Journal of Advanced Research in Computer Technology and Software Applications
Volume 7, Print Issue 1 - 2023, Pg. No. 1-7
Peer Reviewed Journal

Research Article

Forecasting Football Matches: An Analysis on Predictive Models and Performance Evaluation Techniques for Betting

Savya Khanna[1], Utkarsh Bhagat[2]

[1,2]CSE Department, DAV Institute of Engineering & Technology, Jalandhar, Punjab, India.

I N F O

E-mail Id:
technicalpapers01@gmail.com
Orcid Id:
https://orcid.org/0009-0002-5709-0070

How to cite this article:
Khanna S, Bhagat U. Forecasting Football Matches: An Analysis on Predictive Models and Performance Evaluation Techniques for Betting. *J Adv Res Comp Tech Soft Appl* 2023; 7(1): 1-7.

Date of Submission: 2023-05-01
Date of Acceptance: 2023-06-12

A B S T R A C T

Football is one of the foremost widely followed sports in the world, thus fans, coaches, the media, gamblers are all interested in understanding the game and forecasting the results. It is impossible to predict the result of a football match, yet the football industry has expanded all through time. The un predictable nature of football games, as well as the expansion of the betting industry, point to the creation of predictive models to assist punters. In this work, we create a machine learning approach for predicting the outcome of a football match using a collection of data from past matches and the attributes of the players on both sides. Several hypotheses were investigated, the experimental findings demonstrated that performance in the setting of competitive football is supported by data.

Keywords: Data Mining, Sports Betting, Feature Selection, Distribution, Football

Introduction

Football is one of the most popular sports in the world. Sport has changed a lot over the past few decades as the capital comes from joint ventures. Making a profit in sports betting is very difficult, especially when it comes to football. Predicting the result is a difficult task because there are many factors that influence the game. Due to the nature of the game, the team can lose to a ridiculous group, which makes any bet difficult Table .

The unpredictability of the game makes it difficult to bet without doing some analysis of the game data.

There are now some websites that give free football tips. These sites make predictions based on team strength, number of goals[1] and math.[2] Although these websites provide a lot of information, they are unreliable and betting on these estimates can result in huge losses. An example website is forebet.[2]

Taking the 2017/18 Premier League as an example, the website can only predict correctly 2 out of 10 games per round, which shows the risk of relying on these predictions.

So, there is an opportunity to create a better predictor to show the most likely outcome of a football match and the reliability of the outcome, which leads to more knowledge about gambling.

A football match can be predicted by analyzing historical data from the previous season. The availability of competitive information in many football sports is expanding, indicating information gathering. In this article, a Machine Learning (ML) algorithm will be used to predict the outcome of a football match with various features and the behavior of all players of both teams.

The remainder of this paper is organized as follows. Chapter 2 describes related activities. After examining

and processing the data in Chapter 3, several models were built using different learning machines, models were tested in Chapter 4. Finally, Chapter 5 concludes the article and looks to the future.

Context and State of the Art

Sports betting involves risking a certain amount of money by trying to predict the outcome of a sporting event, profiting if the prediction is correct, or losing all if the prediction is correct. While there are many types of sports bets, in this article we will focus on the most common type of football, 1X2 bets. There are three possible outcomes: 1 - home team wins, X - draw and 2 - away team wins.

The winnings from a particular bet are calculated by multiplying the winnings by the odds of the bet. An odd number is a value greater than 1 that indicates the probability of that event.

The higher the probability, the lower the odd number (i.e., closer to 1). Since a football match can have three different outcomes, the chance of getting the result by chance is 1/3, the way bookies set their odds. Odds are the odds which make up the average loss. The use of machine learning in sports has increased significantly in recent years and has led to decisions in this area.

For example, ML was successfully applied by a German research team at the 2014 World Cup to study rival teams and monitor their players to support the manager's decisions.[3] Gomez et al.[4] used statistics from more than 4900 matches over 13 seasons (2000/01 to 2012/13) to develop a predictive model. This model was tested in 2013/14 season with 54.29% accuracy and 20% profit in 7 rounds, a total of 70 matches.

An important stage of this research is to determine the variables that will occur before the match starts, such as the average number of goals scored by the team.

Despite the uncertainty of the sport, the football world has produced surprises from time to time, for example Leicester City's Premier League title in 2015/16. An in-depth study[5] was conducted to try to understand what caused this confusion and what the future could be. It turns out that this was done because of the quality of the Leicester goalkeeper and his goalscoring efficiency on offense. Another important thing is that many Leicester City players make a lot of passes with a success rate of more than 80%.

In this study, a model has also been developed to predict the number of shots and goals scored by the team in the match. It has been seen that model containing information about the type of shot (e.g. shots from the opponent, shots from the middle to the penalty area) achieve more useful results. There is also a case study in which soccer matches are predicted using multiple factors.[6]

The study used is a multilayered sensor and experiment using data from the 2015/2016 season of the Spanish Super League. The success rate is 61%. In another study[7], logistic regression was used to predict Premier League matches for the 2015/16 season. The prediction model only predicts the home team's wins and losses, without considering possible draws. The rating is around 69. This concluded that the variable with the greatest impact on the forecast was data protection at home and abroad.

Another study[8] used data from the 2010/ 11 season of Italian Serie A, using 300 games for training and 80 games for testing. One conclusion from the study is that a team that uses more jump balls will either equalize or lose matches. Finally, the study[9] explores machine learning to predict the outcome of football based on the game and the characteristics of the players. A simulation study covering all matches from Europe's top five leagues and minor leagues from 2006 to 2018 shows combined teams achieving results Key benefits 1.58% per game. Analysis conducted by shows that studies with lower standards are associated with a lack of differences in what is best for the player and the game itself. Also, the model must be learned using games from different seasons because the teams in each season are very different.

Data Description

The data to be used in this study corresponds to a total of 1900 football matches played in 5 seasons between 2013/2014 and 2018/2019. These competitions are associated with England's top football league known as the Premier League. Of the 1900 matches recorded, 861 (45.3%) were won by the home team, 470 (24.7%) drawn, 569 (29.3%) by the away team.

In addition to free statistics for football matches, information is collected on the performance of both teams, including goals, shots, corners, substitutions, pieces, yellow and red cards, the difference between a game, the final result of the match and the referee. In addition to this information, data from the sofafa.com website contains information describing each athlete's personality and skills (e.g., passing accuracy, agility, reflexes, aggression) and football team quality statistics are also used. Most of these variables are scored on a scale from 0 to 100. Other variables related to team performance, such as the overall rating and rating of the team's offensive, midfield and defense, are considered constant for each team throughout the season.

After collecting the data, examine the data for potential problems. However, no important or conflicting data was found, so the data count is still 1900 after clearing the data.

File Handling

New variants have been created to better predict the end of the game. As in[4], the difference refers to the home team's

home wins and the away team's wins. In the research articles examined, only the data on the goals scored by the team were used, the data on the recognized goals were not used.

Therefore, we decided to count the goals recognized by the group, as these statistics can improve the forecast. The variables produced are the average number of goals conceded by the home team in home matches and the average number of goals conceded by the away team in away matches.

In order to make predictions before the football match starts, it is necessary to combine the prediction model with the data available before each match. Since the data extracted relates to the final time of each game, such as the number of goals and shots for each team, this data cannot be used directly to train the prediction model alone. For this, this information needs to be changed.

The solution that works in this case is to consider the average of the available data, such as the team's average goals before a game or the average of shots. Calculate the average of the following attributes for each match and each team: Goals, Attacks, Baskets, Fouls, Yellow and Red Cards.

Therefore, for a match, the average number of goals scored by the home team is calculated from the matches the team played at home before that season. For the away team, the average is also calculated based on the matches played by the away team in the season in which the match is played. The mean values were calculated for the remaining variables in the same way.

Data Exploration

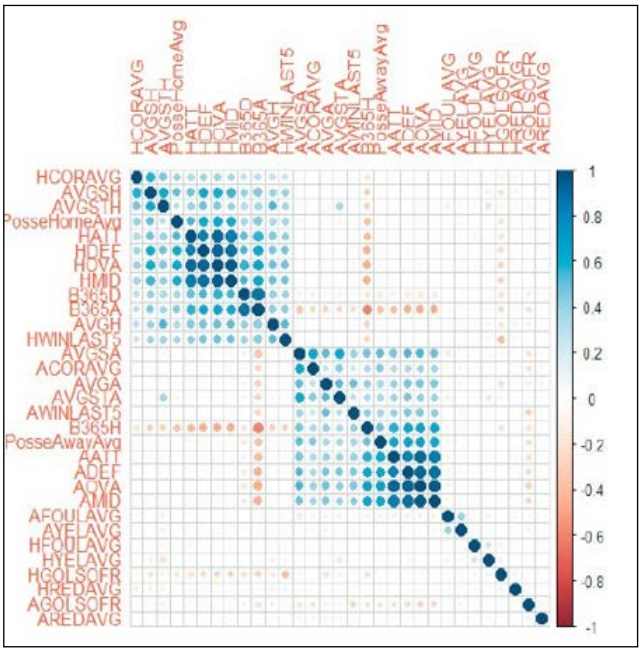

Figure 1.Matrix correlations between variables

Then, out of a total of 31 variables available, it identified the most relevant variable, which one of the variables more clearly predicted the target's behavior, any changes could be excluded. To analyze the relationship between the variables, a correlation matrix was created from all the variables shown in Figure 1.

It can also be concluded that the B365H variant is negatively associated with mutations such as B365A, HATT or HMID. This is normal because the higher the home team difference, the lower the opposing team. Similarly, the home team is better, i.e., the higher the HATT and HMID values, the less chance the home team has to win. The correlation matrix is also important for identifying differences to be extracted from the original data. During classification, variables associated with other variables should be removed.[10]

This should be done to avoid overestimating the significance of these changes, thereby affecting the estimation of the outcome. If there are two variables, one of them repeats and does not add any relevant information to the learning model. Therefore, the difference between HATT and AATT corresponds to the contrast between playing indoors and outdoors, the correlation greater than 0.9 is removed.

An important step before estimating is to try to identify a variable that helps predict the target behavior/variable more easily, this is the FTR variable.

Therefore, several charts were created to analyze the relationship between each variable and the FTR values: Win, Draw and Lose. This chart allows us to identify the four variables that are most useful in predicting the ending of the game. These variants are B365H, B365A, AOVA, HWINLAST5 shown in Figure 2. In addition to the accepted variable that can provide a better prediction, the variables that have the least potential to predict the main target are also identified. There was a difference between HREDAVG, AREDAVG, HCORAVG and ACORAVG (Figure 4). It has little to do with training the predictive model. Therefore, these 4 variables were not used in the training of the prediction model.

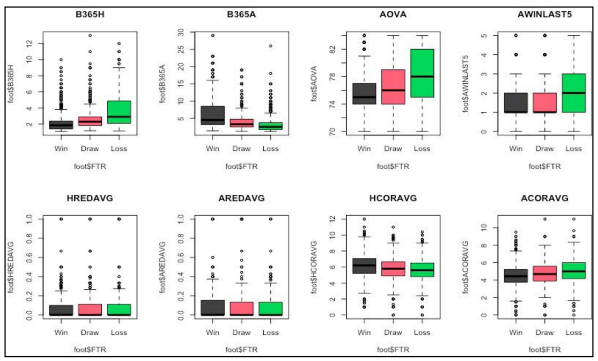

Figure 2.Box plot of 4 variables associated with the best and worst characteristics of the target

Data Mining

Then the data is divided into two parts as training and testing. In this study, it is argued that the test should include every game in a season, as the performance of the team is different every season. Teams may not be at their previous level at the start of the season and may wear out or meet their goals by the end of the season, causing some teams to underperform. These features can lead to unexpected results, so the classification model must be tested each season to be reliable.

Regarding the training process, there may be a risk of overfitting if the model is trained with too few samples, i.e., the model overfits the training data and potentially bad prediction.

For this, 4 seasons were used for training and 1 season for testing. The 4 training seasons (2013/2015 - 2016/2017) correspond to a total of 1520 games, the test season (2018/2019) 380 games.

In order to find the best classification model, many algorithms with different properties have been tried to find which one is best for the data. Next, the algorithms and R software libraries used:

- Naive Bayes (NB) – e1071 package
- K Neighbors (KNN) – kknn package
- Support Vector Machine (SVM) – svm method from package e1071.
- C5.0 (Decision Trees) – C50 package.
- Xgboost – xgboost package.
- Multinomial Logistic Regression (MLR) – from the package nnet Multinomial method.
- Artificial Neural Network (ANN) - nnet method of nnet package.

Before testing different algorithms10, normalize the data using the "zscore" method to eliminate the risk of large changes in value.

After normalizing the data, the most important variables were identified to predictoutcomes. This is done so that the forecasting model uses only the relevant variables and guarantees a better forecast. Boruta algorithm[11] was used to determine the best differences. Boruta is a heuristic variable selection algorithm based on the Random Forest algorithm designed to find the most variables in a dataset. The results of running the algorithm show correlation and inequality.

This algorithm is used because it does not find the best solution but tries to find all the differences with the relevant data, removing any changes that may adversely affect the prediction model. Boruta's algorithm removes 7 random variables, leaving 18 variables for the design of the distribution.

Prediction Results - Level 1

To obtain a different classification model, different criteria are used, such as the accuracy of the model and the percentage of the predicted game for the group to draw and go home and leave to win.

Each time the bet turns out right or wrong, it is also considered as the result to be obtained. Since the betting model is involved in the decision of the game, it is important to calculate the winnings to verify the success of the model. Results are calculated at a price of 2 Euros per bet. Note that there are 380 test games with a total bet of €760 to be played in 9 months. In case of losing the bet, the profit will be reduced by 2 EUR.If true, the result is calculated according to the equation:

$Proffait = betamountbt$ for example, in a game with draw and draw odds of 1.5, the bet value will always be 2 EUR. The result will be equal to 2x1.5-2 = 1.

Table 1, shows the prediction results of eight models developed using the selected algorithms. It should be noted that all algorithms are profitable. However, the best match cannot be obtained beyond what is found in scientific literature. The best model is correct for only 3.57% of the relationship, which is a very low value.

Table 1.Fore cast results with 18 variables

Algo-rithm	Accuracy	Profit	% Victories Home Team	Draws	% Victories Away Team
Bayes	53,42%	17,40€	51,87%	30,95%	73,79%
KNN	57,63%	78,02€	78,07%	15,48%	55,05%
RF	59,21%	85,20€	75,40%	21,43%	60,55%
SVM	61,32%	95,06€	88,77%	3,57%	58,72%
C5.0	55,26%	42,52€	72,73%	23,81%	49,54%
Xg-boost	59,47%	72,80€	77,54%	10,71%	66,06%
RLM	57,63%	32,56€	78,07%	5,95%	62,34%
RNA	50,00%	18,28€	58,29%	30,95%	50,46%

As the results did not meet expectations, it was decided to develop a new forecasting model.

Prediction Results - Phase 2

In the second phase, we decided to test all combinations using 18 preselection's to get the best score. However, the total number of combinations to be tested with 18 variants would be more than 260,000.

This number was too high for this method to be fair, so it was decided to first identify the most significant difference between the 18 options using the reverse feature option

"rfe" method of the software "caret" package R. The first algorithm used. All given variables define the values of the variables and then complete the iterations, eliminating some variables leaving only the most significant of each iteration. Finally, the most important variables are those used in the experiment that produce the best results. The most important features identified by the algorithm are: B365H, B365D, B365A, AVGH, AGVA, HOVA and AGOLSOFR. In this way, 11 characters are left out of the first 18 characters, so in this case, the total number of connections needs is 2048 and this method is possible.

In this new method, all combinations are tested using 7 characters, only the remaining 11 characters are different, and 8 patterns are always created using all combinations. Table 2 shows the best results obtained with each combination of variables.

This program aims to improve existing skills by combining two popular modern methods of forecasting, expected goals and team offensive and defensive scores. This is thanks to the huge amount of data currently being recorded on football matches.

Different machine learning models will be tested, and different design models and hypotheses will be explored to get effective predictive models.

To generate predictions, we need to achieve several goals:

First, we need to find and clean up good data to use in our model.To do this, we need to find the necessary information. This will give us access to a wide variety of stats for us to use, compared to most studies done on the subject in the past that only determine the end of each game.The main way we're going to do this is to model project expectations

Table 2.Better predictions with the various combinations of variables

No. of variables combined	Algorithm Accuracy	Profit	% Victories	Home Team	% Victor	Away Team
1	SVM	62,62%	134,50€	51,87%	30,95%	73,79%
2	SVM	63,68%	157,22€	90,91%	4,76%	62,39%
3	SVM	63,95%	160,12€	91,44%	3,57%	63,30%
4	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
5	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
6	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
7	RF	63,95%	191,58€	80,21%	28,57%	63,30%
8	RF	65,26%	203,24€	81,28%	29,76%	65,14%
9	RNA	62,89%	181,72€	82,35%	27,38%	56,88%
10	SVM	62,36%	123,36€	88,77%	8,33%	58,72%
11	SVM	62,11%	115,06€	88,77%	7,14%	58,72%

Results Analysis

Method used to evaluate differences between performance differences. In any case, the best model outperforms the first model with a success rate of 61.32%.

Algorithms to arrive at the best model in most cases are SVM, RF, Xgboost and RNA. The best model was obtained by testing a combination of 8 variables, of which 7 were identified as most important, so there are 15 variables in total. The top model uses the RF algorithm and achieves a 65.26% rating and a 26.74% margin.

Compared to the first model, the success rate increased by about 4%. The percentage of correct bets for the home team decreased by 7%, but the draw increased by 26% and the percentage of correct bets for the team's victory was around 7%. Revenue increased 14%. This increase is evidenced by the increase in the number of correct guesses. Betting on a draw tends to have more chances than betting on a win and can therefore be more profitable.

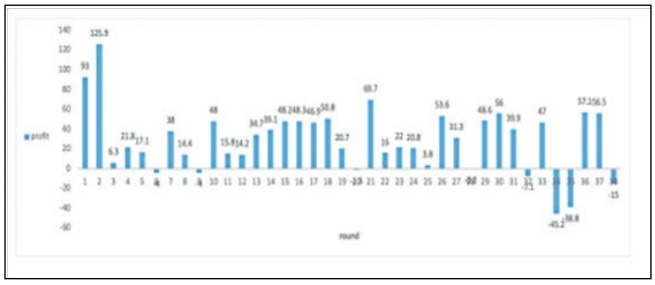

Figure 3.Profit of best model by round

The model performed well throughout the season. While 30 rounds have been won out of 38 rounds this season, only 8 rounds have been lost.

This gives an average of 26 margins.78% compete with business analytics. This program aims to improve existing skills by combining two popular modern methods of forecasting, expected goals and team offensive and defensive scores. This is thanks to the huge amount of data currently being recorded on football matches.

Different machine learning models will be tested, different design models and hypotheses will be explored to get effective predictive models.

To generate predictions, we need to achieve several goals:

First, we need to find and clean up good data to use in our model.To do this, we need to find the necessary information. This will give us access to a wide variety of stats for us to use, compared to most studies done on the subject in the past that only determine the end of each game.The main way we're going to do this is to model project expectations to better understand how the team is doing so we can make better predictions about the future. To build this model, we will test different types of machine learning and algorithms for best performance. We may use information about shots and goals, such as the location of the pitch or the angle to the goal, to predict a team's expected goals in a match and reduce the risk of gambling. A key part of the project will be to set up a viable machine learning and testing pipeline to be able to test new algorithms with new capabilities.

Objectives Finally, our model will be evaluated

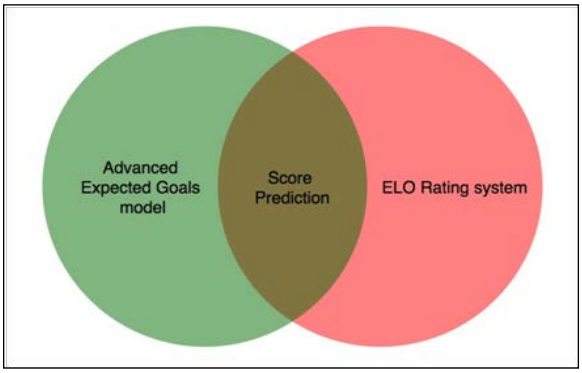

Figure 4.The main purpose of the project is to generate forecasts. It is easy to compare this with other models that lead to the general operations shown in Figure 1

by predictive models, including bookmaker competition, using different performance indicators. The success of this project will be the creation of a classification model that can predict future game outcomes and a regression model that can predict past game scores with better prediction performance than different methods.

Machine Learning Techniques

In this section, we provide an overview of popular machine learning techniques and their classification and replication.

Tracking Learning is a task of learning a program that maps input data to output data based on input output pairs. Classification occurs when the output is a category, and regression occurs when the output is a continuous number.

In our case, we want to predict the result of the team (home win / draw / away win) or the number of goals scored (conjunction) by the team, so we just need to focus on monitoring the learning of the machine learning. We noted some of the popular training tracks in Figure 2.1 and will now illustrate them in more detail.

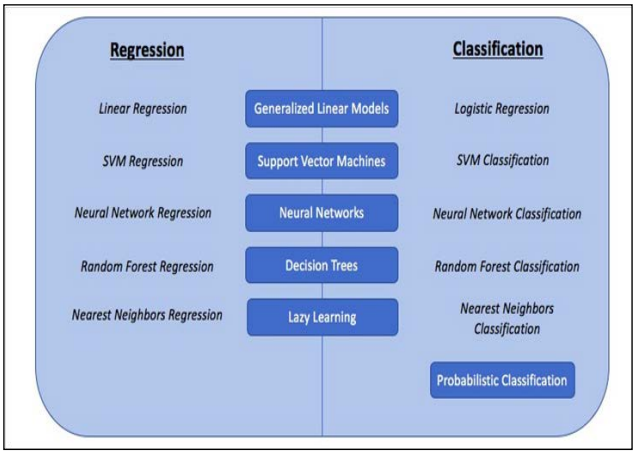

Figure 6.Overview of supervised machine learning techniques

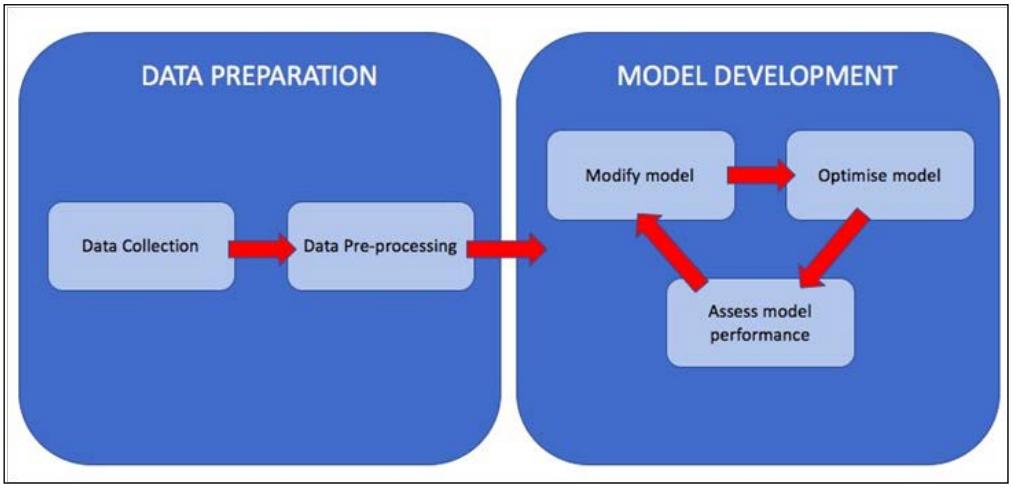

Figure 5.Project work flow

Conclusion

This paper covers the process of developing a model for forecasting football match outcomes in order to facilitate gambling on sports. Data from two separate sources was used, one for historical match statistics and the other for team statistics. The data analysis and processing allow the major points of the parameters utilized in the model to be drawn.

The prediction model will be incorporated into a system that supports decisions in the future, which will first assess gambling threat based on the likelihood of the prediction model's result. This ensures that punters understand the risks connected with their bets and, as a result, are more driven to benefit from sports betting.

References

1. Soccer Vista Football Results, Predictions and Betting Options, 2018. Available from: http://www.soccervista.com. Last visited: January 10, 2022.
2. Forebet, football numbers, tips, stats, preview, 2018. Available from: https://www.forebet.com. Last accessed: January 10, 2022.
3. Bojanova I. IT enhances football at world cup 2014. IT Specialist 2014; 16(4): 12-7. [Google Scholar]
4. Gomes J, Portela P, Santos MF. Decision support to predict the outcome of a football match. 2015, pp. 348-353. [Google Scholar]
5. Ruiz H, Hwj Chim P, Wei X, et al. Leicester City tales? Comparing the performance of the 15/16 and 16/17 EPL seasons using new football analysis tools. In the Proceedings of the 23rd ACM SIGKDD International Conference on Information Discovery and Data Mining, 2017, p. 1991-2000.
6. Cañizares PC, Merayo MG, Nunez M, Suarez-Paniagua V.A multi-agent system architecture for statistics managing and soccer forecasting. IEEE International Conference on Computational Intelligence and Applications (ICCIA), p. 107-111, 572-576. [Google Scholar]
7. Prasetio, D., 2016, August. Predicting football match results with logistic regression. In 2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA), pp. 1-5. [Google Scholar]
8. Zuccolotto P, Carpita M, Sandri M, et al. Football mining with R. In using Data Mining with R. 2014.
9. Stübinger J, Mangold B, Knoll J. Education in football: gambling results based on player traits. Applied Sciences 2020; 10(1): 46. [Google Scholar]
10. Dasgupta, N., 2018. Practical big data analytics: Hands-on techniques to implement enterprise analytics and machine learning using Hadoop, Spark, NoSQL and R. Packt Publishing Ltd. [Google Scholar]
11. Kürsa MB, Rudnicki W., Use Boruta button for special options. J Stat Software 2010; 36(11): 1-13.