FGV EBAPE

RAP
BRAZILIAN JOURNAL OF PUBLIC ADMINISTRATION

Forum: Practical Perspectives

Tax evasion identification using open data and artificial intelligence

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 – Brazil
[2] Universidade Federal de Goiás / Escola de Engenharia Elétrica, Mecânica e de Computação, Goiânia / GO – Brazil
[3] Universidade Federal de Goiás / Instituto de Informática, Goiânia / GO – Brazil

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.

Identificação de evasão fiscal utilizando dados abertos e inteligência artificial

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.

DOI: http://dx.doi.org/10.1590/0034-761220210256x
ISSN: 1982-3134
Article received on July 17, 2021 and accepted April 06, 2022.
[Translated version] Note: All quotes in English translated by this article's translator.

ACKNOWLEDGEMENTS

Acknowledgments to the Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) and the Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG) for funding this research.

1. INTRODUCTION

Tax evasion is an important topic for governments around the world (Bethencourt & Kunze, 2019). According to the Brazilian Union of National Treasury Attorneys (Sindicato Nacional dos Procuradores da Fazenda Nacional [Sinprofaz], 2019), in Brazil, it is estimated that tax collection could be 23.1% higher without evasion, almost 8% of the country's GDP. If there were no evasion, the tax burden could be reduced by almost 30%, without a drop in revenue. In Goiás, Brazil, the situation is the same. Several works such as Vieira (2015) and Esteves (2013) show a similar situation in tax evasion and revenue recovery.

Smart computational tools can increase the productivity of inspection bodies both in revenue recovery and preventive action, anticipating tax evasion actions (Nasution, Salim, & Budhiarti, 2020).

Obtaining data to create a machine learning model is a well-known problem in the field of data science (Nasution et al., 2020). Many data used to identify tax evasion are not easily provided. They are usually confidential information of companies and their partners, protected by Brazilian law (Matos, Macedo, & Monteiro, 2015). An alternative, not widely explored yet, is the use of open data. Those data that should be freely available for everyone to use and republish as they wish (Auer et al., 2007). The Brazilian Federal Revenue, for example, discloses the registration information of companies in Brazil in the form of open data.

This study aims to use open data, from various sources, to build machine learning models focused on tax evasion detection. No non-public data was used in this research, making the approach generic and applicable to any unit of the federation.

One of the main contributions of this study is the use of open data in the problem of identifying tax evasion. Most works in this field need sensitive data. Another contribution is the use of graph neural networks, GNNs (Zhang, Song, Huang, Swami, & Chawla, 2019), in the tax evasion problem. Here, they are compared to classical machine learning approaches, such as the Random Forest algorithm (Ho, 1995) and Multilayer Neural Networks (Haykin, 2007).

This work focused on separating companies with a potential evader profile from those that do not. The company's presence in the state's active debt was used as a label for the "evader" behavior. For the labelling of the "non-evading" bias, filters were created based on the experience of tax auditors.

The organization of this article is as follows: an introduction; section 2, which presents theoretical foundations and some related works; section 3, which shows the proposed approach; section 4, which presents the results obtained and an analysis of them; and, finally, section 5, which contains the conclusion of the work.

2. BACKGROUND

This section presents some important concepts for a better reading of the work, as well as some related works.

2.1 Tax evasion and default

For Carrazza (2020), tax evasion is the act practiced by someone who adopts illegal conduct intending to not pay or reduce taxes due or, even, postpone their collection. It is legally defined in Brazil by Law No. 4,729, of July 14, 1965 (Lei nº 4.729, de 14 de julho de 1965).

Tax default is the failure to pay fees, taxes, or contributions. It alone does not characterize evasion, since default due to lack of knowledge of tax rules is not intended to defraud tax inspection (Projeto de Lei 6520/2019). However, the Brazilian Federal Supreme Court (STF) ruled a case of tax default as a crime, according to STF Information No. 963/2019 (Supremo Tribunal Federal [STF], 2019).

As the differentiation between default and evasion is still in the maturing phase in Brazil, the work focuses on identifying defaulting companies with a profile of bad taxpayers, a concept that is already well-defined.

Several works in the literature have already proposed techniques for detecting financial fraud or tax evasion. There are comprehensive surveys that describe many of them (Barman et al., 2016; West & Bhattacharya, 2016). Research in this area can be divided into traditional tax audit methods, machine learning-based audit methods and graph-based methods (Ruan, Yan, Dong, Zheng, & Qiana, 2019).

Manual case selection, complaint-based selection and selections using computer tools are three traditional methods that are frequently used. Manual selection of cases or based on complaints is time-consuming and requires specialized experience from the tax auditor (Ruan et al., 2019). According to Sinprofaz (2019), qualified professionals are expensive and less than necessary.

Matos et al. (2015) and Wu, Ou, Lin, Chang, and Yen (2012) used association rules to screen tax returns and identify frequent fraud patterns. Assylbekov et al. (2016) and Liu, Pan, and Chen (2010) adopted clustering for tax inspection and identification of anomalies that can lead to tax fraud. Noguera, Quesada, Tapia, and Llàcer (2014) approached an agent-based model for a simulation of tax compliance that combined mechanisms of social influence with rational choices.

There are also studies that used graphs to identify tax evasion, as did Beutel, Akoglu, and Faloutsos (2015) and Dreżewski, Sepielak, and Filipkowski (2015), who used data from social and banking networks to identify fraud carried out in different ways. Direct or indirect. Finally, Ruan et al. (2019) and Zha et al. (2019) propose the use of graph neural networks for more complex scenarios.

All the works mentioned above use confidential data. The present study is similar to the more recent works by Ruan et al. (2019) and Zha et al. (2019), with the use of open data as a differential.

2.2 Open data

According to Auer et al. (2007), open data should be freely available for everyone to use and republish, without restrictions by copyright, patents, or other control mechanisms. There is also the definition given by the Open Definition[1], which says that a dataset is open when it can be accessed, used, shared, and replicated by anyone.

Many governments use the open data concept and technologies to make publicly available data. In Brazil, the data.gov.br portal alone offers more than 10,000 types of open data. There are also open data portals for states and municipalities.

Open data are an instrument for exercising citizenship (Ribeiro & Almeida, 2011), as well as a source for the development of the most diverse research purposes, notably in data science, as presented by Prado (2020).

When a lot of data is available, and someone is looking for non-trivial patterns, computational machine learning mechanisms can be used.

2.3 Machine Learning

Machine learning, considered part of artificial intelligence, is the study and application of computational algorithms that automatically improve through experience and the use of data (Mitchell, 1997). Machine learning technics build intelligent models based on data samples to perform prediction or decision support tasks on new data.

This study uses the machine learning approach called Supervised Learning, which learns from a set of desired input-output pairs. The input is a set of data about an entity, and the output is the desired classification for it. Training consists of mapping this input-output relationship, allowing the model to predict the output of new entities that were not part of the training set.

Ruan et al. (2019) and Zha et al. (2019) are examples of research which use Supervised Learning in the process of generating prediction models.

Techniques such as Random Forest (Ho, 1995) and Multilayer Neural Networks (Haykin, 2007), used in this study, are commonly applied in supervised learning with good results. On the other hand, association, and clustering rules, mentioned in section 2.1, are unsupervised approaches.

This work models the identification of tax evasion as a supervised learning problem, specifically classification, since it is intended to classify companies between reputable and in default.

3. PROPOSED APPROACH

The proposed approach is to model the identification of tax evaders as a binary classification between the potential "default" and "reputable" profiles. The dataset construction activities and data analysis followed the process model for data mining called CRISP-DM (Wirth & Hipp, 2000). Figure 1 shows the process steps. Sections 1, 2 and 3 of this paper correspond to the first step (Business Understanding).

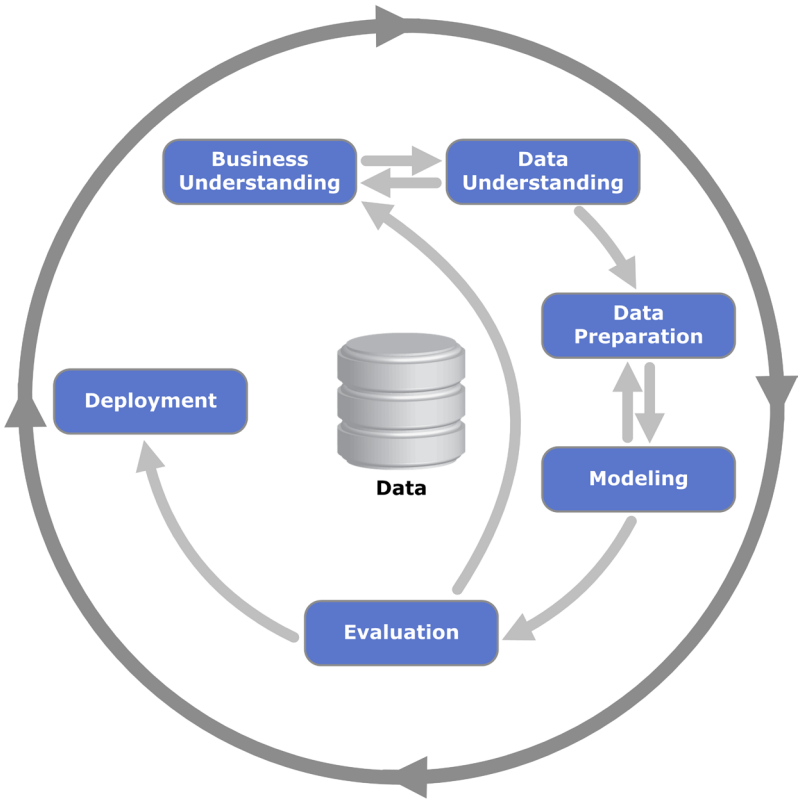
[1] The Open Definition sets out principles that define "openness" in relation to data and content. Retrieved from http://opendefinition.org/

3.1 Data Understanding

This work used the following data sources:

- **Public Data of the National Register of Legal Entities (CNPJ)**[2] – stores the public data of all the registered Legal Entities (mainly companies and institutions), and their partners, registered by the Brazilian IRS.

FIGURE 1 STEPS OF THE CRISP-DM PROCESS



Source: Adapted from CRISP-DM (Wirth & Hipp, 2000).

- **Sintegra**[3] – Integrated Information System on Interstate Transactions with Goods and Services. It has data of companies that operate in multiple states of the federation.
- **Tax Administrative Council of the State of Goiás**[4] – contains the processes of companies assessed and judged in the administrative sphere of the state of Goiás.

[2] Public Data of the National Register of Legal Entities (CNPJ). Retrieved from https://cutt.ly/Mcn6RM0
[3] Sintegra. Retrieved from http://www.sintegra.gov.br/
[4] Tax Administrative Council of the State of Goiás. Retrieved from https://cutt.ly/Vcn617O

Altogether, the used datasets contain 1.6 million companies in the state of Goiás. For the composition of the "default" profile, we considered the companies registered in the Active Debt of the State of Goiás, in a total of 193,987 companies. For the composition of the "suitable" profile, we applied filters based on the experience of tax auditors, and it was possible to extract 617,622 companies. Note that the classes are unbalanced. The "suitable" class is 3.2 times greater than the "default" class.

3.1.1 Data balancing

To match the number of companies in each of the classes, we extracted a random sample of the "reputable" class, which contains the same number of companies in the "default" class. There was heterogeneity between the types of companies existing in the registry, thus, the balancing of the data must consider the segmentation of specific groups of companies. In this work, we segmented by the opening date and the category of the company (if it is registered as an individual microentrepreneur - MEI).

3.2 Data preparation

The following activities are part of this step:

- Elimination of non-categorical alphabetical variables, such as trade name, company name, etc.
- Formatting and simplification of categorical data. Data such as opting for the national tax simplification program (Simples Nacional) or if it is an individual microentrepreneur (MEI) have more than one value with the same meaning, for instance.
- Binarization of categorical variables, for the non-correlation of magnitude in categorical data.
- Removal of null fields and samples with more than 50% null values in features.

At this stage, we observed that corporate ties may indicate the evader profile. A company can have as a partner another company with an evading profile or an individual who is part of the corporate structure of an evading company. There is also the link between the parent company and its subsidiaries, which can be obtained by the CNPJ numbers of the companies.

For separation of the validation dataset, we used a threshold date, January 1, 2020. Thus, for training and testing, only companies open until that date were considered, totaling 375,062 companies, divided equally between the two classes. Then, we split this amount in 80% (300,050) for training and 20% (75,012) for testing.

3.3 Modeling

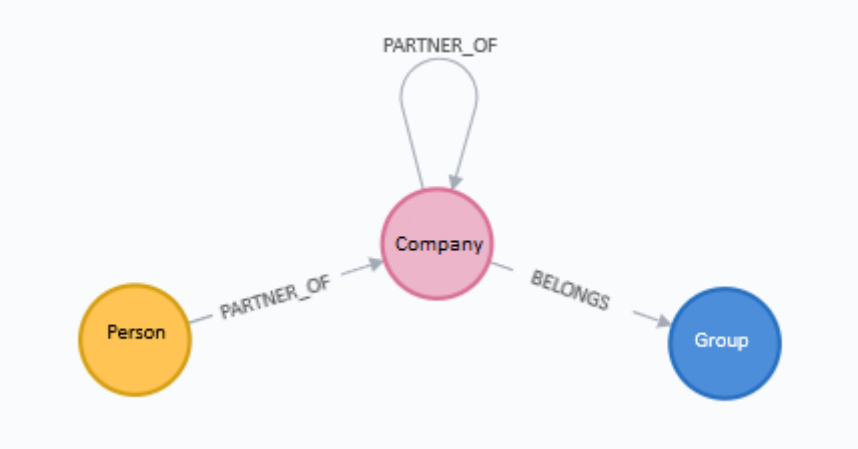
The different relationships that a company may have with different companies and people (corporate structure) introduce complexity in the tabular representation of data. For example, two CNPJs of companies from the same group are not numerically close. A solution to this type of problem is the use of a model of graph-based neural networks, the GNNs, as described by Kipt et al. (2016). In them, graphs represent the input data, which do not belong to a Euclidean space. In this work, we adopted a specific type of GNN called R-GCN (Zhang et al., 2019).

For comparison, we also built models based on Multilayer Neural Networks and Random Forest. The central idea of Multilayer Neural Networks is the mathematical representation of neurons

interconnected in several layers, to simulate a natural neural network (Haykin, 2007). Random Forest, on the other hand, is a learning method that operates through the construction of various decision trees at the time of training. Decision trees are techniques that use a tree-like decision model, in which the assembly of the tree structure maps the algorithm's ability to classify events (Ho, 1995).

Figure 2 shows the structure of the graph created for the experiment. Each node in the figure represents a node type in the R-GCN input graph. Company nodes are the only ones that contain detailed data, and the other node types are limited to the item identifier or name. Even so, the graph can represent, through the relationships, information that is not easily representable through tabular data.

FIGURE 2 INPUT HETEROGENEOUS GRAPH FOR R-GCN



Source: Elaborated by the authors.

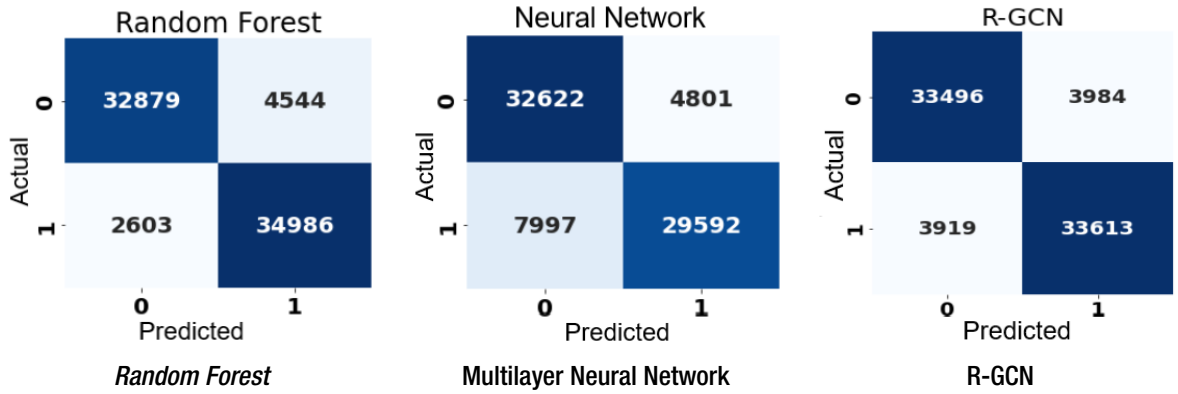
3.4 Results Evaluation

The evaluation consisted of the analysis of confusion matrices and accuracy metrics, which consist of dividing the number of samples correctly classified by the total number of samples and the area under the ROC curve (AUC).

For binary classifications, the confusion matrix has a dimension of 2 × 2, in which the lines represent the predicted classes and the columns the actual classes. The stronger the main diagonal of the matrix, the better the classification process. The ROC curve shows the behavior between the true-positive rate and the false-positive rate. AUC is the area under the ROC curve and provides a numerical value for comparing the efficiency of the classifier.

Figure 3 shows the confusion matrices with the results for each model, and Table 1 shows the other metrics.

FIGURE 3 CONFUSION MATRICES WITH TEST RESULTS FOR THE THREE PROPOSED MODELS



Random Forest Multilayer Neural Network R-GCN

Note: 0: class of potentially reputable companies; 1: class of potentially defaulters.
Source: Elaborated by the authors.

TABLE 1 RESULTS OBTAINED WITH THE CLASSIC AND R-GCN MODELS

Model		Training	Test
Random Forest	Accuracy AUC	99,99% 99,99%	90,47% 96,85%
Neural Network	Accuracy AUC	82,73% 82,75%	82,93% 82,96%
R-GCN	Accuracy AUC	89,10% 95,67%	89,13% 95,64%

Source: Elaborated by the authors.

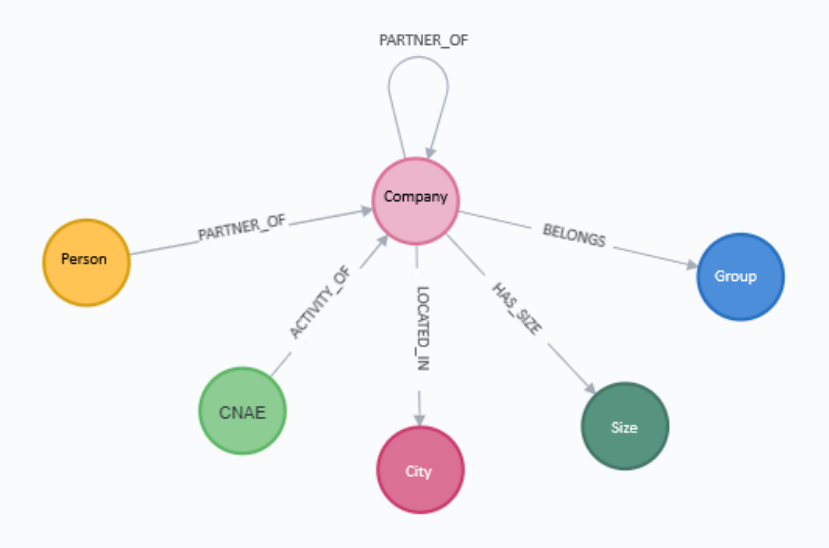
The "Test" column of the Table 1 shows that the models with Random Forest and R-GCN obtained similar results, with slightly better accuracy with the model based on Random Forest. However, R-GCN obtained a more uniform result, with similar amounts of false positives and false negatives. It is also possible to observe that the model with R-GCN had training metrics values similar to those of test and validation, an indication that the model is more generalist.

3.4.1 Models' fine-tuning

During the analysis of the results, we observed that the R-GCN gives more importance to the relationships (edges) between the nodes of the graph than to the attributes of each node. Thus, the hypothesis raised was that the transformation of node attributes (such as Company Size, Legal Nature, Municipality, etc.) into separate nodes could evidence such attributes for the R-GCN. Therefore, we've created a new graph for this purpose, shown in Figure 4.

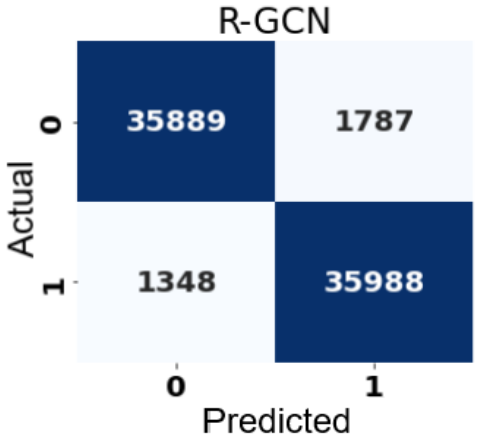
Figure 5 shows the confusion matrix and Table 2 shows the other metrics for this new experiment with R-GCN. There is a significant improvement with the proposed changes. With that, we raised a new hypothesis: as the non-Euclidean characteristics are few for this dataset (only the links between companies, partners and CNAEs), it may be possible to improve the classical models by transforming the data, as much as possible, into tabular ones.

FIGURE 4 **MODIFIED HETEROGENEOUS GRAPH**



Source: Elaborated by the authors.

FIGURE 5 **MODIFIED R-GCN CONFUSION MATRIX**



Source: Elaborated by the authors.

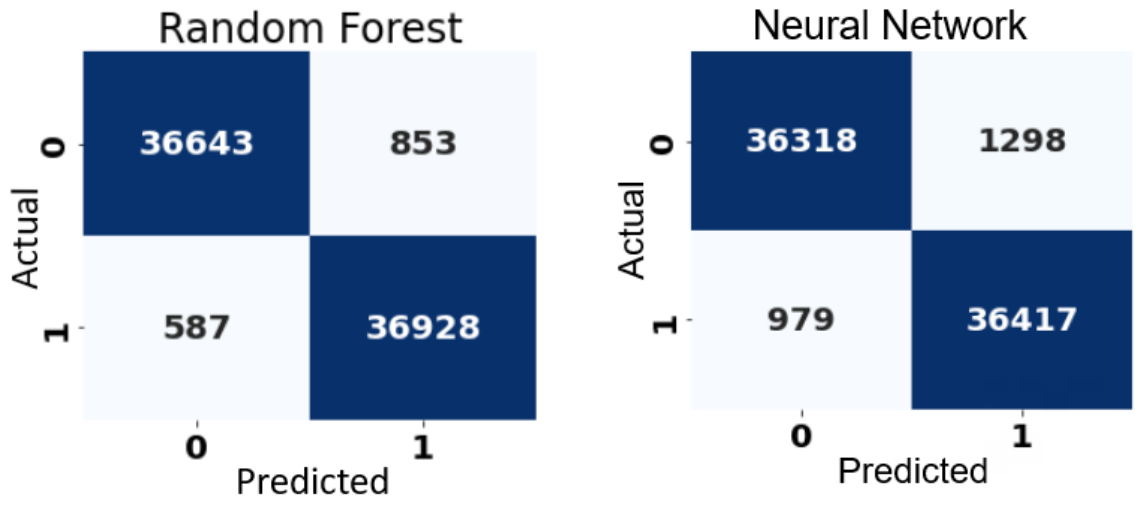
The use of tabular data entry contradicts the initial hypothesis of this work, since, in this way, the classical methods may have an accuracy similar to GNN or even better. After such transformation, we carried out a new experiment and the results can be seen in Figure 6 and Table 2. The Random Forest models and the neural network obtained a significant improvement, with even more positive results than those presented by the second model with R-GCN.

TABLE 2 RESULTS OBTAINED AFTER FINE-TUNING

Model		Training	Test
Random Forest	Accuracy AUC	**99,99%** **99,99%**	**98,08%** **99,65%**
Neural Network	Accuracy AUC	98,15% 99,73%	96,96% 99,22%
R-GCN	Accuracy AUC	95,96% 99,17%	95,82% 99,15%

Source: Elaborated by the authors.

FIGURE 6 CONFUSION MATRIX RESULTING FROM THE FINE-TUNING OF RANDOM FOREST AND NEURAL NETWORK



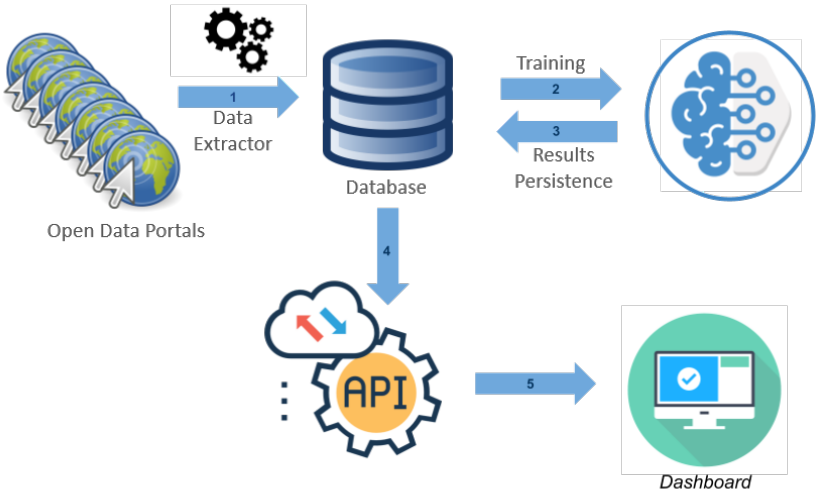
(a) *Random Forest* Multilayer Neural Network

Source: Elaborated by the authors.

3.5 Deployment

The last step of the CRISP-DM process is deployment. Figure 7 presents the architecture of the proposed system. There is an automated process to extract open data in portals. It periodically updates the data aimed at improving the intelligent models. The trained models, in turn, are executed for all companies in the state of Goiás and provide a probability of belonging to the class of defaulters. Such predictions are available to other systems, through an Application Program Interface (API). For this purpose, we've created a web dashboard (Figure 8), with interactive filters that allow users to visualize the predictions.

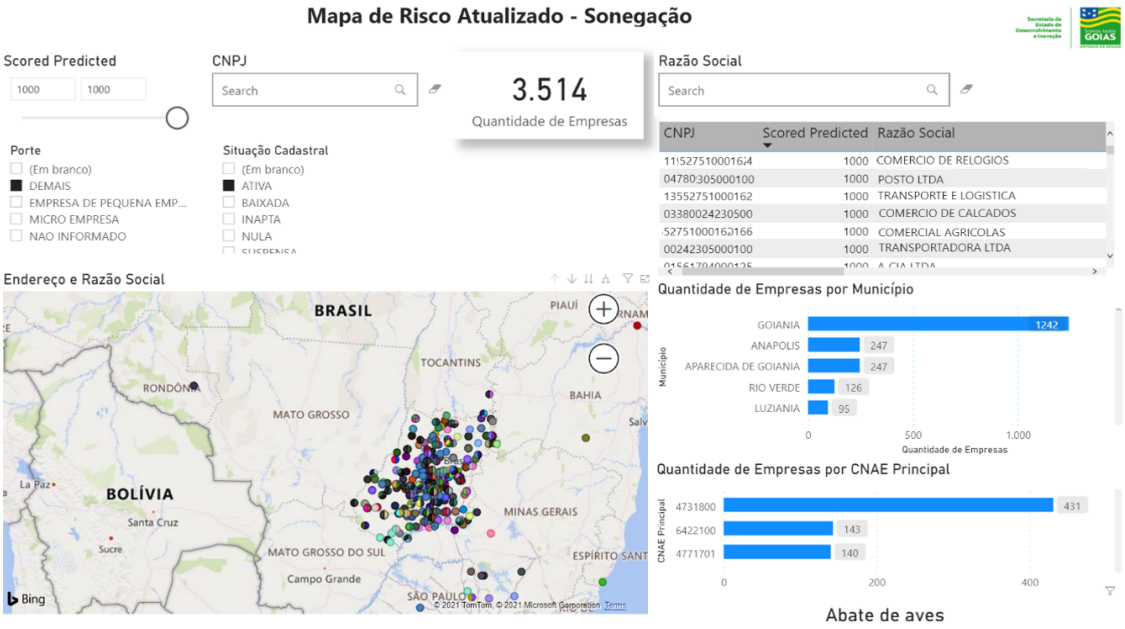
FIGURE 7 ARCHITECTURE OF THE PROPOSED SOLUTION



Source: Elaborated by the authors.

In the dashboard, someone can filter by the probability index of being a "default" company, which varies from zero to one thousand. The "Map" functionality presents the geographic location of the companies, facilitating the analysis.

FIGURE 8 EVASION RISK MAP DASHBOARD



Source: Elaborated by the authors.

4. CONCLUSION

This work produced intelligent models with open data to identify companies that practice tax evasion. We've compared three models and did some fine-tuning to obtain accuracy above 98%, reinforcing the initial hypothesis of the work, that it is possible to evaluate the profile of companies based on open data.

The conclusion was that the use of relational data, represented in graphs, is equivalent to the tabular data used in classifications for the situations presented in such a work. The experiments also allowed a better understanding of the data, making it possible, therefore, to represent the relational characteristics (previously presented only in graphs) also through tabular data. Thus, the classic Random Forest model obtained an improvement of almost 8% in accuracy, being chosen for the construction of the final solution.

The scientific contribution of this work is to show the feasibility of using public data to deal with the problem of tax evasion, a proposal not observed in any other work. The use of Neural Networks for Heterogeneous and Relational Graphs, even not obtaining the best results, contributed to the improvement of the data used in the other techniques.

The authors are not aware, after extensive literature review, that there is another text that used this technique for the problem of tax evasion.

The result of the research carried out in this work exceeded the expectations of the authors and the tax auditors. As a result, the system is already in use by auditors and tax delegates.

REFERENCES

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. Retrieved from https://doi.org/10.1504/IJTMCC.2016.084561

Bethencourt, C., & Kunze, L., (2019, April). Tax evasion, social norms, and economic growth. *Journal of Public Economic Theory*, *21*(2), 332-46. Retrieved from 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, February). The application of social network analysis algorithms in a system supporting money laundering detection. *Information Sciences*, *295*, 18-32. Retrieved from 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* (Undergraduate Thesis). 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. Retrieved from 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. Retrieved from 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* (Master Thesis). 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. Retrieved from 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, March). Identifying suspicious groups of affiliated-transaction-based tax evasion in big data. *Information Sciences*, *477* 508-32. Retrieved from https://doi.org/10.1016/j.ins.2018.11.008

Sindicato Nacional dos Procuradores da Fazenda Nacional. (2019, June). Sonegação no Brasil – uma estimativa do desvio da arrecadação do exercício de 2018. *Quanto Custa Brasil*. Retrieved from 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: Author. Retrieved from https://www.stf.jus.br/arquivo/informativo/documento/informativo963.htm

West, J., & Bhattacharya, M. (2016, March). Intelligent financial fraud detection: a comprehensive review. *Computers & Security*, *57*, 47-66. Retrieved from 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, August). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, *39*(10), 8769-77. Retrieved from 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
Master in Computer Science; Assistant Professor at the Federal Institute of Goiás (IFG).
E-mail: otavio.xavier@ifg.edu.br

Sandrerley Ramos Pires

https://orcid.org/0000-0002-7273-1334
Ph.D. in Electrical Engineering; Adjunct Professor at the School of Electrical, Mechanical and Computer Engineering at the Federal University of Goiás (UFG). E-mail: sandrerley@ufg.br

Thyago Carvalho Marques

https://orcid.org/0000-0002-5434-5421
Ph.D. in Electrical Engineering; Associate Professor at the School of Electrical, Mechanical and Computer Engineering at the Federal University of Goiás (UFG). E-mail: thyago@ufg.br

Anderson da Silva Soares

https://orcid.org/0000-0002-2967-6077
Ph.D. in Electronic and Computer Engineering; Associate Professor at the Institute of Informatics at the Federal University of Goiás (UFG). E-mail: andersonsoares@ufg.br