Sports Results Prediction Model Using Machine Learning

Alden Obradović [1], Dino Kečo [1]

[1] *International Burch University, Francuske revolucije bb, Sarajevo, Bosnia and Herzegovina*

***Abstract* – Within the past few decades, more and more people use Machine learning technology to predict sports performance. This paper aims to research the possibility of sports results prediction with good precision.**

Machine learning methods are often used to create model that will produce a representative output of the probable match result. This can assist coaches and managers in evaluating player performance, identifying skills, predicting injuries, and evaluating game strategy. Thanks to the good accuracy of the forecast, one of the constantly evolving areas is sports betting.

Finally, considering the current research review, a framework is proposed to develop a system based on artificial intelligence to predict the outcome of a football match.

This research provides information on how system learning algorithms have proven to be useful and how they can produce success. This research also reviews the research of data mining systems to predict athletic performance and evaluate the strengths and weaknesses of each system.

***Keywords* – machine learning, sport results, prediction.**

DOI: 10.18421/SAR73-03
https://doxi.org/10.18421/SAR73-03

Corresponding author: Alden Obradović ,
International Burch University, Francuske revolucije bb, Sarajevo, Bosnia and Herzegovina
Email: aldxen.obradovic@stu.ibu.edu.ba

Received: 27 March 2024.
Revised: 03 August 2024.
Accepted: 10 August 2024.
Published: 27 September 2024.

1. Introduction

Sport has always followed people's daily lives - from a kind of spiritual ritual to military training and as a form of entertainment. In addition to the sport itself, sports betting was always part of entertainment, since the time of the ancient Greeks who presented the Olympic Games and the earliest records of betting on athletic competitions. From the Greeks, sports betting spread to ancient Rome, where it was eventually legalized. The Romans bet the most on gladiator games. Although gladiator games were stopped, gambling survived and spread to other kingdoms and sports. During medieval times, some religious leaders tried to pass laws prohibiting sports betting, driving it underground. Despite these efforts, sports betting persisted and even flourished as new sporting events were introduced in the world.

Later in history gambling became very popular in England in the form of horse racing betting. The English spread this practice to the rest of the world, especially in the United States, where it quickly became a favorite pastime of many people.

A sport betting is very popular today, especially in Europe, which has become the largest market for sports betting, making sports betting not just a fun activity, but also a billion-dollar business.

With the development of information technologies and the advent of the Internet, sports betting have largely spread and continue to grow from year to year. The Internet has made information more available, enabling betting from the comfort of your own home, live betting and the like [1]. Predicting sports results is an interesting and challenging problem due to the unpredictable nature of sports and the seemingly infinite number of potential factors that can affect results, and sports betting participants are most interested in this type of problem.

According to Bunker and Susnjak the topic of predicting sports results using machine learning is becoming increasingly popular [2].

The prediction accuracy of the models for team sports is about 70%, with some models slightly below 60% and some slightly above 80%, depending on the sport. There are several studies related to individual sports such as horse racing, swimming, golf and tennis, where the accuracy of model prediction is about 65%.

To create the most accurate model possible, it is necessary to know the domain, i.e. the sport itself, in order to be able to better determine which are the important features that contribute to more accurate prediction. It is assumed that the model itself can achieve a prediction accuracy of around 60% due to a lack of domain knowledge.

2. Literature Review

The first research related to predicting outcomes in sports began in the late 20th century, and a large number of papers began to appear in the early 21st century. Generally speaking, the most commonly used methods of predicting outcomes in sports are certainly neural networks [3], [4].

As part of the research, author analyzed a large number of papers related to predicting outcomes in team sports such as basketball, football, American football, cricket and baseball and drew certain conclusions. The author concluded that researchers mainly use classification methods of supervised machine learning, and the main problem is the impossibility of comparing the results because there are practically no papers that use the same set of data. As a method of validation in most cases, the method of dividing the data set is used.

Fewer researchers also use the cross-checking method, but sporting events are not completely independent, and therefore cross-checking is not suitable for sports predictions. In general, the best results were obtained using smaller data sets.

Research has shown that the use of new, alternative machine learning algorithms can achieve good, in some cases better, prediction results. The main challenge related to predicting sports outcomes is to propose a universal method that will be able to successfully predict the outcomes of multiple sports. Since the testing of the model will be performed on football, special emphasis is given to works related to predicting the outcome in football.

As already mentioned, large amounts of structured and unstructured data related to sporting events are generated on a daily basis. In addition to increasing the amount of data, the number of relevant databases containing various sports statistics is also increasing.

Analyzing the works of other researchers, it is easy to see that the official pages of sports organizations are mostly used as a source of information. Papers that use at least one machine learning algorithm in predicting sports outcomes or papers that extract useful facts and regularities from the input data set were considered. Possible reasons for excluding papers were difficulty in interpreting the data, poor work methodology, or lack of clarity.

3. Methodology Section

The basic assumption of the prediction model is the existence of data. Data preparation is perhaps the most difficult step of machine learning, and the reason for this is the diversity and specificity of the data set, primarily oriented to the analyzed problem. However, there are common steps during the data preparation phase.

Different authors also suggest a different number of data set preparation steps, however regardless of the number of data set preparation steps, the final goal is the same. The preparation of a data set can thus be divided into several phases such as data collection, data cleaning, identification and calculation of missing values, data transformation, division of data into a learning set and a test set regardless of the validation method used, and finally the use of the data set itself.

Each sporting event can be described by a set of statistics. The peculiarity of the sport is that each sport is characterized by specific statistics. Specifically, the statistical parameters of football cannot be used in basketball, or the statistical parameters of basketball cannot be used in baseball, etc. However, there is one statistical parameter that is unique for all sports, and that is the final result, regardless of whether it is defined by goals, points, sets, etc. Ultimately, sporting events represent processes with the same goal, which is to be more successful in that statistical parameter that defines the final outcome.

Predicting outcomes based on a single parameter is very difficult, so a sporting event needs to be described using a larger set of parameters. Despite the fact that each sport is characterized by a specific set of statistical parameters, there are statistical parameters used in several sports, and this applies primarily to ball sports such as football, basketball, handball, water polo, etc. The analyzed set of statistical parameters for a sports event is not solely dependent on the sport itself but also on the competition, the country, and the capabilities of the event organizers.

3.1. Dataset

The data used in the research are taken from several publicly available source, platforms that offer a large number of data sets and resources for learning and applying machine learning in the form of competitions.

Focus of the research is English Premier League seasons 2000 – 2020. Data for the seasons 2000 – 2019 is retrieved from Kaggle.com, and the last season, season 2019/2020 is created based on sport result pages like livescore.com and rezultati.com.
Format that data was in is .csv and .json, and was later filtered using Python programming language. Python was chosen because it is a high-level interpreted language with simple syntax. Given these features, Python is easy to understand and suitable for fast programming and prototyping.
Once the data is filtered it is imported into MS SQL database using MS SQL Management Studio.

3.2. Data Preparation

To have the best dataset for further analysis and to get the best prediction at the end, data needed to be prepared. All rows that contained empty data were removed as well as the columns that contained data that was not needed, like the betting odds.
Since dataset was gathered from several different sources team names were not the same. Synchronization between names is done in the SQL using a team table that will be explained in the following chapter.

3.3. Data Model

Data model that was used for this research is presented in Figure 1.

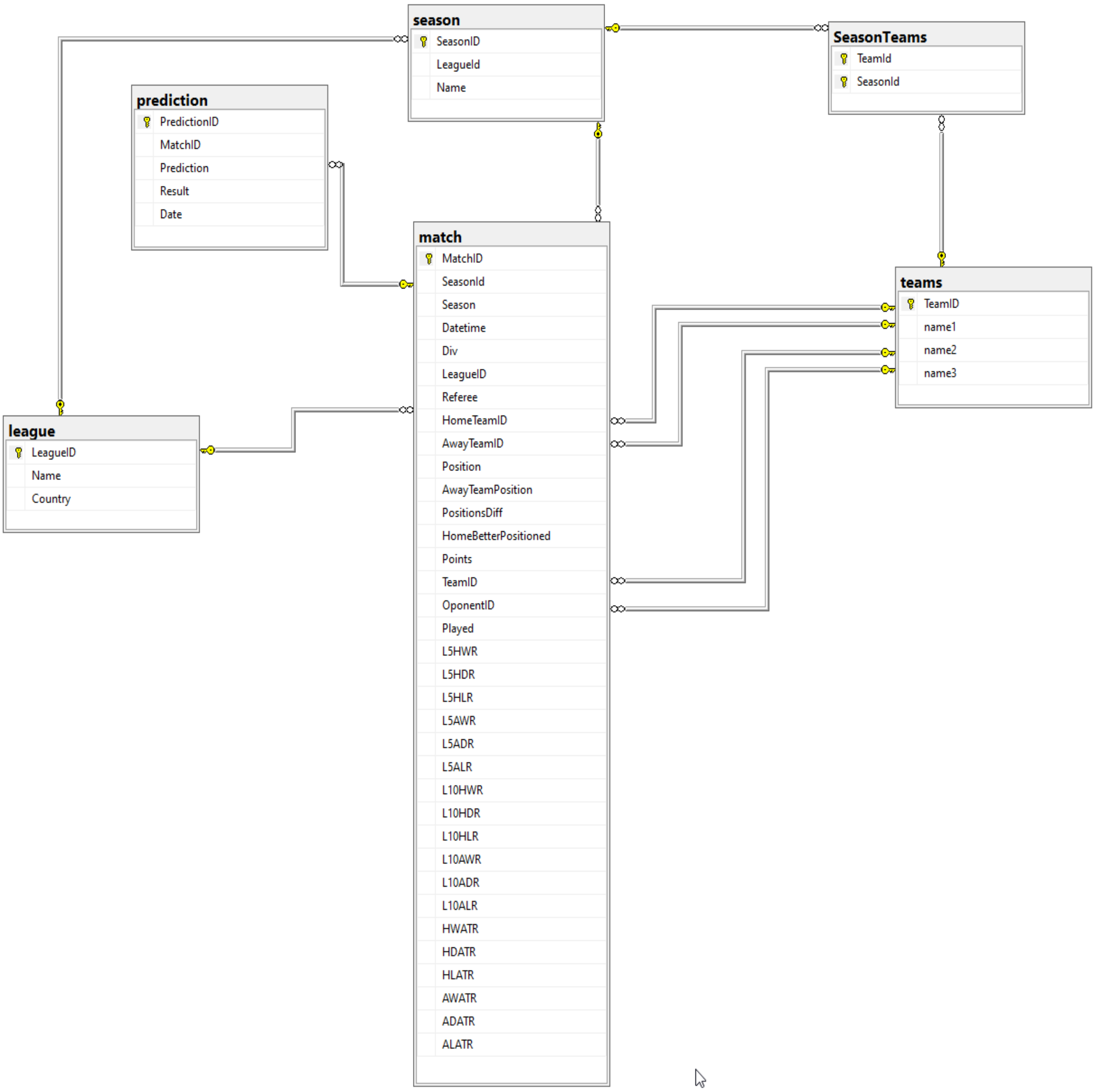

Figure 1. Data model

3.4. Prediction Process

Sport prediction is usually treated as a classification problem, with one class (win, lose, or draw) to be predicted. Prediction process is presented in the next figure, Figure 2.

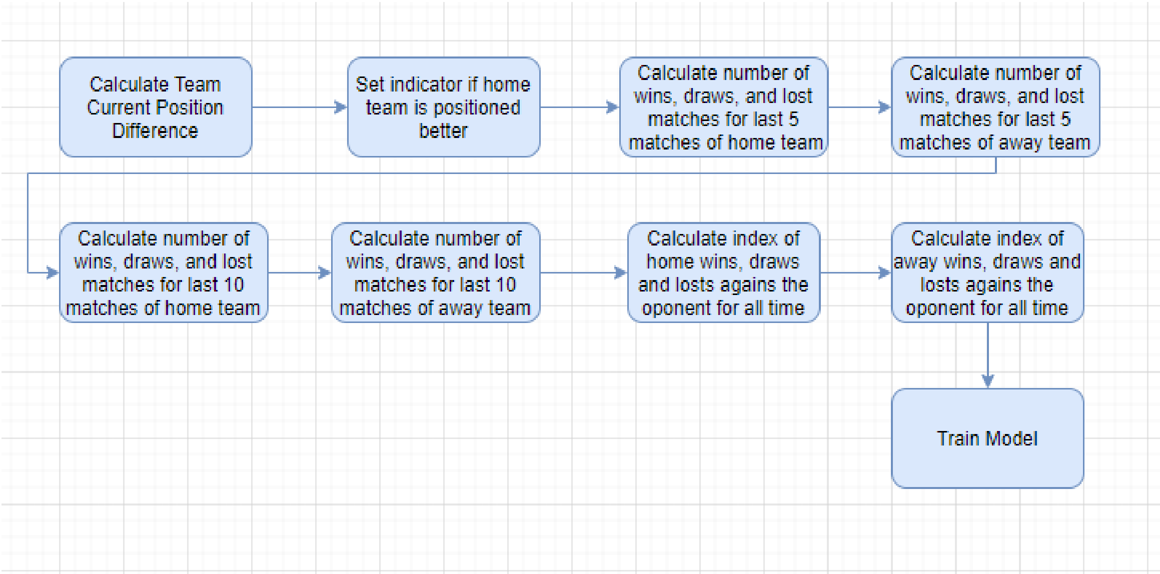

Figure 2. Prediction process

To obtain predictions on the wanted match user is using the following program. User enters necessary information and by clicking calculate in the background prediction process starts.
For training and testing submitted dataset Weka Explorer is used.

4. Results

In this chapter results for different prediction algorithms will be shown. Several algorithms were used in order to create several different models, so that these models could be compared and applied in a later period of program development. In Table 1 comparison between different models can be seen with focus on F-measure, ROC area and Accuracy measures.

Table 1. Results across different models

Prediction model	**F-measure**	**ROC area**	**Accuracy (%)**
Logistic Regression	0,590	0,786	60.15
Decision tree LMT	0,587	0,783	59.88
Neural network	0,590	0,798	60.95
SVM	0,585	0,713	59.88

Besides the previously shown results, confusion matrix is represented below.

	Logistic Regression Confusion Matrix					Decision tree Confusion Matrix			
	Home	Away	Draw			Home	Away	Draw	
Home	755	139	98		Home	752	132	108	
Away	187	367	87		Away	189	366	86	
Draw	224	122	172		Draw	222	126	170	
	Neural Network Confusion Matrix					SVM Confusion Matrix			
	Home	Away	Draw			Home	Away	Draw	
Home	790	137	65		Home	758	142	92	
Away	189	378	74		Away	186	368	87	
Draw	241	134	143		Draw	231	125	162	

5. Discussion

Predicting sports outcomes is a popular field for many researchers. There are a large number of newspaper articles that deal with the analysis of the popularity of sports, and the most current and extensive research related to finding the most popular sport is given by Sourav [12]. The author used 15 criteria to find the most popular sport related to the number of spectators, Internet presence, number of professional/amateur players, competitiveness, number of countries where the sport is popular, etc. The most popular sport is football, followed by cricket, basketball, hockey, tennis, volleyball, table tennis, baseball, etc.

According to Bunker and Thabath, a good knowledge of the domain is one of the most important factors when predicting sports results because the results themselves, i.e. statistics related to the match or match itself, are usually not a sufficiently comprehensive data set [13].

Most sports offer two or three final outcomes. Looking mathematically and without going into the complexity of the sport, it is easy to conclude that sports with two possible outcomes are easier to predict, more precisely there is a higher probability that the final outcome will be guessed. For example, in football, three possible outcomes are very likely, making it difficult to predict outcomes. Also, football is a team sport in which the final outcome is influenced by several parameters. A small number of goals are not in favour of football.

Considering the above reasons, mathematically the highest probability for an accurately predicted outcome are individual sports with two possible outcomes such as tennis, chess, most martial arts, but not individual sports in which more competitors participate such as gymnastics, athletics, swimming and other similar sports.

As stated earlier, looking at the strict definition of an individual sport, it should be easiest to predict the outcomes of individual sports in which two competitors compete and two outcomes are possible. In this case, the specifics of the competition, budget constraints, league competitiveness, competition format and other similar factors are not considered.

In addition to specific metrics related to the game itself, external factors affect the result of the game, and they are usually not considered in the model itself, and therefore most machine learning models in predicting sports results achieve an average accuracy of 70%. To improve these results, it is necessary to collect more data describing external factors and include them in the model.

Furthermore, historical match data should be subjected to the calculation of average values, so that the last record shows the average values of all matches before that record.

All algorithms showed rather poor accuracy. Most algorithms have accuracy around 60%.

The test results showed that, in this project, both linear algorithms (logistic regression and linear support vector machine) and nonlinear approach classification algorithms (k nearest neighbors and RBF support vector machine) have similar results. The best results were obtained when using neural networks.

Showed results are in accordance with previous results gained in several researches where it was shown that the average prediction of results in football is between 50% and 70% [5], [6], [7], [8], [9], [10], [11].

6. Conclusion

Without adequate model testing, it is not possible to know how reliable and applicable the model is. In this project, testing was conducted on each model to obtain comparable results.

The previously separated test data set is used by the model classifying the input space (X) data into a set of labels (Y) and then comparing its classifications with the correct answers. The measure obtained is called accuracy. Precision and recall values are also calculated in the testing process. Precision measures how many of the predicted positives are real positives, so precision can give an insight into how many negative positives there are - the higher the value of precision, the fewer negative positives. The response shows how many positives the model has labeled as positive, that is, the response measure indicates how many false negatives there are.

Machine learning is an important part of scientific activity today because it is a way to use huge amounts of data generated from a multitude of digital devices based on which new information is created. The accessibility of Python and the availability of various packages specializing in data science allows the use of machine learning methods in almost all sciences, and proves to be a useful way of acting outside the information and computer science circles. Today, machine learning is no longer used only to predict or classify, but can also be used to create new content through, for example, GAN algorithms. However, the application of machine learning could expand with the development of ideas such as the concept of open science that would allow even greater access to data and thus open space for new research based on machine learning.

References:

[1]. Hing, N., Smith, M., Rockloff, M., Thorne, H., Russell, A. M., Dowling, N. A., & Breen, H. (2022). How structural changes in online gambling are shaping the contemporary experiences and behaviours of online gamblers: An interview study. *BMC Public Health*, *22*(1), 1620.

[2]. Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, *73*, 1285-1322.

[3]. Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural networks*, *1*(1), 17-61.

[4]. Zhang, G. P. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *30*(4), 451-462.

[5]. McCabe, A., & Trevathan, J. (2008). Artificial intelligence in sports prediction. In *Fifth International Conference on Information Technology: New Generations (itng 2008)*, 1194-1197. IEEE.

[6]. Buursma, D. (2010). Predicting sports events from past results. In *14th Twente Student Conference on IT*, 21.

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

[8]. 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.

[9]. Prasetio, D. (2016). Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 1-5. IEEE.

[10]. 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.

[11]. Knoll, J., & Stübinger, J. (2020). Machine-learning-based statistical arbitrage football betting. *KI-Künstliche Intelligenz*, *34*(1), 69-80.

[12]. Das, S. (2022). Top 10 most popular sports in the world. *Sports Browser*.

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