SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I12P114

Volume 11 Issue 12, 146-153, December 2024
© 2024 Seventh Sense Research Group®

Original Article

Enhancing Student Academic Performance Forecasting in Technical Education: A Cutting-edge Hybrid Fusion Method

K. Rajesh Kannan[1], K. T. Meena Abarna[1], S. Vairachilai[2]

[1]*Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India.*
[2]*School of Engineering and Information Technology, Sanskriti University, Uttar Pradesh, India.*

[1]*Corresponding Author : rajeshlpm88@gmail.com*

Received: 19 October 2024 Revised: 23 November 2024 Accepted: 11 December 2024 Published: 30 December 2024

Abstract - *Forecasting early-stage student performance within higher education is important to the academic community, offering a proactive framework to mitigate student attrition. However, gauging and prognosticating students' achievements in the Indian context are beset by formidable challenges due to the vast student populace and the deeply entrenched educational system. Each institution in India employs distinct criteria to assess student progress, lacking a standardized mechanism to oversee and appraise developmental trajectories. The past decade has witnessed diverse exploration of machine learning methodologies in educational research. Nonetheless, student performance prediction grapples with substantial obstacles, particularly when contending with imbalanced datasets. This research work adopts a dual-phase methodology to grapple with this quandary. Initially, conventional classification algorithms are deployed on a dataset encompassing the academic journeys of 4424 students. Subsequently, innovative hybrid machine learning (ML) algorithms are harnessed to yield more refined prognostications. The outcomes furnished by the proposed model furnish a platform for informed early decision-making of the advancement of higher education institutions. This streamlines the prediction of students' performance and empowers the educational domain to tackle these challenges with a more robust and insightful approach.*

Keywords - *Academic performance, Cross-validation, Artificial Intelligence, Hybrid ML algorithms.*

1. Introduction

Its students' academic accomplishments significantly shape any educational institution's achievement. Within the educational journey across various levels, students grapple with two prominent challenges: high rates of academic failure and dropout incidents in diverse courses. Nurturing top-tier university graduates today is arduous, and upholding robust student academic excellence remains pivotal. Individuals with subpar academic performance are more prone to delayed graduation or abandoning their college pursuits. The global march towards unlocking human potential fundamentally relies on education. Traditional and generic methods constitute the framework for assessing student progress within the Indian educational structure. Universities prioritize scholastic triumphs and extracurricular engagements as metrics for appraising student proficiency. Within India's educational landscape, institutions underscore the significance of students' academic records in determining their eligibility for higher education admission. Early anticipation of student performance facilitates proactive interventions and the implementation of measures to enhance their academic standing. By pinpointing the underlying issue—financial constraints, health concerns, or other factors—such foresight permits effective management of these predictions [1].

Artificial intelligence and diverse machine learning algorithms have found application in advanced domains like virtual reality, visual analysis, speech recognition, and knowledge exploration within the academic sector. Knowledge exploration can be accomplished by employing various machine-learning methods, including classification. Among the prominent areas of research, predicting student performance stands out as a significant endeavor, aiming to unearth valuable patterns that can facilitate early decision-making for educational institutions [2]. A pivotal factor that could contribute to enhancing a student's academic achievements revolves around the capability to predict their academic grades. Previous studies have illuminated that different machine learning methodologies effectively forecast student academic performance. Nonetheless, the quest for analogous research addressing the challenge of imbalanced classification in predicting students' grades is notably intricate [3].

2. Related Works

A comprehensive review of existing literature within this academic domain is carried out to identify potential research voids concerning the prediction of student progress. The primary outcomes of this investigation are delineated in this section.

Forecasting academic achievement and predicting student attrition holds a position of significant importance within the higher education sector. Numerous scholars have explored this realm, employing conventional Machine Learning classification algorithms.

Academic factors such as grades secured at the intermediate level and the Grade Point Average (GPA) and Cumulative Grade Point Average (CGPA)at the first-year course completion are the most often utilized factors used as projected variables for evaluating and forecasting students' academic accomplishment at the higher education level, according to our literature study on predicting students' academic accomplishment progress using various machine learning approaches [4-5].

The examined research endeavors are closely aligned with our study. As cited in [6], this research analyzes several classification algorithms and concludes that feature selection in any data set plays a critical role in performance prediction. Nevertheless, the development of a predictive model for imbalanced datasets within the academic domain remains largely unexplored.

In this regard, a study from [7] used several SMOTE techniques for balancing, such as Borderline SMOTE, SMOTE Tomek, SVM SMOTE and SMOTE ENN, to improve the prediction and dropout accuracy. Despite the utilization of various classification algorithms such as XG-Boost, K-Nearest-Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM), XG-Boost outperforms the rest with an accuracy of 93.29% in this study.

Coping with the increasing volume of data within educational institutions to facilitate optimal decision-making poses a significant challenge. Because of this, researchers highlight the many difficulties in obtaining, analyzing, and utilizing data in education [9]. These challenges could stem from issues related to methodology, data protection, training, and more. In this research work [8], video and data mining methods are also utilized in 2020 to forecast students' behavior.

The utilization of Random Forest yielded an accuracy of 88.3% in predicting outcomes across 772 instances. In this research work [10], Suresh et al. considered several factors, including educational records, parental education qualifications and financial status, student medical history, and student conduct.

Additionally, Nave Bayes was employed to compute the student attrition rate. Utilizing the AI-based Multi-Layered Perceptron (MLP) algorithm, this work [11] suggests a method to forecast students' academic achievement in the fundamentals of computer programming courses.

The investigation encompasses the analysis of interrelated factors, incorporating multiple variables to gauge a student's likelihood of achieving subpar performance in the introductory programming course. These factors encompass student activity logs and personal information accessible through the student learning management system, as well as grades attained during the learning process, including quizzes, assignments, midterms, and final exams, as well as other data collected through surveys.

3. System Model

The objective of the proposed work is to construct a system model by employing suitable feature selection methods and classification algorithms to enhance the forecasting measures. The approach undertaken in this study comprises four distinct phases: data acquisition, feature extraction and preprocessing, data representation, and classification for student categorization.

The initial phase focuses on gathering data and implementing feature engineering techniques. Phase II involves the implementation of stratified-K fold cross-validation to partition the data set into training and testing subsets.

Phase III involves using classification algorithms to establish a predictive model, while the final phase pertains to evaluation. Detailed depictions of the proposed methodology's architecture and algorithm are illustrated in Figures 1 and 4, respectively.

3.1. Data Collection

This research used a public dataset from several disjointed databases by the Polytechnic Institute of Portalegre, Portugal [12]. The dataset comprises 4482 instances, encompassing a diverse range of courses selected by students between 2008-09 and 2018-19.

The dataset includes academic performance records from the first two semesters and demographic, socio-economic, and academic trajectory details available at the point of enrollment. These pieces of information are harnessed to construct classification models that predict student achievements and dropout occurrences.

As the standard course duration concludes, the problem becomes a multiclass categorization task involving the graduate, dropout, and enrolled labels. More details can be found in [12] for a comprehensive dataset overview.

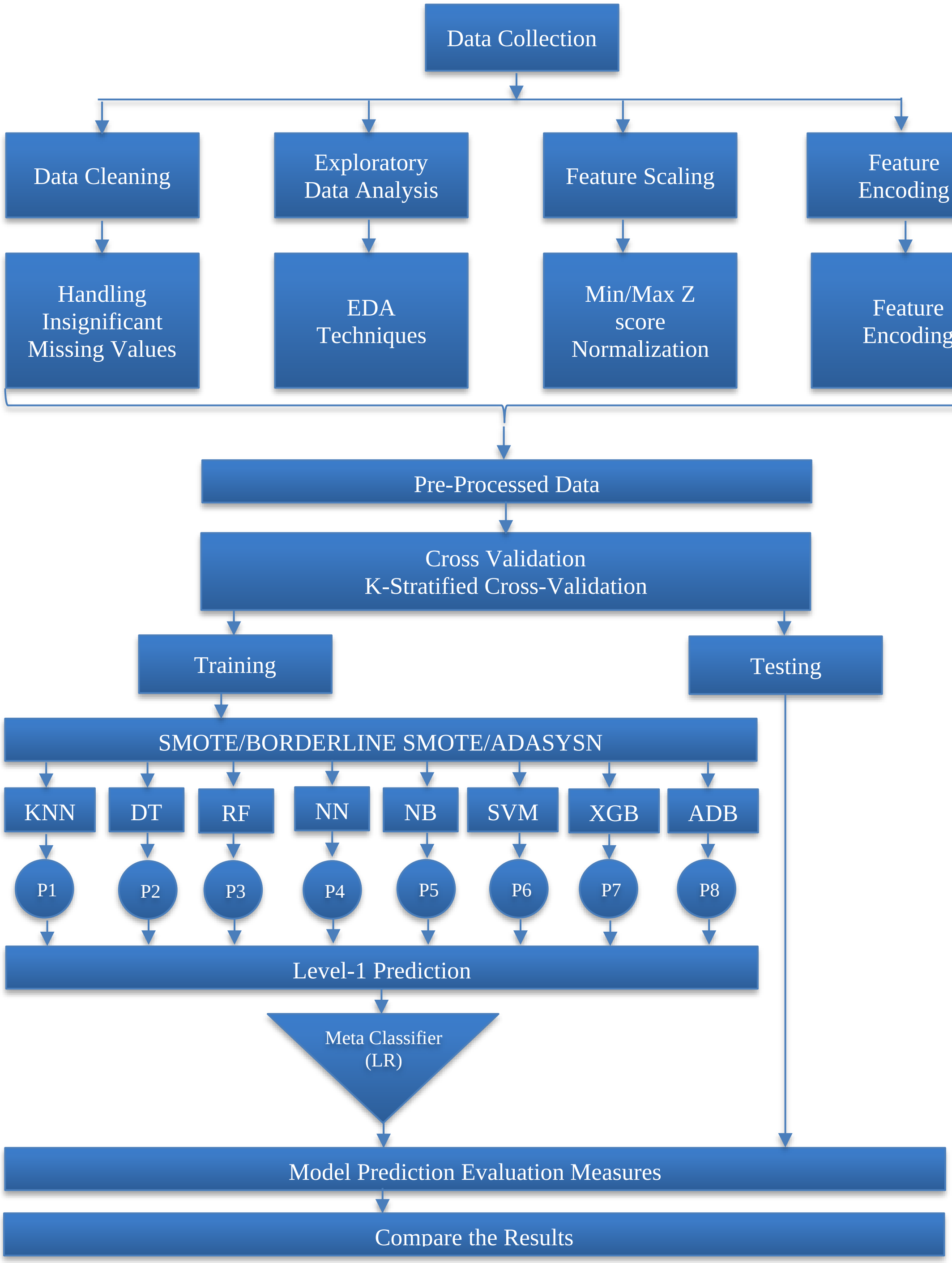

Fig. 1 Architecture of proposed model

3.2. Data Preprocessing and Feature Selection

The presence of missing values is a common challenge encountered in numerous real-world concrete datasets. These gaps have the potential to introduce bias into the results of machine learning (ML) models and/or diminish the overall accuracy of the model. Depending on the specific dataset approaches to addressing missing values encompass options such as eliminating corresponding rows or columns or substituting them with arbitrary values.

Machine learning algorithms cannot directly handle qualitative or categorical data when our features fall into this category. Consequently, it is necessary to convert such qualitative data into a numerical format before feeding it into the predictive model. One approach to achieve this is label encoding, wherein each distinct categorical variable is assigned a corresponding integer value. However, it is important to note that only numerical data was utilized for the present study.

Data scientists [14] examine and analyze data sets and epitomize their key properties using exploratory data analysis (EDA), which regularly employs data visualization techniques. It empowers researchers to analyze valuable patterns, identify inconsistencies, and validate assumptions by discerning strategies for refining data assets to enhance accuracy. Exploratory Data Analysis (EDA) is commonly employed to investigate potential data revelations beyond specific modeling or hypothesis-testing tasks. It facilitates a more comprehensive exploration of latent patterns among variables within the dataset, including their interrelationships. Furthermore, exploratory data analysis (EDA) aids in evaluating the appropriateness of selected statistical methods for data analysis. As depicted in Figure 2, visualisation techniques were utilized to analyze specific features in the study.

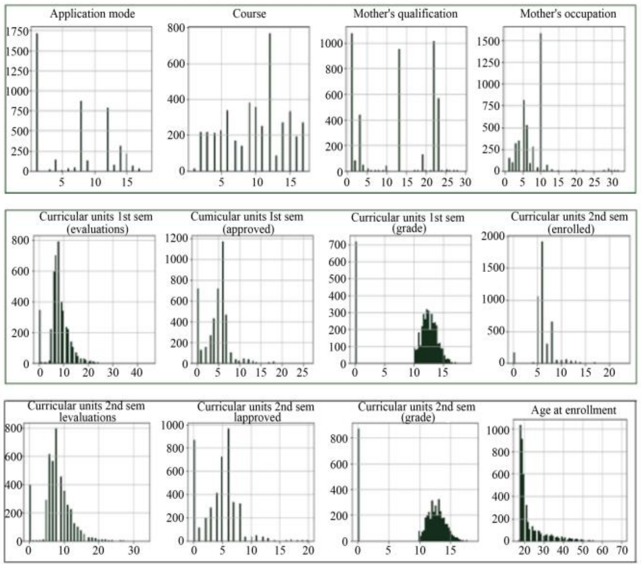

Fig. 2 Attribute analysis using visualization techniques

In machine learning, the data preparation step commonly involves a technique called "normalization." This procedure entails adjusting all attributes within a dataset to a consistent scale. However, it is important to note that not all datasets utilized in machine learning necessitate normalization or standardization. The decision to apply normalization depends on the specific characteristics of the dataset, such as the presence of outliers. When outliers are present, Z-score normalization is appropriate, whereas Min-Max normalization can be employed in cases where outliers are absent. Normalization becomes essential when there are variations in the ranges of features.

3.3. Data Preprocessing and Feature Selection

The fundamental concept underlying the resampling technique called cross-validation involves partitioning the dataset into two distinct sets: a training set and a test set. During this process, the model is trained using the training data, and then predictions are generated using the untouched test data.

This approach helps determine whether the model has avoided overfitting to the training data, and its predictive capability can be ascertained by assessing performance on the unseen test data, aiming for high accuracy [13]. Given our dataset's highly imbalanced nature, this work has employed a stratified K-fold cross-validation with a value of '5' for K. This method ensures that the proportions of all categories are represented in roughly equal amounts, enhancing the validity of the validation process.

3.4. Unbalanced Dataset

Imbalanced data refers to datasets with an uneven distribution of target classes for the dependent variable. This means one class label has many observations while another class has significantly fewer, as illustrated in Figure 3. The task at hand involves a three-category classification challenge deliberately skewed toward one of the classes. Precisely, within the dataset, "Graduate" constitutes 50% of the total records (2209 out of 4424), "Dropout" makes up 32% (1421 out of 4424), and "Enrolled" comprises 18% (approximately 794 out of 4424). This imbalance may lead to an elevation in prediction accuracy for the majority class but at the expense of reduced performance for the minority class.

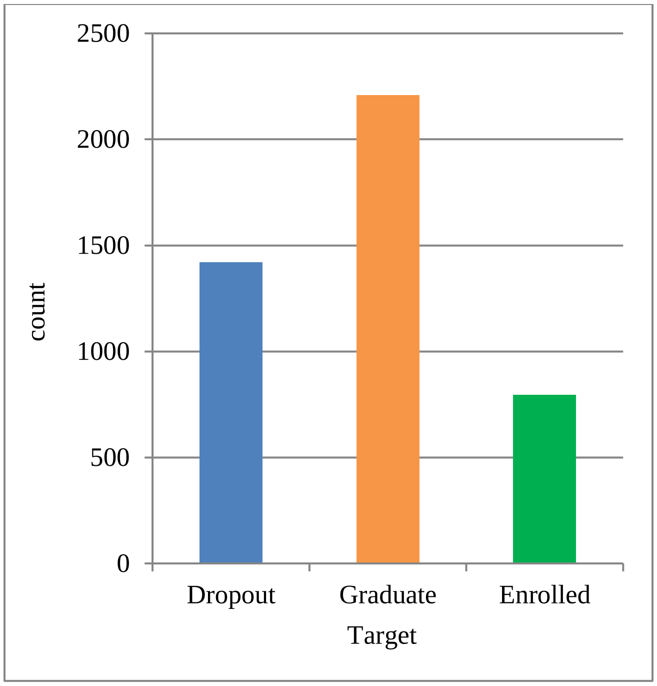

Fig 3(a). Datasets representation PRE-SMOTE

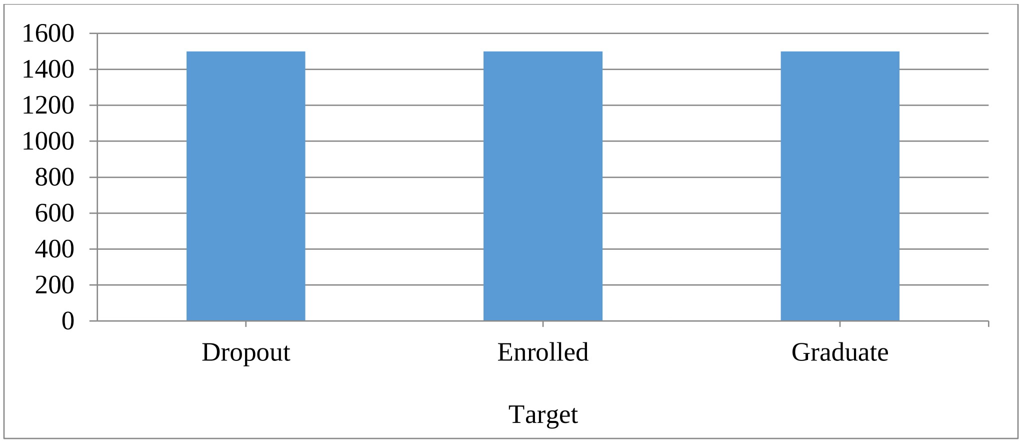

Fig 3(b). Datasets representation POST-SMOTE

To address the issue of imbalanced data, the Synthetic Minority Over Sampling Technique (SMOTE) was employed. Traditionally applied before training and testing, this resampling method helps alleviate imbalances. However, duplicate values could distort predictions using training and test data. To maintain accuracy, SMOTE was exclusively applied to the training dataset. Figures 3a and 3b depict the alterations before and after the implementation of SMOTE.

3.5. Machine Learning Classification Models.

For predicting student academic performance, the modelling stage encompasses nine distinct Machine Learning classification algorithms: K-Nearest Neighbors (KNN), Support Vector Machine, Decision Tree Classifier, Naive Bayes, Random Forest, Neural Networks, XG Boost, and AdaBoost as Level-1 base classifiers, with Logistic Regression serving as the Meta classifier. The main objective of this prediction process is to create an enhanced hybrid model that utilizes information related to population characteristics, economic conditions, financial factors, and educational data, all aimed at predicting student performance.

3.6. Stacking Classifier

Stacking [14] [1] classifier is an ensemble learning method that incorporates multiple classification models to improve the veracity and robustness of predictions. The functioning of a stacking classifier involves training multiple base classifiers on a shared dataset, each utilizing distinct algorithms or hyperparameters. The forecasting generated by these foundational classifiers is then amalgamated through a higher-level classifier, ultimately producing the ultimate forecast. The higher-level classifier, employed within the stacking classifier, can include various classification algorithms like Logistic Regression. This stacking classifier undergoes a two-stage training process. During stage I, the foundation-level classifiers are trained using the training set. Subsequently, the higher-level classifier in stage II is trained to utilize the forecasted probabilities from the foundation level classifiers using the identical training dataset. Stacking classifiers excel in handling complex datasets marked by nonlinear relationships between features and the target variable. Furthermore, they can enhance prediction robustness by leveraging the strength of integrated foundation level classifiers.

3.7. Performance Evaluation

In this research work, evaluation metrics are statistical measures employed to assess the effectiveness of a model. Within this study, we have incorporated five frequently utilized assessment metrics. Accuracy quantifies the proportion of accurate predictions made by the trained model, commonly applied to classification tasks. It represents the ratio of correctly predicted instances to the total cases. Precision gauges the model's capacity to correctly identify positive cases, calculated as the ratio of correctly predicted positive observations to the total positive observations. Recall reflects the model's capability to anticipate all positive cases, quantifying the proportion of correctly predicted positive observations relative to the total number of positive cases in the dataset. The F1 Score combines recall and accuracy in a harmonic manner that proves valuable, especially in scenarios with imbalanced class distributions. The ROC-AUC curve is widely employed for a graphical representation of classifier performance. The area under the curve (AUC) offers a comprehensive performance. The indicator, encompassing various threshold values, showcases the ratio of the observed true positive rate to the observed false positive rate. Recall reflects the model's capability to anticipate all positive cases, quantifying the proportion of correctly predicted positive observations relative to the total number of positive cases in the dataset. The F1 Score, which harmonically combines recall and accuracy, proves valuable, especially in scenarios with imbalanced class distributions.

The ROC-AUC curve is widely employed for a graphical representation of classifier performance. The area under the curve (AUC) offers a comprehensive performance indicator encompassing various threshold values, showcasing the ratio of the observed true positive rate to the observed false positive rate.

```
Algorithm: Student Data Classification
Input: Training dataset containing 4424 students' data with 34 attributes.
    1. Begin
    2. Import necessary library packages and select the dataset
    3. Perform data preprocessing
        3.1 Handle insignificant/missing values.
        3.2 Select appropriate EDA techniques for visual representation of the data.
        3.3 Perform feature scaling with Z-score normalization.
        3.4 Apply feature encoding (if required).
    4. Apply stratified K-fold cross-validation (K=10)
        4.1 Split data into training and testing datasets using stratified 10-fold cross-validation.
    5. Apply SMOTE only to the training dataset to avoid duplicates in testing
    6. Use classification models to predict the results
        6.1 Utilize classification models (KNN, LR, DT, RF, SVM, NN, NB, XGBoost, and AdaBoost) as
    single classifiers.
        6.2 Build a stacking classifier
    7. Evaluate the accuracy of well-known classification models using evaluation measures
    8. End
Output: Three-category classification (Enrolled, Dropout, Graduate)
```

Fig. 4 Methodology for the proposed student performance prediction model

Accuracy = (True Positives (TP) + True Negative (TN)) / (Total Number of Predictions) (1)

Precision = (True Positives (TP)) / (True Positives (TP) + False Positives (FP)) (2)

Recall = (True Positives (TP)) / (True Positives (TP) + False Negatives (FN)) (3)

F1 Score = 2 * (Precision * Recall) / (Precision + Recall) (4)

The outcome summary for the eight distinct classification algorithms is presented in Figure 5, serving as input at level 1 for the classifier, which Logistic Regression constructs.

4. Experimental Results

The ROC-AUC curve, as described in [15], is a widely used assessment measure in machine learning for binary classification tasks. It visually presents the equilibrium between True Positives and False Positives instances across diverse classification thresholds. The Area under the Curve delivers a comprehensive performance appraisal across all possible thresholds, while the ROC curve illustrates the TPR versus FPR across various thresholds. An AUC score of 1 signifies a flawless classifier, whereas an AUC of 0.5 indicates randomness.

The ROC-AUC curve holds significant utility, visually depicting classifier performance that clarifies the trade-off between TPR and FPR. This proves especially valuable when consequences are tied to false negatives and false positives and when selecting the optimal threshold, which holds significance. Generally, a classifier positioned higher and to the left on the ROC curve is deemed superior, as it boasts higher TPR and lower FPR. The ROC-AUC curve is an often-employed yardstick in machine learning contests and real-world scenarios where the balance of false negatives and false positives is pivotal. It is worth noting that choosing an ideal evaluation metric hinges on the precise problem and model objectives. While accuracy might prevail in some cases, precision and recall could take precedence in others. The ROC-AUC curve is one facet within the arsenal of machine learning evaluation tools, underscoring the need to select the most fitting metric tailored to the specific challenge.

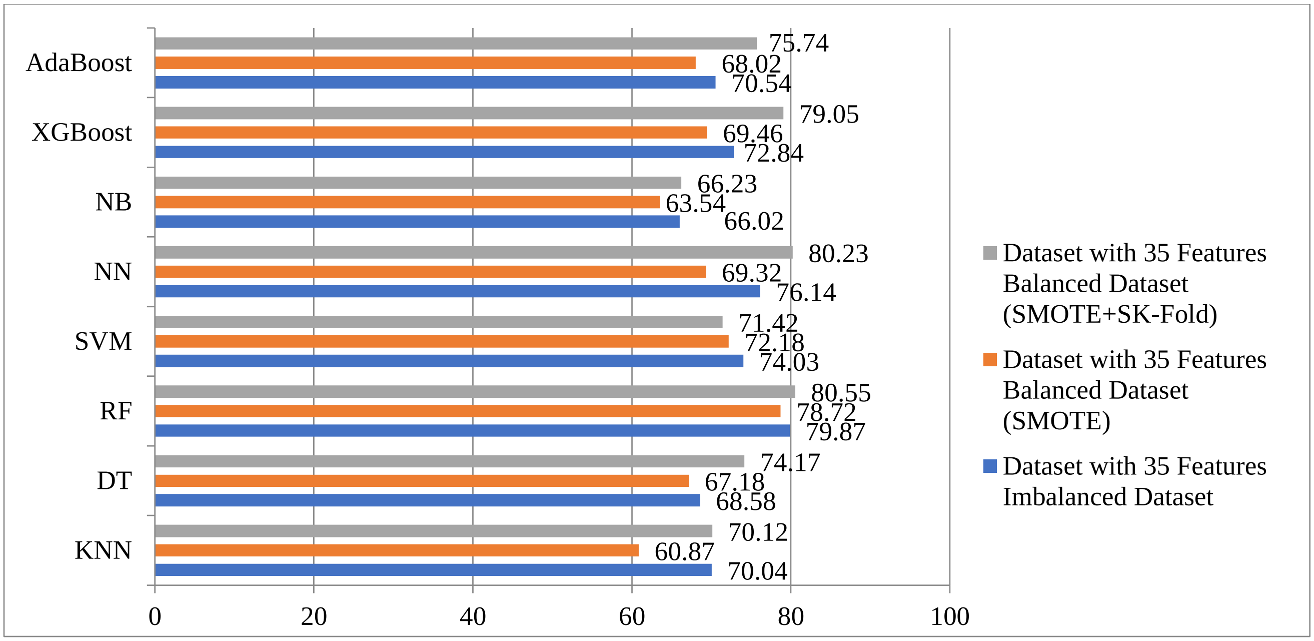

Fig. 5 Comparing the accuracy of level-1 classification algorithms

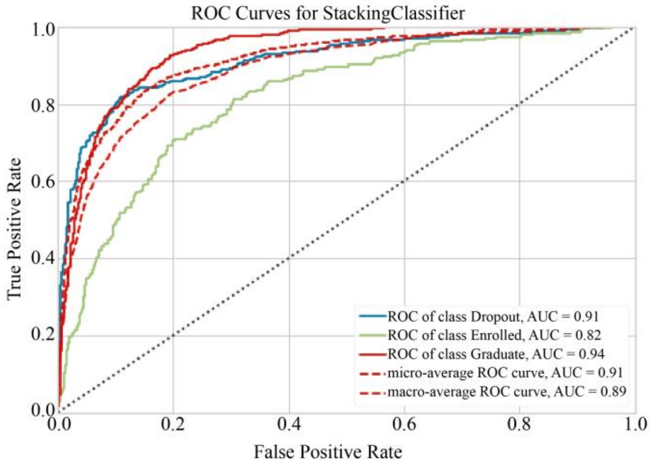

Fig. 6 ROC-AUC curves for hybrid machine learning stacking classifier across all models

5. Conclusion and Future Scope

The Hybrid algorithms stem from amalgamating the strengths of multiple classification algorithms to surmount the limitations inherent in individual ones. As evident in the outlined methodology of this study, an exploration of eight diverse ML classification algorithms has been undertaken. The conclusive outcomes underscore the commendable performance of these classification algorithms, particularly in conjunction with stratified 5-fold cross-validation. Concurrently, the ROC-AUC curve establishes that the suggested stacking model approach produces improved outcomes across the three distinct categories: dropout (0.91%), enrolled (0.82%), and graduate (0.94%).

In our future work, the integration of graph neural network (GNN) is strongly recommended due to its relevance in educational environments. Student interactions within such settings are pivotal, influencing their academic progress. GNNs offer the capacity to model these intricate relationships, thereby encapsulating the interconnections among students. This approach is more robust for performance prediction than relying solely on individual student data. GNNs further excel in managing substantial datasets, aligning with the data-rich nature of educational contexts. Notably, student performance encompasses multifarious contextual factors encompassing the educational milieu, teacher calibre, and socioeconomic backdrop.GNNs adeptly integrate these contextual nuances into the predictive framework, enhancing outcomes' precision. In summation, opting for GNNs proves judicious in forecasting student performance owing to their prowess in managing heterogeneous data, assimilating contextual insights, and accommodating voluminous datasets.

References

[1] Reynold A. Rustia et al., "Predicting Student's Board Examination Performance Using Classification Algorithms," *ICSCA '18: Proceedings of the 2018 7th International Conference on Software and Computer Applications*, Kuantan, Malaysia, pp. 233-237, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Hanan Abdullah Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," *IEEE Access*, vol. 8, pp. 55462-55470, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Siti Dianah Abdul Bujang et al., "Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 1970-1989, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ramin Ghorbani, and Rouzbeh Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899-67911, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Akhilesh P. Patil, Karthik Ganesan, and Anita Kanavalli, "Effective Deep Learning Model to Predict Student Grade Point Averages," *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Coimbatore, India, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[6] Praveena Chakrapani, and D. Chitradevi, "Academic Performance Prediction Using Machine Learning: A Comprehensive & Systematic Review," *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, Chennai, India, pp. 335-340, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Matloob Khushi et al., "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960-109975, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Raza Hasan et al., "Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques," *Applied Sciences*, vol. 10, no. 11, pp. 1-20, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Christian Fischer et al., "Mining Big Data in Education: Affordances and Challenges," *Review of Research in Education*, vol. 44, no. 1, pp. 130-160, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Aishwarya Suresh, H.S. Sushma Rao, and Vinayak Hegde, "Academic Dashboard—Prediction of Institutional Student Dropout Numbers Using a Naive Bayesian Algorithm," *Computing and Network Sustainability*, vol. 12, pp. 73-82, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ivan Nunes da Silva et al., *Artificial Neural Network Architectures and Training Processes*, Artificial Neural Networks : A Practical Course, Springer International Publishing, pp. 21-28, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[12] Valentim Realinho et al., "Predicting Student Dropout and Academic Success," *Data*, vol. 7, no. 11, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Juan D. Rodriguez, Aritz Perez, and Jose A. Lozano, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 569-575, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[14] Oleg Uzhga-Rebrov, and Peter Grabusts, "Comparative Evaluation of Four Methods for Exploratory Data Analysis," *2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, Riga, Latvia, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Shahzad Ali Khan, and Zeeshan Ali Rana "Evaluating Performance of Software Defect Prediction Models Using Area Under Precision-Recall Curve (AUC-PR)," *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*, Lahore, Pakistan, pp. 1-6, 2019. [CrossRef] [Google Scholar] [Publisher Link]