Journal of Innovative Applied Mathematics and Computational Sciences

J. Innov. Appl. Math. Comput. Sci. **4**(2) (2024), 153–174. DOI: 10.58205/jiamcs.v4i2.1848

http://jiamcs.centre-univ-mila.dz/

A Bayesian approach for predicting match outcomes: FIFA World Cup 2026

Rotimi Ogundeji ✉[1], Anuoluwapo Aleem [2] and David Obute [3]

[1,3]Department of Statistics, Faculty of Science, University of Lagos, Akoka, Nigeria.
[2]Department of Statistics and Data Science, Fox School of Business, Temple University, USA

Received 13 October 2024, Accepted 21 December 2024, Published 27 December 2024

Abstract. One of the biggest international football competitions, the FIFA World Cup provides teams with an exciting and unpredictable stage on which to display their skills. Predicting match outcomes isn't easy due to the numerous factors involved, like team strategy, player performance, and even unpredictable elements like weather or injuries. Traditional statistical methods in the frequentist framework (such as regression model, machine learning and Monte Carlo simulation) might not fully capture these complexities. This study applied the Bayesian logistics regression and gradient boosting model. to predict possible match outcomes in the forthcoming FIFA World Cup 2026. The Bayesian framework provides a probabilistic and adaptable base that adjusts to tournament dynamics and incorporates prior knowledge, while gradient boosting captures complex non-linear correlations. Key variables include player form, team dynamics, and strategic differences. Data were collected from FIFA's official site and Kaggle, covering historical match data, player statistics and team rankings. Data preprocessing, including median imputation for missing values and feature engineering were carried out. The dataset is split into train-test-validate sets, and the two models evaluated exhibited high predictive accuracy. The study identified top contenders, highlighted offensive and defensive strengths, noted feature importance. The findings emphasize the potential of machine learning in sports analytics. The results identified the leading contenders for the 2026 FIFA World Cup, listing them in order of superiority. Results aim to contribute to the field of sports analytics, offering valuable insights into the complex dynamics influencing success in high-stakes football tournaments. From the literatures, this study on the application of Bayesian logistics regression and gradient boosting model is one of the rare applications to sport analytic.

Keywords: Bayesian logistics regression; gradient boosting model; FIFA World Cup; machine learning, sports analytics.

2020 Mathematics Subject Classification: 62C10, 62C12, 62P25. MSC2020

1 Introduction

In the world of sports, predicting outcomes is as thrilling as the games themselves. The 2026 FIFA World Cup is a global event that captures the attention of millions worldwide. Besides

✉Corresponding author. Email: rogundeji@unilag.edu.ng

ISSN (electronic): 2773-4196

the excitement and entertainment, the World Cup also has a significant impact on the world in various ways [7]. Economically, it can boost a country's economy [16]. Predicting match outcomes isn't easy due to the numerous factors involved, like team strategy, player performance and even unpredictable elements like weather or injuries [9]. Traditional statistical methods in the frequentist framework (such as regression model, machine learning and Monte Carlo simulation) might not fully capture these complexities. This study applied Bayesian logistic regression method has the potential for modelling such complex data based on its robustness and probabilistic method of prediction.

The anticipation and excitement surrounding past FIFA world cups have sparked a surge of interest in advanced predictive analytics, particularly employing Bayesian methodologies. The study reviewed key studies and contemporary research on a Bayesian approach for predicting match outcomes, shedding light on the unique challenges and opportunities in predicting sports or game outcomes. Foundational work on the Bayesian hierarchical model for ranking NCAA basketball teams, such as [13], has paved the way, demonstrating the adaptability and efficacy of Bayesian methods in sports team rankings. [19] proposed a Bayesian methodology for predicting match outcomes by computing the probabilities of wins, draws, and losses for each match, as well as simulating the entire competition to estimate group-stage classification probabilities and tournament-winning chances for each team. [1] applied a Bayesian hierarchical model to predict football results, demonstrating the versatility of Bayesian frameworks and offering insights into their ability to provide fine distinctions and accurate predictions, specifically tailored to the complexities of football matches in tournaments of global magnitude. Insights into the research by [20] highlight the use of Bayesian networks to predict football results in the English Premier League. Other related studies using Bayesian approaches in sports analytics-particularly for predicting match outcomes in high-stakes tournaments and their capacity to seamlessly incorporate prior knowledge and adapt to dynamic scenarios-include [3–6,12,17,21–23,25,26]. Conversely, non-Bayesian frameworks in sports analytics are exemplified by the works of [2,8,10,27].

Bayesian modeling has emerged as a powerful tool in sports analytics, offering a robust framework that integrates prior knowledge with empirical data. Bayesian approach is a methodology that offers a dynamic and probabilistic method to prediction where evidence is updated with data [18]. In this study, the Bayesian approach is used to update and make predictions for the 2026 FIFA World Cup matches. This involves a blend of statistical theory and football analysis. Our understanding of marginal probability could definitely play a role in interpreting the results.

In this study, the prior is initially based on team rankings, player stats, etc. As the World Cup progresses and more data are available, the prior is updated to get the corresponding posterior. Thus, the updated prediction of the match outcome.

This study aims to leverage a Bayesian approach to predict the outcomes of the 2026 FIFA World Cup. This will be achieved by constructing a comprehensive dataset on the teams and players participating in the tournament, developing a robust machine learning model to analyze the data, and ultimately obtaining and interpreting probabilities for each match outcome. The overall objective is not merely to predict match outcomes but also to gain deeper insights into the power and flexibility of Bayesian statistics.

2 Materials and methods

2.1 Research design

The framework used to build the predictive model for this study [17] is shown in Figure 1.

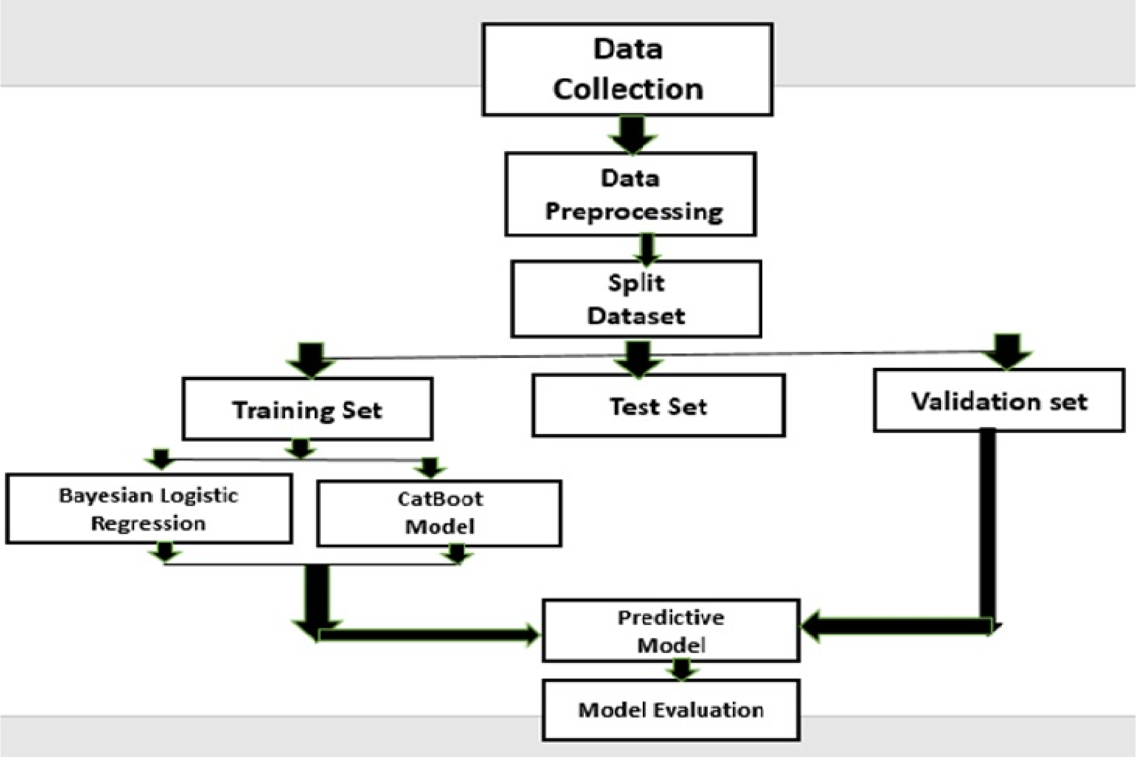


Figure 1: Research methodology framework

2.1.1 Data collection

The data used for this research are secondary data collected from two major sources: the Fédération Internationale de Football Association (FIFA), the official site of the international governing body of football, and Kaggle. The dataset provides a comprehensive collection of historical match data from various tournaments, including the FIFA World Cup, World Cup qualifiers, continental tournaments, and regional championships. It includes features such as match dates, respective locations, total FIFA points, team rankings, player performance metrics, historical team and player statistics, match outcomes, and additional contextual factors. Overall, the dataset encompasses approximately 23,921 match results.

2.1.2 Data preprocessing

Data preprocessing is a phase in this framework that involves meticulous handling of missing values, addressing outliers, and rectifying inaccuracies to ensure the dataset's quality. Additionally, it encompasses transforming the data into a standardized format suitable for predictive models. By optimizing the secondary data through preprocessing, the goal is to enhance the accuracy and reliability of the predictive models, laying the foundation for subsequent analyses. Preprocessing includes data cleaning, data transformation, data reduction, data integration, and other related tasks.

2.1.3 Data wrangling

Data wrangling is the comprehensive process of cleaning and transforming secondary data into a structured and usable format. This involves handling missing values and ensuring data consistency to make the dataset more accessible and easier to analyze. Data wrangling is instrumental in creating a well-prepared dataset that aligns with the requirements of predictive models, facilitating accurate predictions and providing meaningful insights into the dynamics of World Cup matches.

2.1.4 Feature engineering

Feature engineering is crucial for improving the predictive accuracy and interpretability of datasets. It involves creating new attributes or modifying existing ones to extract meaningful insights and enhance model performance. By applying domain expertise and using data manipulation techniques, feature engineering enables the representation of complex relationships and patterns present in the dataset.

2.1.5 Data encoding

Data encoding is a pivotal process that involves transforming categorical variables into a format suitable for machine learning models. In this study, categorical attributes such as team names or match locations were converted into numerical representations to ensure compatibility with Bayesian Logistic Regression and CatBoost machine learning models. However, CatBoost employs internal mechanisms to effectively handle categorical variables, eliminating the need for explicit encoding.

By efficiently representing categorical information in numerical format, the dataset becomes well-suited for accurate predictions, ultimately enhancing the overall success of the modeling process and providing deeper insights into the complex dynamics of World Cup matches.

2.1.6 Data splitting

This phase involves dividing the historical match data into distinct subsets, typically training and testing sets. The training set is used to train the models, while the testing set, also referred to as the validation or hold-out set, is used to validate their performance and generalization to new and unseen data. For this research, an 80:20 dataset split was employed, as it is considered an optimal train-test ratio based on online sources. Specifically, the dataset was split into 80% training data and 20% test data, in alignment with the Pareto Principle.

2.1.7 Training data

Training data, also known as the learning set, is the portion of the split dataset used to teach a predictive model, enabling it to learn patterns and relationships within the historical match data [5]. The effectiveness of the models in forecasting match outcomes is improved and refined during the training process, preparing them for subsequent evaluation and application to new data.

2.1.8 Test data

Test data, also known as validation data, is the portion of the split dataset used to evaluate the performance of the predictive model. It serves as an independent set of examples not seen by the models during training, enabling an unbiased assessment of their predictive accuracy and effectiveness in forecasting World Cup match outcomes.

2.2 Predictive models

Predictive modeling can be defined as the process of creating or selecting a model to best predict the probability of an outcome [14]. It involves using algorithms to make predictions or classifications based on data [11]. The predictive models used in this study to predict FIFA World Cup match outcomes (win, draw, or lose) are Bayesian Logistic Regression and Gradient Boosting Trees. These models, used for classification tasks, predict categorical outcomes or labels and belong to the category of supervised learning models.

2.2.1 Bayesian logistic regression

Bayesian Logistic Regression is a probabilistic approach to logistic regression that incorporates Bayesian principles. Like traditional logistic regression, it is used for binary classification tasks, where the goal is to predict the probability of an observation belonging to one of two classes. However, Bayesian Logistic Regression introduces a Bayesian framework, which allows for modeling uncertainty in the estimates of the model parameters. Logistic regression, a statistical technique employed for multiclass classification tasks like predicting win, lose, or draw outcomes, models the probability of each outcome based on predictor variables. It involves fitting a logistic curve to the data, where this curve depicts the probability of each outcome as a function of the predictors [5, 15]. The logistic function is used to model the probability of an observation belonging to the positive class:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}}. \tag{2.1}$$

Where Y is the binary outcome, $X_1, \ldots, X_n$ are the input features, and $\beta_0, \beta_1, \ldots, \beta_n$ are the model parameters.

Bayesian Logistic Regression incorporates prior beliefs about the parameters by specifying prior distributions. The priors are denoted as:

$$p(\beta_0), p(\beta_1), \ldots, p(\beta_n). \tag{2.2}$$

For this study, Gamma distribution priors are considered, especially when modeling rates or counts in terms of historical match results, player statistics, team rankings, etc.

The likelihood function represents the probability of observing the data given the model parameters. For logistic regression, this involves the product of Bernoulli likelihood for each observation:

$$P(Y|X,\beta) = \prod_{i=1}^{N} P(Y_i|X_i,\beta). \tag{2.3}$$

Where N is the number of observations.

Bayes' theorem is used to update the prior beliefs based on observed data, yielding the posterior distribution:

$$P(\beta|Y,X) \propto P(Y|X,\beta) \cdot P(\beta). \tag{2.4}$$

The posterior distribution represents the updated beliefs about the parameters given the observed data. Markov Chain Monte Carlo (MCMC) methods, such as Gibbs sampling or Metropolis-Hastings, are often employed to draw samples from the posterior distribution, allowing for estimation of the parameter values.

2.2.2 Gradient boosting trees (GBTs)

In this study, the gradient boosting algorithm employed is the CatBoost model. CatBoost (Categorical Boosting) is a machine learning algorithm belonging to the family of Gradient Boosting Trees (GBTs), designed to handle categorical features efficiently. Its proof involves showing how the algorithm minimizes a specific loss function by adding trees sequentially and updating the model parameters. In the context of this study, the ability of CatBoost to handle categorical features efficiently makes it a suitable choice for tasks involving diverse types of input features. CatBoost is a powerful machine learning algorithm designed for a variety of classification tasks, such as binary or multiclass classification problems. It has shown outstanding performance in predicting outcomes across different categories, such as win, loss, or draw, in events like the World Cup. In contrast to logistic regression, which relies on a logistic curve for modeling probabilities, CatBoost utilizes gradient boosting on decision trees to improve predictive accuracy by efficiently managing categorical variables. One notable feature of CatBoost is its ability to handle categorical features without the need for explicit encoding, thanks to its built-in mechanisms for effectively managing such variables. Additionally, CatBoost's automatic handling of missing data is another attribute that contributes to its effectiveness, as it eliminates the need to fill in missing values, unlike the Logistic Regression model [24].

The Gradient Boosting Model can be mathematically represented as:

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i). \tag{2.5}$$

where $\hat{y}$ is the predicted output for the i-th instance, K is the number of trees, and $f_k(x_i)$ is the output of the k-th tree for the i-th instance.

2.3 Model evaluation

2.3.1 Confusion matrix

This is an $N \times N$ matrix structure used for evaluating the performance of a classification model, where N is the number of classes that are predicted. It is applied to the test dataset produced after data splitting, in which the true values are known.

Table 1: Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

The confusion matrix is typically organized into a grid with four quadrants, representing the possible outcomes of a binary classification problem, as shown in Table 1.

2.3.2 True positive (TP)

This refers to the number of cases that a model predicts correctly such that both the "Truth" label and the "Predicted" label are positive in a confusion matrix.

2.3.3 True negative (TN)

This refers to the number of cases that a model predicts correctly such that both the "Truth" label and the "Predicted" label are negative in a confusion matrix.

2.3.4 False positive (FP)

This refers to the number of cases that a model predicts incorrectly such that the "Truth" label is negative, but the "Predicted" label is positive in a confusion matrix.

2.3.5 False negative (FN)

This refers to the number of cases that a model predicts incorrectly such that the "Truth" label is positive, but the "Predicted" label is negative in a confusion matrix.

2.3.6 Precision

Precision is the ratio of the number of true positives to the total number of positive predictions made by the classifier. This is represented mathematically as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}. \tag{2.6}$$

2.3.7 Recall

Recall is the ratio of the number of true positives to the total number of instances that should have been predicted as positive by the classifier. This is represented mathematically as:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}. \tag{2.7}$$

2.3.8 F1-score

This is the harmonic mean of precision and recall. This is represented mathematically as:

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{2.8}$$

2.3.9 Accuracy

Accuracy, also known as error rate or the micro average of F1-score, is defined as the percentage of correct predictions out of all predictions made by a trained machine learning model. This is represented mathematically as:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}. \tag{2.9}$$

2.3.10 Receiver operating curve (ROC) scores

The Receiver Operating Curve (ROC) is used to evaluate the accuracy of a model with two possible outcomes (binary). The ROC curve is a plot of the true positive rate versus the false positive rate, where the false positive rate is on the horizontal axis and the true positive rate is on the vertical axis. The area under the curve (AUC) helps to give an accuracy score to the model. The advantage of the AUC-ROC score over classification accuracy is that, while classification accuracy tests metrics on predicted classes, the AUC of the ROC curve tests accuracy based on predicted scores.

3 Results

3.1 Data source and collection

The data utilized in this study were sourced from:

(i) Kaggle Notebooks: A popular platform for data science and machine learning enthusiasts, which hosts a diverse range of datasets relevant to football analytics. This includes historical match results, player statistics, team rankings, and other pertinent information. Dataset URL: https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017

(ii) FIFA: Data were extracted directly from the Fédération Internationale de Football Association (FIFA) website, including information on national teams, player attributes, match schedules, and tournament history. This data was accessed through web scraping techniques, ensuring that the dataset remained up-to-date and aligned with the latest developments in international football.
FIFA Rankings URL: https://inside.fifa.com/fifa-world-ranking/men
The datasets from various sources were merged, standardized, and validated to ensure uniformity and precision. Particularly, data from the FIFA website played a crucial role in scrutinizing and overseeing the combined dataset, thereby enhancing its reliability.

In this study, the analysis and model implementation were done using statistical software/tools like R and Python.

3.2 Data structure

The dataset consists of 23,921 rows and 25 columns, where each row represents a unique observation, such as a match instance or player profile, and each column corresponds to a specific attribute or feature. The dataset's layout allows for in-depth analysis and investigation of various football-related topics, such as player performance, team interactions, and match results. Some of the attributes include, but are not limited to, the data structure in Appendix H.

3.3 Exploratory data analysis (EDA)

The EDA involved a systematic examination of the dataset through summary statistics, visualizations, and exploratory techniques as presented below.

3.3.1 Top 10 teams on the FIFA ranking

The data obtained indicates that Brazil is currently the highest-ranked team as of June 6th, 2022. Following Brazil in the rankings are Belgium, France, Argentina, and England, making up the top five teams.

Table 2: Top 10 teams on the FIFA ranking

Team	Date	Rank
Brazil	2022-06-06	1
Belgium	2022-06-14	2
France	2022-06-13	3
Argentina	2022-06-05	4
England	2022-06-14	5
Italy	2022-06-14	6
Spain	2022-06-12	7
Portugal	2022-06-12	8
Mexico	2022-06-14	9
Netherlands	2022-06-14	10

3.3.2 Top ten teams with the highest wins

The top ten teams that win matches or boast the highest win percentage both at home and away are shown below. On average, Brazil, Spain, and France rank among the top teams, as illustrated in Figure 2.

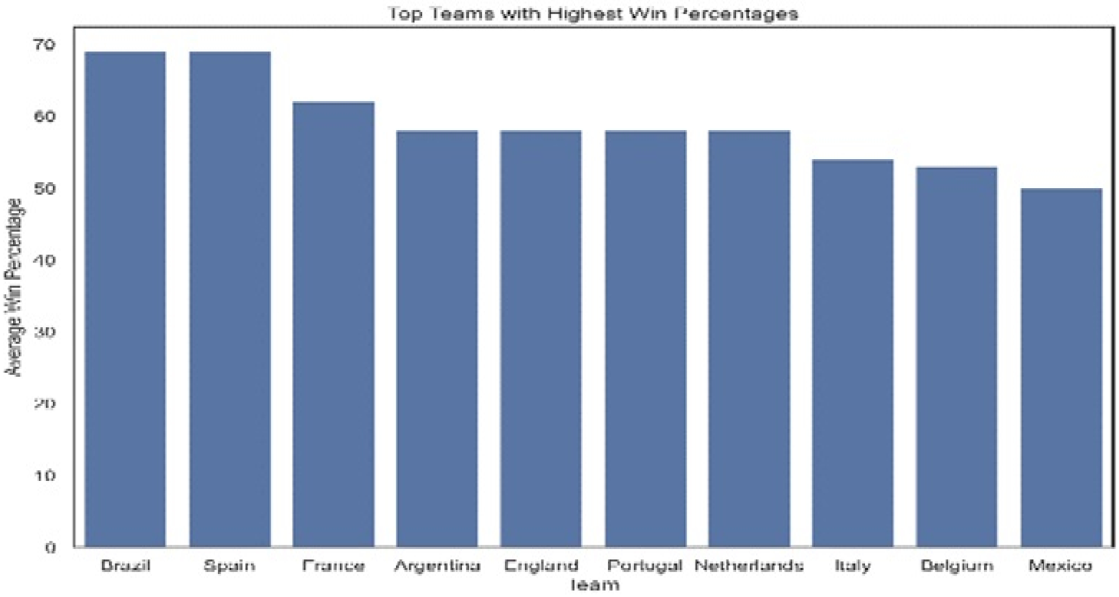


Figure 2: Top teams with highest win percentages

3.3.3 Most offensive teams

Argentina is the team with the greatest attacking potential, with France, England, Brazil, and Portugal completing the top five, as shown in Figure 3.

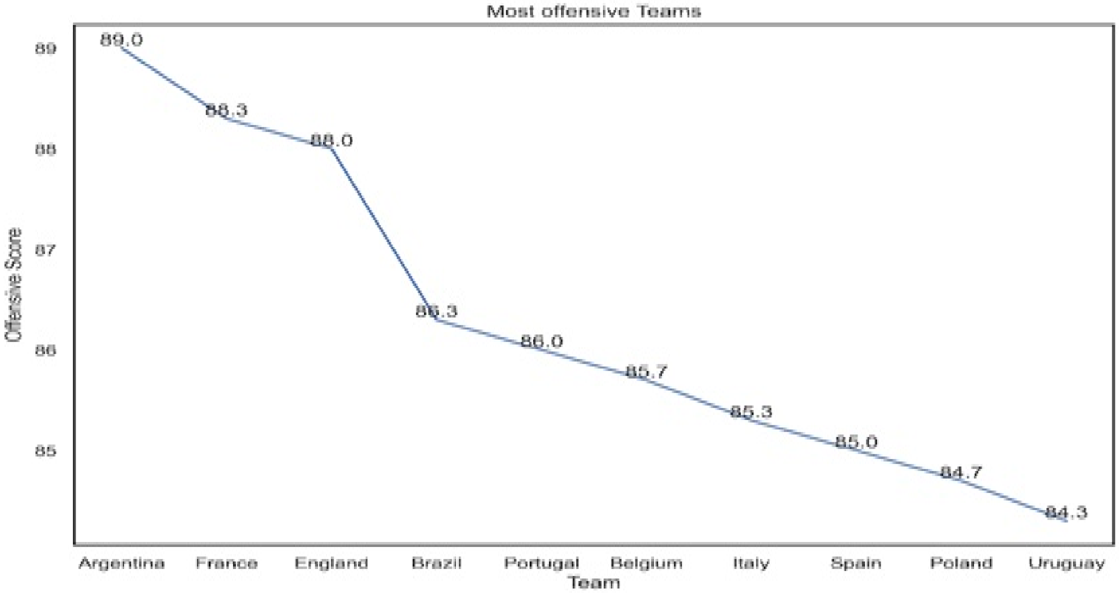


Figure 3: Teams with the greatest offensive potential

3.3.4 Most defensive teams

Spain is the team with the greatest defensive potential, with Portugal, Netherlands, England, and Brazil completing the top five, as shown in Figure 4.

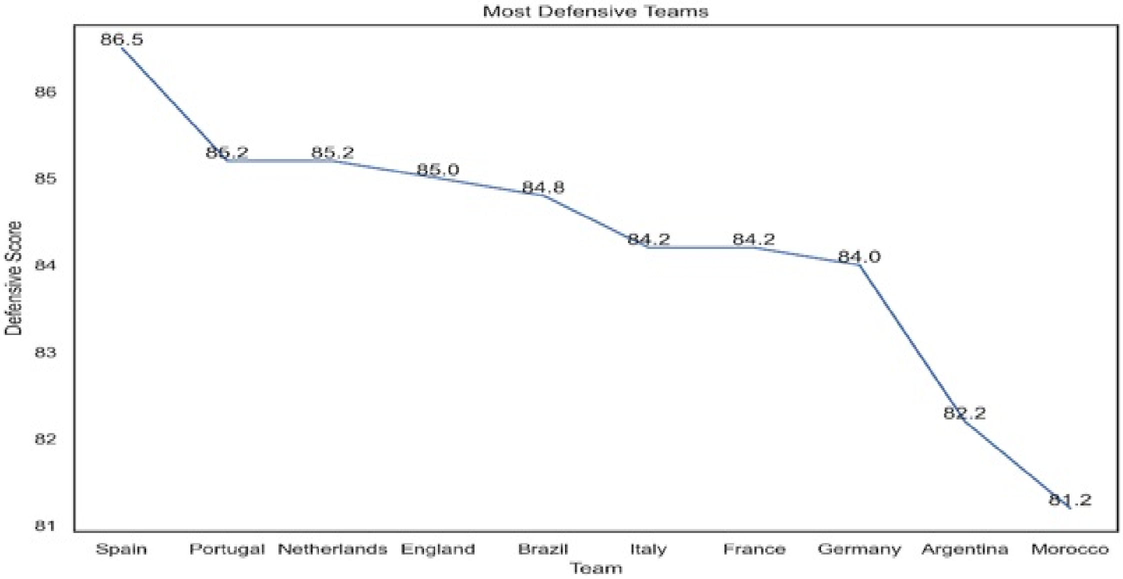


Figure 4: Teams with the greatest defensive potential

3.3.5 Distribution of the match results

The most prevalent result among the match outcomes considered as the target variable is a Win, accounting for 49.2% of the total, followed by Lose at 28.3%, and then Draws at 22.5% (Figure 5).

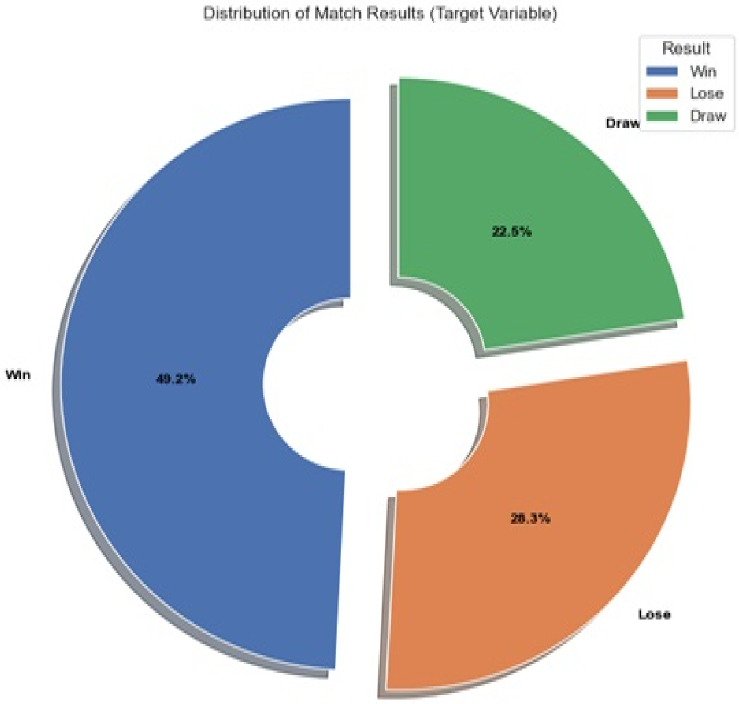
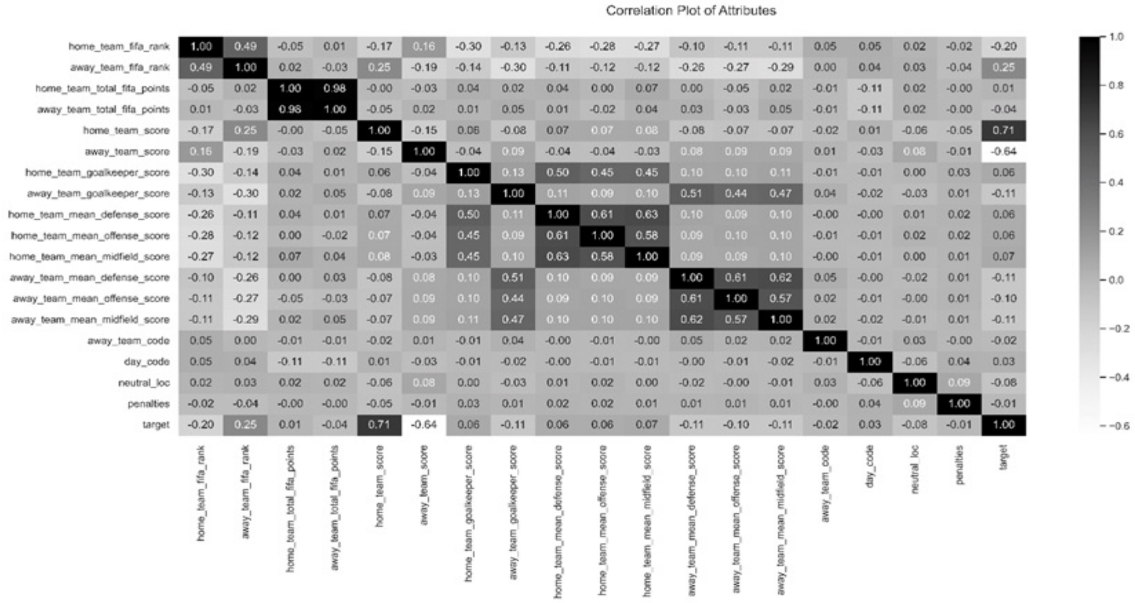


Figure 5: Match results distribution

3.4 Correlation plots

Pairwise correlation of team and match attributes (Figure 6).



	home_team_fifa_rank	away_team_fifa_rank	home_team_total_fifa_points	away_team_total_fifa_points	home_team_score	away_team_score	home_team_goalkeeper_score	away_team_goalkeeper_score	home_team_mean_defense_score	home_team_mean_offense_score	home_team_mean_midfield_score	away_team_mean_defense_score	away_team_mean_offense_score	away_team_mean_midfield_score	away_team_code	day_code	neutral_loc	penalties	target
home_team_fifa_rank	1.00	0.49	-0.05	0.01	-0.17	0.16	-0.30	-0.13	-0.26	-0.28	-0.27	-0.10	-0.11	-0.11	0.05	0.05	0.02	-0.02	-0.20
away_team_fifa_rank	0.49	1.00	0.02	-0.03	0.25	-0.19	-0.14	-0.30	-0.11	-0.12	-0.12	-0.26	-0.27	-0.29	0.00	0.04	0.03	-0.04	0.25
home_team_total_fifa_points	-0.05	0.02	1.00	0.98	-0.00	-0.03	0.04	0.02	0.04	0.00	0.07	0.00	-0.05	0.02	-0.01	-0.11	0.02	-0.00	0.01
away_team_total_fifa_points	0.01	-0.03	0.98	1.00	-0.05	0.02	0.01	0.05	0.01	-0.02	0.04	0.03	-0.03	0.05	-0.01	-0.11	0.02	-0.00	-0.04
home_team_score	-0.17	0.25	-0.00	-0.05	1.00	-0.15	0.06	-0.08	0.07	0.07	0.08	-0.08	-0.07	-0.07	-0.02	0.01	-0.06	-0.05	0.71
away_team_score	0.16	-0.19	-0.03	0.02	-0.15	1.00	-0.04	0.09	-0.04	-0.04	-0.03	0.08	0.09	0.09	0.01	-0.03	0.08	-0.01	-0.64
home_team_goalkeeper_score	-0.30	-0.14	0.04	0.01	0.06	-0.04	1.00	0.13	0.50	0.45	0.45	0.10	0.10	0.11	-0.01	-0.01	0.00	0.03	0.06
away_team_goalkeeper_score	-0.13	-0.30	0.02	0.05	-0.08	0.09	0.13	1.00	0.11	0.09	0.10	0.51	0.44	0.47	0.04	-0.02	-0.03	0.01	-0.11
home_team_mean_defense_score	-0.26	-0.11	0.04	0.01	0.07	-0.04	0.50	0.11	1.00	0.61	0.63	0.10	0.09	0.10	-0.00	-0.00	0.01	0.02	0.06
home_team_mean_offense_score	-0.28	-0.12	0.00	-0.02	0.07	-0.04	0.45	0.09	0.61	1.00	0.58	0.09	0.10	0.10	-0.01	-0.01	0.02	0.02	0.06
home_team_mean_midfield_score	-0.27	-0.12	0.07	0.04	0.08	-0.03	0.45	0.10	0.63	0.58	1.00	0.09	0.09	0.10	-0.00	-0.01	0.00	0.01	0.07
away_team_mean_defense_score	-0.10	-0.26	0.00	0.03	-0.08	0.08	0.10	0.51	0.10	0.09	0.09	1.00	0.61	0.62	0.05	-0.00	-0.02	0.01	-0.11
away_team_mean_offense_score	-0.11	-0.27	-0.05	-0.03	-0.07	0.09	0.10	0.44	0.09	0.10	0.10	0.61	1.00	0.57	0.02	-0.01	-0.00	0.01	-0.10
away_team_mean_midfield_score	-0.11	-0.29	0.02	0.05	-0.07	0.09	0.11	0.47	0.10	0.10	0.10	0.62	0.57	1.00	0.02	-0.02	-0.01	0.01	-0.11
away_team_code	0.05	0.00	-0.01	-0.01	-0.02	0.01	-0.01	0.04	-0.00	-0.01	-0.00	0.05	0.02	0.02	1.00	-0.01	0.03	-0.00	-0.02
day_code	0.05	0.04	-0.11	-0.11	0.01	-0.03	-0.01	-0.02	-0.00	-0.01	-0.01	-0.00	-0.01	-0.02	-0.01	1.00	-0.06	0.04	0.03
neutral_loc	0.02	0.03	0.02	0.02	-0.06	0.08	0.00	-0.03	0.01	0.02	0.00	-0.02	-0.00	-0.01	0.03	-0.06	1.00	0.09	-0.08
penalties	-0.02	-0.04	-0.00	-0.00	-0.05	-0.01	0.03	0.01	0.02	0.02	0.01	0.01	0.01	0.01	-0.00	0.04	0.09	1.00	-0.01
target	-0.20	0.25	0.01	-0.04	0.71	-0.64	0.06	-0.11	0.06	0.06	0.07	-0.11	-0.10	-0.11	-0.02	0.03	-0.08	-0.01	1.00

Figure 6: Pairwise correlation of team and match attributes

3.5 Machine learning model development

Machine learning models are developed to predict World Cup outcomes. The logistic regression model and CatBoost, a Gradient Boosting Tree model, are explored as techniques for predictive models in the multi-class classification task.

3.5.1 Logistic regression

Firstly, a list of features is created to include the attributes chosen for analysis. These attributes include various aspects related to both home and away teams, such as FIFA rankings, scores, tournament details, and geographical information (See Appendix A for the Python codes). After identifying and separating categorical and numerical features, the categorical features were encoded to convert them into a numerical format suitable for the logistic regression model. Target encoding was used to encode the categorical features, replacing categorical values with the mean of the target variable in each category. This encoding method allows the logistic regression model to effectively use the features in its calculations (See Appendix B for the Python codes). The data was split into training and testing sets with a 20% test size using 'train_test_split'. Stratification based on the target variable maintained class distribution, and 'random_state=42' ensured reproducibility. This allows for evaluating the model's performance on unseen data. After training, predictions were made on the testing data using the 'predict' method, resulting in y_pred. (See Appendix C for the Python codes)

3.5.2 Logistic regression model evaluation

The model was evaluated using several metrics to assess its performance. The accuracy, precision, recall, and F1 score were calculated using the 'accuracy_score', 'precision_score', 'recall_score', and 'f1_score' functions, respectively. These metrics provide insights into the model's overall effectiveness in predicting World Cup outcomes. The obtained results indicate a high level of performance, with an accuracy of approximately 99.2685%, precision of 99.2720%, recall of 99.2685%, and F1 score of 99.2692% (See Appendix D for the Python codes). These metrics demonstrate the model's exceptional ability to accurately classify World Cup outcomes based on the provided attributes. The study used a confusion matrix to provide insight into the performance of the predictive model. In the case of the logistic regression model, the confusion matrix reveals that it accurately classified 1,343 match outcomes as Wins, 1,078 match outcomes as Draws, and 2,329 match outcomes as Losses. However, the model also made errors by incorrectly predicting 24 match outcomes as Wins when they were actually Losses, and 11 match outcomes as Losses when they were actually Wins (Figure 7). These impressive findings serve as strong evidence of the model's predictive prowess.

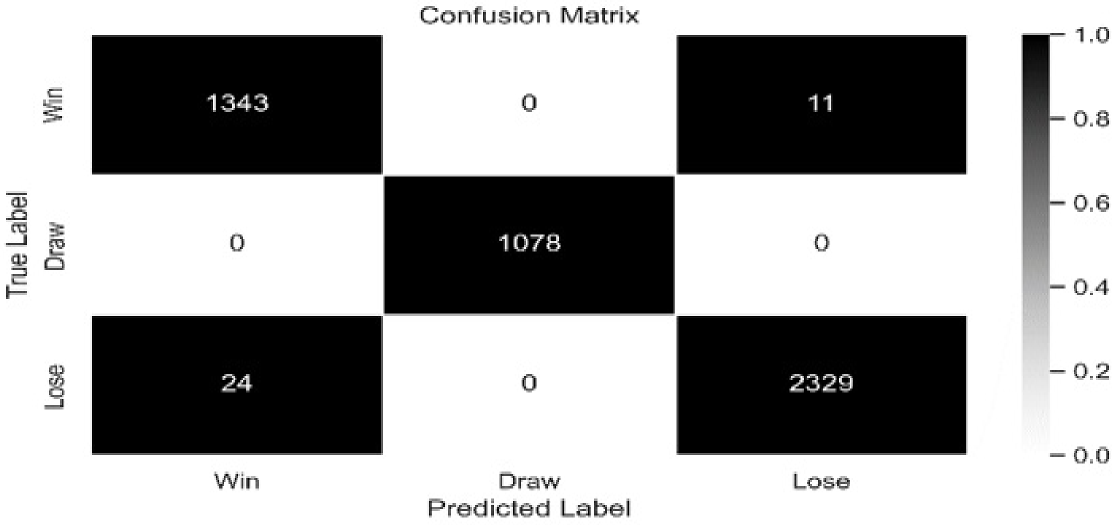
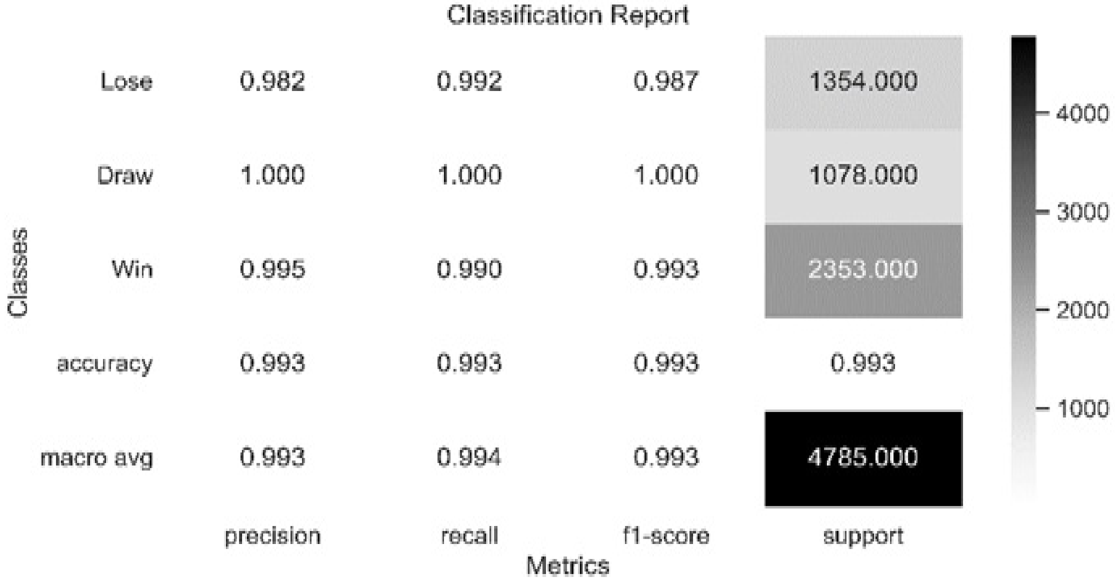


Figure 7: Confusion matrix plot to provide more insights into the model's predictive capabilities

A classification report is a text report that provides a detailed overview of the model's performance across different classes (Win, Draw, and Lose). The report includes metrics such as precision, recall, F1-score, and support for each class, offering a good understanding of the model's predictive capabilities (Figure 8).


Classification Report
Classes
Metrics
4000
3000
2000
1000

	precision	recall	f1-score	support
Lose	0.982	0.992	0.987	1354.000
Draw	1.000	1.000	1.000	1078.000
Win	0.995	0.990	0.993	2353.000
accuracy	0.993	0.993	0.993	0.993
macro avg	0.993	0.994	0.993	4785.000

Figure 8: Classification report

3.5.3 CatBoost

The study also explored CatBoost, a powerful machine learning algorithm designed for a variety of classification tasks, such as binary or multiclass classification problems, predicting outcomes across different categories like Win, Loss, or Draw in events such as the World Cup. The same list of features created for the Logistic Regression model was used for the

CatBoost model. The data was then split into training and testing sets with a 20% test size using 'train_test_split'. Stratification based on the target variable maintained class distribution, and 'random_state=42' ensured reproducibility. The CatBoost classifier was used to train the model, and the model was fitted to the training data, with its performance evaluated on both the training and testing sets. To prevent overfitting and ensure optimal model generalization, the early stopping mechanism was employed with a tolerance of 20 rounds. After training, evaluation results were obtained to assess the model's performance. The training and validation log loss metrics were recorded, and the iteration with the minimum validation log loss was identified as the best iteration. The corresponding validation log loss at this iteration was also noted. These results provide insights into the model's effectiveness in predicting World Cup outcomes, with lower log loss values indicating better performance (See Appendix E for the Python codes). The model was retrained with the optimal early stopping rounds, 629. Using the CatBoost Classifier with specified hyperparameters, it was fitted to the training data. Performance was evaluated on both training and testing sets. Early stopping was set to the best iteration found previously (See Appendix F for the Python codes).

3.5.4 CatBoost model evaluation

The model was evaluated using several metrics to assess its performance. The accuracy, precision, recall, and F1 scores were calculated using the 'accuracy_score', 'precision_score', 'recall_score', and 'F1_score' functions, respectively. These metrics provide insights into the model's overall effectiveness in predicting World Cup outcomes. The obtained results indicate a high level of performance, with an accuracy of approximately 99.1431%, precision of 99.1603%, recall of 99.1431%, and F1 score of 99.1452%. These metrics demonstrate the model's exceptional ability to accurately classify World Cup outcomes based on the provided attributes. The confusion matrix, as a model evaluation tool, provides insight into the performance of a predictive model. In the case of the CatBoost model, the confusion matrix reveals that it accurately classified 2,316 match outcomes as Wins, 1,078 match outcomes as Draws, and 1,350 match outcomes as Losses. However, the model also made errors by incorrectly predicting 4 match outcomes as Wins when they were actually Losses, and 37 match outcomes as Losses when they were actually Wins (Figure 9). These impressive findings serve as strong evidence of the model's predictive prowess.

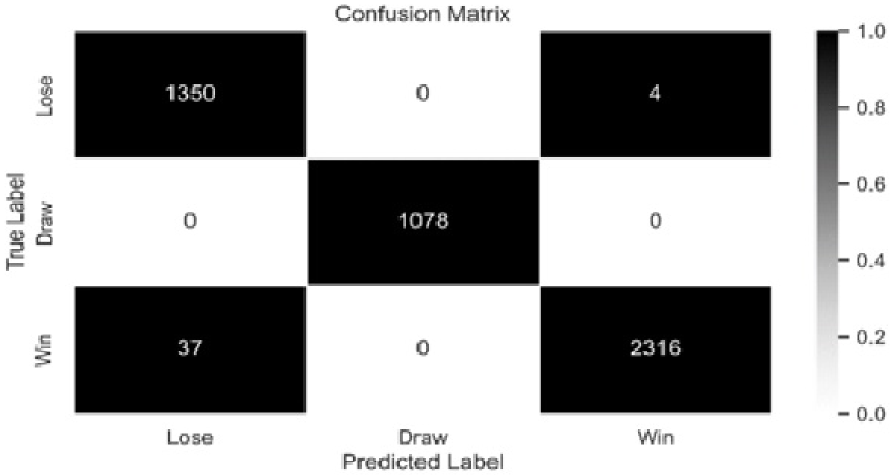
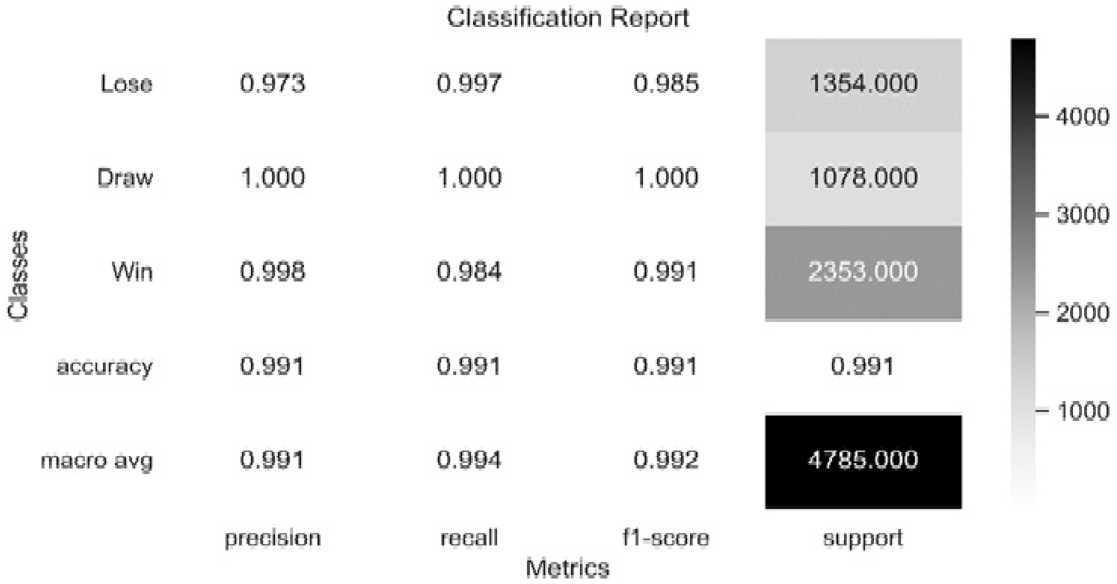


Figure 9: CatBoost model confusion matrix

A classification report is a text report that provides a detailed overview of the model's performance across different classes (Win, Draw, and Lose). The report includes metrics such as precision, recall, F1-score, and support for each class, offering a good understanding of the model's predictive capabilities (Figure 10).



Classification Report

Classes	precision	recall	f1-score	support
Lose	0.973	0.997	0.985	1354.000
Draw	1.000	1.000	1.000	1078.000
Win	0.998	0.984	0.991	2353.000
accuracy	0.991	0.991	0.991	0.991
macro avg	0.991	0.994	0.992	4785.000

Metrics
4000
3000
2000
1000

Figure 10: CatBoost model classification report

3.6 Outcomes of the predictions made by the machine learning models

It is crucial to recognize that there are constraints in accurately forecasting the victor of the World Cup, owing to various factors:

(i) Absence of team qualifying match data: The absence of team qualifying match data impacted the prediction model's ability to accurately determine the eventual winner of the 2026 World Cup. This data, which plays a crucial role in assessing the teams' performance and capabilities, was unfortunately unavailable to the model.

(ii) Absence of knockout stage data: The prediction model lacked data on the group stage and knockout stage, a key factor in determining the ultimate champion.

(iii) Other unaccounted factors: There might be other significant factors influencing the outcome of the World Cup that were not incorporated into the model.

3.6.1 Identifying the top contenders for the World Cup

The machine learning model employed a data-driven approach to identify potential contenders for the 2026 World Cup. The model pinpointed the leading contenders for the 2026 FIFA World Cup, listing them in order of superiority as: (i) Argentina (ii) Brazil (iii) Spain (iv) France (v) The Netherlands (See Appendix G for the Python codes).

4 Conclusion

In this study, the focus was on two specific machine learning algorithms: Bayesian logistic regression and CatBoost. Both models demonstrated an impressive ability to distinguish between win, loss, and draw results in World Cup matches. This high level of accuracy highlights the effectiveness of machine learning in extracting valuable insights from historical data and using them to make actionable predictions in the field of sports analytics. The study identified top contenders, highlighted offensive and defensive strengths, and noted feature importance. The findings emphasize the potential of machine learning in sports analytics. The results identified the leading contenders for the 2026 FIFA World Cup, listing them in order of superiority. The results aim to contribute to the field of sports analytics, offering valuable insights into the complex dynamics influencing success in high-stakes football tournaments. The results obtained confirmed the effectiveness of the developed models in tackling the challenge of predicting the FIFA World Cup outcomes. The outcome of this study would be very useful and add value to the sports analytics industry, including sports analysts, sports enthusiasts, betting agencies, team management, etc.

From the literature, this study on the application of Bayesian logistic regression and gradient boosting models is one of the rare applications to sports analytics. Future research stemming from this study could involve predicting the best player and goalkeeper awards among the potential winners of the 2026 FIFA World Cup, using Bayesian logistic regression and CatBoost models.

Declarations

Funding

There are no funding sources to declare for this study.

Authors' contributions

All the authors have contributed equally to this paper.

Conflict of interest

The authors have no conflicts of interest to declare.

References

[1] G. Baio and M. A. Blangiardo, *Bayesian Hierarchical Model for the Prediction of Football Results,* Journal of Applied Statistics, **35**(2) (2010), 253–264. DOI

[2] R. P. Bunker and F. Thabtah, *A Machine Learning Framework for Sport Result Prediction,* Applied Computing and Informatics, **15**(1) (2019), 27–33. DOI

[3] A. C. Constantinou and N. E. Fenton, *Pi-Football: A Bayesian Network Model for Forecasting Association Football Match Outcomes,* Knowledge-Based Systems, **36** (2012), 322–339. DOI

[4] A. C. Constantinou and N. E. Fenton, *Determining the Level of Ability of Football Teams by Dynamic Ratings Based on the Relative Discrepancies in Scores Between Adversaries,* Journal of Quantitative Analysis in Sports, **9**(1) (2013), 37–50. DOI

[5] A. C. Constantinou, *Dolores: A Model that Predicts Football Match Outcomes from All Over the World,* Machine Learning, **108**(10) (2019), 1007. DOI

[6] N. Danisik, P. Lacko and M. Farkas, *Football Match Prediction Using Players Attributes,* (2018), 201–206. Link

[7] R. Giulianotti and R. Stebbins, *Football: A Sociology of the Global Game,* Contemporary Sociology, **29** (2000), 842. Link

[8] M. Gifford and T. Bayrak, *A Predictive Analytics Model for Forecasting Outcomes in the National Football League Games Using Decision Tree and Logistic Regression,* Decision Analytics Journal, **8** (2023), 100296. DOI

[9] D. Goldblatt, *The Ball is Round: A Global History of Soccer,* Penguin Publishing Group, (2008), 1008. Link

[10] T. Horvat and J. Job, *The Use of Machine Learning in Sport Outcome Prediction: A Review,* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, (2020). DOI

[11] M. Hughes and J. Franks, *Analysis of Passing Sequences, Shots and Goals in Soccer,* Journal of Sports Sciences, **23**(5) (2005), 509–514. DOI

[12] C. Igiri, *An Improved Prediction System for Football Match Result,* IOSR Journal of Engineering, **4** (2014), 12–20. Link

[13] M. Ingram, *A Point-Based Bayesian Hierarchical Model to Predict the Outcome of Tennis Matches,* Journal of Quantitative Analysis in Sports, **15** (2019). DOI

[14] D. Johansen, C. Gurrin and D. Havard, *Towards Consent-Based Lifelogging in Sport Analytic,* (2015), 335–344. DOI

[15] J. Ma, F. Stingo, and B. Hobbs, *Bayesian Predictive Modeling for Genomic Based Personalized Treatment Selection,* Journal of Biometrics, **72** (2015). DOI

[16] V. Matheson, *Mega-Events: The Effect of the World's Biggest Sporting Events on Local, Regional and National Economies,* International Association of Sports Economists, Working Papers (2006). Link

[17] B. Min, J. Kim and H. Eom, *A Compound Framework for Sports Results Prediction: A Football Case Study,* Knowledge-Based Systems, **21** (2008), 551–562. DOI

[18] B. Olawale and M. Oladapo, *Bayesian Analysis of 2014 FIFA World Cup Matches Played and Goals Scored,* International Journal of Modern Mathematical Sciences, **16**(1) (2018), 25–36. Link

[19] F. Owramipur, *Football Result Prediction with Bayesian Network in Spanish League,* International Journal of Computer Theory and Engineering, **5** (2013), 812–815. Link

[20] N. Razili and A. Mustapha, *Predicting Football Matches Result Using Bayesian Networks for English Premier League (EPL)*, IOP Conference Series: Materials Science and Engineering, **226**(1) (2017), 012099. DOI

[21] C. Sandvoss, *A Game of Two Halves: Football Fandom, Television and Globalization*, (2003). DOI

[22] M. Stein, *Bring it to the Pitch: Combining Video and Movement Data to Enhance Team Sport Analysis*, IEEE Transactions on Visualization and Computer Graphics, (2017). DOI

[23] J. Sugden and A. Tomlinson, *FIFA and The Contest for World Football: Who Rules the People's Game?*, (1998). Link

[24] A. Suzuki, *A Bayesian Approach for Predicting Match Outcomes: The 2006 (Association) football world cup*, Journal of the Operational Research Society, **61**(10) (2010), 1530–1539. Link

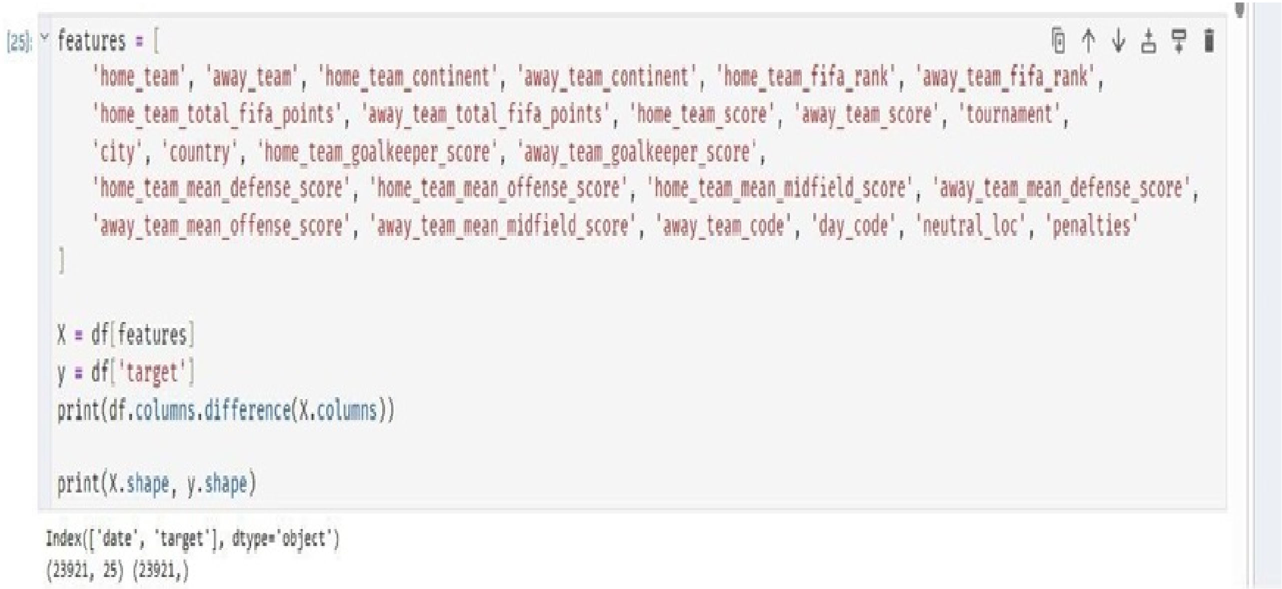
[25] S. Vaidya, *Football Match Winner Prediction*, International Journal of Computer Applications, **154** (2016), 31–33. Link

[26] M. Veenman, A. M. Stefan and J. Haaf, *Bayesian Hierarchical Modeling: An Introduction and Reassessment*, Behavior Research Methods, **56**(1) (2023). DOI

[27] F. Wunderlich and D. Memmert, *Forecasting the Outcomes of Sports Events: A Review*, European Journal of Sport Science, **21**(7) (2020), 944–957. DOI

Appendix A

Creation of the features and target variables for building the model

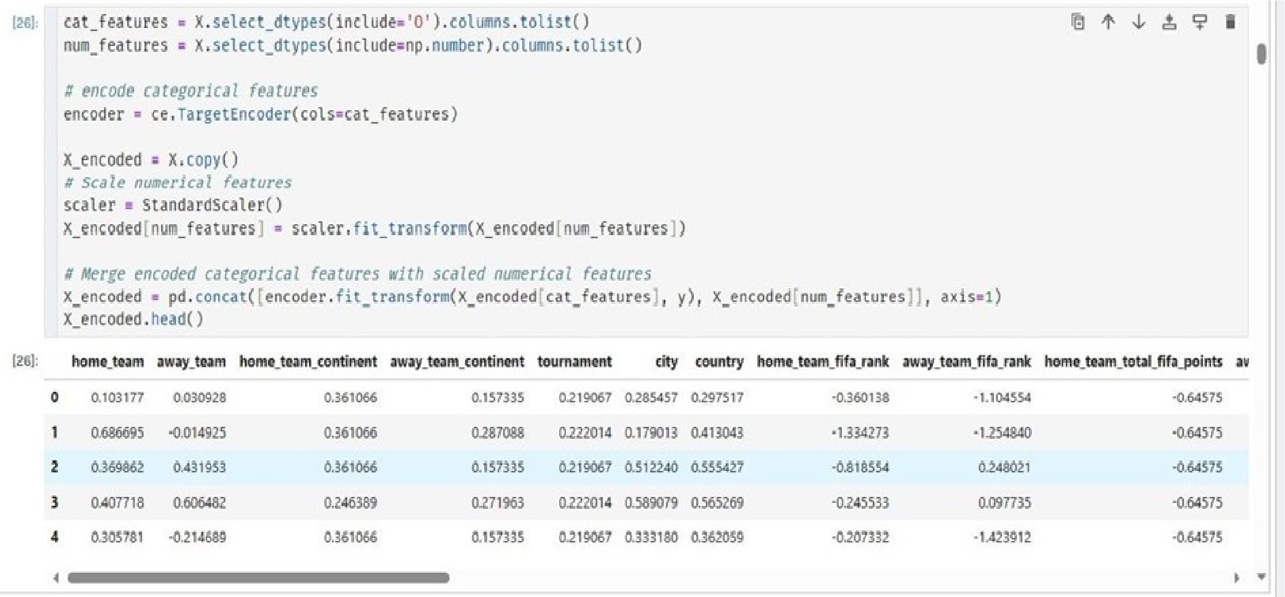


```
features = [
    'home_team', 'away_team', 'home_team_continent', 'away_team_continent', 'home_team_fifa_rank', 'away_team_fifa_rank',
    'home_team_total_fifa_points', 'away_team_total_fifa_points', 'home_team_score', 'away_team_score', 'tournament',
    'city', 'country', 'home_team_goalkeeper_score', 'away_team_goalkeeper_score',
    'home_team_mean_defense_score', 'home_team_mean_offense_score', 'home_team_mean_midfield_score', 'away_team_mean_defense_score',
    'away_team_mean_offense_score', 'away_team_mean_midfield_score', 'away_team_code', 'day_code', 'neutral_loc', 'penalties'
]

X = df[features]
y = df['target']
print(df.columns.difference(X.columns))

print(X.shape, y.shape)
```

```
Index(['date', 'target'], dtype='object')
(23921, 25) (23921,)
```

Appendix B

Encoding and scaling the data before applying logistic regression

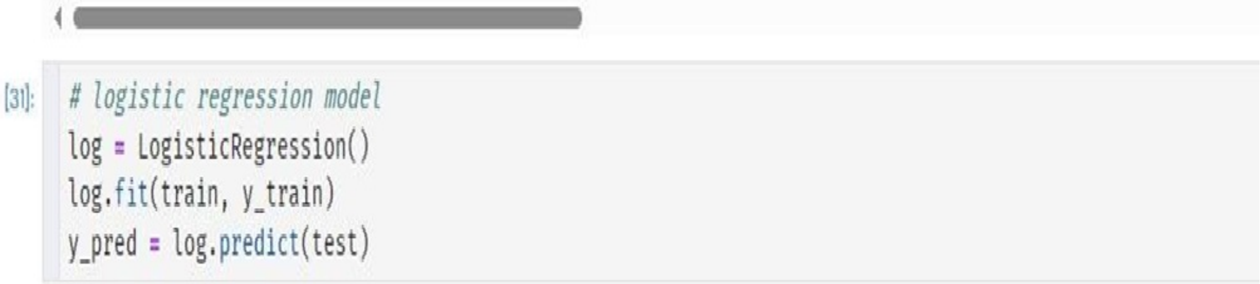


```
[26]: cat_features = X.select_dtypes(include='O').columns.tolist()
      num_features = X.select_dtypes(include=np.number).columns.tolist()

      # encode categorical features
      encoder = ce.TargetEncoder(cols=cat_features)

      X_encoded = X.copy()
      # Scale numerical features
      scaler = StandardScaler()
      X_encoded[num_features] = scaler.fit_transform(X_encoded[num_features])

      # Merge encoded categorical features with scaled numerical features
      X_encoded = pd.concat([encoder.fit_transform(X_encoded[cat_features], y), X_encoded[num_features]], axis=1)
      X_encoded.head()
```

```
[26]:   home_team  away_team  home_team_continent  away_team_continent  tournament      city   country  home_team_fifa_rank  away_team_fifa_rank  home_team_total_fifa_points  av
     0   0.103177   0.030928             0.361066             0.157335    0.219067  0.285457  0.297517            -0.360138            -1.104554                     -0.64575
     1   0.686695  -0.014925             0.361066             0.287088    0.222014  0.179013  0.413043            -1.334273            -1.254840                     -0.64575
     2   0.369862   0.431953             0.361066             0.157335    0.219067  0.512240  0.555427            -0.818554             0.248021                     -0.64575
     3   0.407718   0.606482             0.246389             0.271963    0.222014  0.589079  0.565269            -0.245533             0.097735                     -0.64575
     4   0.305781  -0.214689             0.361066             0.157335    0.219067  0.333180  0.362059            -0.207332            -1.423912                     -0.64575
```

Appendix C

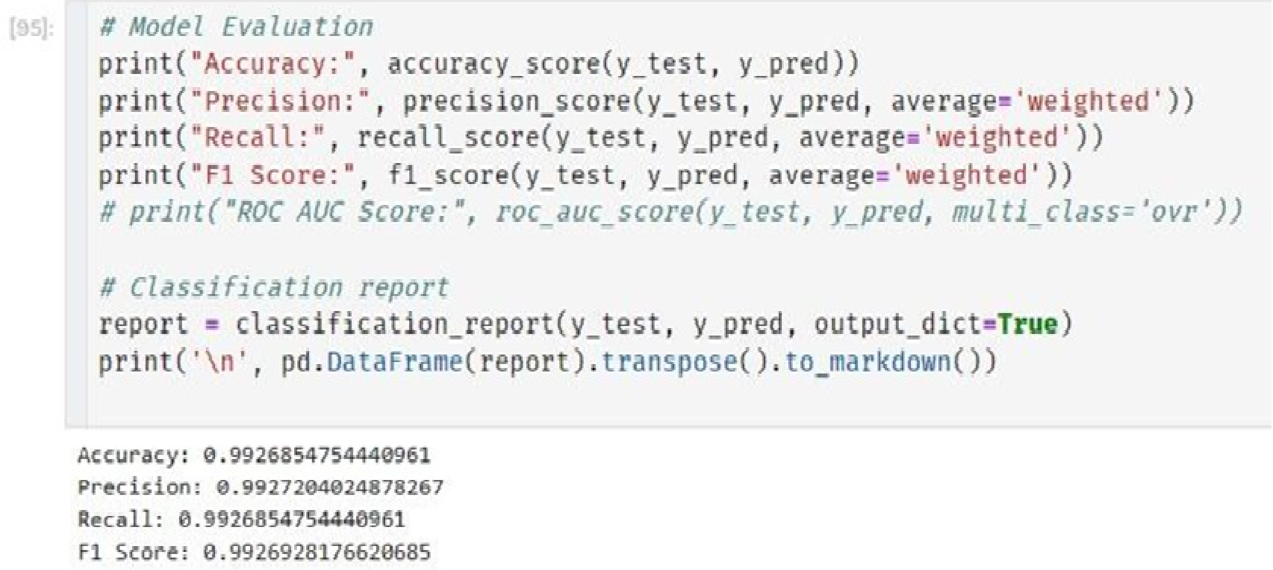
Calling the logistic regression model and fitting it to the data before making predictions on the test set



```
[31]: # logistic regression model
      log = LogisticRegression()
      log.fit(train, y_train)
      y_pred = log.predict(test)
```

Appendix D

Model performance using various classification evaluation metrics

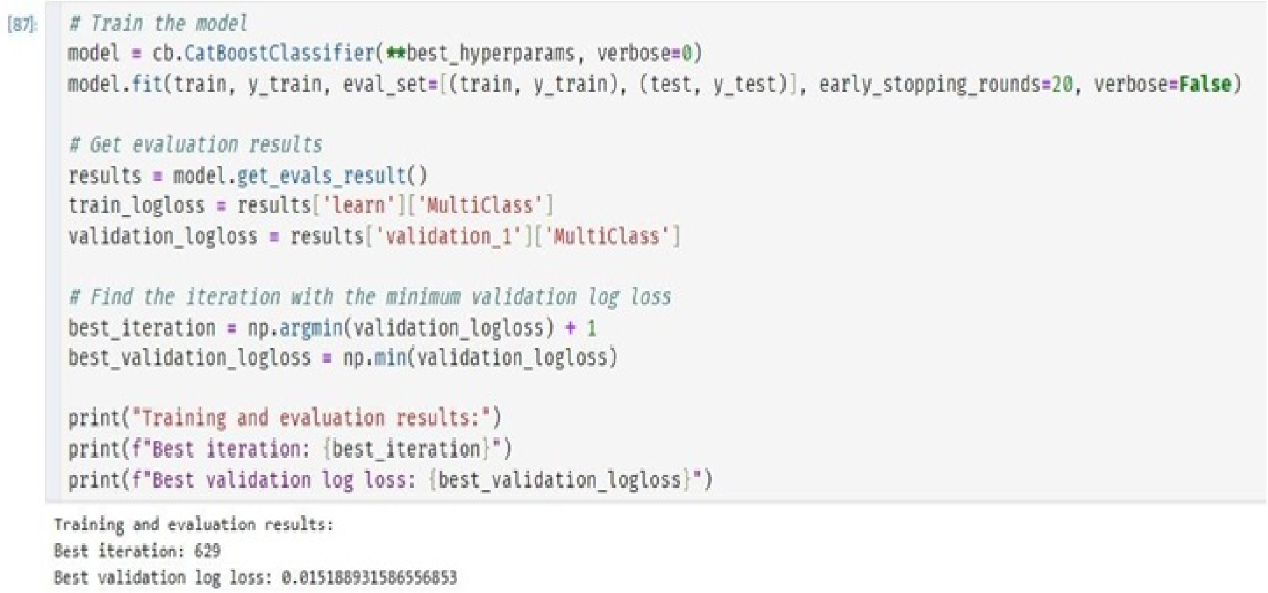


```
[95]:
# Model Evaluation
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred, average='weighted'))
print("Recall:", recall_score(y_test, y_pred, average='weighted'))
print("F1 Score:", f1_score(y_test, y_pred, average='weighted'))
# print("ROC AUC Score:", roc_auc_score(y_test, y_pred, multi_class='ovr'))

# Classification report
report = classification_report(y_test, y_pred, output_dict=True)
print('\n', pd.DataFrame(report).transpose().to_markdown())
```

```
Accuracy: 0.9926854754440961
Precision: 0.9927204024878267
Recall: 0.9926854754440961
F1 Score: 0.9926928176620685
```

Appendix E

Training and evaluating the CatBoost algorithm using the iteration with minimum validation log loss

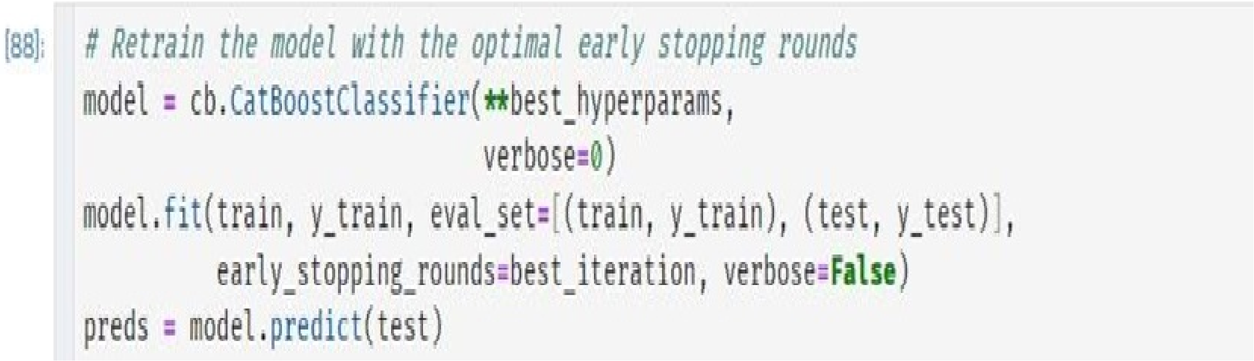


```
[87]:
# Train the model
model = cb.CatBoostClassifier(**best_hyperparams, verbose=0)
model.fit(train, y_train, eval_set=[(train, y_train), (test, y_test)], early_stopping_rounds=20, verbose=False)

# Get evaluation results
results = model.get_evals_result()
train_logloss = results['learn']['MultiClass']
validation_logloss = results['validation_1']['MultiClass']

# Find the iteration with the minimum validation log loss
best_iteration = np.argmin(validation_logloss) + 1
best_validation_logloss = np.min(validation_logloss)

print("Training and evaluation results:")
print(f"Best iteration: {best_iteration}")
print(f"Best validation log loss: {best_validation_logloss}")
```

```
Training and evaluation results:
Best iteration: 629
Best validation log loss: 0.015188931586556853
```

Appendix F

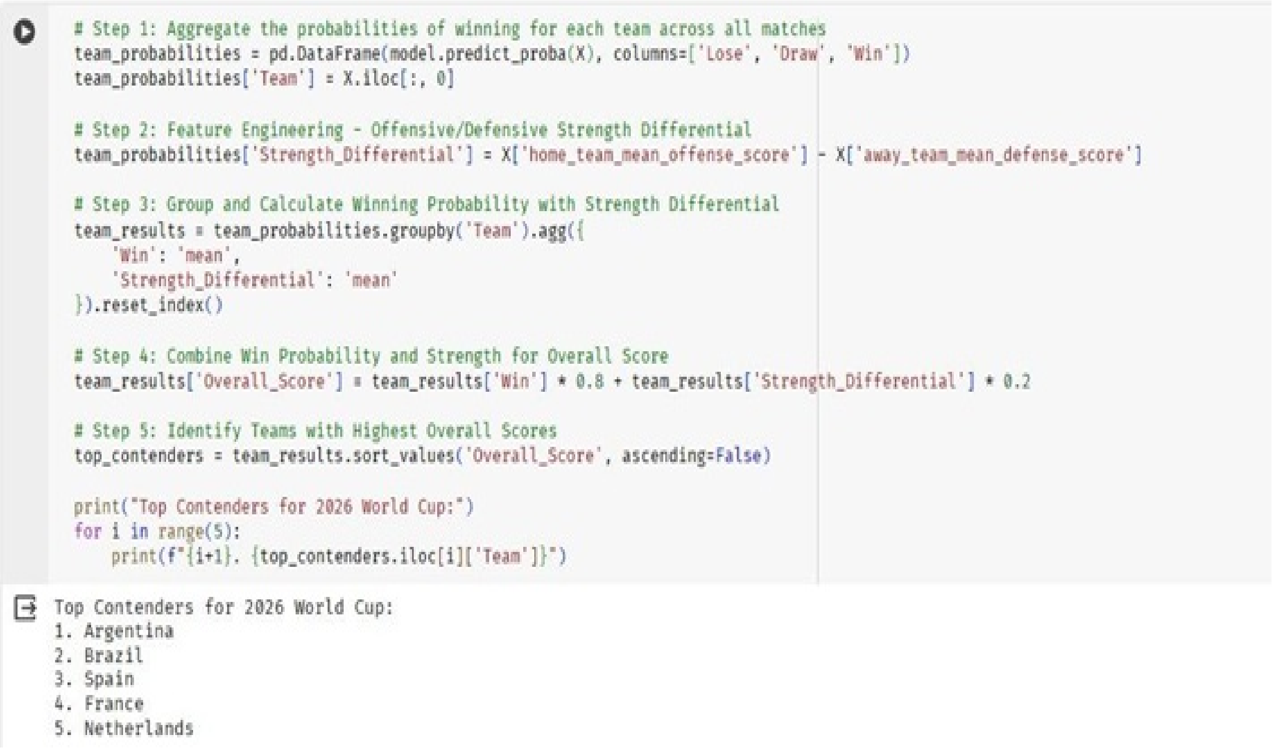
Retraining the CatBoost model with the best iteration obtained from the validation log loss



```
[88]: # Retrain the model with the optimal early stopping rounds
      model = cb.CatBoostClassifier(**best_hyperparams,
                                    verbose=0)
      model.fit(train, y_train, eval_set=[(train, y_train), (test, y_test)],
                early_stopping_rounds=best_iteration, verbose=False)
      preds = model.predict(test)
```

Appendix G

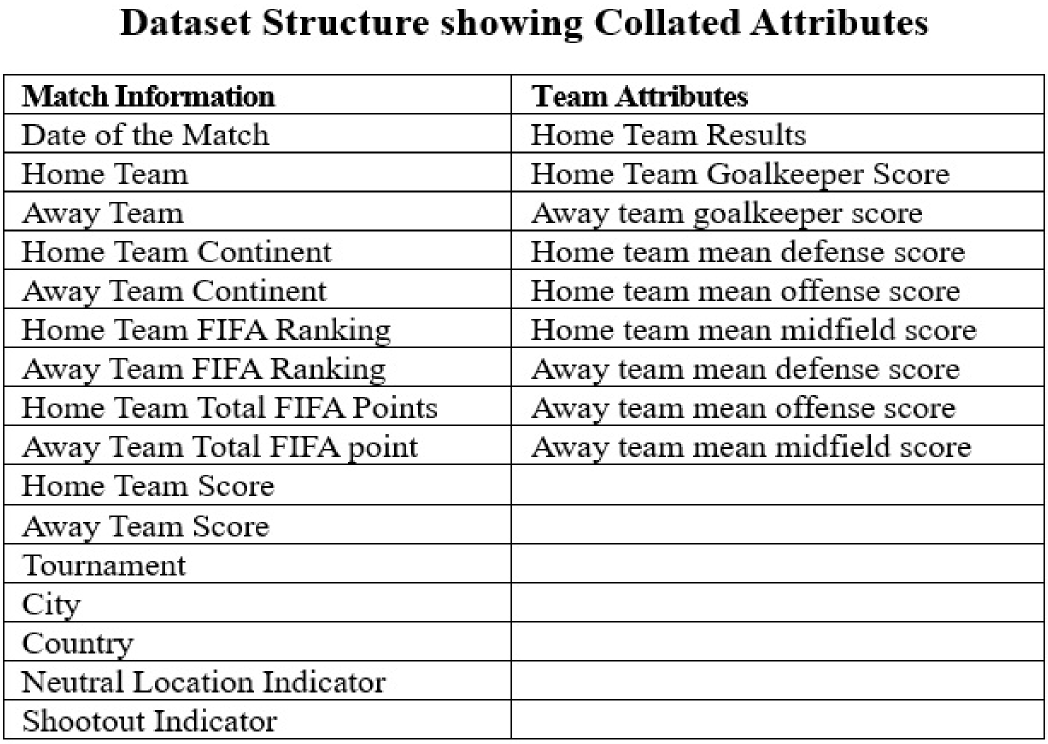
Retraining the CatBoost model with the best iteration obtained from the validation log loss



```
# Step 1: Aggregate the probabilities of winning for each team across all matches
team_probabilities = pd.DataFrame(model.predict_proba(X), columns=['Lose', 'Draw', 'Win'])
team_probabilities['Team'] = X.iloc[:, 0]

# Step 2: Feature Engineering - Offensive/Defensive Strength Differential
team_probabilities['Strength_Differential'] = X['home_team_mean_offense_score'] - X['away_team_mean_defense_score']

# Step 3: Group and Calculate Winning Probability with Strength Differential
team_results = team_probabilities.groupby('Team').agg({
    'Win': 'mean',
    'Strength_Differential': 'mean'
}).reset_index()

# Step 4: Combine Win Probability and Strength for Overall Score
team_results['Overall_Score'] = team_results['Win'] * 0.8 + team_results['Strength_Differential'] * 0.2

# Step 5: Identify Teams with Highest Overall Scores
top_contenders = team_results.sort_values('Overall_Score', ascending=False)

print("Top Contenders for 2026 World Cup:")
for i in range(5):
    print(f"{i+1}. {top_contenders.iloc[i]['Team']}")
```

```
Top Contenders for 2026 World Cup:
1. Argentina
2. Brazil
3. Spain
4. France
5. Netherlands
```

Appendix H



Dataset Structure showing Collated Attributes

Match Information	Team Attributes
Date of the Match	Home Team Results
Home Team	Home Team Goalkeeper Score
Away Team	Away team goalkeeper score
Home Team Continent	Home team mean defense score
Away Team Continent	Home team mean offense score
Home Team FIFA Ranking	Home team mean midfield score
Away Team FIFA Ranking	Away team mean defense score
Home Team Total FIFA Points	Away team mean offense score
Away Team Total FIFA point	Away team mean midfield score
Home Team Score	
Away Team Score	
Tournament	
City	
Country	
Neutral Location Indicator	
Shootout Indicator	