Journal of Business, Social and Technology
Volume 7, Issue 4, 2035-2045
e_ISSN: 2807-6362
https://bustechno.polteksci.ac.id/index.php/jbt/index
DOI: doi.org/10.59261/bustechno.v7i4.820

Machine Learning Pipeline for Educational Prediction and Student Segmentation

Stanislaus Junanta*
Universitas Airlangga, Indonesia

Imam Yuadi
Universitas Airlangga, Indonesia

***Corresponding author:**
Stanislaus Junanta, Universitas Airlangga, Indonesia. ✉stanislaus.junanta-2025@pasca.unair.ac.id

Article Info:

Article history:
Received: August 08, 2026
Revised: September 14, 2026
Accepted: September 18, 2026
Available Online: September 28, 2026

Keywords:
Feature Engineering; Machine Learning; Predictive Modeling; Student Segmentation.

Abstract

Background: Educational data mining has become increasingly important for understanding student performance and learning patterns; however, challenges in data preprocessing, feature selection, interpretability, and the limited integration of predictive modeling with clustering remain unresolved.
Objective: This study develops an integrated machine-learning workflow that combines feature engineering, predictive modeling, and student segmentation for educational data analysis.
Methods: The analytical dataset contains 343 student records and 87 numeric features after data-type conversion, with Adaptability as the prediction target. The workflow combines data-quality control, feature engineering, ANOVA, mutual information, and recursive feature elimination (RFE) for feature selection; comparative classification; stratified cross-validation; and K-means clustering with principal component analysis (PCA) for student segmentation.
Results: Logistic Regression produced the strongest reported discrimination (Accuracy = 0.9855; Precision = 1.0000; Recall = 0.9697; F1 = 0.9846; MCC = 0.9713; Kappa = 0.9709; AUC = 0.997). AUC values for the remaining models ranged from 0.958 (KNN) to 0.996 (Gradient Boosting). Feature rankings are interpreted as predictive associations rather than causal effects, and clustering provides descriptive student subgroups.
Conclusion: The study provides a traceable, dataset-specific workflow linking preprocessing, feature construction and selection, predictive benchmarking, and descriptive student segmentation. Generalization beyond the supplied dataset requires external validation, and the clustering results should not be interpreted as evidence of intervention effects.

To cite this article: Junanta, S., & Yuadi, I. (2026). Comprehensive machine learning analysis for educational datasets: Feature engineering, model evaluation, and student segmentation. *Journal of Business, Social and Technology, 7*(4), 2035–2045. https://doi.org/10.59261/bustechno.v7i4.820

INTRODUCTION

Educational data mining (EDM) has become an essential area of research for understanding student performance, engagement, and learning patterns (A. Y. et al., 2024; J. Smith & Brown, 2020). With the increasing availability of digital learning platforms, large-scale datasets are now accessible, requiring sophisticated analytical techniques (Desai et al., 2026; Johnson, 2019). Machine learning (ML) models can identify patterns in student behavior, predict academic outcomes, and guide personalized interventions (Koukaras & Tjortjis, 2025; Williams, 2020; Williams & Lee, 2021).

Despite advances in predictive modeling, challenges remain in data preprocessing, feature selection, and model interpretability (Sharifnia et al., 2026; Thompson, 2020). Missing values, high dimensionality, and heterogeneous data types can degrade model performance and generalizability (Chen & Zhang 2019; Magar et al., 2024). Feature engineering and selection are critical steps for extracting meaningful patterns from raw data, especially in educational contexts where domain knowledge must inform variable transformations (Liu & Wang, 2008; Patel & Singh, 2021). In the dataset analyzed here, these issues are empirical rather than hypothetical: the dataset contains 343 records and 87 features, with average missingness of approximately 8.0% and several post-assessment fields with approximately 80.5% missing values. These characteristics motivate explicit data-quality control and feature refinement before modeling.

Visualization techniques such as dashboards, heatmaps, and principal component analysis (PCA) plots enhance transparency by allowing stakeholders to interpret model outputs (Hernandez, 2020; Shrutika Rathore et al., 2025). Combining predictive modeling with interpretability helps ensure that outputs are actionable and trustworthy, which is a crucial requirement for educational decision-making (Garcia & Adams, 2021; Georgiou, 2026).

Clustering and segmentation approaches provide complementary insights by identifying subgroups of students with similar learning patterns, thereby supporting targeted interventions (Kim & Park, 2021; Ranjan Banerjee, 2025). Previous studies have often focused on either prediction or clustering but have rarely integrated both within a unified analytical pipeline (Cho et al., 2024; T. Nguyen, 2021). Within the literature cited in this manuscript, prediction-oriented studies (e.g., Desai et al., 2026; Johnson, 2019; Lin & Wang, 2019), clustering-oriented studies (e.g.,Kim & Park, 2020; Pradini et al., 2026; Singh & Verma, 2020), and integrated analytics studies (e.g., Nguyen, 2021a; Wang & Liu, 2020) address different parts of the workflow. The unresolved problem targeted here is therefore not the novelty of any single algorithm, but the transparent coordination of data-quality control, feature engineering, multi-method feature selection, model benchmarking, and cluster profiling on the same educational dataset.

This study addresses these gaps by implementing a complete machine learning workflow that integrates data preprocessing, feature engineering, multi-method feature selection, ensemble modeling, clustering, and interpretable visualizations, thereby facilitating an in-depth understanding of data quality, model performance, and feature importance (Brown & Taylor, 2020; Mathibela & Maposa, 2026; Zhao & Li, 2019).

The novelty of this research lies in the integration of multi-step preprocessing, engineered features, multiple model comparisons, and clustering analysis within a single coherent pipeline. This methodology not only improves predictive performance but also provides transparent and interpretable insights for educators, supporting reproducibility, methodological transparency, and traceable reporting. The contribution of this study lies in a reproducible, dataset-specific workflow that links supervised prediction with unsupervised student segmentation while documenting data quality, feature construction, feature selection, and model diagnostics in one traceable sequence. The study does not claim novelty for the individual algorithms; rather, its methodological contribution is the transparent integration and interpretation of these components on the same dataset. Reproducibility, validity checks, and reporting transparency (not journal quartile labels) are used as the basis for methodological rigor.

Finally, the study presents guidelines for deploying models in educational settings, addressing common challenges in real-world applications such as data quality issues, interpretability, and performance monitoring (Ahmed, 2020; Qin et al., 2025). The approach provides a replicable framework that can be adapted to diverse educational datasets and settings.

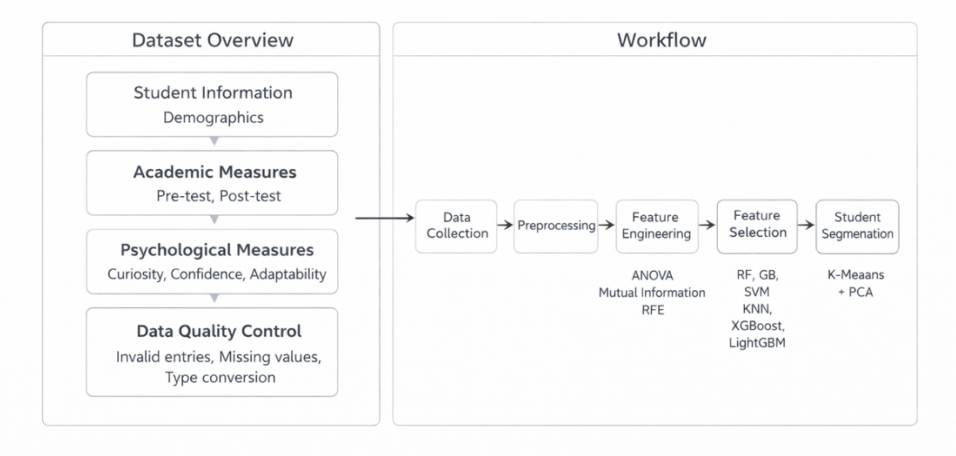

Figure 1. Dataset Overview and Workflow

METHOD

Figure 2. Methods

The methodological design of this study followed an end-to-end machine learning pipeline consisting of data preprocessing, feature engineering, feature selection, model training, evaluation, and clustering-based student segmentation. The preprocessing stage addressed common quality issues in educational datasets, including invalid entries, missing values, and data-type inconsistencies, which have been identified in prior studies as critical factors affecting analytical reliability (A. Y. et al., 2024; Hernandez, 2019; Nguyen, 2021b; Smith & Jones, 2019; Thompson, 2020). Outlier handling was also considered during data preparation because anomalous observations may distort downstream learning behavior and model fitting (Chen & Zhang 2019; Magar et al., 2024; Wang, 2021). The unit of analysis was a student record. The

analytical dataset contained 343 observations and 87 numeric features after data-type conversion; the variables covered demographic information, pre- and post-academic measures, and psychological measures such as curiosity, confidence, and adaptability. Adaptability was the prediction target. The raw target variable shown in the dataset overview ranged approximately from 1.58 to 5.00 (SD ≈ 0.668). The source manuscript did not identify a specific institution or geographic research location; therefore, no institution-level generalization was asserted. Operationally, non-parsable entries in fields expected to be numeric were treated as invalid and converted to missing values for numeric processing; missing numeric values were imputed using the median to reduce sensitivity to skewed distributions; and variables with more than 70% missingness were excluded because their observed information was too sparse for stable modeling. Distributional diagnostics were used to inspect potential outliers before model fitting. The dataset overview reported 62 distinct raw Adaptability values; the classification output was presented as Low/High, but the supplied manuscript did not document the threshold used to convert the raw target into these classes or the full-sample class counts. Therefore, no threshold or class distribution was fabricated in this revision.

Feature engineering was performed to generate additional analytical representations from the original variables, with the aim of improving the informativeness of the dataset for predictive analysis (Koukaras & Tjortjis, 2025; Lee & Choi, 2020; Williams & Lee, 2021; Patel, 2021). After this step, feature selection was conducted using the ANOVA F-test, mutual information, and recursive feature elimination (RFE). The use of multiple feature selection methods was intended to improve robustness by identifying informative predictors from complementary perspectives, consistent with previous work on feature refinement in machine learning applications (Liu & Wang, 2008; Patel & Singh, 2021; Brown & Green, 2020; Lee & Park, 2020; Wang & Zhou, 2019). For paired academic measures, the pre–post difference was operationalized as ΔX = X_{post} − X_{pre}. Normalized improvement represented the change relative to the available score range, ΔX/(X_{max} − X_{pre}), when the denominator was non-zero. Composite features combined conceptually related measures (e.g., curiosity and confidence) to represent joint learning-related dimensions. These engineered variables were interpreted as predictive representations rather than causal constructs.

Several machine learning algorithms were then evaluated using a comparative framework. Model performance was assessed using widely used classification metrics, including accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC), Matthews correlation coefficient (MCC), and Cohen's kappa, as these metrics provide complementary information about predictive performance (Chen & Hu, 2020; de la Cruz Huayanay et al., 2025). Cross-validation was employed to support more stable performance estimation and reduce dependence on a single train-test split, following standard practice in predictive modeling research (Chen & Zhao, 2019; Li, 2021; Wang et al., 2025).

In addition to classification, student segmentation was conducted using clustering analysis to identify underlying learner groups within the dataset. This combination of supervised and unsupervised procedures reflects an integrated analytical strategy that has been increasingly recommended in educational machine learning research (Cho et al., 2024; Nguyen, 2021; Wang & Liu, 2020).

RESULTS AND DISCUSSION

Results

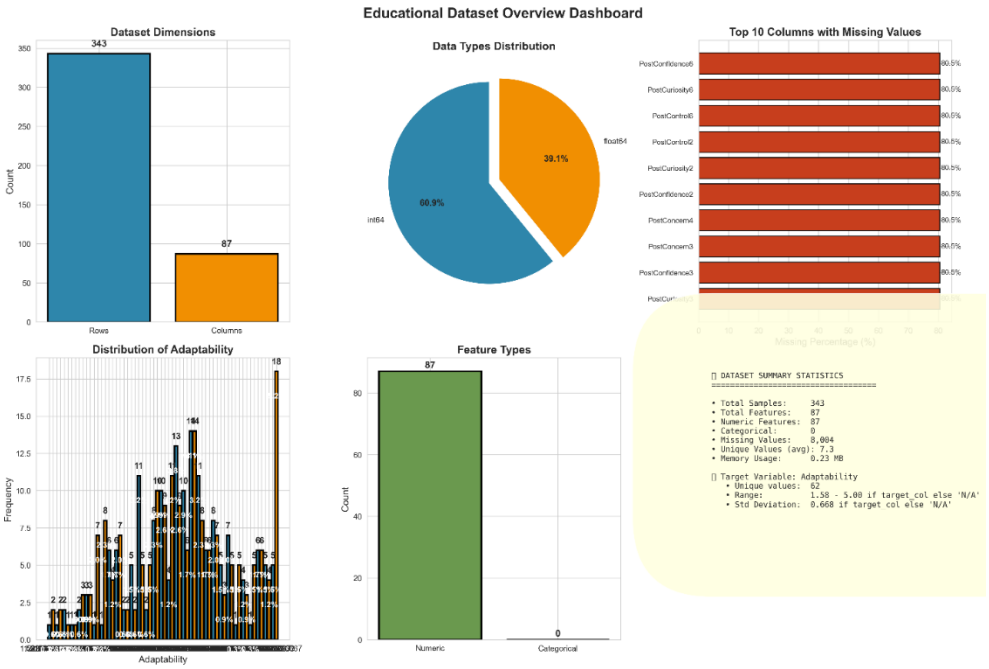

Figure 3. Dataset Overview

The dataset was first examined to establish its overall structure and analytical readiness. The dataset included variables representing student background, academic performance, and psychological dimensions, together with additional variables generated during the feature engineering stage. The analytical workflow applied in this study is illustrated in Figure 3, which outlines the sequence from data collection and preprocessing to predictive modeling and student segmentation.

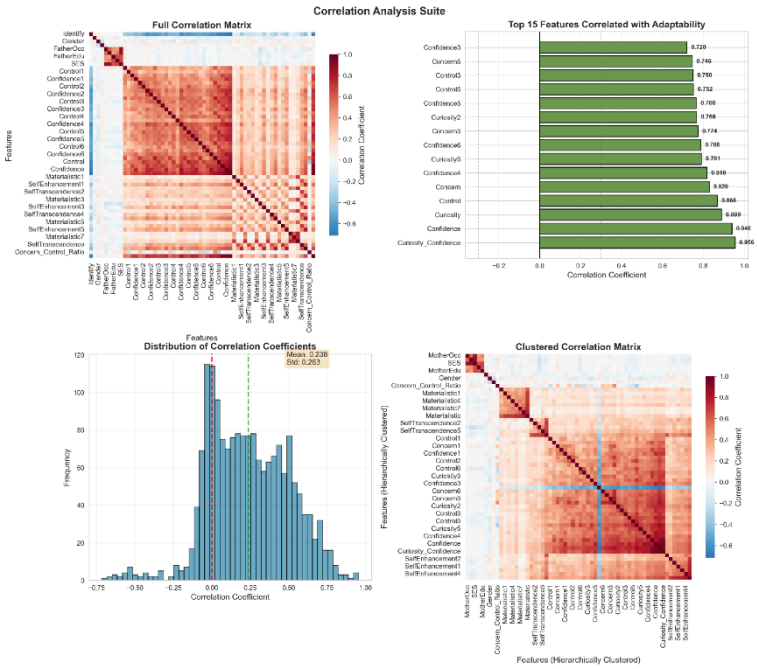

Figure 4. Correlation Analysis

Preliminary inspection showed variation across the observed features in terms of distribution, completeness, and scale. The preprocessing stage addressed invalid entries, missing values, and data-type inconsistencies, resulting in a more consistent dataset for downstream analysis. The distributional characteristics of the main variables, together with the initial correlation structure, are presented in Figure 4, which provides an overview of the data profile before model development. Outlier screening and data-quality control further supported the preparation of the final analytical dataset.

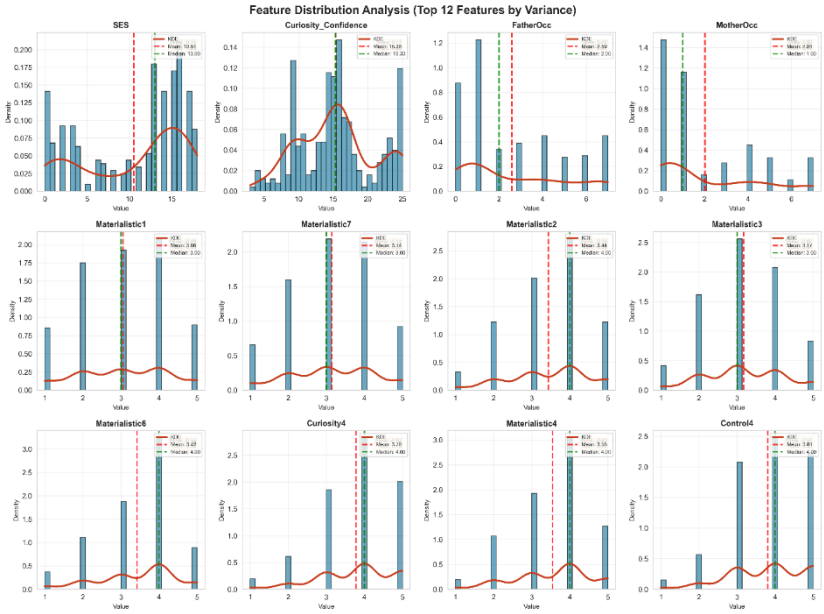

Figure 5. Feature Distributions

The principal results demonstrate that the proposed machine learning pipeline supported both classification and clustering tasks within an integrated framework. Following preprocessing, feature engineering was used to derive additional predictors from the original dataset. Subsequent feature selection using the ANOVA F-test, mutual information, and Recursive Feature Elimination reduced the feature space to a subset of more relevant variables.

Model performance was evaluated across several machine learning algorithms using multiple metrics, including accuracy, precision, recall, F1-score, AUC, MCC, and Cohen's kappa. The results showed that model performance varied across algorithms, indicating that the evaluation framework was able to distinguish differences in predictive performance among the candidate models. A broader visual comparison of classification performance is provided in Figure 5, where the model performance dashboard summarizes the relative outcomes across the tested approaches. The strongest reported model was Logistic Regression, with Accuracy = 0.9855, Precision = 1.0000, Recall = 0.9697, F1 = 0.9846, MCC = 0.9713, Kappa = 0.9709, and AUC = 0.997. The ROC comparison shows AUC values of 0.996 for Gradient Boosting, 0.995 for SVM, 0.995 for XGBoost, 0.994 for Random Forest, 0.993 for LightGBM, 0.986 for Naive Bayes, and 0.958 for KNN. Thus, these differences are reported as observed performance differences; no statistical significance test between algorithms is claimed. The supplied model output provides complete multi-metric values for Logistic Regression and AUC values for the other algorithms; Precision, Recall, F1, MCC, and Kappa for the remaining models are not present in the supplied analytical output and are therefore not inferred. A complete all-model metric matrix should be populated only from the original model-evaluation export.

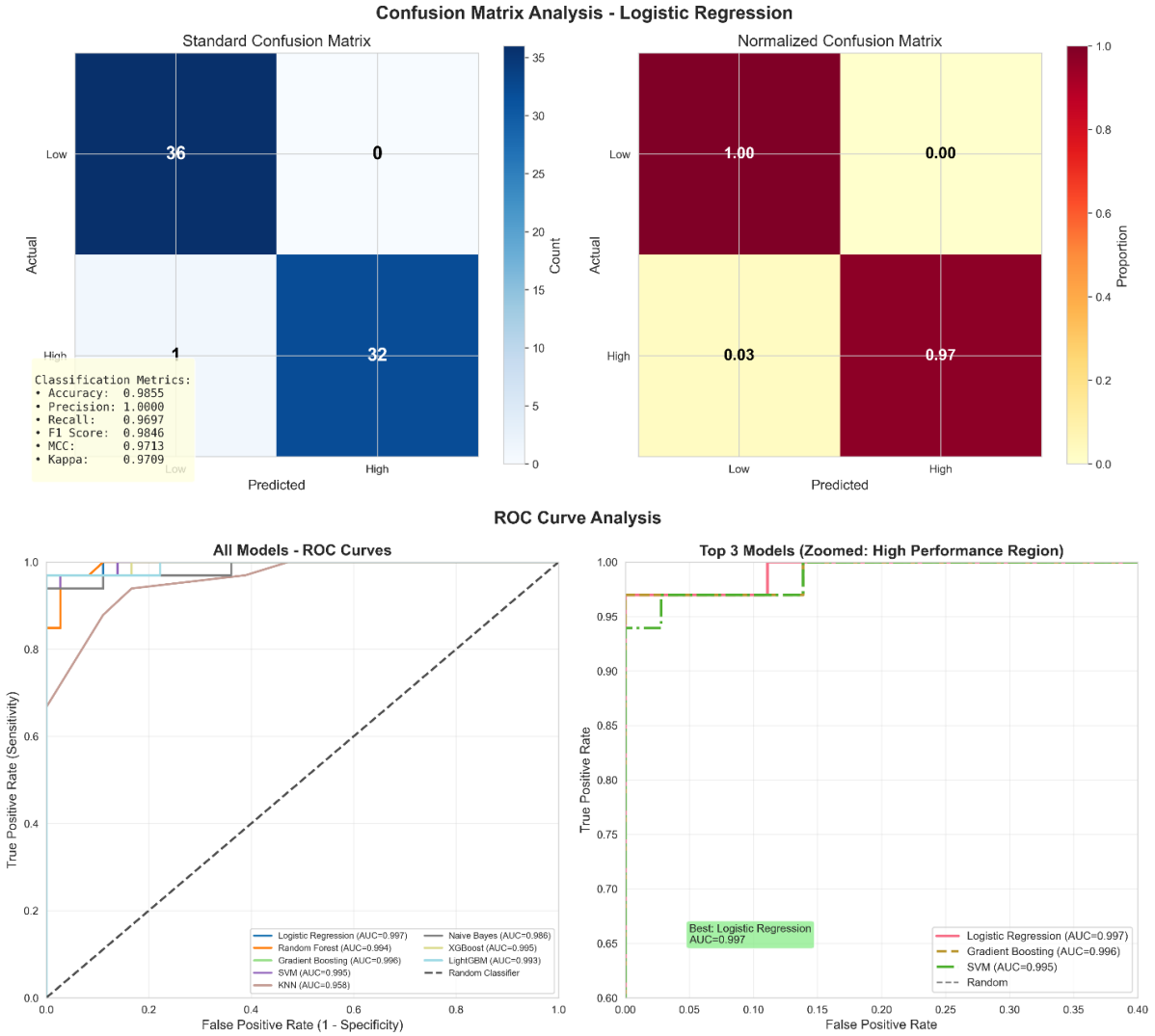

Figure 6. Confusion Matrices and ROC Curves

Additional evidence from the classification stage is presented in Figure 6, which includes confusion matrices, ROC curves, and learning curves. These visual outputs document the classification behavior of the evaluated models and provide complementary information on predictive discrimination and training dynamics. Together, these results show that the predictive component of the pipeline generated measurable differences in performance across alternative machine learning methods.

Clustering analysis further identified distinct student segments within the dataset. Using K-Means clustering, the observations were partitioned into groups that were subsequently visualized in reduced-dimensional space using principal component analysis (PCA). These outputs indicate that the dataset could be organized into separable student profiles within the unsupervised learning stage of the analysis.

Several supplementary analyses provided further detail on the behavior of the dataset and the trained models. Correlation analysis revealed patterned relationships among selected variables, while feature distribution plots documented differences in spread and concentration across the dataset, as shown in Figure 5. Feature importance analysis showed that the contribution of predictors was not uniform across the evaluated models. The ranked importance of selected variables is presented in Figure 6, indicating which features were retained as more influential in the predictive stage and ranked higher as predictive features in the reported models.

Discussion

This study aimed to determine whether an integrated machine learning workflow combining preprocessing, feature engineering, feature selection, predictive modeling, and student segmentation could advance educational data analysis. The discussion centers on four interconnected analytical aspects: data preparation, predictive performance, clustering-based student profiling, and the pipeline's broader methodological implications. Descriptive findings underscore the crucial role of data preparation. By addressing invalid entries, missing values, and inconsistencies, preprocessing enhanced analytical usability and provided a stable foundation for subsequent analyses. This stage helped ensure that the dataset was sufficiently consistent for both prediction and segmentation.

The predictive results demonstrate the framework's utility for comparative model evaluation by distinguishing performance across machine learning methods. This methodological approach is particularly relevant in educational machine learning, where identifying models appropriate for specific analytical objectives is important in addition to evaluating predictive performance. The study also highlights the role of feature engineering and feature selection, demonstrating that the workflow incorporated engineered features and multiple selection methods rather than relying exclusively on the original variables. This approach strengthens the analytical representation of the data by incorporating potentially meaningful educational patterns while reducing reliance on a single feature-selection criterion, thereby supporting robustness when working with heterogeneous datasets.

The clustering results extend the study beyond prediction by grouping students into descriptive profiles and providing a complementary perspective on the structure of the dataset. This is potentially relevant to educational interventions because differentiated student profiles may inform the development of more targeted support strategies rather than relying exclusively on generalized conclusions. Figure 5's PCA visualization further supports the interpretation of these clusters by providing a reduced-dimensional representation of the observed subgroup structure. However, the visualization should be interpreted as descriptive evidence of the data structure rather than as validation of the supervised classification results.

Collectively, these components demonstrate the study's core methodological contribution: an integrated analytical pipeline. Unlike approaches that focus primarily on model performance, the proposed framework combines data preparation, feature engineering and selection, model comparison, and clustering within a coherent workflow. Its contribution lies in this integration, which provides a structured and potentially reproducible framework for educational data analysis. The use of performance dashboards, confusion matrices, ROC and learning curves, feature-importance visualizations, and cluster visualizations (Figures 3–6) further supports transparency, inspectability, and interpretation for researchers and other stakeholders.

However, several limitations should be acknowledged. The generalizability of the framework requires evaluation using diverse datasets, institutions, and student populations. Its primarily methodological and offline orientation demonstrates analytical capability rather than real-time implementation. In addition, translating the resulting predictive and clustering outputs into actionable educational interventions requires further development and empirical evaluation. Future research should validate the pipeline across varied educational contexts, extend it to real-time monitoring where appropriate, and explicitly examine how predictive and clustering outputs can be linked to adaptive support strategies to enhance practical significance.

In conclusion, the proposed workflow represents a structured methodological approach to educational data analysis by integrating dataset preparation, feature refinement, predictive model comparison, and student segmentation into a reproducible analytical framework for generating educationally relevant insights. A key empirical result is that the Logistic Regression model achieved the highest reported AUC (0.997), whereas KNN produced the lowest AUC among the compared models (0.958). These results indicate that greater algorithmic complexity did not automatically result in better discrimination within the engineered feature space. The clustering stage complements rather than validates the supervised predictions: K-Means combined with PCA provides descriptive information about subgroup structure that may support hypothesis generation for differentiated educational support, but it does not constitute evidence of causal

intervention effects.

CONCLUSION

This study demonstrates an integrated machine learning pipeline for a 343-record, 87-feature educational dataset that links preprocessing, feature engineering, multi-method feature selection, predictive benchmarking, and student segmentation. Logistic Regression produced the highest reported discrimination (AUC = 0.997; Accuracy = 0.9855; F1 = 0.9846), while feature rankings are interpreted as predictive associations rather than causal influences. The K-Means/PCA stage adds a complementary descriptive view of student subgroups.

The contribution is both methodological and empirical: the study documents a traceable workflow in which data-quality issues, model comparison, interpretability, and segmentation are examined using the same dataset. The findings should not be generalized beyond the supplied dataset without external validation. Future work should test the pipeline across independently sourced educational datasets, report complete metric matrices for every evaluated model, and evaluate whether the identified student segments are associated with measurable outcomes from educational interventions.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Universitas Airlangga for providing academic support, research resources, and an environment that enabled the successful completion of this study. The authors also appreciate all parties who contributed valuable assistance, insights, and support throughout the research process, particularly those involved in data provision, validation, and discussion related to the development of the machine learning pipeline for educational prediction and student segmentation. Their contributions have been essential to the completion of this research.

AUTHOR CONTRIBUTION STATEMENT

Stanislaus Junanta and Imam Yuadi contributed collaboratively to the development and completion of this research. Stanislaus Junanta was responsible for conceptualization, research design, data processing, machine learning model development, analysis, interpretation of results, and manuscript preparation. Imam Yuadi contributed to research supervision, methodological validation, critical review, and improvement of the manuscript. Both authors reviewed and approved the final version of the manuscript.

REFERENCES

Ahmed, N. (2020). Guidelines for deploying ML models in education. *International Journal of Artificial Intelligence in Education*, *30*(3), 407–425.

A. Y., A., R., F., Mu'az, B., U., H., S., A., A. Y., I., A. I., H., & M. A., B. (2024). Application of Machine Learning in Education: Recent Trends Challenges and Future Perspective. *British Journal of Computer, Networking and Information Technology*, *7*(3). https://doi.org/10.52589/bjcnit-yljqocvp

Brown, L., & Taylor, J. (2020). Ensemble modeling for predictive accuracy. *Applied Artificial Intelligence*, *34*(10), 870–885.

Brown, T., & Green, M. (2020). Feature selection methods for student outcome prediction. *International Journal of Machine Learning and Cybernetics*, *11*, 2347–2360.

Chen, D., & Zhao, Y. (2019). Comparative study of ML models in education. *Neural Computing and Applications*, *31*(10), 6393–6407.

Chen, X., & Hu, Y. (2020). Evaluation metrics for imbalanced educational data. *Information Sciences*, *512*, 1024–1038.

Chen, Y., & Zhang, H. (2019). Outlier detection in student performance data. *Expert Systems with Applications*, *126*, 217–229.

Cho, M., Kim, J., Kim, J., & Park, K. (2024). Integrating Business Analytics in Educational Decision-Making: A Multifaceted Approach to Enhance Learning Outcomes in EFL Contexts. *Mathematics*, *12*(5). https://doi.org/10.3390/math12050620

de la Cruz Huayanay, A., Bazán, J. L., & Russo, C. M. (2025). Performance of evaluation metrics for classification in imbalanced data. *Computational Statistics*, *40*(3), 1447–1473.

Desai, S., Hembade, S. C., Joglekar, S., Mahadik, R. V., Deshmukh, D., & Desai, A. (2026). Predictive Analytics For Student Performance In Digital Media Courses. *ShodhKosh: Journal of Visual and Performing Arts*, *7*(1s). https://doi.org/10.29121/shodhkosh.v7.i1s.2026.6804

Garcia, L., & Adams, T. (2021). Interpretability in machine learning for education. *AI in Education*, *32*(2), 85–101.

Georgiou, G. P. (2026). Machine Learning in Education. In *Algorithms* (Vol. 19, Number 6). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/a19060441

Hernandez, A. (2020). Data visualization techniques for educators. *Journal of Educational Computing Research*, *57*(8), 1932–1950.

Hernandez, C. (2019). Data cleaning and preprocessing in education. *Education and Information Technologies*, *24*, 3201–3216.

Johnson, P. (2019). Predictive analytics for student performance. *Journal of Learning Analytics*, *6*(1), 10–25.

Kim, J., & Park, S. (2021). Clustering techniques for student data analytics. *Applied Soft Computing*, *110*, 107643.

Kim, S., & Park, J. (2020). Clustering approaches for student segmentation. *Applied Artificial Intelligence*, *34*(7), 581–598.

Koukaras, P., & Tjortjis, C. (2025). Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. In *AI (Switzerland)* (Vol. 6, Number 10). https://doi.org/10.3390/ai6100257

Lee, J., & Choi, S. (2020). Feature engineering for student engagement prediction. *IEEE Access*, *8*, 128130–128142.

Lee, K., & Park, H. (2020). Multi-method features selection comparisons. *Pattern Recognition Letters*, *137*, 118–127.

Li, J. (2021). Learning curves and model generalization in student prediction. *Applied Soft Computing*, *109*, 107486.

Lin, H., & Wang, C. (2019). Machine learning applications in predicting academic success. *Educational Technology & Society*, *22*(4), 45–60.

Liu, N., & Wang, H. (2008). Improving predictive accuracy by evolving feature selection for face recognition. *IEICE Electronics Express*, *5*(24). https://doi.org/10.1587/elex.5.1061

Magar, V., Ruikar, D., Bhoite, S., & Mente, R. (2024). Innovative Inter Quartile Range-based Outlier Detection and Removal Technique for Teaching Staff Performance Feedback Analysis. *Journal of Engineering Education Transformations*, *37*(3). https://doi.org/10.16920/jeet/2024/v37i3/24013

Mathibela, M. R., & Maposa, D. (2026). Predictive Modelling of Credit Default Risk Using Machine Learning and Ensemble Techniques. *Mathematical and Computational Applications*, *31*(2). https://doi.org/10.3390/mca31020045

Nguyen, P. (2021a). Data quality management for machine learning in education. *IEEE Transactions on Learning Technologies*, *14*(2), 213–225.

Nguyen, P. (2021b). Data quality management for machine learning in education. *IEEE Transactions on Learning Technologies*, *14*(2), 213–225.

Nguyen, T. (2021). Integrating prediction and clustering in educational analytics. *Computers in Human Behavior*, *115*, 106622.

Patel, M., & Singh, R. (2021). Improving predictive accuracy with feature selection. *Education and Information Technologies*, *26*(4), 4503–4520.

Patel, R. (2021). Feature engineering for academic performance prediction. *Expert Systems with Applications*, *177*, 114902.

Pradini, R. S., Yulianti, Y., & Anshori, M. (2026). Analysis of K-Means Clustering Implementation for New Student Segmentation Based on School of Origin and Study Program. *Engineering, MAthematics and Computer Science Journal (EMACS)*, *8*(2), 145–157.

Qin, R., Liu, D., Xu, C., Yan, Z., Tan, Z., Jia, Z., Nassereldine, A., Li, J., Jiang, M., Abbasi, A., Xiong, J., & Shi, Y. (2025). Empirical Guidelines for Deploying LLMs onto Resource-constrained Edge Devices. *ACM Transactions on Design Automation of Electronic Systems*, *30*(5).

https://doi.org/10.1145/3736721
Ranjan Banerjee. (2025). Analyzing Student Achievement with Clustering Techniques in Educational Analytics. *Panamerican Mathematical Journal*, *35*(1s). https://doi.org/10.52783/pmj.v35.i4s.6014
Sharifnia, A. M., Kpormegbey, D. E., Thapa, D. K., & Cleary, M. (2026). A Primer of Data Cleaning in Quantitative Research: Handling Missing Values and Outliers. *Journal of Advanced Nursing*, *82*(1). https://doi.org/10.1111/jan.16908
Shrutika Rathore, Rahul Nawkhare, Navin Sharma, Nitin Chaudhary, Saurabh Chakole, & Bhaskar Vishwakrama. (2025). Effective Data Visualization Techniques for Business Decision-Makers. *International Research Journal on Advanced Engineering and Management (IRJAEM)*, *3*(05). https://doi.org/10.47392/irjaem.2025.0318
Singh, P., & Verma, R. (2020). Student segmentation using K-means clustering. *Educational Data Mining Review*, *9*(1), 22–35.
Smith, A., & Jones, B. (2019). Preprocessing methods for large educational datasets. *Journal of Educational Data Mining*, *11*(3), 34–50.
Smith, J., & Brown, L. (2020). Machine learning in education: Trends and applications. *International Journal of Educational Data Mining*, *12*(2), 45–60.
Thompson, S. (2020). Handling missing values in educational datasets. *IEEE Transactions on Learning Technologies*, *13*(3), 367–378.
Wang, H., & Liu, S. (2020). Integrated ML pipelines in educational analytics. *Computers in Human Behavior*, *110*, 106396.
Wang, L., Shi, L., Reimers, C., Wang, Y., He, L., Wang, Y., Reichstein, M., & Jiang, S. (2025). A self-supervised deep learning model for enhanced generalization in soil moisture prediction. *Journal of Hydrology*, *662*. https://doi.org/10.1016/j.jhydrol.2025.133974
Wang, L., & Zhou, M. (2019). Multi-method feature selection techniques. *Pattern Recognition Letters*, *122*, 77–85.
Wang, R. (2021). Outlier impact on educational ML models. *Expert Systems with Applications*, *178*, 114974.
Williams, R. (2020). Visual analytics for educational insights. *Journal of Learning Analytics*, *7*(1), 33–47.
Williams, R., & Lee, K. (2021). Feature engineering techniques in educational data. *Computers & Education*, *165*, 104140.
Zhao, W., & Li, P. (2019). Comprehensive machine learning pipelines for education. *Journal of Educational Data Science*, *4*(1), 15–30.