ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240718091147

Subject review
Received: 2024-07-18, Accepted: 2025-01-20

Supervised Machine Learning Methods in Predicting English Premier League Game Outcomes

Mateo Vujčić, Tomislav Horvat*, Dejan Barešić, Dražen Crčić

Abstract: This paper explores the application of machine learning (ML) models to predict soccer game outcomes in the English Premier League. Utilizing data from fbref.com, various Features and target variables were analyzed to assess their influence on game results. The study implemented multiple ML algorithms, including logistic regression, Naive Bayes, decision trees, k-nearest neighbors, random forest, AdaBoost, and multilayer perceptron. The results highlighted significant variations in prediction accuracy across different teams and models, with the Random Forest model achieving the highest average accuracy. The findings underscore the importance of careful algorithm selection and data processing to enhance prediction precision. This research contributes to the field of sports analytics, providing insights that can be applied to improve tactical planning and decision-making in sports. Future work will focus on optimizing models and exploring additional data sources to further increase prediction accuracy.

Keywords: English Premier League; machine learning; predicting outcomes; soccer; supervised learning; Train&Test; validation methods

1 INTRODUCTION

Predicting outcomes in various fields is an increasingly essential and intriguing task. Whether it's making economic decisions, diagnosing health conditions, or devising sports strategies, the advantage provided by foreseeing future events is invaluable. Accurate predictions necessitate a thorough understanding and appropriate interpretation of relevant data, which can be structured, semi-structured, or unstructured. This introduction highlights the importance of outcome prediction and emphasizes the need for access to relevant data to enhance prediction accuracy.

The ability to predict outcomes is vital across diverse segments of life. In the business sector, predictions enable companies to make strategic decisions, adapt to market conditions, and minimize risks. In healthcare, accurate predictions contribute to timely diagnoses and personalized treatment plans, significantly improving patient care. In sports, data analytics is crucial for tactical planning and optimizing performance. This underscores the general significance of outcome prediction and the crucial role of relevant data in achieving precise predictions.

Access to relevant data is fundamental to achieving optimal results, regardless of the application area. Such data provides deeper insights into current conditions and forms the basis for informed decision-making. In business, relevant data includes information about consumer habits; in medicine, it encompasses biological and genetic markers; and in sports, performance analysis is key. Thus, success in outcome prediction relies on the efficient collection, analysis, and interpretation of relevant data.

The diversity of data presents both challenges and opportunities for proper collection, analysis, and interpretation. Understanding the different types of data is crucial for developing effective methods for outcome prediction. Structured data consists of organized information stored in defined structures like tables or databases, with clearly specified attributes and relationships. These data allow for easy searching, filtering, and analysis, forming the basis for quantitative and statistical analyses in business and science.

Semi-structured data, while partially organized, lack strictly defined attributes or relationships. Formats such as XML, JSON, and YAML represent these data, offering greater flexibility compared to strictly structured data but requiring more complex analysis methods. Unstructured data, on the other hand, includes information without clear organization, such as text documents, audio recordings, images, and videos. These data do not follow predefined patterns, necessitating the use of advanced processing techniques to extract relevant information [1].

The variety of data types offers valuable information but also presents processing challenges. For outcome prediction, combining different data types is often used for deeper insights and thorough analysis.

Predicting the outcomes of sports events has evolved into a significant area of research, especially in the realm of high-profile sports leagues like the EPL. In modern times, the vast amount of data collected from various sports events is used not only for detailed analysis but also for anticipating future game outcomes or even results.

After the introduction and a short motivation, section 2 provides an overview of existing papers from other researchers, focusing on studies related to the prediction of outcomes in sports, not exclusively limited to soccer. The obtained results, set of used features, validation methods will be analyzed and specific conclusions will be drawn. In Section 3 the input dataset and features of presented model will be described. Section 4 will outline the methodology, explaining the techniques and processes applied in the research. In Section 4, a predictive model will also be explained and presented. Section 5 presents and discusses the research results, while Section 6 concludes with a summary of the findings and suggestions for future research.

2 RELATED RESEARCH

Predicting outcomes in sports certainly represents a very interesting, but also challenging area. In this section, an overview of papers related to the outcome prediction in sports will be given, with an emphasis on soccer.

Tab. 1 provides an analysis of various studies conducted by researchers focusing on predicting sports outcomes [2-17]. These studies have primarily utilized supervised machine learning (ML) classification methods. However, the comparison of results is often challenging due to the use of different datasets across studies. The most commonly used validation method is data partitioning, with fewer researchers employing cross-validation. Due to the interconnected nature of sports events, cross-validation is not always suitable for sports predictions [18]. Best results are typically achieved using smaller datasets. This review analyzed 16 studies from the period 2008 to 2024, focusing on soccer. The most frequently studied league is the English Premier League, accounting for 55.56 % of the research. All studies presented in Tab. 1 utilize historical results, while some additionally incorporate data such as team rankings, weather conditions, individual player statistics, or even betting odds.

Table 1 Analyzed papers categorized by the publication year

#	Year	Input Dataset	Algorithms Used	Best Result (%)
[2]	2008	English Premier League (EPL) (2002-2007)	Neural Network	58.40
[3]	2011	Dutch Soccer Leagues (last 15 seasons)	Linear Regression	57.00
[4]	2011	Champions League	Neural Network, LogitBoost	68.80
[5]	2013	Spanish Soccer League (2008)	Bayesian Network	92.00
[6]	2014	EPL (2014)	Logistic Regression	93.00
[7]	2015	Dutch Eredivisie (2000-2012)	LogitBoost	56.10
[8]	2016	EPL (2010-2015)	Logistic Regression	69.50
[9]	2018	Spanish Soccer League (2012-2016)	Logistic Regression	71.63
[10]	2020	Multiple European Leagues (2013-2017)	Random Forest	75.62
[11]	2021	5 European Leagues (2006-2017)	Integrated Model	81.77
[12]	2021	UEFA Champions League (2016-2017)	Logistic Regression	81.00
[13]	2021	EPL (2005-2006)	Decision Tree, Neural Network	88.00
[14]	2022	Top 5 European Leagues (2014-2017)	Bayesian Network	92.01
[15]	2022	EPL (2018-2019)	SVM, Bayesian Network	61.32
[16]	2022	EPL (2019-2021)	Logistic Regression, k-NN, Random Forest	70.00
[17]	2024	World Cup (2022)	Random Forest	69.20

As can be seen in Tab. 1, the algorithms examined include logistic regression, Naive Bayes, decision trees, *k*-nearest neighbors (*k-NN*), random forest, LogitBoost, and multilayer perceptron (MLP) [19]. Each algorithm was tested on datasets comprising historical game results, team statistics, and factors related to home and away games. Most of these studies employ multiple ML algorithms, meaning the analyzed sample includes over 50 different prediction results. However, comparing these results is challenging or nearly impossible due to the use of different datasets and leagues of varying competitiveness. Only the best results are considered when different feature sets or datasets are used by the authors.

Fig. 1 shows a box plot of the range of results obtained by the analyzed studies in predicting soccer game outcomes. The box plot provides a detailed insight into the accuracy of different ML algorithms in predicting soccer outcomes. The results show significant variations in accuracy among the algorithms. Logistic regression is noted for its wide range of accuracy, indicating its adaptability to different datasets and processing methods. Naive Bayes shows consistent and stable results, while decision trees and k-NN have a broader accuracy range, suggesting their performance may depend on specific conditions and data features. Random forest often achieves high accuracy, making it a reliable choice for predicting soccer outcomes. LogitBoost also shows good results but does not reach the highest accuracy level of some other algorithms. Multilayer perceptron has consistent accuracy, indicating a potential need for further model fine-tuning or feature adjustment.

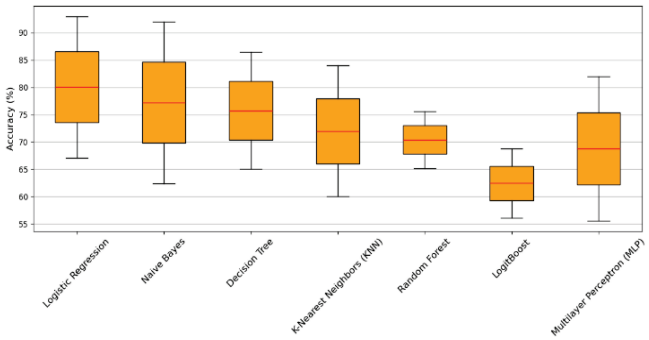

Figure 1 Box Plot Analysis of Results from Other Studies

These results underscore the importance of careful algorithm selection and data processing methods to achieve optimal outcomes in sports prediction. Using appropriate features and model adjustments is crucial for achieving high accuracy and reliability in predictions. Additionally, other researchers emphasize the significance of feature selection and feature extraction. Feature selection is often performed using specific methods and based on expert experience, aiming to identify the most relevant features that contribute to the predictive model performance. Feature extraction involves transforming raw data into informative and non-redundant features, which can enhance the predictive power. The analysis also shows that while some algorithms can achieve high accuracy under specific conditions, their overall success in practice is often determined by consistency and adaptability to different datasets. As noted, a large amount of data related to sports events is generated daily, both structured and unstructured. With the growing volume of data, the number of relevant databases containing various sports statistics also increases. Reviewing other researchers' work makes it clear that the main sources of information are mostly official sports organization websites. The study analyzed several studies focused on predicting sports outcomes or extracting useful information and patterns related to sports. Furthermore, as the availability and volume of sports data increase, so does the interest in using this data

for analytical purposes and predicting future outcomes. Sports organizations often provide detailed statistics and historical data through their official channels, enabling researchers to access valuable information for modeling and analysis. Many studies also use these data sources to develop and test various ML algorithms, striving to improve prediction accuracy in sports. There are also many other research studies, both scientific and review articles, not strictly related to predicting soccer outcomes, that provide valuable results and conclusions in predicting sports outcomes.

In addition to the mentioned papers concerning the outcome prediction in soccer, there is also a substantial body of review and scientific papers dedicated to predicting outcomes in various sports disciplines. These papers encompass a wide range of methodologies and approaches to enhance the accuracy and reliability of sports predictions. Furthermore, numerous scientific studies focus on the analysis of datasets specifically tailored for predicting sports outcomes, highlighting the importance of data quality and feature selection in developing robust predictive models. The authors in the study [20] and [21] provided an overview of the comparison of using ML algorithms and derived models in predicting outcomes in sports for various disciplines. The study [22] provides a comprehensive review of computer vision in sports. Studies [23-28] propose models for predicting outcomes in sports, specifically basketball and soccer. Study [29] discusses the long-term influence of technical and physical performance indicators, and situational variables on finale game outcome in soccer, while study [30] focuses on variations in the physical demands and technical performance of professional soccer teams.

Mentioned studies contribute to the growing knowledge base, offering valuable insights and advancements in sports analytics and outcome prediction.

3 DATA COLLECTION

The dataset was scraped from *fbref.com* platform to ensure it included all necessary information for an in-depth analysis of team performance, results, and key parameters throughout the season [31]. The *fbref.com* platform offers a robust database of soccer statistics, essential for evaluating various aspects of soccer games.

The features were chosen for their potential to influence game outcomes, providing valuable insights into different factors that could affect the game. The target variables represent the direct outcomes that the predictive models aim to predict, serving as a measure of the models' accuracy.

3.1 Dataset

The dataset used in this research includes features, presented in Tab. 2, and target variables, listed in Tab. 3, which are relevant to predicting soccer game outcomes. These features were selected for their potential influence on game results. Detailed data from English Premier League (EPL) games were carefully collected from the *fbref.com* platform, which provides extensive statistics on game outcomes, team performance, and individual player metrics. This database serves as a comprehensive foundation for the analysis conducted in this study. The selected features capture various factors that influence soccer game outcomes, while the target variables correspond to the results that the predictive models aim to predict.

In particular, the target variable in this analysis is the game outcome (win/loss/draw), which is derived from the combination of the features presented in Tab. 2. The model utilizes a variety of features, including but not limited to expected goals (*xg*), expected goals against (*xga*), possession percentage, and opponent team strength, to predict the final game result. The choice of features was motivated by their relevance and potential impact on the game outcome, allowing for a more robust and accurate prediction model.

Table 2 Features included in the dataset

Features	Description
venue_code	Encoded value representing the game venue (home or away)
opp_code	Encoded value representing the opposing team
hour	Hour of the day when the game was played
day_code	Encoded value representing the day of the week the game was played
xg	Expected goals for the team
xga	Expected goals against the team
poss	Possession percentage of the team

Table 3 Target variables included in the dataset

Target Variable	Description
gf	Goals for the team
ga	Goals against the team
sh	Total shots by the team
sot	Shots on target by the team

Expected goals (*xg*) is a metric used to assess the quality of scoring opportunities by considering factors such as shot location, angle, assist type, distance from goal, and defensive pressure. ML or regression models use historical data to estimate the likelihood of a shot resulting in a goal, assigning an *xg* value between 0 and 1, with higher values indicating greater goal-scoring probability. Similarly, expected goals against (*xga*) estimates the number of goals a team is likely to concede, based on the quality of shots taken by the opposing team. Like *xg*, *xga* incorporates factors such as shot location, type, defensive pressure, and goalkeeper positioning. The cumulative *xga* values reflect a team's defensive performance.

4 METHODOLOGY

ML has become a crucial tool for predicting sports outcomes, aiding coaches and managers in forecasting game results, assessing player or team performance, detecting injuries, and identifying emerging talents. Furthermore, such data is invaluable to the general public, especially in sports betting, offering a basis for more informed decisions [32].

The EPL, known for its high level of competitiveness, offers a wealth of data that is extremely valuable for analysts. Analyzing factors such as team formations, historical encounters, player injuries, and current form is essential for applying supervised ML methods to predict game outcomes [33]. Individual sports with binary outcomes, such as tennis or chess, are relatively easier to predict compared to soccer, which involves multiple variables and presents a much

greater challenge. However, studying unexpected sports events, like Leicester City's surprising Premier League title win in the 2015/2016 season, highlights the unpredictability of sports competitions and the potential for surprises on the field.

4.1 Prediction Tools

Scikit-learn is an open-source ML library for Python, known for its simple yet effective implementation of a wide range of ML algorithms and model evaluation tools. Due to its broad functionality in model creation, training, and evaluation, the library has gained popularity among researchers and practitioners in the field of ML [34].

Logistic regression, found in the *sklearn.linear_model* module, is used for binary and multi-class classification. It estimates the probability of class membership using a logistic function, transforming input features into probability estimates. This algorithm is particularly useful due to its simplicity and interpretability. Algorithms such as GaussianNB, MultinomialNB, and BernoulliNB from the *sklearn.naive_bayes* module are based on Bayes' theorem with an assumption of conditional independence between features. These algorithms are particularly effective for large datasets and are often used in classification problems such as spam filtering and sentiment analysis. The *sklearn.tree* module includes decision tree algorithms used for classification tasks. This algorithm splits data into subsets based on attribute values through a series of decision rules, resulting in easily interpretable structures. Despite their tendency to overfit, decision trees can be very effective when used with properly tuned parameters.

The *sklearn.neighbors* module's *k-NN* algorithm identifies the k-nearest neighbors to each data instance for classification tasks. This method is known for its simplicity and intuitiveness, making it highly effective for multi-class problems and complex data distributions [35]. Part of the *sklearn.ensemble* module, random forest is an ensemble method that creates multiple decision trees from randomly selected features and data samples. This method ensures diversification among the trees and reduces the risk of overfitting, making it widely used in fields such as biomedicine, ecology, and economics for classification and regression problems. The *sklearn.neural_network* module's *MLPClassifier* uses a structure of multiple layers of perceptrons for classification tasks. MLP employs supervised learning techniques and is trained using the backpropagation algorithm, making it effective for solving complex classification problems. These algorithms and tools from the *Scikit-learn* library facilitate the development and testing of ML models in various applications, enabling researchers and practitioners to focus on innovation and solving specific problems in the field.

4.2 Steps and Implementation

The process of developing a ML model using Python libraries, particularly *Scikit-learn*, involves several key steps. Block diagram of the proposed model is shown in Fig. 2. Through this methodology, the entire model development process is explained, from initial data processing to final model evaluation. To begin, required libraries for data manipulation and various ML models from *Scikit-learn* are imported. Game data is then loaded from a CSV file, setting the first column as the index. The date column is converted to datetime format, and a target variable target is created to indicate a win (1) or loss (0). Categorical variables venue and opponent are encoded, the hour is extracted from the time column, and the day of the week is derived from the date column.

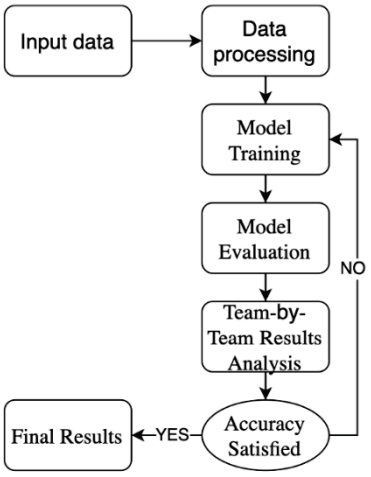

Figure 2 Block Diagram of the Proposed Prediction Model

Features for training the model are defined, and columns for calculating rolling averages are identified. The *rolling_averages* function sorts the data by date, computes rolling averages for specified columns, adds new columns with rolling averages, and removes rows without sufficient data. This function is applied to each team using the groupby and apply methods. The additional team index is then removed, and row indices in the *matches_rolling* data frame are reset.

Input dataset is split into training (1160 games before January 1, 2022) and testing sets (1132 games after that date). Various models for training are defined, including logistic regression, Naive Bayes, decision trees, k-NN, random forest, AdaBoost with logistic regression as the base model, and multilayer perceptron [19].

As already mentioned, several ML algorithms were used to predict match outcomes based on features outlined in Tab. 2. The target variable represents match results, where *W* indicated a win and *L* indicated a loss. Two approaches were used, Model-Level Accuracy and Team-Level Accuracy.

1) Model-Level Accuracy - The overall accuracy of each model was calculated as the ratio of correct predictions to total predictions across all teams.

$$Accuracy = \frac{N_{correct\_predictions}}{N_{total\_predictions}}. \quad (1)$$

2) Team-Level Accuracy - The accuracy for each team was calculated as the ratio of correct predictions to the total predictions for that specific team.

$$Accuracy = \frac{N_{gms\_predicted} - N_{mssd\_predictions}}{N_{gms\_predicted}}. \quad (2)$$

The difference between goals scored (*gf*) and goals conceded (*ga*), although used for rolling averages and form indicators, did not directly affect the accuracy calculation, which was based on predicting match outcomes (win/loss). Results, including team-level accuracy for each model, are summarized in Tab. 4. The average accuracy of each model is also presented, showing the overall performance across teams.

A table to store results for each model and each team is created, with column names as model names and indices as unique teams from the test dataset. Each model is trained using training data, and outcomes for each team in the test dataset are predicted. Accuracy is calculated as the ratio of correctly predicted outcomes to the total number of games, and accuracy results are stored in the *team_results* table. Finally, average accuracy per team for each model is stored in the results table. The accuracy results of each model for each team and the average accuracy per model are printed using the print function.

Through this methodology, the application of ML techniques to real-world data is demonstrated. This approach enables the development of efficient models and a deeper understanding of the dynamics of sports competitions. The learned methods can be applied to various analytical challenges in many industries, opening possibilities for innovation and improving data-driven decision-making.

5 RESULTS

Fig. 3 highlight the complexity of predicting sports event outcomes, where factors such as team form, tactical decisions, and random events during the game can significantly impact prediction accuracy. The highest average accuracy per model is achieved by the Random Forest model at 77.01 %, indicating a satisfactory level of efficiency. However, the *k-NN* model shows the lowest average accuracy at 63.79 %, leaving room for improvement, especially in cases of teams with lower prediction accuracy.

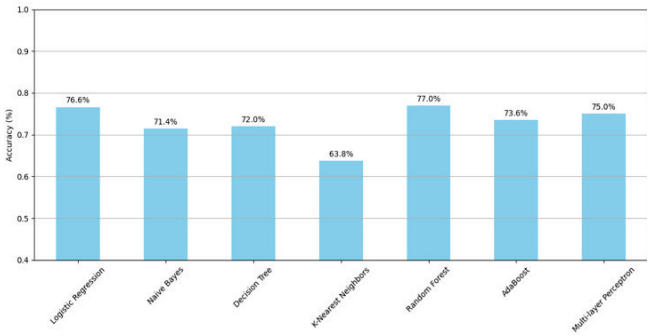

Figure 3 Algorithm Accuracy Comparison

Despite the best model achieving an accuracy of 77.01 %, a significant number of incorrect predictions were observed. This indicates the need for further analyses and potential improvements in the models to increase precision and reduce prediction errors. Improvements could include feature revision, parameter optimization, and testing alternative modeling approaches to achieve more reliable and accurate results.

Tab. 4 clearly illustrates the differences in accuracy of various ML models. The accuracy percentages above the bars allow for easy comparison of each model's effectiveness. An observed accuracy ranges from 47.06 % to 100.00 % suggests that there is room for improvement. Through further analysis of the results, it can be identified which models have the most potential for improvement and which optimization methods are most suitable for increasing overall accuracy and reducing prediction error.

Table 4 Detailed Accuracy of Algorithms

Team	*LR* (%)	*NB* (%)	*DT* (%)	*k-NN* (%)	*RF* (%)	*AB* (%)	*MLP* (%)
Arsenal	82.46	71.93	66.67	47.37	82.46	75.44	77.19
Aston Villa	70.69	63.79	72.41	53.45	72.41	68.97	70.69
Bournemouth	82.86	80.00	77.14	80.00	82.86	82.86	82.86
Brentford	67.24	63.79	68.97	68.97	62.07	68.97	63.79
Brighton	62.07	58.62	63.79	53.45	67.24	60.34	65.52
Burnley	86.36	72.73	68.18	68.18	86.36	90.91	86.36
Chelsea	64.29	50.00	67.86	66.07	64.29	66.07	64.29
Crystal P.	84.21	73.68	70.18	66.67	73.68	78.95	82.46
Everton	77.97	72.88	69.49	61.02	71.19	66.10	72.88
Fulham	65.79	63.16	73.68	68.42	78.95	65.79	68.42
Leeds United	75.86	81.03	81.03	67.24	75.86	81.03	68.97
Leicester City	74.14	72.41	72.41	68.97	72.41	70.69	74.14
Liverpool	77.19	70.18	66.67	59.65	82.46	68.42	75.44
Man. City	79.41	76.47	69.12	47.06	76.47	79.41	75.00
Man. United	74.14	63.79	70.69	51.72	77.59	67.24	74.14
Newcastle	73.68	59.65	61.40	54.39	71.93	63.16	73.68
Norwich City	89.47	84.21	78.95	73.68	84.21	84.21	84.21
Nottingham	74.29	77.14	80.00	71.43	85.71	74.29	71.43
Southampton	80.70	80.70	80.70	71.93	82.46	77.19	85.96
Tottenham	81.36	71.19	62.71	59.32	81.36	69.49	77.97
Watford	95.24	90.48	95.24	76.19	100.00	95.24	85.71
West Ham	77.19	77.19	70.18	61.40	71.93	71.93	77.19
Wolves	65.52	67.24	68.97	70.69	67.24	65.52	67.24

Additionally, the analysis of results from testing the ML model designed to predict EPL soccer outcomes is presented. Tab. 4 provides a detailed performance breakdown for each team, focusing on prediction accuracy. This visualization offers clear insights into the model's efficiency and identifies trends and anomalies in team performance. These results represent the average performance of the model for each club.

Analyzing the results presented in the table, significant variations in prediction accuracy between different teams are observed. Watford stands out with exceptionally high accuracy of 100.00 % for the Random Forest model, suggesting that the models successfully predicted their outcomes, possibly indicating certain predictability in their tactics or styles of play. Conversely, the lowest accuracy of 47.06 % was recorded for Manchester City using the *k-NN* model, indicating challenges in the models adapting to the variability characterizing their games.

6 CONCLUSION

The task of predicting outcomes across various domains has become increasingly important and captivating. From

making economic decisions and diagnosing medical conditions to formulating sports strategies, the ability to anticipate future events offers immense value. Achieving accurate predictions requires a deep understanding and proper analysis of relevant data.

In this study, the focus was on predicting soccer game outcomes in the EPL using various ML algorithms. The dataset was meticulously gathered from the *fbref.com* platform, encompassing a range of features and target variables relevant to soccer game outcomes. Features were chosen for their potential to influence game results, and the target variables were used to measure the models' accuracy.

The methodology section detailed the process of developing ML models using Python and the *Scikit-learn* library. The models tested included logistic regression, Naive Bayes, decision trees, *k-NN*, random forest, AdaBoost, and multilayer perceptron (*MLP*). Each model was evaluated based on its prediction accuracy for different teams in the league.

The results showed significant variations in prediction accuracy between different teams and models. The accuracy results ranged from 47.06 % (*k-NN* for Manchester City) to 100.00 % (Random Forest for Watford). Random Forest achieved the highest average accuracy at 77.01 %, demonstrating its effectiveness, while *k-NN* showed the most room for improvement with an average accuracy of 63.79 %.

These findings underscore the importance of careful algorithm selection and data processing methods to achieve optimal outcomes in sports prediction. Future research should aim to improve model accuracy by refining feature selection, optimizing hyperparameters, and exploring alternative modeling techniques. Additionally, integrating diverse types of data, such as real-time game statistics, player biometrics, and contextual factors like weather or crowd influence, could enhance prediction capabilities. Advanced processing techniques, including deep learning or hybrid models, may further boost performance and uncover hidden patterns in complex datasets. Such developments could lead to more reliable and insightful tools for decision-making in sports analytics [36].

In conclusion, this study demonstrates the potential of ML models to predict soccer game outcomes, highlighting the need for ongoing refinement and development. Accurate outcome prediction is vital across various fields, and the insights gained from this research can be applied to improve decision-making processes in sports and beyond. As data availability and computational methods continue to advance, the ability to predict outcomes with higher precision will become increasingly feasible, providing valuable benefits across multiple domains.

7 REFERENCES

[1] Spinner, J. (2018). *Web scraping with Python: Collecting more data from the modern web*. Sebastopol: O'Reilly Media.

[2] McCabe, A., & Travathan, J. (2008). Artificial Intelligence in Sports Prediction. In *Fifth International Conference on Information Technology: New Generations*. https://doi.org/10.1109/ITNG.2008.203

[3] Buursma, D. (2011). Predicting sports events from past results: Towards effective betting on football matches. In *14th Twente Student Conference on IT*.

[4] Hucaljuk, J., & Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *Proceedings of the 34th International Convention MIPRO*.

[5] Owramipur, F., Eskandarian, P., & Mozneb Sadat, F. (2013). Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. *International Journal of Computer Theory and Engineering, 5*(5), 812-815. https://doi.org/10.7763/IJCTE.2013.V5.802

[6] Igiri, C., & Nwachukwu, E. (2014). An Improved Prediction System for Football a Match Result. *IOSR Journal of Engineering, 4*(12), 12-20. https://doi.org/10.9790/3021-04124012020

[7] Tax, N., & Joustra, Y. (2015). Predicting the Dutch Football Competition Using Public Data: A Machine Learning Approach. *Transactions on Knowledge and Data Engineering, 10*(10), 1-13.

[8] Prasetio, D., & Harlili, D. (2016). Predicting football match results with logistic regression. In *International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*. https://doi.org/10.1109/ICAICTA.2016.7803111

[9] Zaveri, N., Shah, U., Tiwari, S., Shinde, P., & Teli, L. K. (2018). Prediction of Football Match Score and Decision Making Process. *International Journal on Recent and Innovation Trends in Computing and Communication, 6*(2), 162-165.

[10] Knoll, J., & Stübinger, J. (2020). Machine-learning-based statistical arbitrage football betting. *KI - Künstliche Intelligenz, 34*(1), 69-80. https://doi.org/10.1007/s13218-019-00610-4

[11] Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. *Applied Sciences, 10*(1), 46. https://doi.org/10.3390/app10010046

[12] Ievoli, R., Palazzo, L., & Ragozini, G. (2021). On the use of passing network indicators to predict football outcomes. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2021.106997

[13] Azeman, A. A., Mustapha, A., Razali, N., Nanthaamomphong, A., & Wahab, M. H. A. (2021). Prediction of football matches results: Decision forest against neural networks. In *The 8th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. https://doi.org/10.1109/ECTI-CON51831.2021.9454789

[14] Razali, N., Mustapha, A., Mustapha, N., & Clemente, F. M. (2021). A Bayesian approach for major European football league match prediction. *International Journal of Nonlinear Analysis and Applications, 12*(Special Issue), 971-980.

[15] Rodrigues, F., & Pinto, Â. (2022). Prediction of football match results with machine learning. *Procedia Computer Science, 204*, 463-470. https://doi.org/10.1016/j.procs.2022.08.057

[16] Ren, Y., & Susnjak, T. (2022). Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index. In *School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand*.

[17] Demartino, R. M., Egidi, L., & Torelli, N. (2024). Alternative ranking measures to predict international football results.

[18] Horvat, T., Havaš, L. & Srpak, D. (2020). The Impact of Selecting a Validation Method in Machine Learning on Predicting Basketball Game Outcomes. *Symmetry, 12*(3), 431-446. https://doi.org/10.3390/sym12030431

[19] Kelleher, J. D., & Tierney, B. (2021). *Znanost o podacima*. Zagreb: Mate d.o.o. (in Croatian)

[20] Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley interdisciplinary reviews, 1*, e1380. https://doi.org/10.1002/widm.1380

[21] Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F. M., & Baca, A. (2023). Machine learning application in soccer: a systematic review. *Biology of Sport, 40*(1), 249-263. https://doi.org/10.5114/biolsport.2023.112970

[22] Naik, B. T., Hashmi, M. F., & Bokde, N. D. (2022). A Comprehensive Review of Computer Vision in Sports: Open Issues, Future Trends and Research Directions. *Applied Sciences, 12*(4429). https://doi.org/10.3390/app12094429

[23] Horvat, T., Job, J., Logožar, R., & Livada, Č. (2023). A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games. *Symmetry, 15*(4), 798. https://doi.org/10.3390/sym15040798

[24] Horvat, T., Job, J., & Medved, V. (2018). Prediction of Euroleague Games based on Supervised Classification Algorithm k-Nearest Neighbours. In Pezarat-Correia, P., Vilas-Boas, J., Rivera, O. et al. (Eds.), *Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support* (pp. 203-207). Sevilla: SCITEPRESS. https://doi.org/10.5220/0006893502030207

[25] Horvat, T., & Job, J. (2019). Importance of the training dataset length in basketball game outcome prediction by using naive classification machine learning methods. *Elektrotehniški vestnik, 86*(4), 197-202.

[26] Bilek, G., & Ulas, E. (2019). Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators. *Journal of Performance Analysis in Sport, 19*, 930-941. https://doi.org/10.1080/24748668.2019.1684773

[27] Hsu, Y.-C. (2021). Using Convolutional Neural Network and Candlestick Representation to Predict Sports Match Outcomes. *Applied Sciences, 11*(6594). https://doi.org/10.3390/app11146594

[28] Zhang, Q., Zhang, X., Hu, H., Li, C., Lin, Y., & Ma, R. (2022). Sports Match Prediction Model for Training and Exercise Using Attention-Based LSTM Network. *Digital Communications and Networks, 8*, 508-515. https://doi.org/10.1016/j.dcan.2021.08.008

[29] Zhou, C., Calvo, A. L., Robertson, S., & Gómez, M.-Á. (2021). Long-Term Influence of Technical, Physical Performance Indicators and Situational Variables on Match Outcome in Male Professional Chinese Soccer. *Journal of Sports Sciences, 39*, 598-608. https://doi.org/10.1080/02640414.2020.1836793

[30] Akyildiz, Z., Nobari, H., González-Fernández, F. T., Praça, G. M., Sarmento, H., Guler, A. H., Saka, E. K., Clemente, F. M., & Figueiredo, A.J. (2022). Variations in the Physical Demands and Technical Performance of Professional Soccer Teams over Three Consecutive Seasons. *Scientific Reports, 12*, 2412. https://doi.org/10.1038/s41598-022-06365-7

[31] McDaniel, J. (2020). *Web Scraping with Python*. Birmingham: Packt Publishing Ltd.

[32] Alpaydin, E. (2021). *Strojno učenje*. Zagreb: Mate d.o.o. (in Croatian)

[33] Mohalder, R. N., Hossain, M. A., & Hossain, N. (2024). Classifying the Supervised Machine Learning and Comparing the Performances of the Algorithms. *International Journal of Advanced Research, 12*(1), 422-438. https://doi.org/10.21474/IJAR01/18138

[34] PMF. (2022). 6. Simpozij studenata doktorskih studija PMF-a. Zagreb. (in Croatian)

[35] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. https://doi.org/10.1007/978-0-387-84858-7

[36] Herbinet, C. (2018). Predicting Football Results Using Machine Learning Techniques. London.

Authors' contacts:

Mateo Vujčić, bacc. ing. comp.
University North,
Trg dr. Žarka Dolinara 1, 48000 Koprivnica, Croatia
Armed Forces of Croatia,
Zagrebačka ul. 2, 43000 Bjelovar, Croatia
mateo7.vujcic@gmail.com

Tomislav Horvat, PhD, Assistant Professor
(Corresponding author)
University North,
104. brigade 3, 42000 Varaždin, Croatia
tomislav.horvat@unin.hr

Dejan Barešić, PhD, Assistant Professor
Croatian Military Academy "Dr. Franjo Tuđman",
Ilica 256b, 10000 Zagreb, Croatia
dejan.baresic@morh.hr

Dražen Crčić, mag. ing. el. eng.
University North,
104. brigade 3, 42000 Varaždin, Croatia
drazen.crcic@unin.hr