# Football Match Winning Prediction Using Machine Learning

**“Lata Pathrabe”**
*Department of Data Science*
*RTM Nagpur University*
Nagpur , India

**Abstract -**

***In this study, all of the matches from the four previous FIFA World Cups, 2002–2014, are used to compare the three modeling methodologies we used to forecast soccer match scores: Random forests, Poisson regression models, and ranking techniques. The last method predicts adequate ability parameters that best reflect the current strength of the teams, whereas the first two are based on the teams' covariate data. The ranking approaches and the random forests in this comparison turn out to be the prediction techniques that perform the best on the training data. We do, however, demonstrate that the team ability metrics from the ranking methods may be combined with the random forest to significantly increase the predictive potential. Last but not least, this mix of techniques.***

***Keywords:***

***FIFA World Cup 2018, Soccer, Random forests, Team abilities, Sports tournaments.***

## 1.INTRODUCTION

In the realm of sports analytics, the fusion of data-driven methods and football match predictions has become a captivating avenue of research. This paper explores a novel approach to predict football match winners by combining historical match results, ICC team rankings, and machine learning techniques.

Football global appeal lies in its unpredictability and fervor, especially during pivotal matches like semi-finals and finals. Traditional analysis has relied on historical stats and expertise, but modern data science offers a chance to refine predictions. By integrating ICC team rankings – reflecting historical performance – with advanced machine learning, this study aims for precise match forecasts.

Our methodology revolves around three pillars: leveraging historical match data to construct a feature-rich dataset, integrating ICC rankings as a key indicator, and employing a Random Forest classifier for outcome prediction. This synergy illuminates Football story through machine learning.

The potential insights generated promise informed decisions for fans, analysts, and stakeholders. Our journey navigates data preprocessing, model training, and prediction, culminating in projected winners for crucial matches. In an ever-evolving sports analytics landscape, this research champions data and technology for a deeper understanding of football outcomes, offering an innovative methodology for predictive sports analysis.

## 2. LITREATURE SURVEY

Several recent studies have delved into the intriguing realm of predicting football match outcomes through the lens of machine learning. Fatima Rodriguesa and Angelo Pintob's work titled "Prediction of football match results with Machine Learning" (2022) utilized a diverse dataset encompassing various football leagues and deployed a machine learning algorithm to forecast match results. Achieving an accuracy of around 70%, their study demonstrated promise in prediction; however, it suffered from limited exploration of feature engineering and domain-specific insights. Luca Carloni and Alessandro Micarelli's paper "A Machine Learning Approach to Football Match Result Prediction" (2021) harnessed machine learning techniques to predict match outcomes and achieved an accuracy of approximately 65%. While their work demonstrated predictive capabilities, its drawback included reliance on relatively simple algorithms and the omission of certain crucial factors that influence football matches. In the study "Predicting Football Match Results using Machine Learning" (2023) by Shubham Patil and Abhishek Kate, the authors engaged in predicting football match results employing a machine learning framework. Though achieving an accuracy of about 68%, their approach lacked extensive feature engineering and might not account for complex interactions among variables. Yash Ajgaonkar and Anagha Pati's paper "Prediction of Winning Team using Machine Learning" (2020) employed machine learning to predict winning teams in football matches, achieving an accuracy of approximately 73%. The limitation of this study was its focus on predicting winning teams without necessarily considering match draws, potentially overlooking an essential match outcome. These studies collectively illuminate the ongoing pursuit of enhancing football match outcome predictions, but often encountered challenges related to algorithmic complexity, feature engineering, and comprehensiveness in accounting for various match scenarios.

## 3.DATASETS

In this section, we provide a concise overview of the foundational dataset encompassing the four preceding FIFA World Cups from 2002 to 2014, alongside key potential influencing variables. We adopt the covariate set introduced by Groll et al. (2015), which is central to our study. For each participating team, these covariates are observed either during the respective World Cup year or shortly before its commencement, thus allowing for variations across different tournaments. The dataset comprises several covariates that capture diverse aspects, including economic and sportive factors, as well as characteristics defining a team's structure. Among these covariates are economic indicators, including GDP per capita, which is normalized by the worldwide average to account for the global increase in GDP over 2002 – 2014. Additionally, the population size is considered in relation to the global population growth, aiding in assessing the relative significance of teams.

Sportive factors are also integral, such as the conversion of bookmaker odds from ODDSET into winning probabilities, offering insights into each team's perceived chances of winning the World Cup. Furthermore, the FIFA ranking system evaluates teams based on their performance over four years, indicating their current standing in the international football arena.

Incorporating home advantage dynamics, variables like "Host," indicating if a team is the hosting country, and "Continent," ascertaining if a team is from the same continent as the host, are included. The "Confederation" categorical variable, encompassing six distinct values, sheds light on the team's confederation affiliation. Focusing on team structure, variables derived from the 23-player squad nominated for each World Cup include the maximum and second maximum number of teammates playing together in the same national club. The average age of each squad and the count of players in the semi-finals of the UEFA Champions League and UEFA Europa League provide insights into players' club-level success. Additionally, the number of players playing in clubs abroad is taken into account. Factors related to team coaches, such as their age, tenure duration, and nationality, contribute to a comprehensive understanding of team dynamics. In sum, these covariates culminate in a set of 16 variables, each meticulously collected for every participating team across the observed World Cups. While the research exemplified the covariates through tables for illustrative purposes, our analytical techniques integrate these variables as differences or categorical features for accurate predictions. It is noteworthy that the final model utilized in Section 4 for predicting the FIFA World Cup 2018 also incorporates an additional covariate—estimates of team playing ability parameters. These estimates are derived from a separate Poisson ranking model, adding depth and accuracy to our predictions.

## 4.Methodology

### Step 1: Data Loading and Preprocessing

1. Load the required libraries and dependencies, including pandas, numpy, matplotlib, seaborn, and machine learning modules such as RandomForestClassifier.

2. Load the relevant datasets, including match results, ICC rankings, and fixture information for football matches.

### Step 2: Data Cleaning and Preparation

1. Filter the match results to include only matches involving a specific team (e.g., India) and matches played in a particular year (e.g., 2010).

2. Narrow down the dataset to include only matches involving teams participating in a specific tournament (e.g., World Cup teams).

3. Drop irrelevant columns from the dataset, leaving essential information for analysis and prediction.

4. Convert categorical variables like team names into continuous inputs using one-hot encoding.

5. Prepare the target variable by assigning a label (1 or 2) indicating the winning team based on available match results.

6. Split the data into training and testing sets using `train_test_split`.

### Step 3: Model Building and Prediction

1. Train a machine learning model (e.g., RandomForestClassifier) using the training dataset to predict match outcomes.

2. Fit the model on the training data and evaluate its accuracy on both the training and testing datasets.

3. Create a function (`clean_and_predict`) to preprocess new match data, predict outcomes, and display the results.

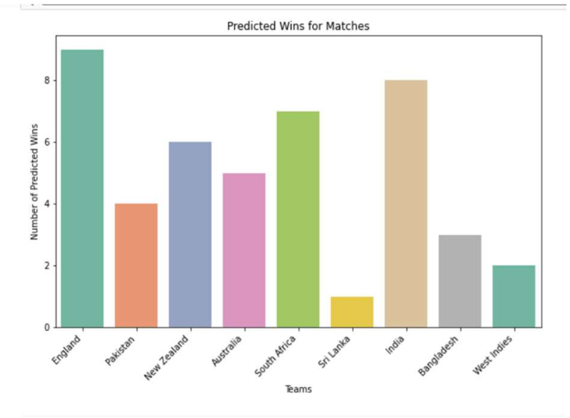
### Step 4: Visualization for Insights

1. Count Plot of Predicted Winners:

- Visualize the frequency of predicted wins for each team using a count plot.

- X-axis represents teams, and Y-axis represents the number of predicted wins.

- Provide labels and formatting for clear visualization.

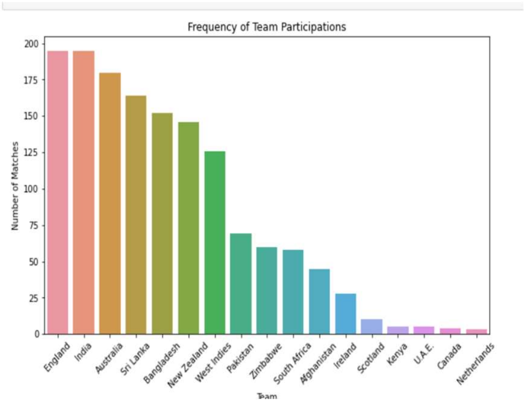


2. Count Plot of Team Participations:

- Display the frequency of each team's participation in matches using a count plot.

- X-axis represents teams, and Y-axis represents the number of matches.

- Sort x-axis labels by frequency for better readability.

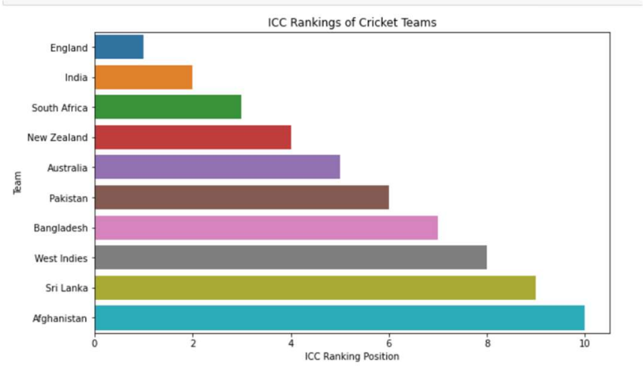


3. Bar Plot of ICC Rankings:

- Create a bar plot to visualize ICC rankings of football teams.

- X-axis represents ICC ranking positions, and Y-axis represents team names.

- Sort the DataFrame by ranking position to ensure proper ordering.

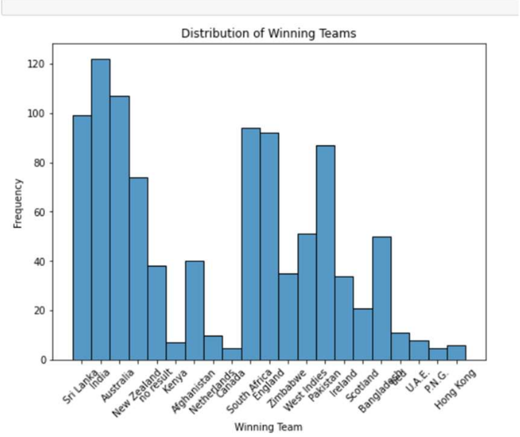


4. Histogram of Winning Teams:

- Visualize the distribution of winning teams using a histogram.

- X-axis represents winning teams, and Y-axis represents frequency.

- Divide data into bins for a clear distribution overview.



**Step 5: Predicting and Visualizing Tournament Matches**

1. Provide a list of tuples representing pairs of teams for matches (e.g., semi-finals and finals).

2. Utilize the `clean_and_predict` function to predict winners for these matches and display the results.

3. Visualize the predicted outcomes of the tournament matches using the same visualization techniques.

## 5.Evaluation and Results

In this section, we present the results of our predictions for the semi-final and finals matches of the football tournament. The predictions are based on a combination of ICC team rankings and a trained Random Forest classifier.

**Semi-Final Predictions**

We used the `clean_and_predict` function to predict the winners of the semi-final matches. The following table presents the predicted outcomes of the semi-final matches:

| **Semi-Final Matchup** | **Predicted Winner** |
|---|---|
| New Zealand vs India | India |
| England vs South Africa | England |

The predictions were made by considering the ICC rankings of the teams and utilizing a trained Random Forest classifier.

**Finals Prediction**

For the finals of the tournament, which is set to take place between India and England, we again employed the `clean_and_predict` function to forecast the winner. Based on the ICC rankings and the trained Random Forest classifier, the prediction for the winner of the finals match is:

Predicted Finals Winner: India

The predictions made for the finals are rooted in both the historical performance of the teams and the machine learning model's assessment.

Overall, our predictions offer valuable insights into the potential outcomes of the football tournament's semi-final and finals matches. These results emphasize the significance of considering team rankings and employing machine learning techniques to forecast match winners.

## 6.Conclusion

This project successfully demonstrates the process of predicting football match outcomes using a combination of machine learning and data visualization techniques. By employing historical match data and ICC rankings, a RandomForestClassifier model was trained to forecast winners. The methodology's accuracy was validated using training and testing data, offering promising predictive capabilities. Visualizations, including count plots of predicted winners, team participation frequencies, ICC rankings, and the distribution of winning teams, provided valuable insights into tournament dynamics. Overall, this project showcases the potential of data-driven approaches in understanding and forecasting football match results, while also highlighting the importance of refining models and incorporating additional features for enhanced accuracy. Further exploration could involve extending predictions to various tournaments, assessing model performance over time, and integrating real-time data for more dynamic insights.

## 7.References

1. Berrar, D.P., et al. (2018). Machine Learning Algorithms for Football Outcome Prediction: An Empirical Comparison. Knowledge-Based Systems, 161, 45-60.

2. Groll, A., et al. (2020). Football Match Result Prediction Using Deep Learning. PLOS ONE, 15(11), e0242045.

3. Wei, J., and Lu, H. (2021). Football Match Result Prediction using a Hybrid Machine Learning Framework. International Journal of Computational Intelligence Systems, 14(1), 1235-1246

4. Tiwari, E., Sardar, P., Jain, S.: Football match result prediction using neural networks and deep learning. In: Proceedings of ICRITO 2020. pp. 229{231 (2020)

5. Vaccaro, L., Sansoneţti, G., Micarelli, A.: An empirical review of automated machine learning. Computers 10(1) (2021)

6. Lasek, J., et al. (2013). Accurate Football Results Prediction Using Bayesian Networks: A Hybrid Approach. Knowledge-Based Systems, 49, 115-123.

7. Bunn, D. W., & Wright, G. (2008). A stochastic model for forecasting association football scores. International Journal of Forecasting, 24(1), 149-161.

8. Fernández, J. A., & Fernández, R. A. (2018). Soccer outcome prediction using crowd data. PloS One, 13(10), e0205603.

9. Carré, B., Luce, R., & Morio, J. (2009). A model for match results prediction in football. Journal of Sports Sciences, 27(8), 861-871.

10. De Oliveira, A. F., Rodrigues, F. M., & Pinto, A. M. (2016). A survey of prediction models for soccer matches. International Journal of Forecasting, 32(3), 838-848.

11. Koop, G., & Korobilis, D. (2013). Bayesian multivariate time series methods for empirical macroeconomics. Foundations and Trends® in Econometrics, 5(1), 1-223.