ISSN 2347 - 3983
Volume 8. No. 1.1, 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter1181.12020.pdf
https://doi.org/10.30534/ijeter/2020/1181.12020

Football Match Outcome Prediction by Applying Three Machine Learning Algorithms

AzrelAiman Azeman, Aida Mustapha, Salama A Mostafa, Samah W.G. AbuSalim,
Mohammed Ahmed Jubair and Mustafa Hamid Hassan
Faculty of Computer Science and Information Technology, UniversitiTun Hussein Onn Malaysia, Parit Raja,
86400 BatuPahat, Johor, Malaysia
CI170068@siswa.uthm.edu.my, {aidam, salama}@uthm.edu.my,
{samahwa.salim, mohamed.a.jubair,mustafa.hamid.alani, }@gmail.com

ABSTRACT

Football prediction has become an interesting problem, and researchers are trying to find solutions that are different from each other. The idea of modeling and building a smart model that studies the available data and expects the outcome of the game has become more popular and widespread in the past few years. Professionally different predictive methods have been developed to assess the characteristics that cause a soccer team to lose a match or win a match. Machine Learning (ML) is one of the branches of artificial intelligence that is concerned with designing algorithms that allow computers to have the advantage of learning without programming the rules for each issue. These algorithms consist of a series of commands, and instructions necessary to direct the machine or computer to how the tasks should be carried out, as the algorithms play the role of the mastermind in the machine because of its polarization of data, collection, analysis and finally relying on the analyzed data to determine how the task should be performed. The algorithms used in machine learning rely on a set of graphical models and decision tools such as the decision tree, natural language processing, and artificial neural networks for the task of automating analyzed data and processing; thus, motivating the machine to make decisions and carry out the tasks assigned to it with precision and ease. This paper implements three ML algorithms; Decision Forest (DF), Artificial Neural Network (ANN) and Support Vector Machine (SVM) to predict the match result. The result showed that DF achieved the highest accuracy of 89%, ANN achieved the accuracy of 72%, and SVM achieved the accuracy of 70%. It is worth noting that the prediction results are obtained based on 10-folds cross-validation.

Key words : Artificial Neural Network, Decision Forest, Football, Match Outcome Prediction, Support Vector Machine.

1. INTRODUCTION

Football is one of the most widely practiced sports around the world since 1930 and is the most popular among different sports. Because of this, FIFA has been established, organizing and developing the basic laws of football and hosting them once every four years. With the increasing popularity of the game, local and international tournaments appeared through league and cup competitions in neighboring European countries such as the Netherlands, Germany and France, reaching South American countries such as Argentina. And Brazil [1]. Both football fans and the administrative team are curious about knowing the results of football, which leads them to rely on the various programs and applications that are used in predicting the results of the game, which have become available with more than one feature [2].

Football prediction has become a focus of attention in many scientific types of research, due to the fact that it is difficult to have high confidence in the outcome of the prediction, due to the difficulty in formulating many factors that significantly affect the game such as weather, teamwork, skills and many other factors [3]. The problem is also faced by sports experts, as it is very difficult to predict the outcomes of football matches [1-4]. The usual time for a soccer match takes 90 minutes, meanwhile, many unpredictable things happen, such as hitting a player or kicking a player off the red card, or early goals from stationary kicks that make any prediction based on logic fail.

The issue of predicting the outcome of a football match requires clear luck, and the elements that play a role in it cannot be claimed that everyone has counted all of them, as it is possible for the weak team to beat the strong team [2], [5]. In this way, it is not surprising that much research has been done on predicting football results. The research began to predict football results in early 1977 by [6]. Stefani [6] developed a model called the least squared model that measured both the power of the home and the away team using the goal score distribution matrix.

However, with the introduction of Bayesian Networks, the first football prediction model has only begun in the work of [7]. The work of [7] suggested a Bayesian network to take into account differences in time for all properties simultaneously (Dynamic) also known as Dynamic Bayesian Networks

(DBNs). As a result, the offensive and defensive strength of both the home and away teams will change over time.
This paper focuses on Machine Learning (ML) approach due to its proven applicability in predicting. The three ML algorithms Decision Forest (DF), Artificial Neural Network (ANN) and Support Vector Machine (SVM) are selected to predict the match result. The three ML algorithms for the prediction of football results implement the attributes provided by football matches results for English Premier League (EPL) session 2005-2006 dataset in the section of historical data for English Football Results.

The rest of this paper is organized as follow. A summary of earlier work on football predictions is presented in Section 2. In Section 3, the experiments including the dataset, the Machine Learning as well as the results in terms of predictive accuracies are presented. Finally, the conclusions are in provided in Section 4.

2. RELATED WORK

Many efforts have been targeted towards improving the accuracy of the prediction result of the football match. The researchers proposed several models via implement different ML algorithms. Razali et al. [2] build a Bayesian hierarchical model that predicts football results. Their model relied on the goals that both teams scored in each match. Min et al. [8] provided a dynamic system for predicting the outcomes of football matches. This dynamic structure known as the FRES system consists of two main components: theorem based on rules and the Bayesian network component. Therefore, the FRES method is a mixture of two methods that work together to predict the outcomes of football matches. In addition, the FRES method has also been introduced in-game time-series approach, which allows prediction more practical. Nevertheless, the FRES program requires sufficient professional expertise in order to be well controlled.

Constantinou [9] has developed a football prediction model called pi-rating to produce prognoses on the outcome of football matches, whether home win, draw or away win for EPL matches during the 2010/2011 seasons, which combines objective information and subjective information such as team strength, team form, psychological effect and fatigue. Koopman and Lit [10] are expanding work by Maher [11] on the Poisson distribution, demonstrating the offensive and defensive power of the goal score distribution. Koopman and Lit [7] are developing a statistical model for the study and estimation of the outcomes of football matches, which assumes a bivariate distribution of Poisson with coefficients of intensity that vary randomly over time.

3. METHODOLOGY

The methodology used in this research is Knowledge Discovery in Database (KDD). KDD is not an easy process that stops when collecting and managing data, but rather extends to analysis, searching for cognitive patterns, predicting and exploring the vast and increasing amount of data to gain access to knowledge in the various databases [12]. Figure 1 shows the KDD methodology.

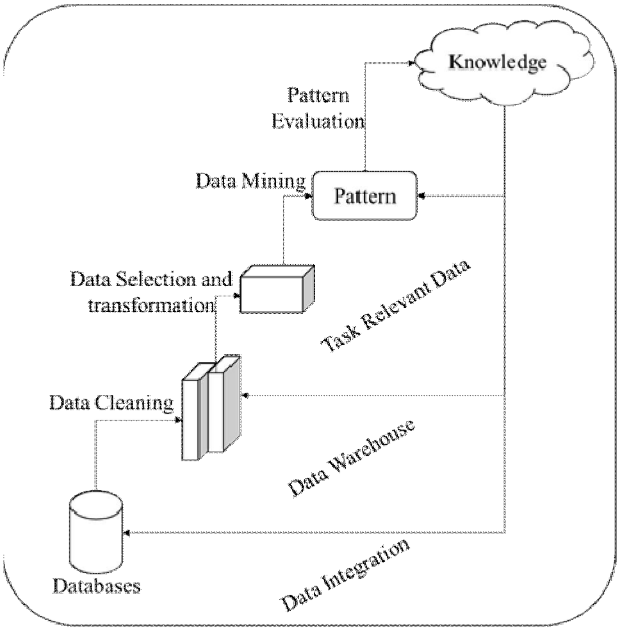

Figure1: Knowledge Discovery in Database (KDD)

As displayed in Figure 1, several steps are used in the KDD process which is used as a basic algorithm to extract trends, patterns, and correlations, and it is also a process of discovering implicit knowledge from a data. Data integration is a process in which similar and related data is collected from multiple data sources and combined. Data integration is applied to provide accurate data and to identify heterogeneous data. Next, data cleaning is performed, and at this point, annoying Noise data that is of no importance is removed, as well as conflicting data and inconsistent data are deleted. Then, the data selection process through which the appropriate data is selected and retrieved from the data set. Subsequent, Data conversion applies to convert data into custom forms suitable for search and retrieval procedures by way of completion summary or collection operations.

It is worth to mention, there are two main steps in the data transformation: Data mapping and code generation. Succeeding, the data mining steps are used to clever methods applied to extract useful patterns as possible. At the pattern evaluation stage, really important patterns that represent the knowledge base for using some important metrics are identified. Finally, knowledge representation step is the stage in which what the human mind wants and which the beneficiary sees is done. Knowledge is understood as the relationships and patterns between data items. This process detects relationships and patterns that were not previously detected between data elements, and these patterns must be clear and useful so that one can use it [13]. The concept of Data Mining is used as a technology for the knowledge discovery stage of the KDD process [14]. This basic stage uses the visual method to assist the beneficiary in understanding and interpreting the results of data mining.

3.1Dataset

There is the possibility of predicting the results of the matches by following the teams and the ability of the players, the team's advantage, its strategies, the coach's management of his team and more factors related to the football game, and the person must follow all these factors to all teams continuously in order to be able to give the correct expectations for the results of the games. Table 1 shows the main factors that use for prediction the football outcome via Bayesian Networks. Similar to [15], this research considered the English Premier League for the seasons of 2010-2011, 2011-2012 and 2012-2013 that was collected from the Football Data UK website at http://www.football-data.co.uk/englandm.php. The league includes 20 teams in which each team plays each other twice in a season (one at home and one away).

Table 1: Dataset for Football Match Outcome Prediction

	Home Team	Away Team	FT HG	FT AG	FT Score
E0	Aston Villa	Bolton	2	2	D
E0	Everton	Man United	0	2	A
E0	Fulham	Birmingham	0	0	D
E0	Man City	West Brom	0	0	D
E0	Middlesboro	Liverpool	0	0	D
E0	Portsmouth	Tottenham	0	2	A
E0	Sunderland	Charlton	1	3	A

3.2 Algorithms

This paper applies three algorithms to assess the accuracy of the match result prediction; the first algorithm is ANN [16], it a supervised learning algorithm, which requires a labelled dataset. After choosing suitable model, the model will be training by supplying the labelled dataset and the model as an input to train the model or to modify the hyperparameters of the model in order to improve the accuracy, time and other parameters. The learned model can be used to predict new input values.

The second algorithm is two-class Support Vector Machines (SVM) [17], is a supervised learning model that looks at data and sorts it into one of the two categories. SVM is based on the idea of creating a hyperplane that divides the data set into two classes in the best way. After defining the model parameters, train the model by using one of the training modules, and providing a tagged dataset that includes a label or outcome column.

The third algorithm is Multiclass Decision Forest which, is one of the most powerful and fully automated machine learning techniques. It almost does not need any data preparation, or any modelling expertise, and enables analysts to obtain effective models. Decision Forest [18] works by creating multiple decision trees and then voting on the most popular output categories. Voting is a form of aggregation, with each tree determining the forest's classification, resulting in a frequency map for unmeasured stickers. The aggregation process groups these graphs and normalizes the results to obtain the "likelihood" of each label. Random forests are very fast and efficiently operate on large databases and can handle thousands of income variables. It is able to deal with unbalanced data containing missing values and maintains accuracy when there is a large percentage of missing data.

3.2 Evaluation Metrics

Following evaluation methodology in various prediction models such as in medical diagnosis [19], sentiment analysis [20], air pollution [21], and stock market prediction [22], the evaluation metrics used in the experiments include the accuracy, precision, and recall.

- Accuracy is the total number of correct predictions divided by the total number of input samples. The formula for calculating accuracy is shown in Eq. 1.

$$\text{Accuracy} = \frac{\text{No. of correct prediction}}{\text{Total no. of predictions made}} \quad (1)$$

- Precision is the number of True Positives divided by the number of True Positives and False Positives. Put another way, it is the number of positive predictions divided by the total number of positive class values predicted. The formula for calculating precision is shown in Eq. 2.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

- The recall is the number of True Positives divided by the number of True Positives and the number of False Negatives. Put another way it is the number of positive predictions divided by the number of positive class values in the test data. The formula for calculation of Recall is shown in Eq. 3.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)$$

4. RESULTS AND DISCUSSION

The purpose of this experiment is to compare the performance of Artificial Neural Network (ANN) and Two-Class Support Vector Machine (SVM) and Multiclass Decision Forest (DF) algorithm in football match outcome prediction dataset. The prediction results are evaluated based on 10-folds cross-validation. Tables 2 represents the overall prediction results.

Table 2:Results of 10-folds Cross-Validation

Metrics	ANN	DF	SVM
Overall accuracy	0.715789	0.889474	0.701111
Avg. accuracy	0.810526	0.926316	0.783172
Overall precision	0.715789	0.889474	0.701111
Avg. precision	0.638775	0.875734	0.619254
Overall recall	0.715789	0.889474	0.701111
Avg. recall	0.645029	0.856140	0.620274

The results showed that Multiclass Decision Forest is more accurate and precision result than Neural Network Algorithm and Two-Class Support Decision Machine. From Table 2, the result shows that the DF has the highest accuracy (89%), followed by the ANN (72%) and SVM (70%). In terms of precision and recall, the highest percentage still belongs to the DF (89%), followed by ANN (72%) and SVM (70%). Overall, DF works very well with the football dataset, highly likely due to the type of features in the dataset, which are in discrete numbers. Meanwhile, ANN and SVM models are more suitable with continuous values.

5. CONCLUSIONS

Our main objective of building an expected goals model by exploring different Machine Learning (ML) techniques has been accomplished. This paper presented an analysis for football prediction in 2005/2006 season of EPL based on three optimization algorithms, which are Artificial Neural Networks, Support Vector Machine, and Decision Forest. The performance of these algorithms is evaluated by calculating three popular measurement terms; accuracy, precision, and recall. After evaluating all measures, decision forest algorithm as the best solution for the football prediction result. This algorithm excelled in prediction performance and robustness and exhibited faster calculation time compared to SVM and ANN. Prediction calculation based on 10-folds cross-validation. The result shows that DF has achieved the highest accuracy of 89%, the ANN achieves the second-highest accuracy of 72% and the SVM achieves the lowest accuracy of 70%.

Future research should focus on improving the decision forest algorithm to increase its prediction accuracy. Although the results showed considerably high accuracy in terms of predicting the match outcome, accuracy alone does not provide any insights regarding strategies to win the game due to limited features in the dataset. For instance, one weakness of the home goals and away goals is the data does not give information on the position of the opposing team's players at the time of the shot. This is important because having a player between the ball and the goal will dramatically reduce the probability of the shot resulting in a goal. In addition, there is no information regarding the type of passes made in the game. Having passes, data would allow the model to better estimate match outcome in a game.

ACKNOWLEDGEMENT

This paper is supported by Research Fund E15501, Research Management Centre, UniversitiTun Hussein Onn Malaysia.

REFERENCES

1. B. Gianluca, and M. Blangiardo. **Bayesian hierarchical model for the prediction of football results**, *Journal of Applied Statistics,*vol. 37, no. 2, 253-264, 2010.
2. N. Razali, A. Mustapha, F. M. Clemente, M. F. Ahmad, M. A. Salamat. **Pattern Analysis of Goals Scored in Malaysia Super League 2015**, *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 2, pp. 718-724, 2018.
3. A. A. Ogunseye, O. Balogun, A. A. Ogunseye, F. S. P. Global. **Artificial Neural Network Approach to Football Score Prediction**,*Journal of Artificial Intelligence*, vol. 1, no. 1, 2019.
4. S. Johannes, B. Mangold, and J. Knoll. **Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics**, *Applied Sciences,*vol. 10, no. 1, pp. 46, 2020.
5. R. Ashiqur. **A deep learning framework for football match prediction**, *SN Applied Sciences*, vol. 2, no. 2, pp. 165, 2020.
6. R. T. Stefani.**Football and basketball predictions using least squares**, *IEEE Transactions on Systems, Man, and Cybernetics,*no. 7, vol. 2, pp. 117-21, 1977.
7. R. Havard, and O. Salvesen. **Prediction and retrospective analysis of soccer matches in a league**, *Journal of the Royal Statistical Society: Series D (The Statistician),*vol. 49, no. 3, pp. 399-418, 2000.
8. B. Min, J. Kim, C. Choe, H. Eom, R. B. McKay. **A compound framework for sports results prediction: A football case study**, *Knowledge-Based Systems,*vol. 21, no. 7, pp. 551-562, 2008.
9. C. Anthony, N. Fenton, and M. Neil. **Pi-football: A Bayesian network model for forecasting Association Football match outcomes**, *Knowledge-Based Systems,*vol. 36, pp. 322-339, 2012.
10. K. S. Jan, and R. Lit.**A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League**, *Journal of the Royal Statistical Society: Series A (Statistics in Society),*vol. 178, no. 1, pp. 167-186, 2015.
11. M. Michael. **Modelling association football scores**, *Statistica Neerlandica,*vol. 36, no. 3, pp. 109-118, 1982.
12. G. Michael, and L. Gruenwald.**A survey of data mining and knowledge discovery software tools**, *ACM SIGKDD explorations newsletter*, vol. 1, no. 1, pp. 20-33, 1999.
13. G. M. Monir, and A. Hammad.**Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects**, *International Journal of Construction Management*, pp. 1-15, 2020.

14. U. Muzaffer.**Advancement in computing: Implications for tourism and hospitality**, *Scandinavian Journal of Hospitality and Tourism*, vol. 4, no. 3, pp. 208-224, 2004.
15. N. Razali, A. Mustapha, F. A. Yatim, and R. Aziz. **Predicting football matches results using Bayesian networks for English premier league (EPL)**, *IOP conference series: Materials Science and Engineering*, vol. 226, no. 1, 2017.
16. M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. Abd Ghani, S. A. Mostafa. **Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images**, *Computers & Electrical Engineering*, vol. 70,871-882, 2018.
17. B. J. Chelliah, S. Kalaiarasi, A. Anand, G. Janakiram, B.Rathiand N. K. Warrier.**Classification of Mushrooms using Supervised Learning Models**. International Journal of Emerging Technologies in Engineering Research (IJETER), 6(4),2018.
18. A. B. Annasaheb and V. K. Verma.**Data mining classification techniques: A recent survey**. International Journal of Emerging Technologies in Engineering Research, 4(8), pp 51-54,2016.
19. S. A. Mostafa, A. Mustapha, S. H. Khaleefah, M. S. Ahmad, and M. A. Mohammed. **Evaluating the performance of three classification methods in diagnosis of Parkinson's disease**, in *Proc. of the International Conference on Soft Computing and Data Mining*, 2018, Springer, Cham.
20. I. S. Makki and F. Alqurashi.**An Adaptive Model for Knowledge Mining in Databases EMO_MINE for Tweets Emotions Classification**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 7, no. 3, pp. 52–60, 2018. https://doi.org/10.30534/ijatcse/2018/04732018
21. M. H. Hamid Hassan, S. A. Mostafa, A. Mustapha, M. Z. Saringat, R. Darman, and M. A. Jubair. **A Statistical Risk Assessment Method of Dynamic Environments: A Case Study of Air Pollution**, *AUS Journal*, 2019.