RBPO
REVISTA BRASILEIRA DE PLANEJAMENTO E ORÇAMENTO

Previsão do Cumprimento da Lei de Responsabilidade Fiscal por Municípios do Mato Grosso utilizando Random Forest

Micaelly Cristine de Moura Santos micaellycristine01@gmail.com
Profissional da área de dados, com graduação em Ciências Econômicas pela Universidade Federal de Mato Grosso (UFMT) e uma graduação em Análise e Desenvolvimento de Sistemas pelo Centro Universitário Cidade Verde (UniFCV). Micaelly também concluiu um MBA em Data Science e Analytics pela Universidade de São Paulo (USP/Esalq).

Introduction

The Fiscal Responsibility Law (LRF), established in 2000, aims to regulate responsible fiscal management in Brazilian municipalities, ensuring that expenses do not exceed revenues while promoting transparency and oversight of public accounts (Complementary Law No. 101, of May 4, 2000). The LRF sets limits for personnel expenses, indebtedness, and spending on health and education, aiming to promote financial sustainability.

Fiscal management in municipalities involves the administration of financial resources, covering revenue, expenses, and public investments. Unlike companies that generate income from sales, municipalities rely on taxes and government transfers and must comply with specific legal obligations, such as reporting to the Audit Court (Fernandes, 2011).

In Mato Grosso, several municipalities face challenges in meeting the LRF limits, impacting economic and social development. Compliance with the LRF is essential for responsible public resource management and fiscal balance.

This study aims to use the “Random Forest” model to predict compliance with the LRF in Mato Grosso’s municipalities for 2022, using historical data from 2017 to 2022. The model’s dependent variable is compliance

with the LRF, while the independent variables include all available data up to 2022, excluding the current year's "applied limit."

The "Random Forest" algorithm is chosen for its robustness in handling large datasets and nonlinear relationships (Breiman, 2001). Our findings indicate that the model accurately predicts LRF compliance, with key variables identified as indicators of financial health.

This paper includes sections on related studies, methodology, results, conclusions, and future work suggestions. It explores previous machine learning applications in prediction, outlines the data collection and treatment process, and discusses model evaluation metrics like accuracy and ROC curve. Finally, it suggests future research directions to enhance prediction capabilities.

Contributions from Related Studies

This study highlights crucial research on the topic, such as the work conducted by Cláudia Patrícia S. Pimentel, Fábio M. F. Lobato, and Antonio F. L. Jacob Jr., titled "Application of Machine Learning Techniques with Variable Selection in Forecasting Public Revenues of 8 Capitals: Report of the most relevant preliminary results with data from São Luís" (Pimentel, Lobato, & Jacob Jr., 2023). This study stands out by addressing the selection of exogenous variables in the context of the prerogatives of Art. 12 of the Fiscal Responsibility Law (LRF), seeking discoveries about the effects of price and quantity on the prediction of monthly revenues.

The results presented by the authors were obtained through the application of Machine Learning techniques, with an emphasis on the use of Random Forest. The clarity and relevance of the results highlight the practical utility of these methods in detecting anomalies, leading to effective actions by the Municipal Manager. The ability to identify significant events underscores the importance of these techniques in the efficient management of public finances.

Another relevant study addresses the application of Computer Science in public management, conducted by Pedro Fernandes Freitas and titled "Models for Forecasting ICMS Revenue in Rio de Janeiro Using Deep Learning" (Freitas, 2023). The focus of this study is to assist public management in the planning and execution of financial resources collected through ICMS in the state of Rio de Janeiro. Using Long Short-Term Memory (LSTM) Recurrent Neural Network models, the author aims to predict future ICMS revenue values, providing a crucial basis for the state's budget planning.

Despite the valuable contributions of these studies, it is crucial to note the scarcity of research employing Random Forest to make predictions in the context of the LRF. This observation highlights the lack of studies that explore this specific approach in the literature related to the LRF. Thus, the proposal to apply Random Forest techniques to predict LRF indicators stands out as a significant contribution to filling this gap in research, demonstrating the innovation and originality of this study in the context of LRF prediction.

Materials and Methods

The data used in this study were extracted from the State Court of Auditors of Mato Grosso (TCE-MT, 2017-2022) and the Brazilian Institute of Geography and Statistics (IBGE, 2017-2022). The TCE-MT provided information related to the compliance with the Fiscal Responsibility Law (LRF) by the municipalities of Mato Grosso, while the IBGE provided relevant socioeconomic and demographic data for the analysis. The combination of these data from TCE-MT and IBGE allowed for a comprehensive and in-depth analysis of the compliance with the LRF by the municipalities of Mato Grosso, considering not only the fiscal aspects but also the socioeconomic characteristics that may influence fiscal performance.

For data collection, the technique of web scraping was utilized using the Selenium library in Python. This technique enables the automatic collection of information from official websites, such as the State Court of Auditors of Mato Grosso (TCE-MT) and the Brazilian Institute of Geography and Statistics (IBGE). The analyzed variables include:

Territorial unit area
Guarantee Concession
Executive Branch Personnel Expenditure
Legislative Branch Personnel Expenditure
Net Consolidated Debt
Basic Education Development Index (IDEB)
Municipal Human Development Index (IDHM)
Mesoregion
Credit Operation
Per Capita GDP
Population
Primary Result
Average Monthly Salary

Python was employed for preprocessing and treating the collected data to prepare it adequately for the utilization of the Random Forest model. Preprocessing steps are essential for dealing with unstructured data and optimizing the performance and effectiveness of the model. The Random Forest model, on the other hand, was applied to make predictions based on historical data.

As explained by Breiman (2001), Random Forest is a machine learning algorithm consisting of a set of decision trees. Each tree is constructed from a random sample of the training data. During the tree construction, the algorithm evaluates different variables and divides the data into subsets based on these variables. This process is repeated several times to create a set of trees.

According to Breiman (2001), it is important to divide the sample into two parts, one for training and the other for testing, to assess the generalization capacity of the Random Forest model, i.e., its ability to make accurate predictions on previously unseen data. Separating a portion of the data for testing allows us to assess how the model performs on previously unseen data, enabling us to realistically estimate the model's performance in real-world situations. This sample division is crucial to avoid overfitting or underfitting problems of the model. In the case of the Random Forest model applied in this study, the available information covered the period from 2017 to 2021 and also included data for the year 2022. These data were used to train the algorithm and allowed the model to learn from historical information. With this training, the model was able to make an accurate prediction for the year 2022.

When evaluating the results of the Random Forest model, it is essential to consider the model's accuracy and the relative importance of explanatory variables. According to Liaw and Wiener (2002), the model's accuracy is a measure of the precision of predictions compared to actual values. The higher the accuracy, the better the model's performance in correctly classifying municipalities regarding compliance with the LRF.

In this work, the analysis of variable importance in the model is fundamental, as it allows determining which variables have the greatest impact on predicting compliance with the LRF. This assessment is generally performed through the "feature importance" module of Random Forest. Feature importance is a metric that quantifies each variable's contribution to the model's decision-making process. Variables with higher importance have a more significant effect on determining municipalities' compliance with the law.

It is worth noting that in the context of this study, multicollinearity among variables could be a challenge to be faced if the analysis method were, for example, a linear regression model. However, one advantage of the Random Forest model is its ability to handle highly correlated variables. This is because the model uses only a random subset of variables at each stage of model construction, reducing the probability of selecting highly correlated variables together.

Similarly, heteroscedasticity, a problem to be avoided when using prediction methods that seek to minimize the sum of squared residuals (OLS), has little impact on the Random Forest model. This is because decision trees are built in a non-parametric and non-linear manner and do not assume a linear relationship between independent and dependent variables. Thus, Random Forest offers greater flexibility in modeling error variation without the need to assume a specific distribution for the error. Furthermore, the Random Forest model is less susceptible to outliers and non-normal distributions, as trees are built iteratively and do not require specific assumptions about the data distribution.

In some data analysis models, as explained by Liaw and Wiener (2002), there are some measures of model fit quality, such as accuracy and the area under the Receiver Operating Characteristic (ROC) curve. These measures allow the assessment of the model's efficiency. For example, accuracy measures the proportion of correct classifications relative to the total number of observations. The area under the ROC curve is a measure of the model's discriminatory ability, indicating how well the model distinguishes between positive and negative classes.

In the specific case of the Random Forest model, being a non-parametric and non-linear machine learning model, there is no single universally adopted measure to assess its model fit quality. However, some commonly used metrics include accuracy, sensitivity, specificity, and the area under the ROC curve. All mentioned metrics are widely used in evaluating the quality of machine learning models, including the Random Forest model.

After collecting data from (TCE-MT, 2017-2022) and (IBGE, 2017-2022), they were merged into a single "DataFrame," enabling data analysis and processing for the application of a Machine Learning model. The data processing involved the following steps, as presented in Table 1.

Table 1: Data Processing Steps

Step	Description
Selection of Numeric Columns	Identified columns containing numerical data that needed to be converted to the "float" type. A function was defined to handle this treatment.
Conversion of Columns to "Category"	Selected columns to be converted to the "category" type. This included the "alerta_90" column containing "yes" or "no" values, and the "mesorregião" column representing the geographical mesoregion of each municipality. Conversion to the "category" type facilitated the interpretation and analysis of categorical variables.

Source: Self-prepared

In the application of the "Random Forest" model, the independent and dependent variables were defined. The independent variables, represented by the dataset X, include all municipality characteristics except for the columns "limite aplicado" of the EXECUTIVE and LEGISLATIVE branches. These variables will be used to make predictions based on the available data.

The dependent variable, represented by the variable Y, indicates whether the Fiscal Responsibility Law (LRF) was complied with by each municipality in 2022. Compliance criteria are based on the "limite aplicado" columns of the EXECUTIVE and LEGISLATIVE branches, where values must be less than or equal to 54% and 6%, respectively. The variable Y was defined as a binary variable, where the value 1 indicates that the LRF was complied with, and the value 0 indicates that it was not.

To ensure the accuracy and validity of the model results, as mentioned earlier, it was essential to divide the data into training and testing sets. For this purpose, the "train_test_split" method (TRAIN TEST SPLIT, 2023) available in the Sklearn (SCIKIT-LEARN, 2023) module was used. Sklearn is a Python library that provides a variety of machine learning algorithms, data preprocessing tools, model evaluation metrics, and utilities to facilitate machine learning application development (PEDREGOSA, 2011).

Firstly, null values of the dependent variable were removed from the X dataset to avoid testing the model with unknown data. The result of this filtering was stored in two variables: one containing only the rows of X where Y values were non-null, and another containing only the non-null values of Y.

Subsequently, the data was split into training and testing sets using the "train_test_split" method (TRAIN TEST SPLIT, 2023), with the resulting variables from the filtering process as parameters. It was defined that 35 municipalities (or 25% of the total dataset) would be allocated for testing, and the remaining would be used for training. To ensure that the separation between sets was always the same, the random number seed was fixed with the value 42 to guarantee reproducibility. Additionally, the sets were stratified based on the dependent variable (Y). Stratification is a method used in dividing data into training and testing sets, where the proportion of each class of the dependent variable is preserved in the resulting sets. This is important to avoid bias and ensure that the model is trained and evaluated in a representative manner for all classes of the dependent variable. Finally, the "shuffle" parameter was set to "True" to randomly shuffle the data before splitting into training and testing sets.

Next, two variables were created to handle missing values in the dependent variable. The first variable stored the rows of the X dataset where there were missing values in the dependent variable, while the second variable stored only the missing values of the dependent variable. These variables were used to separate the municipalities with missing values in the dependent variable for the training set. This approach ensures that these municipalities are not present in the test set and are used exclusively for model training.

Then, the training data was combined into a single dataset. This step was important to ensure that the model was trained with all available information, including data with missing values in the dependent variable and data without missing values. For this purpose, variables were created to receive the information from the filtered training datasets, combining data with missing values in the dependent variable with data without missing values. Thus, the model could use all this information during the training process. The number of variables in each set can be seen in Table 2.

Table 2: Number of Variables in Each Set

Set	Dimension
Training Set (X_train)	(106, 187) - 106 observations (municipalities) with 187 predictor variables each
Dependent Variable Training Set (Y_train)	(106,) - A vector with 106 values corresponding to the dependent variable
Test Set (X_test)	(35, 187) - 35 observations (municipalities) with 187 predictor variables each
Dependent Variable Test Set (Y_test)	(35,) - A vector with 35 values corresponding to the dependent variable

Source: Self-prepared

In the model application stage, a pipeline (PIPELINE, 2023) was employed, a tool available in "Sklearn," which consists of a sequence of data transformations to prepare input data for the machine learning algorithm. The pipeline comprised the following stages: (1) Column Transformer; (2) Imputation; (3) Normalization; (4) Dimensionality Reduction; (5) Classifier.

In stage (1), the "Column Transformer" was responsible for transforming categorical data present in a specific column of the dataset (COLUMN TRANSFORMER, 2023). Categorical columns were selected to apply the necessary transformations to each type of variable, choosing the "One Hot Encoder" technique, widely used in classification problems. This technique transforms each unique value into a new binary column, capturing important information contained in these categorical variables (ONE HOT, 2023).

For stage (2), Imputation, various imputation techniques were tested to fill missing values in the data. Three different imputation methods were tested: "Simple Imputer," "KNN Imputer," and "Iterative Imputer." The parameters of each method were optimized using "Bayes Search CV" (BAYES SEARCH CV, 2023) to select the best combination of hyperparameters for each technique.

Normalization techniques were tested in stage (3) to scale the data. Six normalization techniques were evaluated: "Min Max Scaler," "Robust Scaler," "Standard Scaler," "Normalizer," "Max Abs Scaler," and "Power Transformer." Similar to the previous stage, the normalization techniques were tested using "Bayes Search CV" to find the best combination of hyperparameters for each technique.

In stage (4), Dimensionality Reduction, two techniques were tested: "Principal Component Analysis" (PCA) and "Select KBest" to simplify the model by removing redundancies and noise in the data, resulting in a more compact and relevant representation of variables.

The fifth stage, Classifier, employed the "Random Forest Classifier" algorithm for classification and prediction. Additionally, the "Self-Training" technique was applied with "Random Forest" to enhance the predictive capability of the model. "Self-Training" is a semi-supervised learning approach where the initial model is trained using labeled data available.

To optimize the model's performance, various parameters of each element of the pipeline were tested using "Bayes Search CV," a hyperparameter search technique used to maximize the model's performance.

Following the creation of the pipeline, a parameter search space containing possible values for hyperparameters for each stage of the pipeline was defined. Then, a "Bayes Search CV" object was created, encompassing the entire transformation pipeline along with the possible hyperparameter values defined in the parameter search space. The objective was to find the configuration that maximized the accuracy of the model validation.

Within this object, 150 iterations were executed, testing and evaluating different combinations of hyperparameters. During each iteration, the algorithm adjusted the hyperparameter values based on the results obtained up

to that point, aiming to maximize the model's accuracy. The metric used to evaluate the model's performance was accuracy, which measures the proportion of correct predictions to the total predictions. Additionally, three-fold cross-validation (cv=3) was applied, dividing the training set into three parts, and the model was trained and evaluated on each of them.

After training the optimized model, predictions were made on the test and training sets to evaluate the model's performance. Predictions were obtained using the "predict" method of "Bayes Search CV," and variables were defined to store the model's predictions on the test and training sets.

Several evaluations were conducted to analyze the model's performance. Accuracy, AUC (Area Under the ROC Curve), precision, and recall metrics were calculated using appropriate functions from the "Sklearn" module. The importance of columns was determined using the "Select KBest" method and the "feature_importance" method of "Random Forest."

Identifying the most influential variables provides a deeper understanding of the factors affecting compliance with the limits established by the Fiscal Responsibility Law. These analyses contribute to assessing the model's predictive capabilities and understanding the underlying relationships between variables.

Results and Discussion

The best parameters for the model were selected using the "Bayes Search CV" method (Table 3). Different combinations of techniques and parameters were tested to optimize the classification system's performance with the "best_params" command.

Table 3: Selection of the best parameters by "Bayes Search CV"

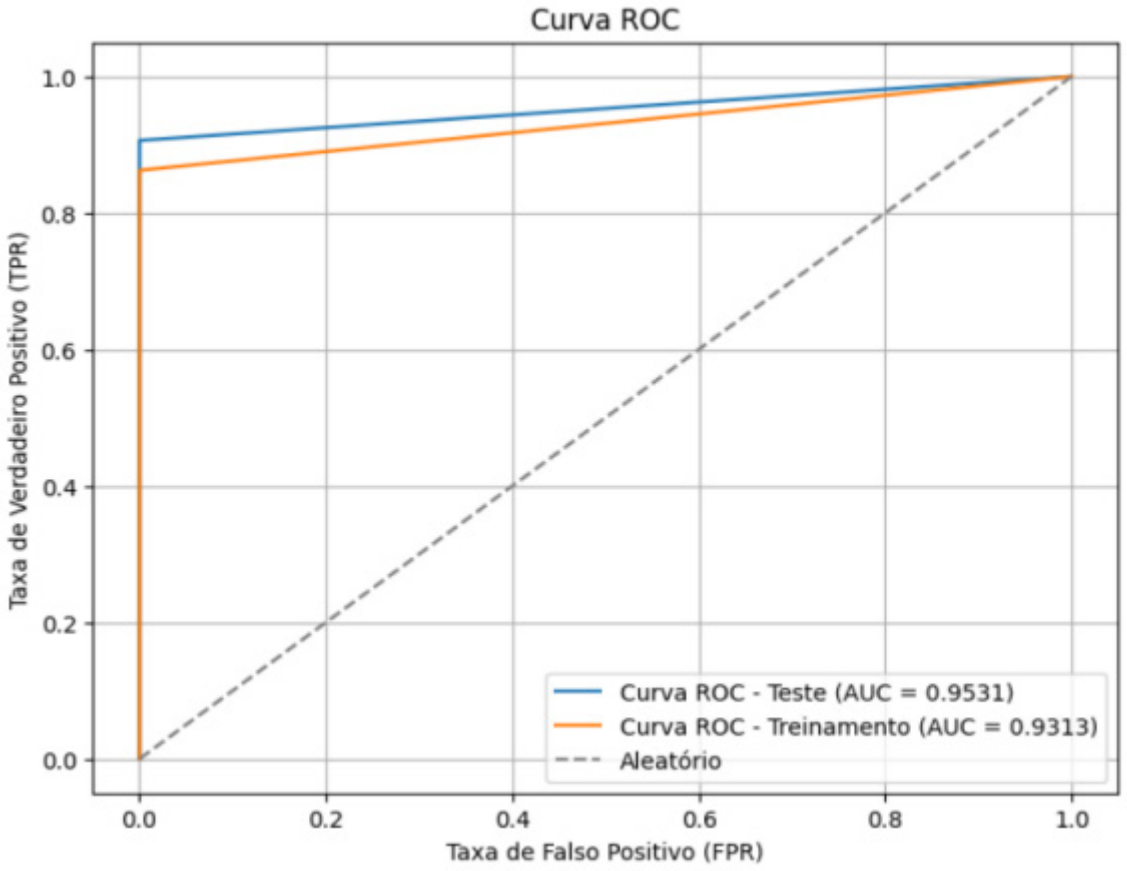
Parameter	Selection
Categorical Variables	One Hot Encoder
Imputation	Simple Imputer
Normalization	Power Transformer
Dimensionality Reduction	Select KBest
Classification Algorithm	Random Forest Classifier with Self-Training

Source: Self-prepared

The parameters listed in Table 3 represent the final configuration of the optimized model used for predictions based on the test data. The best validation result from the "best_score_" of "Bayes Search CV" during hyperparameter search was 0.7735, indicating that the model correctly classified, on average, 77.35% of the samples in cross-validation.

As shown in Figure 1, the model achieved 87.35% accuracy during training and 91.42% accuracy on the test set, demonstrating its effectiveness in predicting compliance with the Fiscal Responsibility Law (LRF) by municipalities. Additionally, Figure 1 presents the ROC curve, illustrating the model's performance in class discrimination.

Figure 1: ROC Curve for Model Performance Evaluation



Source: Self-prepared

The gray dashed line in Figure 1 represents random performance, indicating the model's lack of discriminative ability. The ROC curve's distance from this line towards the upper left corner signifies better performance. In this case, the curve is well away from the random performance line, demonstrating the model's good discriminative power. The AUC values are high, with 93.12% on the training set and 95.31% on the test set, indicating strong discrimination in both datasets.

Overall, the ROC curve and AUC values highlight the model's ability to effectively discriminate between classes, correctly classifying most instances in both training and test sets. This reinforces its predictive accuracy regarding compliance with the LRF by municipalities.

The model's precision and recall metrics are as follows: precision: 1.0 and recall: 0.90625. A precision of 1.0 indicates no false positives, while a recall of 0.90625 shows that 90.62% of positive samples were correctly identified, demonstrating the model's effectiveness in identifying relevant samples.

The most important columns of the prediction model, obtained through the Random Forest algorithm, are presented in Table 4 in descending order of importance. This result represents the relevance of the variables in the prediction process and allows understanding which characteristics have the greatest impact on determining compliance with the Fiscal Responsibility Law (LRF) by municipalities.

Table 4: Importance of Columns in Descending Order

Column	Importance
EXECUTIVE__2022__alerta_90_YES	0.30651088592188025
EXECUTIVE__2022__alerta_90_NO	0.2384279794132314
DEBT__2019__%_maximum_limit	0.16869928165185205
DEBT__2018__%_maximum_limit	0.1431549706726989
DEBT__2019__total_debt	0.13180806761345445
EXECUTIVE__2018__alerta_90_YES	0.011398814726882908
EXECUTIVE__2022__alerta_90_nan	0.0
LEGISLATIVE__2022__alerta_90_nan	0.0
DEBT__2022__total_debt	0.0
DEBT__2022__%_maximum_limit	0.0

Source: Self-prepared

The variables presented in Table 1 can be related to compliance with the Fiscal Responsibility Law (LRF) by municipalities in different ways. The first relevant variable is "EXECUTIVE__2022__alerta_90_YES", which indicates whether the Municipal Executive exceeded 90% of the legal limit established by the LRF in 2022. With an assigned importance of 30.65%, this variable draws attention to possible violations of established fiscal rules. When its value is "YES", it indicates an alert regarding compliance with the LRF by the Municipal Executive.

On the other hand, the variable "EXECUTIVE__2022__alerta_90_NO" is equally relevant, with an importance of 23.84%. It indicates that the Municipal Executive did not reach 90% of the legal limit established by the LRF in 2022. This information suggests compliance with the LRF since the municipality is below the established limit. The prediction of compliance with the LRF benefits from this variable, as it demonstrates that the Executive is aligned with the established fiscal guidelines.

Furthermore, variables related to debt also play a fundamental role in predicting compliance with the LRF by municipalities. The variable "DEBT__2019__%maximum_limit", with an importance of 16.86%, evaluates the Net Consolidated Debt in relation to the maximum limit allowed by the LRF in 2019. Values close to or above this limit may indicate a higher risk of non-compliance with the legislation. Similarly, the variable "DEBT__2018%_maximum_limit", with an importance of 14.31%, evaluates this proportion in the year 2018.

Another relevant variable is "DEBT__2019__total_debt", with an importance of 13.18%. This variable represents the total value of Net Consolidated Debt for a municipality in 2019. High debt values may indicate a situation of excessive indebtedness, which can compromise municipal finances and hinder compliance with the requirements established by the LRF.

Finally, the variable "EXECUTIVE__2018__alerta_90_YES", with an importance of 1.14%, is similar to the variable "EXECUTIVE__2022__alerta_90_YES". It indicates whether the Municipal Executive reached or exceeded 90% of the legal limit established by the LRF in 2018. Although it has a lower importance in the model, it still provides relevant information about compliance with the LRF by the Municipal Executive.

It is important to emphasize that the selection of these variables is based on their predictive ability, being chosen by the model according to their contribution to explaining the variability of the data and discriminating between classes. These variables provide insights into the key factors influencing compliance with the Fiscal Responsibility Law by municipalities.

Overall, the model developed in this study demonstrates promising performance in predicting compliance with the Fiscal Responsibility Law by municipalities. By leveraging machine learning techniques and optimizing hyperparameters through Bayes Search CV, the model achieves high accuracy, AUC, precision, and recall, indicating its ability to make accurate predictions. Furthermore, the identification of important features provides valuable insights into the factors influencing compliance with fiscal regulations, enabling policymakers to make informed decisions.

In conclusion, the model represents a valuable tool for assessing and monitoring compliance with fiscal responsibility regulations, assisting policymakers in identifying municipalities at risk of non-compliance and implementing targeted interventions to ensure fiscal sustainability and accountability. Further research and refinement of the model could enhance its predictive capabilities and contribute to more effective fiscal governance at the municipal level.

Final Remarks

In this study, a Random Forest model was employed to predict compliance with the Fiscal Responsibility Law (LRF) by municipalities in Mato Grosso in the year 2022. By analyzing the most important variables identified by the model, it was possible to understand which characteristics exerted the greatest influence on this prediction. The results showed that the model accurately estimated compliance with the limits established by the LRF by municipalities, highlighting the relevance of this approach for the analysis of public finances and fiscal compliance by municipalities.

Based on the results obtained, alerts related to total personnel expenses, indicating whether they are close to or exceeded 90% of the established limit, were the most influential characteristics in compliance with the LRF

limits by municipalities. Additionally, variables associated with municipal debts, such as the maximum limit and the total amount of net consolidated debt, also exerted significant influence on the prediction of compliance with the LRF. These indicators reflect the magnitude of the municipality's indebtedness and its ability to comply with its legal and fiscal obligations, making them essential indicators for assessing the financial health of municipalities and their compliance with LRF guidelines.

It is important to emphasize that this work serves as a starting point for the application of machine learning techniques in the area of public finances. The use of the Random Forest model provided valuable insights into which variables are most relevant for predicting compliance with the LRF.

Thus, the combination of machine learning techniques and data analysis can offer a more comprehensive and accurate approach to monitoring and analyzing the fiscal compliance of municipalities.

Suggestions for Future Work

While the application of Random Forest has proven effective in predicting Fiscal Responsibility Law (LRF) indicators in the context of this study, there are opportunities to explore other machine learning techniques. For example, models based on Artificial Neural Networks (ANNs) may be considered, given their ability to capture complex patterns in datasets. Additionally, time series approaches, such as ARIMA models or Recurrent Neural Networks (RNNs), may offer additional insights into financial data modeling. Investigating the applicability of these techniques in conjunction with Random Forest could further enrich prediction analyses, providing a more comprehensive understanding of the dynamics of public revenues in compliance with the LRF.

References

ACCURACY SCORE. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html. Accessed on May 1, 2023.

BAYES SEARCH CV. Available at: https://scikitoptimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html. Accessed on May 1, 2023.

BERGSTRA, J., Yamins, D., & Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Proceedings of the 30th International Conference on Machine Learning (ICML-13) (pp. 115-123).

BRASIL. Lei Complementar nº 101, de 4 de maio de 2000. Estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências. Brasília, DF: Presidência da República, 2000.

BRASIL. Ministério da Economia. Secretaria do Tesouro Nacional. Manual de Demonstrativos Fiscais: Aplicado à União, aos Estados, ao Distrito Federal e aos Municípios. Brasília, DF: Secretaria do Tesouro Nacional, 2020.

BREIMAN, L. (2001). "Random Forest"s. Machine learning, 45(1), 5-32.

COLUMN TRANSFORMER. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html. Accessed on May 1, 2023.

FEATURE IMPORTANCES. Available at: https://scikitlearn.org/stable/auto_examples/ensemble/plot_forest_importances.html. Accessed on May 1, 2023.

FERNANDES, B. H. P. (2011). Lei de Responsabilidade Fiscal: Análise de Impactos no Setor Público Municipal. Dissertation (Master's in Accounting and Control) - University of São Paulo, São Paulo.

FREITAS, P. F. (2023). Models for Forecasting ICMS Revenue in Rio de Janeiro Using Deep Learning. Bachelor's thesis, Institute of Exact Sciences, Bachelor's in Information Systems, Federal University of Juiz de Fora.

GUJARATI, D. N., & Porter, D. C. (2009). Basic Econometrics. McGraw-Hill.

HAIR Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2014). Multivariate Data Analysis (6th ed.). Bookman.

HOSMER Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

IBGE. Cidades, 2017-2022. Available at: https://cidades.ibge.gov.br. Accessed on May 1, 2023.

ITERATIVE IMPUTER. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html. Accessed on May 1, 2023.

JOHNSON, R. A., & Wichern, D. W. (2017). Applied Multivariate Statistics (12th ed.). São Paulo: Cengage Learning.

JUPYTER NOTEBOOK. Available at: https://jupyter.org. Accessed on May 1, 2023.

LIAW, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22.

KNN IMPUTER. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.impute.KNNImputer.html. Accessed on May 1, 2023.

MAX ABS. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html. Accessed on May 1, 2023.

MENARD, S. (2002). Applied Logistic Regression Analysis. Sage.

MIN MAX. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html. Accessed on May 1, 2023.

NORMALIZER. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html. Accessed on May 1, 2023.

NUMPY. Available at: https://numpy.org. Accessed on May 1, 2023.

ONE HOT. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. Accessed on May 1, 2023.

PEDREGOSA, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

PANDAS. Available at: https://pandas.pydata.org. Accessed on May 1, 2023.

PCA. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.decomposition.PCA.html. Accessed on May 1, 2023.

Pimentel, Lobato, & Jacob Jr. (2023). Application of Machine Learning Techniques with Variable Selection in Forecasting Public Revenues of 8 Capitals: Report of the most relevant preliminary results with data from São Luís. In XIV Computer on the Beach, March 30 to April 1, 2023, Florianópolis, SC, Brazil.

PIPELINE. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html. Accessed on May 1, 2023.

PIRES, J. A. (2009). Lei de Responsabilidade Fiscal: Comentários e Casos Práticos. São Paulo: Atlas.

POWER TRANSFORMER. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html. Accessed on May 1, 2023.

PRECISION SCORE. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.metrics.precision_score.html. Accessed on May 1, 2023.

PYTHON. Available at: https://www.python.org. Accessed on May 1, 2023.

"RANDOM FOREST". Available at: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed on May 1, 2023.

RECALL SCORE. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.metrics.recall_score.html. Accessed on May 1, 2023.

ROBUST SCALER. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html. Accessed on May 1, 2023.

ROC AUC SCORE. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html. Accessed on May 1, 2023.

ROC CURVE. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.metrics.roc_curve.html. Accessed on May 1, 2023.

SCIKIT-LEARN. Available at: https://scikit-learn.org/stable/. Accessed on May 1, 2023.

SELECT KBEST. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html. Accessed on May 1, 2023.

SELENIUM. Available at: https://selenium-python.readthedocs.io. Accessed on May 1, 2023.

SELF TRAINING. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.semi_supervised.SelfTrainingClassifier.html. Accessed on May 1, 2023.

SIMPLE IMPUTER. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html. Accessed on May 1, 2023.

STANDARD SCALER. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Accessed on May 1, 2023.

TCE-MT. Limites da LRF, 2017-2022. Available at: https://cidadao.tce.mt.gov.br/home/limitesLrf. Accessed on May 1, 2023.

TRAIN TEST SPLIT. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.train_test_split. Accessed on May 1, 2023.

VAN ROSSUM, G., & Drake, F. L. (2009). Python 3 Reference Manual. Create Space.

Forecasting Compliance with the Fiscal Responsibility Law by Municipalities in Mato Grosso Using Random Forest

Resumo Este estudo desenvolve um modelo preditivo usando Random Forest para prever o cumprimento da Lei de Responsabilidade Fiscal (LRF) pelos municípios de Mato Grosso em 2022. A análise foi baseada em dados de 2017 a 2021 e informações disponíveis para 2022. A variável dependente foi o cumprimento da LRF, e as variáveis independentes incluíram dados disponíveis de 2017 a 2022, exceto o valor do "limite aplicado" em 2022. O modelo alcançou 91,42% de precisão, demonstrando alta capacidade preditiva. As informações foram obtidas do IBGE e do TCE-MT. Limitações envolvem a exclusão de certas variáveis e dependência da precisão dos dados. O uso do Random Forest facilita a análise das variáveis que afetam a conformidade fiscal, apoiando o monitoramento financeiro municipal e a tomada de decisões.

Palavras-chave: Lei de Responsabilidade Fiscal; Aprendizado de Máquina; Random Forest; Previsão; Finanças Municipais.

Abstract: This study develops a predictive model using Random Forest to forecast compliance with the Fiscal Responsibility Law (LRF) by municipalities in Mato Grosso in 2022. The analysis was based on data from 2017 to 2021 and information available for 2022. The dependent variable was LRF compliance, and independent variables included data from 2017 to 2022, except the "applied limit" value in 2022. The model achieved 91.42% accuracy, showing high predictive capability. Data were sourced from IBGE and TCE-MT. Limitations include the exclusion of some variables and reliance on data accuracy. The use of Random Forest facilitates the analysis of factors affecting fiscal compliance, supporting municipal financial monitoring and decision-making.

Keywords: Fiscal Responsibility Law; Machine Learning; Random Forest; Prediction; Municipal Finances.