Journal of Soft Computing and Artificial Intelligence
JSCAI, 1(2): pp:86-91 (2020), www.dergipark.org.tr/en/pub/jscai
JSCAI
ISSN: 2717-8226
Received: October 27 2020 Accepted: November 19 2020

Applying Decision Tree Techniques to Classify European Football Teams

[1]Bünyamin Fuat Yıldız

[1]M.Sc. (Econ.), 24 Crooks Ave Apt 229, Clifton, NJ 7011-1614 United States

Corresponding author: Bünyamin Fuat Yıldız e-mail: bunyaminfuatyildiz@yahoo.com

ABSTRACT Machine learning techniques are powerful tools used in all aspects of science. However, these techniques are relatively new in sports. This study was carried out to measure the accuracy of decision trees in the classification of football teams. We applied five types of decision tree algorithms to classify elite football teams in Spain, Italy, and England to determine whether decision tree techniques are robust in classifying elite football teams. The findings show that the accuracy rate is above 77 percent for each of the decision trees. The key qualities that cause branching in decision trees may constitute a criterion for the targeting of football authorities. More research is required to determine which machine learning techniques are more efficient in classifying football teams.

KEYWORDS: Machine Learning, Decision Trees, Football, Classification, Sports.

1. INTRODUCTION

The role of machine learning has received increased attention across several disciplines in the last three decades. Recently scholars have applied machine learning techniques in several sports from ice hockey [1,2] to basketball [3-6]. There is a considerable amount of literature in the field of baseball employing machine learning techniques [7-10]. Besides, machine learning methods have begun to examine the issue of football from various aspects [11-16].

Since the beginning of human existence, people have classified something. When people classify things, they arranged and defined them based on some parameters that they have in general. In the same way, the machines classify the data according to their characteristics. Also, it allows researchers to understand certain qualities and differences in the subject area of interest. Likewise, football clubs have certain characteristics. It is among the aims of football authorities to determine the characteristics of successful clubs and to set goals in this direction.

Empirical studies with the appropriate tools to guide soccer teams are needed to identify the characteristics of successful clubs. In this way, clubs can set new targets according to the performance criteria that constitute key distinctions through decision trees. The main reason for this study is therefore established. The other contribution of this study that there are no previous works done in this context. Accordingly, there might be new applications to extends the current literature further. Taken together, this study assesses the performance of decision tree techniques to classify European football teams in Serie A, La Liga, and Premier League—which is the first extensive examination that provides new insights into literature. We divided football teams into three groups: a) the top tiers of the leagues which finished the season pretty well to qualify European Championship League and Euro League, b) the

teams which are below top-tiers and above-average denoted as top-half, c) Teams that finish the leagues in 11th place and below. The reasoning for this type of categorization is that we perceive several commonalities in football teams in preliminary review which is also supported by Table 1. This paper uses 10-year league data consist of 600 observations. The remaining part of the work proceeds as follows. Section 2 will give information both about methodologies and the sources of data. The third section presents the preliminary statistics regarding each class of teams and the performance of each decision tree algorithms. The final section provides a summary and recommendations for further research.

2. METHOD AND DATA

2.1. Decision Trees

Classification procedure of data by machine learning divided mainly into the two-stage process. [17]. At the primary stage, the learning process constructs a model from the received knowledge. If the employer provides classification information in the learning stage, it is called supervised learning; contrarily, it is called unsupervised learning—in which classes are derived from a dataset without prior classification information [18]. A distinct advantage of using decision trees (DT) is that provides the availability to the employer to conduct both supervised and unsupervised learning. Therefore, they are commonly used for knowledge discovery [19].

The process of discovering general rules starts with DT from tuples contain classes [20]. Figure 1 displays an illustration of DT. It is clear from the top of Figure 1 that the root is a unique node— has no precursor. The remaining nodes in the DT possess precisely one precursor. There are two forms of nodes to identify: the first one is the leaves, in other sayings, the terminal nodes have no descendants; the second one is internal nodes that have more than one descendant. The test results must be equivalent to the total quantity of branches emanate from that node due to any test result associated with a single branch [20]. Lastly, the leaves contain results (i.e. classes, numbers).

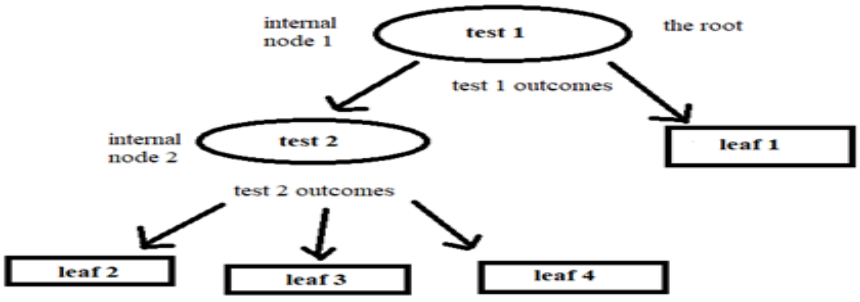

Figure 1. An Illustration of Decision Trees. Adapted from "Evolutionary Decision Trees in Large-Scale Data Mining," by [20].

Decision trees obtained through the R Software. We use the RWeka package to implement the C4.5 algorithm and Logistic Model Trees (LMT), which respectively use gain ratio and logistic regressions as division rules [21]. [22] provides detailed background information regarding the C4.5 algorithm; whereas [23] demonstrates the essentials of LMT. Cart and rpart packages are used to obtain results for the CART algorithm [24]. This simple but effective method is based on binary division by using the Gini index while creating decision rules. The random forests (RF) are another classification algorithm used in this study. Unlike the previously mentioned algorithms, the inference of overall error is proportional to the strength of each tree. It works by bringing together the predictions produced by bootstrapping a large number of independent decision trees [25]. Our reference guide for RF application is the randomForest package that [26] has introduced into the literature. The last empirical approach is regression trees based on binary division, which is similar to the Classification and Regression Trees (CART) algorithm. The classification process is based on the squared error difference between real and estimated values for the assigned dependent variable [27]. In this context, the tree library created by [28] was used for the regression tree classification process—which is mainly based on the squared error differences between real and estimated values for the assigned dependent variable.

2.2. Dataset

The data set used in the study consists of observations from Serie A, Premier League, and La Liga covering the 2009-2010 and 2018-2019 seasons obtained from the whoscored.com. There are 200 observations from each league. The separation of the data as a training dataset and test dataset was created randomly. The training data set represents 70 percent of the data. The test dataset constitutes 30 percent of the data. The classifications in the test data set are estimated by the rules learned from the training data.

Our dependent variable is denoted as "class". Teams under consideration are labeled as "top-half" in case that they completed the league in the top 10 but not well enough to participate in European Cups. The team under consideration is labeled as "top-half", in case that they finished the league in the top 10 but not well enough to participate in European Cups. The team that finished the league lower than 10th place is classified as "below-average". The number of independent variables is four. The first one is gperg which indicates the goal scored per game. The second variable is denoted as pass, reflects the successful pass percentage of the teams. The variable denoted as 'poss' shows the average ball possession percentage of football teams. Eventually, the gapg presents the number of goals conceded per game.

3. EMPIRICAL RESULTS

Table 1 below illustrates the mean value of subjected variables for each class of teams.

What stands out in the table is that the qualified teams have scored 1.86 goals per game; while top-half and below-average teams scored 1.33 and 1.07 goals respectively. The possession rate of football teams is provided in the second column, which shows qualified teams' highest mean with 54 percent. The football teams from other classes have possession of less than 50 percent. The qualified teams also

have the highest pass accuracy, which is above 80 percent, whereas the remaining teams have success below them. One notable difference between below-average football teams with the remaining classes is that they conceded a 1.57 goal per game.

Table 1. Preliminary statistics for each class of teams

Class	gperg	poss	pass	gapg
qualified	1.87	0.54	0.82	1.01
top-half	1.33	0.50	0.78	1.30
bel_ave	1.07	0.47	0.75	1.57

Turning now to the performance of each DT algorithms, Table 2 provides the accuracy rates which compares the predictions with the actual values. The calculated 95 percent confidence interval range and p-values respectively provided in the second and third row. In the case of overall accuracy rates, the random forest (RF) has the highest success in general with 79 percent. The C4.5 and CART algorithms, which have the lowest success, have 77 percent. However, if we compare the accuracy rates for qualified football teams, LMT has the highest percentage of accuracy. Unfortunately, there are disturbing results in the diagnosis of the top-half teams. The accuracy rate of All DT algorithms is below 70 percent. The most likely cause of the low accuracy rate in the top-half football teams is the need for additional data to separate them from the other classes.

Table 2. Accuracy of Decision Tree Algorithms

	C4.5	CART	LMT	Ran_Forest	Reg_Trees
Accuracy	0.77	0.77	0.78	0.79	0.78
95% CI	(0.70,0.83)	(0.70,0.83)	(0.72-0.84)	(0.72, 0.85)	(0.72,0.84)
P-Value	1.529e-13	1.529e-13	3.551e-15	9.495e-16	3.551e-15
Acc. qualified	0.84	0.86	0.88	0.85	0.86
Acc. top-half	0.59	0.62	0.65	0.68	0.63
Acc. bel_ave	0.87	0.85	0.85	0.87	0.87

4. CONCLUSION

The classification performances of the C4.5, CART, LMT, RF, and Regression Trees algorithms were evaluated for European football teams. The data set randomly separated into two subgroups. 70 percent of the data were employed to training algorithms while the remaining 30 percent used for testing. We used accuracy as criteria for the performance evaluation of the algorithms. The results of this study have shown that decision trees have a good performance in classifying football clubs. The performance of RF is the most successful based on the accuracy criterion, the rest of the DT algorithms

have also achieved more than 77 percent accuracy. The football officials can detect and progress towards the key branching factors aroused from DTs — which differentiate qualifying football teams from the rest. Nonetheless, due to the page limitations, we are unable to include all the visuals of each decision tree. We can share datasets and codes for those who are further interested.

Notwithstanding the strong accuracy rates of DT, several issues remain. It is recommended that further research be undertaken in the following ways. A further study with more focus on other machine learning techniques, such as locally weighted naïve Bayes or OneR, would be recommended. Besides, the researchers who want can obtain different results by modifying the algorithms used in this study or by changing the selection procedure of the training data. For instance, for the CART algorithm, we determined the smallest divisional value as 4. Researchers can change this value or use 80 percent of the current data set for training and 20 percent for testing. Moreover, it would be interesting if the DT techniques should also be evaluated whether there will be effective tools for football betting.

Conclusively, it is essential to participate in tournaments organized by the Union of European Football Associations (UEFA) for all football teams in the European continent. Football clubs that participate in the European Cups prosper financially by earning millions of euros in income. Therefore, football attracts the attention of many disciplines from sociology to economics due to the high impact it generated. Empirical studies made due to the teams' desire to win increase their importance. Football clubs that do not close themselves to innovations and progress in the light of science will be successful. Concerning the consequences of the machine learning techniques for football fans is that there might be a convergence of quality between football clubs that will draw more audience.

REFERENCES

[1] Mulholland J, Jensen ST. Predicting the draft and career success of tight ends in the National Football League.Journal of Quantitative Analysis in Sports. 2014;10(4):381–96. https://doi.org/10.1515/jqas-2013-0134.

[2] Joash Fernandes, C., Yakubov, R., Li, Y., Prasad, A. K., & Chan, T. C. (2019). Predicting plays in the National Football League. Journal of Sports Analytics, 1-9.

[3] Lorenzo Calvo J, Menéndez García A, Navandar A. Analysis of mismatch after ball screens in Spanish professional basketball. International Journal of Performance Analysis in Sport. 2017;17(4):555–62. https://doi.org/10.1080/24748668.2017.1367999.

[4] Leicht AS, Gomez MA, Woods CT. Team performance indicators explain outcome during women's basketball matches at the Olympic Games. Sports. 2017 Dec;5(4):96. https://doi.org/10.3390/sports5040096 PMID:29910456

[5] Cene E, Parim C, Özkan B. Comparing the performance of basketball players with decision trees and TOPSIS. Data Science and Applications. 2018;1(1):21–8.

[6] Horvat T, Havaš L, Srpak D. The Impact of Selecting a Validation Method in Machine Learning on Predicting Basketball Game Outcomes. Symmetry. 2020;12(3):431.

[7] Freiman MH. Using random forests and simulated annealing to predict probabilities of election to the Baseball Hall of Fame. Journal of Quantitative Analysis in Sports. 2010;6(2). https://doi.org/10.2202/1559-0410.1245.

[8] Oh Y, Kim H, Yun J, Lee JS. Using Data Mining Techniques to Predict Win-Loss in Korean Professional Baseball Games. Journal of Korean Institute of Industrial Engineers. 2014;40(1):8–17. https://doi.org/10.7232/JKIIE.2014.40.1.008.

[9] Tolbert B, Trafalis T. Predicting major league baseball championship winners through data mining. Athens Journal of Sports. 2016;3(4):239–52. https://doi.org/10.30958/ajspo.3.4.1

[10] Koseler K, Stephan M. Machine learning applications in baseball: A systematic literature review.Applied Artificial Intelligence. 2017;31(9-10):745–63. https://doi.org/10.1080/08839514.2018.1442991

[11] Myers BR. A proposed decision rule for the timing of soccer substitutions. Journal of Quantitative Analysis in Sports. 2012;8(1). https://doi.org/10.1515/1559-0410.1349.

[12] Folgado H, Duarte R, Marques P, Sampaio J. The effects of congested fixtures period on tactical and physical performance in elite football. Journal of sports sciences. 2015;33(12):1238–47.

[13] Schauberger G, Groll A. Predicting matches in international football tournaments with random forests. Statistical Modelling. 2018;18(5-6):460–82. https://doi.org/10.1177/1471082X18799934.

[14] Young CM, Luo W, Gastin P, Tran J, Dwyer DB. The relationship between match performance indicators and outcome in Australian Football. Journal of Science and Medicine Sport. 2019;22(4):467-71 https://doi.org/10.1016/j.jsams.2018.09.235 PMID:30352743

[15] Baboota R, Kaur H. Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. 2019;35(2):741-55.

[16] Min DK. Contribution analysis of scoring in the soccer game: using decision tree. The Korean Data & Information Science Society. 2019;30(6):1385–97. https://doi.org/10.7465/jkdi.2019.30.6.1385

[17] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011

[18] Michie D, Spiegelhalter DJ, Taylor CC. Machine learning. Neural and Statistical Classification. 1994;13:1–298.

[19] Rokach L, Maimon OZ. Data mining with decision trees: theory and applications. World scientific; 2008.

[20] Kretowski M. Evolutionary Decision Trees in Large-Scale Data Mining. Springer International Publishing; 2019. https://doi.org/10.1007/978-3-030-21851-5

[21] Hornik K, Buchta C, Zeileis A. Open-source machine learning: R meets Weka. Computational Statistics. 2009;24(2):225–32.

[22] Quinlan JR. Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research.1996;4:77–90. https://doi.org/10.1613/jair.279

[23] Ripley B. (2019). tree: Classification and Regression Trees. R package version 1.0-40. https://CRAN.R-project.org/package=tree

[24] Therneau T, Atkinson B, Ripley B. (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

[25] Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

[26] Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.

[27] Loh WY. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011;1(1):14–23. https://doi.org/10.1002/widm.8.

[28] Ripley B. (2019). tree: Classification and Regression Trees. R package version 1.0-40. https://CRAN.R-project.org/package=tree