Spor ve Performans Araştırmaları Dergisi
Journal of Sports and Performance Researches
http://dergipark.gov.tr/omuspd

Araştırma Makalesi

Geliş Tarihi/Received : 03.01.2019
Kabul Tarihi/Accepted : 28.11.2019
DOI: 10.17155/omuspd.507797

THE EFFECTIVENESS OF DIFFERENT MACHINE LEARNING ALGORITHMS ON BASKETBALL PLAYERS' SHOOTING PERFORMANCE

Serpil KILIÇ DEPREN[1]

ABSTRACT

The main purpose of this study is to determine which factors have an important role in National Basketball Association (NBA) players' shooting accuracy. To achieve this purpose, player-based raw-dataset for each match on the 2014-2015 NBA season is used in this study. Seven different machine learning algorithms are applied and also 10-fold cross-validation with 10-repeat process is performed to avoid the overfitting problem. Nine independent variables and one binary dependent variable are included in the analysis. According to the results of the analysis, k-nearest neighbor algorithm is the best machine learning algorithm among other algorithms that are used in the analysis in order to predict whether basketball player can make a shot or not. Shot Distance, distance of closest defense player and touch time are identified as the most important factors affecting player's successful field goal accuracy. Since the successful field goal performance is very influential in winning the game, the results of this study can be used as a guide for training programs to basketball players and team coaches.

Keywords: Basketball games, classification techniques, machine learning algorithms, performance analysis

FARKLI MAKİNE ÖĞRENMESİ ALGORİTMALARININ BASKETBOL OYUNCULARININ ATIŞ PERFORMANSI ÜZERİNDEKİ ETKİNLİĞİ

ÖZET

Bu çalışmanın temel amacı, National Basketball Association (NBA) oyuncularının atış isabeti üzerinde hangi faktörlerin önemli bir rolü olduğunu belirlemektir. Bu amaca ulaşmak için, çalışmada 2014-2015 NBA sezonunda oynanan her bir maç için oyuncu bazlı ham veri seti kullanılmıştır. Yedi farklı makine öğrenme algoritması uygulanmış ve aynı zamanda aşırı uyum problemini önlemek için 10 kat çapraz geçerlilik prosedürü 10 defa tekrar edilmiştir. Analizde dokuz adet bağımsız değişken ve bir ikili bağımlı değişken kullanılmıştır. Bir basketbol oyuncusunun başarılı bir atış yapıp yapamayacağını tahmin etmek için kullanılan algoritmalar arasında en başarılı makine öğrenme algoritması k-en yakın komşu algoritmasıdır. Atış Mesafesi, en yakın savunma oyuncusunun mesafesi ve temas süresi oyuncunun başarılı bir atış yapmasını etkileyen en önemli faktörler olarak tanımlanır. Oyuncuların atış performansı oyunu kazanmada çok etkili olduğu için, bu çalışmanın sonuçları basketbol oyuncularına ve takım koçlarına antrenman programları için bir rehber olarak kullanılabilir.

Anahtar Kelimeler: Basketbol oyunu, makine öğrenmesi, performans analizi, sınıflama teknikleri

Serpil KILIÇ DEPREN: 0000-0003-4737-7131

[1] Yıldız Technical University, Faculty of Art and Sciences, Department of Statistics,İstanbul, Turkey, serkilic@yildiz.edu.tr

INTRODUCTION

Nowadays, machine learning techniques are implemented in every part of life by both the government and other profit/non-profit organizations. These techniques are used in credit risk modeling in finance and insurance, determining factors affecting students', schools' or countries' achievement in education, the effectiveness of a cure or medicine in health and determining/monitoring the performance of players in sports areas. Thus, the areas that are needed to be improved and the pain points of the processes can be determined and action plans can be created by the authorities.

In a basketball game, there is one major question, which is "Which team is going to win the game?". In order to estimate this question, players' performance should be measured and predict their performance according to the previous games' statistics. In this point of view, researchers are always trying to find factors affecting teams' and players' performance. There are many studies dealing with measuring teams' and players' performance using different techniques in the literature [1-3].

Researchers examined different sports branches in terms of players' or teams' performance in the literature [4-10].

Sampaio, Janeira, Ibáñez and Lorenzo [11] examined the game performance of the players' positions (which are guard, forward and center) using game-related statistics. In this context, they used three different basketball league: National Basketball Association (NBA, superior level) in the USA, Associacio´n de Clubs de Baloncesto (ACB, one of the best European leagues) in Spain and Liga de Clubes de Basquetebol (LCB, inferior level) in Portugal. In order to understand the performance dissimilarities of the players' position, linear discriminant analysis was applied. As a result, it was revealed that the differences between positions seemed to be wider in the NBA league than others. The ACB players' game-related statistics were not also too diverged.

In the study of Ibáñez, et al. [12] the long-season success of basketball teams' participating in the Spanish Basketball League (LEB1) was measured using discriminant analysis. The sample consisted of 870 games played between the 2000-2001 and 2005-2006 regular seasons. Game-related statistics such as assists, blocks, steals, rebounds, successful free throws and successful field goals, etc. were used in this study. Results of the

discriminant analysis showed that assists, steals, and blocks had a significant effect on team success.

In the literature, statistical models such as regression models and mixed models were also used to assess players' or teams' performance. Puente, Coso, Salinero, and Abián-Vicén [13] studied on identified basketball game performance in 2015. Accuracy in 2-point (and also 3-point) field goals was taken as a dependent variable in the regression model. As a result, the most important factors that affect team success were 2-point field goals, number of assists, defensive rebounds, and steals. Casals and Martinez [14] used four-level (individual, team, division, and conference) mixed models to measure performance with two different mixed models. For this purpose, points made and win score were used as dependent variables while season period, home advantage, difference of team quality, quality factor of a game, rest days, game started, player momentum, player's wage relative to team salary, teams fighting for the playoffs, player position, age, contract condition, minutes played and usage percentage were used as independent variables. Results showed that minutes played and usage percentage had an effective factor on field goal success and winning score. In another study, analysis of variance and multiple ridge regression were used to assess shooting performance of basketball players during FIBA EuroBasket 2015 [15]. This study showed that the number of successful two-point shots was the most important predictor of the points scored by the eight best teams.

In this study, factors affecting players' shooting accuracy are determined using different machine learning techniques with the 10-fold and 10-repeat algorithm. The remainder of this paper is organized as follows: In section 2, the information about the dataset and methodology are briefly described. Section 3 introduces the results and findings of player-based raw-dataset. In section 4, discussions of the study are presented. Lastly, in section 5, the suggestions about goal performance for winning the game are given to basketball players and team coaches.

DATASET AND METHODOLOGY

Dataset, which consists of 10 variables, is obtained from the NBA website (www.nba.com). 904 games and 127,752 observations in the 2014-2015 season are taken into consideration in the analysis. The dependent variable is a binary variable and it is coded as 0 if the player does not make a successful shot and coded as 1 if the player makes a

successful shot. Independent variables in the study are location (away or home), game period (from 1 to 7), game clock (minutes), shot clock (seconds), dribbles, touch time (seconds), shot distance (feet), points type (2-point or 3-point shot) and distance of closest defensive player (feet).

To assess predictive model for players' shooting accuracy, the most common machine learning algorithms are used, which are Logistic Regression (LR), Linear Discriminant Analysis (LDA), C5.0, k-Nearest Neighbor (k-NN), Naive Bayes (NB), Multilayer Perceptron (MLP) and Multivariate Adaptive Regression Splines (MARS) by using 10-fold cross-validation with the 10-repeat process.

Descriptive statistics of the continuous variables are given in Table 1.

Table 1. Descriptive statistics

	Mean	Max	Min	StdDev
Shot Number	6.5	38.0	1	4.71
Shot Clock (sec)	12.5	24.0	0	5.63
Dribbles	2.0	32.0	0	3.48
Touch Time (sec)	2.8	23.9	0	2.98
Shot Distance (feet)	13.6	47.2	0	8.89
Closest Defense Distance (feet)	4.1	53.2	0	2.76

According to Table 1, average number of shot attempts of a player, average attack time of a team and average touch time before making a shot is 6.5 attempts, 12.5 seconds and 2.8 seconds, respectively. The average number of dribble attempts before making a shot, shot distance and closest defense player distance are 2.0, 13.6 feet (4.2 meters) and 4.1 feet (1.3 meters), respectively. In addition, the ratio of a successful shot of a basketball player is 45%, on average. Also, 27% of all shooting attempts are 3-point shots while 73% of them are 2-point shots.

Linear Discriminant Analysis

LDA is commonly used with the goal of reducing dimensionality. With LDA, the linear combination of variables is transformed dataset that maximizes the ratio of the between-class variance to the within-class variance. Thus, it is obtained to statistically distinguish between multiple classes. LDA assumes that the cases of each class have Multivariate Normal distribution with the means and the covariance matrix. The LDA process can be

summarized as 5 steps:

1. The d-dimensional mean vectors for the different classes are calculated.
2. In-between-class and within-class scatter matrices are calculated.
3. The eigenvectors (e_1, e_2, ..., e_d) and corresponding eigenvalues (λ_1, λ_2, ..., λ_d) for the scatter matrices are calculated.
4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a dxk dimensional matrix W.
5. Use this dxk eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: Y=XxW (where X is an nxd-dimensional matrix representing the n samples, and y are the transformed nxk-dimensional samples in the new subspace).

Logistic Regression

Logistic regression is a probabilistic classification model that helps researchers to predict probabilities of different class values based on the relationships between the dependent variable and independent variable(s) [16]. The logistic regression model can be defined as;

$$logit\,(P(y=1)) = log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where $y_i \in \{0,1\}$ is the value of the binary response variable that is predicted by the values of the independent variable xi (i=1,2,…,k). P(y=1) indicates the probability of the sample belonging to class 1 and β_i represents the regression coefficient of xi.

Naïve Bayes

NB is a special form of Bayes' theorem which is based on the probability of an event. The NB classifier assumes that the values of the features are conditionally independent given the value of class variable. With NB algorithm, the training dataset from experiments predicts the class information and identifies the classification for which the classification membership is not available [17]. The process has 4 steps:

1. The dataset is converted into a frequency table.
2. Create a Likelihood table by finding the probabilities.
3. In order to find the posterior probability for each class, the Naive Bayesian equation is used.
4. The class with the highest posterior probability is the outcome of the prediction.

Multivariate Adaptive Regression Splines

MARS is a nonparametric multiple regression method of high-dimensional and correlated data under nonlinearity. This method is constructed in a two-stage process which is called the forward and the backward stages. In the forward stage process, all the basis functions produced using independent variables are iteratively added and found knots to improve predicting [18]. This continues until the complex model is the largest that contains many basis functions.

BFs are given by

$$-(x-t)_+^q = \begin{cases} (\mathrm{t}-\mathrm{x})^q, & \text{if } x < t \\ 0, & otherwise \end{cases}$$

$$(x-t)_+^q = \begin{cases} (\mathrm{t}-\mathrm{x})^q, & \text{if } x \geq t \\ 0, & otherwise \end{cases}$$

where q (≥0) is the power that determines the degree of polynomial piecewise function. If q=1, the splines are linear. BFs or splines are constructed as a separate piecewise polynomials of degree. The intervals of these splines are called knots or nodes, t.

The backward stage process is applied to prevent overfitting from this complex model, the best model is also obtained by removing some basis functions which indicate a small increase in the residual square error [19-20].

The general form of the MARS model can be expressed as;

$$f(x_i) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X)$$

β_0 is the intercept and β_m is corresponding coefficients that are estimated using the least-squares method and $B_m(X)$ is the m-th basis function.

k-Nearest Neighbor

The k-NN algorithm is commonly used to predict discrete response value that finds the closest neighbors among the variables in order to cluster the data using the distance between the data points. Thus, the performance of the algorithm is based on not only the variables used in the study but also the parameter k [21-22]. There are 4 steps in the k-NN algorithm.

1. The similarity is calculated by distance function such as Euclidian, Manhattan, Minkowski or Weighted distances.
2. Determine the k value. If it is 5, the algorithm searches for the 5 observation closest to the reference one.
3. Check the list of classes with the shortest distance and count the amount of each class that appears.
4. Takes as correct class the class that appeared the most times.

C5.0

C5.0 is regarded as a supervised machine learning algorithm based on decision trees. C5.0 is defined as an improvement of widely used ID3 and C4.5 algorithms. C5.0 decision tree gives a more accurate and efficient algorithm than all these classifiers [23]. C5.0 can classify more than two groups that represent a decision tree or correlation rules. The algorithm uses the following steps:

1. The entropy and measurement degree are calculated.
2. Calculate the information gain statistics.
3. Use SplitInfor function for every variable.
4. Use the gain division SplitInfor to obtain the entropy-based information gain value.

The feature with the highest information gain is selected to split the data into multiple subgroups.

Multilayer Perceptron

MLP approach is one of the Neural Network algorithms that consists of an interconnected network of neurons and synapses. MLP has three components, which are an input layer, hidden layer(s) and an output layer. This is a 4-step algorithm. At 1st step weights, which have an important role in carrying information from one neuron to another, are randomly determined. At 2nd step, the independent variable(s) propagate forward using different functions to produce output for each hidden layer. At 3rd step, the error is propagated backward by updating the weights and biases obtained from 3rd step. At last step, Errors are computed for each output and hidden layer. Then, weights and biases are updated and returned to 2nd step. The steps are repeated until the overall error is minimized [24].

Model Performance and Validation

Although there are many performance measures used for classification in the literature, commonly used performance measures are True Positive, True Negative and Correct Classification Rate, Precision, F-Measure, and Matthews Correlation Coefficient (MCC) to identify the significant variables in this study. True positive rate, which is also called as sensitivity, is a proportion of positive samples that are classified correctly. True negative rate, which is also called as specificity, is a proportion of negatively classified samples that are classified correctly. Precision is the proportion of correct predictions for the positive classified samples. MCC is a discrete version of Pearson's correlation coefficient that takes values between -1 and 1. As the MCC value is close to 1, there is a strong and same directional relationship, and there is a strong and inverse relationship with the MCC value -1. The F-measure is the weighted average of precision and sensitivity.

RESULTS

As a result of the study, the accuracy of the algorithms is compared with True Positive, True Negative, False Negative, False Positive, Correct Classification Rate, Precision, F-Measure and MCC statistics in order to find out the best performing algorithm. Model performance measures on classification are given in Table 2.

Table 2. Model performance statistics of the algorithms used

	True Positive	True Negative	Correct Classification Rate	False Positive	False Negative	Precision	MCC	F-Measure
LR	0.486	0.712	0.609	0.514	0.288	0.486	0.204	0.486
LDA	0.486	0.711	0.608	0.514	0.289	0.486	0.202	0.486
C5.0	0.373	0.833	0.623	0.627	0.167	0.373	0.232	0.373
k-NN	0.575	0.760	0.676	0.425	0.240	0.575	0.341	0.575
NB	0.584	0.578	0.581	0.416	0.422	0.584	0.162	0.584
MLP	0.272	0.884	0.604	0.728	0.116	0.272	0.196	0.272
MARS	0.418	0.784	0.617	0.582	0.216	0.418	0.217	0.418

According to Table 2, the Naïve Bayes algorithm has the highest true positive rate, which means that with this algorithm 58.4% of all successful shots are correctly classified. This ratio is 48.6%, 48.6%, 37.3%, 57.5%, 27.2% and 41.8% for LR, LDA, C5.0, k-NN, MLP and MARS algorithms, respectively. When algorithms' true negative rates are examined, it is seen that MLP has the highest true negative rate with 88.4%, which means 88.4% of all unsuccessful shots are classified correctly. In all algorithms, the true negative rate is above 70%, except NB.

Correct classification rates of all algorithms are between 60% and 70% level. k-NN has the highest correct classification rate, which is 67.6%. This means that with this algorithm 67.6% of all shots (including both successful and unsuccessful shots) are classified correctly. This ratio is 60.9%, 60.8%, 62.3%, 58.1%, 60.4% and 61.7% for LR, LDA, C5.0, NB, MLP and MARS algorithms, respectively.

Similar to the correct classification, true positive and true negative rates, the k-NN algorithm has the highest MCC value, which is 0.341. In addition, it has the second-highest value in terms of F-Measure statistics. Algorithms used in this research, except naïve Bayes, are generally suffering from the correct classification performance statistics in this dataset, which causes lower MCC statistics. It can be said that true positive ratios of the algorithms are not very high. Thus, it is obvious that different player-based statistics such as training time, age, and # of the match they've played in their career should be included in the dataset in order to increase the accuracy of the prediction.

Variable importance for each algorithm is given in Table 3.

Table 3. Variable importance

Variables	LR	LDA	C5.0	NB	MLP	MARS
Location (away or home)	8	7	6	-	9	-
Game Period	9	6	7	-	8	-
Game Clock	10	9	-	7	5	-
Shot Clock	4	8	5	8	6	-
Dribbles	5	5	-	3	4	-
Touch Time	3	3	3	5	3	3
Shot Distance	1	4	1	2	1	1
Position	7	-	-	4	7	-
Points Type (2-point or 3-point shot)	6	1	4	1	10	-
Distance of Closest Defensive Player	2	2	2	6	2	2

- : the variable is not statistically significant
k-NN: the importance of each variable is set equal while running the algorithm

In Table 3, 1 represents the most important factor and 10 represents the least important factor in shooting accuracy. In the k-NN algorithm, the importance of each variable is set equal while running the algorithm, so it is not given in Table 3.

According to Table 3, shot distance, distance of the closest defensive player and touch time are determined as the 3 most important factors in all algorithms. Location, game clock, shot clock and position are the factors that have little effect on players' shooting accuracy.

The correct classification rate of the MARS algorithm is 61.7%, and the major difference from other algorithms is that it reaches this level using only 3 independent variables. Shot distance and distance of the closest defense player are the most important factors in players' shooting accuracy.

Points type is the most important factor affecting players' shooting accuracy in LDA and NB algorithms, which is a very different finding than others this result can be a clue that the data can be split into 2 subsets such as 2-point and 3-point shot and then algorithms can be run.

DISCUSSION

NBA players' shooting performance has been modeled with this research using different machine learning algorithms. There are nine independent variables in this research

and in some algorithms, a few variables are not used because they have no statistically significant effect on shooting performance.

According to the logistic regression algorithm, shot distance, distance of closest defense player and touch time are the most important variables on players' shooting accuracy. On the other hand, game clock, game period and location are not as important as other variables. Player's shooting performance is declining in these conditions: if the player is playing at opposite team's stadium, as the touch time increases, as the shot distance increases and as the game time increases. The player's position is also important in shooting accuracy. If the player is playing as a point guard, shooting accuracy of the player is relatively higher than others.

In LDA, the most important variables on discrimination are point's type, the distance of closest defense player and touch time. Contrary to the LR algorithm, player's position has no significant effect on shooting accuracy. As a result, correct classification rate of LDA reaches 60.8% with eight independent variables while correct classification rate of LR reaches 60.7% with nine independent variables.

The results of the C5.0 decision tree algorithm are similar to LR results in terms of variable importance. However, C5.0 algorithm uses only seven independent variables and its correct classification rate reaches 62.3%. In this algorithm, if the shot distance is higher than 2.6 meters and the shot clock is lower than 3.6, the probability of a successful shot of a player is 22.3%. If the shot distance is between 1.5 meters and 2.5 meters and closest defense player is higher than 0.9 meters, the probability of a successful shot of a player is 78.0%. If the shot distance is lower than 1.5 meters and the closest defense player is lower than 0.9 meters, then touch time will be the most influential variable of the variables. In this situation, if touch time is lower than 1.6 seconds, the probability of a successful shot of a player is 29.6%. In conclusion, the important cut-offs are 1.5, 0.9 and 1.6 for shoot distance, the closest defense player and touch time, respectively.

Point type, shot distance and dribbles are the most important factors on shooting accuracy in the NB algorithm. In this algorithm, the closest defense player is not as significant as in other algorithms.

MLP algorithm with 2 hidden layers shows that shot distance, the distance of the closest defensive player and touch time are the most important variables affecting shot accuracy. This result is similar to the results of LR and C5.0 algorithms. However, the true positive rate of the MLP algorithm is the lowest among the other algorithms.

In the MARS algorithm, only three variables and their cross-interaction terms are used to predict a player's shooting accuracy and this algorithm's correct classification rate reaches 61.7%. The most important variables in this algorithm are similar to LR and C5.0 algorithms. These variables are shot distance, distance of the closest defensive player and touch time. Player's shooting accuracy has been affected negatively when shot distance is higher than 2.2 meters or the closest defensive player is lower than 1.4 meters or touch time is lower than 1.2 seconds. In addition to this, if the shot distance is lower than 2.2 meters and distance of the closest defensive player is lower than 1.4, Player's shooting accuracy has been affected negatively. In order to increase shooting accuracy, shot distance should be lower than 2.2 meters and distance of the closest defensive player should be higher than 1.4 meters and touch time should be higher than 3.8 seconds. Furthermore, correct classification rate of MARS algorithm is 61.7% with only three variables.

CONCLUSION

All the shot logs in the 2014-2015 NBA season is analyzed to find out the best machine learning algorithm in terms of shooting accuracy of basketball players in this study. In addition, for different machine learning algorithms, factors that have a significant effect on shooting accuracy are determined. It is shown that different factors affecting shooting accuracy can be important for each machine learning algorithms. However, the k-NN algorithm is the best among other alternatives used in this study and the most important factors for shooting accuracy of a basketball player are shot distance, distance of closest defense player and touch time. The results of this study can be used as guidance for training programs for basketball players and team coaches.

REFERENCES

1. Hughes M, and Franks IM. Notational analysis of sport systems for better coaching and performance in sport. London: Routledge, 2004.
2. Leite N, Baker J, and Sampaio J. Paths to expertise in Portuguese national team athletes. Journal of Sports Science and Medicine, 2009; 8(4): 560-566.

3. Ortega E, Villarejo D, and Palao J. Differences in game statistics between winning and losing rugby teams in the six nations tournament. Journal of Sports Science and Medicine, 2009; 8(4): 523-527.

4. Bartlett R. Performance analysis: can bringing together biomechanics and notational analysis benefit coaches? International Journal of Performance Analysis in Sport, 2001; 1(1): 122-126.

5. Hughes M, and Bartlett R. The use of performance indicators in performance analysis. Journal of Sports Sciences, 2002; 20: 739-754.

6. Trninic S, Dizdar D and Luksic E. Differences between winning and defeated top quality basketball teams in final of European club championship. Collegium Antropologicum, 2002; 26(2): 521-531.

7. Hughes M, and Franks IM. The essentials of performance analysis – An introduction. London: Routledge, 2008.

8. Tsamourtzis E, Karypidis A, and Athanasiou N. Analysis of fast breaks in basketball. International Journal of Performance Analysis in Sport, 2005; 5(2): 17-22.

9. Csataljay G, O'Donoghue P, Hughes M, et al. Performance indicators that distinguish winning and losing teams in basketball. International Journal of Performance Analysis in Sport, 2009; 9(1): 60-66.

10. Zuccolotto P, Manisera M, and Sandri M. Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. International Journal of Sports Science and Coaching, 2017; 13(4): 569-589.

11. Sampaio J, Janeira M, Ibáñez S, et al. Discriminant analysis of game-related statistics between basketball guards, forwards and centres in three professional leagues. European Journal of Sport Science, 2006; 6(3): 173-178.

12. Ibáñez SJ, Sampaio J, Feu S, et al. Basketball game-related statistics that discriminate between teams' season-long success. European Journal of Sport Science, 2008; 8(6): 369-372.

13. Puente C, Coso JD, Salinero JJ, et al. Basketball performance indicators during the ACB regular season from 2003 to 2013. International Journal of Performance Analysis in Sport, 2015; 15(3): 935-948.

14. Casals M, and Martinez AJ. Modelling player performance in basketball through mixed models. International Journal of Performance Analysis in Sport, 2013; 13(1): 64-82.

15. Gryko K, Mikołajec K, Maszczyk A, et al. Structural analysis of shooting performance in elite basketball players during FIBA EuroBasket 2015. International Journal of Performance Analysis in Sport, 2018; 18(2): 380-392.

16. Hosmer D, and Lemeshow S. Applied Logistic Regression (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc., 2000.

17. Yang CC, Soh CS and Yap VV. A non-intrusive appliance load monitoring for efficient energy consumption based on Naive Bayes classifier. Sustainable Computing-Informatics and Systems, 2017; 14: 34-42.

18. Kılıç Depren S. Prediction of Students' Science Achievement: An Application of Multivariate Adaptive Regression Splines and Regression Trees. Journal of Baltic Science Education, 2018; 17(5): 887-903.

19. Nieto PG, Garcia-Gonzalo E, Anton JA, et al. A comparison of several machine learning techniques for the centerline segregation prediction in continuous cast steel slabs and

evaluation of its performance. Journal of Computational and Applied Mathematics, 2017; 330(1): 1-19.

20. Ayyıldız E, Purutçuoğlu V, and Weber GW. Loop-based conic multivariate adaptive regression splines is a novel method for advanced construction of complex biological networks. European Journal of Operational Research, 2018; 270(3): 852-861.
21. Jiang S, Pang G, Wu M, et al. An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications, 2012; 39(1): 1503-1509.
22. Liu S, and Meng L. Re-examining factor structure of the attitudinal items from TIMSS 2003 in cross-cultural study of mathematics self-concept. Educational Psychology, 2010; 30(6): 699-712.
23. Khun M, and Johnson K. Applied Predictive Modeling. New York: Springer, 2013.
24. Han J, Kamber M, and Pei J. Data Mining Concepts and Techniques. Waltham: USA: Elsevier Inc., 2012.