

An Integrated K-Means and SVM Framework for Binary Prediction of Student Grades

DOI: <http://dx.doi.org/10.35889/progresif.v22i3.3717>

Creative Commons License 4.0 (CC BY –NC)



Rosellie Simbulan^{1*}, Arief Hermawan², Donny Avianto³

¹Program Magister Teknologi Informasi, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

²Departemen Sains & Teknologi, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

*e-mail Corresponding Author: rosyliesimbultan.24@gmail.com

Abstract

Selecting students for academic competitions and enrichment programs is often based on teachers' subjective observations, which may overlook students with strong academic potential. This study proposes a Hybrid K-Means+SVM framework for predicting Grade A achievement using academic records of 281 Grade X Informatics students. To address the limited separability of homogeneous grade distributions, the prediction task was reformulated as a binary classification problem (Grade A vs. Non-A). K-Means clustering generated three soft-cluster distance features ($K = 3$; Silhouette Score = 0.35), which were combined with academic features and classified using an RBF-kernel SVM. Class imbalance was handled using SMOTE. The proposed model outperformed the baseline SVM and achieved an AUC-ROC of 0.8541 on the test set. Probability threshold tuning further improved the minority-class F1-score from 0.5714 to 0.6154, precision from 0.44 to 0.50 and accuracy from 0.8310 to 0.8592. The results demonstrate the effectiveness of the proposed approach for identifying high-achieving students from educational data.

Keywords: Educational Data Mining; K-Means Clustering; Support Vector Machine; Student Performance Prediction; Class Imbalance

Abstrak

Pemilihan siswa untuk mengikuti kompetisi akademik dan program pengayaan sering kali didasarkan pada pengamatan subjektif guru sehingga berpotensi mengabaikan siswa yang memiliki kemampuan akademik tinggi. Penelitian ini mengusulkan kerangka kerja Hybrid K-Means+SVM untuk memprediksi pencapaian Grade A berdasarkan data akademik 281 siswa kelas X Informatika. Untuk mengatasi rendahnya separabilitas pada distribusi nilai yang homogen, tugas prediksi direformulasi menjadi klasifikasi biner (Grade A dan Non-A). K-Means digunakan untuk menghasilkan tiga fitur jarak soft-cluster ($K = 3$; Silhouette Score = 0,35), yang kemudian digabungkan dengan fitur akademik dan diklasifikasikan menggunakan SVM berkernel RBF. Ketidakseimbangan kelas ditangani menggunakan SMOTE. Model yang diusulkan mengungguli SVM dasar dan mencapai AUC-ROC sebesar 0,8541 pada data uji. Optimasi threshold probabilitas meningkatkan F1-score kelas minoritas dari 0,5714 menjadi 0,6154, presisi dari 0.44 menjadi 0.50 dan akurasi dari 0.8310 menjadi 0.8592. Hasil penelitian menunjukkan efektivitas pendekatan yang diusulkan untuk mengidentifikasi siswa berprestasi berdasarkan data akademik.

Kata kunci: Educational Data Mining; K-Means Clustering; Support Vector Machine; Prediksi Performa Siswa; Ketidakseimbangan Kelas

1. Introduction

Educational Data Mining (EDM) has become an important research area for supporting evidence-based educational decision-making. One of its primary applications is student academic performance prediction, which enables schools to identify students who require academic

intervention as well as students with high achievement potential who may benefit from enrichment programs [1], [2]. The increasing availability of digital academic records has encouraged the adoption of machine learning techniques for educational prediction tasks because such methods can discover complex patterns that are difficult to identify through conventional statistical analysis [2], [3]. Accurate prediction models are therefore valuable for improving learning outcomes, optimizing resource allocation, and supporting data-driven educational management.

This study focuses on academic performance prediction using data from 281 Grade X Informatics students at a senior high school in Yogyakarta, Indonesia, during the 2024/2025 academic year. Based on the school assessment policy, students are categorized into Grade A (≥ 86), Grade B (80–85), and Grade C (< 80). The dataset presents two important challenges. First, the final grades are relatively homogeneous, ranging only from 75 to 92, which makes class boundaries difficult to separate. Second, the dataset is highly imbalanced, with only 41 Grade A students compared to 240 students in Grades B and C (5.9:1 ratio). Such conditions may reduce the effectiveness of conventional multi-class classification approaches because the minority class tends to be underrepresented and difficult to identify accurately [4][5]. Consequently, a prediction framework capable of improving minority-class detection while preserving discriminative information is required.

Several studies have investigated machine learning methods for student performance prediction. Alsariera et al. [3] compared Decision Tree, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, and Support Vector Machine (SVM) using academic, demographic, and behavioral attributes, reporting competitive performance from SVM. Khairy et al. [6] employed Random Forest, Decision Tree, and K-Nearest Neighbor to predict examination outcomes using academic history and learning engagement indicators. Yağcı [5] evaluated multiple classification algorithms, including Decision Tree, Random Forest, Naïve Bayes, and SVM, and highlighted the difficulty of maintaining predictive performance when score distributions are highly homogeneous. Similarly, Novianto and Suhirman [7] compared K-Nearest Neighbor and SVM for predicting students' study duration, further reinforcing SVM's relative effectiveness over instance-based methods in educational prediction tasks. Beyond supervised learning, Pamungkas et al. [8] applied K-Means clustering to group students according to academic characteristics, while Wu [9] demonstrated that cluster-derived distance information can be incorporated as additional predictive features to improve classification performance. Although these studies have demonstrated the potential of machine learning and clustering techniques in educational prediction, several research gaps remain. First, binary reformulation of homogeneous student grade prediction has received limited attention despite its potential to alleviate multi-class separability problems. Second, the integration of K-Means soft-cluster distance features with SVM classifiers remains underexplored in student grade prediction research. Third, threshold optimization for improving minority-class detection under severe class imbalance conditions is rarely incorporated into educational prediction frameworks.

To address these gaps, this study proposes a Hybrid K-Means and Support Vector Machine (SVM) framework for binary prediction of student grades. The proposed approach reformulates the original three-grade classification problem into a binary prediction task consisting of Grade A and Non-A categories, thereby focusing the model on the most educationally relevant decision boundary. K-Means clustering is employed to generate soft-cluster distance features that capture latent academic performance structures, while SVM is used as the classifier due to its effectiveness in handling non-linear decision boundaries and relatively small datasets [3], [5]. Furthermore, probability threshold optimization is applied to improve the detection of minority-class students. The main contributions of this study are: (1) reformulating homogeneous student grade prediction into a binary classification problem; (2) integrating K-Means soft-cluster distance features within an SVM-based prediction framework; and (3) applying threshold optimization to improve Grade A detection in imbalanced educational data. Accordingly, this study aims to develop and evaluate the proposed Hybrid K-Means+SVM model, quantify the contribution of cluster-derived features through an ablation study, and assess the effectiveness of threshold optimization for identifying high-achieving students.

2. Method

This study adopts a computational performance evaluation approach to assess the effectiveness of a hybrid K-Means Clustering and Support Vector Machine (SVM) framework for binary student grade prediction. The proposed framework was implemented through eight

consecutive stages. The process begins with data collection and preprocessing, followed by feature engineering and binary label generation. K-Means clustering is then employed to generate soft-cluster distance features, which are integrated with engineered academic features to form a hybrid feature set. Subsequently, an SVM classifier is developed and optimized using Grid Search. Model performance is evaluated through cross-validation and hold-out testing using Accuracy, F1-Score, and AUC-ROC metrics. Finally, the trained model is applied to predict the probability of achieving Grade A and to identify students requiring academic intervention. The overall research workflow is presented in Figure 1.

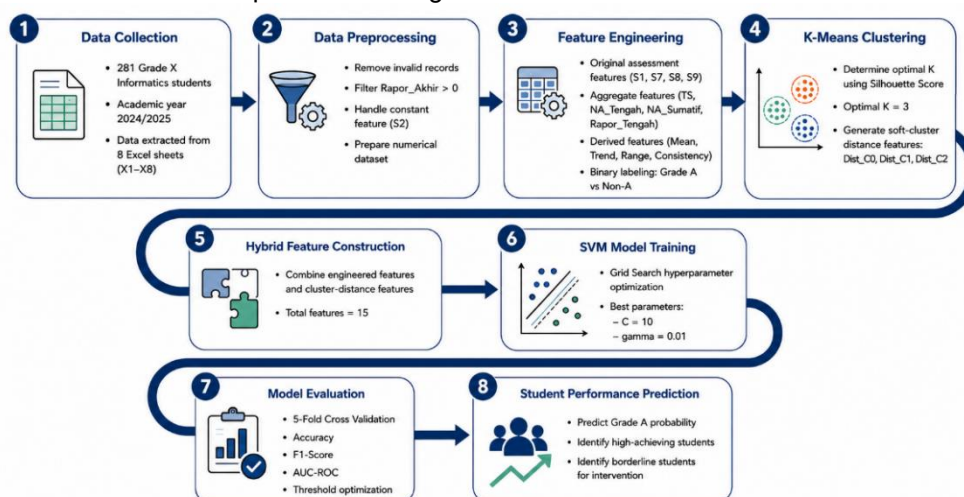


Figure 1. Research workflow of the proposed Hybrid K-Means and SVM framework for binary student grade prediction

2.1 Dataset and Research Variables

The dataset consists of academic assessment records of 281 Grade X Informatics students from a senior high school in Yogyakarta, Indonesia, during the 2024/2025 academic year. The data were collected from official school assessment documents across eight classes (X1–X8). After removing records with invalid final grades ($Rapor_Akhir \leq 0$), 281 student records were retained for analysis. Final report card grades ranged from 75 to 92, with an average score of 81.2.

In the Indonesian secondary school assessment system, students receive periodic formative assessment scores throughout the semester. Variables S1–S9 represent weekly formative assessment scores obtained from nine instructional topics in the Grade X Informatics course. Each score was recorded after the completion of a weekly learning session and reflects students' mastery of the corresponding topic. These variables provide a longitudinal representation of students' academic progress and serve as the primary indicators used for feature engineering and predictive modeling.

The proposed framework utilizes 12 baseline features derived from assessment scores and engineered academic indicators. The variable S2 was excluded because it contained no variation across students. To capture latent academic performance patterns, K-Means clustering was applied to generate three soft-cluster distance features (Dist_C0, Dist_C1, and Dist_C2). These features were combined with the baseline variables to form the hybrid feature set used by the SVM classifier. Table 1 summarizes all research variables.

The prediction task was formulated as a binary classification problem. Students with $Rapor_Akhir \geq 86$ were labeled as Grade A (1), while those with $Rapor_Akhir < 86$ were labeled as Non-A (0).

Table 1. Description of Research Variables

Variable	Type	Description
S1-S9	Numeric	Weekly formative assessment scores obtained from Topics 1–9 of the Grade X Informatics course
NA_Tengah	Numeric	Average mid-semester assessment score
NA_Sumatif	Numeric	Average summative score

Variable	Type	Description
TS	Numeric	Mid-Semester Summative Assessment Score
Rapor_Tengah	Numeric	Mid-semester report card grade
S_Akhir_Mean	Engineered	Mean of S7, S8, and S9
Trend_Akhir	Engineered	Difference between S_Akhir_Mean and S1
Range_Akhir	Engineered	Difference between maximum and minimum scores of S7–S9
Konsistensi	Engineered	Standard deviation of S1, S7, S8, and S9
Dist_C0	Soft Cluster	Distance to Cluster 0 centroid
Dist_C1	Soft Cluster	Distance to Cluster 1 centroid
Dist_C2	Soft Cluster	Distance to Cluster 2 centroid
Grade Category	Binary Target	0 = Non-A (<86), 1 = Grade A (≥86)

2.2 Data Preprocessing and Feature Engineering

The dataset was cleaned by removing invalid records and converting all selected variables into numerical format. The variable S2 was excluded because it contained no variation across students and therefore provided no discriminative information. To enrich student representation, four engineered features were generated: S_Akhir_Mean, Trend_Akhir, Range_Akhir, and Konsistensi. These features capture recent academic achievement, performance trends, score variability, and assessment consistency, respectively.

The final baseline feature set consisted of original assessment variables, aggregate academic indicators, and the engineered features. In addition, two candidate variables (Gap_Rapor and S_All_Mean) were excluded because they were linearly redundant with existing features.

2.3 K-Means Clustering and Soft-Cluster Feature Generation

In the first phase, K-Means Clustering was employed to discover latent academic performance patterns among students and to generate additional soft-cluster features for the classification model. The clustering process utilized seven academic variables consisting of S1, S7, S8, S9, NA_Sumatif, TS, and Rapor_Tengah. Prior to clustering, all variables were standardized using StandardScaler to ensure comparable feature scales [10].

The optimal number of clusters was determined using the Silhouette Score criterion by evaluating several candidate values of K. The highest Silhouette Score was obtained at K = 3, which was therefore selected as the optimal cluster configuration. Instead of using cluster labels directly, this study transformed the clustering results into soft-cluster distance features. For each student, the Euclidean distance to every cluster centroid was calculated, generating three additional variables: Dist_C0, Dist_C1, and Dist_C2 [11].

K-Means partitions observations into K clusters by iteratively minimizing within-cluster sum of squared Euclidean distances[12]. Rather than using discrete cluster labels:

$$Dist_i(x) = \sqrt{\sum_{m=1}^d (x_m - \mu_{im})^2} \quad (1)$$

where d denotes the number of clustering variables. These distance-based features capture the degree of similarity between each student and all cluster profiles and are subsequently integrated with the engineered academic variables to form the hybrid feature set used by the SVM classifier.

2.4 Hybrid K-Means-SVM Model Development

The hybrid input space consists of 15 features and 12 baseline academic variables combined with three K-Means soft-cluster distance features (Dist_C0, Dist_C1, Dist_C2). These features set was fed into an SVM classifier embedded within an Imbalanced-Learn Pipeline. To address the 5.9:1 class imbalance, SMOTE ($k_neighbors = 3$) was applied exclusively within each cross-validation fold to prevent data leakage [13], [14]. The SVM classifier employs an RBF kernel with $class_weight = balanced$ to impose greater misclassification penalties on minority-class instances. The decision function is expressed as[15]:

$$f(x) = \text{sign}(w^T x + b) \quad (2)$$

where w denotes the weight vector and b represents the bias term. Since the student performance data are not linearly separable, the RBF kernel is employed to project the data into a higher-dimensional feature space. The kernel function is defined as[16]:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (3)$$

where γ controls the influence radius of each support vector.

Hyperparameter optimization was performed using GridSearchCV with 5-fold stratified cross-validation across 36 combinations: $C \in \{0.01, 0.1, 1, 10, 100, 1000\}$ and $\gamma \in \{\text{scale, auto, 0.001, 0.01, 0.1, 1.0}\}$, with F1-Score as the optimization criterion[17].

2.5 Model Evaluation and Validation

Model performance was evaluated using Accuracy, F1-Score, and AUC-ROC, with the minority-class (Grade A) F1-Score serving as the primary metric due to class imbalance (240 Non-A vs. 41 Grade A students).

Performance estimation was conducted using stratified 5-fold cross-validation. An ablation study compared a baseline SVM model using academic features only with the proposed Hybrid K-Means+SVM model incorporating soft-cluster distance features.

Hyperparameter optimization was performed using GridSearchCV with stratified 5-fold cross-validation to identify the optimal values of the SVM regularization parameter (C) and RBF kernel coefficient (γ). Probability threshold tuning was subsequently optimized via Out-of-Fold (OOF) on the training set to maximize the minority-class F1-Score.

Model interpretability was assessed using permutation feature importance, while robustness was evaluated through repeated stratified 10×5-fold cross-validation. Performance stability was measured using the coefficient of variation (CoV) of the F1-Score. All experiments were implemented in Python 3.10 using scikit-learn and imbalanced-learn with a fixed random seed of 42.

3. Results And Discussion

3.1 Dataset Characteristics and Sample Data

This study utilizes academic assessment data from 281 Grade X Informatics students at senior high school during the 2024/2025 academic year. The dataset was collected from eight class sheets (X1–X8) and contains formative, summative, mid-semester, and report card assessment components. After data cleaning, records with missing or invalid final report card scores were excluded, resulting in 281 valid student records for analysis.

Table 2 presents a sample of the academic attributes used as prediction inputs. These variables include weekly formative assessment scores (S1, S7, S8, and S9), aggregate academic indicators (NA_Tengah, NA_Sumatif, TS, and Rapor_Tengah), and the final report card score (Rapor_Akhir), which is used to generate the target label.

Table 2. Sample of Student Academic Data Used for Prediction

No	Kelas	S1	S7	S8	S9	TS	NA_Sumatif	Rapor_Tengah	Rapor_Akhir
1	X1	80	80	90	100	75	90	83	78
2	X1	80	92	93	95	75	92	83	83
3	X1	80	85	90	87	75	88	83	79
4	X1	80	80	75	95	75	86	83	81
5	X1	80	80	75	95	79.56	86	85	83

The prediction task is formulated as a binary classification problem, where students with a final report card score of 86 or higher are assigned to the Grade A category, while the remaining students are categorized as Non-A. This formulation is intended to identify students who qualify for the highest achievement category and may be considered for academic enrichment programs.

Figure 2 illustrates the distribution of final report card scores in the dataset. The scores range from 75 to 92, with an average value of 81.2. Most students are concentrated between 80 and 86, indicating a relatively homogeneous academic performance distribution. Only a limited number of students achieved scores above the Grade A threshold.

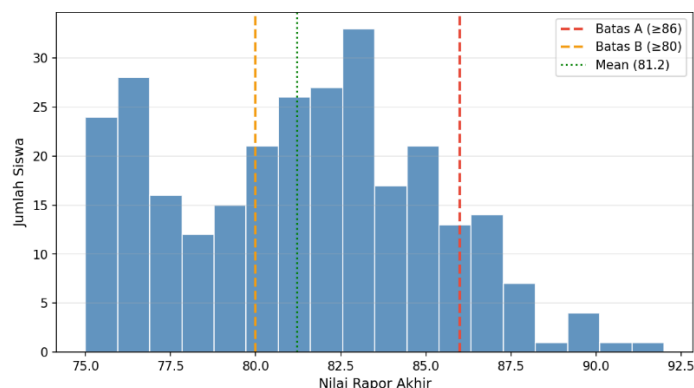


Figure 2. Distribution of Final Report Card Scores

The narrow score range indicates that adjacent grade categories possess highly similar characteristics, making multi-class classification more challenging. Consequently, transforming the problem into binary classification allows the model to focus on a single decision boundary that is more educationally meaningful and easier to learn.

Figure 3 presents the class composition after label transformation. Of the 281 students, 240 students (85.4%) belong to the Non-A category, while 41 students (14.6%) belong to the Grade A category, resulting in a class imbalance ratio of approximately 5.9:1.

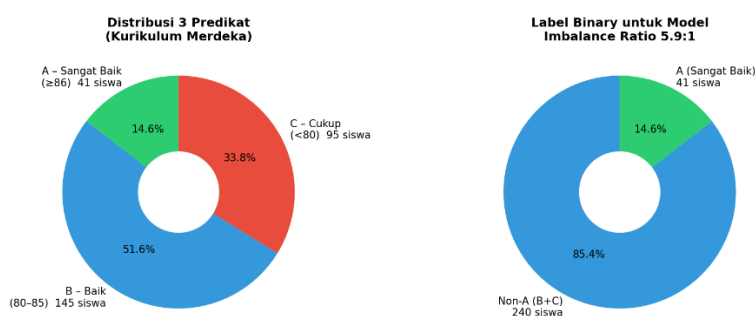


Figure 3. Distribution of Grade Categories and Binary Class Labels

The prediction process in this study begins with feature engineering to derive additional academic indicators from the original assessment variables. The processed features are then clustered using K-Means to generate soft-cluster distance attributes, which are subsequently combined with the academic features and used as inputs to the SVM classifier. The classifier produces the probability that each student belongs to the Grade A category, which serves as the basis for the final prediction results discussed in the following sections.

3.2 K-Means Clustering and Soft Cluster Feature Generation

The optimal number of clusters was determined using the Silhouette coefficient for $K = 2-8$. As shown in Figure 4, the highest Silhouette score (0.35) was obtained at $K = 3$, indicating that three clusters provided the most appropriate grouping structure for the dataset.

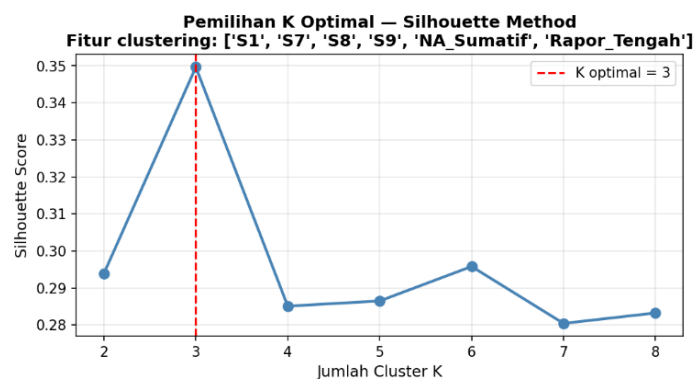


Figure 4. Silhouette Score for Determining the Optimal Number of Clusters

To interpret the clusters, descriptive analysis was conducted using students' final grades and performance trends. Figure 5 and Table 3 summarize the academic characteristics of the three clusters.

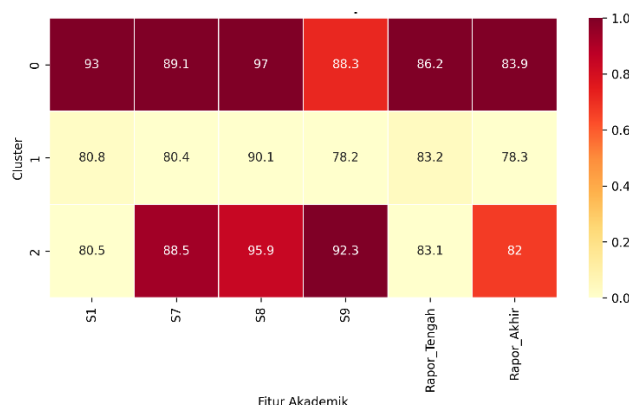


Figure 5. Academic Performance Profiles of the Three K-Means Clusters

Cluster C0 ($n = 48$) exhibited the highest mean final grade (83.9) and the largest proportion of Grade A students (37.5%), despite showing a declining performance trend. Cluster C1 ($n = 87$) represented lower-performing students with an average final grade of 78.3 and only 2.3% Grade A representation, although their recent performance showed improvement. Cluster C2 ($n = 146$), the largest cluster, showed moderate academic performance with an average final grade of 82.0 and a Grade A proportion of 14.4%.

Table 3. K-Means Cluster Profiles (K=3)

Cluster	Label	n	Mean Rapor_Akhir	Grade A (%)	Performance Trend
C0	High-Declining	48	83.9	37.5	S_Akhir < S1
C1	Low-Improving	87	78.3	2.3	S_Akhir > S1
C2	Medium-Improving	146	82.0	14.4	S_Akhir > S1

The concentration of Grade A students in Cluster C0 suggests that K-Means successfully captured meaningful academic performance patterns despite the relatively homogeneous score distribution. To preserve richer cluster information, soft-cluster features were generated by computing each student's Euclidean distance to the three cluster centroids, producing the variables Dist_C0, Dist_C1, and Dist_C2. These features were combined with the 12 baseline variables to form the final 15-feature hybrid dataset used by the SVM classifier.

3.3 Ablation Study of Hybrid K-Means and SVM

To evaluate the contribution of K-Means-derived soft-cluster features, an ablation study was conducted by comparing a baseline SVM model using 12 academic features with the

proposed Hybrid K-Means+SVM model incorporating three additional cluster-distance features (Dist_C0, Dist_C1, and Dist_C2). Both models were evaluated under identical settings using stratified 5-fold cross-validation.

Table 4 presents the performance comparison between the baseline SVM and the proposed Hybrid K-Means+SVM model.

Table 4. Ablation Study Results

Model	Accuracy	F1-Score	AUC-ROC
SVM without K-Means	0.8008	0.5331 ± 0.0843	0.8320
Hybrid K-Means+SVM	0.8043	0.5392 ± 0.0582	0.8434

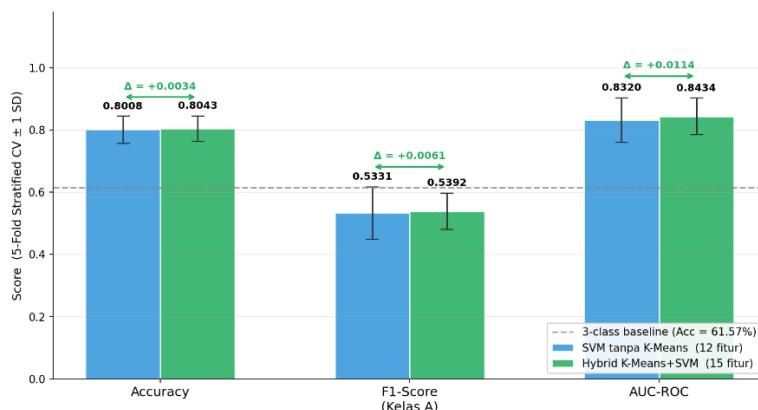


Figure 6. Ablation Study: SVM Baseline vs. Hybrid K-Means+SVM

The Hybrid K-Means+SVM model outperformed the baseline SVM across all evaluation metrics. AUC-ROC increased from 0.8320 to 0.8434, while F1-score improved from 0.5331 to 0.5392. Although the performance gains were modest, they indicate that the cluster-derived features provided complementary information beyond the original academic variables.

In addition, the standard deviation of the F1-score decreased from 0.0843 to 0.0582, representing a 30.9% reduction in performance variability across cross-validation folds. This result suggests that the soft-cluster features improved not only predictive performance but also model stability.

These findings support the use of distance-based cluster representations to capture latent academic patterns that are not fully reflected by raw assessment scores. Therefore, the Hybrid K-Means+SVM configuration was selected for subsequent hyperparameter optimization and final evaluation.

3.4 Hyperparameter Optimization of Hybrid K-Means+SVM

To determine the optimal Hybrid K-Means+SVM configuration, hyperparameter tuning was performed using GridSearchCV with 5-fold stratified cross-validation on the training set ($n = 211$). The search evaluated six values of the regularization parameter $C \in \{0.01, 0.1, 1, 10, 100, 1000\}$ and six values of the RBF kernel parameter $\gamma \in \{\text{scale}, \text{auto}, 0.001, 0.01, 0.1, 1.0\}$, resulting in 36 parameter combinations. Because the dataset exhibits a 5.9:1 class imbalance, optimization was based on the F1-score of the minority class (Grade A).

Figure 7 presents the mean cross-validation F1-score for all parameter combinations. The optimal configuration was identified as $C = 10$ and $\gamma = 0.01$, achieving a cross-validation F1-score of 0.5683. The selected parameters provide a balance between decision boundary flexibility and generalization, enabling the model to capture meaningful patterns without excessive sensitivity to individual training samples. These hyperparameters were subsequently used for all final model evaluations reported in Section 3.5.

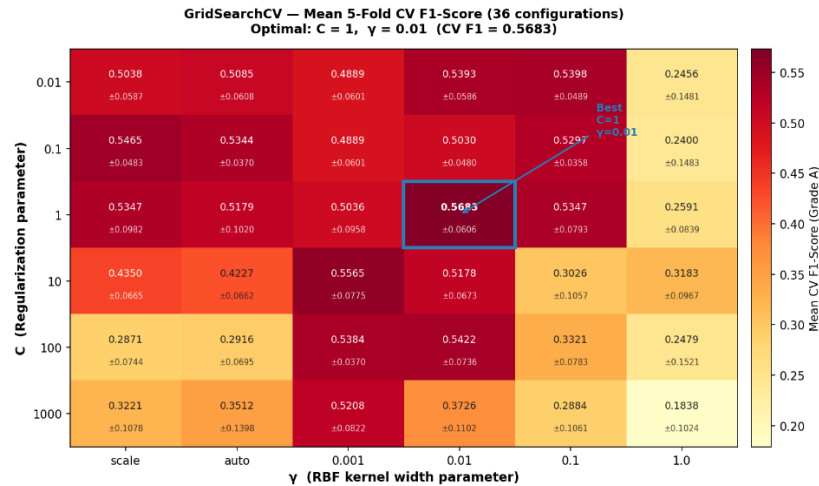


Figure 7. GridSearchCV

3.5 Final Model Performance Evaluation

The optimized Hybrid K-Means+SVM model ($C = 10$, $\gamma = 0.01$) was evaluated on the held-out test set and validated using 5-fold stratified cross-validation. The performance results are summarized in Table 5.

Table 5. Performance of the Optimized Hybrid K-Means+SVM ($C = 10$, $\gamma = 0.01$)

Metric	Test Set (Thr = 0.50)	Test Set (Thr = 0.44)	5-Fold CV (Training Set)
Accuracy	0.8310	0.8592	0.7952
Precision (Grade A)	0.4400	0.5000	—
Recall (Grade A)	0.8000	0.8000	—
F1-Score (Grade A)	0.5714	0.6154	0.5683 ± 0.0606
AUC-ROC	0.8541	0.8541	0.8678 ± 0.0444

AUC-ROC is threshold-independent and identical across both thresholds.

The model achieved a test-set AUC-ROC of 0.8541 and an F1-score of 0.5714 at the default threshold (0.50), indicating good discrimination between Grade A and Non-A students. Cross-validation results were consistent with test performance (CV AUC = 0.8678 ± 0.0444 ; CV F1 = 0.5683 ± 0.0606), suggesting stable generalization without evidence of substantial overfitting.

To improve minority-class detection, probability threshold tuning was performed using out-of-fold (OOF) probabilities on the training set to avoid test-set contamination. As shown in Table 6, lowering the threshold from 0.50 to 0.44 improved precision from 0.44 to 0.50 and accuracy from 0.8310 to 0.8592, with an improved F1-score from 0.5714 to 0.6154. Therefore, 0.44 was selected as the operational threshold for subsequent analyses. It should be noted that the optimal threshold was identified via OOF cross-validation on the training set and subsequently applied to the test set for evaluation, and this choice is discussed as a study limitation in Section 4.

Table 6. Threshold Comparison on the Test Set

Threshold	Precision (Grade A)	Recall (Grade A)	F1-Score (Grade A)	Accuracy	Note
0.20	0.22	1.00	0.36	0.62	Maximum Recall
0.44	0.50	0.80	0.6154	0.8592	Optimal Threshold
0.50	0.44	0.80	0.5714	0.8310	Default Threshold
0.70	0.75	0.30	0.43	0.8732	Maximum Precision

Table 7 presents the confusion matrices at thresholds 0.50 and 0.44. The optimized threshold improved Grade A detection while maintaining comparable overall accuracy. In the educational context, prioritizing recall is appropriate because failing to identify a high-achieving student may be more consequential than incorrectly flagging a Non-A student for further review.

Table 7. Confusion Matrix of the Hybrid K-Means+SVM at Default and Optimal Thresholds (Test Set, n = 71)

Actual Class	Threshold = 0.50: Predicted Non-A	Threshold = 0.50: Predicted Grade A	Threshold = 0.44: Predicted Non-A	Threshold = 0.44: Predicted Grade A
Non-A (n = 61)	51 (83.6%)	10 (16.4%)	53 (87%)	8(13%)
Grade A (n = 10)	2 (20.0%)	8 (80.0%)	2(20.0%)	8 (80.0%)

Figure 8 shows the ROC curve of the optimized model. The AUC-ROC value of 0.8541 indicates good discriminative ability and confirms that the proposed Hybrid K-Means+SVM framework effectively distinguishes Grade A from Non-A students.

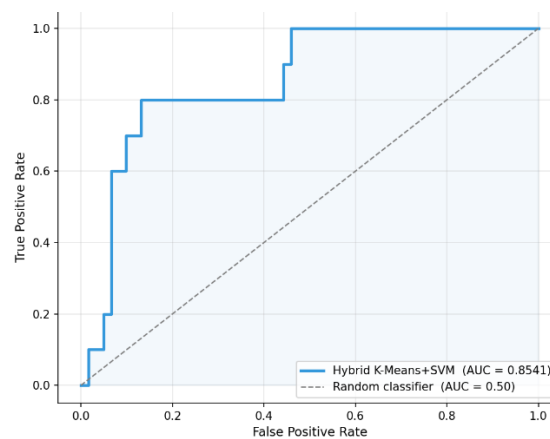


Figure 8. ROC Curve

3.6 Model Interpretation and Robustness

To improve model interpretability, permutation feature importance was applied using the decrease in AUC-ROC after randomly shuffling each feature. Figure 9 presents the importance ranking of all hybrid features.

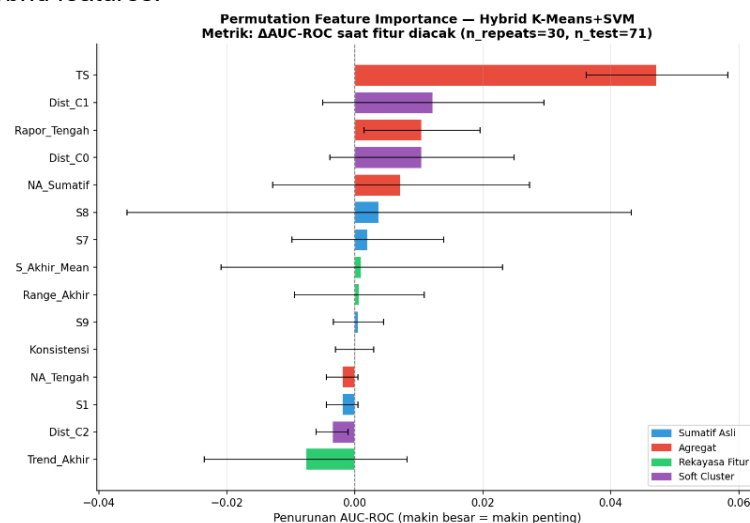


Figure 9. Permutation Feature Importance

The results indicate that TS (Sumatif Tengah Semester) was the most influential predictor, producing the largest decrease in AUC-ROC (+0.0472) when permuted. In addition, two cluster-derived features (Dist_C1 and Dist_C0) ranked among the most important variables, indicating that the K-Means soft-cluster features contributed meaningful predictive information beyond the original academic attributes.

The learning curve in Figure 10 shows that the training and validation F1-scores gradually converged as the training size increased, with only a small gap at the largest sample size. This pattern suggests good generalization and no substantial overfitting.

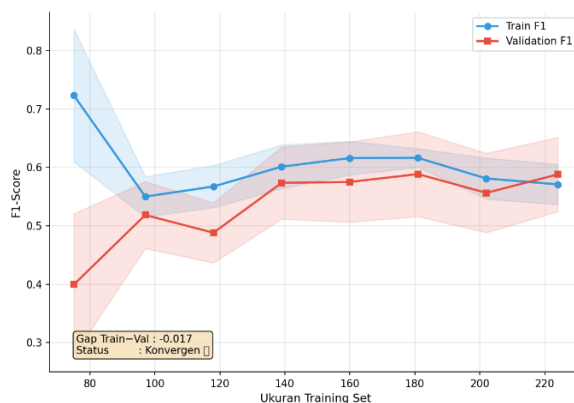


Figure 10. Learning Curve

Model robustness was further evaluated using repeated stratified 10×5-fold cross-validation (50 folds). The results are summarized in Table 8.

Table 8. Repeated Cross-Validation Results

Metric	Mean ± SD	Interpretation
F1-Score	0.5602 ± 0.0663	Stable
AUC-ROC	0.8759 ± 0.0442	Highly consistent
CoV F1	11.8%	Stable

The relatively low variability across repeated folds and a CoV below 15% indicate that the proposed Hybrid K-Means+SVM model is sufficiently stable and robust for student performance prediction. Permutation importance confirmed the contribution of cluster-derived features to predictive performance, while the learning curve and repeated cross-validation results demonstrated satisfactory convergence and stability. These findings support the suitability of the proposed approach for binary prediction of student academic achievement.

3.7 Student Prediction Analysis

The final Hybrid K-Means+SVM model was applied to all 281 students using the optimized threshold of 0.44. The prediction summary is presented in Table 9.

Table 9. Full-Dataset Prediction Summary

Metric	Value
Total Students	281
Actual Grade A	41 (14.6%)
Predicted Grade A	74 (26.3%)
TP	35
TN	201
FP	39
FN	6

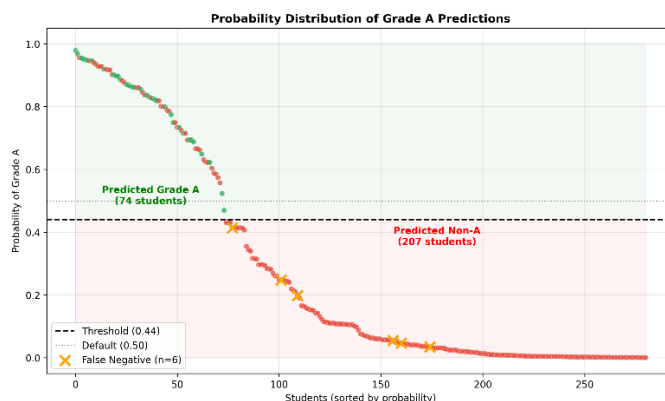


Figure 11. Probability Distribution of Grade A Predictions

Figure 11 shows the probability distribution of Grade A predictions across all students. Using the optimized threshold, the model correctly identified 35 of 41 Grade A students, resulting in only six false negatives. The model predicted more Grade A students than were actually observed, reflecting a strategy that prioritizes the identification of high-achieving students and reduces the risk of overlooking potential Grade A achievers.

The false-negative students had slightly lower TS scores than the correctly identified Grade A students (75.8 vs. 77.0 on average), consistent with the feature importance analysis that identified TS as the strongest predictor. Overall, the results demonstrate the practical value of the proposed model for identifying students with strong Grade A potential and supporting early academic intervention.

3.8 Discussion

The primary objective of this study was to develop and evaluate a Hybrid K-Means+SVM framework for identifying Grade A students from homogeneous and imbalanced educational data. The results indicate that the proposed approach successfully achieved this objective. The Hybrid K-Means+SVM model achieved an AUC-ROC of 0.8541 on the test set, outperforming the baseline SVM model in both AUC-ROC (0.8434 vs. 0.8320) and F1-score (0.5392 vs. 0.5331). This improvement was accompanied by a 30.9% reduction in F1-score variance (SD: 0.0843 $\rightarrow$ 0.0582), indicating enhanced model stability. The concentration of 37.5% of Grade A students within Cluster C0 derived from K = 3 clustering (Silhouette Score = 0.35) confirms that K-Means captured meaningful latent performance structures despite the narrow score range of 75–92. Permutation importance analysis confirmed that cluster-derived features (Dist_C1, Dist_C0) ranked among the top predictors alongside TS, validating their complementary contribution beyond the original academic variables. Probability threshold optimization (0.50 $\rightarrow$ 0.44) maintained Grade A recall at 0.80 while improving precision from 0.44 to 0.50 and the minority-class F1-score from 0.5714 to 0.6154, enabling correct identification of 35 of 41 Grade A students across the full dataset, consistent with recommendations emphasizing AUC-ROC and threshold-based evaluation for imbalanced datasets[18].

The observed improvements confirm the theoretical properties of each methodological component. The concentration of 37.5% of Grade A students within Cluster C0, despite the narrow score range of 75–92**, reflects the convergence mechanism of K-Means, whereby the iterative minimization of within-cluster squared distances, which Liu[19] demonstrated to be effective for uncovering latent performance regularities in educational evaluation data. The optimal configuration of C = 10 and γ = 0.01 reflects the maximum margin principle of the RBF kernel[20]. A higher C value reduces misclassification tolerance to fit the minority-class boundary more precisely, while a lower γ widens the kernel radius to accommodate the sparse, overlapping feature distribution arising from homogeneous scores. The sustained recall of 0.80 under a 5.9:1 class imbalance is consistent with SMOTE's linear interpolation mechanism [21], whereby synthetic samples generated between existing minority instances and their K nearest neighbors shift the decision boundary toward Grade A identification. Model robustness was further demonstrated through repeated cross-validation, yielding an AUC-ROC of 0.8759 ± 0.0442 and a coefficient of variation below 15%, indicating stable predictive performance across multiple data partitions.

The findings are also consistent with previous Educational Data Mining studies reporting competitive SVM performance for educational prediction [3], [5], while extending them by demonstrating that cluster-derived distance features improve performance in imbalanced settings [8], [9]. Unlike prior work focused on multi-class grade prediction, the binary reformulation adopted here effectively addressed homogeneous score distributions and overlapping class characteristics, which have been identified as common challenges in educational prediction problems [22].

This study makes three methodological contributions to the Educational Data Mining literature. Recent reviews have highlighted the growing importance of feature engineering, hybrid learning architectures, and robust evaluation strategies for educational prediction tasks [23], [24]. First, it empirically demonstrates that binary reformulation of a multi-class grade prediction problem improves minority-class separability in homogeneous and overlapping score distributions, a challenge widely recognized in EDM research [22]. Second, it establishes that K-Means soft-cluster distance features constitute a principled feature engineering strategy that improves not only predictive accuracy but also model stability, as evidenced by a 30.9% reduction in F1-score variance and a coefficient of variation below 15% across repeated cross-validation, a stability dimension rarely reported in prior educational prediction studies. Third, it demonstrates that probability threshold optimization can be systematically integrated as a decision-support component in educational prediction pipelines, enabling practitioners to explicitly control the precision-recall trade-off according to institutional priorities, which is particularly critical when the cost of overlooking high-achieving students is high [18].

From a practical perspective, the proposed framework supports data-driven identification of high-achieving students for enrichment programs and scholarships without relying solely on subjective teacher observations [23]. The high recall after threshold optimization minimizes the risk of overlooking Grade A candidates. Nevertheless, limitations apply: the dataset was collected from a single school ($n=281$), the model relied exclusively on academic variables whereas behavioral and LMS engagement data may further improve performance [24], and threshold optimization was performed via OOF cross-validation on the training set, mitigating the risk of test-set contamination. Future studies should validate the framework on external datasets, investigate multidimensional learning analytics approaches [22], and apply Explainable Artificial Intelligence (XAI) techniques to improve model transparency.

4. Conclusion

This study proposed and evaluated a Hybrid K-Means+SVM framework for predicting Grade A achievement using academic records of 281 Grade X Informatics students. The prediction task was reformulated as a binary classification problem (Grade A vs. Non-A) to address the challenges of homogeneous score distributions and severe class imbalance (5.9:1 ratio). Three methodological contributions were advanced: (1) binary reformulation of a multi-class grade prediction problem to improve minority-class separability; (2) K-Means soft-cluster distance features as a principled feature engineering strategy within an SVM-based classification pipeline; and (3) probability threshold optimization as a systematic decision-support mechanism for precision-recall control in imbalanced educational prediction settings.

The results confirm that the proposed framework effectively combines unsupervised and supervised learning. K-Means clustering ($K = 3$; Silhouette Score = 0.35) successfully captured meaningful latent performance structures, as evidenced by the concentration of 37.5% of Grade A students within Cluster C0, and generated three soft-cluster distance features that improved both predictive performance and model stability relative to the baseline SVM. The Hybrid K-Means+SVM achieved an AUC-ROC of 0.8541 on the test set and 0.8678 ± 0.0444 in 5-fold cross-validation, indicating good discriminative ability and stable generalization. Model stability was further confirmed by a 30.9% reduction in F1-score variance (SD: 0.0843 $\rightarrow$ 0.0582) and a repeated stratified 10 $\times$ 5-fold cross-validation AUC-ROC of 0.8759 ± 0.0442 with a coefficient of variation below 15%.

Probability threshold optimization further improved classification effectiveness. Adjusting the decision threshold from 0.50 to 0.44 increased precision from 0.44 to 0.50, improved the Grade A F1-score from 0.5714 to 0.6154, and increased overall accuracy from 0.8310 to 0.8592 while maintaining a recall of 0.80. When applied to the full dataset, the model correctly identified 35 of 41 Grade A students, demonstrating its practical potential for supporting data-driven student

selection and academic enrichment programs with a reduced risk of overlooking high-achieving students.

Collectively, the findings confirm that integrating K-Means soft-cluster distance features with SVM constitutes an effective, stable, and practically applicable approach for student performance prediction in homogeneous and imbalanced educational datasets. Nevertheless, several limitations should be acknowledged: the dataset was collected from a single school ($n = 281$), the model relied exclusively on academic variables, and threshold optimization was performed via OOF cross-validation on the training set, mitigating the risk of test-set contamination. Future research should validate the proposed framework on larger and more diverse external datasets, incorporate behavioral and learning engagement variables to capture multidimensional learning dynamics, and apply Explainable Artificial Intelligence (XAI) techniques to enhance model transparency and broaden its educational applicability.

References

- [1] M. H. Roslan and C. J. Chen, "Educational Data Mining for Student Performance Prediction : A Systematic Literature Review (2015-2021)," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 05, pp. 207–225, 2022, doi: 10.3991/ijet.v17i05.27685.
- [2] A. Namoun and A. Alshantiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques : A Systematic Literature Review," *Appl. Sci.*, vol. 11, no. 1, pp. 1–28, 2021, doi: 10.3390/app11010237.
- [3] Y. A. Alsariera, Y. Baashar, G. Alkaws, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/4151487.
- [4] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling , Oversampling , and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information*, vol. 14, no. 1, pp. 1–15, 2023, doi: 10.3390/info14010054.
- [5] M. Yağcı, "Educational data mining : prediction of students ' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 11, pp. 1–19, 2022, doi: 10.1186/s40561-022-00192-z.
- [6] D. Khairy, N. Alharbi, M. A. Amasha, and M. F. Areed, "Prediction of student exam performance using data mining classification algorithms," *Educ. Inf. Technol.*, vol. 29, no. 16, pp. 21621–21645, 2024, doi: 10.1007/s10639-024-12619-w.
- [7] E. Novianto and S. Suhirman, "Comparison of K-Nearest Neighbor Classification Methods and Support Vector Machine in Predicting Students' Study Period," *J. Educ. Sci.*, vol. 33, no. 1, pp. 32–45, 2024, doi: 10.33899/edusj.2023.144865.1408.
- [8] L. Pamungkas, N. A. Dewi, and N. A. Putri, "Classification of Student Grade Data Using the K-Means Clustering Method," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 13, no. 1, pp. 86–91, 2024, doi: 10.32736/sisfokom.v13i1.1983.
- [9] M. Wu, "K-Means Clustering-Based Feature Generation for Student Performance Prediction," *ICCK Trans. Educ. Data Min.*, vol. 2, no. 1, pp. 14–28, 2026, doi: 10.62762/TEDM.2026.716076.
- [10] A. Fadil and Z. Fatah, "Algoritma K-Means Clustering Untuk Menentukan Siswa Unggulan," *J. Ris. Sist. Inf.*, vol. 2, no. 1, pp. 67–75, 2025, doi: 10.69714/p26gcf27.
- [11] E. D. Lestari and F. Kurniasari, "Learning Analytics-Driven Student Performance Prediction Using Ensemble Machine Learning Models," *Artif. Intell. Learn.*, vol. 2, no. 1, pp. 1–16, 2026, doi: 10.63913/ail.v2i1.5.
- [12] A. A. Wani, "Comprehensive analysis of clustering algorithms : exploring limitations and innovative solutions," *PeerJ Comput. Sci.*, vol. 10, pp. 1–45, 2024, doi: 10.7717/peerj-cs.2286.
- [13] G. Dimić and L. Pecić, "Proposed Approach for Overcoming the Impact of Unbalanced Distribution in Predicting Students ' Performance," *TEM J.*, vol. 13, no. 4, pp. 2839–2849, 2024, doi: 10.18421/TEM134.
- [14] S. Malik *et al.*, "Enhancing education quality with hybrid clustering and evolutionary neural networks in a multi phase framework," *Sci. Rep.*, vol. 15, pp. 1–18, 2025, doi: 10.1038/s41598-025-04622-z.
- [15] N. I. M. Talib, N. A. A. Majid, and S. Sahran, "Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine," *Appl. Sci.*,

- vol. 13, no. 5, pp. 1–19, 2023, doi: 10.3390/app13053267.
- [16] T. Hongli, “Educational data mining for student performance prediction: feature selection and model evaluation,” *J. Electr. Syst.*, vol. 20, no. 3, pp. 1063–1074, 2024, doi: 10.52783/jes.3434.
- [17] R. Guido, S. Ferrisi, and D. Lofaro, “An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications : A Review,” *Information*, vol. 15, no. 4, pp. 1–36, 2024, doi: 10.3390/info15040235.
- [18] S. Priya, T. Ankit, and D. Divyansh, “Student Performance Prediction Using Machine Learning,” in *Computing Technologies and Applications*, vol. 40, IOS Press, 2021, pp. 167–174. doi: 10.3233/APC210137.
- [19] R. Liu, “Data Analysis of Educational Evaluation Using K-Means Clustering Method,” *Comput. Intell. Neurosci.*, vol. 2022, p. 10, 2022, doi: 10.1155/2022/3762431.
- [20] K. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, “Exploring Kernel Machines and Support Vector Machines : Principles , Techniques , and Future Directions,” *Mathematics*, vol. 12, no. 24, p. 3935, 2024, doi: 10.3390/math12243935.
- [21] D. Elreedy, A. F. Atiya, and F. Kamalov, “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.
- [22] Z. Shou, M. Xie, J. Mo, and H. Zhang, “Predicting Student Performance in Online Learning : A Multidimensional Time-Series Data Analysis Approach,” *Appl. Sci.*, vol. 14, no. 6, p. 2522, 2024, doi: 10.3390/app14062522.
- [23] C. Duan, X. Wan, W. Zhang, Y. Wang, J. Wang, and M. Zhang, “Predicting Student Performance Using Machine Learning Techniques : A Systematic Literature Review,” *Int. Conf. Comput. Sci. Technol. Educ.*, pp. 362–368, 2025, doi: 10.1109/CSTE64638.2025.11092243.
- [24] M. Murithi and D. Marangu, “A Systematic Review of the Machine Learning-Based Educational Data Mining Models for Student Performance Prediction Performance Analysis of Students OR Learning Curve Prediction,” *Int. J. Comput. Appl. Technol. Res.*, vol. 14, no. 10, pp. 59–67, 2025, doi: 10.7753/IJCATR1410.1012.