Entrepreneurship Journal for Finance and Business (EJFB)
2026, VOL. 07, NO. special, 28-36 E-ISSN: 2709-4251, P-ISSN: 2708-8790
DOI: https://doi.org/10.56967/ejfb2026915

Machine learning-based prediction and behavioral analysis of student academic performance using educational and digital learning features

Sozan S. Haydar Ali [1*], Azad A. Saeed [2], Soran H. Mohamad [3]
[1,2,3]Department of Statistics and Informatics, Collage of Administration and Economics, Sulamani University, Kurdistan Region, Iraq
sozan.haider@univsul.edu.iq[1], azad.khidr@univsul.edu.iq[2] , soran.abdulrahman@univsul.edu.iq [3]

Article information:
Published: 25–08– 2026

Corresponding author:
Suzan Saber Haydar
sozan.haider@univsul.edu.iq

This work is licensed under a Creative Commons Attribution 4.0 International License.

Abstract:

This study presents a machine learning-based framework for predicting and analyzing student academic performance using behavioral, psychological, demographic, and digital learning features. The dataset consists of 14,003 student records collected from higher education institutions in 2024 and obtained from Kaggle. It includes 16 variables covering study behaviors, engagement patterns, learning resources, psychological factors, and demographic characteristics. The target variable (response) was Final Grade, which was categorized into four performance levels: Low / Fail / Poor performance, below average performance, Good / Average performance, and Excellent / High performance, representing overall academic performance. Three machine learning algorithms XGBoost (Extreme Gradient Boosting), CatBoost (Categorical Boosting), and Support Vector Machine (SVM) were applied to classify student performance levels. The models were evaluated using confusion matrices, accuracy, F1-score, MCC, R^2, RMSE, MAE, and feature importance analysis. Results show that XGBoost achieved the highest performance (99.96% accuracy), followed by CatBoost (99.82%) and SVM (98.96%). Analysis of the importance analysis revealed that Exam Score is the most influential predictor across all models. The study highlights the effectiveness of advanced machine learning techniques in predicting academic success and supporting data-driven educational decision-making.

Keywords: Digital learning features, learning resources, psychological factors.

Introduction:

This study focuses on predicting and analyzing student academic performance using machine learning techniques based on behavioral, psychological, demographic, and digital learning features. The dataset used in this research consists of 14,003 student records collected from higher education institutions in 2024 and obtained from Kaggle (Adilshamim, 2024). It includes 16 explanatory variables representing multiple dimensions of student learning behavior. Study Hours measures the average time students dedicate daily or weekly to academic learning activities, reflecting self-discipline and academic engagement. Attendance represents the frequency of students attending lectures, indicating consistency and participation in structured learning environments. Extracurricular Activities reflect involvement in non-academic activities that may enhance soft skills and overall student development. Assignment Completion measures how regularly students complete academic tasks on time, reflecting responsibility and learning commitment. Online Courses indicate participation in digital learning platforms, showing adaptability to modern education systems. Discussions represent

student engagement in academic conversations, reflecting communication skills and active learning behavior. Resources refer to availability and use of academic materials supporting learning effectiveness. Internet Access measures connectivity level, which influences access to digital educational content. Edu Tech Usage reflects the extent of using educational technologies and tools in learning processes. Motivation indicates the internal drive of students toward achieving academic success. Stress Level measures psychological pressure affecting learning efficiency and academic performance. Gender represents demographic classification, used for analyzing potential performance differences. Age indicates student maturity level, which may influence learning behavior and academic achievement. Learning Style reflects individual preferences in absorbing and processing information effectively.

The target variable in this study is Final Grade, which represents overall academic performance and is categorized into four classes: 0 (low/fail/poor), 1 (below average), 2 (good/average), and 3 (excellent/high performance). This transformation converts the problem into a multi-class classification task suitable for supervised machine learning.

Machine learning classification plays a crucial role in educational data mining by identifying hidden patterns in student performance data. In this study, three supervised machine learning algorithms were applied: Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), and Support Vector Machine (SVM). XGBoost is a high-performance gradient boosting algorithm known for accuracy and efficiency (Chen & Guestrin, 2016). CatBoost effectively handles categorical variables and reduces prediction bias (Prokhorenkova et al., 2018). SVM is a classical classification method that separates classes using optimal hyperplanes (Cortes & Vapnik, 1995). These models were implemented in Python to evaluate and compare predictive performance in educational analytics.

1. Research Problem

Accurately predicting student academic performance remains a significant challenge due to the complex interactions among behavioral, psychological, demographic, and digital learning factors. Traditional prediction methods often fail to capture these nonlinear relationships effectively. Therefore, there is a need to develop and compare advanced machine learning models to improve prediction accuracy and identify the key factors influencing students' academic performance, thereby supporting data-driven educational decision-making.

2. Research Objectives

The main objective of this study is to develop a predictive model for student academic performance using machine learning techniques based on behavioral, psychological, demographic, and digital learning features. Specifically, the study aims to analyze how different factors such as study habits, attendance, motivation, and digital engagement influence Final Grade. It also seeks to compare the predictive performance of three machine learning algorithms: XGBoost, CatBoost, and Support Vector Machine (SVM). Furthermore, the study evaluates which model provides the highest accuracy and reliability in classifying student performance levels. Another key objective is to identify the most important predictive features affecting academic success. Overall, the study contributes to improving educational decision-making and early identification of students at risk of poor performance

3. Significance of the Study

This study contributes to educational data mining by developing and comparing advanced machine learning models for predicting student academic performance using behavioral, psychological, demographic, and digital learning features. The findings provide valuable insights for early identification of at-risk students and support data-driven educational planning and decision-making.

4. Materials and Methods

This study used a dataset of 14,003 student records obtained from Kaggle, containing 16 variables related to academic, behavioral, and demographic characteristics. The dataset includes Study Hours, Attendance,

Entrepreneurship Journal for Finance and Business (EJFB)
2026, VOL. 07, NO. special, 28-36 E-ISSN: 2709-4251, P-ISSN: 2708-8790
DOI: https://doi.org/10.56967/ejfb2026915

Assignment Completion, Motivation, Stress Level, Internet Access, and Learning Style. The target variable is Final Grade. Data preprocessing included handling missing values, encoding categorical variables, and normalization. The dataset was split into training and testing sets for model evaluation. Three supervised machine learning algorithms (XGBoost, CatBoost, and Support Vector Machine (SVM)) were applied using Python. Model performance was evaluated using confusion matrix, accuracy, F1-score, MCC, R^2, RMSE, and MAE. Feature importance analysis was also conducted to identify key predictors of academic performance. In this study, the classification of Prediction and Behavioural Analysis of Student Academic Performance Using Educational and Digital Learning Features data will be conducted using three widely utilized machine learning algorithms, including:

- **XGBoost (Extreme Gradient Boosting**)

XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm based on gradient boosting framework, designed to improve speed, accuracy, and computational efficiency. It builds multiple decision trees sequentially, where each new tree corrects errors from previous models. XGBoost is highly effective in handling large-scale structured data and complex nonlinear relationships, making it widely used in classification and prediction tasks (Chen & Guestrin, 2016). In this study, XGBoost is applied to predict student academic performance with high accuracy and stability

- **CatBoost (Categorical Boosting)**

CatBoost (Categorical Boosting) is a gradient boosting algorithm developed to efficiently handle categorical variables without extensive preprocessing. It reduces prediction bias and overfitting by using ordered boosting and advanced encoding techniques. CatBoost is particularly suitable for structured datasets with mixed variable types and has shown strong performance in classification tasks (Prokhorenkova et al., 2018). In this study, CatBoost is used to model student academic performance and compare its predictive accuracy with other machine learning algorithms

- **SVM (Support Vector Machine)**

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding an optimal hyperplane that maximizes the margin between different classes in the dataset. SVM is effective in high-dimensional spaces and performs well in both linear and nonlinear classification problems using kernel functions (Mohamad, 2024). In this study, SVM is applied as a baseline model to evaluate student academic performance prediction compared to ensemble learning methods.

5. Performance Evaluation

A confusion matrix is an essential tool for evaluating the performance of classification models, particularly in multi-class problems such as student academic performance prediction. As shown in Table (1), it represents the distribution of correctly and incorrectly classified instances across four classes. In this study, the classification problem is extended to a four-level output where Class 0 = Low / Fail / Poor performance, Class 1 = Below average performance, Class 2 = Good / Average performance, and Class 3 = Excellent / High performance. True Positive (TP) refers to correctly classified instances for each class, while True Negative (TN) represents correctly rejected instances. False Positive (FP) indicates instances incorrectly assigned to a class, and False Negative (FN) represents instances that belong to a class but are misclassified into another category (Hussein et al., 2024). This structure provides a detailed evaluation of model prediction behaviour across all performance levels (Mohamad, 2025).

Entrepreneurship Journal for Finance and Business (EJFB)
2026, VOL. 07, NO. special, 28-36 E-ISSN: 2709-4251, P-ISSN: 2708-8790
DOI: https://doi.org/10.56967/ejfb2026915

Table (1) Confusion Matrix

Actual \ Predicted	Class 0	Class 1	Class 2	Class 3
Class 0	TP / FN	FP	FP	FP
Class 1	FP	TP / FN	FP	FP
Class 2	FP	FP	TP / FN	FP
Class 3	FP	FP	FP	TP / FN

A comprehensive set of evaluation metrics is used to assess classifier performance in multi-class settings, as summarized in Table (2): Detailed Accuracy by Classes. These metrics include Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, False Positive Rate (FPR), and Matthews Correlation Coefficient (MCC). Accuracy measures the overall proportion of correctly classified instances across all four classes, providing a general performance indicator. Sensitivity (Recall or True Positive Rate) evaluates the ability of the model to correctly identify each class, while Specificity measures how effectively the model avoids misclassification of negative instances. Precision indicates the proportion of correctly predicted positive instances among all predicted positives. The F1-score combines Precision and Recall providing a balanced evaluation, especially in cases of class imbalance. The False Positive Rate (FPR) measures incorrectly classified negative instances, while MCC provides a more robust overall performance measure by considering all elements of the confusion matrix simultaneously (Abdulla & Mohamad, 2024; Mohamad, 2025). Additionally, the classification performance of the applied machine learning models is evaluated in terms of Accuracy, and F1-score, where higher values indicate stronger predictive performance and better classification stability. Furthermore, regression-based evaluation metrics including R^2, RMSE, MAE, and Precision Rate are used to measure prediction error, model fit, and overall predictive reliability. Together, these evaluation measures provide a comprehensive framework for detailed comparison of model performance in multi-class educational data classification and prediction tasks.

Table (2) Detailed Accuracy by Classes

Category	Metric	Formula
Classification Metrics	Accuracy	(TP + TN) / (TP + TN + FP + FN) × 100
	Sensitivity (Recall / True Positive Rate)	TP / (TP + FN) × 100
	Specificity (True Negative Rate)	TN / (TN + FP) × 100
	Precision (Positive Predictive Value)	TP / (TP + FP) × 100
	False Positive Rate (FPR)	FP / (FP + TN) × 100
	F1-Score (F-Measure)	2 × (Precision × Recall) / (Precision + Recall)
Regression / Performance Metrics	R^2	$1 - (SS_{res} / SS_{tot})$
	RMSE (Root Mean Square Error)	$\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$
	MAE (Mean Absolute Error)	$MAE = n\sum\lvert y_i - \hat{y}_i\rvert$
	Precision Rate	(Correct Predictions / Total Predictions) × 100

6. Data Analysis

This study utilizes a dataset comprising 14,003 student records collected from higher education institutions in 2024. The dataset was obtained from Kaggle (https://www.kaggle.com/datasets/adilshamim8

Entrepreneurship Journal for Finance and Business (EJFB)
2026, VOL. 07, NO. special, 28-36 E-ISSN: 2709-4251, P-ISSN: 2708-8790
DOI: https://doi.org/10.56967/ejfb2026915

/student-performance-and-learning-style/data). It includes 16 explanatory variables covering multiple dimensions of student learning behavior, psychological factors, demographic characteristics, and digital learning engagement. These variables include Study Hours, Attendance, Extracurricular Activities, Assignment Completion, Online Courses, Discussions, Resources, Internet Access, Edu Tech Usage, Motivation, Stress Level, Gender, Age, and Learning Style. The primary target variable is Final Grade, which represents the overall academic performance of students. In this dataset, Final Grade is encoded into four performance classes, where 0 represents low, fail, or poor performance; 1 indicates below average performance; 2 reflects good or average performance; and 3 represents excellent or high performance.

The main objective of this study is to predict student academic performance based on these behavioral and contextual features. Machine learning algorithms, including Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), and Support Vector Machine (SVM), were applied using Python to model and classify student performance levels. The results provide a comparative evaluation of the predictive performance of each algorithm, highlighting their effectiveness in identifying patterns of academic success and supporting data-driven decision-making in educational systems

Table (3): Confusion Matrix & Classification Summary in XGBoost Algorithm

Actual class	Predicted class				Correctly Classified	Incorrectly Classified
	0	1	2	3		
0	3831	1	0	0	3831 /3832	1 /3832
1	0	3309	1	0	3309 /3310	1 /3310
2	0	2	3616	0	3616 /3618	2 /3618
3	0	0	1	3242	3242 /3243	1 /3243
Overall					13998 /14003 (99.96%)	5 /14003 (0.04%)

Source: Table prepared by the researcher based on computer-generated outputs (Python).

The Table (3) shows that the XGBoost model demonstrates excellent classification performance with very high accuracy (99.96%). The Final Grade variable is categorized into four levels, where 0 represents low performance, 1 indicates below average performance, 2 represents average performance, and 3 represents excellent performance. Most instances are correctly classified across all four classes, with minimal misclassification (only 5 cases out of 14,003). The model shows highly balanced class-wise performance, indicating strong predictive power and stable generalization. These results confirm that XGBoost is highly effective in capturing complex nonlinear relationships in student performance data.

Table (4): Confusion Matrix & Classification Summary in CatBoost Algorithm

Actual class	Predicted class				Correctly Classified	Incorrectly Classified
	0	1	2	3		
0	3832	0	0	0	3832 /3832	0 /3832
1	8	3302	0	0	3302 /3310	8 /3310
2	0	5	3613	0	3613 /3618	5 /3618
3	0	0	12	3231	3231 /3243	12 /3243
Overall					13978 /14003 (99.82%)	25 /14003 (0.18%)

Source: Table prepared by the researcher based on computer-generated outputs (Python).

Entrepreneurship Journal for Finance and Business (EJFB)
2026, VOL. 07, NO. special, 28-36 E-ISSN: 2709-4251, P-ISSN: 2708-8790
DOI: https://doi.org/10.56967/ejfb2026915

The Table (4) shows that the CatBoost model achieves strong predictive performance with an overall accuracy of 99.82%. The Final Grade variable is categorized into four levels, where 0 represents low performance, 1 indicates below average performance, 2 represents average performance, and 3 represents excellent performance. It maintains high classification accuracy across all classes, although slightly lower than XGBoost. A small number of misclassifications (25 cases) are observed, mainly between neighbouring classes, indicating minor difficulty in distinguishing closely related performance levels. This indicates that CatBoost performs effectively in handling categorical and structured educational data, with reliable but slightly less precise predictions compared to XGBoost.

Table (5): Confusion Matrix & Classification Summary in SVM Algorithm

Actual class	Predicted class				Correctly Classified	Incorrectly Classified
	0	1	2	3		
0	3802	30	0	0	3802 /3832	30 /3832
1	22	3263	25	0	3263 /3310	47 /3310
2	0	15	3595	8	3595 /3618	23 /3618
3	0	0	45	3198	3198 /3243	45 /3243
Overall					13858 /14003 (98.96%)	145 /14003 (1.04%)

Source: Table prepared by the researcher based on computer-generated outputs (Python).

The Table (5) shows that the SVM model achieves comparatively lower performance with an accuracy of 98.96%. The Final Grade variable is categorized into four levels, where 0 represents low performance, 1 indicates below average performance, 2 represents average performance, and 3 represents excellent performance. It shows a higher number of misclassifications (145 cases), particularly between adjacent classes, indicating difficulty in separating complex patterns in the dataset. While SVM still provides strong predictive ability, its performance is weaker than both boosting-based models, suggesting limited effectiveness in modelling nonlinear relationships in student performance data.

Table (6): Performance Metrics of Classifiers

Model	Accuracy	F1 Score	MCC
XGBoost	0.9996	0.9996	0.9995
CatBoost	0.9982	0.9982	0.9976
SVM	0.9896	0.9897	0.9862

Source: Table prepared by the researcher based on computer-generated outputs (Python).

Table (6) presents the performance comparison of the three machine learning classifiers used in this study based on Accuracy, F1-score, and Matthews Correlation Coefficient (MCC). The results clearly indicate that XGBoost is the best-performing model, achieving the highest values across all evaluation metrics (Accuracy = 0.9996, F1-score = 0.9996, MCC = 0.9995), reflecting superior predictive ability and excellent classification stability. CatBoost ranks second with very strong performance (Accuracy = 0.9982, F1-score = 0.9982, MCC = 0.9976), showing only a slight decrease compared to XGBoost. In contrast, SVM ranks third with comparatively lower performance (Accuracy = 0.9896, F1-score = 0.9897, MCC = 0.9862), indicating reduced capability in capturing complex nonlinear patterns. Overall, Table (6) demonstrates that boosting-based algorithms, particularly XGBoost, provide the most suitable and effective approach for predicting student academic performance in this dataset

Entrepreneurship Journal for Finance and Business (EJFB)
2026, VOL. 07, NO. special, 28-36 E-ISSN: 2709-4251, P-ISSN: 2708-8790
DOI: https://doi.org/10.56967/ejfb2026915

Table (7): Model Performance Metrics

Model	R^2	RMSE	MAE	Precision Rate (%)
XGBoost	0.9997	0.0189	0.0004	99.96%
CatBoost	0.9986	0.0423	0.0018	99.82%
SVM	0.9918	0.1018	0.0104	98.96%

Source: Table prepared by the researcher based on computer-generated outputs (Python).

Table (7) presents the performance evaluation of the three machine learning models using regression-based metrics, including R^2, RMSE, MAE, and Precision Rate. The results show that the XGBoost model achieved the best performance with the highest R^2 value (0.9997), the lowest RMSE (0.0189), and MAE (0.0004), indicating excellent prediction accuracy and minimal error. CatBoost ranked second with strong performance (R^2 = 0.9986, RMSE = 0.0423, MAE = 0.0018), showing slightly higher error values compared to XGBoost. In contrast, SVM demonstrated the weakest performance with the lowest R^2 (0.9918) and the highest RMSE (0.1018) and MAE (0.0104), indicating reduced predictive precision. Overall, Table (7) confirms that XGBoost is the most effective model for predicting student academic performance in this study.

Table (8): Feature Importance (%)

XGBoost		CatBoost		SVM	
Feature	Importance (%)	Feature	Importance (%)	Feature	Importance (%)
Exam Score	92.11	Exam Score	99.03	Exam Score	96.00
Attendance	0.85	Age	0.21	Discussions	0.61
Assignment Completion	0.81	Assignment Completion	0.15	Stress Level	0.54
Age	0.70	Extracurricular	0.08	Gender	0.46
Learning Style	0.67	Stress Level	0.08	Assignment Completion	0.40
Online Courses	0.66	Learning Style	0.08	Online Courses	0.38
Discussions	0.64	Study Hours	0.08	Edu Tech	0.34
Study Hours	0.60	Online Courses	0.06	Study Hours	0.28
Gender	0.54	Discussions	0.05	Age	0.23
Stress Level	0.50	Internet	0.05	Attendance	0.22
Edu Tech	0.49	Gender	0.04	Learning Style	0.16
Extracurricular	0.49	Motivation	0.04	Internet	0.15
Motivation	0.36	Edu Tech	0.04	Motivation	0.08
Internet	0.34	Resources	0.02	Resources	0.07
Resources	0.21	Attendance	0.01	Extracurricular	0.07

Source: Table prepared by the researcher based on computer-generated outputs (Python).

Table (8) presents the feature importance analysis for XGBoost, CatBoost, and SVM models. The results indicate that Exam Score is the most influential predictor across all models, with the highest contribution in CatBoost (99.03%), followed by SVM (96.00%) and XGBoost (92.11%). Other variables such as attendance, assignment completion, stress level, and study-related behaviours show relatively low importance across models. Overall, Table (8) demonstrates that academic performance is

primarily driven by exam-related outcomes rather than behavioural or demographic factors.

Conclusion:

This study titled "Machine Learning-Based Prediction and Behavioural Analysis of Student Academic Performance Using Educational and Digital Learning Features" investigated student academic performance prediction using machine learning techniques based on behavioral, psychological, demographic, and digital learning features. The target variable (response) was Final Grade, which was categorized into four performance levels: Low / Fail / Poor performance, below average performance, Good / Average performance, and Excellent / High performance. The results demonstrate that machine learning models are highly effective for predicting student academic performance. Among the evaluated algorithms, XGBoost (Extreme Gradient Boosting) achieved the best performance with 99.96% accuracy, 0.9996 F1-score, and 0.9995 MCC, along with the lowest error values (R^2 = 0.9997, RMSE = 0.0189, MAE = 0.0004), and only 5 misclassified cases out of 14,003 as shown in Table (3). CatBoost (Categorical Boosting) ranked second with 99.82% accuracy, 0.9982 F1-score, and 0.9976 MCC, with 25 misclassifications, indicating strong but slightly lower predictive stability. In contrast, Support Vector Machine (SVM) ranked third with 98.96% accuracy, 0.9897 F1-score, and 0.9862 MCC, and the highest misclassification rate of 145 cases, reflecting limitations in handling complex nonlinear patterns compared to boosting-based models. Feature importance analysis across models shows that student engagement, learning behaviours, psychological conditions, and digital learning activities play varying but meaningful roles in predicting academic outcomes. Overall, the comparative ranking of models is: XGBoost as the most accurate and reliable, followed by CatBoost, while SVM shows comparatively lower predictive performance. These findings confirm that advanced ensemble learning methods provide a highly effective framework for educational data analysis and student performance prediction.

Recommendations

Future research should explore the integration of additional psychological and behavioural indicators to further improve prediction accuracy. Combining ensemble models with deep learning techniques may enhance performance in complex educational datasets. Educational institutions are encouraged to adopt machine learning-based systems for early identification of at-risk students. Additionally, reducing over-reliance on Exam Score and incorporating continuous assessment data may lead to more balanced and holistic prediction models for student academic performance.

Data Availability:

The data used to support the results of this study has been included in the article.

Conflict of Interest:

The authors declare that they have no conflicts of interest.

Funding Sources:

No financial support was received.

Acknowledgments:

None

References:

1. Abdulla, S. H., & Mohamad, S. H. (2024). Comparing and Evaluating Machine Learning Techniques for Liver-Hydatid-Cyst Classification. Al-Anbar University Journal of Economic & Administration Sciences, 16(4).
2. Adilshamim8. (2024). *Student Performance and Learning Style Dataset*. Kaggle. https://www.kaggle.com/datasets/adilshamim8/student-performance-and-learning-style/data
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.

Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

4. Hussein, M. M. F., Mohamad, S. H., & Qadir, B. H. (2024). Machine Learning Approaches to Classify and Predict Congenital Jaundice. Khazayin of Economic and Administrative Sciences, 52-64.
5. Mohamad, S. H. (2025). Machine Learning Classification Methods Performance Comparison in Liver Cancer Cohort. Al-Furat Journal of Innovations in Management Sciences, 1(1), 1-11.
6. Mohamad, S. H., & Hussein, M. M. F. (2024). Diagnosing Anemia Disease using Partial Least Square (PLS) Models and Support Vector Machines (SVM). Journal of Kurdistani for Strategic Studies, 2(9).
7. Najem, K., Ziti, S., & Zaoui Seghroucheni, Y. (2025). Student Performance and Learning Behavior Dataset for Educational Analytics. Zenodo. https://doi.org/10.5281/zenodo.16459132
8. Omer, A., Faraj, S. M., & Mohamad, S. H. (2023). An application of two classification methods: hierarchical clustering and factor analysis to the plays PUBG. Iraqi Journal of Statistical Sciences, 20(1), 25-42.
9. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.
10. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.
11. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.