Application of Machine Learning Algorithms for Automatic Classification of Problems Football*

Adão Baptista Pereira Lopes

IT Department, School of Science and Technology, University of Évora
Rua Romão Ramalho, 59, 7000-671 Évora, Portugal
abpl@uevora.pt

Abstract. Nowadays, companies or organisations, including those in sport area, show a great demand for data analysis. There is a great deal of interest in forecasting the results of the various types of sport, especially football, where we can find various types of forecasting. Since the forecast results, the number of goals, prediction if both teams score goals or not, the objective is to reach the correct forecast. It is exactly on the slope «both teams score goal or not» that focuses on this project, looking through the application various Machine Learning (ML) techniques, to promote the help of forecast and allow to verify if both teams score goals or not. While it is true that the area of technology has evolved in a fast way and constant, there is a massive growth of data in the sport, so the project intends to analyse the Portuguese championship data (LIGA NOS) from the 1994/1995 to the present time. Similarly, it aims to present several classification algorithms such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Linear Regression (LR), and demonstrate what best fits in this case study, through the evaluation methods of each of the models. Benchmarking prediction performance of the study problem will be based on the Confusion matrix, in micro-average, macro-average and F1 Score.

Keywords: Machine Learning · Automatic Classification· Football· k-Nearest Neighbor · Support Vector Machine

1 Introduction

In Sport, football is one of the most exciting, which makes thousands of fans around the world happy. One of the reasons associated with this manifest collective interest, is related to the fact that it is an unpredictable game, with countless variables that can decide a game regardless of the teams economic power.

One of the recent facts that proves this, is the occurred in the period 2015/2016 where against all possibilities or prediction Leicester City, a team that had recently been promoted to Premier League, was the winner of the English championship, where it is known by the existence of several giants of European and world football.

* This work was developed within the scope of the Discipline (Automatic Classification and Core Methods)

Lately, there are several projects [12][26][22] that resort to the help of ML techniques, to find patterns that can help through data analysis to predict the Game results. In this article, we try to use several techniques of Machine Learning, for implementation of model classification in a football match of the Portuguese championship (Liga NOS). The central question of this study is the application of this classification model to predict the possibility for both teams to score goals.

2 State of Arts

The learning process has been fascinating psychologists, philosophers, and scientists from all areas of knowledge, which makes it difficult to define this concept. In the computational area, the great motivation is to put the computer at such level that it can think like a human brain[20]. Artificial Intelligence (AI) thus appears as an area of computing study, which aims to build devices and programs that simulate the human capacity to think, make prior decisions and solve certain problems[29]. The ML study field consists of having computer programs which can mimic human learning behavior[20]. ML may be understood as a subarea of the AI, whose purpose is to study systems, and retain as much learning as possible, or learn from data[7].

Related Works Nowadays there are several projects and researches in the scope of using AI, properly ML, to predict game data from football. There is a vast literature to cover this problem, part of the literature found, the forecasting strategies are based on models of RL, KNN and SMV[10].

A research was carried out, using the logistic regression model, predicting the results of the Barclays Premier League games of 2015/2016, it has been concluded that there are several significant variables in the result forecast, with emphasis on the defensive power of both teams, the visiting team and the team visited[22]. Choosing only the most significant variables, can maximise the forecast accuracy by 18%.

FIFA 2015 and the European Championship, served as a knowledge base, for research on the soccer results forecast[26]. In this project, real data were collected as well as virtual data of the players at the physical level (Acceleration, Strength, Speed, etc.), in addition to data related to the soccer technique (Dribbles, Pass Precision, etc.), of the players in the game of PlayStation (FIFA 2015). The collection of virtual data was justified as a form of time saving precision, and to allow comparison with actual data. This strengthened the data set to increase the algorithm[26]. The algorithms applied in this project are: Linear Regression, SVM, RBF, and Logistic regression.

The Logistic Regression applied to a Premier League data set Season 2014-2015, rated 9 features (Home and Guest, Short goals, Odds, Attack Strength, Player Performance Index, Directors Performance Index and team victories), generating an accuracy of about 95%[12], therefore quite high. It is important to

note that even high precision for predicting football data, is not reliable, derived from the fact that a football game depends on many factors not to be missed.

However, none of the projects addresses the prediction problem if both teams score goals or not. On the other hand, most of the projects of this nature, are more concerned with predicting the exact outcome of the game[25], but others address the study of goal differences in a game[18][28][14]. Projects of this nature are the ones that most resemble the central issue of this study.

Machine Learning (ML) Can be understood as the study of algorithms, with the ability to improve their performance in a situation based on predecessor experience. This area is strongly related to the recognition standards and statistics. The purpose of ML is to develop computational techniques that investigates the simulation of the human learning process, and to build systems capable of acquiring automatic knowledge[6].

Commonly, ML algorithms deal with previous experiences. The University of California has created a repository centre [1], with sets of data of various natures available. There are many other available data repositories, KDnuggets[2] is used as an example. Before predicting or classify something, it is necessary to train the algorithm before. And for this it is necessary to choose the most relevant characteristics for the case under study. Selecting the most relevant features, and eliminating the irrelevant ones, is to improve ML[16].

Supervised and Unsupervised Algorithms There are two learning techniques methods commonly used in ML: (I) through supervised algorithms, in which the inputs and outputs in the learning set is known. In the course of learning, the model tune the variables, in order to map the corresponding inputs and outputs[19].The crucial point of this algorithm is the ability to make predictions with a high level of performance; is to construct a classifier that, in a correct way, can constitute a class for new examples. In (ii) unsupervised algorithms, the program analyses the data, verifies which can be grouped. There is no target result[19].

3 Classification Algorithms

For the development of this study, the framework used is Scikit-Learn (Python)[3]. It is a Python module that integrates an immense range of algorithms of ML for both supervised and unsupervised problems. This package focuses on bringing the ML to experts, using a high level language (Python). It provides easiness of usability, a high performance, has a lot of official documentation and it is consistent[21]. They are listed followed the algorithms / classification techniques:

[1] $http://archive.ics.uci.edu/ml/datasets.html$
[2] $https://www.kdnuggets.com/datasets/index.html$
[3] $http://scikit-learn.sourceforge.net.$

Support Vector Machine (SVM) Is known to be a classification technique very powerful, especially for data with bulky dimensions[3]. The basic idea can be emphasised geometrically. If the data is in a space, the algorithm finds the hyper plane that separates the data with the feasible margin. With this hyper plane, it is presumed to classify the data[8].

In other words, the algorithm separates the data points using a line if it is two dimensions, in case for three dimensions it uses a hyper plane. This line is chosen in such a way that it will be more important than the data into two categories. SVM is a classification technique, which seeks to find a model where the separation between classes has a bigger possible margin.

The SVM were originally designed for binary classification[2]. This technique is applied to classify problems of different natures, such as Economy[11], Sports (Football)[1], Medicine[8][9], where it is of good use to detect the patterns of various diseases, especially cancer.

Decision Tree (DT) The data structure is composed of a set of elements, which stores information in nodes, which are called Trees in computer science. It has a main node that is usually called root.

$$Entropy = \sum_{i=1}^{k} P_i log_2 P_i \tag{1}$$

At the hierarchical level it is the largest node, and the connections from the root node, are designated of children. These son nodes, can have their own children, and so on. If the node does not have any children it is termed as a leaf node or terminal node.

By knowing these definitions, it becomes easier to understand what a TD is. It is a tree that has rule in its nodes, and the decision process is represented by the leaf of the tree[30]. Summarising, in a TD the decision is a path travelled from the root node to a leaf node.

Generally, it is one of the most commonly used algorithms for solving Ranking problems. Categorisation is done using some techniques. The Formula 2[27]presents the Gini technique, and Formula1[27] presents the Entropy. To problem of this study were used these two categories. The same result, both in the accuracy of the model, as in the confusion matrix, and in the other evaluation model.

$$Gini = 1 - \sum_{i=1}^{k} P_i^2 \tag{2}$$

Linear Regression (LN) Linear regression is another algorithm for classification that tries to find a straight line that traverses a scattered chart of points. This line will be as close as possible to all points, or, find the best line. This line is called the regression line[23].

k-Nearest Neighbours (KNN) K-Nearest Neighbours (KNN) This technique is a method for classifying which is based on the closest learning examples in the attributes dimension. The training examples are placed in the multidimensional attribute space multidimensional[5].

The most important parameter in this technique is K. It defines how many units of distance the elements may be to be considered neighbours[17]. The KNN is a supervised learning algorithm. The ecumenical idea of this algorithm is to find the k examples labelled closest to the unclassified example[6].

Briefly, KNN is a simple algorithm that predicts unknown data points, based on the proximity of its neighbours. The value of k is a critical factor, because the accuracy of the forecast depends on it. By way of example we see that the classification is totally different, depending on the K for 3 or is 20. To determine the calculation of the nearest distance of a point, this algorithm uses the basic distance functions of Euclidean, represented by the Formula3, which is one of several possibilities for calculating distances.

It is advisable, when deciding the best value of K, to make a comparison of different K's values by performing several tests to find the ideal K. The value of K to choose, is the one that has the highest hit rate. In the training phase, the KNN algorithm requires little effort. On the other hand, the cost of marking a new unclassified example is a little high. At worst, case this new example has to be compared with all the data contained in the training set[6].

The KNN algorithm uses instance-based learning. This means, it uses the training data set to sort out the unknown data points. Although KNN is a classification algorithm, it is largely used to predict and make estimates. Taking into historical values account, the ideal values of K in most cases are in the range of 3 to 10[23].

$$Euclidean = \sqrt{\sum_{i=1}^{K}(x_i - y_i)^2} \tag{3}$$

In this way, we can observe that there is not only a single appropriate value of K[6]. The choice of this value will depend very much on the nature of the study. It is also very important to emphasise that for the same case study, the best value of K depends very much on the numbers of the chosen characteristics, and its relevance to the case under study.

Using odd values of K is more appropriate, using it in case of the majority of class to avoid tie situations[6]. Alternatively, there is another approach to solve this problem. Consists of assigning weights to each of the nearest K neighbours. They are sorted in ascending order, for the determination of the classification, and the class of the most similar examples has a greater weight, than classes with little similarity[6].

This classification method is applied to problems of different nature, for example in Medicine[15][4] for the identification of risk factors in the prostate cancer, based on clinical and demographic variables; in Agriculture[24] for predicting the climate, estimating soil water parameters, to simulate precipitation

and other meteorological variables; in the Financial area[31][13] for the stock market, which includes predicting the discovery of market tendencies , plan investment strategies, identifying the best actions. It can still be added, exchange rate, banks bankruptcy, credit management classification, loan management etc. "[6].

4 Dataset

Figure 1 below is intended to represent the data set of the Portuguese championship from the 1994/1995 season to the current season in 2018/2019. In total there is an average of 300 games per year, being that in the 21 season, one reaches an accumulated of about 7000 games. At the data set from the year 1994 to 2000 we found only 4 characteristics (Home Team (HT); Away Team (AT); Goals of the home team (FTHG), and Away Team Goals (FTAH).

From the 2000/2001 season until 2016/2017, there are two more characteristics which are the goals in the break of home team and the visiting team (HTHG, HTAG). In the last two seasons, it has been observed that the data set is more complete, encompassing all the game statistics, such as corners, yellow cards, fouls committed, shots, among other characteristics. The data set was developed from the available data repository in Football Betting, Scores and Results Service (FBSRS)[4].

Div	Date	HomeTeam	AwayTeam	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	Ambas Marcam
P1	06/08/17	Aves	Sp Lisbon	0	2 A	0	1 A		12	12	
P1	06/08/17	Setubal	Moreirense	1	1 D	1	0 H		6	12	
P1	07/08/17	Feirense	Tondela	1	1 D	0	1 A		12	13	
P1	07/08/17	Portimonense	Boavista	2	1 H	0	1 A		12	5	
P1	07/08/17	Rio Ave	Belenenses	1	0 H	1	0 H		11	10	
P1	08/08/17	Maritimo	Pacos Ferreira	1	0 H	0	0 D		7	4	
P1	09/08/17	Benfica	Sp Braga	3	1 H	2	1 H		15	6	

Fig. 1. The Set of Data (Portuguese Championship).

From the DCP[5] data set obtained from the FBSRS, were created some variables, with the purpose of enriching the statistical knowledge base, in order to allow the data to change dynamically each game. Initially, were discarded several attributes that do not present correlation level with other relevant characteristics to the study in question. A title for example we can mention, the odds of the different betting houses that were ignored, and other attributes that do not have data for all games, for example red and yellow card.

In total, this approach will use six variables (HT), Away team (AT), Home team attack power (HTAP), Away Team Goal (ATAP), Home Team average home goals (HTAHG) and Away Team Average Away Goals (ATAAG).

[4] $http://www.football-data.co.uk/portugalm.php$
[5] Portuguese Championship Dataset

The first two variables (HT and AT), are obtained from the set of DCP data. It is of enormous importance, having as a source the past data for classification[18]. The HTAP and ATAP are obtained through Sofifa[6]. Are range of fixed values from 0 to 100, which is assigned to each team according to the present season 2018/2019. The teams that are not in the NOS league is assigned the value 50.

HTAHG and ATAAG have an extremely important detail. It benefits the home team, since in football, the home team has a better average goal statistic. This can be justified, because the home team has greater support from the fans, which can lead to player's concentration, the mentality of facing a home game, factors that conjugates imply that the teams have a better result at home, with the rare exception of some cases. The average goal difference between Benfica vs Porto, is different from the goal average between Porto and Benfica. In this case, HTAGH (Benfica), is calculated from all matches played by Benfica at home against Porto, and goals scored, where the total number of goals is divided by total number of games between them.

5 Classifiers Evaluation

In order to know how good the classification model is, it is essential to calculate accuracy of the test data forecast. In addition to precision there are other methods to evaluate how well a model behaves, using the confusion matrix (MC) and classification report (CR) (Micro_Average, Macro_Average, F1-Score).

To calculate the evaluation methods (Recall (4a), Accuracy (4b) and Precision (4c) of classifiers the following formulas are used:

$$Recall = \frac{TP}{TP + FN}, Accuracy = \frac{TP + TN}{Total}, Precision = \frac{TP}{TP + FP}, \quad (4)$$

The first quadrant of MC belongs to True Positive (TP), the second quadrant is False Positive (FP), the third False Negative quadrant (FN), and the last quadrant is True Negative (TN).

Parameters and Precision - (SVM) SVM The accuracy of this algorithm depends on the Parameters values and their combinations. The big question is to find the best hyper plane in the separation of data sets. Accuracy depends on the C parameters, Gamma, and the Kernel. The analysis of these values in this case of study, a range of values was created for each parameter getting following conclusion:

- Using the interval [0.001, 0.01, 0.1, 1, 10, 100] for parameter C, the best value is 0.01;
- Use the range [0.0001, 0.001, 0.01, 0.1] for the gamma parameter, the best value is 0.0001;
- For the kernel parameter, linear is better than rbf.

[6] $https://sofifa.com/teams?lg=308\&col=sa\&sort=asc\&showCol=ta$

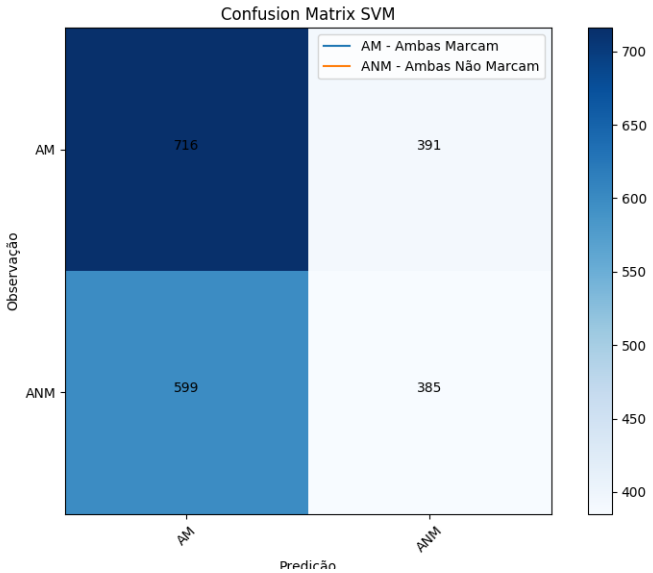

Fig. 2. Confusion Matrix - SVM

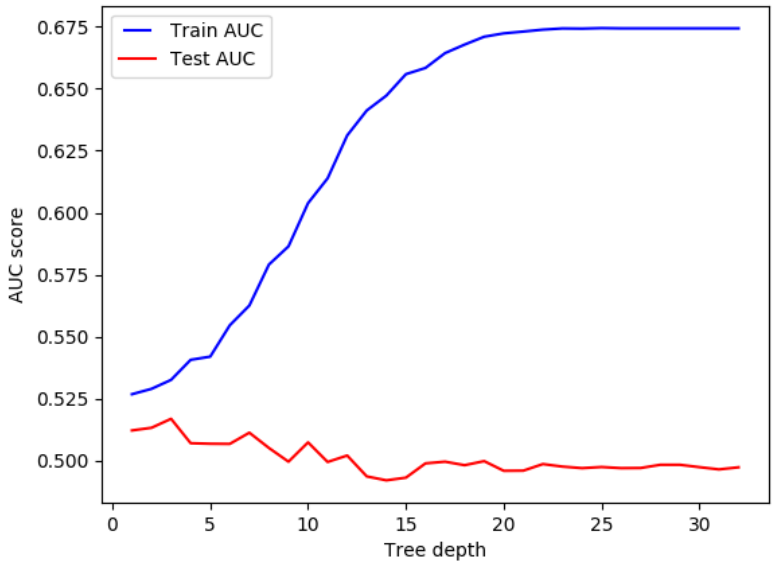

Fig. 3. Maximum Depth and Minimum Sample Sheet

Accuracy and Parameters -(DT) The first parameter to adjust is the maximum depth (max_depth). This indicates how deep the tree can be. The deeper the tree, the more divisions it has and it captures more data information. The tree was set at a depth ranging from 1 to 32, and the training is performed and tested, which depth has the best score Area Under the Curve (AUC). According to Figure 3 on the left, we can see that the depth that had the best AUC score is 3.

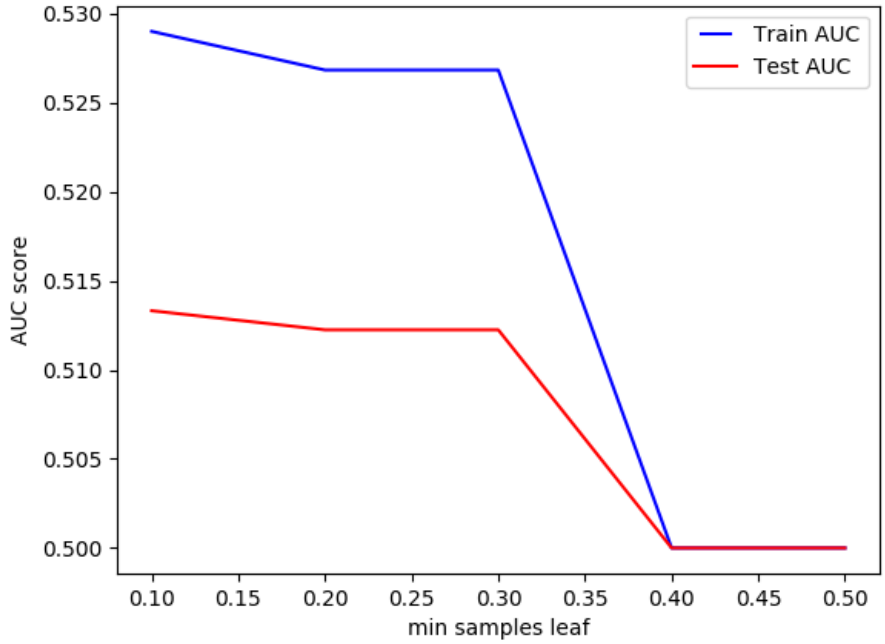

Fig. 4. Maximum Depth and Minimum Sample Sheet

The second parameter is min_samples_leaf. It is the minimum number of samples needed to be on a leaf node. This parameter describes the minimum number of samples on the sheets. According to Figure 3 on the right, it is observed that 5 is the best result.

Parameters and Precision - (KNN) The KNN algorithm classifies examples, taking in view of the neighbouring K class more adjacent. If K is equal to 1, then the new data is sorted with the same class as the most adjacent example. And, in case K is greater than 1, then the classes of K are considered more adjacent to perform the classification.

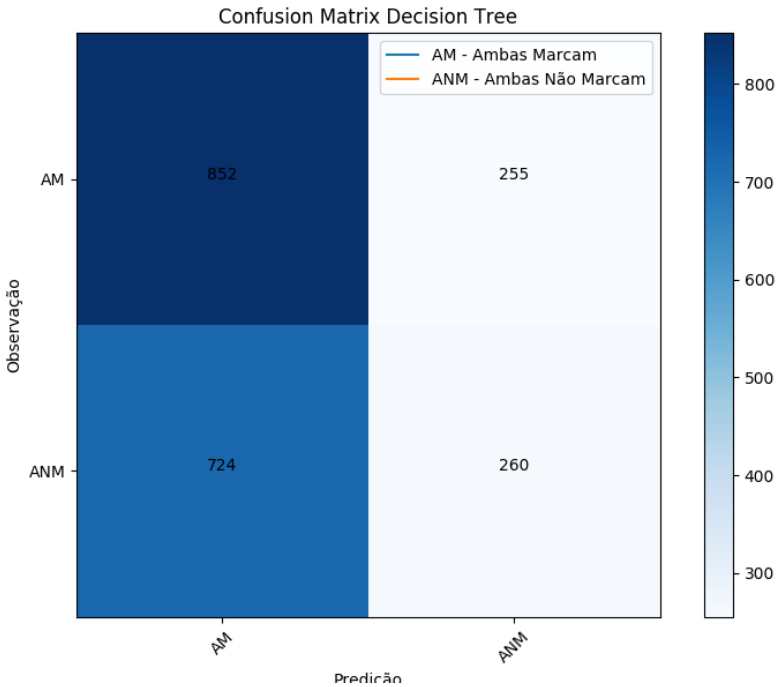

Fig. 5. Confusion Matrix - DT

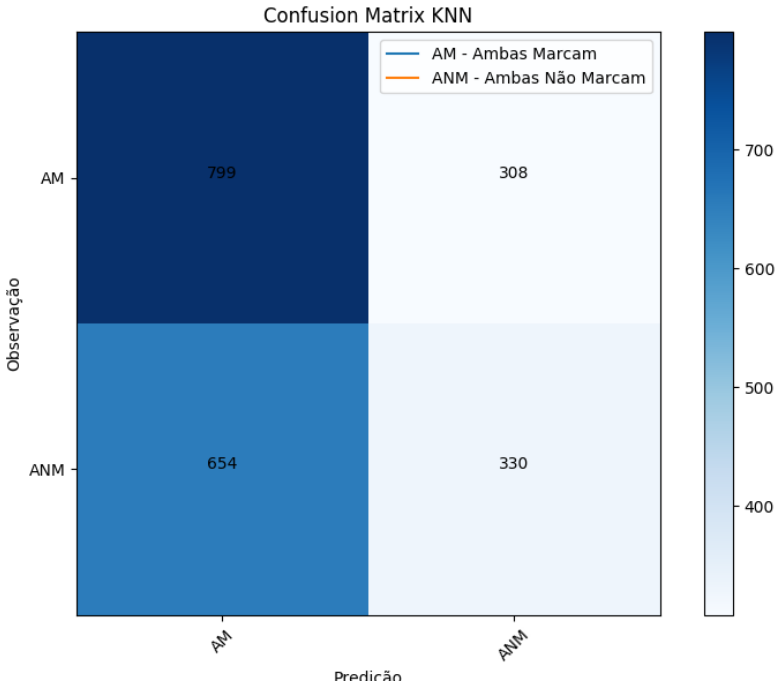

Fig. 6. Confusion Matrix - KNN

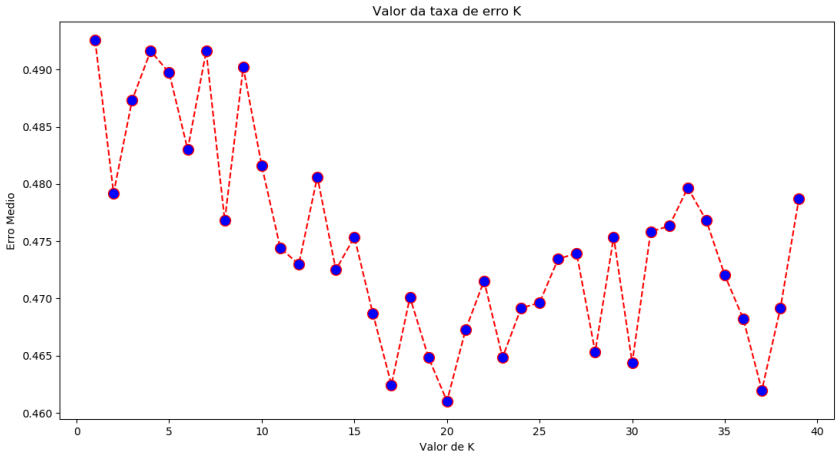

Fig. 7. K-value most suitable for the KNN algorithm

Since to have a good classification with this algorithm, it is essential to choose the most ideal K parameter, two ways were applied to find the best K, as we can see in Figure 7, where the highest value ideal of K is 20.

6 Case of Study

Table 1. Algorithm Classification Report applied to the case study

Algorithms	Target	Precision	Recall	F1_Score	Support
SVM	*Both do not mark*	0,54	0,65	0,59	1107
	Both mark	0,50	**0,39**	**0,44**	984
	Micro Average	0,53	0,53	0,53	2091
	Macro Average	0,52	0,52	0,51	2091
	Weighted avg	0,52	0,53	0,52	2091
DT	*Both do not mark*	0,54	**0,77**	**0,64**	1107
	Both mark	0,50	0,26	0,35	984
	Micro Average	0,53	0,53	0,53	2091
	Macro Average	0,52	0,52	0,49	2091
	Weighted avg	0,52	0,53	0,50	2091
KNN	*Both do not mark*	**0,55**	0,72	0,62	1107
	Both mark	**0,52**	0,34	0,41	984
	Micro Average	**0,54**	**0,54**	**0,54**	2091
	Macro Average	**0,53**	**0,53**	**0,52**	2091
	Weighted avg	**0,53**	**0,54**	**0,52**	2091

The problem to be studied with the various techniques of classification algorithms, is whether or not both teams score or not in a game. Based on past records, the model will predict each round the result if both teams score in the NOS League. As shown in Table 2 prediction was made of 18Th day of the Portuguese league, and the classification model with best result is the KNN.

Obtained Result and Algorithms Evaluation In this case study, several percentages were applied in the division between training data and test. The best percentage was 30% for the test data and 70% for the training data. Thus, there are 4877 samples in the training set, and 2091 samples in the test set.

Initially, using SVM with the default values, without choosing the best value of the parameter, the results obtained according to table 1, demonstrate that with the SVM algorithm applied in this project, there is a prediction accuracy of about 52.65%, as can be seen in Figure 2. But according to Table 2, applying this method to the 18th NOS League Match day, we obtain a hit rate of 44.44%, in 9 games, having been successful in 4 games.

Then, when the best values of C, Gamma and Kernel, we get an improvement of 0.5%. Figure 2 on the left shows the MC before finding the best value of each of the parameters. The figure of the right shows the MC after been applied the best value of the parameters. An improvement in the accuracy of the algorithm can be observed. Applying again to the 18th Match day of NOS League, we obtain an accuracy of 55.56%, where in 9 games you get it in 5 games.

Of all the classification algorithms applied to this problem, the LR is the algorithm that had the worst predictive accuracy, although according to Table 2 had obtained better accuracy than SVM.

Table 2. Accuracy of the 18th NOS League Match

Home Team	**Away Team**		**Both Mark**	**KNN**	**SVM**	**Decision Tree**	**Linear Regression**
Rio Ave	Feirense	0-0	0	0	1	1	0
Boavista	Portimonense	0-2	0	0	0	0	0
Aves	Guimaraes	2-1	1	1	1	1	0
Santa Clara	Maritimo	0-1	0	0	0	1	0
Sp Lisbon	Moreirense	2-1	1	0	0	0	0
Belenenses	Tondela	2-2	1	0	0	0	0
Setubal	Benfica	0-1	0	0	0	0	0
Chaves	Porto	1-4	1	1	0	1	1
Nacional	Sp Braga	0-3	0	0	1	0	1
			100%	**77,78%**	44,44%	55,56%	55,56%

The DT classification algorithm obtained the second best prediction accuracy if both teams score. Table 1 shows that it has a precision of 53.18%. According to Table 2, applying this method to the 18Th League NOS, we get a precision of 55.56%, which means that in 9 games, hit 5 games.

According to Table 1, with K equal to 20, can get a prediction accuracy of 53.99%. Applying this algorithm to the 18Th Journey we had the best classification with about 77.18%, which means that in 9 games, 7 games were hit in right way. This It can be considered that, for this case study where there are many unpredictable variables, getting this result is very good. Was not found no design with similarity that could be compared with this model implemented.

7 Critic and Conclusion

The costs of error or adequately measure the performance of classifiers through the error rate or precision, assumes a preponderant role in ML, since the real objective is to construct classifiers with low rate of error in new examples[19].

In this case study, it is not easy to find the ideal characteristics to minimise the error rate, since, as already mentioned, in football there are many unpredictable variables, such as the meteorological factors, players and coaches, among other characteristics which cannot be defined. If it were possible, it would be extremely important for the classifiers improve their performances.

The development of the present study made possible the acquisition of knowledge, analysis, application of LM techniques, and implementation of classifiers for answer to the future games of the Portuguese league or any national competition. In addition, it also allowed to discover among the algorithms classification which is better suited to the nature of this project, and has concluded that K-Nearest Neighbor managed to have the best prediction accuracy of the results, with about 54%.

Given the importance of the project subject, it is proposed as future work the study, discovery, development and implementation of new features which may be relevant, in order to increase the precision of the built classifier model. Also, for future work, the need to conduct a deep study, modifying the data set, adding two characteristic (Date of play and defensive power of the teams), and to study the importance of these Highlighted problem in the study. You can also extend the horizon of this project, instead only make the forecast if both teams and can apply the study to predict which of the teams wins the game, and also see if the two teams together score more than 2.5 goals, which is one of the new facets in the betting houses.

Briefly, it is concluded that the object of study can add more features and continue with the same accuracy or worse. It has to take into consideration the relevance of the new feature in relation to the study in question. The power of the half field of the teams was added and the precision kept same in all classifiers., which makes it possible to record that this characteristic has no relevance if the teams score goals or not. Finally, it is proposed a deeper study of the LR algorithm, comparing with the implemented model, to see if it will obtain better results. Performing another test with the 19th NOS League, again the KNN algorithm in 9 matches hit 7 games, while SVM hit 5 games and the DT hit 4 games, which attests that in this case study, the KNN is the best prediction algorithm.

References

[1] N. Ancona, G. Cicirelli, A. Branca, and A. Distante. Goal detection in football by using support vector machines for classification. In *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, volume 1, pages 611–616. IEEE, 2001.

[2] H. Chih-Wei and L. Chih-Jen. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, March 2002.

[3] V. Cortes, C.and Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[4] B. Deekshatulu, P. Chandra, et al. Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia Technology*, 10:85–94, 2013.

[5] B. Faria, L. Reis, N. Lau, and G. Castillo. Machine learning algorithms applied to the classification of robotic soccer formations and opponent teams. In *2010 IEEE Conference on Cybernetics and Intelligent Systems*, pages 344–349, June 2010.

[6] C. Ferrero. *Algoritmo kNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia*. PhD thesis, Universidade de São Paulo, 2009.

[7] P. Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

[8] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

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

[11] Z. Huang, H. Chen, C. Hsu, W. Chen, and S. Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, 37(4):543–558, 2004.

[12] C. Igiri and E. Nwachukwu. An improved prediction system for football a match result. *IOSR Journal of Engineering (IOSRJEN)*, 4:12–20, 2014.

[13] S. Imandoust and M. Bolandraftar. Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications*, 3(5):605–610, 2013.

[14] D. Karlis and I. Ntzoufras. Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, 20(2):133–145, 2008.

[15] I. Kuncheva, Ludmila. Editing for the k-nearest neighbors rule by a genetic algorithm. *Pattern Recognition Letters*, 16(8):809–814, 1995.

[16] P. Langley et al. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, volume 184, pages 245–271, 1994.

[17] J. Lope, D. Maravall, and J. Martin. Robust high performance reinforcement learning through weighted k-nearest neighbors. *Neurocomputing*, 74(8):1251–1259, 2011.

[18] M. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[19] J. Monard, M.and Baranauskas. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, 1(1):32, 2003.

[20] B. Natarajan. *Machine learning: a theoretical approach*. Elsevier, 2014.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[22] D. Prasetio and D. Harlili. Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–5, Aug 2016.

[23] P. Rudin. Football result prediction using simple classification algorithms, a comparison between k-nearest neighbor and linear regression, 2016.

[24] L. Samaniego and K. Schulz. Supervised classification of agricultural land cover using a modified k-nn technique (mnn) and landsat remote sensing imagery. *Remote Sensing*, 1(4):875–895, 2009.

[25] H. Schmidt. Uso de técnicas de aprendizado de máquina no auxílio em previsão de resultados de partidas de futebol., 2017.

[26] J. Shin and R. Gasparyan. A novel way to soccer match prediction. *Stanford University: Department of Computer Science*, 2014.

[27] M. Shouman, T. Turner, and R. Stocker. Using decision tree for diagnosing heart disease patients. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, pages 23–30. Australian Computer Society, Inc., 2011.

[28] R Stefani. Predicting score difference versus score total in rugby and soccer. *IMA Journal of Management Mathematics*, 20(2):147–158, 2009.

[29] J. Texeira. *Inteligência artificial*. Pia Sociedade de São Paulo-Editora Paulus, 2014.

[30] D. Wu. Supplier selection: A hybrid model using dea, decision tree and neural network. *Expert Systems with Applications*, 36(5):9105–9112, 2009.

[31] Q. Yu, A. Sorjamaa, Y. Miche, A. Lendasse, E. Séverin, A. Guillén, and F. Mateo. Optimal pruned k-nearest neighbors: Op-knn application to financial modeling. In *Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on*, pages 764–769. IEEE, 2008.