**SPORMETRE**
The Journal of Physical Education and Sport Sciences
Beden Eğitimi ve Spor Bilimleri Dergisi

**DOI: 10.33689/spormetre.1381602**

Geliş Tarihi (Received): 26.10.2023 Kabul Tarihi (Accepted): 02.04.2024 Online Yayın Tarihi (Published): 30.04.2024

# PERFORMANCE COMPARISON OF MACHINE LEARNING METHODS IN TURKISH SUPER LEAGUE MATCH RESULT PREDICTIONS

**Duygu Topcu[1], Özgül Vupa Çilengiroğlu[2*]**
[1]Dokuz Eylül Univerisity, The Graduate School of Natural and Applied Sciences, İZMİR
[2]Dokuz Eylül University, Faculty of Science, İZMİR

,

**Abstract:** The aim of this study is to determine, examine, interpret and compare the performances of the models formed by the most effective variables in predicting the results of the matches played in the Turkish Super League, using machine learning methods. For this purpose, 743 matches of 23 teams in the Turkish Football Super League were examined using data from the 2018-2021 seasons. The winning and losing situations of the teams were modeled using machine learning methods such as logistic regression, decision trees and random forest. The performances of the models were compared according to sensitivity, specificity, accuracy and F-score criteria. When the machine learning methods and models were compared, it was determined that the best model with 67.4% accuracy was the classification and regression trees (CART) with the variables "pozitive passing percentage of the opponent team", "offensive power of the home team" and "defensive power of the opponent team".

**Key Words:** Logistic regression, decision trees, random forest, offensive power, defensive power

# TÜRKİYE SÜPER LİGİ MAÇ SONUÇ TAHMİNLERİNDE MAKİNE ÖĞRENME YÖNTEMLERİNİN PERFORMANS KARŞILAŞTIRILMASI

**Öz:** Bu çalışmanın amacı, Türkiye Süper Ligi'nde oynanan maçların sonuçlarının tahmin edilmesinde en etkili değişkenlerin oluşturduğu modellerin performanslarını makine öğrenmesi yöntemlerini kullanarak belirlemek, incelemek, yorumlamak ve karşılaştırmaktır. Bu amaçla Türkiye Futbol Süper Liginde 2018-2021 sezonlarındaki veriler kullanılarak 23 takımın 743 maçı incelenmiştir. Takımların kazanma ve kaybetme durumları, lojistik regresyon, karar ağaçları ve rassal orman gibi makine öğrenme yöntemleri kullanılarak modellenmiştir. Modellerin performansları duyarlılık, seçicilik, doğruluk ve F-puanı kriterlerine göre karşılaştırılmıştır. Makine öğrenme yöntemleri ve modelleri karşılaştırıldığında "rakip takımın olumlu pas yüzdesi", "ev sahibi takımın hücum gücü" ve "rakip takımın savunma gücü" değişkenleri ile sınıflandırma ve regresyon ağaçları (CART) %67.4 doğrulukla en iyi model olarak belirlenmiştir.

**Anahtar Kelimeler:** Lojistik regresyon, karar ağaçları, rassal orman, savunma gücü, hücum gücü

** Bu çalışma, Özgül Vupa Çilengiroğlu danışmanlığında yürütülen Duygu Topcu'ya ait yüksek lisans tezinden üretilmiştir.*

## INTRODUCTION

All over the world, football is a sport that attracts the attention of different age groups as well as people from different social and cultural backgrounds (Bunker and Susnjak, 2019). It is estimated that football, the world's most popular sport, has more than 4 billion fans and spectators worldwide. It is played in more than 200 countries in the world and considered as a very difficult game to predict the result due to many unpredictable and influencing factors

[*] *Sorumlu Yazar: Özgül VUPA ÇİLENGİROĞLU, Doç. Dr., E-mail: ozgul.vupa@deu.edu.tr*

(Andrews et al., 2021). Predictions about football matches can be made using home, stadium, team strength, winning percentage, offensive or defensive strength, etc. Football predictions enables club staff to make the right decision regarding training and player management and to prepare teams for their future games based on their performance (Jawade et al., 2021). Predicting the results of a match and analyzing the parameters that affect the outcome is a common application of both machine learning and statistical methods in the field of sports analytics (Barron et al., 2020; Samba, 2019).

Machine learning is an approach that aims to discover information from uncertain, previously unknown, and thought to be beneficial information or knowledge from data (Witten and Frank, 2005). Nowadays, with the growth of data, machine learning has become one of the most important methods used in big data in computer science. These methods can be listed as logistic regression (LR), decision trees (DT), support vector machines (SVM), random forest (RF), Naive Bayes, k-nearest neighbors (KNN), clustering and artificial neural networks (ANN).

In recent years, machine learning methods have been used in many fields and have become a tool. Machine learning is now used to make humans' lives easier in a variety of areas, including engines, computer and software systems, digital machines, phones, sport game applications, the betting industry, medicine, health-care, security, entertainment, physical science, and computer engineering (C´wiklinski et al., 2021; Çimen, 2019; Singla and Singh, 2020). Moreover, machine learning is also considered one of the determinants of the future of data science. With this thought, adapting data science to sports, which is the most popular field in the world, has been an inevitable thought for data scientists (Andrews et al., 2021; Çali et al., 2013).

In the literature, it has been determined that different machine learning methods are used for football and other sports branches. When the articles containing machine learning algorithms used to predict sports results (football, basketball and so on) were reviewed, it was found that researchers mostly used data segmentation, k-cross evaluation, LR, DT and ANN (Horvat and Job, 2020; Özdemir and Ballı, 2020; Taspinar et al., 2021). In addition, since it produces models that can predict match results by taking into account the basic indicators (player, coach strategy, training, weather condition…), the suitability of machine learning methods in studies in the field of sports has also been demonstrated by different literature studies (Lotfi and Rebbouj, 2021). Football, one of the most popular sports, has been widely used in the leagues of different countries with many machine learning methods since 2020 (Ajgaonkar et al., 2021; Carloni et al., 2021; Coşkuner, et al., 2020; C´wiklinski et al., 2021; Taspinar et al., 2021).

Many factors affect the prediction of the football match result. These factors are the offensive and defensive powers of the teams, the field conditions (air temperature, season, etc.), the red or yellow cards received, the characteristics of the players, the number of assists and passes, missed shots, the tackles, the shots on target and the number of foreign players.
Hucaljuk and Rakipovic (2011) focused on estimating football scores based on various factors (number of injuries and goals, team design) using Bayesian Network, KNN, ANN, and RF methods in pre-2011 UEFA Champions League data. The most successful algorithm was ANN with 68.8% prediction accuracy.

Yezus (2014) focused on modeling English Premier League games using a set of 9 features as input data with KNN and RF methods. Model success rates were found to be 55.8% for KNN

and 63.4% for RF. Ulmer and Fernandez (2014) modeled the 2002-2012 English Premier League seasons for 4560 matches using machine learning algorithms. 50% accuracy for prediction was found with the SVM and RF models. Igiri et al. (2014) focused on developing a predictive model of English Premier League football match results for 110 matches played in the 2014-2015 season. Their proposed model system was implemented based on both ANN and logistic regression techniques, with 85% and 93% prediction accuracy, respectively. In the classification, goals scored by home and away teams, corners, attacking power, player and manager performances and winning streaks of the teams were used.

According to Karaoğlu's (2015) study, 16 football leagues with data from the 2013-2015 seasons in Europe were evaluated, and match results were estimated using machine learning methods such as Naive Bayes, multilayer perceptrons, logit boost, Bayesnet, decision trees, zeroR and C4.5 algorithms. As a result of the study, the decision tree algorithm achieved the highest success in 11 of the 16 leagues evaluated. This model performed 52% prediction accuracy.

Vaidya et al. (2016) modeled the English Premier League for 658 matches between the 2004-2015 seasons with LR, RF and Bayes algorithms and found prediction accuracy of 49%, 47% and 47%, respectively. Prasetio and Harlili (2016) calculated the matches between the 2010-2016 seasons in the Premier League with 68% and 69.5% accuracy by using LR method according to the defensive values of the home and away teams. Tüfekci (2016) predicted the results of 1222 matches in the Turkish Super League between the 2009-2013 seasons with SVM, bagging and RF algorithms. Results showed that RF achieved 70.61% prediction accuracy.

Herbinet (2018) predicted match results and scores with Naive Bayes, RF and SVM, using 25000 Champions League match data of 11 countries between 2008-2016. Ganesan and Harini (2018) modeled the SVM, XGBoost, and LR methods using 65 different features (the away team goals, venue, scores, and home team) and 5 seasons data to predict Barclays' English Premier League match results. Zaveri et al. (2018) used LR, RF, ANN, Naive Bayes and SVM methods to predict football match results for the 2012-2017 seasons. In the established models, it was found that LR reached the best prediction accuracy of 71.6%, and RF reached the second prediction accuracy of 69.9%.

Herold et al. (2019) emphasized that machine learning algorithms have the potential to create a revolutionary impact in the field of football analytics through tactics and the characteristics of the home and away teams. Bilek and Ulaş (2019) examined the winning of the match with ANOVA, k-means clustering and DT in 760 games in the English Premier League 2017-2018 season. It has been found that the most influential factor in each decision tree is the first score. It was determined that the rate of scoring the first goal in winning against DT was 0.45, 0.62 and 0.86 against strong, balanced and weak opponents, respectively. When the opposing team is not taken into account, this ratio is found to be 0.67. Alfredo and Isa (2019) studied English Premier League 2007-2017 season data and football match predictions with 3800 match data using tree-based model algorithms for 15 features. They found the best accuracy with RF with 68.55% and the worst accuracy with the C5.0 algorithm. Tewari et al. (2019) used LR, XgBoost and SVM models to predict the match outcome in the English Premier League. They decided that XgBoost is the best model according to the F score criterion.

Coşkuner et al.(2020), 18 teams and 612 matches in the Turkish Super League for the 2017-2018 season were examined. The number of goals scored, total shots, shots on target, number

of encounters with the ball, percentage of possession, total passes, correct passes, correct pass percentage, and field (home/away) features were discussed with LR. In the 2017-2018 Turkish Super League, the outcome of football matches was successfully predicted at a rate of 65%. Yıldız (2020) used DT, LR, regression trees, and RF methods for the classification of football teams with 400 match data obtained from the 2009-2019 seasons for Premier and La Liga Leagues. The results indicated that decision trees were able to classify football clubs with an accuracy rate of over 77%.

Ajgaonkar et al. (2021) decided that among the SVM, RF and Bayesian methods used on 3000 match data to predict English Premier League match results, SVM was the method that gave the best model with the highest accuracy (67%). Andrews et al. (2021) found that the logistic regression they used to predict match results in the 2015-2018 Premier League data gave better results than SVM and XG Boost. C'wiklinski et al. (2021) found 82% accuracy with the RF method, using the 2016-2019 seasons of the 8 most popular leagues of the UEFA rankings to predict player transfer success. Taşpınar et al. (2021) predicted the match results with 89.63% accuracy using LR for 2027 football match results of the Serie A League 2014-2015 and 2019-2020 seasons. Manish et al. (2021), the performances of 572 football players in the English Premier League 2018-2019 season were evaluated to predict the match outcomes with linear regression, SVM, ANN, Xgboast according to player positions. It has been determined that linear regression gives the best results in all positions according to $R^2$, MSE, RMSE, MAE criteria.

Rodrigues and Pinto (2022) calculated football match results for different machine learning algorithms over 1900 matches played in the English Premier League between the 2013-2019 seasons, especially taking into account the betting variable. In the models built by taking into account 18 variables, it was found that SVM was the best for predicting the match result with 61.32%. Haruna et al. (2022) examined LR, SVM, RF, KNN and Naïve Bayes models in English Premier League data, taking into account 760 match results between the 2011-2013 seasons. KNN was determined to be the best model in predicting the match result with an accuracy criterion of 83.95%.

The Super League, referred to as Turkish Professional Football League, is the top-tier professional football league in Türkiye. The aim of this study was to determine, analyze, and interpret the most effective variables in predicting the outcomes of matches played in the Turkish Super League employing machine learning methods, specifically decision trees, random forest, and logistic regression. The machine learning method was also decided depending on the performance criteria (accuracy, sensitivity, specificity, and F-score).

## METHODS

*Study Design and Sample*

Türkiye's first official league was formed in 1959, with 16 teams participating in the inaugural season under the name Milli League. Today, the league called “Super League” is the highest level football league in Türkiye. The Super League, which is affiliated with the Turkish Football Federation (TFF), is implemented as a double-stage league method in which 19 teams play two matches against each other in a season. This team number may vary from year to year. There are 38 weeks in the Super League, which lasts for nine months. Data related to the study were obtained directly from the public database on the official website of the TFF. A total of 743 match data were analyzed, taking into account the wins and losses from all matches played by 23 teams in 3 seasons (TFF, 2022).

According to the match results of the TFF 2018-2021 seasons, the team's win and loss situation was determined as a binary dependent variable (Y). While analyzing the collected data, the variables of home team in the match were named starting with "the home team" and the variables of the away team starting with "the away team". The percentages of success rates included in the data were obtained by considering the data of the teams in the season they are in.

The independent variables are abbreviated as follows: “Tackle success rate of the home team - TSRH”, “Shot on target percentage of the away team - STPA”, “Pozitive pass percentage of the home team - PPPH ”, “Pozitive pass percentage of the away team - PPPA”, “Whether the away team has received a red card - WRCA”, “The number of foreign players in the home team - NFPH”, “Offensive power of the home team - OPH” and “Defensive power of the away team - DPA”. Among these variables, TSRH, STPA, PPPH, PPPA variables are continuous, NFPH variable is discrete and WRCA, OPH and DPA variables are categorical variables.

*Statistical Analysis*

In the analysis of sports data, the descriptive statistics and frequency tables for variables were obtained. Super League match results were made by looking at performance criteria using machine learning algorithms (LR: Logistic Regression, DT: Decision Tree, RF: Random Forest). The study was carried out in IBM SPSS Statistics 24 and R 4.2.2. A margin of error of 5% was used for all tables and statistical tests.

Decision trees, consisting of the steps of "tree formation", "pruning" and "best (optimum) tree detection, are a method of separating large numbers of data into smaller datasets according to certain splitting criteria (Gini index, Information gain index, Twoing algorithm, Entropy, Chi-Square). Decision trees consist of roots, nodes, and branches. The root node, which contains all the variables in the dataset, branches and creates new nodes. When new branching does not occur, the tree becomes optimal. In addition, in order for the tree to be the best decision tree, it must be evaluated with independently selected test data after each pruning process (Wu and Kumar, 2009). While binary division is used in "The Classification and Regression Tree (CART)" algorithm, "The Chi-Squared Automatic Interaction Detector (CHAID)" algorithm has the feature of splitting into more than two subgroups (Díaz-Pérez and Cejas, 2016). In "The Quick, Unbiased, Efficient Statistical (QUEST)" algorithm, separate times are allocated for deciding the independent variable that will give the optimum division and the point at which the optimum division will be achieved during branching (Kuzey, 2012). CHAID and QUEST algorithm were used for both categorical (Chi-Square test) and continuous (F test) dependent variables. The obtained results in decision trees were expressed as a percentage.

The random forest (RF) method is an ensemble method that combines multiple decision trees to make predictions. In this method, each tree is first created independently and then combined to make a more accurate prediction. (Freund and Schapire, 1996).

Logistic regression (LR) is a modeling method used to determine the cause and effect relationship of the categorical dependent variable with other independent variables. With this modeling method, the estimated values of the dependent variable are calculated as probabilities and classification is made. In LR model obtained with the logit function, the maximum likelihood method is used for the estimation of the coefficients, Wald test is used

for the significance of the coefficients, and the odds ratio is used for the interpretation of the coefficients (Yavuz and Çilengiroğlu, 2020).

In Karaoğlu's (2015) study, offensive and defensive powers were calculated using the number of goals scored and conceded in the match as a variable. $Oi=OAGi \div LOAG$ and $Di=OYGi \div LOAG$. Where $Oi$=offensive power of team i, $Di$=defensive power of team i, $OAGi$=Average number of goals scored by team i, $OYGi$=Average number of goals conceded by team i, $LOAG$=Average goals scored/conceded by a team in the league. $LOAG=TG \div (TS*N)$. Where LOAG=Average goals scored by a team in the league, TG=Total goals scored in the league, TS=Number of teams in the league, N=Number of weeks evaluated.

All matches of the season were taken into account for the relevant calculations. The new variable values obtained for the offensive and defensive power of each team were categorized as above average and below average. From these categorical variables, the offensive power of the home team and the defensive power of the away team in the match were used in the models.

## RESULTS

The descriptive statistics and frequency tables of the variables affecting the match results (wins and losses) of the 23 teams playing in the Turkish Super League between 2018-2021 from TFF were given in Table 1. A total of 743 matches were examined. Of these matches, 444 resulted in victories and 299 in defeats. The mean, standard deviation, minimum and maximum values of the continuous variables (TSRH, STPA, PPPH, PPPA) were calculated. The frequency values for the discrete variables (WRCA, NFPH, OPH, DPA) and chi-square test results were given. The chi-square test was used to find the relationship between the categorical variables and the match result (wins and losses).

**Table 1** Descriptive statistics and frequencies of the variables

| **Variables** | **Win f(%) (n=444)** | **Loss f(%) (n=299)** | **Total f(%) (n=743)** | **p-value** | **Variables** | **Min-Max** | **Mean±Std** |
|---|---|---|---|---|---|---|---|
| **WRCA** | | | | | **TSRH** | | |
| **None** | 368 (82.9) | 274 (91.6) | 642 (88.4) | 0.001* | | 46.7-53.0 | 50.01±1.52 |
| **Available** | 76 (17.1) | 25 (8.4) | 101 (13.6) | | | | |
| **OPH** | | | | | **STPA** | | |
| **<Mean** | 185 (41.7) | 205 (68.6) | 390 (52.5) | 0.000* | | 38.0-52.9 | 45.38±3.10 |
| **>Mean** | 259 (58.3) | 94 (31.4) | 353 (47.5) | | | | |
| **DPA** | | | | | **PPPH** | | |
| **<Mean** | 171 (38.5) | 171 (57.2) | 342 (46.0) | 0.000* | | 73.8-85.6 | 80.01±2.93 |
| **>Mean** | 273 (61.5) | 128 (42.8) | 401 (54.0) | | | | |
| **NFPH** | | | | | **PPPA** | | |
| **≤6** | 118 (39.5) | 142 (27.3) | 260 (35.0) | 0.001* | | 73.80-85.60 | 80.02±2.89 |
| **>7** | 181 (60.5) | 378 (72.7) | 483 (65.0) | | | | |

*: p value is obtained from chi-square test < alpha=0.05

The situation of the away team receiving a red card during the match was found to be 13.6% (WRCA). The percentage of the away team receiving a red card (17.1%) in the match win was calculated to be higher than the percentage in the match loss (8.4%). In addition, a statistically significant relationship between the away team's red card and winning and losing the match was found with 95% confidence (p=0.001<0.05). The offensive power (OPH) of the home team was calculated above the average (58.3%) when the match was a win, but

below the average (68.6%) when the match was a loss. In addition, a statistically significant relationship was found between the offensive power of the home team and the winning and losing of the match with 95% confidence ($p=0.000<0.05$). Similar results can be made for the defense power of the away team. The defensive power (DPA) of the away team was calculated above the average (61.5%) when the match was a win, but below the average (57.2%) when the match was a loss. In addition, a statistically significant relationship was found between the defensive power of the away team and the winning and losing of the match with 95% confidence ($p=0.000<0.05$). The number of foreign players in the home team (NFPH) is considered categorically according to whether it is less than or more than 6. The mean tackle success rate of the home team (TSRH) was found to be 50.01. The minimum and maximum values of this variable were given as 46.7 and 53.0, respectively. Looking at the pass percentages, the maximum and minimum values of the home team (PPPH) and the away team (PPPA) were the same, with a minimum of 73.8 and a maximum of 86.6. It was also seen that there was no difference between the means for both cases in terms of pass percentages. The average shooting percentage on the away teams target was calculated as 45.38. The number of foreign players in the home team varies between 2 and 11. Some categories were combined for analysis (Table 1).

Model combinations were developed by taking into account the correlation of variables thought to be effective in determining whether a match would win or lose. With this first elimination, 13 models consisting of variables related to the match result but not related to each other were established, taking into account the correlation matrix. Later, these models were reduced to five basic models by using the decision tree classification method, one of the machine learning algorithms. Logistic regression, decision trees (CART, CHAID, QUEST), and random forest were used to generate these models. The data set was 85% training and 15% test data. The model was trained with the training data, and the prediction was made with the test data. Each model consists of 3 variables (Table 2).

**Table 2** Models derived from machine learning algorithms for match results

| Model | Variable 1 | Variable 2 | Variable 3 |
|---|---|---|---|
| **Model 1** | Pozitive pass percentage of the home team - PPPH | Offensive power of the home team - OPH | Defensive power of the away team - DPA |
| **Model 2** | Pozitive pass percentage of the away team - PPPA | Offensive power of the home team - OPH | Defensive power of the away team - DPA |
| **Model 3** | Shot on target percentage of the away team - STPA | Pozitive pass percentage of the home team - PPPH | Defensive power of the away team – DPA |
| **Model 4** | Tackle success rate of the home team -TSRH | Shot on target percentage of the away team - STPA | Whether the away team has received a red card - WRCA |
| **Model 5** | Pozitive pass percentage of the away team - PPPA | The number of foreign players in the home team - NFPH | Offensive power of the home team - OPH |

Repeated measurements were made in order to make comparisons in these models, which were established with 85% training and 15% test data. The accuracy, sensitivity, specificity and F-score values of the models, which were repeated 30 times by selecting different samples, were calculated. These calculated values for performance comparisons were averaged (Table 3).

Considering the performance criteria, it was decided that random forest and CART for model 1, CART for model 2, model 3 and model 5, and finally random forest for model 4 were the best. When the performance criteria for choosing the best model among the models were examined, it was determined that model 2 came to the fore. Although model 5 got high values

in sensitivity and specificity, it could not show the same performance when accuracy and F-score values were taken into account. For this reason, model-2 was decided to be used. Variables for the CART model were determined as “the pozitive pass percentage of the away team, PPPA”, “the offensive power of the home team, OPH,” and “the defensive power of the away team, DPA”. The tree formed as a result of this algorithm had 3 branches and a total of 9 nodes. Five of these nine nodes were terminal nodes. When the tree structure was examined, if the offensive power of the home team was below the average and the pozitive pass percentage of the away team was below 82.5%, the team's winning rate was 55.6%. If the offensive power of the home team was below the average and the pozitive pass percentage of the away team was below is above 82.5%, the team's winning rate droped to 22.7%. On the other hand, if the offensive power of the home team was above the average, if the defensive power of the away team was above the average, the team's winning rate was 78.9%. If the defensive power of the away team was below the average and the good pass percentage of the away teamwas over 78.35%, the team's winning rate droped to 58.1%.

**Table 3** Performance criteria for models

| | Model | Accuracy | Sensitivity | Specificity | F-Score |
|---|---|---|---|---|---|
| 1 | **CART** | ***0.676** | 0.815 | 0.582 | ***0.759** |
| | **CHAID** | 0.620 | 0.721 | 0.523 | 0.692 |
| | **QUEST** | 0.640 | 0.767 | 0.573 | 0.715 |
| | **Logistic Reg.** | 0.640 | 0.618 | 0.550 | 0.669 |
| | **Random F.** | 0.652 | ***0.824** | ***0.633** | 0.732 |
| 2 | **CART** | ***0.674** | ***0.948** | ***0.763** | ***0.777** |
| | **CHAID** | 0.627 | 0.855 | 0.617 | 0.728 |
| | **QUEST** | 0.656 | 0.870 | 0.623 | 0.754 |
| | **Logistic Reg.** | 0.615 | 0.574 | 0.522 | 0.637 |
| | **Random F.** | 0.668 | 0.876 | 0.665 | 0.759 |
| 3 | **CART** | ***0.672** | ***0.897** | 0.616 | ***0.774** |
| | **CHAID** | 0.635 | 0.790 | 0.558 | 0.720 |
| | **QUEST** | 0.627 | 0.827 | 0.547 | 0.727 |
| | **Logistic Reg.** | 0.620 | 0.607 | 0.531 | 0.652 |
| | **Random F.** | 0.645 | 0.843 | ***0.628** | 0.734 |
| 4 | **CART** | ***0.627** | 0.778 | 0.532 | 0.713 |
| | **CHAID** | 0.603 | 0.704 | 0.508 | 0.676 |
| | **QUEST** | 0.604 | 0.806 | 0.512 | 0.708 |
| | **Logistic Reg.** | 0.596 | 0.601 | 0.500 | 0.639 |
| | **Random F.** | 0.625 | ***0.907** | ***0.611** | ***0.741** |
| 5 | **CART** | ***0.666** | ***0.963** | ***0.812** | ***0.772** |
| | **CHAID** | 0.621 | 0.848 | 0.612 | 0.724 |
| | **QUEST** | 0.663 | 0.882 | 0.645 | 0.761 |
| | **Logistic Reg.** | 0.614 | 0.590 | 0.509 | 0.648 |
| | **Random F.** | 0.654 | 0.893 | 0.661 | 0.756 |

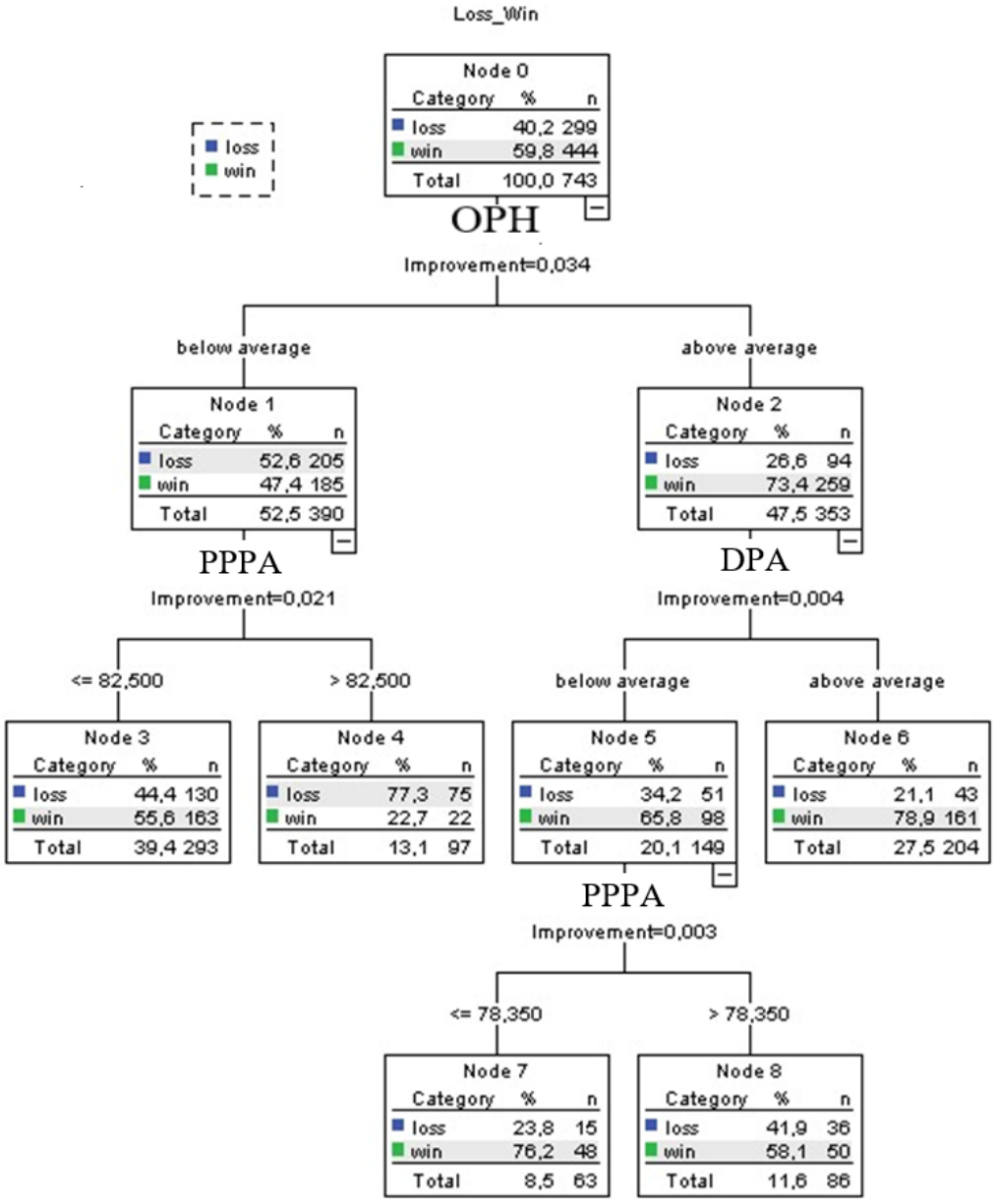


**Figure 1** CART algorithm for model 2

## 4. DISCUSSION

Machine learning algorithms are used in many disciplines and fields. According to reports, its usage area is quite wide in sports, especially football, which is popular in many countries of the world. However, the application of machine learning algorithms in sports science has gone beyond simply generating reports, allowing results to be predicted in real time, suggested alternatives, and new ideas implemented more quickly. Many scenarios can be created that can be applied in the field of sports, especially in classification and prediction. Inferences made about player performance and team efficiency, examination of factors to determine match results, sponsor bets and sports advertisements are the basic examples of these scenarios.

Classification used in machine learning methods is used to categorize new data samples, often into predefined classes or categories. In sports analytics, it helps organize and make sense of the information generated during each season, including data related to teams, matches, and players. Classification can be used to predict the outcomes of sports matches. By analyzing historical data and various factors, it's possible to estimate the likelihood of a team winning, losing, or drawing a match. In addition, it can assist in evaluating the performance of individual players and predictive models can be built to estimate the risk of player injuries. Finally, classification can help coaches analyze the strengths and weaknesses of their teams and opponents.

In this paper, a model is proposed to predict the outcome of football matches in the Turkish Super League. In modeling, the data set of the past seasons (2018-2021) was trained in various machine learning classifiers (Logistic regression, Decision Trees, Random Forest) and then tested. Comparisons between the algorithms were made by considering various performance criteria (sensitivity, specificity, accuracy and F-Score).

When the studies that modeled the football data by using decision trees, logistic regression and random forest methods in predicting the match results of recent years were examined, it was determined that there were different accuracy values depending on the leagues, variables and season range. Igiri et al. (2014) used ANN and LR models in football match predictions in the 2014-2015 season English Premier League data. The best models were found ANN model with 85% accuracy and LR model with 93% accuracy. However, the small number of variables in the models was given as a constraint. Karaoğlu (2015) established the DT model with 52% accuracy in the 2013-2015 Europa League. Vaidya et al. (2016) determined LR and RF models in predicting match results in the English Premier League between the 2004-2015 seasons, with accuracy values of 49% and 47%, respectively. Prasetio and Harlili (2016) found match result predictions for home and away teams in the 2010-2016 Premier League with LR models with 68% and 69.5% accuracy, respectively. Zaveri et al. (2018) proposed a solution for the prediction of football match results with the LR method (71.6%) and the RF method (69.9%) using Spanish La Liga data. Since these years, more studies have been found on decision trees, especially as visuality has come to the fore. Bilek and Ulaş (2019) found the prediction of the match with DT model according to 760 match outcomes in the English Premier League 2017-2018 season. Alfredo and Isa (2019) showed the football match prediction using the C5.0 algorithm (64.87%) and the RF method (68.55%) using English Premier League with 3800 match data. Yıldız (2020) showed that the accuracy rate was above 77% for each of the decision trees had a good performance in classifying football clubs. Andrews et al. (2021) found that the LR they used to predict match results in 2015-2018 Premier League data gave better results than SVM and XG Boost with 82% accuracy. C´wiklinski et al. (2021) predicted the player transfer success with the RF method (82%) using 8 most popular UEFA Leagues data. Taspinar et al. (2021) found the most effective features for predicting match results with the logistic regression method (89.63%) using Serie A League data. Hu and Fu (2022) used LR, Gradient Boosting Decision Tree and RF models to predict the match outcome in 2776 match data in the 2018-2022 season Premier and La Liga Leagues. In their study, they showed that RF was the best model according to $R^2$ (61.5%) and accuracy (63.8%) values.

Unlike the Europa League, Champions League and other leagues, Tüfekci (2016) and Çoşkuner et al. (2020) predicted the match results by using machine learning algorithms in the Turkish Super League research. Tüfekci (2016) calculated the results of the matches played in the Turkish Super League between the 2009-2013 seasons with 70.61% accuracy rate with the

random forest model. Coşkuner et al. (2020), using field features and match statistics, explained the match results with the 2017-2018 season Turkish Super League data with the logistic regression model with 65% accuracy rate.

Apart from these models, there are studies using other machine learning methods to predict match results. Hucaljuk and Rakipovic (2011) used the ANN model with 68.8% accuracy on Champions League data, Yezus (2014) used KNN and RF models with accuracy values of 55.8% and 63.4%, respectively, on English Premier League data. Ulmer and Fernandez (2014) established SMV and RF models with 50% accuracy in the English Premier League for the 2002-2012 seasons. Unlike accuracy value, Tewari et al. (2019) established the XgBoost model with the F score value to predict the match result in the English Premier League. Ajgaonkar et al. (2021) used the SVM model in the English Premier League with an accuracy value of 67%. Same year, Manish et al. (2021) found a linear regression model according to the $R^2$ value in predicting match results in the English Premier League for the 2018-2019 season. In the English Premier League, Rodrigues and Pinto (2022) used the SVM model with 61.32% accuracy in the 2013-2019 season, Haruna et al. (2022) used the KNN model to predict the match result with 83.95% accuracy in the 2011-2013 season.

In this study conducted for the Turkish Super League, the prediction of the match result was made depending on various variables by using DT, RF and LR models, which are among the machine learning methods. It was decided that the best model was CART for the variables "pozitive pass percentage of the away team", "offensive power of the home team" and "defensive power of the away team" with 67.4% accuracy from decision tree models in predicting match results for the years 2018-2021 in Turkish Super League. CART algorithm was found suitable. This study determined that the CART method is superior not only in terms of accuracy but also in terms of other performance criteria. Furthermore, it has been demonstrated that variables other than the independent variables used in the literature can be used.

## CONCLUSION

The designed model contains statistical results that help predict the winning team according to the chosen parameters. It can be taken in account for more seasons and different variables to increase the accuracy of the model. Showing the results in percentages with the decision tree models is more useful as it makes it easier for coaches, managers and football players to understand and interpret this model.

In future work, it is suggested that the proposed method can be used on other sports as well as other classification methods and compared with the methods in this article.

## REFERENCES

Ajgaonkar, Y., Bhoyar, K., Patil, A., & Shah, J. (2021). Prediction of winning team using machine learning. *I. J. of Engineering Research&Technology (IJERT) Special Issue*, *3*(3), 461-466.

Alfredo, Y. F., & Isa, S. M. (2019). Football match prediction with tree based model classification. *International Journal of Intelligent Systems and Applications*, *11*(7), 20-28. https://doi.org/10.5815/ijisa.2019.07.03

Andrews, S. K., Narayanan, K. L., Balasubadra, K., & Josephine, M. S. (2021, July). Analysis on sports data match result prediction using machine learning libraries. In *Journal of Physics: Conference Series* (Vol. 1964, No. 4, p. 042085). IOP Publishing.

Barron, D., Ball, G., Robins, M., & Sunderland, C. (2020). Identifying playing talent in professional football using artificial neural networks. *Journal of Sports Sciences*, *38*(11-12), 1211-1220. https://doi.org/10.1080/02640414.2019.1708036

Bilek, G., & Ulas, E. (2019). Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators. *International Journal of Performance Analysis in Sport*, *19*(6), 930-941. https://doi.org/ 10.1080/24748668.2019.1684773

Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, *73*, 1285-1322. https://doi.org/10.48550/arXiv.1912.11762

Carloni, L., De Angelis, A., Sansonetti, G., & Micarelli, A. (2021). A machine learning approach to football match result prediction. In *HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23* (pp. 473-480). Springer International Publishing. https://doi.org/10.1007/978-3-030-78642-7_63

Coşkuner, Z., Büyükçelebi, H., & Kurak, K. (2020). Analysis of in-game variables in Turkish Super League. *The J. of Germenica Physical Education and Sports Science, 1*(1), 46-54.

C´wiklinski, B., Giełczyk, A., & Choras, M. (2021). Who will score? A machine learning approach to supporting football team building and transfers. *Entropy, 23*(90), 1-12. https://doi.org/10.3390/e23010090

Çali, A., Gelecek, N., & Subasi, S. S. (2013). Non-specific low back pain in male professional football players in the Turkish super league. *Science & Sports*, *28*(4), e93-e98.

Çimen, E.A. (2019). *Prediction of the football match results with using machine learning algorithms*. Ms Thesis, Çankaya University, Computer Engineering Department.

Díaz-Pérez, F. M., & Bethencourt-Cejas, M. (2016). CHAID algorithm as an appropriate analytical method for tourism market segmentation. *Journal of Destination Marketing & Management*, *5*(3), 275-282. https://doi.org/10.1016/j.jdmm.2016.01.006

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. Machine learning: *Proceedings of the Thirteenth International Conference*, pp.148-156.

Ganesan, A., & Harini, M. (2018). English football prediction using machine learning classifiers. *I. J. of Pure and Applied Mathematics, 118*(22), 533-536.

Haruna, U., Maitama, J. Z., Mohammed, M., & Raj, R. G. (2021, November). Predicting the outcomes of football matches using machine learning approach. In *International Conference on Informatics and Intelligent Applications* (pp. 92-104). Cham: Springer International Publishing.

Herbinet, C. (2018). *Predicting football results using machine learning techniques*. Individual Project Report, Imperial College, Department of Computing Imperıal College of Science, Technology and Medicine, London.

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., & Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, *14*(6), 798-817. https://doi.org/ 10.1177/1747954119879350

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1380. https://doi.org/10.1002/widm.1380

Hu, S., & Fu, M. (2022, August). Football match results predicting by machine learning techniques. In *2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)* (pp. 72-76). IEEE.

Hucaljuk, J., & Rakipović, A. (2011, May). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO* (pp. 1623-1627). IEEE.

Igiri, R., Peace, C., Nwachukwu, A., & Okechukwu, E. (2014). An improved prediction system for football a match result. *IOSR journal of Engineering*, *4*(12), 12-20.

Jawade, I., Jadhav, R., Vaz, M. J., & Yamgekar, V. (2021). Predicting football match results using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *8*(7), 177-180.

Karaoğlu, B. (2015). Modelling sports games using machine learning. *TMMOB Elektrik Mühendisleri Odası,* 5(9), 1-5.

Kuzey, C. (2012). *Measuring the effect of knowledge workers on organization performance by using support vector machines and decision trees in data mining and an application.* Phd Thesis, İstanbul University, İstanbul.

Lotfi, S., & Rebbouj, M. (2021). Machine learning for sport results prediction using algorithms. International Journal of Information Technology and Applied Sciences. International *Journal of Information Technology, 3*(3), 148-155. https://doi.org/10.52502/ijitas.v3i3.114

Manish, S., Bhagat, V., & Pramila, R. (2021). Prediction of football players performance using machine learning and deep learning algorithms. *2nd International Conference for Emerging Technology (INCET), IEEE*, pp.1-5.

Özdemir, E., & Ballı, S. (2020). Prediction of Turkish Men's Basketball Super League game results with machine learning methods. *Journal of Engineering Sciences and Design, 8*(3), 740-752. https://doi.org/10.21923/jesd.723109

Prasetio, D. (2016, August). Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICAICTA.2016.7803111

Rodrigues, F., & Pinto, Â. (2022). Prediction of football match results with Machine Learning. *Procedia Computer Science*, *204*, 463-470. https://doi.org/10.1016/j.procs.2022.08.057

Samba, S. (2019). *Football result prediction by deep learning algorithms*. Ms Thesis, Tilburg University, School of Humanities and Digital Sciences Department of Cognitive Science & Artificial Intelligence, The Netherlands.

Singla, R., & Singh, A. (2020). Sports prediction using machine learning. *Journal of Emerging Technologies and Innovative Research (JETIR), 7*(10), 2759-2465.

Tewari, A., Parwani, T., Phanse, A., Sharma, A., & Shetty, A. (2019). Soccer analytics using machine learning. *International Journal of Computer Applications, 181*(50), 54–56. https://doi.org/10.5120/ijca2019918773

TFF, (2022). Turkish Football Federation Official Site. https://www.tff.org/

Taşpınar, Y. S., Çınar, İ., & Koklu, M. (2021). Improvement of Football Match Score Prediction by Selecting Effective Features for Italy Serie A League. *MANAS Journal of Engineering*, *9*(1), 1-9. https://doi.org/10.51354/mjen.802818

Tüfekci, P. (2016). Prediction of football match results in turkish super league games. In *Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015* (pp. 515-526). Springer International Publishing.

Ulmer, B., Fernandez, M., & Peterson, M. (2013). Predicting soccer match results in the english premier league. *Doctoral dissertation, Doctoral dissertation, Ph. D. dissertation, Stanford*.

Vaidya, S., Sanghavi, H., & Gevario, K. (2016). Football match winner prediction. *International Journal of Computer Applications, 154*(3), 31-33.

Witten, I. H., & Frank, E. (2005). *Data mining, practical machine learning tools and techniques*. Second Edition. Elsevier. ISBN: 9780080477022

Wu, X., & Kumar, V. (2009). CART: *Classification and regression trees, top ten algorithms in data mining*. First Edition. New York: Chapman and Hall.

Vupa Çilengiroğlu, Ö., & Yavuz, A. (2020). Comparison of predictive performance of logistic regression and CART methods for life satisfaction data. *European J Sci Tec*, *18*, 719-727. https://doi.org/10.31590/ejosat.691215

Yezus, A. (2014). *Predicting outcome of soccer matches using machine learning*. Mathematics and Mechanics Faculty Term Paper, Saint-Petersburg State University.

Yıldız, B. F. (2020). Applying decision tree techniques to classify European Football Teams. *Journal of Soft Computing and Artificial Intelligence*, *1*(2), 86-91.

Zaveri, N., Shah, U., Tiwari, S., Shinde, P., & Teli, L. K. (2018). Prediction of football match score and decision making process. *International Journal on Recent and Innovation Trends in Computing and Communication*, *6*(2), 162-165.