Predictive Modeling of Football Results in the WWIN League of Bosnia and Herzegovina

Ervin Vladić
International Burch University
Sarajevo, Bosnia and Herzegovina
ervin.vladic@stu.ibu.edu.ba

Dželila Mehanović
International Burch University
Sarajevo, Bosnia and Herzegovina
dzelila.mehanovic@ibu.edu.ba

Elma Avdić
International Burch University
Sarajevo, Bosnia and Herzegovina
elma.avdic@ibu.edu.ba

Abstract

Predictive modeling in football has emerged as a valuable tool for enhancing decision-making in sports management. This study applies machine learning techniques to predict football match outcomes in the WWIN League of Bosnia and Herzegovina. The aim is to evaluate the effectiveness of various models, including Support Vector Machines (SVM), Logistic Regression, Random Forest, Gradient Boosting, and k-Nearest Neighbors (kNN), in accurately predicting match results based on key features such as shots on target, possession percentage, and home/away status. By (1) gathering and analyzing match data from three seasons, (2) comparing the performance of machine learning models, and (3) drawing conclusions on key performance factors, we demonstrate that SVM achieves the highest accuracy at 83%, outperforming other models. These insights contribute to football management, allowing for data-driven strategic planning and performance optimization. Future research will integrate additional factors such as player injuries and weather conditions to improve the predictive models further.

Keywords

Football match prediction, machine learning, WWIN league, Support Vector Machines, Random Forest

1 Introduction

Accurate predictions of match outcomes can inform a wide range of decisions, from tactical adjustments to player acquisitions, and can improve engagement for fans and stakeholders. While predictive modeling has been extensively applied to top-tier football leagues like the English Premier League, there is limited research on regional leagues such as the WWIN League of Bosnia and Herzegovina. The specificity of the country that is Bosnia and Herzegovina and the WWIN League, which has not been researched in the sphere of sports research, provides context for this step.

The WWIN League of Bosnia and Herzegovina was established in the year 2000 and the same year the WWIN was formed by the merging of three leagues, it became a league covering the entire territory of Bosnia and Herzegovina. Originally, the league consisted of 16 clubs, and, from the 2016-2017 season, the league contains 12 clubs which makes the level of the league higher [25]. The winner is the team that has the most points by the completion of thirty-three rounds; this position will grant a team a place in the UEFA Champions League qualifications [10]; the remaining two teams and the winner of the cup will compete for a place in the UEFA Conference League. Since the founding of the WWIN League of Bosnia and Herzegovina, team with the highest number of titles was HŠK Zrinjski from Mostar who emerged as the winner eight times, followed by Sarajevo which won four times, Zeljeznicar and Borac both won three times, Siroki Brijeg won two times and Leotar and Modrica both won once [12]. Depending on which entity association they belong to, the teams that occupy the last two places in the league at the end of the season are relegated to the league below, with two teams from the First League of the Federation of BiH and the First League of the RS being promoted in their stead. To elevate football in our country to the highest level, we must support in-depth analyses of matches and the factors influencing their outcomes. This will enable coaches to fine-tune strategies for future games, help commentators provide more insightful commentary, and allow fans to develop a deeper understanding and get more pleasure from the match.

The study aims to evaluate the performance of various ML models, including Support Vector Machines (SVM), Logistic Regression, Random Forest, Gradient Boosting, and k-Nearest Neighbors (kNN), in predicting match results. By examining key features such as shots on target, possession percentage, and home/away status, we conduct an analysis based on match data from three seasons of the WWIN League, encompassing 400 matches and key performance metrics.

The remainder of the paper is structured as follows: Section II provides an overview of related work in football ML-based prediction. Section III describes the methodology, including the dataset and models used. Section IV presents the results and analysis of models performance, with a discussion on the practical implications of the findings for football management. Finally, Section V concludes the paper and outlines directions for future research.

2 Literature Review

The prediction of the results of football matches has been recently studied extensively because of its relation to betting and decision-making in sports. Studies examining the employment of ML methods are primarily focused on large European leagues, where extensive and highly detailed data is available. The application of these techniques to regional football leagues, such as the WWIN League of B&H, remains underexplored.

Rodrigues and Pinto [15] used a variety of ML methods, including Naive Bayes, K-nearest neighbors, Random Forest, and SVM, to predict the match outcomes based on previous match data and player attributes. Their studies revealed excellent results in terms of soccer betting profit margins, with the Random Forest approach obtaining a success rate of 65.26% and a profit margin of 26.74%. Rahman [13] dedicated his work to employing deep learning frameworks especially Deep Neural Networks (DNNs) for football match outcome prediction, particularly during FIFA World Cup 2018. The study classified match outcomes

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.1642

with 63.3% accuracy with DNN architectures with LSTM or GRU cells. Baboota and Kaur [3] used machine learning approaches to predict English Premier League match results. The models compared included Support Vector Machines, Random Forest; and Gradient Boosting. From their study , they ascertained that Gradient Boosting outperformed other models in accuracy and overall predictiveness. Authors in [16] used machine learning techniques, notably SVM and Random Forest Classifier, to predict English Premier League (EPL) football match results. They got 54.3% accuracy with SVM and 49.8% with Random Forest after evaluating data from 2013/2014 to 2018/2019 seasons. Another study [8] employed a few machine learning algorithms to predict matches of the English Premier League season 2017-2018. Models including Linear Regression, SVM, Logistic Regression, Random Forest, and Multinomial Naïve Bayes classier show that the K-nearest neighbors give the best accurate predictions.

In summary, while existing studies have demonstrated the effectiveness of machine learning in football matches prediction, there remains a gap in the application of these techniques to regional leagues like the WWIN League, due to the availability and quality of data. The characteristics of these leagues, such as smaller datasets and potentially different factors influencing match outcomes, require a tailored approach. In lesser-known football leagues models might perform differently due to variations in competitive structures and gameplay strategies, as well. The study of Munđar and Šimić [11] in which they developed a simulation model using the Poisson distribution to predict the seasonal rankings of teams in the Croatian First Football League, highlighted the predictive power of statistical models and demonstrated the significance of home advantage in determining match outcomes, which is also an important factor in the WWIN League.

3 Materials and Methods

In this section, we describe the study conducted, detailing the data collection and feature selection processes, the machine learning models applied, the evaluation metrics used to assess model performance, and the approach taken to analyze key features influencing match outcomes. As a result of providing numerous procedures that are declared in this section, we represent the graphical illustration of our methodology. The steps involved in predicting the outcomes of the WWIN League of Bosnia and Herzegovina, including data collection, preprocessing, model development, and algorithm evaluation.

3.1 Dataset

The authors created the dataset for this study by consolidating information from rezultati.com [14], 1XBET [1], and Sofascore [24]. The unique dataset represents the seasons 2021/2022, 2022/2023, and 2023/2024 of WWIN League of Bosnia and Herzegovina. The platforms provide a wide range of football match data so it is easy to find important information for examination. The dataset includes key match facts as date, day of the week, time, home team, away team, final as well as half-time goals scored in the game, referee details, shots taken at goal as well as corner kicks resulting from these attempts on target, bookings made during play by both teams and other relevant performance indicators.

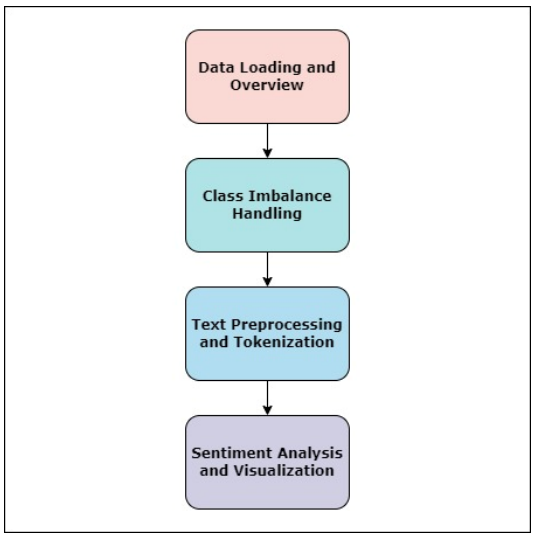

Figure 1: Workflow diagram

Table 1: Class Distribution

Match Type	Count
Home Win	301
Away Win	142
Draw	151

The table sums up a type of match result in terms of its frequency in the dataset.

In the recorded 594 matches, 301 ended in home team victories, 142 in away team victories and 151 were tied. The following pie chart describes the percentage distribution of the match outcome. Curiously, home wins are in the majority, comprising 50.7% of all

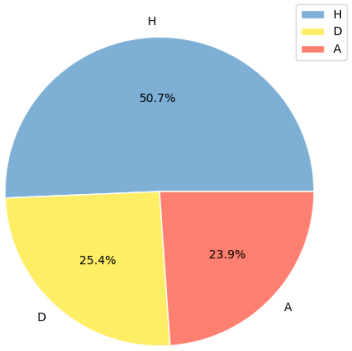

Figure 2: Class distribution of the dataset

matches. However, away victories contribute to approximately 23.9% of all recorded match results, while 25.4% contribute to draw results. The Fig.2 depicts the frequency of each of the match outcomes.

3.2 Machine Learning Prediction

In football, the concept of machine learning prediction entails developing models to forecast match outcomes based on the teams'

and players' histories and other attributes [5]. These models employ such methods as regression analysis, classification, and neural nets to determine the results given the data fed as the input.

3.2.1 Models initialisation, preprocessing, training and testing. While implementing Logistic Regression, we have set the max_iter =1000 and random_state = 42. Again, with the same classifier, the kernel argument was assigned a linear value while the random_state was set to 42 to keep the results predictable. Gaussian Naive Bayes was employed with no modification of its settings because of the model's simplistic nature. For Random Forest, we used the default parameters since the algorithm is capable of changing the setting on its own based on the complexity of given data. We initiated the Gradient Boosting with the default parameters so that the gradients could easily learn and an ensemble could be formed. Last but not least, we left all the parameters of k-Nearest Neighbors (kNN) for default value because the algorithm can find the optimal number of neighbors appropriate for the distribution of the data.

Following that, we proceed with the process of dividing this gathered data into two sets: the training and the testing ones. We split the data into training, where 70% of the data was allocated and the testing data where 30% was allocated.

Subsequently, the phase of model preprocessing is created for which it is essential to filter data effectively to ensure proper model training. In the case of feature transformation, we used scikit-learn's ColumnTransformer [17] to empower the numeric features normalization via the StandardScaler [23] while transforming the categorical variables into the binary format by the use of the OneHotEncoder [18]. This technique pays a lot of attention to ensuring that feature types are standard as well as harmonious. This method ensures consistency by creating a pipeline where preprocessing processes and the model are joined in the same line of work. This means that there is always uniformity in the training and the testing of the model, hence a manageable variability. Assuming the pipeline has been defined and is ready to proceed, we proceed to the next step of model training.

3.2.2 Models in Detail. In this study, many supervised learning classifier techniques that have proven valuable in the sports area for predictive purposes are employed. Logistic Regression is a statistical technique that predicts the probability of a binary classification, using a sigmoid function to map outputs to a [0,1] probability space. Coefficients indicate the strength and direction of relationships between variables, with positive values increasing the likelihood of an event and negative values decreasing it [9].

Random Forest extends the bagging method by generating multiple decision trees using randomly selected data samples. Each tree operates independently, and the final prediction is the average result across all trees, reducing overfitting and improving accuracy in classification tasks [4].

SVM aims to find the best hyperplane to separate data points by class, maximizing the margin between them. It handles nonlinear boundaries by transforming the input data into a higher-dimensional space [2].

Naïve Bayes, based on Bayes' theorem, assumes feature independence, making it fast and easy to implement, especially in applications like spam detection and text classification. Despite the simplicity of this assumption, it performs well in practice [26].

Gradient Boosting combines multiple weak learners (typically decision trees) to create a stronger predictive model, improving accuracy by focusing on correcting errors from previous models [6].

k-Nearest Neighbors (kNN) is an instance-based learning method that classifies data by identifying the majority label among the k closest points. Though simple, it can be computationally expensive as it requires storing all training data and performing real-time comparisons [7].

3.2.3 Evaluation Metrics. Last but not the least, the trained models are assessed by metrics such as accuracy of the models [19], precision of the models [21], the recall of the models [22], and F1-score value of the models [20]. This evaluation enables one to analyze how well each of the models is likely to perform in terms of match outcome prediction.

4 Results and Discussion

In this study, we employed six different classifiers to predict football match outcomes and conducted a comparative analysis of their performance. The effectiveness of each classifier was evaluated based on its accuracy, providing a clear comparison of their predictive capabilities across the dataset.

4.1 Model Performance

Among the classifiers employed, SVM predicted the most accurate results at 83% This model performed almost well, with balanced precision and recall across all three classes (A, D, and H), showing that it can predict match outcomes. In comparison, Random Forest achieved a lower accuracy of 65%, with especially evident deficits in precision and recall for class 'D'. Logistic Regression performed worse than Support Vector Machines, with accuracy of 77%. Despite its simplicity and computational efficiency, Gaussian Naive Bayes had the lowest accuracy of any classifier tested, at 39%. This model struggled to predict class 'D', with low accuracy and recall scores. Random Forest, an established ensemble learning approach, performed not so good, with an accuracy of 54%. This model has generally balanced accuracy and recall across all classes, making it an acceptable alternative for predicting match results. Gradient Boosting, another ensemble learning technique, has a little higher accuracy than Random Forest at 64%. While Gradient Boosting is recognized for its ability to manage complicated connections, it produced poorer recall ratings, especially for class 'D'. Lastly, k-Nearest Neighbors (kNN) resulted in 51% accuracy, showing that the classifier was relatively poor, they had relatively fair precision and recall with all the classes.

For making the match predictions, we employed the following classification models – Logistic Regression, Support Vector Machine, Gaussian Naive Bayes, Random Forest, Gradient Boost and k-Nearest Neighbors. We obtained the results varying from 39% to 83%, in which Support Vector Machines were the most effective. Our findings are partially consistent with prior research because classifiers like Support Vector Machines, Logistic Regression, and Random Forest have manifested robustness in predicting the match outcome across datasets. Nevertheless, the results are not in conformity with some emerging works, particularly concerning the efficacy of Gaussian Naive Bayes, which performed poorly in our study in contrast to other research results. It should be noted that results may vary significantly between different studies depending on the quality, the quantity, and the nature of the data that had been used for creating the models of Gaussian Naive Bayes.

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	77%	75%	74%	74%
SVM	83%	86%	83%	84%
Gaussian NB	39%	47%	42%	36%
Random Forest	54%	43%	46%	43%
Gradient Boosting	64%	64%	59%	60%
kNN	51%	49%	49%	49%

Table 2: Model Performances

The Table 2 shows how accurate various machine learning models are in predicting WWIN League of Bosnia and Herzegovina match outcomes.

4.1.1 Key factors influencing match outcomes. While this study does not perform formal feature analysis, the observed performance trends allow us to draw conclusions about the key factors influencing match outcomes. In line with prior research, home advantage emerged as a critical factor, with teams winning at home in over 50% of cases (Table 1) which reinforces the psychological and tactical advantages that come with playing on familiar ground.

Offensive metrics, particularly shots on target, also revealed themselves as strong predictors of success. Teams that generated more attempts on goal were significantly more likely to win, reinforcing the widely accepted view that aggressive, forward-driven play translates directly into better results. This trend mirrors observations from other football leagues, where offensive intensity is often directly correlated with victory.

4.1.2 Limitations and future work. Despite the promising results, this study has several limitations. First, the dataset used does not account for external factors such as player injuries, weather conditions, or team morale, all of which can influence match outcomes. Future research should incorporate these factors to improve the accuracy of predictions. Second, while SVM performed well in this context, more advanced models such as deep learning could potentially offer even better predictive performance, particularly when dealing with larger datasets.

Future work will explore the integration of additional domain-specific features, such as player statistics, team form, and environmental conditions, to further refine the predictive models. We will also experiment with more complex algorithms, such as neural networks, to capture the intricate relationships between features that may be missed by traditional machine learning models.

5 Conclusion

This study demonstrates that machine learning, particularly SVM, effectively predicts football match outcomes in the WWIN League of Bosnia and Herzegovina. Support Vector Machine has been found to be the highest accurate classifier with 83% of accuracy rate on match result prediction. SVM has moderate accuracy and recall with all three outcome classes: Home Win, Away Win, and Draw, indicating football prediction applicability. However, it has also revealed that other classifiers' performances are varying with Logistic Regression producing 77% of accuracy and Gaussian Naïve Bayes a poor 39% accuracy. Both Random Forest and Gradient Boosting, which are ensemble learning algorithms, have similar levels of accuracy; 54% and 64% respectively. While further refinement of the models is needed, the current findings establish a strong foundation for data-driven decision-making in football management. Future work should incorporate additional factors such as player injuries and weather conditions to enhance predictive accuracy.

References

[1] 1XBET. 2007–2024. 1xbet. Retrieved May 26, 2024, from https://1xlite-579542.top/en?tag=s_245231m_5435c_. (2007–2024).

[2] Mariette Awad and Rahul Khanna. 2015. Support vector machines for classification. In *Efficient Learning Machines*. Rahul Khanna, editor. Apress, 39–66. DOI: 10.1007/978-1-4302-5990-9_3.

[3] Rahul Baboota and Harleen Kaur. 2019. Predictive analysis and modeling football results using a machine learning approach for the english premier league. *International Journal of Forecasting*, 35, 2, 741–755. DOI: 10.1016/j.ijforecast.2018.01.003.

[4] Leo Breiman. 2001. Random forests. *Machine Learning*, 45, 5–32. DOI: 10.1023/A:1010933404324.

[5] Rory P. Bunker and Fadi Thabtah. 2019. A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15, 1, 27–33.

[6] Stefanos Fafalios, Pavlos Charonyktakis, and Ioannis Tsamardinos. 2020. *Gradient Boosting Trees*. Gnosis Data Analysis PC, (Apr. 2020).

[7] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003*. Springer Berlin Heidelberg, Catania, Sicily, Italy, (Nov. 2003), 986–996.

[8] Ishan Jawade, Rushikesh Jadhav, Mark Joseph Vaz, and Vaishnavi Yamgekar. 2021. Predicting football match results using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, 8, 7, 177. https://www.irjet.net.

[9] Daniel Jurafsky and James H. Martin. 2023. *Logistic Regression*. Stanford University, 5. https://web.stanford.edu/~jurafsky/slp3/5.pdf.

[10] Haris Kruskic. 2019. Uefa champions league explained: how the tournament works. Bleacher Report. Retrieved from https://bleacherreport.com/articles/2819840-uefa-champions-league-explained-how-the-tournament-works. (2019).

[11] Dušan Munđar and Diana Šimić. 2016. Roatian first football league: teams' performance in the championship. *roatian Review of Economic, Business and Social Statistics 2*, 2, 1, 15–23. https://hrcak.srce.hr/file/245359.

[12] Prva Liga BiH. 2022. Osvajači trofeja. Retrieved from https://plbih.ba/osvajaci-trofeja/. (2022).

[13] Ashiqur Rahman. 2020. A deep learning framework for football match prediction. *SN Applied Sciences*, 2, 2, 165. DOI: 10.1007/s42452-019-1821-5.

[14] 2006–2024. Rezultati. Retrieved May 26, 2024, from https://www.rezultati.com/. (2006–2024).

[15] Fátima Rodrigues and Ângelo Pinto. 2022. Prediction of football match results with machine learning. *Procedia Computer Science*, 204, 463–470. DOI: 10.1016/j.procs.2022.08.057.

[16] Sayed Muhammad Yonus Saiedy, Muhammad Aslam HemmatQachmas, and Dr. Amanullah Faqiri. 2020. Predicting epl football matches results using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology*, 5, 3, 83–91. http://www.ijeast.com.

[17] scikit-learn. 2024. Columntransformer. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html. (2024).

[18] scikit-learn. 2024. Onehotencoder. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. (2024).

[19] scikit-learn. 2024. Sklearn.metrics.accuracy_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html. (2024).

[20] scikit-learn. 2024. Sklearn.metrics.f1_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. (2024).

[21] scikit-learn. 2024. Sklearn.metrics.precision_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html. (2024).

[22] scikit-learn. 2024. Sklearn.metrics.recall_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html. (2024).

[23] scikit-learn developers. 2024. Sklearn.preprocessing.standardscaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. (2024).

[24] Sofascore. 2024. Sofascore. Retrieved May 26, 2024, from https://www.sofascore.com/. (2024).

[25] SportMonks. 2022. Premier league api bosnia. Retrieved from https://www.sportmonks.com/football-api/premier-league-api-bosnia/. (2022).

[26] Geoffrey I. Webb. 2016. Naïve bayes. In *Encyclopedia of Machine Learning and Data Mining*. Claude Sammut and Geoffrey I. Webb, editors. (Jan. 2016), 1–2. DOI: 10.1007/978-1-4899-7502-7_581-1.