Football Match Result Prediction

P7: Manav Hemantkumar Patel, Manan Manojkumar Tiwari, Anshul Khairnar

Department of Computer Science, North Carolina State University, Raleigh, 27695

(mhpatel4, mtiwari3, akhairn) @ ncsu.edu

Abstract

This research focuses on predicting international soccer match winners through advanced machine learning and statistical analysis. By analyzing historical team performance data, including team attributes, past results, and external factors, we aim to develop a robust predictive model. Our study also explores the impact of evolving team dynamics and key player statistics. Ultimately, this research offers insights into the factors influencing international soccer match outcomes, contributing to the field of sports analytics and data-driven decision-making in sports.

1 Introduction

International soccer is a thrilling arena of competition where the outcome of matches remains uncertain, driven by a multitude of factors. This research project delves into the fascinating world of predicting match winners through the fusion of advanced machine learning techniques and meticulous statistical analysis. By mining historical data encompassing team attributes, past results, and external variables, we aim to construct a potent predictive model. Beyond this, we explore the fluid dynamics within teams and the impact of key player performance on match results. This study promises not just to predict outcomes but to unravel the complex web of factors influencing international soccer matches. It serves as a beacon in the realm of sports analytics, offering actionable insights for decision-makers in the sports industry, thus marking a significant leap forward in the application of data-driven strategies to the beautiful game.

1.1 Project Idea

Our project goal is to predict the winner of an international soccer match given past results of the international teams. Our project aims to utilize advanced machine learning techniques. We will gather and pre-process data on various team attributes, past match results, and external factors like home-field advantage to build predictive models. Moreover, we will explore the impact of evolving team dynamics and the importance of key player statistics. The ultimate objective is to develop a robust and accurate prediction system that can offer valuable insights into the factors influencing international soccer match outcomes. This research not only has practical applications in sports analytic but also offers an opportunity to showcase the potential of data-driven decision-making in the world of sports.

2 Method

2.1 Split Data

We plan to split the data into training data(70 percent) and testing data(30 percent). After splitting, we have 4544 data entries as training data and 1948 data entries as testing data from results.csv, 6676 data entries as training data and 2861 as testing data from goalscorers.csv and 118 data entries as training data and 50 data entries as testing data from shoutouts.csv.

2.2 Decision Tree Model

We use decision tree models to develop classification systems that predict or classify future observations based on a set of decision rules. We have data divided into classes: home team wins or away team wins. We use this data to build rules that we use to classify new matches with maximum accuracy.

2.3 Random Forest Model

Once we are done with creating a decision tree model, we will create a Random Forest Model which combines multiple decision trees and gives us better results and accuracy.We use the same data which we used for decision tree and use this model to classify new matches with their results accurately.

2.4 Naive Bayes Model

The final model which we have trained and tested our data on is the Naiive Bayes Model. It is a conditional probability model which calculated the conditional probability of all the attribute values and gives us results based on that.

3 Experiment

3.1 Dataset Description

This dataset includes 44,341 results of international football matches starting from the very first official match in 1872 up to 2023. The matches range from FIFA World Cup to FIFI Wild Cup to regular friendly matches. The matches are strictly men's full internationals and the data does not include Olympic Games or matches where at least one of the teams was the nation's B-team, U-23 or a league select team. The original dataset is split into 3 csv files goalscorers.csv, results.csv, and shootouts.csv. The main data set we will use is the results.csv. This data set has 9 attributes before any processing.

3.2 Data Preparation and Pre-processing

The information from the data set will be extracted via a CSV file. We would start by first cleaning the information taken that was extracted. We would not need the data from 1872 to predict match results in 2023. To get more accurate results in recent times we would first clean the data set to only include matches between 2013 to 2023. To further clean the data we removed all duplicate entries and entries that included missing data. This was appropriate for this data set because we already had a large number of entries and removing a few entries with missing values would not impact the accuracy of the model greatly. Also we only consider the matches played among the 75 teams who played most matches in the world. The attributes in the results.csv consists of many categorical values, so we convert these categorical values into binary using one hot encoding. Hence every categorical value has its own attribute and since we have 75 countries and each country is present as the home team and away team, in total we have more than 150 attributes added after encoding.

4 Results

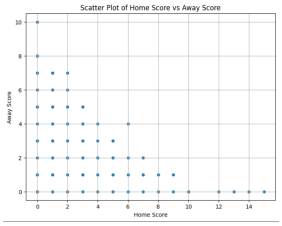

Figure 1: Home vs Away

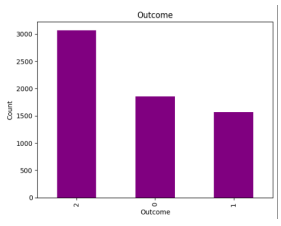

Figure 2: Outcome

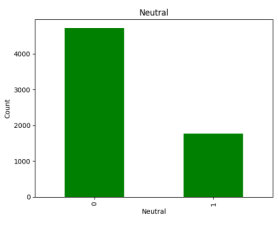

Figure 3: Neutral

We evaluated the performance of the match predictor using a decision tree model, random forest model, and a Naive Bayes model. Validation accuracy and further metrics are mentioned in the table below. Under construction

Model	Accuracy
Decision Tree	98%
Random Forest	96.71%
Logistic Regression	96.52%
Naïve Bayes	65.23%
K-Nearest Neighbor	74.78%

Figure 4: Model Results

5 Conclusion

We will address the fundamental challenge of predicting football math scores by employing a variety of models such as Decision Tree, Random Forest, and Naive Bayes.The Decision tree model produced a FillInPercent% validation accuracy, the Random Forest model produced a FillInPercent% validation accuracy, and the Naiive Bayes produced a FILL% validation accuracy. Placeholder for other models we may use.

6 References

[1] Lotfi, Said & Rebbouj, Mohamed, "Machine Learning for sport results prediction using algorithms", *International Journal of Information Technology*, 3. 148-155.
[2] D. Prasetio and D. Harlili, "Predicting football match results with logistic regression",*International Conference On Advanced Informatics: Concepts, Theory And Application* pp. 1-5.
[3] A. A. Azeman, A. Mustapha, N. Razali, A. Nanthaamomphong and M. H. Abd Wahab, "Prediction of Football Matches Results: Decision Forest against Neural Networks," *18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology.* pp. 1032-1035.

7 Appendix

7.1 Old Plan

7.1.1 Data Set

Our project will classify different types of mental health conditions. It contains over 2000 surveys and each survey has a text and nine questions linked to that survey, complied by Arxiv

7.1.2 Project Idea

Our project goal is to classify and predict mental health problems in US college students. The project will help to classify mental health disorders which are

impacting college students. We will compare the performance of our model using our testing dataset to determine its accuracy.

7.1.3 Relevant Papers

[1] Jetli Chung, Jason Teo, "Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges", *Applied Computational Intelligence and Soft Computing*, vol. 2022, Article ID 9970363, 19 pages, 2022.
[2] Pedrelli, P., Nyer, M., Yeung, A., Zulauf, C., & Wilens, T. (2015). College Students: Mental Health Problems and Treatment Considerations. Academic psychiatry : *the journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*, 39(5), 503–511.
[3] Sahlan, Fadhluddin & Mohammad Nizam, Faris Hamidi & Misrat, Muhammad & Zamzuri, Muhammad & Wani, Sharyar & Gulzar, Yonis. (2021). Prediction of Mental Health Among University Students. *International Journal on Perceptive and Cognitive Computing*. 7. 85-91.

7.1.4 Division of Work

Manan - Research relevant paper ideas to build models and reference papers for testing purposes. Manav - Implement preprocessing of our dataset and divide dataset into training and test dataset. Anshul- Use NLP techniques to tokenize the surveys and generate feature matrix for each survey. All- Build different models to find the most accurate way to classify match outcomes.

7.2 New Plan

7.2.1 New Division of Work

Manan - Research and work upon training and testing dataset to build different types of models.
Manav - Implement preprocessing of our dataset and divide dataset into training and test dataset.
Anshul- Jupyter Notebook for exploratory analysis
All- Build multiple different models to find the most accurate way to predict match outcomes. Jupyter notebooks for results.