RESEARCH ARTICLE

A Gradient Boosting-Based Feature Selection Framework for Predicting Student Performance

Shaoyuan Weng[1,*], **Yuanyuan Zheng**[1], **Chao Zhang**[1] **and Zimeng Liu**[1]

[1] Xiamen Institute of Software Technology, Xiamen 361024, China

Abstract

In educational data mining, accurate prediction of student performance is important for supporting timely intervention for at-risk students. However, educational datasets often include irrelevant or redundant features that could reduce the performance of prediction models. To tackle this issue, this study proposes a gradient boosting-based feature selection framework that can automatically identify and obtain the most important features for student performance prediction. The proposed framework leverages the gradient boosting model to calculate feature importance and refine the feature subset, aiming to achieve comparable or superior prediction performance using fewer but important input features. To ensure a robust evaluation of the results, we apply a 10-fold cross-validation strategy with 10 repetitions. Experimental results on the Mathematics and Portuguese Language course datasets demonstrate that the proposed framework is able to consistently outperform the baseline models in terms of the evaluation metrics used. These findings highlight the effectiveness of the proposed feature selection for student performance, which makes it a reliable tool for data-driven educational analytics.

Keywords: feature selection, gradient boosting, student performance prediction, educational data mining.

1 Introduction

In the big data era, with the fast development of digital learning environments, large amounts of educational data can be continuously collected from the learning management systems [1], online assessments [2], and student interaction platforms [3]. Such abundance of data has brought new opportunities for Educational Data Mining (EDM), which applies data-driven analytical methods to discover patterns of learning behaviors, understand educational processes, and improve learning outcomes from educational data [4–6]. One of the most important applications in EDM is student performance prediction and it has become particularly important for supporting early-warning systems [7], dropout prevention [8], and the design of adaptive learning interventions [9]. For example, accurate prediction of student performance is able to allow educators to identify at-risk students early, provide them with personalized support, and optimize teaching strategies to boost engagement and academic achievement [10, 11].

To accurately predict student performance, a wide range of machine learning models have been employed. For the single regressors/models, the Multi-Layer

Submitted: 14 October 2025
Accepted: 22 October 2025
Published: 28 November 2025

Vol. 1, **No.** 1, 2025.
10.62762/TEDM.2025.414136

***Corresponding author:**
✉ Shaoyuan Weng
wengshaoyuany@163.com

Citation
Weng, S., Zheng, Y., Zhang, C., & Liu, Z. (2025). A Gradient Boosting-Based Feature Selection Framework for Predicting Student Performance. *ICCK Transactions on Educational Data Mining*, 1(1), 25–35.

Perceptrons (MLP) and Support Vector Machines (SVM) have demonstrated strong predictive capability for student performance prediction [12]. For ensemble models, the Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) often achieve higher generalization performance across diverse educational datasets [13, 14]. However, educational data often consist of mixed-type features (numerical, categorical, and other types) that can reduce the performance of the above machine learning models. In this context, the Categorical Boosting (CatBoost) has gained increasing attention due to its ability to process categorical features. Early studies showed that CatBoost is able to outperform the traditional models such as the MLP, SVM, RF, and XGBoost in student performance prediction using the tested educational datasets [15]. However, feature redundancy and irrelevance remain persistent challenges in educational datasets, as such features may reduce the performance of the prediction models [16]. For example, many student-related features, such as demographic or social features are collected into the educational datasets while irrelevant or redundant features may increase training time and obscure model interpretability issues especially when analyzing complex, multi-source educational data [26]. Therefore, an effective feature selection method is essential to identify the most important features and improve the model performance in student performance prediction.

Among the feature selection methods, they can be roughly divided into three types: filters, wrappers, and embedded methods [17]. Compared with filter and wrapper methods, embedded feature selection exhibits distinct advantages for stronger interaction with the underlying model and more efficient computation. Among the embedded methods, the tree-based models are widely employed for feature selection [18]. For example, Zhou et al. [19] enhanced a decision tree for this purpose, and their results demonstrated that the proposed model achieved superior performance in terms of accuracy, recall, and F1-score. Similarly, Fan et al. [20] utilized the RF for feature selection and showed that it improved the prediction of student performance.

Given that the CatBoost is a tree-based ensemble model and exhibits strong efficacy in handling categorical features, such features are common in educational datasets, this study proposes a gradient boosting-based feature selection framework for student performance prediction. The proposed framework leverages the CatBoost's inherent gradient boosting strategy to quantify and rank feature importance, thereby enabling the automatic selection and iterative refinement of the most relevant predictive features using the threshold. For robust evaluation, a 10-fold cross-validation strategy with multiple repetitions is implemented across the educational datasets.

The main contributions of this study are summarized as follows:

(1) We propose a gradient boosting-based feature selection framework for student performance prediction.

(2) The framework automatically identifies and retains the most informative features.

(3) A comprehensive experimental evaluation is conducted on the benchmark educational datasets using repeated 10-fold cross-validation. The results demonstrate that the proposed framework achieves superior prediction performance compared to baseline models.

(4) This study provides insights into the impact of feature selection on educational data mining tasks, offering an approach that supports more reliable and data-driven student performance prediction.

2 Methods

In this section, we first introduce the CatBoost and its feature importance mechanism, followed by the proposed feature selection framework for student performance prediction.

2.1 CatBoost

The CatBoost is a gradient boosting algorithm developed to address common limitations of traditional boosting methods in handling categorical features [21]. Unlike the conventional boosting models, the CatBoost uses an ordered boosting strategy: this strategy mitigates the prediction drift by ensuring that each model iteration is trained exclusively using target statistics derived from preceding data samples. By using this design, the model generalization is enhanced while effectively reducing overfitting.

Given a dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^{N}$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ is an m-dimensional feature vector and y_i is the corresponding target feature, the CatBoost constructs an additive model by sequentially fitting weak base estimators using the decision trees to the

residuals of previous models. The prediction function after T iterations can be expressed as:

$$\hat{y}_i = \sum_{t=1}^{T} \eta f_t(\mathbf{x}_i), \tag{1}$$

where $f_t(\cdot)$ denotes the decision tree at iteration t, and η is the learning rate controlling the contribution of each tree.

For each iteration, the CatBoost minimizes a differentiable loss function $L(y_i, \hat{y}_i)$ by updating the model according to the negative gradient of the loss with respect to the prediction:

$$g_i^{(t)} = -\frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}. \tag{2}$$

The new tree $f_t(\cdot)$ is trained to approximate the gradient $g_i^{(t)}$ for correcting the residual errors of prediction results from the previous iterations. This iterative refinement enable the CatBoost to capture complex nonlinear relationships between the input features and target feature.

2.2 Feature selection using the CatBoost

The aim of the proposed CatBoost-based feature selection framework is to improve prediction performance of baseline models for student performance prediction. The framework comprises three main steps: (1) extracting feature importance using the CatBoost, (2) selecting feature subsets based on a threshold, and (3) evaluating the model via 10-fold cross-validation with 10 repeated experiments.

Specifically, a CatBoost model is first trained on the educational dataset to compute feature importance scores. These features are then ranked by their importance values, and their cumulative contribution to the model is quantified. Only features whose cumulative importance meets or exceeds a predefined threshold are retained in the final subset. Finally, this reduced feature subset is used as input to the baseline models, including MLP, SVM, RF, XGBoost, and CatBoost for performance comparison. To ensure result robustness and mitigate the impact of randomness in data partitioning, each baseline model is evaluated using repeated 10-fold cross-validation.

For the calculation of feature importance, the importance score for feature j is computed as the total reduction in the loss function achieved by the splits across decision trees involving that feature, averaged over all decision trees:

$$I_j = \frac{1}{T} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_{t,j}} \Delta L_{s,t}, \tag{3}$$

where T is the number of decision trees, $\mathcal{S}_{t,j}$ is the set of splits on feature j in tree t, and $\Delta L_{s,t}$ is the reduction in the loss function contributed by split s in tree t.

Subsequently, to ensure the sum of all feature importance scores equals 100% for facilitating intuitive interpretation of relative importance, the initially computed feature importance scores are normalized using the following formula:

$$I_j = \frac{I_j}{\sum_{k=1}^{m} I_k}, \tag{4}$$

ensuring that $\sum_{j=1}^{m} I_j = 100\%$.

To select the most relevant subset of features, we adopt a cumulative importance thresholding approach. Specifically, features are first sorted in descending order based on their normalized importance scores. The cumulative importance of the top-k features is then calculated as:

$$C_k = \sum_{j=1}^{k} I_j^*. \tag{5}$$

Features are retained in the final subset as long as the cumulative importance does not exceed a predefined threshold τ. In this study, the threshold is set to $\tau = 0.95$, meaning that only the top features whose combined importance accounts for 95% of the total are selected:

$$\mathcal{F}_{\text{selected}} = \{f_j \mid C_j \leq 0.95\}. \tag{6}$$

where $\mathcal{F}_{\text{selected}}$ is the final selected features that are used to train the base models for student performance prediction.

This feature selection strategy ensures the retention of the most influential features while eliminating those that contribute marginally to prediction accuracy.

In addition, the feature importance derived from the Mathematics and Portuguese Language course datasets is presented in Figures 1 and 2, respectively. The data descriptions can be found in Table 1. The values are all based on the percentage, meaning the total of all feature importance scores sum up to 100%.

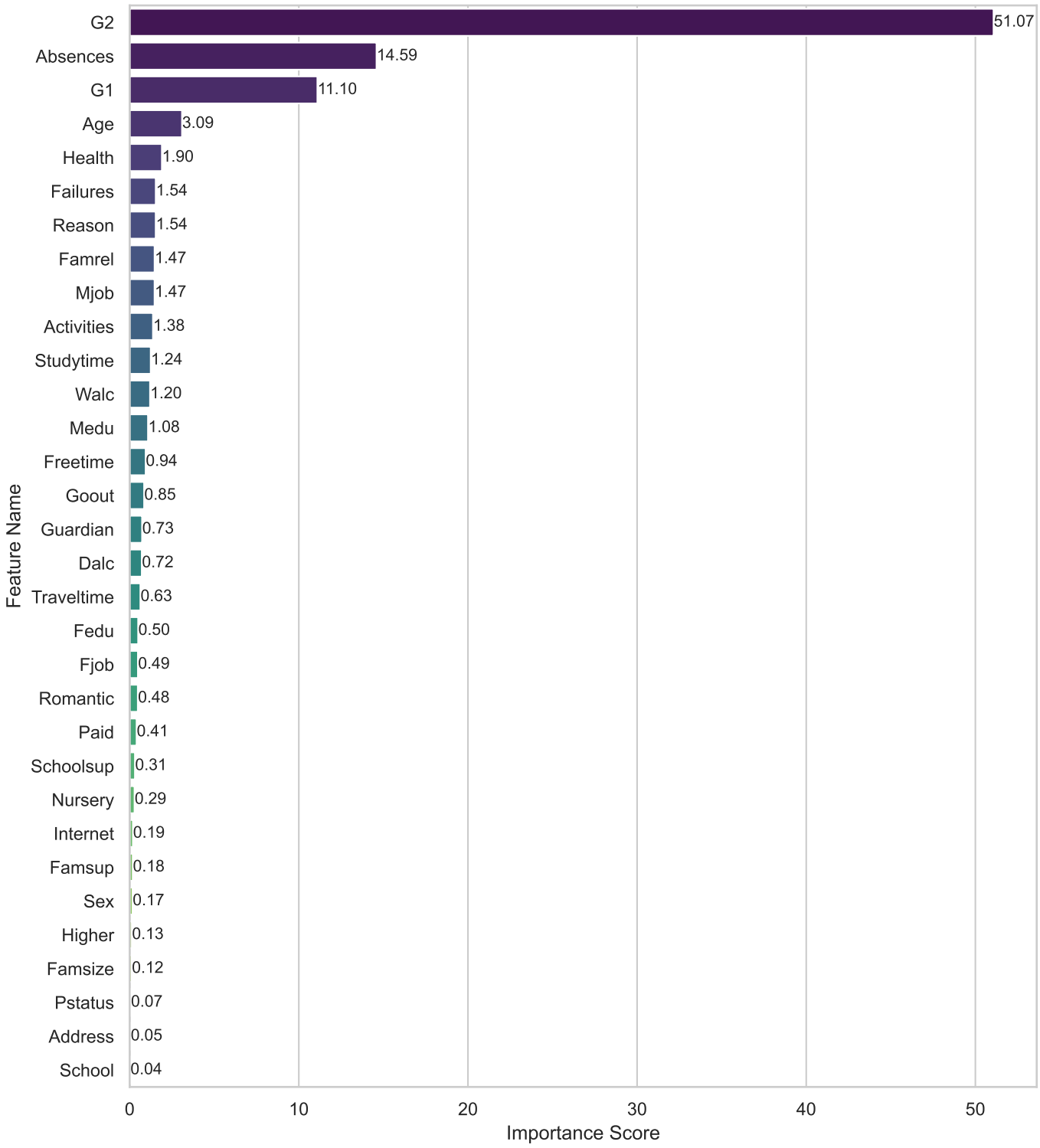

Figure 1. Feature importance based on the Mathematics course dataset.

The analysis of feature importance, together with the statistical characteristics of the educational datasets, underscores the crucial role of feature selection. The important attributes such as previous grades (G1 and G2) and school absences exhibit dominant influence in predicting final grades. Notably, as shown in Figure 1, the feature of the second-period grade (G2) achieves a very high importance score of 51.07%, aligning well with the dataset's logic that final grades largely reflect cumulative learning progress. Similarly, G1 (which indicates prior performance) and the number of absences (which indicates class participation) also contribute substantially to the prediction. In contrast, features such as school type, address, and

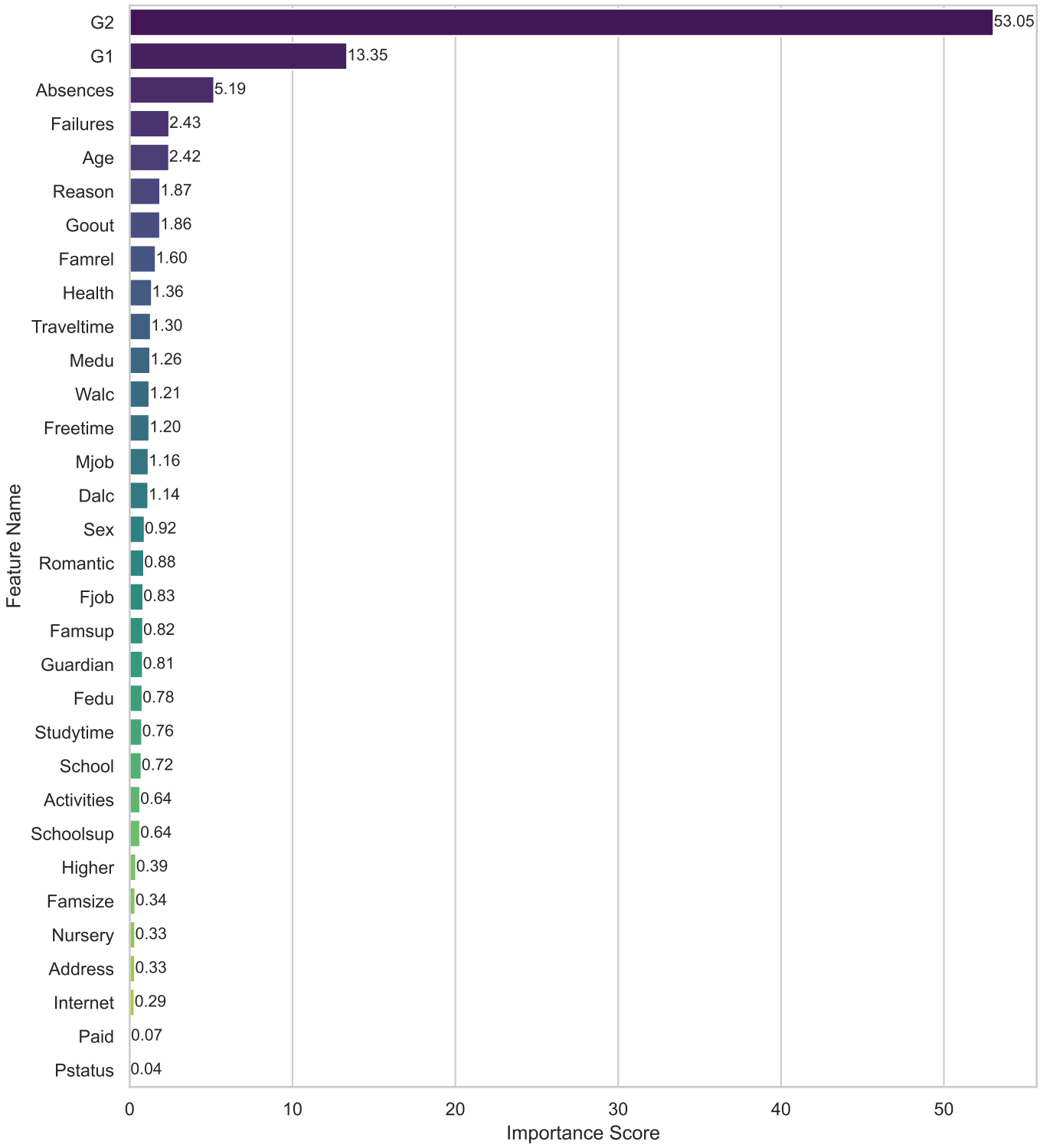

Figure 2. Feature importance based on the Portuguese Language course dataset

sex demonstrate minimal relevance. For example, the importance of school is only 0.04%, suggesting negligible differences between the two schools, while sex has an importance of 0.17%, indicating that gender plays a limited role in determining academic outcomes for these courses.

2.3 The proposed framework for student performance prediction

The proposed framework consists of three main stages:

(1) Feature importance extraction: Train the CatBoost model to compute and normalize feature importance scores.

(2) Cumulative threshold-based feature selection: Select features whose cumulative importance accounts for up to τ (95% is used in this paper) of the total.

(3) Model Evaluation with repeated 10-fold cross-validation: Assess the prediction performance of baseline models using 10-fold cross-validation repeated 10 times to ensure robustness and generalization.

Details of the CatBoost-based feature selection framework with repeated 10-fold cross-validation can also be found in Algorithm 1.

Algorithm 1: CatBoost-based feature selection framework with repeated 10-fold cross-validation

Input: Educational dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$;
Cumulative importance threshold $\tau = 0.95$;
Baseline models
$\mathcal{M} = \{\text{MLP}, \text{SVM}, \text{RF}, \text{XGBoost}, \text{CatBoost}\}$;
Number of folds $K = 10$;
Number of repetitions $R = 10$.
Output: Selected feature subset $\mathcal{F}_{\text{selected}}$;
Evaluation metrics of each model.

Step 1: Feature importance extraction
Train the CatBoost on $\mathcal{D}$ to compute the raw feature importance scores $\{I_j\}_{j=1}^{m}$ for each feature f_j;
Normalize feature importance so that all the feature importance scores are added up to 100%.

Step 2: Cumulative threshold-based feature selection
Sort all features $\{f_j\}$ in descending order;
Compute cumulative importance until the selected features satisfying $C_k \leq \tau$.

Step 3: Repeated 10-fold cross-validation evaluation
foreach $M \in \mathcal{M}$ **do**
  for $r = 1$ *to* R **do**
    Randomly shuffle $\mathcal{D}$;
    Partition $\mathcal{D}$ into 10 folds $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{10}\}$;
    for $k = 1$ *to* 10 **do**
      Training set $\mathcal{D}_{\text{train}} = \mathcal{D} \setminus \mathcal{D}_k$;
      Test set $\mathcal{D}_{\text{test}} = \mathcal{D}_k$;
      Train M using $\mathcal{F}_{\text{selected}}$ on $\mathcal{D}_{\text{train}}$;
      Predict $\hat{y}$ on $\mathcal{D}_{\text{test}}$;
      Compute evaluation metrics.
    end
  end
  Average the evaluation metrics for the given prediction model;
end
return $\mathcal{F}_{\text{selected}}$ and the final averaged metrics for each M

3 Experimental results and analysis

This section details the experimental setup to evaluate the effectiveness of the proposed CatBoost-based feature selection framework. First, we describe the educational datasets used in the experiments and the evaluation metrics adopted to quantify model performance. Subsequently, we present a detailed analysis of the experimental outcomes to validate the proposed framework.

3.1 Data description

To evaluate the performance of the proposed framework for student performance prediction, two benchmark educational datasets were sourced from the UCI Machine Learning Repository [22]. These datasets specifically capture student academic achievements in two distinct courses: Mathematics and Portuguese Language. The Mathematics course dataset includes 395 student samples, while the Portuguese Language dataset consists of 649 samples. Both datasets contain 32 input attributes representing various student-related factors and one output variable corresponding to the final exam grade. A comprehensive statistical summary of the attributes is presented in Table 1. Note that numeric features represented by codes (e.g., Medu and Failures) are explicitly defined to facilitate clear interpretation of their values. For binary and nominal features, discrete categories are separated using semicolons to ensure consistent formatting.

3.2 Evaluation metrics

The student performance prediction is assessed using four evaluation metrics, namely the Mean Absolute Error (MAE), Standard Deviation of prediction errors (SD), Root Mean Squared Error ($RMSE$), and Robustness (MAC) [23]. The objective of the evaluation is to minimize MAE, SD, and $RMSE$, while maximizing MAC, which measures the alignment between the true and predicted outcomes. The mathematical definitions of these metrics are as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{7}$$

Table 1. Statistical properties of the student performance dataset.

Feature	Symbol	Type	Values
School	School	Binary	Gabriel Pereira; Mousinho da Silveira
Sex	Sex	Binary	Female; Male
Age	Age	Numeric	15–22 (in years)
Address	Address	Binary	Urban; Rural
Family Size	Famsize	Binary	≤ 3 (small); > 3 (large)
Parent Status	Pstatus	Binary	Living together; Living apart
Mother's Education	Medu	Numeric	0 (none) – 4 (higher education)
Father's Education	Fedu	Numeric	0 (none) – 4 (higher education)
Father's Job	Fjob	Nominal	Teacher; Health care; Civil services (e.g., administrative, police); At home; Other
Mother's Job	Mjob	Nominal	Teacher; Health care; Civil services (e.g., administrative, police); At home; Other
School Choice Reason	Reason	Nominal	Close to home; School reputation; Course preference; Other
Guardian	Guardian	Nominal	Mother; Father; Other
Travel Time	Traveltime	Numeric	1 (< 15 min); 2 (15–30 min); 3 (30 min–1 hour); 4 (> 1 hour)
Study Time	Studytime	Numeric	1 (< 2 hours); 2 (2–5 hours); 3 (5–10 hours); 4 (> 10 hours)
Past Failures	Failures	Numeric	0 (no failures); 1 (1 failure); 2 (2 failures); 4 (≥ 3 failures)
School Support	Schoolsup	Binary	Yes; No
Family Support	Famsup	Binary	Yes; No
Paid Extra Classes	Paid	Binary	Yes; No
Extracurricular Activities	Activities	Binary	Yes; No
Nursery Attendance	Nursery	Binary	Yes; No
Pursue Higher Education	Higher	Binary	Yes; No
Home Internet Access	Internet	Binary	Yes; No
Romantic Relationship	Romantic	Binary	Yes; No
Family Relationship Quality	Famrel	Numeric	1 (very bad) – 5 (excellent)
Post-School Free Time	Freetime	Numeric	1 (very low) – 5 (very high)
Socializing with Friends	Goout	Numeric	1 (very low) – 5 (very high)
Weekday Alcohol Consumption	Dalc	Numeric	1 (very low) – 5 (very high)
Weekend Alcohol Consumption	Walc	Numeric	1 (very low) – 5 (very high)
Current Health Status	Health	Numeric	1 (very bad) – 5 (very good)
School Absences	Absences	Numeric	0–93 (number of days)
First-Period Grade	G1	Numeric	0–20 (grading scale: 0 = lowest, 20 = highest)
Second-Period Grade	G2	Numeric	0–20 (grading scale: 0 = lowest, 20 = highest)
Final Grade (Target)	G3	Numeric	0–20 (grading scale: 0 = lowest, 20 = highest)

Table 2. Experimental results for the prediction of Mathematics course in the final exam using the MLP, SVM, RF, XGBoost, CatBoost, and their feature selection-based variants.

Model	MAE	SD	$RMSE$	MAC
MLP	1.1666	1.6804	1.7070	0.9764
FS-MLP	**1.1173**	**1.6569**	**1.6751**	**0.9770**
SVM	1.1997	1.9412	1.9881	0.9673
FS-SVM	**1.1865**	**1.9233**	**1.9642**	**0.9685**
RF	0.9655	1.5279	1.5498	0.9800
FS-RF	**0.9525**	**1.5086**	**1.5284**	**0.9808**
XGBoost	1.0357	1.6577	1.6781	**0.9767**
FS-XGBoost	**1.0076**	**1.6404**	**1.6661**	0.9766
CatBoost	1.0790	1.5857	1.6073	0.9792
FS-CatBoost	**1.0444**	**1.5366**	**1.5594**	**0.9804**

Table 3. Experimental results for the prediction of Portuguese Language course in the final exam using the MLP, SVM, RF, XGBoost, CatBoost, and their feature selection-based variants.

Model	MAE	SD	$RMSE$	MAC
MLP	0.8174	1.2387	1.2520	0.9890
FS-MLP	**0.8089**	**1.2312**	**1.2434**	**0.9893**
SVM	0.8208	**1.3529**	**1.3658**	**0.9869**
FS-SVM	**0.8185**	1.3587	1.3731	0.9867
RF	0.8338	1.2889	1.2988	0.9884
FS-RF	**0.8265**	**1.2710**	**1.2814**	**0.9886**
XGBoost	**0.8893**	1.4006	1.4131	0.9863
FS-XGBoost	0.8957	**1.3878**	**1.3960**	**0.9866**
CatBoost	**0.8327**	**1.2895**	**1.3006**	**0.9882**
FS-CatBoost	0.8332	1.3035	1.3124	0.9881

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2}, \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{9}$$

$$MAC = \frac{\left(\boldsymbol{y}^T\hat{\boldsymbol{y}}\right)^2}{\left(\boldsymbol{y}^T\boldsymbol{y}\right)\left(\hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}}\right)}, \tag{10}$$

where n denotes the number of test samples, y_i represents the true value of the i-th sample, $\hat{y}_i$ denotes its predicted value, $e_i = \hat{y}_i - y_i$ is the corresponding prediction error, and $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$ is the average prediction error. Lower values of MAE, SD, and $RMSE$ indicate higher prediction accuracy and stability, while higher MAC values imply stronger robustness.

For model parameterization, all baseline and proposed models were configured using their default hyper-parameters (consistent with standard implementation practices to avoid bias from arbitrary tuning). The threshold for feature selection was set to 95%, indicating that only the top features contributing to 95% of the total cumulative importance were retained. To ensure the reliability and generalizability of results, all experiments were repeated 10 times with 10-fold cross-validation [24].

3.3 Experimental results

Table 2 presents the results for the prediction of final exam in Mathematics course using the baseline models (MLP, SVM, RF, XGBoost, CatBoost) and their feature selection-enhanced variants (FS-MLP, FS-SVM, FS-RF, FS-XGBoost, FS-CatBoost). The best results between with and without the use of FS are highlighted in bold. As we can see from the table, the proposed CatBoost-based feature selection framework consistently improves the prediction performance across all baseline models except the XGBoost in terms of MAC with comparative results (0.9767 and 0.9766). Lower values of MAE, $RMSE$, and SD indicate higher accuracy and stability, while higher MAC values confirm stronger prediction robustness. For example, the FS-CatBoost achieves lower MAE (1.0444) and $RMSE$ (1.5594) than the baseline CatBoost (1.0790 and 1.6073, respectively). Similarly, the FS-RF yields the best results overall ($MAE = 0.9525$ and $RMSE = 1.5284$), highlighting the effectiveness of feature selection when combined with ensemble learning. Among all the results, the FS-RF has the best performance in terms of MAE, $RMSE$, SD, and MAC, with the values of 0.9525, 1.5086, 1.5284, and 0.9808, respectively.

Table 3 shows the prediction performance for the Portuguese Language course using baseline models and their feature selection-enhanced variants. Best results between models with and without feature selection are highlighted in bold. The results show that the feature selection framework generally improves the prediction performance of baseline models in most cases. For example, the MLP and RF are improved in terms of all metrics used. In addition, the XGBoost is also improved in terms of SD, $RMSE$, and MAC. Although SVM and CatBoost do not improve, their performance remains comparable to the baseline. This indicates that with fewer features, the models can still achieve high prediction performance.

Furthermore, Figures 3 and 4 visualize the experimental results reported in Tables 2 and 3, respectively. Each figure presents the comparison

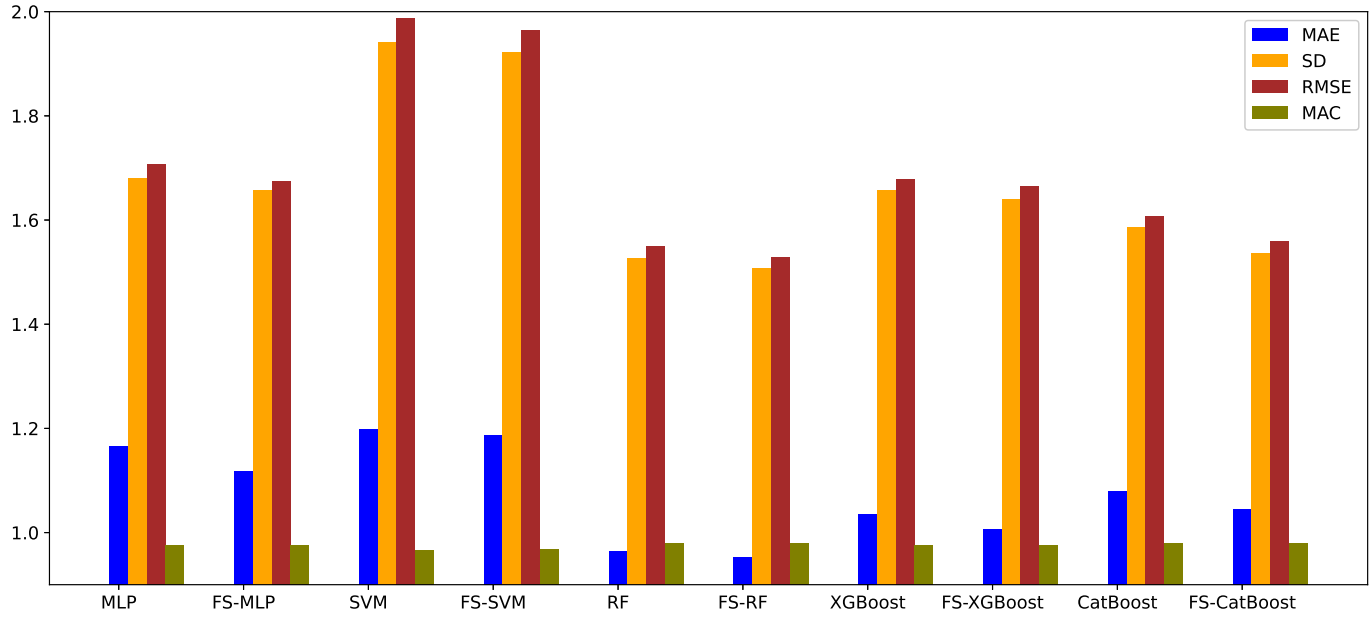

Figure 3. Experimental results based on the results in Table 2.

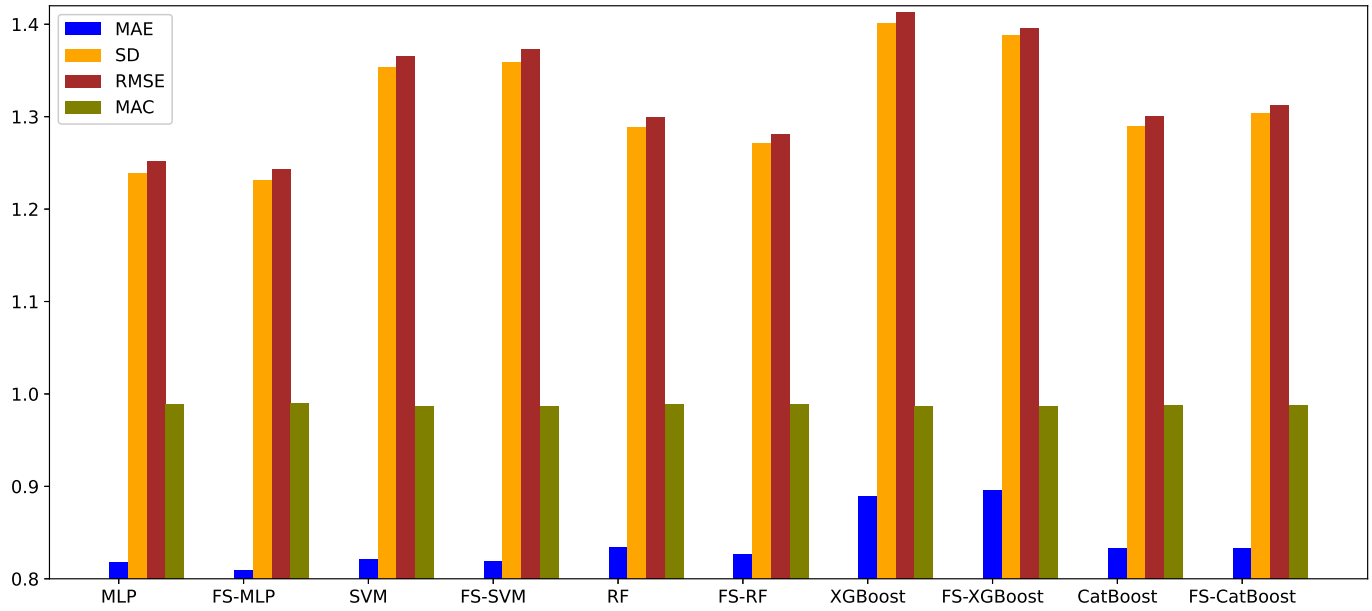

Figure 4. Experimental results based on the results in Table 3.

of the baseline models (MLP, SVM, RF, XGBoost, CatBoost) and their feature selection-enhanced variants (FS-MLP, FS-SVM, FS-RF, FS-XGBoost, FS-CatBoost) across four metrics (MAE, SD, $RMSE$, and MAC).

4 Conclusion

This study presented a CatBoost-based feature selection framework for student performance prediction. This framework utilized the gradient boosting with feature importance analysis to automatically identify the most influential features for student performance prediction. Through extensive experiments on two benchmark educational datasets (Mathematics and Portuguese Language courses), in most cases, the proposed approach demonstrated superior prediction performance compared to baseline models. The results confirmed that the selective inclusion of high-importance features is effective to enhance student performance prediction.

Nevertheless, the proposed framework has some limitations. First, it focuses primarily on tabular data and does not consider video or audio data. Second, while the feature selection process is effective, it

relies exclusively on CatBoost's internal importance metrics. Future research will address these limitations and extend the framework in three key directions. First, we will collect and incorporate diverse types of educational datasets (e.g., video and voice data) to improve the student performance models. Second, we will explore additional feature selection methods to enhance the feature subset quality. Finally, we will investigate the integration of explainable artificial intelligence techniques [25] to boost model transparency.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported by the Fujian Provincial Young and Middle-aged Teachers' Educational Research Project (Science and Technology Category), China under Grant JAT241390.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Bradley, V. M. (2021). Learning Management System (LMS) use with online instruction. *International Journal of Technology in Education*, *4*(1), 68-92.

[2] Heil, J., & Ifenthaler, D. (2023). Online Assessment in Higher Education: A Systematic Review. *Online Learning*, *27*(1), 187-218.

[3] Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, *122*, 119-135. [CrossRef]

[4] Rabelo, A., Rodrigues, M. W., Nobre, C., Isotani, S., & Zárate, L. (2024). Educational data mining and learning analytics: A review of educational management in e-learning. *Information Discovery and Delivery*, *52*(2), 149-163. [CrossRef]

[5] Kalita, E., Oyelere, S. S., Gaftandzhieva, S., Rajesh, K. N., Jagatheesaperumal, S. K., Mohamed, A., ... & Ali, T. (2025). Educational data mining: A 10-year review. *Discover Computing*, *28*(1), 81. [CrossRef]

[6] Fan, Z., Gou, J., & Wang, C. (2025). An error complementarity-based iterative learning approach via categorical boosting for student performance prediction. *Engineering Applications of Artificial Intelligence*, *161*, 112192. [CrossRef]

[7] Cao, W., & Mai, N. (2025). Predictive Analytics for Student Success: AI-Driven Early Warning Systems and Intervention Strategies for Educational Risk Management. *Educational Research and Human Development*, *2*(2), 36-48. [CrossRef]

[8] Hemdanou, A. L., Sefian, M. L., Achtoun, Y., & Tahiri, I. (2024). Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models. *Computers and Education: Artificial Intelligence*, *7*, 100301. [CrossRef]

[9] Öz, E., Bulut, O., Cellat, Z. F., & Yürekli, H. (2025). Stacking: An ensemble learning approach to predict student performance in PISA 2022. *Education and Information Technologies*, *30*(6), 7753-7779. [CrossRef]

[10] Bañeres, D., Rodríguez-González, M. E., Guerrero-Roldán, A. E., & Cortadas, P. (2023). An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education*, *20*(1), 3. [CrossRef]

[11] Maiya, A. K., & Aithal, P. S., (2023). A Review based Research Topic Identification on How to Improve the Quality Services of Higher Education Institutions in Academic, Administrative, and Research Areas. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *8*(3), 103-153. [CrossRef]

[12] Ilani, M. A., & Banad, Y. M. (2024). EDMNet: unveiling the power of machine learning in regression modeling of powder mixed-EDM. *The International Journal of Advanced Manufacturing Technology*, *135*(5), 2555-2570. [CrossRef]

[13] Hong, Y. Z., Rani, M. N. A., Radzuan, N. F. M., Yen, L. H., & Nagalingam, S. A. R. A. S. V. A. T. H. I. (2024). An early warning system for students at risk using supervised machine learning. *Journal of Engineering Science and Technology*, *19*(1), 131-139.

[14] Cheng, B., Liu, Y., & Jia, Y. (2024). Evaluation of students' performance during the academic period using the XG-Boost Classifier-Enhanced AEO hybrid model. *Expert Systems with Applications*, *238*, 122136. [CrossRef]

[15] Fan, Z., Gou, J., & Weng, S. (2025). Complementary CatBoost based on residual error for student performance prediction. *Pattern Recognition*, *161*, 111265. [CrossRef]

[16] Zaffar, M., Hashmani, M. A., Savita, K. S., & Khan, S. A. (2021). A review on feature selection methods for improving the performance of classification in educational data mining. *International Journal of Information Technology and Management*, *20*(1-2), 110-131. [CrossRef]

[17] Song, Y. W., Wang, J. S., Qi, Y. L., Wang, Y. C., Li, S., Song, H. M., & Shang-Guan, Y. P. (2025). PF-PSS:

a double-layer parallel embedded feature selection method for cancer gene expression data. *Journal of Big Data, 12*(1), 136. [CrossRef]

[18] Huang, Y., Chen, G., Gou, J., Fan, Z., & Liao, Y. (2025). A hybrid feature selection and aggregation strategy-based stacking ensemble technique for network intrusion detection. *Applied Intelligence, 55*(1), 28. [CrossRef]

[19] Zhou, H., Zhang, J., Zhou, Y., Guo, X., & Ma, Y. (2021). A feature selection algorithm of decision tree based on feature weight. *Expert Systems with Applications, 164*, 113842. [CrossRef]

[20] Fan, Z., Gou, J., & Weng, S. (2024). A feature importance-based multi-layer catboost for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering, 36*(11), 5495-5507. [CrossRef]

[21] 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]

[22] Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. [CrossRef]

[23] Fan, Z., Gou, J., & Weng, S. (2024). An unbiased fuzzy weighted relative error support vector machine for reverse prediction of concrete components. *IEEE Transactions on Artificial Intelligence, 5*(9), 4574-4584. [CrossRef]

[24] Wong, T. T., & Yeh, P. Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering, 32*(8), 1586-1594. [CrossRef]

[25] Guleria, P., & Sood, M. (2023). Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Education and Information Technologies, 28*(1), 1081-1116. [CrossRef]

[26] Bosch, N. (2021). AutoML feature engineering for student modeling yields high accuracy, but limited interpretability. *Journal of Educational Data Mining, 13*(2), 55-79. [CrossRef]

Shaoyuan Weng received the master's degree from the College of Computer Science and Technology, Huaqiao University, Xiamen, China, in 2021. She is currently a lecturer with the Xiamen Institute of Software Technology. Her main research interests include in the areas of analysis of data mining and collaborative task. (Email: wengshaoyuany@163.com)

Yuanyuan Zheng received the master's degree from the College of Computer Science and Technology, Huaqiao University, Xiamen, China, in 2021. She is currently a lecturer with the Xiamen Institute of Software Technology. Her main research interests include in the areas of analysis of data mining and collaborative task. (Email: zhengyuanyuan1025@163.com)

Chao Zhang received his M.S. degree from the School of Computer Science and Technology, Harbin University of Science and Technology, in 2018. He is currently a lecturer at Xiamen Institute of Software Technology. His research interests include data analysis and data mining (Email: zhangchao636@163.com)

Zimeng Liu is currently the Director of the Teaching and Research Section for Big Data Technology and Cloud Computing Technology Application at the Xiamen Institute of Software Technology. He is also a Senior Engineer (Information System Project Manager), Senior Technician, and Dual-Qualified Teacher. (Email: liuzimenglzm@163.com)