SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

APPLICATION OF MACHINE LEARNING MODELS IN THE CONTEXT OF BOLSA FAMÍLIA: AN APPLIED STUDY IN RIO GRANDE DO NORTE AND PARAÍBA

APLICAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA NO CONTEXTO DO BOLSA FAMÍLIA: UM ESTUDO APLICADO NO RIO GRANDE DO NORTE E PARAÍBA

APLICACIÓN DE MODELOS DE APRENDIZAJE DE MÁQUINA EN EL CONTEXTO DE BOLSA FAMÍLIA: UN ESTUDIO APLICADO EN RIO GRANDE DO NORTE Y PARAÍBA

Luiz Fernando da Cunha Silva[1], Maria Eduarda Bandeia Hora de Vasconcelos[2], Verônica Maria Lima Silva[3], Samara Martins Nascimento Gonçalves[4]

e768077
https://doi.org/10.47820/recima21.v7i6.8077

PUBLISHED: 06/2026

ABSTRACT

This study investigates the application of Machine Learning techniques to support decision-making in public administration, focusing on predicting family eligibility for the Bolsa Família Program in the Brazilian states of Rio Grande do Norte and Paraíba. De-identified microdata from the Cadastro Único database (2016–2018) were used to train and evaluate predictive models. After data preprocessing, class balancing with the Synthetic Minority Over-sampling Technique, and dimensionality reduction using SelectKBest, five ML models were implemented: K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, and a Recurrent Neural Network. The results show that tree-based models, neural networks, and Support Vector Machines achieve robust performance in both states, with accuracy values of up to 90%. Random Forest, XGBoost, and Recurrent Neural Networks were more stable in RN, while Support Vector Machine achieved the best performance in PB, indicating regional differences in data separability. Feature selection effectively reduced model complexity without loss of accuracy, highlighting income, household structure, access to basic services, and family size as key determinants of eligibility. Overall, the findings confirm the feasibility of using ML models as decision-support tools for social policy management, contributing to more efficient monitoring and allocation of public resources. Despite limitations related to data availability and scope, this study provides empirical evidence of the potential of Artificial Intelligence to support evidence-based policymaking in a transparent and ethical manner.

KEYWORDS: *Machine Learning. Public Administration. Bolsa Família.*

RESUMO

Este estudo investiga a aplicação de técnicas de Aprendizado de Máquina para apoiar a tomada de decisão na administração pública, com foco na predição da elegibilidade de famílias ao Programa Bolsa Família nos estados brasileiros do Rio Grande do Norte e da Paraíba. Microdados desidentificados da base do Cadastro Único (2016–2018) foram utilizados para treinar e avaliar modelos preditivos.

[1] Bachelor in Information Systems from the Universidade Federal Rural do Semi-Árido (UFERSA), Collaborative Researcher at the Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos-SP.
[2] Bachelor's degree student in Data Science and Artificial Intelligence at the Universidade Federal da Paraíba (UFPB), João Pessoa-PB.
[3] PhD in Electrical Engineering from the Universidade Federal de Campina Grande (UFCG). Professor at the Universidade Federal da Paraíba (UFPB), João Pessoa-PB.
4 PhD in Computer Science from the Universidade Federal do Ceará (UFC). Professor at the Universidade Federal Rural do Semi-Árido (UFERSA), Angicos-RN.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

Após o pré-processamento dos dados, o balanceamento de classes com a técnica Synthetic Minority Over-sampling Technique (SMOTE) e a redução de dimensionalidade utilizando o método SelectKBest, cinco modelos de Aprendizado de Máquina foram implementados: K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost e Rede Neural Recorrente (RNN). Os resultados demonstram que modelos baseados em árvores, redes neurais e máquinas de vetores de suporte alcançaram desempenho robusto em ambos os estados, com acurácia de até 90%. Random Forest, XGBoost e Redes Neurais Recorrentes mostraram-se mais estáveis no contexto do Rio Grande do Norte, enquanto o Support Vector Machine apresentou o melhor desempenho na Paraíba, indicando diferenças regionais na separabilidade dos dados. A seleção de atributos reduziu efetivamente a complexidade dos modelos sem perda de desempenho, destacando renda, estrutura domiciliar, acesso a serviços básicos e tamanho familiar como fatores determinantes da elegibilidade. De modo geral, os resultados confirmam a viabilidade do uso de modelos de Aprendizado de Máquina como ferramentas de apoio à gestão de políticas sociais, contribuindo para um monitoramento mais eficiente e para uma melhor alocação de recursos públicos. Apesar das limitações relacionadas à disponibilidade e ao escopo dos dados, este estudo fornece evidências empíricas sobre o potencial da Inteligência Artificial para apoiar a formulação de políticas públicas baseadas em evidências, de forma transparente e ética.

PALAVRAS-CHAVE: Aprendizado de Máquina. Administração Pública. Bolsa Família.

RESUMEN

Este estudio investiga la aplicación de técnicas de aprendizaje automático para apoyar la toma de decisiones en la administración pública, centrándose en la predicción de la elegibilidad familiar para el Programa Bolsa Família en los estados brasileños de Rio Grande do Norte y Paraíba. Se utilizaron microdatos anonimizados de la base de datos Cadastro Único (2016–2018) para entrenar y evaluar modelos predictivos. Tras el preprocesamiento de datos, el balanceo de clases con la técnica Synthetic Minority Over-sampling Technique y la reducción de dimensionalidad mediante SelectKBest, se implementaron cinco modelos de aprendizaje automático: K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost y una red neuronal recurrente. Los resultados muestran que los modelos basados en árboles, las redes neuronales y las máquinas de vectores de soporte logran un rendimiento robusto en ambos estados, con valores de precisión de hasta el 90 %. Random Forest, XGBoost y las redes neuronales recurrentes fueron más estables en RN, mientras que la máquina de vectores de soporte logró el mejor rendimiento en PB, lo que indica diferencias regionales en la separabilidad de los datos. La selección de características redujo eficazmente la complejidad del modelo sin pérdida de precisión, destacando los ingresos, la estructura familiar, el acceso a servicios básicos y el tamaño de la familia como determinantes clave de la elegibilidad. En general, los resultados confirman la viabilidad del uso de modelos de aprendizaje automático como herramientas de apoyo a la toma de decisiones para la gestión de políticas sociales, contribuyendo a una supervisión y asignación más eficientes de los recursos públicos. A pesar de las limitaciones relacionadas con la disponibilidad y el alcance de los datos, este estudio proporciona evidencia empírica del potencial de la inteligencia artificial para respaldar la formulación de políticas basadas en evidencia de manera transparente y ética.

***PALABRAS CLAVE**: Aprendizaje automático. Administración pública. Bolsa Família.*

INTRODUCTION

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

The use of Artificial Intelligence (AI) techniques, and in particular Machine Learning (ML) models, has been transforming the public administration landscape on an international scale, promoting innovative solutions that strengthen oversight, fraud prevention, and the promotion of transparency in public policies (Desordi; Bona, 2020). In Brazil, this scenario becomes even more relevant, especially in the context of the Bolsa Família Program (PBF), which benefits millions of Brazilian families and therefore requires robust mechanisms for monitoring, registry verification, and decision-making support (Azevedo *et al*., 2021). Thus, it is important to explore the impact that tools such as algorithms and computational methods can play in improving administrative efficiency, automating bureaucratic tasks, and enhancing the reliability of governmental decision-making processes.

In this context, the use of ML presents significant potential to handle large volumes of administrative data, identify complex patterns, and anticipate behaviors that could be difficult to detect manually. As highlighted by Géron (2021), ML is a field of study in Computer Science that seeks to develop algorithms capable of learning relationships from data, thus becoming a promising alternative for automating processes that are traditionally bureaucratic and susceptible to human inconsistencies. Consequently, computational learning techniques can contribute not only to the analysis of information from the *Cadastro Único*, but also to the improvement of public policies aimed at the management of social benefits.

In the specific case of income transfer programs, such as the PBF, the application of ML shows potential to support the identification of eligible families, registry monitoring, and the analysis of inconsistencies in administrative records. However, the adoption of these techniques requires additional care, since automated decisions can generate relevant social impacts (Alshehhi; Cheaitou; Rashid, 2022). Therefore, it is essential that the models employed are rigorously evaluated, interpretable, and aligned with ethical principles and social responsibility.

While AI usage in the public sector has advanced, gaps remain in applying predictive models to *Cadastro Único* microdata and cross-state comparisons. Most studies prioritize fraud detection over individual family classification using socioeconomic variables (Azevedo *et al*., 2021). Thus, evaluating ML model performance and generalization across regional scenarios using real program data is essential.

In view of the above, and considering the importance of building and validating future tools that deal with AI, this work proposes an applied study on the implementation and evaluation of different ML models aimed at predicting eligibility for the PBF, considering the states of Rio Grande do Norte (RN) and Paraíba (PB). The specific objectives include: (i) selecting and implementing models suitable for the classification problem; (ii) defining and applying metrics for comparing the algorithms; (iii) analyzing the performance of the models based on data from each state; and (iv)

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

discussing the potentialities, limitations, and practical implications of the use of ML in public administration.

The paper is structured as follows: Section 1 presents the related studies; Section 2 describes the adopted methodology; Section 3 presents and discusses the obtained results; and Section 4 brings together the conclusions, limitations, and suggestions for future work.

1. RELATED WORK

The use of AI techniques, particularly ML models, has expanded significantly in the public sector, driven by the increasing availability of administrative data and the need to improve governmental decision-making processes (Desordi; Bona, 2020; Alshehhi; Cheaitou; Rashid, 2022). This section presents a series of works related to this research, addressing topics that range from the importance of using AI techniques in the context of public administration to ethical issues, which become particularly relevant when considering the data context of the PBF.

Azevedo *et al*. (2021) propose an approach based on Benford's Law to assist in the detection of fraud in governmental social assistance programs, with a specific focus on the PBF. To this end, using data available from the Brazilian Government Transparency Portal, the authors developed a method that analyzes the conformity of municipal PBF payments with the distribution predicted by Benford's Law, thus combining statistical hypothesis and summation tests to identify municipalities with payments outside the expected pattern. Based on this, the authors highlight that only when the values are aggregated by municipality does adherence to Benford's Law become evident, revealing regions suspected of fraud that could be further investigated.

Alshehhi, Cheaitou, and Rashid (2022) presented a systematic literature review on frameworks for the adoption of AI practices in the public sector, including applications that use ML to improve governmental processes. In this way, the authors classified the adoption of AI in this sector into four categories, including: regulatory, normative, application, and evaluative, highlighting implementation barriers, ethical challenges, and the need for citizen-focused strategies.

Within the same line of research, Desordi and Bona (2020) discussed how AI can contribute to efficiency in public administration by analyzing real cases of the use of intelligent systems by Brazilian governmental agencies. Based on this, the study explored initiatives such as the use of algorithms for data cross-checking, identification of irregularities, and process optimization in courts of accounts and in the oversight of parliamentary expenditures. Thus, by addressing the impact of AI on reducing bureaucracy and improving internal control, the authors reinforce that ML and intelligent systems are fundamental tools for strengthening oversight, fraud prevention, and even for promoting transparency in public policies.

Tan *et al*. (2023) presents an interpretative analysis of the factors that influence the adoption of AI and algorithm-based decision-making for fraud detection in public policies. The article identifies

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

13 variables organized in a hierarchical chain with five layers, namely: trust, interoperability, benefits, data governance, and digital governance. According to the authors, the relevance of building trust between citizens and administrators is evident as a way to facilitate the acceptance of tools such as AI for detecting fraud in social benefits. Furthermore, barriers such as technical limitations, concerns about transparency, governance, and algorithmic explainability are identified as the main sources of discussion regarding the challenges faced by ML initiatives in the public sector.

Finally, Caiza *et al*. (2024) conducted a review on the impact of AI on governmental decision-making, mapping recent advances and the challenges faced in adopting this practice. According to the authors, AI can automate bureaucratic tasks, improve administrative efficiency, enhance transparency, and foster public trust, which are fundamental factors for the management of income transfer programs. In addition, the study highlights important barriers such as algorithmic bias, transparency, social acceptance, accountability, and the need for robust ethical and regulatory guidelines.

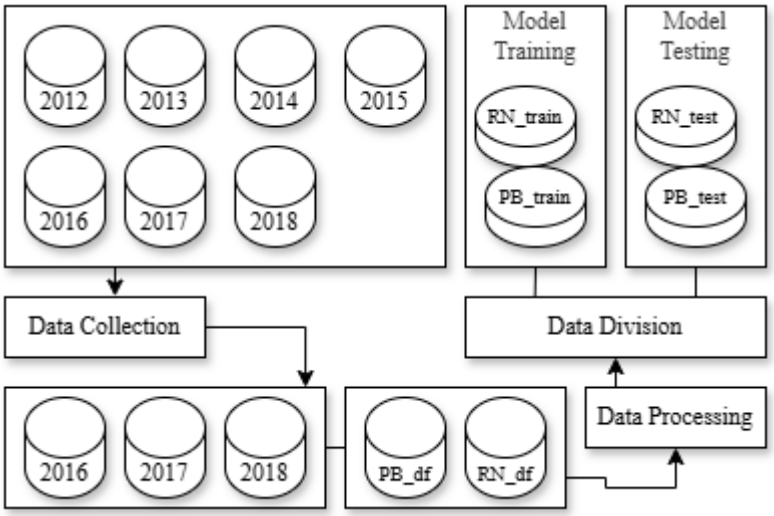
In light of the presented overview, it can be observed that although there have been relevant advances in the use of AI techniques in public administration, the literature still lacks studies that apply ML algorithms directly to microdata from the Cadastro Único to perform individual household classification, especially in comparative analyses across different state contexts. The predominance of research focused on fraud detection, aggregated analyses, or conceptual discussions reveals a methodological and applied gap in the use of predictive models to support the management of the PBF. In this sense, the main contribution of this study is to fill this gap by implementing, comparing, and evaluating different ML models on household data from the states of RN and PB, demonstrating the feasibility and potential of these techniques to enhance socioeconomic analysis and support more effective decision-making in public administration.

2. METHODOLOGY

This study was developed from the analysis of data from the Unified Registry of Beneficiary Families of the PBF, also known as the Cadastro Único, focusing on the states of RN and PB. To this end, the methodological process, illustrated in Figure 1, involved everything from organizing and cleaning the original data to the necessary pre-processing for the application of ML models, respecting the characteristics of each state context. The adopted methodology is grounded in a quantitative and experimental approach, supported by the use of computational techniques for data processing and modeling, as detailed in the following subsections.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

Figure 1. Flowchart of the adopted methodological process, from the collection of Cadastro Único data to the training and evaluation of the ML models



Source: Authors (2026).

2.1. Data Collection

The initial stage of this proposal consisted of the collection and organization of microdata related to families registered in the Cadastro Único for Federal Government Social Programs. For this purpose, the data were obtained through the Federal Government Open Data Portal, which provides de-identified datasets segmented by year, covering in total the period from 2012 to 2018.

For this work, the files covering the period from 2016 to 2018 were selected, and this choice was motivated by two main factors. First, the most recent data made publicly available correspond to this time interval, as no updated public datasets for subsequent years were available at the time this study was conducted. Second, the large volume of records presents in the Cadastro Único databases-imposed limitations related to the available computational resources for conducting the experiments. Thus, selecting a specific time interval made it possible to carry out the processing, training, and evaluation of the ML models without compromising the quality of the analyses performed. Accordingly, for each state considered, the data were compiled and loaded into dataframes for subsequent processing, resulting in a total of 296,062 records for RN and 182,184 for PB.

The collected datasets include a broad set of 32 variables related, as presented in Table 1, to family identification, household characteristics, family composition and dynamics, as well as socioeconomic information such as income, access to basic services, and housing conditions. The indicator of participation in the PBF, *marc_pbf*, was used as the target variable in the modeling stages. The diversity and granularity of these variables make the Cadastro Único a rich source for

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

the application of ML techniques, enabling detailed analyses of family profiles and the conditions that influence their eligibility for the program.

Table 1. Variables of the registered families database in the Cadastro Único

Variable	Description
cd_ibge	Municipality code (IBGE)
dat_cadastramento_fam	Family registration date
dat_alteracao_fam	Last family record update
vlr_renda_media_fam	Per capita income (R$)
dat_atualizacao_familia	Sensitive data update date
dat_atual_fam	Last modification date
cod_local_domic_fam	Household location
cod_especie_domic_fam	Household type
qtd_comodos_domic_fam	Number of rooms
qtd_comodos_dormitorio_fam	Number of bedrooms
cod_material_piso_fam	Floor material
cod_material_domic_fam	Wall material
cod_agua_canalizada_fam	Piped water availability
cod_abaste_agua_domic_fam	Water supply source
cod_banheiro_domic_fam	Bathroom availability
cod_escoa_sanitario_domic_fam	Sanitation system type
cod_destino_lixo_domic_fam	Waste disposal method
cod_iluminacao_domic_fam	Lighting type
cod_calcamento_domic_fam	Pavement condition
cod_familia_indigena_fam	Indigenous family flag
ind_familia_quilombola_fam	Quilombola family flag
nom_estab_assist_saude_fam	Health facility name
cod_eas_fam	Health facility code (EAS/MS)
nom_centro_assist_fam	Social assistance center
cod_centro_assist_fam	Assistance center code
Ind_parc_mds_fam	Traditional/specific group flag
peso.fam	Family sample weight
id_familia	Family unique identifier
estrato	City size classification
classf	Administrative area class
qtde_pessoas	Family size

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

marc_pbf	Bolsa Família indicator

Source: Authors (2026).

2.2. Data Preprocessing

The preprocessing stage was essential to ensure the quality, consistency, and suitability of the data for the subsequent modeling steps. Initially, an exploratory analysis was conducted to identify variables with a high proportion of missing values, formatting inconsistencies, or low relevance to the object of study. Based on these criteria, the variables *dat_cadastramento_fam*, *dat_atualizacao_familia*, *nom_estab_assist_saude_fam*, *dat_alteracao_fam*, *ind_parc_mds_fam*, *cod_eas_fam*, *nom_centro_assist_fam* e *cod_centro_assist_fam* were removed from the dataset, as their incompleteness and specific characteristics did not contribute to the classification process.

In addition to attribute removal, techniques for handling extreme values were also applied. In this sense, considering the variable *qtde_pessoas*, which represents the number of family members, an asymmetric distribution was observed, with a few records containing very high values. To mitigate the impact of these outliers and preserve data representativeness, all records with more than six people were grouped into a single category. This approach made it possible to smooth the variable's distribution, reduce variability caused by rare cases, and improve model stability during training.

2.3. Data Splitting

After the preprocessing stage, the data were divided into two subsets, with 70% allocated for training and 30% for the testing phase, following a practice widely adopted in ML experiments. In this way, this split aims to ensure that the models are evaluated on data not seen during training, allowing for a more realistic estimation of their generalization capability.

Furthermore, during the analysis of the distributions, a significant imbalance was identified between the classes of the target variable, with approximately 56% of the records for RN classified as beneficiaries and 43% as non-beneficiaries, and 61% of the records for PB being beneficiaries and 39% non-beneficiaries. Since classification models may exhibit biased performance when trained on imbalanced datasets, especially with respect to the recall of the minority class, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. This technique generates synthetic samples of the underrepresented class, contributing to the balancing of the training set and to improving model recall (Chawla *et al*., 2002). As a result of the balancing process, the RN dataset came to contain 117,564 records in each class, while for PB, the balanced dataset totaled 130,058 records per class.

For illustrative purposes, Table 2 presents a simplified example of the SMOTE technique applied to a hypothetical subset of the training set. This subset comprises records from two classes, where 0 represents non-beneficiary families and 1 represents beneficiary families. As shown in Table

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

2 (Panel A), class 1 is initially underrepresented relative to class 0, characterizing an imbalanced scenario. Upon applying SMOTE (Panel B), new synthetic instances of the minority class are generated through interpolation between existing samples. This procedure balances the dataset while preserving the distribution in the feature space and avoiding the simple duplication of records (Chawla *et al*., 2002).

Table 2. Training set examples: before and after SMOTE

Income	No. of People	Rooms	Class
Panel A: Before SMOTE			
1500	4	5	0
1200	3	4	0
600	5	2	1
Panel B: After SMOTE			
1500	4	5	0
1200	3	4	0
600	5	2	1
700	4	3	1

Source: Authors (2026).

2.4. Machine Learning Models

For the modeling stage, different ML algorithms widely used in classification tasks were selected, allowing the comparison of approaches with distinct learning characteristics. Among them is the Random Forest (RF), an ensemble method based on the combination of multiple decision trees, whose objective is to increase prediction robustness and reduce overfitting issues through the aggregation of independent models (Géron, 2021). The Support Vector Machine (SVM) was also employed, which identifies an optimal hyperplane to separate classes in the feature space, showing good performance even in scenarios with linear separability (Géron, 2021).

In addition, the K-Nearest Neighbors (KNN) algorithm was used, an instance-based method that classifies new observations by considering their proximity to the nearest neighbors, being sensitive to the local distribution of the data (Faceli *et al.*, 2025). Complementarily, Extreme Gradient Boosting (XGBoost) was evaluated, a model based on gradient boosting that stands out for its high computational efficiency, ability to handle complex interactions among features, and strong performance on large datasets (Chen; Guestrin, 2016).

Finally, this study also included the use of a Recurrent Neural Network (RNN), which has an architecture capable of modeling dependencies among attributes by using recurrent connections between its units (Koutník; *et al.*, 2014). Although traditionally applied to sequential data, an RNN

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

can capture complex and nonlinear patterns present in structured datasets, offering a complementary approach to classical classification methods. The inclusion of this model made it possible to evaluate the potential of deep neural network–based techniques in the context of the PBF, thereby expanding the diversity of the investigated algorithms.

2.5. Model Training

The experiments in this study were conducted using two complementary training approaches, with the aim of evaluating the impact of dimensionality and feature relevance on algorithm performance. In the first approach, all remaining attributes after preprocessing were used directly in the models, allowing the analysis of classifier behavior in a scenario with greater informational diversity. In the second approach, the SelectKBest method from the Scikit-learn library was employed to select the variables with the highest statistical relevance to the target variable, reducing dataset dimensionality and seeking to improve the algorithms' generalization capability (Tislenko; Gaidel; Kupriyanov, 2022).

In both approaches, the models were trained exclusively using the training set previously balanced via SMOTE, ensuring greater class equity during learning. Model evaluation was performed using the test set, which was preserved without intervention to ensure an unbiased assessment of predictive capability. To measure performance, classical classification metrics were adopted, including accuracy, precision, recall, and F1-score, which allowed a comprehensive evaluation of both the balance between false positives and false negatives and the overall effectiveness of each model in identifying beneficiary and non-beneficiary families.

Accuracy measures the total proportion of correct classifications relative to the total number of samples, providing an overall view of model performance (Géron, 2021). However, in scenarios with possible class imbalance, this metric alone may mask deficiencies in identifying the class of greatest interest (Sujon; *et al*., 2025). Furthermore, precision indicates the proportion of instances correctly classified as positive relative to the total number of positive predictions made by the model (Sujon; *et al*., 2025). In the context of this study, this metric is associated with the administrative cost of classifying non-eligible families as beneficiaries, which may lead to improper allocation of resources.

Recall, in turn, measures the model's ability to correctly identify all positive instances (Sujon; *et al*., 2025). This metric is particularly relevant from a social perspective, since low recall values imply the occurrence of false negatives, that is, potentially eligible families that fail to be identified by the model. The F1-score, on the other hand, corresponds to the harmonic mean between precision and recall, and is adopted as a balanced metric between both aspects (Sujon; *et al*., 2025). In this way, it allows simultaneous evaluation of the model's efficiency in avoiding both false positives and false negatives.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

Finally, in addition to aggregated metrics, the confusion matrix was used to analyze in detail the distribution of correct and incorrect classifications (Zeng, 2025), enabling a clearer interpretation of model behavior with respect to beneficiary and nonbeneficiary classes. This complementary analysis is fundamental for understanding the practical impact of the errors made (Zeng, 2025) and for supporting the discussion on the applicability of the models in the context of public administration. Thus, the combination of quantitative metrics and confusion matrix analysis allows for a more transparent evaluation aligned with the requirements of responsibility and social impact regarding the use of ML in public administration.

3. RESULTS

This section presents and discusses the results obtained from the experiments conducted with the ML models applied to the PBF data. The analysis is organized into four main subsections: the first addresses the results related to the state of RN (item A), the second presents the performance of the models trained with PB data (item B), the third provides a comparative discussion between the two states (item C), and the fourth discusses the policy implications of the obtained results (item D). In each subsection, the performance of the algorithms, the effects of feature selection, and the implications of the findings for understanding the socioeconomic patterns associated with program eligibility and public policy management are examined.

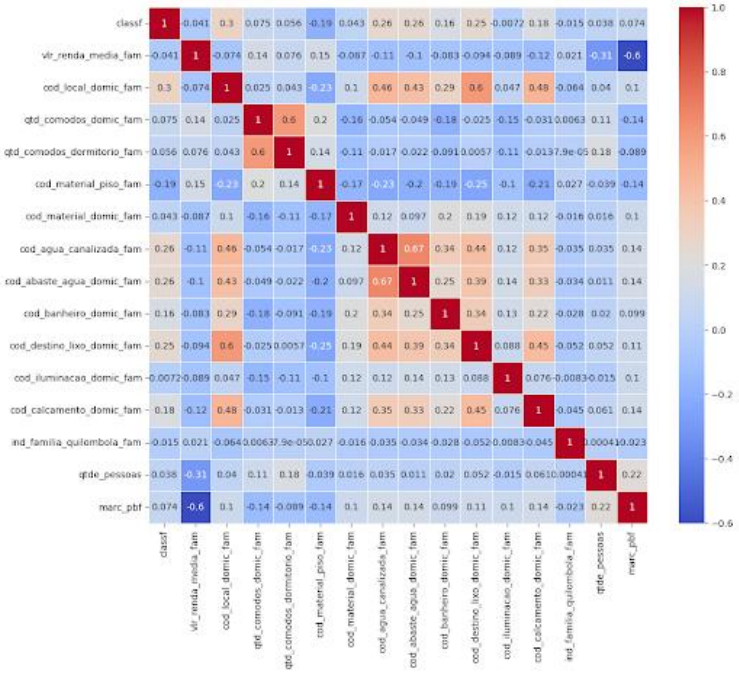
3.1. Results for Rio Grande do Norte

To evaluate the performance of the ML models applied to the PBF context in the state of RN, experiments were conducted under two distinct configurations: (i) using all variables available after data preprocessing; and (ii) using only the most relevant variables identified by the SelectKBest method. This approach made it possible to compare the impact of dimensionality reduction on the algorithms and to verify whether feature selection could improve or at least maintain, model performance.

The five variables selected through the application of SelectKBest were: vlr_rendamedia_fam, qtd_comodos_domic_fam, cod_agua_canalizada_fam, cod_abaste_agua_domic_fam, and qtde_pessoas. In this context, Figure 2 represents the correlation matrix among the study variables, in which red colors indicate positive correlations, while blue colors represent negative correlations; thus, the more intense the color, the stronger the relationship between variables. Accordingly, it can be observed that the selected attributes show a strong relationship with the classification variable *marc_pbf*, justifying their choice for the second stage of experiments.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

Figure 2. Correlation matrix among the variables of Rio Grande do Norte



Source: Authors (2026).

Table 3 presents the results obtained from training with all variables. In this scenario, the models showed consistent performance, with accuracy ranging between 82% and 90%. Regarding the best performances, it was observed that models based on decision tree algorithms, namely RF and XGBoost, stood out, achieving an accuracy of 90% and an average F1 score of 89%, highlighting a good balance between precision and recall. In addition, the RNN model also presented competitive results, with 89% accuracy. On the other hand, KNN, although yielding satisfactory results, showed inferior performance compared to the other models, with an accuracy of 82% and an average F1-score of 80%.

Table 3. Classification results for RN (All features)

Model	Accuracy	Precision	Recall	F1-Score
KNN	82%	0.82	0.81	0.80
SVM	89%	0.89	0.88	0.90
RF	90%	0.91	0.89	0.89
XGBoost	90%	0.91	0.89	0.89
RNN	89%	0.89	0.89	0.89

Source: Authors (2026).

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

Table 4 considers the results obtained after selecting the most relevant attributes for the PBF prediction context. In this case, the results remained close to the previous scenario, but with some important variations. With respect to RF, XGBoost, and RNN, it was observed that they continued to achieve the highest results, reaching 90% accuracy and an average F1-score of 89%. SVM maintained stable performance, with an accuracy of 89%, while KNN showed a slight improvement in accuracy, reaching 83% compared to the scenario using all variables, although it still remained the lowest-performing model.

Table 4. Classification results for RN (Best features)

Model	Accuracy	Precision	Recall	F1-Score
KNN	83%	0.78	0.83	0.84
SVM	89%	0.90	0.88	0.89
RF	90%	0.91	0.89	0.89
XGBoost	90%	0.91	0.89	0.89
RNN	89%	0.90	0.89	0.89

Source: Authors (2026).

In general, the results show that feature selection was effective as a simplification strategy without causing significant performance losses. In particular, RF, XGBoost, and RNN proved to be the most robust models for the RN context, maintaining high performance both with the full set and with the reduced set of attributes. These findings indicate that the socioeconomic and household variables selected by SelectKBest capture a large part of the relationship between the analyzed factors and families' eligibility for the program, reinforcing the potential of dimensionality reduction as a useful tool in predictive modeling.

3.2. Results for Paraíba

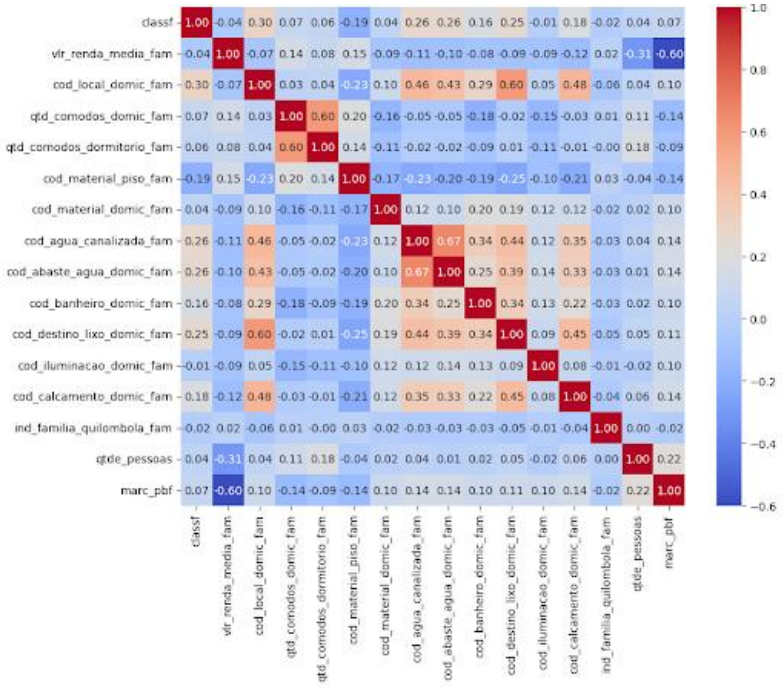
Similarly to the procedure adopted for the state of RN, the ML models were evaluated considering the context of the PBF in PB. Thus, two experimental scenarios were analyzed: (i) training using all variables available in the dataset; and (ii) training using only the most relevant features according to the SelectKBest method.

Similarly to the previous method, the SelectKBest function was used, which selects the top K features based on a statistical test. The following five variables were selected: *vlr_rendamedia_fam*, *qtd_comodos_domic_fa*, *cod_agua_canalizada_fam*, *cod_abaste_agua_domic_fam*, and *qtde_pessoas*. Figure 3 shows the correlation matrix among the

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

variables of the study in the state of Paraíba, where red colors indicate positive correlations and blue colors represent negative correlations; the more intense the color, the stronger the relationship between the variables. In this sense, it can be observed that the variables selected by the algorithm exhibit a strong correlation with the classification variable *marc_pbf*.

Figure 3. Correlation matrix among the variables of Paraíba



Source: Authors (2026).

Table 5 presents the results obtained from training using all variables. In this scenario, accuracy ranged from 82% to 90%. It can be observed that tree-based models, such as RF and XGBoost, as well as the RNN, achieved very similar performances, with 87% accuracy and average F1-scores ranging between 81% and 89%, indicating a consistent balance between precision and recall. The SVM achieved the best overall performance in this setting, reaching 90% accuracy and average F1-scores of 85% and 93% for Class 0 and Class 1, respectively, standing out as the most effective model in this scenario. On the other hand, KNN exhibited inferior performance compared to the other models, with 82% accuracy and an average F1-score of 81%, although still presenting acceptable results.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

Table 5. Classification results for PB (All features)

Model	Accuracy	Precision	Recall	F1-Score
KNN	82%	0.77	0.80	0.81
SVM	90%	0.91	0.88	0.89
RF	87%	0.88	0.85	0.86
XGBoost	87%	0.87	0.85	0.86
RNN	87%	0.87	0.85	0.85

Source: Authors (2026).

In Table 6, the results obtained using only the most relevant features are presented. In this case, it can be observed that the models maintained stable performance, with only slight variations compared to the previous scenario. The SVM once again stood out, maintaining an accuracy of 90%, confirming its robustness under dimensionality reduction. The RF, XGBoost, and RNN models also exhibited competitive performance, with accuracies ranging between 85% and 87%, reinforcing the consistency of these algorithms in the analyzed context. The KNN, in turn, showed a slight decrease in performance compared to the scenario using all variables, achieving an accuracy of 80% and average F1-scores of 76% and 83% for Class 0 and Class 1, respectively, remaining the least effective model among those evaluated.

Table 6. Classification results for PB (Best features)

Model	Accuracy	Precision	Recall	F1-Score
KNN	80%	0.75	0.81	0.80
SVM	90%	0.91	0.88	0.89
RF	85%	0.86	0.83	0.85
XGBoost	87%	0.87	0.85	0.86
RNN	87%	0.88	0.85	0.86

Source: Authors (2026).

Overall, the results obtained for the state of PB confirm the robustness of the SVM, which stood out in both scenarios. Tree-based models and the RNN also demonstrated consistent performance, reinforcing their suitability for classification tasks in the analyzed socioeconomic context. As in the case of RN, dimensionality reduction proved to be a valid strategy, capable of simplifying the models without compromising predictive performance.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

3.3. Discussion of the Results

The comparative analysis between the results for RN and PB reveals relevant patterns regarding the behavior of ML models in the context of predicting eligibility for the PBF. Overall, it was observed that the algorithms exhibited consistent performance for both states, although with specific differences that reflect particular characteristics of the state-level datasets and the relationship between socioeconomic variables and the indicator of participation in the program.

In both states, decision tree–based models, especially RF and XGBoost, demonstrated robust performance, maintaining accuracies between 87% and 90% in virtually all scenarios. This behavior is expected, as these models tend to handle heterogeneous data, nonlinear relationships, and interactions between variables typical of socioeconomic datasets efficiently. In addition, the RNN showed performance close to that of the tree-based models, reinforcing its potential to capture complex patterns even in tabular data, albeit at a higher computational cost.

The SVM, in turn, stood out mainly in PB, achieving a total accuracy of 90%. This difference compared to RN suggests that class separability in the feature space is more pronounced in the PB dataset, possibly due to a stronger correlation between the selected variables and the target variable, marc_pbf, in that state. In both contexts, however, the SVM showed high performance, reinforcing its sensitivity to highly discriminative combinations of attributes.

KNN, on the other hand, consistently presented the lowest accuracy and F1-score values in both states, both when using all features and after dimensionality reduction. This behavior reinforces the method's limitations in datasets with large volumes, high variability, and uneven class distribution, a scenario commonly found in Cadastro Único data. The fact that KNN showed a slight improvement after feature selection in RN and a decrease in PB also highlights its direct dependence on the spatial structure of the data, which varies between states.

When analyzing the two training approaches, using all variables and using the selected features, it can be observed that dimensionality reduction did not compromise the performance of the models in either state. This indicates that the variables selected by SelectKBest efficiently synthesize the main socioeconomic determinants associated with beneficiary status, which is aligned with the objective of the study to identify relevant attributes to support decision-making processes and monitoring actions.

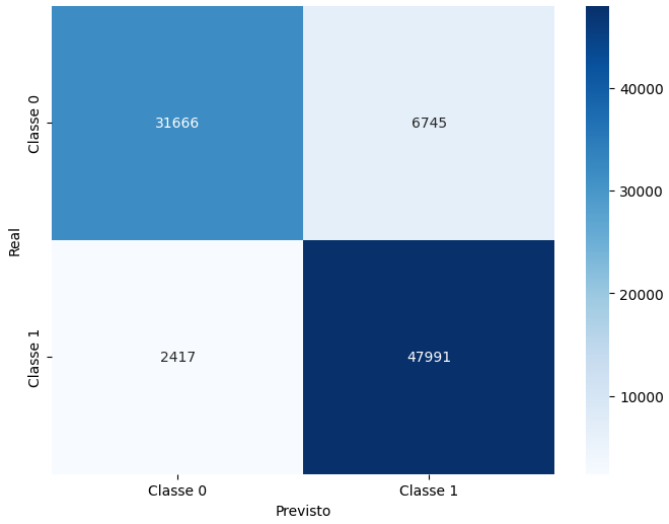
The choice of the best model for each state was guided not only by quantitative performance, but also by practical aspects related to stability, generalization capability, and suitability to the context of the analyzed data. In the case of RN, RF with feature selection presented the best balance among accuracy, precision, and recall, in addition to high robustness to noise and heterogeneous variables, which are typical characteristics of socioeconomic datasets from Cadastro Único. This property makes the model particularly suitable for operational scenarios, in which small variations in the data should not significantly compromise predictions (Géron, 2021).

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

For the state of PB, SVM stood out as the most effective approach, achieving the best overall performance, which indicates greater class separability in the selected feature space (Géron, 2021). From a practical standpoint, this suggests that the socioeconomic profile of families in PB exhibits more well-defined patterns, favoring models based on optimal decision margins. Thus, adopting different models for each state reflects a more realistic and adaptive strategy, respecting regional particularities and reinforcing the use of ML as a tool to support the management and oversight of the PBF.

Figures 4 and 5 present the confusion matrices obtained for the highest-performing models evaluated in the states of RN and PB. Each matrix explicitly shows the absolute distribution of correct and incorrect classifications, allowing a detailed analysis of the models' behavior with respect to the nonbeneficiary class, Class 0, and the beneficiary class, Class 1 (Zeng, 2025). In the case of RN, a high number of true positives is observed, totaling 47,991, indicating a strong ability of the model to correctly identify beneficiary families, while the number of false negatives remains relatively low, with 2,417 records.

Figure 4. Confusion matrix of the best model for RN

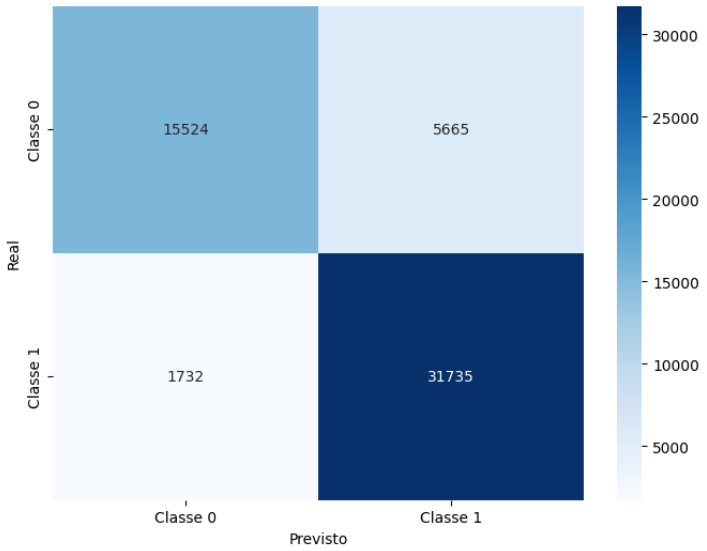


Source: Authors (2026).

For PB, a similar pattern is observed, with a predominance of true positives, totaling 31,735 records, and low values of false negatives, 1,732, reinforcing the good performance of the models in identifying the class of greatest interest in the context of the program. The analysis of the confusion matrices complements the quantitative metrics presented earlier, demonstrating that the models correctly prioritize the identification of PBF beneficiaries, a fundamental aspect for applications supporting the management and oversight of public policies.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

Figure 5. Confusion matrix of the best model for PB



Source: Authors (2026).

In addition to aggregated performance metrics, the analysis of the errors made by the models provides a deeper understanding of their behavior in the context of the PBF. Based on the confusion matrices, it can be observed that the best-performing models exhibit low rates of false negatives, that is, cases in which potentially eligible families are classified as non-beneficiaries. This aspect is particularly relevant from a social perspective, since false negatives may lead to the undue exclusion of families in vulnerable situations.

On the other hand, the occurrence of false positives, although it implies additional administrative costs by indicating non-eligible families as beneficiaries, tends to be less critical than excluding families that meet the program's criteria. Therefore, the results indicate that the models adequately prioritize sensitivity to the class of interest, favoring the recall of Class 1 (Zeng, 2025), which is aligned with the objective of the study to support public management through the effective identification of families potentially eligible for the PBF.

Finally, the comparison between RN and PB shows that although there are differences in absolute performance among the models, the overall behavior of the algorithms remains similar: tree-based models and RNNs prove to be consistent, SVM stands out when the variables exhibit greater separability, and KNN maintains inferior performance. This uniformity reinforces the applicability of ML models in the context of the PBF and demonstrates that, regardless of the particularities of each state, the socioeconomic patterns captured by the models are sufficiently stable to support predictive solutions aimed at improving the management and oversight of the program.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

3.4. Policy Implications

The findings of this study reinforce the potential of ML models as decision-support tools for public administration, particularly in the context of large-scale social programs such as PBF. As discussed in the literature, the adoption of AI in the public sector can enhance administrative efficiency, transparency, and oversight, provided that its use is aligned with governance and ethical frameworks (Alshehhi; Cheaitou; Rashid, 2022), (Caiza *et al*., 2024).

From an operational perspective, predictive models can support the prioritization of administrative actions, such as registry verification, targeted social assistance visits, and data updating processes. By identifying families with a higher probability of eligibility, public managers may allocate limited resources more efficiently, reducing costs associated with un targeted inspections while improving program coverage. This use of ML as a complementary mechanism aligns with the application-oriented dimension of AI adoption frameworks in the public sector (Alshehhi; Cheaitou; Rashid, 2022).

Moreover, ML-based predictions can contribute to preventive auditing strategies. As highlighted by Tan *et al*. (2023), trust and transparency are central to the acceptance of algorithmic decision-making in public policies. In this context, models may be employed to flag atypical or low-confidence classifications for further human analysis, supporting early identification of inconsistencies without replacing administrative judgment.

The comparative analysis between RN and PB further indicates that regional socioeconomic characteristics influence model performance, reinforcing the need for adaptive and context-aware solutions rather than a single uniform predictive approach. This observation is consistent with prior studies that emphasize the importance of interoperability and data governance in the deployment of AI systems across heterogeneous administrative contexts (Tan *et al*., 2023), (Caiza *et al*., 2024).

Finally, the responsible integration of ML into social policy management requires explainability, human oversight, and continuous evaluation. Automated predictions should remain auditable and interpretable to avoid reinforcing social biases or excluding vulnerable populations due to data limitations. When guided by ethical principles and transparent governance structures, ML systems can strengthen evidence-based policy making while preserving the central role of human decision makers in public administration (Alshehhi; Cheaitou; Rashid, 2022), (Caiza *et al*., 2024).

4. CONCLUSION

This study investigated the potential of ML techniques to support decision-making processes in public administration, using Cadastro Único data from the PBF referring to the states of RN and PB. The objectives of developing, training, and evaluating predictive models capable of classifying families according to their eligibility for the program were fully achieved, demonstrating the feasibility of applying AI as a support instrument in the management of social policies.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

The experimental results indicate that ML models are capable of capturing complex relationships among socioeconomic and household variables commonly found in administrative databases. In the case of RN, RF, XGBoost, and RNN models stood out for their robustness and stability, achieving accuracy levels of up to 90%. In PB, SVM achieved superior performance, also reaching 90% accuracy, suggesting a higher degree of class separability in the feature space for that state. In both contexts, tree-based models and RNN consistently presented competitive results, reinforcing their suitability for large-scale socioeconomic classification tasks.

An important methodological contribution of this work lies in the evaluation of dimensionality reduction through feature selection. The results demonstrate that a reduced subset of variables—mainly related to income, household structure, access to basic services, and family size—was sufficient to maintain high predictive performance. This finding not only simplifies the models but also enhances their interpretability, which is a critical requirement for applications in public administration. Moreover, it reinforces the relevance of these attributes as key determinants associated with PBF eligibility, offering empirical evidence aligned with social policy criteria.

Despite the promising results, some limitations must be acknowledged. The analysis relies on Cadastro Único data, which, although comprehensive, are subject to registration inconsistencies, outdated information, and potential biases inherent to administrative records. Furthermore, the restriction to the period between 2016 and 2018, imposed by public data availability and computational constraints, limits the generalization of the findings to different temporal scenarios, especially in light of recent socioeconomic changes. Another limitation concerns the geographic scope restricted to two states, which may affect the direct transferability of the models to other regions with distinct demographic and socioeconomic profiles. Additionally, although the SMOTE technique was employed to mitigate class imbalance, synthetic oversampling may introduce artifacts that do not fully represent real-world dynamics; therefore, the models should be interpreted strictly as decision-support tools rather than as autonomous decision-makers.

As directions for future work, it is recommended to extend the analysis to other Brazilian states and longer time horizons, increasing the robustness and representativeness of the proposed models. The integration of updated and continuously maintained databases, ideally in collaboration with governmental agencies, may further enhance predictive performance and practical applicability. Additionally, future studies may explore explainable AI techniques to improve transparency and trust in model predictions, as well as hybrid approaches that combine predictive modeling with domain-specific rules. When adopted responsibly and ethically, ML-based systems have strong potential to strengthen evidence-based public management, improve oversight mechanisms, and support more efficient and equitable social policies (Tan *et al.*, 2023), (Caiza *et al.*, 2024).

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

REFERENCES

ALSHEHHI, K.; CHEAITOU, A.; RASHID, H. Adoption Frameworks for Artificial Intelligence in the Public Sector: A Systematic Review of Literature. **Proc. 3rd South Amer. Int. Ind. Eng. Oper. Manag. Conf**, [*s. l.*], p. 919–929, 2022. DOI 10.46254/SA03.20220211. Disponível em: https://doi.org/10.46254/SA03.20220211. Acesso em: 8 abr. 2026.

AZEVEDO, C. S.; GONÇALVES, R. F.; GAVA, V. L.; SPINOLA, M. M. A Benford's Law Based Methodology for fraud detection in social welfare programs: Bolsa Familia Analysis. **Physica A**: Statistical Mechanics and its Applications, [*S. l.*], v. 576, p. 125626, 2021. DOI 10.1016/j.physa.2020.125626. Disponível em: https://doi.org/10.1016/j.physa.2020.125626. Acesso em: 9 abr. 2026.

CAIZA, G. Navigating Governmental Choices: A Comprehensive Review of Artificial Intelligence's Impact on Decision-Making. **Informatics**, [*s. l.*], v. 11, n. 64, ed. 3, 2024. DOI 10.3390/informatics11030064. Disponível em: https://doi.org/10.3390/informatics11030064. Acesso em: 11 abr. 2026.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **J. Artif. Intell. Res**, [*s. l.*], v. 16, 2002. DOI 10.1613/jair.953. Disponível em: https://doi.org/10.1613/jair.953. Acesso em: 11 abr. 2026.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min**, [*s. l.*], p. 785-794, 2016. DOI 10.1145/2939672.2939785. Disponível em: https://doi.org/10.1145/2939672.2939785. Acesso em: 16 abr. 2026.

DESORDI, Danubia; BONA, Carla Della. A inteligência artificial e a eficiência na administração pública. **Revista de Direito**, [*S. l.*], v. 12, n. 02, p. 01–22, 2020. DOI: 10.32361/202012029112. Disponível em: https://periodicos.ufv.br/revistadir/article/view/9112. Acesso em: 9 abr. 2026.

FACELI, Katti; LORENA, Ana C.; GAMA, João; ALMEIDA, Tiago Agostinho De; CARVALHO, André C. P. L. F. de. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina**. 3. ed. Rio de Janeiro: LTC, 2025. *E-book.* p.iii. ISBN 9788521639213. Disponível em: https://app.minhabiblioteca.com.br/reader/books/9788521639213/. Acesso em: 12 abr. 2026.

GÉRON, Aurélien. **Mãos à Obra**: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow. 2. ed. Rio de Janeiro: Alta Books, 2021. 640 p. ISBN 8550815489.

KOUTNÍK, J.; GREFF, K.; GOMEZ, F.; SCHMIDHUBER, J. A Clockwork RNN. **Proc. 31st Int. Conf. Mach. Learn. (ICML)**, Beijing, v. 32, 2014. DOI 10.48550/arXiv.1402.3511. Disponível em: https://doi.org/10.48550/arXiv.1402.3511. Acesso em: 16 abr. 2026.

SUJON, K. M.; HASSAN, R.; CHOI, K.; SAMAD, M. A. Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. **Journal of Big Data**, [*s. l.*], v. 12, n. 268, 2025. DOI 10.1186/s40537-025-01313-4. Disponível em: https://doi.org/10.1186/s40537-025-01313-4. Acesso em: 16 abr. 2026.

TAN, E. et al. Artificial intelligence and algorithmic decisions in fraud detection: An interpretive structural model. **Data & Policy**, Reino Unido, e25, ed. 5, p. 919–929, 2023. DOI 10.1017/dap.2023.22. Disponível em: https://doi.org/10.1017/dap.2023.22. Acesso em: 8 abr. 2026.

TISLENKO, M. D.; GAIDEL, A. V.; KUPRIYANOV, A. V. Comparison of feature selection algorithms for Data classification problems. **2022 VIII International Conference on Information Technology and Nanotechnology (ITNT)**, Samara, p. 1-5, 2022. DOI 10.1109/ITNT55410.2022.9848765. Disponível em: https://doi.org/10.1109/ITNT55410.2022.9848765. Acesso em: 16 abr. 2026.

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

ZENG, G. Invariance Properties and Evaluation Metrics Derived from the Confusion Matrix in Multiclass Classification. **Mathematics**, [*s. l.*], v. 13, ed. 16, p. 2609, 2025. DOI 10.3390/math13162609. Disponível em: https://doi.org/10.3390/math13162609. Acesso em: 16 abr. 2026.