ENHANCING PREDICTIVE ANALYTICS WITH STACKING AND SYNTHETIC MINORITY OVER-SAMPLING BALANCING

Loice Yator Kutoh[1]
Department of Applied Sciences
Rift Valley Technical Training Institute
Eldoret, Kenya
eckutoh@gmail.com
info@rvti.ac.ke

Chrispus Zacharia Oroni[2]
School of Information Science and Technology
Dalian Maritime University
Dalian, China
chrisoroni12@gmail.com

Serem Kenneth Kipruto [3]
Department of Business and Liberal Studies
Rift Valley Tehnical Training Institute
Eldoret, Kenya
seremken454@gmail.com

Abstract

In the sphere of education, the assessment of student advancement is of utmost importance for educational establishments. These institutions sometimes encounter hurdles such as appraising student performance, conducting academic scrutiny, and foreseeing prospective learning requisites. To tackle these impediments, academic intervention schemes have been put into operation. Accurate prediction regarding student performance not only facilitate the early identification of struggling learners but also assist higher education institutions in their decision-making processes. Consequently, this study introduces a predictive model for the evaluation of scholastic accomplishment. It utilizes XGBoost for the selection of pertinent attributes and employs Synthetic Minority Oversampling Technique (SMOTE) to rectify class imbalances. Furthermore, it employs a stacking ensemble methodology encompassing base models like K-Nearest Neighbors, Support Vector Machines, Random Forest, Decision Trees, Multi-Layer Perceptron, and Logistic Regression to enhance predictive precision. This investigation provides an exhaustive theoretical assessment of models, showcasing practical applications of machine learning in education. With an exceptional 98% precision in the prediction of student performance, this research underscores the capacity of educational institutions to elevate outcomes. This study substantially propels predictive modeling in the realm of education, making a valuable contribution to the field of research.

Key words*: Predictive analytics, stacking model, educational data mining, SMOTE, ensemble learning*

Introduction

In the field of education, the assessment of students' progress and achievements plays a crucial role, serving as an indispensable tool within educational institutions (Albreiki et al., 2021). Today's educational landscape finds institutions operating in highly competitive and demanding environments, presenting numerous challenges such as the delivery of high-quality education, the

[1] Corresponding Author
[2] Corresponding Author
[3] Co-Author

establishment of effective systems for evaluating student performance, conducting comprehensive academic analyses, and anticipating the future learning needs of students (Albreiki et al., 2021). In response to these challenges, academic institutions have implemented academic intervention plans, designed to support students in navigating complex learning situations. These plans extend their benefits to school administrators and education stakeholders, aiding in the development of strategies grounded in predictive assessments of student performance, both at the point of entry and throughout subsequent semesters (Albreiki et al., 2021).

As the educational landscape continues to evolve, the education sector has demonstrated keen interest in predicting student success across pre-tertiary and higher education settings (Nawang et al., 2021). The development and application of predictive models entail the collection and analysis of variables correlated with the outcomes to be predicted. These variables encompass a broad spectrum of factors, including students' prior educational experiences, the utilization of e-learning platforms, socio-demographic characteristics, data from their interactions on social media platforms, cognitive and behavioral patterns, formative assessments, participation in extracurricular and co-curricular activities, the school environment, family dynamics, and more (Nawang et al., 2021; Albreiki et al., 2021). Educational data mining has gained prominence as it explores the application of data mining tools for analyzing educational data within higher education institutions, emphasizing how data mining, machine learning, and statistical techniques can be employed to scrutinize educational information. This includes data from universities, learning management systems, and intelligent tutoring systems, unlocking insights for enhancing the quality of educational offerings and support systems (Mengash, 2020; Akgün, 2020; Alamgir et al., 2023).

Educational data mining offers valuable insights for academic institutions, researchers, and students alike, shedding light on various aspects of student performance, learning methodologies, and educational experiences (Romero & Ventura, 2010). These insights empower educators with improved resources for lesson planning and material assessment (Rabelo et al., 2023). Moreover, they provide educational scholars with deeper insights into student behavior within the classroom environment and the profound impact of learning settings on students' educational journeys (Aulakh et al., 2023), ultimately contributing to enhanced student engagement and retention rates. Accurate predictions of student performance are of paramount importance as they enable the early identification of struggling students (Romero & Ventura, 2010). Additionally, they offer valuable insights into the learning processes and academic progress of students for higher education institutions (Uylaş, 2018). However, predicting student success is a complex task due to the multitude of factors that can influence academic achievement, driving extensive research efforts to identify critical determinants (Aulakh et al., 2023; Márquez-Vera et al., 2012).

Predictive analytics encompasses various methodologies aimed at making well-informed forecasts about uncertain future events (Liu, 2015). In education, this discipline plays a crucial role in anticipating academic achievements, evaluating teaching methodologies, and assessing administrative metrics such as student retention and course enrollments. It's essential to distinguish between two fundamental modeling paradigms: explanatory and predictive modeling. Explanatory modeling leverages existing data to predict forthcoming values or classifications based on observed variables, while predictive modeling focuses on creating models to forecast values or categories of novel data, with a specific emphasis on future events. Predictive modeling is tailored to project future outcomes. In educational data mining, models are often evaluated primarily on their predictive accuracy, especially their ability to forecast learner performance (Liu, 2015).

In recent years, educational institutions have increasingly utilized machine learning techniques to gain insights into students' academic trajectories, predict future behaviors, proactively address potential challenges, and foster collaborative environments among students and educators (Atif, 2013). Predictive analytics within higher education enhances teaching methodologies and encourages reflective practices in students, enabling a deeper understanding of their learning patterns (Dervenis et al., 2022; Atif, 2013). Continuous monitoring is crucial for forecasting students' academic achievement, necessitating frequent updates to accommodate shifts in their educational trajectory. This is because students often begin and complete new courses, either successfully or unsuccessfully, within predetermined timeframes. Predicting student success involves monitoring their development throughout their academic careers, rather than relying solely on a single assessment (Xu et al., 2017). Students have the flexibility to enroll in courses immediately or defer them to future semesters, and they may also choose to retake courses they had previously failed. The ensemble progressive prediction approach allows for a continual improvement of results, utilizing previous forecasts to inform current predictions.

In this article, our primary objective is to construct a robust predictive model for assessing students' academic performance. To achieve this, we employ a feature selection method based on XGBoost (XGB) to identify and extract relevant features for our modeling efforts. Additionally, recognizing the dataset's imbalance, we apply the Synthetic Minority Over-Sampling Technique (SMOTE) to address this issue effectively by generating synthetic samples for the minority class. To further enhance our model's predictive capabilities, we adopt a stacking ensemble approach, integrating ensemble learning techniques from various base models, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), Multi-Layer Perceptron (MLP), and Logistic Regression (LR). Leveraging the strengths of these diverse algorithms, we aim to enhance the accuracy and robustness of our prediction framework. This approach not only allows us to harness the unique strengths of each base model but also results in a more powerful and reliable predictive model for assessing students' academic performance.

The forthcoming sections of the paper are structured as follows: Section 2 engages in an extensive exploration of the established body of literature, offering essential background information. In Section 3, we expound upon the methodology we have put forth, emphasizing the significance of diverse attributes. Progressing to Section 4, we furnish a comprehensive scrutiny of the results generated by our proposed approach, followed by an exhaustive discourse. Lastly, within Section 5, we wrap up our paper while also delineating prospects for future research. This arrangement guarantees a coherent and systematic dissemination of information across the remaining portions of the document.

The domain of forecasting student academic achievement has been the focus of extensive research, with various methodologies for selecting relevant features and employing classification algorithms. These investigations have generated valuable findings but have also highlighted areas for enhancement, especially concerning predictive precision and the management of class imbalance problems.

One notable trend in recent research is the adoption of ensemble algorithms within the educational data mining domain. These ensemble models have demonstrated superior performance when compared to traditional single models. For instance, Mastour et al. (2023) conducted a comprehensive study comparing classical models like Logistic Regression (LR), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN) with ensemble models, including Voting, Bagging (BG), Random Forests (RF), Adaptive Boosting (ADA), eXtreme Gradient Boosting (XGB), and Stacking. Their findings underscored the superiority of ensemble

machine learning models, signaling a shift towards leveraging the collective intelligence of multiple base models. While ensemble models exhibit remarkable predictive power, it's crucial to consider model interpretability. Decision Trees and rule-based classifiers, offer transparency by elucidating the underlying reasoning behind predictions. Additionally, alternative techniques like clustering algorithms or association rule mining can be explored to shed light on performance determinants (Lang et al., 2015). The choice of classification and regression methods is vast, ranging from neural networks and Bayesian networks to support vector machines, logistic regression, and linear regression, all of which hold potential for solving the student performance prediction problem (Lang et al., 2017).

Several prior studies have contributed valuable insights to the field of predicting student performance. Rahman et al. (2017) applied information gain and wrapper feature selection methods in conjunction with Naïve Bayes, Decision Tree, and Artificial Neural Networks, with the Artificial Neural Network achieving the highest accuracy at 79.37%. Kumar & Pal (2011) explored Decision Trees, Neural Networks, and K-Nearest Neighbors to predict student results, with Decision Trees emerging as the most accurate. Zaffar et al. (2017) delved into the impact of feature selection algorithms on student performance analysis, discovering that principal component analysis combined with the random forest classifier yielded the best results.

González-Brambila et al. (2018) sought to predict the academic performance of engineering students using Decision Trees but achieved only 63.19% accuracy. K-Nearest Neighbors (kNN), renowned for its simplicity (Howard, W.R., 2007), classifies cases based on similarity measures. Schalk et al. (2011) observed the accuracy of Support Vector Machine and Neural Network models but identified challenges related to adaptability and overfitting, largely attributable to missing values in datasets, leading them to explore Random Forest as a more adaptable alternative.

Recent research by Mulyana et al. (2023) achieved an impressive 89% accuracy in classifying successful and dropout students using Random Forest. Qin et al. (2017) leveraged the Gini index to rank factors affecting student performance and employed Random Forest for prediction, achieving an accuracy of 83.82%. Begum &Padmannavar (2023) introduced Bayesian-optimized KNN and Random Forest models, yielding accuracies of 87% and 73%, respectively, in forecasting student success. Sawangarreerak&Thanathamathee (2020) introduced a combined sampling technique to enhance the classification of university student depression data, achieving an overall accuracy of 94.17%. In a different investigation conducted by Alamri et.al (2020) aimed at forecasting student performance, they employed Support Vector Machines (SVM) and Random Forest (RF). The outcomes of the study demonstrated that both SVM and RF algorithms achieved an accuracy rate of 93%. Also, In the study conducted by Nugroho et al. (2020), which assessed student academic performance using the K-Nearest Neighbors (KNN) algorithm, the findings were notably evident and indicated a favorable level of accuracy.

Haryawan&Sebatubun (2020) employed Multi-Layer Perceptron (MLP) for predicting student failure, achieving an accuracy rate of 86.9%. Lagman (2015) applied logistic regression to predict student graduation, focusing on early identification of students at risk of delayed graduation. Yohannes & Ahmed (2018) evaluated final cumulative grade point averages using Neural Networks, Support Vector Regression, and Linear Regression, showcasing the potential of data mining techniques for predicting academic performance. Mativo& Huang (2014) examined Support Vector Machine and Multiple Linear Regression on a smaller dataset, demonstrating that the SVM model exhibited higher accuracy in identifying students with low grades.

Building on this body of research, our study proposes a pioneering approach that capitalizes on the strengths of multiple base models, including Decision Trees (DT), Logistic Regression

(LR), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP), within a stacked ensemble framework. By combining the insights from these diverse base models, our innovative ensemble model aims to elevate predictive accuracy, especially when compared to single models.

To address the challenge of class imbalance frequently encountered in educational datasets, our study integrates the Synthetic Minority Over-Sampling Technique (SMOTE). This technique generates synthetic samples for the minority class, ensuring a balanced dataset and reducing potential biases in predictions. Furthermore, our research employs the eXtreme Gradient Boosting (XGB) algorithm for feature selection. XGB is well-regarded for its efficacy in identifying and extracting relevant features from complex datasets, thus enhancing the overall performance of predictive models. In addition to predictive accuracy, our study evaluates model performance using various metrics, including accuracy, precision, ROC curve and F1-score, providing a comprehensive assessment of our framework's capabilities.

In summary, this study contributes to the field of educational data mining by introducing an innovative predictive framework. Our research addresses existing challenges related to predictive accuracy, class imbalance, and model interpretability. The proposed ensemble model, enriched with SMOTE balancing and XGB feature selection, has the potential to significantly advance the prediction of student academic performance. By amalgamating the strengths of diverse base models, we aspire to provide educational institutions, administrators, and students with a comprehensive and accurate tool for assessing academic success.

Methodology

This section elaborates on the study's framework, encompassing a comprehensive description of the dataset employed, the methodologies applied in data preprocessing, and the techniques utilized for model evaluation.

Framework

In this section, we explore the detailed framework employed to create a robust predictive model, as depicted in Fig. 1.

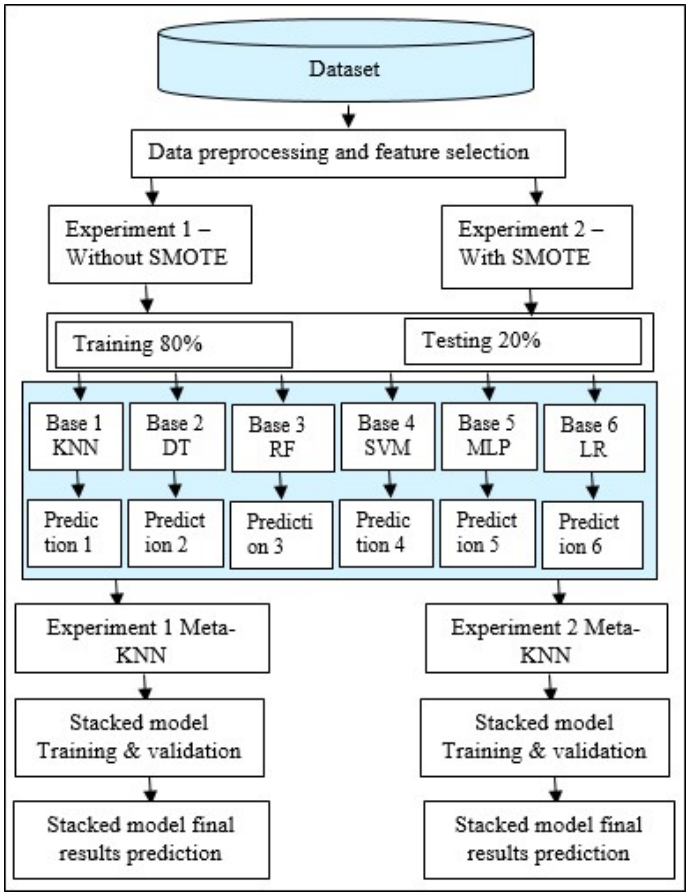

Fig.1 Implementation of stacking predictive model

This systematic process consists of a series of well-planned stages aimed at ensuring the model's efficacy and reliability. The initial step involves Data Preparation, a crucial phase where the dataset undergoes careful preprocessing. This includes essential tasks such as separating the features (X) from the target variable (y) and addressing any data intricacies. Categorical variables are transformed, and missing data is meticulously managed. An examination of the class distribution within the target variable (y) is conducted to identify any disparities, which are important considerations in predictive modeling. To tackle this issue, the Synthetic Minority Over-Sampling Technique (SMOTE) is expertly employed. SMOTE proficiently generates synthetic samples by interpolating between existing minority class samples, effectively addressing class imbalances. Subsequently, the dataset is divided into training and testing sets, facilitating model evaluation. A variety of base machine learning models are enlisted, including Decision Trees, k-Nearest Neighbors, Random Forest, Logistic Regression, Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP) classifiers. Each base model is meticulously trained on the training dataset and then utilized to make predictions on the testing dataset. To enhance predictive performance, a meta-model, KNN, is introduced for stacking purposes. Stacking involves combining the predictions from the base models to improve accuracy and enhance generalization. Performance assessment encompasses a range of metrics, including accuracy,

precision, recall, and the F1-score. The ensemble model is then employed for predictions on new, unseen data, followed by a thorough analysis that includes evaluating confusion matrices, ROC curves, and other relevant metrics. This methodical approach underscores our commitment to developing a potent predictive model that can be reliably applied in practical scenarios.

Dataset

The dataset comprises 120 samples, representing students enrolled in an Information Communication Technology diploma course at a publicly-funded tertiary institution in Kenya.

Table 1. Family social – economic status and parental level of education

SN	Feature	Data type	Values
1	Student Social status (Cheng & Kaplowitz, 2016)	Categorical	[0,1]
2	Family Size (Adongo et al., 2022)	Categorical	[0,1,2]
3	Parental absence (Liu & Hannum, 2023)	Categorical	[0,1,2]
4	Family economic level (Li & Qiu, 2018)	Categorical	[0,1,2,3,4]
5	Parental employment status (Lehti, 2023)	Categorical	[0,1,2]
6	Parents education level (Tantoh, 2023)	Categorical	[0,1,2,3,4]
7	Family relationship (Li & Qiu, 2018; Kleinkorres et al., 2023; Prakhov et al., 2020)	Categorical	[0,1,2,3,4]

To gather this valuable data, the study utilized a carefully designed questionnaire as the primary data collection instrument. The research employed a comprehensive survey, consisting of a total of 30 unique items organized into three distinct sections. The first section focused on socio-economic factors and the educational background of participants, as detailed in Table 1.

Table 2. Utilization of learning resources, sleep patterns, geographical proximity, and class attendance

SN	Feature	Data type	Values
1	Usage of library resources (Wang & Wang, 2023; (Hanaysha et al., 2023)	Categorical	[0,1,2]
2	Usage of ICT resources (Wang & Wang, 2023; (Hanaysha et al., 2023)	Categorical	[0,1,2]
3	Class attendance (Lucey & Grydaki, 2022; (Hanaysha et al., 2023)	Categorical	[0,1,2]
4	Sleep hours (Desjardins & Grandbois, 2022; (Chan et al., 2023)	Categorical	[0,1,2,3]
5	Usage of reference materials (Wang & Wang, 2023)	Categorical	[0,1,2]
6	Geographical distance (Arrival time, transport means) (Banda et al., 2023)	Categorical	[0,1]

The second section aimed to collect information on various aspects such as the use of learning resources, sleep patterns, geographical proximity, and class attendance, as outlined in Table 2. Lastly, the third section was designed to capture students' academic performance, specifically concerning the grades they achieved in specific courses during the initial module of their first year, as indicated in Table 3. These questionnaire items were drawn from previous research, as indicated in the respective tables.This systematic structuring of features enables a comprehensive examination of students' characteristics and their academic performance in the context of the Information Communication Technology diploma program. In essence, this dataset stands as a valuable asset for conducting in-depth analyses and research inquiries aimed at comprehending the intricate connections between demographic, economic, and educational variables, and how they collectively impact the academic achievements of students in the field of Information Communication Technology.

Data preprocessing

Data preprocessing is a crucial step in any data analysis or modeling project, as it significantly impacts the accuracy and reliability of subsequent analyses and predictions. This study followed several key steps in the data preprocessing phase.

Table 3. Course grades in the first module

SN	Feature (Rajendran et al., 2022)	Data type	Values
1	Gender	Categorical	[0,1]
2	System analysis and design (SAD)	Numeric	[0-100]
3	Object oriented programming (OOP)	Numeric	[0-100]
4	Database management system (DBMS)	Numeric	[0-100]
5	Visual programming (VP)	Numeric	[0-100]
6	Quantitative techniques (QT)	Numeric	[0-100]
7	Computer applications (CA)	Numeric	[0-100]
8	Final score (Pass: P, Referred: R, Course Requirements not Met: C)	Categorical	

Data cleaning

Upon thorough examination, we found that there were no missing values in the dataset. Therefore, there was no need for any data imputation or handling of missing information.

Data transformations

We performed a data transformation step by using a technique called one-hot encoding. This method was applied to convert categorical variables into a numerical format, making them suitable for analysis and modeling.

Feature selection

The original dataset contained 30 different features, and we recognized the need to reduce dimensionality. To achieve this, we employed XGBoost for feature selection. The XGBoost algorithm is a highly regarded distributed gradient-boosted decision tree (GBDT) machine learning framework known for its versatility. It offers a wide range of objective functions, making it suitable for various tasks, including regression, classification, and ranking. What sets XGBoost apart is its remarkable predictive prowess, linear modeling capabilities, and proficiency in tree-based learning methods. It stands out as the preferred choice for assessing accuracy in diverse scenarios. Notably, when compared to existing gradient boosting algorithms, XGBoost demonstrates superior speed and efficiency, as evidenced by a study conducted by Priscilla and Prabha (2021). This distinction highlights its significance in the field of machine learning and predictive analytics. XGBoost goes a step further by incorporating a second-order Taylor expansion of the loss function and introducing a regularization term into the objective function, effectively mitigating overfitting. The XGBoost algorithm leverages decision trees as its foundational learners to construct a potent predictive model, resulting in an ensemble that yields compelling results. The prediction model of XGBoost can be succinctly expressed as (Eq1) - (Eq4):

$$y_i = \sum_{k=1}^{K} f_k(x_i) \quad (1)$$

Where *K* represents the number of decision trees, f_k represents the prediction result of the kth tree and y_i represents the prediction result of sample *i*. The objective function of XGBoost algorithm is:

$$J(\theta) = \sum_{i=1}^{n} l(y_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

Where the first term is the training error of sample *i*, and the second term is the cumulative sum of all regularization terms of decision tree, where

$$\Omega(f) = \sqrt{T} + \frac{1}{2}\lambda||\omega||^2 \qquad (3)$$

Where T is the number of leaf nodes of each decision tree, ω is the weight of leaf node, and γ and λ is the corresponding coefficient. According to the change of objective function before and after node splitting, XGBoost algorithm chooses to split nodes or not and sets threshold value of splitting nodes. It uses the gain to evaluate the advantage and disadvantage of each feature and the split point of each feature. And then it uses the approximate algorithm or greedy algorithm to select the feature and split node. The gain is calculated as follows：

$$\square = \frac{1}{2}[\frac{\square_{\square}{}^2}{H_L + \lambda} + \frac{\square_{\square}{}^2}{H_R + \lambda} + \frac{(\square_{\square} + \square_{\square})^2}{H_L + H_R + \lambda}] - \gamma \qquad (4)$$

Where $\square_{\square}$is the sum of the first step degree of all samples on the left leaf nodes after splitting, and $\square_{\square}$is the sum of the second step degree of that. and $\square_{\square}$ are the sum of the first-order and second-order degrees of all samples on the right leaf nodes after splitting respectively. Therefore, tthrough this rigorous selection process based on XGB, we retained only 15 features, as shown in Fig. 2. The primary goal of this reduction was to improve model efficiency and performance by focusing on the most relevant and informative features.

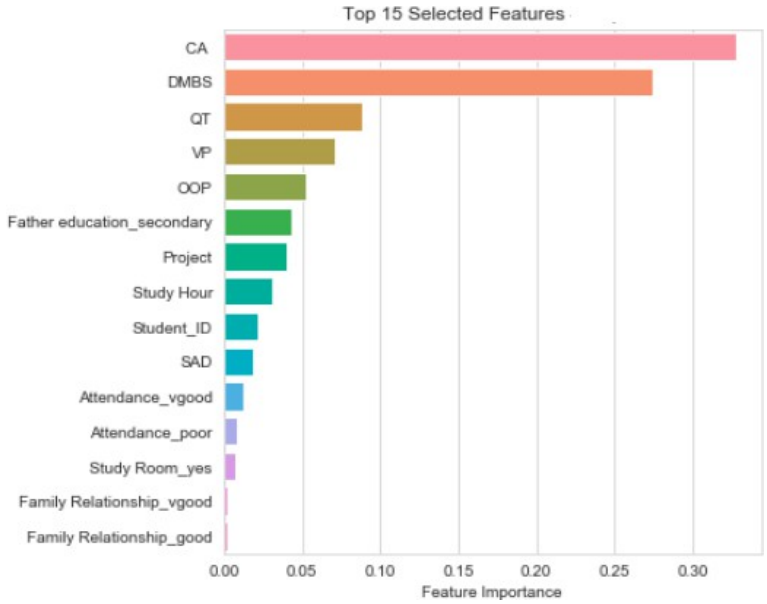

Fig.2. Feature selection based on XGB

Handling class imbalance

In numerous real-world applications, we frequently encounter imbalanced classification challenges wherein the dataset exhibits a significant disproportion in the distribution of instances across two distinct classes. This discrepancy manifests as one class being the predominant or majority class, housing a substantial volume of data, while the other class represents the minority, comprising only a limited number of instances. Effectively mitigating this class imbalance poses a primary challenge, with a specific focus on enhancing the performance of the underrepresented minority class. An effective approach to rectify imbalanced datasets involves the generation of supplementary synthetic data points to bolster the minority class, thereby mitigating its scarcity of representative data. The Synthetic Minority Over-Sampling Technique (SMOTE), first introduced

by Chawla et al. (2002), stands out as a widely embraced method within the academic literature for executing this oversampling strategy. SMOTE operates by creating novel synthetic instances for the minority class by drawing insights from the existing dataset. It begins by identifying data points associated with the minority class and selecting a single point from this subset. Subsequently, SMOTE pinpoints the K-nearest neighbors of the chosen point within the minority class. Synthetic instances are then synthesized by interpolating along the line segments connecting the chosen point and its K-nearest neighbors. The adoption of SMOTE in this study seeks to rectify the dataset's imbalance, as visually depicted in Fig. 3. This equates to an adjustment ensuring that the minority class attains a more balanced representation, rendering it akin in size to the majority class. This balancing methodology holds paramount significance as it contributes to assessing the overall data quality, particularly in terms of how adeptly the artificially generated data points mirror the genuine underlying distribution characterizing the minority class. In essence, SMOTE serves as an invaluable tool for effectively addressing the challenges posed by class imbalance, facilitating the creation of a more robust and balanced dataset primed for subsequent analytical and modeling endeavors.

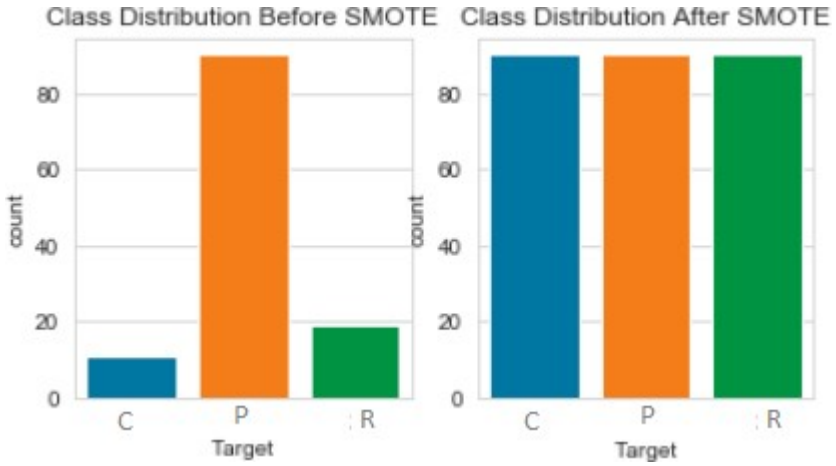

Fig.3 Class distribution before and after applying SMOTE balancing approach

Data splitting

This research encompassed two distinct experiments to investigate the phenomenon under scrutiny. In the initial experiment, the dataset remained unaltered, with no attempts made to address its inherent imbalance. The original dataset was subjected to a division, allocating 80% of its composition to the training set and reserving the remaining 20% for validation purposes. This partitioning yielded sample sizes of 96 and 24 for the training and validation sets, respectively. Subsequently, the second experiment was conducted on a dataset that had been meticulously balanced, rectifying the previously observed class imbalance. In alignment with the established protocol, this balanced dataset was also subjected to a division into training and validation subsets, adhering to the same proportions. Consequently, 80% of the balanced dataset was dedicated to the training set, with the remaining 20% designated for validation purposes. This approach ensured that both experiments maintained consistency in terms of dataset partitioning, facilitating a comparative analysis of results and outcomes.

Cross validation

Cross-validation serves as a valuable method for evaluating the effectiveness of a machine learning algorithm and predicting its performance on new data drawn from the same underlying distribution, as outlined by King et al. (2021). In this study, the training process employed a cross-validation approach with 10 iterations, aimed at attaining the optimal model. This technique enables the assessment of the machine learning algorithm's capabilities and its potential to generalize to unseen data points. By repeatedly splitting the dataset into distinct training and validation subsets, with each iteration using a different portion for validation while the rest is employed for training, a more robust and reliable model is developed. The application of 10 repetitions in the cross-validation process enhances the model's robustness by subjecting it to a diverse range of training and validation data combinations. This approach ultimately contributes to the selection of the most suitable model that exhibits strong predictive performance across various scenarios.

Modelling evaluation

Several distinct independent models were enlisted for the modeling process, encompassing Decision Trees, k-Nearest Neighbors, Random Forest, Logistic Regression, Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP) classifiers. Additionally, in an endeavor to enhance predictive prowess, a meta-model, typically KNN, was introduced to facilitate the stacking of predictions. The modeling phase was executed on two separate datasets: one with balanced class distribution and another with an unbalanced distribution. A comparative analysis was subsequently conducted, contrasting the individual models against the proposed stacked model. This approach aimed to discern the advantages and performance gains achieved through stacking. To evaluate the efficacy of these models, a comprehensive set of metrics was employed. These metrics encompassed Accuracy, Precision, the F1-Score, and the ROC Curve as represented in (Eq5) – (Eq10). Each of these metrics plays a distinctive role in gauging the model's performance in various aspects. Accuracy provides an overall assessment of correct predictions, Precision evaluates the proportion of true positives among all positive predictions, and the F1-Score offers a balance between Precision and Recall. The ROC Curve graphically illustrates the model's ability to discriminate between positive and negative classes across different thresholds. In summary, the utilization of diverse models, including stacking, on both balanced and unbalanced datasets, alongside a multifaceted set of performance metrics, contributed to a comprehensive evaluation of their predictive capabilities and highlighted the strengths and weaknesses of each approach.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{F-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

TP signifies the count of correctly predicted positive instances, while TN represents the number of correctly predicted negative instances. On the other hand, FP denotes the count of instances falsely classified as positive, and FN indicates instances falsely classified as negative. The Receiver Operating Characteristic (ROC) curve serves as a valuable tool for assessing model performance across various threshold settings in classification problems. The ROC curve presents a graphical depiction of the True Positive Rate (TPR) on the y-axis and the False

Positive Rate (FPR) on the x-axis. It allows for a visual examination of a model's ability to distinguish between positive and negative classes at different decision thresholds. Furthermore, the Area Under the Characteristic (AUC) corresponds to the area beneath the ROC curve. A higher AUC value signifies superior model performance, indicating a model's proficiency in correctly classifying instances across a spectrum of threshold values. Essentially, a larger AUC underscores the model's effectiveness and its capacity to make accurate predictions across a range of decision boundaries.

$$TPR = \frac{TP}{TP+FN} \quad (9)$$

$$FPR = \frac{FP}{FP+TN} \quad (10)$$

Results and discussion

This section presents the outcomes achieved concerning the composite models, both with and without dataset balancing. These composite models involve the application of stacking, utilizing Decision Trees, k-Nearest Neighbors, Random Forest, Logistic Regression, Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP) classifiers, in conjunction with KNN as the meta-classifier. Additionally, the section provides an overview of the parameter selection.

Hyper-parameter selection

Hyperparameters hold a significant role in shaping the model's performance. Fine-tuning these hyperparameters is crucial to elevate the model's efficiency on the dataset, leading to an overall enhancement in its quality. In this investigation, we employed a hybrid strategy that amalgamates both grid search and 10-fold cross-validation techniques to optimize the machine learning model's hyperparameters. The grid search approach is a thorough method that explores diverse hyperparameter combinations, subjecting the model to multiple training iterations. The selection of the optimal hyperparameter set is contingent on achieving the highest performance observed during these training sessions. Meanwhile, the 10-fold cross-validation method partitions the dataset into ten independent subsets, employing nine for training without duplication and reserving one for testing in each cycle. This iterative process is repeated ten times, yielding ten unique combinations, and their outcomes are averaged to yield the final result. By integrating the grid search and cross-validation methodologies, we systematically assessed all hyperparameter combinations within the grid search using the robust 10-fold cross-validation technique. This comprehensive approach effectively bolsters the model's capacity for generalization and adept performance on unseen data, a concept elucidated by Pedregosa et al. (2011).

Base learner's and stacked model results

As the efficacy of the ultimate stacking model hinges on the performance of its underlying base learners, this research commenced with an assessment of the aptness of these base learners. The initial phase involved an appraisal of the predictive proficiency of these base learners in isolation, entailing their application to generate predictions for the test dataset. Detailed in Table 4 are the configurations of hyperparameters that produced optimal outcomes for the four base learners.

Table 4. Hyperparameter setting

Model	Setting
Random Forest	The maximum depth is set at 7, with a minimum requirement of 1 sample per leaf node, a minimum of 2 samples for node division, and a total of 50 trees in the ensemble
MLP	It involved utilizing the default hidden layer size of 100, employing the relu activation function. For weight optimization, the chosen solver was adam, while the penalty and learning rate were kept at their default settings.
KNN	The configuration includes a choice of 7 neighbors and the specification of distance weighting with p values of 1 and 2.
LR	Solver=liblinear, Penalty=L2, tolerance convergence set to 1e-4 and the maximum number of iterations to converge is set to 100

DT	It involved using the "gini" criterion for splitting nodes, choosing the "best" splitting strategy, not limiting the maximum depth of the tree and setting a minimum of 1 sample per leaf.
SVM	The settings include using the radial basis function (rbf) as the kernel function, setting gamma to 'scale,' and using a penalty coefficient (C) of 1

Table 5. The evaluation results of the base and stacked model before and after balancing

	Before SMOTE application				After SMOTE application			
Model	Accuracy	Precision	F1	ROC	Accuracy	Precision	F1	ROC
Stacked	0.83	0.83	0.83	0.96	0.98	0.98	0.98	1.00
RF	0.83	0.84	0.83	0.95	0.96	0.97	0.96	1.00
KNN	0.88	0.94	0.89	0.95	0.87	0.89	0.87	0.97
SVM	0.75	0.64	0.69	0.96	0.89	0.92	0.89	1.00
MLP	0.71	0.80	0.75	0.90	0.85	0.87	0.85	0.97
DT	0.79	0.82	0.80	0.84	0.94	0.95	0.94	0.98
LR	0.75	0.81	0.77	0.92	0.80	0.81	0.80	0.98

Table 5 provides a comprehensive overview of the model performances on the test dataset. Notably, the stacked ensemble model emerges as the top performer, showcasing superior predictive capabilities, closely followed by the K-Nearest Neighbors (KNN) and Random Forest (RF) models. In contrast, the Multi-Layer Perceptron (MLP) model demonstrates the least favorable performance among the base learners. Due to the varying performance profiles of these base learners, the selection of the KNN model as the meta-learner for the stacking model in this research was a thoughtful choice, emphasizing its potential to significantly enhance the ensemble's predictive power and resilience.

Fig.4 offers a comprehensive insight into the stacked model's proficiency in predicting student performance across three distinct categories: "course requirements not met" (C), "pass" (P), and "Referred" (R). The outcomes are highly informative. In the "course requirements not met" segment (C), the model displayed exceptional accuracy, both pre- and post-data balancing. Prior to balancing, it accurately predicted all three instances in this category, registering a flawless record devoid of false positives or false negatives. Subsequent to data balancing, the model's prowess in this category remained undiminished, sustaining a perfect prediction streak. Transitioning to the "pass" grouping (P), the model initially rendered predictions for 18 students. While it correctly identified 15, it encountered challenges with three cases. Nevertheless, after the implementation of the SMOTE data balancing methodology, significant improvement became evident. The "Referred" (R) category showcased remarkable performance. Initially, the model made predictions for three students, correctly classifying two and making one erroneous prediction. Following data balancing, this category witnessed noteworthy progress, with the model achieving flawless predictions. The ROC curve's evaluation illuminated the model's prowess. Pre-SMOTE, the ROC curve exemplified a performance level of 96%, signifying a high true positive rate and a minimal false positive rate, as depicted in Fig 4(c). Post-SMOTE, the ROC curve scaled to an exceptional 100% performance level (refer to Fig 4(d). This highlighted the model's enhanced ability to differentiate between classes, particularly in the "pass" and "Referred" contexts. Furthermore, crucial metrics, including accuracy, F1 score, and precision, witnessed substantial enhancements. Across the board, values ascended from 83% to a commendable 98%. These findings underscore the vital role played by the SMOTE data balancing strategy in amplifying the model's predictive acumen and overall efficacy.

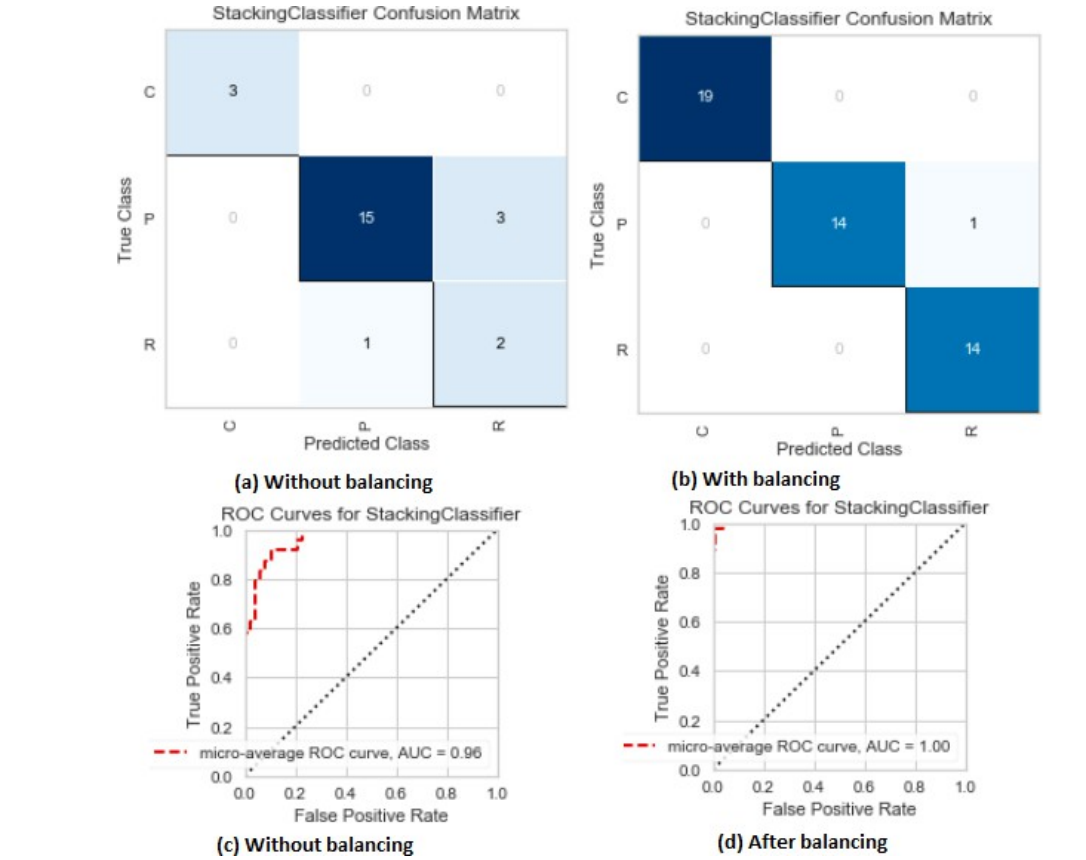

Fig. 4. Confusion matrices and ROC curves for stacked model, highlighting the outcomes both before and after the application of SMOTE for data balancing

Furthermore, Fig.5 presents the outcomes of student performance prediction across three distinct categories, specifically using the independent K-Nearest Neighbors (KNN) model, both pre- and post-application of the SMOTE technique. Before the data balancing procedure, it is noteworthy that all three students categorized as "C" were accurately classified. In a similar vein, all three students belonging to the "R" category received correct predictions. However, in the case of the "P" category (pass), while 15 students were correctly predicted, three errors were encountered. Following the data balancing process, a comparable pattern emerged, with all predictions being correct for groups "C" and "R." Slight variations were observed in the "P" category, where 11 predictions were accurate, while four false predictions occurred. Notably, there was a marginal improvement in the ROC Curve value, which increased from 95% to 97%. Conversely, there was a minor dip in performance metrics such as the F1 score, accuracy, and precision, as evidenced in Table 5. Accuracy declined from 88% to 87%, while precision saw a decrease from 94% to 89%. The F1 score also experienced a slight reduction, moving from 89% to 87%. These subtle fluctuations suggest that while the SMOTE approach contributed to enhanced class balance, it had a nuanced impact on certain predictive metrics for the KNN model.

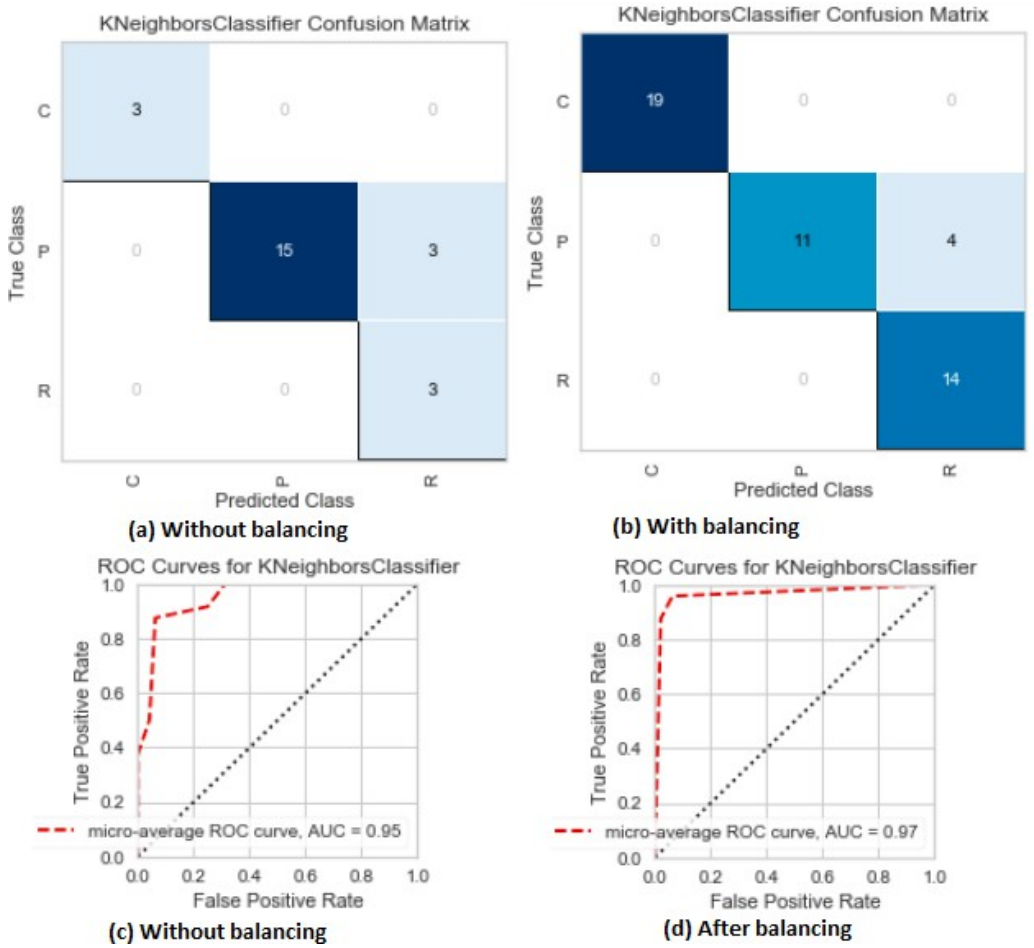

Fig. 5. Confusion matrices and ROC curves for KNN model, highlighting the outcomes both before and after the application of SMOTE for data balancing

Fig.6 provides an insightful depiction of student performance prediction using the standalone Decision Tree (DT) model across three distinct categories, mirroring the earlier analyses. Before implementing data balancing, all students in the "C" category received accurate predictions, which is consistent with the previous observations. However, there were minor discrepancies in the "P" and "R" categories, with three and two false predictions out of 18 and three instances, respectively. Intriguingly, after applying the data balancing technique, a noticeable enhancement in predictive accuracy was achieved. This was particularly evident in the remarkable reduction of false predictions across all categories. Specifically, each category experienced only one false prediction, reflecting a substantial improvement in predictive precision and reliability. The ROC curve's performance also witnessed significant progress, elevating from an initial 84% to an impressive 98%. This substantial increase in the ROC curve value signifies an enhanced ability of the model to distinguish between different classes, indicating an elevated true positive rate while maintaining a low false positive rate. This remarkable improvement in predictive performance can be attributed to the harmonization of class distribution through data balancing. By ensuring a more equitable representation of each class, the model became less skewed toward the majority

class, resulting in a more balanced and accurate prediction across all categories. This, in turn, translated into notable enhancements in key metrics such as the F1 score, accuracy, and precision. In Table 5, these improvements are elucidated further, with the F1 score witnessing an ascent from 80% to 94%, accuracy elevating from 79% to 94%, and precision improving from 82% to 95%. These metrics collectively demonstrate the profound impact of data balancing on refining the model's predictive prowess and reinforcing its reliability in classifying students' performance accurately.

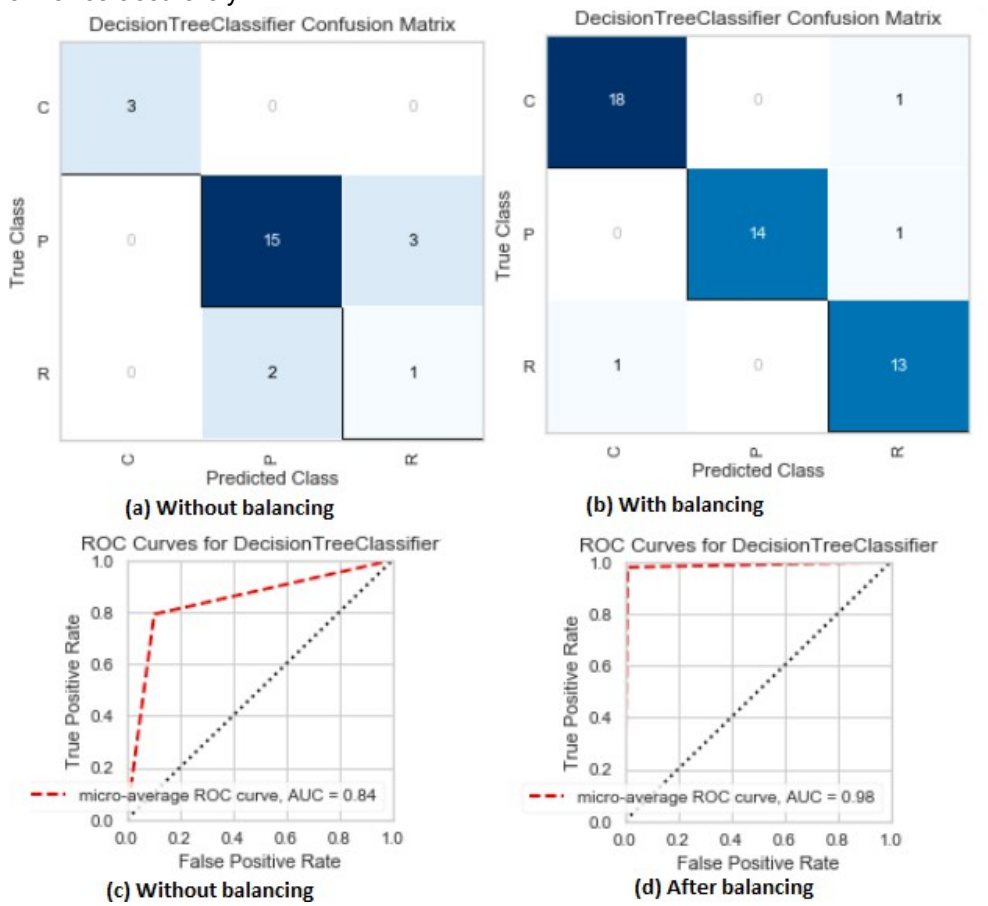

Fig. 6. Confusion matrices and ROC curves for DT model, highlighting the outcomes both before and after the application of SMOTE for data balancing

Fig.7 delves into the outcomes of predicting student performance using the Random Forest (RF) base model across three distinct categories, both before and after applying the SMOTE technique. The insights provided by this illustration are intriguing. Prior to implementing data balancing, there were noticeable discrepancies in the "P" and "C" categories. Specifically, the model made one erroneous prediction in the "P" group out of 18 instances and committed three errors in the "C" category out of a total of three students. Furthermore, the "R" category exhibited challenges, with two incorrect classifications out of three students. Post the application of the data balancing approach, a remarkable improvement was discernible. Notably, both the "C" and "R" minority groups benefited from this enhancement, as all predictions within these categories

proved to be correct. This underscores the significance of balancing the dataset, as it eradicated errors in the "C" and "R" categories. However, it's worth noting that a single incorrect prediction persisted within the "P" group. While this indicates room for further refinement, the overall improvement in predictive accuracy is undeniable. The sharp increase in the F1 score, accuracy, and precision metrics provides robust evidence of this enhancement. Specifically, these metrics surged from initial values of 83%, 83%, and 84% to impressive levels of 96%, 96%, and 97%, respectively. The significant boost in these metrics can be attributed to the harmonization of class distribution achieved by the SMOTE technique. By equalizing the representation of each class, the model's predictions became more balanced and precise, significantly reducing errors. This balancing act is also reflected in the remarkable rise of the ROC Curve performance, escalating from an initial 95% to a perfect 100%. This heightened ROC Curve value underscoring the model's newfound capability to effectively differentiate between classes, particularly in the "pass" and "course requirements not met" categories, while maintaining an exceptionally low false positive rate.

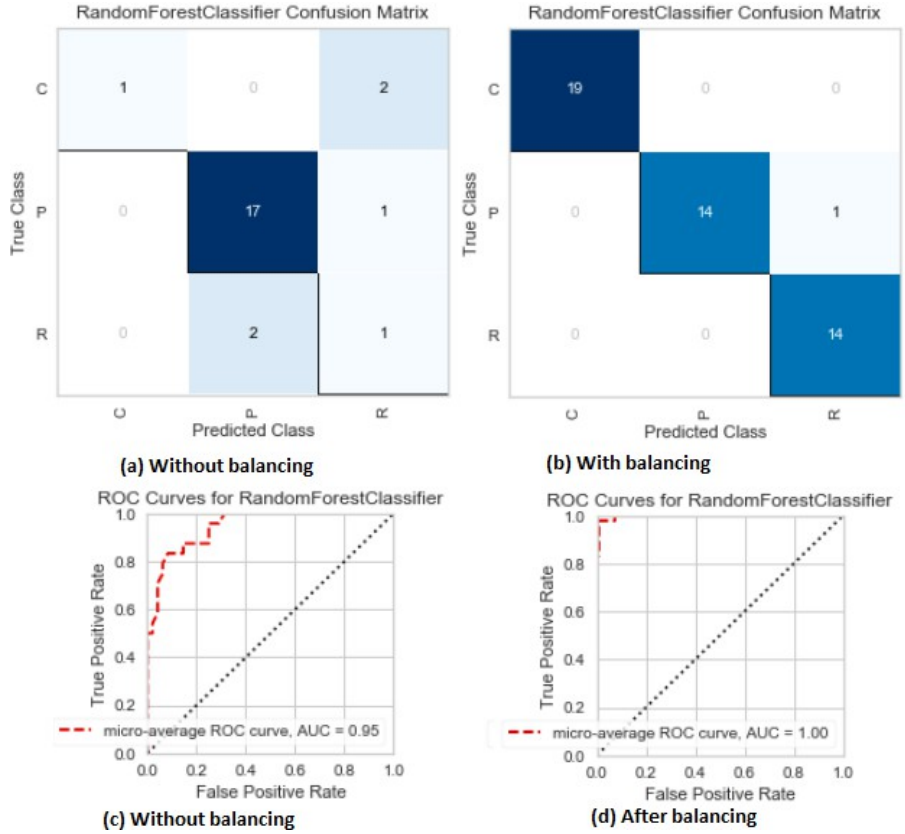

Fig.7. Confusion matrices and ROC curves for RF model, highlighting the outcomes both before and after the application of SMOTE for data balancing

Fig.8 offers a comprehensive view of the SVM-based predictive model's performance in forecasting student performance across three different categories. This assessment encompasses both pre-balancing and post-balancing scenarios, yielding intriguing insights. Before applying the data balancing technique, the model demonstrated noteworthy precision in the "C" (course requirements not met) and "P" (pass) categories. In these groups, all student

predictions were accurate, indicating an impressive performance. However, the same could not be said for the "R" (referred) category, where none of the predictions were correct. This deficiency underscored the challenge of handling imbalanced data. Conversely, after implementing data balancing, a discernible transformation took place. Notably, the minority classes, specifically the "C" and "R" groups, benefited from this adjustment, as all predictions within these categories proved to be correct. This marked improvement highlights the significance of data balancing in addressing the inherent challenges associated with imbalanced datasets. In contrast, the "P" group witnessed four erroneous predictions out of 15 students, indicating a slight decrease in precision. However, it's essential to consider the broader perspective. The data balancing technique significantly improved overall performance, as evidenced by the substantial surge in key metrics. When comparing the pre-balancing and post-balancing scenarios, an unmistakable trend emerges. The F1 score, accuracy, and precision metrics experienced remarkable growth. Specifically, these metrics catapulted from initial values of 69%, 75%, and 81% to highly commendable levels of 89%, 89%, and 92%, respectively. This transformation underlines the effectiveness of data balancing in refining predictive accuracy and precision. Furthermore, the ROC Curve, a crucial indicator of a model's discriminatory ability, exhibited notable improvement. Pre-balancing, the ROC curve registered a respectable 96%. However, post-balancing, this performance metric soared to a perfect 100%. This signifies the model's enhanced capacity to effectively distinguish between classes, particularly in the "pass" and "course requirements not met" categories, while maintaining an impressively low false positive rate.

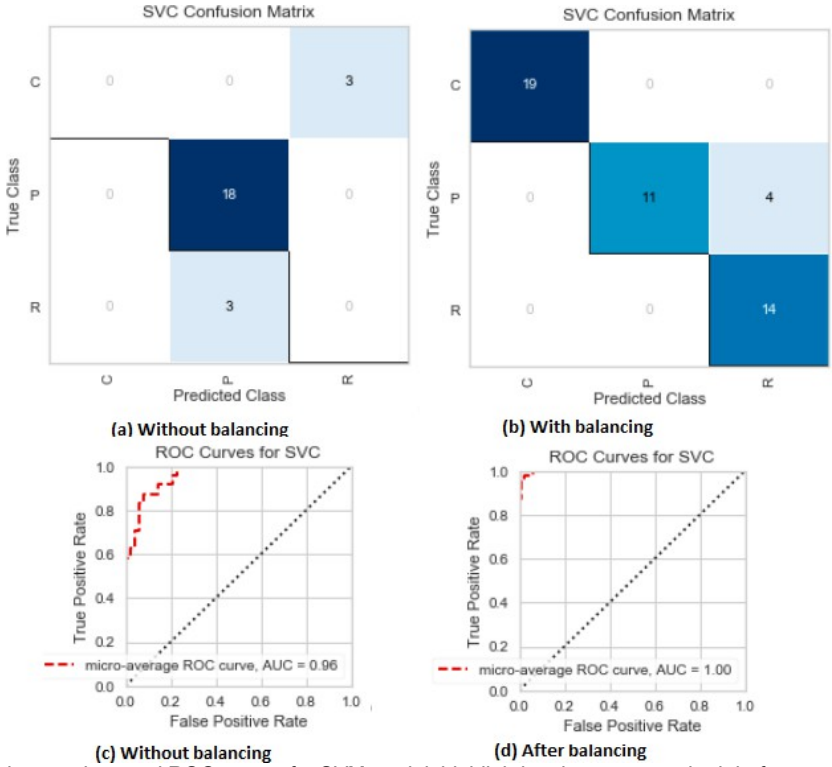

Fig.8. Confusion matrices and ROC curves for SVM model, highlighting the outcomes both before and after the application of SMOTE for data balancing

Furthermore, Logistic regression serves as a versatile statistical technique, applicable not only to binary classification but also to multiclass classification challenges. As depicted in Fig.9, the outcomes of logistic regression are of particular interest. Before applying data balancing, the logistic regression model exhibited a minor misclassification, erroneously predicting one sample within the "course requirements not met" (C) category out of three. Within the "pass" (P) category, which comprised 18 samples, 15 were accurately classified, while three were misclassified. In the "Referred" (R) category, the model accurately predicted only one out of three samples. Following the implementation of the SMOTE approach to address data imbalance, notable improvements emerged. In the "course requirements not met" (C) category, all samples were correctly predicted, signifying a substantial enhancement in model performance. However, within the "pass" (P) category, one misclassification persisted among the 15 samples, showcasing a slight improvement. The most remarkable progress was observed in the "Referred" (R) category, with only two out of 14 samples being incorrectly classified. These enhancements were accompanied by notable increases in key performance metrics. The F1 score, accuracy, and precision all exhibited marked improvements, rising from initial values of 77%, 75%, and 81% to new levels of 80%, 80%, and 81%, respectively. Additionally, the ROC curve demonstrated a notable ascent, surging from an initial 92% to a commendable 98%. These findings underscore the effectiveness of the SMOTE data balancing technique in enhancing the logistic regression model's accuracy, precision, and overall predictive capabilities. Lastly, Fig.10 provides insights into the performance of the Multilayer Perceptron (MLP) model in predicting student performance. Before the application of data balancing through SMOTE, some misclassifications were evident. Notably, one misclassification occurred within the "course requirements not met" (C) group, where one sample was incorrectly classified. In the "pass" (P) group, which encompassed 18 samples, three were inaccurately classified. Additionally, the "Referred" (R) group saw two samples out of three being incorrectly classified. However, a substantial improvement was witnessed after the implementation of the SMOTE technique. Specifically, within the "course requirements not met" (C) category, all samples were correctly predicted, showcasing a significant enhancement in predictive accuracy. For the "pass" (P) category, the improvement was noteworthy, with 13 out of 15 samples being accurately classified and only two misclassifications remaining. Similarly, within the "Referred" (R) category, the misclassifications reduced significantly, with only one out of 14 samples being incorrectly classified. These improvements extended to crucial performance metrics.

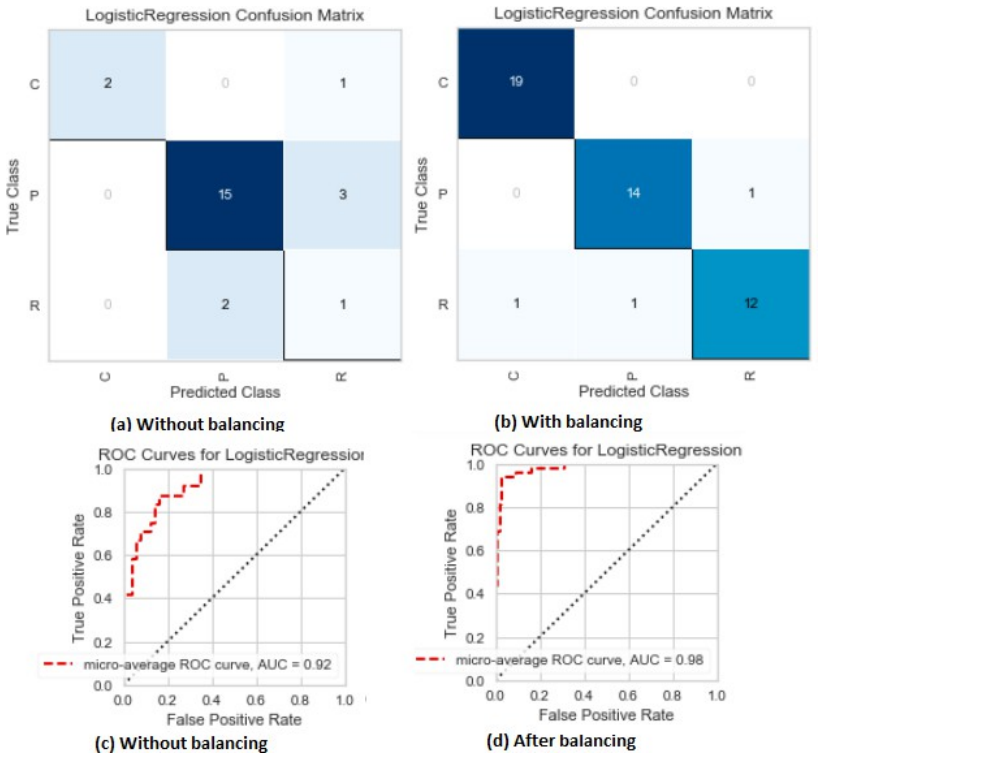

Fig.9. Confusion matrices and ROC curves for LR model, highlighting the outcomes both before and after the application of SMOTE for data balancing

The F1 score, accuracy, and precision exhibited substantial increases, rising from their initial values of 75%, 71%, and 80% to new levels of 85%, 85%, and 87%, respectively. Furthermore, the ROC curve, a pivotal indicator of model discrimination and performance, displayed remarkable progress. It ascended from its initial level of 90% to an impressive 97%, signifying the model's enhanced ability to distinguish between different student performance categories. These findings underscore the efficacy of the SMOTE data balancing technique in elevating the MLP model's predictive accuracy, precision, and overall performance.

Model comparisons

Prior to data balancing, the base models displayed a range of performance levels, each with its own strengths and weaknesses. For instance, the K-Nearest Neighbors (KNN) model exhibited commendable accuracy and precision in certain categories but grappled with false predictions in others. Similarly, Decision Trees (DT) showcased a comparable pattern, demonstrating misclassifications in specific categories. Random Forest (RF) encountered difficulties in the "pass" category, where misclassifications occurred, but it showcased a significant turnaround after data balancing. Support Vector Machines (SVM) faced particular challenges in correctly categorizing students in the "Referred" group before the introduction of data balancing techniques. Multilayer Perceptron (MLP) exhibited misclassifications across categories before data balancing, but its performance notably improved afterward. Logistic Regression (LR) struggled with some misclassifications, notably within the "pass" category. Nonetheless, the application of the SMOTE data balancing technique wrought transformative improvements across all models. Notably, the stacked model emerged as a standout performer, boasting remarkable

accuracy, ROC curve performance, precision, and F1 score. The stacked model harnessed the enhanced predictive capabilities of the base models, resulting in a synergistic and robust predictive engine, as visually depicted in Fig.11. In summation, the stacked model surpassed the individual base models across multiple performance metrics. Its accuracy, precision, and F1 score exhibited substantial gains, underlining its prowess in predicting student performance with accuracy and reliability. The ROC curve, a vital indicator of the model's discriminatory capabilities, testified to the stacked model's ability to effectively differentiate between diverse student performance categories. This suggests that the stacked model excelled not only in correctly identifying students' performance outcomes but also in minimizing false positive and false negative predictions, a crucial aspect of predictive modeling. It's imperative to acknowledge the instrumental role played by data balancing techniques, particularly SMOTE, in enhancing the performance of all models. By rectifying the class imbalances present in the dataset, SMOTE facilitated a more equitable representation of each class.

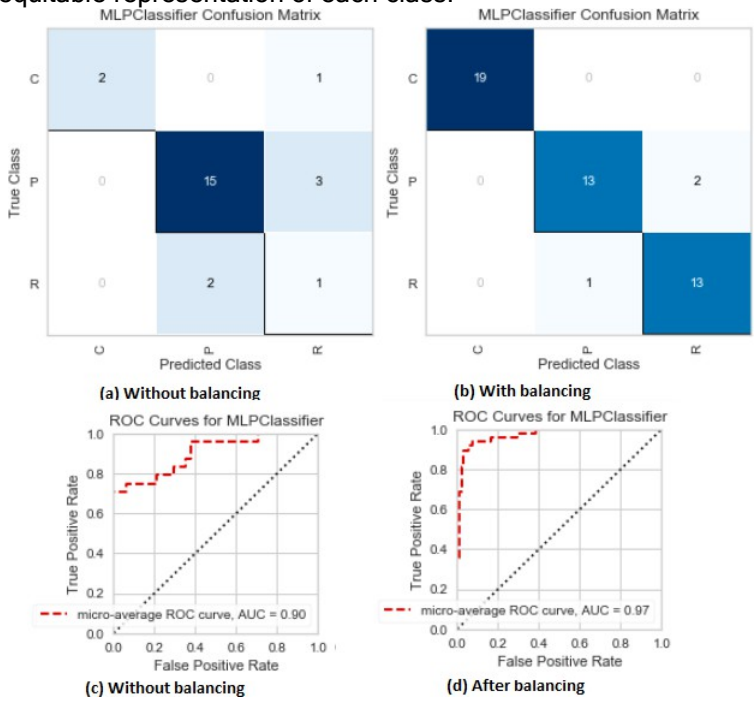

Fig.10. Confusion matrices and ROC curves for MLP model, highlighting the outcomes both before and after the application of SMOTE for data balancing

This leveling of the playing field enabled all models, including the stacked model, to make predictions that were more balanced and precise, significantly mitigating errors. In a nutshell, the stacked model, fortified by data balancing techniques, emerged as the most formidable predictor of student performance. Its ability to harness the collective strength of the base models and deliver accurate, balanced, and reliable predictions underscores its efficacy in practical applications.

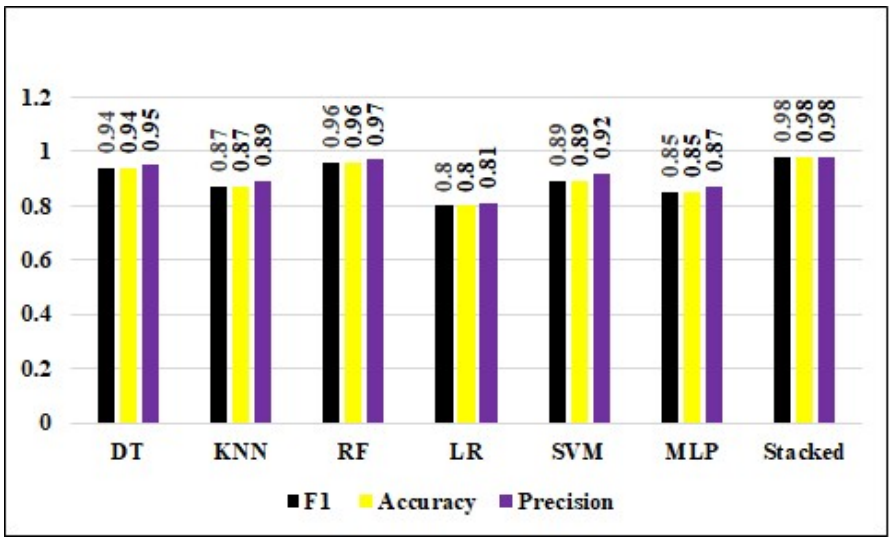

Fig. 11. A comprehensive comparison of the models' performance after the application of the SMOTE data balancing technique

A comparative analysis of the proposed stacking approach versus prior studies

Several prior studies evaluated the performance of single machine learning models such as Decision Trees, Naïve Bayes, Neural Networks, and Support Vector Machines. For example, Rahman et al. (2017) applied information gain and wrapper feature selection methods in conjunction with Naïve Bayes, Decision Tree, and Artificial Neural Networks, achieving an accuracy of 79.37%. Kumar & Pal (2011) explored Decision Trees, Neural Networks, and K-Nearest Neighbors to predict student results, with Decision Trees emerging as the most accurate. These single models demonstrated reasonable predictive power. However, in contrast to these studies, our research leverages an ensemble stacking model, as highlighted by Mastour et al. (2023), which consistently outperforms these single models. In addition, Mastour et al. (2023) conducted a comprehensive study comparing classical models like Logistic Regression (LR), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN) with ensemble models, including Voting, Bagging (BG), Random Forests (RF), Adaptive Boosting (ADA), eXtreme Gradient Boosting (XGB), and Stacking. Their findings underscored the superiority of ensemble machine learning models, signaling a shift towards leveraging the collective intelligence of multiple base models. Our study takes this a step further by introducing a stacked ensemble approach, as discussed in depth, which combines Decision Trees, Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Machines, and Multi-Layer Perceptron within a unified framework.

Many previous studies encountered class imbalance issues within educational datasets. González-Brambila et al. (2018) sought to predict the academic performance of engineering students using Decision Trees but achieved only 63.19% accuracy, highlighting the challenges of class imbalance. Our research addresses this challenge by integrating the Synthetic Minority Over-Sampling Technique (SMOTE). This technique generates synthetic samples for the minority class, ensuring a balanced dataset and reducing potential biases in predictions. Zaffar et al. (2017) delved into the impact of feature selection algorithms on student performance analysis, discovering that principal component analysis combined with the random forest classifier yielded

the best results. However, our study introduces the concept of feature selection using eXtreme Gradient Boosting (XGB), building on these previous insights. Our use of XGB highlights a sophisticated approach to identifying and extracting relevant features, contributing to enhanced model efficiency and interpretability. Our research evaluates model performance using a comprehensive set of metrics, including Accuracy, Precision, ROC curve, and F1-score. This comprehensive assessment provides a more nuanced understanding of our framework's capabilities compared to studies that rely on a limited set of metrics, as exemplified in various prior studies. Recent research by Mulyana et al. (2023) achieved an impressive 89% accuracy in classifying successful and dropout students using Random Forest. However, our study surpasses this achievement by achieving a remarkable 98% accuracy when dealing with balanced data, as vividly depicted in Table 5 and Fig.11. This outcome solidifies the potential real-world impact of our predictive model in enhancing student outcomes and resource allocation, as exemplified by Sawangarreerak&Thanathamathee (2020). In summary, our research builds upon the foundations laid by previous studies in the field of predicting student academic performance. It introduces an innovative stacked ensemble model, as highlighted by Mastour et al. (2023), which surpasses the predictive power of single models and demonstrates robustness in handling class imbalance. The use of XGBoost for feature selection and comprehensive model evaluation metrics further enhances the contributions of our study. The achievement of 98% accuracy with balanced data, as compared to prior studies, solidifies our model's potential for practical implementation in educational contexts, marking a significant advancement in the field.

Practical and theoretical contributions

The article makes substantial contributions to the realm of education and predictive modeling, encompassing both theoretical and practical dimensions. Firstly, it conducts an extensive theoretical evaluation of various predictive models, including Decision Trees, k-Nearest Neighbors, Random Forest, Logistic Regression, Support Vector Machines, and Multi-Layer Perceptron. This in-depth analysis sheds light on the strengths and weaknesses of these models when applied to the task of predicting student performance. Such theoretical insights provide a foundational understanding of model suitability within educational contexts. Additionally, the article introduces a crucial theoretical concept: feature selection using XGBoost (XGB), a formidable machine learning framework. This contribution underscores the critical role of selecting pertinent features for accurate student performance modeling. The utilization of XGB for feature selection showcases a sophisticated approach to identifying and extracting the most influential attributes from the dataset. This not only elevates the model's efficiency but also imparts invaluable insights into the key determinants of student success.

Furthermore, the incorporation of XGB-based feature selection into the methodology enhances the practical robustness of the predictive model. This process ensures that only the most informative variables are considered, diminishing noise and bolstering the model's generalization capabilities. This practical application of feature selection streamlines the creation of an efficient and effective predictive framework, which is paramount for practical implementation within educational settings. The article doesn't stop at feature selection but goes on to introduce the concept of a stacking ensemble approach, which combines the predictive capabilities of multiple base models. This theoretical framework elevates the accuracy and reliability of student performance prediction, showcasing the potential advantages of ensemble learning within educational contexts.

Moreover, the article extends its contributions into practical domains by offering insights into data preprocessing techniques. It elucidates the importance of techniques such as one-hot encoding and the Synthetic Minority Over-Sampling Technique (SMOTE) for handling imbalanced

datasets, thereby enhancing the robustness and accuracy of the predictive models in real-world educational applications. Furthermore, the article demonstrates the practical application of machine learning and data mining techniques in the education sector. It elucidates how these methods can be effectively employed to predict student performance, identify at-risk students early on, and provide targeted academic support. Lastly, the article underscores the real-world impact of predictive modeling in education. Achieving an impressive 98% accuracy in predicting student performance with balanced data, it implies that educational institutions can utilize these methods to enhance student outcomes, allocate resources more efficiently, and offer timely support to students in need.

In a nutshell, this article bridges the theoretical and practical realms in the field of education, delivering valuable insights for both researchers and educational practitioners. Its contributions span model evaluation, feature selection, ensemble learning, data preprocessing, practical implementation, and tangible outcomes, collectively enriching the landscape of predictive modeling in education.

Conclusion

In our extensive investigation, we compared six individual machine learning models to a stacking model to forecast the performance of higher education students, utilizing both balanced and imbalanced datasets. Across various performance metrics such as Accuracy, the ROC curve, Precision, and the F1-Score, the stacking model with data oversampling via SMOTE consistently outperformed all others. This establishes it as the most efficient method for predicting student performance, underscoring the vital role of machine learning in education, particularly when dealing with complex data requirements and the need for precision.

In summary, our ensemble stacking model, enhanced by SMOTE data rebalancing, is a powerful tool for predicting higher education student performance. It validates the effectiveness of this ensemble approach, especially with imbalanced datasets, offering educators a valuable tool to identify and support students who may need additional academic assistance. With a remarkable 98% accuracy, this approach has the potential to significantly impact educational interventions and enhance student academic achievements, bridging the gap between data science and education, and highlighting machine learning's transformative role in academia.

Acknowledgment

Appreciations for all the students who participated in the research.

Ethical approval

I affirm that this manuscript is entirely my own work, created independently. It does not incorporate any research findings previously published or authored by others. I am the sole author of this manuscript, and I take full legal responsibility for the accuracy of this statement.

Conflict of interest

The corresponding author confirms that there are no conflicts of interest associated with this manuscript.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

Akgün, E. (2020). Science Mapping Research on Educational Data Mining: A Bibliometric Review of International Publications. Future Visions Journal, 4(14), 1-17. https://doi.org/10.29345/futvis.160

Alamgir, Z., Akram, H., Karim, S., & Wali, A. (2023). Enhancing Student Performance Prediction via Educational Data Mining on Academic data. Informatics in Education. https://doi.org/10.15388/infedu.2024.04

Albreiki, B., Zaki, N., &Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. Education Sciences, 11(9), 552. https://doi.org/10.3390/educsci11090552

Atif, A. (2013). Learning Analytics in Higher Education: A Summary of Tools and Approaches. Learning & Technology Library (LearnTechLib).

Aulakh, K., Roul, R. K., & Kaushal, M. (2023). E-learning enhancement through educational data mining with Covid-19 outbreak period in backdrop: A review. International Journal of Educational Development, 101, 102814. https://doi.org/10.1016/j.ijedudev.2023.102814

Banda, L. O. L., Liu, J., Banda, J. T., & Zhou, W. (2023). Impact of ethnic identity and geographical home location on student academic performance. Heliyon, 9(6), e16767. https://doi.org/10.1016/j.heliyon.2023.e16767

Begum, S., &Padmannavar, S. S. (2023). Student Performance Analysis using Bayesian Optimized Random Forest Classifier and KNN. International Journal of Engineering Trends and Technology, 71(5), 132-140. https://doi.org/10.14445/22315381/ijett-v71i5p213

Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Cheng, S. T., & Kaplowitz, S. A. (2016). Family economic status, cultural capital, and academic achievement: The case of Taiwan. International Journal of Educational Development, 49, 271-278. https://doi.org/10.1016/j.ijedudev.2016.04.002

Dervenis, C., Kyriatzis, V., Stoufis, S., &Fitsilis, P. (2022). Predicting Students' Performance Using Machine Learning Algorithms. Proceedings of the 6th International Conference on Algorithms, Computing and Systems. https://doi.org/10.1145/3564982.3564990

Desjardins, S., & Grandbois, M. (2022). Sleep parameters associated with university students' grade point average and dissatisfaction with academic performance. Sleep Epidemiology, 2, 100038. https://doi.org/10.1016/j.sleepe.2022.100038

González-Brambila, S. B., Sánchez-Guerrero, L., Ardón-Pulido, I., Figueroa-González, J., & González-Beltrán, B. (2018). Predicting Academic Performance of Engineering Students After Approving a Mathematics Leveling Course using Decision Trees. Research in Computing Science, 147(12), 171-181. https://doi.org/10.13053/rcs-147-12-16

H. Alamri, L., S. Almuslim, R., S. Alotibi, M., K. Alkadi, D., Ullah Khan, I., & Aslam, N. (2020). Predicting Student Academic Performance using Support Vector Machine and Random Forest. 2020 3rd International Conference on Education Technology Management. https://doi.org/10.1145/3446590.3446607

Hanaysha, J. R., Shriedeh, F. B., &In'airat, M. (2023). Impact of classroom environment, teacher competency, information and communication technology resources, and university facilities on student engagement and academic performance. International Journal of Information Management Data Insights, 3(2), 100188. https://doi.org/10.1016/j.jjimei.2023.100188

Haryawan, C., &Sebatubun, M. M. (2020). Implementation of multilayer perceptron for student failure prediction. JUTI: JurnalIlmiahTeknologiInformasi, 18(2), 125. https://doi.org/10.12962/j24068535.v18i2.a990

Howard, W.R. (2007), "Pattern Recognition and Machine Learning", Kybernetes, Vol. 36 No. 2, pp. 275-275. https://doi.org/10.1108/03684920710743466

Kim, Y., Li, T., Kim, H. K., Oh, W., & Wang, Z. (2023). Socioeconomic status and school adjustment trajectories among academically at-risk students: The mediating role of parental school-based involvement. Journal of Applied Developmental Psychology, 87, 101561. https://doi.org/10.1016/j.appdev.2023.101561

King, R. D., Orhobor, O. I., & Taylor, C. C. (2021). Cross-validation is safe to use. Nature Machine Intelligence, 3(4), 276-276. https://doi.org/10.1038/s42256-021-00332-z

Kleinkorres, R., Stang-Rabrig, J., & McElvany, N. (2023). Comparing parental and school pressure in terms of their relations with students' well-being. Learning and Individual Differences, 104, 102288. https://doi.org/10.1016/j.lindif.2023.102288

Kumar, B., & Pal, S. (2011). Mining Educational Data to Analyze Students Performance. International Journal of Advanced Computer Science and Applications, 2(6). https://doi.org/10.14569/ijacsa.2011.020609

Lagman, A. C. (2015). Embedding Logistic Regression Model in Decision Support Software for Student Graduation Prediction. Proceedings Journal of Interdisciplinary Research, 2, 104-110. https://doi.org/10.21016/irrc.2015.au05ef81o

Lang, C., Siemens, G., Wise, A., &Gasevic, D. (Eds.). (2017). Handbook of Learning Analytics. https://doi.org/10.18608/hla17

Lehti, H. (2023). Parental Unemployment and Children's Educational Outcomes – A Literature Review. International Encyclopedia of Education (Fourth Edition), 118-127. https://doi.org/10.1016/b978-0-12-818630-5.01019-8

Li, Z., & Qiu, Z. (2018). How Does Family Background Affect Children's Educational Achievement? Evidence from Contemporary China. The Journal of Chinese Sociology, 5(1). https://doi.org/10.1186/s40711-018-0083-8

Liu, R. (2015). Variations in Learning Rate: Student Classification Based on Systematic Residual Error Patterns across Practice Opportunities. Retrieved from https://api.semanticscholar.org/CorpusID:11413105

Liu, R., & Hannum, E. (2023). Parental Absence and Student Academic Performance in Cross-National Perspective: Heterogeneous Forms of Family Separation and the Buffering Possibilities of Grandparents. International Journal of Educational Development, 103, 102898. https://doi.org/10.1016/j.ijedudev.2023.102898

Lucey, S., & Grydaki, M. (2022). University Attendance and Academic Performance: Encouraging Student Engagement. Scottish Journal of Political Economy, 70(2), 180-199. https://doi.org/10.1111/sjpe.12334

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2012). Predicting Student Failure at School Using Genetic Programming and Different Data Mining Approaches with High Dimensional and Imbalanced Data. Applied Intelligence, 38(3), 315-330. https://doi.org/10.1007/s10489-012-0374-8

Mastour, H., Dehghani, T., Moradi, E., & Eslami, S. (2023). Early Prediction of Medical Students' Performance in High-Stakes Examinations Using Machine Learning Approaches. Heliyon, 9(7), e18248. https://doi.org/10.1016/j.heliyon.2023.e18248

Mativo, J. M., & Huang, S. (2014). Prediction of Students' Academic Performance: Adapt a Methodology of Predictive Modeling for a Small Sample Size. 2014 IEEE Frontiers in Education Conference (FIE) Proceedings. https://doi.org/10.1109/fie.2014.7044287

Mengash, H. A. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. IEEE Access, 8, 55462-55470. https://doi.org/10.1109/access.2020.2981905

Mulyana, A. F., Puspita, W., &Jumanto, J. (2023). Increased Accuracy in Predicting Student Academic Performance Using Random Forest Classifier. Journal of Student Research Exploration, 1(2), 94-103. https://doi.org/10.52465/josre.v1i2.169

Nawang, H., Makhtar, M., & Hamza, W. M. A. F. W. (2021). A Systematic Literature Review on Student Performance Predictions. International Journal of Advanced Technology and Engineering Exploration, 8(84). https://doi.org/10.19101/ijatee.2021.874521

Nugroho, A., Riady, O. R., Calvin, A., &Suhartono, D. (2020). Identification of Student Academic Performance Using the KNN Algorithm. Engineering, MAthematics and Computer Science (EMACS) Journal, 2(3), 115-122. https://doi.org/10.21512/emacsjournal.v2i3.6537

Priscilla, C. V., & Prabha, D. P. (2021). A two-phase feature selection technique using mutual information and XGB-RFE for credit card fraud detection. International Journal of Advanced Technology and Engineering Exploration, 8(85). https://doi.org/10.19101/ijatee.2021.874615

QIN, F., ZHU, L. Q., CHENG, Z. K., & ZHANG, Q. (2017). Research on Student Performance Evaluation Based on Random Forest. DEStech Transactions on Engineering and Technology Research. https://doi.org/10.12783/dtetr/eeta2017/7762

Rabelo, A., Rodrigues, M. W., Nobre, C., Isotani, S., & Zárate, L. (2023). Educational data mining and learning analytics: a review of educational management in e-learning. Information Discovery and Delivery. https://doi.org/10.1108/idd-10-2022-0099

Rahman, L., Setiawan, N. A., &Permjereari, A. E. (2017). Feature selection methods in improving accuracy of classifying students' academic performance. 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE). https://doi.org/10.1109/icitisee.2017.8285509

Rajendran, S., Chamundeswari, S., & Sinha, A. A. (2022). Predicting the academic performance of middle- and high-school students using machine learning algorithms. Social Sciences & Humanities Open, 6(1), 100357. https://doi.org/10.1016/j.ssaho.2022.100357

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618. https://doi.org/10.1109/tsmcc.2010.2053532

Sawangarreerak, S., &Thanathamathee, P. (2020). Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression. Information, 11(11), 519. https://doi.org/10.3390/info11110519

Schalk, P. D., Wick, D. P., Turner, P. R., & Ramsdell, M. W. (2011). Predictive assessment of student performance for early strategic guidance. 2011 Frontiers in Education Conference (FIE). https://doi.org/10.1109/fie.2011.6143086

Tantoh, M. C. (2023). Parental Level of Education and its Implications of their Expectations towards their Children Academic Performance. International Journal of Psychology and Cognitive Education, 2(1), 1-20. https://doi.org/10.58425/ijpce.v2i1.109

Uylaş, N. (2018). Semi-Supervised Classification in Educational Data Mining: Students' Performance Case Study. International Journal of Computer Applications, 179(26), 13-17. https://doi.org/10.5120/ijca2018916549

Wang, Y., & Wang, Y. (2023). Exploring the relationship between educational ICT resources, student engagement, and academic performance: A multilevel structural equation analysis based on PISA 2018 data. Studies in Educational Evaluation, 79, 101308. https://doi.org/10.1016/j.stueduc.2023.101308

Xu, J., Han, Y., Marcu, D., & Van der Schaar, M. (2017). Progressive Prediction of Student Performance in College Programs. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). https://doi.org/10.1609/aaai.v31i1.10713

Yohannes, E., & Ahmed, S. (2018). Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study. International Journal of Computer Applications, 180(40), 39-47. https://doi.org/10.5120/ijca2018917057

Zaffar, M., Hashmani, M. A., & Savita, K. S. (2017). Performance analysis of feature selection algorithm for educational data mining. 2017 IEEE Conference on Big Data and Analytics (ICBDA). https://doi.org/10.1109/icbdaa.2017.8284099

Zhang, Y., & Liu, Y. (2021). The Research of Predicting Student's Academic Performance Based on Educational Data. 2021 5th International Conference on Computer Science and Artificial Intelligence. https://doi.org/10.1145/3507548.3507578