*Available online at:* https://journal.upgripnk.ac.id/index.php/saintek
**Jurnal Pendidikan Informatika dan Sains**
**15(1), 2026, 88-97**
*E-ISSN: 2407-1536*
*P-ISSN: 2089-2802*

# EDUCATIONAL DATA MINING FOR INFORMATICS GRADE PREDICTION USING A HYBRID K-MEANS FRAMEWORK

**Roselilie Simbulan*, Arief Hermawan, Donny Avianto**
Magister of Technology Information, Universitas Teknologi Yogyakarta, Daerah Istimewa Yogyakarta, Indonesia
*email: rosyliesimbulan@gmail.com

**Received: 2026-05-01 Accepted: 2026-06-10 Published: 2026-06-30**

**Abstract**

This study proposes a two-phase hybrid Educational Data Mining (EDM) framework that integrates K-Means clustering with supervised classification to predict students' final grade categories in Informatics within the Kurikulum Merdeka competency-based assessment system. The dataset consists of 281 Grade X students, all of whom achieved scores above the Minimum Achievement Criterion (KKTP = 75), with prediction focused on three grade categories: A (≥86), B (80–85), and C (75–79). Four summative assessment scores (S1, S7, S8, and S9) were used as input features. In the first phase, K-Means generated three clusters (Silhouette Score = 0.3198), and the resulting cluster labels were added as an additional feature. In the second phase, Random Forest and Logistic Regression were optimized using Grid Search with 5-Fold Stratified Cross-Validation, while SMOTE was employed to address class imbalance. The results show that Logistic Regression outperformed Random Forest, achieving a test accuracy of 59.65% and a Macro F1-Score of 0.5899, whereas Random Forest achieved 49.12% accuracy and a Macro F1-Score of 0.4799 and exhibited signs of overfitting. Feature importance analysis identified S7, S8, and S9 as the most influential predictors, while the cluster-derived feature contributed more strongly to Random Forest than to Logistic Regression. These findings suggest that well-regularized linear models may generalize better than ensemble methods on small datasets with narrow score distributions. The proposed framework is best positioned as a screening-support tool for early formative intervention in competency-based educational settings.

**Keywords:** Educational Data Mining, K-Means Clustering, Random Forest, Logistic Regression, Grade Prediction.

**How to cite (in APA style):** Simbulan, R., Hermawan, A., & Avianto, D. (2026). Educational data mining for informatics grade prediction using a hybrid K-means framework. *Jurnal Pendidikan Informatika Dan Sains, 15*(1), 88–97. https://doi.org/10.31571/saintek.v15i1.10728


DOI: 10.31571/saintek.v15i1.10728

## INTRODUCTION

The rapid growth of academic data in educational institutions has accelerated the adoption of Educational Data Mining (EDM) to extract meaningful patterns and support data-driven decision-making. Prior studies consistently show that classification techniques dominate EDM applications, with algorithms such as Decision Tree and Random Forest widely used for predicting student academic performance (Namoun & Alshanqiti, 2021; Roslan & Chen, 2022). Ensemble methods, particularly Random Forest, have demonstrated superior predictive performance compared to single classifiers across various educational datasets (Evangelista & Sy, 2022; Khairy et al., 2024).

However, this performance advantage is not universal: on small or class-imbalanced datasets, simpler linear classifiers such as Logistic Regression may achieve comparable or superior generalization due to lower model complexity and reduced susceptibility to overfitting (Yağcı, 2022).

In recent years, hybrid approaches that integrate unsupervised and supervised learning have gained increasing attention due to their ability to capture both latent data structures and predictive relationships. Studies have shown that combining clustering techniques such as K-Means with classification algorithms can significantly improve prediction accuracy (Malik et al., 2025; Sayed & Fouad, 2025). Clustering methods enable the identification of hidden group patterns within student data, which can be leveraged as additional features to enhance model performance (Wu, 2026). This integration has proven effective in improving both prediction accuracy and model robustness (Men & Zhao, 2025).

Despite these advancements, class imbalance remains a critical challenge in educational datasets, often leading to biased models that favor majority classes. Techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been widely adopted to address this issue by improving minority class representation without significantly degrading overall performance (Dimić & Pecić, 2024; Wongvorachan et al., 2023). In addition, hyperparameter optimization methods such as Grid Search play an essential role in maximizing model performance and ensuring optimal parameter selection (Yanyu et al., 2025).

However, most existing studies focus on general performance prediction or binary outcomes (e.g., pass/fail), with limited attention to fine-grained grade classification in specific educational contexts. In particular, research addressing the Informatics subject within the Indonesian Kurikulum Merdeka framework remains scarce. This curriculum introduces a competency-based assessment structure with multiple summative components, requiring more nuanced prediction approaches beyond conventional models.

To address this gap, this study proposes a two-phase hybrid framework integrating K-Means Clustering with supervised classification, incorporating SMOTE for class imbalance handling and Grid Search for hyperparameter optimization, and evaluated comparatively using Random Forest and Logistic Regression to assess whether ensemble complexity yields a meaningful advantage over a simpler linear approach on a small, imbalanced educational dataset. The study utilizes summative assessment data from Grade X Informatics senior high school students during the 2024/2025 academic year. The objectives of this study are: (1) to identify the most influential features affecting students' final grade categories, (2) to evaluate the contribution of cluster-derived features to prediction performance, and (3) to compare the generalization capabilities of an ensemble-based classifier (Random Forest) and a linear classifier (Logistic Regression) under the constraints of limited sample size and class imbalance.

## METHOD

This study adopts a quantitative approach using a two-phase hybrid machine learning framework to predict students' final grade categories in Informatics. The dataset was collected from the academic assessment system of senior high school student at Yogyakarta, consisting of 281 Grade X students from eight classes (X1–X8) in the 2024/2025 academic year. All students achieved scores above the minimum competency threshold (KKTP = 75), resulting in a relatively narrow score distribution, which is a common characteristic in educational datasets with academically filtered cohorts (Yağcı, 2022).

The predictor variables include four summative assessment scores (S1, S7, S8, and S9), while S2 was excluded due to near-zero variance, as features with minimal variation do not contribute to model discrimination (Bulut et al., 2025). The target variable is categorized into three classes: A (≥86), B (80–85), and C (75–79). The dataset characteristics are summarized in Table 1.

**Table 1. Dataset Characteristics**

| Characteristic | Description |
|---|---|
| Total students | 281 students (8 classes) |
| Score range | 75–92 |
| Grade A | 41 students (14.59%) |
| Grade B | 145 students (51.60%) |
| Grade C | 95 students (33.81%) |
| Features | S1, S7, S8, S9 |
| KKTP | 75 |

The proposed method consists of two main phases. In the first phase, K-Means Clustering is applied to identify latent group structures in student performance data. K-Means partitions the dataset into K clusters by minimizing intra-cluster variance, typically measured using Within-Cluster Sum of Squares (WCSS) (Pamungkas et al., 2024). Prior to clustering, all features are standardized using z-score normalization to ensure equal contribution in distance calculations, as K-Means is sensitive to feature scaling (Wongoutong, 2024). The optimal number of clusters is determined using the Silhouette Score (Wu, 2026). To assess the robustness of this choice, K-Means is additionally benchmarked against Gaussian Mixture Models (GMM) and DBSCAN, two alternative clustering algorithms commonly applied in educational data mining contexts. The resulting cluster labels are then appended as an additional feature to enrich the dataset.

In the second phase, classification is performed using Random Forest (RF) and Logistic Regression (LR), enabling a direct comparison between an ensemble-based and a linear classification approach. Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and random feature selection, and aggregates predictions through majority voting (Scornet, 2023). Logistic Regression, by contrast, estimates class probabilities through a linear combination of predictors, offering a simpler and more interpretable alternative that is generally less prone to overfitting on small datasets. Comparing these two approaches allows the study to assess whether the added complexity of an ensemble method yields a meaningful performance advantage over a simpler linear model given the limited sample size. The integration of clustering-derived features into classification models has been shown to enhance predictive performance in educational data mining tasks (Sayed & Fouad, 2025).

Prior to model training, the dataset was partitioned into training (80%) and held-out test (20%) sets using stratified sampling to preserve class proportions. All subsequent preprocessing, class balancing, and hyperparameter tuning were performed exclusively on the training set; the test set remained untouched until final model evaluation. To address class imbalance (A:B:C = 41:145:95), the Synthetic Minority Over-sampling Technique (SMOTE) was applied within a unified preprocessing-and-modeling pipeline, ensuring that synthetic samples were generated only from the training portion of each cross-validation fold and never influenced the corresponding validation fold. This pipeline-based implementation prevents data leakage and ensures that cross-validation performance estimates reflect genuine generalization capability. SMOTE generates synthetic samples for minority classes using interpolation between nearest neighbors, thereby improving class distribution and model sensitivity to minority classes (Dimić & Pecić, 2024; Wongvorachan et al., 2023).

Hyperparameter optimization is conducted using Grid Search with 5-fold Stratified Cross-Validation performed on the training set, ensuring that class proportions were preserved across folds. For Random Forest, the parameter search space and optimal values are presented in Table 2. For Logistic Regression, the search space included the regularization strength $C \in \{0.01, 0.1, 1, 10, 100\}$ and penalty type (L1, L2), optimized using the same cross-validation procedure.

**Table 2. Parameter and Optimal Value**

| Model | Parameter | Values Tested | Optimal Value |
|---|---|---|---|
| Random Forest | n_estimators | 100, 200, 300 | 200 |
| | max_depth | 3, 5, 7, None | 7 |
| | min_samples_split | 2, 5, 10 | 5 |
| | min_samples_leaf | 1, 2, 4 | 1 |
| | max_features | sqrt, log2 | sqrt |
| | class_weight | balanced | balanced |
| Logistic Regression | C | 0.01, 0.1, 1, 10, 100 | 0.1 |
| | penalty | L1, L2 | L2 |
| | solver | saga | saga |

Model performance is evaluated using Accuracy, Precision, Recall, and F1-Score, computed separately on the cross-validation folds and on the held-out test set to assess generalization. Considering the imbalanced nature of the dataset, macro-averaged F1-Score is used as the primary evaluation metric, as it assigns equal importance to all classes regardless of their distribution (Opitz, 2024).

## RESULTS AND DISCUSSION

Exploratory data analysis was conducted on 281 student records to examine the distribution and relationships among variables. The final report card grades ranged from 75 to 92, indicating a narrow distribution where all students exceeded the minimum competency threshold (KKTP = 75). Such constrained distributions are commonly observed in academically filtered educational datasets and tend to reduce model discriminative capacity (Yağcı, 2022; Zhang et al., 2025). The distribution is concentrated in the 80–85 range with a mean of approximately 81.2, and the class distribution is moderately imbalanced, with Grade B as the majority class (51.60%), followed by Grade C (33.81%) and Grade A (14.59%), as shown in Figure 1.

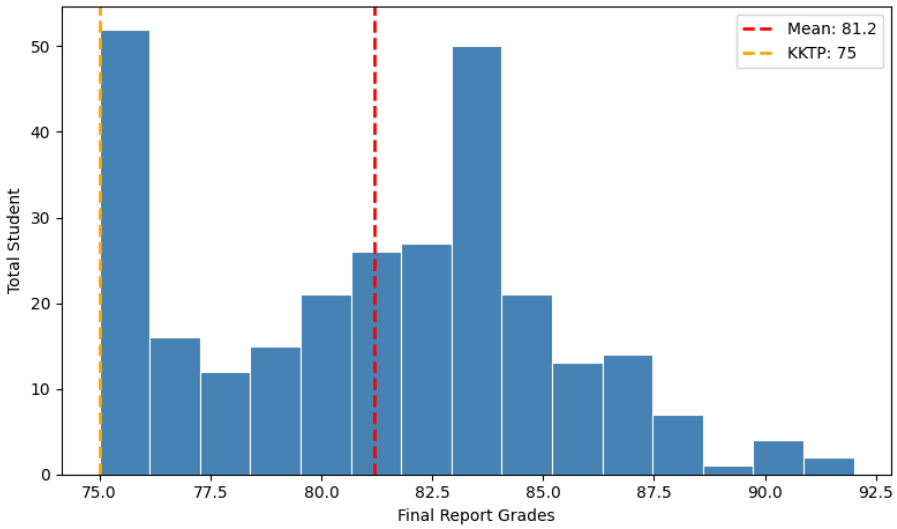


**Figure 1. Distribution of Final Report Card Grades with Mean and KKTP Threshold**

Figure 1 presents the distribution of final report card scores across all 281 students. The histogram shows a narrow, right-skewed distribution concentrated between 75 and 92, with a mean of 81.2 (SD = 3.98). All scores lie above the KKTP threshold of 75, confirming that every student met the minimum competency requirement. This limited variability illustrates the core

methodological challenge addressed in this study: differentiating fine-grained grade categories (A, B, C) within an already homogeneous, high-performing cohort.

Correlation analysis shows that end-of-semester assessments (S7, S8, S9) have stronger relationships with final grades compared to mid-semester assessment S1. This finding indicates that performance in later learning stages plays a more dominant role in determining final outcomes, consistent with prior studies(Alsariera et al., 2022; Khairy et al., 2024).

In the first phase, K-Means clustering was applied to identify latent group structures. The optimal number of clusters was determined as K=3 based on the Silhouette Score (0.3198), consistent with clustering practices in educational datasets where three performance levels are commonly identified (Pamungkas et al., 2024; Raj & Vidyaathulasiraman, 2021). The clustering result is visualized in Figure 2.

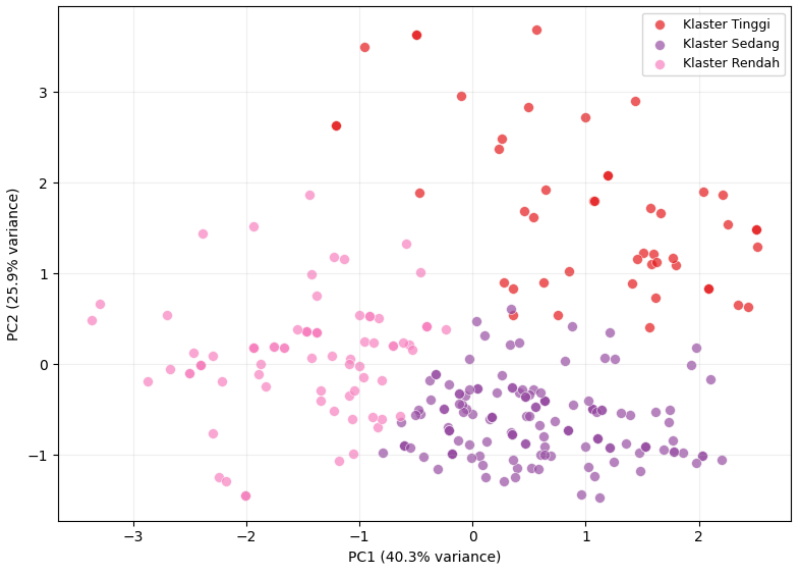


**Figure 2. Visualization of Student Clusters in PCA Feature Space Using K-Means (K=3)**

As shown in Figure 2, the two principal components jointly explain 66.2% of the total variance (PC1 = 40.3%, PC2 = 25.9%). PC1 is primarily driven by S7, S8, and S9 (loadings 0.53–0.61), representing a general achievement axis along which the Low cluster is clearly separated, while the High and Medium clusters substantially overlap due to their similarly strong performance on these three features. PC2, in contrast, is dominated by a contrast between S1 and S9 (loadings 0.85 and −0.47, respectively), which separates the High cluster (high S1, lower S9) from the Medium cluster (low S1, higher S9) along the vertical axis. This explains the moderate overall separation observed: clusters are well-distinguished along one dimension but overlap along another, a pattern consistent with the narrow, homogeneous score distribution noted in Figure 1. The cluster profiles are summarized in Table 3.

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

| Cluster | Label | n | Mean S1 | Mean S7 | Mean S8 | Mean S9 | Mean Final Report |
|---|---|---|---|---|---|---|---|
| 2 | High | 48 | 93.02 | 89.15 | 97.17 | 87.77 | 83.71 |
| 1 | Medium | 146 | 80.52 | 88.5 | 95.90 | 92.45 | 82.08 |
| 0 | Low | 87 | 80.80 | 80.39 | 90.08 | 78.24 | 78.34 |

Table 3 shows that the medium cluster dominates, aligning with the majority class distribution. Although the Silhouette Score is moderate, it remains acceptable for datasets with overlapping

distributions (Odoom et al., 2026; Wu, 2026). Importantly, cluster-derived features can enhance predictive performance even when cluster separation is not optimal.

To further evaluate the suitability of the clustering approach, additional experiments were conducted using Gaussian Mixture Models (GMM) and DBSCAN. For DBSCAN, parameters were optimized via grid search across a range of eps and min_samples values, with the best configuration identified at eps = 0.1 and min_samples = 3. The comparative results are presented in Table 4.

**Table 4. Comparison of Clustering Methods**

| Method | Clusters | Noise | Silhouette Score |
|---|---|---|---|
| K-Means | 3 | 0 | 0.3198 |
| GMM | 3 | 0 | 0.2134 |
| DBSCAN | 17 | 217 | 1.0000 |

The results indicate that K-Means achieved the highest meaningful cluster quality among the evaluated methods. Although GMM produced the same number of clusters, its Silhouette Score was lower, suggesting weaker separation between student groups. DBSCAN yielded a perfect Silhouette Score; however, it identified 217 observations (77.2% of the dataset) as noise and generated 17 small clusters. Such a clustering structure is difficult to interpret in an educational context and provides limited utility for subsequent classification tasks.

Pairwise Adjusted Rand Index (ARI) analysis further confirms substantial divergence in cluster assignments across methods: K-Means and GMM show only modest agreement (ARI = 0.2987), while DBSCAN's partition shows minimal correspondence with either K-Means (ARI = 0.0927) or GMM (ARI = 0.0000), the latter indicating no agreement beyond chance. This demonstrates that DBSCAN's apparent clustering quality is not merely difficult to interpret, but fundamentally incongruent with the partitions produced by the other two methods.

These findings suggest that the relatively homogeneous distribution of student scores limits the formation of highly separated clusters regardless of the clustering algorithm employed. Consequently, K-Means was retained as the clustering component of the proposed framework because it provided the most balanced trade-off between cluster interpretability, cluster coverage, and compatibility with the downstream classification stage.

In the second phase, two classification algorithms, namely Random Forest (RF) and Logistic Regression (LR) were evaluated using the augmented dataset. SMOTE was applied exclusively within the training folds through a pipeline-based implementation to prevent data leakage, ensuring that synthetic samples never propagated into validation or test folds. This within-fold balancing approach is particularly critical given the severe minority-class constraint in this dataset. Class A contains only 41 training-eligible samples, which required setting k_neighbors=4 in SMOTE rather than the default of 5. The classification performance is summarized in Table 5.

**Table 5. Classification Performance Comparison**

| Model | CV Accuracy | Test Accuracy | Macro F1-Score | Gap |
|---|---|---|---|---|
| Random Forest | 58.94% | 49.12% | 0.4799 | 9.82% |
| Logistic Regression | 53.98% | 59.65% | 0.5899 | -5.67% |

The Table 5 shows the Random Forest model demonstrated a moderate gap between cross-validation accuracy (58.94%) and test accuracy (49.12%), amounting to 9.82 percentage points. This gap reflects the limited sample size (n = 281) relative to the three-category target, which leaves each cross-validation fold with relatively few examples per class even after SMOTE balancing, making

the held-out test estimate more variable than the cross-validation average. Logistic Regression, by contrast, shows no indication of overfitting: its test accuracy (59.65%) marginally exceeds its CV accuracy (53.98%), a pattern consistent with the small size of the held-out test set (n = 57) rather than any underlying instability.

Several factors explain the models' moderate accuracy. The dataset reflects an academically pre-filtered cohort: all 281 students exceeded the minimum competency threshold (KKTP = 75), producing a narrow score range (75–92) that inherently limits separability between adjacent grade categories. The three-category formulation (A, B, C) further imposes a finer-grained task than binary classification, compounded by the relatively small size of Category A (n = 41, 14.6%). The feature set was also deliberately restricted to four summative scores (S1, S7, S8, S9) and the cluster-derived feature, excluding components directly used to compute the target category in order to avoid circular prediction. The classification report shows misclassification spread fairly evenly across all three categories (F1-scores between 0.44 and 0.52 under Random Forest), suggesting that the difficulty arises from overlapping scores near category boundaries rather than systematic bias toward any one grade level. The effect of within-fold SMOTE balancing is most evident in Class A recall: Logistic Regression achieves 0.75 for this minority class (14.6% of the dataset), compared to 0.50 under Random Forest, a difference unlikely without synthetic oversampling and consistent with LR's overall higher Macro F1-Score (0.5899 vs. 0.4799).

Random Forest's test accuracy (49.12%) falls below a threshold that would support its use in autonomous, high-stakes grading decisions. Logistic Regression performs more reliably (59.65% test accuracy, no overfitting), making it the more defensible choice for this dataset, though still below a level suitable for unsupervised decision-making. Its more appropriate role is as a screening-support tool: with a recall of 0.42 for Category C under Random Forest, the model can help flag a meaningful subset of lower-performing students for early formative intervention, where the cost of a false positive is low.

To further characterize how each model attributes predictive value across features, importance rankings from both classifiers are compared in Table 6. For Random Forest, importance is derived from impurity reduction; for Logistic Regression, the mean absolute standardized coefficient across the three classes is reported as a comparable measure of feature influence.

**Table 6. Feature Importance Comparison (Random Forest vs. Logistic Regression)**

| Rank | Feature | Description | RF Importance | LR Mean \|Coefficient\| |
|---|---|---|---|---|
| 1 | S7 | End-semester Summative 1 | 0.3450 | 0.4241 |
| 2 | S8 | End-semester Summative 2 | 0.2259 | 0.3526 |
| 3 | S9 | End-semester Summative 3 | 0.1777 | 0.2953 |
| 4 | S1 | Mid-semester Summative 1 | 0.1260 | 0.2896 |
| 5 | Cluster | K-Means cluster label | 0.1254 | 0.0379 |

As shown in Table 6, both classifiers agree that S7, S8, and S9 are the three most influential predictors, jointly dominating the importance structure of both models and confirming the central role of end-of-semester assessments in determining final grades (Random Forest: approximately 75% of cumulative importance). The two models diverge, however, in how strongly the cluster-derived feature contributes: under Random Forest, the cluster feature (0.1254) is nearly as influential as S1 (0.1260), whereas under Logistic Regression its mean coefficient magnitude (0.0379) is roughly seven times smaller than S1's (0.2896). This divergence likely reflects the two models' differing capacities to exploit the cluster label. Random Forest can partition on it flexibly within a tree structure, while Logistic Regression treats it as a single linear term, potentially under-utilizing the information it encodes.

The per-class coefficients of Logistic Regression further provide a directional interpretation unavailable from Random Forest's importance scores alone: all four assessment scores show positive

coefficients for Category A, near-zero coefficients for Category B, and negative coefficients for Category C, indicating a consistent, monotonic relationship in which higher assessment performance increases the likelihood of the higher grade category and decreases the likelihood of the lower one. This supports previous findings that hybrid approaches improve model performance by capturing latent structures (Sayed & Fouad, 2025; Wu, 2026), while also illustrating a practical advantage of retaining a linear classifier alongside the ensemble-based approach.

From a practical perspective, these results highlight the importance of emphasizing end-of-semester assessments in instructional strategies. The clustering approach also enables more nuanced student profiling, supporting targeted interventions in competency-based learning environments.

However, this study has limitations. The dataset is relatively small and homogeneous, limiting generalizability. Additionally, the narrow score range reduces class separability, making fine-grained classification inherently challenging. Future research should explore larger datasets and more advanced models to improve predictive performance.

## CONCLUSION

This study proposes a two-phase hybrid Educational Data Mining framework that integrates K-Means clustering with supervised classification. The framework is comparatively evaluated using Random Forest and Logistic Regression to predict students' final grade categories within the competency-based assessment structure of the Kurikulum Merdeka. The comparative evaluation reveals that Logistic Regression outperforms Random Forest on this constrained dataset, achieving higher test accuracy (59.65% vs. 49.12%) and Macro F1-score (0.5899 vs. 0.4799) without evidence of overfitting. In contrast, Random Forest exhibits a 9.82-percentage-point gap between cross-validation and test performance, suggesting overfitting given the dataset's limited size and three-category target structure.

Feature importance analysis from both classifiers consistently identifies end-of-semester assessments (S7, S8, S9) as the dominant predictors, jointly accounting for approximately 75% of cumulative importance under Random Forest and the largest mean coefficient magnitudes under Logistic Regression. The cluster-derived feature generated by K-Means contributes meaningfully to the Random Forest model (importance = 0.1254, nearly equivalent to S1 at 0.1260), though its contribution is substantially smaller under Logistic Regression (mean |coefficient| = 0.0379), suggesting that ensemble-based models are better able to exploit non-linear cluster structure. Within-fold SMOTE balancing further proved critical for minority-class handling: Logistic Regression achieved a Class A recall of 0.75 compared to 0.50 under Random Forest, confirming that pipeline-based oversampling meaningfully improves sensitivity to underrepresented grade categories.

From a practical perspective, the proposed framework is most appropriately positioned as a screening-support tool rather than an autonomous grading system. Its capacity to flag lower-performing students supports early formative intervention in competency-based learning environments. The findings also demonstrate that simpler, well-regularized models can achieve stronger generalization than ensemble methods when datasets are small and score distributions are narrow, an important consideration for future EDM applications in similar educational contexts.

This study is limited by its reliance on a single academically filtered dataset with a narrow score range (75–92) and moderate class imbalance, which constrains model separability and generalizability. Future research should incorporate more diverse datasets spanning a broader performance range, explore alternative resampling strategies, and investigate advanced hybrid architectures to improve predictive performance in competency-based educational settings.

## REFERENCES

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and evaluation of different machine learning algorithms for predicting student

performance. *Computational Intelligence and Neuroscience*, 2022. https://doi.org/10.1155/2022/4151487

Bulut, O., Tan, B., Mazzullo, E., & Syed, A. (2025). Benchmarking variants of recursive feature elimination: Insights from predictive tasks in education and healthcare. *MDPI (Information), 16*(476), 1–21. https://doi.org/https://doi.org/10.3390/info16060476

Dimić, G., & Pecić, L. (2024). Proposed approach for overcoming the impact of unbalanced distribution in predicting students ' performance. *TEM Journal, 13*(4), 2839–2849. https://doi.org/10.18421/TEM134

Evangelista, E. D. L., & Sy, B. D. (2022). An approach for improved students ' performance

prediction using homogeneous and heterogeneous ensemble methods. *International Journal of Electrical and Computer Engineering (IJECE), 12*(5), 5226–5235. https://doi.org/10.11591/ijece.v12i5.pp5226-5235

Khairy, D., Alharbi, N., Amasha, M. A., & Areed, M. F. (2024). Prediction of student exam performance using data mining classification algorithms. *Education and Information Technologies, 29*(16), 21621–21645. https://doi.org/10.1007/s10639-024-12619-w

Malik, S., Mahanty, C., Mohanty, J., Vaghela, K., Narmadha, T., Sivaranjani, R., Bhutto, J. K., Khan, A., & Zewdie, A. (2025). Enhancing education quality with hybrid clustering and evolutionary neural networks in a multi phase framework. *Scientific Reports, 15*(21323), 1–18. https://doi.org/https://doi.org/10.1038/s41598-025-04622-z

Men, Y., & Zhao, M. (2025). Hybrid machine learning techniques for improving student management and academic performance. *International Journal of Information and Communication Technology, 26*(17). https://doi.org/10.1504/ijict.2025.10071317

Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *MDPI (Applied Sciences), 11*(237), 1–28. https://doi.org/https://doi.org/10.3390/app11010237

Odoom, S., Osei, E. O., Effah, E. Q., Bakariwie, A., & Rufai, A. A. (2026). Intelligent profiling beyond grades using a stacking ensemble framework for student success prediction. *Springer Nature, 5*(242), 1–19. https://doi.org/https://doi.org/10.1007/s44217-026-01277-4

Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics, 12*(2018), 820–836. https://doi.org/https://doi.org/10.1162/tacl_a_00675

Pamungkas, L., Dewi, N. A., & Putri, N. A. (2024). Classification of Student Grade Data Using the K-Means Clustering Method. *Jurnal Sisfokom (Sistem Informasi Dan Komputer), 13*(1), 86–91. https://doi.org/10.32736/sisfokom.v13i1.1983

Raj, S. A. P., & Vidyaathulasiraman. (2021). Determining optimal number of K for e-learning groups clustered using K-medoid. *(IJACSA) International Journal of Advanced Computer Science and Applications, 12*(6), 400–407. https://doi.org/https://dx.doi.org/10.14569/IJACSA.2021.0120644

Roslan, M. H., & Chen, C. J. (2022). Educational data mining for student performance prediction : A systematic literature review (2015-2021*). International Journal of Emerging Technologies in Learning, 17*(05), 207–225. https://doi.org/https://doi.org/10.3991/ijet.v17i05.27685

Sayed, H. E. F., & Fouad, M. M. (2025). Hybrid machine learning models for enhanced student performance prediction. *SciNexuses, 2*, 167–183.

Scornet, E. (2023). Trees, forests, and impurity-based variable importance in regression. *Institute of Mathematical Statistics (IMS) Association Des Publications de l'Institut Henri Poincaré., 59*(1), 21–52. https://doi.org/https://doi.org/10.1214/21-AIHP1240

Srinivasulu, A., & Palanisamy, V. (2024). Data-Driven revolution in academic support for mathematics underachievers through random forest individual and hybrid model. *Journal*

*of Artificial Intelligence and System Modelling, 02*(03), 1–21. https://doi.org/https://doi.org/10.22034/jaism.2024.469529.1048

Wongoutong, C. (2024). The impact of neglecting feature scaling in k-means clustering. *Public Library of Science (PLOS), 19*(12), 1–19. https://doi.org/10.1371/journal.pone.0310839

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *MDPI (Information), 14*(54), 1–15. https://doi.org/https://doi.org/10.3390/info14010054

Wu, M. (2026). K-Means clustering-based feature generation for student performance prediction. *ICCK Transactions on Educational Data Mining, 2*, 14–28. https://doi.org/https://doi.org/10.62762/TEDM.2026.716076

Yağcı, M. (2022). Educational data mining: prediction of students ' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(11). https://doi.org/10.1186/s40561-022-00192-z

Yanyu, G., Jizu, L., Cliff, D., & Huayun, D. (2025). Classification and prediction of miners' emergency response competence under sudden events based on seed k-means and stacking learning algorithm. *Engineered Science, 38*(1877), 1–26. https://doi.org/https://dx.doi.org/10.30919/es1877

Zhang, X., Zhang, Y., Chen, A. L., Yu, M., & Lihao, Z. (2025). Optimizing multi label student performance prediction with GNN-TINet : A contextual multidimensional deep learning framework. *Public Library of Science (PLOS) ONE, 20*(1), 1–25. https://doi.org/10.1371/journal.pone.03