An Optimized Ensemble Framework for Predicting Student Achievement in Educational Environments

Mansi Rawat; Sahil Parasar ; Dr. Meenu Vijarania; Dr.Swati Gupta
School of Engineering and Technology
K.R. Mangalam University Gurugram, India

Abstract—This study introduces an enhanced ensemble framework for predicting student performance early, using only pre-midterm behavioral indicators like study habits and attendance instead of exam scores. Although the baseline models such as Random Forest initially showed high accuracy, they were less validated for early predictions. To overcome this, a pipelined approach was developed that combines K-Means clustering with an ensemble of Random Forest, Gradient Boosting, and XGBoost models. By aggregating ordinal probability boosting to capture class label more accurately, the final model achieved 84.60% accuracy and a 0.87 recall rate in labeling at-risk students. These results showcase that behavioral data alone can support effective early-warning systems, offering educators an adaptable and intuitive methodology for early-stage performance improvement intervention.

Keywords—Educational Data Mining, Ensemble Learning Models, Student Performance Prediction, Early Alert System, At-Risk Students

I.Introduction

The explosion of data in education has provided a unique opportunity to enhance the performance of students. There is a large portion of the behaviour and engagement data that is collected by educational institutions but cannot be utilized effectively, especially when it comes to detect “at-risk” students before their grades start to decline.

Teachers are using mid-term or final test results to track progress of a student. The major problem in this system is that, by the time these assessments are gathered, its often too late for any intervention. There is a desperate need of predictive algorithms that can identify underperforming students using early indications such as attendance, study habits and background data before the performance of a student start to decline.

Educational data mining has brought machine learning tools like Random Forest and Logistic Regression in this field, but most of the models that are being used today still rely heavily on the test data from the past. This creates a “Catch-22” in which a student’s performance cannot be predicted until they have already failedanexam.

In order to close this gap, this study suggest an ideal ensemble architecture that ignores past exam scores of mid-terms. By collectively implementing k-means clustering and ensemble learning we can combine identical learning pattern and obtain accurate predictions. Moreover, to respect the natural hierarchy of grades, our hybrid categorization method uses ordinal modelling to respect the natural hierarchy of grades.

The main objective of this project is to provide a scalable solution that can give useful early warning symbols to show that we can effectively support students' achievement even before the first term paper is graded.

II.Related Work

Several machine learning techniques for forecasting student performance have been investigated in earlier research in Educational Data Mining (EDM). In order to forecast results without depending on previous academic grades, a recent study [1] used ensemble approaches, combining Random

DOI: https://doi.org/10.5281/zenodo.19563033

Forest for classification and Gradient Boosting for regression. Using RandomizedSearchCV, 5-fold cross-validation, and SHAP for interpretability, the model was further improved by hyperparameter tweaking. Despite achieving good prediction accuracy, the technique lacked behavioral signs and customization and instead concentrated on academic aspects.

In a different study, [2] used regression analysis and clustering to create a customized prediction framework that divided students into comparable groups. Compared to conventional one-size-fits-all models, this approach allowed for more customized insights and addressed ethical issues in educational institutions. Nevertheless, in order to increase prediction accuracy, the study did not fully utilize sophisticated ensemble or hybrid modeling methodologies.

Additionally, baseline machine learning techniques have been extensively researched. For instance, in addition to correlation-based feature selection, [3] assessed many models, such as k-Nearest Neighbors, Random Forest, Decision Tree, Logistic Regression, and Neural Networks. This study showed that binary classification consistently shows improved performance over multi-class prediction because of the class imbalance. These approaches perform well as a baseline, but they don't have the complexity needed for accurate early-stage prediction using behavioral attributes.

In the study, [4] a more aggregated view is offered which studied recent progress in EDM, identifying the efficiency of ensemble approaches, techniques for handling class imbalance like SMOTE, and the growing significance of interpretable AI methods like LIME and SHAP. The research also emphasized the necessity of incorporating many data sources, such as demographic and behavioral aspects, while maintaining data transparency and privacy.

Additionally, [5] emphasized the significance of family support and parental impact on student achievement and showed the powerful predictive power of Random Forest models. Deeper understanding of feature significance and model behaviour was obtained by the use of many explainability strategies. However, rather than enhancing predictive performance through integrated or hybrid frameworks, interpretability continued to be the key focus.

The prediction of student success in EDM has been improved by a number of research. Effective feature selection techniques, such Information Gain with Decision Tree or Chi-Square with Random Forest, can greatly increase accuracy while lowering feature redundancy, according to a study [6]. Similar to this, [7] stressed the need of preprocessing and prompt intervention while concentrating on the early detection of at-risk kids using classification algorithms. Another study [8] contrasted ensemble methods with conventional models, emphasizing the importance of behavioral predictors, hyperparameter adjustment, and scalability for big datasets. Furthermore, [9] showed that behavioral interaction patterns might be more predictive than time-based metrics alone by combining K-means clustering with Random Forest to discover learner archetypes.

The proposed study builds on previous research by combining behavioral characteristics, ensemble regression, clustering-based student segmentation using KMeans, correlation-based feature selection, and hybrid classification using ordinal probability boosting. This method, which relies only on behavioral and demographic characteristics without using midterm results, achieves a dependable hold-out test accuracy of 84.60% and offers a strong early-stage prediction framework and a useful early-warning system.

A. Gap Analysis:

- The majority of previous research, such as [1], [3], [6], [7] and [8], mostly concentrates on binary (pass/fail) or restricted multi-class categorization, ignoring a fine-grained six-class grading system (A–F), which limits in-depth study of academic performance.
- While solely pre-midterm behavioral and demographic factors are still understudied, many research, like [1] and [8], significantly rely on academic or past-grade data, which limits their efficacy for real early prediction.
- While some studies use ensemble methods [1], [4] or clustering [2], [9], these approaches are seldom combined into a single pipeline that combines ordinal probability modeling, ensemble regression,

DOI: https://doi.org/10.5281/zenodo.19563033

hybrid classification, and segmentation based on clustering.

- Prediction accuracy is the main emphasis of current methods, with little attention paid to creating useful early-warning systems that offer educators and students tailored and useful advice.
- Additionally, the ordinal structure of grade categories is not taken into account by the majority of previous models, which results in less accurate forecasts that miss the natural evolution of student performance levels.

III.Methodolgy

Using just pre-midterm behavioral and demographic characteristics, this study suggests a multi-stage framework for early student performance prediction. To provide precise and useful predictions, the whole pipeline combines preprocessing, feature engineering, clustering, ensemble regression, hybrid classification, and an ordinal probability boosting method.

A.Dataset Description

The study made use of the Student Performance Dataset from Kaggle, which was reduced from 25,000 entries to 15,000 records after eliminating the duplicates. The midterm or subject-specific scores (including English, science, and math) are strategically dropped from the dataset, which only consist of demographic characteristics and early-semester behavioral to guarantee accurate early prediction features.

Key features include age, gender, school_type, parent_education, study_hours, attendance_percentage, internet_access, travel_time, extra_activities, and study_method. The model first uses a continuous total score as the target variable, which is then converted into a detailed six-grade classification system (A, B, C, D, E, and F). Compared to traditional binary or limited multi-class approaches, this grading structure offers richer and more detailed information.

B. Data Preprocessing and Feature Engineering

In the data preprocessing stage, categorical variables were converted using one-hot encoding, and numerical features were standardized using StandardScaler. To more effectively capture student effort and consistency, a new feature, effective_study_time, was engineered, as defined in Equation (1):

$$\text{effective}_{\text{study}_{\text{time}}} = \text{study}_{\text{hours}} \times \left(\frac{\text{attendance\_percentage}}{100}\right) \quad (1)$$

As defined in Equation (1), this feature offers a more precise representation of student effort by considering both study time and attendance. Following this, features showing an absolute correlation of 0.3 or higher with the target variable were selected through correlation-based feature selection. As a result, study hours and effective study time were identified as the most significant indicators. Midterm scores were excluded to maintain the goal of early prediction.

C. Student Clustering

Students were grouped using KMeans clustering based on study_method, attendance_percentage, parent_education, and extra_activities for personalized modeling. Silhouette score analysis was applied to find the optimal number of clusters (k ranging from 2 to 9). This segmentation improves the model's ability to handle different behavioral patterns and helps identify similarities among students.

D. Baseline Models

The preprocessed dataset was evaluated using Logistic Regression, K-Nearest Neighbors, and Random Forest for benchmarking. The models achieved accuracy rates of 95%, 78.03%, and 99.57%, respectively. These models—Random Forest in particular—performed well, but they are not strong enough for early prediction in the absence of exam-related variables.

E. Proposed Global Ensemble Regression Model

The continuous total score was predicted using an ensemble regression model that used Random Forest, Gradient Boosting, and XGBoost (each with 300 estimators). Their outputs were averaged to determine the final anticipated score. A ($\geqslant$95), B ($\geqslant$85), C ($\geqslant$ 70), D ($\geqslant$55), E ($\geqslant$40), and F (<40) were the predetermined criteria used to map these scores to grades.

DOI: https://doi.org/10.5281/zenodo.19563033

F. Hybrid Classifier and Ordinal Probability Boosting

By include predicted_score as an input feature and training a tailored Random Forest Classifier (400 estimators, max_depth = 15), a hybrid classifier was created. An ordinal logistic model (LogisticIT) was used to create ordered probabilities in order to maintain grade order. The final boosted predictions were obtained by retraining the model after they were included as features.

G. Final Ensemble and Early-Warning System

By include predicted_score as an input feature and training a tailored Random Forest Classifier (400 estimators, max_depth = 15), a hybrid classifier was created. An ordinal logistic model (LogisticIT) was used to create ordered probabilities in order to maintain grade order. The final boosted predictions were obtained by retraining the model after they were included as features.

H. Evaluation Methodology

To preserve class balance, an 80/20 stratified train-test split was employed. Accuracy, precision, recall, F1-score (macro and weighted), and confusion matrix were used to assess performance. To guarantee repeatability, fixed random seeds were used in every experiment.

The whole suggested pipeline, including data preparation, feature engineering, clustering, ensemble regression, hybrid classification, ordinal probability boosting, and the final early-warning system, is depicted in Fig. 1.

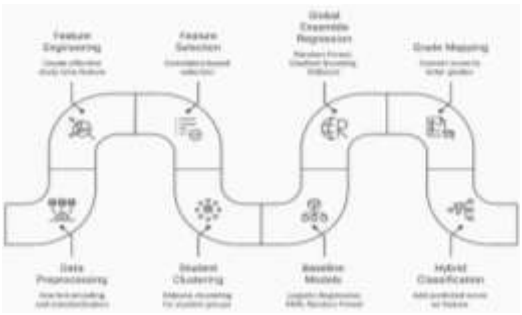

Fig. 1. Proposed EDM pipeline for early student performance prediction

III.Result and Discussion

A.Performance Comparison of Baseline and Proposed Models

Three baseline models were created using the same preprocessed features and an 80/20 stratified hold-out split in order to assess the efficacy of the suggested framework. Table 1 shows the performance metrics of all models, and Fig. 2. visually compares their accuracies.

TABLE.1
PERFORMANCE COMPARISON ON HOLD – OUT TESET SET

Model	Performance Comparison		
	Accuracy (%)	Macro F1-score	Notes
Logistic Regression (Baseline)	95.00	0.94	Strong linear baseline
K-Nearest Neighbors (Baseline)	78.03	0.77	Distance-based classifier
Random Forest (Baseline)	99.57	0.99	Best performing baseline
Global Ensemble Regression + Thresholds	77.34	0.77	Initial regression-based approach
Hybrid Classifier	82.84	0.81	Regression score added as extra feature
Tuned Hybrid Classifier	91.07	0.90	Improved RF with deeper trees
Ordinal-Probability Boosted Hybrid (Proposed)	84.60	0.85	Final model (hold-out test)

On unseen data, the suggested ordinal-probability boosted hybrid model produced a dependable 84.60% accuracy with a macro F1-score of 0.85. Using just early behavioral data, it maintained balanced performance while making a large improvement over the regression baseline (77.34%).

DOI: https://doi.org/10.5281/zenodo.19563033

Fig. 2. Accuracy Comparison of Baseline and Proposed Models on the Hold-out Test Set

As shown in Figure 4, the proposed Ordinal-Probability Boosted Hybrid model strikes a strong balance between performance and reliability, outperforming the initial regression approach and matching the best baseline models.

B.Detailed Analysis of the Proposed Model

All six grade classes (A–F) exhibit balanced performance in the final model. Significantly, it attained a high recall for the failing grade (F-class recall = 0.87), which is essential for identifying at-risk pupils early on.

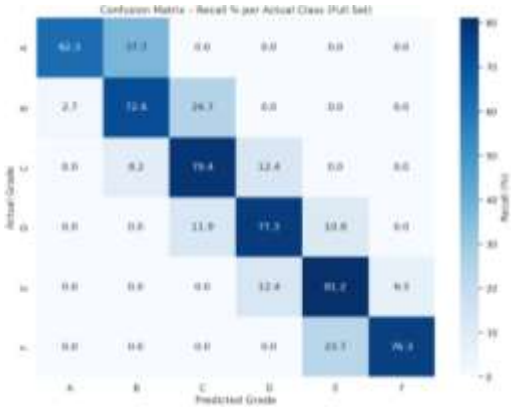

Fig. 3. Confusion Matrix (Recall % per Actual Class) of the Proposed Ordinal-Probability Boosted Hybrid Model on the Hold-out Test Set

The model performs exceptionally well on the failing grade F (76.3%), as seen in Fig. 3. and exhibits great recall across all grades. Additionally, the confusion matrix validates the efficacy of the ordinal probability boosting strategy by demonstrating low misclassification across neighboring classes. The six-class grading structure offers more precise and useful information than conventional binary or restricted multi-class systems.

C.Practical Early-Warning System and Dashboard

The creation of a useful early-alert dashboard that converts model predictions into instantly applicable information for educators and learners is a significant practical contribution of this study. In addition to forecasting the final grade, the algorithm also determines a risk level and produces tailored, practical suggestions.

Through this dashboard, educators can quickly identify students who are in danger and provide guidance such as increasing attendance, extending study hours, or altering study methods. This study and effective use of this data identified 6,422 high-risk kids out of 15,000 in the present dataset, which demonstrates its scalabilities for useful educational application.

D.Discussion

The results show that behavioral and demographic features alone are sufficient for early prediction, even without mid-term scores. The benefits of combination of ordinal modeling, clustering, ensemble regression and hybrid classification are highlighted through the evolution from baseline model (Random Forest, KNN, and Logistic Regression) to the recommended hybrid ensemble.

This six-class grading system and early warning dashboard closes the significant gap of previous literature, and final model achieves a balance between interpretability and accuracy. All things considered, the suggested approach offers a workable, scalable, and teacher-ready way to enhance student outcomes through early and focused intervention.

V.Conclusion

Using just pre-midterm behavioral and demographic characteristics, this study offers an efficient methodology for predicting early student success. The suggested model obtained a dependable accuracy of 84.60% with a macro F1-score of 0.85 by combining feature engineering, KMeans clustering, ensemble regression, hybrid classification, and ordinal probability boosting.

DOI: https://doi.org/10.5281/zenodo.19563033

The adoption of a six-class grading system (A–F) in conjunction with a useful early-warning system that offers tailored recommendations is a significant contribution that increases the model's actionability for teachers. The findings show that precise early prediction may be made without depending on exam-based characteristics, allowing for prompt intervention.

All things considered, the suggested framework provides a teacher-ready, scalable, and interpretable way to enhance student outcomes. Future work will focus on improving generalizability, use explainable AI techniques, and deploying the system in real educational environments.

Refernece

[1] N. Al-Shanableh et al., "Forecasting Students' Academic Performance in Educational Data Using Machine Learning Techniques," International Journal of Information and Communication Technology Education, vol. 22, no. 1, Jan. 2026, doi: 10.4018/IJICTE.399756.

[2] A. Angeioplastis, J. Aliprantis, M. Konstantakis, and A. Tsimpiris, "Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques," Computers, vol. 14, no. 3, Mar. 2025, doi: 10.3390/computers14030083.

[3] S. O. Semerikov, O. V. Bondarenko, P. P. Nechypurenko, T. A. Vakaliuk, and I. S. Mintii, "Student elective course selection patterns and satisfaction determinants identified through educational data mining," Sci. Rep., vol. 16, no. 1, Dec. 2026, doi: 10.1038/s41598-026-37712-7.

[4] E. Kalita et al., "Educational data mining: a 10-year review," Dec. 01, 2025, Springer Science and Business Media B.V. doi: 10.1007/s10791-025-09589-z.

[5] T. Niu, T. Liu, Y. T. Luo, P. C. I. Pang, S. Huang, and A. Xiang, "Decoding student cognitive abilities: a comparative study of explainable AI algorithms in educational data mining," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-12514-5.

[6] S. Malik et al., "Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-92324-x.

[7] O. Scheuer and B. M. Mclaren, "Educational Data Mining," Springer, 2011.

[8] M. M. Islam, F. H. Sojib, M. F. H. Mihad, M. Hasan, and M. Rahman, "The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system," Telematics and Informatics Reports, vol. 18, Jun. 2025, doi: 10.1016/j.teler.2025.100203.

[9] //Mrcis Org, S. Johnson, and M. Richardson, "Evaluating the Learning Outcomes of Metaverse-Based Educational Platforms using Learning Analytics and Educational Data Mining." [Online]. Available: https://mrcis.org

DOI: https://doi.org/10.5281/zenodo.19563033