International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 544–562
https://doi.org/10.21449/ijate.1167705
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

A comparative study of ensemble methods in the field of education: Bagging and Boosting algorithms

Hikmet Şevgin [1,*]

[1]Van Yüzüncü Yıl University, Faculty of Education, Department of Educational Sciences, Van, Türkiye

ARTICLE HISTORY
Received: Aug. 27, 2022
Revised: July 10, 2023
Accepted: Aug. 26, 2023

Keywords:
Educational Data Mining,
Ensemble Learning,
Bagging,
Boosting.

Abstract: This study aims to conduct a comparative study of Bagging and Boosting algorithms among ensemble methods and to compare the classification performance of TreeNet and Random Forest methods using these algorithms on the data extracted from ABİDE application in education. The main factor in choosing them for analyses is that they are Ensemble methods combining decision trees via Bagging and Boosting algorithms and creating a single outcome by combining the outputs obtained from each of them. The data set consists of mathematics scores of ABİDE (Academic Skills Monitoring and Evaluation) 2016 implementation and various demographic variables regarding students. The study group involves 5000 students randomly recruited. On the deletion of loss data and assignment procedures, this number decreased to 4568. The analyses showed that the TreeNet method performed more successfully in terms of classification accuracy, sensitivity, F1-score and AUC value based on sample size, and the Random Forest method on specificity and accuracy. It can be alleged that the TreeNet method is more successful in all numerical estimation error rates for each sample size by producing lower values compared to the Random Forest method. When comparing both analysis methods based on ABİDE data, considering all the conditions, including sample size, cross validity and performance criteria following the analyses, TreeNet can be said to exhibit higher classification performance than Random Forest. Unlike a single classifier or predictive method, the classification or prediction of multiple methods by using Boosting and Bagging algorithms is considered important for the results obtained in education.

1. INTRODUCTION

The retrieval of information that needs to be obtained in order to make speculations concerning an event or situation from a community instead of a single person definitely provides the opportunity to make stronger inferences with poorer error rate. In the daily life as well, the attempt to obtain a greater deal of information that can be gained regarding an event or situation, and the overall evaluation of the collected data, is ultimately the result of attempting to reach a more precise conclusion. However, during a decision phase yielding important results, the opinions of experts who are thought to help make decisions are consulted. For example, the opinions of several specialists are asked before a life- threatening operation. In addition, ensemble- based decision- making processes are also administered to elect a manager or to

*CONTACT: Hikmet Şevgin ✉ hikmetsevgin@gmail.com Van Yüzüncü Yıl University, Faculty of Education, Department of Educational Sciences, Van, Türkiye

e-ISSN: 2148-7456 /© IJATE 2023

decide on a new law (Polikar, 2012). Likewise, ensemble methods performs analysis methods and, in this respect, it has received increasing attention in recent years with its use with various multiple classification systems, data mining methods and machine learning algorithms (Do-Nascimento et al., 2019; Lee et al., 2010; Zhang & Ma, 2012). The methods that were initially used to reduce the variance of classification and predictive analyses and to increase the accuracy of classification were then successfully utilized for various purposes such as feature selection and the determination of confidence interval (Abeel et al., 2010; Kumari, 2012; Saeys et al., 2008; Zhang & Ma, 2012).

Technological advancement and novel statistical algorithms have allowed for a better understanding of data mining and improved its use. The emergence and development of ensemble learning in the last quarter can be regarded as a reflection of this process. On account of the combination of basic statistical methods to generate ensemble learning methods, the results with high classification success and precise prediction as well as low error variance have been obtained (Bauer & Kohavi, 1999; Hansen & Salamon, 1990; Onan, 2015; Opitz & Shavlik, 1996; Polikar, 2006; Sagi & Rokach, 2018) and, in this respect, its use has recently increased in various areas such as health, economy, banking, agriculture, engineering, business and education (Akman, 2010; Şevgin & Önen 2022).

There have been several studies employing ensemble methods encountered in the literature (Abidi et al., 2020; Baskin et al., 2017a; Baskin et al., 2017b; Dietterich, 2000; Dietterich, 2002; Freund & Schapire, 1996; Friedman, 2001; Kapucu & Cubukcu, 2021; Kausar et al., 2020; Li et al., 2022; Mousavi & Eftekhari, 2015; Pong-Inwong & Kaewmak, 2016; Steinki & Mohammad, 2015; Wang et al., 2018). It is worth noting that the researchers who conduct studies on data mining and machine learning have fallen behind in discovering the success of Ensemble-based learning methods in terms of classification and prediction-based decision-making (Polikar, 2012). Nevertheless, with the studies carried out in recent years, it has been seen that a great deal of knowledge and literature have been obtained especially in the field of education (Abdar et al., 2018; Abellán & Castellano, 2017; Aggarwal et al., 2021; Almasri et al., 2019; Ashraf et al., 2021; Ashraf et al., 2020; Arun et al., 2021; Guo et al., 2021; Karalar et al., 2021; Keser & Aghalarova, 2022; Kotsiantis et al., 2010; Injadat et al., 2020a; Injadat et al., 2020b; Premalatha & Sujatha, 2021). This comparative study focusing on Bagging and Boosting (Akman, 2010; Zhou, 2012) algorithms that are the most well-known Ensemble methods may contribute to the literature and, particularly the field of educational data mining, in order to list and utilize the concept of Ensemble Learning and its methods among advanced statistical methods in the field of education.

In the field of education, both in the phase of various and big data processing that poses opportunities for the construction of education within the Ministry and in the analysis process of multidimensional, complicated and noisy data obtained from students and teachers through large- scale tests, it is of importance to achieve strong and non-deviating outputs. Indeed, the use of ensemble methods can be considered as flexibility (Strobl et al., 2009) for the data analysis in the noisy data by its nature that we often call traditional which do not provide various assumptions required for the parametric methods. Thus, the achievement of the output with lower error variances in the field of education can be contributed. Considering the situations where decisions regarding students such as fail- pass or successful- unsuccessful are made or variables that affect student achievement are examined, the realization of analyses with high classification and prediction success and poor error rate may ensure the results in terms of high classification/ decision validity. It is clear that the use of ensemble methods in education serves to obtain results with high classification and prediction success and to gain results with high classification/ decision validity. Therefore, it is considered important to utilize ensemble methods to obtain evidence concerning classification/ precision validity in the procedures to be performed for classification and prediction.

1.1. Ensemble Learning

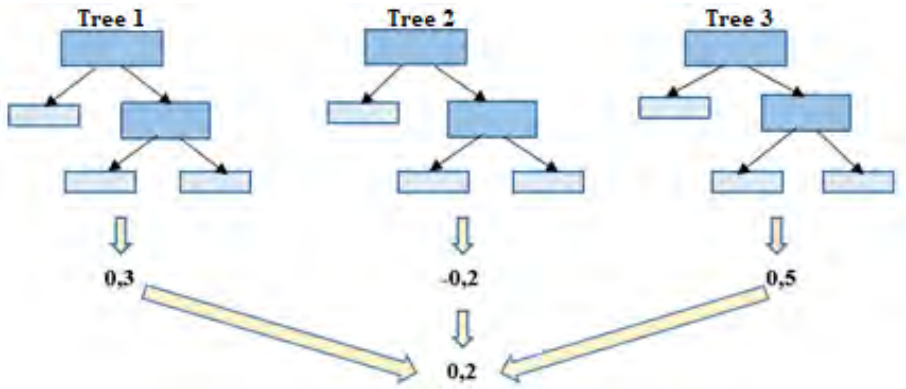
Recently, in the process of statistically synthesizing the data obtained through scientific research, the idea of combining multiple methods to produce a new model based on classification or prediction has been emphasized by the researchers and been the subject of publications in recent years. Tukey is the first researcher who has introduced the concept of ensemble learning (1977) where he had used linear regression model to fit the original data as first step and then again linear regression model to fit the residual as a second step (Sagi & Rokach, 2018). Later, in the 1990s, Hansen and Salamon shared the outputs of neural network ensembles. In addition, in 1996, Breiman first proposed ideas for the Bagging algorithm and in the same year, Freund and Schapire came up with the first boosting algorithm. Subsequently, the AdaBoost algorithm was introduced by Freund and Schapire (1996) as a result of combining multiple weak classifiers to build one strong classifier. Moreover, certain studies on the development of ensemble methods using boosting algorithms such as Gradien Boosting presented by Friedman et al. (2000) and Multiple Additive Regression Trees (MART) proposed by Friedman and Meulman (2003) have been encountered. In the meantime, numerous ensemble methods which perform ensemble learning by using Bagging and Boosting algorithms have been developed (Kumari, 2012; Polikar, 2006; Schapire, 2003; Zhou, 2012).

The fact that the information is obtained from the narration of more than one person who witnessed the same event rather than of a person, in other words, the information gathered from ensembles provide more reliable results with high accuracy. Learning in this way is called ensemble learning (Polikar, 2012). Similarly, the combination of the predictions of several base estimators is generally better than the prediction of one best predictor. A group of predictive methods is gathered under the title of Ensemble and the process of making predictions from the ensemble is called Ensemble Learning (Geron, 2019). To sum, Ensemble methods can be considered as the combination of multiple methods to produce outputs with higher success (Quinlan, 1996), that is outputs with higher levels of reliability (Akman, 2010; Maclin & Opitz, 1997) in contrast to the outputs based on classification and prediction obtained from single methods. These methods, combined together to give an ensemble, can be a decision tree (C&RT, C5, CHIAD, ID3, QUEST) as well as such methods as MARS, YSA, SVA (Chen & Guestrin, 2016; Clarke et al., 2009; Freund & Schapire, 1996; Friedman, 2001; Friedman & Meulman, 2003; Quinlan, 1996; Sutton, 2005; Zhou, 2012). The algorithms that combine these methods and give an ensemble are Boosting, Bagging, Stacking, Max Voting, Averaging, Weighted Averaging and Blending algorithms (Baskin et al. 2017a; Zhang & Ma, 2012; Zhou, 2012). Of these algorithms, Bagging and Boosting are the most elaborated and known ensemble learning algorithms (Akman, 2010; Zhou, 2012). Within the scope of this study, Bagging and Boosting algorithms are included.

As stated above, although Bagging and Boosting algorithms can be applied to several methods, it has been seen that they are mostly used together with decision trees in the literature. In certain sources, however, ensemble methods are referred as Tree- based Ensemble Methods (Akman, 2010). The TreeNet method, which creates ensembles using classification and regression trees (C&RT) with the boosting algorithm, and the Random Forest (Breiman, 2001), which creates ensembles using C&RT with the bagging algorithm, are included in the present study. In certain sources, although Random Forest is considered as an Ensemble method independently due to the fact that it creates random subspaces to do a random selection of a subset of features to use to grow each tree (Geron, 2019; Han et al., 2012), it is also included in the Bagging title since it utilizes Bagging algorithm in the formation of ensemble (Clarke et al., 2009; Nisbet et al., 2009). Hastie et al. (2009) stated that Random Forest method was a modification of the Bagging algorithm. The main factor choosing TreeNet and Random Forest methods for the current study is that both methods are Ensemble methods that combine single decision trees (classification and regression trees - C&RT) with Bagging and Boosting algorithms and combine the outputs

obtained from each of them into a single output. An example representing the working principle of ensemble methods is presented in Figure 1 below:

Figure 1. *The illustration of the working principle of Ensemble Model*



In Figure 1, the value of each tree is combined to produce the final value of the ensemble. The combination process differs since Bagging and Boosting algorithms use their own techniques. During the consolidation process, boosting algorithm iteratively constructs a series of decision trees being trained whereas Bagging algorithm consists of simple random sampling with replacement. These algorithms and the analyses that use them are respectively elaborated below.

1.1.1. *Boosting*

In Boosting algorithm, each model is constructed on the incorrectly predicted data of the previous model (Friedman, 2001). In other words, each model learns from the errors of the previous model. This is realized by weighting the data points and the whole process continues sequentially (Friedman & Meulman, 2003). Then, the weak learners are eliminated one by one and the strong learner is reached (Polikar, 2012). The last model is yielded from the weighted average of all models (Zhou, 2012).

Boosting algorithm [Rokach (2019)].

Input: I (a weak inducer), S. (a training set) and k (the sample size for the first classifier)

Output: M_1, M_2, M_3

1: $S_1 \leftarrow$ Randomly selected $k < m$ instances from S without replacement;

2: $M_1 \leftarrow I(S_1)$

3: $S_2 \leftarrow$ Randomly selected instances (without replacement) from S - S_1 such that half of them are correctly classified by M_1.

4: $M_2 \leftarrow I(S_2)$

5: $S_3 \leftarrow$ any instances in S - S_1 - S_2 that are classified differently by M_1 and M_2.

As shown above, boosting algorithm has an iterative characteristic. The algorithm generates three classifiers. The sample S_1, which is used to train the first classifier M_1, is randomly selected from the original data set. The second classifier, M_2, is trained on a sample M_2, half of which consists of instances that are incorrectly classified by M_1, and the other half is composed of instances that are correctly classified by M_2. The last classifier, M_3, is trained with instances that the two previous classifiers disagree on (Rokach, 2019).

The error rate of the M_i model is calculated using the given the formula below:

$$error(M_i) = \sum_{j=1}^{d} wj \times error(X_j) \quad (1)$$

In this equation, *error* (X_j) is the classification error of Xj. If the group is incorrectly classified, error (Xj) = 1, otherwise it is 0 (zero) (Han et al., 2012). If the performance of the classifier, M_i, is poor, the classification error exceeds 0.5, in which case M_i is abandoned. Instead, the operation is retried by generating a new Si training data (Han et al., 2012). The error rate of M_i affects the updating of the weights of the training set. If the observations are correctly classified, the weighting of observations is multiplied by the value obtained from the equation below:

$$\frac{error(M_i)}{(1-error(M_i)} \quad (2)$$

When the weights of all correctly classified observations are updated, the weights of all observations (including those that are incorrectly classified) are normalized so that their sum remains the same as before. As a result, the weights of misclassified observations are increased and the weights of correctly classified observations are reduced. The lower the error rate is for a classifier, the higher the accuracy rate is (Han et al., 2012). The weight calculated for each M_i classifier is represented by the equation below:

$$log\,\frac{1-error(M_i)}{error(M_i)} \quad (3)$$

Based on boosting algorithm, various alternatives such as AdaBoost (Adaptive Boosting – Freund & Schapire, 1996), Gradient Boosting (Friedman, 2001), XGBoost (Chen & Guestrin, 2016) have been developed to determine the weights used in the training and classification phases of the boost iteration. However, AdaBoost and Gradient Bosting are commonly used algorithms (Sinharay, 2016).

1.1.2. *Bagging*

Bagging is an abbreviation for Bootstrap-Aggregating. It was first proposed by Leo Breiman in 1996. It is a simple, yet effective method for generating an ensemble of classifiers. The ensemble classifier that is created by this method consolidates the outputs of various learned classifiers into a single classification and this results in a classifier whose accuracy is greater than the accuracy of each individual classifier (Rokach, 2019). Bootstrap in the bagging algorithm is represented as resampling (Breiman, 1996). In this method, each classifier in the ensemble is trained on a sample of instances taken with replacement (allowing repetitions) from the training set. All classifiers are trained using the same learning algorithm. Therefore, some of the original instances may appear more than once in a training set, and some may not be included at all (Efron & Tibshirani, 1993).

Bagging Algorithm [Rokach (2019)].

Input: *I* (a base inducer), *T* (the number of iterations), *S* (the original training set), μ (the sample size).

1: $t \leftarrow 1$

2: Repeat

3: $S_t \leftarrow$ a sample of μ instances from *S* with replacement.

4: Construct classifier M_t using *I*, with S_t as the training set.

5: $t \leftarrow t + 1$

6: until $t > T$

The Bagging algorithm works as shown above. The classifiers are all trained using the same learning algorithm. The algorithm receives an induction algorithm *'I'* which is used for training all members of the ensemble. The stopping criterion in line six terminates the training when the ensemble size reaches *'I'*. One of the main advantages of bagging is that it can be implemented

easily in a parallel mode by training the various ensemble classifiers on different processors (Rokach, 2019).

The most important feature that distinguishes the Bagging algorithm from the Boosting algorithm is that sampling with replacement is used. That is, it is likely to use a sample more than once in the Bagging algorithm. However, in Boosting algorithm, the sample that has been used is not used again. The common feature of the Bagging and Boosting algorithms is that in both algorithms, they generate the last classifier through multiple voting for classification models, and the last estimator through the average of parameter estimates for regression models (Ferreira & Figueiredo, 2012). In this respect, it has been considered important in terms of using the data obtained in the field of education in the analysis of classification and prediction. Besides, unlike the results obtained by a single method, the use of results obtained through more than one method has also been regarded noteworthy in terms of the reliability and validity of the results obtained. Finally, it has been thought that it may contribute to the field in terms of using novel methods built on Bagging and Boosting algorithms in education. In fact, it has been seen that both the Boosting and Bagging algorithms are included in certain studies conducted in the field of education. However, this study is remarkable in terms of the fact that it elaborates the concept of *'Ensemble Learning'* entitled under data mining and machine learning and compares the methods based on the most known algorithms, Bagging and Boosting, on the data in the field of education. Therefore, "The purpose of the study is to conduct a comparative study of Bagging and Boosting algorithms among ensemble methods and to examine the classification performance of both methods on the data obtained in the field of education through TreeNet and Random Forest". To this end, answers to the following questions have been sought:

1) Do the performance measurements of TreeNet and Random Forest methods using Bagging and Boosting algorithms obtained according to each sample size based on 3,5 and 10-fold cross validity on the ABİDE data using Bagging and Boosting algorithms differ?

2) Is there a difference between TreeNet and Random Forest method using Bagging and Boosting algorithms on the ABİDE data based on the comparison of RMSE, MSE, MAD and MRAD values?

2. METHOD

The study was designed with quantitative research and a relational survey model was used with a descriptive approach. The relational model allows researchers to obtain information regarding a large group by examining a sample (Leedy & Ormrod, 2005).

2.1. Data Set

The data set of the study consists of mathematics scores of ABIDE (Academic Skills Monitoring and Evaluation) 2016 administered to 8th grade students. ABIDE implementation includes Turkish, Mathematics, Science and Social Studies achievement tests prepared for 8th grade students. However, the Mathematics achievement test was focused in the current research. For the data of 5000 students randomly recruited from the data set, data deletion was carried out for the demographic data and the values were assigned to the obtained from the scales through (MCAR) regression since it is below %5 for the loss data (Tabachnick & Fidell, 2015). As a result of the deletion of loss data and assignment procedures, this number decreased to 4568. The dependent variable (students' maths achievement), which is a continuous variable, was dual-categorized by considering the first quarter of %25 (low maths achievement) and the fourth quarter of %25 (high maths achievement). 2284 (1034 female and 1250 male) students, 1142 in the first quarter and 1142 in the fourth quarter, constitute the sample of the study. Those in the first quarter with maths scores between 343,10- 440,14 refer to the students with low

maths achievement whereas those in the fourth quarter with maths scores between 556,62-776,02 refer to the students with high maths achievement.

2.1.1. *Measurement tools*

The current research consists of mathematics achievement test in ABİDE implementation, demographic information collected by student survey and the variables collected at the scale level that are the attitude towards the school, peer bullying, parental approach, liking of mathematics course, self-efficacy perception towards the mathematics course, the value given to the mathematics course and teacher's instructional activities.

Prior to the data analysis through ensemble methods, the reliability, validity and multiple connection problems of the scales used in the research were examined. With the purpose of determining the reliability coefficient, McDonald's (ω) reliability index was employed instead of Chronbach Alpha reliability index due to the fact that the factor loads of the items were not equal (Yurdugül, 2006). McDonald's (ω) reliability index of the scales ranged from 0.77 (parental approach) the lowest to 0.94 (teacher's instructional activities) the highest and these values can be said to be at acceptable levels. In order to prove the validity of the scale, exploratory factor analysis was performed and it was found that each scale had one dimensional and that factor loads of the items varied between 0.369 the lowest and 0.875 the highest. Since the factor loads related to the items are above the acceptable minimum value, 0.30 (Çokluk et al., 2012), it can be said that they are above the acceptable value. Moreover, Tolerance and VIF values were examined for multi connection problem, and it was revealed that Tolerance values ranged between 0.520 and 0.916 and VIF values varied between 1.091and 1.922. Since these values are higher than 0.100 for Tolerance and lower than 10 for VIF (Schroeder et al., 1990), it can be stated that there is no multi connection problem.

2.2. Data Analysis

In the research, the data set was divided into four data sets as 250, 500, 1000 and 2000 in terms of sample size through simple random sampling without replacement. The observations in each data set were assigned to the data set in a way that they were subjected to 3-fold, 5-fold and 10-fold cross validation.

In this study, in the context of ensemble methods, performance criteria based on sample size were compared for TreeNet analysis method using Boosting algorithm and Random Forest method using Bagging algorithm in the background. In data analysis, the educational version of the SPM 7.0 statistical package program and open source Phyton-based Orange package 3.34 version were utilized. In addition, the evaluation of performance criteria yielded by confusion matrix was made through the test data and the 2nd category (Successful) was considered as the focus group.

2.2.1. *TreeNet*

The TreeNet method is based on stochastic gradient boosting algorithm to determine the weights used in the training and classification phases of the incremental iteration (Padmaja et al., 2016). Stochastic gradient boosting, developed by Friedman (2002), is used to address a regression task by optimizing the mean squared error. It is a non- parametric method where each successive learner is trained following the pseudo - residual errors of the preceding learner, thus finding solutions to classification and regression problems (Friedman, 2002; Hastie et al., 2009). The TreeNet (TM Salford Systems, inc.) method has various titles due to commercial concerns such as Multiple Additive Regression Trees-MART (TM Jerill, inc.), Boosted Regression Trees-BRT (TM Stat Soft, inc.), Gradient Boosting Trees (GBT) and Gradient Boosting Model (GBM) (Elish & Elish, 2009; Hill & Lewicki, 2006). TreeNet is successfully applied in science fields where complex relationships of numerous variables are modelled by

adding classification trees when the dependent variable is categorical and the regression trees are added when the variable is continuous (Şevgin & Önen, 2022).

2.2.2. *Random forest*

Random Forest method is a special modification of Bagging algorithm (Amrieh et al., 2016; Hastie et al., 2009). It was created as a result of the application of the Random Subspace technique proposed by Ho (1998) on the Bagging method (Biau, 2012). In the bagging method, decision trees are generated by selection from the data set independently of one another through bootstrap technique. However, the Random Subspace method does a random selection of a subset of features to use to grow each tree (Akman, 2010). In Random Forest method, each decision tree that generates the decision forest is created by bootstrap sampling randomly selected from the original data set with replacement. The Random Forest proposed by Breiman (2001) is a non-parametric method applied in science fields where complex relationships of numerous variables are modelled by adding classification trees to regression trees through bootstrap sampling method when the dependent variable is two- class or multi- class (Biau & Scornet, 2016; Geneur et al., 2017).

Recent studies have shown that ensemble learning methods outperform traditional regression methods (Elith et al., 2006). It can be said that TreeNet and Random Forest are among best performing ensemble methods (More detailed information for these two methods, see Breiman, 2001; Friedman, 2002).

2.2.3. *Confusion matrix*

A confusion matrix is a technique for summarizing the performance of a classification algorithm. A confusion matrix is a two-dimensional matrix ("actual" and "predicted"), indexed in one dimension by the true class of an object and in the other by the class that the classifier assigns (Ting, 2017) and it allows easily discovering whether the system mixes the two classes (Şevgin, 2020). Table 1 presents an example of confusion matrix for a two - class classification task.

Table 1. *Confusion matrix.*

		Predicted Class		
		Unsuccessful	Successful	Total
Actual Class	Unsuccessful	TN	FP	TN+FP
	Successful	FN	TP	FP+TP
	Total	TN+FN	FP+TP	TN+FN+FP+TP

(TP: True Positive TN: True Negative FP: False Positive Fn: False Negative)

Confusion matrices represent counts from predicted and actual values. It is applied to binary classification. In this regard, the confusion matrix represents true positive (TP) values, false positive (FP) values, true negative (TN) values and false negative (FN) values (Ting, 2017). The output for True Positive and True Negative shows the instances predicted accurately. However, False Positive and False Negative represent the instances predicted incorrectly. Accuracy is calculated as the sum of two accurate predictions (TP + TN) divided by the total number of data sets (P + N). The best accuracy is 1.0, and the worst is 0.00. Ideally, the sum of TP and TN should have an approximate value to the total of the pattern and the sum of FP and FN values should be close to zero (Han et al., 2012).

2.2.4. *Performance criteria for the categorical dependent variable*

In this research, accuracy- percentage- sensitivity- precision ratios, AUC value of ROC curve and F1 score were used as performance criteria. The formulas are given below:

Accurate classification rate indicates how well the method used in classification problems predicts the class distributions of the data and is often expressed as a percentage.

$$Accurate\ Classification\ Rate = \frac{(\mathrm{TP+TN})}{(\mathrm{TP+FP+TN+FN})} \quad (4)$$

Specificity refers to the probability of a negative test result, conditioned on the individual truly being negative and it takes a value between 0 and 1. This value is usually expressed as a percentage.

$$Specificity = \frac{(TN)}{(TN+FP)} \quad (5)$$

Sensitivity represents how well a test can identify true positives and it reveals a value between 0 and 1. This value is usually expressed as a percentage.

$$Sensitivity = \frac{(TP)}{(TP+FN)} \quad (6)$$

The numerical value of accuracy represents the proportion of true positive results in the selected population and yields a value between 0 and 1. This value is usually expressed as a percentage.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (7)$$

The F- score (also known as the F1- score or F-measure) is defined as the harmonic mean of precision and recall scores of a model in order to ensure a balanced measure of overall classification performance.

$$F1 - Score = 2x\frac{sensitivity\ x\ precision}{sensitivity+precision} \quad (8)$$

2.2.5. *Performance criteria for the continuous dependent variable*

The RMSE, MSE, MAD, and MRAD values which give error values for numerical prediction, allow data mining and machine learning methods to be examined and compared to one another.

RMSE (Root Mean Square Error): RMSE measures the average difference between a statistical model's predicted values and the actual values. The RMSE value is the measurement of how close the predictions are to the actual values. A low RMSE value refers to a better model performance.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \quad (9)$$

MSE (Mean Square Error): MSE is defined as mean or average of the square of the difference between actual and estimated values. Unlike RMSE, MSE is computed without taking the square root. The MSE value quantifies the size of prediction errors and a low MSE value means a better model performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (10)$$

MAD (Mean Absolute Deviation): MAD is a measure of the average absolute distance between each data value and the mean of a dataset. The MAD value measures the size of prediction errors, yet, unlike RMSE and MSE, it can be more sensitive to larger extreme outliers since it does not take the square of the deviation.

$$MAD = \frac{\sum_{i=1}^{n}|\chi_i - \bar{x}|}{n} \quad (11)$$

MRAD (Mean Relative Absolute Deviation): MRAD is the average distance between each data point and the mean. MRAD provides an independent assessment of the scale of the measured

values by calculating the prediction errors to the actual values. Besides, it is useful or comparing values measured in different times.

$$MRAD = \frac{\frac{(\sum_{i=1}^{n}|(x_i - \bar{x})|)}{n}}{\bar{x}*100} \quad (12)$$

2.2.6. *Cross validation*

Cross validation, also being referred to as rotation estimation, is a resampling technique used in statistical modelling and machine learning to evaluate the performance and generalization ability of two or more models. Cross validation involves dividing the existing dataset into k subsets, training the model on a subset of the data, and evaluating its performance on the remaining fold(s) (Olson & Delen, 2008). In K-fold cross-validation, the full data set is randomly divided into various subsets of k of approximately equal size. The classification model is trained and tested k times. In the present study, 3-fold, 5-fold and 10-fold cross-validity was applied to evaluate the performance of the methods. In other words, a cross-validity was performed in which one- third, one- fifth and one- tenth of the data set were considered as test data.

3. RESULTS

In this section, the TreeNet method using the boosting algorithm in the background and the Random Forest method using the Bagging algorithm are examined in different sample sizes, 3-fold, 5-fold and 10-fold cross validity rates. At each sample size and each cross-validity rate, the number of trees that is required by the TreeNet and Random Forest methods to generate the optimal model is presented in Table 2.

Table 2. *The number of trees where Treenet and random forest models are established.*

		250	500	1000	2000
TreeNet	3K	648	312	762	484
	5K	446	465	700	475
	10K	561	426	739	465
Random Forest	3K	526	258	589	461
	5K	433	436	547	417
	10K	489	438	628	423

Table 2 represents the number of trees needed to determine the optimal number of trees in the area under the ROC curve for TreeNet (Hastie et al., 2009). For Random Forest, the value with the lowest error rate in the decision forest refers to the number of trees needed for the most appropriate model to be established (Huffer and Park, 2020; Probst and Boulesteix, 2017).

3.1. Findings on the TreeNet and Random Forest Methods by Sample Size

The classification performances yielded by both analysis methods as a result of 3-fold cross validation for each level of the sample size taken from the study group are presented in Table 3 as a percentage. In Table 3, it is seen that for both analysis methods with 3-fold cross-validity, they received the same value in terms of accurate classification rate in 500 sample size although TreeNet method was higher than Random Forest method in 250, 1000 and 2000 sample sizes. In terms of specificity, TreeNet method was found to be higher in 250 smaple sizes whereas Random Forest was revealed to be higher in 500, 1000 and 2000 sample sizes. In terms of sensitivity, it is seen that TreeNet method is higher than Random Forest method in all sample sizes. In terms of accuracy, it is seen that the TreeNet method is higher in the sample size of 250 and 1000 and the Random Forest method is higher in the sample size of 500 and 2000. However, in terms of F1- score, it has been revealed that the TreeNet method is higher than the

Random Forest method in all sample sizes. In terms of AUC value, it has been found that the Random forest method is higher in the sample size of 250 and, however, that TreeNet method is higher in the sample sizes of 500, 1000 and 2000.

Table 3. *Percentages of classification performance by sample sizes for 3-Fold Cross validity.*

			250	500	1000	2000
3K	TreeNet	Accurate Classification Rate	%76.80	%71.40	%77.20	%77.20
		Specificity	%74.56	%70.59	%76.24	%78.35
		Sensitivity	%78.68	%72.24	%78.18	%76.00
		Accuracy	%78.68	%70.24	%76.33	%77.10
		F1- score	%78.68	%71.23	%77.25	%76.54
		AUC value	%83.98	%80.84	%85.77	%84.80
	Random Forest	Accurate Classification Rate	%72.80	%71.40	%74.10	%76.15
		Specificity	%67.54	%72.94	%77.03	%79.52
		Sensitivity	%77.21	%69.80	%71.11	%72.62
		Accuracy	%73.94	%71.25	%75.21	%77.28
		F1- score	%75.54	%70.52	%73.10	%74.88
		AUC value	%80.71	%81.35	%83.61	%84.79

The classification performances obtained by both analysis methods as a result of 5-fold cross validation for each level of the sample size taken from the study group are presented in Table 4 as a percentage.

Table 4. *Percentages of classification performance by sample sizes for 5-Fold Cross validity.*

			250	500	1000	2000
5K	TreeNet	Accurate Classification Rate	%71.20	%75.20	%75.20	%77.10
		Specificity	%67.54	%74.90	%75.45	%77.67
		Sensitivity	%74.26	%75.51	%74.95	%76.51
		Accuracy	%73.19	%74.30	%74.95	%76.66
		F1- score	%73.72	%74.90	%74.95	%76.58
		AUC value	%80.65	%82.53	%84.47	%85.30
	Random Forest	Accurate Classification Rate	%75.20	%74.20	%74.20	%76.55
		Specificity	%71.93	%75.69	%77.22	%79.53
		Sensitivity	%77.94	%72.65	%71.11	%73.44
		Accuracy	%76.81	%74.17	%75.37	%77.47
		F1- score	%77.37	%73.40	%73.18	%75.41
		AUC value	%82.79	%81.75	%83.92	%84.90

In Table 4, it is seen that for both analysis methods with 5-fold cross-validity, the Random Forest method is higher in the accurate classification rate in the sample size of 250 and that the TreeNet method is higher in the sample size of 500, 1000 and 2000. In terms of specificity, it has been demonstrated that Random Forest method is higher in all sample sizes. In terms of sensitivity, it is seen that the Random Forest method is higher in the sample size of 250 and the TreeNet method has been found to be higher in the sample sizes of 500, 1000 and 2000. Moreover, in terms of accuracy, it is seen that the Random Forest method is higher in sample size of 250 and the TreeNet method is higher in 500, 1000 and 2000 sample sizes. As for F1-

score, it is seen that the Random Forest method is higher in the sample size of 250 and TreeNet method is higher in the sample sizes of 500, 1000 and 2000. In terms of AUC value, it has been revealed that the Random Forest method is higher in the sample size of 250 and TreeNet method is higher in the sample sizes of 500, 1000 and 2000.

The classification performances obtained by both analysis methods as a result of 10-fold cross validation for each level of the sample size taken from the study group are presented in Table 5 as a percentage.

Table 5. *Percentages of classification performance by sample sizes for 10-Fold Cross validity*

			250	500	1000	2000
10K	TreeNet	Accurate Classification Rate	%75.20	%74.40	%76.60	%77.20
		Specificity	%76.32	%73.73	%75.84	%78.35
		Sensitivity	%74.26	%75.10	%77.37	%76.00
		Accuracy	%78.91	%73.31	%75.84	%77.10
		F1- score	%76.52	%74.19	%76.60	%76.54
		AUC value	%83.42	%82.92	%84.83	%84.80
	Random Forest	Accurate Classification Rate	%75.20	%74.20	%73.70	%76.35
		Specificity	%71.05	%75.69	%76.24	%79.82
		Sensitivity	%78.67	%72.65	%71.11	%72.72
		Accuracy	%76.43	%74.16	%74.58	%77.56
		F1- score	%77.53	%73.40	%72.80	%75.07
		AUC value	%83.12	%83.17	%83.49	%85.01

In Table 5, it has been demonstrated that both methods receive the same value in the sample size of 250 in terms of correct classification rate with 10-fold cross-validity; however, it has been seen that the TreeNet method is higher compared to the Random Forest method in the sample sizes of 500, 1000 and 2000. Nevertheless. in terms of specificity, it has been found that the TreeNet method is higher in the sample size of 250 and that the Random Forest method is higher in the sample size of 500, 1000 and 2000. As for sensitivity, it has been indicated that the Random Forest method is higher in the sample size of 250 and that TreeNet method is higher in the sample sizes of 500, 1000 and 2000. In terms of accuracy, it is seen that the TreeNet method is higher in the sample size of 250 and 1000 and the Random Forest method is higher in the sample size of 500 and 2000. Furthermore, In terms of F1-score, it is seen that the Random Forest method is higher in the sample size of 250 and TreeNet method is higher in the sample sizes of 500, 1000 and 2000. Finally, in terms of AUC value, it has been revealed that the TreeNet method is higher in the sample sizes of 250 and 1000 and the Random Forest method is higher in the sample sizes of 500 and 2000.

3.2. Findings on the TreeNet and Random Forest Methods Based on RMSE, MSE, MAD and MRAD Performance Measurements

The classification performances of RMSE, MSE, MAD and MRAD values obtained by both analysis methods for each level of sample size taken from the study group are presented in Table 6. As shown in Table 6, it is seen that the TreeNet method yields lower error values than the Random Forest method in all sample sizes. It has been shown that the error values of the TreeNet method, in itself, increase in all metrics towards the sample sizes of 250, 500 and 1000, and decrease in the sample size of 2000. In the Random Forest method. however, it has been revealed that the error values obtained in all metrics decrease as the sample size increases.

Table 6. *RMSE. MSE. MAD and MRAD Values of Both Methods in Each Sample Size.*

		250	500	1000	2000
TreeNet	RMSE	46.45	61.67	71.65	71.10
	MSE	2158.32	3803.72	5133.75	5056.28
	MAD	36.00	48.71	57.29	56.47
	MRAD	0.075	0.102	0.120	0.118
Random Forest	RMSE	96.65	93.35	92.91	90.32
	MSE	9342.81	8714.05	8633.24	8156.97
	MAD	83.72	79.61	78.70	75.69
	MRAD	0.175	0.166	0.165	0.159

4. DISCUSSION and CONCLUSION

In the current study. Bagging and Boosting algorithms were elaborated and the classification performances of TreeNet and Random Forest methods using these algorithms were compared through a real data set from a large-scale national assessment. In this section, the results yielded from both methods and the usefulness of both analysis methods in education have been discussed.

As the first result of the research, it was found that the performance measurements of TreeNet and Random Forest methods varied based on each sample size under 3, 5 and 10-fold cross validity. In its broadest sense, the TreeNet method yielded high values in accuracy, sensitivity rate, F1-score and AUC value in large samples whereas it takes high values in specificity and accuracy in smaller samples while it takes high values in specificity and accuracy in smaller samples. Furthermore, the Random Forest method takes high values in large samples in terms of specificity and accuracy although it yields high values in the smaller samples in the accuracy, sensitivity, F1-score and AUC value. In the performance measures listed above, it can be said that the Random forest method performs better in specificity and accuracy; however, the TreeNet method have a better performance in other metrics. Märker et al. (2011) noted that the TreeNet method performed better compared to the Random Forest method in terms of AUC value, Cohen's Kappa statistics and R2 value. In contrast, Mi et al. (2017) and Padmaja et al. (2021) reported in their study that the Random forest method performed better than the TreeNet method.

As the second result of the research, it has been found that with the increase in the number of samples within the TreeNet method the metric values expressing the error increase by the sample sizes of 1000 and 2000 and that it yield similar values in the sample sizes of 1000 and 2000. Instead of generating new classes through random selection from the data set, the Boosting algorithm learns from the errors and determines with which samples the incorrect classification process is performed and makes selections on these samples. In other words, considering that the Boosting algorithm acts sequentially with an iterative working principle with the logic of learning from errors by using the whole sample, the amount of error it produces in low data is reflected as less until the optimum number of trees is reached. In addition, as for the Random Forest method, it has been seen that the metric values that express the correct error from 250 samples to 2000 samples are reduced. Considering that the Bagging algorithm acts with an iterative working principle with the logic of learning from errors in order to use the random sample it yields from the data set to put back into place, it can be said that it can be said that these values decrease with the increase of the data it pulls randomly until it reaches the optimum number of trees to establish the final model. Finally, at all error rates for each sample size of the same data set, the TreeNet method has been shown to produce lower values than the Random Forest method. In this respect, it can be said that the TreeNet method produces more unbiased (Robust) results and performs better than the Random Forest method. Indeed, Padmaja et al. (2016) reported in their studies that the TreeNet method was more successful than the

Random forest method. In the same vein, in the study conducted by Subasi et al. (2022), it was reported that Stochastic Gradient Boosting method (another literature use of the TreeNet method) performed better compared to the Random Forest, Support Vector Machines, K-nearest neighbours algorithm and artificial neural networks for RMSE, MSE, MAE and RAE performance criteria. Moreover, Tuğ-Karaoğlu and Okut (2020) have stated that the Boosting algorithm is more successful than the Bagging algorithm in their study and the same authors have also drawn attention to the above-mentioned issues as the source of success. Likewise, Dietterich (2000b), Machová (2006) and Quinlan (1996) stated in their study that the Boosting algorithm was more successful than the Bagging algorithm.

When both analysis methods are compared internally, taking into account all conditions including sample size, cross-validity and performance criteria, it can be said that the TreeNet method shows higher classification and prediction performance than the Random Forest method. Märker et al. (2011) stated in their studies that the TreeNet method performed better than the Random Forest method in terms of classification performance. Similarly, Hastie et al. (2009) reported that boosting-based algorithms gave better results than bagging-based algorithms in most problem situations.

In conclusion, these conclusions have been yielded by the mathematics achievement test of the ABİDE implementation administered to 8th grade students. Further studies with higher actual and artificial data are recommended for the comparability of the results. Furthermore, it is recommended to use both analysis methods to give flexibility to the analysis of data sets obtained in the field of education, especially data that do not show parametric features.

Acknowledgments

This study was partly presented as an oral presentation at the Measurement and Evaluation Congress in Education and Psychology on 01-04 September 2021 in Ankara.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Hikmet Şevgin https://orcid.org/0000-0002-9727-5865

REFERENCES

Abdar, M., Zomorodi-Moghadam, M., & Zhou, X. (2018, 12-14, November). *An ensemble-based decision tree approach for educational data mining* [Conference presentation]. In 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), Kaohsiung, Taiwan. https://doi.org/10.1109/BESC.2018.8697318

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics, 26*(3). 392-398. https://doi.org/10.1093/bioinformatics/btp630

Abidi, S.M.R., Zhang, W., Haidery, S.A., Rizvi, S.S., Riaz, R., Ding, H., & Kwon, S.J. (2020). Educational sustainability through big data assimilation to quantify academic procrastination using ensemble classifiers. *Sustainability, 12*(15), 6074. https://doi.org/10.3390/su12156074

Aggarwal, D., Mittal, S., & Bali, V. (2021). Significance of non-academic parameters for predicting student performance using ensemble learning techniques. *International Journal of System Dynamics Applications*, *10*(3), 38-49. https://doi.org/10.4018/IJSDA.2021070103

Akman, M. (2010). *An overview of data mining techniques and analysis of Random Forests method: An application on medical field* [Unpublished master's thesis]. Ankara University.

Almasri, A., Celebi, E., & Alkhawaldeh, R.S. (2019). EMT: Ensemble meta-based tree model for predicting student performance. *Hindawi,* 1-13. https://doi.org/10.1155/2019/3610248

Amrieh, E.A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application, 9*(8), 119-136. http://dx.doi.org/10.14257/ijdta.2016.9.8.13

Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Procedia Computer Science*, *167*, 1471-1483. https://doi.org/10.1016/j.procs.2020.03.358

Ashraf, M., Salal, Y.K., & Abdullaev, S.M. (2021). *Educational Data Mining Using Base (Individual) and Ensemble Learning Approaches to Predict the Performance of Students*. In Data Science. Springer. https://doi.org/10.1007/978-981-16-1681-5_2

Arun, D.K., Namratha, V., Ramyashree, B.V., Jain, Y.P., & Choudhury, A.R. (2021, 27-29, January). *Student academic performance prediction using educational data mining* [Conference presentation]. In 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India. https://doi.org/10.1109/ICCCI50826.2021.9457021

Baskin, I.I., Marcou, G., Horvath, D., & Varnek, A. (2017a). *Bagging and boosting of classification models*. Tutorials in Chemoinformatics, 241-247. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119161110.ch15

Baskin, I.I., Marcou, G., Horvath, D., & Varnek, A. (2017b). *Bagging and boosting of regression models*. Tutorials in Chemoinformatics, 249-255. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119161110.ch16

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging. Boosting and variants. *Machine Learning. 36*(1), 105-139. https://doi.org/10.1023/A:1007515423169

Biau, G. (2012). Analysis of a Random Forest. *Journal of Machine Learning Research*, *13*(2012), 1063-1095. https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

Biau, G., & Scornet, E., (2016). A random forest guided tour. *An Official Journal of the Spanish Society of Statistics and Operations Research, 25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123-140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016, 13, August). *Xgboost: A scalable tree boosting system* [Conference presentation]. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA. http://dx.doi.org/10.1145/2939672.2939785

Clarke, B., Fokoue, E., & Zhang, H.H. (2009). *Principles and theory for data mining and machine learning.* Springer Science & Business Media. https://doi.org/10.1007/978-0-387-98135-2

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Multivariate statistics for social sciences: SPSS and LISREL applications* (2th edition). Pegem Academy.

Do-Nascimento, R.L., Fagundes, R.A., & Maciel, A.M. (2019, 15-18, July). *Prediction of School Efficiency Rates through Ensemble Regression Application* [Conference

presentation]. In 2019 IEEE 19th International Conference on Advanced Learning Technologies, Maceio, Brazil. https://doi.org/10.1109/ICALT.2019.00050

Dietterich, T.G. (2000a). Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. *Lecture Notes in Computer Science, 1857*, 1-15. https://doi.org/10.1007/3-540-45014-9_1

Dietterich, T.G. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40*(2), 139-157. https://doi.org/10.1023/A:1007607513941

Dietterich, T.G. (2002). Ensemble learning. *The Handbook of Brain Theory and Neural Networks, 2*(1), 110-125. https://courses.cs.washington.edu/courses/cse446/12wi/tgd-ensembles.pdf

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Elish, M.O., & Elish, K.O. (2009, 24-27, March). *Application of treenet in predicting object-oriented software maintainability: A comparative study.* In 2009 13th European Conference on Software Maintenance and Reengineering, Kaiserslautern, Germany. https://doi.org/10.1109/CSMR.2009.57

Ferreira, A.J., & Figueiredo, M.A. (2012). Boosting algorithms: A review of methods, theory, and applications. *Ensemble machine learning* (1th edition, 35-85). Springer. https://doi.org/10.1007/978-1-4419-9326-7_2

Freund, Y., & Schapire, R.E. (1996, 3-6, July). *Experiments with a new boosting algorithm* [Conference presentation]. Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari Italy.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*, *28*(2), 337-407. https://doi.org/10.1214/aos/1016218223

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics, 29*(5) 1189-1232. https://www.jstor.org/stable/2699986

Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J.H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*(9), 1365-1381. https://doi.org/10.1002/sim.1501

Geneur, R., Poggi, J.M., Tuleao Malot, C., & Villa-Vialaneix, N. (2017). Random forest for big data. *Big Data Research*, *9*, 28-46. https://doi.org/10.1016/j.bdr.2017.07.003

Geron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (1th edition). O'Reilly Media.

Guo, J., Bai, L., Yu, Z., Zhao, Z., & Wan, B. (2021). An AI-application-oriented in-class teaching evaluation model by using statistical modeling and ensemble learning. *Sensors, 21*(1), 241. https://doi.org/10.3390/s21010241

Han, J., Kamber, M., & Pei, J., (2012). *Data mining: concepts and techniques* (3th edition). Elsevier.

Hansen, L.K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(10), 993-1001. https://doi.org/10.1109/34.58871

Hastie, T., Tibshirani, R. & Friedman, J.H. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer. https://doi.org/10.1007/978-0-387-21606-5

Hill, T., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining* (1th edition). StatSoft, Inc.

Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832-844. https://doi.org/10.1109/34.709601

Huffer, F.W., & Park, C. (2020). A Simple Rule for Monitoring the Error Rate of Random Forest for Classification. *Quantitative Bio-Science, 39*(1), 1-15.

Injadat, M., Moubayed, A., Nassif, A.B., & Shami, A. (2020a). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems, 200*, 105992. https://doi.org/10.1016/j.knosys.2020.105992

Injadat, M., Moubayed, A., Nassif, A.B., & Shami, A. (2020b). Multi-split optimized bagging ensemble model selection for multi-class educational data mining. *Applied Intelligence, 50*(12), 4506-4528. https://doi.org/10.1007/s10489-020-01776-3

Kapucu, C., & Cubukcu, M. (2021). A supervised ensemble learning method for fault diagnosis in photovoltaic strings. *Energy,* 227, 1-12. https://doi.org/10.1016/j.energy.2021.120463

Karalar, H., Kapucu, C., & Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education, 18*(1), 1-18. https://doi.org/10.1186/s41239-021-00300-y

Kausar, S., Oyelere, S., Salal, Y., Hussain, S., Cifci, M., Hilcenko, S., ... & Huahu, X. (2020). Mining smart learning analytics data using ensemble classifiers. *International Journal of Emerging Technologies in Learning, 15*(12), 81-102. https://www.learntechlib.org/p/217561/

Keser, S.B., & Aghalarova, S. (2022). HELA: A novel hybrid ensemble learning algorithm for predicting academic performance of students. *Education and Information Technologies, 27*(4), 4521-4552. https://doi.org/10.1007/s10639-021-10780-0

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, *23*(6), 529-535. https://doi.org/10.1016/j.knosys.2010.03.010

Kumari, G. T. (2012). A Study of Bagging and Boosting approaches to develop meta-classifier. *Engineering Science and Technology: An International Journal, 2*(5), 850-855.

Leedy, P.D., & Ormrod, J.E. (2005). *Practical research (Vol. 108).* Saddle River.

Lee, S.L.A., Kouzani, A.Z., & Hu, E. J. (2010). Random forest based lung nodule classification aided by clustering. *Computerized Medical Imaging and Graphics*, *34*(7), 535-542. https://doi.org/10.1016/j.compmedimag.2010.03.006

Li, B., Yu, Q., & Peng, L. (2022). Ensemble of fast learning stochastic gradient boosting. *Communications in Statistics-Simulation and Computation*, *51*(1), 40-52. https://doi.org/10.1080/03610918.2019.1645170

Machová, K., Puszta, M., Barčák, F., & Bednár, P. (2006). A comparison of the bagging and the boosting methods using the decision trees classifiers. *Computer Science and Information Systems*, *3*(2), 57-72. https://doi.org/10.2298/CSIS0602057M

Maclin, R., & Opitz, D. (1997, 27-31, July). *An empirical evaluation of bagging and boosting* [Conference presentation]. *AAAI-97:* Fourteenth National Conference on Artificial Intelligence, Rhode Island.

Märker, M., Pelacani, S., & Schröder, B. (2011). A functional entity approach to predict soil erosion processes in a small Plio-Pleistocene Mediterranean catchment in Northern Chianti, Italy. *Geomorphology, 125*(4), 530-540. https://doi.org/10.1016/j.geomorph.2010.10.022

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *Peer J, 5*, e2849.

Mousavi, R., & Eftekhari, M. (2015). A new ensemble learning methodology based on hybridization of classifier ensemble selection approaches. *Applied Soft Computing*, *37*, 652-666. https://doi.org/10.1016/j.asoc.2015.09.009

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications* (1th edition). Academic Press.

Olson, D.L., & Delen, D. (2008). *Advanced data mining techniques.* Springer Science & Business Media.

Onan, A. (2015). On the performance of ensemble learning for automated diagnosis of breast cancer. R. Silhavy R. Senkerik, Z. K. Oplatkova, Z. Prokopova, & P. Silhavy (eds.), *In Artificial Intelligence Perspectives and Applications: Proceedings of the 4th Computer Science On-line Conference, Vol 1* (pp. 119-129). Springer International Publishing.. https://doi.org/10.1007/978-3-319-18476-0_13

Opitz, D.W., & Shavlik, J.W. (1996). Generating accurate and diverse members of a neural network ensemble. *Advances in Neural Information Processing Systems*, *8*, 535-541.

Padmaja, B., Prasad, V.R., & Sunitha, K.V.N. (2016). TreeNet analysis of human stress behavior using socio-mobile data. *Journal of Big Data*, *3*(1), 1-15. https://doi.org/10.1186/s40537-016-0054-3

Padmaja, B., Srinidhi, C., Sindhu, K., Vanaja, K., Deepika, N.M., & Patro, E.K.R. (2021). Early and accurate prediction of heart disease using machine learning model. *Turkish Journal of Computer and Mathematics Education, 12*(6), 4516-4528.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine, 6*(3). 21-45. https://doi.org/10.1109/MCAS.2006.1688199

Polikar, R. (2012). Ensemble learning. *In Ensemble machine learning* (1th edition pp. 1-34). Springer. https://doi.org/10.1007/978-1-4419-9326-7_1

Premalatha, N., & Sujatha, S. (2021, 15-17, September). *An Effective Ensemble Model to Predict Employment Status of Graduates in Higher Educational Institutions* [Conference presentation]. In 2021 Fourth International Conference on Electrical, Computer and Communication Technologies Erode, India. https://doi.org/10.1109/icecct52121.2021.9616952

Probst, P., & Boulesteix, A.L. (2017). To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research, 18*(1), 6673-6690. http://jmlr.org/papers/v18/17-269.html

Rokach, L. (2019). *Ensemble learning: Pattern classification using ensemble methods* (2th edition). World Scientific. https://doi.org/10.1142/9789811201967_0003

Pong-Inwong, C., & Kaewmak, K. (2016, 14-17, October). *Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration* [Conference presentation]. In 2016 2nd IEEE international conference on computer and communications, Chengdu, China. https://doi.org/10.1109/CompComm.2016.7924899

Quinlan, J.R. (1996, 4-8, August). *Bagging, boosting, and C4. 5* [Conference presentation]. In 13th National Conference on Artificial Intelligence, Portland, Oregon, USA.

Saeys, Y., Abeel, T., & Peer, Y.V.D. (2008). Robust feature selection using ensemble feature selection techniques. W. Daelemans, B. Goethals & K. Morik (Eds.), *Machine learning and knowledge discovery in databases* (pp 313-325) Springer. https://doi.org/10.1007/978-3-540-87481-2_21

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery. 8*(4). e1249. https://doi.org/10.1002/widm.1249

Schapire, R.E. (2003). The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification,* 149-171. https://doi.org/10.1007/978-0-387-21579-2_9

Schroeder, M.A., Lander, J., & Levine-Silverman, S. (1990). Diagnosing and dealing with multicollinearity. *Western Journal of Nursing Research, 12*(2), 175-187. https://doi.org/10.1177/019394599001200204

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. Educational Measurement: *Issues and Practice, 35*(3), 38-54. https://doi.org/10.1111/emip.12115

Skurichina, M., & Duin, R.P. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications, 5*(2), 121-135. https://doi.org/10.1007/s100440200011

Steinki, O., & Mohammad, Z. (2015). Introduction to ensemble learning. *Available at SSRN, 1*(1), 1-9. http://dx.doi.org/10.2139/ssrn.2634092

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323. https://doi.org/10.1037/a0016973

Subasi, A., El-Amin, M.F., Darwich, T., & Dossary, M. (2022). Permeability prediction of petroleum reservoirs using stochastic gradient boosting regression. *Journal of Ambient Intelligence and Humanized Computing, 13,* 3555-3564. https://doi.org/10.1007/s12652-020-01986-0

Sutton, C.D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of Statistics, 24*, 303-329. https://doi.org/10.1016/S0169-7161(04)24011-1

Şevgin, H. (2020*). Predicting the ABIDE 2016 science achievement: The comparison of MARS and BRT data mining methods [Unpublished Doctoral Thesis].* Gazi University.

Şevgin, H., & Önen, E. (2022). Comparison of Classification Performances of MARS and BRT Data Mining Methods: ABİDE-2016 Case. *Education and Science, 47*(211). http://dx.doi.org/10.15390/EB.2022.10575

Tabachnick, B.G., & Fidell, L.S. (2015). *Using multivariate statistics* (6th edition). (M. Baloğlu, Trans.). Nobel Publications. (Original work published 2012).

Ting, K. M. (2017). Confusion matrix. In C. Sammut & G. I. Webb (Eds.) *Encyclopedia of Machine Learning and Data Mining* (pp. 260–260). Springer.

Tuğ Karoğlu, T.T., & Okut, H., (2020). Classification of the placement success in the undergraduate placement examination according to decision trees with bagging and boosting methods. *Cumhuriyet Science Journal, 41*(1), 93-105. https://doi.org/10.17776/csj.544639

Wang, Z., Wang, Y., & Srinivasan, R.S. (2018). A novel ensemble learning approach to support building energy use prediction. *Energy and Buildings, 159*, 109-122. https://doi.org/10.1016/j.enbuild.2017.10.085

Yurdugül, H. (2006). The comparison of reliability coefficients in parallel, tau-equivalent, and congeneric measurements. *Ankara University Journal of Faculty of Educational Sciences*, *39*(1), 15-37. https://doi.org/10.1501/Egifak_0000000127

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: methods and applications.* Springer. https://doi.org/10.1007/978-1-4419-9326-7

Zhou Z.H. (2012). *Ensemble methods: foundations and algorithms.* Chapman and Hall/CRC.