Prediction of Football Match Outcomes Based on Bookmaker Odds by Using k-Nearest Neighbor Algorithm

Engin Esme and Mustafa Servet Kiran

***Abstract*—Making predictions for sport competitions, which are followed by a large number of people, has always been an interesting field for sport fans, bettors, researchers etc. despite the complexity and uncertainty of many factors. The result of a sport competition is affected by many independent variables and factors. The number of variables that are included in the calculation affects the accuracy of the prediction. It is difficult for an ordinary punter to cope with these factors, which are in high numbers and have high complexity. On the other hand, a bookmaker must consider all the factors that might affect the result. When a bookmaker determines the odds with some delicate calculations, it is actually digitizing all the above-mentioned complex factors. In this way, the consistency of the betting odds of past competitions becomes a good indicator to be able to make predictions. In this study, a prediction model has been proposed for football, which is more common than the other branches of sport. In this model, the basic design approach was to measure the similarity between the competitions based on the betting odds. The model was enhanced with the performance data obtained using the previous games. Adding a risk analysis option to the model significantly decreased the margin of error in the predicted games. The super league of Turkey competition was used to test the model in which the k-nearest neighbor algorithm was preferred as the estimation technique.**

***Index Terms*—Bookmaker odds, estimation, football, k-nearest neighbor algorithm**

I. Introduction

Sports activities, which provide social, economic, and political acquisitions as a part of universal culture, have become an important part of human life. Today, sport is defined as a competitive activity whose rules are predefined. Since sport enables the player and viewer to experience satisfaction, it has become a field that has attracted the interest of millions of people.

Observing the performance parameters of competitions including techniques, tactics, conditions and motivation as well as the availability of historical data has enabled us to make predictions for the competition results. Meanwhile, the analyses also guide the trainer and sportsmen, and improve the performance of the game. For this reason, researchers have developed several models that calculate the possibility of predicting the possible results in various sports competitions [1]-[4]. Comparing the strength of the rival is one of the most basic methods used to predict the results of sport competitions. The Elo system developed by Arpad Elo and its derivatives provide certain criteria by assessing the performance of the rivals in individual or team games [5]-[7]. The possibility hypothesis presents various methods based on the historical competition data. The competitions of a tournament are individual events that are independent of each other. In other words, the result of a competition does not affect the result of another. The Poisson distribution is a discrete distribution used in possibility hypotheses and has been used in prediction models that are based on the scores [8], [9]. Some researchers handle the scores from the very beginning of the league, while some others use the results from several recent games played before the competition for which the predictions will be made [10]-[12]. Sports competitions are multi-variate systems. The scores or player performance may be good indicators; however, it is impossible that these factors provide adequate data for prediction. The place where the game is played, weather conditions, the intensity of the audience, special players, the position of the players, cards, injured players, the live broadcast status of the competition and similar parameters affect the result of the competition. For this reason, when researchers develop prediction models, they use statistical data as well as some subjective data such as psychological status, fatigue and motivation [13]. Hughes grouped the performance analysis models under four categories, which are empirical models, dynamic systems, statistical techniques and artificial intelligence [14].

Artificial Intelligence algorithms have become an alternative to other solutions in the classification and prediction requirements of systems in the field of economics, health, defense, industry etc. In sports competitions, there are several studies in which artificial intelligence techniques are used for the purpose of predicting the scores and results, and measuring performance. Joseph *et al.* suggested the expert Bayesian network to predict the results of the competitions (home, draw, away) for the Tottenham Hotspur football team for the period 1995-1997. In this study, data such as whether prominent players would play in a game or not, their positions on the field, the attacking force of the team and average team quality were used as the variables. The average classification success of the model was determined to be 59.21% [15]. Constantinou *et al.* designed a model that calculated the possibility of three situations for football competitions {p(H),p(D),p(A)}. In this empirical study, the results of 6244 competitions in the English Premier League for the period 1993/94-2009/10 were used as the training dataset and the results of the 380 games played in 2010/11 were predicted. The Bayesian network was designed for 4 components, which were the team strength, team form, psychological impact and fatigue. When the standard

Manuscript received November 17, 2017; revised February 3, 2018. This work was supported in part by the Scientific Research Project of Selcuk University.

The authors are with the Selcuk University, Engineering Faculty, Department of Computer Engineering, Konya, Turkey (E-mail: eesme@selcuk.edu.tr, mustafaservetkiran@gmail.com).

doi: 10.18178/ijmlc.2018.8.1.658

profitability measure was used, a profit was obtained between 2.87-9.48% [13]. Hucaljuk and Rakipovic tested the machine-learning algorithm to make predictions for three possible results for the championship league competition. Six different artificial intelligence algorithms were trained using 20 features derived from the data including the form of the team according to the last six games, the results of the previous competitions with the rivals, the status of the teams in ranking, the number of the injured players and the average number of goals scored per game for each team. A 65% success was achieved using the most successful classifier [16]. Aslan and Inceoglu trained two LVQ network models, whose entry vectors were different to predict the results of the competitions for the 2001/02 season of the Italy Serie A league. In this study, the first half of the season was used as the training dataset and the second half was used as the test data. According to the win, draw, or lose results in the first half of the season, the home and away field points were computed and a grading was obtained. The success obtained from the trained networks was 51.29% and 53.25%, respectively [17].

Bookmakers consider many other criteria, such as the weather conditions, the psychological status of the players, whether the team is in crisis or not and specialist viewpoints as well as the present statistical data, when it is making predictions for sports competitions. When defining the fixed odds, bookmaker specialists digitalize all these criteria. For this reason, in order to predict fixed odds, they may be used as an important indicator. The researchers assess the betting odds in the algorithms they develop. Strumbelj used the odds, which were determined by the bookmakers for certain competitions in various branches of sports. Four different methods, basic normalization, Shin probabilities [18] and two types of regression approaches, were compared in the study. As a result of this study, it was determined that the Shin probability method could provide more accurate predictions with the rates of some bookmakers when compared with those of other bookmakers [19]. Dixon and Coles computed the positive expected value parameter in order to detect the games on which they could bet in the prediction model they proposed. The probabilities computed using the Poisson distribution were compared with the bookmaker odds and the games to bet on were determined. In this study, the games to bet on during the 1995-1996 season were selected using the data obtained for the British premier league for 1992-1995 [8]. Odachowski and Grekow worked on a prediction model based on the change in the betting odds of the approaching competitions. The 10-hour period before the competition was examined as 3 periods. In this process, the odds were sampled at 10-min intervals. 360 features were produced from the statistical data and from the changes between the odds and the start value, end value, the difference between, the angle with the horizontal axis and standard deviation. Using the binary classification approach, 6 different classification algorithms were tested. In this study, the relationship between the betting odds changes were used and a 70% success was achieved [20].

In the present study, an approach based on the classification function of learning algorithms has been developed to predict the results of football competitions. The main idea, the structure of the algorithm used in the model and the feature vectors are explained in the second section of the paper. In the third section, the test results obtained in the study are explained. An interpretation of the test results is given in the fourth section.

II. Materials And Methods

Some bettors consider the past competitions in which similar conditions existed when they are making a prediction for a competition. In this context, the k-nearest neighbor algorithm has been preferred in the present study since the similarities between the competitions have been investigated in the developed model. The purpose of the designed model is to reveal the past competitions that have similar characteristics with the competition to predict. The model developed attempts to foresee the outcome of the competition (home, draw, away) using some of the historical statistical data and the betting odds, as-explained in Section 2.1. In addition, it performs a risk analysis based on the bookmakers' odds. According to the analysis, if the three possible results have a probability of being realized close to each other, the model avoids the prediction. The values for the k parameter between 1 and 20 in the algorithm have been studied in all the tests.

A. k-Nearest Neighbor Algorithm

The k-nearest neighbor algorithm (kNN) is a sample-based classifier algorithm reported by Fix and Hodges in 1951 [21]. Sample-based classifier methods are based on predicting the class of the new pattern over the patterns whose classes are already known in the training set. The kNN method seeks for similarity in the patterns of the new pattern and in the training set. Firstly, all the patterns are made to overlap with a point in the n-dimensional space. The similarity is found by computing the distance between the feature vectors of the patterns whose classes will be found, and the patterns in the training set. The distance between the feature vector of the pattern X^u whose class is not known, and the feature vectors X_i^j of all the patterns in the training set are computed as shown in Equation 1.

$$dist_i = \left\| X^u - X_i^j \right\| \quad (1)$$

Various metrics are used to compute the linear distance between two points [22]. In the k-nearest neighbor algorithm approach, the class of the pattern is accepted as the class of the closest neighbor. If the class of the pattern whose class is defined and the class of the nearest neighbor are different, an error occurs. In order to reduce the error rates, the k-nearest neighbor approach is preferred. The class of the pattern is determined by making a generality calculation between the nearest k-number neighbors. Taking k as a small value increases the effect of the nearest neighbors. The steps of the algorithm are explained as follows:

Define the k value
Compute the distance between the new sample and all the samples in the training set
Rank the distances from the lowest to the highest
Define the nearest k-number neighbors
Define the class of the new sample with majority voting

B. Dataset

In the present study, the football data between the 2010/11 and 2015/16 seasons for the super league of Turkey were used. Eighteen teams compete in the super league of Turkey. The competitions last for 34 weeks, 17 weeks in the first half and 17 weeks in the second half. A total of 306 competitions are performed over the 34 week season with 9 in each week. The data for the 2010/11-2014/15 seasons were used only to compute the frequency feature. The first half of the 2015/16 season was used as the training set and the second half of the season was used as the test set. Using 8 different types of data that may be computed for all the training and test samples, a total of 17 features were obtained, 10 of which were based on betting odds and 7 of which were based on other statistical data. The data for the competitions were obtained from www.mackolik.com.

1) The name of the team (name): The brand values of the teams are important indicators used to predict the results of competitions. Some teams like Barcelona, Chelsea and Galatasaray have become popular due to their long-term success and are well-known by a large number of people. Even a person who is not interested in football may make an accurate prediction based on the name of the team. For this reason, each team in the league was assigned numbers between 1 and 18, and was included in the similarity survey.
2) The cost of the team (cost): This is the value computed using the market value of the footballers in the team. Although high transfer expenses do not always bring high success, there is a significant relationship between the market value of the team and their performance on the field [23].
3) Standard deviation and probability based on fixed odds (Bet): The probabilities based on full time result odds defined by the bookmaker for the competitions were scaled between 0-1, and the normalized probabilities were obtained. If the fixed odds declared by the bookmaker for n number of possible results of a competition {Home, Draw, Away} is shown by:

$$o = (o_1, o_2, \dots, o_n) \tag{2}$$

and if the inverse of the betting odds are scaled in the range of 0-1, the probabilities of the possible outcomes are obtained using Equation 3.

$$P_{odds_i} = \frac{o_i^{-1}}{\sum_{i=1}^{n} o_i^{-1}} \tag{3}$$

where, i is the outcome of a competition {Home, Draw, Away} and $\overline{P_{odds}}$ is the average of the full time result probabilities. The standard deviation was calculated using Equation 4.

$$\sigma_{odds} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_{odds_i} - \overline{P_{odds}}\right)^2} \tag{4}$$

4) The relative frequency and standard deviation (frequency): This is the frequency percentage of the betting odds on the full time match declared by the bookmaker in the super league of Turkey since 2010. A total of 352 different fixed odds have been written in the super league in the last 5 years. The frequency percentages for each betting odd were calculated using Equation 5.

$$P_{freq_i} = \frac{f_i}{\sum_{i=1}^{n} f_i} \tag{5}$$

where f is the frequency of the betting odds on the outcomes, P_{freq} is the frequency percentage of the full time result betting odds and $\overline{P_{freq}}$ is taken as the average of the frequency percentage. The standard deviation was calculated using Equation 6.

$$\sigma_{freq} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_{freq_i} - \overline{P_{freq}})^2} \tag{6}$$

5) Consistency: This is the Euclidean distance between the probabilities of a competition computed using betting odds and the probabilities computed using the frequency of the betting odds. The calculation given in Equation 7 measures the consistency between the two predictions.

$$\left\|P_{odds} - P_{freq}\right\| = \sqrt{\sum_{i=1}^{n}\left(P_{odds_i} - P_{freq_i}\right)^2} \tag{7}$$

6) The marks: This is the scoring based on the points received by the teams as a result of the competitions. The team winning a competition receives 3 points, the loser receives 0, and when the game ends in a draw, both teams receive 1 point. The points received by the players in the competition to be predicted were calculated using the following pseudocode. In this code, $team_i$ represents the i^{th} team and $mark_i$ represents the point of the $team_i$.

Initial value of the teams' marks are zero
Repeat each of the previous events of each team in the season
if $team_i$ *wins* $mark_i$ += 3
if $team_i$ *draws* $mark_i$ += 1
if $team_i$ *loses* $mark_i$ += 0
until the predicted event

7) The intersection point: This is the set of data computed by considering the common rivals of the teams in the previous competitions. The n number of competitions in which the common rivals are away played in recent times is defined. The betting odds for the defined competition $odds = \{o_{\text{Home}}, o_{\text{Draw}}, o_{\text{Away}}\}$ is equalized according to the betting odds for the away team (o_{Away}) using Equation 8.

$$odds_i = \frac{o_i}{o_{\text{Away}}}, \{i = 1,2,3\} \tag{8}$$

The addition of the multiplication of the number of the goals of the teams by the inverse of the betting odds was computed according to Equation 9. This sum provides us with a scoring according to the past competitions of the teams with their common rivals. The multiplication of the expected probability of the past competition by the number of goals affect the superiority of the team in a positive way in the case the expectation is realized. In the case the expectation does not happen, it affects the superiority of the team in a negative way. The purpose here is to support the team that is favorite

in the case the expected probability occurs, and to support the team that is not favorite in the case the expected probability does not occur. The intersection point of the team is given below:

$$\text{Intersection}_{\text{team}} = \sum_{i=1}^{n} \left(\text{scored}_{\text{Home}}^{i} * \left(odds_{\text{Home}}^{i} \right)^{-1} - conceded\text{Home}i * odds\text{Away}i-1 \right) \quad (9)$$

where, $\text{scored}_{\text{Home}}^{i}$ represents the goals of the home team scored in the previous competitions, $conceded_{\text{Home}}^{i}$ represents the number of goals conceded, $\left(odds_{\text{Home}}^{i}\right)^{-1}$ represents the probability of the home team winning in the previous competition, $\left(odds_{\text{Away}}^{i}\right)^{-1}$ represents the probability of the away team winning in the previous competition, n represents the number of past competitions with the common rivals and i = (1, 2, .., n) represents the past competition in the ith order. This sum is computed for two players in the competition to be predicted and named the "team intersection point".

The intersection point of the competition is the difference between the intersection points of the teams. A difference close to zero shows that the teams are equal to each other, if it is positive, this shows that the home team is stronger and if it is negative, this shows that the away team is stronger. The intersection point of the competition according to Equation 10 is as follows:

$$\text{Intersection}_{\text{match}} = \text{Intersection}_{\text{Home}} - \text{Intersection}_{\text{Away}} \quad (10)$$

In this equation, the intersection point of the home team in the competition to be predicted is shown by $\text{Intersection}_{\text{Home}}$ and that of the away team is shown by $\text{Intersection}_{\text{Away}}$, and the intersection point of the competition is shown by $\text{Intersection}_{\text{Match}}$.

8) The rating of the competition: This is a scoring based on the number of the goals, and is the difference between the goals scored and conceded from the start of the season until the competition in question. The number of the goals scored and conceded was calculated using the following pseudocode.

Initial values of the teams' ratings are zero
Repeat each previous events of each team in the predicted event
Total goals scored by the team$_{Home}$ += goals scored by the team$_{Home}$ in the event$_i$
total goals conceded by the team$_{Home}$ += goals conceded by the team$_{Home}$ in the event$_i$
total goals scored by the team$_{Away}$ += goals scored by the team$_{Away}$ in the event$_i$
total goals conceded by the team$_{Away}$ += goals conceded by the team$_{Away}$ in the event$_i$
until the predicted event

The difference between the ratings for the home and away teams gives the rating of the competition, as given in Equation 11. If the goal superiority of the team is shown with the rating:

$$\begin{aligned} \text{rating}_{\text{Home}} &= tgs_{\text{Home}} - tgc_{\text{Home}} \\ rating_{Away} &= tgs_{\text{Away}} - tgc_{\text{Away}} \\ \text{rating}_{\text{Match}} &= \text{rating}_{\text{Home}} - \text{rating}_{\text{Away}} \end{aligned} \quad (11)$$

C. Risk analysis

The risk level expresses the closeness of the probability of winning by the players of the competition to be predicted. The standard deviation of the betting odds foreseen for the probable results of a competition odds = {o_{Home}, o_{Draw}, o_{Away}} is an indicator of the σ_{odds} risk level. As the probability values of the possible results move close to each other, the standard deviation moves closer to zero, and as the probability values move away from each other, the standard deviation become larger. In the case the standard deviation of the odds of a competition is smaller than the defined threshold value, avoiding the prediction increases the rate of the competitions for which accurate predictions are made.

$$\text{class}_i = \begin{cases} min\ odds_{\text{Team}}^{i}, & \sigma_{odds}^{i} \geq \rho \\ \text{do not guess}, & \sigma_{odds}^{i} < \rho \end{cases} \quad (12)$$

where class_i refers to the result of the predicted competition i^{th} $\{\text{home}, \text{draw}, \text{away}\}$, $min\ odds_{\text{Team}}^{i}$ refers to the favorite team of the competition, σ_{odds}^{i} shows the standard deviation of the fixed odds of the competition and ρ refers to the threshold value.

III. Experimental Results

In this study, the second half competition results in the super league of Turkey were predicted using the data obtained during the first half of the 2015/16 season. The success of the developed model was measured using the probability declared by the bookmaker as a reference. In other words, with a standard viewpoint, the bettor bets on the highest outcome in a competition based on the odds declared by the bookmaker. This approach reveals the number of the competitions predicted accurately by the bookmaker. For this reason, two types of betting games, which were full time result and double chance, were used to compare the success of the accurate predictions.

Full time result betting is a game in which predictions are made on how a competition will end $\{\text{home}, \text{draw}, \text{away}\}$. Full time result betting odds, on the other hand, reflects the probability values foreseen for the possible results of the competition.

TABLE I: The Prediction Results in the Case the Result with the Highest Probability Is Chosen according to the Bookmaker Odds.

Bet type	Number of events predicted	Number of correct predictions	Percentage of correct predictions
Full time result	153	85	55.56
Double chance	153	120	78.43

Double chance betting enables the bettor to make two predictions simultaneously. However, the betting odds are smaller. We must consider that the commission of the bookmaker is also included in these rates. The success of the prediction obtained according to the odds declared by the bookmaker is given in Table I. According to the results,

55.56% of the full time results of the competitions may be predicted accurately in a selection based on the betting odds without having an idea on any of the competitions.

D. Making predictions by ignoring the risk analysis option

In our empirical study, a prediction approach was developed using the kNN algorithm with 17 features from 8 types of data produced from the betting odds and the scores of the previous competitions. It is a wonder how well the features represent the system. For this reason, the predictions made using the features individually (one-by-one) in both game types were examined and the success of the highest accurate classification obtained for each feature is given in Table II.

TABLE II: THE PREDICTION RESULTS OBTAINED USING THE FEATURE VECTORS SEPARATELY.

		Full time result			Double chance		
Features	Number of events predicted	k	Number of correct predictions	Percentage of correct predictions	k	Number of correct predictions	Percentage of correct predictions
Name	153	9	69	45.10	1	119	77.78
Cost	153	19	73	47.71	11	119	77.78
Bet	153	**14**	**83**	**54.25**	9	120	78.43
Frequency	153	14	75	49.02	20	123	80.39
Intersection	153	11	79	51.63	8	123	80.39
Rating	153	4	76	49.67	7	118	77.12
Mark	153	2	81	52.94	**1**	**127**	**83.01**
Consistency	153	2	69	45.10	2	117	76.47

TABLE III: THE PREDICTION RESULTS OBTAINED WITHOUT RISK ANALYSIS.

	Full time result			Double chance		
Number of events predicted	k	Number of correct predictions	Percentage of correct predictions	k	Number of correct predictions	Percentage of correct predictions
153	18	89	**57.52**	5	132	**86.27**
153	19	86	56.21	6	130	84.97
153	20	86	56.21	3	127	83.01
153	12	85	55.56	7	127	83.01
153	15	85	55.56	9	126	82.35

TABLE IV: THE PREDICTION RESULTS IN THE CASE THE RESULT WITH THE HIGHEST PROBABILITY IS CHOSEN ACCORDING TO THE BOOKMAKER ODDS.

	Full time result			Double chance		
Threshold Value	Number of events predicted	Number of correct predictions	Percentage of correct predictions	Number of events predicted	Number of correct predictions	Percentage of correct predictions
4	12	12	100	10	10	100
3	15	15	100	27	27	100
2	19	19	100	31	31	100
1	18	18	100	39	39	100
0.9	24	22	91.67	37	37	100
0.8	28	25	89.29	32	32	100
0.7	42	33	78.57	45	44	97.78
0.6	111	72	64.86	111	98	88.29
0.5	153	87	56.86	145	125	86.21
0.4	153	88	57.52	153	132	86.27
0	153	88	57.52	153	132	86.27

Comparing the results in Table I and Table II, it was observed that the features are extremely close to the success of the bookmaker in the full time result game; in the double chance game, some of the features make better predictions than the bookmaker, even on their own. For this reason, it is possible to claim that these features are suitable for prediction. As expected, the use of the features together has increased the success of the prediction in the developed model.

In the first step of the empirical study, the prediction was made for all of the 153 competitions played in the second half of the season without using a risk analysis. The best 5 prediction results obtained using the k-values between 1-20 in the full time result and double chance games are listed in Table 3. The best prediction success obtained in the full time result game was 57.52%, which means a better result at a rate of nearly 1.96%; the best prediction success obtained in the double chance game was 86.27%, which means a better result at a rate of 7.84%. Fig. 1 shows the difference between the correct predictions of the bookmaker and the proposed model without using the risk analysis.

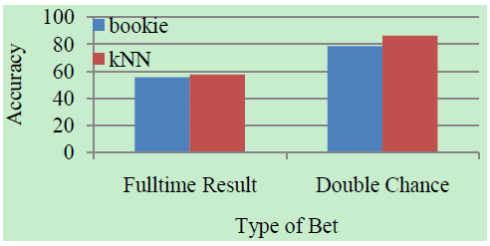

Fig. 1. A comparison of the performance of the bookmaker and model without risk analysis.

E. Making predictions Using the Risk Analysis Option

The standard deviation of the betting odds of all the competitions in the super league of Turkey during the 2015/16 season was within the $0.32 \leq \sigma_{odds} \leq 4.99$ range.

Selecting the risk as high means making predictions for teams that are equal to each other, in other words, more competitions are predicted. Selecting the risk as low means avoiding making predictions for teams that are equal to each other, in other words, fewer competitions are predicted.

The high values of the threshold value mean low risk and the low values of the threshold value mean high risk. The prediction results made in the full time result and double chance games for the different threshold values ranging between 0-4 are given in Table IV and Table V, respectively. Increasing the threshold value decreases the number of competitions for which the predictions were made and increased the number of competitions that were predicted accurately. In this way, it is possible to catch competitions that are known as "*banco*" by bettors. For example, according to Table IV, predictions were made for 19 competitions among the 153 competitions for the threshold value in the full time result game, and all of these were predicted accurately. Again, in the full time result game, 42 predictions were made among the 153 competitions for a threshold value of 0.7, and 33 of them were predicted accurately. In this way, the accurate prediction rate was improved up to 21.05%. On the other hand, in the double chance game, whose results are shown in Table IV, nearly all of the competitions were predicted accurately for a threshold value of 0.7 or higher. Fig. 2 shows the success rates of the predictions using various threshold values.

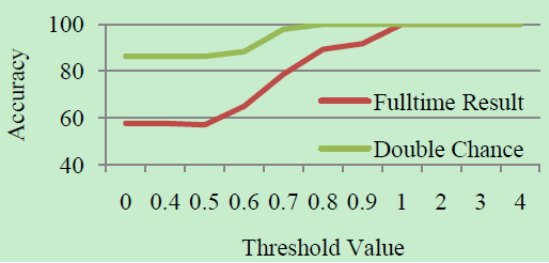

Fig. 2. The performance of the model using various threshold values.

IV. Conclusions

Statistical analyses are made on a great deal of factors such as the fined players in the competitions, the injured players, the motivational status of the team and management affairs reported in the sports pages of newspapers, internet sites with sports content and in the sports programs of television and radio channels. Foreseeing the result of a competition may be a way of earning money and satisfaction for a bettor. However, it is also important for a team, its players and the management in terms of developing tactics and increasing performance.

Researchers have developed several models using statistical data or artificial intelligence techniques to make predictions on sports events that have attracted the attention of a large number of people. Sports events are multi-variate systems and the success of the prediction increases in a direct proportion with the number of variables analyzed. It is impossible for a bettor to follow many variables about the competitions. However, many bookmakers may make more successful analyses using the specialists in their organizations. The fact that the betting odds, which appear as a result of the intense analyses made by the bookmaker, being numerical data representing all these variables constitutes the basic idea of this study. Another basis of this study is to make a prediction by determining how past competitions with similar conditions ended. The kNN algorithm, which is a similarity-based method, was preferred in this study because the similarity was investigated with past competitions. A total of 17 feature vectors have been derived from 8 types of data, which could be computed for all the training and test samples. The success of the prediction results using the developed model was assessed by taking the probability rates of the bookmaker as reference. Without using risk analysis, the developed model made better predictions at a rate of 1.96% according to full time result betting and at a rate of 7.84% according to double chance betting. When the results of the prediction were examined, it was observed that the number of the erroneous predictions was high in the competitions that had high-risk levels. For this reason, the option "make prediction according to the risk level" was added to the model. Avoiding predictions for risky competitions decreased the number of competitions for which predictions were made and increased the accurate prediction rate; adjusting the risk level at a very low level ensured that the competitions known as "banco" were selected. The most important handicap of the model is its success depending on the consistency of the predictions of the bookmaker. If the bookmaker does not declare the betting odds in accordance with the possible expectation or if it does not distribute its commission to the odds in a fair manner, the success of the model is affected negatively. However, this situation paves the way to opportunities for bettors who follow the sports events in a regular manner. In this context, the situation called rate chicane appears, which will be the subject matter of another research study.

References

[1] G. Cheng, Z.Y. Zhang, M. N. Kyebambe, and N. Kimbugwe, "Predicting the outcome of NBA playoffs based on the maximum entropy principle," *Entropy-Switz*, vol. 18, 2016.

[2] S. Lessmann, M. C. Sung, and J. E. V. Johnson, "Alternative methods of predicting competitive events: An application in horserace betting markets," *Int J Forecasting*, vol. 26, pp. 518-536, 2010.

[3] D. F. Percy, "Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes," *J Oper Res Soc*, vol. 66, pp. 1840-1849, 2015.

[4] D. Spanias and W. J. Knottenbelt, "Predicting the outcomes of tennis matches using a low-level point model," *Ima J Manag Math*, vol. 24, pp. 311-320, 2013.

[5] A. E. Elo and S. Sloan, *The Rating of Chess Players, Past and Present*, Ishi Press International, 2008.

[6] L. M. Hvattum and H. Arntzen, "Using ELO ratings for match result prediction in association football," *Int J Forecasting*, vol. 26 pp. 460-470, 2010.

[7] C. K. Leung and K.W. Joseph, "Sports data mining: predicting results for the college football games," *Procedia Comput Sci*, vol. 35, pp. 710-719, 2014.

[8] M. J. Dixon and S. G. Coles, "Modelling association football scores and inefficiencies in the football betting market," *Appl Stat-J Roy St C*, vol. 46, pp. 265-280, 1997.

[9] D. Karlis and N. L. tzoufras, "Analysis of sports data by using bivariate Poisson models," *J Roy Stat Soc D-Sta*, vol. 52, pp. 381-393, 2003.

[10] T. Y. Cheng, D. G. Cui, Z. M. Fan, J. Zhou, and S. W. Lu, "A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system," in *Proc. Fifth International Conference on Computational Intelligence and Multimedia Applications*, 2003, pp. 308-313.

[11] B. Hunter, "Football fortunes," *Oldcastle Books*, 2005.

[12] T. X. Cui, J. P. Li, J. R. Woodward, and A. J. Parkes, "An ensemble based genetic programming system to predict english football premier league games," in *Proc. IEEE Conf Evol Adapt*, 2013, pp. 138-143.

[13] A. C. Constantinou, N. E. Fenton, and M. Neil, "PI-football: A Bayesian network model for forecasting Association Football match outcomes," *Knowl-Based Syst*, vol. 36, pp. 322-339, 2012.
[14] M. Hughes and I. Franks, *The Essentials of Performance Analysis*, Routledge, 2007.
[15] A. Joseph, N. E. Fenton, and M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques," *Knowl-Based Syst*, vol. 19, pp. 544-553, 2006.
[16] J. Hucaljuk and A. Rakipović, "Predicting football scores using machine learning techniques," in *Proc. the 34th International Convention MIPRO*, 2011, pp. 1623-1627.
[17] B. G. Aslan and M. M. Inceoglu, "A comparative study on neural network based soccer result prediction," in *Proc. the 7th International Conference on Intelligent Systems Design and Applications*, 2007, p. 545.
[18] H. S. Shin, "Measuring the incidence of insider trading in a market for state-contingent claims," *Econ J*, vol. 103, pp. 1141-1153, 1993.
[19] E. Strumbelj, "On determining probability forecasts from betting odds," *Int J Forecasting*, vol. 30, pp. 934-943, 2014.
[20] K. Odachowski and J. Grekow, "Predicting the final result of sporting events based on changes in bookmaker odds," *Front Artif Intel Ap*, vol. 243, pp. 278-287, 2012.
[21] E. Fix and J. L. Hodges, "Discriminatory analysis-nonparametric discrimination: consistency properties," *DTIC Document*, 1951.
[22] P. Perner, *SpringerLink (Online service)*, in *Proc. 7th International Conference on Machine Learning and Data Mining in Pattern Recognition*, New York, NY, USA, September 3, 2011.
[23] M. Pinnuck and B. Potter, "Impact of on-field football success on the off-field financial performance of AFL football clubs," *Account Finance*, vol. 46, pp. 499-517, 2006.

Engin Eşme was born in Turkey in 1981. He received his bachelor degree at the Electronics Department of Technical Education Faculty in 2002 from Sakarya University. He received his M.S. degree from Selcuk University in 2006. He is currently a Ph.D. in computer engineering at Selcuk University. His research interest is in the area of machine learning and their applications.

Mustafa Servet Kıran was born in Osmaniye, Turkey in 1983. He received his bachelor, master and Ph.D. degrees in Computer Engineering Department at Selcuk University. His PhD thesis is based on swarm intelligence and its applications. He still works at the same department as an Associate Professor and his research topics are artificial intelligence, machine learning, evolutionary computation and swarm intelligence.