


Research

Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study

Alice Villar¹  · Carolina Robledo Velini de Andrade² 

Received: 27 June 2023 / Accepted: 14 September 2023

Published online: 04 January 2024

© The Author(s) 2024 

Abstract

Utilizing a dataset sourced from a higher education institution, this study aims to assess the efficacy of diverse machine learning algorithms in predicting student dropout and academic success. Our focus was on algorithms capable of effectively handling imbalanced data. To tackle class imbalance, we employed the SMOTE resampling technique. We applied a range of algorithms, including Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), as well as boosting algorithms such as Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), CatBoost (CB), and Light Gradient Boosting Machine (LB). To enhance the models' performance, we conducted hyperparameter tuning using Optuna. Additionally, we employed the Isolation Forest (IF) method to identify outliers or anomalies within the dataset. Notably, our findings indicate that boosting algorithms, particularly LightGBM and CatBoost with Optuna, outperformed traditional classification methods. Our study's generalizability to other contexts is constrained due to its reliance on a single dataset, with inherent limitations. Nevertheless, this research provides valuable insights into the effectiveness of various machine learning algorithms for predicting student dropout and academic success. By benchmarking these algorithms, our project offers guidance to both researchers and practitioners in their choice of suitable approaches for similar predictive tasks.

Keywords Dropout prediction · Supervised classification algorithms · Unsupervised classification algorithms · Class imbalance · Boosting algorithms · Hyperparameters · Learning analytics

1 Introduction

Predicting student behavior is a crucial task for educational institutions, as it can help to improve curriculum design and plan academic support interventions that are timely and personalized. As Chung and Lee [1] highlighted, “at-risk students” who drop out of school due to difficulties are more likely to adopt antisocial behaviors or face challenges in the labor market, making it harder for them to adapt to society.

Machine learning (ML) techniques, such as predictive modeling, have the potential to improve student retention by allowing educators to recognize students' weaknesses and provide learning metrics at any stage of educational progress [2]. The application of these techniques can be used to aid in the development of early warning systems, which can detect students who are at risk of dropping out in advance and offer the necessary support [3].

✉ Alice Villar, alsaldanhavillar@gmail.com; alsaldanha.gana@gmail.com; ✉ Carolina Robledo Velini de Andrade, rvacarolina01@gmail.com | ¹Faculty of Computer Science, University of Essex, Colchester, Essex, UK. ²Faculty of Applied Mathematics, University of Sao Paulo, Sao Paulo, Sao Paulo, Brazil.



In this comparative study, we explore various methods for predicting student success and dropout in higher education institutions, building upon the work of Martins et al. [4], “Early Prediction of Student’s Performance in Higher Education: a Case Study.” Martins et al. used machine learning classification models to predict students who might be at risk of failing to complete their degrees on time at the Polytechnic Institute of Portalegre (IPP) in Portugal.

While their study focused solely on supervised classification algorithms, we expand upon their work by incorporating unsupervised classification algorithms into our analysis. We apply resampling methods such as SMOTE and ADASYN to enhance the representation of the minority class in the dataset. Additionally, we employ various machine learning algorithms, including cost-sensitive learning algorithms, ensemble algorithms, and unsupervised anomaly detection algorithms. Furthermore, we leverage SMOTE and ADASYN with the unsupervised classification algorithm Isolation Forest to identify patterns and clusters within the data.

In this paper, we present a comparative study that utilizes the dataset collected by Martins et al. [4].

Specifically, we aim to address the following research questions:

- RQ1: how effective are different resampling techniques SMOTE and ADASYN in addressing class imbalance in the context of dropout prediction?
- RQ2: which machine learning algorithms are most effective in predicting student dropout in our dataset?
- RQ3: how do boosting algorithms (Gradient Boosting, Extreme Gradient Boosting, CatBoost, and LightGBM) compare to traditional machine learning algorithms in predicting student dropout?
- RQ4: what are the key factors that contribute to student success or failure, as revealed by the application of SHAP (SHapley Additive exPlanations)?

The remainder of this paper is organized as follows. In Sect. 2, we provide a condensed analysis of related work, focusing on the class imbalance problem in the context of dropout prediction. In this section, we also discuss the findings of Martins et al. [4], which serve as a foundation for our comparative study. In Sect. 2, we offer a concise review of related research, with a specific focus on addressing the class imbalance problem within dropout prediction. Additionally, we discuss the key findings of Martins et al. [4], which serve as the basis for our comparative study. Later in Sect. 3, we initiate our exploration with a bibliometric analysis, examining the knowledge structure surrounding the efficacy of boosting algorithms in forecasting student academic success. Subsequently, we transition into a systematic literature review, concentrating on the most highly cited papers in this domain. Section 4 describes the methodology used to build and evaluate the Machine Learning (ML) models, including the data description, methods used to handle the unbalanced dataset, and the procedures for training and evaluating the classification models. In Sect. 5, we present the experimental results of Supervised and Unsupervised Machine Learning models. In this section, we compare F1-scores of Supervised ML Algorithms before and after Hyperparameter Optimization with Optuna and apply the SHAP (SHapley Additive exPlanations) method to explain the output of the algorithms that performed best, specifically LightGBM and CatBoost with Optuna. Section 6 presents a comprehensive assessment of supervised machine learning algorithms’ performance in predicting student outcomes. Section 7 concludes the paper and suggests future research directions.

2 Analyzing class imbalance in dropout prediction and building on prior research

Section 2 provides a condensed analysis of the related work. The first subsection discusses the class imbalance problem and the application of unsupervised machine learning techniques in classification, particularly in the context of dropout prediction. The second subsection provides a summary of Martins et al. ’s [4] research, which serves as a foundation for our comparative study.

2.1 Addressing class imbalance in machine learning for classification

In educational data mining (EDM), class imbalance is a common issue, particularly when dealing with student retention data. This occurs because the number of students who drop out is significantly smaller than the number of students who stay in school or have good academic performance.

Class imbalance negatively impacts the accuracy of predictive models. Due to the scarcity of data points on the minority class, they tend to be biased towards the majority class, resulting in poor predictions for the minority class. [2, 5]. In such cases, standard classifier algorithms may not perform well, as they usually have a bias towards the majority class.

To address this issue, a number of techniques have been proposed in the literature. As noted by Islam et al. [6], in the past few decades, three popular approaches have been used to handle imbalanced data: data-driven, algorithm-based, and hybrid methods.

- Data-driven methods focus on balancing the distribution of classes in the training set by either oversampling the minority class or undersampling the majority class (increasing the representation of the minority class or decreasing the representation of the majority class). Oversampling techniques include Random Over-Sampling Examples (ROSE), Synthetic Minority Over-Sampling Technique (SMOTE), and Adaptive Synthetic (ADASYN), among others. Undersampling techniques include Random Under-Sampling Examples (RUSE) and Tomek Links (TL).
- Algorithm-based techniques, on the other hand, modify the learning algorithms to improve their performance on imbalanced data. Examples of such techniques include cost-sensitive learning, anomaly detection, and ensemble-based methods like bagging and boosting.
- Hybrid methods combine data-driven and algorithm-based techniques to improve the performance of machine learning models on imbalanced data. For example, hybrid methods may use data-driven techniques to balance the distribution of classes in the training set and then apply algorithm-based techniques to further improve the performance of the model.

Class imbalance is a common problem in predicting student performance, and various techniques have been proposed in the literature to address it. Rastrollo-Guerrero et al. [7] conducted a review of nearly 70 papers to identify the modern techniques commonly used for predicting student performance. They found that supervised learning, particularly the Support Vector Machine algorithm, was the most widely used and provided accurate results. Decision Tree (DT), Naïve Bayes (NB), and Random Forest (RF) were also well-studied algorithms that produced good results.

According to Rastrollo-Guerrero et al. [7], unsupervised learning is often considered an unattractive technique for researchers due to its low accuracy in predicting students' behavior in certain cases. However, the authors suggest that this can serve as an incentive for further research to improve these techniques and obtain more reliable results.

Our study aims to contribute to this area by exploring the effectiveness of unsupervised machine learning using Isolation Forest (IF) for predicting student dropout, while also addressing class imbalance in the data. Despite our efforts, our study found that IF did not perform well in predicting student dropout. These findings highlight the need for continued research into alternative machine learning algorithms and techniques that can effectively handle class imbalance and improve dropout prediction.

Recent research has demonstrated the effectiveness of boosting algorithms in predicting student dropout [8, 9]. These studies are in line with our findings, which demonstrate that boosting algorithms outperformed traditional classification methods.

2.2 Summary of Martins et al. [4] research

In this section, we briefly describe the purpose of their study, the methods they used, the datasets they worked with, and their findings. In addition, it shows the include limitations of their study and how our comparative study builds upon their work.

Martins et al. [4] aimed to develop a system that could identify students with potential difficulties in their academic path at an early stage, so that strategies to support the students could be put into place. Their research focused on building a system that generalizes to any course at IPP, rather than focusing on a specific field of study. The dataset included information from students enrolled in several courses from the four different schools belonging to IPP. Additionally, their paper only relied on information available at the time of enrollment and did not include any information on academic performance after enrollment. Another unique aspect of the paper was the use of a third intermediate class (relative success) in addition to the usual approach of restricting the model categories to failure/success. This allowed for different interventions for academic support and guidance for students who are at moderate risk versus those who are at high risk of being unsuccessful.

In their work, Martins et al. [4] highlighted the limitations of using accuracy as a performance metric for models trained on imbalanced datasets. This is because accuracy may give a false sense of good performance, as it tends to favor the

majority class and can ignore the performance of the minority classes. To address this issue, Martins et al. used single-class metrics to evaluate the performance of the model on each class separately.

Specifically, they used the F1 measure, which takes into account the balance between precision and recall, as the performance metric for the three classes in the dataset. By computing F1 scores for each class, Martins et al. were able to gain insights into the performance of the model for both the majority and minority classes. They were then able to use the average F1 score for the three classes as the metric for hyperparameter tuning, ensuring that the model is optimized to perform well on all classes.

Additionally, Martins et al. computed the accuracy of the optimized model as an overall metric, which provides a useful summary of the model's performance. By using both F1 scores and accuracy as performance metrics, Martins et al. were able to gain a more complete understanding of the performance of their model, especially for imbalanced datasets.

Martins et al. [4] aimed to improve classification performance using different algorithms. Initially, they applied four algorithms, namely DT, SVM, RL, and RF, to the dataset. Then, they applied the same four algorithms with hyperparameter tuning using grid search and compared the results. To deal with the class imbalance challenge, they applied data sampling techniques SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) to the dataset prior to model training. They found that SMOTE outperformed ADASYN. Thus, they applied SMOTE and used the same four algorithms with hyperparameter tuning using grid search and compared the results. Furthermore, Martins et al. [4] applied four boosting algorithms, namely Gradient Boosting, Extreme Gradient Boosting, CatBoost, and LogitBoost, using SMOTE. They then applied the same four boosting algorithms with hyperparameter tuning using randomized grid search and compared the results, using SMOTE.

In the study conducted by Martins et al. [4], the findings suggest that these boosting algorithms outperform standard methods when dealing with particular classification tasks. Among the four boosting algorithms assessed, Extreme Gradient Boosting emerged as the top classifier, although very similar to Gradient Boosting. However, despite the use of boosting algorithms, Martins et al. found that these models still struggled to accurately classify the minority classes in the imbalanced dataset. This is a common issue with imbalanced datasets, where the majority class dominates the dataset and makes it difficult for the model to learn from the minority class samples.

3 Systematic review and bibliometric analysis: boosting algorithms in student success prediction

In this section, we conducted a comprehensive bibliometric analysis using the Bibliometrix R-package. Table 1 presents the assessment criteria and Table 2 outlines our search strategy methodology. Our search yielded a total of 203 relevant articles.

Expanding on the knowledge acquired from the Most Globally Cited Documents among the 203 articles, we conducted a Systematic Literature Review (SLR) detailed in Sect. 3.2. Our focus was on the 15 most frequently cited papers in the field. These papers illuminate the present state of research regarding boosting algorithms, particularly Gradient Boosting (GB) and Extreme Gradient Boosting (XGBoost), in the context of student performance. These algorithms consistently exhibited superior performance, highlighting their potential as valuable tools for advancing educational research and enhancing student outcomes.

Table 1 Assessment criteria

Eligibility criteria	Reasoning
Period: published papers from 2017 to 2023	This study included publications from 2017 onwards to consider the most recent and up-to-date research in your field
Document type: article	Ensures that we are primarily evaluating peer-reviewed documents
Language: English	The search strategy was restricted to English because abstracts are available in this language
Chosen search locations within the text: titles and abstracts	Effectively targets the most relevant information in each paper
Database source: dimensions	Dimensions is a comprehensive academic database that provides access to a diverse range of scholarly publications, making it a suitable platform for conducting systematic literature reviews

Table 2 Search strategy methodology

Topic	Search Query	Reasoning
Boosting algorithms for student performance	("boosting algorithms" OR "gradient boosting" OR "xgboost" OR "catboost" OR "lightgbm") AND ("academic success" OR "student")	This search is designed to find relevant literature on the performance of boosting algorithms on academic success and student outcomes

3.1 Overview of the literature landscape

This subsection presents two informative graphs: Annual Scientific Production (Fig. 1) and the Most Globally Cited Documents (Fig. 2), both based on our dataset of 203 documents. Additionally, Table 3 provides key statistics, revealing a notable annual growth rate of 78.18% and an average document age of just 1.42 years, emphasizing the timeliness of research in this domain. Notably, each document in our dataset received an average of 6.241 citations, underscoring the significance and influence of the literature.

3.2 Systematic literature review: leveraging boosting algorithms for precision in predicting student performance

This section presents our Systematic Literature Review (SLR) based on the 15 most frequently cited papers in the field (Table 4). This table provides additional details for each document in the ranked list of our bibliometric research (Fig. 2), including the full title, the total number of citations per year (TC per year), and the normalized citation impact scores (normalized TC).

In this section, we present the findings of our Systematic Literature Review (SLR) focused on the 15 most frequently cited papers in the field, which are presented in Table 4. This table provides comprehensive details for each document in our bibliometric research (see Fig. 2). The information includes the full title, the total number of citations per year (TC per year), and the normalized citation impact scores (normalized TC).

In each of the following studies, boosting algorithms, particularly Gradient Boosting (GB) and Extreme Gradient Boosting (XGBoost), consistently showcased superior performance, underscoring their potential as valuable tools for enhancing educational research and improving student outcomes.

1. Predicting student satisfaction with MOOCs [10]

In their study, Hew et al. aimed to measure MOOC success based on student satisfaction, rather than conventional completion rates. Employing machine learning and sentiment analysis, the study examined 249 MOOCs and student perceptions. Notably, their research revealed that the gradient boosting trees model outperformed other candidate models, demonstrating acceptable values for accuracy, precision, recall, F1 score, and Cohen’s kappa, leading to its selection for the study.

2. Predicting academic performance in public schools [11]

Fernandes et al. developed classification models using the Gradient Boost Machine (GBM) algorithm to predict student performance and compared their predictive capabilities at different points in the academic cycle. They began with a descriptive statistical analysis of datasets from the State Department of Education in the Federal District of Brazil, creating two datasets—one with attributes collected before the school year commenced and another with additional variables like ‘absences,’ ‘grades,’ and ‘school subjects’ gathered two months into the school year. The choice of GBM classification models stemmed from its ability to combine multiple decision trees into a potent predictive model. Notably, the accuracy of these models was significantly enhanced through the boosting technique, which played a pivotal role in leveraging data and decision trees to offer valuable insights into student academic outcomes.

3. Identifying behavioral phenotypes of loneliness in college students [12]

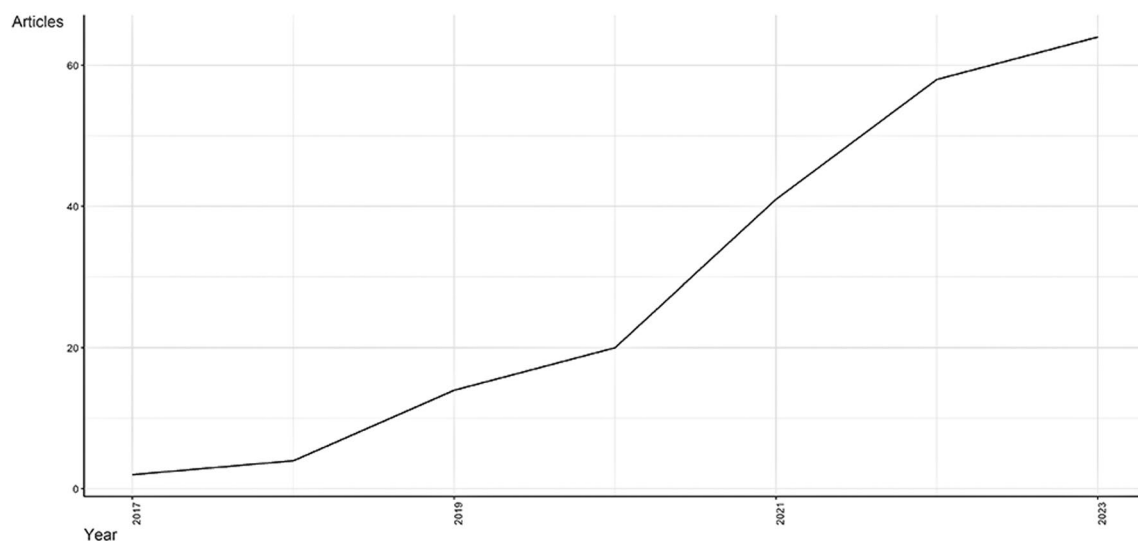


Fig. 1 Annual scientific production

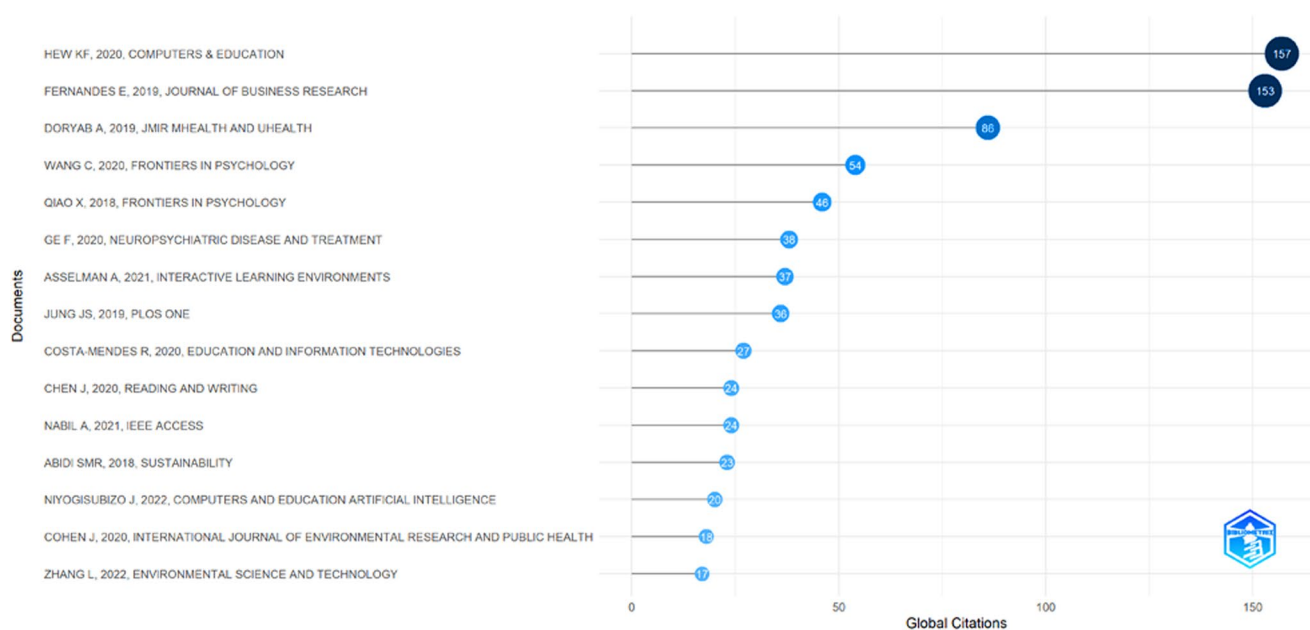


Fig. 2 Most globally cited documents

Doryab et al. employed various analytic methods to understand loneliness in college students. They used machine learning classification to infer the level of loneliness and changes in loneliness levels, employing an ensemble of gradient boosting and logistic regression algorithms with feature selection. Notably, the boosting algorithm, gradient boosting, was chosen for its ability to perform well on noisy datasets and learn complex nonlinear decision boundaries. In their research, gradient boosting demonstrated superior performance in detecting outcomes related to loneliness, which highlights its effectiveness in analyzing and predicting behavioral phenotypes associated with loneliness and social isolation among college students.

Table 3 Main information about data

Main information about data	Results
Timespan	2017:2023
Sources	143
Documents	203
Annual growth rate %	7818
Document average age	142
Average citations per doc	6241
Authors	743
Authors of single-authored docs	15
Single-authored docs	16
International co-authorships %	1576

4. Anxiety levels during COVID-19 in Chinese University students [13]

Wang et al. explored anxiety severity among Chinese non-graduating undergraduate students during the COVID-19 pandemic using machine learning techniques. The study applied machine learning techniques, specifically Extreme Gradient Boosting (XGBoost), to predict both the anxiety level and changes in anxiety levels four weeks later, using initial SAS (a sociodemographic questionnaire and the Self-Rating Anxiety Scale) scores from students' first test. Remarkably, the XGBoost model demonstrated superior prediction accuracy when compared to the traditional multiple stepwise regression model in evaluating the anxiety status of university students.

5. Educational assessment process data analysis [14]

Qiao and Jiao addressed the application of data mining methods to analyze educational assessment process data. Among the supervised techniques used, the results indicated that gradient boosting outperformed the other methods, achieving the highest classification accuracy with a Kappa value of 0.94 and an overall accuracy of 0.96. Overall, all four methods demonstrated satisfactory classification accuracy, but the boosting algorithm, gradient boosting, exhibited superior performance in accurately analyzing the educational assessment process data, highlighting its efficacy in this context.

6. Predicting psychological state among Chinese undergraduate students [15]

Ge et al. aimed to develop predictive models for identifying adolescents at high risk of suicide during the COVID-19 using machine learning techniques. The gradient boosting algorithm demonstrated one of the highest accuracy rates among the models. Their study included 2009 students, with a response rate of 80.36%. The Extreme Gradient Boosting (XGBoost) algorithm demonstrated remarkable performance by accurately predicting 97.3% of students with probable anxiety and 96.2% of students with probable insomnia symptoms.

7. Enhancing student performance prediction with ensemble methods [16]

Asselman et al. aimed to improve the accuracy of predicting student performance by harnessing Ensemble Learning methods in the context of Knowledge Tracing (KT). They introduced a new Performance Factors Analysis (PFA) approach utilizing various models, including Random Forest, AdaBoost, and XGBoost. These models were evaluated across three distinct datasets, and the experimental results revealed that the scalable XGBoost algorithm outperformed the other models, leading to a substantial enhancement in performance prediction compared to the original PFA algorithm.

8. Predicting suicidal risk in Korean adolescents [17]

Jung et al. sought to develop predictive models for identifying adolescents at high risk of suicide using machine learning methods. They employed five different machine learning techniques, including logistic regression (LR),

Table 4 List of the 15 most cited articles

Author	Title	Total citations	TC per year	Normalized TC
1. Hew et al. [10]	What predicts student satisfaction with MOOCs: a gradient boosting trees supervised machine learning and sentiment analysis approach	157	3925	795
2. Fernandes et al. [11]	Educational data mining: predictive analysis of academic performance of public school students in the capital of Brazil	153	3060	687
3. Doryab et al. [12]	Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: STATISTICAL analysis, data mining and machine learning of smartphone and fitbit data	86	1720	386
4. Wang et al. [13]	Chinese college students have higher anxiety in new semester of online learning during COVID-19: a machine learning approach	54	1350	273
5. Qiao & Jiao. [14]	Data mining techniques in analyzing process data: a didactic	46	767	267
6. Ge et al. [15]	Predicting psychological state among Chinese undergraduate students in the COVID-19 epidemic: a longitudinal study using a machine learning	38	950	192
7. Asselman et al. [16]	Enhancing the prediction of student performance based on the machine learning XGBoost algorithm	37	1233	556
8. Jung et al. [17]	Prediction models for high risk of suicide in Korean adolescents using machine learning techniques	36	720	162
9. Costa-Mendes [18]	A machine learning approximation of the 2015 Portuguese high school student grades: a hybrid approach	27	675	137
10. Chen et al. [19]	Synergistic effects of instruction and affect factors on high-and low-ability disparities in elementary students' reading literacy	24	600	122
11. Nabil et al. [20]	Prediction of students' academic performance based on courses' grades using deep neural networks	24	800	360
12. Abidi et al. [21]	Prediction of confusion attempting algebra homework in an intelligent tutoring system through machine learning techniques for educational sustainable development	23	383	133
13. Niyogisubizo et al. [22]	Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization	20	1000	678
14. Cohen et al. [25]	A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions	18	450	091
15. Zhang et al. [23]	Haze air pollution health impacts of breath-borne VOCs	17	850	577

random forest (RF), support vector machine (SVM), artificial neural network (ANN), and extreme gradient boosting (XGB). Through their analysis, they found that factors such as sadness, violence, substance use, and stress were associated with an increased risk of suicidal ideation/attempt among adolescents. Using 26 predictor variables, the machine learning models demonstrated comparable accuracy to logistic regression, with XGB exhibiting the highest accuracy at 79.0%, followed by SVM at 78.7%, LR at 77.9%, RF at 77.8%, and ANN at 77.5%.

9. Predicting high school academic scores [18]

Costa-Mendes et al. aimed to predict high school academic scores using machine learning techniques. They employed several algorithms, including Random Forest (RF), support vector machine (SVM), artificial neural network (ANN), and extreme gradient boosting (XGB), stacking them as an ensemble. The results of their study demonstrated that all machine learning algorithms outperformed the classical multiple linear regression model in terms of predictive accuracy for high school academic scores. Specifically, XGB showed superior performance to RF, particularly when specific conditions related to the determinant of the relationship between matrices (Det [R]) were met.

10. Disparities in elementary students' reading literacy [19]

Chen et al. examined disparities in reading literacy among elementary students and identify the key factors contributing to these disparities. They employed four machine learning algorithms: logistic regression, support vector machine (SVM), decision tree (DT), and extreme gradient boosting (XGBoost). These algorithms were utilized to classify and predict high- and low-proficiency readers while pinpointing the most influential factors in differentiating between high and low achievers. The study found that, compared to logistic regression and DT, both linear SVM and XGBoost exhibited outstanding and robust classification performances. Consequently, linear SVM and XGBoost were selected to identify the most important features or factors responsible for distinguishing high-achieving students from their low-achieving counterparts.

11. Early-stage prediction of student academic performance [20]

Nabil et al. addressed the challenge of early-stage prediction of student academic performance, with a focus on courses like "Programming" and "Data Structures" known for high failure and dropout rates. The study utilized various predictive models, including a Deep Neural Network (DNN), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), Support Vector Classifier, and K-Nearest Neighbor. They also tackled the issue of imbalanced datasets, employing resampling methods like SMOTE, ADASYN, ROS, and SMOTE-ENN to mitigate this challenge. Results from the experiments highlighted the effectiveness of the proposed DNN model in predicting students' performance in data structure courses and identifying students at risk of failure early in the semester, achieving an accuracy of 89%. This accuracy surpassed that of other models like decision tree, logistic regression, support vector classifier, and K-nearest neighbor. Notably, Gradient Boosting (GB) also exhibited strong performance, achieving one of the best accuracies at 91%, highlighting its effectiveness in addressing the imbalanced dataset issue and enhancing prediction results.

12. Identifying struggling students [21]

Abidi et al. used machine learning techniques to identify students who struggled with mastering algebra skills assigned as homework within an Intelligent Tutoring System (ITS). Gradient boosting, along with other models, exhibited high accuracy in classifying students attempting algebra homework. The study focused on ASSISTments, an ITS, and employed various machine learning techniques to analyze skill-builder data. Seven candidate models were utilized, including Naïve Bayes (NB), Generalized Linear Model (GLM), Logistic Regression (LR), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), and Gradient Boosted Trees (XGBoost). The boosting algorithm, XGBoost, along with RF, GLM, and DL, exhibited high accuracy in classifying confused students attempting algebra homework within the ITS. The study emphasized the importance of leveraging machine learning methods to identify and support students who struggle with specific skills, ultimately contributing to their knowledge and promoting their role in sustainable development, with XGBoost showing one of the highest accuracies at 84.9%.

13. Predicting student dropout [22]

Niyogisubizo et al. proposed a novel stacked ensemble method for predicting student dropout, a pressing research topic. The authors proposed a novel stacked ensemble method, combining Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to achieve this prediction. Employing a dataset spanning from 2016 to 2020 from Constantine the Philosopher University in Nitra, their approach outperformed base models, as demonstrated by testing accuracy and the area under the curve (AUC) evaluation metrics. Particularly, XGBoost exhibited an AUC score of 0.978, emphasizing its strong performance in identifying students at risk of dropping out of university classes.

14. Predicting suicide risk based on language in therapy sessions [25]

Cohen et al. addressed the pressing issue of rising adolescent suicide rates by exploring the feasibility of using machine learning to identify suicide risk based on language samples from therapy sessions. This study involved the testing of natural language processing machine learning models within outpatient therapy sessions for adolescents. Data collection encompassed language samples, standardized depression and suicidality scale scores, and therapists' impressions of the client's mental state. Among the Machine Learning (ML) models tested, Extreme Gradient Boosting (XGBoost) demonstrated the best performance with an area under the curve (AUC) of 0.78 during external validation.

15. Monitoring health effects during haze pollution [23]

Zhang et al. investigated the potential of using breath-borne volatile organic compounds (VOCs) as biomarkers for monitoring health effects related to air pollution, specifically during haze pollution episodes. The study involved 47 healthy college students, and their exhaled breath samples were collected and analyzed for VOCs before, during, and after two separate haze pollution episodes using gas chromatography-ion mobility spectrometry (GC-IMS). Machine learning, specifically the Gradient Boosting Machine (GBM) model, was employed to differentiate between different exposure periods. The GBM model demonstrated excellent performance in distinguishing pre- and on-exposure to haze pollution, achieving precision rates of 90–100% for both haze episodes.

4 Methodology used to build and evaluate the machine learning models

This methodology section starts with an overview of the methodology used for building and evaluating unsupervised machine learning models, followed by detailed subsections that explain the data description and preprocessing, class imbalance treatment, model training and evaluation procedures, and experimental design. Finally, the section ends with a presentation and discussion of the experimental results.

4.1 Overview

The methodology employed in this study involved several sequential steps and algorithms to address the research objectives. An overview of our methodology is presented in Fig. 3, outlining the key steps undertaken.

The first step involved applying the SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) methods to resample the dataset. These techniques are commonly used for handling imbalanced datasets. To compare their performance, a decision tree classifier was employed, and the results were evaluated. Based on this comparison, the best-performing resampling technique was selected for further analysis.

Next, the preprocessed dataset was subjected to classification using three algorithms: Decision Tree (DT), Support Vector Machines (SVM), and Random Forest (RF). These algorithms were chosen for their effectiveness in classification tasks. The performance of each algorithm was evaluated to identify any significant differences.

To enhance the classification performance, boosting algorithms were employed, namely Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), CatBoost (CB), and LightGBM (LB). These algorithms are known for their ability to improve predictive power. By incorporating these boosting techniques, the study aimed to optimize the classification results obtained from the initial algorithms.

Furthermore, the study employed the Isolation Forests algorithm, an unsupervised anomaly detection technique, to identify outliers or anomalies within the dataset. This step focused on identifying instances that deviated significantly from the majority of the dataset.

To optimize the performance of all the algorithms used in this study, hyperparameter tuning was conducted using Optuna. Optuna is a hyperparameter optimization framework that assists in finding the best combination of hyperparameters for machine learning models. This step aimed to fine-tune the algorithms' performance by optimizing their respective hyperparameters. To ensure the reliability of our findings, we integrated Optuna with the cross-validation function, `cross_val_score()`. This approach not only enabled us to achieve hyperparameter optimization but also allowed for a rigorous assessment of the models' performance.

Additionally, the SHAP (SHapley Additive exPlanations) framework was applied to gain insights into the contribution of each feature in predicting the model's output. Specifically, SHAP was employed with the CATBoost (CB) and LightGBM (LB) algorithms to analyze the importance of individual features in the classification task. This analysis helped to enhance the understanding of how each feature influenced the model's predictions and identify key factors driving the classification results.

In the final phase of our study, we presented and thoroughly analyzed the experimental results obtained through various methods and algorithms. To evaluate the discriminative capabilities of our top-performing models, LightGBM (LB) and CatBoost (CB), we utilized ROC (Receiver Operating Characteristic) and AUC (Area Under the Curve) curves. Additionally, we conducted further assessments to explore potential significant differences in the results achieved by the LightGBM and CatBoost algorithms. To accomplish this, we employed the Statistical Tests for Algorithms Comparison (STAC) platform.

Table 5 compares the analysis methods we use in the present paper and the methodology used by Martins et al. [4], which was presented in Sect. 2.2.

4.2 Data description

In this paper, we present a comparative study with Martins et al. [4], utilizing the same dataset that they collected. The dataset is in a CSV file format and contains information about 4424 students, with 35 columns. The dataset covers a range of factors that may influence a student's academic success, including demographic, socio-economic, and academic factors. The data contains variables related to demographic factors (age at enrollment, gender, marital status, nationality, address code, special needs), socio-economic factors (whether the student works or not, the educational and employment background of their parents, whether they received a student grant or have student debt) and student's academic path (admission grade, retention years at high school, their admission grade, how many years they spent in high school, the order of choice for their enrolled course, and the type, of course, they took in high school).

Martins et al. [4] classified each record into three categories based on how long it took the student to obtain their degree: Success, Relative Success, and Failure. *Success* means the student obtained their degree in due time, *Relative Success* means they took up to three extra years to obtain their degree, and *Failure* means they took more than three extra years to obtain their degree or did not obtain it at all. These categories correspond to different levels of risk, with low-risk students having a high probability of succeeding, medium-risk students who might benefit from institution measures to succeed, and high-risk students having a high probability of failing.

The imbalanced distribution of the records among the three categories presented challenges for developing accurate predictive models that could effectively identify high-risk students who might benefit from interventions. As a result, predictive models that are trained on such imbalanced datasets tend to be biased towards the majority class and may have lower accuracy in predicting the minority classes. As shown in Fig. 4, in the distribution, "Failure" accounts for 32.15% of total records, and "Relative Success" accounts for 17.96% of total records, while the majority class, "Success", accounts for 49.89% of the records.

4.3 Resampling methods to address class imbalance

Class imbalance is a common challenge in classification tasks, such as predicting student dropout rates. To address this issue, we explore different resampling methods to improve the accuracy of our models in this study. Specifically, we compare the performance of two commonly used methods, SMOTE and ADASYN, to address the class imbalance

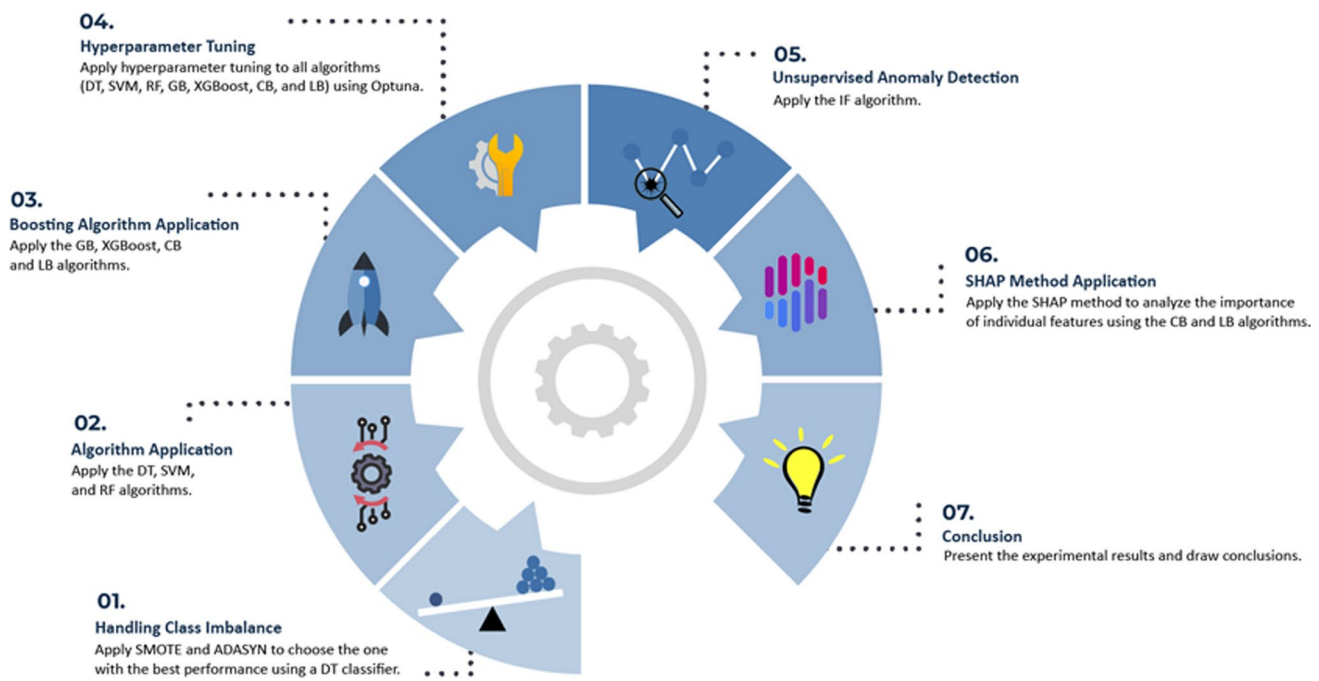


Fig. 3 Methodology to build and evaluate the machine learning models

in our dataset. We use decision trees to compare their performance, and it is worth noting that we do not use Logistic Regression due to the presence of multicollinearity in our data, as explained in Sect. 4.4.

Our findings indicate that both SMOTE and Adasyn resampling methods effectively improve the accuracy of our models. However, SMOTE slightly outperforms Adasyn in addressing the class imbalance issue, as shown in Table 6. It is important to note that the dataset we used in this study had no null and no missing data. This means that all the data points in the dataset were complete and did not require any imputation or handling of missing values.

4.4 Navigating challenges of data preprocessing

Martins et al. [4] employed a rigorous data preprocessing method to handle anomalies and unexplainable outliers, resulting in a final dataset consisting of 3623 records and 25 independent variables. In contrast, we decided not to remove anomalies in the preprocessing stage for several reasons, such as the potential impact on the representativeness of the dataset, the potential loss of valuable information, and the potential difficulty of distinguishing between

Table 5 Comparison of Methodologies in Martins [4] and our study

Aspect of methodology	Martins [4]	Our Study
Traditional algorithms used	DT, SVM, RL, RF	DT, SVM, RF
Hyperparameter tuning	Applied grid search	Applied Optuna
Sampling methods	SMOTE, ADASYN	SMOTE, ADASYN
Comparison of sampling methods	Compared SMOTE with ADASYN using logistic regression	Compared SMOTE with ADASYN using decision trees
Boosting algorithms	GB, XGBoost, CB, and LB	GB, XGBoost, CB, and LB
Anomaly detection	Not mentioned	Isolation forests
Feature importance analysis	Not mentioned	SHAP—CatBoost and LightGBM
Multicollinearity analysis	Not mentioned	VIF method for detecting Multicollinearity
ROC and AUC analysis	Not mentioned	ROC and AUC curves—CatBoost and LightGBM
Statistical tests	Not mentioned	Used STAC platform and employed paired samples t-test

true anomalies and rare but meaningful data points. As such, we have decided that the benefits of retaining anomalies outweighed the potential drawbacks.

Our preprocessing analysis revealed that there was significant Multicollinearity in our dataset. This means that predictor variables were highly correlated with each other, making it difficult to distinguish the individual effects of each variable on the outcome. Specifically, we employed the Variance Inflation Factor (VIF) method, which is a widely-used approach for detecting Multicollinearity in regression analysis. The VIF method calculates the degree of correlation between predictor variables, and if the correlation is high, it indicates that Multicollinearity is present in the data.

A VIF of 1 indicates that there is no multicollinearity, while a VIF of greater than 1 indicates that multicollinearity may be present. A VIF of 5 or greater is often considered high. Table 7 shows the variables with a VIF higher than 10. Table 8 shows the variables with a VIF between 5 and 10.

When Multicollinearity is present in a dataset, the regression coefficients can become unreliable or even meaningless, ultimately affecting the model's accuracy and validity. To mitigate this issue, we opted not to use Logistic Regression, which is not recommended in the presence of Multicollinearity and can lead to biased results. Instead, we used a range of machine learning algorithms that are better suited for handling Multicollinearity, such as decision trees, random forests, and gradient-boosting models.

5 Supervised ML algorithm comparison and Optuna hyperparameter optimization

This section presents the results obtained by applying three machine learning algorithms, namely Decision Trees, Support Vector Machines, and Random Forest, to predict student dropout and failure in our case study. Additionally, the section includes the results of Hyperparameter Optimization using Optuna.

5.1 Supervised algorithms

This section of the paper will present the F1-score results for seven different supervised machine learning algorithms: Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, Extreme Gradient Boosting, CatBoost, and LightGBM. The F1-score metric will be used to evaluate the performance of each algorithm. Additionally, the Optuna framework will be applied to optimize the hyperparameters of each model, and the F1-score after the application of Optuna will also be included in the table. The tables will provide a clear comparison of the performance of each algorithm, both before and after hyperparameter optimization.

The F1-score is a suitable evaluation metric for imbalanced datasets, as it balances precision and recall. In this study, the F1-score was chosen to assess the model's accuracy, considering both false positives and false negatives. By combining precision and recall, the F1-score provides a comprehensive measure of the model's performance. Higher F1 scores indicate better classification performance for a specific class.

Fig. 4 Data distribution

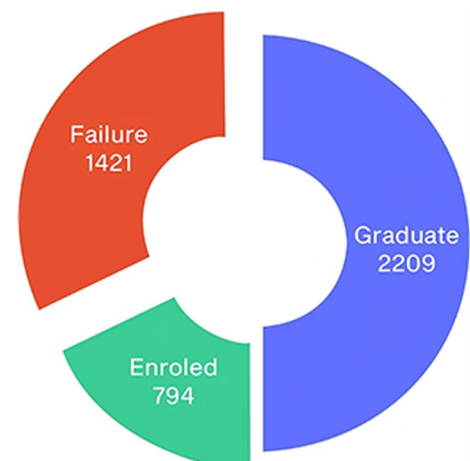


Table 6 SMOTE and ADASYN: performance comparison

Target	Description	Smote F1-score	Adasyn F1-score
0	Graduate	0.75	0.74
1	Dropout	0.78	0.73
2	Enrolled	0.69	0.64

5.1.1 Traditional algorithms

This section provides an overview of the results obtained through the application of traditional algorithms, specifically Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), to predict student dropout and failure within our case study. Additionally, we present the outcomes of our Hyperparameter Optimization using Optuna. DT, RF, and SVM are machine learning algorithms used for various classification and regression tasks, but they differ in their underlying principles and how they make decisions.

a) Decision trees

Decision Trees are a supervised learning algorithm that partitions the data into subsets based on the values of different input features. It makes decisions by following a tree-like structure of binary decisions. Table 9 shows the F1 scores obtained from the Decision Tree (DT) model applied to three different classes: Graduate, Dropout, and Enrolled. The results reveal that the DT model achieved an F1 score of 0.75 for the Graduate class, 0.78 for the Dropout class, and 0.69 for the Enrolled class. These scores represent the RF model's ability to accurately classify instances from each respective class.

To further optimize the Decision Tree model's performance, the Optuna framework was applied. The "f1_weighted" metric was used to evaluate the hyperparameters, which takes into account the F1 scores for each class, considering the number of samples in each class and assigning higher weight to the classes with more samples. The hyperparameter optimization process was run for 20 trials. Comparing the F1 scores obtained from the DT model and the DT model with

Table 7 VIF score higher than 10

Variables	VIF
Mother's occupation	110,643,811
Age at enrollment	199,640,064
Curricular units 1st sem (credited)	1,639,464,884
Curricular units 1st sem (enrolled)	1,755,537,684
Curricular units 1st sem (evaluations)	2,009,324,388
Curricular units 1st sem (approved)	3,428,874,105
Curricular units 1st sem (grade)	2,747,421,131
Curricular units 2nd sem (credited)	1,310,292,464
Curricular units 2nd sem (enrolled)	159,080,573
Curricular units 2nd sem (evaluations)	1,802,847,137
Curricular units 2nd sem (approved)	2,686,446,034
Curricular units 2nd sem (grade)	2,506,334,656
Unemployment rate	2,021,456,508

Table 8 VIF score between 5 and 10

Variables	VIF
Marital status	7,414,206,571
Course	7,977,765,035
Daytime/evening attendance	9,563,568,198
Nacionality	7,194,048,942
Father's occupation	8,187,867,963
Tuition fees up to date	7,310,689,245

Optuna, it is observed that the F1 score for the three classes has increased slightly. This suggests that the hyperparameter optimization process was successful in improving the model's performance, especially for the Dropout class.

b) Random forest

Random Forest is an ensemble learning technique designed to enhance accuracy and mitigate overfitting by combining multiple decision trees. It achieves this by constructing numerous decision trees, each trained on a subset of the data, and then consolidating their predictions. This method, rooted in decision trees, finds utility in classification and regression tasks, distinguished for its robustness and adeptness in managing high-dimensional data.

Table 10 shows the F1 scores obtained from the Random Forest (RF) model applied to three different classes: Graduate, Dropout, and Enrolled. The results show that the Random Forest model achieved an F1 score of 0.84 for the Graduate class, 0.87 for the Dropout class, and 0.81 for the Enrolled class. These scores represent the RF model's ability to accurately classify instances from each respective class. To further optimize the RF model's performance, the Optuna framework was applied, using the "f1_weighted" metric to evaluate the hyperparameters. The hyperparameter optimization process was run for 20 trials. Comparing the F1 scores obtained from the RF model and the RF model with Optuna, it is observed that the F1 score remained practically the same. This suggests that the hyperparameter optimization process did not have a significant impact on the model's performance.

c) Support vector machines

Support Vector Machine (SVM) seeks the optimal hyperplane to separate data into distinct classes by maximizing the margin, which represents the distance between the hyperplane and the nearest data points of each class. This process entails mathematical optimization, where SVM strategically chooses support vectors, the data points closest to the hyperplane, to define the decision boundary. Table 11 presents the F1 scores obtained from the Support Vector Machines (SVM) model applied to three different classes: Graduate, Dropout, and Enrolled. Additionally, the table includes the results after applying the Optuna framework to optimize the SVM model's hyperparameters.

The initial SVM model achieved an F1 score of 0.79 for the Graduate class, 0.76 for the Dropout class, and 0.68 for the Enrolled class. These scores represent the SVM model's ability to accurately classify instances from each respective class. To further improve the SVM model's performance, the Optuna framework was employed to optimize its hyperparameters. Optuna conducts a search for the best combination of hyperparameters to enhance the model's performance. The results reflect the F1 scores after applying Optuna to the SVM model. However, the F1 scores for the SVM model with Optuna remain the same as the initial SVM model. The F1 score for the Graduate class is 0.79, 0.76 for the Dropout class, and 0.68 for the Enrolled class. This indicates that the hyperparameter optimization process did not have a substantial impact on the model's performance in this case.

5.1.2 Boosting algorithms

This section presents the results obtained by applying four boosting algorithms, namely Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), CatBoost (CB), and LightGBM (LB) to predict student dropout and failure in our case study. Additionally, the section includes the results of Hyperparameter Optimization using Optuna. These four algorithms belong to the family of gradient boosting algorithms. They share the common purpose of improving predictive performance by combining weak learners through an ensemble approach, and they have each introduced unique optimizations to enhance their performance and usability. Additionally, each of these algorithms has introduced distinct optimizations to enhance their performance and usability.

Table 9 F1-Scores for Decision Tree with and without Optuna

Target	Description	Decision tree	Decision tree + Optuna
0	Graduate	0.75	0.80
1	Dropout	0.78	0.81
2	Enrolled	0.69	0.71

Table 10 F1-Scores for Random Forest with and without Optuna

Target	Description	Random forest	Random forest + Optuna
0	Graduate	0.84	0.84
1	Dropout	0.87	0.87
2	Enrolled	0.81	0.80

Table 11 F1-Scores for support vector machines with and without Optuna

Target	Description	Support vector machines	Random-Forest + Optuna
0	Gtaduate	0.79	0.79
1	Dropout	0.76	0.76
2	Enrolreel	0.68	0.68

a) Gradient boosting (GB)

Gradient Boosting (GB) serves as the foundational algorithm in this group of gradient boosting algorithms. It constructs an ensemble of decision trees sequentially, with each tree designed to rectify the errors of the preceding one. During training, it minimizes a loss function by adjusting the weights of data points.

Table 12 displays the F1 scores obtained from the Gradient Boosting (GB) model applied to three different classes: Graduate, Dropout, and Enrolled. Additionally, it includes the results after applying the Optuna framework to optimize the Gradient Boosting model's hyperparameters. The initial Gradient Boosting model achieved an F1 score of 0.83 for the Graduate class, 0.85 for the Dropout class, and 0.76 for the Enrolled class. These scores represent the GB model's ability to accurately classify instances from each respective class.

To further improve the GB model's performance, the Optuna framework was employed to optimize its hyperparameters. However, the F1 scores for the Gradient Boosting model with Optuna remain the same as the initial Gradient Boosting model. This suggests that the hyperparameter optimization process did not have a substantial impact on the model's performance in this case.

b) Extreme gradient boosting (XGBoost)

XGBoost represents an optimized and highly efficient implementation of gradient boosting, featuring several enhancements such as regularization, parallel processing, and a tree-pruning algorithm. These improvements contribute to its speed and often result in higher predictive accuracy compared to traditional gradient boosting.

Table 13 displays the F1 scores obtained from the Extreme Gradient Boosting (XGBoost) model applied to three different classes: Graduate, Dropout, and Enrolled. Additionally, the table includes the results after applying the Optuna framework to optimize the Gradient Boosting model's hyperparameters. The initial XGBoost model achieved an F1 score of 0.86 for the Graduate class, 0.88 for the Dropout class, and 0.83 for the Enrolled class. These scores represent the XGBoost model's ability to accurately classify instances from each respective class.

To further improve the XGBoost model's performance, the Optuna framework was utilized to optimize its hyperparameters. Notably, the F1 scores for the XGBoost model with Optuna differ from the initial XGBoost model. The F1 score for the Graduate class is 0.81, 0.82 for the Dropout class, and 0.75 for the Enrolled class. In this scenario, the hyperparameter optimization process led to a slight decrease in the F1 scores for the XGBoost model.

c) CatBoost (CB)

CatBoost (CB) is another gradient boosting algorithm tailored for categorical feature support. It can automatically handle categorical variables without one-hot encoding. CB also incorporates robust handling of missing data and employs ordered boosting to enhance its performance.

Table 12 F1-Scores for gradient boosting with and without Optuna

Target	Description	Gradient boosting	Random forest + Optuna
0	Graduate	0.83	0.83
1	Dropout	0.85	0.85
2	Enrolled	0.76	0.76

Table 13 F1-Scores for extreme gradient Boosting with and without Optuna

Target	Description	Extreme Gradient Boosting	Gradient Boosting + Optuna
0	Graduate	0.86	0.81
1	Dropout	0.88	0.82
2	Enrolled	0.83	0.75

Table 14 presents the F1 scores obtained from the CB model applied to three different classes: Graduate, Dropout, and Enrolled. It also includes the results after applying the Optuna framework to optimize the CB model's hyperparameters. The initial CB model achieved an F1 score of 0.86 for the Graduate class, 0.87 for the Dropout class, and 0.82 for the Enrolled class. These scores reflect the CatBoost model's ability to accurately classify instances from each respective class.

The results show that the F1 scores for the CB model with Optuna differ from the initial CatBoost model. The F1 score for the Graduate class is 0.86, 0.88 for the Dropout class, and 0.84 for the Enrolled class. This indicates that the hyperparameter optimization process had an impact on the model's performance, resulting in slightly higher F1 scores for the Dropout and Enrolled classes.

d) LightGBM (LB)

LightGBM (LB) is a machine learning algorithm prized for its remarkable speed and memory efficiency. This is achieved through its histogram-based tree-building approach and gradient-based one-sided sampling. LB proves invaluable for handling large datasets, where optimizing training times and resource utilization are critical factors.

Table 15 presents the F1 scores obtained from the LB model applied to three different classes: Graduate, Dropout, and Enrolled. The table also includes the results after applying the Optuna framework to optimize the LightGBM model's hyperparameters. The initial LB model achieved an F1 score of 0.86 for the Graduate class, 0.87 for the Dropout class, and 0.83 for the Enrolled class. These scores reflect the LB model's ability to accurately classify instances from each respective class.

The F1 scores for the LB model with Optuna remain the same as the initial LB model. The F1 score for the Graduate class is 0.86, 0.88 for the Dropout class, and 0.83 for the Enrolled class. This indicates that the hyperparameter optimization process did not have a substantial impact on the model's performance in this case.

5.2 Unsupervised outlier detection with isolation forest

Isolation Forest (IF) is an anomaly detection algorithm frequently employed in machine learning and data mining to identify outliers or anomalies within a dataset. Its fundamental concept revolves around the isolation of anomalies by randomly selecting features and partitioning the data into subsets until the anomalies are singled out.

In our case study, we applied the Isolation Forest (IF) algorithm to assess its efficacy in outlier detection and classification within a dataset featuring three classes: failure, Enrolled, and Graduate. Our objective was to ascertain whether the minority classes, namely dropout and enrolled, would be identified as outliers. In order to evaluate the IF algorithm, we utilized the SMOTE test dataset, which consisted of 1325 instances. The application of the algorithm involved two stages. The first stage focuses on the overall results with the contamination as default. In the second stage, a contamination rate of 0.5 was applied to further evaluate the algorithm's performance.

In the first stage, the following results were obtained:

Table 14 F1-Scores for CatBoost with and without Optuna

Target	Description	CatBoost	Cat-Boost + Optuna
0	Graduate	0.86	0.86
1	Dropout	0.87	0.88
2	Enrolled	0.82	0.84

Table 15 F1-Scores for LightGBM with and without Optuna

Target	Description	LightGBM	Light-GBM + Optuna
0	Graduate	0.86	0.86
1	Dropout	0.87	0.88
2	Enrolled	0.83	0.83

1. Outlier detection:

Total number of outliers: The IF algorithm successfully identified 197 instances as outliers, representing data points that significantly deviated from the majority of the dataset. The outliers account for approximately 14.86% of the total dataset.

2. Minority classes:

Dropout Prediction and Enrolled Prediction: The IF algorithm predicted instances from the minority classes a total of 129 times. This highlights the algorithm's ability to identify and classify instances within these minority classes as outliers. These instances constitute approximately 9.73% of the total dataset.

3. Graduates:

Accurate Classification: The IF algorithm identified and classified instances as graduates 68 times. These instances account for approximately 5.13% of the total dataset.

In the second stage, the following results were obtained:

1. Anomalies Identified:

Dropout Prediction: The IF algorithm identified 357 instances of dropouts as anomalies, which represents 26.94%

Enrolled Prediction: the IF algorithm identified 155 instances of enrolled students as anomalies, which represents 11.69%.

2. Misclassifications:

The IF algorithm classified 181 instances as anomalies when they were actually graduates, which represents 13.66%. These indicate instances where the algorithm considered them to be anomalous, but in reality, they belong to the graduate class.

The performance of the IF algorithm in outlier detection and classification was promising. The algorithm showed effectiveness in identifying most of the minority classes as outliers. However, further analysis and refinement may be necessary to address the misclassification of graduates as anomalies while maintaining high accuracy in identifying other anomalies within the dataset.

5.3 SHAP method application

In order to gain a deeper understanding of the contribution of each feature in predicting the model's output, we employed the SHAP ("SHapley Additive exPlanations") method. This method provides insights into the importance and influence of different features on the model's predictions.

We applied the SHAP method in conjunction with two algorithms: LightGBM (LB) and CatBoost. By leveraging SHAP with the LB algorithm (Fig. 5) and the CatBoost algorithm (Fig. 6), we obtained visual representations of the top ten variables and their respective importance scores. The outputs from both the LB and CatBoost algorithms demonstrated similar results, with the same variables consistently appearing at the top in terms of their contribution to the model's predictions. This consistency indicates the robustness and reliability of these variables in influencing the model's output across different algorithms.

By examining the top ten variables identified through SHAP analysis, we can gain valuable insights into the factors that have the most significant impact on the model's predictions. These insights can guide us in understanding the underlying patterns and relationships within the dataset and help us make informed decisions related to feature selection, model interpretation, and potential areas for further investigation.

6 Discussion: performance evaluation and model comparison in predicting student outcomes

Table 16 summarizes the performance of the various supervised machine learning algorithms in the study. The models were evaluated based on their F1 scores for three target classes: Graduate, Dropout, and Enrolled. Boosting algorithms, specifically LightGBM (LB) and CatBoost (CB) with Optuna, consistently outperformed traditional classification algorithms in predicting student outcomes accurately, as evidenced by their higher F1 scores across the target classes. The success of boosting algorithms can be attributed to their ability to handle class imbalance effectively. Imbalanced class distributions are common in educational datasets, where minority classes, such as dropouts, require careful attention. Boosting algorithms, by nature, have mechanisms that address class imbalance issues and focus on learning from the minority classes, resulting in improved predictive performance.

To assess the discriminative capabilities of our top-performing models, LightGBM (LB) and CatBoost (CB), we employ ROC (Receiver Operating Characteristic) and AUC (Area Under the Curve) curves. The AUC value is a metric that ranges from 0 to 1, with higher values indicating superior model performance in classification tasks. In a broader context, an AUC value of 0.5 suggests random guessing (no discrimination), meaning the model struggles to distinguish between positive and negative cases. An AUC in the range of 0.7 to 0.8 is generally considered acceptable, while an AUC between 0.8 and 0.9 is regarded as excellent. Moreover, an AUC exceeding 0.9 is classified as outstanding, as indicated by Hosmer et al. [24].

Figure 7 (ROC Curve for Catboost—one-vs-one multiclass) and Fig. 8 (ROC Curve for LightGBM—one-vs-one multiclass) graphically represent the performance of CatBoost and LightGBM in a multi-class classification task. In such a scenario, ROC curves and AUC values are calculated for each class against every other class, providing insight into the model's ability to differentiate one class from the rest.

Our findings, depicted in Figs. 7 and 8, reveal an AUC exceeding 0.9. The position of the ROC curve on the graph reflects the model's performance. When the curve resides close to the top-left corner, it signifies that the model boasts a high true positive rate and a low false positive rate across various decision thresholds, a highly desirable trait. An AUC surpassing 0.9 is typically considered outstanding, indicating exceptional classification performance. This result underscores the exceptional discriminative prowess of these models, demonstrating their proficiency in distinguishing between different classes or groups within the dataset and affirming their reliability for the intended task.

To further evaluate potential significant differences between the results obtained using the LightGBM and Catboost algorithms, we employ the Statistical Tests for Algorithms Comparison (STAC) platform. This method serves to determine whether the disparity between the algorithm yielding the best results and other algorithms is statistically meaningful.

Based on the collected data, the STAC platform generates a decision tree to identify the most appropriate statistical test for the user-provided data. As presented in Fig. 9, the STAC assistant advised us the use of a specific statistical test: the paired samples t-test.

The paired sample t-test is employed to assess whether the mean difference between two sets of observations is statistically different from zero. In this type of t-test, each subject or entity is measured twice, yielding pairs of

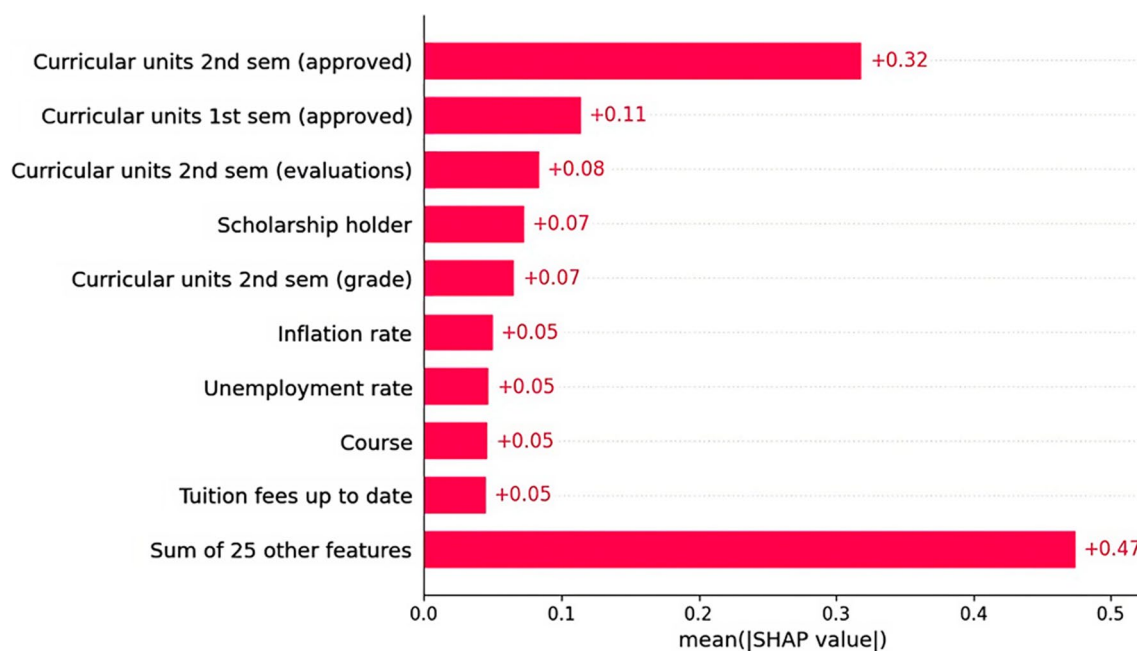


Fig. 5 SHAP Method using LightGBM (LB)

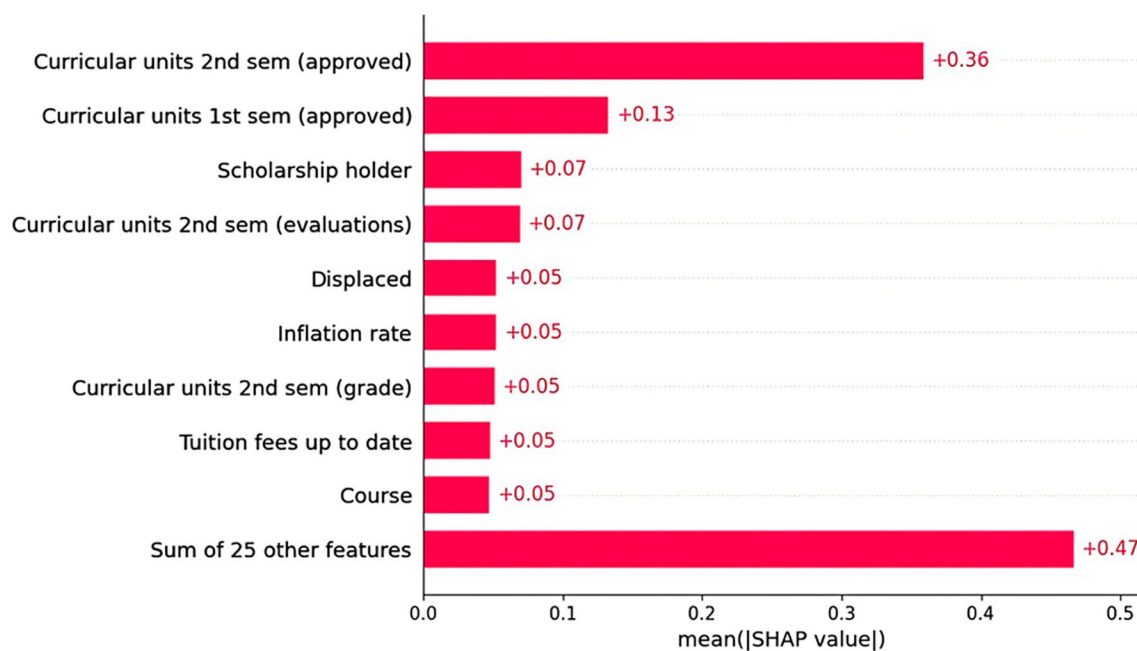


Fig. 6 SHAP Method using CatBoost (CB)

observations. In our specific case, we utilize the paired sample t-test to compare the f1-scores of the LightGBM and Catboost algorithms, which are the top-performing algorithms in our models. Our objective is to determine whether there exists a statistically significant difference between their results or if one outperforms the other.

The outcome of this analysis yielded a p-value of 0.003, using a significance level of 0.05. This indicates that the null hypothesis is accepted, implying that the f1-scores of the LightGBM and Catboost algorithms have statistically identical average values. Consequently, we can confidently conclude that there is no substantial difference between the f1-scores of these two models.

Table 16 Comparative performance of supervised machine learning algorithms

Supervised Algorithms	Target	FI-Score	FI-Score + Optuna
Decision tree	0	0.75	0.80
	1	0.78	0.81
	2	0.69	0.71
Random forest	0	0.84	0.84
	1	0.87	0.87
	2	0.81	0.80
Support vector machine	0	0.79	0.79
	1	0.76	0.76
	2	0.68	0.68
Gradient boosting	0	0.83	0.83
	1	0.85	0.85
	2	0.76	0.76
Extreme gradient boosting	0	0.86	0.81
	1	0.88	0.82
	2	0.83	0.75
CatBoost	0	0.86	0.86
	1	0.87	0.88
	2	0.82	0.84
LightGBM	0	0.86	0.86
	1	0.87	0.88
	2	0.83	0.83

7 Conclusion

This article presents a comprehensive study comparing supervised and unsupervised machine learning algorithms for predicting student dropout and academic success, with a primary focus on addressing class imbalance through SMOTE resampling. Notably, SMOTE significantly improved model accuracy. The study highlighted the exceptional performance of boosting algorithms, particularly LightGBM and CatBoost, surpassing traditional classification methods. However, these boosting algorithms often lack interpretability, which led us to employ the SHAP method for insightful feature analysis. We also explored the Isolation Forest algorithm for outlier detection, yielding promising results, although further refinement is needed to avoid misclassification.

These findings offer valuable guidance to researchers and practitioners in selecting suitable approaches for predicting student outcomes, optimizing model performance, and designing targeted interventions for at-risk students. Future research could explore alternative ensemble methods and advanced optimization techniques to enhance predictive model accuracy and interpretability in educational contexts.

As presented in our Systematic Literature Review (SLR) and corroborated in the present paper, Boosting algorithms consistently demonstrate their superiority in predicting various student performance aspects across diverse educational settings. Their consistent outperformance of traditional methods like logistic regression and decision trees underscores their potential as valuable tools for advancing educational research and enhancing student outcomes.

In the discussion section, we employ the ROC and AUC curves for the top-performing algorithms, LightGBM (LB) and CatBoost (CB). AUC values exceeded 0.9, which indicated outstanding classification performance, confirming these models' proficiency in distinguishing between different classes within the dataset. To assess differences between LightGBM and CatBoost, we employ the Statistical Tests for Algorithms Comparison (STAC), which shows that there is no substantial difference in their predictive performance.

Fig. 7 ROC Curve for Catboost
(one-vs-one multiclass)

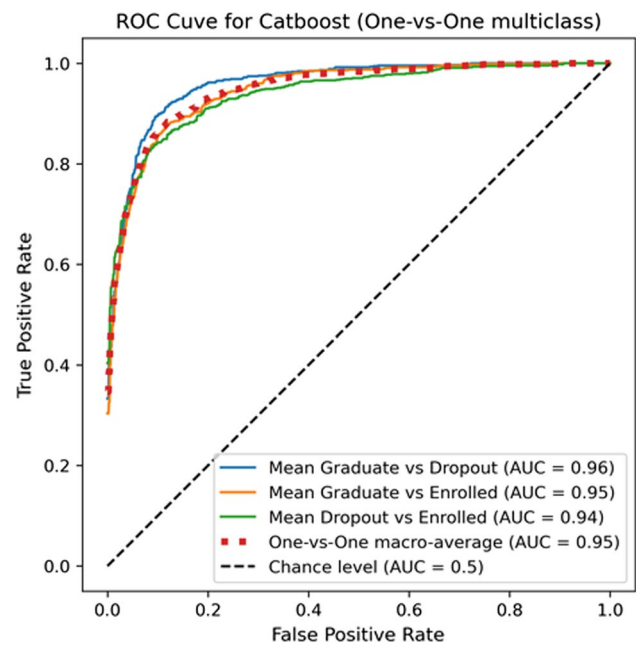
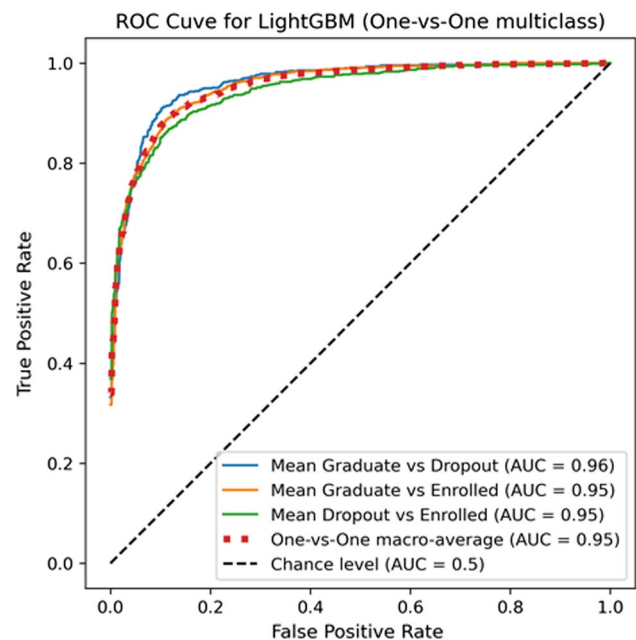


Fig. 8 ROC Curve for LightGBM
(one-vs-one multiclass)



It's important to acknowledge that our study has limitations, primarily its reliance on a single dataset, which may limit generalizability to other contexts. Future research avenues could explore hybrid ensemble methods, combining boosting algorithms with traditional classifiers, and investigate advanced optimization techniques beyond Optuna for hyperparameter fine-tuning to unlock further performance improvements.

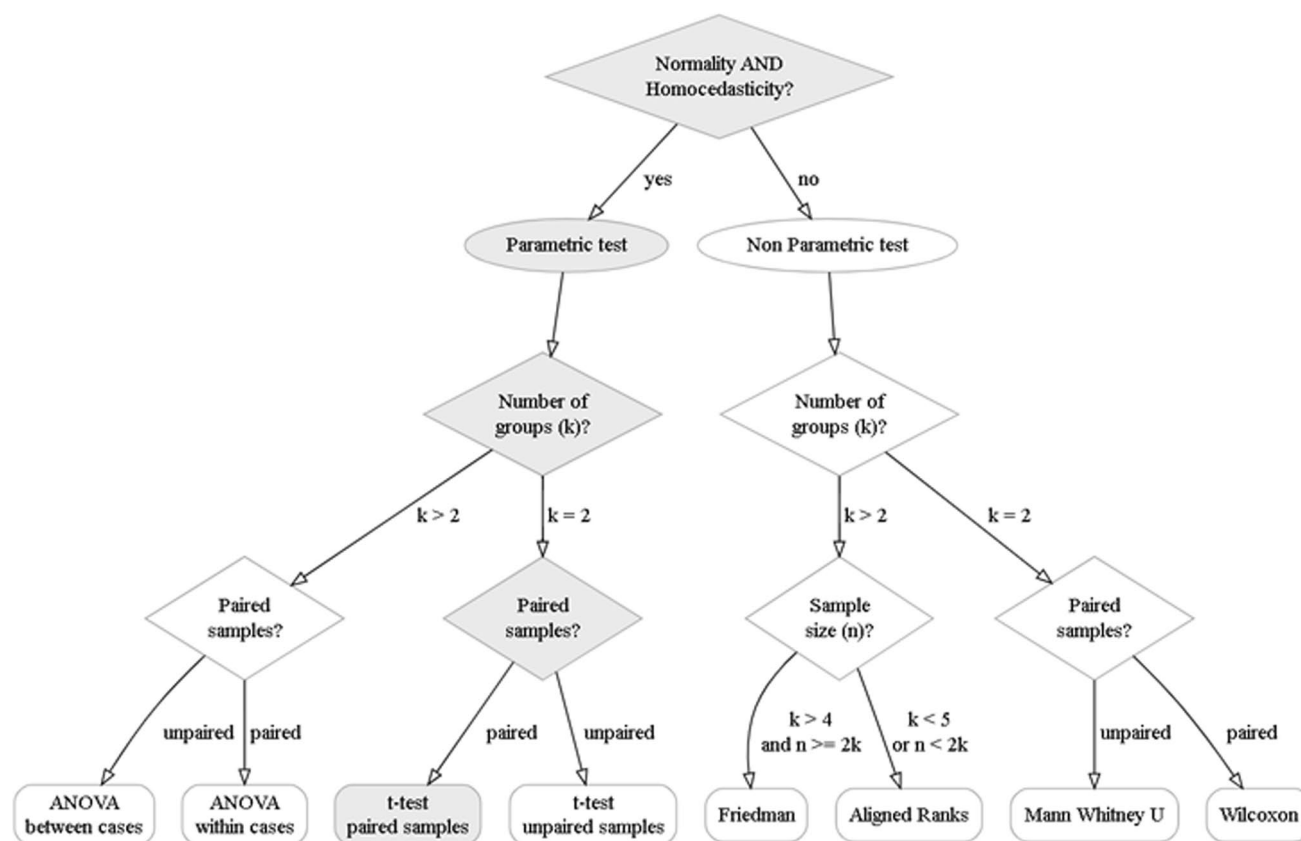


Fig. 9 STAC assistant decision tree

Acknowledgements We would like to sincerely thank graphic designer Juan Leite Farias for his contributions to the artwork's graphic design for this paper.

Author contributions Conception or design of the work: AV and CdA. Data collection: AV and CdA. Data analysis and interpretation: CdA and AV. Drafting the article: AV. Critical revision of the article: AV and CdA. Final approval of the version to be published: AV and CdA.

Funding The authors received no funding for this work.

Data availability The authors utilized a dataset from a higher education institution, which is publicly available. They have also made the dataset and necessary code accessible through a GitHub, b repository, ensuring transparency and reproducibility of their research project.

Code availability <https://github.com/alicevillar/SML-Comparative-Study>.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests We declare that we have no significant competing financial, professional, or personal interests that might have influenced the performance or presentation of the work described in this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Chung JY, Lee S. Dropout early warning systems for high school students using machine learning. *Child Youth Serv Rev*. 2019;96:346–53.
2. Gkontzis AF, Kotsiantis S, Panagiotakopoulos CT, Verykios VS. A predictive analytics framework as a countermeasure for attrition of students. *Interact Learn Environ*. 2022;30(6):1028–43.
3. Berens J, Schneider K, Görtz S, Oster S, Burghoff J. Early detection of students at risk—predicting student dropouts using administrative student data and machine learning methods. *SSRN J*. 2018. <https://doi.org/10.2139/ssrn.3275433>.
4. Martins MV, Tolledo D, Machado J, Baptista LM, Realinho V. Early prediction of student's performance in higher education: a case study. In: Rocha Á, Adeli H, Dzemyda G, Moreira F, Correia AMR, editors. *Trends and applications in information systems and technologies*, vol. 9. Berlin: Springer International Publishing; 2021. p. 166–75.
5. Thammasiri D, Delen D, Meesad P, Kasap N. A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition. *Expert Syst Appl*. 2014;41(2):321–30.
6. Islam A, Belhaouari SB, Rehman AU, Bensmail H. KNNOR: An oversampling technique for imbalanced datasets. *Appl Soft Comput*. 2022;115:108288.
7. Rastrollo-Guerrero JL, Gómez-Pulido JA, Durán-Domínguez A. Analyzing and predicting students' performance by means of machine learning: a review. *Appl Sci*. 2020;10(3):1042.
8. Hamim T, Benabbou F, Sael N. Student profile modeling using boosting algorithms. *Int J Web-Based Learn Teach Technol*. 2022;17(5):1–13.
9. Tenpipat W, Akkarajitsakul K. Student dropout prediction: a KMUTT case study. In: 2020 1st international conference on big data analytics and practices (IBDAP). IEEE. 2020. pp. 1–5.
10. Hew KF, Hu X, Qiao C, Tang Y. What predicts student satisfaction with MOOCs: a gradient boosting trees supervised machine learning and sentiment analysis approach. *Comput Educ*. 2020;145:103724.
11. Fernandes E, Holanda M, Victorino M, Borges V, Carvalho R, Van Erven G. Educational data mining: predictive analysis of academic performance of public school students in the capital of Brazil. *J Bus Res*. 2019;94:335–43.
12. Doryab A, Villalba DK, Chikeral P, Dutcher JM, Tumminia M, Liu X, Dey AK. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. *JMIR mHealth uHealth*. 2019;7(7):e13209.
13. Wang C, Zhao H, Zhang H. Chinese college students have higher anxiety in new semester of online learning during COVID-19: a machine learning approach. *Front Psychol*. 2020;11:587413.
14. Qiao X, Jiao H. Data mining techniques in analyzing process data: a didactic. *Front Psychol*. 2018;9:2231.
15. Ge F, Zhang D, Wu L, Mu H. Predicting psychological state among Chinese undergraduate students in the COVID-19 epidemic: a longitudinal study using a machine learning. *Neuropsychiatr Dis Treat*. 2020;16:2111–8.
16. Asselman A, Khaldi M, Aammou S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interact Learn Environ*. 2021. <https://doi.org/10.1080/10494820.2021.1928235>.
17. Jung JS, Park SJ, Kim EY, Na KS, Kim YJ, Kim KG. Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. *PLoS ONE*. 2019;14(6):e0217639.
18. Costa-Mendes R, Oliveira T, Castelli M, Cruz-Jesus F. A machine learning approximation of the 2015 Portuguese high school student grades: a hybrid approach. *Educ Inf Technol*. 2021;26(2):1527–47.
19. Chen J, Zhang Y, Hu J. Synergistic effects of instruction and affect factors on high-and low-ability disparities in elementary students' reading literacy. *Read Writ*. 2021;34:199–230.
20. Nabil A, Seyam M, Abou-Elfetouh A. Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*. 2021;9:140731–46.
21. Abidi SMR, Hussain M, Xu Y, Zhang W. Prediction of confusion attempting algebra homework in an intelligent tutoring system through machine learning techniques for educational sustainable development. *Sustainability*. 2018;11(1):105.
22. Niyogisubizo J, Liao L, Nziyumva E, Murwanashyaka E, Nshimyumukiza PC. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization. *Comput Educ Artif Intell*. 2022;3:100066.
23. Zhang L, Li X, Chen H, Wu Z, Hu M, Yao M. Haze air pollution health impacts of breath-borne VOCs. *Environ Sci Technol*. 2022;56(12):8541–51.
24. Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied logistic regression*, vol. 398. Hoboken: John Wiley & Sons; 2013.
25. Cohen J, Wright-Berryman J, Rohlf L, Wright D, Campbell M, Gingrich D, Pestian J. A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. *Int J Environ Res Public Health*. 2020;17(21):8187.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.