suspicious groups of affiliated-transaction-based tax evasion in big data. *Information Sciences*, *477* 508-32. Recuperado de https://doi.org/10.1016/j.ins.2018.11.008

Sindicato Nacional dos Procuradores da Fazenda Nacional. (2019, junho). Sonegação no Brasil – uma estimativa do desvio da arrecadação do exercício de 2018. *Quanto Custa Brasil*. Recuperado de http://www.quantocustaobrasil.com.br/artigos/sonegacao-no-brasil-uma-estimativa-do-desvio-da-arrecadacao-do-exercicio-de-2018

Supremo Tribunal Federal. (2019) *Informativo STF*. Brasília, DF: Autor. Recuperado de https://www.stf.jus.br/arquivo/informativo/documento/informativo963.htm

West, J., & Bhattacharya, M. (2016, março). Intelligent financial fraud detection: a comprehensive review. *Computers & Security*, *57*, 47-66. Recuperado de https://doi.org/10.1016/j.cose.2015.09.005

Wirth, R., & Hipp, J. (2000). Crisp-DM: towards a standard process model for data mining. In *Proceedings of the 4º international conference on the practical applications of knowledge discovery and data mining*, Manchester, UK.

Wu, R. S., Ou, C. S., Lin, H. Y., Chang, S. I., & Yen, D. C. (2012, agosto). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, *39*(10), 8769-77. Recuperado de https://doi.org/10.1016/j.eswa.2012.01.204

Zha, Z. (2020). A reliable tax auditor assistant for exploring suspicious transactions. In *Proceedings of the WWW'20: Companion Proceedings of the Web Conference*, Taipei, Taiwan.

Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network. In *Proceedings of the 25º ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK.

Otávio Calaça Xavier

https://orcid.org/0000-0002-7826-4730
Mestre em Ciência da Computação; Professor Assistente no Instituto Federal de Goiás (IFG).
E-mail: otavio.xavier@ifg.edu.br

Sandrerley Ramos Pires

https://orcid.org/0000-0002-7273-1334
Doutorado em Engenharia Elétrica; Professor Adjunto na Escola de Engenharia Elétrica, Mecânica e de Computação da Universidade Federal de Goiás (UFG). E-mail: sandrerley@ufg.br

Thyago Carvalho Marques

https://orcid.org/0000-0002-5434-5421
Doutor em Engenharia Elétrica; Professor Associado na Escola de Engenharia Elétrica, Mecânica e de Computação da Universidade Federal de Goiás (UFG). E-mail: thyago@ufg.br

Anderson da Silva Soares

https://orcid.org/0000-0002-2967-6077
Doutor em Engenharia Eletrônica e Computação; Professor Associado do Instituto de Informática da Universidade Federal de Goiás (UFG). E-mail: andersonsoares@ufg.br