On the Evaluation of Classification Methods Applied to Requests for Revision of Registered Debts

Helton Souza Lima[1], Damires Yluska de Souza Fernandes[1], Thiago José Marques Moura[1] and Daniel Sabóia[2]
[1]Instituto Federal da Paraíba, 720 Avenida Primeiro de Maio, João Pessoa, Brazil
[2]Procuradoria-Geral da Fazenda Nacional, Esplanada dos Ministérios Bloco P, Brasília, Brazil

Keywords: Government Data, Tax, Supervised Learning, Imbalanced Data.

Abstract: Tax management is a complex problem faced by governments around the world. In Brazil, in order to help solving problems in this area, data analytics has been increasingly used to support and enhance tax management processes. In this light, this work proposes an approach which uses supervised learning in order to classify requests of an administrative service. The requests at hand are named as Requests for Revision of Registered Debt (R3Ds). The service underlying such requests is offered by the Brazil's National Treasury Attorney-General's Office and usually deals with a high volume of registrations. The experimental evaluation accomplished in this work presents some promising results. The obtained classification models present good levels of accuracy, area under ROC curve and recall. Four evaluation scenarios have been experimented, including imbalanced and balanced data. The Random Forest model achieves the best results in all the evaluated scenarios.

1 INTRODUCTION

Failure to comply with tax obligations may have a negative impact on the quality of life of citizens. This is due to the fact that without tax revenue it is not possible to maintain essential public services, such as health services, sanitation, mobility, security, education, among others (Mathews et al., 2018). Once the legal deadline for paying a tax has expired, the debt can be claimed by the government through the Judiciary, i.e., by the system of courts of justice in a country. Particularly in Brazil, according to the country's National Treasury Attorney-General's Office (hereafter called as *PGFN abbreviated from* "*Procuradoria-Geral da Fazenda Nacional*"), the Federal Active Debt [1] (FAD), in early 2019, accumulated 2.4 trillion reals (Brazilian currency), from 4.9 million debtors[2].

Brazilian tax enforcement processes take too long and may have a low resolution rate. According to the Brazil's National Council of Justice[3], the average processing time for a tax enforcement process is usually about 8 years. These processes represent 39% of total pending cases, and 70% of pending executions, with a congestion rate of 87%. This means, for instance that, in 2019, for every hundred tax enforcement proceedings, only 13 of them were closed. Thereby, debts usually reach the Judiciary after the administrative means of collection are exhausted, what implies in a hard task to recover their tax.

In this context, Artificial Intelligence (AI) techniques have been progressively used to support and improve some Brazilian tax enforcement processes (Souza and Siqueira, 2020). Specifically in the area of tax justice, there is an initiative of the National Council of Justice on using AI that aims to reduce the time for the outcome of tax enforcement processes[4].

[1] https://www.gov.br/pgfn/pt-br/assuntos/divida-ativa-da-uniao

[2] https://www.gov.br/pgfn/pt-br/acesso-a-informacao/institucional/pgfn-em-numeros-2014/pgfn-em-numeros-2020/view

[3] https://www.cnj.jus.br/wp-content/uploads/2020/08/WEB-V3-Justi%C3%A7a-em-N%C3%Bameros-2020-atualizado-em-25-08-2020.pdf

[4] https://www.cnj.jus.br/cnj-usara-automacao-e-inteligencia-artificial-para-destravar-execucao-fiscal/

Lima, H., Fernandes, D., Moura, T. and Sabóia, D.
On the Evaluation of Classification Methods Applied to Requests for Revision of Registered Debts.
DOI: 10.5220/0010498403350342
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 335-342
ISBN: 978-989-758-509-8

The *PGFN* currently offers the Request for Revision of Registered Debt[5] (hereafter called as R3D), which is a service available since 2018. It is an administrative claim, that allows taxpayers to request a reanalysis of the situation of their debts registered at FAD. It is an important way for reducing the rate of new tax enforcement processes, aiming to avoid the judicialization of erroneous processes. According to the Federal Services Monitoring Panel[6], R3D is the most requested service in the light of the *PGFN*, which highlights the high volume of requests to be analyzed by the institution: approximately 44 thousands were registered in 2019, involving nearly 44 billions reals. Enhancing activities related to administrative tax processes may lead to an increase of tax recovery.

There is a dataset prepared by the PGFN that includes a lot of information about R3Ds. Understanding this dataset and analyzing it can indeed generate important insights for the PGFN. Particularly, classifying the likelihood of an R3D being approved or rejected can help PGFN to improve its processes and streamline results. Considering this, the dataset is labeled with two possible classes: approved R3D or rejected R3D. Nevertheless,it has been realized that the two classes have a level of imbalance that must be addressed.

With this scenario in mind, we define three main problems that have guided this work, as follows: (i) the need to indicate the likelihood for an R3D to be approved or rejected based on the use of supervised classification models; (ii) to evaluate some supervised classification models regarding important measures with respect to the context of this scenario and (iii) to analyze strategies and apply some of them to deal with the imbalance of existing classes.

Thus, historical data of the R3Ds are used to train some supervised classification models. The five generated models are evaluated with respect to the measures Accuracy (ACC), Recall (REC) and area under the ROC curve (AUC). To this end, experimental scenarios have been defined taking into account hold out and cross-validation strategies as well as imbalanced versus balanced data. The results obtained are promising and demonstrate good scores for the evaluated metrics. In particular, the model produced with the Random Forest method has obtained the best measure scores. Regarding the use of class balancing strategies, there has been no change in relation to the results of the obtained models.

This paper is organized as follows: Section 2 provides some theoretical background; Section 3 describes some related works; Section 4 presents the applied methodology; Section 5 discusses the results which have been obtained, and Section 6 concludes the paper and suggests some future work.

2 THEORETICAL BACKGROUND

In this section, we provide some concepts regarding the tax management business domain in our country and also some principles with respect to Supervised Learning.

2.1 Request for Revision of Registered Debt

The Request for Revision of Registered Debt (R3D) is an administrative claim that allows taxpayers to ask for a reanalysis of the situation of their debts. It can be used in cases of payment, instalment, suspension of request under judicial decision, administrative decision, judicial deposit, offset, correction of statement, filling the statement inaccurately, formal defect in the credit constitution, decay or prescription, issues related to situations where the active debt enrolment is prohibited and any extinction or suspension cause of tax or non-tax debt.

Once the request for revision is granted, its registration may be cancelled or rectified. The demand for the debt may also be suspended. The task of analysing and answering R3Ds is actually a time-consuming task. Nowadays it is accomplished in about 30 days. And it is completely human-dependent.

2.2 Cross Industry Standard Process for Data Mining

The Cross Industry Standard Process (CRISP-DM) is a methodology which is usually used by data scientists in order to ensure quality on knowledge discovery project results (Chapman et al., 1999). The process is tool-independent and can be used across various business domains. It is based on iterative and incremental principles.

In this light, in order to extract knowledge from data of a given domain, the CRISP-DM guides data

[5] https://www.gov.br/pt-br/servicos/solicitar-revisao-de-divida-inscrita

[6] http://painelservicos.servicos.gov.br/

scientists to (i) identify and give a solution to a problem with the use of data mining techniques, (ii) understand the underlying data and their relationships, (iii) extract a suitable dataset, (iv) create machine learning models in order to solve the identified problem, (v) evaluate the performance of the obtained new models, and (vi) demonstrate how these models can be used and, eventually, be deployed in the given business context. We use this process in the light of our problem domain, i.e., with respect to the R3D classification problem.

2.3 Supervised Learning

Machine Learning is an area of the Artificial Intelligence (AI) whose objective is the construction of systems capable of acquiring knowledge automatically (Rezende, 2005). A subarea of Machine Learning (ML), named Supervised Learning, is composed of systems able to provide predictions based on previous specific situations stored on a dataset (Mitchell, 1997).

In supervised learning, one predictive task is classification. Classification algorithms predicts qualitative values, which will be assigned in predefined categories (Mohri et al., 2018). In this work, we deal with a two-class classification problem, thus we aim to learn a class from its positive and negative examples.

In the light of this work, an example (instance) is positive in case of a rejected R3D (request). On the other hand, negative examples regard accepted requests. For two-class problems a variety of performance measures has been proposed. For a positive example, if the prediction is also positive, this is a true positive (TP); if a prediction is negative for a positive example, this represents a false negative (FN). For a negative example, if the prediction is also negative, we have a true negative (TN), and we have a false positive (FP) if we predict a negative example as positive (Alpaydin, 2010).

The measures used in this work are Accuracy (ACC), Recall (REC) and Area Under Receiver Operating Characteristic Curve (AUC). They are defined in accordance with the following formulas (Hossin and Sulaiman, 2015):

$$ACC = (TP + TN) / (TP + TN + FP + FN) \qquad (1)$$

$$REC = TP / (TP + FN) \qquad (2)$$

AUC is calculated through the plot of the ROC curve, where the TPR is in y-axis and the FPR is in x-axis (3)

Some reasons for choosing such measures are described as follows.

The Accuracy (ACC) measures the ratio of correct predictions over the total number of instances evaluated. Accuracy is the most used evaluation measure in practice either for binary or multi-class classification problems. It is easy to compute and easy to understand by human (Hossin and Sulaiman, 2015).

In addition to accuracy, the AUC measure may be used to present an overall view of a binary classification model performance. It describes the relationship between sensitivity (recall) and specificity measures. The AUC has been proven theoretically and empirically better than the accuracy metric for evaluating some classifiers performance (Huang and Ling, 2005; Alpaydin 2010).

One point that deserves attention is the cost involved in making incorrect predictions: it is less costly to predict a rejection when the request should be accepted than to predict an approval when the request should be rejected. In the dataset used in this work, the positive value (1) indicates a rejected request, and the negative value (0) indicates an accepted one. This is the reason why the recall measure (REC) is the most important (not exclusively) one in the evaluation accomplished in this work. Classifiers with a large recall don't have a high index of false negatives (Harrington, 2012).

The supervised classification methods used in this work are Artificial Neural Networks (ANN), Naive Bayes (NB), Random Forest (RF) and Support Vector Machines (SVM). They are briefly described as follows.

- The Naive Bayes classifier is inspired by Thomas Bayes Theorem. It estimates the classification of new examples through a probabilistic algorithm (Rish, 2001). It is called "naïve" for making no assumption among the classes.
- Support Vector Machines classify data by building a separating hyperplane to distinguish and identify two types of different classes. To this end, they determine points between two domain universes, usually drawing a line (or vector) and differentiating the data on both sides (Gonzalez et al., 2005).
- Artificial Neural Networks are models inspired by the human brain. They are composed by a net of interconnected units called Perceptrons (Mitchell, 1997), which are organized in layers. The network receives the training examples and uses error functions to calculate weights in order to maximize the correct prediction.

- Random Forests are a combination of decision trees. Each tree has a different behaviour by the effect of a randomly function applied in all trees in the forest (Breiman, 2001). For every classification, the majority vote of all trees determines the models' classification.

These methods have been chosen due to some characteristics.

Regarding a NB classifier, one of the major advantages is its short computational time for training. NB provides the probability of an instance to belong to a class, rather than simply providing a classification (Kotsiantis et al., 2007). This is an information that must add value to the prosecutor's decision. Thus, it is desirable to be achieved in our approach.

The SVM method has been considered interesting since it usually fits the available data well without overfitting (Bhavsar and Panchal, 2012).

With respect to ANNs, they outperform other methods in many different business domains (Paliwal and Kumar, 2009). One of the important advantages of this method is that it can automatically approximate any nonlinear mathematical function. This aspect is useful when the relationship among the variables is not known.

Random forests are fast and easy to implement. They produce highly accurate predictions and can handle a very large number of input variables without overfitting (Biau, 2012). They can also provide the most important variables of the dataset considered for the model. They can be useful on a future dimensionality reduction task.

Another usual issue in classification tasks regards imbalanced classes. A two-class dataset is said to be imbalanced when one minority class is under-represented with regard to the majority class (Japkowicz and Stephen, 2002). The application of re-sampling techniques to obtain a more balanced data distribution is an effective solution to the imbalanced class problem (He and Ma, 2013).

Among a diverse set of re-sampling methods, we briefly describe the two ones used in this work: Random Undersampling and SMOTE. The former removes a random set of majority class examples. It is one of the simplest re-sampling approaches. Although it can eliminate useful examples, it requires less computational effort (Branco et al., 2016). The latter, which means Synthetic Minority Oversampling TEchnique, over-samples the minority class by generating new artificial data. The synthetic data are created using an interpolation strategy that introduces a new example along the line segment joining a seed example and a user-defined number of nearest neighbours (Chawla et al., 2002);

These methods have been used and evaluated in several related works (Branco et al., 2016).

3 RELATED WORKS

In this section, we briefly resume some relevant and related work which applies machine learning in the data domain of tax management.

One of the works regards classifying companies as contumacious tax debtors or not (Soares and Cunha, 2020). In this work, the dataset used was built from a data warehouse system of a brazilian city. The work aimed to help tax auditors on prioritizing the taxpayers that have higher risks of service tax default. They evaluated LightGBM, Logistic Regression and Random Forest models with respect to accuracy and AUC measures. Results were considered better than their previous work.

The work of Dias and Becker (2017) conducted a study to classify invoices as potential audit candidates or not. It used data extracted from the electronic invoice system of Porto Alegre city finance secretary, in Brazil. Results were considered as promising since they presented a high precision rate using the SVM method.

Another related work aimed to help decision-making in government taxes audit plans by using historical data from previous audits (Ippolito and Lozano, 2020). It tried to predict service tax crimes against the tax system of the city of São Paulo, Brazil. The target variable contained the information whether the taxpayer committed a crime against tax system or not, in previous tax audits. Six algorithms were applied: Neural Networks, Naive Bayes, Decision Trees, Logistic Regression, Random Forest and Ensemble Learning. Random Forest yielded the highest scores in the majority of the performance metrics utilized.

López et al., (2019) used data from the Spanish Revenue Office, with the goal of identifying taxpayers who evade tax. Their study applied Neural Networks and reached a good level of correct predictions.

Another recent work proposed a customized loss function, assigned to a social cost, to evaluate the performance of some models (Battiston et al., 2020). The proposition was validated through the use of a dataset provided by the Italian Revenue Agency, with information of income tax of more than 600 thousand individuals over 5 years. The Random Forest model was considered the best classifier, achieving the lowest value for the defined loss function.

Silva et al., (2015) worked on building predictive models on the results of specific claims in a tax administration process in the Brazilian Federal Revenue (BFR). This is the most similar work to ours. It classified credit compensation requests as "granted" or "rejected". The dataset included information built from several transactional and analytical BFR's systems. Random Forest was identified as the algorithm selected for the deployment phase with the argument that it was more accurate in the most important class: it is less costly to predict a rejection when the request should be granted than to predict a grant when the request should be dismissed.

Comparing these works with ours, some different aspects are identified as follows. One aspect is that, differently from the works of López et al., (2019) and Battiston et al., (2020), this work does not deal with fraud detection. Another aspect is that our work deals with historical data filled with manual analysis in order to label the target variable. It is not set by specific automatic business rules like the ones of two brazilian cities (Soares and Cunha, 2020; Dias and Becker, 2017). The third aspect is that this is the first work that deals with this *PGFN's* specific dataset, with its own characteristics and business rules. For example, the size of the dataset, with 70.780 cases is significantly bigger than the 151 cases of tax crime detection presented in Ippolito and Lozano (2020). Other example regards the fact that the dataset used in this work represents all regions of Brazil and not only one specific jurisdiction such as the work of Silva et al., (2015). Futhermore this work observes the effects of class balancing methods on the performance of the models, and none of the related works registered this observation.

4 METHODOLOGY

In the following subsections we present details on how the steps of the CRISP-DM methodology is applied in this work. The steps applied are: Business Understanding, Data understanding, Data preparation, Modelling and Evaluation.

4.1 Business Understanding and Research Questions Definitions

This initial phase focuses on understanding the business objectives and is used to define some research questions. The *PGFN*'s business main objectives are to improve taxpayer assistance and also to increase tax recovery. In order to help achieving these objectives, our approach has been specified to assist decision-making of analysts of the Requests for Revision of Registered Debts. Thereby, there should be an increase of the assertiveness of the requests' results as well as a decrease of the response time of the requests answering.

With this scenario in mind, besides que questions presented in Section 1, some additional ones are included as follows:

- Q1 - In order to allow a better understanding of the factors that influence decisions, what are the main statistics, relationships, and correlations between the variables?
- Q2 - Are there any anomalies or unexpected behaviours that require attention from the central administration?

4.2 Data Understanding

We have collected the dataset from the *PGFN*. The dataset has been created by a team composed of domain experts and systems analysts, that gathered data from several *PGFN* data sources, including transactional and analytical systems. The available historical data of the R3Ds have been included, by considering the period of November 2018 and June 2020.

The dataset has 23 independent variables and a total amount of 70.780 R3Ds instances, containing a nationwide representation. Personal or business identification information and any other variable considered as sensitive were disregarded.

The independent variables regard the following information: (i) the request itself; (ii) the taxpayer; (iii) some of the taxpayer's relationship in the real world; (iv) information describing the debt (e.g., value, age, type, and situation); and (iv) some history of actions and situations associated with PGFN processes. The dataset also contains the analysis result of the request, i.e., the dependent variable indicating approval or rejection. For the sake of security and confidentiality, details regarding the variables are not mentioned in this work.

Each variable was analysed with respect to its main statistics, in order to observe the data distribution, maximum and minimum values, existence of outliers, temporal distribution and correlation with other variables. Tasks concerned with cleaning or transformation were verified and executed to assure a better model creation. Despite these issues, no missing values were detected, and no outliers were removed.

With respect to the target variable, the dataset has a 70/30% proportion between the two classes. Even though it's not a strong imbalance problem, we decided to apply some re-sampling techniques in order to observe the behaviour of the classification methods.

4.3 Data Preparation

The data preparation phase usually covers all activities to construct the final dataset from the collected data. The transformations made to the data involved the following actions:

- A normalization of all data in a standard scale between 0 and 1.
- Two pairs of variables presented a correlation coefficient equals to 1, i.e., they presented the same values for every dataset example. Since this situation was not expected, one variable of each pair was removed.

4.4 Modelling and Evaluation

In the modelling step, the classification models are created according to four experimental scenarios. For each scenario, the measures evaluated are Accuracy (ACC), AUC and Recall (REC).

The classification methods which have been applied are: Neural Networks, Naive Bayes, Random Forest and Support Vector Machines. All the models are trained using the default parameters from SciKit-Learn library. These parameters are as follows::

- Multilayer Perceptron: activation='relu', hidden_layer_sizes=(100,100), learning_rate='constant', max_iter=4000, solver='adam', and tol=0.0001.
- GaussianNB: priors='None', and var_smoothing='1e-09'.
- Random Forest: bootstrap=True, criterion='gini', min_samples_leaf=1, min_samples_split=2, and n_estimators=100.
- SVC: C=1.0, cache_size=200, decision_function shape='ovr', degree=3, kernel='rbf', shrinking=True, and tol=0.001.

The first scenario of the modelling step is a random stratified hold-out, using 80% of the available data for the training set and 20% for the test set.

The second scenario is built considering a 10-fold cross-validation, using the stratified shuffle split method. It is defined, for every iteration, the same 80% of the available data for the training set and 20% for the test set.

In the third scenario, we include balancing methods. Thus, at this one, a 10-fold cross-validation is executed applying a Random Undersampling class balancing method at each iteration. In the fourth scenario, a 10-fold cross-validation is executed applying the SMOTE technique at each iteration.

5 RESULTS AND DISCUSSION

The Data Understanding step brings some results, by means of answering the questions defined at the Business Understanding step (Section 4.1). Thus, in order to answer Q1, the correlation matrix has been plotted. It shows low correlation among most of the variables, except for two pairs of variables that presented a correlation coefficient equals to 1. One variable of each pair has been removed due to such high correlation.

In order to answer Q2, through some statistical analysis, it is possible to identify some anomalies. One of them regards 110 registered debts in a peculiar situation: each one of them is composed by more than 20 requests. This situation shows a possibility of using a R3D service just to postpone the debt's payment. Therefore, it requires attention from the central administration to better evaluate cases like that.

In the Modelling and Evaluation steps, the results obtained in the first scenario (random hold-out) are presented in Table 1. The highest scores for each measure are presented in bold. The first scenario brings these results: The Random Forest model showed the highest ACC and AUC among the evaluated models, followed by Neural Networks, SVM and Naive Bayes. Regarding REC, the SVM achieved a slightly (only 1%) higher rate than the Random Forest.

Table 1: Random stratified hold-out results.

Classifier	ACC	AUC	REC
Neural Networks	81%	88%	84%
Naive Bayes	60%	72%	49%
Random Forest	**88%**	**94%**	92%
SVM	69%	72%	**93%**

The results obtained in the second, third and fourth scenarios are presented in Table 2, including the mean and standard deviation obtained for each metric. The results after applying the class balancing techniques are presented with an arrow up when the measure has more than one percent of variation.

Table 2: Results before and after applying class balancing methods in 10-fold cross-validation scenarios.

	Unbalanced Scenario			**After Under Sampling**			**After SMOTE**		
Classifier	*Mean (Std Dev) ACC*	*Mean (Std Dev) AUC*	*Mean (Std Dev) REC*	*Mean (Std Dev) ACC*	*Mean (Std Dev) AUC*	*Mean (Std Dev) REC*	*Mean (Std Dev) ACC*	*Mean (Std Dev) AUC*	*Mean (Std Dev) REC*
Neural Networks	82% (±0,4%)	89% (±0,4%)	87% (±2,5%)	79% ↓ (±0,9%)	88% (±0,3%)	78% ↓ (±4,4%)	81% (±0,7%)	89% (±0,4%)	82% ↓ (±2,3%)
Naive Bayes	58% (±1,4%)	71% (±0,3%)	45% (±3,6%)	55% ↓ (±2,6%)	71% (±0,3%)	38% ↓ (±5,8%)	53% (±1,3%)	71% (±0,3%)	33% ↓ (±2,8%)
Random Forest	**88%** (±0,3%)	**95%** (±0,2%)	**92%** (±0,2%)	**87%** (±0,4%)	**94%** (±0,2%)	**87%** ↓ (±0,5%)	**88%** (±0,3%)	**94%** (±0,2%)	**91%** (±0,3%)
SVM	70% (±0,6%)	71% (±0,4%)	**92%** (±2,3%)	66% ↓ (±2,6%)	72% (±0,5%)	65% ↓ (±8,5%)	64% ↓ (±1,8%)	72% (±0,4%)	60% ↓ (±6,3%)

The second scenario (cross-validation with unbalanced data) confirms Random Forest with higher scores of ACC, AUC and REC, followed by the same order of models presented in the first scenario. Although SVM presented a lower ACC comparing to Neural Networks, it has a higher REC, and can be considered a better estimator to this study.

The third and fourth scenarios show that the application of Random Under Sampling and SMOTE techniques decreased the ACC and REC. It can be explained that, in both techniques, there is an increase on the representation of the negative class. The negative class is the minority class in this work. Then, the models tend to increase the predictions on this class, and the number of False Negatives and True Negatives also increase. Consequently, it may decrease ACC and REC. Weiss and Provost (2003) concluded that, when ACC is the priority performance measure, the best class distribution for learning tends to be near the natural class distribution, and when AUC is the priority performance metric, the best class distribution for learning tends to be near the balanced class distribution. With respect to standard deviations, the application of class balancing techniques did not cause significant changes.

6 CONCLUSIONS AND FUTURE WORK

This work has presented an approach to predict if R3Ds should be accepted or rejected. The evaluation of the created classification models indicates promising results mainly with regards to the Random Forest model. It achieves the best performance in terms of the most important measures considered in this work (ACC, AUC and REC). Cross-validation strategies have been used and show that the Random Forest model performs a good generalization. The class balancing techniques employed in this work do not improve the models' performance. This is due to the kinds of data we deal with, i.e., increasing the number of false negatives cases is costly than increasing the number of false positives cases.

The solution provided by this work may be useful to support decisions of the prosecutor who registers the result of a request application. It may not only increase the decision assertiveness but also decrease the response time.

As future work we point out some tasks to be done: (i) to experiment different hyper-parameters for the algorithms with the best performances (Random Forest and Neural Networks); (ii) to apply XGBoost method or other one evaluated with good performance on financial data (Pugliese et al., 2020); (iii) to reduce the number of variables used in training models, and then checking the impact of them on the observed created models; and (iv) to deploy the classification model which best fits the real PGFN scenario.

REFERENCES

Alpaydin, E. (2010). *Introduction to machine learning*. MIT press.

Battiston, P., Gamba, S., and Santoro, A. (2020). Optimizing Tax Administration Policies with Machine Learning. *University of Milan Bicocca Department of Economics, Management and Statistics Working Paper*, (436).

Bhavsar, H., and Panchal, M. H. (2012). A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 1(10), 185-189.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1), 1063-1095.

Branco, P., Torgo, L., and Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys* (CSUR), 49(2), 1-50.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (1999). The CRISP-DM user guide. In *4th CRISP-DM SIG Workshop in Brussels in March* (Vol. 1999).

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Dias, M., and Becker, K. (2017). Identificação de Candidatos à Fiscalização por Evasão do Tributo ISS. In *Proceeding of the 5th Symposium on Knowledge Discovery, Mining and Learning.*

Gonzalez, L., Angulo, C., Velasco, F., and Catala, A. (2005). Unified dual for bi-class SVM approaches. *Pattern Recognition*, 38(10), 1772-1774.

Harrington, P. (2012). *Machine learning in action.* Manning Publications.

He, H., and Ma, Y. (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. In *International Journal of Data Mining and Knowledge Management Process*.

Huang, J., and Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3), 299-310.

Ippolito, A., and Lozano, A. C. G. (2020). Tax Crime Prediction with Machine Learning: A Case Study in the Municipality of São Paulo. In *22nd International Conference on Enterprise Information Systems* (pp. 452-459).

Japkowicz, N., and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5), 429-449.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), 3-24.

López, C. P., Rodríguez, M. J. R., and Santos, S. L. (2019). Tax fraud detection through neural networks: an application using a sample of personal income taxpayers. *Future Internet*, 11(4), 86.

Mathews, J., Mehta, P., Kuchibhotla, S., Bisht, D., Chintapalli, S. B., and Rao, S. K. V. (2018). Regression analysis towards estimating tax evasion in Goods and Services Tax. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 758-761). IEEE.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, 1st edition.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Paliwal, M., and Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, 36(1), 2-17.

Pugliese, V. U., Hirata, C. M., and Costa, R. D. (2020). Comparing Supervised Classification Methods for Financial Domain Problems. In *22nd International Conference on Enterprise Information Systems* (pp. 440-451).

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).

Rezende, S. O. (2005). *Sistemas inteligentes: fundamentos e aplicações*. 1. ed. Editora Manole.

Russel, S., and Norvig, P. (2013). *Artificial intelligence: a modern approach*. Pearson Education Limited.

Silva, L. S., Carvalho, R. N., and Souza, J. C. F. (2015). Predictive models on tax refund claims-essays of data mining in brazilian tax administration. In *International Conference on Electronic Government and the Information Systems Perspective* (pp. 220-228). Springer, Cham.

Soares, G. V.; Cunha, R. C. L. V. (2020). Predição de Irregularidade Fiscal dos Contribuintes do Tributo ISS. In: *Anais do Simpósio Brasileiro de Banco de Dados*.

Souza, K. L. C. M. and Siqueira, M. (2020). A inteligência artificial na execução fiscal brasileira: limites e possibilidades. In *Revista de Direitos Fundamentais e Tributação*, *Volume 1*, *n. 3, p. 17-44*, 2020.

Weiss, G. M., and Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. In *Journal of artificial intelligence research*, 19, 315-354.

Wirth, R., and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39). London, UK: Springer-Verlag.