e-ISSN: 2655-867X
p-ISSN: 2655-8661
Volume 8 No. 2 August 2026

An Explainable CatBoost Framework Optimized with Optuna for Predicting Students' Mathematics Proficiency

Andri Feriansyah[1)], Ahmad Labudi[1)], Yayu Nur Faidah[2)], Agung Rachmat Raharja[1)]
[1)] Informatics, Bandung University
[2)] Health Information Management, Bandung University
Email: andriferiyansyah@universitasbandung.ac.id; ahmadlabudi@universitasbandung.ac.id; yayu.nrfdh@universitasbandung.ac.id; agungrachmat@universitasbandung.ac.id.

Accepted:	Accepted After Revision:	Published:
15 July 2026	14 August 2026	31 August 2026

Abstract

Predicting students' mathematics proficiency is an important task in educational data mining because it supports educational assessment and data-informed decision making. This study proposes an explainable machine learning framework by integrating CatBoost, Optuna, and SHAP to classify students into different levels of mathematics proficiency. The study uses the Student Performance in Mathematics dataset from Kaggle, which contains 1,000 student records with demographic, socioeconomic, and academic characteristics. Mathematics scores were grouped into three proficiency categories (Low, Medium, and High), transforming the problem into a multiclass classification task. The dataset was divided into training and testing subsets using an 80:20 stratified split. Hyperparameter optimization was performed on the training data using Optuna with 30 optimization trials and five-fold stratified cross-validation, while the testing dataset was reserved exclusively for the final evaluation. The optimized CatBoost model was compared with XGBoost and LightGBM under identical experimental settings. Performance was evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix. The optimized CatBoost model achieved the highest observed performance among the evaluated models, with an accuracy of 80.00%, a weighted F1-score of 80.07%, and a ROC-AUC of 0.9190. SHAP analysis identified reading score, writing score, and gender as the most influential predictors of mathematics proficiency. The results indicate that the proposed framework provides competitive classification performance and model interpretability within the studied dataset.

Keywords: CatBoost, Optuna, SHAP, Explainable Artificial Intelligence.

1 INTRODUCTION

The importance of mathematics cannot be undermined when it comes to education since it plays a vital role in the development of logical and analytical skills, which are crucial in both academic study and practical decision making. Through these skills, individuals are able to examine information critically, reason out, and make deductions. In this respect, proficiency in mathematics is one of the indicators of academic success. Research conducted lately has shown that mathematics education plays an important role in developing critical thinking skills, and cognitive traits play a crucial role in determining academic success. With regard to EDM developments, educational institutions can predict academic performance of the students through a data driven approach [1], [2], [3].

Educational data is growing rapidly, and as a result, machine learning methods are increasingly being used to predict the academic achievements of students, which has led to Educational Data Mining (EDM) becoming a significant field of study [4]. The machine learning models are capable of detecting intricate relations between different demographic, socioeconomic, and academic variables and therefore predict with higher accuracy than many traditional statistical methods when applied on big educational datasets [5]. Moreover, prediction models can assist teachers in classifying students according to their mathematics proficiency, thereby supporting educational assessment and informed decision-making.

Numerous research papers have used several algorithms in machine learning such as Decision Tree, Random Forest, Support Vector Machine, XGBoost, and LightGBM for predicting academic performance of the students, proving that machine learning can indeed help make educational decisions [5]. Despite

being quite effective, these algorithms had been either configured using default settings or traditional optimization methods in most of the previous research papers, which might hinder the capacity of achieving maximum performance by the models[6]. Moreover, in most research papers, prediction accuracy was prioritized over model interpretation, preventing us from knowing the influence of certain features on the prediction outcomes [7].

This framework has gained a lot of traction due to its ability to process categorical features without additional preprocessing efforts, thus making it efficient when dealing with tabular data featuring various types of information [8]. However, tuning hyperparameters is also an essential step, since the effectiveness of CatBoost is based on proper selection of the parameters of the model, rather than using the default parameter configuration. Recent researches have demonstrated that by using certain optimization frameworks, including Optuna, one can easily find optimal hyperparameters, decreasing the effort involved in manual tuning at the same time [9]. Thus, combining CatBoost and Optuna proves to be an effective solution.

Highly accurate machine learning models tend to be hard to interpret since the process of decision-making cannot be observed, hence the term "black-box" models. Such models have prompted a need for explainable artificial intelligence (XAI) especially in areas where decision making is critical and requires high transparency and accountability [10]. Explainable artificial intelligence is a technique which offers ways through which users can understand how prediction models arrive at their output results [11]. There are a number of techniques under XAI, but the SHapley Additive exPlanation (SHAP) method has been among the popular methods used in interpreting machine learning models due to its ability to quantify the impact of every feature on the predictions [12].

While the body of literature on student performance prediction is constantly expanding, a majority of the existing studies concentrate on enhancing the accuracy of predictions based on traditional machine learning algorithms, while little is said about the optimization and interpretation of such models. Moreover, recent reviews indicate that explainable artificial intelligence (XAI) methods remain underused in educational prediction research [13]. Furthermore, recent studies have reported that integrating hyperparameter optimization into machine learning models can improve predictive performance compared with default parameter settings [12]. Furthermore, recent studies have reported that integrating hyperparameter optimization into machine learning models can improve predictive performance compared with default parameter settings [9].

Inspired by those findings, the paper proposes the framework, which combines CatBoost, Optuna, and SHAP for predicting students' mathematics proficiency. The effectiveness of the suggested framework is estimated based on the Student Performance in Mathematics dataset obtained on Kaggle platform and includes the following metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix. Additionally, in order to test the effectiveness of optimized CatBoost model, the latter is compared to XGBoost and LightGBM, while SHAP is used for explaining the contribution of each feature to classification results.

2 LITERATURE REVIEW

2.1 Educational Data Mining

Although there has been an increasing body of literature focusing on predicting student performance, the majority of these studies pay greater attention to predictive accuracy than to the interpretability and explainability of machine learning models in educational contexts. Nevertheless, such combined solutions are rare in educational data mining [7]. Being encouraged by this limitation in the existing literature, this paper suggests the framework that integrates CatBoost, Optuna, and SHAP in order to predict students' proficiency in mathematics. This framework will be tested using the Student Performance in Mathematics dataset from Kaggle, and its performance will be measured using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix. For validation purposes, the optimized CatBoost model will be compared with other similar machine learning algorithms (XGBoost and LightGBM), and SHAP will be utilized for explaining the contribution of each predictor to the final results of the model [13].

2.2 CatBoost for Classification

CatBoost is a gradient boosting technique which has been known to perform well with tabular data, especially that which includes categorical variables. This capability to handle categorical data directly eliminates the need for heavy data pre-processing hence reducing complexity in modeling. CatBoost has also been found to perform competitively in various classification tasks especially where the data is heterogeneous [14] [15]. Furthermore, CatBoost uses ordered boosting and target statistics techniques to eliminate prediction biases hence improving generalization in the model [13] [14].

2.3 Hyperparameter Optimization Using Optuna

The performance of any machine learning algorithm is highly dependent on the set of parameters it uses since different parameters may result in varying prediction accuracies. In order to come up with the optimum machine learning algorithm, auto hyperparameter tuning has emerged as a common practice that ensures minimum restrictions of manually trying out several algorithms in addition to making the process of conducting machine learning experiments more repeatable [15], [16]. Optuna is one of the many optimization frameworks that have drawn much attention due to its efficient adaptive sampling and pruning capabilities that make it more efficient compared to other methods such as grid and random searches [17].

2.4 Explainable Artificial Intelligence Using SHAP

Despite the fact that machine learning models have high performance, most of them can be called black-box models because the process of obtaining predictions using these models remains unclear. That is why today the issue of Explainable Artificial Intelligence (XAI) gains increasing significance. The goal of XAI is to increase the transparency, credibility, and reliability of machine learning models through providing an explanation for predictions made by these models [18]. There are different techniques that fall under the definition of XAI. One of them is SHapley Additive exPlanations (SHAP), which calculates the contribution of every variable to a particular prediction according to Shapley value from the game theory and allows for both local and global model interpretation [19]. In education, SHAP becomes an increasingly popular technique, which is used for explaining predictions [20].

2.5 Related Studies

Recent works have used machine learning algorithms for predicting academic achievement of students with the help of algorithms like Decision Tree, Random Forest, Support Vector Machine, XGBoost, and LightGBM showing encouraging results on different education datasets [21]. Recently, another algorithm named CatBoost has received considerable popularity due to its capability to process both categorical and numeric variables along with good prediction performances on educational datasets [22]. But majority of these works focus on the aspect of prediction accuracy while optimization and interpretation of models remains less considered. Hence, there is a requirement for an explanation-based framework that can predict accurately [23].

2.6 Research Gap

Previous literature on the evaluation of machine learning models usually uses the default setting of hyperparameters or standard methods of optimization, while a rather small number of papers integrate CatBoost and adaptive hyperparameter optimization and explainable machine learning approaches. Moreover, usually, there is no interpretation of the models developed, which makes the analysis of the impact of predictors difficult.

This research suggests integrating CatBoost, Optuna, and SHAP for predicting the level of mathematical proficiency of the students. The Optuna library will be used for tuning the hyperparameters of the CatBoost model. SHAP will provide the interpretation of the prediction made by identifying the factors having the greatest influence on mathematics proficiency.

3 RESEARCH METHODS

3.1 Research Framework

The current research work introduces a predictive modeling technique to predict students' mathematics proficiency via the CatBoost algorithm trained using Optuna and explained by SHAP (SHapley Additive exPlanations). The entire research process is comprised of six phases, including data gathering, data pre-processing, exploratory data analysis, statistical analysis, model building, and explanation. Figure 1 represents the entire research framework used in the current study.

Figure 1. Research Framework

3.2 Dataset

The above-mentioned experiments were performed based on the Student Performance in Mathematics dataset collected by Kaggle. The dataset contains 1,000 records of students with eight features representing students' demographic information, socio-economic status, and academic achievement. Five features are categorical, which include gender, race/ethnicity, parental level of education, lunch, and test preparation course, whereas three other features are numeric such as math score, reading score, and writing score.

In order to create a classification model, mathematics score variable in the initial dataset was modified by creating a new target variable Mathematics Proficiency. If a student had a mathematics score less than 60, he/she was classified as Low; for those whose score was between 60 and 79, their category was Medium; and for those whose score was 80 or more – their category was High. Consequently, the original regression task was transformed to the multiclass classification problem, which allowed predicting students' mathematics proficiency levels instead of numeric values.

3.3 Data Preprocessing

Prior to developing the prediction model, the dataset was preprocessed to ensure its suitability for analysis. The preprocessing included checking for missing values and duplicate records, and no data quality issues were identified; therefore, no further cleaning was required. Subsequently, the original Math Score was recoded into a new target variable, Mathematics Proficiency, consisting of three classes (Low, Medium, and High), thereby transforming the original regression problem into a multiclass classification task. Finally, the dataset was divided into training (80%) and testing (20%) subsets using stratified sampling with a random seed of 42 to preserve the class distribution. The testing dataset was reserved exclusively for the final model evaluation and was not used during model training or hyperparameter optimization.

3.4 Exploratory and Inferential Analysis

Following the preprocessing stage, exploratory data analysis (EDA) was done to have a better insight into the data set. This involved conducting descriptive statistics, data distribution, and correlation analysis in order to determine the features of the variables and how they relate before developing the models.

In an effort to establish the importance of each of the predictors, inferential statistical analysis was done. The Chi-Square test was conducted to establish the relationship between categorical predictors and the response variable, whereas the Kruskal-Wallis test was used to find the differences among the numerical predictors across the three mathematics proficiency levels. Those predictors whose significance level was less than 0.05 were regarded as statistically significant.

3.5 CatBoost Model Development

CatBoost was selected as the primary classification algorithm because of its ability to handle categorical and numerical features without extensive preprocessing. Hyperparameter optimization was performed using Optuna on the training dataset through 30 optimization trials. During each trial, five-fold stratified cross-validation was applied, and the objective function was to maximize the mean classification accuracy across the validation folds. The search space consisted of iterations, tree depth, learning rate, L2 leaf regularization, bagging temperature, and random strength, as presented in Table 1.

Table 1. Hyperparameter Search Space

Hyperparameter	Search Space
Iterations	100–500
Tree Depth	4–10
Learning Rate	0.01–0.30 (log scale)
L2 Leaf Regularization	1–10
Bagging Temperature	0–5
Random Strength	1–10

The hyperparameter configuration with the highest mean cross-validation accuracy was selected as the final CatBoost model. The optimized hyperparameter values are presented in Table 2. The resulting configuration consisted of 469 iterations, a tree depth of 7, a learning rate of 0.03645, an L2 leaf regularization value of 4.69031, a bagging temperature of 0.05525, and a random strength of 4.78462. The testing dataset was excluded from the hyperparameter optimization process and was used only for the final performance evaluation, thereby preventing data leakage.

Table 2. Best Hyperparameters Obtained by Optuna

Hyperparameter	Best Value
Iterations	469
Tree Depth	7
Learning Rate	0.03645
L2 Leaf Regularization	4.69031
Bagging Temperature	0.05525
Random Strength	4.78462
Loss Function	MultiClass
Random Seed	42

3.6 Model Evaluation

The performance of the improved CatBoost algorithm was evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix. Different evaluation metrics were selected to assess the classification performance of the model from multiple perspectives. Since this study addressed a multiclass classification problem, the ROC-AUC was computed using the one-vs-rest (OvR) strategy with the weighted-average approach. To evaluate the effectiveness of the proposed method, the optimized CatBoost model was compared with two popular gradient boosting frameworks, namely XGBoost and LightGBM. All models were trained and evaluated using identical training and testing datasets under the same evaluation protocol.

To further assess whether the observed differences in model performance were statistically robust, the Friedman test was used to compare the three models across the repeated cross-validation evaluations. When significant differences were detected, pairwise Wilcoxon signed-rank tests were performed to compare CatBoost with XGBoost and LightGBM. Holm correction was applied to adjust for multiple pairwise comparisons. The statistical analyses were conducted for Accuracy, Weighted F1-score, and ROC-AUC, with a significance level of $\alpha = 0.05$.

To quantify the uncertainty of the repeated cross-validation estimates, 95% confidence intervals were calculated for the main performance metrics. The confidence intervals were reported for Accuracy, weighted F1-score, and ROC-AUC for each evaluated model. These intervals were used to provide an estimate of the uncertainty surrounding the observed model performance and to facilitate a more cautious comparison among CatBoost, XGBoost, and LightGBM.

To further assess whether the observed differences in model performance were statistically robust, paired performance estimates obtained from the repeated stratified cross-validation were compared between CatBoost and the baseline models, namely XGBoost and LightGBM. Appropriate statistical comparison procedures were applied to the repeated evaluation results for Accuracy, weighted F1-score, and ROC-AUC. The statistical comparisons were used to determine whether the observed performance differences between CatBoost and the baseline models were consistent across the repeated data partitions.

3.7 Explainable AI

Following the evaluation of the model, SHAP (SHapley Additive exPlanations) was used to enhance the interpretability of the optimized CatBoost model. SHAP was used to determine the importance of each variable to the outcomes of the classifier, as well as the effect of each of the predictors on the prediction. The analysis was carried out using a SHAP summary plot and a SHAP feature importance plot. This explanatory analysis is an addition to the predictiveness of the model that has been proposed, as it provides insights about the predictors of students' math skills.

4 RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis (EDA)

EDA was performed on the data to analyze the features of the dataset prior to developing the predictive model. The Student Performance in Mathematics Dataset comprises of 1,000 student records and includes eight different variables, with five categorical variables being gender, race/ethnicity, parental level of education, lunch, and test preparation course; while three numerical variables are math score, reading score, and writing score.

Upon an initial look at the data, it was clear that the dataset was complete without any missing or duplicated records, meaning that there was no need for any further data cleaning. The descriptive statistics of the numerical variables are listed in Table 3 below.

Table 3. Descriptive Statistics of Numerical Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
Math Score	67.81	15.25	15	100
Reading Score	70.38	14.11	25	100
Writing Score	69.14	15.03	15	100

Among the three academic variables, reading score has the highest average (70.38), followed by writing score (69.14) and math score (67.81). The relatively large range of scores indicates that the dataset represents students with varying levels of academic performance, making it suitable for predictive analysis.

To build a classification model, the math score was converted into three proficiency levels: Low (score < 60), Medium (60–79), and High (≥ 80). This categorization allows the model to predict students' mathematics proficiency instead of estimating numerical scores.

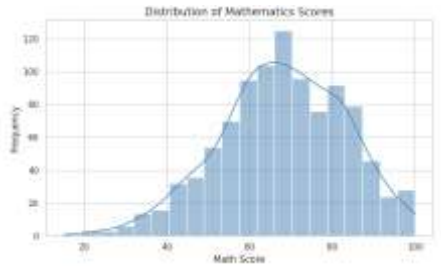

Figure 2. Distribution of Mathematics Scores

From the histogram, it can be seen that the majority of students obtained mathematics scores between 60 and 80, with the highest frequency observed between 65 and 70. Overall, the scores were concentrated around the middle range, with fewer observations at the lower and upper ends of the distribution. This distribution provides an overview of the students' mathematics performance and indicates that most students achieved scores within the moderate to relatively high range.

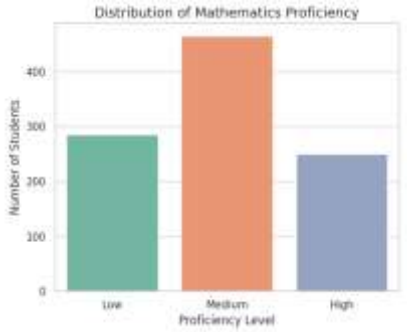

Figure 3. Distribution of Mathematics Proficiency

The Medium class includes the highest proportion of students (46.5%), followed by the Low class (28.5%) and the High class (25.0%). Although the class frequencies are not equal, the distribution does not indicate a severe class imbalance. Therefore, resampling techniques were not applied in this study. Instead, stratified sampling was used during the train-test split to preserve the original class distribution in both subsets.

To investigate the relationships among the numerical features, Pearson correlation analysis was performed, as presented in Figure 4.

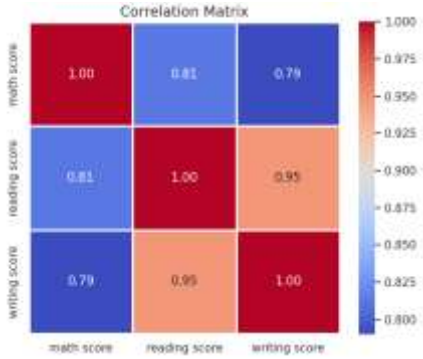

Figure 4. Correlation Matrix of Numerical Variables

As it can be seen from the matrix, there are positive relations between all academic features. The strongest correlation was observed between Reading Score and Writing Score (r = 0.949), while Mathematics Score was positively correlated with Reading Score (r = 0.812) and Writing Score (r = 0.790). his indicates a strong positive association between reading, writing, and mathematics scores in the dataset

From the analysis, it can be concluded that the dataset is complete and does not contain missing or duplicated records. Although the class frequencies are not equal, the distribution does not indicate a severe class imbalance. Stratified sampling was therefore used to preserve the class distribution during model development. The observed variation among the academic scores also provides sufficient variability for the subsequent multiclass classification analysis

4.2 Inferential Statistical Analysis

Prior to the development of the predictive model, inferential statistical analysis was conducted to examine the association between the predictors and Mathematics Proficiency. For categorical variables, the Chi-Square test was used to assess their association with Mathematics Proficiency, while the Kruskal–Wallis test was applied to examine differences in the numerical variables across the three proficiency groups. The significance level was set at 0.05. The results are summarized in Table 4.

Table 4. Results of Inferential Statistical Analysis

Variable	Statistical Test	Statistic	p-value	Decision
Gender	Chi-Square	21.633	0.00002	Significant
Race/Ethnicity	Chi-Square	90.043	<0.001	Significant
Parental Level of Education	Chi-Square	22.232	0.01397	Significant
Lunch	Chi-Square	118.793	<0.001	Significant
Test Preparation Course	Chi-Square	10.938	0.00422	Significant
Reading Score	Kruskal–Wallis	564.843	<0.001	Significant
Writing Score	Kruskal–Wallis	532.143	<0.001	Significant

As shown in Table 4, all predictor variables showed statistically significant associations or differences with Mathematics Proficiency ($p < 0.05$). Among the categorical variables, Lunch produced the largest Chi-Square statistic ($\chi^2 = 118.793$), followed by Race/Ethnicity ($\chi^2 = 90.043$). Gender, Parental Level of Education, and Test Preparation Course also showed statistically significant associations with Mathematics Proficiency ($p < 0.05$). Among the numerical variables, Reading Score produced the largest Kruskal-Wallis statistic ($H = 564.843$), followed by Writing Score ($H = 532.143$).

4.3 CatBoost Model Development with Optuna

CatBoost was selected as the primary classification algorithm because it can process categorical and numerical variables within a unified modeling framework. In addition, its ordered boosting mechanism is designed to reduce prediction bias during model training, making it suitable for tabular classification problems.

Hyperparameter optimization was performed using the Optuna framework to identify an effective configuration for the CatBoost classifier. Optuna employs an adaptive search strategy that enables the optimization process to explore the predefined hyperparameter space efficiently. In this study, the optimization was conducted through 30 trials using five-fold stratified cross-validation, with classification accuracy as the optimization objective. The testing dataset was not involved in the optimization process and was reserved exclusively for the final model evaluation. The resulting optimal hyperparameter configuration is presented in Table 2**.**

The optimized CatBoost model was obtained using 469 iterations, a tree depth of 7, a learning rate of 0.03645, an L2 leaf regularization value of 4.69031, a bagging temperature of 0.05525, and a random strength of 4.78462. Using this optimized configuration, the model achieved an accuracy of 80.00% on the held-out testing dataset. The optimized model was subsequently used for comparison with XGBoost and LightGBM and for the SHAP-based explainability analysis. Using the optimized configuration, the CatBoost classifier achieved an accuracy of 80.0% on the held-out testing dataset. The resulting model was subsequently used for the final performance comparison with other gradient boosting algorithms and for the explainability analysis using SHAP. Model performance was evaluated using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC-AUC.

4.4 Model Performance Evaluation

The optimized Cat Boost model was evaluated on the held-out testing dataset and compared with XGBoost and LightGBM using the same evaluation metrics. As shown in Table 5, CatBoost achieved the highest observed performance, with an Accuracy of 0.8000, Precision of 0.8052, Recall of 0.8000, F1-score of 0.8007, and ROC-AUC of 0.9190. The confusion matrix of the CatBoost model is presented in Figure 5.

Table 5. Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
CatBoost	0.8000	0.8052	0.8000	0.8007	0.9190
XGBoost	0.7750	0.7823	0.7750	0.7761	0.8830
LightGBM	0.7900	0.7934	0.7900	0.7902	0.8942

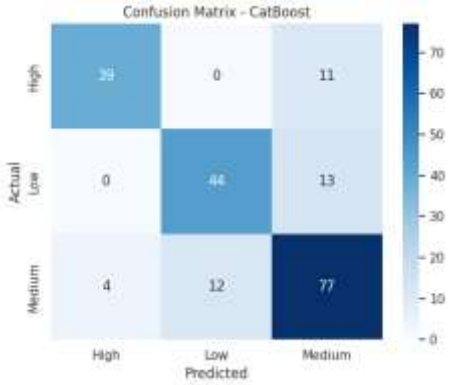

Figure 5. Confusion Matrix of the CatBoost Model

The confusion matrix shows that 39 of 50 High-level students, 44 of 57 Low-level students, and 77 of 93 Medium-level students were correctly classified. The classification errors were mainly observed between adjacent proficiency categories.

To further assess the robustness of the model performance beyond the original 80:20 train-test split, repeated stratified 10-fold cross-validation with 5 repetitions was performed. As presented in Table 6, CatBoost achieved the highest mean Accuracy (0.8020 ± 0.0453), Weighted F1-score (0.8018 ± 0.0455), and ROC-AUC (0.9184 ± 0.0267). XGBoost and LightGBM obtained lower mean Accuracy values of 0.7712 ± 0.0443 and 0.7670 ± 0.0436, respectively.

Table 6. Repeated Stratified Cross-Validation Performance

Model	Accuracy (Mean ± SD)	95% CI	Weighted F1-score (Mean ± SD)	95% CI	ROC-AUC (Mean ± SD)	95% CI
CatBoost	0.8020 ± 0.0453	0.7895–0.8145	0.8018 ± 0.0455	0.7892–0.8144	0.9184 ± 0.0267	0.9110–0.9258
XGBoost	0.7712 ± 0.0443	0.7590–0.7835	0.7710 ± 0.0441	0.7588–0.7833	0.8993 ± 0.0262	0.8921–0.9066
LightGBM	0.7670 ± 0.0436	0.7549–0.7791	0.7668 ± 0.0435	0.7547–0.7788	0.8990 ± 0.0272	0.8915–0.9065

Figure 6 illustrates the distribution of Accuracy across the 50 repeated evaluations for each model. CatBoost shows the highest central performance compared with XGBoost and LightGBM, while the

distributions indicate variability across different data partitions.

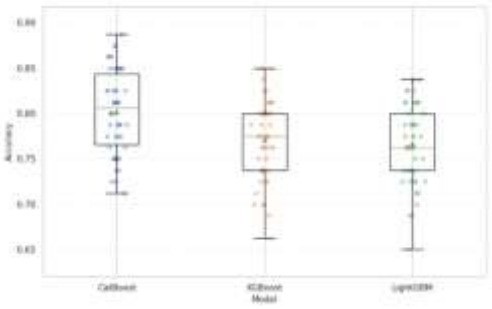

Figure 6. Accuracy Distribution Across Repeated Stratified Cross-Validation

To further examine the observed performance differences, the Friedman test indicated statistically significant differences among the three models for Accuracy, Weighted F1-score, and ROC-AUC ($p < 0.001$). Pairwise Wilcoxon signed-rank tests with Holm correction indicated higher performance of CatBoost compared with XGBoost and LightGBM across the evaluated metrics (adjusted $p < 0.001$). These results indicate that the observed differences were statistically significant under the applied repeated cross-validation evaluation procedure. However, the statistical findings should be interpreted within the context of the present dataset and evaluation design and should not be considered evidence of universal superiority of CatBoost. The detailed statistical comparisons are presented in Table 7.

Table 7. Statistical Comparison of Model Performance

Comparison	Metric	Mean Difference	Adjusted p-value	Result
CatBoost vs XGBoost	Accuracy	0.0307	< 0.001	Significant
CatBoost vs LightGBM	Accuracy	0.0350	< 0.001	Significant
CatBoost vs XGBoost	Weighted F1-score	0.0308	< 0.001	Significant
CatBoost vs LightGBM	Weighted F1-score	0.0351	< 0.001	Significant
CatBoost vs XGBoost	ROC-AUC	0.0191	< 0.001	Significant
CatBoost vs LightGBM	ROC-AUC	0.0194	< 0.001	Significant

Overall, the results show that CatBoost achieved the highest observed performance on the held-out test set and maintained the highest average performance across repeated cross-validation. The statistical comparison further indicates that the performance differences between CatBoost and the two baseline models were significant under the applied evaluation procedure.

4.5 Explainable AI Analysis Using SHAP

To improve the interpretability of the optimized CatBoost model, SHAP was applied to quantify the contribution of each feature to the model predictions. The analysis was performed on 200 testing instances using seven predictor variables across three Mathematics Proficiency classes. The global feature importance based on mean absolute SHAP values is presented in Table 8.

Table 8. Feature Importance Based on Mean Absolute SHAP Values

Rank	Feature	Mean Absolute SHAP Value
1	Reading Score	0.5844
2	Gender	0.5363
3	Writing Score	0.4915
4	Lunch	0.2133
5	Race/Ethnicity	0.1877
6	Test Preparation Course	0.0901
7	Parental Level of Education	0.0728

As shown in Table 8, Reading Score had the highest mean absolute SHAP value (0.5844), followed by Gender (0.5363) and Writing Score (0.4915). These features therefore made the largest contributions to the CatBoost model predictions, while Test Preparation Course and Parental Level of Education showed comparatively smaller contributions.

Figure 7 presents the SHAP summary plot for the High Mathematics Proficiency class. The plot illustrates the magnitude and direction of individual feature contributions to the model output. Reading Score and Writing Score show relatively wide SHAP distributions, indicating substantial variation in their contributions across individual observations. Positive SHAP values represent contributions toward the High proficiency prediction, whereas negative values represent contributions away from this class.

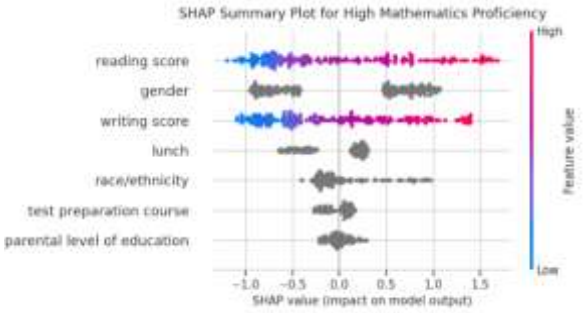

Figure 7. SHAP Summary Plot for High Mathematics Proficiency

Figure 8 presents the global feature importance based on mean absolute SHAP values. Overall, Reading Score, Gender, and Writing Score were the

three most influential features in the CatBoost model, while the remaining variables provided comparatively smaller contributions. These results improve the interpretability of the classification model by showing which predictors contributed most strongly to its predictions.

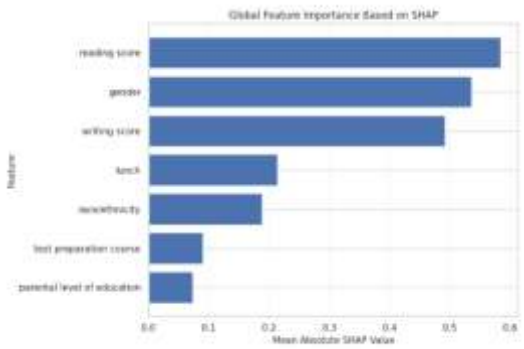

Figure 8. Global Feature Importance Based on SHAP

4.6 Discussion

The results demonstrate that the optimized CatBoost model achieved the highest observed performance among the evaluated models. On the held-out testing dataset, CatBoost obtained an Accuracy of 80.00%, a weighted F1-score of 80.07%, and a ROC-AUC of 0.9190, outperforming XGBoost and LightGBM under the same evaluation conditions. To further assess robustness, repeated stratified 10-fold cross-validation with five repetitions was performed. CatBoost maintained the highest mean Accuracy (0.8020 ± 0.0453), weighted F1-score (0.8018 ± 0.0455), and ROC-AUC (0.9184 ± 0.0267).

The statistical analysis further indicated differences among the three models across Accuracy, weighted F1-score, and ROC-AUC. The Friedman test showed statistically significant differences ($p < 0.001$), while the pairwise Wilcoxon signed-rank tests with Holm correction indicated higher performance for CatBoost compared with both XGBoost and LightGBM across the evaluated metrics (adjusted $p < 0.001$). These findings indicate that the observed performance advantage of CatBoost was consistent under the repeated cross-validation evaluation procedure. However, the statistical results should be interpreted within the context of the present dataset and evaluation design rather than as evidence of universal superiority of CatBoost.

The SHAP analysis provided additional insight into the prediction mechanism of the optimized CatBoost model. Reading Score, Gender, and Writing Score were the three most influential features based on mean absolute SHAP values, with Reading Score showing the highest contribution, followed by Gender and Writing Score. Lunch, Race/Ethnicity, Test Preparation Course, and Parental Level of Education showed comparatively smaller contributions. These results indicate the relative importance of these variables to the model predictions, but they should not be interpreted as evidence of causal relationships.

An important consideration is that Reading Score and Writing Score were collected during the same assessment period as the Mathematics Score in the original dataset. Therefore, the proposed framework should be interpreted as a mathematics proficiency classification model rather than a genuine early prediction system. Its primary value lies in supporting interpretable educational assessment based on contemporaneously available academic information.

Overall, the integration of CatBoost, Optuna, and SHAP provides a framework that combines predictive performance with model interpretability. Nevertheless, the findings are based on a single educational dataset, and the contribution of individual modeling components was not independently evaluated. Future research should therefore validate the framework using larger and independent educational datasets and consider predictors that are available before mathematics assessment, such as demographic, socioeconomic, attendance, or learning-related variables.

5 CONCLUSION

In this study, an explainable machine learning approach for classifying students' mathematics proficiency was proposed by integrating CatBoost, Optuna, and SHAP. The Student Performance in Mathematics dataset was first examined through descriptive and inferential statistical analyses to characterize the predictors and their associations with mathematics proficiency. Using Optuna for hyperparameter optimization, the CatBoost model achieved the highest observed performance among the evaluated models on the held-out testing dataset, with an accuracy of 80.00%, a weighted F1-score of 80.07%, and a ROC-AUC of 0.9190. Compared with the XGBoost and LightGBM baseline models under the same evaluation settings, CatBoost obtained higher values across the reported performance metrics. Furthermore, repeated stratified 10-fold cross-validation with five repetitions showed that CatBoost maintained the highest mean Accuracy (0.8020 ± 0.0453), weighted F1-score (0.8018 ± 0.0455), and ROC-AUC (0.9184 ± 0.0267). The SHAP analysis identified Reading Score, Gender, and Writing Score as the three features with the highest mean absolute SHAP values, while Lunch, Race/Ethnicity, Test

Preparation Course, and Parental Level of Education showed relatively smaller contributions to the model's predictions. Since Reading Score and Writing Score were collected during the same assessment period as the Mathematics Score, the proposed framework should be interpreted as a mathematics proficiency classification model rather than an early prediction system. Overall, the integration of CatBoost, Optuna, and SHAP provides a combination of predictive performance and model interpretability within the studied dataset.

The statistical comparison indicated significant differences in model performance under the applied repeated cross-validation procedure; however, these findings should be interpreted within the context of the present dataset and evaluation design rather than as evidence of universal superiority of CatBoost. Future research should validate the framework using larger and independent educational datasets and external validation settings to further assess its generalizability. Future studies may also investigate nested cross-validation for model selection and optimization, as well as predictor variables available before mathematics assessment, such as demographic, socioeconomic, attendance, or learning behavior data, to support genuine early prediction. Further research may also examine alternative ensemble methods, deep learning, and other explainable artificial intelligence techniques.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, 2020, doi: 10.1002/widm.1355.

[2] Q. Xie *et al.*, "Cognitive style and Students' academic achievement: a meta-analysis," *Front. Educ. (Lausanne).*, vol. 10, no. October, 2025, doi: 10.3389/feduc.2025.1634732.

[3] M. Wang, M. E. E. Mohd Matore, and R. Rosli, "A systematic literature review on analytical thinking development in mathematics education: trends across time and countries," *Front. Psychol.*, vol. 16, no. June, pp. 1–10, 2025, doi: 10.3389/fpsyg.2025.1523836.

[4] W. Xiao, P. Ji, and J. Hu, "A survey on educational data mining methods used for predicting students' performance," *Engineering Reports*, vol. 4, no. 5, p. e12482, May 2022, doi: https://doi.org/10.1002/eng2.12482.

[5] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022, doi: 10.1186/s40561-022-00192-z.

[6] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 2, 2024, doi: 10.1007/s44163-023-00079-z.

[7] A. Abukader, A. Alzubi, and O. R. Adegboye, "Intelligent System for Student Performance Prediction: An Educational Data Mining Approach Using Metaheuristic-Optimized LightGBM with SHAP-Based Learning Analytics," *Applied Sciences (Switzerland)*, vol. 15, no. 20, pp. 1–25, 2025, doi: 10.3390/app152010875.

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Section 4, pp. 6638–6648, 2018.

[9] M. Hassanali, M. Soltanaghaei, T. Javdani Gandomani, and F. Zamani Boroujeni, "Software development effort estimation using boosting algorithms and automatic tuning of hyperparameters with Optuna," *Journal of Software: Evolution and Process*, vol. 36, no. 9, p. e2665, 2024, doi: https://doi.org/10.1002/smr.2665.

[10] V. Hassija *et al.*, "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence," *Cognit. Comput.*, vol. 16, no. 1, pp. 45–74, 2024, doi: 10.1007/s12559-023-10179-8.

[11] Md. T. Hosain, J. R. Jim, M. F. Mridha, and M. M. Kabir, "Explainable AI approaches in deep learning: Advancements, applications and challenges," *Computers and Electrical Engineering*, vol. 117, p. 109246, 2024, doi: https://doi.org/10.1016/j.compeleceng.2024.109246.

[12] G. Türkmen, "The Review of Studies on Explainable Artificial Intelligence in Educational Research," *Journal of Educational Computing Research*, vol. 63, no. 2, pp. 277–310, 2025, doi: 10.1177/07356331241310915.

[13] E. Ben George, "Explainable AI Methods for Predicting Student Grades and Improving Academic Success," *Journal of Information Systems Engineering and Management*, vol. 10, no. 23s, pp. 117–126, 2025, doi: 10.52783/jisem.v10i23s.3680.

[14] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, p. 94, 2020, doi: 10.1186/s40537-020-00369-8.

[15] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, Mar. 2023, doi: https://doi.org/10.1002/widm.1484.

[16] J. M. Tan *et al.*, "Hyperparameter optimization: Classics, acceleration, online, multi-objective, and tools," *Mathematical Biosciences and Engineering*, vol. 21, no. 6, pp. 6289–6335, 2024, doi: 10.3934/mbe.2024275.

[17] H. Almarzooq and U. bin Waheed, "Automating hyperparameter optimization in geophysics with Optuna: A comparative study," *Geophys. Prospect.*, vol. 72, no. 5, pp. 1778–1788, Jun. 2024, doi: https://doi.org/10.1111/1365-2478.13484.

[18] M. Mersha, K. Lam, J. Wood, A. K. AlShami, and J. Kalita, "Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction," *Neurocomputing*, vol. 599, 2024, doi: 10.1016/j.neucom.2024.128111.

[19] M. Saarela and V. Podgorelec, "Recent Applications of Explainable AI (XAI): A Systematic Literature Review," *Applied Sciences (Switzerland)*, vol. 14, no. 19, 2024, doi: 10.3390/app14198884.

[20] L. C. Nnadi, Y. Watanobe, M. Rahman, and A. M. John-otumu, "Scopus - Document details - Prediction of Students' Adaptability Using Explainable AI in Educational Machine Learning Models," 2024.

[21] Z. Ersozlu, S. Taheri, and I. Koch, "A review of machine learning methods used for educational data," *Educ. Inf. Technol. (Dordr).*, vol. 29, no. 16, pp. 22125–22145, 2024, doi: 10.1007/s10639-024-12704-0.

[22] Z. Fan, J. Gou, and S. Weng, "Complementary CatBoost based on residual error for student performance prediction," *Pattern Recognit.*, vol. 161, p. 111265, 2025, doi: https://doi.org/10.1016/j.patcog.2024.111265.

[23] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive Models for Educational Purposes: A Systematic Review," *Big Data and Cognitive Computing*, vol. 8, no. 12, pp. 1–42, 2024, doi: 10.3390/bdcc8120187.