Predicting Bundesliga Match Outcomes Using Machine Learning Models

A Historical Analysis of German Football League Results
from 1993/94 to 2021/22 Season

Veldhuis, T.G.
School of Information & Technology
Dalarna University, DU
Borlänge, Sweden
h22twave@du.se

***Abstract*— This paper presents a comprehensive analysis of predicting Bundesliga match outcomes using data mining models. The study explores both competition-based and team-based analysis approaches and evaluates various machine learning models, including logistic regression, decision trees, random forests, support vector machine (SVM) classifier, and neural networks (MPL). The models are assessed using performance metrics such as accuracy and F1-score. The results highlight the effectiveness of the Logic Regression model for the team based prediction network model, which achieves the highest accuracy and F1-score among the tested models. The research also investigates the impact of dataset characteristics on model performance, revealing challenges in evaluating team-based analysis due to smaller dataset sizes. The findings contribute to the field of sports prediction and have implications for the betting industry and team performance analysis. Further research opportunities include refining team-based analysis methodologies, exploring advanced models, incorporating additional factors for improved prediction accuracy, and inclusion of multiple competitions for improved results.**

Keywords – machine learning; data mining; sport result prediction; neural networks

I. INTRODUCTION

The Bundesliga, Germany's premier professional football league, has long been renowned for its fierce competition, passionate fans, and global appeal [1]. With its rich history and highly competitive nature, the league presents an intriguing opportunity for exploring the application of machine learning to predict match outcomes. By analyzing a dataset containing the historical Bundesliga results the 1993/94 season to the 2021/22 season, this project aims to uncover patterns and trends that can provide valuable insights into the factor or factors influencing match outcomes [2] [3] [4]. Such insights hold significant importance for both research and business purposes, offering an edge to betting companies while contributing to our understanding of football dynamics.

Accurately predicting the outcome of football matches remains a challenging task, primarily due to the complex interplay of various factors such as team performance, player statistics, and home/away advantage [5]. By harnessing the power of machine learning, we can leverage historical data to uncover hidden patterns and establish predictive models. This project addresses the need for a systematic and data-driven approach [6] to Bundesliga match outcome prediction, offering potential benefits to the football industry, researchers, and betting companies alike. So the following research questions:

- Can machine learning algorithms accurately predict whether the points will stay at home or be taken by the away team?
- Can predictive models be tailored to account for team-specific characteristics and overall competition levels?
- How do different machine learning models compare in their predictive performance for Bundesliga match outcomes?

The aim of this project is to develop Machine Learning model that can predict Bundesliga match outcomes based on historical data. By answering the research questions above, we strive to uncover meaningful insights into the dynamics of the league and provide practical tools for predicting match results.

This project contributes to the field by conducting a comprehensive analysis of historical Bundesliga match results, identifying patterns and trends that influence outcomes. It evaluates and compares different machine learning algorithms to determine the most effective approach for predicting match results [7]. Additionally, the project investigates the impact of team-specific characteristics and overall competition levels on predictive performance. The practical applications of this research offer valuable insights for the football industry, researchers, and betting companies, enabling improved decision-making and potential profitability.

The proposed solution involves leveraging a large dataset of Bundesliga match results from the 1993/94 season to the 2021/22 season, sourced from a football betting website. Machine learning algorithms will be trained on this dataset to identify relationships. Various models, such as logistic regression, decision trees, random forests, SVM, and neural networks [7], will be implemented and evaluated to determine their effectiveness in predicting match outcomes.

The motivation behind this project lies in the potential to improve match outcome prediction in the Bundesliga [8] [9] [10]. By applying machine learning techniques to vast historical data, we can harness the power of data-driven insights to enhance decision-making and improve accuracy. The outcomes of this research hold value not only for betting companies seeking a competitive edge but also for the football industry as a whole [11].

II. LITERATURE REVIEW

The growing interest in applying machine learning techniques to sports prediction has led to extensive research in various sports, including football. The Bundesliga, Germany's top professional football league, presents an intriguing context for exploring the predictive capabilities of machine learning algorithms [12]. By analyzing a comprehensive dataset of historical Bundesliga match results, researchers aim to uncover patterns and trends that influence match outcomes, providing valuable insights into the factors at play [13].

Conventional methods for predicting football match outcomes often rely on subjective assessments, which can be biased and unreliable [7]. To overcome this limitation, researchers have turned to machine learning as a data-driven approach to enhance predictive accuracy [14]. By leveraging historical data, machine learning models can identify hidden patterns and establish predictive models that consider factors such as team performance, player statistics, and home/away advantage [15].

The literature reviewed demonstrates the potential benefits of machine learning in sports outcome prediction. For example, [16] applies an adaptive neural network model to predict sports match results with high accuracy and reliability, achieving an error rate of 0.001 compared to other models with error rates around 0.1. [14] introduces a self-created two-phase machine learning framework for predicting football match results, involving web scraping for data extraction and utilizing various ML algorithms, which shows good performance with an accuracy-based Return on Investment (ROI) of 0.58. Similarly, [15] constructs a predictive model for the English Premier League using machine learning, emphasizing feature engineering and exploratory data analysis, resulting in a gradient boosted classifier achieving an accuracy of 0.519.

In the context of improving sports outcome prediction, [4] integrates adaptive weighted features and machine learning techniques, showing promising results in predicting basketball game scores with RMSE scores consistently below 13.5 for all models used. These studies collectively highlight the need for a systematic and data-driven approach to sports outcome prediction, which aligns with the objectives of the current project.

Building on previous research, this project aims to develop a machine learning model that accurately predicts Bundesliga match outcomes based on historical data. By addressing specific research questions, such as the accuracy of machine learning algorithms in predicting home or away team points and tailoring predictive models to account for team-specific characteristics and competition levels, this project seeks to contribute to the field of sports outcome prediction [2] [4] [5] [6] [7] [8] [9] [10] [11] [12]. By utilizing a large dataset of Bundesliga match results and implementing various machine learning models, including logistic regression, decision trees, random forests, SVM, and neural networks, which have been used in other research studies [2] [3] [5] [8] [17], this project aims to provide practical tools for predicting match results and enhancing decision-making in the football industry.

Additionally, some studies have employed feature selection techniques [7] [18] [19], which involve identifying and selecting the most relevant and informative features from a given dataset. Feature selection helps improve model performance by eliminating irrelevant or redundant features, reducing overfitting, enhancing interpretability, and facilitating faster and more efficient modeling processes by reducing data dimensionality.

Overall, the literature review establishes a foundation for the current project, highlighting the effectiveness of machine learning in predicting sports outcomes and emphasizing the value of data-driven approaches in enhancing accuracy and reliability. Insights gained from previous studies conducted in various sports contexts inform the development of a machine learning model specifically tailored to predict Bundesliga match outcomes, offering potential benefits to the football industry, researchers, and betting companies alike [1] [2] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19].

III. METHOD DESCRIPTION

A. *The Dataset*

The dataset used in this project is sourced from the website ‘football-data.co.uk’, consisting of comprehensive data on Bundesliga match results from the 1993/94 season to the 2021/22 season.

The provided dataset consists of CSV-formatted football match data, including league division, match date and time, home and away team names, full-time and half-time goals, results (home win, draw, or away win), match statistics (attendance, shots, corners, fouls, yellow/red cards), and betting odds from different bookmakers. It covers multiple seasons and leagues, which are divided into multiple files, providing comprehensive information for analyzing match outcomes, team performances, and statistics.

During the initial analysis, it is crucial to highlight that the dataset did not exhibit any significant outlier issues. Surprisingly, the presence of outliers, particularly in the half-time scores where one team had a clear advantage, actually contributed to enhancing the accuracy of match outcome predictions. Though it has to be noted that some abbreviations and odds from specific bookmakers are no longer in use, as indicated by the data collected in earlier seasons and for this reason were removed from the dataset.

To clean and prepare the dataset, the following steps were taken. Irrelevant and redundant/highly correlated features, such as bookmakers' data, were removed. Features with less than 70% of rows filled were also eliminated, to prevent the interpretation of unusable match statistics. Full-time home and away goals were excluded to prevent revealing the result, because these would else be used by the prediction algorithm to predict the scores 100% of the time. Result categories (H, D, A) were converted to numerical values: 1 for home win and draw, and 0 for away win. Null values were filled with zeros, as they had minimal impact and most of the NaNs were already removed in the previous steps.

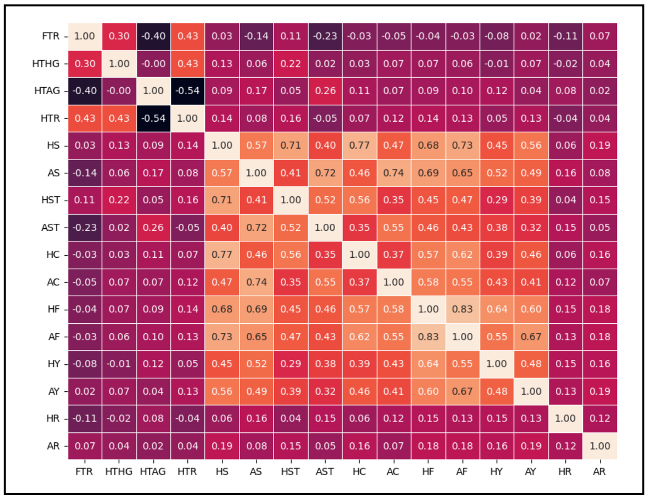

Figure 1. Correlation Matrix

In addition to providing a comprehensive description of the dataset, a correlation matrix was generated to perform descriptive analysis. This matrix offers valuable insights into the potential significance of certain features for conducting more focused team-based analysis.

In the end included the dataset variables like teams, half-time goals, result, shots, shots on target, woodwork hits, corners, fouls, free kicks conceded, offsides, yellow and red cards, and bookings points. This dataset still had a size of almost nine thousand rows and 19 features. For team based analysis the dataset was separated based on the home team and the away team was used as the dummy variable, which created around forty to fifty additional columns for each small dataset. The league-based analysis, was also implemented with additional dummy variables but this time for home and away teams which resulted in around about one hundred columns to be added to the final dataset.

B. *Data Mining Method*

This project aims to predict the outcomes of Bundesliga football matches using data mining models. The models employed include logistic regression, decision trees, random forests, support vector machine (SVM) classifier, and neural network (MLP classifier). For the first part of the modelling, the logistic regression model was executed, this was done to create a baseline model to assess the expected performance on the most basic model of all. Also is this model used to assess the impact of individual variables on importance and influence on the variance of the results. The models were selected based on their usage in the previous studies discussed in the literature review.

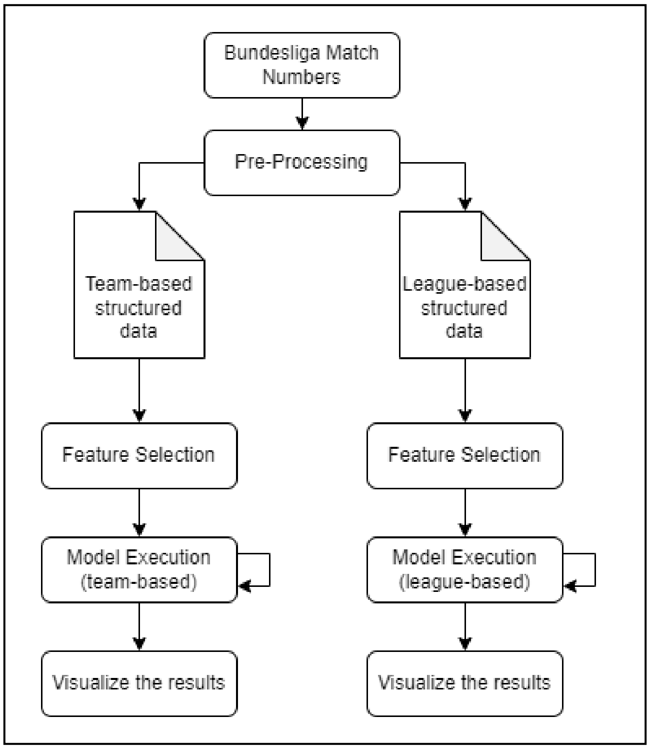

Figure 2. Diagram of process

All the models have been used with the standard parameters applied, this is done to get an overall overview of the expected results of the different models. Based on this baseline the models could then be selected for the different kind of predictions and optimized to get a final model that has the best accuracy within the current context.

Prior to executing the models, a backward feature selection technique was applied to identify and prioritize the crucial features from the dataset. This process aimed to enhance the modeling approach concerning both team-based and league-based models. This methodology was adopted based on previous studies (also mentioned in literature review), which demonstrated good results.

The dataset is divided into training and testing sets for model invocation and training, which is done by the K-fold cross validation. The training set is used to train the models on historical match data, including features like team statistics, and match-specific factors, along with the target variable (match outcome). The models learn from this data to establish relationships and patterns for predicting match outcomes. The testing set is then used to evaluate the trained models' performance on unseen data.

To enhance model robustness and mitigate overfitting, cross-validation techniques are employed. K-fold cross-validation, a common method, divides the dataset into k subsets (folds) of equal size. The models are trained and evaluated k times, using a different fold as the testing set each time and the remaining folds as the training set. This process provides a comprehensive assessment of the models' performance and helps identify any issues related to data variability or model instability.

In summary, this project utilizes data mining models like logistic regression, decision trees, random forests, SVM, and neural network to predict Bundesliga match outcomes. The models' parameters are the data is divided into training and testing sets. Cross-validation techniques, such as k-fold cross-validation, are used to assess and validate the models' predictive capabilities.

IV. Results and Analysis

The results of our machine learning models for predicting Bundesliga match outcomes are presented below. We evaluated the models using various metrics such as accuracy and F1-score.

For the competition-based analysis, where models were trained on the entire dataset without separating it per team, the following accuracy and F1-score were obtained:

TABLE I. Competition-based analysis

Model	Accuracy	F1-score
Logistic Regression	0.804	0.869
Decision Tree	0.718	0.791
Random Forest	0.802	0.868
SVM	0.777	0.836
Neural Network	0.729	0.805

For the team-based analysis, where models were created based on scores and match information of each home team separately, the following accuracy statistics were obtained:

TABLE II. Team-based analysis

Model	Accuracy	F1-score
Logistic Regression	0.773	0.818
Decision Tree	0.704	0.754
Random Forest	0.739	0.790
SVM	0.757	0.804
Neural Network	0.717	0.767

In the tables you can see that there is a lot of consistency in the metrics of the models for the competition-based analysis and the team-based analysis. Several factors can be used as reasoning for the observed consistency. Firstly, it is suggested that the similarity or overlap in the datasets used in both analyses plays a significant role. If the datasets share similar characteristics, patterns, and distributions, it is expected that the models' performances would align closely. Another factor contributing to the consistency is the utilization of shared features. When the features employed in both analyses are similar or derived from the same set of variables, the models have access to equivalent information and patterns within the data. The similarity in model architectures used in both analyses is also considered as a potential factor. If the models utilized in both scenarios have similar structures or belong to the same type, it can result in comparable outcomes.

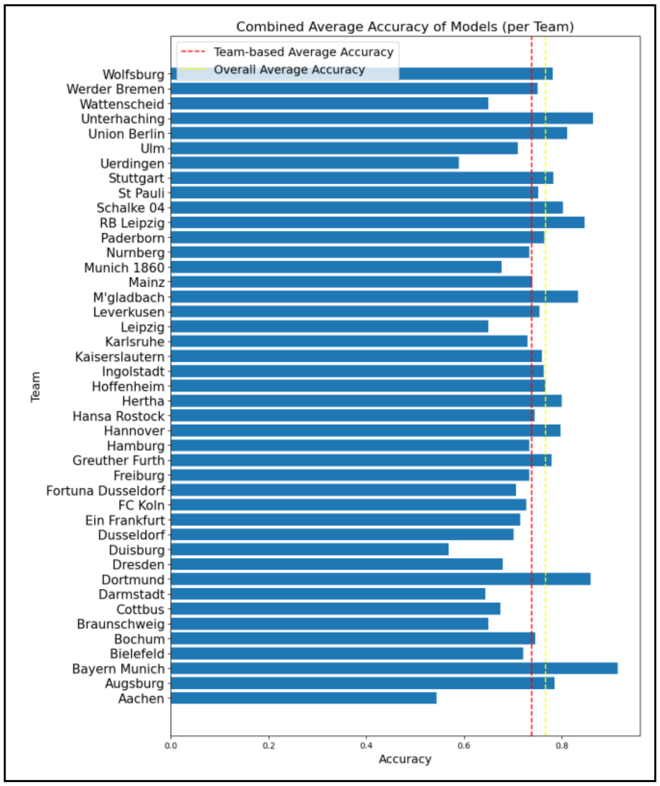

Figure 3. Accuracy of models.

Based on these results, the logistic regression model demonstrated the best performance, achieving the highest accuracy and F1-score among the tested models for the team-based analysis as well as for the competition-based analysis. The Random Forest and SVM models demonstrated satisfactory performance.

The Multilayer Perceptron (MLP) neural network and Decision Tree models had slightly lower performance compared to the other models. The Multilayer Perceptron (MLP) neural network is a powerful type of artificial neural network composed of interconnected layers of neurons. It is utilized for modeling complex non-linear relationships between input features and the target variable, enabling accurate predictions in a variety of domains. MLPs are trained through backpropagation, adjusting weights to minimize the difference between predicted and actual outputs, ultimately providing valuable insights through their ability to capture intricate patterns. Decision Trees, known for their simplicity, may struggle with intricate relationships and high-dimensional data. However, both models still provided valuable insights into the data and could be improved with further fine-tuning.

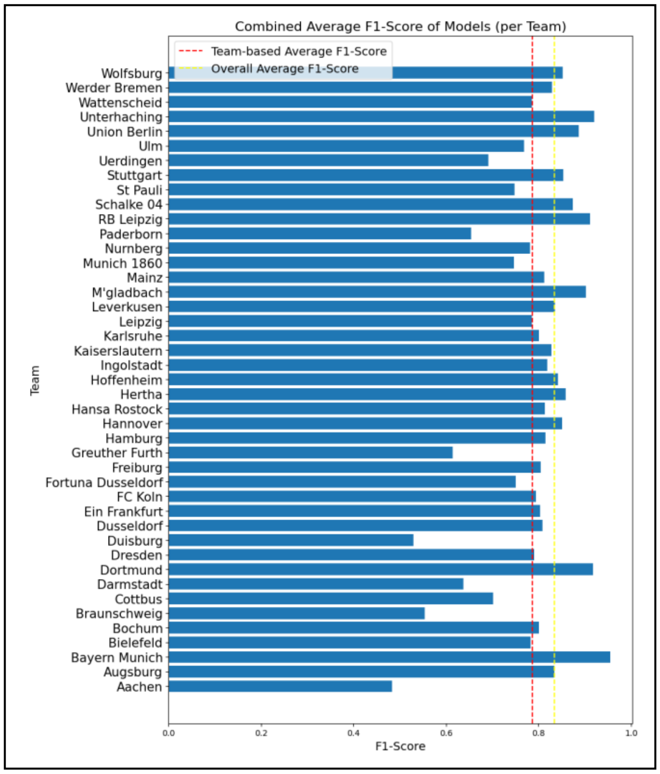

Figure 4. F1-score of models.

In analyzing the results, we observed that the accuracy of the machine learning models in predicting Bundesliga match outcomes was relatively low compared to other sports prediction models. This aligns with previous studies on football outcome prediction using machine learning, which also reported moderate accuracy. The complexity of football as a sport, with numerous unpredictable factors influencing match outcomes, contributes to the challenges in achieving higher prediction accuracy.

Even though the already promising, can be even further improved for the betting industry where even small improvements in prediction accuracy can lead to significant profits. Additionally, the analysis of player statistics and match performance data can provide insights into how teams can enhance their future performance.

V. Discussion

In our research, a critical decision was made to transition from competition-based analysis to team-based analysis. To ensure methodological consistency, both approaches were evaluated using k-fold cross-validation to assess their performance. While this technique proved effective for competition-based analysis due to the ample dataset, it encountered difficulties when applied to team-based analysis. The smaller dataset size resulted in less accurate predictions and decreased overall performance.

The performance decline was primarily attributed to the inclusion of teams that negatively impacted the prediction quality, thereby lowering the evaluation scores. Nonetheless, considering the available features, the models still achieved favorable performance.

The variation in performance between the two approaches can be attributed to disparities in data availability and variability. The competition-based analysis benefited from a larger dataset, enabling the models to capture broader patterns and trends. In contrast, the team-based analysis focused on individual teams and had a substantially smaller dataset, leading to diminished predictive capabilities.

Our findings underscore the importance of further investigating and refining the evaluation method. Potential strategies include removing teams with low statistics or employing up-sampling techniques. However, since the current research sufficiently addresses the formulated research questions, these refinements were not pursued.

Additionally, expanding the selection of competitions in future studies could yield even better predictive results. A larger dataset would contribute to increased accuracy and enhance the predictive capabilities of the models.

VI. CONCLUSION

In conclusion, this research aimed to predict Bundesliga match outcomes using data mining models and analyze the impact of competition-based and team-based analysis approaches. Various machine learning models, including logistic regression, decision trees, random forests, support vector classifier (SVC), and neural networks, were utilized and evaluated using metrics such as accuracy and F1-score.

With logistic regression giving a good prediction with a accuracy of around 0.8, it was able to predict the end result quite good. But it must be taken into consideration that this is a two category result where the game statistics are given to make a prediction, which is not useful in actual score prediction. So the prediction is accurately, but the features used for this prediction are not really useful.

Predictive models can be customized to incorporate team-specific characteristics and overall competition levels. Team-based analysis allows for the consideration of factors such as playing style, team dynamics, and individual player performance, providing insights into each team's unique patterns. Competition-based analysis captures the overall competition levels and interactions between teams, offering a broader context for predictions. Combining both approaches enables models to capture team nuances while considering the larger competition landscape. Customization depends on available data, competition nature, and analysis objectives.

In a comparative analysis of machine learning models for predicting Bundesliga match outcomes, the logistic regression model consistently outperformed other models in terms of accuracy and F1-score. The Random Forest and SVM models also demonstrated satisfactory performance. However, the Multilayer Perceptron (MLP) neural network and Decision Tree models showed slightly lower performance but still provided valuable insights into the data. The logistic regression model emerged as the top performer, but further research and fine-tuning are needed to enhance the accuracy given the inherent complexity and unpredictability of football.

This research contributes to the field of sports prediction by demonstrating the effectiveness of machine learning models in Bundesliga match outcome prediction. The results hold implications for the betting industry, where even small improvements in prediction accuracy can lead to significant profits. Additionally, the analysis of player statistics and match performance data can provide valuable insights for teams seeking to enhance their future performance.

Moving forward, further research could focus on refining team-based analysis methodologies, exploring advanced models, and incorporating additional factors such as team formations, player injuries, and recent form. These endeavors have the potential to enhance prediction accuracy and contribute to the development of more robust and reliable models for predicting Bundesliga match outcomes. Also clustering could be a great addition to put teams in clusters of the level of the team in the competition to make predictions more precise for these teams based on the historical results.

In summary, this research demonstrates the potential of data mining models in predicting Bundesliga match outcomes and provides insights into the impact of different analysis approaches. The findings lay the foundation for future research and practical applications in the betting industry and team performance analysis, paving the way for advancements in sports prediction and decision-making processes.

References

[1] The New York Times. (2022, August 12). Bayern Munich and the Myth of Competition. The New York Times. https://www.nytimes.com/2022/08/12/sports/soccer/bayern-munich-bundesliga.html

[2] Min, B., Kim, J., Choe, C., Eom, H. and McKay, R.B., 2008. A compound framework for sports results prediction: A football case study. Knowledge-Based Systems, 21(7), pp.551-562.

[3] Kanters, M.A., Bocarro, J. and Casper, J., 2008. Supported or pressured? An examination of agreement among parent's and children on parent's role in youth sports. Journal of sport behavior, 31(1).

[4] Lu, C.J., Lee, T.S., Wang, C.C. and Chen, W.J., 2021. Improving Sports Outcome Prediction Process Using Integrating Adaptive Weighted Features and Machine Learning Techniques. Processes, 9(9), p.1563.

[5] Bunker, R.P. and Thabtah, F., 2019. A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), pp.27-33.

[6] Heuer, A. and Rubner, O., 2012. Towards the perfect prediction of soccer matches. arXiv preprint arXiv:1207.4561.

[7] Haghighat, M., Rastegari, H., Nourafza, N., Branch, N. and Esfahan, I., 2013. A review of data mining techniques for result prediction in sports. Advances in Computer Science: an International Journal, 2(5), pp.7-12.

[8] Kozak, J. and Głowania, S., 2021. Heterogeneous ensembles of classifiers in predicting Bundesliga football results. Procedia Computer Science, 192, pp.1573-1582.

[9] Xu, H., 2021, March. Prediction on Bundesliga games based on decision tree algorithm. In 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) (pp. 234-238). IEEE.

[10] AKTUĞ, Z.B., Serkan, İ.B.İ.Ş., Hasan, A.K.A. and KILIÇ, F., 2022. The estimation of German Football League (Bundesliga) team ranking via artificial neural network model. Turkish Journal of Sport and Exercise, 24(1), pp.22-29.

[11] Bullock, G.S., Hughes, T., Sergeant, J.C., Callaghan, M.J., Riley, R. and Collins, G., 2021. Methods matter: clinical prediction models will benefit sports medicine practice, but only if they are properly developed and validated. British journal of sports medicine, 55(23), pp.1319-1321.

[12] Hucaljuk, J. and Rakipović, A., 2011, May. Predicting football scores using machine learning techniques. In 2011 Proceedings of the 34th International Convention MIPRO (pp. 1623-1627). IEEE.

[13] Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), 27-33.

[14] Carloni, L., De Angelis, A., Sansonetti, G., & Micarelli, A. (2021). A machine learning approach to football match result prediction. In HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II (pp. 473-480). Springer International Publishing.

[15] 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.

[16] Li, H. (2020). Analysis on the construction of sports match prediction model using neural network. Soft Computing, 24(11), 8343-8353.

[17] Kasera, M., & Johari, R. (2021). Prediction using machine learning in sports: a case study. In Data Analytics and Management: Proceedings of ICDAM (pp. 805-813). Springer Singapore.

[18] Ani, R., Harikumar, V., Devan, A.K. and Deepa, O.S., 2019, May. Victory prediction in League of Legends using Feature Selection and Ensemble methods. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 74-77). IEEE.

[19] Horvat, T. and Job, J., 2020. The use of machine learning in sport outcome prediction: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(5), p.e1380.