

Comparative Analysis of K-Means and Random Forest for Student Academic Performance Prediction

1. Mustapha Malami Idina

Department of Computer Science, Abdullahi Fodio University of Science and Technology Aleiro, Nigeria.

2. Mubarak Jibril Yeldu

Department of Computer Science, Abdullahi Fodio University of Science and Technology Aleiro, Nigeria.

DOI: doi.org/10.5281/zenodo.23121738

Abstract

The increasing availability of educational data has created opportunities for machine learning (ML) techniques to support academic performance analysis, early identification of students at risk of poor outcomes, and evidence-based educational decision-making. This study presents a comparative analysis of K-Means clustering and Random Forest (RF) for student academic performance prediction. Although both techniques can discover useful patterns in educational datasets, they differ fundamentally in their learning paradigms: K-Means is an unsupervised clustering algorithm that identifies groups of students with similar characteristics, whereas Random Forest is a supervised ensemble-learning algorithm capable of directly predicting predefined academic outcomes. The study adopts the UCI Student Performance dataset, which contains demographic, social, school-related, behavioural, and academic variables. The proposed methodology involves data preprocessing, feature transformation, exploratory analysis, K-Means clustering, Random Forest classification, and comparative evaluation using accuracy, precision, recall, F1-score, cluster quality, and computational considerations. Recent studies indicate that tree-based methods, particularly Random Forest and related ensemble models, remain widely used in student-performance prediction, while clustering provides useful insights into heterogeneous student profiles. The analysis demonstrates that the two algorithms should not be regarded as direct substitutes: K-Means is primarily valuable for discovering latent student groups, whereas Random Forest is more appropriate when the objective is explicit prediction of academic-performance categories. The study therefore highlights the complementary rather than identical roles of clustering and supervised learning in educational data mining.

Keywords: Academic Performance Prediction, K-Means, Random Forest, Educational Data Mining, Machine Learning.

Article Information

Article Type:
Research Article

Received:
30 August 2026

Accepted:
20 September
2026

Published:
03 October 2026

1:0. INTRODUCTION

Student academic performance is an important indicator of educational achievement and institutional effectiveness. Universities and schools increasingly collect large amounts of educational data through student information systems, learning management systems, examinations,



Copyright © 2026, Authors retain copyright. Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. <https://creativecommons.org/licenses/by/4.0/> (CC BY 4.0 deed)

attendance records, and online learning platforms. These data provide opportunities for educational institutions to identify patterns associated with academic success and failure and to develop interventions for students who may require additional support.

Educational Data Mining (EDM) and learning analytics have consequently become important research areas for applying machine learning to educational problems. Recent systematic reviews show increasing use of machine learning, deep learning, clustering, and explainable artificial intelligence for student-performance prediction and academic decision support (Chek, et al, 2026, García-García, et al, 2026, Mones et al, 2026). A comprehensive review of student-performance prediction found that tree-based models and neural networks are among the dominant predictive approaches, while explainability is increasingly considered important for translating predictions into meaningful educational decisions.

Among machine-learning techniques, Random Forest has received considerable attention because it combines multiple decision trees and can model nonlinear relationships between student characteristics and academic outcomes. Recent educational studies continue to employ Random Forest alongside other supervised algorithms for classification and prediction tasks (Lou, et al., 2025, Colvin, 2025, Malik et al., 2025)

In contrast, K-Means is an unsupervised clustering algorithm. Rather than learning from predefined academic labels, it divides students into groups according to similarities among their characteristics. Recent research demonstrates that K-Means can be used to identify groups such as high-, medium-, and low-performing students and to reveal behavioural patterns associated with academic achievement (Durachman, 2025, Abdul Rahman, 2025, Qadir et al., 2025).

The distinction between these two approaches is important. K-Means does not inherently predict a target variable in the same way as Random Forest. Instead, it discovers latent structures in the data. Random Forest, on the other hand, requires labelled training data and directly learns the relationship between predictor variables and an outcome. Therefore, comparing them provides insight into the difference between student segmentation and academic-performance prediction.

The objective of this study is therefore to comparatively examine K-Means and Random Forest for student academic-performance analysis, with particular attention to their methodological differences, predictive capability, interpretability, and practical educational applications.

2. Literature Review

2.1 Machine Learning in Educational Data Mining

Machine learning has become an important component of educational data mining because educational datasets contain heterogeneous variables describing student demographics, academic history, behaviour, engagement, and learning activities. Recent reviews indicate that machine-learning applications in education include academic-performance prediction, dropout-risk identification, automated assessment, learning progression modelling, and student-support systems (Jdaitawi et al., 2026).

A systematic review of higher-education academic-performance prediction identified personal, academic, and environmental variables as important predictors. The review reported the use of regression, Random Forest, Support Vector Machines, gradient boosting, and deep-learning approaches. It also emphasized accuracy, precision, recall, F1-score, and error-based measures as common evaluation metrics (Yuan et al., 2026).

Similarly, Mones synthesized 58 empirical studies and found extensive use of Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, and Artificial Neural Networks for identifying students at risk of poor academic outcomes.

2.2 Random Forest for Academic Performance Prediction

Random Forest is an ensemble algorithm that constructs multiple decision trees using randomly selected subsets of observations and features. The final prediction is produced through aggregation

of the individual trees. Its ability to capture nonlinear relationships and interactions makes it suitable for educational datasets containing demographic, behavioural, and academic variables.

(Colvin, 2025), for example, conducted a comparative study involving generalized linear regression, decision trees, and Random Forest regression for predicting end-of-course examinations. Their results demonstrated that model performance depends strongly on the specific dataset and prediction task, emphasizing the importance of empirical comparison rather than assuming that one algorithm will always outperform others.

(Malik et al, 2025) similarly examined feature selection and classification techniques for student-performance prediction. Their study reported strong performance from ensemble and tree-based methods after feature-selection procedures, demonstrating the importance of selecting informative predictors in educational datasets. Recent work has also extended Random Forest-based student prediction toward explainability. A study incorporated Random Forest and other models with SHAP, LIME, ELI5, and Shapash to explain predictions and identify variables associated with academic performance.

2.3 K-Means for Student Performance Analysis

K-Means is an unsupervised learning algorithm that partitions observations into KK clusters by minimizing the distance between observations and their assigned cluster centroids. For student data, the algorithm can identify groups of students with similar academic, behavioural, demographic, or engagement characteristics.

Recent educational research has applied clustering to distinguish different learner profiles. (Qadir et al, 2025) for example, used clustering to examine variations in student performance and identify groups of high-, intermediate-, and low-performing learners. (Rahman 2025) similarly used K-Means to investigate student behavioural patterns involving study hours, attendance, and tutoring activities.

The main advantage of K-Means in this context is that it can reveal structures that may not be known before analysis. However, because K-Means does not use predefined target labels, the resulting clusters do not automatically correspond to official academic categories. Researchers must interpret the clusters after formation.

2.4 Research Gap

Existing research demonstrates extensive use of supervised machine learning for academic prediction and increasing interest in clustering for student segmentation. However, K-Means and Random Forest perform fundamentally different tasks. Treating clustering output as equivalent to supervised prediction can therefore produce misleading comparisons.

A useful comparative analysis should evaluate not only predictive accuracy but also the purpose of each method, required data labels, interpretability, computational requirements, and practical educational applications. This study addresses this gap by examining K-Means as a student-segmentation technique and Random Forest as a supervised academic-performance prediction technique.

3. Methodology

3.1 Research Design

The study adopts a quantitative experimental design based on educational data mining. The overall workflow consists of: Dataset Acquisition, Data Cleaning, Feature Transformation, Exploratory Analysis, K-Means Clustering, Random Forest Prediction, Evaluation, Comparative Analysis. The experiment can be implemented using Python, Pandas, NumPy, Scikit-learn, and Matplotlib.

3.2 Dataset

The study uses the UCI Student Performance dataset, a widely used benchmark dataset for educational data mining. The dataset contains student information from two Portuguese secondary schools and provides separate Mathematics and Portuguese-language datasets. Variables include

demographic characteristics, family background, school information, study habits, social factors, absences, and academic grades. The final grade, G3, ranges from 0 to 20 and can be used as the academic-performance outcome.

The Mathematics dataset contains 395 students, while the Portuguese dataset contains 649 students. The datasets contain 33 attributes. Importantly, UCI warns that G3 has a strong relationship with G1 and G2 because G1 and G2 are earlier-period grades. Consequently, excluding G1 and G2 produces a more meaningful early-prediction experiment because those variables otherwise provide information very close to the target.

Table 1. Major Variables Used in the Study

Variable	Description	Role
Age	Student age	Predictor
Sex	Student sex	Predictor
Address	Urban rural residence	Predictor
Medu	Mother's education	Predictor
Fedu	Father's education	Predictor
Studytime	Weekly study time	Predictor
Failures	Previous class failures	Predictor
Schoolsup	Additional educational support	Predictor
Famsup	Family educational support	Predictor
Internet	Internet access	Predictor
Absences	Number of absences	Predictor
G1	First-period grade	Optional predictor
G2	Second-period grade	Optional predictor
G3	Final grade	Target

3.3 Data Preprocessing

Preprocessing involves removing duplicate observations, checking missing values, encoding categorical variables, scaling numerical variables where appropriate, and examining outliers.

Categorical variables are transformed using numerical or one-hot encoding. Numerical variables are standardized before K-Means because clustering based on Euclidean distance can be disproportionately influenced by variables with larger numerical scales.

For Random Forest, scaling is generally not essential because decision trees split observations based on thresholds. However, using a consistent preprocessing pipeline helps ensure methodological comparability.

To formulate the prediction task, G3 can be transformed into academic-performance categories. For example:

- i. Low performance: $G3 < 10$
- ii. Pass medium performance: $10 \leq G3 < 15$
- iii. High performance: $G3 \geq 15$

The exact categorization should be stated clearly in the final experimental implementation and justified according to the educational grading context.

3.4 K-Means Model

K-Means seeks to minimize the within-cluster sum of squared distances:

$$J = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where K represents the number of clusters, C_k represents cluster k , x_i is a student observation, and μ_k is the centroid of cluster k .

The number of clusters can be selected using the Elbow Method and Silhouette Score. After clustering, the academic-performance distribution within each cluster is examined to determine whether the clusters represent meaningful student profiles.

For example, the resulting clusters may reveal groups characterized by:

- i. strong academic engagement and high grades,
- ii. moderate engagement and intermediate achievement,
- iii. weak engagement, high absenteeism, or previous failures.

These interpretations must be based on the actual experimental results rather than assumed before analysis.

3.5 Random Forest Model

Random Forest is trained using the labelled academic-performance categories. Multiple decision trees are constructed using bootstrap samples and randomly selected feature subsets.

For classification, the prediction can be represented as:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_n(x)\}$$

where $T_i(x)$ represents the prediction of tree I .

The model can be optimized through hyper parameters such as: number of trees, maximum tree depth, minimum samples per split, minimum samples per leaf, maximum number of selected features.

A stratified train-test split and cross-validation should be used to reduce the possibility that the reported results depend on a particular partition of the data.

4. Evaluation Metrics

Different evaluation procedures are required because K-Means and Random Forest have different objectives.

4.1 Random Forest Metrics

The following classification metrics are recommended:

Accuracy

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

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score

$$F1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Because academic-performance datasets can contain unequal numbers of students in different performance groups, macro-averaged precision, recall, and F1-score should be reported rather than accuracy alone.

K-Means Metrics

K-Means should primarily be evaluated using clustering metrics, including: Silhouette Score, Davies–Bouldin Index, Within-Cluster Sum of Squares, cluster size distribution, post-clustering academic-performance profiles.

A high Silhouette Score indicates that observations are relatively close to members of their own clusters and separated from other clusters.

5. Comparative Analysis

Table 2. Comparison of K-Means and Random Forest

Criterion	K-Means	Random Forest
Learning type	Unsupervised	Supervised
Requires target labels	No	Yes
Main purpose	Student segmentation	Performance prediction
Academic target directly predicted	No	Yes
Handles nonlinear relationships	Limited	Strong
Feature scaling	Important	Generally unnecessary
Interpretability	Cluster profiles	Feature importance tree decisions
Main evaluation	Silhouette, Davies-Bouldin	Accuracy, Precision, Recall, F1
Detects unknown groups	Yes	No
Suitable for early prediction	Indirectly	Directly
Handles mixed predictors	Requires preprocessing	Generally flexible
Main educational use	Student profiling	Risk/performance prediction

The comparison shows that K-Means and Random Forest should not be interpreted as identical predictive models. K-Means addresses the question:

What groups of students exist in the dataset?

Random Forest addresses a different question

Given the characteristics of a student, which academic-performance category is the student likely to belong to?

This distinction is particularly important in educational decision-making.

6. Results and Discussion

The experimental results should be reported separately for clustering and supervised prediction. Because K-Means and Random Forest optimize different objectives, reporting only a single accuracy measure for both algorithms would not constitute a methodologically sound comparison.

For K-Means, the analysis should identify the optimal number of clusters using the Elbow Method and Silhouette Score. Each cluster should then be profiled according to academic grades, study time, absenteeism, previous failures, parental education, and other relevant characteristics. This enables educators to identify distinct learner profiles without requiring predefined labels.

For Random Forest, the model should be evaluated using the held-out test data. Accuracy, precision, recall, macro-F1, confusion matrix, and feature importance should be reported. The model's ability to distinguish low-performing students is particularly important because educational interventions generally require identification of students who may need additional support.

Recent literature supports the usefulness of tree-based approaches for student-performance prediction. showed that predictive performance varies across algorithms and datasets, emphasizing the importance of empirical validation. (Malik et al, 2025) also demonstrated that feature-selection strategies can substantially influence the performance of student-performance classifiers. At the same time, clustering offers a complementary perspective. Recent K-Means studies demonstrate its ability to identify student behavioural and performance profiles that can support monitoring and targeted educational strategies.

An important methodological issue is data leakage. Since G1 and G2 are strongly related to G3, including them in a model intended for early prediction can make performance appear artificially

strong. The UCI dataset documentation explicitly highlights this relationship. Therefore, a rigorous experiment should present at least two scenarios:

Scenario A: Early prediction: G1 and G2 excluded.

Scenario B: End-of-year prediction: G1 and G2 included.

This distinction makes the resulting findings more meaningful for educational intervention.

Recent research has also emphasized fairness and explainability in academic-performance prediction. For example, a (Malik et al, 2025) study using the UCI data examined fairness across demographic and socioeconomic variables alongside Random Forest and other predictive models. Consequently, future implementations should not evaluate models solely on predictive performance but should also examine whether predictions systematically differ across student groups.

7. Practical Implications

The comparative findings have several implications for educational institutions.

First, Random Forest can support early-warning systems by identifying students who are likely to experience poor academic outcomes. Such systems could help instructors identify students requiring additional academic support.

Second, K-Means can support student segmentation. Institutions could use clustering to identify groups of students with similar learning behaviours, attendance patterns, study habits, and academic characteristics.

Third, the algorithms can be combined. For example:

Student Data, K-Means, Student Profile Identification, Random Forest, Academic Performance Prediction, Targeted Intervention.

In this hybrid architecture, K-Means provides information about student profiles, while Random Forest performs supervised prediction. This combination may provide richer information than relying on either technique independently.

Fourth, recent research emphasizes the growing importance of explainable AI in educational prediction. Predictive systems should therefore provide understandable reasons for predictions rather than simply producing an academic-risk label. (Malik et al, 2025)

8. Conclusion

This study presented a comparative analysis of K-Means and Random Forest for student academic-performance analysis. The two techniques represent fundamentally different machine-learning paradigms. K-Means is an unsupervised algorithm designed to discover groups of students with similar characteristics, whereas Random Forest is a supervised ensemble algorithm designed to predict predefined outcomes. The analysis indicates that K-Means is particularly useful for student segmentation, behavioural profiling, and discovery of hidden patterns, while Random Forest is more directly suited to academic-performance prediction and early identification of students at risk of poor outcomes. Consequently, the two algorithms should not be compared solely on predictive accuracy because K-Means does not inherently perform supervised prediction. The UCI Student Performance dataset provides an appropriate benchmark for comparative educational-data-mining experiments, but researchers must carefully address the strong relationship between G1, G2, and G3 to avoid information leakage. (Malik et al, 2025) Overall, a rigorous experimental study should report clustering quality for K-Means and classification performance for Random Forest, while also considering explainability, fairness, feature importance, and practical educational usefulness. Future research could develop a hybrid K-Means–Random Forest framework in which student profiles identified through clustering are incorporated into supervised academic-performance prediction.

Article Publication Details

This article is published in the **Globenix Multidisciplinary Journal**, ISSN 3139-0870 (Online). In Volume 2, Issue 4 (July - August) – 2026. The journal is published and managed by **Globenix Publisher**.

Copyright © 2026, Authors retain copyright. Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. <https://creativecommons.org/licenses/by/4.0/> (CC BY 4.0 deed)

Acknowledgements

We sincerely thank the editors and the reviewers for their valuable suggestions on this paper.

Funding

The authors declare that no funding was received for this work.

References

1. Chek, M. Z. A., Ismail, I. L., Jamal, N., Zulkifli, Z. H., Nariswari, R., & Asrulsani, M. S. (2026). *International Journal of Research and Innovation in Social Science*, 10(3), 72–80
2. Durachman, Y., & Abdul Rahman, A. W. (2025). <https://doi.org/10.63913/ail.v1i1.48>
3. García-García, F. J., Gómez-Núñez, M. I., & Molla-Esparza, C. (2026).. <https://doi.org/10.1007/s10734-026-01709-y>
4. Jdaitawi, M. M., Kan'an, A. F., Al-Mawadieh, R., Hamadneh, B. H., Alsabatin, H., Alkafaween, S., & Alomar, O. (2026). *Frontiers in Education*, 11. <https://doi.org/10.3389/educ.2026.1751900>
5. Kord, A., Aboelfetouh, A., & Shohieb, S. M. (2026).. *Journal of Computing in Higher Education*, 38, 38-76. <https://doi.org/10.1007/s12528-024-09426-0>
6. Lou, Y., & Colvin, K. F. (2025).. 4, 112. <https://doi.org/10.1007/s44217-025-00502-w>
7. Malik, S., Patro, S. G. K., Mahanty, C., et al. (2025).. *Scientific Reports*, 15, 8738. <https://doi.org/10.1038/s41598-025-92324-x>
8. *Journal on Efficiency and Responsibility in Education and Science*, 19(1), 1-14. <https://doi.org/10.7160/eriesj.2026.190101>
9. Qadir, R., Meghji, A. F., & Mahoto, N. A. (2025).. *Journal of Computing & Biomedical Informatics*. <https://doi.org/10.56979/1001/2025/1120>
10. Ziz, M., Faris, S., & Mansouri, K. (2026). *IEEE Access*, 14, 53267-53291. <https://doi.org/10.1109/ACCESS.2026.3680924>
11. UCI Machine Learning Repository. (2014). *Student performance* [Dataset]. University of California, Irvine. <https://doi.org/10.24432/C5TG7T>

Publisher's Note

GLOBENIX PUBLISHER remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of GLOBENIX PUBLISHER and/or the editor(s). GLOBENIX PUBLISHER disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.