Football Match Probability Prediction

Title of the session (you can be creative highlighting your findings)

Sergio S. Duarte, 20211020013, Sol D. Cely, 20212020026, Franco J. Guzmán, 20211020155

***Abstract*—Provide a summary of the session. What was done, what measurements were taken, brief methods, what calculations, brief conclusion. The Abstract should be approximately 250 words or fewer, italicized, in 10-point Times (or Times Roman.) Please leave two spaces between the Abstract and the heading of your first section. It should briefly summarize the essence of the paper and address the following areas without using specific subsection titles. Objective: Briefly state the problem or issue addressed, in language accessible to a general scientific audience. Technology or Method: Briefly summarize the technological innovation or method used to address the problem. Results: Provide a brief summary of the results and findings. Conclusions: Give brief concluding remarks on your outcomes. Detailed discussion of these aspects should be provided in the main body of the paper.**

***Index Terms*—KNearest, XGB, Random, etc.**

I. Introduction

Football one of the most popular and followed sports worldwide, not only captivates millions of fans but also moves enormous amounts of money through sports betting and related financial markets. In this context, the ability to accurately predict the outcome of football matches has become a primary goal for both fans and professionals in the betting industry. Predicting football match results is a complex task that involves a variety of factors, from the historical performance of teams (Goals, Rating), to match conditions (home, is cup, etc.), and the strategies employed by coaches[1].

This project focuses on exploring solutions to predict football match outcomes using the Football Match Probability Prediction dataset, which includes records of over 150,000 historical matches played between 2019 and 2021[2]. The objective is to develop models capable of predicting whether a team will win, draw, or lose in their upcoming match.

In the context of the Dataset to be used, it is essential to understand the fundamentals of European football, especially its leagues and cups. European leagues are annual competitions where teams face each other twice over a full season, playing one match at home and one away against each rival[3]. In addition to national leagues, European teams participate in various national and international cup competitions. National cups are direct elimination tournaments within each country, where teams compete in knockout matches until only one winner remains. At the European level, competitions begin with a qualifying phase based on performance in national leagues, followed by a group stage where teams play home and away matches[4].

Previously, several studies have focused on predicting football match outcomes. One notable work is that of Chandra B, Jennet Shinny and Keshav Adhitya (2024) called "Prediction of Football Player Performance UsingMachine Learning Algorithm", which utilizes machine learning techniques and data mining for this purpose. The primary objective of the study is to develop accurate models capable of reliably predicting match results, integrating key features such as historical statistics and specific match conditions. Advanced methods like logistic regression, SVM, and Bayesian networks are employed and analyzed to optimize predictive models, aiming to provide valuable insights for strategic and tactical management in professional football[5].

A second study used as a reference is the one conducted by Rory Bunker, Calvin Yeung, and Keisuke Fujii, called "Machine Learning for Soccer Match Result Prediction" , which analyzes the use of machine learning in predicting football match outcomes. This study provides an overview of the current state and potential future developments in this area, discussing available datasets, types of models and features, and ways to evaluate model performance. The findings indicate that gradient-boosted tree models, such as CatBoost, applied to football-specific ratings, like pi-ratings, are the most effective when datasets only include goals as features[6].

Also, the third study is the article "Football Match Prediction with Tree Based Model Classification" by Yoel Alfredo and Sani Isase, which investigated the prediction of football match outcomes using tree-based classification models, including C5.0, Random Forest, and Extreme Gradient Boosting. Using historical data from ten seasons of the English Premier League (2007/2008 - 2016/2017) and fifteen initial features, a "backward wrapper" feature selection method was applied to optimize model accuracy. The results showed that Random Forest achieved the highest accuracy at 68.55% among the tested models, suggesting the need to explore additional methods to improve prediction accuracy[7].

Finally, the "Football Match Predictor" repository by the author with the pseudonym 'aziztitu' is also taken as a reference, which uses machine learning to predict the outcome of football matches based on halftime statistics. The repository uses structured data from the top five European leagues, covering results from the last nine years. The data was preprocessed by removing missing values and selecting key features through statistical tests and collinearity analysis. Three classification models were implemented: Naive Bayes, Random Forest, and Logistic Regression, achieving an accuracy of 70% with the Logistic Regression and Random Forest models, and 65% with Naive Bayes[8]. Additionally, the solutions provided on Kaggle were used as assistance in solving the problem[9][10].

II. Materials and Methods

This study focuses on investigating predictions of football match outcomes through advanced machine learning methodologies. A rigorous methodology will be employed for predictive modeling and model evaluation, exploring techniques to enhance prediction accuracy.

For the project, the Python programming language will be used along with the Pandas, Scikit-learn, and XGBoost libraries. Pandas will facilitate data manipulation and analysis, allowing for efficient cleaning and transformation of datasets. Scikit-learn will be employed to implement machine learning algorithms and validation techniques, providing tools for modeling and evaluating models in classification and regression tasks. XGBoost, in its implementation as a regressor and classifier, will be used to build and train high-performance models. These combined tools will enable the development of a robust and effective workflow for data preparation, model building, and evaluation. This study begins with data collection, where football match records are acquired from Kaggle, distributed across two separate files: one for training data (football _train.csv) and another for target data (football_target.csv"). These datasets are loaded using the Pandas library.

In the Data Preprocessing, to improve predictions, initially, teams or rows with fewer than 10 matches were removed, as these records often contain many missing values or only one recorded match, which could negatively impact predictions. Subsequently, columns containing the word "coach" were eliminated, halving the number of NaN values. Finally, rows with less than 50% and 70% of available information were dropped. Additionally, team IDs and names were separated into a dictionary for simplification, and team names were removed from the DataFrame.

TABLE I
NaN Values Analysis

Action to Remove	**Total NaN Values**	
	Initial	**After Removal**
Teams with less than 10 matches	1717256	376114
Columns with *coach*	376114	187936
Rows with less than 50% of information	187936	187936
Rows with less than 70% of information	187936	61268
Total rows of the *df*	77551	
Total columns of the *df*	168	

Continuing with the data preparation, several key transformations are applied to the football_train_df DataFrame. First, the date columns are formatted by converting them to datetime data type for easier temporal manipulation and analysis. Next, the league_name column is categorized to optimize storage and performance during processing. Then, label encoding is applied to the target and is_cup variables using Scikit-learn's LabelEncoder, transforming text labels into numeric values more suitable for machine learning models. Finally, the score column is split into two new columns, home_score and away_score, representing the home and away scores respectively, facilitating match result analysis.

Too, in the process of Feature Engineering for the football dataset, new variables or features are created from existing data to enhance the predictive capability of models using sklearn. Initially, core columns such as home and away team names, match date, league name, whether it's a cup match, scores for both teams, and the target outcome (target) are selected. Subsequently, additional features are computed, including days since the home team's last match, average home goals over the last 10 matches, average goals conceded by opponents in the last 10 matches, and similar metrics for the away team. Also incorporated are variables like the month of the match, counts of wins, draws, and losses in the last 10 matches for both home and away teams, and the average team and opponent ratings in those matches. These additional variables aim to capture historical patterns and recent team performances that could predict future match outcomes.

Where F equals to force, m to mass and a to acceleration.

III. Results

Show plots of any data collected and describe with words what your plots are showing. Describe the relationship between variables and time. Remember to number all your figures. This is the most critical part affect the technical achievement.

No picture, table, schematic, or graph should appear without a name (generally of the form Fig.1 o Table 1). None should appear without a reference to them by name in the main body of the writing. All figures and tables must be discussed in the text, including what it is, significant observations, and analysis.

Capitalize "Table" and "Fig." any time they are accompanied by specific table or figure numbers. Examples: "The measured data are plotted in Fig. 2. The figure shows a linear relationship in....". "The table shows ..." vs. "The data of Table 3..."

Student	Max Temperature
aabbbccc	35°
eeeddd	54°
eeeddd	54°

TABLE II
Temperature measurements performed for session 1.

Use your word processor to make "real tables" (i.e., boxed in, etc.). Center all tables and include a heading and caption with the appropriate table number below each table. For example, "Table 1: Temperature measurements performed for session 1."

Figures must be centered, and the figure number and caption is centered beneath the figure. For example, "Figure 1".

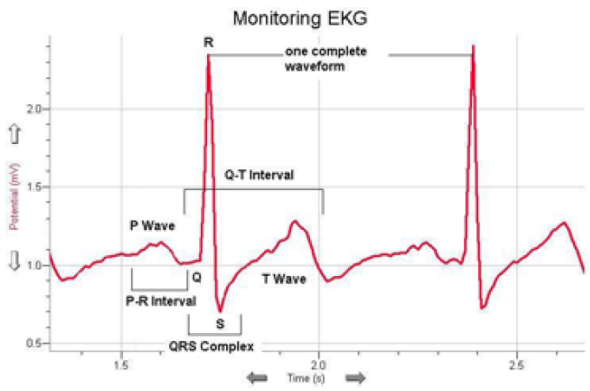

Fig. 1. Illustrations, graphs, and photographs may fit across both columns, if necessary. Your artwork must be in place in the article.

Always spell out table or Table. Give abbreviation of Figure, i.e., Fig., when used in the middle to end of sentence, but spell it out when used at the very start of the sentence.

All graphs must be done with a computer (i.e., spreadsheet software such as Microsoft Excel or even Matlab.). Do not include hand drawn graphs unless specifically instructed to do so.
Include a leading zero when a number's magnitude is less than 1 (use 0.83 instead of writing .83).
Use your word processor for Greek symbols for common engineering quantities as β, π, γ ,Ω .

IV. Discussion and Summary

Discuss any interesting result related to the materials used or to any claim from the introduction. Discuss your measurements using engineering terms (accuracy, precision, resolution, etc). Give technical conclusions. Restate the main objectives and how or to what degree they were achieved. What principles, laws and/or theory were validated by the experiment? Describe some applications of your results and comment any possible recommended future work.

References

[1] TecScience, "Can science predict soccer match results? - TecScience Predicting soccer results machine learning — Tec Science," TecScience, 30 de marzo de 2023. https://tecscience.tec.mx/en/science-communication/predicting-soccer-results-machine-learning/

[2] "Football Match probability Prediction — Kaggle." https://www.kaggle.com/competitions/football-match-probability-prediction/data

[3] J. Rincon, "Beginner's guide to European Soccer," Loyolan, March 6, 2023. Accessed: July 13, 2024. [Online]. Available at: https://www.laloyolan.com/sports/beginners-guide-to-european-soccer/article_7f585b4c-fe62-5ccd-a6e9-1026dc65669f.html

[4] Bundesliga, "How is European soccer structured with leagues and cup competitions?," bundesliga.com - The Official Bundesliga Website, 21 de febrero de 2020. https://www.bundesliga.com/en/faq/what-are-the-rules-and-regulations-of-soccer/how-is-european-soccer-structured-with-leagues-and-cup-competitions-10568

[5] Chandra B, Jennet Shinny D, Keshav Adhitya M et al. Prediction of Football Player Performance Using Machine Learning Algorithm, 01 March 2024, PREPRINT (Version 1) available at Research Square https://doi.org/10.21203/rs.3.rs-3995768/v1

[6] R. Bunker, C. Yeung, y K. Fujii, "Machine Learning for Soccer Match Result Prediction," Nagoya University. [En línea]. Disponible en: https://arxiv.org/pdf/2403.07669

[7] Yoel F. Alfredo, Sani M. Isa, "Football Match Prediction with Tree Based Model Classification," International Journal of Intelligent Systems and Applications (IJISA), Vol.11, No.7, pp.20-28, 2019. DOI: 10.5815/ijisa.2019.07.03

[8] Aziztitu, "football-match-predictor," GitHub. https://github.com/aziztitu/football-match-predictor/blob/master/README.md

[9] "Football Match probability Prediction — Kaggle." https://www.kaggle.com/competitions/football-match-probability-prediction/code

[10] Masterofdeception, "Football Prediction by XGBoost," Kaggle, 9 de julio de 2023. https://www.kaggle.com/code/masterofdeception/football-prediction-by-xgboost