The Implementation of Educational Data Mining in Predicting Students' Academic Achievement in Mathematics at a Private Elementary School

Hendra Tjahyadi* and Krismon N. L. Tude

Faculty of Computer Science, Universitas Pelita Harapan, Tangerang, Indonesia
Email: hendra.tjahyadi@uph.edu (H.T.); rellytude2@gmail.com (K.N.L.T.)
*Corresponding author

Manuscript received August 21, 2024; revised September 18, 2024; accepted October 14, 2024; published January 20, 2025

***Abstract*—This paper examines the use of Educational Data Mining (EDM) to predict the academic performance of elementary students specifically in Mathematics. It explores ten Machine Learning classifiers, comprising eight base learners (Linear SVM, Logistic Regression, Medium KNN, Wide NN, Fine Decision Trees, Bilayered NN, Fine KNN, and Medium NN) as well as two ensemble learners (Ensemble Subspace Discriminant and Ensemble Boosted Trees) within the MATLAB environment. The analysis utilizes a dataset featuring 33 academic and demographic features of 280 students. To mitigate the imbalanced distribution in class data, resampling techniques such as Random Under-Sampling Boost (RUSBoost), Synthetic Minority Oversampling Technique (SMOTE), and hybrid combinations of both are employed. The experimental outcomes demonstrate that the hybrid-sampling SMOTE-RUSBoosted Trees algorithm achieves the highest accuracy of 75% on testing data, indicating the efficacy of combining oversampling and under-sampling techniques for modeling imbalanced datasets. This finding underscores the potential of EDM in the elementary education sphere to bolster data-driven interventions and enhance students' Mathematics achievement.**

***Keywords*—educational data mining, mathematics achievement, ensemble learning, imbalanced class, resampling methods**

I. Introduction

Mathematics holds a special significance in Indonesian education, being crucial for scientific and technological comprehension [1–3]. Despite its importance as the cornerstone of scientific and technological understanding, it remains widely regarded as the most challenging subject among Indonesian students [4]. The latest Indonesian National Assessment in 2021, covering over 259 thousand schools and 6.5 million students, revealed that two-thirds of students had not achieved minimum competency in numeracy [5]. Evaluations from the Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) also indicate deficiencies in logic and reasoning among Indonesian students during Mathematics assessments [6, 7]. Consequently, continuous efforts are being undertaken by government in Indonesia to enhance Mathematics learning, spanning from elementary to higher education levels.

Numerous factors likely influenced students' Mathematics achievement [8]. These include content knowledge, perceptions [9], negative attitudes toward Mathematics [10], and environmental factors like social, family, and school environments [11, 12]. A recent study by Alyahyan and Düştegör [13] identified five key predictors of student performance: academic achievement, demographics, learning environment, psychological traits, and learning activities. Schools can use these factors to design effective approaches for improving Mathematics achievement, especially for at-risk students.

In the process of learning over long periods of time, a large amount of data about students is inevitably collected by the schools. Therefore, there is a great potential to utilize data mining to the increasing data for the beneficial of schools' practices. Utilizing educational data collected over time, schools can employ data mining techniques, known as Educational Data Mining (EDM), to extract valuable insights for educational practices. EDM explores diverse data features to enhance understanding of students and their learning contexts, offering more precise insights than traditional statistical methods [14].

A recent systematic literature review spanning from 2015 to 2021 underscores the challenge of imbalanced data distribution in predicting student performance [15]. This issue can lead to bias and errors in binary or multi-class classifications, yet comprehensive methods to address this are lacking. Therefore, exploring the potential of Educational Data Mining (EDM) in predicting achievement in imbalanced classes holds promise for improving model accuracy. However, there is a shortage of studies focusing on implementing EDM in the field of Mathematics education. Existing literature indicates that EDM is predominantly applied at the university level, with limited usage in secondary and primary schools [13, 15–17]. Moreover, most studies concentrate on outcomes such as final GPA, dropout rates, and passing rates in higher education, rather than specifically targeting mathematical achievement at the elementary level.

From the discussion above, it can be concluded that despite the importance of identifying factors influencing elementary school students' mathematics achievement to design effective approaches, the available school data has not been fully utilized. Additionally, there is limited literature on the use of EDM in Mathematics, particularly at the elementary school level. Therefore, the objective of this paper is to explore the use of machine learning-based EDM to predict elementary school students' mathematics achievement, with a focus on improving model accuracy despite the presence of imbalanced data

The rest of the paper is structured as follows: Section II explores related works and identifies research gaps in implementing EDM for predicting students' Mathematics achievement. Section III details the methodology of machine

doi: 10.18178/ijiet.2025.15.1.2228

learning-based EDM implementation. Section IV presents the results, followed by a discussion on the evaluation and interpretation of these results. Section V provides the research conclusion, and finally Section VI discusses the limitation of the research and proposes future direction.

II. Related Works and GAP

Researchers commonly utilize EDM to predict students' academic achievement, given its ability to reveal information that enhances decision-making [18]. Predicting student achievement involves two main factors: attributes and prediction methods [19]. Attributes encompass student properties such as final grades, gender, and demographics, while prediction methods include algorithms like Regression Models, Decision Trees, Naïve Bayes, Artificial Neural Networks and Support Vector Machines. Supervised learning techniques are predominantly used for predicting students' achievement [19], while unsupervised learning remains less explored in this domain [20].

Several studies have applied various algorithms to predict students' performance in subjects such as Mathematics and English in high school. For instance, employing algorithms like Decision Trees, Naïve Bayes, Neural Networks, and Support Vector Machines resulted in significant correlations between English and Mathematics performance, with Naïve Bayes exhibiting the highest accuracy in predicting Mathematics performance [21]. Other studies have analyzed the influence of students' backgrounds, social behaviors, and coursework completion on predicting secondary school students' Mathematics performance, revealing significant impacts of these factors [22]. Furthermore, algorithms like Support Vector Machines and K-Nearest Neighbors have been employed to estimate university students' grades in final Mathematics exams, with Support Vector Machines achieving slightly better results with correlation coefficient of 0.96, while the KNN achieved correlation coefficient of 0.95 [23].

To enhance the performance of algorithms like Random Forest, researchers have proposed hybrid approaches such as the Improved Random Forest Classifier, resulting in higher accuracy in predicting student performance [24, 25]. Additionally, ensemble or hybrid learning models have been found to be more effective in accurately predicting students' academic performance compared to individual learning models [26, 27]. To address imbalanced data, researchers have utilized resampling methods in combination with ensemble classifiers, with Random Forest as an ensemble classifier achieving the best results [28]. It is concluded that utilizing more ensemble methods in student grade prediction is crucial for improving prediction accuracy.

Despite these advancements, the literature review indicates that there are still opportunities to discover improved algorithms for predicting students' achievement in Mathematics. Given the likelihood of imbalanced class distribution, with middle-achieving students dominating, this study aims to identify the best algorithm using hybrid classification learner approaches combined with resampling methods. Additionally, due to the limited application of EDM in elementary school mathematics, this research addresses this gap by implementing EDM to predict students' Mathematics Achievement at the elementary level.

III. Methodology

Fig. 1 outlines the research methodology, which begins with raw data collection from an Elementary school. This is followed by pre-processing steps involving data cleaning, feature encoding, and scaling. The dataset is then split into training and testing sets. Next, the training data is input into classification algorithms through two processes: direct input and input after resampling. This allows for evaluating the effectiveness of resampling in handling imbalanced data. Finally, the best algorithm is chosen based on evaluation metrics.

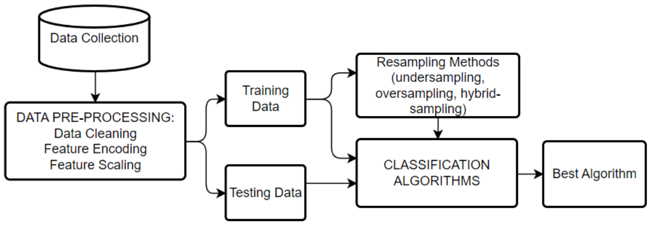

Fig. 1. Flow diagram of the research methodology.

A. Environment

For this research, MATLAB R2023b software is employed on a laptop equipped with an AMD Ryzen 5 5500U processor and 8 GB of RAM. MATLAB is selected due to its user-friendly data visualization capabilities and the availability of toolboxes for statistical analysis and data mining. The Data Cleaner toolbox is utilized for data preprocessing and transformation, while the Classification Learner toolbox facilitated the creation and evaluation of prediction models.

B. Data Collection

The main data source for this research comprises raw data extracted from a database containing academic and non-academic records of students at a private Elementary school in Tangerang, Indonesia. These records encompassed the academic years 2017/2018 and included the academic performance of 280 Grade IV students, along with their Mathematics achievements upon graduating in 2020/2021.

C. Data Mining Techniques

1) Data selection

The data selection process entails identifying pertinent attributes from both academic and non-academic records that could impact student performance in Mathematics. A total of 33 features are under analysis in this research, as presented in Table 1.

The features will be utilized to predict students' Mathematics achievement levels, corresponding to their Report Card Grade 6 upon graduation. These features may influence Mathematics performance through cognitive, behavioral, and environmental pathways. For instance, attendance, concentration, and extra lessons directly impact cognitive engagement, while variables such as place of birth, religion, and daily language may indirectly shape performance by influencing cultural attitudes and learning habits. Likewise, parental education and occupation can shape the home learning environment, affecting motivation and academic support. Students' Mathematics achievement, designated as the class label for prediction, is categorized into three groups: High, Medium, and Low. Students scoring 79 or below are classified as Low, those scoring between 80 and 95 are considered Medium, and those scoring 95 or above are classified as High. Based on this categorization, from 280 students, it is obtained that 10% are classified as low, 67% as middle and 23% as high, indicating an imbalance in the datasets.

Table 1. Features and description

Group of Features	Features	Types
Academic Records	Religion, Civics, Science, Social Studies, Indonesian, English, Mandarin, Art and Craft, Computer, Physical Education, Religion	Numerical
	Extracurricular types, Extracurricular scores, Scout activity scores	Categorical
Student Demographics	Attendance, Age, Total sibling, Distance to school, Gender, Body-mass index, Activities afterschool, Extra-lesson afterschool, Extra-lesson period, Concentration level	Numerical
	Place of Birth, Students' religion, Daily language, Blood type, Both Parents' background of education, Both Parent's job	Categorical

2) Data pre-processing

The Data Pre-Processing involves ensuring data quality and suitability for data mining techniques. This includes data cleaning, feature encoding, feature scaling, model validation, and handling imbalanced data.

a) Data cleaning

Missing values are addressed using linear interpolation and moving averaging techniques to smooth the dataset before modeling.

b) Feature encoding

Since machine learning algorithms cannot work with categorical data, the categorical data should be converted into numerical form. Categorical data is converted into numerical form using label encoding and one-hot encoding, resulting in a dataset of 280 rows and 48 columns.

c) Feature scaling

Z-square normalization is applied to both training and testing datasets to standardize the range of features.

d) Handling imbalanced-data

The datasets obtained indicate an imbalance that could results in compromised reliability of the prediction model. The dominance of the majority class may skew the model's inclination towards it, diminishing the predictive accuracy for minority classes. In this study, imbalanced class labels are addressed through resampling techniques such as SMOTE for oversampling and RUSBoost for undersampling, and a combination of both for hybrid-sampling. Resampling is only applied to the training set to maintain model reliability, while the testing set is used for evaluation.

e) Cross validation

Model validation assesses how well the independent features in the dataset generalize the analysis results. This research employs both random hold-out and 5-fold cross-validation techniques. Random hold-out divides the 280x33 dataset into a 70% training set and a 30% testing set. Meanwhile, in 5-fold cross-validation, the model undergoes training on four parts and validation on the remaining part, repeated five times for each different part. The use of 5-fold cross-validation will optimize the utilization of the available data, thereby mitigating overfitting, providing a more reliable model, and improving accuracy, particularly for the use of relatively small datasets.

3) Machine learning algorithm

In this research, ten classification learners are compared using the MATLAB Classification Learners Application. These algorithms will be utilized in two distinct experiments: one with resampling techniques applied on the training set (such as RUSBoost, SMOTE, and their combination) and another without resampling.

Table 2. Classification algorithms

No.	Model Type	Learner	Code
1	Decision Trees	Fine Trees	DT
2	K-Nearest Neighbour	Fine KNN	F-KNN
3	K-Nearest Neighbour	Medium KNN	M-KNN
4	Support Vector Machine	Linear SVM	L-SVM
5	Logistic Regression	Logistic Regression	LR
6	Neural Network	Wide NN	W-NN
7	Neural Network	Medium NN	M-NN
8	Neural Network	Bilayered NN	B-NN
9	Ensemble Subspace	Subspace Discriminant	ESD
10	Ensemble Boosting	Boosted Trees	EBT

Each classification learner automates the selection of hyperparameter values using MATLAB to minimize errors and provide the model with optimized hyperparameters [29]. Table 2 displays the ten classification algorithms that will be evaluated in this study.

a) Decision Trees (DT)

Decision Trees construct decision nodes connected by branches from the root to the leaf nodes. Each decision node undergoes statistical testing, generating branches with respective outcomes directing towards another node or the final decision. Hyperparameter optimization involves specifying parameters like the maximum number of splits or split criteria. This research utilizes the Fine Trees learner, which sets the maximum number of splits to 100 [30]. Additionally, DT is employed in ensemble learning (EBT) for predictive modeling.

b) K-Nearest Neighbor (KNN)

K-Nearest Neighbor determines a class based on its proximity to neighboring data points. Various distance metrics, such as Euclidean, City Block, Cosine, Chebyshev etc., are used to calculate the distance between points. KNN requires specifying the number of neighbors (k) for classification decisions. This research employs Fine KNN with 1 neighbor and Medium KNN with 10 nearest neighbors.

c) Support Vector Machine (SVM)

SVM is supervised algorithm for classification and regression, based on structural risk minimization. It identifies a hyperplane with maximum margin to separate classes without assuming data distribution. Linear SVM is utilized in this study.

d) Discriminant analysis

Discriminant Analysis swiftly generates classes based on Gaussian distributions. The fitting function estimates parameters for each class. In this research, Discriminant Analysis is employed as an ensemble model, optimizing subspace dimensions of features.

e) Logistic Regression (LR)

LR predicts categorical variables' probability based on predictor variables. It's used for supervised parameter learning, encompassing linear and logistic regression, and linear discriminants like SVM for classification. LR is applied as a single learner in this study.

f) Neural Network (NN)

NN mimics the human brain's structure to solve complex problems. It consists of input, hidden, and output layers interconnected by nodes. Increasing layer size enhances model flexibility for better classification. Wide NN, Medium NN, and Bilayered NN are utilized in this research.

g) Ensemble learning

Ensemble Learning combines multiple learners to enhance model performance through Bagging, Boosting, and Stacking. Bagging aggregates results through voting, Boosting corrects weak learners' mistakes, and Stacking creates a meta-classifier. Ensemble classifiers like ESD, EBT, and ERT are implemented in this study, focusing on ensemble subspace dimension manipulation for improved prediction accuracy in student academic performance.

4) Evaluation and interpretation

The results will be assessed and interpreted within the framework of the research objectives, particularly focusing on identifying the algorithm that performs best in predicting students' Mathematics achievement. Evaluation will utilize metrics such as Accuracy, Precision, Recall, F-Measure, and Receiver Operating Characteristics Curve (ROC) to compare the results.

a) Accuracy

It represents the proportion of correctly classified data points out of the total number of data points, as depicted in (1). This ratio is determined through the confusion matrix illustrated in Fig. 2, which contrasts the model's estimates with the actual values.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

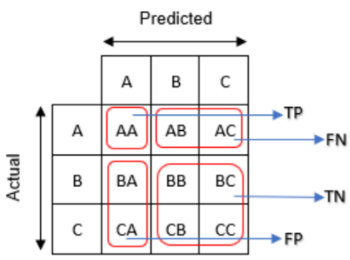

Fig. 2. Confusion matrix 3×3 [31].

b) Precision

It is the ratio of true positive instances to the total number of instances predicted as positive, as depicted in (2). This metric indicates the proportion of relevant items among those selected.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

c) Recall

It is the ratio of true positive instances to the sum of positive instances and false negative instances. This metric is also known as the true positive rate.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

d) F-Measure

It assesses precision and recall criteria jointly to yield more accurate and sensitive outcomes. It represents the harmonic mean of precision and recall values.

$$F - Measure = \frac{2 \,.\, Precision \,.\, Recall}{Precision+Recall} \tag{4}$$

e) Receiver Operating Characteristics Curve (ROC) – Ranking Metric

It plots True Positive Rate (TPR) on the y-axis and False Positive Rate (FPR) on the x-axis, creating a curve that depicts the balance between detection and false alarm rates. This method is widely used to assess the performance of an imbalanced learner. Additionally, a single numerical metric, called Area Under the Curve (AUC), will be utilized to rank the algorithms.

IV. Result and Discussion

Three scenarios of experiment are described based on the experiment using MATLAB Classification Learner Application on the pre-processed dataset. The first scenario is comparing the classification algorithms without resampling methods. The second scenario is comparing the algorithms with resampling methods (oversampling, undersampling, and hybrid-sampling). The last scenario aims to compare the top performed classification algorithms in the first and the second scenarios to find the best algorithm to predict students' Mathematics achievement.

A. Classification Algorithm without Resampling

In implementing the classification technique, all ten chosen algorithms are employed to train and test the dataset. Each algorithm undergoes training and testing with validation through both random hold-out and 5-fold cross-validation techniques. Table 3 presents the performance comparison of all the algorithms, while Table 4 provides detail evaluation metrics for the top three algorithms.

From Table 3, it is apparent that nearly all learners experience a decline in accuracy when tested on the testing set. Notably, ESD stands out as the top performer with an accuracy of 73%, followed by L-SVM and EBT, both achieving 70% accuracy. This observation underscores the superior performance of ensemble learners, particularly ESD, compared to single algorithms like LR, KNN, NN, and DT. These results reaffirm the prevailing notion in literature that ensemble learners consistently outperform single learners in terms of accuracy.

Table 4 reveals the strengths and weaknesses of the top three algorithms - ESD, L-SVM, and EBT- in classifying students into three different labels. However, both ESD and EBT exhibit superior performance compared to L-SVM, as evidenced by the highest percentage (highlighted in bold) of each metric for all class labels.

Specifically, ESD demonstrates better Accuracy in classifying students across all labels, particularly High-class and Middle-class. Furthermore, comparing the F-Measure indicates ESD as a more balanced learner, as it represents the harmonic mean between Recall and Precision. However, these three algorithms still struggle in classifying Low-class due to low Recall and Precision, although EBT shows relatively better performance. The higher accuracy with lower Recall, Precision, and F-Measure may stem from the imbalanced distribution of Low labels, impacting the True Negative (TN) value, thereby inflating accuracy.

Table 3. Classification algorithms without resampling

Model Type	Learner	Code	Validation Accuracy	Test Accuracy
Ensemble	Subspace Discriminant	ESD	80%	73%
SVM	Linear SVM	L-SVM	78%	70%
Ensemble	Boosted Trees	EBT	73%	70%
Logistic Regression	Logistic Regression	LR	76%	68%
KNN	Medium KNN	M-KNN	72%	68%
NN	Wide NN	W-NN	76%	67%
Decision Trees	Fine Trees	DT	69%	67%
NN	Bilayered NN	B-NN	72%	64%
KNN	Fine KNN	F-KNN	64%	64%
NN	Medium NN	M-NN	78%	62%

Table 4. Comparison of ESD, L-SVM, and EBT

Learner (Code)	Class Label	Accuracy	Recall	Precision	F-Measure
Subspace Discriminant (ESD)	High	**85%**	**75%**	**65%**	**70%**
	Low	**88%**	43%	33%	38%
	Middle	**73%**	**75%**	83%	**79%**
Linear SVM (L-SVM)	High	83%	74%	61%	67%
	Low	87%	25%	11%	15%
	Middle	70%	72%	**85%**	78%
Boosted Trees (EBT)	High	82%	70%	61%	65%
	Low	**88%**	**44%**	**44%**	**44%**
	Middle	70%	**75%**	79%	77%

Fig. 3–Fig. 5 display the AUC-ROC scores of the algorithms at various classification thresholds. A higher AUC-ROC score indicates better performance in distinguishing True Positive Rate from False Positive Rate. The AUC-ROC scores reveal that ESD excels in differentiating High-class with a score of 0.87, L-SVM performs better in distinguishing Low-class with a score of 0.844, and EBT is superior in distinguishing Middle-class. Comparing the AUC-ROC scores suggests that each algorithm excels in different classes, with L-SVM exhibiting slightly more balanced scores across all classes.

Overall, considering all evaluation metrics (Accuracy, Recall, Precision, F-Measure, AUC-ROC), it can be concluded that ESD outperforms both L-SVM and EBT by consistently achieving higher scores. ESD surpasses L-SVM and EBT in Accuracy, Recall, and F-Measure, particularly for the High-class and Middle-class labels. In the subsequent section, ESD will be further compared with other classification algorithms to determine the best algorithm for predicting students' Mathematics achievement.

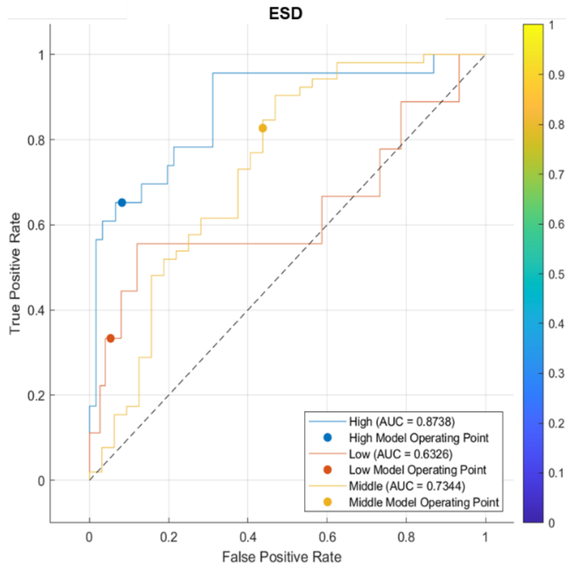

Fig. 3. AUC-ROC Curve of ESD.

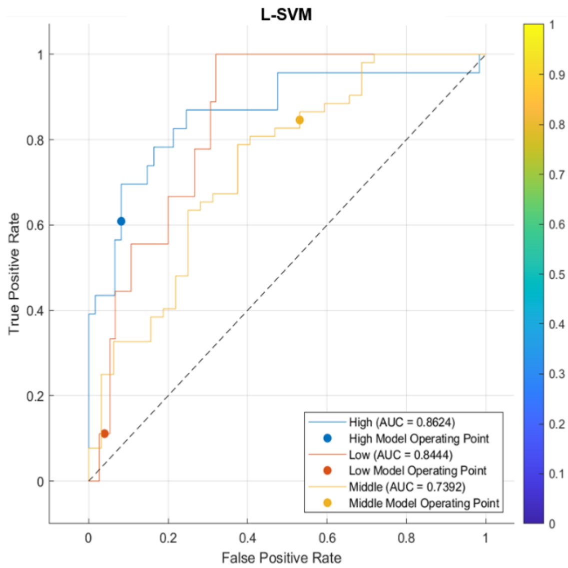

Fig. 4. AUC-ROC Curve of L-SVM.

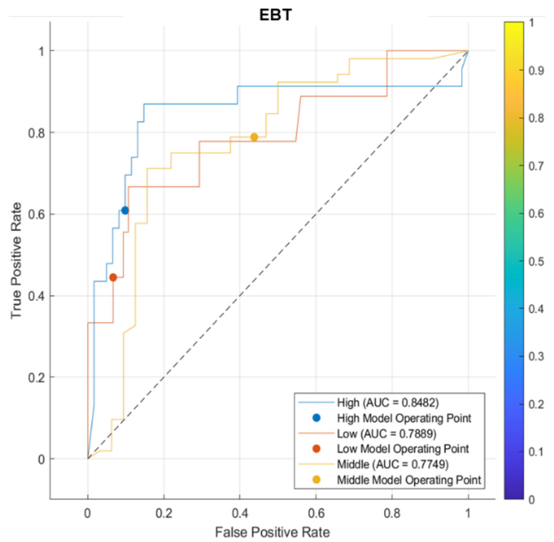

Fig. 5. AUC-ROC Curve of EBT.

B. Imbalanced Classification with Resampling

1) Imbalanced classification with SMOTE oversampling technique

SMOTE oversampling technique is commonly used to address imbalanced classification by augmenting the dataset with synthetic minority samples. In this research, the SMOTE oversampling technique is applied to the 70% training set to create various SMOTE balanced training sets, which will then be used to construct the predictive model. Following the application of these different SMOTE-balanced training sets, Table 5 presents a comparison of the performance of all algorithms.

Overall, implementing the ten algorithms on various SMOTE-balanced training sets significantly improved the validation accuracy to a range of 89% to 99% from 64% to 80% (Table 3). However, there is not much improvement in test accuracy, which remained in the range of 62% to 73% before and after using different SMOTE-balanced training sets, with only three algorithms showing improvement: EBT from 70% to 71%, DT from 67% to 68%, and M-NN from 62% to 65%. Importantly, Table 5 demonstrates that both ensemble learners continue to outperform single learners on the balanced training set. Notably, SMOTE-Subspace Discriminant (S-ESD) achieved 73% test accuracy, and SMOTE-Boosted Trees (SBT) achieved 71% test accuracy, indicating superior performance on different balanced training sets compared to other classification algorithms. These results corroborate findings from [15] and [28] that employing ensemble methods with resampling techniques for predicting students' grades yields superior performance on imbalanced datasets compared to single classifiers.

Table 5. Classification on different SMOTE-balanced training set

Model Type	Learner	Code	Validation Accuracy	Test Accuracy
Ensemble	Subspace Discriminant	S-ESD	89%	**73%**
SVM	Linear SVM	S-LSVM	91%	68%
Ensemble	Boosted Trees	S-EBT	97%	**71%**
Logistic Regression	Logistic Regression	S-LR	95%	67%
KNN	Medium KNN	S-KNN	89%	65%
NN	Wide NN	S-NN	99%	65%
Decision Trees	Fine Trees	S-DT	92%	68%
NN	Bilayered NN	S-NN	99%	63%
KNN	Fine KNN	S- KNN	99%	62%
NN	Medium NN	S-NN	99%	65%

Table 6 and Fig. 6 and Fig. 7 provide a detailed comparison of the top two classification algorithms on different SMOTE balanced-training sets using various evaluation metrics on the testing set. Generally, S-ESD exhibits relatively higher scores on multiple metrics compared to S-EBT, although the differences are not significant. Based on Table 6, it is evident that S-ESD demonstrates better performance in predicting High and Middle-classes, as indicated by metrics such as Accuracy, Precision, and F-Measure. Conversely, S-EBT shows superior performance in classifying Low-class across all metrics, albeit with lower scores. Comparing the AUC-ROC score also leads to the same conclusion, with S-ESD excelling in differentiating the positive class of High and Middle-classes, while S-EBT performs better in distinguishing the positive class for Low-class. Overall, considering scores from all evaluation metrics, S-ESD outperforms S-EBT, particularly in predicting Middle and High-classes. In the subsequent section, S-ESD will be further compared with other algorithms to determine the best algorithm for predicting students' Mathematics achievement.

Table 6. Comparison between ERT and ESD on different SMOTE balanced-training set

Learner (Code)	Class Label	Accuracy	Recall	Precision	F-Measure
S-ESD	High	**85%**	**75%**	**65%**	**70%**
	Low	88%	43%	33%	38%
	Middle	**73%**	75%	**83%**	**79%**
S-EBT	High	82%	70%	61%	65%
	Low	**89%**	**50%**	**44%**	**47%**
	Middle	71%	**81%**	75%	78%

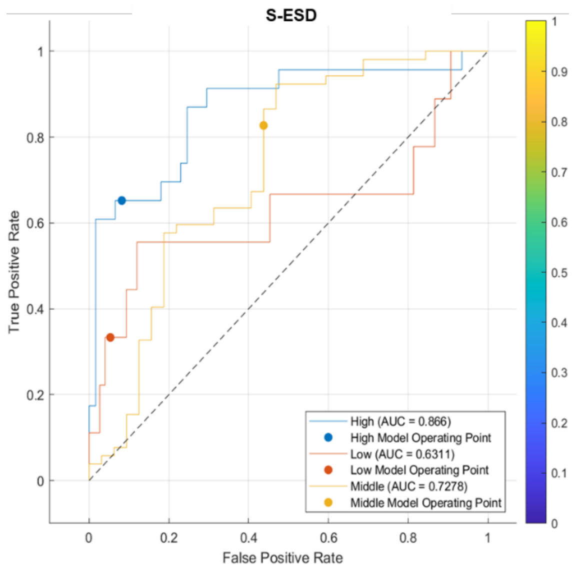

Fig. 6. AUC-ROC Curve of S-ESD.

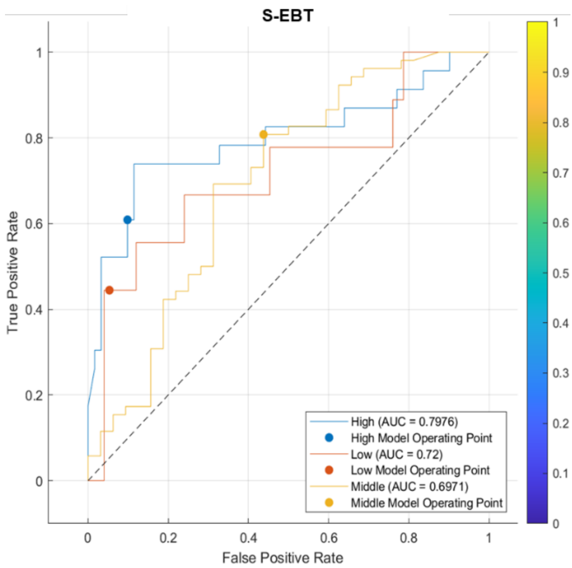

Fig. 7. AUC-ROC Curve of S-EBT.

2) Imbalanced classification with undersampling technique

In the undersampling technique, the dataset is resampled by randomly removing instances from the majority class to balance the class distribution. In this study, RUSBoosted-Trees (ERT) with Decision Trees as the base learner is chosen for its efficiency compared to other learners.

Table 7. Imbalanced classification with undersampling technique (ERT)

Validation Accuracy	Test Accuracy	Class Label	Accuracy	Recall	Precision	F-Measure
65%	64%	High	82%	63%	87%	73%
		Low	82%	35%	78%	48%
		Middle	64%	84%	52%	64%

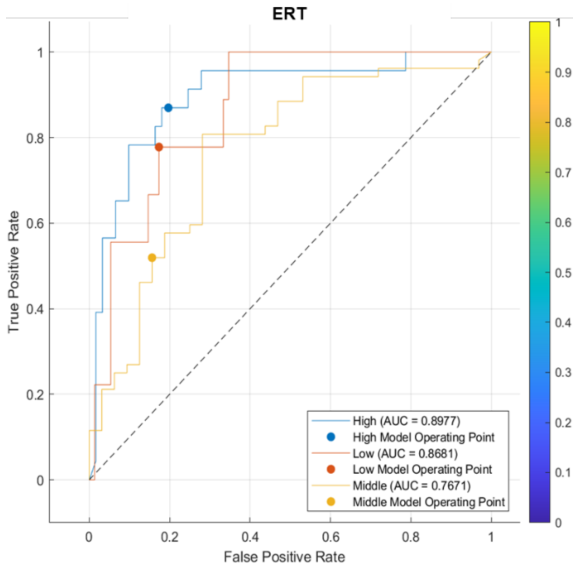

Fig. 8. AUC-ROC Curve of ERT.

Table 7 and Fig. 8 show the performance of ERT algorithm. Table 7 indicates that the 64% test accuracy closely corresponds with the 65% model validation accuracy, showing consistent performance in predicting testing data. The ERT algorithm excels in classifying High-class and Low-class but struggles with Middle-class due to its under-sampling technique, which prioritizes performance improvement on the minority class by reducing the majority class. Despite being less sensitive in identifying Low-class (with a low Recall score), it demonstrates good Precision, suggesting better detection of exact Low-class. This is supported by the AUC-ROC score in Fig. 8, which highlights ERT's effectiveness in distinguishing High and Low- classes but reveals less efficacy in distinguishing Middle-class, scoring 0.76. To enhance the ERT model and enable better comparison, exploring hybrid-sampling, which combines both undersampling and oversampling, would be a compelling approach to consider.

3) *Imbalanced classification with hybrid-sampling technique*

The combination of SMOTE oversampling and ERT undersampling, referred to as hybrid sampling, is employed to increase the number of minority samples and decrease the number of majority samples, thereby mitigating sample imbalance. As the undersampling technique is solely utilized by the RUSBoosted-Trees algorithm (ERT), the hybrid sampling technique is implemented by applying ERT to the different SMOTE-balanced training set (S-ERT).

Table 8 and Fig. 9 display the performance of S-ERT. It is evident that applying hybrid-sampling yields improvement as demonstrated in Table 8.

Table 8. Imbalanced classification with hybrid-sampling (S-ERT)

Validation Accuracy	Test Accuracy	Class Label	Accuracy	Recall	Precision	F-Measure
92%	75%	High	88%	76%	83%	79%
		Low	87%	43%	67%	52%
		Middle	75%	84%	73%	78%

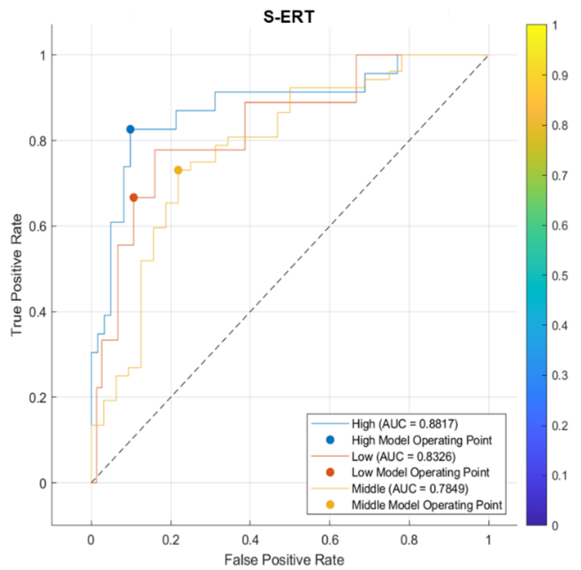

Fig. 9. AUC-ROC Curve of S-ERT.

The validation accuracy has increased from 65% to 92%, and the testing accuracy has improved from 64% to 75%. This highlights the significant enhancement in accuracy achieved through hybrid-sampling for the ERT model. Table 8 indicates that S-ERT performs relatively well in classifying High and Middle-class but exhibits lower sensitivity in identifying Low-class, as evidenced by low Recall scores. Additionally, the AUC scores in Fig. 9 reveal that S-ERT is better at distinguishing High and Low-classes compared to Middle-class, with AUC scores of 0.88 for High-class, 0.83 for Low-class, and 0.78 for Middle-class. This evaluation will be compared with other algorithms in the next section to determine the best algorithm for predicting students' Mathematics achievement.

C. *Best Algorithm in Predicting Students' Mathematics Achievement*

To determine the best algorithm, a thorough comparison of all evaluation metrics is conducted. The best algorithm is selected by identifying the highest score among the chosen classification algorithms under two previous scenarios, resulting in four different algorithms: (i) classification without resampling methods (ESD), and classification algorithms with resampling methods: (ii) S-ESD applying oversampling technique, (iii) ERT employing undersampling technique, and (iv) S-ERT implementing hybrid-sampling technique. Table 9 contrasts the performance of these four algorithms, taking into account evaluation metrics.

Analysis of Table 9 reveals varying scores across different algorithms in the evaluated metrics. Each metric highlights distinct advantages and limitations, suggesting that a holistic approach is crucial. Therefore, not only accuracy but also F-Measure, which accounts for the harmonic average of Recall and Precision, and AUC-ROC are considered to ensure consistency among the evaluation metrics, enhancing the model's reliability for real-world implementation.

Table 9. Comparison of classification algorithms without resampling and with resampling method

Types	Learner (Code)	Validation Accuracy	Test Accuracy	Class Label	AUC - ROC	Accuracy	Recall	Precision	F-Measure
Ensemble Classification Algorithm without resampling	Subspace Discriminant **(ESD)**	80%	73%	High	87%	85%	75%	65%	70%
				Low	63%	88%	43%	33%	38%
				Middle	73%	73%	75%	83%	**79%**
Ensemble with Over-sampling	SMOTE - Ensemble Subspace Discriminant **(S-ESD)**	89%	73%	High	87%	85%	75%	65%	70%
				Low	63%	88%	43%	33%	38%
				Middle	73%	73%	75%	83%	**79%**
Enemble with Under-sampling	Ensemble RUSBoosted Trees **(ERT)**	63%	64%	High	**90%**	82%	63%	**87%**	73%
				Low	**87%**	82%	35%	**78%**	48%
				Middle	77%	64%	**84%**	52%	64%
Ensemble with Hybrid-resampling	SMOTE - Ensemble RUSBoosted Trees **(S-ERT)**	**92%**	**75%**	High	88%	**88%**	**76%**	83%	**79%**
				Low	83%	**87%**	43%	67%	**52%**
				Middle	**78%**	**75%**	**84%**	73%	78%

Table 9 indicates that employing resampling techniques, particularly involving the Ensemble RUSBoosted Trees algorithm, yields superior and consistent performance across several evaluation metrics compared to Subspace Discriminant. RUSBoosted Trees demonstrate competence in both undersampling (ERT) and hybrid-sampling scenarios (S-ERT), recording the highest scores in various metrics. Meanwhile, S-ERT showcases exceptional performance in several metrics, including Validation and Test Accuracy, AUC score for Middle-class, and Accuracy in observation across all classes, Recall for High and Middle-classes, and F-Measure for High and Low-classes. Furthermore, a closer examination of Table 9, focusing on ERT and S-ERT, underscores S-ERT's superiority across multiple evaluation metrics. Consequently, in predicting students' Mathematics achievement at the Elementary School level, SMOTE-Ensemble RUSBoosted Trees (S-ERT) emerges as the optimal algorithm due to its superior accuracy and balanced results in identifying achievement levels across all categories. However, it is worth noting that S-ERT exhibits weakness in sensitivity to classify all Low-class, as evidenced by lower Recall scores. Thus, further confirmation and evaluation are recommended for Low-labeled students before intervention

Although resampling methods have enhanced accuracy, the issue of imbalanced class distribution in the testing set may persist. Therefore, improving the model's performance could involve utilizing a larger dataset, including testing sets encompassing diverse years or departments within the school, and exploring feature selection or employing additional ensemble learners combined with different resampling methods to enhance prediction accuracy.

D. Comparison with Existing Literature

The results of this study align with prior research, in particular to manage imbalanced educational datasets and improving prediction accuracy using hybrid sampling methods. For example, Bujang *et al.* [15] achieved a maximum accuracy of 62.0% for grade prediction using combination of SMOTE, Random Forest Feature Selection, and Multi-Layer Perceptron (MLP) on small datasets. In comparison, our hybrid model, combining SMOTE and RUSBoost, achieved a higher accuracy of 75%, highlighting its effectiveness in elementary Mathematics.

Similarly, Ghorbani and Ghousi [28] found that using SMOTE alone improved sensitivity for underrepresented classes and increased accuracy. While they identified SVM-SMOTE as the best-performing algorithm, our study found that SMOTE-Ensemble RUSBoosted Trees (S-ERT) produced the best results, indicating a divergence in findings and highlighting the adaptability of the proposed hybrid approach in addressing different datasets.

Additionally, prior studies like Al-Shehri *et al.* [23] showed that Support Vector Machines slightly outperformed KNN for high school performance prediction, which matches our findings. In our study, SVM reached 70% accuracy, while KNN was slightly lower at 67%. However, using resampling and ensemble techniques significantly enhanced performance, supporting Rawat and Malhan's [27] conclusion that ensemble models perform better than single algorithms on complex educational data.

The strong performance of the ensemble models in our study, particularly the SMOTE-RUSBoosted Trees, also aligns with Livieris *et al.* [26], who reported that ensemble methods achieved higher precision and recall than single models in predicting secondary school outcomes.

Overall, these findings show that the proposed hybrid ensemble model not only addresses class imbalance effectively but also provides a more reliable and accurate framework for predicting elementary Mathematics performance compared to previous methods.

V. Conclusion

Predicting students' Mathematics achievement in elementary school is crucial for early intervention. This study compares 10 classification algorithms using MATLAB to identify the best predictor. Resampling techniques, including SMOTE oversampling and RUSBoost undersampling, are applied due to class imbalance. SMOTE-RUSBoosted-Trees, a hybrid-sampling technique, emerges as the top performer with 75% accuracy, surpassing

RUSBoosted-Trees. This underscores the efficacy of ensemble learning combined with resampling for predicting students Mathematics achievement in unbalanced multi-class scenarios.

VI. Limitation and Future Direction

One limitation of the study is its reliance on a dataset of 280 students, which may not be large enough to generalize the findings across various educational contexts. Since the dataset is from one school, it may lack diversity in demographics, socio-economic backgrounds, and educational settings, limiting the results' applicability to other populations. Although the study used resampling techniques to address class imbalance, their effectiveness can vary depending on the dataset's specific characteristics. Highly skewed class distributions can lead to biased predictions for underrepresented classes.

Future research should collect data from a larger and more diverse sample of students across different schools and regions to enhance the generalizability of the findings and allow for a more robust analysis of factors influencing Mathematics achievement. Additionally, exploring other academic and non-academic features, such as emotional and psychological factors and extracurricular activities, could provide a more nuanced understanding of academic success. Advanced machine learning techniques like deep learning could offer improved predictive capabilities. Moreover, identifying the most influential factors affecting students' academic achievement in Mathematics at the elementary school level could guide targeted interventions and educational strategies.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Hendra Tjahyadi contributed to the research design, data analysis, revision, and finalization of the manuscript. Krismon N. L. Tude conducted the literature review, data collection, data analysis, and wrote the entire paper draft. All authors have approved the final version of this paper.

Acknowledgement

The authors are thankful for the support from the Faculty of Computer Science Universitas Pelita Harapan for providing MATLAB for data preprocessing and analysis. Appreciation is also extended to the principal of a private school in Tangerang, Indonesia, for granting access to its academic records for this research, with the condition that the school's name remains confidential. To uphold this agreement, the names of both the principal and the school are omitted from this acknowledgment.

References

[1] L. Ayebale, G. Habaasa, and S. Tweheyo, "Factors affecting students' achievement in Mathematics in secondary schools in developing countries: A rapid systematic review," *Stat J IAOS*, vol. 36, no. S1, pp. 73–76, Jan. 2020. doi: 10.3233/SJI-200713

[2] F. Stelzer, S. Vernucci, Y. Aydmune, M. Valle, M. L. Andres, and I. M. Introzzi, "Mathematics achievement in the last year of primary school. Longitudinal relationship with general cognitive skills and prior mathematics knowledge," *European Journal of Psychology of Education*, pp. 1–17, May 2023. doi: 10.1007/S10212-023-00700-W/TABLES/2

[3] E. A. Es, M. Cashen, T. Barnhart, and A. Auger, "Learning to notice Mathematics instruction: Using video to develop preservice teachers' vision of ambitious pedagogy," *Cogn. Instr.*, vol. 35, no. 3, pp. 165–187, Jul. 2017. doi: 10.1080/07370008.2017.1317125

[4] M. Abdurrahman, *Pendidikan Bagi Anak Berkesulitan Belajar*, Jakarta: Rineka Cipta, 2009.

[5] Yang Ditunggu, Ini Hasil Lengkap Asesmen Nasional 2021. (Jun. 22, 2023). [Online]. Available: https://www.detik.com/edu/edutainment/d-6011654/yang-ditunggu-ini-hasil-lengkap-asesmen-nasional-2021

[6] TIMSS 2015 and TIMSS Advanced 2015 International Results—TIMSS Reports. (Jun. 22, 2023). [Online]. Available: https://timss2015.org/

[7] Publications - PISA. (Jun. 22, 2023). [Online]. Available: https://www.oecd.org/pisa/publications/pisa-2018-results.htm

[8] A. Al Rashedi, A. Alnuaimi, M. Badri, and G. Yang, "Examining the relationships of factors influencing student mathematics achievement," *International Journal of Innovation in Education*, vol. 6, no. 1, p. 12, 2020. doi: 10.1504/IJIIE.2020.10027787

[9] M. S. Sumantri and R. Satriani, "The effect of formative testing and self-directed learning on mathematics learning outcomes," *International Electronic Journal of Elementary Education*, vol. 8, no. 3, pp. 507–524, 2016.

[10] A. A. Lipnevich, F. Preckel, and S. Krumm, "Mathematics attitudes and their unique contribution to achievement: Going over and above cognitive ability and personality," 2015. doi: 10.1016/j.lindif.2015.12.027

[11] A. Karakolidis, V. Pitsia, and A. Emvalotis, "Mathematics low achievement in Greece: A multilevel analysis of the Programme for International Student Assessment (PISA) 2012 data," *Themes in Science & Technology Education*, vol. 9, no. 1, pp. 3–24, 2016.

[12] B. Mayer, A. Blume, C. Black, and S. Stevens, "Improving student learning outcomes through community-based research: The poverty workshop," *Teach Sociol*, vol. 47, no. 2, pp. 135–147, Apr. 2019. doi: 10.1177/0092055X18818251

[13] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1–21, Dec. 2020. doi: 10.1186/S41239-020-0177-7/TABLES/15

[14] S. Alturki, N. Alturki, and H. Stuckenschmidt, "Using educational data mining to predict students' academic performance for applying early interventions," *Journal of Information Technology Education: Innovations in Practice*, vol. 20, pp. 121–137, 2021. doi: 10.28945/4835

[15] S. D. A. Bujang *et al.*, "Imbalanced classification methods for student grade prediction: A systematic literature review," *IEEE Access*, vol. 11, pp. 1970–1989, 2023. doi: 10.1109/ACCESS.2022.3225404

[16] X. Du, J. Yang, J. L. Hung, and B. Shelton, "Educational data mining: A systematic review of research and emerging trends," *Inf Discov Deliv*, vol. 48, no. 4, pp. 225–236, Oct. 2020. doi: 10.1108/IDD-09-2019-0070/FULL/XML

[17] A. Hellas *et al.*, "Predicting academic performance: A systematic literature review," *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, pp. 175–199, Jul. 2018. doi: 10.1145/3293881.3295783

[18] H. A. 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. doi: 10.1109/ACCESS.2020.2981905

[19] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Comput Sci*, vol. 72, pp. 414–422, 2015. doi: 10.1016/J.PROCS.2015.12.157

[20] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A systematic review of deep learning approaches to educational data mining," *Complexity*, vol. 2019, 2019. doi: 10.1155/2019/1306039

[21] M. A. Saleh, S. Palaniappan, N. Ali, and A. Abdalla, "Predicting student performance using data mining techniques in Libyan high schools," *Edukasi*, vol. 15, no. 2, pp. 91–100, Nov. 2021. doi: 10.15294/EDUKASI.V15I2.30068

[22] C. C. Kiu, "Data mining analysis on student's academic performance through exploration of student's background and social activities," in *Proc. 2018 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018*, Oct. 2018. doi: 10.1109/ICACCAF.2018.8776809

[23] H. Al-Shehri *et al.*, "Student performance prediction using support vector machine and K-nearest neighbor," in *Proc. Canadian*

Conference on Electrical and Computer Engineering, Jun. 2017. doi: 10.1109/CCECE.2017.7946847
[24] S. Huang and J. Wei, "Student performance prediction in Mathematics course based on the random forest and simulated annealing," *Sci Program*, vol. 2022, 2022. doi: 10.1155/2022/9340434
[25] E. G. Bayirli, A. Kaygun, and E. Öz, "An analysis of PISA 2018 Mathematics assessment for Asia-Pacific countries using educational data mining," *Mathematics 2023*, vol. 11, no. 6, p. 1318, Mar. 2023. doi: 10.3390/MATH11061318
[26] I. E. Livieris, K. Drakopoulou, V. T. Tampakas, T. A. Mikropoulos, and P. Pintelas, "Predicting secondary school students' performance utilizing a semi-supervised learning approach," *Journal of Educational Computing Research*, vol. 57, no. 2, pp. 448–470, Apr. 2019. doi: 10.1177/0735633117752614
[27] K. S. Rawat and I. V. Malhan, "A hybrid classification method based on machine learning classifiers to predict performance in educational data mining," *Lecture Notes in Networks and Systems*, vol. 46, pp. 677–684, 2019. doi: 10.1007/978-981-13-1217-5_67
[28] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020. doi: 10.1109/ACCESS.2020.2986809
[29] Hyperparameter Optimization in Classification Learner App—MATLAB & Simulink. (Jan. 08, 2024). [Online]. Available: https://www.mathworks.com/help/stats/hyperparameter-optimization-in-classification-learner-app.html
[30] Choose Classifier Options—MATLAB & Simulink. (Dec. 16, 2023). [Online]. Available: https://www.mathworks.com/help/stats/choose-a-classifier.html
[31] K. Wabang, O. D. Nurhayati, and Farikhin, "Application of the Naïve Bayes classifier algorithm to classify community complaints," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 872–876, Nov. 2022. doi: 10.29207/RESTI.V6I5.4498