Fundamental Journal of Mathematics and Applications, 4 (4) (2021) 232-244
Research Article

Fundamental Journal of Mathematics and Applications

Journal Homepage: www.dergipark.org.tr/en/pub/fujma
ISSN: 2645-8845
doi: https://dx.doi.org/10.33401/fujma.951665

Modelling Sport Events with Supervised Machine Learning

İrem Barman[1] **and İbrahim Demir**[1*]

[1]*Department of Statistics, Faculty of Science and Arts, Yıldız Technical University, İstanbul, Turkey*
[*]*Corresponding author*

Article Info

Keywords: *Decision tree, k-nearest neighbor, Modelling sport events, Naive Bayes, Random forest, Supervised machine learning, Support vector machines*
2010 AMS: *xxxxx, xxxxx*
Received: *13 June 2021*
Accepted: *1 October 2021*
Available online: *1 December 2021*

Abstract

It has been very important to understand the change of multivariable systems to make predictions accordingly. The goal of supervised machine learning is to build a model of changing classes of observations depending on various variables and to make predictions about the coming situations. Due to the fact that sports are followed by the whole world modelling sports events and studies about predicting the results of future matches have gained importance. In this study, match statistics of the teams in the Turkey Super League were used, and it was examined how successfully the outcome of the match was predicted using a decision tree, random forest, k-nearest neighbor, naive Bayes, support vector machine. According to the tests done in Turkey Super League, the support vector machine performs the best.

1. Introduction

Nowadays, machine learning has been one of the fastest growing fields of computer science with the data growth and large scaled applications. In general, machine learning is unfoldment of the hidden structure in the data. Machine learning is the field of research devoted to the formal study of learning systems. This is a highly interdisciplinary field which borrows and builds upon ideas from statistics, computer science, engineering, cognitive science, optimization theory and many other disciplines of science and mathematics [1].

There are several applications for Machine Learning (ML), the most significant of which is data mining [2]. Machine learning is examined under three main headings: Supervised machine learning, unsupervised machine learning, reinforcement machine learning. Machine learning algorithms differ from each other according to the way to get the required results. Supervised machine learning has a algorithm that process by matching the outputs and inputs. While the system was being trained, the outputs and inputs of each sample in data set are given. The problem with supervised machine learning is tackled as classification problem[3].

The training set given for unsupervised leaning is the unlabeled dataset. Unsupervised learning aims at clustering, probability density estimation, finding association among features, and dimensionality reduction. In general, an unsupervised algorithm may simultaneously learn more than one properties listed above, and the results from unsupervised learning could be further used for supervised learning [3].

Score prediction with machine learning has frequently been the subject of academic studies because of collecting data of past sport events is easier than other fields with sport being followed by the whole world and the opportunities provided by technology. Although the most of the studies is about football, basketball, baseball and American football, there are also studies for different sports. Analyses such as athlete injury risks, ticket sales prediction, success evaluation, score prediction

Email addresses and ORCID numbers: barmaniem@gmail.com, 0000-0002-2857-014X (İ. Barman), idemir@gmail.com, 0000-0002-2734-4116 (İ. Demir)

are performed. In addition to these, machine learning are utilized for specifying game strategies from sports data and making selection athlete, trainer, equipment [4].

In modelling sport events, generally supervised machine learning algorithms such as regression models based Poisson, decision tree, neural network and models completely explained by data are used.

2. Literature

There are many statistical and machine learning studies about modelling sports events. Linear models based on probability distribution ise used in some of these. Another its part is tackled as a classification problem in form of predicting the match outcome with machine learning algorithms. Harville [5] has developed one of the first linear models based on past matches data. He used the model to predict match results. Knorr-Held [6] and Koning [7] developed models based on different statistical eventuations with match results in the form of win-lose-draw and predicted future match results.

The number of goals scored by teams in a football match are Poisson distributed and Poisson variables of these teams depend offensive power of one of these teams and defensive power of other team. Maher [8] developed the first Poisson-based regression model depending the number of goals scored and conceded by the teams in the match and he used for score prediction in football. Crowder, Dixon et.al. [9] added the time factor to the Poisson-based regression model and used to predict football match results. After Karlis and Ntzoufras [10, 11] have done a study on score prediction in football, they made evaluations for outcome estimation per applying the Poisson-based model to water polo.

In Poisson-based models and time series models, outcome estimations based on team strength determined by data based on goals scored and conceded by the teams in the past matches, and in linear regression model, outcome estimations based directly on the match results are made. Especially, Poisson-based models and time series models are used often for outcome prediction in football matches. However factors such as foul, shot, corner, offside point, percentage of passes, possession of the ball, goal attempt, players morale and the position of the team in the league affect the match results, accordingly these models are inadequate in outcome estimation. Bayesian network models aim to predict match results by determining the offensive and defensive strength of teams by considering different factors. Bayesian network models give successful results because of they also pay regard to the relationships between variables. Rue and Salvesen [12], have used Bayesian linear model for English Premier League. Baio and Blangiargo [13] brought a Bayesian approach to the Poisson based model and they have established the score prediction model in football.

In modelling studies in field of sport, when the variable selection based on expert knowledge was made, the ideas that the models give more successful results have became prominent. Josephs, Fenton et.al. [14]'s study is a good example for this. Using 2006 FIFA World Cup data, the matches that ended together were extracted, and the winnings of the home team or the visiting team were estimated using data based machine learning algorithms and Bayesian network based on expert knowledge. While achieving 60% success with machine learning algorithms, 76.9% success was achieved with Bayesian networks based on expert knowledge. Thats why Bayesian network has been stated that it is more successful in modelling. Huang [15] modeled the Tottenhnam Hotspur's which is Englang Premier League team game between 1995-1997 with Bayesian network. In her study, while Bayesian networks created with the experts of the subject made prediction with 59% success, the other networks remained in the range of $40-50\%$. Similarly, Constantinou et.al [16] developed Bayesian networks model based on expert knowledge with English Premier League seasonal data. The model has been successful with high accuracy rate. Points that the teams will collect during the season has been estimated via expert knowledge based Bayesian network model developed in the study made by Contantinou ve Fenton [17]. It has been established that the model can be used to predict both the season ranking and the outcome of individual matches. Karabiyik and Yet [18] developed Bayesian network model based on expert knowledge which called FutBa for Turkey Super League. The model predicts the outcome of past and future matches with an accuracy of $60-70\%$.

The use of artificial neural networks in modelling sports events is also quite common. In Purucker [19]'s study, one of the first examples, the matches belonging to the first eight weeks of the NFL 1994 season were modeled with artificial neural networks and it was intended to predict the winner of the NFL games. Back-propagation network model predicts with 60.7% accuracy. Kahn [20] developed artificial neural networks model with data from 2003 NFL season in first 14 week and that model has achieved 75% success in predicting seasonal averages at 14th week. McCabe and Trevathan [21] modeled rugby matches with artificial neural networks. Tests were applied out for four different rugby leagues in their study. The model succeeded between 54% and 65%.

Various machine learning algorithms have been used in modelling sport events. Hamadoni [22] used logistic regression and support vector machine algorithms in NFL for predict which team winning game from 2003 to 2005 seasons. In comparison made by Hamadoni, support vector machine algorithm has shown a test success rate of 67.08% for 2003, 61.37% for 2004, 65.83% for 2005. Similarly Sierra, Forco et.al. [23] have developed support vector machine model and logistic regression model to predict the NFL game results. Linear support vector machine algorithm was more successful than logistic regression

model and other support vector machine models. Smith, Lipscomb et.al. [24] used naive Bayes algorithms to predict which team wins in MLB (Major League Baseball) and they tested with passed matches data between 1967 and 2006. The model predicted winners 80% correctly. Hucaljuk ve Rakipovic [25] modeled 96 matches in UEFA Champions League using 30 features with different machine learning algorithms. Models have shown success rate between 60% and 50%. Cao [26] modeled with NBA data from 2006 to 2010 seasons. Logistic regression made predictions with 67.82% accuracy, support vector machine made predictions with ½67.22%½ accuracy, multilayer perceptron neural network made predictions with 66.67% accuracy and naive Bayes made predictions with 65.82% accuracy. Yezus [27] modeled English Premier League ames using 9 features with k-nearest neighbors and random forest algorithms. Models have shown success rate 55.8% and 63.4% respectively. Ulmer and Hernandez [28] also modeled English Premier Leauge. Linear classification model they developed has achieved 48% success. Support vector machine model predicted with 50% accuracy and random forest model made predictions with 50% accuracy. Karaoğlu [29] modeled 16 different league with data from 2013-2014 season and 2015-2016 season using machine learning algorithms. The best model performance was 52%. Vaidya, Sanghavi et.al. [30] also modeled English Premier League using data from 2006 to 2010 seasons with logistic regression, random forest and naive Bayes algorithms. The models have shown 49.37%, 47.11% and 47.11% success respectively. Soto Valero [31] aimed to assess the predictive capabilities of four machine learning methods for predicting outcomes in MLB regular season games and he used k-nearest neighbors, artificial neural network, support vector machine algorithms. As a result of him study, it was revealed that the classification algorithms make more successful predictions than regression models. The support vector machine model from four machine learning algorithms was the most successful model by making predictions with approximately 60% accuracy.

In this study decision tree, random forest, k-nearest neighbors, naive Bayes, support vector machine will be used from supervised machine learning algorithms in the modelling of Turkish Super League. This paper has a unique structure due to the variety of algorithms used in this study and the lack of modelling studies in Turkey Football League.

3. Method

Each observation used by machine learning algorithms is represented with the same variable set. Variables in the variable set can have different structures such as categorical, continuous or bivalent. When one of the variables is considered as output and the others are also considered as inputs, if the label of output variable of the observations is known it is supervised machine learning method. Supervised machine learning methods are called as classification algorithms because of supervised machine learning methods works with labeled dataset. The main purpose of supervised machine learning is that build model that can predict the class values of the test set with training set.

In supervised machine learning, the training data with labeled output variable is processed with machine learning algorithms and thus a prediction model is created. The model is tested using unlabeled test data and the class labels of observations in test data is predicted. This process has been shown in Figure 3.1.

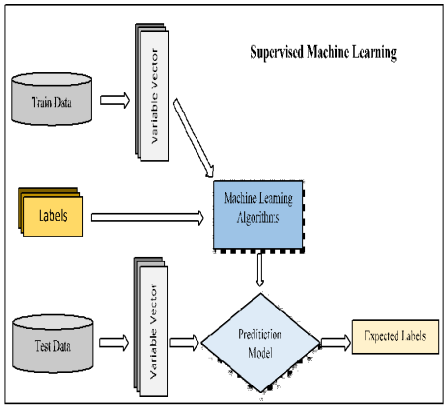

Figure 3.1: Supervised machine learning process

In this study, decision tree, random forest, k-nearest neighbor, naive Bayes, support vector machine widely used have been concentrated.

3.1. Decision tree

Decision tree is one of the commonly used supervised machine learning methods in classification. In this case, its properties such as simplicity of implementation, understandability, no using parameters, availability for mixed data types, being faster than other methods plays a role. Despite these positive properties, decision tree algorithm contains some problems. Not

possible to obtain outputs containing more than one feature, causing changeable results, being sensitive to small alterations, giving complex output for numerical dataset are some of these problems.

In decision trees, it is aimed to create a tree according to the features of the data in the training set. Decision trees decide which class new data belongs by determining separation rules based on historical data. It acts on questions and answers and it creates rules by combining asked questions with answers. It can also be said that the resulting tree is set of rules consisting of many if-then. When it is decided which variable in the data to start asking question, the related variable creates the root node of the tree. Starting from the root node, new node are created according to the answers received by asking questions whose answer in the dataset. Then, each node is split up two node or more. If new question can not be asked after the node are created, branching is over [32].
As a result of this process, a classification tree as in Table 1 is obtained.

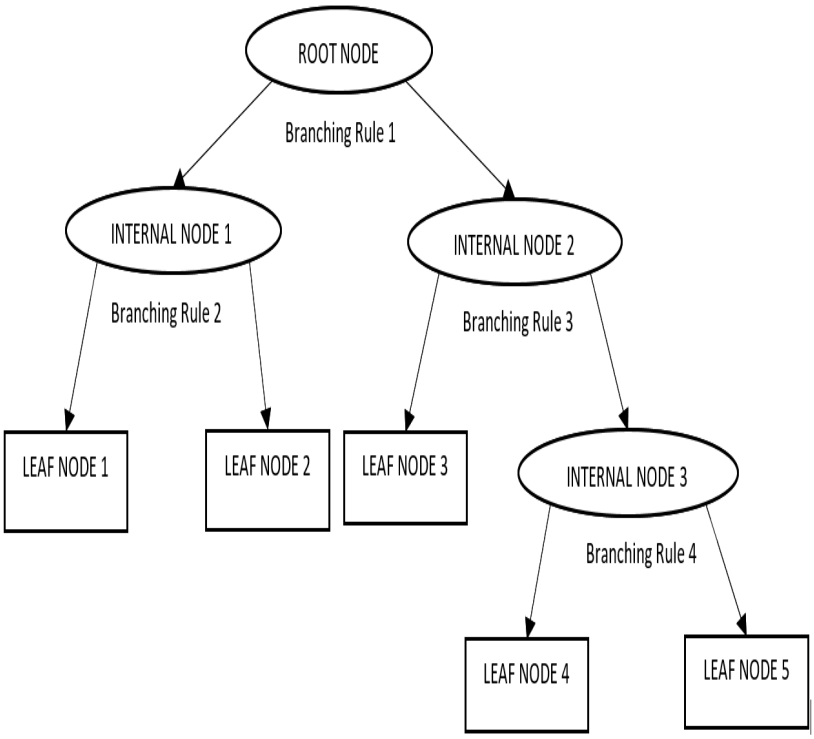

Figure 3.2: Decision tree structure

There are various decision tree algorithms that differ according to the criteria they use. First, Automatic Interaction Detector (AID) was used in decision tree applications. Then many algorithms were developed. The main ones are CART (Classification and Regression Trees), CHAID (Chi-Squared Automatic Interaction Detector), Exhaustive CHAID (Chi-Squared Automatic Interaction Detection), ID3 (Iterative Dichotomiser 3), C4.5, MARS (Multivariate Adaptive Regression Splines), QUEST (Quick, Unbiased, Efficient Statistical Tree), C5.0, SLIQ (Supervised Learning in Quest), SPRINT (Scalable Parallelizable Induction of Decision Trees).

3.2. Random forest

In multivariate different data groups, the success of classifier varies. In this situation, it can be efficient to use single classifier in terms of results. Ensemble algorithms uses sets of classifier together instead of a classifier to resolve the problem. The most common of these are bagging, boosting and random forest.
Random forest comprises of the union of many decision trees. Trees are formed by drawing independently from each other with boostrap method. After the trees are voted by one, the winning classifier is chosen. Random forest determines variables to use in branching by selecting the randomly choosed m pieces from all the data in the data set. m is usually taken as the square root of the variable number. The advantage of random forest over the bagging algorithm is that it puts randomness into training stage [33].

Random forest algorithm can be used for purposes such as finding error rate of the algorithm, determining the importance levels of variable, specifying the outliers and detecting the missing value.

In random forest, if the original data set does not have a test set of its own, from the original dataset n samples are selected by the bootstrap method. 2/3 of each sample is used to create tree and the remaining 1/3 is used to calculate the error rate. If the original data set has its own test set, error rate of this test can also be calculated with the set. The variable that gives the best information among the randomly selected variables from the training set is used as the branching variable. These are performed simultaneously and iterated until the most successful tree is obtained. The flow chart of this process is given in Figure 3.3 below.

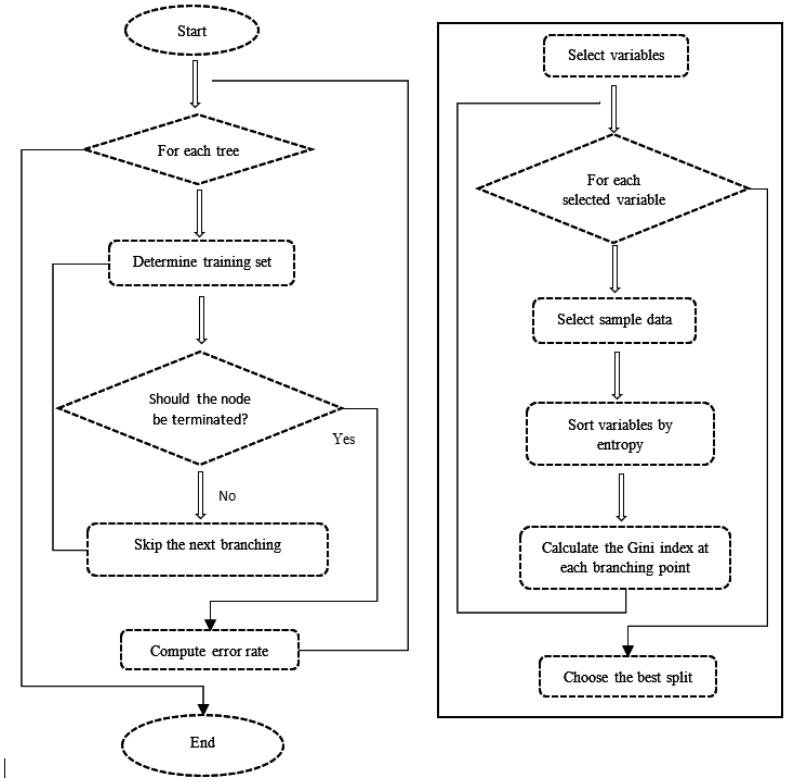

Figure 3.3: The random forest flow chart, [34]

3.3. k – nearest neighbor

k-nearest neighbor is a non-parametric, sample based classification algorithm. k-nearest neighbor is predicated on the principle that samples which are close to each other will be similar. When an sample without a class label is given, a sample space based on the classifier k created with k-nearest neighbor algorithm is created and if which class is repeated mostly in this sample space, this observation is assigned that class label. It would be appropriate to visualize it as in Figure 3.3.

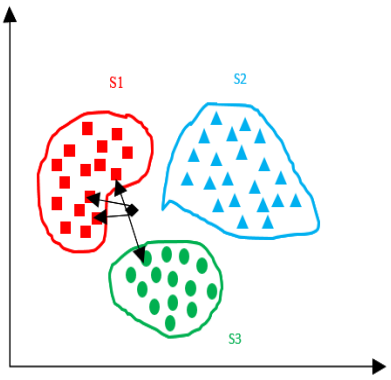

Figure 3.4: k-nearest neighbor classification visualization

The chose of k is very important. However there is no optimal k value, its value varies according to the nature of the problem. The chose of k effects the performance of k-nearest neighbor algorithm. If k is small value, it causes classification error. This problem can be solved by choosing a greater k. If k is great value, the proportion of classes at a specified distance will decrease

and the instances of other classes will gain the majority. This is also causes a classification error again. It can be solved by choosing a smaller k [35].

3.4. Naive Bayes algorithms

Naive Bayes networks are the simplest Bayesian networks. Naive Bayes classifier is composed without a cyclical relationship with only one parent and several children with a strong assumption of independence among child nodes in the context of their parent.

Accordingly, for example, the naive bayes network for the data set consisting of one output variable and four input variables should be handled as in the Table 2. This input variables affect the output variable independently of each other.

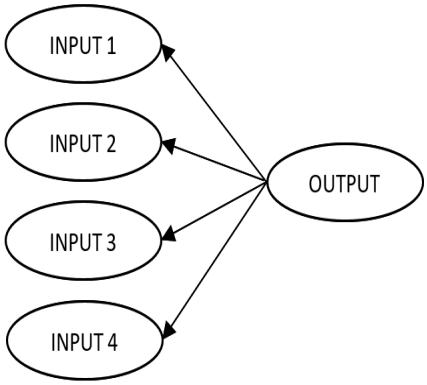

Figure 3.5: Naive Bayes algorithm network

The naive Bayes is based on Bayes Theorem. As part of the theorem, naive Bayes algorithm makes predictions by calculating the probabilities of classes with the probabilities obtained from a labeled training set. Bayes Theorem is shown below.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

where, Y: Some hypothesis, such that data tuple X belongs to specified class C, X: Some evidence, describe by measure on set of attributes, $P(Y|X)$: The posterior probability that the hypothesis H holds given the evidence X, $P(Y)$: Prior probability of H, independent on X, $P(X|Y)$: The posterior probability that of X conditioned on H. The algorithm calculates the following two probabilities and compares them.

$$R = \frac{P(m|X)}{P(n|X)} = \frac{P(m)P(X|m)}{P(n)P(X|n)} = \frac{P(m)\prod P(X|m)}{P(n)\prod P(X|n)}.$$

As a result of comparing these probabilities, predicted class label is the class of higher probability. If $R > 1$, prediction is m, otherwise prediction is n. The major advantage of the naive Bayes classifier is its short computational time for training. In addition, since the model has the form of a product, it can be converted into a sum through the use of logarithms with significant consequent computational advantages. It does not need any complicated iterative parameter estimation schemes, so can be applied to large data set. Easy interpretation of knowledge representation. It is also easy to present and understand due to the interpretation of knowledge representation is easy. It may not be best classifier in any particular application, but it does well and robust. Otherwise, naive Bayes has some disadvantages. Theoretically, naive Bayes classifier have minimum error rate comparing to other classifier, but practically it is not always true, because of assumption of class conditional independence and the lack of available probability data. It has less accurate compare to other classifier [36]-[38].

3.5. Support vector machine

Support vector machines are one of the most powerful supervised machine learning algorithms. In addition, it has many application fields such as classification, regression, variable selection, detection of outliers. Support vector machine is algorithm that revolve around the notion of margin and create optimal separating hyperplane according to this. Margin is the distance between the hyperplane and the closest observation point on any side of the plane. The classifier chooses hyperplane in accordance the data used to detect optimum margin. It is examined as cases where data is separated completely linear and data can not be separated linearly [39]. SVM searches for the optimal separating hyperplane that correctly classifies the data as shown in Figure 3.6.

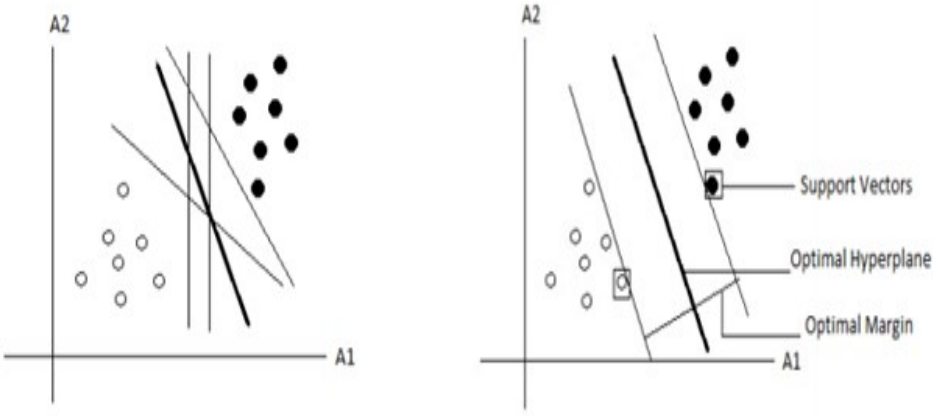

Figure 3.6: The optimal seperating hyperplane for linearly seperable data [38]

In real world problem, linearly separated data are not encountered. In this case support vector machine can not use linear learning and it is extended to learn non-linear decision onto a high dimensional sample space using Kernel functions. The common Kernel functions are used in support vector machine are linear, polynomial, radial, sigmoid. In case of the data can not be separated linearly, the support vector machine illustration designed using Kernel functions is shown in Figure 3.7.

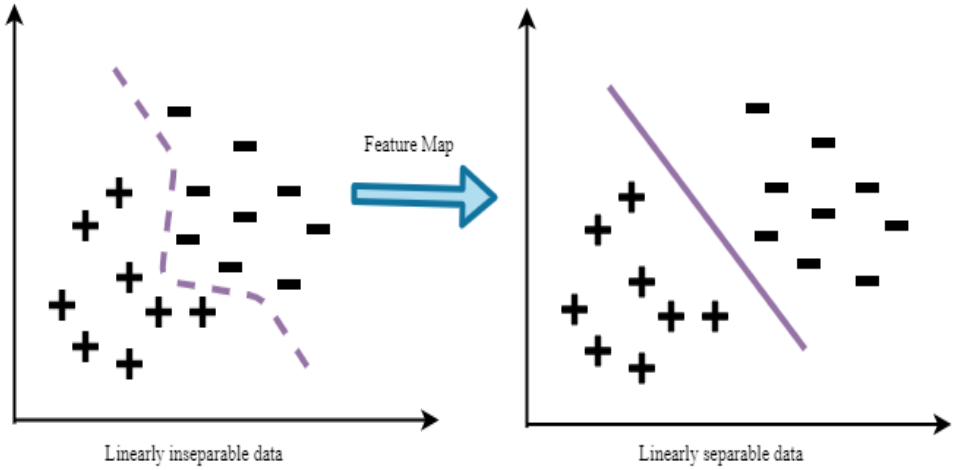

Figure 3.7: Illustration of nonlinear support vector mavhine consept [41]

Support vector machine has a sturdy theoretical basis. It requires only examples and is insensitive to the size. Support vector machine is the best algorithm to classify the members of two classes in the training set. It is less prone to overfilling than other methods. However, it is computationally expensive and tedious. It learns extremely slow. Its results may be poorly interpreted. [36, 40]

4. Data management

Modelling Turkey football matches with supervised machine learning algorithms using Turkey Super Lig data and comparison of prediction success of the models are aimed. The study is carried out in R Studio.

4.1. Data collection

The data is taken from WhoScored.com. Turkey Super Lig 2018-2019 season statistics have been used. There are 612 observations belong to 306 matches played during Turkey Super Lig 2018-2019 season. Each observation consists of 14 variables. These variables are team name and the match results in form of win-lose-draw, home ownership, number of goal, number of offside, percentage of pass, crossing, ball hawking, number of shot, goal conversion rate, possession of the ball, number of red-card, number of yellow-card, number of foul.

4.2. Data arrangement

Choosing the variables to be used in organizing data is very important. It can be explained with two subject. The first one is the requirement for the model to be simple. Second, using significant variables in the model can show a more meaningful model performance.

The first thing to look at is the correlation between variables. In consequence of determination of correlation between number of goal and goal conversion rate with percentage of pass and possession of the ball, goal conversion rate and possession of the ball variables have left out of analysis. After controlling correlation, choosing the variables should be made. There is various variable selection method can be used. The methods can be organized into three categories, depending on how they combine the feature selection search with the construction of the classification model: filter methods, wrapper methods and embedded methods [42].

Forward selection method has been used in the study. Forward selection algorithm works in the way deciding the final model, starting from empty model, adding each variable. The model involved all variables has been decided. The variables in the processed data set are given below.

```
     result            team               goal        homeownership
 Min.   :0.0000   Length:612         Min.   :0.000   Min.   :0.0
 1st Qu.:0.0000   Class :character   1st Qu.:0.000   1st Qu.:0.0
 Median :1.0000   Mode  :character   Median :1.000   Median :0.5
 Mean   :0.9363                      Mean   :1.338   Mean   :0.5
 3rd Qu.:2.0000                      3rd Qu.:2.000   3rd Qu.:1.0
 Max.   :2.0000                      Max.   :7.000   Max.   :1.0

     ofsaid            pass           crossing         hawking
 Min.   : 0.000   Min.   :0.5400   Min.   : 1.000   Min.   : 4.00
 1st Qu.: 1.000   1st Qu.:0.7400   1st Qu.: 6.000   1st Qu.:13.00
 Median : 2.000   Median :0.7900   Median : 8.000   Median :16.00
 Mean   : 1.894   Mean   :0.7765   Mean   : 8.572   Mean   :16.43
 3rd Qu.: 3.000   3rd Qu.:0.8200   3rd Qu.:11.000   3rd Qu.:19.00
 Max.   :11.000   Max.   :0.9200   Max.   :28.000   Max.   :35.00

      shot              red              yellow           foul
 Min.   : 2.00   Min.   :0.0000   Min.   : 0.000   Min.   :  4.0
 1st Qu.: 9.00   1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.: 11.0
 Median :12.00   Median :0.0000   Median : 2.000   Median : 14.0
 Mean   :12.74   Mean   :0.1307   Mean   : 2.248   Mean   : 14.1
 3rd Qu.:16.00   3rd Qu.:0.0000   3rd Qu.: 3.000   3rd Qu.: 16.0
 Max.   :34.00   Max.   :2.0000   Max.   :20.000   Max.   :141.0
```

Figure 4.1: Descriptive statistics of the variables

Dependent variable:
result (0: Lose, 1: Win, 2: Draw)

Independent variables:
team (Teams in Turkey Super Lig 2018-2019 season)
home ownership (0: No, 1: Yes)
goal (The number of goals a team scored in the match)
offside (The number of offside a team has committed in the match)
pass (The pass rate of a team in the match)
crossing (The number of opponents passed by a team player without losing the ball)
hawking (The number of times a team takes the ball from opponent)
shot (The number of shot for a team in a match)
red (The number of red card a team committed in the match)
yellow (The number of yellow card a team committed in the match)
fault (The number of fouls a team has committed in the match)

5. Model building

It is aspired to classify the match results belonged to the teams in form of win/lose/draw through the instrument of other match statistics. Hence classification algorithms from supervised machine learning algorithms that suitable for the purpose of study are used. At this stage prediction model will be developed using decision tree, random forest, k-nearest neighbors, naive Bayes, support vector machine algorithms. Data set consisting of 612 observations is divided into 75% training set and 25% test set to use in modelling.

5.1. Decision trees algorithms

The variables given in the data management phase have been used in analysis. The CART algorithm from decision tree have been applied to the data. Because of the dependent variable is categorical, the Gini criterion was chosen as the classification scale. The tree structure that minimizes the error rates and maximizes the classification success was specified applying 10-fold cross validation test. As a consequence of the analysis, it is determined that the most important variable affecting the class variable which is dependent variable is goal. Team and pass variables supervene on goal. When classification success was controlled, the accuracy rate was found to be 60.1%. Respectively, sensitivity values of the classes are 52.2%, 79.3%, 41.3% and selectivity values of the classes are 80.7%, 86.6%, 73.8%. These are given in Table 1.

	Lose	Win	Draw
Sensitivity	0.5227	0.7937	0.4130
Selectivity	0.8073	0.8667	0.7383

Table 1: Decision tree sensitivity and selectivity values

5.2. Random forest algorithms

Although the random forest algorithm is onerous, because of data management is done in previous stages, the algorithm has been simply applied to the data. After the necessary libraries are loaded in R Studio, the codes are written using the parameters in random forest package to train model. Then the fitting codes is written. When classification success was controlled, the accuracy rate was found to be 60.1%. Respectively, sensitivity values of the classes are 61.3%, 79%, 34% and selectivity values of the classes are 75.2%, 87.9%, 78.3%. These are given below in table. The results obtained are given in Table 2.

	Lose	Win	Draw
Sensitivity	0.6136	0.7903	0.3404
Selectivity	0.7523	0.8791	0.7830

Table 2: Random forest sensitivity and selectivity values

5.3. k-nearest neighbors algorithms

The variables given in the data management phase have been used in analysis. As a result of cross validation control, the optimum value for k is decided to be 7. Right after classification, it is fixed that the accuracy rate of the model is 42.4%. Respectively, sensitivity values of the classes are 30.6%, 58%, 41.4% ve and selectivity values of the classes are 72.5%, 67.3%, 73.2% These are given below in table. The results obtained are given in Table 3.

	Lose	Win	Draw
Sensitivity	0.3065	0.5800	0.4146
Selectivity	0.7253	0.6736	0.7321

Table 3: k-nearest neighbors sensitivity and selectivity values

5.4. Naive Bayes algorithms

After data management phase, football data is separated as training set and test set. In the system the data has been classified with R Studio packages required for naive Bayes algorithm based upon the labeled class data. Then the classes of the unlabeled observation in test set has been predicted via the naive Bayes model. The classification success is determined as 55.6%. Respectively, sensitivity values of the classes are 33.3%, 90.7%, 37.9% and selectivity values of the classes are 72.5%, 86.9%, 74.7%. These are given below in Table 4.

	Lose	Win	Draw
Sensitivity	0.3333	0.9074	0.3788
Selectivity	0.7250	0.8687	0.7471

Table 4: Naive bayes sensitivity and selectivity values

5.5. Support vector machine algorithms

The system is modeled with the train data set using support vector machines algorithm. After the modelling of the system is completed, the basis function of support vector machines is improved in accordance with the model, radial kernel function has been used. Train and test errors vary related to gamma and cost values used in the function. With the calculations, the best gamma and cost values were found. The ideal situation would be that there be no training errors and minimal test errors. The aim is to make both minimums and thus the best classification model will be developed. Support vector machine training was repeated using best gamma and cost values. The model was accepted and controls were made with the test set. Accuracy rate was found to be 61.4% when the classification success was monitored(control). Sensitivity rates for lose, win, draw are 49.2%, 86%, 42.4%, and the selectivity rates are 85.6%, 86.5%, 72.5%, respectively. The classification results obtained are shown in Table 5.

	Lose	Win	Draw
Sensitivity	0.3333	0.9074	0.3788
Selectivity	0.7250	0.8687	0.7471

Table 5: Support vector machine sensitivity and selectivity values

5.6. Model comparison

In this study, we aimed to apply supervised machine learning algorithms to the field of sports and to evaluate the performance of the models through it. While evaluating the success of the models, accuracy, sensitivity and selectivity measures were used. The output which we get from these metrics is the complexity matrix. In the complexity matrix, there is a case of comparing the predicted groups made by a classification algorithm with the real-case groups. A complexity matrix is shown in the table. TP defines true-positive, TN indicates the true-negative which represents the correct number of classified samples. FP means false-positive. It shows that the positively predicted samples are in the negative class. FN means false negative and gives the number of samples that were predicted negatively when in fact in the positive class. Generally, the diagonal of the complexity matrix gives the number of correctly estimated samples, while the others give the number of incorrectly estimated samples.

		Prediction	
		Positive	Negative
Fact	Positive	TP	FN
	Negative	FP	TN

Table 6: Confusion matrix

The accuracy rate used in measuring model success is the ratio of the number of correctly classified samples to the total number of samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}.$$

Sensitivity is indicated the ratio of the number of correctly classified positive samples to the total number of positive samples.

$$Sensitivity = \frac{TP}{TP+FN}.$$

The selectivity criterion used to measure the success of the model is explained as the ratio of the number of correctly classified negative samples to the total number of negative samples.

$$Selectivity = \frac{TN}{TN+FP}.$$

6. Performance measure

The confusion matrices of the models were given in detail during the model building phase. At this stage, the accuracy rates are shown in the table below in order to compare the models. As a result of the tests, the best performing model was support vector machine, while the worst performing was k-nearest neighbor.

	Accuracy Rate
Decision tree model	0.6013072
Random forest model	0.6013072
k-Nearest neighbor model	0.4248366
Naive Bayes model	0.5555556
Support Vector Machine Model	0.6143791

Table 7: Performance of classification algorithms

In order to evaluate the models correctly, it is necessary to pay attention to the working area. Football is a sport that affected by a wide variety of factors. At the same time, according to experts, Turkey Super League teams are not very stable as team success. Teams that are the leaders of the league may suddenly show low performance and risk falling to lower ranks. In addition, there are major differences between the top three and the last three of the league. All of these pose a big problem in predicting the match results. Therefore, 61.4% success in terms of this study is a satisfactory result. Such as the English Premier League teams which are more stable and more homogeneous structure in themselves can show better model success within the same scope studies.

7. Results

Machine learning algorithms have diversified in progress of time and have been applied to different fields. Correspondingly, studies on which of the algorithms are more successful have a great place in literature. Although there are so many studies on the success of the algorithms, no consensus could be reached on this issue. There may be various reasons of that, but the main reason is even if a study is not an unique, each study has its idiosyncratic structure. Factors such as the data source used in the study, the preprocessing method, parameters used in the algorithms form the structure of the study and these affect the performance of the process. Therefore, it is normal for studies to give different results, assessments should be made by considering these.

In this study, modelling sport events via classification methods and comparison of the prediction success of the models are aimed. In line with this purpose, the five of supervised machine learning algorithms are commonly used in literature were applied. These are decision tree, random forest, k-nearest neighbor, naive Bayes and support vector machine algorithms. As a result of the test, it has been determined that support vector machines have the best performance with 64% success rate.

Considering this study and experimental studies in the literature, it is not possible to talk about the superiority of any algorithm over another. However, comparison of the models in solving a problem contributes both of the study and academic literature in the terms of consequences.

8. Discussion

The use of machine learning is getting common of sports area. Therefore, a lots of algorithm been developed and improved has a wide place in literature. Linear models, Poisson-based models, time series models used frequently in the beginning. As a result of the inadequacy of these models in predicting outcomes, Bayesian networks have gained importance and then Bayesian networks based on expert knowledge have showed off themselves. Joseph, Fenton et.al. predicted which of the home team and guest team will win with 2006 FIFA World Cup data's by removing the draw status from the data's. In their study while other supervised machine learning algorithms have shown 60% success, Bayesian networks were 76.9% successful. In this regard one of studies in Turkey is FutBa Model based on expert knowledge is developed by Karabiyik and Yet in 2019. The model has estimated with success rates in the range $60-70\%$. If the studies concentrated on supervised machine learning is viewed, Hucalijuk and Rakipovic's study draw attention. They modeled UEFA Champions League games and they made estimation in the range of $50-60\%$. Secondly, Yezus used k- nearest neighbor and random forest algorithms in 2014 to predict English Premier League results. The models have shown 55.8% and 68.4% estimation success. The same year Ulmer and Fernandez modeled English Premier League in their study. Linear classification model 48%, support vector machines 50%, random forest 50% were successful. Vaidya, Sanghavi et.al. modeled English Premier League for 2006-2010 seasons with logistic regression, random forest and naive Bayes algorithms. The accuracy of these models was 49.37%, 47.11%, 47.11% respectively. These are some examples about soccer from literature, nevertheless there is no consensus could be reached on the success of the models. In this study, sport events were modeled with five supervised machine learning algorithms which are decision tree, k-nearest neighbor, naive Bayes, support vector machine. After the test, it has been observed that support vector machine with 61.4% have the best performance. Consequently, it is not possible to talk about the superiority of any algorithm over another.

Acknowledgements

The authors would like to express their sincere thanks to the editor and the anonymous reviewers for their helpful comments and suggestions.

Funding

There is no funding for this work.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

References

[1] Z. Ghahrami, *Unsupervised Learning* Advanced Lectures on Machine Learning Springer, 2004.
[2] F.Y. Osisanwo, J.E.T. Akinsola, O. Awodele, J.O. Hinmikaiye, O. Olakanmi, J.Akinjobi, *Supervised machine learning algorithms: Classification and comparison*, IJCTT International Journal of Computer Trends And Technology, **48** (2017), 128-138.
[3] W.L. Chao, *Machine Learning Tutorial*, DISP Lab, Graduate Institute of Communication Engineering, National Taiwan University, 2011, https://tcxsproject.com.br/dev/Biblioteca%20Livros%20Hacker%20Gorpo%20Orko/Machine%20Learning%20Tutorial.pdf.
[4] Cao, Chenjie, *Sports data mining technology used in basketball outcome prediction*, Masters Dissertation, Technological University Dublin, 2012, https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis.
[5] D. Harville, *Predictions fot national football league games via linear model methodology*, J. Amer. Stat. Ass., **75** (1980), 516-524.
[6] Knorr-Held, *Dynamic rating of sport teams the statistican*, **49** (2000), 261-276.
[7] R.H. Koning, *Balance in competition in dutch soccer*, J. Royal Stat. Soci.: Ser. Statistician, **49** (2000), 419-431.
[8] M.J. Maher, *Modeling association football scores*, Statistica Neerlandica, **36** (1982), 109-110.
[9] M. Crowder, M. Dixon, A. Ledford, M. Robinson, *Dynamic modelling and prediction of English football league matches for betting*, J. Royal Stat. Soci.: Ser. Statistican, **51** (2002), 157-168.
[10] D. Karlis, L. Ntzoufras, *On modelling soccer data*, Student, **3** (2000), 229-244.
[11] D. Karlis, L. Ntzoufras, *Analysis of sports data by using bivariate poisson models*, J. Royal Stat. Soci.: Ser. Statistican, **52** (2003), 381-393.
[12] H. Rue, Ø. Salvesen, *Prediction and retrospective analysis of soccer matches in A league*, J. Royal Stat. Soci.: Ser. Statistican, **49** (2000), 399-418.
[13] G. Baio, M. Blangiardoi, *Bayesian hierarchical model for the prediction of football results*, J. App. Statistics, **37** (2010), 253-264.
[14] A. Joseph, N.E. Fenton, M. Neil, *Predicting football results using Bayesian nets and other machine learning techniques*, Knowledge-Based Systems, **19** (2006), 544-553.
[15] K.Y. Huang, *A neural network method for prediction od 2006 world cup football game*, The 2010 International Joint Conference on Neural Network, 2010.
[16] A.C. Constantinou, N.E. Fenton, M. Neil, *Pi-football: A bayesian network model for forecasting association football match qutcomes*, Knowledge-Based System, **36** (2012), 322-339.
[17] A.C. Constantinou, N.E. Fenton, *Towards smart-data: Improving predictive accuracy in long-term football team performance*, Knowledge Based System, **124** (2017), 93-104.

[18] M. Karabiyik, B. Yet, *Football analytics with Bayesian networks: The FutBA model*, Pamukkale University Journal of Engineering Sciences, **25** (2019), 121-131.
[19] M.C. Purucker, *Neural network quarterbacking potential*, IEEE, **15** (1996), 9-15.
[20] J. Kahn, *Neural Network Prediction of NFL Football Games*, Lecture Notes, Fall 2003, 1-19, https://docplayer.net/21763052-Neural-network-prediction-of-nfl-football-games-joshua-kahn.html.
[21] A. McCabe, J. Trevathan, *Artifical intelligence in sports prediction*, The Fifth International Conference on Information Technology: New Generations, Las Vegas, USA, 2008, 1194-1197.
[22] B. Hamadani, *Predicting The Outcome of NFL Games Using Machine Learning*, Stanford University, 2006, http://cs229.stanford.edu/proj2006/BabakHamadani-PredictingNFLGames.pdf.
[23] A. Sierra, J. Forco, C. Fierro, *Football Futures*, 2011, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.374.9764&rep=rep1&type=pdf.
[24] L. Smith, B. Lipscomb, A. Simkins, *Data mining in sports predicting Cy young award winners*, J. Com. Sci. in Colleges, **22** (2007), 115-121.
[25] J. Hucalijuk, A. Rakipovic, *Predicting Football Scores Using Machine Learning Techniques*, MIPRO 2011, 2011, 1623-1627.
[26] Cao, *Sports data mining technology used in basketball outcome prediction*, Masters Dissertation, Technological University Dublin, Ireland, 2012 https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis.
[27] A. Yezus, *Predicting Outcome of Soccer Matches Using Machine Learning*, Mathematics and Mechanics Faculty Term Paper, Saint-Petersburg State University, 2014, https://www.math.spbu.ru/SD_AIS/documents/2014-12-341/2014-12-tw-15.pdf.
[28] B. Ulmer, M. Fernandez, *Prediction Soccer Match Results in the English Premier League*, Stanford University, 2014, http://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf.
[29] B. Karaoğlu, *Modeling sports matches with machine learning*, EMO Sci. J., **5** (2015), 1-5.
[30] S. Vaidya, H. Sanghavi, K. Gevario, *Football match winner prediction*, Int. J. Comp. Appl., **154** (2016), 31-33.
[31] C. Soto Valero, *Prediction Win-Loss qutcomes in MLB regular season games – A comparative study using data mining methods*, I. J. Comp. Sci. in Sport, **15** (2016), 91-112.
[32] K. J. Archer, R. V. Kimes, *Empirical characterization of random forest variable importance measures*, Computational Statistics & Data Analysis, **52** (2008), 2249-2260.
[33] L. Breiman, *Random forest*, Machine Learning, **45** (2001), 5-32.
[34] L. Breiman, *Manual-Setting Up, Using, And Understanding Random Forests*, University of California, Berkeley https://docplayer.net/44149058-Manual-setting-up-using-and-understanding-random-forests-v4-0.html.
[35] T. Cover, P. Hart, *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory, **13** (1967), 21-27.
[36] J. Han, M. Kamber, J. Pei, *Data Mining Concepts and Techniques*, 2011, Morgan Kaufmann.
[37] S. B. Kotsiantis,*Supervised machine learning: A riview of classification techniques*, Informatica, **31** (2007), 249-268.
[38] H. Bhavsar, A. Ganatra, *A comparative study of training algorithms for supervised machine learning*, International Journal of Soft Computing and Engineering, **2** (2012), 74-81.
[39] T. G. Dietterich, E. B. Kong, *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms*, Department of Computer Science, Oregon State University, Corvallis, 1995, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.2702&rep=rep1&type=pdf.
[40] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh , Q. Yang , H. Motoda , G. J. McLachlan , A. Ng , B. Liu , P. S.Yu , Z. Zhou , M. Steinbach , D. J. El & Dan Steinberg, *Top 10 Algorithms in Data Mining*, Knowledge Information System, **14** (2008), 1-37.
[41] A. E. Mohamed, *Comparative study of four supervised machine learning techniques for classification*, Int. J. App. Sci. Tech., **7** (2017), 5-18.
[42] Y. Saeys, I. Inza, P. Larranaga, *A review of feature selection techniques in bioinformatics*, Bioinformatics, **23** (2007), 2507-2517.