50 YEARS
eISSN 1981-1829
Ciência e Agrotecnologia
Science and Agrotechnology

Automatic classification of rural environmental registry records using machine learning algorithms

Classificação automática de registros de cadastro ambiental rural usando algoritmos de aprendizado de máquina

Fernando Elias de Melo Borges[1*], Danton Diego Ferreira[2*], Max Deivid Alves do Nascimento[1], Antônio Carlos de Sousa Couto Júnior[3], Ricardo Rodrigues Magalhães[2], Luis Antonio Coimbra Borges[4]

ABSTRACT

The Rural Environmental Registry (CAR, from its Portuguese acronym) is a mandatory public online registry that consolidates information about a property's area, land use, native vegetation, legal reserve, and vegetation conservation areas. The objective of CAR is to monitor land use and combat deforestation. Currently, CAR registry analyses are conducted manually, requiring specialized personnel and a time-consuming verification process — a challenge amplified by Brazil's almost 8 million CAR registries and over 8 million rural properties. To address this problem, the use of machine learning techniques for automatic verification is crucial for Brazil's environmental management. This study aims to develop machine learning-based strategies to classify CAR registries into two categories: 'Approved', referring to correct registries, and 'Canceled', denoting registries with inconsistencies. Additionally, interpretable machine learning tools were used to determine the impact of each feature on the automatic classification. The best-performing model was the Gradient Boosting Machine, achieving classification accuracy above 90% (using a real test dataset containing more than 6,000 registries) and demonstrating correlations between its feature importance scores and the inconsistencies identified in the literature. These results demonstrate the feasibility of using this approach for registry classification and interpretation.

Index terms: Explainable artificial inteligence; enviromental management; data science.

RESUMO

O Cadastro Ambiental Rural (CAR) é um registro público obrigatório, online, que consolida informações sobre a área do imóvel, uso do solo, vegetação nativa, reserva legal e áreas de preservação da vegetação. O objetivo do CAR é monitorar o uso da terra e combater o desmatamento. Atualmente, as análises dos registros do CAR são realizadas manualmente, exigindo profissionais altamente qualificados e um processo de verificação demorado, um desafio ampliado pelos quase 8 milhões de registros de CAR e mais de 8 milhões de imóveis rurais no Brasil. Para enfrentar esse problema, o uso de técnicas de aprendizado de máquina para verificação automática é crucial para a gestão ambiental brasileira. Este estudo tem como objetivo desenvolver estratégias baseadas em aprendizado de máquina para classificar registros do CAR em duas categorias: "Aprovado", referente a registros corretos, e "Cancelado", denotando registros com inconsistências. Além disso, foram utilizadas ferramentas de aprendizado de máquina interpretável para determinar o impacto de cada atributo na classificação automática. O modelo com melhor desempenho foi a Gradient Boosting Machine, alcançando acurácia de classificação acima de 90% e demonstrando correlações entre seus escores de importância das variáveis e as inconsistências identificadas na literatura. Esses resultados demonstram a viabilidade do uso dessa abordagem para classificação e interpretação de registros.

Termos para indexação: Inteligência artificial explicável; gestão ambiental; ciência de dados.

Introduction

The Rural Environmental Registry (CAR, from the Portuguese acronym) is a mandatory, self-declaratory online registry established by the Forest Code (Brasil, 2012). CAR was developed to monitor land use, combat illegal deforestation, support compliance with environmental regulations, and sustainable land management in Brazil (Brasil, 2012); (Almeida, Sousa Júnior & Espindola, 2026); (Silva et al., 2025). The registry integrates comprehensive information on rural properties, including documentation, land use, vegetation types, hydrological features, and geospatial boundaries.

The Brazilian government has implemented incentive mechanisms to encourage farmer adherence to CAR requirements, such as tax reductions, facilitated rural credit, and preferential agricultural insurance terms (Brasil, 2012). Furthermore, because CAR data are publicly accessible, they enable extensive research across domains such as environmental planning, sustainability, and land-use studies.

Agricultural Sciences

Ciênc. Agrotec., 50:e008226, 2026
http://dx.doi.org/10.1590/1413-7054202650008226

Editor: Renato Paiva

[1]Universidade Federal de Lavras/UFLA, Escola de Engenharia/EENG, Lavras, MG, Brasil
[2]Universidade Federal de Lavras/UFLA, Departamento de Automática/DAT, Lavras, MG, Brasil
[3]Universidade Federal de Lavras/UFLA, Agência de Inovação Zetta, Lavras, MG, Brasil
[4]Universidade Federal de Lavras/UFLA, Departamento de Ciências Florestais/DCF, Lavras, MG, Brasil

*Corresponding author: fernandoelias.mb@gmail.com; danton@ufla.br

Received in May 19, 2026 and approved in September 16, 2026

All the contents of this journal, except where otherwise noted, is licensed under a Creative Commons Attribution BY.

Beyond these incentives, addressing inconsistencies in CAR registries is critical for effective environmental monitoring and regulatory compliance. The self-declaratory nature of the data can lead to conflicts, including overlapping property claims (Furumo et al., 2024). Verification is conducted by Brazilian state environmental agencies; however, this process is currently manual and requires substantial time and specialized personnel due to Brazil's vast scale – over 8.3 million existing registries (Brasil, 2026), approaching the total number of rural properties in Brazil (Brasil, 2022). Table 1 summarizes key CAR statistics.

Critically, fewer than 42% of registries have been analyzed, and less than 10% have achieved full regularization (CAR report accessed in August of 2026 (Brasil, 2026). This verification gap threatens data quality and regulatory efficacy. To address this challenge, we propose a novel application of machine learning to classify irregular registries. Our work encompasses the entire data mining pipeline, from collection to classifier evaluation.

Data classification, a fundamental machine learning task, is widely applied across diverse domains such as healthcare (Ray et al., 2025), structural engineering (Yu, Yu & Li, 2025), bank fraud detection (Ali et al., 2022), and port security (Hocek, Yay & Yazir, 2025). These studies highlight the versatility of classification models in addressing heterogeneous problem domains.

Leveraging the capability of automated data classification using machine learning and addressing the challenges associated with the manual verification of CAR records, the main contribution of this work is the development of a workflow that enables the automated classification of these records. This workflow encompasses data preparation, classification, and an analysis of the impact of each variable on the automated classification process, aiming to determine whether the automated decision-making aligns with existing literature.

Another significant contribution lies in applying machine learning techniques to detect record and transaction irregularities. Similar methodologies have been reported in the literature, including the detection of fraudulent financial transactions (Thanh & Huy, 2025) and data quality assessment in healthcare services (Jarmakovica, 2025). Beyond detection, several studies (Lokanan, 2025; Mazumder, 2026) leverage machine learning to rank feature importance, thereby enhancing interpretability in model-driven decision-making.

Rural environmental registry

CAR was established under the 2012 Forest Code Law (Brasil, 2012) and implemented in 2014. The principal objectives of CAR are to monitor land use and combat illegal deforestation. It is important to note that CAR serves as an instrument for environmental regularization and cannot be used for land tenure regularization.

Since its inception, the Brazilian government has introduced policies to encourage farmers to register in CAR and maintain environmental compliance. Key incentives include tax exemptions or reductions, preferential agricultural insurance terms, reduced financing fees for machinery, among others (Brasil, 2012).

The public availability of CAR data has enabled research across multiple fields. For example, Ramos et al. (2025) employed an innovative method to map rural properties using UAV (Unmanned Aerial Vehicle) to obtain maps to improve the analysis of land use and coverage though the CAR. Arruda et al. (2025) evaluated the CAR impact on the deforestation and environmental compliance in the Brazilian Central-West. The authors achieved results varying in accord with the land size and reduce in the deforestation on small rural properties. Additionally, Silva et al. (2023) investigated the green grabbing (English term for Brazilian term "grilagem") on the MATOPIBA region. Therefore, these works present the importance of a public platform such as CAR in the Brazilian environmental management.

Nevertheless, CAR's implementation presents structural and operational challenges, particularly related to data quality. As a self-declaratory system in which farmers or their representatives submit information, registries may contain inaccuracies or irregularities. For instance, Furumo et al. (2024) demonstrated how overlapping property boundaries within CAR registries can generate conflicts and create significant obstacles for governmental oversight.

Data classification

Data classification is a supervised machine learning task, meaning that classifier modeling is based on pre-labeled data provided by a specialist or obtained from a consolidated system (Han, Kamber & Pei, 2011). The primary goal of data classification is to predict labels for unseen data, with the label set depending on the specific application. For example, in healthcare, a classifier might categorize patients as healthy or ill; in banking, transactions may be classified as legitimate or fraudulent.

Table 1: CAR numbers obtained on August 20, 2026.

Information	Quantity	Total registries (%)
Number of CAR registries (Rural properties)	8,337,574	100
Registries that have undergone some analysis	3,490,918	41.9
Registries with complete analysis	819,515	9.8

There are several types of classification models, each with distinct advantages and limitations. The classifiers applied in this work are described in the following subsections.

Logistic regression

Logistic regression is a linear model used to estimate the *a posteriori* probability that an input instance belongs to a particular class. This estimation is obtained by applying a linear function to the input data, followed by the logistic (sigmoid) transformation. Given x as the input vector and θ as the parameter vector, the classification function for the binary case is defined by Equation (1):

$$g(\mathrm{x},\theta) = \frac{1}{1+\exp\left(-\mathrm{x}\theta^{T}\right)} \tag{1}$$

Where x is the augmented input vector in the format [1, x] (including 1 as the bias term), x is the original feature vector, and θ is the parameter vector of the logistic model.

Parameter estimation is typically performed using the maximum likelihood method, which maximizes the model's representational capacity. For computational convenience, the negative log-likelihood is minimized, transforming parameter estimation into an optimization problem. The objective function is given in Equation (2):

$$l(\theta) = -\sum_{i=1}^{N}\left[y_i \log\left(\widehat{y_i}\right) + \left(1-y_i\right)\log\left(1-\widehat{y_i}\right)\right] \tag{2}$$

Where N is the number of samples in the training dataset, y_i is the i-th true label, and $\widehat{y_i}$ is the i-th predicted probability.

Like other linear models, logistic regression has a low computational cost due to the simplicity of its formulation. Another key advantage is its interpretability: relative magnitude of weights (learned during training) reflects predictive importance. The sign of a weight reveals the class to which the attribute contributes, while its magnitude indicates the strength of the contribution.

Multi-layer perceptron

Artificial Neural Networks (ANNs) are mathematical models inspired by the functioning of biological neurons (Haykin, 2007), widely employed in various forms to address a broad range of machine learning problems.

Within the family of ANNs, several architectures exist, each with distinct topologies, characteristics, and application domains. Among these, one of the most well-known and widely applied models is the Multi-Layer Perceptron (MLP), whose weights are typically adjusted using the backpropagation algorithm, optimized via gradient descent. An example of an MLP architecture is shown in Figure 1.

The general input–output relationship of an MLP can be expressed in matrix form, relating the output of layer k to the output of the previous layer $k-1$, as given in Equation (3):

$$u_k = f\left(\mathrm{W}_k \mathrm{u}_{k-1} + \mathrm{b}_k\right) \tag{3}$$

Where u_k denotes the output vector of layer k, W_k is the weight matrix of layer k, b_k is the corresponding bias vector, and u_{k-1} is the output vector from layer $k-1$. For $k-1=0$, $\mathrm{u}_{k-1} = x$, where x is the input feature vector. The function $f(\cdot)$ represents the activation function applied elementwise to the layer's weighted sum.

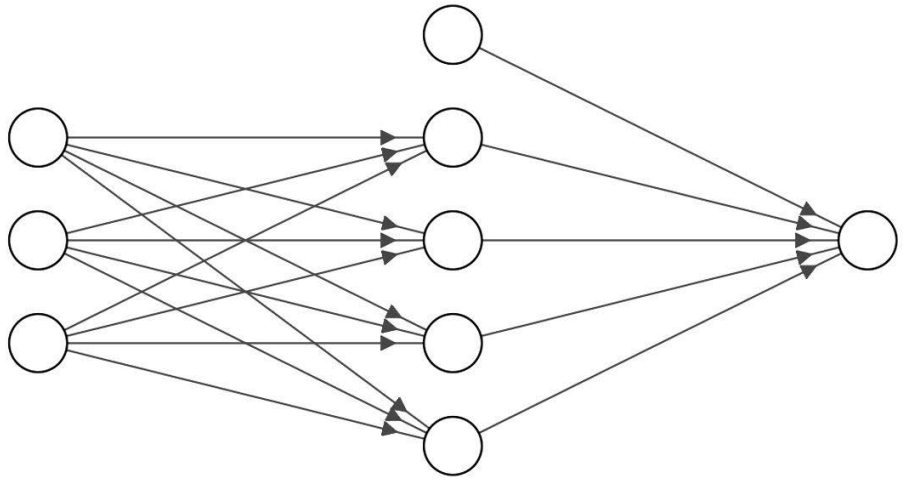

Figure 1: Example of an MLP network architecture.

Random Forest

Random Forest is an ensemble learning method composed of multiple base learners designed to improve predictive performance and robustness (Breiman, 2001). It employs Decision Tree classifiers, a classical machine learning algorithm valued for its simple architecture and fast training process.

A Random Forest model trains a user-defined number of Decision Trees, with hyperparameters such as maximum depth, split quality criterion (e.g., entropy or Gini index), minimum partition size, and others specified by the practitioner. Once the trees are trained, classification is performed through majority voting, whereby the class receiving the highest number of votes is assigned to the input sample.

Each tree is trained on a randomly selected subset of the original dataset, with both instances and features sampled. This sampling is performed *with replacement* (bootstrap sampling), ensuring model diversity. The margin function for Random Forests is given by Equation (4):

$$mg\left(\mathrm{X},Y\right)=av_k I\left(h_k\left(\mathrm{X}\right)=Y\right)-\max_{j\neq Y}\left(av_k I\left(h_k\left(\mathrm{X}\right)=j\right)\right) \quad (4)$$

where h_k denotes the k-th Decision Tree classifier, X is the input vector, Y is the true class label, j is a class different from Y, and $I(\cdot)$ is the indicator function. The term av_k indicates the average taken over all k trees in the ensemble.

Boosting algorithms

Building on the ensemble learning principle also employed in Random Forests, boosting algorithms combine multiple weak learners, models of relatively low complexity, to achieve high predictive performance. Common examples include AdaBoost and Gradient Boosting. Unlike Random Forests, these algorithms are not limited to Decision Trees as base learners.

A key distinction between the algorithms presented here and those discussed in the previous section lies in the ensemble construction and the decision-making process. Random Forests employ bagging, where all classifiers are trained in parallel on bootstrapped datasets and have equal weight in the final decision. In contrast, boosting trains classifiers sequentially, and both resampling and classification are *weighted*, assigning different importance to individual samples and models within the ensemble.

In boosting, misclassified samples are assigned higher weights, ensuring that subsequent learners emphasize harder-to-classify instances. The final classification is obtained via a weighted combination of the predictions from all learners, where each learner's weight is determined by its performance during training.

From the original boosting concept, various ensemble training algorithms have been developed, differing mainly in how these weights are updated. AdaBoost uses a predefined weight-update function, which depends on the variant of the algorithm (Hastie et al., 2009), while Gradient Boosting updates weights using a gradient descent procedure (Friedman, 2002).

In the case of AdaBoost, weight adjustment can be performed using the SAMME.R algorithm, defined in Equation (5):

$$w_i \leftarrow w_i \cdot \exp\left(\frac{K-1}{K}\mathrm{y}_i^T \log\left(\mathrm{p}^{(m)}\left(\mathrm{x}_i\right)\right)\right), i=1,\ldots,n \quad (5)$$

Where w_i is the weight of the i-th training sample, K is the number of classes, y_i is the one-hot encoded true label vector, and $\mathrm{p}^{(m)}(\mathrm{x}_i)$ is the predicted class probability vector at iteration m.

Gradient Boosting, on the other hand, updates model parameters by approximating gradient descent to minimize a specified cost function ψ. The stage-wise update parameter γ_{lm} is computed as in Equation (6):

$$\gamma_{lm}=\arg\min_{\gamma}\left(\sum_{x_{\pi(i)}\in R_{lm}}\Psi\left(y_{\pi(i)},F_{m-1}\left(\mathrm{x}_{\pi(i)}\right)+\gamma\right)\right) \quad (6)$$

Where $F_{m-1}(\cdot)$ is the decision function from the $(m-1)$th iteration, R_{lm} is the region (leaf) of the l-th terminal node in the m-th tree, and $\pi(i)$ denotes the index mapping of the training samples. Gradient Boosting methods are widely used due to their strong generalization capabilities and their ability to compute feature importance rankings.

Support vector machines

Support Vector Machines (SVM) are supervised machine learning models grounded in statistical learning theory (Vapnik, 1998). The fundamental objective of SVM is to find an optimal hyperplane that maximizes the margin between classes. For linearly separable problems, the algorithm identifies this hyperplane to achieve perfect separation.

For datasets that are not linearly separable, SVM applies a transformation of the feature space through kernel functions, effectively increasing the dimensionality of the problem. In the transformed space, a linear decision boundary can be established. Commonly used kernel functions include linear, polynomial, radial basis function (RBF), and hyperbolic tangent (Hastie, Tibshirani & Friedman, 2017).

In its original formulation, SVM is hyperparameter-free, relying solely on the internal optimization of its parameters. This property holds only for datasets that are perfectly separable (linearly or nonlinearly) (Evsukoff, 2020). In real-world scenarios, where classes often overlap, the basic SVM cannot be applied directly. To address this, a soft-margin SVM introduces *slack variables* ξ to relax the separation constraints, allowing for certain misclassifications. The soft-margin SVM is defined by Equation (7):

$$\min_{w} \frac{1}{2}\|w\|^2 + \frac{\lambda}{N}\sum_{i=1}^{N}\xi(i)^k \quad (7)$$

subject to $\xi(i) \geq 0$, $v(i)\, g\,(x(i)) \geq 1 - \xi(i)$

where $\xi(i)$ denotes the slack variable for the i-th sample, N is the number of training instances, λ is the regularization parameter controlling the trade-off between margin maximization and classification error, and w is the weight vector defining the separating hyperplane.

Feature selection

Many real-world datasets contain hundreds—or even thousands—of attributes. Training classifiers with the entire attribute set often demand substantial computational resources, including high processing power, large memory capacity, and extended training time. Moreover, using all available features does not necessarily yield optimal predictive performance. Selecting a relevant subset of attributes can simultaneously reduce computational cost, improve classification accuracy and mitigate overfitting.

Among the various feature selection methods proposed in the literature, Fisher's Discriminant Ratio (FDR) is a widely used linear technique applicable to both pattern classification and feature selection tasks. Its main principle is to minimize intra-class variability while maximizing inter-class separation. In essence, FDR aims to cluster the samples of each class within a compact region while ensuring that different classes remain well-separated in the feature space.

In the feature selection context, FDR evaluates the discriminative power of each attribute through a score J. A higher J value indicates greater relevance of that attribute for classification. The FDR for two-class problems is given by Equation 8 (Duda & Hart, 2001):

$$J = \left(\mu_i - \mu_j\right)^2 \odot \left(\frac{1}{\sigma_i^2 + \sigma_j^2}\right) \quad (8)$$

Where J is the vector of discriminative scores for each attribute, μ_i and μ_j are the mean vectors of classes i and j, σ_i^2 and σ_j^2 are their respective variance vectors, and $\odot$ denotes the element-wise product.

The main advantage of FDR lies in its low computational complexity: the J index can be computed directly from class means and variances, making it suitable for large-scale datasets. Additionally, the FDR is applied only during the training phase. After selecting the best features via FDR using training data, the monitored sample is filtered directly in the operational phase. FDR has been successfully used in different applications for feature selection (Kumar & Anand, 2024).

Material and Methods

All experiments in this work were conducted on a computer equipped with an Intel® Core™ i5-1135G7 processor (2.40 GHz base frequency), 20 GB of RAM, and an NVIDIA® GeForce® MX350 GPU. All software implementations were developed in Python (version 3.12). The Python modules used for this work were: Numpy, Pandas, Scikit-learn, Matplotlib and Mlxtend. To summarize the methodological pipeline, a flowchart representing the main steps of the workflow is shown in Figure 2.

Dataset and preprocessing

For the experimental procedure, the Zetta Innovation Agency at the Federal University of Lavras provided public Cadastro Ambiental Rural (CAR) registries. These registries can be accessed via the SiCAR website (https://consulta.car.gov.br/geoservices), or as a structured dataset provided by the authors (http://www.aia.ufla.br/home/filesdatasets/). The dataset contains information about rural properties, including total area, geospatial polygon vertices, land-use characteristics, vegetation types, the number of rectifications made by the registrant, and responses to an environmental regularization questionnaire. In addition to these declaratory data, spatial overlap information was generated, indicating intersections with other rural properties, indigenous territories, and conservation units.

The dataset features can be grouped into six categories:

- Area data: total property area, property documentation type.
- Geospatial polygon vertices.
- Rectifications of the registry.
- Overlapping area indices.
- Land characteristics: land use and vegetation.
- Environmental regularization questionnaire responses.

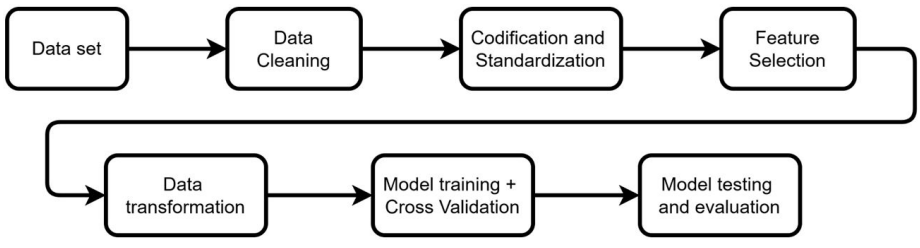

Figure 2: Flowchart representing the steps of this work.

The dataset contains 89 input attributes and a binary target variable indicating registry status: *Approved* (no irregularities) or *Canceled* (irregular registries). The study sample comprises 32,458 observations from 19 Brazilian states (Table 2). These records represent a small but well-labeled subset of the total CAR database, suitable for supervised learning. The limited proportion of labeled data reflects a challenge in data acquisition and also serves as a motivation for applying the proposed approach.

Table 2: Number of records in the dataset by state.

UF	Approved	Canceled	Total
AC	14	322	336
AL	0	100	100
AP	0	258	258
CE	1,456	2,535	3,991
ES	0	540	540
GO	40	578	618
MA	20	781	801
PA	159	3,819	3,978
PB	51	280	331
PE	0	69	69
PI	0	162	162
RJ	160	400	560
RN	0	138	138
RO	323	1,571	1,894
RR	0	19	19
SE	2	87	89
SP	632	4,983	5,615
TO	3,454	9,505	12,959
Total	6,311	26,147	32,458

It is worth noting that all features used in this study were either obtained directly from the public CAR database or derived through spatial analysis. The overlap-related variables were not available as explicit attributes in the CAR database. Instead, they were generated by examining the georeferenced boundary polygons of neighboring properties using Geographic Information System (GIS)-based spatial intersection analysis. Because this information is not automatically provided by the CAR system, each neighboring property pair was manually evaluated to determine the presence or absence of overlapping areas

Regarding the strong data imbalance between Approved and Canceled classes, a resample method was applied (SMOTE) during preliminary tests without significant accuracy improvement. Thus, we decided to proceed with the original dataset, preserving the natural distribution of the classes to avoid introducing synthetic data into the model.

Before preprocessing, the data were split into training and testing sets. The training set was used for exploratory analysis, cleaning, standardization, and model training with cross-validation, while the testing set contained previously unseen data. The split was performed using stratified random sampling without replacement, allocating 80% of the records to training and 20% to testing, in a single iteration. The stratification was based on class balance to avoid biases in the classification process.

This partitioning into training/validation and test sets was performed to ensure that the test data remained entirely unseen during model development, thereby preventing data leakage and optimistic performance bias. Consequently, all cross-validation procedures and statistical tests were conducted strictly on the training set, reserving the test set exclusively for the final evaluation.

Data standardization involved several preprocessing steps. Irrelevant attributes for classification—such as *Federative Unit (FU) of the property*, *city*, and *Property Id* —were removed, along with sensitive information such as the registrant's name. Questionnaire responses were numerically encoded: *Yes* = 2, *No* = 1, and *Not reported* = 0. Higher values were assigned to *Yes* = 2 to establish a 3-point scale that treats *Not reported* as an explicit input variable selected by the user. In this scheme, 0 represents the user-selected absence of information (*Not reported*), 1 represents an explicit negative response (*No*), and 2 reflects the full presence or positive confirmation (*Yes*). For attributes describing rural property area, multiple area values were summed to produce the total area. The target variable, *Condition*, was encoded as 0 for canceled records and 1 for approved records.

After preprocessing, 83 input attributes remained. Feature selection was performed using the Fisher Discriminant Ratio (FDR), which was chosen due to its computational efficiency and ease implementation compared with non-linear or wrapper methods. First, the FDR score vector was computed and sorted in descending order. Attributes were initially tested in groups of 10 (e.g., top 10, top 20) to identify the range with the best performance (e.g., 20–30 attributes). This range was then refined in groups of 5 and finally tuned attribute by attribute. For continuous features exhibiting high skewness, a logarithmic transformation was applied to improve distributional symmetry.

Classifiers modeling and evaluation

As described in Subsection 2.2, six classifiers were employed in this study: AdaBoost (ABC), Gradient Boosting (GBT), Logistic Regression (LRC), Multi-Layer Perceptron (MLP), Random Forest (RFC), and Support Vector Machines (SVM). The use of a diverse set of classifiers enables testing different modeling approaches to improve the likelihood of achieving high classification performance.

For model training, hyperparameter tuning was performed using a Random Search algorithm with 30 iterations due to computational constraints, as a personal desktop workstation was utilized rather than a high-performance computing (HPC) environment. The choice of Random Search was motivated by its lower computational cost compared to methods such as Grid Search or evolutionary algorithms, which require significantly more training time. Alongside hyperparameter optimization, 5-fold cross-validation was applied, and the best model for each classifier was selected based on the highest accuracy. The hyperparameter ranges explored in this work are presented in Table 3.

Table 3: Range of hyperparameters used in this work.

Model	Hyperparameters	Value
ABC	Split criterion	Gini index, Entropy
	Learning rate	[0.01, 1)
	Max depth	[3, 19)
	Number of trees	[25, 525)
GBT	Learning rate	[0.01, 1)
	Max depth	[3, 19)
	Number of trees	[25, 525)
	Loss function	exponential, log
LRC	*C*	[0.01, 100)
	Optimizer	saga
	Regularization	Lasso, Ridge, None
MLP	Hidden layer neurons	[5, 65)
	Activation function	hyperbolic tangent, relu, logistic
	Optimizer	lbfgs, adam
	Learning rate	constant, adaptive
RFC	Split criterion	Gini index, Entropy
	Max depth	[1, 21)
	Number of trees	[25, 525)
SVM	Kernel	rbf, polynomial, linear
	C	[0.01, 100)
	γ	auto, scale

Optimizer 'lbfgs' corresponds to a second-order parameter adjustment method in the quasi-Newton family. The parameter 'scale', when applied, follows the formulation $\gamma = 1/(n_{attributes}\sigma^2(X))$, where $\sigma^2(X)$ is the variance of the training dataset.

Model performance was evaluated using the following metrics: accuracy (ACC), area under the ROC curve (AUC), F1-score, Cohen's kappa index, Balanced Accuracy and precision and recall for each class.

For the training phase, results are reported as the mean ± standard deviation across the 5 folds of cross-validation. For the testing phase, results are presented as absolute values since the test set was evaluated in a single execution.

Additionally, during the training and validation stages, a paired *t*-test was applied to assess whether performance differences between classifiers were statistically significant. Model performance was evaluated using the 5x2 cross-validation paired t-test. This procedure employs 2-fold cross-validation repeated five times to obtain robust statistical comparisons (Dietterich, 1998). The significance level was set at 5%. Finally, all models were applied to the test subset, with numerical results reported in tabular form, cross-validation results summarized in a separate table, and graphical performance shown through ROC curve plots.

Results and Discussion

To provide a clear and structured presentation, the results are organized into two subsections. The first presents the outcomes of the preprocessing stage, while the second reports and discusses the classification results.

Pre-processing Results

The pre-processing step was essential to clear and standardize the raw dataset to standard data to train and test the models. Furthermore, this standardization process can be reproduced in new data during a real API application. Regarding this process, the results can be presented as follows:

- Attributes with no relevant information or sensitive information were removed;
- Attributes with almost all of the samples with constant value in zero were removed;
- After these steps and the categorical attributes codification, the best result with FDR was obtained with 41 input attributes;
- After feature selection, logarithmic transformation was applied to the 33 continuous features (out of the 41 selected features) to reduce skewness.

Regarding the skewness present in the raw data, the logarithm transformation was applied in 33 continuous attributes. In order to show the effect of this transformation, the Figures 3, 4 and 5 show the histogram before and after the logarithm application for the attributes 'Brazilian Rural Tax quantity', 'Documented Area' and 'Total Legal Reserve Area', respectively. Also, it is important to note that this transform can reduce the discrepancy in the scale of these attributes. And, training models with attributes on a similar scale ensures balanced relevance to the classifier model.

Classification results

After the pre-processing step, the data was divided into two subsets: Train + Validation and Test. The first one corresponds to the data used for the training the classifiers using the cross-validation process and the last one refers to the data unknown for the models. To illustrate the numbers of each subset, in Table 4 the numbers of each subset for each class are presented.

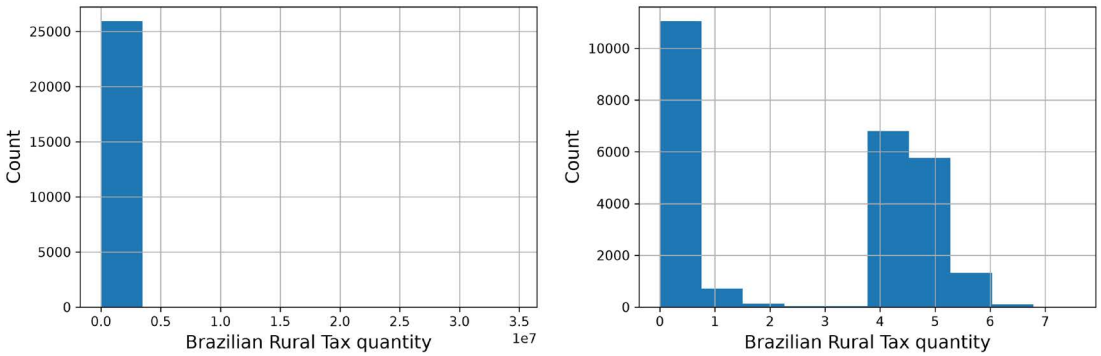

Figure 3: Logarithmic transformation effect on the attribute 'Brazilian Rural Tax quantity.

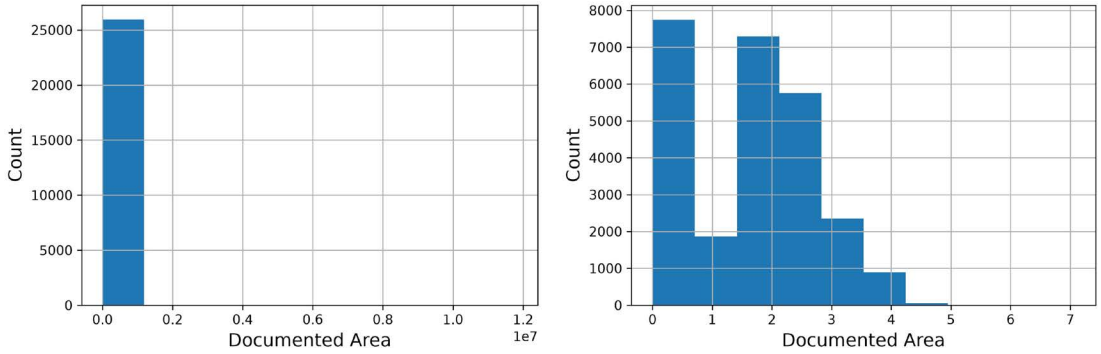

Figure 4: Logarithmic transformation effect on the attribute 'Documented Area'.

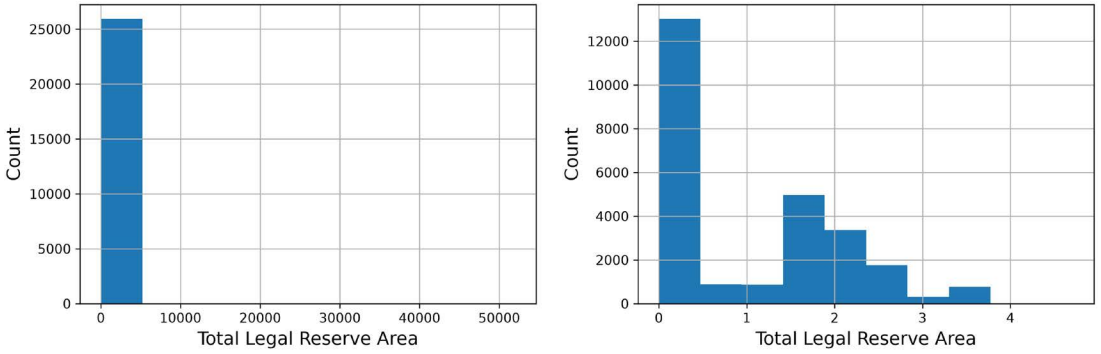

Figure 5: Logarithmic transformation effect on the attribute 'Total Legal Reserve Area'.

Table 4: Numbers of the Training + Validation and Test subsets

Class	Train + Validation	Test	Total
Approved – 1	5418	1263	6311
Canceled – 0	20917	5230	26147
Total	25965	6493	32458

With these subsets, the models were trained and the hyperparameters were adjusted by the Random Search. The values of the hyperparameters for each classifier are displayed in Table 5.

In sequence of the model training, the cross-validation results were obtained and displayed in Table 6. Additionally, paired t-test was applied to assess whether performance differences between classifiers were statistically significant, at the 5% level. The statistical results are presented in Table 7.

Additionally, given the severe class imbalance in the dataset (with Class 0 comprising 80% of the samples), an additional statistical evaluation was performed using balanced accuracy. The procedure followed the same framework as the primary analysis, aiming to assess whether the differences in balanced accuracy between classifiers were statistically

Table 5: Classifiers Hyperparameters.

Model	Hyperparameters	Value
ABC	Split Criterion	Gini Index
	Learning rate	0.0552
	Max Depth	5
	Number of trees	150
GBT	Learning rate	0.1562
	Max Depth	7
	Number of trees	300
	Loss function	exponential
LRC	C	0.6871
	Optimizer	saga
	Regularization	none
MLP	Hidden layer neurons	55
	Activation function	hyperbolic tangent
	Optimizer	lbfgs
	Learning rate	constant
RFC	Split Criterion	entropy
	Max Depth	15
	Number of trees	450
SVM	Kernel	rbf
	C	72.6141
	gamma	scale

significant at a 5% significance level. The results of this t-test are presented in Table 8.

Analysing the results of Tables 5, 6, 7 and 8, some points can be highlighted:

- GBT, ABC, and Random Forest achieved the highest overall metrics, with GBT outperforming the other classifiers across most evaluated parameters.
- Although GBT obtained the best performance, the t-test indicates that GBT results are not significantly different from ABC at the 5% level. This indicates that, during the training phase, ABC and GBT are equivalent.
- The t-test results reveal statistically significant differences between GBT and the other models, with the exception of ABC.
- As shown in Table 8, with the exception of the ABC vs. GBT and ABC vs. MLP classifier pairs, all remaining model comparisons revealed statistically significant differences in balanced accuracy. These results indicate that the ABC, GBT, and MLP models achieved superior performance compared to the other classifiers under the class-imbalanced scenario.
- The best cross-validation results can be attributed to ABC and GBT models, with MLP showing only minor performance differences compared to these two.

Finishing the phase of model tuning and training evaluation, the test with the new data was made. The test subset results are available in Table 9 and the ROC curve plot is shown in Figure 6.

About the results in the test subset, the results reflected the testing results, showing ABC and GBT as the best models. Moreover, it is important to note that RFC achieved the highest precision for Class 1 and recall for Class 0, and the results of MLP performance was slightly lower than GBT and ABC.

Additionally, given the class-imbalanced scenario, evaluating balanced accuracy and Cohen's Kappa coefficient is particularly relevant, as these metrics provide a more robust assessment under such conditions. Notably, the Kappa index can also be directly associated with the quality of the classification performance (Landis & Koch, 1977). Comparing these two metrics reveals that the ABC and GBT models once again achieved the highest scores on the test set.

Table 6: Classification results in cross validation subset.

Subset	Model	ACC	AUC	F1-Score	Kappa	Balanced ACC	Precision Class 0	Precision Class 1	Recall Class 0	Recall Class 1
Cross Validation	ABC	96.7 ± 0.2	93.8 ± 0.5	91.4 ± 0.6	89.4 ± 0.7	93.8 ± 0.5	97.4 ± 0.2	93.9 ± 0.9	98.6 ± 0.2	89.0 ± 1.1
	GBT	96.8 ± 0.3	94.0 ± 0.7	91.6 ± 0.8	89.7 ± 1.0	94.0 ± 0.7	97.5 ± 0.3	93.9 ± 0.4	98.6 ± 0.1	89.5 ± 1.4
	LRC	91.3 ± 0.5	84.4 ± 1.1	76.5 ± 1.5	71.2 ± 1.7	84.4 ± 1.2	93.6 ± 0.5	80.3 ± 1.1	95.7 ± 0.3	73.1 ± 2.3
	MLP	96.2 ± 0.2	93.5 ± 0.5	90.1 ± 0.5	87.8 ± 0.6	93.5 ± 0.5	97.3 ± 0.2	91.3 ± 0.8	98.0 ± 0.2	88.9 ± 1.0
	RFC	91.9 ± 0.8	79.5 ± 2.3	73.9 ± 3.6	69.5 ± 3.9	79.5 ± 2.3	91.0 ± 0.9	98.9 ± 0.5	99.8 ± 0.1	59.1 ± 4.7
	SVM	95.5 ± 0.3	90.9 ± 0.5	87.8 ± 0.8	85.1 ± 1.0	90.9 ± 0.5	96.1 ± 0.2	92.9 ± 0.8	98.5 ± 0.2	83.3 ± 1.0

Table 7: P-value of the paired t-test results for mean accuracy.

-	ABC	GBT	LRC	MLP	RFC	SVM
ABC	-	-	-	-	-	-
GBT	0.2501	-	-	-	-	-
LRC	0.0000	0.0000	-	-	-	-
MLP	0.0897	0.0323	0.0000	-	-	-
RFC	0.0021	0.0012	0.3575	0.0054	-	-
SVM	0.0093	0.0001	0.0000	0.4048	0.0003	-

Table 8: P-value of the paired t-test results for balanced accuracy.

-	ABC	GBT	LRC	MLP	RFC	SVM
ABC	-	-	-	-	-	-
GBT	0.0592	-	-	-	-	-
LRC	0.0000	0.0000	-	-	-	-
MLP	0.5297	0.0406	0.0000	-	-	-
RFC	0.0002	0.0001	0.0071	0.0001	-	-
SVM	0.0106	0.0001	0.0001	0.0003	0.0002	-

Table 9: Classification results in the test subset.

Subset	Model	ACC	AUC	F1-Score	Kappa	Balanced ACC	Precision Class 0	Precision Class 1	Recall Class 0	Recall Class 1
Test	ABC	97.2	99.3	98.3	91.0	94.9	97.9	94.2	98.6	91.3
	GBT	97.2	99.3	98.3	91.1	94.9	97.8	94.7	98.8	91.0
	LRC	91.3	95.3	94.6	71.2	84.4	93.6	80.3	95.7	73.1
	MLP	96.8	99.1	98.0	89.7	94.3	97.7	93.1	98.4	90.3
	RFC	92.1	97.9	95.3	70.3	79.9	91.2	99.1	99.9	59.9
	SVM	96.0	98.3	97.5	86.6	91.6	96.4	93.9	98.7	84.6

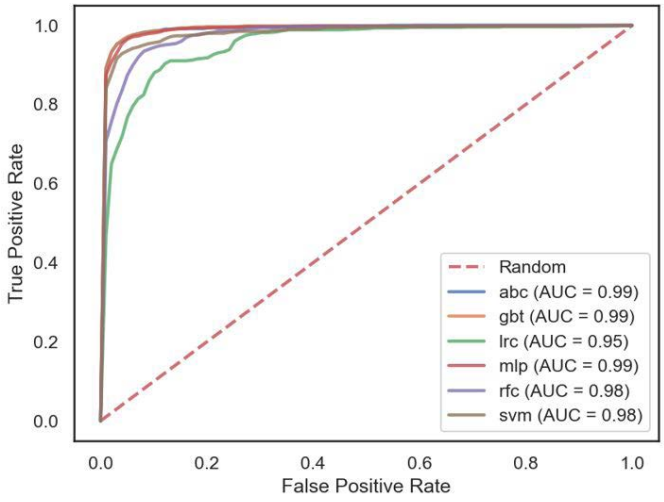

Figure 6: ROC curve for the test subset.

Taking both numerical and graphical test results into account, the ABC and GBT classifiers demonstrated the best overall performance. Although RFC achieved the highest recall for class 0 among all models, it exhibited the lowest recall for class 1. Computational cost is another critical aspect to consider. Being classical machine learning algorithms, all tested models exhibited low computational overhead during training; LRC and RFC achieved the shortest training times, whereas SVM proved to be the most computationally expensive.

Consequently, the selection of the most viable model for practical deployment was refined to ABC and GBT. Given that both classifiers yielded comparable predictive performance on the unseen test set alongside similar computational demands, decision-making required evaluating secondary architectural properties. GBT was ultimately selected due to its inherent interpretability via feature importance rankings, a property lacking in the standard ABC implementation. Interpretability is paramount for post-classification validation, as it enables domain experts to inspect feature attribution and ensure model alignment with real-world decision processes.

In order to analyze the input attributes for the GBT classifier, the feature importance plot is presented in Figure 7. The weights were normalized using min-max scaling to improve visualization of feature importance ranking.

As shown in Figure 7, the most important feature was 'Overlapped Area with Other Rural Properties', which measures the total overlapping area between the recorded property and neighboring rural properties. This finding aligns closely with the work reported by Furumo et al. (2024), who highlighted overlapping CAR registries as a critical issue to address. Another key attribute is the 'Number of Rectifications', corresponding to the frequency of updates made by registrants. The significance of this feature lies in the fact that a registry with a higher number of adjustments is more likely to remain updated; thus, frequent rectifications may indicate proactive correction of inconsistencies.



Figure 7: Normalized feature importance.

While the prediction results show strong potential for real-time deployment, the relatively small sample size relative to the national registry total represents a notable constraint. In particular, the study's regional scope creates a potential susceptibility to context drift, as integrating CAR records from additional states will introduce unobserved data relationships. To mitigate this limitation, retraining the models on newly acquired datasets is planned to preserve continuous classification accuracy.

In summary, the models effectively classified the CAR records, achieving over 90% accuracy on the test set and revealing feature importance rankings that align strongly with current literature.

Conclusions

This study proposed a machine learning model for the automatic classification of Rural Environmental Registry (CAR) records. The Gradient Boosting (GBT) classifier was selected as the best model, achieving metrics above 90% while maintaining interpretability consistent with existing literature. Future work involves implementing this robust method within the CAR platform to support large-scale automated validation and reduce manual workload.

Acknowledgments

The authors gratefully acknowledge the support provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and Zetta Innovation Agency. Their support was essential for the development of this research.

Author contribution

Conceptual idea: Borges, F.E.M.; Ferreira, D.D.; Couto Junior, A.C.S.C.; Methodology design: Borges, F.E.M.; Ferreira, D.D.; Couto Junior, A.C.S.C.; Data collection: Borges, F.E.M.; Ferreira, D.D.; Couto Junior, A.C.S.C.; Data analysis and interpretation: Borges, F.E.M.; Ferreira, D.D.; Couto Junior, A.C.S.C.; Borges, L.A.C.; Nascimento, M.D.A.; Writing and editing: Borges, F.E.M.; Ferreira, D.D.; Couto Junior, A.C.S.C.; Borges, L.A.C.; Nascimento, M.D.A.; Magalhães, R.R.

Data availability statement

Data available upon request to authors.

References

Almeida, J. A. S. D., Sousa Júnior, V. D. P., & Espindola, G. M. D. (2026). Land-tenure dynamics and deforestation pressure in MATOPIBA, Brazil: An assessment of the Rural environmental registry (CAR), 2019-2025. *Land*, *15*(5):700.

Ali, A. et al. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences*, *12*(19):9637.

Arruda, V. et al. (2025). Assessment of the impact of the rural environmental registry on deforestation reduction and environmental compliance in the Brazilian Central-West. *Revista Brasileira de Gestão e Desenvolvimento Regional*, *21*(2):705-730.

Brasil. (2012). Lei nº 12.651, de 25 de maio de 2012. Dispõe sobre a proteção da vegetação nativa; altera as Leis nº 6.938, de 31 de agosto de 1981, nº 9.393, de 19 de dezembro de 1996, e nº 11.428, de 22 de dezembro de 2006; revoga as Leis nº 4.771, de 15 de setembro de 1965, e nº 7.754, de 14 de abril de 1989, e a Medida Provisória nº 2.166-67, de 24 de agosto de 2001; e dá outras providências. *Diário Oficial da União*, Brasília, DF, 28 maio 2012. Available in: <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12651.htm>.

Brasil. Serviço Florestal Brasileiro. (2026). *Painel da regularização ambiental*. Available in: <https://consulta.car.gov.br/>.

Brasil. Receita Federal. (2022). *Dados estatísticos do Cafir*. Available in: <https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/portal-cnir/estatisticas-e-dados-abertos/dados-estatisticos-do-cafir>.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*:5-32.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, *10*(7):1895-1923.

Duda, R. O., & Hart, P. E. (2001). *Pattern classification*. (2nd ed.) New York: John Wiley & Sons. 654p.

Evsukoff, A. G. (2020). *Inteligência computacional: fundamentos e aplicações*. Rio de Janeiro: E-Papers. 272p.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4):367-378.

Furumo, P. R. et al. (2024). Land conflicts from overlapping claims in Brazil's rural environmental registry. *Proceedings of the National Academy of Sciences*, *121*(33):e2407357121.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques.* (3nd ed.) Waltham: Morgan Kaufmann.

Hastie, T. et al. (2009). Multi-class adaboost. *Statistics and its Interface*, *2*(3):349-360.

Hastie, T., Tibshirani, R., Friedman, J. H. (2017). *The elements of statistical learning: data mining, inference, and prediction*. (2nd ed.) New York: Springer. 745p.

Haykin, S. (2007). *Redes neurais: Princípios e prática*. Porto Alegre: Bookman Editora, 898p.

Hocek, H., Yay, S., & Yazir, D. (2025). Comprehensive analysis of ship detention probabilities using binary logistic regression method with machine learning. *Ocean Engineering*, *315*:119889.

Jarmakovica, A. (2025). Machine learning-based strategies for improving healthcare data quality: An evaluation of accuracy, completeness, and reusability. *Frontiers in Artificial Intelligence*, *8*:1621514.

Kumar, R., & Anand, R.S. (2024). Bearing fault diagnosis using multiple feature selection algorithms with SVM. *Progress in Artificial Intelligence*, *13*:119-133.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*:159-174.

Lokanan, M. E. (2025). Enhancing AML compliance: A machine learning approach to suspicious activity detection through routine activity theory. *Journal of Money Laundering Control*, *28*(4-5):680-698.

Mazumder, P. T. (2026). Explainable and fair anti-money laundering models using a reproducible SHAP framework for financial institutions. *Discover Artificial Intelligence*, *6*:262.

Ramos, M. P. et al. (2025). Inovação e controle de qualidade posicional em mapeamento cadastral rural: Um estudo de caso. *Revista Caminhos de Geografia*, *26*(107):14-25.

Ray, N. G. et al. (2025). Isolated middle cerebral artery significantly increases risk of postoperative stroke after transcarotid revascularization in asymptomatic patients. *Journal of Vascular Surgery*, *82*:1288-1298.

Silva, A. A. et al. (2023). Green grabbing in the Matopiba agricultural frontier. *IDS Bulletin*, *54*(1):57-72.

Silva, V. N. da. et al. (2025). A aplicabilidade do código florestal de 2012 em Rondônia: Uma análise da efetividade dos instrumentos CAR e PRA. *IOSR Journal of Business and Management (IOSR-JBM)*, *27*(7):19-27.

Thanh, C. N., & Huy, T. P. (2025). Predicting financial reports fraud by machine learning: The proxy of auditor opinions. *Cogent Business & Management*, *12*(1).

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley, 736p.

Yu, Z., Yu, B., & Li, B. (2025). Enhancing failure mode classification of RC beam-column joints using logistic regression and hybrid sampling strategy. *Engineering Structures*, *327*:119542.