RESEARCH ARTICLE

K-Means Clustering-Based Feature Generation for Student Performance Prediction

Meiting Wu[1,*]

[1] College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China

Abstract

With the development of educational technology and the accumulation of big data, student performance prediction has become a hot topic in the field of education. However, traditional manual statistical methods have limitations in dealing with complex data and are difficult to achieve high-precision prediction. To address this gap, this study proposes a clustering-based feature generation framework to enhance prediction performance. Firstly, the multilayer perceptron (MLP) model is employed to evaluate the effectiveness of the clustering algorithms (K-Means, DBSCAN, and hierarchical clustering) for feature generation. Then, the best clustering algorithm (K-Means) is applied to generate features that are subsequently integrated with original features to construct augmented datasets. Subsequently, grid search is adopted to optimize model hyperparameters, and six machine learning models, including MLP, support vector machine, random forest, Bagging, XGBoost, and CatBoost, are trained and evaluated on datasets with and without clustering-generated features. Experimental results demonstrate that K-Means-based feature generation can effectively improve prediction performance under certain conditions. However, the performance gains are influenced by data characteristics, feature distributions, and model structures. The findings also reveal that clustering-derived features do not universally enhance all machine learning algorithms, highlighting the necessity of selecting appropriate model–feature integration strategies in practical applications.

Keywords: clustering feature generation, academic performance prediction, educational data mining, machine learning, clustering algorithm.

1 Introduction

With the rapid development of educational technology and the continuous accumulation of educational big data, student performance prediction has attracted increasing attention from both academia and educational institutions. Academic performance is a key indicator for evaluating students' learning outcomes and plays an essential role in supporting educational decision-making and resource allocation [1]. Therefore, improving the accuracy and reliability of performance prediction models has become a critical research topic in educational data mining and learning analytics. Traditional statistical analysis methods can provide basic insights into students' learning status. However, as educational data become increasingly complex, high-dimensional,

Submitted: 08 February 2026
Accepted: 04 March 2026
Published: 07 March 2026

Vol. 2, **No.** 1, 2026.
10.62762/TEDM.2026.716076

***Corresponding author:**
✉ Meiting Wu
1416609104@qq.com

Citation
Wu, M. (2026). K-Means Clustering-Based Feature Generation for Student Performance Prediction. *ICCK Transactions on Educational Data Mining*, 2(1), 14–28.

and heterogeneous, these methods often fail to capture nonlinear relationships among learning variables, resulting in limited prediction accuracy. In recent years, machine learning and data mining techniques have demonstrated strong capabilities in extracting valuable knowledge from large-scale datasets and have been widely applied to student performance prediction tasks [2]. Among these techniques, clustering analysis has shown effectiveness in discovering hidden data structures and grouping similar samples, making it a widely used tool in data mining processes [3].

Despite these advancements, existing performance prediction models still face several limitations. Many studies rely primarily on original feature representations and lack targeted processing strategies for heterogeneous student samples, which may reduce model accuracy and robustness [4, 5]. Moreover, clustering techniques are often used solely for student grouping or classification, while their potential for feature generation and representation enhancement remains insufficiently explored.

To address these challenges, this study proposes a student performance prediction framework based on clustering-driven feature generation. Specifically, clustering algorithms are employed to generate cluster labels that capture latent structural information within the data. These clustering-derived features are then integrated with original features to construct enhanced datasets for training and evaluating multiple machine learning models. The effectiveness of the proposed framework is validated using multiple evaluation metrics to examine whether clustering-based feature augmentation can improve prediction performance.

From a practical perspective, improving prediction accuracy is closely related to enhancing educational quality, which remains a fundamental goal of modern education systems [6]. Student academic performance serves as an important indicator of educational quality and provides valuable feedback for teaching and learning improvement [7]. This study proposes the topic of "Research on Learning Performance Prediction Based on Cluster Feature Generation", which has extensive practical applications and academic value for teachers, students and the academic community. The proposed framework offers practical benefits for multiple stakeholders. For educators, clustering-based feature extraction can reveal representative learning patterns and support early teaching intervention [8]. For students, predictive insights can help identify key factors influencing academic outcomes and promote learning motivation. From a research perspective, this study explores the integration of clustering-based feature generation with machine learning prediction models, providing new insights into hybrid feature learning strategies in educational data mining.

2 Related Work

Research on student performance prediction has attracted considerable attention and achieved significant progress in recent years. Early prediction models mainly relied on traditional academic indicators, such as exam scores and course grades. However, as student characteristics become increasingly diverse and learning environments grow more complex, relying solely on academic performance data is often insufficient for accurate prediction. Studies have shown that multiple factors, including learning interests, behavioral patterns, and social interactions, also play important roles in influencing student performance. Consequently, researchers have increasingly focused on extracting multidimensional student features and developing advanced predictive models using data mining and machine learning techniques.

Several studies have applied traditional machine learning approaches to student performance prediction. For example, Pandey et al. [9] employed decision tree algorithms to construct classification rules for student performance prediction, demonstrating their effectiveness in supporting teaching strategy optimization. Xu et al. [10] proposed an incremental learning random forest model that improves prediction accuracy while maintaining model stability during updates, enabling long-term monitoring of student learning behavior. This algorithm not only improves the prediction accuracy but also solves the stability problem after model updates, which is conducive to model iteration optimization and long-term tracking of students' learning behaviors. Yang et al. [11] utilized multiple linear regression to predict graduation performance based on first-year academic results, providing valuable guidance for teaching plan adjustments.

With the development of intelligent optimization and deep learning technologies, more advanced prediction models have been proposed. Waheed et al. [12] combined genetic algorithms with backpropagation neural networks to improve prediction accuracy and reduce error rates. Salah et al. [13] introduced an attention-based long short-term memory network

(Att-LSTM) and a stacking ensemble model to capture temporal and knowledge-based feature information, significantly enhancing prediction performance. Fan et al. [14] developed a classification enhancement model based on dual particle swarm optimization, achieving improvements in accuracy, recall, and model stability. Compared with other algorithms, this model has better accuracy, recall rate, F1 value, and better stability. Ali et al. [15] incorporated K-Means clustering into student performance prediction, demonstrating that grouping similar samples can improve prediction accuracy. Sun et al. [16] proposed a multi-layer feature fusion framework integrating LSTM and attention mechanisms to extract temporal and course-level features simultaneously, achieving superior performance compared with baseline methods. Using the flexible combination of LSTM and the attention mechanism, extract the characteristics of the performance sequence information in both the course and time dimensions synchronously, and construct a calculation method for similar students based on time co-occurrence frequency to fuse the information of similar students to achieve information complementarity. Compared with other baseline methods, this model shows higher performance in terms of accuracy, stability, and timeliness.

In addition, researchers also attach great importance to the prediction research of students' grades [17]. Zaffar et al. [18] proposed a clustering-based feature selection method combining artificial fish swarm optimization, cuckoo search optimization, and non-negative matrix factorization clustering, which significantly improved classification accuracy. Mashagba et al. [19] compared several gradient boosting algorithms, including XGBoost, LightGBM, CatBoost, and NGBoost, with traditional models such as decision trees and random forests, showing that gradient boosting models generally achieved higher prediction accuracy and computational efficiency, with CatBoost demonstrating superior performance. Ani and Khor [20] developed multiple machine learning models, including linear regression and logistic regression, to predict student grades and classify performance levels. Pires et al. [21] conducted a systematic review of student performance prediction methods in higher education, highlighting the growing importance of machine learning approaches. Angeioplastis et al. [22] used learning management system data to predict academic performance using multiple machine learning algorithms and found that incorporating strongly correlated course information significantly improves prediction accuracy. Tapio [23] compared multiple linear regression and random forest regression models, demonstrating the superior nonlinear modeling capability of random forests. Johora et al. [24] proposed a stacked ensemble learning framework integrating SMOTE data balancing techniques and interpretable artificial intelligence methods, such as SHAP and LIME, to improve prediction accuracy and model transparency.

3 Methodology

3.1 Clustering Algorithm

3.1.1 K-Means Clustering Algorithm

The K-Means algorithm is a partition-based clustering method and a classical unsupervised learning technique[25, 26]. Due to its computational efficiency, conceptual simplicity, and strong scalability, K-Means has been widely applied in various domains, including system monitoring[27], image segmentation[28], etc. The algorithm partitions data samples into k clusters by minimizing the intra-cluster variance while maximizing inter-cluster separation. The core idea of K-Means is as follows. First, k samples are randomly selected from the dataset as initial cluster centroids C_i $(1 \leq i \leq k)$. Then, each data sample x is assigned to the nearest cluster centroid according to a distance metric, typically the Euclidean distance. After assigning all samples, cluster centroids are updated by calculating the mean of all samples belonging to each cluster. This iterative process continues until the cluster centroids stabilize or the predefined maximum number of iterations is reached[29, 30]. The formula for calculating the Euclidean distance between a data sample x and a clustering center C_k is as follows:

$$d(x, C_k) = \sqrt{\sum_{i=1}^{n} (x_i - C_{ki})^2} \tag{1}$$

where x is the data sample, C_k is the k-th clustering center, n is the data dimension, and x_i, C_{ki} are the i-th attribute values of the data sample x and the clustering center C_k, respectively.

3.1.2 DBSCAN Clustering Algorithm

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm designed to identify clusters with arbitrary shapes and detect noise in complex datasets [31]. Clusters are formed by iteratively connecting density-reachable core points and their neighboring samples. Points that do not belong to any cluster are

treated as noise [32]. The selection of parameter ε is commonly determined using a k-distance graph derived from the k-nearest neighbor method. The optimal ε value is typically located at the inflection point of the curve. If ε is too small, most samples will remain unclustered; conversely, if ε is too large, multiple clusters may merge into a single cluster. In addition, the determination of MinPts follows the guiding principle that MinPts $\geq$ dim+1, where dim is the dimension of the dataset [33, 34].

3.1.3 *Hierarchical Clustering Algorithm*

Hierarchical clustering is a widely used clustering approach due to its interpretability and deterministic clustering structure. It constructs a hierarchical tree (dendrogram) representing nested clustering relationships among samples. Hierarchical clustering methods can be broadly categorized into agglomerative and divisive strategies. Agglomerative hierarchical clustering begins by treating each data sample as an independent cluster and iteratively merges the two closest clusters until all samples are grouped into a single cluster. In contrast, divisive hierarchical clustering starts with all samples in a single cluster and recursively partitions them into smaller clusters. A key characteristic of hierarchical clustering is the use of linkage criteria to measure inter-cluster proximity during the merging or splitting process. Common linkage methods include single linkage, complete linkage, average linkage, and centroid linkage [35]. These linkage strategies influence cluster formation and can affect clustering performance depending on data distribution characteristics.

3.2 Machine Learning Models

This section introduces the machine learning models employed in this study for student performance prediction. Six representative models are selected, including multilayer perceptron (MLP), support vector machine (SVM), random forest (RF), Bootstrap Aggregating (Bagging), Extreme Gradient Boosting (XGBoost), and Category Boosting (CatBoost). These models cover neural network-based learning, kernel-based learning, and ensemble learning paradigms, providing a comprehensive evaluation of clustering-based feature enhancement strategies.

3.2.1 *MLP Model*

The MLP is a feedforward artificial neural network that simulates the information processing mechanism of biological neural systems. MLP is capable of learning complex nonlinear relationships in data and has been widely applied in prediction tasks [36]. A typical MLP consists of three types of layers: an input layer, one or more hidden layers, and an output layer. Neurons between adjacent layers are fully connected, allowing the network to model complex feature interactions. During forward propagation, each neuron receives weighted inputs from the previous layer and processes them using an activation function. The network parameters are optimized using backpropagation and gradient-based optimization algorithms to minimize prediction error.

3.2.2 *SVM Model*

The SVM is a supervised learning algorithm widely used for classification and regression tasks [37]. When applied to regression problems, SVM is referred to as Support Vector Regression (SVR). The objective of SVR is to determine a regression function that approximates training data within a specified error tolerance while maintaining model simplicity. The SVR optimization objective is defined as:

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2}$$

subject to

$$y_i - (w \cdot x_i + b) \leq \varepsilon + \xi_i, \tag{3}$$

$$(w \cdot x_i + b) - y_i \leq \varepsilon + \xi_i^*, \tag{4}$$

$$\xi_i, \xi_i^* \geq 0, \quad \forall i \tag{5}$$

where w is the weight vector, b is the bias, x_i is the feature vector of the i-th sample, y_i is the target value of the i-th sample, C is the regularization parameter that controls the penalty degree of the error term, ξ_i and ξ_i^* are slack variables that allow some data points to violate the prediction of the model within the ε range, and ε is the tolerance that defines the maximum deviation allowed between the predicted value and the actual value.

3.2.3 *RF Model*

The RF is an ensemble learning method that constructs multiple decision trees and aggregates their prediction results to improve model performance [38]. Each decision tree is trained using bootstrap sampling and random feature selection, which enhances model diversity and reduces overfitting. For regression tasks, RF determines optimal splitting rules by minimizing the mean squared error (MSE) within each tree

node. The final prediction is obtained by averaging the outputs of all decision trees, which improves prediction stability and generalization capability.

3.2.4 Bagging Model

The Bagging is an ensemble learning technique designed to improve prediction accuracy and model robustness. Bagging generates multiple training subsets through random sampling with replacement from the original dataset. Independent base learners are trained on these subsets, and their outputs are aggregated to produce the final prediction. The primary advantage of Bagging is its ability to reduce model variance and improve stability. Model performance can be further optimized by adjusting parameters such as the number of base learners, sampling size, and base model configuration.

3.2.5 XGBoost Model

The XGBoost is an advanced ensemble learning algorithm based on Gradient Boosting Decision Trees (GBDT). Compared with traditional GBDT, XGBoost improves computational efficiency through parallel processing and incorporates regularization mechanisms to reduce overfitting [39]. XGBoost optimizes model performance by expanding the objective function using second-order Taylor approximation, which improves gradient convergence speed and enhances prediction accuracy. Due to its efficiency and strong generalization capability, XGBoost has demonstrated excellent performance in various regression and classification tasks.

3.2.6 CatBoost Model

The CatBoost is a gradient boosting algorithm specifically designed to handle categorical features efficiently [40]. CatBoost employs ordered boosting techniques and specialized encoding methods to reduce prediction bias and prevent overfitting. Additionally, CatBoost utilizes symmetric tree structures to improve training stability and computational efficiency. The algorithm supports multi-threaded and distributed training, making it suitable for large-scale datasets. Due to its strong capability in handling categorical and heterogeneous data, CatBoost has been widely applied in regression, classification, and recommendation systems.

3.3 The proposed method

This study proposes a clustering-driven feature generation framework to enhance student performance prediction. The core idea is to extract latent structural information from student data using clustering algorithms and integrate the obtained clustering labels with original features to improve prediction performance. The overall workflow of the proposed method consists of four main stages:

1. Data Preprocessing. The original datasets are first processed through data shuffling, categorical feature encoding, and dataset partitioning. These steps ensure data quality and improve model training stability.

2. Clustering-Based Feature Generation. Clustering algorithms are applied to the original feature space to group similar samples. Each sample is assigned a clustering label representing its cluster membership. The clustering label is then appended as a new feature to the original dataset.

3. Model Training and Parameter Optimization. Multiple machine learning models are trained using datasets with and without clustering-generated features. Hyperparameters are optimized using grid search and early stopping strategies to improve generalization performance.

4. Performance Evaluation and Comparative Analysis. The predictive performance of each model is evaluated using multiple regression metrics. Comparative experiments are conducted to analyze the effectiveness of clustering-based feature enhancement across different models and datasets.

To determine the most suitable clustering method, three clustering algorithms, including K-Means, DBSCAN, and hierarchical clustering, are initially evaluated using the MLP model as a baseline predictor. Experimental results indicate that K-Means consistently achieves superior prediction performance. Therefore, K-Means is selected as the primary clustering method for subsequent experiments.

4 Experiments

4.1 Evaluation Metrics

To evaluate model prediction accuracy and stability, three widely used regression evaluation metrics are adopted: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). MAE measures the average absolute difference between predicted and actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

where n is the number of samples, y_i is the actual observed value, and $\hat{y}_i$ is the predicted value of the model.

RMSE measures the square root of the average squared prediction error:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

The coefficient of determination evaluates the goodness of fit of the regression model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

4.2 Experimental Data

Two publicly available educational datasets are used to evaluate the proposed framework. Dataset 1 is obtained from the Roycekimmons dataset repository. It contains performance records of 1,000 students, including mathematics, reading, and writing scores. The dataset also includes five categorical attributes describing student demographic and educational background information, including gender, race/ethnicity, parental education level, lunch type, and test preparation course participation. More detailed information can be found from [1].

Dataset 2 is collected from Portuguese secondary school students and is available from the UCI Machine Learning Repository. It contains detailed student information, including demographic characteristics, family background, learning behavior, and academic history. The dataset includes three performance indicators: G1, G2, and G3, representing first-period, second-period, and final grades, respectively. Strong correlations among these variables make the dataset suitable for sequential performance prediction analysis. Compared with Dataset 1, Dataset 2 contains richer behavioral and contextual information and represents a complex multivariate educational dataset. More detailed information can be found from [2].

[1] http://roycekimmons.com/tools/generated_data/exams

[2] https://archive.ics.uci.edu/dataset/320/student+performance

4.3 Parameter Optimization

For the MLP, the key hyperparameters include the learning rate and hidden layer structure. The search space includes learning rates [0.001, 0.01, and 0.1] and hidden layer configurations (50,), (100,), (100, 100), (100, 50), and (100, 50, 50). For the SVM, kernel options include linear, polynomial, and radial basis function (RBF). The regularization parameter C is selected from [0.05, 0.1, 1, 10, 30, 50] are specifically defined for the SVM model. By using the grid search method to find the optimal combination of parameters, the performance of the SVM model on a specific dataset can be significantly improved[41]. For the RF, the number of trees and tree depth are tuned. The number of trees is selected from [50, 100, 200, 400], and maximum tree depth is selected from [None, 10, 20]. For the Bagging, the number of base learners is optimized using candidate values [50, 100, 200, 400, 900]. For the XGBoost, the learning rate [0.001, 0.01, and 0.1] and number of estimators [50, 100, 150, and 200] are optimized. The CatBoost uses the same search space as XGBoost, including learning rate and number of estimators. The optimal parameters for each model are obtained using the grid search method [42].

4.4 Clustering Algorithm Selection

Table 1. Prediction Results of math score of Dataset 1 using different clustering algorithms.

Method	MAE	RMSE	R^2
K-Means	**10.7106**	**13.2205**	**0.2173**
DBSCAN	10.8245	13.3815	0.1953
Hierarchical	10.7814	13.2979	0.2083
Without Clustering	10.7429	13.2836	0.2102

Clustering-based feature generation is implemented by assigning cluster labels to each sample using clustering algorithms. These labels serve as additional features reflecting similarity structures in the data. To determine the most effective clustering approach, K-Means, DBSCAN, and hierarchical clustering are compared using the MLP model. The clustering hyperparameters are set as follows: the number of clusters for K-Means is set to 3; the ε is set to 0.5 and MinPts is set to 7 for DBSCAN; the number of clusters for hierarchical clustering is set to 3. Experimental results for mathematics score prediction on Dataset 1 are summarized in Table 1 (The result with best performance is highlighted in bold). As shown in Table 1, the K-Means clustering algorithm has better performance than the other two clustering algorithms. Therefore, this study selects the K-Means clustering algorithm for subsequent research.

4.5 Experimental Results and Analysis of Dataset 1

Based on the clustering selection results in Section 4.4, K-Means was adopted for clustering-based feature generation. For each model, hyperparameters were optimized using grid search combined with cross-validation. Experiments were conducted for mathematics, reading, and writing score prediction.

Table 2. Prediction Results of math score of Dataset 1 on the Full Feature Set.

Method	MAE	RMSE	R^2
MLP	10.7429	13.2836	0.2102
MLP + K-Means	**10.7106**	**13.2205**	**0.2173**
SVM	**10.8037**	**13.3152**	**0.2147**
SVM+ K-Means	10.8255	13.3262	0.2029
RF	12.0245	14.9112	-0.0028
RF + K-Means	**11.9847**	**14.8729**	**0.0022**
Bagging	12.1587	15.0759	-0.0253
Bagging+ K-Means	**12.1403**	**15.0699**	**-0.0244**
XGBoost	11.3227	14.0304	0.1191
XGBoost+ K-Means	**11.3259**	**14.0215**	**0.1198**
CatBoost	**11.0211**	**13.6181**	**0.1703**
CatBoost + K-Means	11.0570	13.6310	0.1681

Table 2 presents the prediction results for mathematics scores using the full feature set. The introduction of K-Means clustering improves the performance of MLP, RF, Bagging, and XGBoost models. Among these models, MLP achieves the best overall performance, with MAE and RMSE reduced to 10.7106 and 13.2205, respectively, and R^2 increased to 0.2173. However, SVM and CatBoost experience slight performance degradation after incorporating clustering features. This indicates that clustering-based feature enhancement does not universally benefit all prediction models and may depend on the interaction between model structure and data distribution.

The experimental results for reading score prediction are shown in Table 3. After introducing K-Means clustering, the MLP, RF, and Bagging models demonstrate improved prediction accuracy. MLP again achieves the best performance, with reduced MAE and RMSE and improved R^2 values. Conversely, SVM, XGBoost, and CatBoost show marginal decreases in performance after incorporating clustering labels. These results suggest that clustering-based feature augmentation may provide limited benefits for models that already capture feature interactions effectively.

Table 4 summarizes the writing score prediction results. After introducing K-Means clustering, the MLP, SVM, RF, and CatBoost models achieve improved prediction performance. MLP demonstrates the most significant improvement, with MAE reduced to 10.0951 and R^2 increased to 0.3259. In contrast, Bagging and XGBoost show slight performance reductions, further confirming that clustering feature effectiveness varies across models and prediction tasks.

Table 3. Prediction results of the reading score of Dataset 1 on the full set of features.

Method	MAE	RMSE	R^2
MLP	10.5799	12.9232	0.2003
MLP + K-Means	**10.5195**	**12.8620**	**0.2092**
SVM	**10.5988**	**12.9758**	**0.1933**
SVM+ K-Means	10.6069	12.9789	0.1930
RF	11.7204	14.3595	0.0159
RF + K-Means	**11.7057**	**14.3390**	**0.0188**
Bagging	11.8429	14.5025	-0.0038
Bagging+ K-Means	**11.8361**	**14.4984**	**-0.0031**
XGBoost	**11.0322**	**13.5668**	**0.1236**
XGBoost+ K-Means	11.0339	13.5739	0.1226
CatBoost	**10.7551**	**13.2262**	**0.1652**
CatBoost + K-Means	10.7803	13.2316	0.1643

Table 4. Prediction Results of the writing score of Dataset 1 on the full set of features.

Method	MAE	RMSE	R^2
MLP	10.2117	12.5327	0.3145
MLP + K-Means	**10.0951**	**12.4272**	**0.3259**
SVM	10.2587	12.6065	0.3063
SVM+ K-Means	**10.2096**	**12.5273**	**0.3140**
RF	11.3006	13.9697	0.1502
RF + K-Means	**11.2955**	**13.9613**	**0.1513**
Bagging	**11.4173**	**14.1083**	**0.1330**
Bagging+ K-Means	11.4193	14.1099	0.1329
XGBoost	**10.7418**	**13.2095**	**0.2414**
XGBoost+ K-Means	10.7664	13.2039	0.2388
CatBoost	10.4834	12.8972	0.2748
CatBoost + K-Means	**10.4191**	**12.8595**	**0.2796**

Figures 1, 2, 3, 4, 5 and 6 illustrate the performance comparison of six models before and after clustering-based feature enhancement. Overall, the MLP and RF benefit most consistently from K-Means clustering across the three subjects. Other models demonstrate partial improvements depending on the prediction task. The improvements achieved by clustering can be attributed to its ability to reveal latent group structures among students. By introducing the clustering feature as an additional feature, the models can better capture student similarity patterns and improve prediction accuracy. Nevertheless, clustering does not guarantee performance improvement in all

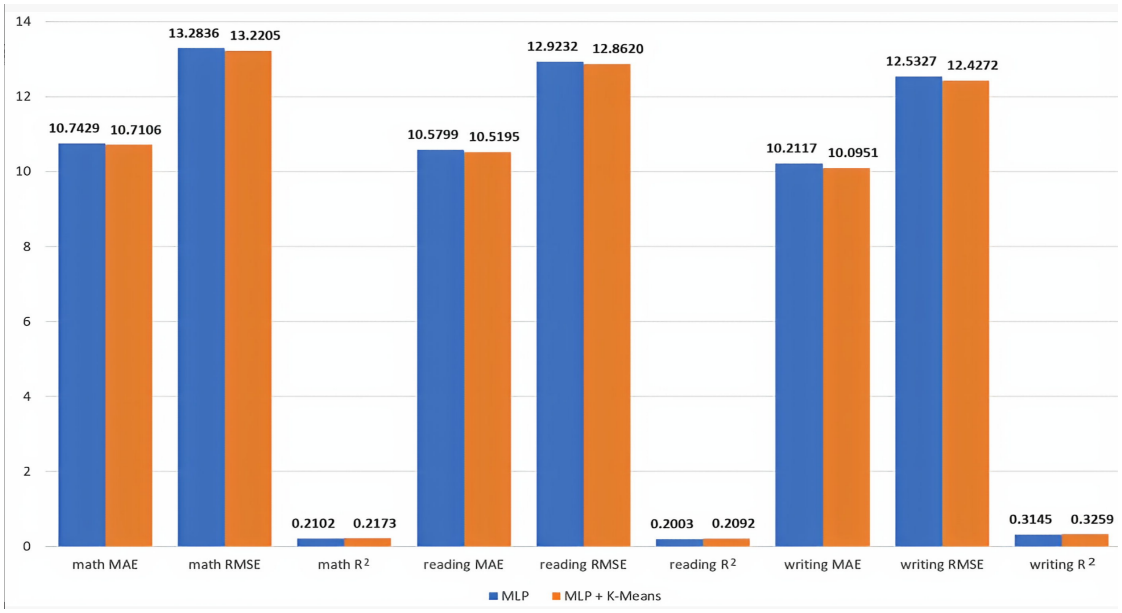

Figure 1. Performance comparison of the MLP model on Dataset 1 before and after optimization.

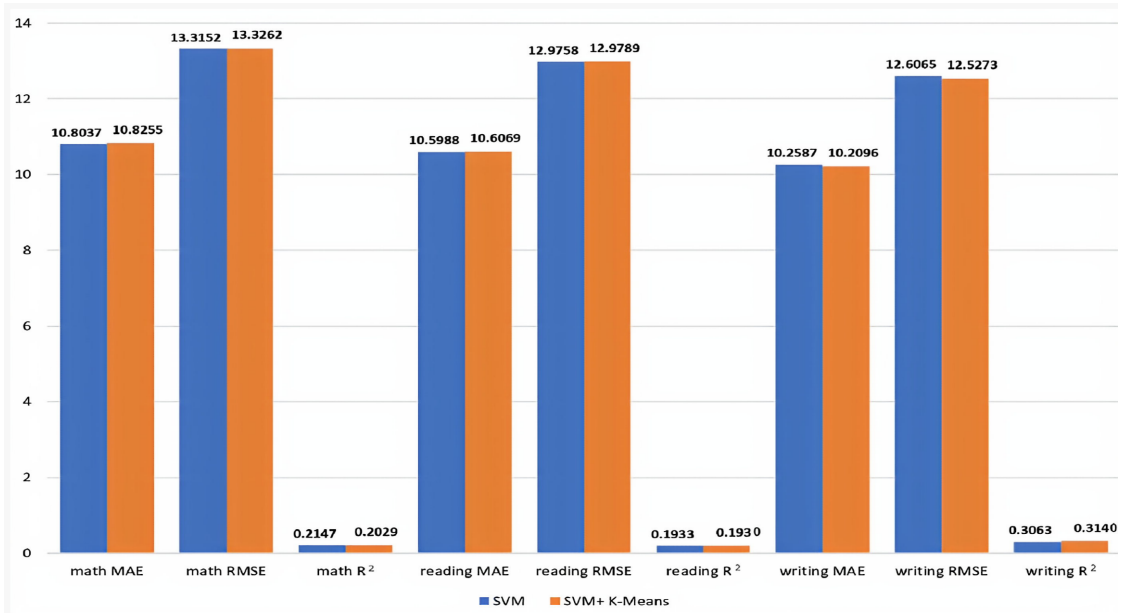

Figure 2. Performance comparison of the SVM model on Dataset 1 before and after optimization.

cases.

4.6 Experimental Results and Analysis of Dataset 2

Dataset 2 includes sequential academic performance indicators (G1, G2, and G3), allowing further evaluation of clustering-based feature enhancement under more complex data conditions. Table 5 presents prediction results for G1 using the full feature set excluding G1 and G2. After introducing K-Means clustering, all six models demonstrate performance improvement. In particular, the Bagging and RF achieve the best performance, indicating that clustering effectively captures latent student learning patterns at early performance stages.

Table 6 presents prediction results for G2 using features including G1 but excluding G2. The results show that the SVM, Bagging, and XGBoost benefit from clustering-based feature enhancement. However, the MLP, RF, and CatBoost experience slight performance decreases after incorporating clustering features. This suggests that when strong predictive indicators such as G1 are available, some models can already capture data patterns effectively, reducing the additional benefit provided by clustering.

Table 7 summarizes prediction results for G3 using features including G1 and G2. Clustering improves the performance of SVM, Bagging, XGBoost, and

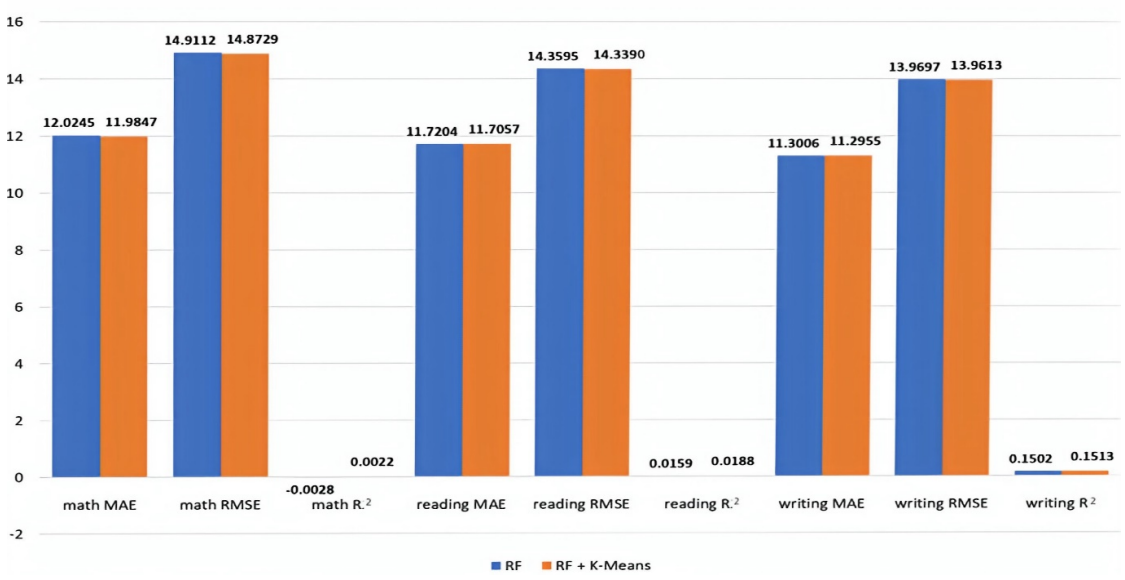

Figure 3. Performance comparison of the RF model on Dataset 1 before and after optimization.

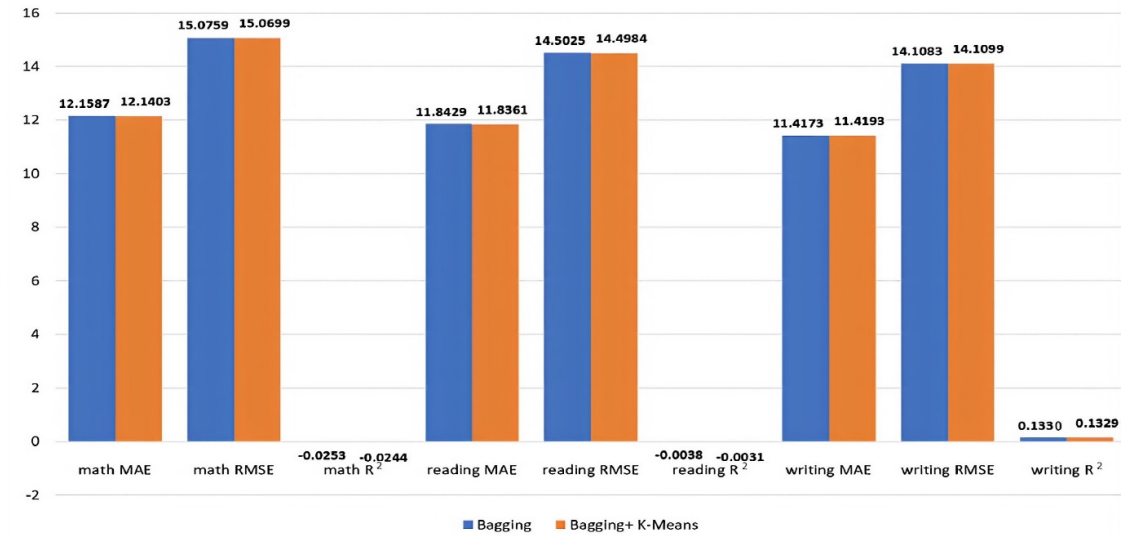

Figure 4. Performance comparison of the Bagging model before and after optimization on Dataset 1.

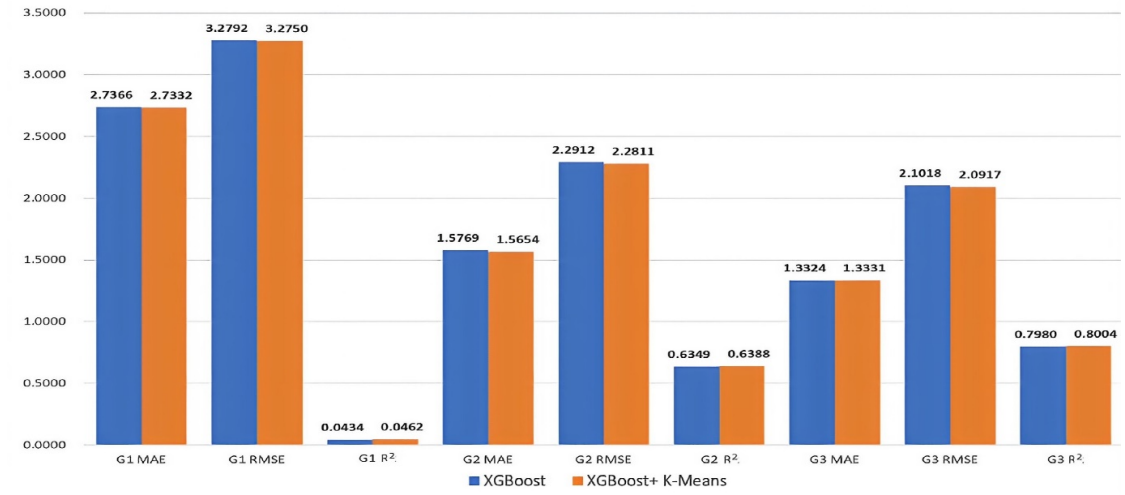

Figure 5. Performance comparison of the XGBoost model before and after optimization on Dataset 1.

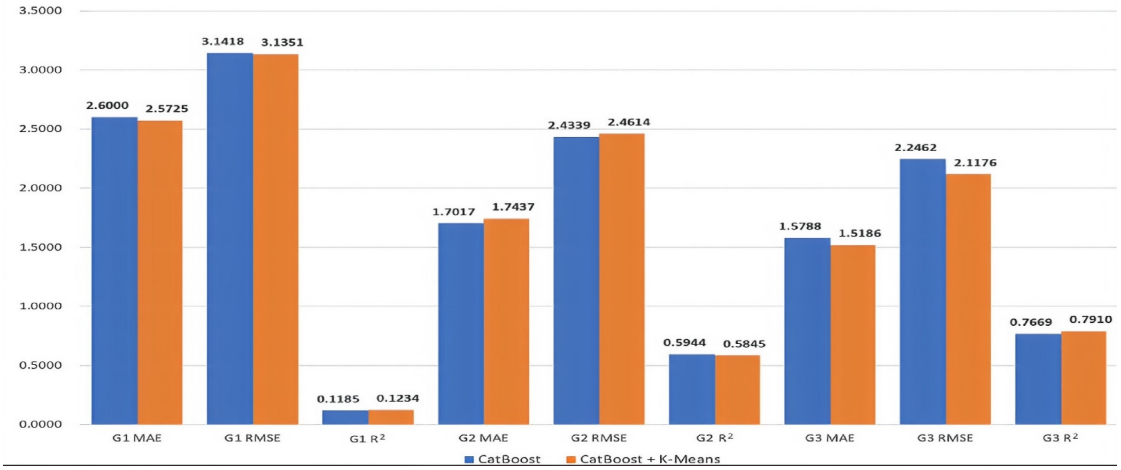

Figure 6. Performance comparison of the CatBoost model before and after optimization on Dataset 1.

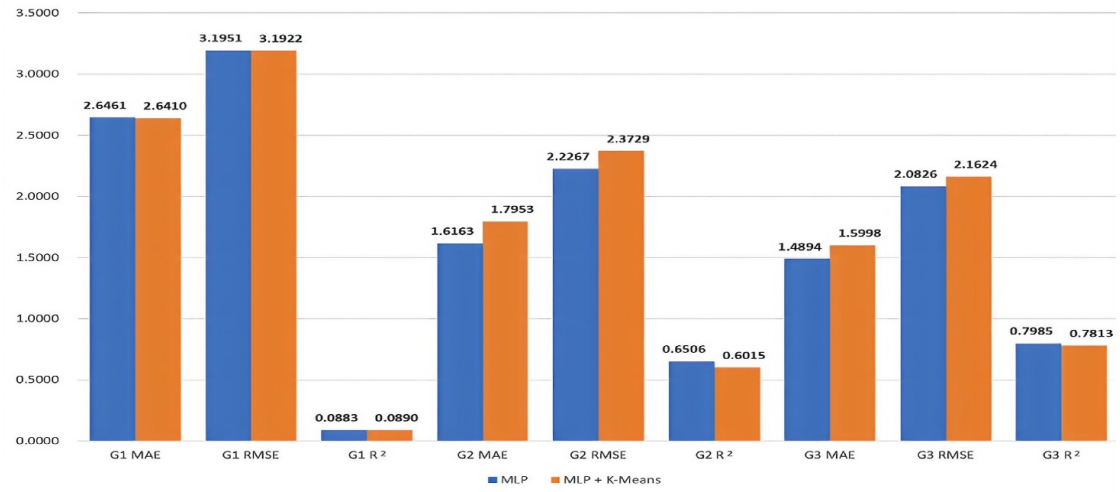

Figure 7. Performance comparison of MLP model before and after optimization on Dataset 2.

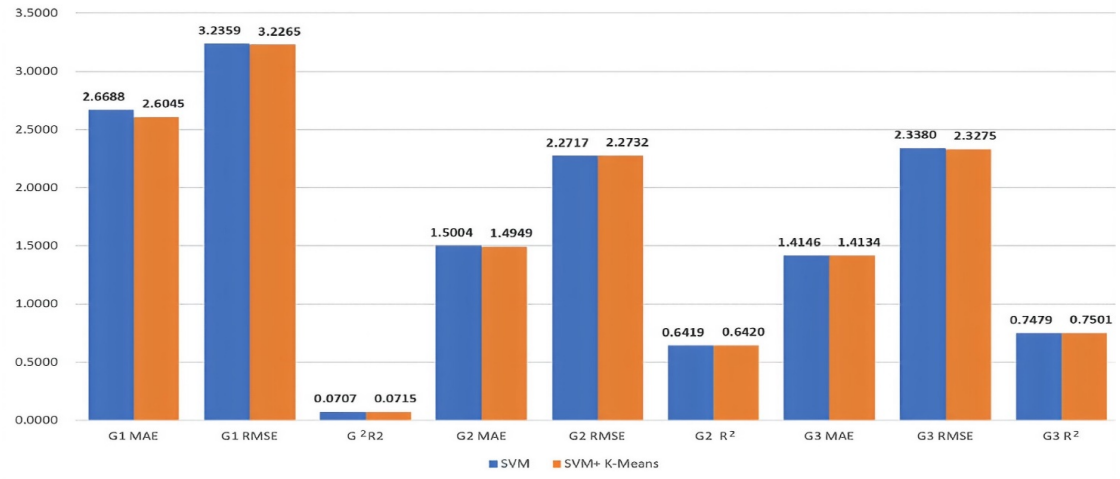

Figure 8. Performance comparison of SVM model before and after optimization on Dataset 2.

CatBoost models. However, the MLP and RF show slight performance degradation, indicating that their inherent nonlinear modeling capability may already sufficiently capture feature interactions.

Figures 7, 8, 9, 10, 11 and 12 illustrate performance comparisons across six models before and after clustering enhancement. Overall, the SVM, Bagging, and XGBoost consistently benefit from clustering

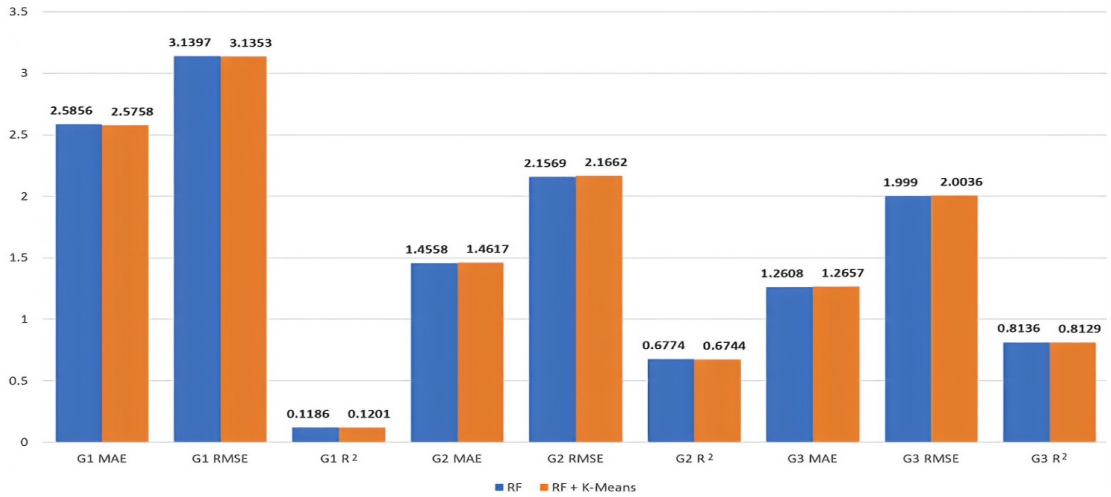

Figure 9. Performance comparison of RF model before and after optimization on Dataset 2.

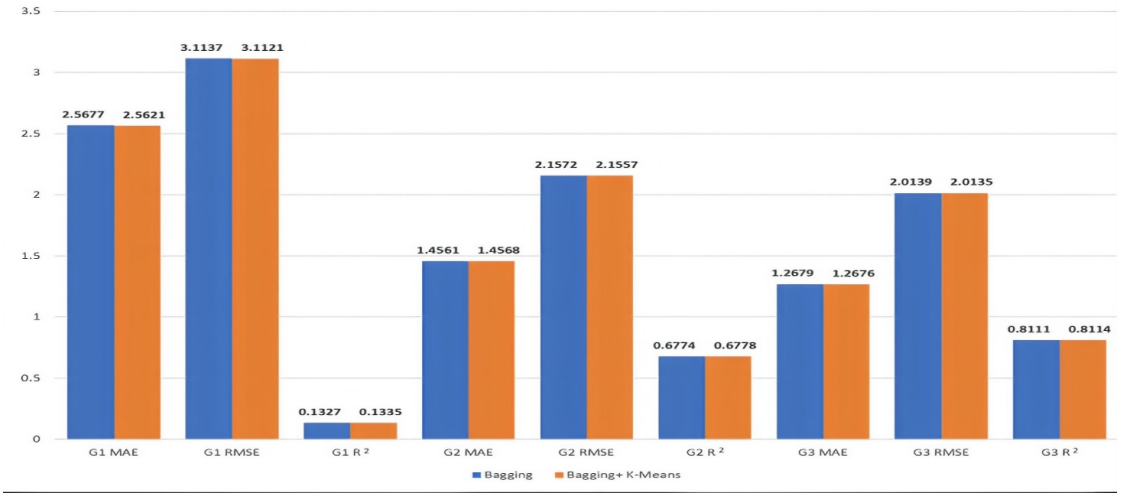

Figure 10. Performance comparison of the Bagging model before and after optimization on Dataset 2.

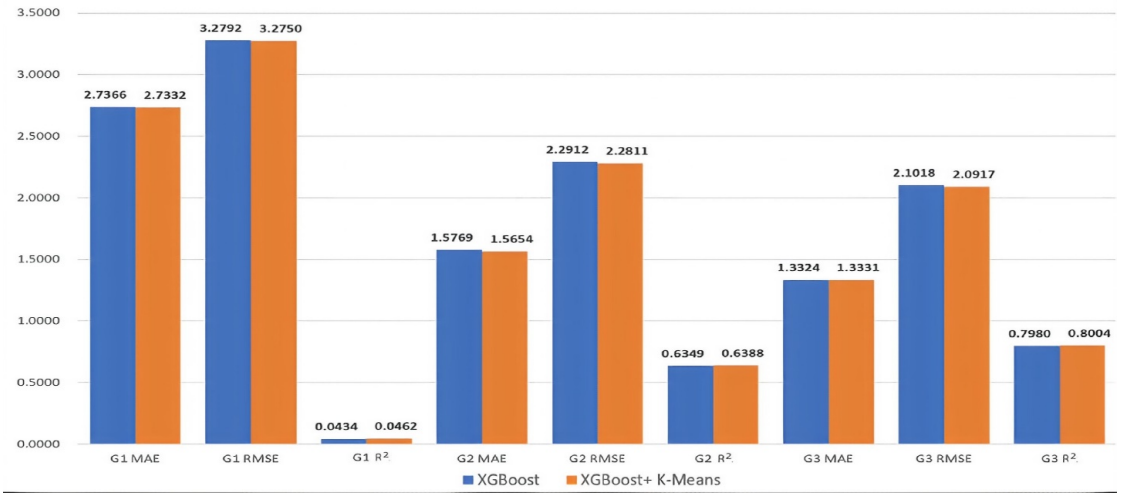

Figure 11. Performance comparison of the XGBoost model before and after optimization on Dataset 2.

features across the three prediction tasks in dataset 2. The MLP, RF, and CatBoost demonstrate improvements in selected tasks but show performance declines in others. The results confirm that clustering-based feature generation can effectively enhance model sensitivity to data distribution patterns.

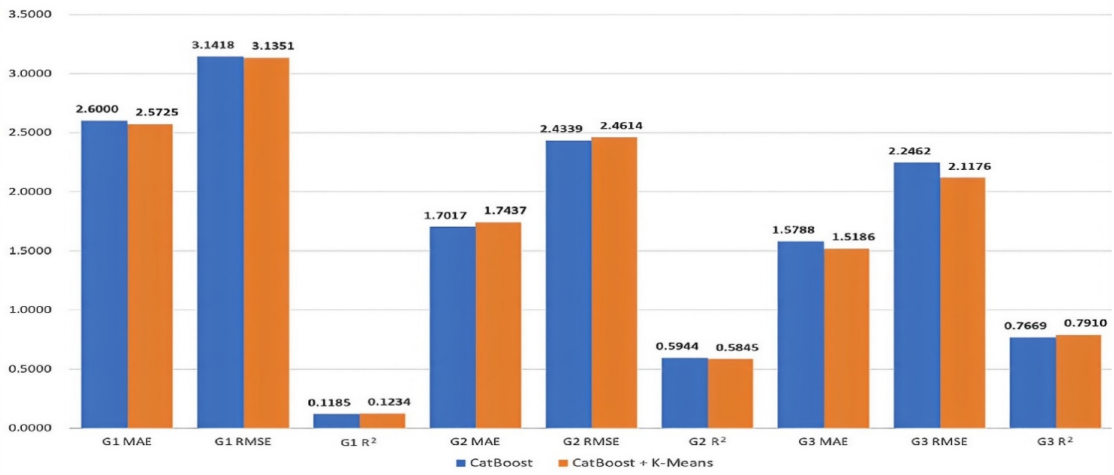

Figure 12. Performance comparison of the CatBoost model before and after optimization on Dataset 2.

Table 5. Prediction Results of G1 in Dataset 2 on the Full Feature Set (excluding G1 and G2).

Method	MAE	RMSE	R^2
MLP	2.6461	3.1951	0.0883
MLP + K-Means	**2.6410**	**3.1922**	**0.0890**
SVM	2.6688	3.2359	0.0707
SVM+ K-Means	**2.6045**	**3.2265**	**0.0715**
RF	2.5856	3.1397	0.1186
RF + K-Means	**2.5758**	**3.1353**	**0.1201**
Bagging	2.5677	3.1137	0.1327
Bagging+ K-Means	**2.5621**	**3.1121**	**0.1335**
XGBoost	2.7366	3.2792	0.0434
XGBoost+ K-Means	**2.7332**	**3.2750**	**0.0462**
CatBoost	2.6000	3.1418	0.1185
CatBoost + K-Means	**2.5725**	**3.1351**	**0.1234**

Table 6. Prediction Results of G2 in Dataset 2 on the Full Feature Set (including G1 but excluding G2).

Method	MAE	RMSE	R^2
MLP	**1.6163**	**2.2267**	**0.6506**
MLP + K-Means	1.7953	2.3729	0.6015
SVM	1.5004	2.2717	0.6419
SVM+ K-Means	**1.4949**	**2.2732**	**0.6420**
RF	**1.4558**	**2.1569**	**0.6774**
RF + K-Means	1.4617	2.1662	0.6744
Bagging	1.4561	2.1572	0.6774
Bagging+ K-Means	**1.4568**	**2.1557**	**0.6778**
XGBoost	1.5769	2.2912	0.6349
XGBoost+ K-Means	**1.5654**	**2.2811**	**0.6388**
CatBoost	**1.7017**	**2.4339**	**0.5944**
CatBoost + K-Means	1.7437	2.4614	0.5845

By revealing latent student group characteristics, clustering improves model generalization ability, particularly in early-stage performance prediction.

Table 7. Prediction results of G3 in dataset 2 on the full feature set (including G1 and G2).

Method	MAE	RMSE	R^2
MLP	**1.4894**	**2.0826**	**0.7985**
MLP + K-Means	1.5998	2.1624	0.7813
SVM	1.4146	2.3380	0.7479
SVM+ K-Means	**1.4134**	**2.3275**	**0.7501**
RF	**1.2608**	**1.9990**	**0.8136**
RF + K-Means	1.2657	2.0036	0.8129
Bagging	1.2679	2.0139	0.8111
Bagging+ K-Means	**1.2676**	**2.0135**	**0.8114**
XGBoost	1.3324	2.1018	0.7980
XGBoost+ K-Means	**1.3331**	**2.0917**	**0.8004**
CatBoost	1.5788	2.2462	0.7669
CatBoost + K-Means	**1.5186**	**2.1176**	**0.7910**

However, similar to Dataset 1, the effectiveness of clustering exhibits strong task and model dependency. Models with strong nonlinear modeling capabilities or datasets containing highly informative sequential features may obtain limited benefits from clustering enhancement.

5 Conclusion

This study proposed a clustering-based feature generation model to improve student performance prediction. After data preprocessing, three clustering algorithms (K-Means, DBSCAN, and hierarchical clustering) were employed to generate additional feature representations. Based on comparative experiments using the MLP model, K-Means was selected and further integrated into six machine learning models, including MLP, SVM, RF, Bagging, XGBoost, and CatBoost, and evaluated on two educational datasets. Experimental results showed

that incorporating K-Means-generated features improved the performance of most models, as evidenced by reduced MAE and RMSE and increased R^2 values. These findings demonstrate that clustering-derived structural information can effectively enhance prediction accuracy. However, the improvement is influenced by data characteristics and model types, indicating that clustering-based feature augmentation should be applied selectively.

In future work, more advanced clustering techniques, such as density-adaptive clustering or fuzzy clustering, can be investigated to capture more complex data structures and improve feature representation quality. Second, dynamic or adaptive clustering strategies can be developed to automatically determine optimal cluster numbers and structures, reducing dependence on manual parameter tuning.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable. This study uses only publicly available, anonymized datasets from UCI and generated sources, which do not involve human subjects or identifiable information; therefore, ethical approval is not applicable.

References

[1] Gonugunta, K. C., & Leo, K. (2024). Role of data-driven decision making in enhancing higher education performance: A comprehensive analysis of analytics in institutional management. *International Journal of Acta Informatica, 3*(1), 149-159.

[2] Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies, 28*(1), 905-971. [CrossRef]

[3] Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development, 60*(3), 383-398. [CrossRef]

[4] Shen, Y. (2024, May). Using long short-term memory networks (LSTM) to predict student academic achievement: dynamic learning path adjustment. In *Proceedings of the 2024 International Conference on Machine Intelligence and Digital Applications* (pp. 627-634). [CrossRef]

[5] Li, M. (2018). A study on the influence of non-intelligence factors on college students' English learning achievement based on C4. 5 algorithm of decision tree. *Wireless personal communications, 102*(2), 1213-1222. [CrossRef]

[6] Abdrakhmanov, R., Zhaxanova, A., Karatayeva, M., Niyazova, G. Z., Berkimbayev, K., & Tuimebayev, A. (2024). Development of a Framework for Predicting Students' Academic Performance in STEM Education using Machine Learning Methods. *International Journal of Advanced Computer Science & Applications, 15*(1). [CrossRef]

[7] Kinash, S., Naidu, V., Knight, D., Judd, M. M., Nair, C. S., Booth, S., ... & Tulloch, M. (2015). Student feedback: a learning and teaching performance indicator. *Quality Assurance in Education, 23*(4), 410-428. [CrossRef]

[8] Liu, Y., Hui, Y., Hou, D., & Liu, X. (2023). A novel student achievement prediction method based on deep learning and attention mechanism. *IEEE Access, 11*, 87245-87255. [CrossRef]

[9] Pandey, M., & Sharma, V. K. (2013). A decision tree algorithm pertaining to the student performance analysis and prediction. *International Journal of Computer Applications, 61*(13), 1-5. [CrossRef]

[10] Xu, Z., Yuan, H., & Liu, Q. (2020). Student performance prediction based on blended learning. *IEEE Transactions on Education, 64*(1), 66-73. [CrossRef]

[11] Yang, S. J., Lu, O. H., Huang, A. Y., Huang, J. C., Ogata, H., & Lin, A. J. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing, 26*, 170-176. [CrossRef]

[12] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior, 104*, 106189. [CrossRef]

[13] Salah Hashim, A., Akeel Awadh, W., & Khalaf Hamoud, A. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 928, No. 3, p. 032019). IOP Publishing. [CrossRef]

[14] Fan, Z., Gou, J., & Wang, C. (2023). Predicting secondary school student performance using a

double particle swarm optimization-based categorical boosting model. *Engineering Applications of Artificial Intelligence, 124*, 106649. [CrossRef]

[15] Ali, Z. M., Hassoon, N. H., Ahmed, W. S., & Abed, H. N. (2020). The application of data mining for predicting academic performance using k-means clustering and naïve bayes classification. *International Journal of Psychosocial Rehabilitation, 24*(03), 2143-2151. [CrossRef]

[16] Sun, D., Luo, R., Guo, Q., Xie, J., Liu, H., Lyu, S., ... & Song, S. (2023). A university student performance prediction model and experiment based on multi-feature fusion and attention mechanism. *IEEE Access, 11*, 112307-112319. [CrossRef]

[17] Nachouki, M., & Abou Naaj, M. (2022). Predicting student performance to improve academic advising using the random forest algorithm. *International Journal of Distance Education Technologies (IJDET), 20*(1), 1-17. [CrossRef]

[18] Zaffar, M., Hashmani, M. A., Savita, K. S., & Rizvi, S. S. H. (2018). A study of feature selection algorithms for predicting students academic performance. International *Journal of Advanced Computer Science and Applications, 9*(5).

[19] Mashagba, E., Al-Saqqar, F., & Al-Shatnawi, A. (2023, March). Using gradient boosting algorithms in predicting student academic performance. In *2023 International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-7). IEEE. [CrossRef]

[20] Ani, A., & Khor, E. T. (2024). Development and evaluation of predictive models for predicting students performance in MOOCs. *Education and Information Technologies, 29*(11), 13905-13928. [CrossRef]

[21] Pires, J. P., Brito Correia, F., Gomes, A., Borges, A. R., & Bernardino, J. (2024). Predicting student performance in introductory programming courses. *Computers, 13*(9), 219. [CrossRef]

[22] Angeioplastis, A., Aliprantis, J., Konstantakis, M., & Tsimpiris, A. (2025). Predicting student performance and enhancing learning outcomes: a data-driven approach using educational data mining techniques. *Computers, 14*(3), 83. [CrossRef]

[23] Tapio, R. (2025). Comparative analysis of multiple linear regression and random forest regression in predicting academic performance of students in higher education. *Asian Research Journal of Mathematics, 21*(4), 170-181. [CrossRef]

[24] Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M., & Akter, F. (2025). An explainable AI-based approach for predicting undergraduate students academic performance. *Array, 26*, 100384. [CrossRef]

[25] Khotimah, B. K., Irhamni, F. I. R. L. I., & Sundarwati, T. R. I. (2016). A Genetic algorithm for optimized initial centers K-means clustering in SMEs. *Journal of Theoretical and Applied Information Technology, 90*(1), 23.

[26] Feng, Y., Zou, J., Liu, W., & Lv, F. (2024). Distributed K-Means algorithm based on a Spark optimization sample. *PLoS One, 19*(12), e0308993. [CrossRef]

[27] Miraftabzadeh, S. M., Colombo, C. G., Longo, M., & Foiadelli, F. (2023). K-means and alternative clustering methods in modern power systems. *IEEE Access, 11*, 119596-119633. [CrossRef]

[28] Yang, S., Li, P., Wen, H., Xie, Y., & He, Z. (2018). K-hyperline clustering-based color image segmentation robust to illumination changes. *Symmetry, 10*(11), 610. [CrossRef]

[29] Ashabi, A., Sahibuddin, S. B., & Salkhordeh Haghighi, M. (2020, December). The systematic review of K-means clustering algorithm. In *Proceedings of the 2020 9th international conference on networks, communication and computing* (pp. 13-18). [CrossRef]

[30] Tan, L. (2015, April). A clustering K-means algorithm based on improved PSO algorithm. In *2015 Fifth International Conference on Communication Systems and Network Technologies* (pp. 940-944). IEEE. [CrossRef]

[31] Li, H., Wang, J., Ren, Y., & Mao, F. (2021). Intercity online car-hailing travel demand prediction via a Spatiotemporal Transformer Method. *Applied Sciences, 11*(24), 11750. [CrossRef]

[32] Zhang, X., Lauber, L., Liu, H., Shi, J., Wu, J., & Pan, Y. (2021). Research on the method of travel area clustering of urban public transport based on Sage-Husa adaptive filter and improved DBSCAN algorithm. *PLoS one, 16*(12), e0259472. [CrossRef]

[33] Song, J., Guo, Y., & Wang, B. (2019). Research on parameter configuration method of DBSCAN clustering algorithm. *Comput. Technol. Dev, 29*(5), 44-48.

[34] Ma, B., Yang, C., Li, A., Chi, Y., & Chen, L. (2023). A faster DBSCAN algorithm based on self-adaptive determination of parameters. *Procedia Computer Science, 221*, 113-120. [CrossRef]

[35] Karypis, G., Han, E., & Kumar, V. (1999). *A hierarchical clustering algorithm using dynamic modeling* (Technical Report No. 99-007). University of Minnesota Digital Conservancy. Available at:https://hdl.handle.net/11299/215363

[36] Ranjeeth, S., Latchoumi, T. P., & Paul, P. V. (2021). Optimal stochastic gradient descent with multilayer perceptron based student's academic performance prediction model. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 14*(6), 1728-1741. [CrossRef]

[37] Burman, I., & Som, S. (2019, February). Predicting students academic performance using support vector machine. In *2019 Amity international conference on artificial intelligence (AICAI)* (pp. 756-759). IEEE. [CrossRef]

[38] Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020, March). Predicting students academic performance using an improved random forest classifier. In *2020 international conference on emerging smart computing and informatics (ESCI)* (pp. 238-243). IEEE. [CrossRef]

[39] Duan, D., Dai, C., & Tu, R. (2021, December). Research on the Prediction of Students' Academic Performance Based on XGBoost. In *2021 Tenth International Conference of Educational Innovation through Technology (EITT)* (pp. 316-319). IEEE. [CrossRef]

[40] Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D., & Khanna, A. (2021). CatBoost—An ensemble machine learning model for prediction and classification of student academic performance. *Advances in Data Science and Adaptive Analysis, 13*(03n04), 2141002. [CrossRef]

[41] Kee, T., & Ho, W. K. (2025). Optimizing machine learning models for urban sciences: a comparative analysis of hyperparameter tuning methods. *Urban Science, 9*(9), 348. [CrossRef]

[42] Zhang, W., Cheng, S., & Lu, F. (2026). A geographic evolutionary framework with multi-task optimization of automatic hyperparameter tuning for spatially stratified machine learning models. *International Journal of Geographical Information Science, 40*(1), 25-48. [CrossRef]