Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

Regression Based Prediction of Roblox Game Popularity Using Extreme Gradient Boosting with Hyperparameter Optimization

Inna Nur Amalina[*1], Norhikmah[2], Dony Ariyus[3], Muhammad Koprawi[4], Rafli Ilham Prasetyo[5]

[1,2,3,4,5]Computer Engineering, Amikom Yogyakarta University, Indonesia

Email: [1]innaamalina317@students.amikom.ac.id

Received : Jan 28, 2026; Revised : Feb 8, 2026; Accepted : Feb 9, 2026; Published : Feb 26, 2026

Abstract

The rapid growth of the digital gaming industry has increased the importance of predicting game popularity on user-generated content platforms such as Roblox, where diverse games and highly variable user engagement patterns create challenges in modeling long-term popularity trends. This study aims to develop a regression-based popularity prediction model using the Extreme Gradient Boosting (XGBoost) algorithm based on user interaction indicators, including visits, likes, dislikes, favorites, and active players. To investigate the effect of model optimization, hyperparameter tuning is performed using GridSearchCV. Model performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Experimental results show that the baseline XGBoost model achieves an R^2 value of 80.74%, indicating strong capability in capturing non-linear popularity patterns. However, the optimized model yields a lower R^2 value of 77.71%, accompanied by slight increases in prediction error metrics, revealing that hyperparameter optimization does not always improve performance for highly skewed popularity data. Feature importance analysis further indicates that interaction-based attributes, particularly likes and dislikes, are the most influential predictors. These findings provide an important contribution to Informatics research by demonstrating the effectiveness of ensemble regression models for digital entertainment analytics while highlighting the need for critical evaluation of optimization strategies rather than assuming universal performance gains.

Keywords : *Game Analytics, Hyperparameter Tuning, Machine Learning, Popularity Prediction, Roblox, XGBoost*

1. INTRODUCTION

The fast pace of the digital gaming industry has made online gaming platforms to be a major industry of entertainment in the world. The most notable systems are Roblox that allows not only to play games but also to create so called user generated content and publish it on the platform [1]. The key feature that makes Roblox an effective sandbox environment is that it could match the type of games to the development of a user, promoting the culture of participation when the participants are the ones who would create and consume the content[2]. Popularity indicators such as the number of visits, likes, dislikes, favorites, and active players reflect user engagement and long-term interest, and have been widely used as proxies for measuring game success on the Roblox platform[2].

Due to the non-linear relationships among popularity indicators such as visits, likes, dislikes, favorites, and active players, conventional statistical methods often fail to capture popularity patterns accurately. Therefore, machine learning-based approaches, particularly ensemble models, have been widely adopted to predict game popularity using user interaction and metadata features [3]. Among these methods, Extreme Gradient Boosting (XGBoost) has demonstrated superior performance due to its high predictive accuracy and ability to handle complex data relationships [4]. XGBoost performance is

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

sensitive to hyperparameter configuration, requiring structured optimization methods such as Grid Search and Randomized Search to achieve stable and generalizable predictions [5], [6].

Previous research has demonstrated the applicability of machine learning techniques in modeling and predicting popularity patterns across various forms of digital content, including online games. Research on platforms such as Steam has explored ensemble learning and model interpretability to analyze popularity patterns [7], while studies on Roblox have applied classification algorithms such as K-Nearest Neighbor to categorize game popularity levels [8]. Although these works demonstrate the feasibility of machine learning for popularity analysis, they primarily focus on classification or algorithm comparison without systematically examining regression-based prediction using optimized gradient boosting models. In addition, prior studies rarely investigate whether hyperparameter optimization consistently improves performance for highly skewed popularity data. This limitation reveals a research gap regarding the behavior of optimized ensemble regression models in user-generated gaming environments.

In predictive modeling tasks, evaluating model performance is essential to ensure reliability and generalization capability. To evaluate regression performance in popularity prediction, this study adopts widely used error-based and explanatory metrics, namely MAE, MSE, RMSE, and R^2 [9] These measures allow assessment of both prediction deviation and explanatory power, providing a comprehensive view of model behavior on skewed popularity data [9], [10]. Previous studies have emphasized that comprehensive evaluation using multiple regression metrics is necessary to accurately assess the effectiveness of machine learning models in popularity prediction tasks [10].

Based on these gaps, this study aims to develop a regression-based Roblox game popularity prediction model using XGBoost integrated with hyperparameter optimization. The main contributions of this research are threefold: (1) applying an optimized XGBoost regression framework to user interaction data on Roblox, (2) empirically evaluating whether hyperparameter optimization improves predictive performance for highly skewed popularity data, and (3) providing feature importance analysis to identify the most influential engagement indicators. These contributions offer both methodological insights into ensemble regression optimization and practical implications for digital gaming analytics.

By integrating hyperparameter optimization into the XGBoost regression framework, this research aims to enhance prediction accuracy and model robustness for Roblox game popularity prediction. Building on prior findings, this study acknowledges that XGBoost regression performance is influenced by hyperparameter selection, especially learning rate and tree depth, which govern the trade-off between model flexibility and generalization ability [11]. In the context of online gaming analytics, machine learning models have also been shown to effectively capture user interaction patterns and behavioral trends, providing valuable insights for predicting engagement and long term popularity [12] Therefore, this study combines hyperparameter-optimized XGBoost with user interaction features to provide a robust and generalizable approach for predicting Roblox game popularity.

2. METHOD

Figure 1 illustrates the research methodology applied in this study to predict the popularity of Roblox games using an optimized Extreme Gradient Boosting (XGBoost) regression model. In this study, a Roblox game dataset is prepared by applying preprocessing procedures such as feature filtering, data cleaning, and numerical transformation. To evaluate model generalization, the dataset is divided into training and testing subsets with a ratio of 80% to 20%. A baseline XGBoost regression model is initially developed using default parameters, which is subsequently improved through hyperparameter optimization using GridSearchCV to obtain the best parameter configuration. After optimization, the XGBoost regression model is applied to the testing dataset to generate popularity predictions. Model performance is then assessed using standard regression evaluation metrics, namely MAE, MSE, RMSE,

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

and R^2, to quantify both prediction error and explanatory power. Finally, feature importance analysis and result visualizations are conducted to interpret the influence of each attribute and assess the overall effectiveness of the proposed model. The overall research workflow is summarized visually in Figure 1 to improve clarity, transparency, and reproducibility of the methodological procedure.

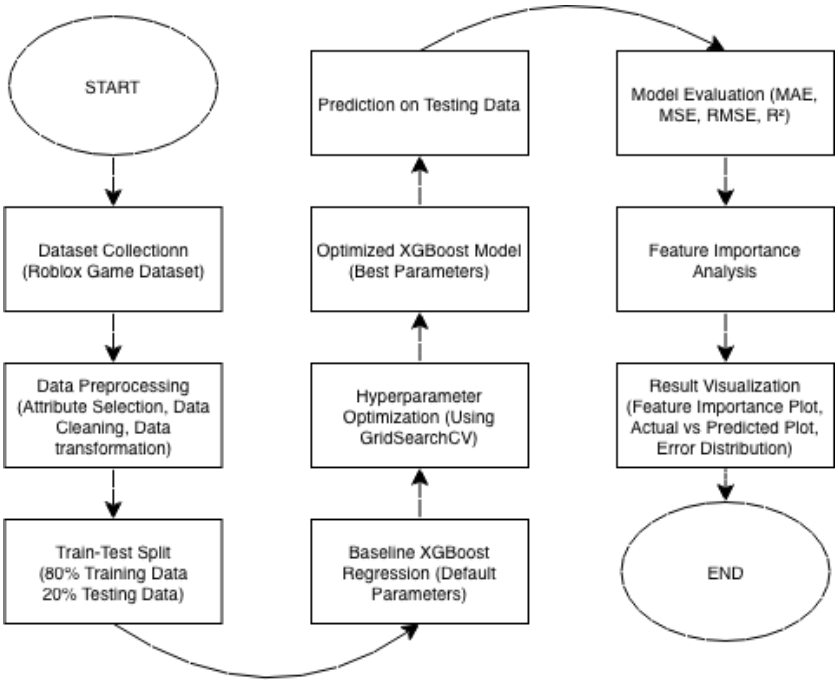

Figure 1. Research Methodology Flowchart

2.1. Data Description

The research is conducted using a Roblox game dataset sourced from an open-access online repository, which provides information on game popularity and user interaction attributes across multiple titles on the Roblox platform. The dataset consists of several key attributes, including game rank, game name, number of active players (Active), total visits (Visits), favorites (Favourites), likes (Likes), dislikes (Dislikes), and rating (Rating).

In this research, the number of visits is selected as the target variable to represent game popularity, while Active, Favourites, Likes, Dislikes, and Rating are used as input features. These attributes were chosen because they directly reflect user engagement and interaction behavior, which are commonly used indicators for evaluating popularity in online gaming environments. Visits is selected as the primary target variable because it represents cumulative user engagement and long-term game exposure, making it a more stable indicator of overall popularity compared to momentary metrics such as active players. The dataset provides sufficient numerical features and variation to support regression-based modeling and is suitable for predicting game popularity using machine learning techniques.

2.2. Data Prepocessing

Prior to model development, the dataset undergoes a preprocessing stage aimed at improving data reliability and ensuring suitability for machine learning analysis. Following established practices in regression modeling research, this study performs data preprocessing through several stages to prepare the dataset for XGBoost regression modeling. A summary of these prepocessing procedures is provided in Table 1 to enhance methodological transparency and reproducibility.

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

Table 1. Summary of Data Prepocessing Steps

Step	Technique	Attributes Involved	Purpose	Outcome
Attribute Selection	Removal of non numeric and irrelevant atributes	Game Name, Rank (removed); Active, Favourites, Likes, Dislikes, Rating (retained); Visits (target)	To retain only informative numerical features that contribute to regression modeling and reduce noise	Dataset contains only predictive engagement features and one target variable
Data Cleaning	Numeric conversion and character removal	Active, Visits, Favourites, Likes, Dislikes, Rating	To convert attributes stored as strings with symbols (e.g., commas, text) into proper numeric format	All features transformed into consistent numeric data types suitable for ML algorithms
Missing Value Handling	Verification and removal of incomplete entries	All selected attributes	To prevent model bias and training errors due to null or inconsistent records	Clean dataset without missing values
Data Transformation	Formatting and normalization of numerical structure	All selected features	To ensure consistent scale and structure for model input and improve learning stability	Structured numerical dataset ready for modeling
Data Splitting	Hold out split (80 : 20)	Entire dataset	To evaluate generalization performance and avoid overfitting	80% training data, 20% testing data
Data Preparation for XGBoost	Feature target separation	X = [Active, Favourites, Likes, Dislikes, Rating]; y = Visits	To organize data into predictor variables and target variable required by XGBoost	Model-ready matrix format for regression

2.2.1. Attribute Selection

Attribute selection consists of selecting and maintaining useful numerical attributes that are useful towards the prediction task. Active, Favourites, Likes, Dislikes, Rating features are taken as input variables and Visits selected as the target variable which is the popularity of the game in this study. Attributes not related to numbers or prediction like the rank and name of games are omitted and this was appropriate as the previous studies have reiterated the need to concentrate on informative attributes so as to [5], [10]

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

2.2.2. Data Cleaning

Data cleaning is performed to handle inconsistent and improperly formatted entries by converting numerical attributes stored as strings and containing non-numeric characters into numeric data types to ensure compatibility with machine learning models and prevent errors during training [13] These characters are removed, and all relevant features are converted to numeric data types to ensure compatibility with regression algorithms such as XGBoost. This step aligns with common data cleaning procedures in machine learning research that stress the need for numeric integrity to prevent errors during model training [14].

2.2.3. Data Transformation

Following data cleaning, transformation procedures are applied to standardize numerical representations and prepare input features for modeling, enabling effective learning by the XGBoost regression algorithm [15].

2.2.4. Data Splitting

The preprocessed dataset is divided into training and testing sets using an 80:20 split to assess the generalization capability of the predictive model and prevent overfitting, as recommended in machine learning evaluation studies [16].

2.2.5. Data Preparation for XGBoost

Data preparation for XGBoost involves organizing the dataset into input features and a target variable in numerical form, ensuring the absence of missing values so that the gradient boosting algorithm can efficiently learn patterns and perform accurate regression tasks [4].

2.3. XGBoost Regression

Extreme Gradient Boosting (XGBoost) is an ensemble algorithm that builds regression models sequentially using gradient-boosted decision trees. It aims to minimize a differentiable loss function while incorporating regularization techniques to prevent overfitting [4]. In this work, we adopt XGBoost regression to examine the non-linear patterns between user interaction data and game popularity. This approach is chosen because of its consistent high performance and ability to scale effectively in a variety of regression prediction problems [4], [17]

2.4. Hyperparameter Optimization (GridSearchCV)

Hyperparameter optimization is performed to systematically explore a predefined parameter grid using GridSearchCV in order to identify the best combination of XGBoost hyperparameters that improves predictive performance and model robustness compared to default settings [6]. The model undergoes a thorough evaluation of various parameter settings, including n_estimators, max_depth, learning_rate, subsample, and colsample_bytree using cross validation to confirm that the selected configuration maintains strong predictive performance on new, unseen data [18], [19].

2.5. Model Evaluation Metrics

The performance of the XGBoost regression model is examined using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics are selected because they provide a comprehensive view of prediction accuracy, the magnitude of errors, and the proportion of variance that the model successfully explains [20]. Mean Absolute Error (MAE) quantifies the typical magnitude of prediction errors, while Mean Squared Error (MSE) emphasizes larger discrepancies. Root Mean Squared Error (RMSE) translates errors into the original data scale, and the coefficient of determination (R^2) shows how effectively the

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

model explains variations in popularity. Using all these metrics together allows for a comprehensive evaluation of the model's predictive performance.

3. RESULT

3.1. Baseline XGBoost Regression Results

This subsection presents the evaluation of the baseline XGBoost regression model in predicting the popularity of Roblox games. For this initial model, default hyperparameter settings were used, and no optimization procedures were applied, providing a reference point for subsequent performance improvements. The input features consist of the number of active players, favorites, likes, dislikes, and rating, while the target variable is the number of visits.

To evaluate the predictive accuracy of the regression model, a set of standard performance metrics was applied, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These metrics are widely used in regression-based machine learning studies to quantify prediction accuracy, error magnitude, and the proportion of variance explained by the model. These metrics are widely used in regression-based machine learning studies to quantify prediction accuracy and variance explanation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}\,i| \quad (1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}\,i)^2 \quad (2)$$

$$RMSE = \sqrt{MSE} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}(yi - \bar{y}i)^2} \quad (4)$$

where $\boldsymbol{yi}$ is the real figure of visits, $\hat{\boldsymbol{y}}\boldsymbol{i}$ is the figure that is predicted, $\bar{\boldsymbol{y}}$ is the mean of the measured figures and n is the overall number of test samples.

The quantitative performance of the baseline XGBoost model is summarized in Table 2**.**

Table 2. Performance of Baseline XGBoost Regression Model

Metric	Value
Mean Absolute Error (MAE)	$\mathbf{1.99 \times 10^8}$
Mean Squared Error (MSE)	$\mathbf{4.80 \times 10^{17}}$
Root Mean Squared Error (RMSE)	$\mathbf{6.92 \times 10^8}$
R-Squared (R^2)	**0.8074**

The results presented in Table 1 indicate that the baseline XGBoost regression model is able to capture a substantial proportion of the variability in Roblox game popularity based on user interaction features. The relatively high coefficient of determination (R^2 = 0.8074) shows that more than 80% of the variance in visit counts can be explained by the model, demonstrating that interaction based indicators contain strong predictive information. The relatively large MAE and RMSE values are consistent with the large numerical scale of visit counts, indicating that absolute errors increase proportionally with popularity magnitude rather than reflecting model instability.

From a practical standpoint, the relatively high R^2 value suggests that the model can be used as an early monitoring tool to identify emerging popular games, while the magnitude of MAE highlights the need for relative rather than absolute error interpretation when dealing with large-scale popularity

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

metrics.From a practical perspective, this level of accuracy suggest that the model can provide a reliable estimation of game popularity trends, which may assist developers in evaluating engangement performance and prioritizing content improvement strategies. Overall, these results confirm that the baseline XGBoost model provides a strong initial performance and serves as a reliable reference for subsequent optimization and comparative analysis.

Alongside the numerical evaluation, the relative contribution of each input feature to the model's predictions was investigated using the feature importance scores provided by XGBoost. The feature importance visualization for the baseline model is presented in Figure 2, illustrating the contribution of each feature in determining the predicted number of visits.

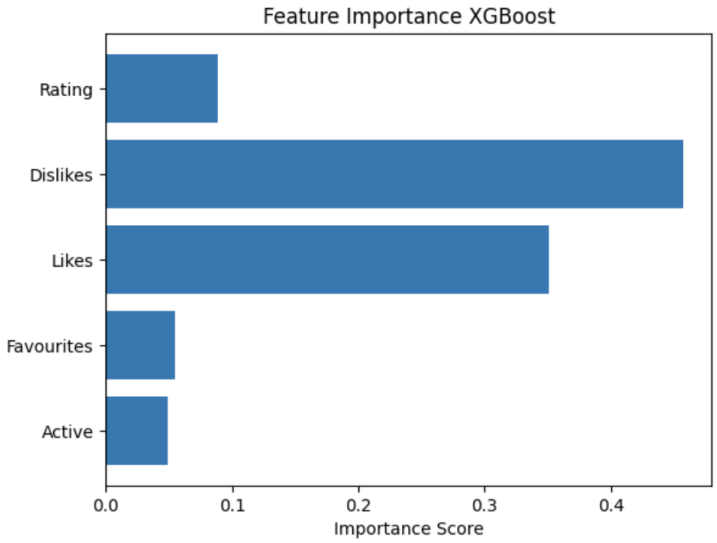

Figure 2. Feature Importance of Baseline XGBoost Regression Model

Figure 2 depicts the feature importance scores generated by the baseline XGBoost regression model, illustrating how each input variable contributes to predicting Roblox game visit counts. Among the features, dislikes has the highest impact with a score of approximately 0.45, followed by likes at around 0.34. The rating feature accounts for roughly 0.10, while favourites and active players show smaller contributions, each with scores below 0.07.

The dominance of dislikes and likes indicates that direct user feedback signals carry more predictive weight than aggregated indicators such as ratings or activity-based measures. This suggests that explicit user reactions reflecting satisfaction or dissatisfaction have a stronger immediate association with visit dynamics compared to passive engagement metrics. Meanwhile, the relatively low contribution of favourites and active players implies that these indicators play a secondary role in visit-based popularity prediction within the Roblox platform. From a practical perspective, this finding highlights the importance of monitoring like–dislike ratios as an early signal of popularity shifts before they are reflected in cumulative visit counts.

3.2. Optimized XGBoost Regression Results

This subsection presents the performance of the XGBoost regression model after the application of hyperparameter tuning, highlighting the improvements achieved through the optimized configuration. To improve predictive performance, the baseline XGBoost model was optimized using the GridSearchCV technique, which systematically evaluates combinations of hyperparameters to identify the optimal configuration based on cross-validation performance.

The hyperparameters optimized in this study include the number of trees (n_estimators), maximum tree depth (max_depth), learning rate (learning_rate), subsampling ratio (subsample), and

column sampling ratio (colsample_bytree). The optimal hyperparameter values obtained from the grid search process are summarized in Table 3**.**

Table 3. Optimal Hyperparameters of XGBoost Regression Model

Hyperparameter	Optimal Value
(***n_estimators***)	300
(**max _depth**)	6
(***learning_rate***)	0.05
(***subsample***)	0.8
(***colsample_bytree***)	0.8

The optimized hyperparameter configuration shown in Table 3 indicates that the XGBoost model benefits from a balanced combination of model complexity and regularization. The selected number of estimators and tree depth suggest that deeper and more numerous trees are required to capture complex interaction patterns among popularity-related features. Meanwhile, the learning rate value enables gradual model updates, helping to stabilize the training process. The subsample and colsample_bytree settings introduce controlled randomness during tree construction, which improves generalization performance by reducing overfitting. This configuration serves as the basis for the optimized XGBoost model evaluated in the subsequent analysis. In the context of Roblox game popularity prediction, this configuration allows the model to better capture complex user interaction patterns while maintaining robustness against extreme variations in engagement metrics.

Using the optimized hyperparameters, the XGBoost regression model was retrained on the training dataset and evaluated on the test set with the same metrics as the baseline model, including MAE, MSE, RMSE, and R^2. The resulting performance of the optimized model is summarized in Table 4.

Table 4. Performance of Optimized XGBoost Regression Model

Metric	Value
Mean Absolute Error (MAE)	$\mathbf{1.9939 \times 10^{8}}$
Mean Squared Error (MSE)	$\mathbf{5.5548 \times 10^{17}}$
Root Mean Squared Error (RMSE)	$\mathbf{7.4531 \times 10^{8}}$
R-Squared (R^2)	**0.7771**

Although hyperparameter optimization aims to enhance generalization performance, the optimized model exhibits a slight reduction in R^2 and increased error metrics compared to the baseline model. Such a behavior is typical in regression problems involving highly skewed distributions, where optimization may prioritize stability over peak value accuracy**,** which can affect the model's sensitivity to viral or exceptionally popular games**.** The results suggest that the hyperparameter-tuned configuration imposes stronger regularization, which helps to mitigate overfitting but may slightly limit the model's ability to accurately capture extremely high visit counts. Such a behavior is typical in regression problems involving highly skewed distributions, where optimization may prioritize stability over peak value accuracy.

Additionally, feature importance was analyzed in the optimized XGBoost model to investigate whether hyperparameter tuning altered the relative influence of each input feature compared to the baseline configuration. The resulting feature importance visualization is shown in Figure 3, illustrating how different popularity indicators contribute to the prediction of game visits in the optimized XGBoost model.

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

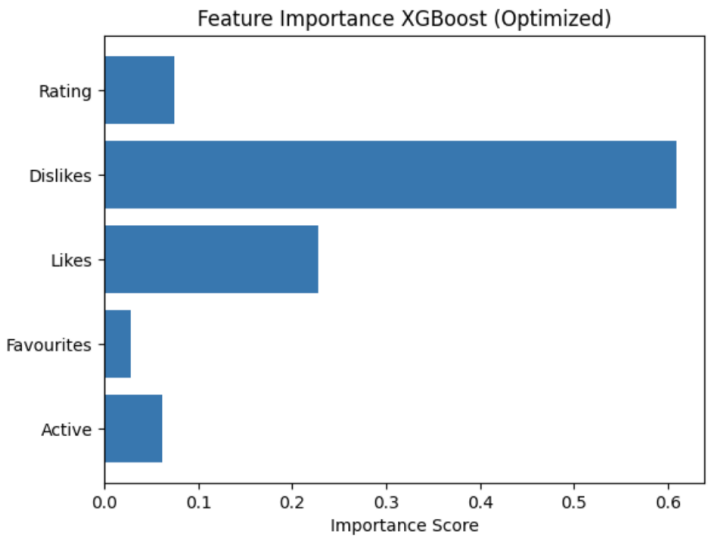

Figure 3. Feature Importance of Optimized XGBoost Regression Model

Figure 3 depicts the feature importance scores of the optimized XGBoost regression model, showing the relative contribution of each input variable in predicting the number of game visits. The results indicate that dislikes are the most influential feature, contributing approximately 0.60 of the total importance score, followed by likes with a score of around 0.23. These findings suggest that user feedback intensity, particularly negative responses, plays a dominant role in shaping popularity dynamics on the Roblox platform. Compared to the baseline model, the optimized configuration further amplifies the importance of dislikes, indicating that hyperparameter tuning increases the model's sensitivity to negative user feedback.

Other interaction-based features, such as rating, active players, and favorites**,** exhibit lower importance scores, indicating a more limited contribution to visit prediction when compared to likes and dislikes. The relatively small importance of rating suggests that aggregated evaluation metrics may be less informative than direct user interaction signals. Overall, the feature importance distribution confirms that engagement-driven attributes are the primary determinants of game popularity in the optimized model, reinforcing the suitability of tree-based boosting methods for modeling complex user-driven popularity dynamics.

3.3. Model Performance Comparison

This subsection provides a comparative evaluation of the baseline XGBoost regression model and the optimized version to examine the effects of hyperparameter tuning on predictive performance. Both models were assessed using the same test dataset and identical metrics MAE, MSE, RMSE, and R^2 to ensure a fair and consistent comparison. The baseline model employs default hyperparameter settings, while the optimized model utilizes the best parameter configuration obtained through the GridSearchCV optimization process.

The comparative results are summarized in Table 5, which presents the quantitative performance differences between the two models. The comparison reveals that the optimized XGBoost model achieves a comparable MAE to the baseline model, indicating a similar average magnitude of prediction errors. However, variations are observed in the MSE and RMSE values, suggesting differences in how each model handles larger prediction errors. Additionally, a slight decrease in the R^2 value is observed after optimization, indicating that while hyperparameter tuning affects error distribution, it does not always guarantee an improvement across all evaluation metrics simultaneously.

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

Table 5. Performance Comparison Between Baseline and XGBoost Models

Metric	Baseline XGBoost	Optimized XGBoost
Mean Absolute Error (MAE)	$\mathbf{1.99 \times 10^{8}}$	$\mathbf{1.9939 \times 10^{8}}$
Mean Squared Error (MSE)	$\mathbf{4.80 \times 10^{17}}$	$\mathbf{5.5548 \times 10^{17}}$
Root Mean Squared Error (RMSE)	$\mathbf{6.92 \times 10^{8}}$	$\mathbf{7.4531 \times 10^{8}}$
R-Squared (R^2)	**0.8074**	**0.7771**

The comparative results in Table 5 reveal that hyperparameter optimization does not uniformly improve all evaluation metrics. While MAE remains relatively stable, the increase in MSE and RMSE indicates that the optimized model is more sensitive to large errors associated with extremely popular games. The decrease in R^2 further suggest that the optimized configuration explains slightly less overall variance than the baseline model. This finding highlights that hyperparameter optimization in gradient boosting does not guarantee superior performance in datasets with highly imbalanced popularity distributions. Practically, this means that default model settings may already provide near optimal performance for large scale engagement data, and excessive tuning could lead to diminishing returns. In the context of Roblox game analytics, this suggests that models optimized for general stability may underrepresent extreme viral popularity spikes, which are critical for early trend detection.

The results indicate that hyperparameter optimization introduces a trade-off between bias and variance, where improvements in certain aspects of model generalization may be accompanied by increased sensitivity to large errors. Such behavior is frequently observed in ensemble regression models when adjusting parameters like tree depth, learning rate, and subsampling ratios. Therefore, performance evaluation should not rely solely on a single metric but consider multiple indicators to obtain a comprehensive understanding of model behavior.

In addition to the numerical comparison, the prediction characteristics of the optimized XGBoost model are further examined using visual analysis techniques, including actual versus predicted value scatter plots and error distribution plots. These visualizations, discussed in the following subsection, provide deeper insights into prediction patterns, model bias, and error dispersion that are not fully captured by aggregate numerical metrics alone.

3.4. Prediction Visualization Results

This subsection presents visual representations of the prediction results generated by the optimized XGBoost regression model. Visualization is employed to provide an intuitive understanding of model behavior by illustrating the relationship between actual and predicted visit values, as well as the distribution of prediction errors across test samples.

Figure 4 shows a scatter plot comparing the actual number of visits with the predicted values produced by the optimized XGBoost model. Each point represents an individual game instance in the test dataset. Points located closer to the diagonal line indicate more accurate predictions, where the predicted visit count closely matches the actual value. The overall distribution of predicted points suggests that the model captures the general trend of visit counts effectively, although larger deviations are observed for games with exceptionally high visits, reflecting the increased challenge of predicting outlier cases.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

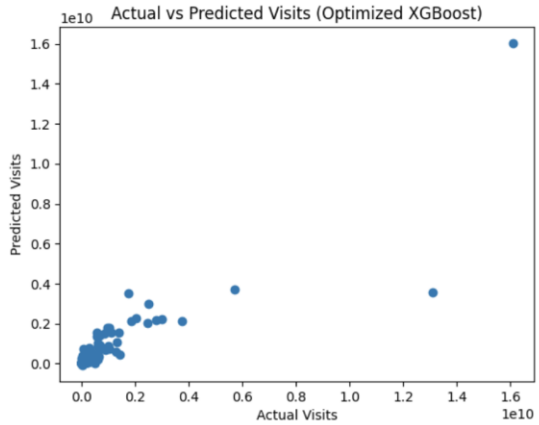

Figure 4. Actual vs Predicted Visits Using Optimized XGBoost

Based on Figure 4, the predicted visit values produced by the optimized XGBoost model show a strong positive relationship with the actual visit counts, as indicated by the clustering of data points along the diagonal direction. This pattern confirms that the model is capable of capturing the overall popularity trend across different game instances. However, larger deviations are observed for games with extremely high visit counts, suggesting that prediction accuracy decreases for highly skewed or outlier data. This pattern is frequently observed in popularity prediction tasks, where a few highly popular items disproportionately influence the data distribution.

The clustering of points along the diagonal confirms that the model captures the general relationship between predicted and actual visits. However, the widening spread of points at higher visit values indicates heteroscedasticity, where prediction variance increases with popularity magnitude. This pattern suggests that extreme popularity cases introduce greater uncertainty, which is consistent with the long tail nature of digital content popularity distributions.

To further examine prediction accuracy and error behavior, Figure 5 illustrates the distribution of prediction errors, defined as the difference between the observed and predicted visit counts. The histogram shows that most prediction errors are concentrated around zero, indicating that the optimized model generally produces unbiased predictions for the majority of samples. However, the presence of a long tail in the error distribution suggests that larger prediction errors occur for a small number of high-visit games, reflecting the inherent challenge of modeling highly skewed popularity data.

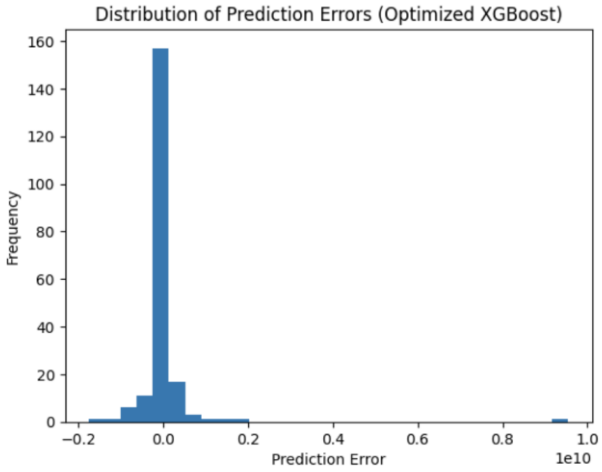

Figure 5. Distribution of Prediction Errors Using Optimized XGBoost

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

As shown in Figure 5, the error distribution of the optimized XGBoost regression model is centered around zero, indicating that the model does not exhibit systematic overestimation or underestimation in predicting game visit counts. The right-skewed tail reflects occasional large positive errors, meaning that the model sometimes underestimates extremely popular games. This behavior suggests that while the model performs consistently for the majority of games, predicting viral-level popularity remains a challenge due to nonlinear amplification effects in user engagement dynamics.

The high concentration of errors near zero suggests stable predictive performance for most samples in the test dataset. Nevertheless, the presence of a right-skewed tail reflects the occurrence of larger prediction errors for a limited number of games with exceptionally high visit values. This phenomenon highlights the inherent difficulty of accurately modeling extreme popularity cases in highly imbalanced and skewed datasets, which is commonly observed in popularity prediction problems. Overall, the visual analysis supports the quantitative evaluation results by demonstrating that the optimized XGBoost model provides reasonable prediction accuracy while maintaining stable error distribution, particularly for the majority of games with moderate visit counts.

4. DISCUSSIONS

This study explores the use of Extreme Gradient Boosting (XGBoost) regression to predict the popularity of Roblox games based on user interaction features, with a particular focus on the impact of hyperparameter tuning. The baseline model was able to explain a substantial portion of the variance in game visit counts, achieving an R^2 value of 0.8074. This confirms that ensemble-based gradient boosting approaches are suitable for popularity prediction tasks where interaction signals dominate the data structure, reinforcing their relevance in digital platform analytics [8].

Rather than reiterating performance metrics, this discussion focuses on interpreting the observed behavior of the XGBoost models under different optimization settings and data characteristics. Although hyperparameter optimization is often expected to enhance model performance, the results of this study show that the tuned XGBoost model does not consistently outperform the baseline across all evaluation metrics. The observed performance trade-off suggests that, in highly skewed popularity datasets, parameter tuning may introduce stronger regularization effects that reduce sensitivity to dominant patterns while increasing stability. This indicates that optimization can shift the balance between bias and variance rather than universally improving model accuracy, especially in long-tail distributions typical of digital content platforms[13]. Similar observations have been reported in previous studies, where optimized models exhibited sensitivity to data characteristics and noise levels, leading to performance trade-offs between bias and variance. [21].

Feature importance analysis further reveals that interaction-based attributes, particularly likes and dislikes, contribute most significantly to the prediction of game popularity. This highlights the dominant role of explicit user feedback as a behavioral signal that directly influences visibility and engagement dynamics. Unlike static metadata features, interaction indicators more accurately reflect user perception and behavioral response, making them stronger predictors of long-term popularity trends. Comparable patterns have been observed in prior research on online gaming behavior and digital content popularity, where engagement-driven features consistently outperform descriptive attributes .

Compared to existing works, the study expands upon existing studies into the issue of limitations in current game popularity prediction literature that, by and large, focused on classification-based methods or algorithm comparisons in an unfocused manner with no explicit hyperparameter optimization, especially on platforms like Roblox and Steam [8][19]. Just as in the larger mobile application market, researchers who perform analyses on large-scale data sets provided by Google Play Store have shown that tree-based ensemble processes, like Random Forest, have always been better predictors of the performance of an application when compared to other more basic algorithms (Naive

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

Bayes and Decision Trees) [22]. Although the most recent works touched upon the improved ensemble models like CatBoost with explainability methods, they focus more on the model interpretability instead of the performance improvement by optimization [7]. To the contrary, the study uses a regression-based paradigm with hyperparameter optimization, allowing a more accurate model of continuous popularity measures, e.g., trip counts, and provide a more detailed view of the popularity dynamics.

Furthermore, studies on XGBoost optimization in other application domains including stock price prediction and environmental forecasting have reported consistent performance gains when applying grid-based hyperparameter search strategies [23]. However, the findings of this study demonstrate that such improvements are domain-dependent and influenced by data distribution characteristics. While the optimized model in this study does not surpass the baseline in all metrics, the alignment of findings across domains highlights the importance of contextual evaluation when applying optimization techniques. This research demonstrates that in user-generated gaming platforms, baseline configurations may already capture dominant popularity patterns, and excessive tuning may lead to overfitting or reduced generalization under certain data conditions [13], [23]. However, tuning parameters is an essential phase in the machine learning workflow for identifying the performance limits of tree-based algorithms such as XGBoost [24].

From a scientific perspective, this research contributes to Informatics by providing empirical evidence that hyperparameter optimization in gradient boosting should be treated as an exploratory performance-boundary analysis rather than a guaranteed enhancement step. This insight is particularly relevant for large-scale digital platforms where popularity follows extreme long-tail distributions. By empirically demonstrating the performance boundaries of optimized XGBoost models, this work advances methodological understanding in data-driven decision support systems within the Informatics domain. Practically, the findings suggest that platform analysts and game developers can rely on interaction-driven features for predictive monitoring, while applying optimization strategies cautiously when dealing with highly skewed popularity metrics**.** Overall, this study strengthens the understanding of machine learning behavior in digital entertainment analytics and highlights the interplay between model complexity, optimization strategies, and data distribution characteristics.

On the whole, this discussion indicates that implementing hyperparameter optimization within the XGBoost regression model could continue serving as an excellent approach to the investigation of the capacities of performance in popularity prediction tasks, despite the trends in improvement not being uniform. The results confirm the appropriateness of XGBoost to predict non-linear variations of popularity and highlight the importance of due consideration of optimization results. To do work even better in the future, it is possible to consider adding more behavioral characteristics, different optimization techniques like Bayesian-based methods, or Hyperopt which were proven to potentially be more effective than the traditional grid-based predictors in improving the predictive capabilities of the proposed models [21] and conduct the cross-platform popularity analysis that can even more promote the robustness and generalizability of the models. These guidelines have potential to enhance new possibilities in developing machine learning applications in digital entertainment analytics and user-generated content ecosystems.

Considering the sports analytics field, scholars have already shown that machine learning algorithms are effective in estimating the variables of match results that center on readily accessible factors in the pre-match that include home-ground advantage and toss decisions [25]. In line with the results of this study, tree-based models, such as Random Forest tend to have a higher accuracy, precision, and recall over probabilistic or simple statistical methods in a dynamic sport environment [25]. This also confirms the use of XGBoost, which is a sophisticated tree-boosting model, to study the non-linear popularity trends of digital games. Overall, these directions position XGBoost optimization not merely

as a performance-enhancing technique, but as a methodological lens for understanding popularity dynamics in complex, user-generated digital ecosystems.

5. CONCLUSION

This study examines the use of an Extreme Gradient Boosting (XGBoost) regression model to predict the popularity of Roblox games using user interaction metrics. By incorporating engagement related features such as visits, likes, dislikes, favorites, and active players, the research aims to capture the complex, non-linear relationships underlying popularity dynamics on user-generated gaming platforms.

Results from the experiments indicate that the baseline XGBoost model explains a substantial portion of the variance in game visit counts, achieving an R^2 of 0.8074. This finding confirms that ensemble-based regression models are effective for popularity prediction tasks involving highly skewed and interaction-driven data. However, the application of hyperparameter optimization using GridSearchCV does not consistently improve model performance. The optimized model shows a slight decrease in predictive accuracy, with the R^2 value declining to 0.7771, accompanied by marginal increases in MAE, MSE, and RMSE. These results suggest that hyperparameter tuning does not universally guarantee performance improvement and may introduce trade-offs when applied to datasets with extreme values and high variance.

Furthermore, feature importance analysis reveals that interaction-based attributes, particularly likes and dislikes, play a dominant role in predicting game popularity. This outcome highlights the significance of direct user feedback mechanisms in shaping long-term engagement patterns on user-generated content platforms. Compared to static or descriptive attributes, interaction indicators provide more representative signals of user perception and behavioral response, making them critical factors in popularity modeling.

Despite its contributions, this study is limited to a single platform and a fixed set of interaction features. Future research may incorporate temporal engagement patterns, textual or visual content features, and alternative optimization strategies such as Bayesian optimization or evolutionary algorithms. Additionally, extending the proposed framework to other digital content platforms could further evaluate its robustness and generalizability, thereby enhancing its contribution to the broader field of digital entertainment analytics and popularity prediction research. Overall, this study contributes to the fields of game analytics and user-generated content analytics by offering empirical insights into the behavior of gradient boosting models and optimization strategies when applied to highly skewed engagement data. By clarifying the strengths and limitations of gradient boosting models under highly skewed engagement data, this study provides practical insights for the development of more reliable game analytics and user-generated content analytics frameworks in large-scale digital platforms.

ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to the academic institution for its support and to the research supervisors for their valuable guidance and suggestions throughout this study. Appreciation is also extended to individuals who assisted in data collection and technical discussions.

REFERENCES

[1] Y. Liu, H. Duan, and W. Cai, "User-Generated Content and Editors in Games: A Comprehensive Survey," Dec. 2024, doi: https://doi.org/10.48550/arXiv.2412.13743.

[2] D. Yi, "Predicting the Popularity Level of Roblox Games Using Gameplay and Metadata Features with Machine Learning Models," *International Journal for Applied Information Management*, vol. 5, no. 1, pp. 30–42, Apr. 2025, doi: 10.47738/ijaim.v5i1.97.

[3] N. H. Nguyen, D. T. A. Nguyen, B. Ma, and J. Hu, "The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity," *Journal of Information and Telecommunication*, vol. 6, no. 2, pp. 217–235, 2022, doi: 10.1080/24751839.2021.1977066.

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Jun. 2016, doi: 10.1145/2939672.2939785.

[5] N. M. Lefi and M. Rahardi, "Hyperparameter Optimization and Feature Selection Analysis on the XGBoost Model for Hepatitis C Infection Prediction," 2025. doi: https://doi.org/10.30871/jaic.v9i6.10876.

[6] "Optimization of XGBoost hyperparameters using grid search and random search for credit card default prediction", doi: https://doi.org/10.35335/mandiri.v14i2.468.

[7] M. T. Syamkalla, S. Khomsah, and Y. S. R. Nur, "Implementasi Algoritma Catboost Dan Shapley Additive Explanations (SHAP) Dalam Memprediksi Popularitas Game Indie Pada Platform Steam," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 4, pp. 777–786, Aug. 2024, doi: 10.25126/jtiik.1148503.

[8] "Analisis Klasifikasi Popularitas Game Roblox Menggunakan Algoritma K-Nearest Neighbor (KNN) (Buana, et al.)", doi: 10.63822/0dwatj19.

[9] G. Airlangga, "Performance Evaluation of Machine Learning Models for Predicting Household Energy Consumption: A Comparative Study," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 8, no. 1, p. 76, Dec. 2024, doi: 10.24014/ijaidm.v8i1.32791.

[10] G. A. Narkunam, K. Kala, and S. Arunpandiyan, "Enhancing Agricultural Forecasting with an Ensemble Learning Approach for Broccoli Yield Prediction ARTICLE INFO ABSTRACT," 2024. [Online]. Available: https://www.jisem-journal.com/

[11] Tarwidi D; Pudjaprasetya SR; Adytia D; Apri M, "An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach," Mar. 2023, doi: 10.1016/j.mex.2023.102119.

[12] N. Rismayanti, "Predicting Online Gaming Behaviour Using Machine Learning Techniques," *Indonesian Journal of Data and Science*, vol. 5, no. 2, Jul. 2024, doi: 10.56705/ijodas.v5i2.166.

[13] S. Mohammed, F. Naumann, and H. Harmouch, "Step-by-Step Data Cleaning Recommendations to Improve ML Prediction Accuracy," in *Advances in Database Technology - EDBT*, OpenProceedings.org, Mar. 2025, pp. 542–554. doi: 10.48786/edbt.2025.43.

[14] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks," Apr. 2021, doi: https://doi.org/10.48550/arXiv.1904.09483.

[15] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," Oct. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ai6100257.

[16] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," *J. Anal. Test.*, vol. 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2.

[17] A. Performa *et al.*, "Performance Analysis of XGBoost Algorithm to Determine the Most Optimal Parameters and Features in Predicting Stock Price Movement," *Jurnal Informatika dan Teknologi Informasi*, vol. 20, no. 1, pp. 91–102, 2023, doi: 10.31515/telematika.v20i1.9329.

[18] N. Alamsyah, B. Budiman, T. P. Yoga, and R. Y. R. Alamsyah, "XGBOOST HYPERPARAMETER OPTIMIZATION USING RANDOMIZEDSEARCHCV FOR ACCURATE FOREST FIRE DROUGHT CONDITION PREDICTION," *Jurnal Pilar Nusa Mandiri*, vol. 20, no. 2, pp. 103–110, Sep. 2024, doi: 10.33480/pilar.v20i2.5569.

[19] D. Morreale¨università, M. Morreale¨università, D. Studi, G. Marconi, A. Rosa, and D. Morreale, "Roblox and the Pervasiveness of Play: What Game-Making Communities Can Teach Us About Participatory Practices in Affinity Spaces," 2024. Accessed: Feb. 03, 2026. [Online]. Available: https://ijoc.org/index.php/ijoc/article/view/21902

[20] X. Ying, "An Overview of Overfitting and its Solutions," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Mar. 2019. doi: 10.1088/1742-6596/1168/2/022022.

Vol. 7, No. 1, February 2026, Page. 675-690
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5648

[21] S. Putatunda and K. Rama, "A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2018, pp. 6–10. doi: 10.1145/3297067.3297080.

[22] S. Abulhaija, S. Hattab, A. Abdeen, and W. Etaiwi, "Predicting Mobile Apps Performance using Machine Learning," *Journal of System and Management Sciences*, vol. 12, no. 6, pp. 300–314, 2022, doi: 10.33168/JSMS.2022.0619.

[23] Sugiarto *et al.*, "Optimizing The XGBoost Model with Grid Search Hyperparameter Tuning for Maximum Temperature Forecasting," *Journal of Applied Data Sciences*, vol. 6, no. 4, pp. 2517–2529, Dec. 2025, doi: 10.47738/jads.v6i4.885.

[24] C. G. L. Pringandana and K. Kusnawi, "A Comparative Analysis of Hyperparameter-Tuned XGBoost and LightGBM for Multiclass Rainfall Classification in Jakarta," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 4, pp. 2467–2483, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4965.

[25] K. Kapadia, H. Abdel-Jaber, F. Thabtah, and W. Hadi, "Sport analytics for cricket game results using machine learning: An experimental study," *Applied Computing and Informatics*, vol. 18, no. 3–4, pp. 256–266, Jun. 2022, doi: 10.1016/j.aci.2019.11.006.