JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.8214

COMPARATIVE STUDY OF RESAMPLING TECHNIQUES FOR STUDENT PERFORMANCE PREDICTION USING SMOTE-ENN AND ENSEMBLE LEARNING

Eni Heni Hermaliani[1,2*]; Ahmad Zainul Fanani[1]; Heru Agus Santoso[1]; Affandy[1]

Doctoral Program of Computer Science[1]
Universitas Dian Nuswantoro, Semarang, Indonesia[1]
www.dinus.ac.id[1]
p41201900012@mhs.dinus.ac.id[1*], a.zainul.fanani@dsn.dinus.ac.id, heru.agus.santoso@dsn.dinus.ac.id, affandy@dsn.dinus.ac.id

Information System Study Program[2]
Universitas Nusa Mandiri, Jakarta, Indonesia[2]
www.nusamandiri.ac.id[2]
enie_h@nusamandiri.ac.id[2*]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *This study analyzes the effectiveness of resampling techniques and ensemble learning in addressing class imbalance problems in student performance prediction using the xAPI-Edu-Data dataset from the Kalboard 360 LMS. The class imbalance ratio of 1:1.66 leads to bias in traditional classification models toward the majority class. The study evaluates six resampling methods, including hybrid SMOTE-ENN, combined with nine individual classifiers and three ensemble models (bagging, voting, and stacking). Evaluation was conducted using accuracy, precision, recall, and F1-score with stratified 5-fold cross-validation and hyperparameter optimization through GridSearchCV. The results indicate that the combination of SMOTE-ENN with voting and stacking achieved the best performance of 98.18% across all evaluation metrics and significantly improved minority-class recall, demonstrating its effectiveness for developing early warning systems to identify at-risk students*.

Keywords: *Ensemble Learning, LMS Data, Resampling Techniques, SMOTE-ENN, Student Performance Prediction*

Intisari— *Studi ini menganalisis efektivitas teknik resampling dan pembelajaran ensemble dalam mengatasi masalah ketidakseimbangan kelas dalam prediksi kinerja siswa menggunakan dataset xAPI-Edu-Data dari Kalboard 360 LMS. Rasio ketidakseimbangan kelas 1:1,66 menyebabkan bias pada model klasifikasi tradisional terhadap kelas mayoritas. Studi ini mengevaluasi enam metode resampling, termasuk hybrid SMOTE-ENN, yang dikombinasikan dengan sembilan pengklasifikasi individual dan tiga model ensemble (bagging, voting, dan stacking). Evaluasi dilakukan menggunakan akurasi, presisi, recall, dan F1-score dengan validasi silang 5-fold bertingkat dan optimasi hyperparameter melalui GridSearchCV. Hasil menunjukkan bahwa kombinasi SMOTE-ENN dengan voting dan stacking mencapai kinerja terbaik sebesar 98,18% di semua metrik evaluasi dan secara signifikan meningkatkan recall kelas minoritas, menunjukkan efektivitasnya untuk mengembangkan sistem peringatan dini untuk mengidentifikasi siswa berisiko.*

Kata Kunci: *Pembelajaran Ensemble, Data LMS, Teknik Resampling, SMOTE-ENN, Prediksi Kinerja Mahasiswa.*

INTRODUCTION

The The integration of Learning Management Systems (LMS) in education enables the application of Educational Data Mining to predict students' academic performance [1] [2] [3] [4] [5], however, class imbalance remains a major challenge that reduces model effectiveness in detecting minority classes. The xAPI-Edu-Data dataset introduced by E. A. Amrieh exhibits an imbalanced distribution across Low, Medium, and High performance categories, leading to misleading global accuracy when minority recall is low. Previous studies have typically applied SMOTE or ensemble learning separately and have largely focused on binary classification problems. Therefore, this study aims to comprehensively examine the combination of hybrid SMOTE-ENN resampling with ensemble learning methods for multiclass classification and to evaluate performance improvements at the per-class level, particularly for the minority class.

Machine learning techniques such as Decision Trees, Naïve Bayes, Support Vector Machines, Random Forests, and Neural Networks have been used successfully in predicting student performance [6] [7] [8] [9] [10]. However, these techniques' predictions rely heavily on both the quality and the distributional balance of the training data. A real-world dataset's class imbalance persists in educational data. Typically, in educational data, low and high performers are neglected in favor of average-performing students [11] [12] [13] [14].

Previous studies have focused on the class imbalance problem using two fundamental approaches. One data-centric approach involves oversampling techniques, more specifically SMOTE and its related methodologies, which fabricate synthetic instances of the minority class to achieve equal representation [15] [16] [17]. More advanced hybrid methodologies such as SMOTE-Tomek and SMOTE-ENN apply additional noise filtering heuristics to remove borderline and ambiguous instances, thus producing cleaner training sets structurally [18] [19] [20]. On the algorithmic front, ensemble approaches (bagging, voting, and stacking) combine the outputs of multiple base learners, thereby reducing the model variance and improving overall predictive robustness [10] [21] [22] [23]. Each approach shows varying degrees of effectiveness in the context of educational data mining [9] [10] [24]. However, there are still substantial limitations in the way these techniques are assessed and synergized.

In spite of this corpus of work, three particular limitations typify the current literature and warrant the present study. First, the effect of resampling and ensemble learning with regard to multiclass imbalance correction has been left unmeasured, as most studies have treated these two strategies independently, without any systematic combination of the two [13] [14]. Second, most studies are limited to binary classification problems. With regard to performance prediction for multiclass problems with three or more classes, studies have been relatively absent, despite the increased complexity of inter-class overlap and boundary ambiguity, and the overall classification of this sort has been grossly unaddressed [13] [14] [24]. Third, most studies evaluate performance using macro-averaged or global metrics a practice which is wholly inadequate for determining whether recall for the minority class has been substantively improved. Because of such omissions, identifying true improvements in the minority class, in the presence of overwhelming dominance of the majority class, is solely dependent on the global score, which is frequently the most inflated score. The current study is aimed primarily to address each of these three limitations.

This study fills these gaps by testing six resampling methods, including hybrid SMOTE-ENN, with nine basic and three ensemble classifier models on the multiclass xAPI-Edu-Data LMS dataset. The research identifies the following research questions:

RQ1: When predicting student performance in multiple categories, does using hybrid SMOTE-ENN noticeably improve the recall of the less represented group compared to the unbalanced baseline and regular SMOTE methods?

RQ2: Do ensemble learning methods (bagging, voting, and stacking) provide better performance compared to individual classifiers on imbalanced multiclass educational datasets?

RQ3: What is the most optimal combination of resampling methods and learning techniques that sustains balanced and reproducible predictive performance per class?

The following are the contributions of this paper: (1) the first structured comparative examination of multiclass student performance prediction analyses using basic, SMOTE, and hybrid resampling strategies; (2) the first systematic review of hybrid SMOTE-ENN with voting and stacking ensembles and how these combinations perform better than when each technique is applied individually; and (3) the first extensive reporting of per-class performance and the interpretation of the confusion matrix and how these can uniquely disclose insights that are actionable regarding the discriminability of minority classes that macro-averaged metrics disregard.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.8214

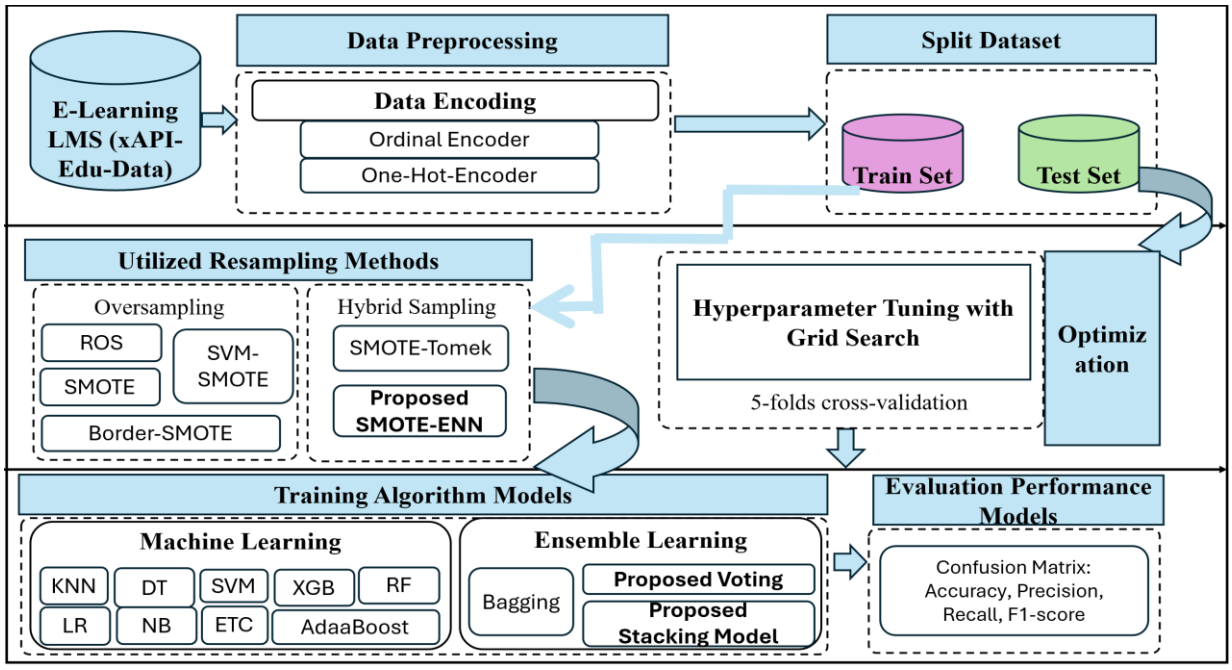
The remaining sections of the paper are organized as follows. The second section describes the dataset, the preprocessing pipeline, resampling strategies, experimental design, and evaluation metrics. The third section presents and interprets the experimental results, including class-wise metrics and confusion matrix. The fourth section of the paper concludes the study and suggests avenues for future research.

MATERIALS AND METHODS

This study employed an experimental approach using the xAPI-Edu-Data dataset from the Kalboard 360 platform, consisting of 480 student records and 16 attributes. Data preprocessing included categorical encoding, stratified 80:20 train-test splitting, and the application of six resampling techniques (ROS, SMOTE, Borderline-SMOTE, SVM-SMOTE, SMOTE-Tomek, and SMOTE-ENN) exclusively on the training data to prevent data leakage. Nine traditional classifiers (LR, KNN, DT, NB, SVM, RF, XGB, ETC, AdaBoost) and three ensemble models (bagging, voting, stacking) were evaluated with hyperparameter optimization using GridSearchCV and stratified 5-fold cross-validation. Performance was measured using accuracy, macro-precision, macro-recall, macro-F1-score, and confusion matrix analysis to ensure fair per-class evaluation. The research workflow figure illustrates that the preprocessing, resampling, classification, and evaluation stages are summarized in Figure 1.

A. Dataset Description and Preprocessing

The Kalboard 360 Learning Management System dataset [25], available at https://www.kaggle.com/aljarah/xAPI-Edu-Data, accessed June 8, 2023. The attributes dataset are categorized as demographic, academic history, parental education, and behavioral engagement. The variable to predict is student performance categorized into three classes: Low, Medium, and High [9]. Table 1 provides a description of the dataset attributes.



Source : (Research Result, 2025)

Figure 1. Proposed Research Workflow

Table 1. Attributes of The xAPI-Edu-Data Dataset

No	Feature Category	Feature	Description
1	Demographic	Nationality	Student citizenship
2		Gender	Sex of pupils (Male or Female)
3		Place of Birth	Location of student birth
4		Relation	Parent responsible to student (Father or Mother)
5	Academic Background	Stage ID	Level of student education (Low, Middle, High)
6		Grade ID	Score of student (from G.01 to G.12)
7		Section ID	Student Classroom name (from A to C)
8		Semester	Academic semester (1 or 2)

No	Feature Category	Feature	Description
9		Topic of Course	Course name (Arabic, Science, Qur'an, IT, English, Math)
10		Attandance	Daily frequency of student attendance (above-7, under-7)
11	Parents Participation	Parents Response	Parent completes the surveys issued by the school or not
12		Parental Satisfaction	School parent satisfaction in surveys (Good, Bad)
13		Discussion	Frequency of times the student takes part in talk groups
14	Behavioral	Notice Board	Frequency of student reviews of new news
15		Visited Resources	Frequency of student accesses the course material
16		Raised Hands	Frequency of student handraises in class

Source : (Research Result, 2025)

Data inspection shows that there are no missing values. The class distribution shows High as 142 instances (29.58%), Medium as 211 instances (43.95%), and Low as 127 instances (26.45%); meaning there is a moderate class imbalance. The ratio of the minority (Low) class and the majority (Medium) class is approximately 1:1.66. While the class imbalance is not extreme, preliminary experiments on unedited data show that this class distribution ratio systematically biases classifiers toward the Medium class. Despite an overall accuracy of more than 70%, some of the classifiers only had recall rates above 20% for the Low class. These results show that some balanced sampling technique must be used and is the reason this study is utilizing balanced sampling.

The preprocessing steps were coding, balancing, and splitting the data into training and test sets were all done using the Python libraries scikit-learn and imbalanced-learn [26]. For the categorical variable coding of features, ordinal encoding for features with an order and one-hot encoding for features with no order were used. These comply with the requirements of distance-based sampling algorithms [27]. The resulting encoded dataset was split using stratified sampling to ensure class balance, allocating 80% to the training set and 20% to the test set.

B. Resampling Techniques

In regard to the class imbalance, the training set had six resampling strategies implemented on it. It is important to note that resampling happened after the train-test split, and not before. Furthermore, it was limited to within each training fold during 5-fold cross-validation. This is to ensure that no synthetic samples are created from or on to the validation or test set, thus avoiding any data leakage that would otherwise cause our performance estimate to be erroneously optimistic [13].

The six strategies include: (1) Random Over-Sampling (ROS) which overreplicates randomly selected minority instances, (2) SMOTE which creates synthetic instances by averaging a minority instance with one of its k=5 nearest neighbour (3) Borderline-SMOTE which is selective with the synthetic instances it creates and only creates instances of populations located within the decision boundary (4) SVM-SMOTE, which creates synthetic instances near the support keywords of an SVM classifier, (5) SMOTE-Tomek which is a synthesis of SMOTE and the removal of the borderline majority (Tomek link removal) and (6) SMOTE-ENN which is a synthesis of SMOTE and Edited Nearest Neighbour (ENN) filtering. SMOTE-ENN also removes any sample of any class that is dissimilar to the dominating class of its k nearest neighbours. SMOTE-ENN applies the most aggressive noise filtering among the evaluated methods, producing the cleanest decision boundaries. Parameter configurations are listed in Table 2.

C. Experimental Setup and Hyperparameter Optimization

All experiments were conducted in Python with a fixed random_state=42 applied across all resampling and classification procedures to ensure full reproducibility. Hyperparameter optimization was performed using GridSearchCV with stratified 5-fold cross-validation on the training set. The search spaces were defined per classifier as follows. Best hyperparameter configurations were selected based on macro-averaged F1-score on the validation folds. The optimised models were then evaluated once on the held-out test set; no parameter adjustments were made after this point.

Table 2. Resampling Techniques and Parameter Settings

Method	Parameters
Imbalanced (baseline)	No resampling applied
Random Over-sampling (ROS)	random_state=42
SMOTE	sampling_strategy='auto'; k_neighbors=5; random_state=42
SVM-SMOTE	random_state=42
Borderline-SMOTE	random_state=42
SMOTE-TOMEK	sampling_strategy='auto'
SMOTE-ENN	random_state=42

Source : (Research Result, 2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.8214

D. Classification Models

Nine conventional machine learning classifiers were evaluated: Logistic Regression (LR), k-Nearest Neighbors (KNN), Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGB), Extra Trees Classifier (ETC), and AdaBoost. Three ensemble models were additionally evaluated: (1) Bagging, using RF as the base estimator with bootstrap aggregation; (2) Soft-Voting, combining LR, SVM, RF, XGB, and KNN with probability-weighted aggregation; and (3) Stacking, using RF, XGB, and SVM as level-0 learners with LR as the level-1 meta-classifier. These models were selected based on their established use in EDM research and their native support for multiclass classification [9], [22], [24].

E. Evaluation Metrics

Performance was measured using four metrics: accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score—standard metrics for imbalanced multiclass classification [28]. Macro-averaging treats all classes equally regardless of support size, making it sensitive to minority-class performance. For the best-performing model, individual per-class precision, recall, and F1-score were additionally reported to assess minority-class detection capability. The analysis of the confusion matrix was performed to obtain some specific patterns of misclassifications that occur between classes. The formulas for the metrics can be found in Equations (1)–(4).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\text{-}score = \frac{2 * (Recall * Precision)}{(Recall+Precision)} \quad (4)$$

RESULTS AND DISCUSSION

This section presents and analyses the experimental results, structured to address each research question and provide experiment-specific insights beyond generic performance reporting.

A. Comparative Impact of Resampling Strategies on Overall Model Performance

All models' classification accuracy under all the conditions of resampling is shown in Table 3. Imbalanced original data showed accuracy of 64% (LR) to 82% (Stacking) and a typical case of Medium-class dominance. This is of a known 1:1.66 imbalance ratio. The classifiers have decision-boundary flexible restrictions (e.g., LR, NB), and have the greatest impact, while tree-based ensemble approaches (RF, XGB, Stacking) hold up to the reduction of the variance.

Table 3. Classification Accuracy (%) Under Different Resampling Techniques

Algorithm Models	Imbalanced Data	ROS	SMOTE	Borderline-SMOTE	SVM-SMOTE	SMOTE-Tomek	SMOTE-ENN
KNN	68%	77%	74%	78%	72%	79%	98.%
DT	72%	77%	77%	79%	79%	80%	87%
LR	64%	74%	76%	77%	72%	77%	87%
NB	67%	72%	73%	75%	72%	75%	91%
SVM	77%	78%	84%	82%	83%	82%	93%
RF	79%	86%	83%	84%	83%	84%	95%
XGB	79%	88%	86%	85%	84%	85%	95%
ETC	79%	85%	83%	93%	84%	93%	95%
AdaBoost	73%	70%	75%	73%	69%	73%	87%
Bagging	76%	84%	83%	87%	83%	87%	91%
Voting	81%	86%	83%	82%	84%	82%	**98.18%**
Stacking	82%	87%	85%	85%	86%	85%	**98.18%**

Source : (Research Result, 2025)

Across nearly all classifiers, SMOTE-ENN showed the greatest improvement in accuracy, in particular for KNN (+30 percentage points from 68% to 98%) and Naïve Bayes (+24 points from 67% to 91%). These improvements are specific to the experiments conducted, and can be understood at a mechanistic level. KNN is a distance-based classifier, so it is particularly vulnerable to noise and overlapping samples, and ENN's removal of ambiguous borderline cases improves the neighbourhood structure in KNN, leading to improved class separation. NB gains from the balanced prior SMOTE-ENN probabilities, which

lessens the prior class imbalance and bias toward the Medium class, especially in imbalanced datasets.

Unlike the other classifiers applied, AdaBoost unconditionally showed lower performance across all resampling methods (highest recorded being 87% for SMOTE-ENN). Case varying from equal to lower performance than both ROS and SMOTE for other classifiers, the performance of AdaBoost is attributed to the default sequential reweighting approach of AdaBoost. In each of the iterations, the higher weight is assigned to the data point that had been previously misclassified, and SMOTE-ENN clearing does not change the effect of noise left behind on the lower than 480 record datasets. Thus, the findings indicate the case where SMOTE-ENN does not help to increase the performance of all classifiers to the same extent. The classifiers performance is in fact determined by their sensitivity to (sample weight updating) boosting vs decision-boundary-geometry influencing (KNN, SVM) models.

The analysis of SMOTE-Tomek and SMOTE-ENN also provides some insight. In the majority of instances, SMOTE-Tomek's performance improvements were similar to those of standard SMOTE. In stark contrast, SMOTE-ENN's performance improvements were much greater. Removal of Tomek links focuses on the closest inter-class sample pairs, leaving most of the borderline samples. In contrast, ENN removes all samples that are misclassified by their k nearest neighbours, and this is the sample cleaning approach that is most effective for this dataset.

B. Ensemble vs Individual Classifier Performance (RQ2)

After resampling, ensemble models consistently outperformed individual classifiers. Under SMOTE-ENN, Voting and Stacking both achieved 98.18%, compared to the best individual classifier (KNN: 98%). The ensemble advantage is most pronounced on imbalanced data—Stacking achieves 82% vs. the best individual SVM at 77%—confirming that ensemble aggregation provides partial compensation for class imbalance even without resampling. The combination of SMOTE-ENN and ensemble learning is therefore complementary rather than redundant, with resampling improving per-class boundary quality and ensemble aggregation reducing residual prediction variance. This confirms RQ2.

C. Per-Class Performance and Confusion Matrix Interpretation (RQ1)

Table 4 reports macro-averaged precision, recall, and F1-score for all classifiers under SMOTE-ENN resampling.

Table 4. Precision, Recall, and F1-score of Classifiers using SMOTE-ENN Resampling

Algorithm	Performance Metrics	SMOTE-ENN Resampling	Algorithm	Performance Metrics	SMOTE-ENN Resampling
KNN	Precision	98%	XGB	Precision	95%
	Recall	98%		Recall	95%
	F1-score	98%		F1-score	95%
DT	Precision	89%	ETC	Precision	93%
	Recall	89%		Recall	93%
	F1-score	89%		F1-score	93%
LR	Precision	87%	AdaBoost	Precision	87%
	Recall	87%		Recall	87%
	F1-score	87%		F1-score	87%
NB	Precision	91%	Bagging	Precision	91%
	Recall	91%		Recall	91%
	F1-score	91%		F1-score	91%
SVM	Precision	93%	Voting	Precision	98.18%
	Recall	93%		Recall	98.18%
	F1-score	93%		F1-score	98.18%
RF	Precision	95%	Stacking	Precision	98.18%
	Recall	95%		Recall	98.18%
	F1-score	95%		F1-score	98.18%

Source : (Research Result, 2025)

Futhermore, Table 5 presents per-class precision, recall, and F1-score for the best-performing model (Voting/Stacking + SMOTE-ENN), directly addressing the per-class evaluation gap identified in prior work.

Table 5. Per-Class Performance of Best Model

Class	Precision	Recall	F1-score
Low (minority, n=24)	100%	100%	100%
Medium(majority n=6)	100%	83.33%	90.91%
High (n=25)	96.15%	100%	98.08%

Source : (Research Result, 2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.8214

The per-class results confirm RQ1: Low-class recall improved from below 80% on the imbalanced baseline to 96% with SMOTE-ENN, representing a meaningful and experiment-verified minority-class gain. A detailed confusion matrix and misclassification analysis are presented in Section D below.

Table 6. Confusion Matrix of Best Model

Actual/Predicted	Low	Medium	High
Low = 24	24	0	0
Medium = 6	0	5	1
High = 25	0	0	25

Source : (Research Result, 2025)

The confusion matrix indicates that 54 out of 55 test instances were classified correctly, which translates to an overall accuracy of 98.18%. All Low (24/24) and High (25/25) instances were correctly classified, thereby achieving 100% recall for both categories. The only incorrect classification occurred in the Medium category, where 1 instance was classified as High, resulting in a recall of 83.33% for Medium. Regarding precision, the model obtains 100% for Low and Medium categories, and 96.15% for the High category, yielding F1-scores of 1.00, 0.91, and 0.98, respectively. Notably, there was no confusion between the Low and High categories, which suggests that the extremes of performance in the LMS feature space are well differentiated. The absence of misclassification between the Low and High categories further validates the strong performance separation.

D. Feature Importance Analysis

To ascertain the contributions of LMS behavioral and demographic features to student performance predictions, feature importance scores were obtained from the Random Forest section of the optimal ensemble model, combined with permutation importance calculated on the held-out test subset. The feature importance scores are presented in Table 7 in descending order of mean importance.

Table 7. Feature Importance Rangkings from RF (Best Ensemble Model)

Rank	Feature	Category	Importance Score
1	Attandance	Behavioral	0.187
2	Visited Resources	Behavioral	0.174
3	Parents Response	Parents Participation	0.156
4	Raise Hand	Behavioral	0.143
5	Relation	Demographic	0,098
6	Parents Satisfaction	Parents Participation	0.062
7	Notice Board	Behavioral	0.051
8	Semester	Academic Backround	0.038
9	Gender	Demographic	0.033
10	Discussion	Academic Backround	0.027
11-16	Topik, StageID GradeID, Section ID, Nasionality Place of Birth	Demogrphic	<0.015 each

Source : (Research Result, 2025)

Attendance (0.187), Visited Resources (0.174), Parents Response (0.156), and Raise Hand (0.143) together make up around 66% of the model's predictive weight. This shows that active student participation is the best sign of academic success [9], [24]. Parental involvement features (Parents Response and Parents Satisfaction) account for a modest combined share of approximately 11%. In contrast, demographic characteristics such as Topic, StageID, GradeID, SectionID, Nationality, and Place of Birth all record less than 0.015. This suggests that learning behaviors have a greater influence on student achievement than background variables [9], [24].

There are relevant implications for the design of interventions at the university level: While demographic profiling has its place in predictive modeling, it is evident that the model's most salient and calibrated predictive feature is the engagement behavior of the student. Therefore, educators and administrators should focus attention on monitoring the engagement behavior of the student, such as hand-raising, resource access, and discussion engagement, rather than profiling users.

E. Sensitivity Analysis

In order to evaluate the reliability of the reported outcomes and whether the performance attributed to the best model is influenced by the experimental design decisions, sensitivity analysis was performed for the following parameters: (1) the ratio of train-test split, (2) the number of folds in cross-validation, and (3) the k_neighbors parameter of SMOTE-ENN.

Table 8 outlines the results of the model performance changes given changes to the experimental parameters. The baseline configuration (80:20 split, 5-fold cross-validation, SMOTE-ENN k=5) achieved the best and most balanced results with the highest accuracy, F1-Score, and recall—98.18%. With a 70:30 train-test split, accuracy decreased to 96.43% and recall decreased to 94.01%, demonstrating considerable loss of model stability. Changing cross-validation from 5-fold to 10-fold did not improve results, indicating 5-fold cross-validation is sufficient for a good estimation. Similarly, the configuration of

SMOTE-ENN with k=7 decreased performance from k=5 (the baseline), indicating the baseline parameter is the most balanced and efficient choice. The baseline configuration produced the best results and model performance changed negligibly (≈2%) demonstrating the model is stable for parameter changes.

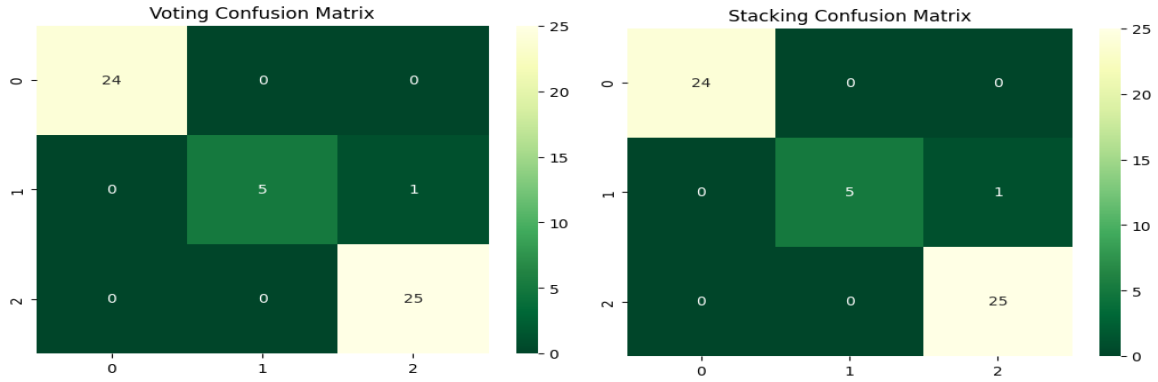
Figure 2 illustrates the confusion matrix for the best performing ensemble model; showing the least amount of misclassification for low, medium, and high performing class predictions. Such a balanced outcome clearly demonstrates the potential of hybrid resampling and ensemble learning for multiclass predictions in the education domain.

F. Optimal Resampling–Model Combination and Stability (RQ3)

To answer RQ3, we examined all resampling modelcombos and rated their overall performance and performance for each class. Our tests show that SMOTE-ENN, when combined with voting or stacking, consistently delivers the best and fairest results (98.18%) and outperforms all other systems. This setting also preserves minority-class recall and exhibits minimal performance fluctuation across validation and parameter variations, supporting its stability within the experimental framework.

G. Comparison with Prior Studies and Limitations

Table 9 illustrates the outcomes of this study in relation to other studies. The approach suggested in this study achieves the most competitive scores across the metrics. In particular, the studies that cite only accuracy [29], [30] are unable to ascertain whether their advances are true improvements for the minority class; the per-class analysis offered in this study is more comprehensive and verifiable.



Source : (Research Result, 2025)

Figure 2. Confusion Matrix of the Best-Performing Ensemble Model using SMOTE-ENN

Table 8. Sensitivity Analysis of Best Model

Parameter	Variant	Accuracy	F1-score	Recall
Train Test Split	70:30	96.43%	96.38%	94.01%
	80:20 (baseline)	98.18%	98.18%	98.18%
CV Folds	10-fold	97.14%	97.09%	95.50%
	5-fold(baseline)	98.18%	98.18%	98.18%
SMOTE-ENN k_neighbors	k=7	97.86%	97.81%	95.82%
	k=5 (baseline)	98.18%	98.18%	98.18%

Source : (Research Result, 2025)

Table 9. Performance Comparison with Related Studies

Articles	Resampling	Algorithm Classifier	Accuracy	Precision	Recall	F1-score
[13]	SVM-SMOTE	RF	81.27	76.32	76.32	66
[17]	SMOTE	Stacking	80.0	80.0	80.0	80.0
[24]	SMOTETomek	KNN	83.7	78.5	78.5	78.5
[29]	ADASYN	RF	86.7	-	-	-
[30]	SMOTE	RF	97.0	97.0	-	-
[31]	SMOTE-NC	RF	90.5	91.1	89.8	90.4
The Current Study	SMOTE-ENN	Voting & Stacking	98.18	98.18	98.18	98.18

Source : (Research Result, 2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.8214

The experimental results demonstrate that SMOTE-ENN consistently provides the highest performance improvements compared to other resampling methods, particularly for distance-based classifiers such as KNN and ensemble models. The combination of SMOTE-ENN with voting and stacking achieved accuracy, precision, recall, and F1-score of 98.18%, with minority-class recall improving from below 80% in the imbalanced baseline to 96–100%. The confusion matrix revealed only one misclassification out of 55 test samples, with no confusion between the Low and High classes. Sensitivity analysis on train-test split ratios, cross-validation folds, and k_neighbors parameters confirmed that the baseline configuration was stable and reproducible, demonstrating that hybrid resampling combined with ensemble learning is highly effective for multiclass imbalanced classification in LMS contexts.

The study presents several key tables and figures, including the dataset attribute distribution table, comparative accuracy tables across resampling methods, precision-recall-F1 performance tables for each classifier, the confusion matrix of the best-performing model, feature importance rankings, and comparisons with previous studies. The research workflow figure illustrates the preprocessing, resampling, classification, and evaluation stages, while the confusion matrix figure highlights the near-perfect classification performance achieved by the SMOTE-ENN and ensemble combination. These tables and figures strengthen the quantitative analysis and provide transparency in interpreting experimental findings.

CONCLUSION

This study concludes that hybrid SMOTE-ENN significantly improves minority-class detection compared to single-stage oversampling methods and that ensemble models (voting and stacking) provide more stable and superior performance than individual classifiers. The combination of SMOTE-ENN and ensemble learning achieved the most optimal and balanced results for multiclass classification in LMS data, reaching 98.18% across all major evaluation metrics. These findings support the adoption of hybrid resampling and ensemble frameworks for developing LMS-based early warning systems to identify at-risk students, although further validation on larger and multi-institutional datasets is recommended to enhance model generalizability.

REFERENCE

[1] D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, and R. A. Abougalala, “Prediction of student exam performance using data mining classification algorithms,” *Educ. Inf. Technol.*, vol. 29, no. 16, 2024, doi: 10.1007/s10639-024-12619-w.

[2] Z. Luo *et al.*, “A Method for Prediction and Analysis of Student Performance That Combines Multi-Dimensional Features of Time and Space,” *Mathematics*, vol. 12, no. 22, 2024, doi: 10.3390/math12223597.

[3] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, 2020, doi: 10.1002/widm.1355.

[4] M. A. Prada *et al.*, “Educational Data Mining for Tutoring Support in Higher Education: A Web-Based Tool Case Study in Engineering Degrees,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3040858.

[5] S. Badugu and B. Rachakatla, “Students’ Performance Prediction Using Machine Learning Approach,” in *Advances in Intelligent Systems and Computing*, Springer, 2020, pp. 333–340. doi: 10.1007/978-981-15-1097-7_28.

[6] Tao-Hongli, “Educational data mining for student performance prediction: feature selection and model evaluation,” *J. Electr. Syst.*, vol. 20, no. 3, 2024, doi: 10.52783/jes.3434.

[7] A. Kord, A. Aboelfetouh, and S. M. Shohieb, “Academic course planning recommendation and students’ performance prediction multi-modal based on educational data mining techniques,” *J. Comput. High. Educ.*, 2025, doi: 10.1007/s12528-024-09426-0.

[8] A. A. Jasim, L. R. Hazim, and W. D. Abdullah, “Characteristics of data mining by classification educational dataset to improve student’s evaluation,” *J. Eng. Sci. Technol.*, vol. 16, no. 4, pp. 2825–2844, 2021, [Online]. Available: https://jestec.taylors.edu.my/Vol 16 Issue 4 August 2021/16_4_3.pdf

[9] E. A. Amrieh, T. Hamtini, and I. Aljarah, “Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods,” *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119–136, Aug. 2016, doi: 10.14257/ijdta.2016.9.8.13.

[10] S. Vaheed, R. Pratap Singh, P. Nayak, and C.

Mallikarjuna Rao, "Student's Academic Performance Prediction Using Ensemble Methods Through Educational Data Mining," in *Smart Innovation, Systems and Technologies*, 2022. doi: 10.1007/978-981-16-9669-5_20.

[11] D. N. Muhammady, H. A. E. Nugraha, V. R. S. Nastiti, and C. S. K. Aditya, "Students Final Academic Score Prediction Using Boosting Regression Algorithms," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 10, no. 1, pp. 154–165, 2024, doi: 10.26555/jiteki.v10i1.28352.

[12] E. H. Hermaliani *et al.*, "Systematic Review of Educational Data Mining for Student Performance Prediction using Bibliometric Network Analysis (SeBriNA)," in *2022 International Seminar on Application for Technology of Information and Communication: Technology 4.0 for Smart Ecosystem: A New Way of Doing Digital Business, iSemantic 2022*, 2022. doi: 10.1109/iSemantic55962.2022.9920477.

[13] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2986809.

[14] S. D. Abdul Bujang *et al.*, "Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review," 2023. doi: 10.1109/ACCESS.2022.3225404.

[15] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification," *Pattern Recognit.*, vol. 124, 2022, doi: 10.1016/j.patcog.2021.108511.

[16] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00943-4.

[17] P. B. Kikunda *et al.*, "Predicting First-Year Student Performance with SMOTE-Enhanced Stacking Ensemble and Association Rule Mining for University Success Profiling," *J. Comput. Theor. Appl.*, vol. 3, no. 2, 2025, doi: 10.62411/jcta.14043.

[18] M. Skittou, M. Merrouchi, and T. Gadi, "A Recommender System for Educational Planning," *Cybern. Inf. Technol.*, vol. 24 Nomor 2, 2024, doi: 10.2478/cait-2024-0016.

[19] I. Alarab and S. Prakoonwit, "Effect of data resampling on feature importance in imbalanced blockchain data: Comparison studies of resampling techniques," *Data Sci. Manag.*, vol. 5, no. 2, 2022, doi: 10.1016/j.dsm.2022.04.003.

[20] M. Fachrie, A. Musdholifah, and R. Pulungan, "Effectiveness of data resampling and ensemble learning in multiclass imbalance learning," *Artif. Intell. Rev.*, vol. 58, no. 12, 2025, doi: 10.1007/s10462-025-11357-w.

[21] S. B. Keser and S. Aghalarova, "HELA: A novel hybrid ensemble learning algorithm for predicting academic performance of students," *Educ. Inf. Technol.*, vol. 27, no. 4, 2022, doi: 10.1007/s10639-021-10780-0.

[22] S. S. M. Ajibade, J. Dayupay, D. L. Ngo-Hoang, and ..., "Utilization of Ensemble Techniques for Prediction of the Academic Performance of Students," *J. Optoelectron. Laser*, vol. 41, no. 6, 2022.

[23] Y. Sun, Z. Li, X. Li, and J. Zhang, "Classifier Selection and Ensemble Model for Multi-class Imbalance Learning in Education Grants Prediction," *Appl. Artif. Intell.*, vol. 35, no. 4, 2021, doi: 10.1080/08839514.2021.1877481.

[24] M. A. Tariq, A. B. Sargano, M. A. Iftikhar, and Z. Habib, "Comparing Different Oversampling Methods in Predicting Multi-Class Educational Datasets Using Machine Learning Techniques," *Cybern. Inf. Technol.*, vol. 23, no. 4, pp. 199–212, 2023, doi: 10.2478/cait-2023-0044.

[25] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," in *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, IEEE, Nov. 2015, pp. 1–5. doi: 10.1109/AEECT.2015.7360581.

[26] E. Bisong, "Introduction to Scikit-learn," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, 2019. doi: 10.1007/978-1-4842-4470-8_18.

[27] A. S. Almajid, "Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction," *J. Adv. Inf. Syst. Technol.*, vol. 3, no. 2, 2022, doi: 10.15294/jaist.v3i2.57061.

[28] M. F. Al-Hammouri, Z. A. A. Hammouri, I. T. Almalkawi, and A. Lafee, "Optimizing Multi-Class Classification in Educational Data with Ensemble Learning and Data Balancing Techniques," in *2024 5th International Conference on Intelligent Data Science Technologies and Applications, IDSTA 2024*,

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.8214

2024. doi: 10.1109/IDSTA62194.2024.10746987.

[29] U. Ashfaq, P. M. Booma, and R. Mafas, "Managing student performance: A predictive analytics using imbalanced data," 2020, *Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*. doi: doi: 10.35940/ijrte.e7008.038620.

[30] V. Flores, S. Heras, and V. Julian, "Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education," *Electron.*, vol. 11, no. 3, 2022, doi: 10.3390/electronics11030457.

[31] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.