EDU REKHA INTERNATIONAL JOURNAL OF ARTS, LAW AND SOCIAL SCIENCE

Journal Homepage: https://edurekhapublisher.com/erijalss/

Volume- 2 Issue -2 (March-April) 2026

ISSN: 3107-5169 (Online)

Frequency: Bimonthly

PAGES: 49-62

ARTICLE TITLE:

Application of Machine Learning Techniques to Predict Students at Risk of Attrition in a Federal University

Milena Ester de Almeida[1], Gilberto Venâncio Luiz[2*] & José Antônio de Babos Mendes[3]

Federal University of Viçosa – Campus Rio Paranaíba, Brazil.

ARTICLE HISTORY

RECEIVED
05-03-2026

ACCEPTED
08-03-2026

PUBLISHED
22-03-2026

Corresponding author:

Gilberto Venâncio Luiz

Federal University of Viçosa – Campus Rio Paranaíba, Brazil.

Abstract

Student dropout in higher education remains a persistent challenge for public universities, generating significant academic, social, and economic impacts. Early student withdrawal compromises the efficiency of public investment in education, reduces graduation rates, and limits students' professional opportunities. In this context, the use of machine learning techniques has shown promising potential for identifying patterns associated with dropout risk and supporting institutional decision-making. This study aimed to apply and compare different supervised machine learning algorithms to predict student dropout in a Brazilian federal university, as well as to identify the main factors associated with academic attrition. The research adopted a quantitative approach using an institutional dataset comprising 20,275 students admitted since 2010 across three campuses of the university. Several classification algorithms were tested, including Neural Network, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Naive Bayes, and Logistic Regression. Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC), along with model interpretability through the SHAP technique. The results indicated strong predictive performance across the models, with Gradient Boosting demonstrating the best overall results. The most influential predictors of dropout were cumulative grade point average and the number of course failures. The findings suggest that machine learning models can support the early identification of at-risk students and contribute to institutional retention policies.

Keywords: Student dropout, Machine learning, Higher education, Predictive analytics.

1. INTRODUCTION

Student attrition in higher education is a recurrent phenomenon in public universities across different countries and produces significant institutional, economic, and social effects. A reduction in the number of enrolled students compromises the financial sustainability of institutions, particularly those in which the allocation of public resources is linked to the number of students formally enrolled in academic programs (Souza et al., 2024). In addition, academic attrition implies a waste of public investment in higher education, without the expected social return being fully achieved (Sandoval-Palis, Naranjo, Gilar-Corbí, 2020).

The impacts of attrition extend beyond the institutional sphere and directly affect students and society. Dropping out of academic programs limits opportunities for insertion into the qualified labor market and contributes to trajectories marked by underemployment, with negative consequences for the living conditions of individuals, their families, and the communities in which they are embedded (Asha et al., 2020). At the institutional level, high attrition rates may affect universities' public image, associating them with difficulties in providing adequate conditions for student persistence and degree completion (Pattanaphanchai, Leelertpanyakul, Theppalak, 2019).

The specialized literature indicates that student attrition results from the interaction among individual, academic, institutional, and socioeconomic factors. Studies such as those by Esteves et al. (2021) point to the combined influence of social, cultural, and economic variables on the decision to withdraw. Complementarily, Fritsch et al. (2015) and Wilhelm and Schlosse (2019) highlight the role of institutional practices, particularly those related to assessment processes and academic management. In a convergent manner, Sandoval-Palis, Naranjo, and Gilar-Corbí (2020) identify demographic characteristics, prior educational background, academic performance, and socioeconomic conditions as recurrent antecedents of attrition. Similar findings are reported by Fior (2021) and Souza et al. (2024), who show a stronger influence of personal and academic variables on the risk of attrition.

Given the multifactorial nature of attrition, the early identification of students at risk has been identified as a central element in the development of institutional persistence policies. In this context, Machine Learning techniques have been employed to model and predict the probability of attrition based on multiple factors, including academic and sociodemographic data (Núñez-Naranjo et al., 2021, Mohammad et al., 2023). Recent evidence indicates that machine learning algorithms outperform traditional statistical methods in predicting student attrition and retention (Utomo, Purwanto, Surarso, 2023).

These models enable the classification of students into different risk levels, providing more precise inputs for the formulation of targeted interventions. The literature highlights that continuous monitoring of academic performance, especially during the initial semesters, contributes to greater predictive accuracy and to the adoption of more appropriate preventive actions (Ameri et al., 2016, Ortiz-Lozano, Rua-Vieites, Bilbao-Calabuig, 2018). In addition, the results derived from these models can guide student advising strategies and personalized support programs, strengthening institutional policies aimed at persistence and retention (Utomo, Purwanto, Surarso, 2023).

Within this framework, the present study seeks to answer the following research question: which supervised machine learning algorithm demonstrates the best predictive performance in forecasting academic attrition at a federal university located in the interior of the state of Minas Gerais, and which institutional and academic variables exert the greatest influence on this process? To this end, the study analyzes data from three campuses of a Brazilian federal university, selected due to its relevance within the higher education system and the persistence of challenges related to student attrition, according to data from the Higher Education Census (INEP, 2023). The objective of the study is to evaluate and compare the performance of different supervised machine learning algorithms in predicting academic attrition, identifying the model with the highest discriminative capacity and the main factors associated with attrition risk.

Predicting attrition risk through Machine Learning techniques enables the early identification of students with a higher propensity to withdraw, facilitating the adoption of institutional persistence policies that are more strongly guided by evidence. The literature indicates that proactive interventions, combined with continuous monitoring of academic performance, contribute to improved retention rates and to students' academic trajectories (Ameri et al., 2016). In particular, monitoring performance during the initial semesters has been associated with increased predictive accuracy and with the implementation of more appropriate preventive actions (Ortiz-Lozano et al., 2018).

2. LITERATURE REVIEW

2.1. Student Attrition in Higher Education

Attrition in higher education constitutes a multifaceted phenomenon resulting from the interaction among individual, academic, institutional, and socioeconomic factors, with relevant consequences for both students and educational institutions (Souza et al., 2024). The literature has emphasized the importance of understanding these factors in an integrated manner, since they operate cumulatively throughout the academic trajectory.

Among socioeconomic factors, financial difficulties stand out as one of the main determinants of withdrawal. Neupane (2024) shows that economic constraints compromise students' ability to continue their studies, particularly in contexts of social vulnerability. Similarly, Bayona-Oré (2022) points out that family background and community support influence academic persistence, with greater obstacles faced by students from lower socioeconomic strata.

At the institutional level, insufficient academic support, limited resources, and low levels of institutional engagement are associated with higher attrition rates (Bonilla-Jurado et al., 2023; Bayona-Oré, 2022). These institutional factors are intertwined with motivational dimensions, such as student demotivation and lack of interest, highlighted by Meedech et al. (2016) and Baulke et al. (2022). In addition, identification with the program and the chosen career has been identified as a relevant element for persistence, insofar as congruence between vocation, professional expectations, and educational experience influences the decision to continue or withdraw (Guimarães et al., 2019; Bayona-Oré, 2022).

Academic factors also exert influence on attrition, especially those related to program difficulty, assessment policies, and teaching quality (Bayona-Oré, 2022). Empirical evidence indicates that low performance in foundational subjects, such as mathematics, English,

chemistry, and psychology, is associated with higher probabilities of withdrawal, constituting one of the main predictors of academic retention (Aulck et al., 2016). Additionally, characteristics such as age and prior educational background are associated with attrition risk, reflecting distinct educational trajectories.

Sociodemographic variables further broaden understanding of the phenomenon by revealing structural inequalities. Barreto et al. (2019) observe higher attrition among both younger students and those of older age, as well as differences associated with program and gender. Santos et al. (2023) complement these findings by linking withdrawal to family income, parental education, type of prior school, and early entry into the labor market. Consistently, Herbaut (2021) indicates that students' social origin is associated with academic failure and withdrawal.

Other studies reinforce the role of additional demographic and economic variables, such as marital status and financial situation, in higher education persistence (Sani et al., 2020; Tayebi et al., 2021). Academic performance, in turn, remains one of the most recurrent indicators of attrition. Rosa et al. (2021) indicate that performance in secondary education and in the initial periods of higher education is associated with withdrawal, while Herbaut (2021) highlights that failures in the first year increase the likelihood of dropping out.

Beyond performance, academic engagement has been identified as a factor associated with persistence. Low attendance at academic activities and limited participation in classes and curricular activities indicate a fragile bond with the program and the institution, thereby favoring withdrawal (Casanova et al., 2018; Esteves et al., 2021).

In summary, the literature shows that attrition in higher education results from the combination and interdependence of academic, financial, institutional, and sociodemographic factors. Studies such as Nurmalitasari et al. (2023) reinforce this perspective by identifying academic performance and personal and family economic conditions as central elements in understanding withdrawal, indicating the need for integrated analytical approaches to address the phenomenon.

2.2. Supervised Machine Learning Techniques

Supervised machine learning, a subfield of artificial intelligence, is based on training algorithms using labeled data with the aim of predicting or classifying future outcomes. This approach enables the identification of patterns in historical data and their application to new datasets, and it is widely used in fields such as health and finance, especially in contexts that require systematic decision support (Burkart & Huber, 2020).

In this type of learning, algorithms are trained on previously known input–output pairs, in which explanatory variables are associated with labels representing the outcome to be predicted (Sem et al., 2019; Narula, 2023). Through this process, models learn underlying relationships in the data and then generalize this knowledge to observations not seen during training. Supervised learning problems are traditionally classified into two main categories, classification, when outcomes are categorical, and regression, when outcomes assume continuous values (Fabris et al., 2017).

In the context of student attrition in higher education, supervised learning has been widely used to classify students according to attrition risk, typically based on binary labels such as "completed the program" or "withdrew from the program." This approach enables early identification of profiles that are more susceptible to attrition, providing support for data-driven institutional interventions.

The literature reports the use of different families of algorithms for predicting academic attrition. Tree-based methods, such as Decision Trees and Random Forests, are frequently employed due to their robustness and relative interpretability (Kanil & Kiran, 2024; Esquivel, 2023). Ensemble techniques, such as Gradient Boosting, AdaBoost, and XGBoost, have also been widely adopted and have shown superior performance in several empirical studies (Villar & Andrade, 2024; Martins et al., 2021; Niyogisubizo et al., 2022). In addition, models based on artificial neural networks and support vector machines have demonstrated strong performance, particularly in datasets with greater complexity and dimensionality (Sulak & Koklu, 2024; Jiménez-Gutiérrez et al., 2024).

Although some studies report high levels of predictive performance, such as elevated accuracy rates or AUC values, the literature emphasizes that these results are strongly conditioned by dataset characteristics, class balance, and the validation strategies adopted (Osemwegie et al., 2023; Narula, 2023). In this regard, there is no consensus on the universal superiority of a single algorithm, and the simultaneous comparison of multiple models is a recurrent practice in order to identify the most suitable approach for a given empirical context.

The development of robust and generalizable predictive models remains one of the main challenges in applying machine learning to student attrition. Issues such as class imbalance, heterogeneity of institutional data, and the need for external validation require rigorous performance assessment (Jiang et al., 2020). In response, several studies adopt comparative approaches, evaluating different supervised algorithms in parallel to select the model that demonstrates the best predictive performance and greatest stability, a strategy that guides the design of the present study.

2.3. Evaluation of Supervised Machine Learning Models

Differences among supervised machine learning algorithms may result in relevant variations in the predictive performance of models. In this context, performance metrics play a central role in evaluating the effectiveness of algorithms across different tasks, enabling systematic comparisons and the identification of their strengths and limitations.

Several studies have analyzed and compared machine learning algorithms using performance metrics that are well established in the literature. Accuracy is one of the most widely used metrics and corresponds to the proportion of correct predictions relative to the total number of predictions made. Its values range from 0 to 1, with higher values indicating greater agreement between model predictions and observed outcomes (Sarraju Pal & Kamilya, 2022; Doulah & Islam, 2023).

Precision measures the proportion of true positives relative to the total number of instances classified as positive by the model. This metric is particularly relevant in contexts in which false positives entail high costs, such as applications in the health sector (Kumar et al., 2017; Kebede et al., 2022). Complementarily, sensitivity, also referred to as recall or the true positive rate, assesses the model's ability to correctly identify actual positive cases, and is widely used in situations in which failure to detect positive events represents a significant concern (Angayarkanni et al., 2023; Kebede et al., 2022).

Specificity, in turn, corresponds to the proportion of true negatives relative to the total number of actual negatives, contributing to the balance between the correct identification of positive and negative cases (Doulah & Islam, 2023; Sarraju Pal & Kamilya, 2022). To integrate the information provided by precision and sensitivity into a single measure, the F1-score is used, defined as the harmonic mean of these two metrics. Higher F1-score values indicate a better balance between correctly identifying positive cases and reducing false alarms, making this metric especially useful in contexts with imbalanced classes (Caruana & Niculescu-Mizil, 2004).

Another widely used metric is the area under the ROC curve (AUC-ROC), which evaluates the model's discriminative ability between classes. The ROC curve relates the true positive rate and the false positive rate across different decision thresholds, while the AUC summarizes the overall performance of the classifier. Higher AUC values indicate a greater capacity to distinguish between the analyzed classes (Kebede et al., 2022; Luiz, 2024).

Finally, the root mean squared error (RMSE) is traditionally used in regression problems and represents the average of the squared differences between predicted and observed values. Lower RMSE values indicate a better model fit to the data, and this metric is used in a complementary manner in studies involving the prediction of continuous variables or the assessment of estimation errors (Doulah & Islam, 2023; Caruana & Niculescu-Mizil, 2004).

3. METHODOLOGICAL PROCEDURES

3.1. Type of Research

This study is characterized as a quantitative and descriptive research design. The quantitative approach was adopted because the research involves the collection and analysis of numerical data, including students' sociodemographic and academic characteristics, in order to test machine learning (ML) models for predicting the probability of attrition. The descriptive nature of the study is evidenced by the objective of observing, describing, and interpreting attrition patterns among university students, without intervention in the variables analyzed.

In this project, a supervised Machine Learning approach will be employed, which relies on a labeled dataset to train the model to make predictions. The focus of the analysis is classification, where the objective is to categorize students into two groups: those who completed the program and those who dropped out. This approach is particularly useful in scenarios where the outcome is binary or multicategorical.

The supervised approach is appropriate for this study because the output variables (attrition or completion) are already known in the historical data, and the objective is to train the model to identify patterns in the input variables (such as sociodemographic characteristics and academic performance) that influence this outcome. The model will learn from the historical data and, after training, will be able to make predictions on new data from currently enrolled students at the institution, identifying students who are at risk of attrition.

3.2. Data Collection and Preparation

The data used in this study were obtained from the academic system of a Federal University located in the interior of the state of Minas Gerais, Brazil, which contains academic and sociodemographic information on 20,275 students who either graduated from or dropped out of the institution between 2010 (the year corresponding to the beginning of the use of the National High School Examination, ENEM) and 2024. Of this total, 2,878 students were from campus 1, 3,406 from campus 2, and 13,900 from campus 3 (the main campus). The database included the variables described below.

Table 1: Research variables and operationalization

Variable	Measurement/Operationalization
Academic status	Categorical variable indicating the student's academic status: graduated or attrition.
Father's education	Ordinal variable representing the highest level of formal education completed by the father.
Mother's education	Ordinal variable representing the highest level of formal education completed by the mother.
Household income	Ordinal variable indicating the household income level.
Race and ethnicity	Categorical variable based on self-reported race and ethnicity, according to the classification of the Brazilian Institute of Geography and Statistics (IBGE).
Year of birth	Continuous variable indicating the student's year of birth.
Type of previous school	Categorical variable indicating whether the student attended a public or private high school.
Field of study	Categorical variable grouping programs into major knowledge areas: Humanities and Social Sciences, Biological and Health Sciences, Exact Sciences, and Agricultural Sciences.
Type of degree program	Categorical variable indicating the type of undergraduate program: bachelor's degree or teaching degree.
Program schedule	Categorical variable indicating the program schedule: full-time or evening.
Total ENEM score	Continuous variable representing the student's overall performance on the National High School Examination (ENEM) at the time of admission.
ENEM Mathematics score	Continuous variable representing the student's performance on the ENEM Mathematics exam.
ENEM Essay score	Continuous variable representing the student's performance on the ENEM Essay exam.

Total course failures	Discrete count variable indicating the total number of course failures throughout the academic trajectory.
Failures due to absenteeism	Discrete count variable indicating the number of course failures attributed to absenteeism.
Cumulative grade point average (CGPA)	Continuous variable representing the student's cumulative academic performance across all completed semesters.
First-semester grade point average (FSGPA)	Continuous variable representing the student's academic performance in the first semester of the program.
Years since admission	Continuous variable measuring the elapsed time, in years, until graduation or attrition.
Context in relation to the COVID-19 pandemic	Categorical variable classifying student admission as occurring in the pre-pandemic or post-pandemic period.

Note. ENEM = National High School Examination; CGPA = cumulative grade point average; FSGPA = first-semester grade point average.

During data preparation, an initial data cleaning procedure was performed, removing records with incomplete information and those corresponding to other admission pathways different from ENEM. In addition, numerical data were normalized to ensure that all variables had the same scale. After these procedures, the final dataset was obtained, consisting of students in two categories for each of the three campuses: those who completed their programs and those who dropped out.

3.3. election of Algorithms and Model Evaluation Metrics

For the predictive analysis of student attrition, the following supervised Machine Learning algorithms were applied, as described in Table 2. These algorithms were selected based on the literature review, which indicated them as the most used and those achieving higher accuracy in classifying students at risk of course withdrawal.

Table 2: Algorithms used in the data analysis

Algorithm	Hyperparameters
Neural Network	The Multilayer Neural Network was structured with one hidden layer containing 100 neurons, ReLU activation, Adam optimizer, regularization $\alpha = 0.0001$, and a maximum of 150 iterations.
Decision Tree	The Decision Tree was induced in binary form, with a minimum of 2 instances per leaf and a maximum depth of 100 levels.
Random Forest	The Random Forest used 10 trees and a minimum split of 5 instances.
Gradient Boosting	Gradient Boosting (scikit-learn) was configured with 100 estimators, a learning rate of 0.10, and a maximum depth of 3.
AdaBoost	AdaBoost employed 50 estimators with a learning rate of 1.0.
Naive Bayes	Naive Bayes was executed using the system's default parameters.
Logistic Regression	Logistic Regression used L2 regularization ($C = 1$).

Predictive modeling was conducted using the Orange Data Mining software (Orange, 2025), adopting a standardized workflow consisting of preprocessing, training, and comparative evaluation of supervised classification algorithms for attrition risk prediction. Preprocessing was performed using the Preprocess module. Categorical variables were encoded using one-hot encoding (one feature per value). Missing values were imputed using the mean for continuous variables and the most frequent category for categorical variables. Numerical variables were standardized to zero mean and unit variance ($\mu = 0$; $\sigma^2 = 1$). All transformations were applied within the cross-validation workflow to prevent information leakage between training and test sets.

Algorithm performance was evaluated using stratified 10-fold cross-validation, implemented through the Test and Score module in Orange Data Mining. In each iteration, nine partitions were used for training and one for validation, preserving the original class proportions. The performance of the classification models was assessed using the metrics AUC, Accuracy (CA), Precision, Recall, F1-score, and the Matthews Correlation Coefficient (MCC). AUC measures the overall discriminative ability of the model, while Accuracy indicates the total proportion of correct classifications. Precision and Recall evaluate, respectively, the reliability of positive predictions and the ability to correctly identify attrition cases. The F1-score summarizes the balance between Precision and Recall. MCC was used as a robust measure of overall model performance, particularly relevant in scenarios with potential class imbalance. The reported metrics correspond to the mean values obtained from stratified cross-validation (10 folds).

The interpretability of the selected model was analyzed using SHAP values (SHapley Additive exPlanations), implemented through the Explain Model module in Orange. The positive class was defined as attrition (1). The ten variables with the highest mean absolute importance were examined, with positive SHAP values interpreted as increasing the predicted probability of attrition and negative values as decreasing it.

3.4. Ethical Issues

It is worth noting that this research complied with applicable ethical principles, ensuring anonymity and confidentiality of student data, which were obtained with institutional authorization and used exclusively for academic and scientific purposes. During the preparation of this study, the authors used the artificial intelligence tools Scispace and Elicit for literature retrieval, NotebookLM for literature organization, reading, and summarization, and ChatGPT 4.0 to improve the text by correcting spelling and grammatical errors. After using these tools, the authors reviewed and edited the content in accordance with the scientific method and assume full responsibility for the content of the publication.

4. RESULTS AND DISCUSSION

4.1. Algorithm Evaluation and Selection

4.1.1. Campus 1 Analysis

In the analysis of Campus 1 data, the evaluation of supervised algorithms using stratified cross-validation (Table 3) indicated that the Gradient Boosting model achieved the best overall performance in predicting academic attrition, with Accuracy (CA) = 0.929 and a Matthews Correlation Coefficient (MCC) = 0.852. Accuracy represents the overall proportion of correct classifications, whereas MCC constitutes a robust measure of overall model performance, as it simultaneously considers true positives, true negatives, false positives, and false negatives, making it particularly informative in scenarios with potential class imbalance.

The Neural Network and Logistic Regression models showed similar performance, with CA = 0.919 and 0.918, respectively, and MCC = 0.831 in both cases. Random Forest achieved CA = 0.906 and MCC = 0.805. The AdaBoost, Decision Tree, and Naive Bayes algorithms exhibited lower performance. The paired statistical comparison based on AUC did not reveal a statistically significant difference between Gradient Boosting and the other best-performing models ($p \geq 0.05$).

Table 3: Performance of supervised models in predicting attrition, Campus 1

Model	AUC	Accuracy (CA)	F1	Precision	Recall	MCC
Gradient Boosting	0,974	0,929	0,929	0,931	0,929	0,852
Neural Network	0,971	0,919	0,919	0,920	0,919	0,831
Logistic Regression	0,965	0,918	0,919	0,920	0,918	0,831
Random Forest	0,958	0,906	0,907	0,908	0,906	0,805
Naive Bayes	0,928	0,868	0,870	0,885	0,868	0,745
AdaBoost	0,889	0,896	0,896	0,896	0,896	0,780
Tree	0,876	0,909	0,909	0,909	0,909	0,808

However, significant differences were observed between lower-performing models and the superior models ($p \leq 0.020$), indicating heterogeneity in the discriminative capacity among the evaluated algorithms. Considering the higher Accuracy and the highest Matthews Correlation Coefficient (MCC), the Gradient Boosting algorithm was selected as the most appropriate model for predicting academic attrition in the analyzed context. Additionally, the sequential boosting algorithm demonstrates greater capacity to model nonlinear relationships and complex interactions among predictor variables (Friedman, 2001), a characteristic that is particularly relevant in multifactorial phenomena such as educational attrition, which involves interdependent academic, institutional, and socioeconomic factors.

4.1.2. Campus 2 Analysis

The comparison of supervised algorithms for Campus 2 (Table 4) also indicated that Gradient Boosting achieved the best overall performance in predicting academic attrition (AUC = 0.960; CA = 0.898; F1 = 0.899; MCC = 0.781). The AUC demonstrated high discriminative capacity, while the Matthews Correlation Coefficient confirmed greater overall agreement between the observed and predicted classes. Neural Network (AUC = 0.949; MCC = 0.749) and Logistic Regression (AUC = 0.948; MCC = 0.749) showed similar performance, although lower across all evaluated metrics.

The Random Forest, AdaBoost, Decision Tree, and Naive Bayes models demonstrated lower discriminative capacity in prediction. However, the paired statistical comparison among the algorithms, based on the mean AUC obtained through cross-validation, indicated that there was no statistically significant difference between Gradient Boosting and the other best-performing models ($p \geq 0.05$). Specifically, comparisons with Neural Network ($p = 0.995$), Logistic Regression ($p = 1.000$), and Random Forest ($p = 1.000$) did not demonstrate statistical superiority of Gradient Boosting, despite its presenting the highest mean AUC (0.960).

Table 4: Performance of supervised models in predicting attrition, Campus 2

Model	AUC	Accuracy (CA)	F1	Precision	Recall	MCC
Gradient Boosting	0,959	0,898	0,899	0,901	0,898	0,781
Logistic Regression	0,948	0,885	0,886	0,888	0,885	0,753
Neural Network	0,947	0,882	0,882	0,883	0,882	0,742
Random Forest	0,939	0,877	0,877	0,877	0,877	0,730

Naive Bayes	0,890	0,814	0,817	0,827	0,814	0,613
AdaBoost	0,846	0,859	0,859	0,859	0,859	0,689
Tree	0,804	0,864	0,864	0,864	0,864	0,701

This result suggests statistical equivalence among the main evaluated algorithms in terms of discriminative capacity. The absence of statistical significance may be attributed to the small magnitude of the observed absolute differences (ΔAUC $\approx$ 0.011–0.012), to the variability inherent in cross-validation folds, and to the potentially limited statistical power of the multiple comparison test employed.

On the other hand, statistically significant differences were identified in comparisons involving lower-performing models, particularly the standalone Decision Tree (Tree), which showed significant differences in relation to Random Forest, Gradient Boosting, Logistic Regression, and Neural Network ($p \leq 0.003$). Additionally, differences were observed between AdaBoost and Naive Bayes ($p = 0.006$), as well as between Random Forest and Neural Network ($p = 0.044$), indicating heterogeneity in discriminative capacity among intermediate and lower-performing models.

Although the comparison based exclusively on AUC did not confirm statistical superiority of Gradient Boosting over its main competitors, the integrated analysis of overall performance metrics reinforces its practical advantage. Gradient Boosting achieved the highest Accuracy (CA = 0.898) and the highest Matthews Correlation Coefficient (MCC = 0.781). Because MCC simultaneously incorporates true positives, true negatives, false positives, and false negatives, it constitutes a particularly informative measure in scenarios with potential class imbalance, reflecting the overall quality of classification more robustly.

Thus, considering the superior consistency observed across performance metrics (AUC, CA, F1, and MCC), combined with the algorithm's ability to model nonlinear interactions and complex relationships among variables (Friedman, 2001), Gradient Boosting was selected as the final predictive model. This decision is based not only on statistical significance but also on the convergence of discriminative performance, overall robustness, and stability across evaluation metrics.

4.1.3. Campus 3 Analysis

For Campus 3, the comparison of supervised algorithms indicated that the Gradient Boosting model achieved the best overall performance in predicting academic attrition (AUC = 0.955; CA = 0.913; F1 = 0.911; MCC = 0.816). The high AUC demonstrates consistent discriminative capacity, while the Matthews Correlation Coefficient indicates greater overall agreement between the observed and predicted classes.

Neural Network exhibited similar performance (AUC = 0.950; MCC = 0.802), although it remained inferior across all evaluated metrics. Random Forest and Logistic Regression showed an additional reduction in discriminative capacity and overall robustness. In contrast, models such as Naive Bayes, AdaBoost, and Decision Tree demonstrated substantially lower performance.

Consistent superiority of Gradient Boosting in the Accuracy and MCC metrics, combined with its ability to model nonlinear relationships and complex interactions among variables, supports its selection as the final predictive model. These findings suggest that approaches based on sequential boosting are more suitable for modeling the multifactorial phenomenon of academic attrition in the analyzed context.

Table 5: Performance of supervised models in predicting attrition, Campus 3

Model	AUC	Accuracy (CA)	F1	Precision	Recall	MCC
Gradient Boosting	0,955	0,913	0,911	0,916	0,913	0,816
Neural Network	0,950	0,907	0,906	0,907	0,907	0,802
Random Forest	0,935	0,899	0,897	0,900	0,899	0,785
Logistic Regression	0,927	0,891	0,889	0,894	0,891	0,770
Naive Bayes	0,879	0,840	0,839	0,839	0,840	0,659
AdaBoost	0,856	0,862	0,862	0,863	0,862	0,710
Tree	0,844	0,879	0,878	0,878	0,879	0,743

Using stratified cross-validation (10 folds), the comparison of supervised algorithms again indicated that the Gradient Boosting model achieved the best overall performance in predicting academic attrition (AUC = 0.955; CA = 0.913; F1 = 0.911; MCC = 0.816). High AUC values indicate consistent discriminative capacity, while the Matthews Correlation Coefficient reflects stronger overall agreement between observed and predicted classes.

Performance of the Neural Network remained close (AUC = 0.950; MCC = 0.802), but still inferior across all evaluated metrics. Random Forest and Logistic Regression showed further reductions in discriminative capacity and global robustness, whereas Naive Bayes, AdaBoost, and Decision Tree produced substantially lower predictive performance.

Overall superiority of Gradient Boosting in Accuracy and MCC, together with its capacity to capture nonlinear relationships and

complex interactions among predictors, supports its adoption as the final predictive model. These results indicate that sequential boosting approaches are particularly suitable for modeling the multifactorial dynamics of academic attrition in the analyzed context (Friedman, 2001).

4.2. Impact of Variables on the Prediction Model

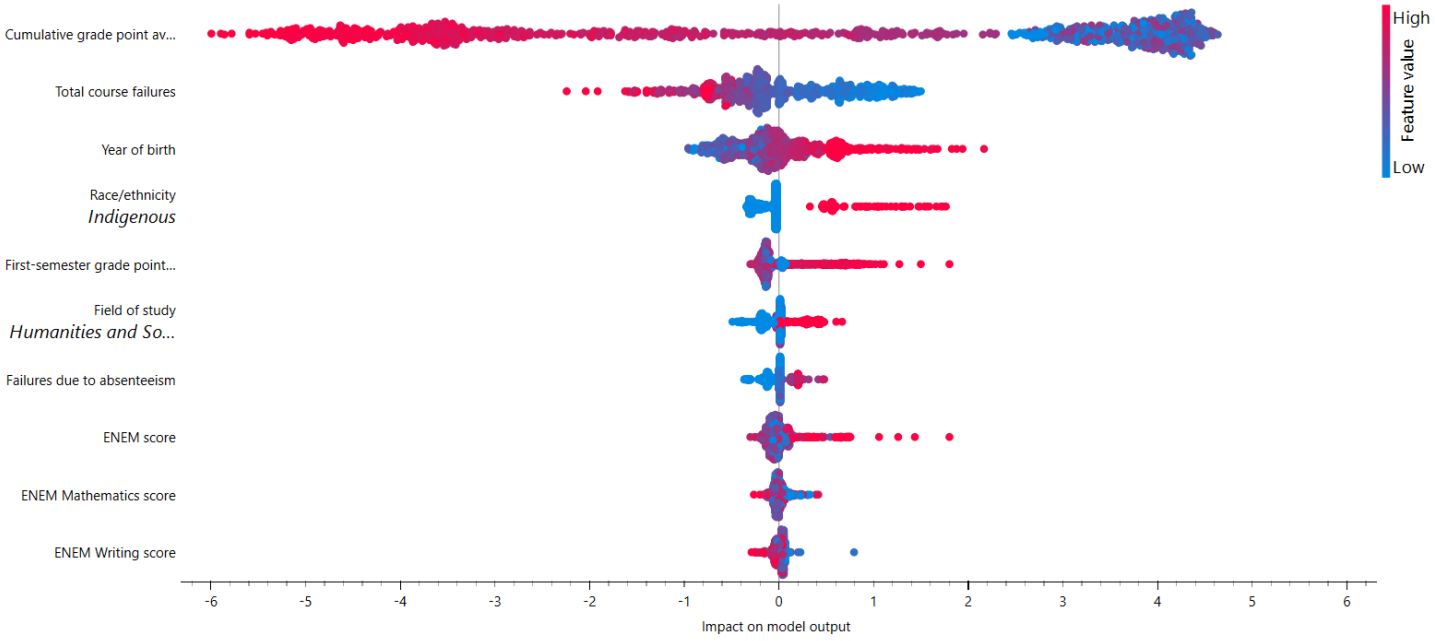
4.2.1. Campus 1 Results

The interpretability analysis using SHAP (Figure 1) values made it possible to identify the variables with the greatest contribution to attrition prediction for Campus 1, considering attrition as the positive class (1). The variables are ordered according to the mean magnitude of their absolute impact on the model.

The variable Cumulative grade point average (cumulative GPA) exhibited the greatest predictive impact. Low GPA values (blue points) are associated with strong positive contributions to the log-odds of attrition, substantially increasing the predicted probability of the event. In contrast, high GPA values (red points) contribute negatively to the model output, reducing the probability of attrition. This pattern indicates a consistent inverse relationship between cumulative academic performance and attrition risk.

In addition, the variable Total course failures also demonstrated a relevant contribution. Higher numbers of failures (red) shift SHAP values toward the positive region, increasing the estimated probability of attrition. Conversely, lower numbers of failures (blue) show a negative or near-zero impact, suggesting a protective effect.

Figure 1: SHAP plot of the Gradient Boosting model for attrition prediction, Campus 1



Note. The plot presents SHAP values (SHapley Additive exPlanations), which indicate the marginal impact of each variable on the model output. The horizontal axis represents the impact on the prediction (positive values increase the predicted probability of attrition; negative values decrease it). The variables are ordered by mean absolute importance. The color scale represents the value of the predictor variable (blue = low values; red = high values).

Regarding demographic characteristics, the Year of birth variable showed a moderate effect. More recent values (younger students) tended to produce positive impacts on the model, whereas earlier birth years (older students) tended to contribute negatively. This pattern suggests a higher predicted probability of attrition among younger students.

Similarly, the variable Race/ethnicity – Indigenous indicated that belonging to the Indigenous category (high values) was associated with positive impacts on the model, increasing the predicted probability of attrition, whereas non-membership showed contributions close to zero or slightly negative.

Another relevant academic indicator, the First-semester grade point average (first-semester GPA), displayed a pattern similar to cumulative GPA, although with smaller magnitude. Lower values contributed positively to attrition prediction, whereas higher values produced a negative effect, indicating that early academic performance exerts a relevant influence on the subsequent risk of withdrawal.

With respect to academic field, the variable Field of study – Humanities and Social Sciences showed a discrete to moderate impact. Belonging to this field was associated with a slight increase in the predicted probability of attrition, although the considerable dispersion of SHAP values suggests heterogeneity within the field. In addition, Failures due to absenteeism showed positive contributions when occurring at higher levels, indicating that recurrent absence constitutes a factor associated with increased predicted attrition risk.

Finally, the ENEM score (overall ENEM score) demonstrated a modest protective effect. Higher scores contributed negatively to the model output, reducing the predicted probability of attrition, whereas lower scores tended to slightly increase the estimated risk. Similarly, the ENEM Mathematics and ENEM Writing scores showed relatively small impacts. Lower values were associated with small positive contributions to the model, whereas higher values tended to reduce the predicted probability of attrition. However, the range of SHAP values indicates that these variables have lower discriminative power compared with university academic performance indicators.

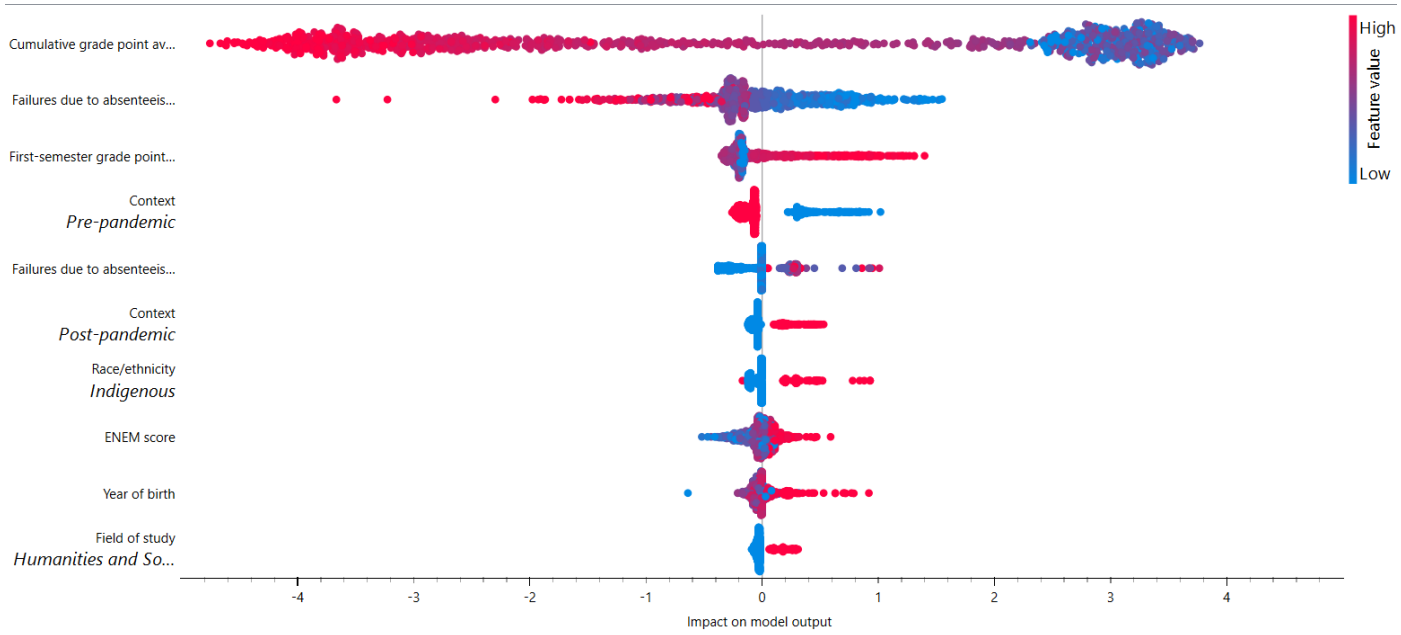
4.2.2. Campus 2 Results

The analysis of SHAP values for Campus 2 (Figure 2) showed that variables related to academic performance remained the main determinants of the predicted probability of attrition, although with changes in their relative order of importance compared with Campus 1.

The variable Cumulative grade point average (cumulative GPA) exhibited the highest mean absolute impact on the model. A consistent inverse pattern can be observed: low GPA values (blue) were associated with positive SHAP values, increasing the log-odds of attrition, whereas higher values (red) contributed negatively, reducing the predicted probability of the event. The range of SHAP values indicates a substantial effect of this variable in distinguishing between students who dropped out and those who did not.

In addition, the variable Failures due to absenteeism showed the second highest predictive contribution. Higher frequencies of absence-related failures shifted SHAP values toward the positive region, increasing the estimated probability of attrition. In contrast, lower levels of absence-related failures showed negative or near-zero impacts. This finding suggests that academic engagement, indirectly measured through attendance patterns, has strong predictive relevance in Campus 2.

Figure 2: SHAP plot of the Gradient Boosting model for attrition prediction, Campus 2



Similarly, the First-semester grade point average (first-semester GPA) also demonstrated a relevant impact. Lower values contributed positively to the predicted risk of attrition, whereas higher academic performance exerted a protective effect. This pattern indicates that initial academic performance continues to influence the subsequent academic trajectory.

With respect to contextual variables, those related to the pandemic period showed distinct effects. In Context – Pre-pandemic, the corresponding category exhibited predominantly negative or near-zero contributions, suggesting a lower predicted probability of attrition during this period. Conversely, Context – Post-pandemic showed positive SHAP values for the active category, indicating an increase in the estimated probability of attrition in the post-pandemic period. This result suggests a possible influence of structural or academic changes associated with the post-pandemic context.

Regarding demographic factors, the variable Race/ethnicity – Indigenous showed positive contributions when present, indicating an increase in the predicted probability of attrition among Indigenous students in Campus 2. However, the dispersion of SHAP values suggests heterogeneity within the group.

Moreover, the ENEM score (overall ENEM score) demonstrated a modest protective effect. Higher scores contributed negatively to the log-odds of attrition, whereas lower scores showed a modest positive impact. The magnitude of this effect was smaller than that observed for internal academic performance indicators.

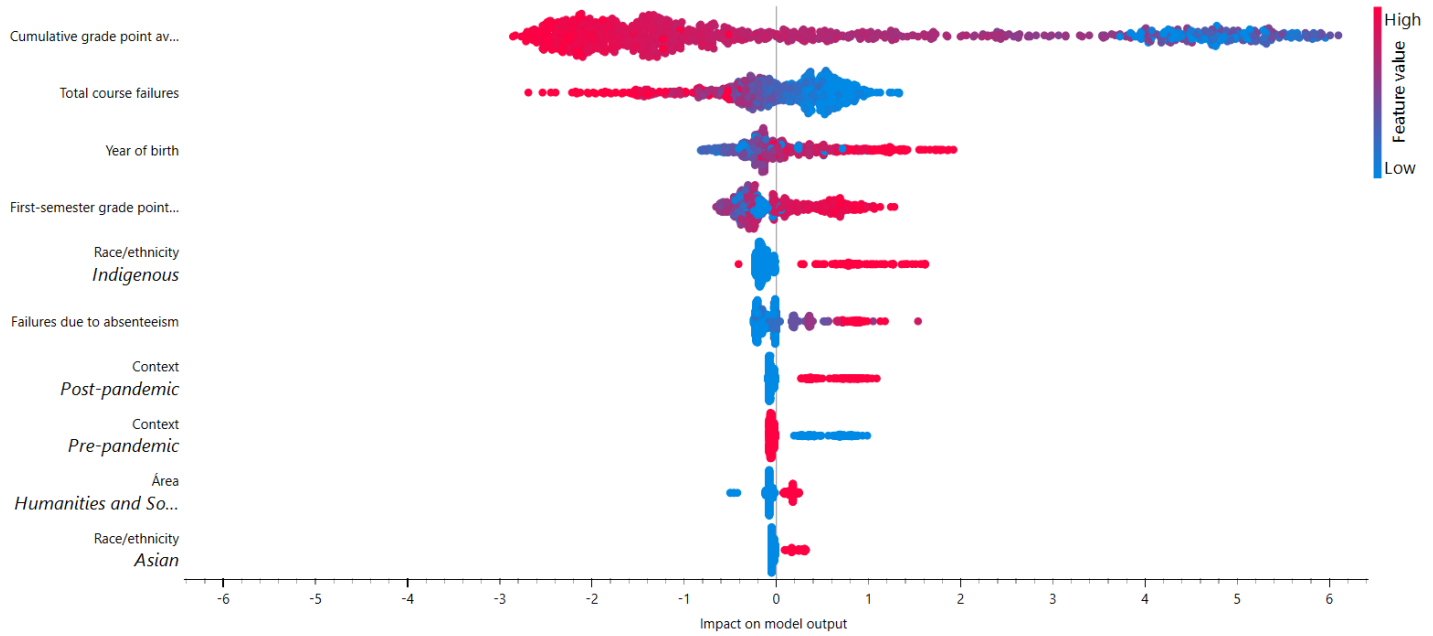
Finally, the Year of birth variable showed a moderate impact. More recent birth years (younger students) tended to increase the predicted probability of attrition, whereas older students showed negative contributions to the model. The variable Field of study – Humanities and Social Sciences exhibited a smaller effect, with a slightly positive contribution to attrition when students belonged to this field, although SHAP values remained concentrated near zero.

Overall, for Campus 2, behavioral variables (absence-related failures) and contextual variables (post-pandemic period) showed greater relative relevance when compared with admission-related variables, while cumulative academic performance remained the primary factor associated with attrition.

4.2.3. Results, Campus 3

In Campus 3, the ordering of variables according to the mean absolute impact of SHAP (Figure 3) values indicates that academic factors continue to play a central role in predicting attrition, although with relevant changes in the hierarchy and relative magnitude of effects when compared with the other campuses.

Figure 3: SHAP plot of the Gradient Boosting model for predicting attrition, Campus 3



The variable Cumulative grade point average (GPA) showed the highest predictive impact. A consistent inverse pattern is observed: low GPA values (blue) were associated with expressive positive SHAP values, substantially increasing the log-odds of attrition, whereas higher values (red) contributed negatively, reducing the predicted probability of the event. The magnitude of the effect suggests a high discriminative capacity of this variable in Campus 3.

The variable Total course failures ranked second in importance. Higher numbers of failures shifted SHAP values toward the positive region, increasing the predicted probability of attrition, whereas lower numbers showed a protective effect. This pattern is consistent with that observed in Campuses 1 and 2.

The variable Year of birth showed a moderate impact, greater than that observed in Campus 2. More recent birth years (younger students) were associated with positive contributions to the model, indicating a higher predicted probability of attrition in this group.

The variable First-semester grade point average showed an effect similar to cumulative GPA, although with smaller magnitude. Lower initial academic performance contributed positively to the estimated risk of attrition. The variable Race/ethnicity, Indigenous showed a positive contribution when present, indicating an increase in the predicted probability of attrition. A dispersion of SHAP values is observed, suggesting internal heterogeneity.

Failures due to absenteeism exerted a positive effect at higher levels, increasing the predicted risk, although with smaller magnitude compared with Campus 2. Regarding the temporal context, Context, Post-pandemic showed a positive contribution, indicating an increase in the predicted probability of attrition in the period following the pandemic. In contrast, Context, Pre-pandemic showed a predominantly negative impact, suggesting a lower estimated risk during this period.

The variable Field of study, Humanities and Social Sciences showed a discrete impact, with a slight positive contribution associated with attrition. Additionally, the variable Race/ethnicity, Asian showed a small magnitude effect, with contributions close to zero, indicating limited influence in the model.

4.3. Discussion of Results

4.3.1. Identification of the Best-Performing Algorithm

The present investigation aimed to identify the supervised algorithm with the best predictive performance and greatest statistical robustness for predicting academic attrition at a federal university located in the interior of Minas Gerais, Brazil, as well as to analyze which institutional and academic variables exert the greatest influence on this prediction. The results obtained allow for a discussion of both the methodological adequacy of machine learning models and the substantive determinants of attrition in light of the theoretical framework.

The findings indicate that Gradient Boosting showed superior overall performance across the three campuses analyzed, considering integrated metrics such as AUC, Accuracy, F1-score, and especially the Matthews Correlation Coefficient (MCC). Although, in some cases, the differences in AUC did not reach statistical significance when compared with models such as Logistic Regression and Neural Networks, the convergence across multiple metrics reinforces its practical superiority.

This result is consistent with studies that identify ensemble techniques based on boosting as particularly suitable for complex educational problems (Villar & Andrade, 2024; Martins et al., 2021; Niyogisubizo et al., 2022). As discussed by Burkart and Huber (2020), algorithms capable of capturing nonlinear relationships and high-order interactions tend to achieve better performance in multifactorial contexts such as higher education attrition. Attrition, as emphasized by Souza et al. (2024), results from the cumulative interaction among academic, institutional, and socioeconomic factors, which explains the advantage of methods capable of modeling complex dependency structures.

On the other hand, the absence of statistically significant differences among the main algorithms in certain campuses confirms the observation of Narula (2023) and Osemwegie et al. (2023), according

to whom there is no universal superiority of a single model, with performance being conditioned by the characteristics of the dataset, class balance, and the validation strategy. Thus, the choice of Gradient Boosting is based not only on isolated statistical significance but also on cross-campus stability and the robustness observed in metrics sensitive to class imbalance, such as the MCC.

From a methodological perspective, the adoption of stratified 10-fold cross-validation meets the recommendations of Jiang et al. (2020) regarding the need for rigorous evaluation of model generalization, reducing the risk of overfitting and increasing the reliability of the results.

4.3.2. Variables Influencing the Prediction of Attrition

The interpretability analysis based on SHAP showed that academic performance, particularly cumulative GPA and first-semester GPA, constitutes the main determinant of the predicted probability of attrition across the three campuses. This evidence is consistent with Aulck et al. (2016), who identified low performance in early courses as a strong predictor of attrition, and with Rosa et al. (2021), who emphasize the importance of academic achievement in high school and during the initial semesters of undergraduate education.

Additionally, Herbaut (2021) notes that course failures during the first year substantially increase the risk of attrition, which is consistent with the relevance of the variables "total course failures" and "failures due to absenteeism" in the estimated models. These results reinforce the perspective that early academic failure acts as a mechanism of progressive disengagement, as discussed by Casanova et al. (2018) and Esteves et al. (2021), who relate low attendance and weak institutional attachment to student withdrawal.

The centrality of academic performance can also be interpreted in light of academic integration theory, according to which persistence is associated with the student's ability to adapt to curricular demands and develop a sense of institutional belonging. Although socioeconomic variables were not modeled in depth in this study, academic performance may operate as a mediating variable of structural inequalities previously described by Herbaut (2021) and Santos et al. (2023).

The variable "year of birth" showed a consistent association with higher risk among younger students, a result partially consistent with Barreto et al. (2019), who identify differentiated patterns of attrition according to age group. This finding may reflect academic maturity, prior experience, or greater vocational clarity among older students.

The variable "Indigenous" showed a positive contribution to the predicted risk of attrition across the three campuses. This result is consistent with the literature on structural educational inequalities (Herbaut, 2021; Bayona-Oré, 2022), suggesting that factors of social vulnerability and unequal access to cultural capital may affect university persistence. However, the heterogeneity observed in SHAP values indicates that this variable does not operate uniformly, requiring complementary qualitative analyses.

In Campuses 2 and 3, the post-pandemic context showed a positive association with attrition, which may reflect institutional and socioeconomic effects resulting from the pandemic. Although the cited literature does not directly address the pandemic period, the findings may be interpreted in light of the discussions of Bonilla-Jurado et al. (2023) on insufficient institutional support and Neupane (2024) on economic vulnerabilities amplified in adverse contexts.

Despite structural heterogeneity among campuses, convergence is observed in the central role of academic performance indicators. The divergences are concentrated in the relative magnitude of behavioral and contextual variables, suggesting that although the phenomenon of attrition shares common determinants, its intensity and configuration depend on the organizational context. This finding reinforces the proposition of Nurmalitasari et al. (2023), according to whom attrition should be analyzed in an integrated and contextualized manner, considering interdependencies between individual and institutional factors.

From a theoretical perspective, the results corroborate multifactorial explanatory models of attrition and reinforce the centrality of academic integration as a structuring axis of the phenomenon. Furthermore, they indicate that internal institutional variables may have greater predictive power than admission variables, expanding the debate on institutional responsibility in student persistence.

Methodologically, the study demonstrates the usefulness of comparative approaches among supervised algorithms, combined with robust metrics such as MCC and AUC, as well as the use of interpretability techniques such as SHAP. This integration helps address recurring criticisms regarding the "black box" nature of machine learning models, as discussed by Burkart and Huber (2020).

From a practical perspective, the findings indicate that policies aimed at early monitoring of academic performance, particularly during the first semester, may constitute effective strategies for preventing attrition. Furthermore, the relevance of failures due to absenteeism suggests the need for actions focused on student engagement and attendance monitoring. The identification of potentially vulnerable groups, such as Indigenous students and younger entrants, reinforces the importance of differentiated and evidence-based institutional policies.

5. CONCLUSION

This study aimed to evaluate and compare the performance of different supervised machine learning algorithms in predicting academic attrition at a federal university located in the interior of Minas Gerais, Brazil, identifying the model with the greatest discriminative capacity and analyzing the main factors associated with attrition risk. The results indicate that the proposed objective was achieved, providing consistent empirical evidence at both the predictive and explanatory levels.

In summary, the findings show that ensemble-based algorithms, particularly Gradient Boosting, demonstrated superior performance or statistical equivalence compared with the most competitive models across the three campuses analyzed. Although pairwise comparisons did not reveal statistically significant differences in terms of AUC relative to Neural Networks and Logistic Regression, Gradient Boosting stood out for its greater consistency across metrics and for the robustness observed in indicators sensitive to class imbalance. These results reinforce the suitability of boosting methods for modeling complex and multifactorial educational phenomena such as academic attrition.

Regarding the factors associated with attrition risk, the results converge on the central role of cumulative academic performance, represented by GPA, and the history of course failures, both total failures and failures due to absenteeism. These variables emerge as the main determinants of predictive decisions across the different contexts

analyzed, supporting the interpretation of attrition as a cumulative process in which persistent academic difficulties precede institutional disengagement. Sociodemographic and institutional variables, such as age, race or ethnicity, field of study, class schedule, and pandemic context, showed moderate or localized impact, suggesting that their influence operates indirectly and is mediated by academic performance and engagement.

From a theoretical perspective, the study reinforces approaches that conceptualize higher education attrition as a multifactorial, dynamic, and interdependent phenomenon, in which individual, academic, and institutional factors interact throughout the university trajectory. The predominance of academic variables as central predictors supports models that emphasize academic integration and academic performance as fundamental axes of student persistence, while also acknowledging the contextual role of social and institutional inequalities.

Methodologically, the study advances by employing a comparative strategy among multiple algorithms, combining stratified cross-validation, paired statistical tests, and complementary performance metrics. The incorporation of interpretability techniques based on SHAP values contributes to understanding the factors underlying the models' predictions, increasing analytical transparency and enhancing the applicability of the results in the context of public management.

From a practical standpoint, the findings provide insights for university management, indicating that attrition prevention policies may benefit from the early monitoring of academic performance and from the identification of students with a history of course failures and absenteeism. The use of robust predictive models, such as Gradient Boosting, may support the implementation of institutional early warning systems, fostering data-driven strategies in public universities.

Despite its contributions, the study presents limitations related to the institutional scope and the absence of direct socioeconomic variables, which restricts the generalizability of the results. Future research may expand the analysis to other institutions, incorporate economic variables, and adopt longitudinal approaches, thereby deepening the understanding of the mechanisms that shape academic attrition. In summary, the results highlight the potential of supervised machine learning techniques to support the analysis and management of academic attrition in public higher education.

ACKNOWLEDGMENTS

The authors would like to thank the Federal University of Viçosa (UFV) for its institutional support and the Minas Gerais Research Support Foundation (FAPEMIG) for funding the undergraduate research scholarship linked to the Institutional Scientific Initiation Scholarship Program (PIBIC) at UFV, which supported the development of this research...

REFERENCES

1. Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, 903–912. https://doi.org/10.1145/2983323.2983351.
2. Angayarkanni, G., & Hemalatha, S. (2023, March). Evaluating the performance of supervised machine learning algorithms for predicting multiple diseases: A comparative study. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 1-7). IEEE. doi: 10.1109/ICACCS57279.2023.10113100
3. Asha, P., Vandana, E., Bhavana, E., & Shankar, K. R. (2020, June). Predicting university dropout through data analysis. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)* (pp. 852-856). IEEE. doi: 10.1109/ICOEI48184.2020.9142882.
4. Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360-3379. https://doi.org/10.1080/10494820.2021.1928235
5. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting student dropout in higher education. *arXiv*, *4*, 16–20. doi: https://doi.org/10.48550/arXiv.1606.06364
6. Balcan M, Nagarajan V, Vitercik E et al. Learning-theoretic foundations of algorithm configuration for combinatorial partitioning problems. In: *Proceedings of the 2017 Conference on Learning Theory*. PMLR, 2017, 213–74.
7. Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245–317.
8. Cardona, T., et al. (2023). Data mining and machine learning retention models in higher education. *Journal of College Student Retention: Research, Theory & Practice*, *25*(1), 51–75.
9. Caruana, R., & Niculescu-Mizil, A. (2004). Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,* 69-78. https://doi.org/10.1145/1014052.1014063
10. Devore, R., Hanin, B., & Petrova, G. (2021). Neural network approximation. *Acta Numerica*, *30*, 327–444. doi:10.1017/S0962492921000052
11. Doulah, S., & Islam, N. (2023). Performance evaluation of machine learning algorithm in various datasets. *Journal of Artificial Intelligence, Machine Learning and Neural Network*, *3*(2), 14–32. doi: https://doi.org/10.55529/jaimlnn.32.14.32
12. Fabris, F., Magalhães, J. P. de, & Freitas, A. A. (2017). A review of supervised machine learning applied to ageing research. *Biogerontology*, *18*, 171–188. doi: 10.1007/s10522-017-9683-y
13. Friedman, J. H. Aproximação de função gulosa: uma máquina de boosting de gradiente. *Annals of Statistic*. 29(5), 1189 – 1232. https://doi.org/10.1214/aos/1013203451
14. Freeman, J., & Simonsen, B. (2015). Examining the impact of policy and practice interventions on high school dropout and school completion rates: A systematic review of the literature. *Review of Educational Research*, *85*(2), 205–248. https://doi.org/10.3102/0034654314554431
15. Hemant, J. (2024). Predicting college dropout likelihood based on high school and college data: A machine learning approach. *MATTER International Journal of Science and*

Technology, *1*, 57–58. https://doi.org/10.20319/icstr.2024.5758
16. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2024). *Censo da Educação Superior*. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior.
17. Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: A brief primer. *Behavior Therapy*, *51*(5), 675–687. Doi: 10.1016/j.beth.2020.05.002
18. Jiménez-Gutiérrez, A. L., et al. (2024). Application of the performance of machine learning techniques as support in the prediction of school dropout. *Scientific Reports*, *14*(1), 3957. https://doi.org/10.1038/s41598-024-53576-1
19. Jung, J. S., Park, S. J., Kim, E. Y., Na, K. S., Kim, Y. J., & Kim, K. G. (2019). Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. *PLoS one*, *14*(6), e0217639. https://doi.org/10.1371/journal.pone.0217639
20. Karabacak E. S. & Yaslan, Y. Comparison of Machine Learning Methods for Early Detection of Student Dropouts, *2023 8th International Conference on Computer Science and Engineering (UBMK)*, Burdur, Turkiye, 2023, pp. 376-381, doi: 10.1109/UBMK59864.2023.10286747.
21. Kebede, M. M., Le Cornet, C., & Fortner, R. T. (2023). In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. *Research Synthesis Methods*, *14*(2), 156–172. doi: 10.1002/jrsm.1589.
22. Kerby, M. B. (2015). Toward a New Predictive Model of Student Retention in Higher Education: An Application of Classical Sociological Theory: An Application of Classical Sociological Theory. *Journal of College Student Retention: Research, Theory & Practice*, *17*(2), 138-161. https://doi.org/10.1177/1521025115578229
23. Kotsiantis, S.B. Decision trees: a recent overview. *Artif Intell Rev* **39**, 261–283 (2013). https://doi.org/10.1007/s10462-011-9272-4.
24. Kumar, S., et al. (2017). Assessment of various supervised learning algorithms using different performance metrics. *IOP Conference Series: Materials Science and Engineering*, *234*(4), 1–7. doi: 10.1088/1757-899X/263/4/042087.
25. Liu, C., et al. (2017). An efficient instance selection algorithm to reconstruct training set for support vector machine. *Knowledge-Based Systems*, *116*, 58–73. doi: https://doi.org/10.1016/j.knosys.2016.10.031
26. Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., & Realinho, V. (2021, March). Early prediction of student's performance in higher education: A case study. In *World Conference on Information Systems and Technologies* (pp. 166-175). Cham: Springer International Publishing. Doi: https://doi.org/10.1007/978-3-030-72657-7_16
27. Mohammad, S., Chowdhury, I. A., Roy, N., Hasan, M. N., & Nandi, D. (2023). Investigation of student dropout problem by using data mining technique. *International Journal of Education and Management Engineering*, *13*(5), 43. Doi: https://doi.org/10.5815/ijeme.2023.05.04
28. Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *Ieee Access*, *9*, 140731-140746. doi: 10.1109/ACCESS.2021.3119596
29. Narula, P. (2023). Analysis of common supervised learning algorithms through application. *Advanced Computational Intelligence: An International Journa*, *10*(1), 2. doi: 10.5121/acii.2023.10303.
30. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*, 100066. Doi: https://doi.org/10.1016/j.caeai.2022.100066
31. Núñez-Naranjo, A. F., Ayala-Chauvin, M., & Riba-Sanmartí, G. (2021). Prediction of university dropout using machine learning. *International Conference on Information Technology & Systems*, *1*, 396–406. Doi: 10.1007/978-3-030-68285-9_38
32. Orange Data Mining. (2024). *Download Orange Data Mining*. https://orangedatamining.com/download/.
33. Ortiz-Lozano, J. M., Rua-Vieites, A., Bilbao-Calabuig, P., & Casadesús-Fa, M. (2020). University student retention: Best time and data to identify undergraduate students at risk of dropout. *Innovations in Education and Teaching International*, *57*(1), 74–85. https://doi.org/10.1080/14703297.2018.1502090.
34. Osemwegie, E. E., & Amadin, F. I. (2023). Student dropout prediction using machine learning. *Fudma Journal of Sciences*, *7*(6), 347–353. https://doi.org/10.33003/fjs-2023-0706-2103
35. Patel, K. K., & Amin, K. (2024). Predictive modeling of dropout in MOOCs using machine learning techniques. *The Scientific Temper*, *15*(2), 2199–2206. https://doi.org/10.58414/SCIENTIFICTEMPER.2024.15.2.32
36. Pattanaphanchai, J., Leelertpanyakul, K., & Theppalak, N. (2019). The investigation of student dropout prediction model in Thai higher education using educational data mining. *Journal of University of Babylon for Pure and Applied Sciences*, *27*(1), 356–367. https://doi.org/10.29196/jubpas.v27i1.2191
37. Salditt, M., Humberg, S., & Nestler, S. (2023). Gradient Tree Boosting for Hierarchical Data. *Multivariate Behavioral Research*, *58*(5), 911–937. https://doi.org/10.1080/00273171.2022.2146638.
38. Sandoval-Palis, I., Naranjo, D., Vidal, J., & Gilar-Corbi, R. (2020). Early Dropout Prediction Model: A Case Study of University Leveling Course Students. *Sustainability*, *12*(22), 9314. https://doi.org/10.3390/su12229314.
39. Sarraju, V., Pal, J. & Kamilya, S. (2022). Performance analysis of supervised learning algorithms on different applications. *CS & IT Conference Proceedings*, *12*(19). doi: 10.5121/csit.2022.121903
40. Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3-29. https://doi.org/10.1177/1536867X20909688.
41. Segura, M., Mello, J., & Hernández, A. (2022). Machine Learning Prediction of University Student Dropout: Does

Preference Play a Key Role? *Mathematics*, *10*(18), 3359. https://doi.org/10.3390/math10183359

42. Sen, P.C., Hajra, M., Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: *Mandal, J., Bhattacharya, D. (eds) Emerging Technology in Modelling and Graphics*. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_11
43. Sulak, S. A., & Koklu, N. (2024). Predicting Student Dropout Using Machine Learning Algorithms. *Intelligent Methods In Engineering Sciences*, *3*(3), 91-98. https://doi.org/10.58190/imiens.2024.103
44. Suthaharan, S. (2016). Supervised learning algorithms. In: *Machine learning models and algorithms for big data classification*. Boston: Springer, 183–206. Doi: 10.1007/978-1-4899-7641-3
45. Utomo, A. P., Purwanto, P., & Surarso, B. (2023). Latest algorithms in machine and deep learning methods to predict retention rates and dropout in higher education: A literature review. *E3S Web of Conferences*, *448* 02034. https://doi.org/10.1051/e3sconf/202344802034