Predicting the 2026 FIFA World Cup Winner through Machine Learning Analysis

Mert Çalışkan
Department of Structural Design
DOLSAR Engineering Inc. Co.
Ankara, Turkiye
mertcaliskan@hacettepe.edu.tr

Faruk Tekbaş
Department of Research & Development
Browder + LeGuizamon & Associates Inc.
Ankara, Turkiye
zubeyirtekbas@hacettepe.edu.tr

***Abstract*—This project presents a comprehensive analysis of machine learning classification models applied to the historical international football games data from August 1993 to June 2022 aiming to predict the winner of the 2026 FIFA World Cup. The methodology involves data pre-processing, feature engineering, and the evaluation of performance metrics for various classification models. The Support Vector Machine was ultimately chosen for its commendable and well-balanced performance. Using the prediction model, the iterative simulation predicts a closely competitive scenario between Belgium and Brazil, with Belgium holding a slightly higher win probability of 50.64%. This study offers valuable insights into predicting the 2026 FIFA World Cup winner, contributing to the understanding of the influences and tournament's potential outcomes.**

Keywords—Machine Learning, Prediction, FIFA World Cup, Simulation

I. INTRODUCTION

The FIFA World Cup, a premier international football tournament held every four years since 1930, captivates audiences worldwide. In the most recent edition, hosted in Qatar from November 20 to December 18, 2022, Argentina emerged victorious. In a thrilling final, they defeated France, the reigning champion of the 2018 FIFA World Cup, with a victory in a penalty shootout following an intense 3-3 draw.

As the entire world is looking forward to the upcoming 23rd edition, which is the 2026 FIFA World Cup, it has been officially announced that the event is scheduled to take place from June to July 19, 2026, in 16 host cities across the three North American countries: Canada, Mexico and the United States [1]. The competition will feature 48 teams instead of the standard 32 for the first time in World Cup history [2].

While the announcement of qualifying teams and groups for the 2026 FIFA World Cup is pending, the prospect of predicting the next champion using machine learning techniques on historical data introduces an intriguing aspect to this globally followed event.

Motivated by these, the primary objectives for this project are to:

- Utilize machine learning techniques to analyse historical data of international games and develop predictive models for determining the winner of 2026 FIFA World Cup
- Conduct a comprehensive evaluation of the developed prediction models, identifying and selecting the best-performing model through assessments of key performance metrics.

II. LITERATURE REVIEW

This section explores previous research on FIFA World Cup prediction highlighting two innovative studies that use distinct methodologies.

One study utilized a radial basis function neural network to predict soccer match results and attribute sensitivities [3]. The research dives into the complexities of big data gathered from sensors, cameras, and analysis systems. The neural network model demonstrated remarkable predictive capabilities, achieving an accuracy of 83.3% for wins and 72.7% for losses. Notably, the study identified 19 influential predictors among the 75 analysed match attributes. In practical terms, it suggests a shift away from conventional data mining tools towards a more distinct understanding of the complex relationships between match attributes and outcomes. The proposed neural network model emerges as a powerful tool, not only for predicting match results but also for highlighting the attribute sensitivities that are crucial for performance optimization.

Another interesting study focused on utilizing social media for accurate predictions during the 2018 tournament [4]. The research collected a large dataset of 38,371,358 tweets during the FIFA World Cup 2018 to propose a prediction system that evaluates teams based on squad and early-stage performances. Using various machine learning algorithms, the study achieved an impressive 87.5% accuracy for game result predictions, confirming the effectiveness of integrating social media data. With the use of social media for match predictions, the study underscored the efficacy of analysing user-generated content on platforms like Twitter. Unlike related works, the model demonstrated high accuracy with a minimal set of features, enhancing accessibility. The approach showcased adaptability for addressing other prediction problems in social networks, emphasizing its versatility. The study concludes by underlining the transformative impact of social media analysis on predicting World Cup match results.

Parallel to our research, these studies underscore the diverse methodologies employed in FIFA World Cup prediction. While one focuses on intricate data analysis and attribute sensitivities, the other explores the predictive potential of social media. In our pursuit, we aim to integrate and advance upon these methodologies, employing a comprehensive approach that incorporates historical data, machine learning models, and feature engineering to enhance the accuracy and depth of our FIFA World Cup outcome predictions.

III. Data Collection

A. Data Source

The dataset used for this project was retrieved from Kaggle, an online data science and machine learning platform [5]. The dataset offers a comprehensive record of international football games played between August 1993 and June 2022, incorporating a total of 23,921 games and 25 crucial metrics for each international game:

- Date: The specific date (day/month/year) of the game, captured in the 'date' column.
- Teams and Their Continents: Information on home and away teams, including the continents they represent. This data is stored in the columns 'home_team', 'away_team', 'home_team_contintent' and 'away_team_continent'.
- FIFA Ranks: The global standing of each team at the time of the game, represented by 'home_team_fifa_rank' and 'away_team_fifa_rank'.
- Total FIFA Points: Cumulative points of each team in the FIFA raking system, available in the columns 'home_team_total_fifa_points' and 'away_team_total_fifa_points'.
- Goals Scored: The number of goals scored by each team during the game, recorded in 'home_team_score' and 'away_team_score'.
- Tournament: The specific tournament in which the game was played, identified by the 'tournament' column.
- Location: Details about the city and country where each game occurred, indicating whether it was a neutral location for each team. This information is stored in the columns 'city', 'country' and 'neutral_location'.
- Penalty Shootouts: An indicator of whether the game extended to penalty shootouts, denoted by 'shoot_out' column.
- Result: The outcome of each game for the home team, provided in 'home_team_result' column.
- Average Player Position Strengths: The average strength of highest-ranked players positions such as goalkeeper, defence, midfield, and offense for each team. These values are contained in columns such as 'goalkeeper_score', 'defense_score', 'midfield_score' and 'offense_score'.

This rich and diverse dataset offers detailed insights into a multitude of factors, including team rankings, position player performances, game outcomes, and other key metrics. These aspects collectively contribute to the prediction of the 2026 FIFA World Cup champion based on historical data."

B. Data Pre-processing

After importing into Python, the dataset went through several pre-processing techniques to ensure data integrity and improve usability.

Duplicate data was carefully examined to ensure uniqueness of the entire dataset. The 'date' column was then transformed from a string format to datetime. Furthermore, a comprehensive check was conducted on team names to ensure consistency throughout the dataset.

A heatmap visualization given in Fig. 1, was employed to identify and visualize the missing data. It revealed a substantial number of missing values related to player position scores in the games. To correct this, the missing values in these columns were imputed using team-specific player position average values. In cases were a team lacked any player position score, the data was treated with a default value of 50, representing the average score that a team could attain.

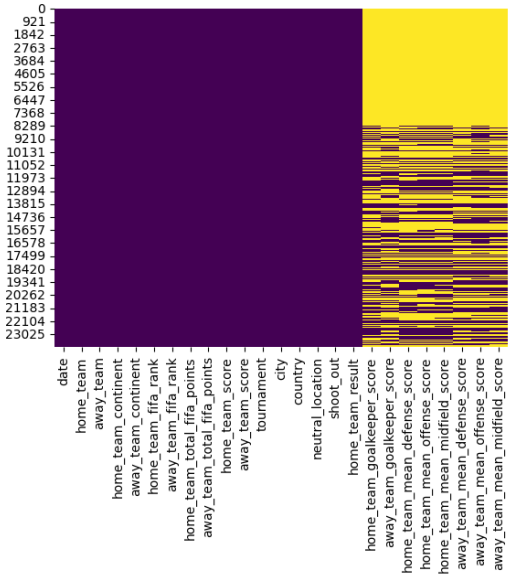

Fig. 1. Heatmap of the dataset.

The pre-processing steps contributed to refining the dataset, specifically addressing data quality issues, and preparing it for utilization in the prediction model. This careful preparation is critical in ensuring the reliability and effectiveness of the model in achieving the objectives of this project.

C. Feature Engineering

During the feature engineering phase, several potential features were introduced to the dataset to improve the prediction of game results. A binary feature, denoted as 'home_advantage', was added to indicate whether the game was played on a neutral venue (0) or if the home team had an advantage (1). This addition aimed to account for the impact of teams playing on their home venue.

Moreover, the target variable, 'is_win' was encoded to represent whether the home team won (1) or lost (0), transforming the problem into a binary classification scenario. The games resulting in a draw are also treated as instances where 'is_win' equals 0.

Correlation matrices were generated to visualize the relationship between numeric features, assisting in identification of prospective features strongly associated with predicting the target variable. Values closer to 1 or -1 suggested a stronger linear (1 = positive; -1 = negative) association between the feature and the target variable, while values closer to 0 indicated a weaker or negligible linear relationship.

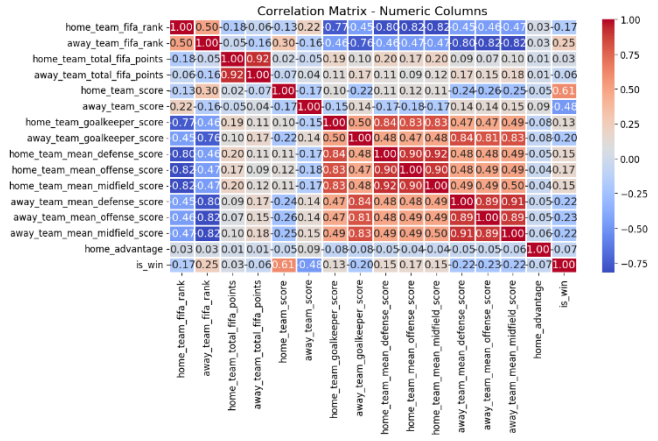

Fig. 2. Correlation matrix of all available features.

The correlation matrix depicted in Fig. 2 revealed various relationships:

- Scores of both home and away teams emerged as the most influential factors in predicting a win.
- Position player scores exhibited weak to moderate correlations.
- FIFA rankings and total points displayed relatively weak correlations.
- Home advantage demonstrated minimal impact on the likelihood of home team winning.

It is important to acknowledge that, while correlation matrix might provide some insights related to the target variable, certain features might still be retained in the prediction model as they could collectively offer meaningful and valuable information.

D. Feature Selection

In the construction of the prediction model, following features were chosen for their potential influence on the game outcome.

- Home and Away Team FIFA Ranks (home_team_fifa_rank' and 'away_team_fifa_rank')
- Home and Away Team Total FIFA Points ('home_team_total_fifa_points' and 'away_team_total_fifa_points')
- Home Advantage ('home_advantage')
- Home and Away Team Goalkeeper Score ('home_team_goalkeeper_score' and 'away_team_goalkeeper_score')
- Home and Away Team Mean Defence Score ('home_team_mean_defense_score' and 'away_team_mean_defense_score')
- Home and Away Team Mean Midfield Score ('home_team_mean_midfield_score' and 'away_team_mean_midfield_score')
- Home and Away Team Mean Offense Score ('home_team_mean_offense_score' and 'away_team_mean_offense_score')

Certain features were deliberately chosen, even though some exhibited weak correlations with the target variable. While the correlation matrix provides insights into linear relationships, it's essential to consider that certain features, might collectively offer meaningful and valuable information for predicting the outcome. For instance, the home advantage, despite demonstrating a weak relationship, was retained in the model due to its potential to influence the outcome.

The choice to exclude the home and away team scores from the selected features was due to the considerations of redundancy and limited impact to the model's predictive performance. These scores exhibited strong correlations with the target variable, making their inclusion likely to cause multicollinearity issues. Also, since the target variable already represents the binary outcome of a win or lose, introducing individual scores might be redundant, as it essentially duplicates information already captured by the target variable.

Additionally, the scikit-learn feature importance attribute was utilized to analyse the importance scores of each feature to interpret the factors contributing to the model's predictions. Feature importance is associated with tree-like structured models such as Decision Trees and Random Forests. The Decision Tree uses Gini impurity measures or information gain to determine the order in which features are used to split the data, providing useful information about the relationship between each future and the target variable. Fig. 3 illustrates the feature importance graph of the selected features, representing their respective contributions.

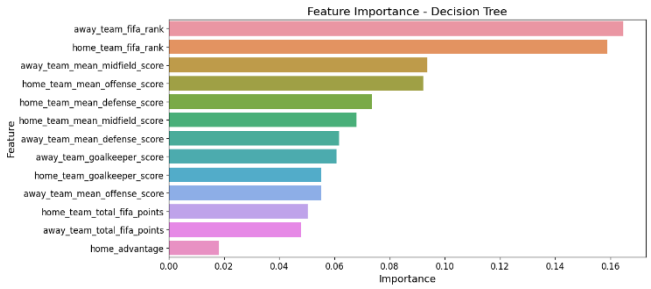

Fig. 3. Feature importance graph of the selected features.

These engineered features are intended to capture the relationships and information that may contribute to improve the predictive performance of the model.

IV. Exploratory Data Analysis

Data analysis was conducted to provide insights and justification of results of the prediction model.

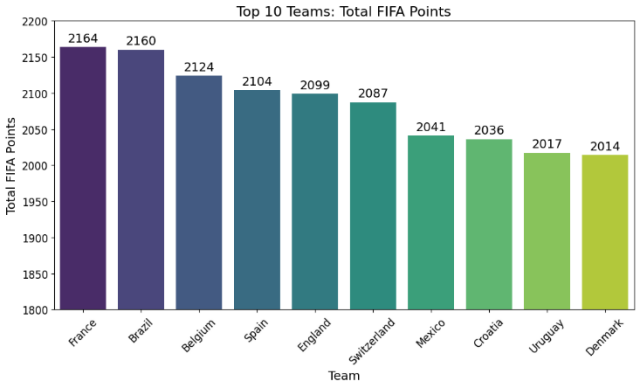

Fig. 4. Top 10 teams based on their total FIFA points.

According to Fig. 4, France leads the total FIFA points with 2164, securing the top position. Brazil closely follows with 2160 points, making it the second-highest ranked team. Notably, Belgium claims the third spot with a total 2124 points.

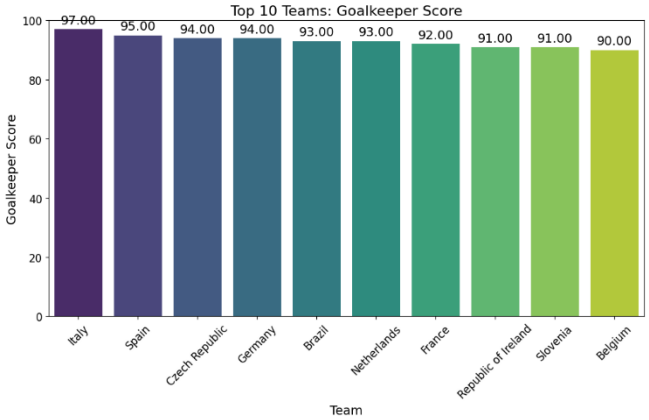

Fig. 5. Top 10 teams based on goalkeeper score.

In terms of the highest goalkeeper scores, as shown in Fig. 5, Italy, Spain and either Czech Republic or Germany emerge as top-performing teams with scores of 97, 95 and 94 points, respectively.

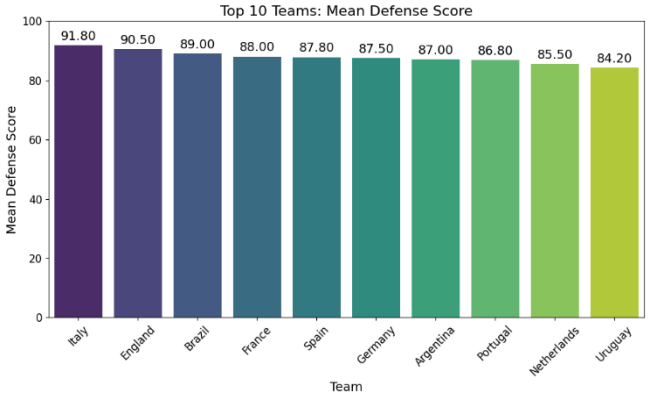

Fig. 6. Top 10 teams based on their defence score.

Illustrated in Fig. 6, Italy, England, and Brazil stand out as top-performing teams with the highest mean defence scores of scores of 91.8, 90.5 and 89.0 points, respectively.

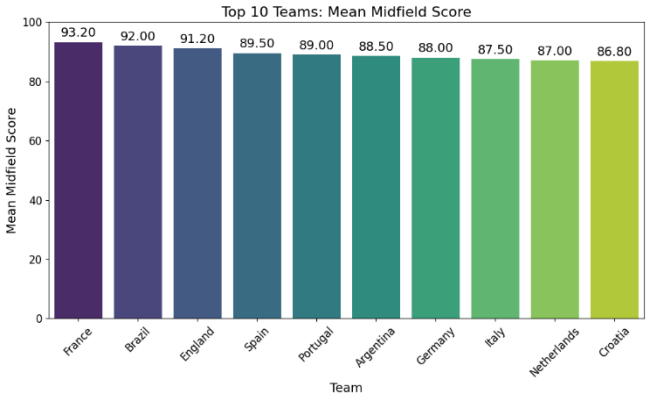

Fig. 7. Top 10 teams based on their midfield score.

As shown in Fig. 7, France, Brazil, and England emerge as top-performing teams with the highest mean midfield scores of 93.2, 92.0 and 91.2 points, respectively.

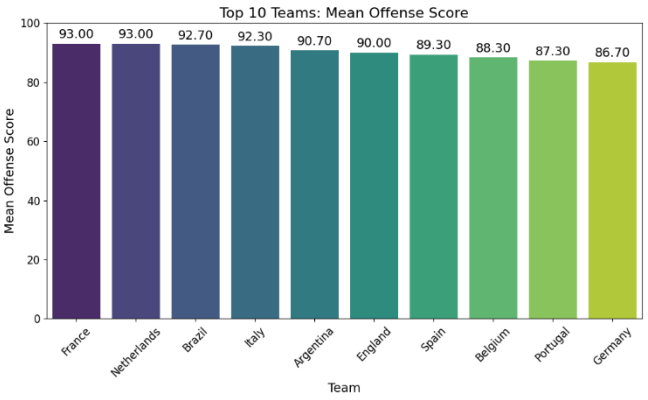

Fig. 8. Top 10 teams based on their offense score.

Illustrated in Fig. 8, France, Netherlands, Brazil, and Italy stand out as top-performing teams with mean offense scores of 93, 93, 92.7 and 92.3 points, respectively.

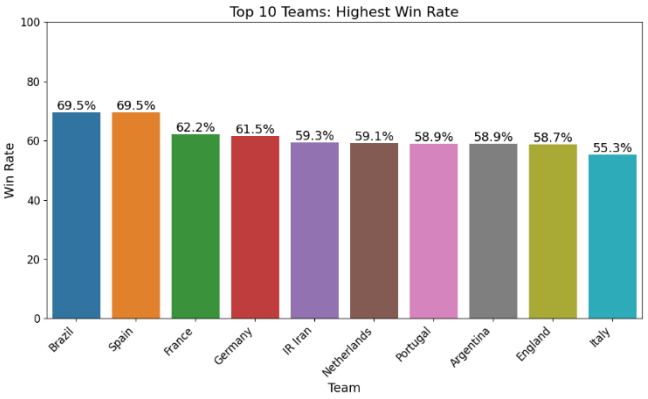

Fig. 9. Top 10 teams based on their highest win rates.

In terms of the highest win rates, illustrated in Fig. 9, Brazil or Spain, France and Germany stand out as top-performing teams with rates of 69.5%, 62.2 % and 61.5%, respectively.

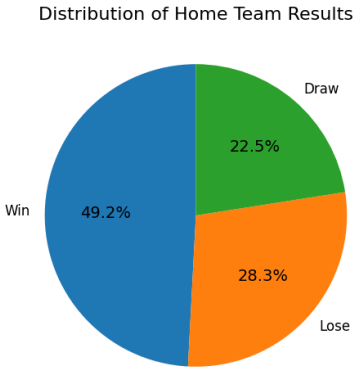

Fig. 10. Distribution of home team results.

According to Fig. 10, the home teams win nearly half of the games at 49.2%. Only 22.5% of games end in a draw, while home teams lose 28.3% of the time.

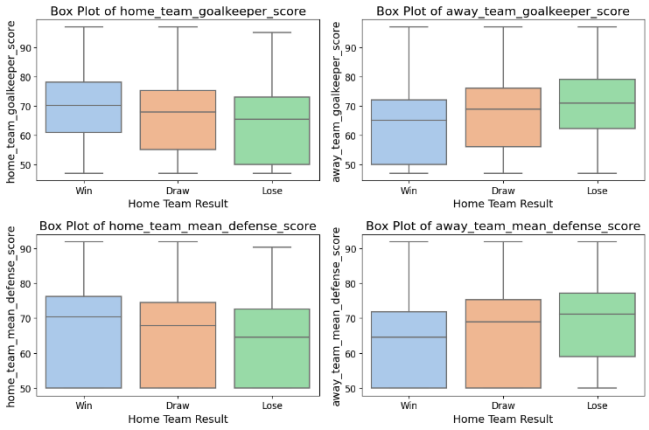

Fig. 11. Home and away team goalkeeper and defence scores vs. home team results.

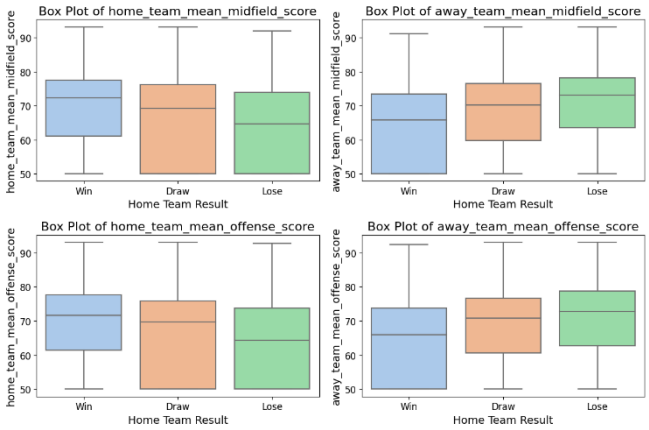

Fig. 12. Home and away team midfield and offence scores vs. home team results.

Based on the boxplots in Fig. 11 and Fig. 12, the higher player position scores correlate with an increased chance of winning the game.

Overall, this data analysis revealed that France leads in FIFA points, followed by Brazil and Belgium. Top-performing teams in various categories include Italy, Spain, Czech Republic/Germany, England, and Netherlands. Higher player position scores correlate with an increased chance of winning.

V. Prediction Model - Methodology

The analysis was conducted using Python 3.12.0 programming language in Visual Studio Code and Google Colab to facilitate collaborative work among authors. Essential libraries, such as NumPy, Pandas, Matplotlib, Seaborn and scikit-learn were utilized for efficient data pre-processing, graphical visualizations, and machine learning tasks.

A. Classification Models

The methods considered for this problem statement were the following.

1) Decision Tree

Decision Tree is a form of classification model that recursively splits the dataset based on feature values. This method is relatively easy to interpret, creating a tree-like structure of decisions. However, they are prone to overfitting, meaning that they can become too specific to the training data and fail to generalize the new data. This can lead to misclassification of instances. The model may capture noise in the training data, causing it to make errors when applied to unseen data.

2) Neural Network

Neural Network is a powerful model capable of handling complicated patterns and relationships in the data. It operates based on a structure inspired by the human brain. However, it requires hyperparameter tuning for optimization. If not properly optimized, it can suffer from overfitting (memorization of the training data, resulting in poor generalization to new data) or underfitting. This model is particularly suitable for tasks such as image recognition, natural language processing, etc. due to their capability of capturing complex patterns.

3) Naive Bayes Classifier

Based on Bayes' theorem, this classifier computes the probability of each instance and classifies according to the highest probability value. It assumes independence among features and if the features are not truly independent, it can lead to miscalculation of probabilities and incorrect classifications.

4) Support Vector Machine (SVM)

Support Vector Machine is a widely used supervised learning method for classification problems. It aims to locate the line at the maximum distance for points of both classes, making it effective in high-dimensional spaces and suitable for small to medium-sized, complex datasets. SVM is also known for handling non-linear decision boundaries through the kernel method. If the selected kernel is not appropriate for the data distribution, the model's decision boundary might not accurately reflect the patterns in the data, resulting in misclassifications.

5) Deep Learning

Deep Learning utilizes neural networks with multiple hidden layers, allowing the model to learn hierarchical features automatically. While it is a strong method, it requires a significant amount of data, processing resources, and careful tuning of hyperparameters for optimal performance. Deep learning is suitable for applications in large-scale and complex datasets, making it well-suited for tasks like image and speech recognition. If the dataset is limited, the model may struggle to learn complex patterns, leading to misclassifications. Also, the complexity of deep learning's structure may make them likely to overfitting.

B. Dataset Training and Testing

The dataset is split into training and testing subsets using the 'train_test_split' function. Following a common practice, 80% of the data is reserved for training the prediction model, while the remaining 20% were used to assess the model's performance. This division ensures a strong assessment of the model's performance.

C. Performance Metrics

In the evaluation of the classification models, the following performance metrics were employed:

1) Confusion Matrix

The Confusion Matrix is a table summarizing the performance of a classification algorithm. This table

distinguishes between correct and incorrect predictions, denoting true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values.

TABLE I. CONFUSION MATRIX

		Predicted Class	
		Class = YES	***Class = NO***
Actual Class	***Class = YES***	TP	FN
	Class = NO	FP	TN

2) Accuracy, Precision, Recall and F-Measure

Accuracy: The most used parameter classifier evaluation. It is defined as the ratio of correctly classified cases against the total number of instances.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Precision: The ability to accurately anticipate true positive cases among the total number of positive cases, especially important when the cost of false positives is high.

$$Precision = (TP) / (TP + FP) \quad (2)$$

Recall: The ability to predict all positive cases in the dataset, especially important when the cost of false negatives is substantial.

$$Recall = (TP) / (TP + FN) \quad (3)$$

F-Measure: The harmonic mean of precision and sensitivity, providing a balanced measure.

$$F\text{-}Measure = (2*TP) / (2*TP + FN + FP) \quad (4)$$

3) Receiver Operating Characteristic (ROC) Curve

The ROC Curve is a significant performance metric, especially for binary classification problems. It plots the true positive rate against the false positive rate, and the area under the curve (AUC) represents the degree of separability. Curves closer to the top-left corner and higher AUC values indicate better performance.

4) Precision vs. Recall Curve

The Precision vs. Recall Curve depicts the trade-off between the precision and recall as the classification threshold varies. A classification model that has a high precision and high recall is the one closes to the upper-right corner of the plot.

VI. PREDICTION MODEL – RESULTS & DISCUSSIONS

The results and discussion of the methodology described in Section IV are provided below.

A. Confusion Matrix

The Confusion Matrix of each considered classification model is presented in the following tables.

TABLE II. CONFUSION MATRIX – DECISION TREE MODEL

		Predicted	
		Win	***Not Win***
Actual	***Win***	1422	950
	Not Win	896	1517

The Decision Tree model accurately predicted 1422 "Wins" and 1517 "Not Wins" but misclassified 950 instances as "Wins" and 896 instances as "Not Wins.".

TABLE III. CONFUSION MATRIX – NEURAL NETWORK MODEL

		Predicted	
		Win	***Not Win***
Actual	***Win***	1927	445
	Not Win	1117	1296

The Neural Network model correctly predicted 1927 "Wins" and 1296 "Not Wins" with 445 instances misclassified as "Wins" and 1117 instances misclassified as "Not Wins.".

TABLE IV. CONFUSION MATRIX – BAYES CLASSIFIER MODEL

		Predicted	
		Win	***Not Win***
Actual	***Win***	1622	750
	Not Win	767	1646

The Bayes Classifier model accurately predicted 1622 "Wins" and 1646 "Not Wins," but it misclassified 750 instances as "Wins" and 767 instances as "Not Wins.".

TABLE V. CONFUSION MATRIX – SUPPORT VECTOR MACHINE MODEL

		Predicted	
		Win	***Not Win***
Actual	***Win***	1620	752
	Not Win	729	1684

The Support Vector Machine model correctly predicted 1620 "Wins" and 1684 "Not Wins," with 752 instances misclassified as "Wins" and 729 instances misclassified as "Not Wins.".

TABLE VI. CONFUSION MATRIX – DEEP LEARNING MODEL

		Predicted	
		Win	***Not Win***
Actual	***Win***	864	1508
	Not Win	245	2168

The Deep Learning model correctly predicted 864 "Wins" and 2168 "Not Wins," with 1508 instances misclassified as "Wins" and 245 instances misclassified as "Not Wins.".

Overall, the Decision Tree model had a notable number of false predictions, especially false positives. Similar to the Decision Tree, Neural Network model struggled with false positives. For the Bayes Classifier and SVM models, the models had a reasonable balance between true positives and false positives. The Deep Learning model had a significant number of false positives, implying the challenges in accurately predicting positive cases.

B. *Accuracy, Precision, Recall and F-Measure*

The table below presents the summary of comparative analysis for each prediction model.

TABLE VII. MODEL PERFORMANCE METRICS

Model	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Decision Tree	60.86	60.69	59.70	60.19
Neural Network	66.96	70.03	58.31	63.63
Bayes Classifier	68.30	67.89	68.38	68.14
Support Vector Machine	69.05	68.97	68.30	68.63
Deep Learning	69.13	70.92	63.95	67.26

The resulting Accuracy, Precision, Recall and F-Measure performance metrics revealed that the Deep Learning model demonstrated the highest accuracy at 69.13% and the highest precision at 70.92%. On the other hand, the Bayes Classifier had the highest recall at 68.30% and the Support Vector Machine had the highest F-Measure at 68.63%.

Specifically, the Decision Tree model demonstrated a moderate level of accuracy and precision, with slightly lower recall, indicating a reasonable but not an exceptional performance. Moreover, the Neural Network exhibited higher accuracy and precision with a notable compromise in recall suggesting proficiency in correctly classifying positive instances but challenges in capturing the entire set of positive cases. The Bayes Classifier consistently performed well across all metrics, demonstrating a balanced classification performance with reasonably high accuracy, precision, and recall. Furthermore, the SVM model exhibited strong and balanced performance across accuracy, precision, recall and F-measure, standing out as a strong classification model. The Deep Learning model achieved high accuracy and precision, with a moderate recall. This signifies the Deep Learning model's effectiveness in making accurate positive predictions with some limitations in capturing the entire positive set.

C. *Receiver Operating Characteristic (ROC) Curve*

The ROC Curve and the AUC for each classification models are given in Fig. 13.

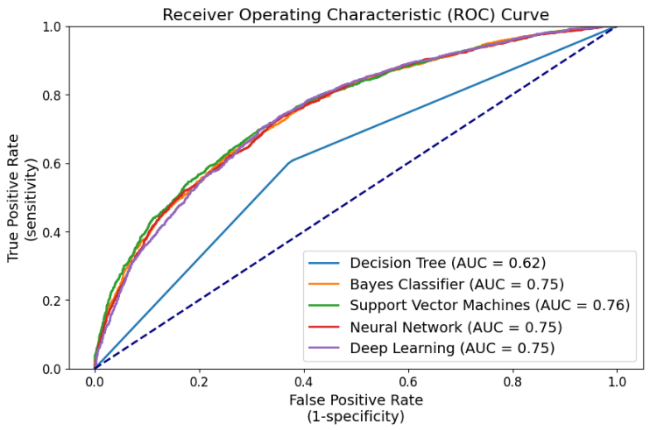

Fig. 13. ROC Curve and AUC results of each classification model.

The ROC Curve and AUC results revealed similar distinctive performances among four of the selected classification models. Notably, the Decision Tree model demonstrated the weakest performance with an AUC value of 0.62 and curve closest to bottom right corner. In contrast, the Bayes Classifier, SVM, Neural Network and Deep Learning models exhibited stronger performances, each with similar AUC values ranging from 0.75 to 0.76. Remarkably, the SVM model achieved a relatively higher AUC of 0.76 and curve closest to top left corner, suggesting its potential as the most optimal choice among the other classification models. This is particularly significant as the SVM model performed better in identifying true positives (correctly classifying actual positives) while minimizing false positives (incorrectly classifying negatives as positive), highlighting its reliability in effectively distinguishing between the classes.

D. *Precision vs. Recall Curve*

The Precision-Recall Curve and the AUC-PR for each classification models are given in Fig. 14.

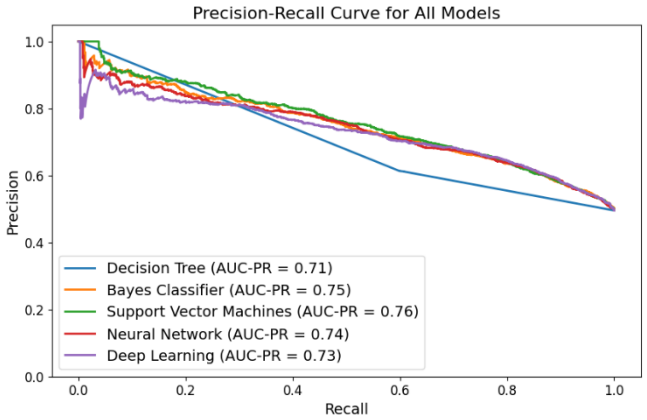

Fig. 14. Precision-Recall curve for all classification models.

Based on the Precision-Recall curves, all four models, except for the Decision Tree, resulted in relatively close to each other in terms of AUC-PR's. As the recall increases (model is identifying more true positives), the precision decreases (model is making more false positives), which means that as the model identifies positive cases, it may also start to include more incorrect negative cases. Notably, the Decision Tree model exhibited the lowest AUC-PR at 0.71 and curve closest to the bottom left corner, indicating the weakest precision-recall balance. All other models, excluding the SVM and the Decision Tree, similarly, achieving an average AUC-PR of 0.74. In contrast, the SVM model emerged as slightly stronger performer in better precision-recall balance, demonstrating the highest AUC-PR at 0.76 and the curve closest to the top right corner. This means that the SVM model performs slightly better across various thresholds, achieving higher precision (identifying more true positives) while maintaining a high level of recall (minimizing false positives).

E. *Model Selection*

Evaluating the overall performance metrics, the SVM model consistently demonstrated strong and balanced results across key indicators such as Accuracy, Precision, Recall, and F-Measure. This commendable result was further reinforced by its performance in the ROC Curve and Precision vs. Recall Curve analyses.

In the ROC Curve, the SVM model achieved a relatively higher AUC, with the curve positioned closest to the top left corner. This emphasizes its proficiency in correctly

classifying the actual positives while minimizing false positives, which is a crucial aspect when predicting the winner of 2026 FIFA World Cup. This capability can be associated to SVM's effectiveness in handling complex data structures, outliers and finding an optimal decision boundary, especially in scenarios where the data has a complex structure or is not linearly separable.

Moreover, the Precision vs. Recall Curve revealed the SVM model's greater precision-recall balance, demonstrating the highest AUC-PR and a curve closest to the top right corner. This signifies the model's good performance in maintaining high precision while identifying true positives, which is also an essential characteristic for accurately predicting the winner of the 2026 FIFA World Cup. This strength aligns with SVM's capacity to handle imbalanced datasets, capture complex decision boundaries, and effectively distinguish positive instances.

In summary, the SVM model consistently demonstrated strong and balanced performance across multiple metrics, making it a reliable and strong choice for the prediction problem. It's effectiveness in correctly classifying instances, maintaining a balance between true positives and false positives, and achieving a commendable precision-recall balance, emphasizes its suitability for the examined classification problem. Consequently, the SVM model was selected for predicting the 2026 FIFA World Cup winner.

VII. Simulation - Methodology

In this phase, utilizing the SVM model, a tournament simulation for the 2026 FIFA World Cup was executed. The simulation comprised a multi-step process aimed at predicting the game outcomes and determine the ultimate champion.

Due to the lack of information regarding qualified teams and groups, an assumption was made that the top 48 teams would qualify based on their total FIFA points. To introduce diversity and account for the unpredictability, teams were randomly shuffled to account for every possible scenario at each iteration and each stage of the tournament. Initially, the teams competed in 12 groups of four, with the top two and eight best third-placed teams progressing to the round of 32, as approved by the FIFA Council [2]. The subsequent stages included the round of 16, quarterfinals, semi-finals, and the championship final.

Given the official announcement that the event will take place in three North American countries [1], the home advantage was factored into the simulation accordingly. This involved considering the impact of teams playing on their home venue during the games.

To ensure the convergence of results and minimize the influence of uncertainty introduced by randomization, the simulation was executed across 10,000 iterations. Upon completion of these iterations, the probability of each team winning the 2026 FIFA World Cup was calculated based on the frequency of their victories in the final stage.

VIII. Simulation – Results & Discussion

The bar chart in Fig. 15 visually represents the probabilities of potential winners as generated by the prediction model.

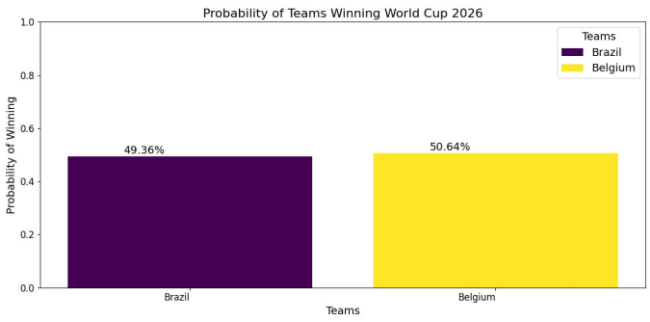

Fig. 15. 2026 World Cup winner probabilities: Brazil vs. Belgium.

The results demonstrate a very close and competitive scenario between the two teams, with Belgium holding a slightly higher win probability of 50.64%, while Brazil follows closely with 49.36%. The prediction model suggests that Brazil and Belgium have nearly equal chances of becoming the champion of the 2026 FIFA World Cup, with neither team significantly favoured over the other.

To assess convergence, the graph of Belgium's winning rate versus the number of iterations is presented in Fig. 16.

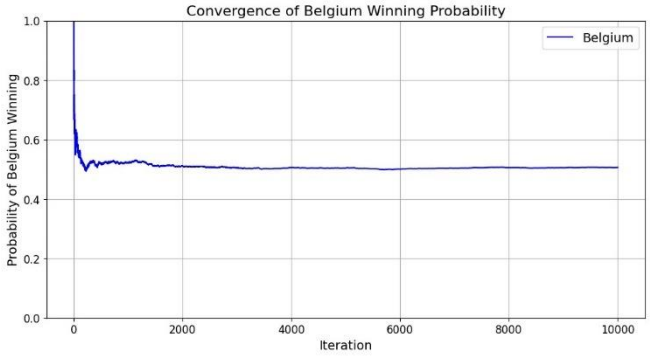

Fig. 16. Graph of Belgium winning rate vs. the number of iterations.

The convergence graph in Fig. 16 indicates that Belgium maintains an almost 50% winning rate at the conclusion of 10,000 iterations. Around the 4,000th iteration, the model seems to have achieved a stable prediction instance, reinforcing the notion that Belgium and Brazil have an almost equal likelihood of winning.

It is crucial to emphasize that uncertainty is introduced to the results due to the randomization of teams and groups. Consequently, if the iterations were repeated, the outcome would likely be similar, with competitive results between Brazil and Belgium, although not precisely identical.

IX. Conclusion

In conclusion, this project aimed to predict the winner of 2026 FIFA World Cup utilizing machine learning techniques based on a dataset retrieved from Kaggle [5], which included international football games from August 1993 to June 2022. Subsequent data pre-processing ensured data integrity, while feature engineering introduced key features such as home advantage and encoded the target variable, enhancing the dataset's predictive potential. Exploratory data analysis showcased the competitive statistics of top football teams, highlighting Brazil, France, Spain, Germany, England, Italy, and Belgium as top performers in various areas.

The classification models, including Decision Tree, Neural Network, Naive Bayes Classifier, Support Vector Machine, and Deep Learning, were evaluated using performance metrics such as accuracy, precision, recall, F-measure, ROC Curve and Prevision vs. Recall Curve. The Support Vector Machine exhibited a strong performance, demonstrating balanced and high-performing classification across various metrics. SVM's ability to correctly identify winners while minimizing false predictions made it the preferred choice for predicting the 2026 FIFA World Cup champion.

The simulation, using the selected SVM model, predicted a showdown between Brazil and Belgium. The results displayed a closely competitive battle, with Belgium holding a slightly higher win probability of 50.64%, while Brazil followed closely at 49.36%. The convergence analysis reinforced the stability of these predictions using 10,000 iterations.

However, it's crucial to acknowledge the uncertainty introduced by randomizing teams and groups during simulations, adding dynamism to the results. The apparent difference between top-performing teams identified in data analysis, such as Brazil, France, Spain, Germany, England, Italy, and Belgium and the predicted winner (either Brazil or Belgium) can be attributed to the dynamic nature of sports and the specific features considered. Common features like total FIFA points and player position scores were examined in both analyses, but the machine learning approach allowed the model to offer a nuanced perspective on the potential 2026 FIFA World Cup winner.

As the football fans are impatiently waiting for the upcoming world cup, this project contributes a machine learning and data-driven perspective to the enthusiasm for the next FIFA World Cup. The only thing left is to witness which team emerges as the winner once the tournament kicks off and to verify the accuracy of the predictions. Football is an inherently unpredictable game, and that's what makes it so thrilling!

X. Recommendations & Future Work

To improve the strength and precision of the prediction model for determining the winner of the 2026 FIFA World Cup, attention to the following aspects in the future works is recommended.

A. Addressing Missing Data

The current dataset lacks the information on international football games played after June 2022, particularly the 2022 FIFA World Cup. Integrating and processing with up-to-date data is crucial. Additionally, the existing dataset contains a significant amount of missing data in player position scores, which was imputed by the average player position scores of each team during the data pre-processing phase.

B. Inclusion of Dynamic Features and External Factors

The current dataset lacks the inclusion of dynamic information, such as injuries or unexpected events. External factors, such as spectator attendance, geopolitical influences, and weather, are also not considered. The addition of these features has the potential to enhance the model's adaptability to changing scenarios, providing a more comprehensive prediction model.

C. Enhanced Data Exploration and Feature Engineering

Implementing a higher level of feature engineering and data exploration could reveal more meaningful information from the existing dataset.

D. Exploring other Machine Learning Classification Models

Evaluate other classification models, such as Random Forest, Logistic Regression, or Gradient Boosting Models, capable of capturing more complex relationships and handling various feature types.

E. Hyperparameter Tuning

Maximize the predictive performance of the classification model by tuning hyperparameters. For instance, using sklearn's GridSearchCV, hyperparameters could be tuned conveniently to identify the optimal model.

F. Inclusion of Year Information

Consider incorporating information about the year into the model, as the performance and dynamics of each team may change over time.

G. Considering Games that Resulted in Draw

Distinct consideration of games resulting in a draw could be implemented. For example, encoding the 'is_win' target variable into three numbers to account for draw games, acknowledging the unique outcomes.

H. Excluding Friendly Games from the Dataset

Consider excluding friendly games from the dataset, as teams may not perform as seriously and intensely in friendly games compared to tournament games.

I. Revisiting the Simulation After the Qualified Teams and Groups are Officially Published

Given the current uncertainty about qualifying teams and groups, the simulations heavily relied on randomization. Revisiting the simulations after the list of qualified teams and their respective groups is officially published is crucial for refining the model's predictions and ensuring accuracy.

XI. Acknowledgment

We would like to express deep gratitude to Prof. Dr. Suat Ozdemir, for their invaluable assistance. Additionally, we extend our thanks to our families for their support and understanding during the challenging phases of the project.

XII. References

[1] FIFA, "2026 FIFA World Cup - Canada/Mexico/USA," FIFA+. Available: https://www.fifa.com/fifaplus/en/tournaments/mens/worldcup/canadamexicousa2026.

[2] FIFA Council, "FIFA Council Approves International Match Calendars," FIFA, Media Releases. [Online]. Available: https://www.fifa.com/about-fifa/organisation/fifa-council/media-releases/fifa-council-approves-international-match-calendars.

[3] A. Hassan, A. R. Akl, I. Hassan, and C. Sunderland, "Predicting wins, losses and attributes' sensitivities in the soccer world cup 2018 using neural network analysis," Sensors, vol. 20, no. 11, pp. 3213, 2020.

[4] A. T. Kabakus, M. Simsek, and Y. Belenli, "The wisdom of the silent crowd: predicting the match results of world cup 2018 through Twitter," International Journal of Computer Applications, vol. 182, pp. 40-45, 2018.

[5] Brenda89, "FIFA World Cup 2022 Dataset," Kaggle. Available: https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022/data.