Exploring the Impact of Artificial Intelligence in Predicting English Premier League Football Matches

[Sports Analytics]

Taiwo Mubarak Oladapo
x22107312
Research in Computing CA2
National College of Ireland

Abstract

This study investigates the prediction of English Premier League (EPL) football games using Artificial Intelligence (AI). It aims to determine how effective AI is at forecasting EPL matches since interest in using it to analyze sports is growing. By developing and testing machine learning models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, and Extreme Gradient Boosting (XGBoost), particularly for English Premier League matches, several issues with earlier research in this field have to be addressed. In order to train AI models on how to forecast matches, the project entails analyzing EPL match data with different evaluation metrics such as accuracy, precision, recall, F1 Score, and confusion matrix. This research intends to enhance the area of sports analytics and better understand what AI can and cannot achieve when forecasting EPL games.

Keywords— Artificial Intelligence, English Premier League, Football Matches, Predictive Models, Machine Learning

1 Introduction

Artificial Intelligence (AI), a branch of research established in the 1950s, is defined as a system's ability to effectively absorb and learn from external inputs as well as adopt the learning outputs to accomplish certain goals and solve problems via adaptability Kaplan & Haenlein (2019). Artificial intelligence (AI) technological advancements have significantly advanced the field of sports analytics. Researchers and sports fans are particularly interested in making predictions about the results of football games. The English Premier League (EPL) is a highly prestigious and widely popular football league, attracting a massive global audience of approximately 4.7 billion people. As the highest tier of the English football league system and the world's largest sports community, the EPL generates significant enthusiasm and interest among football fans worldwide. Consequently, predicting the outcomes of EPL matches has become a phenomenon within the football community, with countless supporters and experts offering predictions through various means before match kick-off Raju et al. (2020).

The research question that guides this study is: To what extent would Artificial Intelligence (AI) particularly machine learning models help in predicting English Premier League (EPL) matches? This research question is worth investigating for several reasons. Firstly, the growing interest in sports analytics, fueled by technological advancements and data availability, has led to a surge in research focused on predicting outcomes in sports events like English Premier League football matches. Accurately predicting match results holds potential applications in areas such as sports betting, fantasy football leagues, and decision-making for team managers and coaches. Artificial intelligence (AI) techniques can enhance the accuracy of these predictions, contributing to improved strategic decision-making. AI has already demonstrated its efficacy in various fields, including sports, through machine learning algorithms, data analysis techniques, and predictive modeling. Exploring the impact of AI in predicting football match outcomes provides valuable insights into the benefits and limitations of these techniques in the realm of sports. The English Premier League, being a highly popular and competitive league with a massive fanbase, serves as an ideal context for investigating the effectiveness of AI methods in predicting match results Fialho et al. (2019). Sports analytics and the useful uses of precise forecasts are gaining popularity. Football is a widely followed sport on all continents, and the Premier League is very competitive and popular. The rising need for precise forecasts may be met by applying AI to match result predictions. The use of AI-based forecasts for decision-making can be advantageous for sports betting businesses, football teams, and fantasy football platforms. Significant advancements have been made in AI technology, notably in the area of machine learning. Numerous prediction challenges have shown decision forests and neural networks to be excellent algorithms. The success of these algorithms in the context of Premier League forecasts may be compared and evaluated, which can advance knowledge and help choose the best strategy Azeman et al. (2021). Due to the necessity of expanding on prior research, considering current AI developments, satisfying the growing need for precise forecasts, and investigating the wider implications of AI in this sector. The study, which focuses on machine learning methods for forecasting match results, lays the groundwork for future research. However, new chances to assess various AI strategies, such as deep learning or ensemble techniques, are provided by breakthroughs in AI. This study is pertinent since sports analysts like sports betting organizations, football teams, and fantasy football platforms heavily rely on sports analytics to make critical decisions. Accurate forecasts are becoming more vital in sports. Therefore, studying how Artificial Intelligence (AI) impacts predictions of English Premier League matches would help in overcoming limitations and provide valuable insights into the broader use of this technology in fast decision-making Choudhary & Sidhu (2020).

The structure of this document is as follows: Section 2 describes the literature review. It entails what other researchers have done in the past regarding this topic. It also explains their findings, comparison and contributions that can be added to the existing knowledge. Section 3 describes the research methodology. It entails the data selection, data preprocessing, data transformation, selected AI algorithms and evaluation metrics. Ethical considerations and the project plan is also described here.

2 Literature Review

The literature review would look at what other researchers have found when it comes to using Artificial Intelligence (AI) to predict English Premier League Matches (EPL). This section would compare and contrast their work and see how it relates to our study. By doing this, it would help to identify any missing information that this research can fill. The previous studies that would be discussed here will help to gain a better understanding of the topic and provide more insights into this research.

2.1 AI Techniques in Sport Analytics

One popular method that researchers have explored is using BP neural network models to predict football match outcomes. These models are good at handling complex input data and they have been used by Guan & Wang (2021) to create more accurate prediction models. A gray fuzzy extreme learning machine prediction combination model based on a neural network was developed, which improved prediction accuracy and overall performance. However, it is important to note that BP neural network models may struggle with the dynamic and complex nature of football match data. To improve prediction accuracy, researchers have also explored alternative data sources beyond traditional performance indicators. For example, using candlestick charts derived from betting market data as features in a predictive model has shown promise. This approach eliminates the need for specific domain knowledge in sports. However, it is essential to consider biases and uncertainties associated with betting market data, because it can introduce inaccuracies into the prediction model Hsu (2020). Moreover, choosing the right supervised learning techniques for training classifiers in football match prediction models has been a topic of interest. Various methods like Logistic Regression, Gradient Boosting Decision Trees, and Random Forest have been compared. The findings revealed that the Random Forest model performed better than other approaches, achieving promising accuracy rates in forecasting match outcomes with a training set accuracy of 66.7% and a test set accuracy of 63.8% Hu & Fu (2022). However, it is crucial to select the appropriate machine learning algorithm based on the dataset's characteristics and the specific problem being addressed, as different algorithms have different strengths and weaknesses in capturing the complexities of football match dynamics.

2.2 Predictive Models in Football

To make accurate predictions about match results, it is important to do a lot of work on the data itself. Baboota & Kaur (2019) emphasised how crucial techniques like feature engineering and exploratory data analysis are. A model that used gradient boosting was developed, which performed reasonably well with a score of 0.2156 on the Ranked Probability Scale (RPS). However, even though their model did okay, it still didn't do as well as the predictions made by betting companies. This highlights the difficulty of beating the accuracy of bookmakers' forecasts. In a simpler way, Beal et al. (2021) came up with a different approach. Information from sports journalists and statistics about matches were used to make predictions. Their models were way more accurate, with a 63.18% improvement compared to the usual methods. This means that by considering the bigger picture, their models can help make better predictions about the result of matches. In a recent study by KINALIOĞLU & KUŞ (2023), a bunch of football games played in

European leagues was looked at. A technique called web scraping was used to gather data and classification methods was used to analyze it. The models did way better than the usual algorithms used for classification, and it was able to predict different types of match outcomes with really high accuracy. In the investigation of predicting football games using tree-based model algorithms such as C5.0, random forest, and extreme gradient boosting, the random forest approach produced the greatest accuracy of 68.55%. The major drawback of the analysis was its feature collection Alfredo & Isa (2019). Different algorithms were used in several investigations, however, the predicted accuracy was only 59% and 58.5% respectively Baboota & Kaur (2019),Sathe et al. (2017). The production of football games has several restrictions and a logistic regression model was utilized which only yielded two outcomes (home or not home) but a football game can end in three outcomes (a home victory, an away victory, or a draw) Rana et al. (2019). Azeman et al. (2021) came up with two prediction algorithms called the Multiclass Neural Network and the Multiclass Decision Forest. Both algorithms were compared and it was discovered that the Multiclass Decision Forest performed better in terms of accuracy. This shows that the decision forest algorithms have a lot of potential when it comes to predicting the outcomes of football matches. Pipatchatchawal & Phimoltares (2021) propose a classification models that use video game ratings of players and teams instead of looking at in-game statistics. These models were tested in experiments and it was found that they did better than other simple classification models in terms of accuracy. The accuracy rates they achieved were 56.5332% and 56.8002% respectively. Elmiligi & Saad (2022) analyzed a large dataset of soccer match results and studied different things like how skilled the players were, how well the teams performed, and whether the match was played at home or away. A special model that combined different methods was proposed, and it did really well. It predicted match results with an accuracy of 46.6% on the test set, which is pretty good. A ranking probability score of 0.216 was achieved, which indicates how well the model ranked the predictions. This study gives us some insights into how effective these kinds of models can be for predicting soccer match results.

Based on the current research question, these past studies help to understand the challenges of predicting soccer match results and how data scientists tackle them. A great emphasis is placed on how feature engineering, data analysis, and model selection are all really important in achieving accurate predictions. However, even with all that effort, it's still hard to beat the predictions made by bookies. This shows that getting better accuracy is really tough. By looking at and comparing these studies, what needs improvement in the field of predicting soccer match results would be discovered.

2.3 Research Niche

The area of research that focuses on how Artificial Intelligence (AI) can be used to predict the outcomes of English Premier League (EPL) football matches is described below:

- The utilization of various prediction models and methodologies: The studies that were looked at tried out different ways to predict the outcomes of EPL matches. Different kinds of models like BP neural networks, gray fuzzy prediction models, extreme learning machines, and supervised learning techniques such as Logistic Regression, Gradient Boosting Decision Trees, and Random Forest were used. Hybrid models that combine machine learning with statistical approaches were also used. But there's still more work to be done to figure out which one works best for predicting EPL match outcomes by exploring and comparing different models.

- Prediction Accuracy Evaluation: Predicted match results were compared with the actual outcomes to see how well the models performed. Other factors were also examined, such as how fast the models processed the data, how accurate the data transformation was, and the scores for analyzing the matches. Ranked Probability Scores (RPS) were used to measure how well the models ranked the predictions. The unique thing about this research is that it focuses specifically on improving the accuracy and performance of prediction models for EPL matches.

- Comparison with bookmakers forecasts: Both sets of studies that were looked at agree that it's really hard to be better than bookmakers when it comes to predicting outcomes accurately. While some models perform less accurately than predictions made by bookmakers, others did a pretty good job. The research in this area focuses on finding ways to make prediction models even better so that they can either surpass or at least be as accurate as bookmakers when it comes to predicting EPL match outcomes.

Considering the area of research already chosen, the primary objective is to create and test prediction models and methods that are specifically designed for predicting English Premier League (EPL) match results. The aim is to improve the accuracy of predictions to make the systems perform better than the bookmaker's forecasts. By improving on this, the research would give useful information about how Artificial Intelligence will be used to predict EPL football matches. This would be a valuable contribution to the field of sports analytics and the use of AI in the sports industry.

3 Research Methods & Specifications

3.1 Research Method

This section would explain the technique that will be used for this research project. The aim is to thoroughly outline all the numerous actions and phases that would be necessary to answer the research question. The Knowledge Discovery in Databases (KDD) technique, which includes the following steps: Data Selection, Data Pre-processing, Data Transformation, Data Mining, and Evaluation/ Interpretation, would be used in this suggested research.

These steps would guide in selecting the right data, preparing it for analysis, transforming the prepared data, mining which involves exploratory data analysis, and evaluating and interpreting the results. This technique would be the basis of our research approach.

3.1.1 Data Selection

The data needed for this proposed study would be gathered from easily available sources on Kaggle. There aren't any ethical issues because the data is publicly available. The dataset consists of 12,026 observations and 8 features. The following is a description of the features that have been chosen for training, testing, and estimation:

- Season End Year: This refers to the year in which the football season ends. It indicates the specific season for which the match data is recorded.

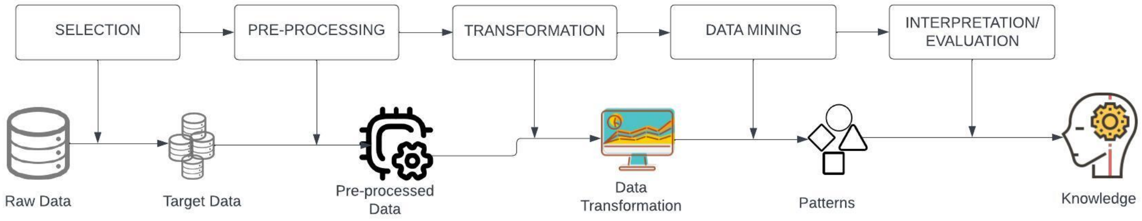

Figure 1: The KDD Lifecycle

- Wk: The ”Wk” stands for ”Week” and it represents the specific week of matches within a given season. It facilitates maintaining orderly organization and record-keeping for the matches. The first weekend of matches of a season, for instance, would be denoted as Wk 1, the second weekend as Wk 2, and so on.

- Date: This feature represents the date when a specific match happened. It provides the exact day, month, and year of the match, which helps in analysing the matches in relation to time and putting them in the right order based on their dates.

- Home: This feature refers to the name of the team that is playing the match at their own stadium. It helps in identifying which team has the advantage of playing on their home ground.

- Home Goals: This feature refers to how many goals the home team scored in a match. It helps in understanding how well the home team performed in terms of their attacking abilities and scoring goals during a particular game.

- Away Goals: This feature refers to how many goals the away team scored in a match. It helps in understanding how well the away team performed in terms of their attacking abilities and scoring goals during a particular game.

- Away: This feature represents the name of the team that is playing the match on the opponent’s ground or away from their own designated home ground. It helps in identifying which team is the visitor.

- FTR: FTR stands for ”Full Time Result” and it helps in identifying how the game ended. It can take one of the following values to describe the outcome of the match:

  - ”H” (Home win): The match was won by the home team.

  - ”A” (Away win) The match was won by the away team.

  - ”D” (Draw): There was no winner as the game was a draw.

3.1.2 Data Pre-processing

The data would undergo pre-processing to ensure it is accurate, reliable, and suitable for analysis. The following steps would also be taken to prepare the data:

- Importing all necessary libraries.
- Reading the dataset into the coding environment.
- Checking for any missing values and handling them accordingly if any are discovered.

3.1.3 Data Transformation

The structured data would be improved at this stage to facilitate analysis. A method known as Exploratory Data Analysis (EDA) would be used to get a better understanding of the data and to discover important patterns. The EDA involves creating new features from existing columns, which would enhance the structure data and provide more insights for the research. Data transformation techniques such as feature selection would be applied to prepare the data for training the Artificial Intelligence (AI) models. This method would make the dataset more useful and of higher quality resulting in better analyses and more accurate predictions when in the modeling phase.

3.1.4 Data Mining

In this stage, data mining techniques would be used to explore the transformed dataset and find meaningful patterns and insights. To correctly predict the results of the English Premier League (EPL), several machine learning techniques would be used. The methods that would be used are Extreme Gradient Boosting (XGBoost), Decision Tree, Logistic Regression, Support Vector Machine (SVM), and Random Forest. These algorithms were chosen based on their ability to handle the complexity and dynamics of football match data and how well they perform in predicting EPL match outcomes. The data collected would be trained and tested on the models. To improve the performance of the models, Cross-validation methods and feature selection techniques would be used. The aim is to create reliable and accurate prediction models that can effectively forecast the results of EPL football games.

3.2 Research Resources

The primary resource for this research would be the premier league match dataset of 1993 to 2023 obtained from Kaggle. These datasets provide detailed information on EPL matches, including season end year, week, date, home team, home goals, away goals, away team, and full-time result. The dataset would be considered the primary resource, while supplementary resources would include books on sports analytics, academic publications, AI approaches, and predictive models. These resources would offer beneficial concepts, tactics, and methods that have been applied in this subject in the past. Python programming language and libraries like Pandas, NumPy, and Scikit-learn will be used to process, analyze, and develop the prediction models. These tools would help with data preparation, transformation, and model construction. It is important to have a reliable computer with sufficient processing capacity and memory to efficiently perform the data analysis and model training tasks for this project.

3.3 Evaluation

The evaluation of the developed prediction models would be performed using appropriate metrics and techniques. The accuracy of the models in predicting EPL match outcomes would be assessed from the EPL dataset. The performance of the models would be assessed using metrics including accuracy, precision, recall, F1-score, and confusion matrix. To verify the robustness of the models and reduce overfitting, cross-validation techniques like k-fold cross-validation would be used. The models would be evaluated on both training and testing datasets to assess their generalization ability. The results would be analyzed and compared with the existing literature and benchmark predictions, including bookmakers' forecasts, to identify the strengths and limitations of the developed models.

3.4 Ethical Considerations of the Research

Throughout the study process, ethical issues would be addressed. There are no significant ethical concerns concerning data privacy and confidentiality because the data utilized in this study are public and contain no sensitive or private information. The study would also adhere to all relevant ethical norms for research, including obtaining all necessary authorizations, conducting objective analysis, and transparently disclosing results. Any limitations or potential biases in the data and models would be acknowledged and underlined to ensure the credibility and integrity of the study conclusions.

3.5 Project Plan

My plan to conduct my research plan is shown in Figure 2.

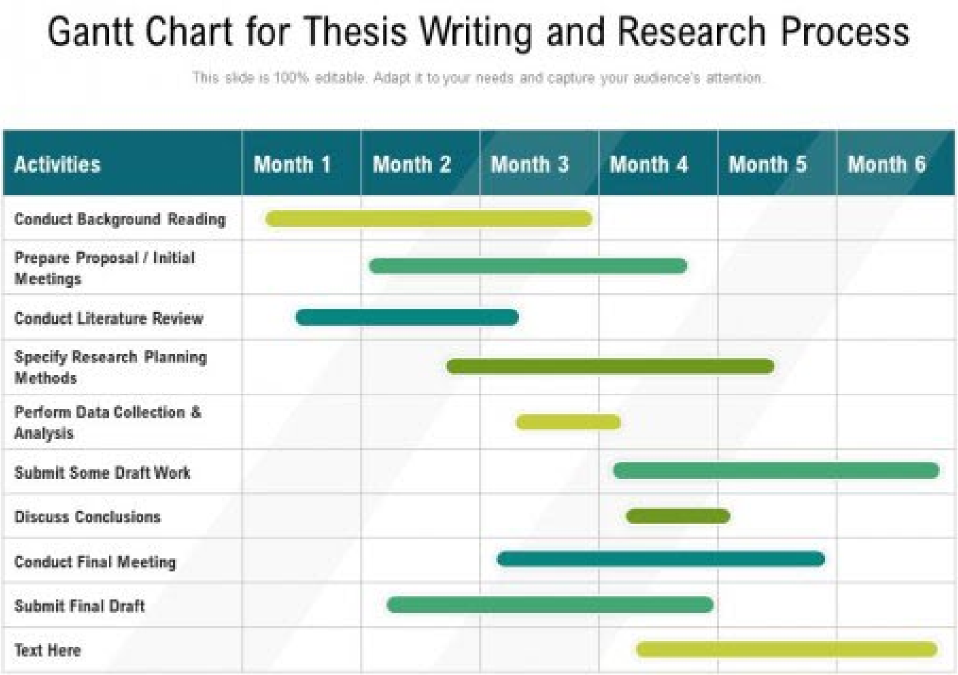

Figure 2: Project Plan for Research Work

References

Alfredo, Y. F. & Isa, S. M. (2019), 'Football match prediction with tree based model classification', *International Journal of Intelligent Systems and Applications* **11**(7), 20–28.

Azeman, A. A., Mustapha, A., Razali, N., Nanthaamomphong, A. & Abd Wahab, M. H. (2021), Prediction of football matches results: Decision forest against neural networks, *in* '2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)', pp. 1032–1035.

Baboota, R. & Kaur, H. (2019), 'Predictive analysis and modelling football results using machine learning approach for english premier league', *International Journal of Forecasting* **35**(2), 741–755.

Beal, R., Middleton, S. E., Norman, T. J. & Ramchurn, S. D. (2021), 'Combining machine learning and human experts to predict match outcomes in football: A baseline model', *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(17), 15447–15451. https://ojs.aaai.org/index.php/AAAI/article/download/17815/17622.
URL: *https://app.dimensions.ai/details/publication/pub.1150866205*

Choudhary, V. & Sidhu, J. S. (2020), 'Premier league match result prediction using machine learning'.

Elmiligi, H. & Saad, S. (2022), Predicting the outcome of soccer matches using machine learning and statistical analysis, *in* '2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)', pp. 1–8.

Fialho, G., Manhães, A. & Teixeira, J. P. (2019), 'Predicting sports results with artificial intelligence–a proposal framework for soccer games', *Procedia Computer Science* **164**, 131–136.

Guan, S. & Wang, X. (2021), 'Optimization analysis of football match prediction model based on neural network', *Neural Computing and Applications* **34**(4), 2525–2541.
URL: *https://app.dimensions.ai/details/publication/pub.1136819893*

Hsu, Y.-C. (2020), 'Using machine learning and candlestick patterns to predict the outcomes of american football games', *Applied Sciences* **10**(13).
URL: *https://www.mdpi.com/2076-3417/10/13/4484*

Hu, S. & Fu, M. (2022), Football match results predicting by machine learning techniques, *in* '2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)', pp. 72–76.
URL: *https://app.dimensions.ai/details/publication/pub.1153661300*

Kaplan, A. & Haenlein, M. (2019), 'Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence', *Business Horizons* **62**(1), 15–25.
URL: *https://www.sciencedirect.com/science/article/pii/S0007681318301393*

KINALIOĞLU, İ. H. & KUŞ, C. (2023), 'Prediction of football match results by using artificial intelligence-based methods and proposal of hybrid methods', *International Journal of Nonlinear Analysis and Applications* **14**(1), 2939–2969.

Pipatchatchawal, C. & Phimoltares, S. (2021), Predicting football match result using fusion-based classification models, *in* '2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE)', pp. 1–6.

Raju, M. A., Mia, M. S., Sayed, M. A. & Riaz Uddin, M. (2020), Predicting the outcome of english premier league matches using machine learning, *in* '2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)', pp. 1–6.

Rana, D., Vasudeva, A. et al. (2019), 'Premier league match result prediction using machine learning'.

Sathe, S., Kasat, D., Kulkarni, N. & Satao, R. (2017), 'Predictive analysis of premier league using machine learning', *IJ Innovative Research in Computer and Communication Engineering* **5**(3), 4121–4124.