ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 325-345

# Comparison of Data Mining Classification Algorithms on Educational Data under Different Conditions *

İlhan KOYUNCU ** Selahattin GELBAL ***

**Abstract**
The purpose of this study was to examine the performance of Naive Bayes, *k*-nearest neighborhood, neural networks, and logistic regression analysis in terms of sample size and test data rate in classifying students according to their mathematics performance. The target population was 62728 students in the 15-year-old group who were participated in the Programme for International Student Assessment (PISA) in 2012 from The Organisation for Economic Co-operation and Development (OECD) countries. The performance of each algorithm was tested by using 11%, 22%, 33%, 44% and 55% of each dataset for small (500 students), medium (1000 students) and large (5000 students) sample sizes. 100 replications were performed for each analysis. As the evaluation criteria, accuracy rates, RMSE values, and total elapsed time were used. RMSE values for each algorithm were statistically compared by using Friedman and Wilcoxon tests. The results revealed that while the classification performance of the methods increased as the sample size increased, the increase of training data ratio had different effects on the performance of the algorithms. The Naive Bayes showed high performance even in small samples, performed the analyzes very quickly, and was not affected by the change in the training data ratio. Logistic regression analysis was the most effective method in large samples but had a poor performance in small samples. While neural networks showed a similar tendency, its overall performance was lower than Naive Bayes and logistic regression. The lowest performances in all conditions were obtained by the *k*-nearest neighborhood algorithm.

*Key Words:* Artificial neural networks, educational data mining, *k*-nearest neighborhood, logistic regression, naive Bayes

## INTRODUCTION

Data mining is used to discover hidden patterns and relationships that help decision making by processing large amounts of data (Bhardwaj & Pal, 2011). A wide variety of methods based on mathematical and statistical algorithms are used to predict, cluster, and reveal relationship networks in many disciplines. Data mining has its roots in machine learning, artificial intelligence, computer science, and statistics (Dunham, 2003). Data mining methods, which are used in a wide range from marketing to engineering, from health sciences to business, have started to be used to examine large and complex educational datasets that have been increasing rapidly with technological developments. Although data mining is applied to a large number of industries and sectors, its applications in the context of education are limited (Ranjan & Malik, 2007).

Predicting student success is the focus of many kinds of research in education. In particular, today, while technology is developing rapidly and gaining more importance in education, there are databases that contain many factors that affect student success. In addition to the course management systems that include rich educational data sources such as Blackboard and Moodle, data is collected at the student, teacher, school, regional and country level in large scale assessments such as Trends In International

* This study is a part of the doctoral dissertation "Comparison of data mining methods in predicting PISA mathematical achievements of students".
** Assist. Prof., Adiyaman University, Faculty of Education, Adıyaman-Turkey, ilhankync@gmail.com, ORCID ID: https://orcid.org/0000-0002-0009-5279
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, sgelbal@gmail.com, ORCID ID: orcid.org/0000-0001-5181-7262

To cite this article:
Koyuncu, İ., & Gelbal, S. (2020). Comparison of data mining classification algorithms on educational data under different conditions . *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 325-345. doi: 10.21031/epod.696664.

*Received: 01.03.2020*
*Accepted: 12.11.2020*

Mathematics And Science Study (TIMMS), Programme for International Student Assessment (PISA), and Progress In International Reading Literacy Study (PIRLS). It is increasingly getting important in recent years to predict and compare students' performances by analyzing large educational datasets. For this purpose, educational data mining (EDM) has emerged as an independent research area in recent years (Baker, 2010).

EDM is a new discipline that emerged in order to apply data mining techniques to educational data (Baker & Yacef, 2009; Huebner, 2013). It can be used in various areas of education, from the effectiveness of teaching programs to predict student success, from educational institutions to the performance of teachers. There are different definitions of EDM in the related literature. According to Baker and Yacef (2009), EDM focuses on the development of new methods to make discoveries from characteristics data obtained from educational settings. EDM is a scientific research area that uses these methods to understand better students and learning environments (Baker & Yacef, 2009). However, Huebner (2013) considers that such definitions are limited, EDM covers an extensive educational area, and the scope and definitions of this area will change with future studies.

Romero and Ventura (2007) stated that data mining in education is an iterative cyclical process consisting of hypothesis creation, testing, and development. In this process, educators and academic specialists have the responsibility to design, plan, and develop educational systems. The outputs (demographic data, course information, academic data, etc.) obtained by the students' use and interaction with these systems can be used in data mining for various purposes (clustering, classification, association, etc.). The useful information discovered can be used by both educators and students (Romero & Ventura, 2007).

Baker (2010) stated that a wide variety of popular methods used in educational data mining are classified under five main categories: Prediction, clustering, discovering relationships, discovery with models, and distillation of data to evaluate individuals. Prediction makes inferences about a single piece of the data by using the other variables making up the majority of the data. An example of this is the use of features such as anxiety, attitude, self-efficacy, etc., in the rest of the data in order to make inferences about students' mathematics performance. Classification of individuals or observations according to a certain categorical variable is one of the most basic prediction techniques in data mining (Baker, 2010). Some popular prediction algorithms are decision trees, logistic regression, support vector machines, artificial neural networks, Bayes algorithms, k-nearest neighborhood, and density estimators based on various kernel functions. In order to evaluate the accuracy of an estimator, criteria such as converted performance metrics based on the error matrix (precision, recall, F criterion, etc.), root mean square error (RMSE), Kappa (Cohen, 1960) concordance coefficient, area under the ROC curve (Egan, 1975) and error rates are used.

In order to test the performance of algorithms in data mining, data is divided into two parts: training and test data. In this method, initial analyses are performed using a specific part of a data set (training data), and a predictive model is created. In the next step, by making use of this model, the prediction is made for individuals or objects in the rest of the data (test data). The reason for testing the performances of methods in data mining in this way is to avoid biased estimates of model error rates. The other methods used for similar purposes are bootstrapping (Efron, 1983) and cross-validation (Lachenbruch & Mickey, 1968) techniques (Michie, Spiegelhalter & Taylor, 1994). However, selecting one-third (33%) of all data as a test dataset and the rest of the data (67%) as training data is often preferred and used mostly for large samples to test the performance of the algorithms. In many studies in the field of data mining, the effect of the train/test ratio (e.g. Brain & Webb, 1999; Çölkesen, & Kavzoglu, 2010; Foody, Mathur, Sanchez-Hernandez, & Boyd, 2006; Heilman, & Madnani, 2015; Shao, Fan, Cheng, Wu & Cheng, 2013; Tadjudin & Landgrebe, 1998; Tayeh et al., 2015) and sample size (e.g. Beleites et al., 2013; Chu et al., 2012; Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012; Heydari, SS, & Mountrakis, 2018; Raudys & Pikelis, 1980; Wharton, 1984) on the performances of the algorithms were assessed. For example, Brain and Webb (1999) showed that when the amount of test data was increased, the error variance decreased, but there was no significant change in bias. Tadjudin and Landgrebe (1998) developed a robust parameter estimation method that reduces the effect of varying test data rates by stating that the limited amount of test data causes errors in classification performance. Foody et al. (2006) stated that even a

90% reduction in the rate of test data did not cause a decrease in some algorithms' performance. Heilman and Madnani (2015) found that increasing test data increased performance, but increasing sample size did not have the same effect. Shao et al. (2013) showed that the minimum rate of test data can be found for some methods. Çölkesen and Kayzoğlu (2010) found in their study that some methods show higher performance in small training sets than others. In the present study, the ideal amount of test and training data are examined for educational data.

In classification studies in the field of education, the performance of methods such as decision trees, support vector machines, logistic regression, neural networks, Bayes algorithms, *k*-nearest neighborhood are examined and compared (e.g., Bahadır, 2013; Barker, Trafalis & Rhoads, 2004; Berens, Schneider, Gortz, Oster, & Burghoff, 2019; Çırak, 2012; Dekker, Pechenizkiy & Vleeshouwers, 2009; Göker, 2012; Hamalainen & Vinni, 2006; Hamalainen & Vinni, 2011; Minaei-Bidgoli, Kashy, Kortemeyer & Punch, 2003; Osmanbegović & Suljić, 2012; Romero, Espejo, Zafra, Romero & Ventura, 2013; Romero, Ventura, Espejo & Hervas, 2008; Shahiri, Husain & Rashid, 2015; Sweeney, Lester, Rangwala, & Johri 2016; Şengür, 2013; Tepehan, 2011; Tezbaşaran, 2016; Tosun, 2007; Yurdakul & Topal, 2015). In addition, methods were compared according to the different number of categories of the dependent variable (Minaei-Bidgoli, Kashy, Kortemeyer & Punch, 2003; Nghe, Janecek & Haddawy, 2007), the data structure (Romero et al., 2008; Romero et al., 2013), amount of missing and noisy data ( Hamalainen & Vinni, 2011) and sample sizes (Hamalainen & Vinni, 2006; 2011).

In the literature, in general, it can be seen that different results are obtained for different data structures. For example, in their study, Kotsiantis et al. (2003) compared some data mining methods; the Naive Bayes algorithm generally yielded better results than any other method. In the study conducted by Tosun (2007), artificial neural networks showed about 92% correct classification performance, while decision trees showed 86% accuracy. In the research conducted by Tepehan (2011) with PISA data, neural networks were as successful as logistic regression. Çırak (2012) found that the correct classification performance (66.1%) of logistic regression analysis was lower than the performance of artificial neural networks (70.16%). Similarly, Bahadır (2013) showed that the prediction performed with artificial neural networks was better with the logistic regression method. Göker (2012) compared many methods to develop a program for predicting students' success before taking an exam and used the Naive Bayes method, which has the highest correct classification rate (87.27%).

Minaei-Bidgoli et al. (2003) have shown that increasing the number of categories of the dependent variable causes significant performance differences in all mining methods, especially in Naive Bayes and k-nearest neighborhood methods. In their study, Nghe et al. (2007) showed that decision trees produce better results than Bayes networks for the different number of categories of the dependent variable. In the study conducted by Barker et al. (2004), when different training and test datasets of different years were combined for the same data structure, different techniques produced the same results, and neural network methods showed good performance when the data of previous years were used as a training set.

In their study, Hamalainen and Vinni (2006) showed that when more variables are added to the model, the support vector machines perform better in small samples; while the number of variables is less, Bayes algorithms show higher performance. Hamalainen and Vinni (2011), while Naive Bayes classifiers are effective for their accuracy in small samples, neural networks and nearest neighborhood classifiers require much larger samples. In another study, Osmanbegović and Suljić (2012) compared Naive Bayes (76.65%), decision trees (73.93%), and artificial neural network (71.20%) methods, and they found that neural networks method took a little time for training the algorithm while other methods did not.

In their study, Sweeney et al. (2016) analyzed students' versatile data with many data mining methods in order to estimate the attendance status of students and found the least erroneous results with Factorization Machines (FM), Random Forests (RF), the Personalized Linear Multiple Regression, and hybrid FM-RF methods. Tezbaşaran (2016) compared the generalized Hebb algorithm and principal component analysis results to confirm the data structure of a scale. She found that the two structures were very similar, and the error and fit indexes were very close to each other. Berens et al. (2019) showed

that the AdaBoost algorithm, which combines regression analysis, neural networks, and decision trees, is effective instead of using a single algorithm in predicting school attendance status through longitudinal data of students attending at two German universities. In the present study, unlike the previous studies, we aimed to compare the performances of Naive Bayes, k-nearest neighborhood, logistic regression, and neural networks classifiers in terms of sample size and test data rate. Therefore, the general structure of these methods will be briefly explained.

One of the most used classification algorithms in data mining is the Naive Bayes method based on Bayes' theorem. This classifier performs comparable performance with decision trees and neural networks classifiers in predicting probabilities of class memberships. The classifier calls "naive" because of the assumption that any value of a property belongs to a class is independent of the probability that other properties' values belong to the same class (Han, Kamber, & Pei, 2011). While this classifier has advantages such as being simple, useful, easy to interpret, and resistant to complexity, it can be used in small data sets and applied to categorical and continuous data (from Gauss distribution). There are disadvantages, such as the fact that the assumption of conditional independence is difficult to provide, and in the categorical data, when the limits of classes are complex, it is difficult to estimate its power (Hamalainen & Vinni, 2011).

Another method most commonly used is the *k*-nearest neighborhood algorithm. This algorithm is mostly used for classification purposes besides estimation and prediction. The method is based on the principle of classifying a new sample according to its similarity with the samples in training data (Larose, 2004). The class in which the sample will be assigned can be the most common class among neighboring samples or a neighboring class distribution. The most important problems to be encountered in calculations are what will be the value of *k* and how to calculate the distance (*d*). Another question that may come to mind is how to weight the sample cases in the training set. This algorithm's advantages are that there are only two parameters (*k* and *d*) in training the model and classification. The classification performance is very well in some problems, and the classification is robust to the complexity and missing data. The most important disadvantage is that there are difficulties in choosing the distance function (*d*) and *k* value (Hamalainen & Vinni, 2011).

Artificial neural networks (ANN) are used to discover relationships and patterns in a data set using certain mathematical and statistical algorithms. As a result of training neural networks, guiding information is obtained in making certain decisions (Sivanandam, Sumathi, & Deepa, 2006). ANN is used effectively in almost every field, especially in computer sciences, engineering, cognitive sciences, neurophysiology, physics, biology, environmental science, and marketing. When applied in educational technologies, it can be problematic if there is not enough numerical data, and it is exactly not known how to train the model (Hamalainen & Vinni, 2011). ANN was developed using the structure of biological cell networks. Neural networks, a subject that has been studied since the 1940s, have been reported in the form of many network architectures in the literature because of the complexity of the structures of real nerve cells and inadequate understanding of their working principles (Sivanandam et al., 2006). Some of the advantages of ANN are that they can easily learn nonlinear boundaries, represent basically different types of classifiers, fully convert variables when they are not discriminatory, robust to complexity (noise), and update themselves with new data. Some disadvantages are that ANNs require more data than typical data sets in education. They are very sensitive to overfitting. They require numerical data, and categorical data should be quantitated (Hamalainen & Vinni, 2011).

Logistic regression analysis is one of the prediction and classification algorithms that are used more than many other data mining methods. This analysis method effectively predicts group memberships when the predicted variable is categorical, and the predictors are categorical, continuous, or a mixture of the two. Discriminant analysis and multiple regression methods seek answers to similar research problems in logistic regression. However, the logistic regression has no strict assumptions such as normality, linearity, homogeneity of variances, etc. (Cox & Snell, 1989; Tabachnick & Fidell, 2013). This analysis method proposed in the early 1960s (Cabrera, 1994) began to take place as a routine package in statistical software since the early 1980s (Peng, Lee, & Ingersoll, 2002). It has become a frequently used method in social sciences and education until today (Cabrera, 1994; Peng & So, 2002). The logistic regression analysis has become popular by means of its advantages, such as being effective in a wide variety of

complex data sets and a lack of assumptions about the distribution of predictive variables. However, in order for the analysis to be effective, it is required that the predictors are well-chosen and have a theoretical basis, there are sufficient samples in variables and category distributions, there is a linear relationship between continuous predictors and logit of the predicted variable, there is no multicollinearity and extreme values, errors and observations are independent of each other. (Tabachnick & Fidell, 2013).

It is possible to come across many studies on the applicability and effectiveness of data mining methods on educational data in the last decade. These researches aim to predict and evaluate student performance in general and to determine the factors affecting performance. However, only a few of these studies addressed the impact of sample size and training data size on the performance of these algorithms, as well as the comparison of data mining algorithms. In addition, studies on EDM and related to Naive Bayes and *k*-nearest neighbor techniques (e.g., Göker, 2012; Yurdakul & Topal, 2015) are limited in Turkey. In the present study, it was aimed to make a comprehensive application by using the data received in PISA (2012) assessment for these deficiencies in the literature.

In addition, data mining techniques have been used in order to predict and classify students' PISA performance in recent studies (e.g., Aksu, & Guzeller, 2016; Bulut, & Yavuz, 2019; Gorostiaga, & Rojo-Álvarez, 2016; Güre, Kayri, & Erdoğan, 2020; Kiray, Gok, & Bozkir, 2015; Martínez-Abad, Gamazo, & Rodríguez-Conde, 2020; Tepehan, 2011). For example, Kiray, Gok, & Bozkir (2015) examined the factors influencing Turkish students' performances in TIMMS 1999 and PISA 2003/2006 studies. Similarly, Aksu and Güzeller (2016) found that CHAID analysis and J.48 decision tree methods in data mining effectively classify Turkish students participating in PISA 2012 study. Moreover, Gorostiaga, and Rojo-Álvarez (2016), proposed a feature selection method in predicting Spanish students' PISA 2009 performance by using data mining techniques in addition to logistic regression. Besides, Bulut and Yavuz (2019) developed "Rattle" which is a R package used to apply data mining with graphical representations by using PISA 2015 data. Martínez-Abad et al. (2020) found that as a data mining technique, decision trees were more effective in explaining inter-school variance when compared to hierarchical linear modeling for PISA 2015 Spanish data. Güre et al. (2020) the performances of multilayer perceptron and random forest methods of data mining in determining factors affecting students' PISA 2015 mathematics literacy. In the literature related to PISA and data mining, the efficiency of different methods in predicting or classifying students' success and development of new techniques or tools were investigated.

As education systems are evaluated worldwide by PISA studies, a careful and systematic way is followed at every stage of the data collection process. Therefore, at the end of each application, a large data pool with high reliability and validity is obtained in terms of measurement and evaluation processes. Since the data of PISA (2012) assessment is used in the present study, the results obtained for the methods are considered to be important for the theory and real-life practice. In addition, in order to increase the reliability of the results obtained from different performance criteria, different data sets were selected by putting with replacement method, and the analyzes were replicated 100 times. Thus, we aimed to obtain results with high precision on real data regarding the methods used in the area of educational data mining.

### *Purpose of the Study*

The aim of this study is to examine the performance of Naive Bayes, k-nearest neighborhood, neural networks, and logistic regression analysis in terms of sample size and training data ratio in classifying students according to their PISA mathematics performance. In accordance with this purpose, the sub-goals are to test whether;

- The performances of algorithms vary for small, medium, and large sample sizes,
- The performances of algorithms vary for different test data ratios,
- There is also a common effect of different sample sizes and test data ratios,

- Some of these algorithms perform better/worse under different conditions or not.

For this purposes, it is sought to find answers to the following research problem: For sample sizes of 500, 1000, and 5000 students, do the performances of Naive Bayes, k-nearest neighborhood, multilayer perceptron methods of artificial neural networks, and logistic regression methods differ for the ratio of test data 11%, 22%, 33%, 44%, and 55% in predicting students' PISA mathematics achievement?

## METHOD

Since it is aimed to determine and explain the performances of Naive Bayes, k-nearest neighborhood, artificial neural networks, and logistic regression algorithms under different conditions, the present study is fundamental research. In this type of studies, it is aimed to produce knowledge by conducting studies based on methodological analysis (Büyüköztürk, Çakmak-Kılıç, Akgün, Karadeniz & Demirel, 2015). Fundamental research aims to add new information to existing knowledge (Karasar, 2005). Research is also quantitative relational research in terms of examining the relationships between methods. Relational studies aim to seek, explain, and discover the relationships between quantitative variables (Fraenkel & Wallen, 2006).

### *Sample*

The research population of the study is 15 years-old students from OECD countries.The samples representing the population for each country were selected by PISA practitioners through stratified random sampling. The total number of people participating in the PISA (2012) assessment from OECD countries is 295416 students. In this study, after the missing data, residual and extreme values were examined and extracted, the target population of 62728 students was obtained. Table 1 shows the distribution of students in the target population by OECD countries.

Table 1. Distribution of The Target Population by OECD Countries

| *Country* | *f* | *%* | *Country* | *f* | *%* | *Country* | *f* | *%* |
|---|---|---|---|---|---|---|---|---|
| Australia | 2982 | 4.75 | Finland | 2001 | 3.19 | Mexico | 6062 | 9.66 |
| Austria | 976 | 1.56 | France | 993 | 1.58 | Holland | 1054 | 1.68 |
| Belgium | 1754 | 2.80 | UK | 2647 | 4.22 | Norway | 1032 | 1.65 |
| Canada | 4910 | 7.83 | Greece | 1190 | 1.90 | New Zeland | 852 | 1.36 |
| Switzerland | 2558 | 4.08 | Hungary | 1088 | 1.73 | Poland | 1010 | 1.61 |
| Chile | 1480 | 2.36 | Ireland | 1237 | 1.97 | Portugal | 1210 | 1.93 |
| Czech Republic | 1339 | 2.13 | Iceland | 780 | 1.24 | Slovakia | 1072 | 1.71 |
| Germany | 833 | 1.33 | Italy | 7479 | 11.92 | Slovenia | 1269 | 2.02 |
| Denmark | 1614 | 2.57 | Japan | 1512 | 2.41 | Sweden | 977 | 1.56 |
| Spain | 5502 | 8.77 | Korea | 1242 | 1.98 | Turkey | 834 | 1.33 |
| Estonia | 1140 | 1.82 | Luxemburg | 1017 | 1.62 | USA | 1082 | 1.72 |
| Total | 62728 | 100.0 | | | | | | |

In data mining, the sample to be used in analysis is expressed as 'medium' when it consists of 1000 subjects, 'small' when it has less than this value, and 'large' when it has more than this value (Michie, Spiegelhalter & Taylor, 1994). In the sample selection, the bootstrapping method recommended by Efron (1983) was used. Accordingly, the samples of the research are 500 (small), 1000 (medium), and 5000 (large) students selected randomly by putting with replacement from the target population. In this sample selection method, the probability of each individual being selected is equal. In order to obtain results with high precision regarding the performance of the methods studied, a total of 180 datafiles consisting of 100 datafiles each including a sample of 500 students, 50 datafiles each including a sample of 1000 students, and 30 datafiles each including a sample of 5000 students were created. As the sample size decreases, the reason why more data files were drawn from all the data is to avoid biased or erroneous generalizations and increase the representativeness of small samples. To prevent the fact that different researchers can obtain different results with the same datasets, the weighted average of analysis

results obtained with these datasets were evaluated by considering standard deviations of 100 replications.

### *Data Collection Instruments*

The data collection tools of this study are mathematics cognitive test developed to measure students' academic performance in PISA (2012) assessment and a student questionnaire prepared to evaluate the students with all their existing characteristics. PISA study is an assessment that examines 15-year-old students' knowledge and skills in mathematics, science, and reading in order to evaluate and compare education systems worldwide in three-year periods (OECD, 2014b). Mathematics cognitive test consists of change and relationships, quantities, distances and shapes, uncertainty and data, tasks, formulation, and interpretation subfields. The test items consist of a mixture of multiple-choice items and items that students create their own answers. In the student questionnaire, students were expected to fill in forms containing various information about themselves, their homes, schools, and learning experiences. Besides the student questionnaire, one of the questionnaires that some countries chose for their students is related to the students' familiarity with the information and communication technologies, and the other is related to students' education processes that question whether they are in preparation for a career for their future or a break during their education process. The student questionnaire consisting of three forms has 53 items in two forms and 54 items in the other. While each of these forms used in the PISA assessment is answered by one-third of the students, there are also students who answer the two forms in addition to the common items in the forms (OECD, 2014a).

### *Data Collection Procedure*

In this study, open-access data obtained by PISA practitioners (OECD, 2014a) were taken from the OECD's public database. Detailed information about the data collection process in PISA assessment can be found in PISA documents (see OECD, 2014a; 2014b).

### *Data Analysis*

In this study, a systematic process was followed in preparing data for the analysis. Firstly, data from OECD countries was drawn from PISA student questionnaire data. The demographic variables and all variables related to mathematics were taken from this existing data file. Then, considering the PISA 2012 technical report published by OECD (2014b), variables consisting of the combination of other variables were taken, and the remaining variables were removed from the file. In the data obtained, all individuals containing missing data related to basic affective variables such as math anxiety and math self-efficacy were excluded from the data. Thus, out-of-school mathematics lessons, class size, basic and applied mathematics experience in school, familiarity with mathematical concepts, time devoted to mathematics lessons, and out-of-school working time consisted of completely missing data. The stratum variable was not interpreted similarly in every country, and in some countries, school type was added as a layer. In this case, when a particular sample is selected from all data, some cells of this independent variable remain empty, and this is especially problematic for logistic regression analysis. A similar situation is valid for the test language variable. For these reasons, when all the mentioned variables above are removed from the analysis and all missing data, and extreme values in the file are deleted, the target population consisting of 35 variables and 62728 students was obtained.

Although data mining algorithms work with a lot of variables, keeping the variables that do not contribute to the classification causes the analysis to take a lot of time and decrease the classification performance. For this purpose, variable (feature) selection, which is a data preprocessing process, is one of the important techniques frequently used in data mining (Blum & Langley, 1997; Liu & Motoda, 2001). Variable selection methods designed according to different evaluation criteria are generally divided into three categories as filtering, winding, and hybrid models (Liu & Yu, 2005). Models other than filtering models require an analysis method to define the significance of variables in classification.

In this study, since different analysis methods were compared, the filtering method, which allows sorting the variables according to gaining the information, was used without requiring an additional analysis method. The filtering method aims to select and evaluate the subset of variables based on the general characteristics of the data, without including any data mining method (Liu and Yu, 2005).

In this study, Information Gain Ranking Filter, Chi-Squared Ranking Filter, Gain Ratio Feature Evaluator, and Symmetrical Uncertainty Ranking Filter in WEKA Version 3.9.0 software (Hall et al., 2009) methods are used to select variables for the analysis. Information Gain Ranking Filter measures the information obtained by classes; Chi-Squared Ranking Filter calculates the Chi-square value according to the class; Gain Ratio Feature Evaluator measures the ratio obtained according to the class; Symmetrical Uncertainty Ranking Filter measures symmetric uncertainty by class and evaluates the importance order of a variable (Frank, Hall & Witten, 2016).

In the present study, the variable selection process was performed on the data belonging to the target population (N = 62728). As the dependent variable, the first of 5 plausible values (PV1MATH) corresponding to students' mathematics performance was used. Plausible values correspond to the ability distribution a student may have, based on the students' responses to the items, and are obtained by subtracting random values from the posterior probability distribution for the Ɵ ability values in the Item Response Theory (IRT) (Wu, 2005). In the simulation study conducted by Wu (2005), it was found that using any of the plausible values alone is sufficient to estimate the population parameters with high accuracy. Therefore, the first plausible value 'PV1MATH' variable was converted to a new variable with two categories that represent the students below and above the medium level (482) according to proficiency levels determined by PISA practitioners (OECD, 2014a). Then, by doing feature selection analyses, the top 10 variables that have the greatest contribution to the classification of students according to their mathematics performance were selected. Then, the first 10 variables that have the greatest contribution to the classification of students according to all filtering methods in terms of mathematics performance were selected. These variables are mathematics self-efficacy, mathematics self-concept, mathematics anxiety, economic, social and cultural status index, openness to the problem solving, country, father's education level (ISCED), mothers' education level (ISCED), teacher behavior: directing students, and calculator use. In this study, all analyzes were performed by using these variables.

After selecting the variables to be used in the analysis, the assumptions and prerequisites of the algorithms were checked. Although logistic regression (LR) analysis does not require any assumptions regarding the distribution of independent variables, the ratio of the number of individuals to the number of variables, the suitability of the expected frequencies, the moderate linear relationship between the continuous variables, the absence of missing and extreme values, and the sufficient model fit values are some preconditions for the analysis (Tabachnick and Fidell , 2013). Although the Naive Bayes (NB) algorithm is based on the conditional independence of all independent variables, this assumption is rarely provided, but this algorithm still yields good results (Hamalainen & Vinni, 2011).

In the *k*-nearest neighborhood algorithm (KNN), choosing the appropriate *k* value and *d* distance criteria are important requirements (Larose, 2004). One of the most used methods for selecting the most appropriate *k* value are taking the square root of the sample size of training data (Dunham, 2003). Some researchers suggest that it is difficult to make a definitive judgment, but they recommend to try values close to this value, to use odd and prime numbers, to use Bayes methods, and *k*-layered cross-validation (Aha, Kibler & Albert, 1991; Ghosh, 2006; Hall, Park & Samworth, 2008). In this study, the square root of the number of students in the training data is taken (Dunham, 2003), and the most appropriate *k* number is selected for each analysis with the *k*-fold cross-validation method (Frank, Hall, and Witten, 2016).

Although the selection of the number of layers is an important issue in the multilayer perceptron (MLP) algorithm of artificial neural networks, reasonable results are obtained in educational data when there is enough numerical data, and the model is well trained (Hamalainen & Vinni, 2011). In this study, in order to select an appropriate number of layers, the values 1 to 5 were tested as the number of layers for the rate of test data of 33%. More than 5 values are not tried because when the number of layers increases, the model becomes complicated, and the analyzes take a lot of time. The experimental design for the number of layers revealed that 3 gives the ideal results. In addition, Akpınar (2014) states that it

will be sufficient to select 3 layers in the solution of many classification problems. Still, it will be useful to examine additional layers if necessary to save time. For these reasons, the number of layers was taken 3 for the multilayer perceptron as the artificial neural network algorithm used in this study.

In the present study, in order to determine the standard conditions that the analysis was performed, some important assumptions and prerequisites were checked, and the following results have been obtained.

- The sample size is sufficient.
- There are no missing and extreme values.
- Continuous variables do not show a significant deviation from the standard normal distribution.
- Variance and covariance matrices are not homogenous.
- Linear relationships between variables are at a low or medium level.
- There is no multicollinearity or singularity problem.
- Conditional independence assumption could not be achieved for the Naive Bayes algorithm.

After the data were prepared for analysis, for each algorithm and datafiles (180 files), the analyses were performed for critical test data ratios 11%, 22%, 33%, 44%, and 55%. Although selecting one-third (33%) of all data as test dataset and the rest of data (67%) as training data is often used in the related literature, we aimed to test the effect of different amounts of test and training data on the performance of algorithms for educational data. For this purpose, one-third of the ideal test data ratio (33%) used in the related literature was drawn from all data. This value was then added and subtracted from 33%, and test values 11%, 22%, 33%, 44%, and 55% were obtained. After the data were prepared for analysis, for each algorithm and datafiles (180 files), 100 replications were performed for a different rate of test data (11%, 22%, 33%, 44%, and 55%) in which training data were randomly selected in the 'Experiment' section of the WEKA Version 3.9.0 software. Therefore, the test data for every replication of each algorithm was selected randomly. A total of 10000 analyzes were carried out for the sample of 500 students (100 datafiles), 5000 for 1000 students (50 datafiles), and 3000 for 5000 students (30 datafiles), and the average of the accuracy rates and RMSE values were reported and interpreted together with the total elapsed times for each algorithm. Selecting different datafiles from whole data and making and averaging 100 replications is to reduce the possible biased and erroneous results that could stem from getting different results for different algorithms. In the analysis, IBM SPSS Statistics 23, Microsoft Office Excel 2016 and WEKA Version 3.9.0 software were used.

In this study, since individuals showed a balanced distribution to the categories of the dependent variable, the accuracy rate, root mean square error (RMSE) values, and total elapsed time of the models were used in the evaluation of the performances of algorithms. The accuracy rate gives the correct classification percentage of a classifier. RMSE is a standard measure of the difference between values estimated by predicted and actual values. It is also a standard measure of accuracy rate that takes into account errors and allows to compare models.

In data mining, hypothesis testing is used to compare different methods and select the method with the least errors. For this purpose, when the assumptions of parametric analyzes are satisfied, the most preferred method is to use the t or F test. In this study, since the RMSE values used in the statistical comparison of methods did not meet the assumptions of parametric methods, the Friedman test was used to compare these values. Binary comparisons of the methods were made with the Wilcoxon Signed Ranks test.

## RESULTS

The findings obtained for the data mining algorithms under different conditions with respect to different evaluation criteria are given in Table 2.

Table 2. Performances of Data Mining Techniques under Different Conditions.

| Sample size | Percent of test data | NB | | | LR | | | MLP | | | KNN | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Accuracy (%)* | *RMSE* | *Time (sec)* | *Accuracy (%)* | *RMSE* | *Time (sec)* | *Accuracy (%)* | *RMSE* | *Time (sec)* | *Accuracy (%)* | *RMSE* | *Time (sec)* |
| 500 | 11 | 75.93 | 0.42 | 0.00 | 75.30 | 0.42 | 0.04 | 72.80 | 0.48 | 0.72 | 71.81 | 0.48 | 0.03 |
| | 22 | 75.87 | 0.42 | 0.00 | 74.69 | 0.43 | 0.03 | 72.62 | 0.48 | 0.59 | 71.44 | 0.48 | 0.02 |
| | 33 | 75.78 | 0.42 | 0.00 | 73.94 | 0.44 | 0.03 | 72.27 | 0.48 | 0.50 | 71.05 | 0.48 | 0.02 |
| | 44 | 75.66 | 0.42 | 0.00 | 72.97 | 0.45 | 0.03 | 71.97 | 0.49 | 0.42 | 70.55 | 0.48 | 0.02 |
| | 55 | 75.41 | 0.43 | 0.00 | 71.62 | 0.47 | 0.02 | 71.56 | 0.49 | 0.35 | 70.06 | 0.48 | 0.01 |
| 1000 | 11 | 76.35 | 0.42 | 0.00 | 76.93 | 0.40 | 0.07 | 73.92 | 0.45 | 1.32 | 72.11 | 0.47 | 0.10 |
| | 22 | 76.27 | 0.42 | 0.00 | 76.70 | 0.40 | 0.06 | 73.65 | 0.46 | 1.16 | 71.97 | 0.47 | 0.09 |
| | 33 | 76.16 | 0.42 | 0.00 | 76.37 | 0.41 | 0.05 | 73.42 | 0.47 | 0.98 | 71.78 | 0.48 | 0.07 |
| | 44 | 76.04 | 0.42 | 0.00 | 75.88 | 0.41 | 0.04 | 73.07 | 0.47 | 0.82 | 71.57 | 0.48 | 0.06 |
| | 55 | 75.94 | 0.42 | 0.00 | 75.15 | 0.42 | 0.04 | 72.66 | 0.48 | 0.66 | 71.18 | 0.48 | 0.05 |
| 5000 | 11 | 76.60 | 0.42 | 0.00 | 78.30 | 0.39 | 0.36 | 76.36 | 0.41 | 6.54 | 74.52 | 0.46 | 2.05 |
| | 22 | 76.60 | 0.42 | 0.00 | 78.25 | 0.39 | 0.30 | 76.20 | 0.41 | 7.44 | 74.34 | 0.46 | 1.71 |
| | 33 | 76.58 | 0.42 | 0.00 | 78.19 | 0.39 | 0.25 | 76.04 | 0.41 | 4.93 | 74.14 | 0.47 | 1.50 |
| | 44 | 76.53 | 0.42 | 0.01 | 78.10 | 0.39 | 0.24 | 75.77 | 0.42 | 4.13 | 73.83 | 0.47 | 1.17 |
| | 55 | 76.48 | 0.42 | 0.01 | 77.96 | 0.39 | 0.25 | 75.53 | 0.42 | 0.35 | 73.50 | 0.47 | 1.03 |

Note: NB: Naïve Bayes, LR: Logistic Regression, MLP: Multilayer Perceptron, KNN: *k*-Nearest Neighborhood, RMSE: Root mean squared error.

According to Table 2, it has been observed that the methods chosen for classifying students according to their PISA mathematics achievement generally show above average or high performance under different conditions. The accuracy rates for all methods range from 70.06 to 78.30. While the NB method showed the highest performance in the sample of 500 students, the LR method showed the highest performance in the samples of 1000 and 5000 students. The MLP method performed less than the NB and LR method in all conditions but higher than the KNN method. The results for comparing the performances of the methods were examined separately according to different evaluation criteria. In Figure 1, the change of the accuracy rates of the methods for different conditions is given.

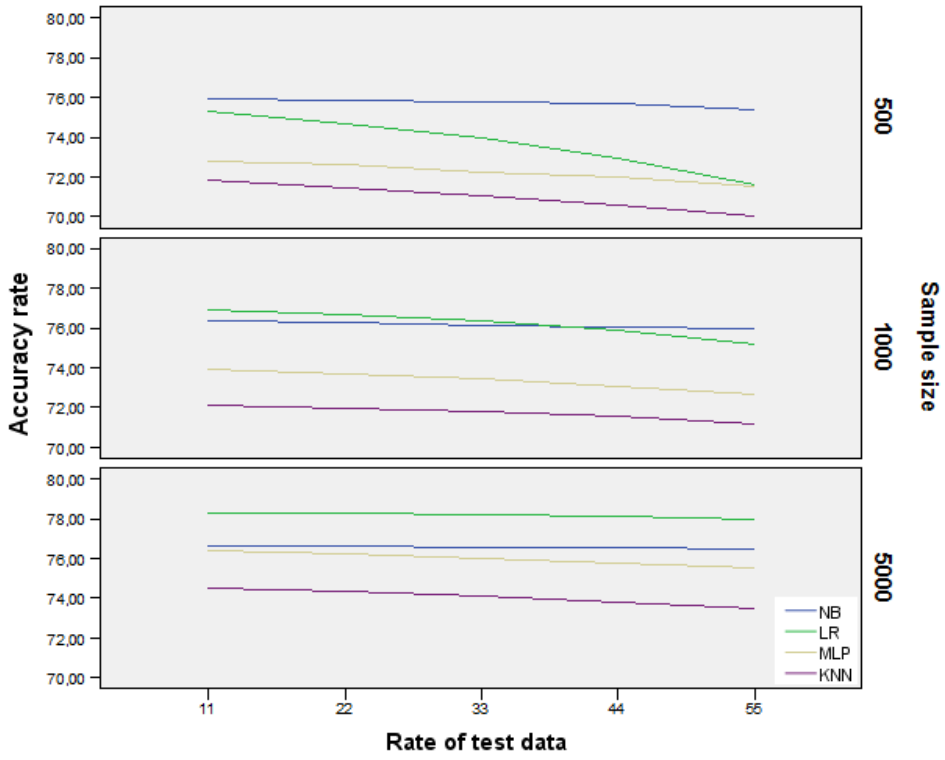


Figure 1. Change of Accuracy Rates of The Algorithms

When Figure 1 is examined, increasing the sample size leads to an increase in the classification performance of all methods, although much less in the NB method. In the samples of 500 and 1000 students, increasing the test data rate causes a significant decrease in the performance of the LR method. While the NB method is not affected by this, other methods are much less affected than the LR method. In the sample of 5000 students, the NB and LR methods are not affected by the increase in the rate of test data, while the MLP and KNN methods decrease slightly, as in other sample sizes. As a result, when the sample size is increased, the LR method is less affected by the change of the test data rate, while the NB method is not. MLP and KNN methods, on the other hand, show lower performance even if the sample size is increased, similarly being affected by increasing the test data rate. In Figure 2, the change of RMSE values of the methods for different conditions is given.

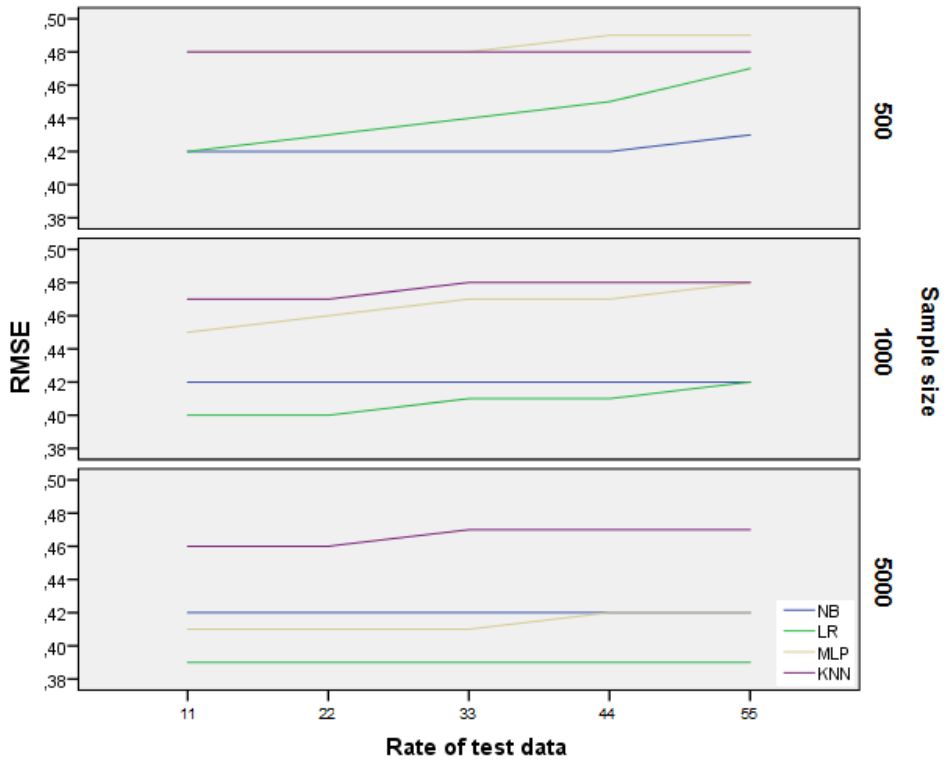


Figure 2. Change of RMSE Values of The Algorithms

For all methods, RMSE values range from approximately 0.39 and to 0.48. In small samples, the least erroneous estimations were made with NB and LR methods. According to Figure 2, the amount of error of the LR method increased significantly when the training data was reduced in small samples. In contrast, other methods were not significantly affected by this situation. In medium-sized samples, the error amount of the methods decreased compared to the small samples except the NB method. The least erroneous results in this sample size were obtained with the LR method. The decrease of the training data rate increased the error amount of the other methods except the NB method. In large samples, the estimation error amounts of the other methods have decreased except for the NB method. The NB method has approximately the same amount of error in all conditions. While increasing the test data rate does not affect the error amount of LR method in large samples, the error amount of MLP and KNN methods increased. The differentiation in RMSE values, which allow comparison of methods under different conditions across different methods, was analyzed by the Friedman test and binary comparisons of the methods were performed with the Wilcoxon test. The results are given in Table 3.

Table 3. Statistical comparison of data mining techniques under different conditions.

| Sample size | Percent of test data | Test statistics (Friedman) | | | Multiple comparisons (Wilcoxon)** |
|---|---|---|---|---|---|
| | | *Chi-Square* | *df* | *p* | |
| 500 | 11 | 257.251* | 3 | 0.000 | 1<3, 1<4, 2<3, 2<4 |
| | 22 | 265.013* | 3 | 0.000 | 1<2, 1<3, 1<4, 2<3, 2<4 |
| | 33 | 275.340* | 3 | 0.000 | 1<2, 1<3, 1<4, 2<3, 2<4, 3<4 |
| | 44 | 284.642* | 3 | 0.000 | 1<2, 1<3, 1<4, 2<3, 2<4, 4<3 |
| | 55 | 271.014* | 3 | 0.000 | 1<2, 1<3, 1<4, 2<3, 2<4, 4<3 |
| 1000 | 11 | 149.705* | 3 | 0.000 | 2<1, 1<3, 1<4, 2<3, 2<4, 3<4 |
| | 22 | 149.149* | 3 | 0.000 | 2<1, 1<3, 1<4, 2<3, 2<4, 3<4 |
| | 33 | 146.351* | 3 | 0.000 | 2<1, 1<3, 1<4, 2<3, 2<4, 3<4 |
| | 44 | 144.013* | 3 | 0.000 | 2<1, 1<3, 1<4, 2<3, 2<4, 3<4 |
| | 55 | 137.068* | 3 | 0.000 | 1<3, 1<4, 2<3, 2<4 |
| 5000 | 11 | 88.729* | 3 | 0.000 | 2<1, 3<1, 1<4, 2<3, 2<4, 3<4 |
| | 22 | 88.052* | 3 | 0.000 | 2<1, 3<1, 1<4, 2<3, 2<4, 3<4 |
| | 33 | 87.632* | 3 | 0.000 | 2<1, 3<1, 1<4, 2<3, 2<4, 3<4 |
| | 44 | 89.022* | 3 | 0.000 | 2<1, 1<4, 2<3, 2<4, 3<4 |
| | 55 | 87.769* | 3 | 0.000 | 2<1, 1<3, 1<4, 2<3, 2<4, 3<4 |

Note: 1: Naïve Bayes, 2: Logistic Regression, 3: Multilayer Perceptron, 4: *k*-Nearest Neighborhood, RMSE: Root mean squared error, df: Degree of freedom

*p<0.001

**p<0.0166 (Calculated based on Bonferroni correction)

According to Table 3, when the sample size increases, the LR method performs analysis with significantly less error than all methods. The NB method, on the other hand, provides significantly less erroneous estimations when the sample size decreases. The KNN method has more errors in medium and large samples compared to other methods at statistically significant level. When the test data rate is increased in small samples, the error amount of MLP method is significantly higher than other methods. In Figure 3, the change of the total elapsed time of the methods for the analysis under different conditions is given.

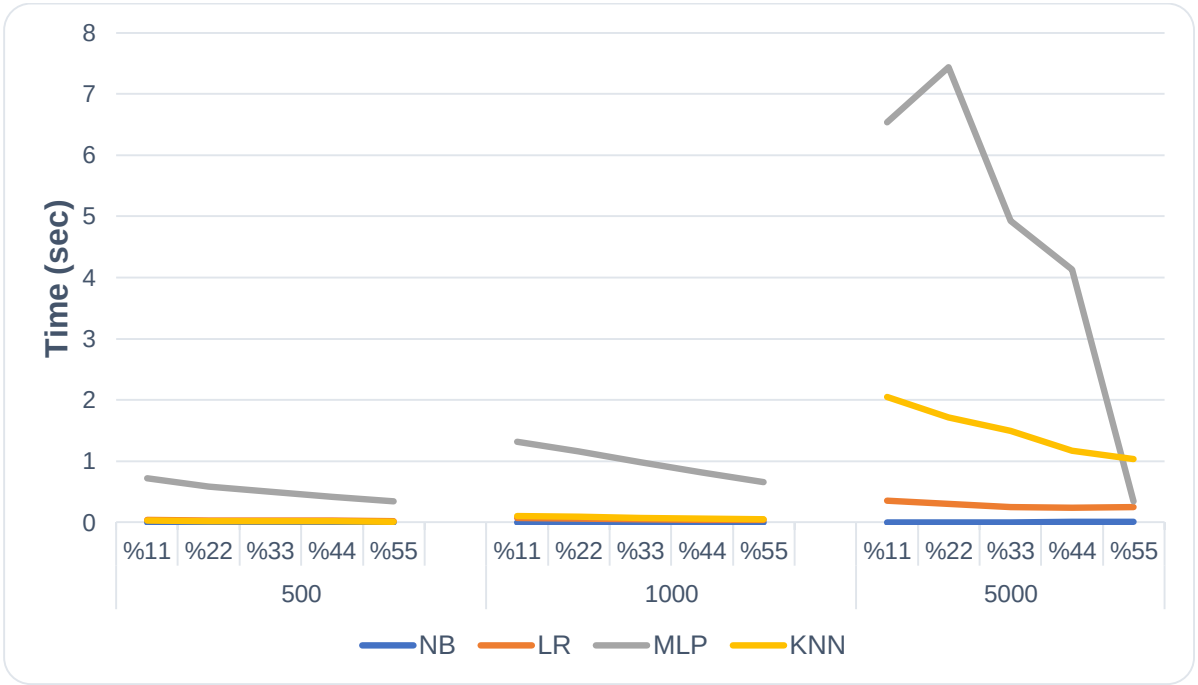


Figure 3. Total Elapsed Time for Each Analysis Under Different Conditions.

According to Figure 3, the NB method performs analyzes without taking any time in almost all conditions. In samples of 500 and 1000 students, LR and KNN methods operate in a much shorter time than MLP method. In small and medium sample sizes, LR and KNN methods carried out analysis in a much shorter time than MLP method. In large samples, KNN method takes more time than other methods for test data rate of 55%. The MLP method takes a lot of time when the test data rate is low, as the training data is high. Due to the k-fold cross-validation method used in the selection of the k value, in larger samples, the KNN method performed analyzes in much longer time than the LR method. However, since the total elapsed times are obtained under standard conditions on a computer with certain features, analysis can be completed in a shorter time on computers with more advanced features.

## DISCUSSION and CONCLUSION

In this study, the performance of different data mining methods for different sample sizes and test data rates were compared on educational data in terms of accuracy rate, RMSE value, and total elapsed time for the analysis. It has been observed that the accuracy rates of the methods vary slightly for different conditions. This situation stems from the data selection and analysis procedure used in the present study. We selected 180 datasets from a huge data dataset of 62728 students by random selection with replacement and replicated each analysis 100 times. Therefore, the average of 10000 analyses for small samples, 5000 analyses for medium samples, and 3000 analyses for large sample sizes were evaluated. The results obtained seem to be close to each other due to these numerous amounts of the analyses. However, statistical hypothesis tests have shown that these seemingly small differences differ significantly.

In small sample sizes, high accuracy rates were obtained, less erroneous estimates were made, and the analyzes were completed in a very short time with the NB method compared to other methods. In addition, the NB method gives acceptable results even with a small amount of training data. In some studies, NB method has been shown to give better results than other methods in small samples (Göker, 2012; Hamalainen & Vinni, 2006; Hamalainen & Vinni, 2011, Kotsiantis et al., 2003; Osmanbegović & Suljić, 2012). However, Nghe et al. (2007) showed that decision trees produce better results than Bayes networks. Data structure might be a preliminary reason for this situation. Hence, it is very important to know which method is the best for a certain data type.

In the study, LR method showed higher performance in all conditions than MLP method. Although this result is different from some research results (Bahadır, 2013; Çırak, 2012; Tepehan, 2011), the most important reason for this situation is that the data structure is suitable for LR analysis. LR method produces less erroneous and higher accuracy estimates than other methods in medium and large samples. In the study conducted by Dekker et al. (2009), the LR method performed better in samples with similar size than the Bayes method.

After NB and LR methods, the highest accuracy rates were obtained by MLP and KNN methods, respectively. In the study of Romero et al. (2013), KNN method performed lower for numerical and categorical data compared to other classifiers. Similarly, in this study, the MLP method gave less erroneous results than the KNN method in medium and large samples. However, the opposite is true in small samples. This was due to the fact that the KNN method has a simpler statistical structure than the MLP method and that the selected k value was more stable in small samples in determining the closest neighborhood. In the MLP method, selecting the number of layers as three was effective in training the network, but in small samples, it yielded a high amount of error.

In this study, KNN method showed lower correct classification performance in all conditions than other methods. However, some studies have shown that the KNN method performs as well as ANN and LR methods (Minaei-Bidgoli et al., 2003; Yurdakul & Topal, 2015). Similarly, Shahiri et al. (2015) compared the studies published in international databases between 2002 and 2015 and found that NB method showed lower performance than KNN and ANN methods in terms of average performance. However, in this study, NB method showed higher classification performance, especially in small and medium-size samples. Some researchers have stated that it is not true to say that a classification method

is best for different conditions and data structures (Romero et al., 2013; Shahiri et al., 2015). Barker et al. (2004), for example, made the classification of students who graduated in different years according to their graduation status and showed that different methods could be effective according to the structure of the data in different years.Barker et al. (2004), on the other hand, made the classification of students who graduated in different years according to their graduation status and showed that different methods could be effective according to the structure of the data in different years. For this reason, the results obtained from the present study have been interpreted within the framework of the structure of the data used and the analysis conditions. Since it is possible to obtain a different result with different data types (Romero et al., 2013), it is important to determine the structure of the data and choose the most appropriate method before the analyses.

Although the rate of test data is generally taken as one-third of all data in the related literature, it has been found that using a general valid rate may not be a proper approach. The results showed that the test data rate is closely related to the number of variables used, sample size, structure of the data and the structure of the method. However, except for the NB method, in general, increasing the rate of test data decreased the performance of the methods and increased the error of the results obtained. Therefore, as increasing the sample size increases classification performance and reduces the amount of error, it will be appropriate to use as much larger sample sizes as possible to achieve high performance from all methods. In many studies, it was found that different train/test ratios (e.g., Brain & Webb, 1999; Çölkesen, & Kavzoglu, 2010; Tadjudin & Landgrebe, 1998; Foody et al., 2006; Heilman, & Madnani, 2015; Shao et al., 2013; Tayeh et al., 2015) have different effects on the performance of the methods. For example, Brain and Webb (1999) showed that error variance decreases when the amount of test data is increased, but there is no significant change in the amount of bias. Similarly, Tadjudin and Landgrebe (1998) stated that the lack of test data caused errors in classification performance. However, Foody et al. (2006) stated in their study that even a 90% reduction in the rate of test data did not cause a decrease in the performance of some algorithms. Heilman and Madnani (2015) found that increasing test data increased performance, but increasing sample size did not have the same effect. Çölkesen and Kayzoğlu (2010) found in their study that some methods show higher performance in small training sets than others.

### *Limitations and Suggestions*

In this study, although the analyses were performed with data for which the conditional independence assumption of the Naive Bayes method was not satisfied, acceptable results were obtained. This result has shown that, as stated by Hamalainen and Vinni (2011), Naive Bayes can perform well even if the conditional independence assumption is not met. In future studies, the acceptability of the results obtained under satisfying this assumption can be examined and compared with the performance of other methods. In the present study, it was seen that the *k* value to be selected for the *k*-nearest neighborhood method affects the classification performance. Accordingly, in other studies, different methods can be used to select the *k* value, or new methods can be developed. In the artificial neural networks method, since many parameters such as the number of layers, the number of nodes in layers, weightings affect the classification performance of the models, the effects of changes in these parameters on the performance of the method can be examined. The results obtained for logistic regression analysis and artificial neural network methods were obtained under the condition that homogeneity of variance-covariance matrices is not satisfied. Although these methods give effective results even when this assumption is violated, the classification performances of the methods can be evaluated and compared under the conditions in which the variance-covariance matrices are homogeneous. The results of the present study are also limited to the PISA 2012 data. For different data types, the performance of the algorithms can be compared in future studies. Besides, a simulation study under similar conditions could be done and compared with the results obtained with student data.

Similar to the results of the present study, it was found that different data types may yield different results (Romero et al., 2013). Therefore, identifying the structure of data and choosing the best analysis might be a solution to this issue. In addition, as a better solution to this problem, the procedure followed by (Göker, 2012; Yurdakul & Topal, 2015) can be used. As a two-step method, this procedure consists

of selecting the method with the lowest error and then reporting the results obtained or performing further analysis with this method.

Using the Naive Bayes method in applications to be carried out under similar conditions will provide better results in a shorter time. Other methods may be preferred to the *k*-nearest neighborhood method to obtain higher classification performance under similar conditions. When the sample size is large, preferring Naive Bayes and logistic regression methods to multilayer perceptron will provide higher classification performance and time-saving.

## REFERENCES


Aha, D. W., Kibler, D. & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning 6*, 37-66.

Aksu, G., & Guzeller, C. O. (2016). Classification of PISA 2012 mathematical literacy scores using decision-tree method: Turkey sampling. *Education and Science, 41*(185), 101-122.

Akpınar, H. (2014). *Veri madenciliği veri analizi.* Papatya Yayınları, İstanbul.

Baker, R. S. J. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112-118.

Baker, R.S.J. & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17.

Bahadır, E. (2013). *Yapay sinir ağları ve lojistik regresyon analizi yaklaşımları ile öğretmen adaylarının akademik başarılarının tahmini* (Doktora tezi, Marmara Üniversitesi, İstanbul). Retrieved from http://tez2.yok.gov.tr/

Barker, K., Trafalis, T. & Rhoads, T. R. (2004). Learning from student data. In *Proceedings of the 2004 Systems and Information Engineering Design Symposium* (pp. 79-86). IEEE.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta, 760*, 25-33.

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk - predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Data Mining, 11*(3), 1-41. https://doi.org/10.5281/zenodo.3594771

Bhardwaj, B. K. & Pal, S. (2011). Data mining: A prediction for performance improvement using classification. *(IJCSIS) International Journal of Computer Science and Information Security, 9*(4), 136-140.

Blum, A. L. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1), 245–271.

Brain, D., & Webb, G. (1999). On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales* (pp. 117-128), December 5-6, Sydney, Australia.

Bulut, O., & Yavuz, H. C. (2019). Educational data mining: A tutorial for the" Rattle" package in R. *International Journal of Assessment Tools in Education*, *6*(5), 20-36.

Büyüköztürk, Ş., Çakmak-Kılıç, E., Akgün, Ö., Karadeniz, Ş. & Demirel, F. (2015). *Bilimsel araştırma yöntemleri.* Ankara: Pegem.

Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. *Higher Education: Handbook of Theory and Research, 10,* 225–256.

Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., Lin, C., & Alzheimer's Disease Neuroimaging Initiative. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, *60*(1), 59-70.

Cox, D. R. & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). London: Chapman and Hall.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37-46.

Çırak, G. (2012). *Yükseköğretimde öğrenci başarılarının sınıflandırılmasında yapay sinir ağları ve lojistik regresyon yöntemlerinin kullanılması* (Yüksek lisans tezi, Ankara Üniversitesi, Ankara). Retrieved from http://tez2.yok.gov.tr/

Çölkesen, I., & Kavzoglu, T. (2010). Farklı boyutta eğitim örnekleri için destek vektör makinelerinin sınıflandırma performansının analizi. In *Proceedings of III. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu* (pp. 161-170), 11 – 13 Ekim, Gebze, Kocaeli, Türkiye.

Dekker, G. W., Pechenizkiy, M. ve Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In *Proceedings of 2nd International Conference on Educational Data Mining* (pp. 41-50). Spain, Cordoba.

Dunham, M.H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education, Inc.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvements on crossvalidation. *J. Amer. Stat. Ass.*, *78*(382), 316–331.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8.

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, *104*(1), 1-14.

Fraenkel, J. R. & Wallen, N. E. (2011). *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill, Inc.

Frank, E., Hall M. A. & Witten, I. H. (2016). *The WEKA workbench: Online appendix for "Data mining: Practical machine learning tools and techniques"* (4th ed.). Morgan Kaufmann.

Ghosh, A. K. (2006). On optimum choice of k in nearest neighbor classification. *Computational Statistics and Data Analysis*, *50*(11), 3113-3123.

Gorostiaga, A., & Rojo-Álvarez, J. L. (2016). On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain. *Neurocomputing*, *171*, 625-637.

Göker, H. (2012). *Üniversite giriş sınavında öğrencilerin başarılarının veri madenciliği yöntemleri ile tahmin edilmesi* (Yüksek lisans tezi, Gazi Üniversitesi, Ankara). Retrieved from http://tez2.yok.gov.tr/

Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020). Analysis of factors effecting PISA 2015 mathematics literacy via educational data mining. *Education and Science*, *45*(202), 393-415.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18.

Hall, P., Park, B. U. & Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, *36*(5), 2135-2152.

Han, J., Kamber, M. & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). MA, USA: Elsevier.

Hamalainen, W. & Vinni, M. (2006). Comparison of machine learning methods for intelligent tutoring systems. In *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 525-534). Springer Berlin/Heidelberg.

Hamalainen, W. & Vinni, M. (2011). *Classifiers for educational technology*. In C. Romero, S. Ventura, M. Pechenizkiy, R.S.J.d. Baker (Eds.), Handbook of educational data mining (pp. 54-74). CRC Press.

Heilman, M., & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 81-85), June 4, Association for Computational Linguistics, Denver, Colorado.

Heydari, S. S., & Mountrakis, G. (2018). Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sensing of Environment*, *204*, 648-658.

Huebner, R. A. (2013). A survey of educational data-mining research. *Research in Higher Education Journal*, *19*, 1-13.

Karasar, N. (2005). *Bilimsel araştırma yöntemi*. Ankara: Nobel Yayın Dağıtım.

Kiray, S. A., Gok, B., & Bozkir, A. S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *Journal of Education in Science, Environment and Health*, *1*(1), 28-48.

Kotsiantis, S. B., Pierrakeas, C. J. & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 267-274). Springer Berlin/Heidelberg.

Lachenbruch, P. A. & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, *10*(1), 1-11.

Larose, D. T. (2004). *K-nearest neighbor algorithm*. In Larose, D.T. and Larose, C.D. (Eds.), *Discovering knowledge in data: An introduction to data mining* (pp. 90-106). Hoboken, NJ, USA John Wiley and Sons, Inc.. https://doi.org/10.1002/0471687545.ch5.

Liu, H. & Motoda, H. (2001). *Feature extraction, construction and selection: A data mining perspective*. Boston: Kluwer Academic Publishers.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 491-502.

Martínez-Abad, F., Gamazo, A., & Rodríguez-Conde, M. J. (2020). Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment. *Studies in Educational Evaluation*, *66*, 100875. https://doi.org/10.1016/j.stueduc.2020.100875

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (1994*). Machine learning, neural and statistical classification*. Ellis Horwood Limited.

Minaei-Bidgoli, B., D.A. Kashy, G. Kortemeyer, & W. Punch. Predicting student performance: An application of data mining methods with an educational web-based system. In *Proceedings of 33rd Frontiers in Education Conference*, (pp. 13-18). Westminster, CO..

Nghe, N. T., Janecek, P. & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, (pp. T2G-7). IEEE.

Organisation for Economic Co-operation and Development (2014a). *PISA 2012 results: What students know and can do - student performance in mathematics, reading and science* (Volume I, Revised edition). PISA, OECD Publishing.

Organisation for Economic Co-operation and Development (2014b). *PISA 2012 technical report*. PISA, OECD Publishing.

Osmanbegović, E. & Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review, 10*(1), 3-12.

Peng, C.Y.J., Lee, K. L. & Ingersoll, G. M. (2002) An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3-14. doi:10.1080/00220670209598786.

Peng, C. Y. J. & So, T. S. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences, 1*(1), 31-70.

Ranjan, J. & Malik, K. (2007). Effective educational process: A data mining approach. *VINE, 37*(4), 502-515.

Raudys, S., & Pikelis, V. (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (3), 242-252.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R. & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education, 21*(1), 135-146.

Romero, C., Ventura, S., Espejo, P. G. & Hervás, C. (2008). Data mining algorithms to classify students. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 8-17). Montréal, Québec, Canada.

Romero, C. & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146.

Romero, C. & Ventura, S. (2013). Data mining in education. *WIREs Data Mining Knowledge Discovery 3*(1), 12-27.

Shahiri, A. M., Husain, W. & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*, 414-422.

Shao, L., Fan, X., Cheng, N., Wu, L., & Cheng, Y. (2013). Determination of minimum training sample size for microarray-based cancer outcome prediction–an empirical assessment. *PloS one, 8*(7), e68579. https://doi.org/10.1371/journal.pone.0068579

Sivanandam, S., Sumathi, S., & Deepa, S. (2006). *Introduction to neural networks using Matlab 6.0*. New Delhi: Tata McGraw-Hill Publishing Company.

Şengür, D. (2013). *Öğrencilerin akademik başarılarının veri madenciliği metotları ile tahmini* (Yüksek lisans tezi, Fırat Üniversitesi, Elazığ). Erişim adresi: http://tez2.yok.gov.tr/

Sweeney, M., Lester, J., Rangwala, H., & Johri, A. (2016). Next-term student performance prediction: A recommender systems approach. *JEDM | Journal of Educational Data Mining, 8*(1), 22-51. https://doi.org/10.5281/zenodo.3554603

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Tadjudin, S., & Landgrebe, D. (1998). *Classification of high dimensional data with limited training samples* (Report No. 56). West Lafayette, Indiana: Purdue University, School of Electrical and Computer Engineering. http://docs.lib.purdue.edu/ecetr/56

Tayeh, N., Klein, A., Le Paslier, M. C., Jacquin, F., Houtin, H., Rond, C., ... & Burstin, J. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Frontiers in Plant Science, 6*(941), 941. https://doi.org/10.3389/fpls.2015.00941

Tepehan, T. (2011). *Türk öğrencilerinin PISA başarılarının yordanmasında yapay sinir ağı ve lojistik regresyon modeli performanslarının karşılaştırılması* (Doktora tezi, Hacettepe Üniversitesi, Ankara). Retrieved from http://tez2.yok.gov.tr/

Tezbaşaran, E. (2016). *Temel bileşenler analizi ve yapay sinir ağı modellerinin ölçek geliştirme sürecinde kullanılabilirliğinin incelenmesi* (Doktora tezi, Mersin Üniversitesi, Mersin). Retrieved from http://tez2.yok.gov.tr/

Tosun, S. (2007). *Sınıflandırmada yapay sinir ağları ve karar ağaçları karşılaştırması: Öğrenci başarıları üzerine bir uygulama* (Yüksek lisans tezi, İstanbul Teknik Üniversitesi, İstanbul). Retrieved from http://tez2.yok.gov.tr/

Wharton, S. W. (1984). An analysis of the effects of sample size on classification performance of a histogram based cluster analysis procedure. *Pattern Recognition, 17*(2), 239-244.

Yurdakul, S. & Topal, T. (2015). Veri madenciliği ile lise öğrenci performanslarının değerlendirilmesi. *XVII. Akademik Bilişim Konferansında* sunulan bildiri. Anadolu Üniversitesi, Eskişehir.

# Veri Madenciliği Sınıflandırma Algoritmalarının Farklı Koşullar için Eğitsel Bir Veride Karşılaştırılması

## *Giriş*

Öğrenci başarısının yordanması eğitimde yapılan birçok araştırmanın odak noktasını oluşturur. Özellikle, teknolojinin hızla geliştiği ve eğitimde daha fazla önem kazandığı günümüzde öğrenci başarısını etkileyen birçok faktörü içinde barındıran veri tabanları bulunmaktadır. Blackboard ve Moodle gibi zengin eğitimsel veri kaynaklarını içeren ders yönetim sistemlerinin yanında, uluslararası düzeyde yapılan TIMMS (Uluslararası Matematik ve Fen Eğilimleri Araştırması), PISA (Uluslararası Öğrenci Değerlendirme Programı) ve PIRLS (Uluslararası Okuma Becerilerinde Gelişim Projesi) gibi çalışmalarda öğrenci, öğretmen, okul, bölge ve ülke düzeyinde bilgiler toplanmaktadır. Elde edilen eğitimsel içerikli veri yığınlarını analiz etmek ve öğrencileri karşılaştırarak başarılarını yordamak son yıllarda gittikçe önem kazanmaktadır. Bu amaçla, eğitsel veri madenciliği (EVM) son yıllarda bağımsız bir araştırma alanı olarak ortaya çıkmıştır (Baker, 2010).

EVM, veri madenciliği tekniklerini eğitim içerikli verilere uygulamak amacıyla ortaya çıkan yeni bir disiplindir (Baker ve Yacef, 2009; Huebner, 2013). Öğretim programlarının etkililiğinden öğrenci başarısının yordanmasına, eğitim kurumlarından öğretmenlerin performansına kadar eğitimin her alanında kullanılabilmektedir. İlgili alan yazında EVM ile ilgili farklı tanımlamalar mevcuttur. Baker ve Yacef (2009), EVM'yi, eğitim ortamlarından elde edilen kendine özgü verilerden keşifler yapmak amacıyla yeni metotların geliştirilmesini merkez alan, öğrencileri ve öğrenme ortamlarını daha iyi anlamak için bu metotları kullanan bilimsel araştırma alanı olarak tanımlamaktadır. Ancak, Huebner (2013) bu şekilde tanımlamaların sınırlı olduğunu, EVM'nin çok geniş bir alanı kapsadığını ve ileride yapılacak çalışmalarla birlikte bu alanın kapsamının ve tanımlarının değişeceğinin belirtmiştir.

Veri madenciliğinde bireylerin ya da gözlemlerin belirli bir kategorik değişkene göre sınıflandırılması en temel yordama tekniklerinden biridir (Baker, 2010). Bazı popüler yordama algoritmaları, karar ağaçları, lojistik regresyon, destek vektör makineleri, sinir ağları, Bayes algoritmaları, k-en yakın komşuluk ve çeşitli kernel fonsiyonlarına dayanan yoğunluk kestiricileridir. Bir kestiricinin doğruluğunu değerlendirmek amacıyla hata matrisine dayanan dönüştürülmüş performans değerlendirme ölçütleri (kesinlik, çağrışım, F ölçütü, vb.), Root mean square error (RMSE), Kappa (Cohen, 1960), ROC eğrisinin altında kalan alan (Egan, 1975) ve yordama hata oranları gibi ölçütler kullanılmaktadır.

Veri madenciliğinde algoritmaların performansını artırmak amacıyla veri öğrenme ve test verisi olmak üzere iki parçaya ayrılır. Bu metotta, bir veri setinin belirli bir bölümü kullanılarak ilk analizler gerçekleştirilir ve bir yordama modeli oluşturulur. Sonraki aşamada, elde edilen bu modelden yararlanılarak verinin kalan kısmındaki bireyler ya da nesneler için yordama işlemi gerçekleştirilir. Yöntemin etkililiğinin test edildiği verinin bu parçasına test verisi denir. Bu veri, tüm verinin belirli bir oranından edildiğinden dolayı test verisi oranı olarak ifade edilir. Veri madenciliğinde yöntemlerin etkililiğinin bu şekilde test edilmesinin nedeni model hata oranlarının yanlı kestirimlerinin önüne geçmektir. Benzer amaçlar için kullanılan diğer yöntemler, önyükleme (Efron, 1983) ve çapraz geçerleme (Lachenbruch ve Mickey, 1968) teknikleridir (Michie, Spiegelhalter ve Taylor, 1994). Ancak, tüm veriden belirli oranda (genellikle 1/3 oranında - %33) test verisi seçilerek bu veri ile yordama işleminin gerçekleştirilmesi sıklıkla tercih edilen ve büyük örneklemler için de çoğunlukla kullanılan etkili bir yöntemdir.

Veri madenciliği yöntemlerinin eğitim verileri üzerinde uygulanabilirliği ve etkililiği üzerine son on yıllık süreçte birçok araştırmaya rastlamak mümkündür (Barker, Trafalis ve Rhoads, 2004; Dekker, Pechenizkiy ve Vleeshouwers, 2009; Kotsiantis, Pierrakeas ve Pintelas, 2003; Hamalainen ve Vinni, 2006; Hamalainen ve Vinni, 2011; Minaei-Bidgoli, Kashy, Kortemeyer ve Punch, 2003; Nghe, Janecek ve Haddawy, 2007; Osmanbegović ve Suljić, 2012; Romero, Espejo, Zafra, Romero ve Ventura, 2013; Romero, Ventura, Espejo ve Hervas, 2008; Shahiri, Husain ve Rashid, 2015). Bu araştırmalar, genel olarak öğrenci performansının yordanması, değerlendirilmesi ve performansı etkileyen faktörlerin belirlenmesi amacı taşımaktadır. Romero ve Venturo (2007) 1995 ve 2005 yılları arasında eğitim alanında yapılan veri madenciliği çalışmalarını derleyerek çeşitli özelliklerine göre sınıflandırmışlardır. Ancak, bu araştırmalardan çok az bir kısmı veri madenciliği algoritmalarının karşılaştırılmasının yanında örneklem büyüklüğü ve eğitim setinin büyüklüğü bu algoritmaların performansına etkisine değinmiştir. Hâlbuki istatistik, mühendislik, sağlık ve sosyal bilimler gibi birçok alanda farklı veri yapısının veri madenciliği algoritmaları üzerindeki etkileri önemli bir araştırma konusu haline gelmiştir. Ayrıca, Türkiye'de EVM ile ilgili uygulamalara ve yukarıda anlatılan yöntemlerden Naive Bayes ve k-en yakın komşuluk tekniklerine yönelik çalışmalar sınırlı düzeydedir. Bu çalışmada, alan yazında görülen bu eksikliklere yönelik PISA (2012) uygulamasında alınan bir veri kullanılarak kapsamlı bir uygulama yapılması hedeflenmiştir.

Bu çalışmanın amacı, öğrencilerin, çeşitli özellikleri bakımından PISA (2012) matematik başarılarını yordamada Naive Bayes, k-en yakın komşuluk, lojistik regresyon ve yapay sinir ağları çok katmanlı algılayıcı yöntemlerinin performanslarının farklı örneklem büyüklükleri (küçük, orta, büyük) ve test verisi oranlarına (%11, %22, %33, %44 ve %55) göre nasıl değiştiğini gözlemlemektir.

## *Yöntem*

Çalışmanın yöntem kısmı burada özetlenmelidir. Naive Bayes, *k*-en yakın komşuluk, yapay sinir ağları ve lojistik regresyon algoritmalarının farklı koşullar altında performanslarının belirlenmesi ve açıklanması hedeflendiğinden, bu çalışma temel bir araştırmadır. Bu tür araştırmalarda metodolojik analize dayalı çalışmalar yaparak bilgi üretilmesi amaçlanmaktadır (Büyüköztürk, Çakmak-Kılıç, Akgün, Karadeniz ve Demirel, 2015). Temel araştırmalar mevcut bilgiye yeni bilgiler eklemeyi amaçlamaktadır (Karasar, 2005). Araştırma aynı zamanda yöntemler arasındaki ilişkileri incelemek açısından nicel ilişkisel araştırmadır. Bu tür çalışmalar nicel değişkenler arasındaki ilişkileri araştırmayı, açıklamayı ve keşfetmeyi amaçlamaktadır (Fraenkel ve Wallen, 2006).

Araştırmanın evreni, PISA uygulamasına katılan OECD ülkelerindeki 15 yaş grubundaki öğrencilerdir. Her bir ülke için evreni temsil eden örneklemler PISA uygulayıcıları tarafından tabakalı tesadüfi örnekleme yoluyla seçilmiştir. OECD ülkelerinden PISA uygulamasına katılan toplam kişi sayısı 295416 kişidir. Bu çalışmada, kayıp veriler, artık ve uç değerler incelenip çıkartıldıktan sonra 62728 kişilik hedef evrene ulaşılmıştır. Araştırmada, incelenen yöntemlerin performanslarına yönelik yüksek kesinlikte sonuçlar elde etmek amacıyla 500 kişilik örneklem (küçük) için 100 veri dosyası, 1000 kişilik örneklem (orta) için 50 veri dosyası, 5000 kişilik örneklem (büyük) için 30 veri dosyası olmak üzere toplam 180 veri dosyası oluşturulmuştur.

Araştırmanın veri toplama araçları, PISA (2012) uygulamasında öğrencilerin matematik alanındaki akademik performanslarını ölçmek amacıyla geliştirilen matematik bilişsel testi ve öğrenciyi var olan tüm özellikleri ile değerlendirmeyi amacıyla hazırlanan öğrenci anketidir. Öğrenci anketinde ise öğrencilerin evleri, okulları, kendileri ve öğrenme deneyimleri hakkında çeşitli bilgileri içeren formları doldurmaları beklenmiştir (OECD, 2014a). Bu çalışmada, PISA uygulayıcıları tarafından takip edilen süreçler (OECD, 2014a) sonucunda elde edilen veri OECD'nin herkese açık veri tabanından alınarak kullanılmıştır.

Verilerin analizinde, öncellikle, PISA (2012) öğrenci anketinden elde edilen veriden öğrencilerin demografik bilgileri ve matematiğe ilişkin tüm değişkenleri alınmıştır. Daha sonra OECD (2014b) tarafından yayınlanan PISA 2012 teknik raporu göz önünde bulundurularak diğer değişkenlerin bileşiminden oluşan değişkenler alınmış ve kalan değişkenler dosyadan çıkartılmıştır. Daha sonra ise

kayıp verilerden oluşan değişkenler, kalan değişkenlere ait tüm kayıp veriler ve uç değerler silindiğinde matematik performansı ile birlikte 35 değişken ve 62728 kişiden oluşan hedef evren elde edilmiştir.

Veri madenciliği yöntemleri çok fazla değişkenle çalışmakla birlikte, sınıflandırmaya katkısı olmayan değişkenlerin analizde bulundurulması yapılacak analizlerin çok zaman almasına ve sınıflandırma performansının düşmesine neden olmaktadır. Bu amaçla, bir veri ön işleme süreci olan değişken seçme veri madenciliğinde sıkça kullanılan önemli tekniklerden biridir (Blum ve Langley, 1997; Liu ve Motoda, 2001). Bu çalışmada, WEKA Version 3.9.0 yazılımında (Hall ve diğerleri, 2009) yer alan Information Gain Ranking Filter, Chi-squared Ranking Filter, Gain Ratio Feature Evaluator ve Symmetrical Uncertainty Ranking Filter metotları kullanılmıştır. Araştırmanın bağımlı değişkeni ve birinci makul değer olan PV1MATH (Plausible Value 1) değişkeni, PISA uygulayıcıları tarafından belirlenen ve öğrencilerin matematikte yeterliğini temsil eden altı düzeyden (OECD, 2014a) orta düzeyin (482) altında ve üstünde yer alan öğrenciler şeklinde iki kategorili bir değişkene dönüştürülmüştür. Daha sonra öğrencilerin matematik performanslarına göre sınıflandırmaya en çok katkı sağlayan ilk 10 değişken çalışmaya dâhil edilmiştir. Bu değişkenler, matematik öz-yeterliği, matematiksel benlik algısı, matematik kaygısı, ekonomik, sosyal ve kültürel statü indeksi, problem çözmeye açık olma, ülke, babanın eğitim düzeyi (ISCED), anne eğitim düzeyi (ISCED), öğretmen davranışı: öğrenciyi yönlendirme ve hesap makinesi kullanımıdır. Bu çalışmada, tüm analizler bu değişkenler kullanılarak gerçekleştirilmiştir.

Araştırmada kullanılacak değişkenlere karar verildikten sonra analizlerin varsayımları kontrol edilmiştir. Bu çalışmada yapılacak analizlere yönelik olarak yapılan varsayım kontrollerinde örneklem büyüklüğünün yeterli olduğu, kayıp ve uç değer olmadığı, normalliğin sağlandığı, varyans-kovaryansların homojen olmadığı, doğrusallığın kısmen sağlandığı, çoklu bağlantı ve tekilliğin olmadığı görülmüştür. Ayrıca, Naive Bayes yöntemi için koşullu bağımsızlık varsayımı sağlanamamıştır. Analizlerde, IBM SPSS Statistics 23, Microsoft Office Excel 2016 ve WEKA Version 3.9.0 yazılımlarından yararlanılmıştır. Model değerlendirilmesinde doğruluk oranı, RMSE değerleri ve modellerin işlem hızları kullanılmıştır. Yöntem karşılaştırma ölçütü olarak kullanılan RMSE değerleri, parametrik yöntemlerin varsayımlarını karşılamadığından, bu değerlerin karşılaştırılmasında Friedman testi kullanılmıştır. Yöntemlerin ikili karşılaştırmaları ise Wilcoxon İşaretli Sıralar testi ile yapılmıştır.

### *Sonuç ve Tartışma*

Bu araştırmada, farklı örneklem büyüklüklerinin ve test verisi oranlarının yöntemlerin performansları üzerinde yarattığı etkiler şu şekildedir:

1. Örneklem büyüdükçe, tüm yöntemlerin doğru sınıflandırma performansları artmış geçerliği ve güvenirliği yüksek sonuçlar elde edilmiştir.

2. Örneklem büyüdükçe, Naive Bayes yönteminin analiz süresi değişmemekle birlikte diğer yöntemlerin analiz işlem süreleri uzamıştır.

3. Test verisi oranı örneklem büyüklüğüne göre yöntemlerin sınıflandırma performanslarında farklı etkiler yaratmıştır.

4. Örneklem büyüdükçe test verisi oranının arttırılmasının yöntemlerin performansları üzerindeki etkisi azalmıştır.

5. Test verisi oranı tüm verinin üçte birinden az olduğunda da yüksek doğru sınıflandırma performansları elde edilmiştir.

6. Örneklem büyüdükçe test verisi oranı tüm verinin üçte birinden fazla olduğunda bile güvenilir sınıflandırma performansları elde edilebilmiştir.

7. Tüm örneklem büyüklükleri için test verisi oranının değişimden en az etkilenen yöntem Naive Bayes yöntemidir.

8. Örneklem büyüklüğünün artmasından en fazla etkilenen yöntem lojistik regresyon analizidir.

9. Tüm koşullarda en düşük doğruluk oranları en yakın komşuluk yöntemi ile elde edilmiştir.

Küçük örneklemlerde, NB yöntemi ile diğer yöntemlere göre, yüksek doğruluk oranları, daha az hatalı kestirimler yapılmış ve analizler çok kısa sürede tamamlanmıştır. Yapılan bazı araştırmalarda da küçük örneklemlerde NB yönteminin diğer yöntemlere göre daha iyi sonuçlar verdiği görülmüştür (Göker, 2012; Hamalainen ve Vinni, 2006; Hamalainen ve Vinni, 2011, Kotsiantis ve diğerleri, 2003; Osmanbegović ve Suljić, 2012). Araştırmada LR yöntemi, tüm koşullarda YSA yöntemine göre daha yüksek performans göstermiştir. Bu bulgu yapılan bazı araştırma sonuçlarından farklı olmakla birlikte (Bahadır, 2013; Çırak, 2012; Tepehan, 2011) bu durumun oluşmasının en önemli nedeni veri yapısının LR analizi için uygun olmasıdır. LR yöntemi orta ve büyük örneklemlerde, daha az hatalı ve daha yüksek doğrulukta kestirimler yapmaktadır. Dekker ve diğerleri (2009) tarafından yapılan çalışmada, benzer büyüklükte örneklemde LR yöntemi Bayes yöntemine göre daha iyi performans göstermiştir.

NB ve LR yöntemlerinden sonra en yüksek doğruluk oranları sırasıyla MLP ve KNN yöntemleri ile elde edilmiştir. Romero ve diğerlerinin (2013) yaptıkları çalışmada, numerik ve kategorik veri için KNN yönteminin diğer sınıflandırıcılara göre daha düşük performans göstermiştir. Benzer şekilde, bu çalışmada, orta ve büyük örneklemlerde, MLP yöntemi KNN yönteminden daha az hatalı sonuçlar vermiştir. Ancak, küçük örneklemlerde tersi bir durum söz konusudur. Bu durum, KNN yönteminin MLP yöntemine göre daha basit bir istatistiksel yapıya sahip olması ve seçilen k değerinin en yakın komşuluğu belirlemede küçük örneklemlerde daha kararlı davranmasından kaynaklanmıştır. MLP yönteminde ise katman sayısının 3 seçilmesi ağın eğitilmesinde etkili olmasına rağmen küçük örneklemlerde hata miktarının fazla olmasına neden olmuştur.

Bu çalışmada, KNN yöntemi diğer yöntemlere göre tüm koşullarda daha düşük doğru sınıflandırma performansı göstermiştir. Ancak, yapılan bazı araştırmalarda KNN yönteminin de en az YSA ve LR yöntemleri kadar performans gösterdiği görülmüştür (Minaei-Bidgoli, Kashy, Kortemeyer ve Punch, 2003; Yurdakul ve Topal, 2015). Benzer şekilde, Shahiri ve diğerleri (2015), 2002 ile 2015 yılları arasında uluslararası veri tabanlarında yayınlanan çalışmaları karşılaştırmış ve ortalama performans açısından NB yönteminin KNN ve YSA yöntemlerine göre daha düşük performans gösterdiği görülmüştür. Ancak, bu çalışmada, NB yöntemi özellikle küçük ve orta büyüklükteki örneklemlerde daha yüksek sınıflandırma performansı göstermiştir. Bazı araştırmacılar, farklı koşullar ve veriler için bir sınıflandırma yönteminin en iyi olduğunu söylemek doğru olmadığını ifade etmişlerdir (Romero ve diğerleri, 2013; Shahiri ve diğerleri, 2015). Barker ve diğerleri (2004) ise farklı yıllarda mezun olan öğrencilerin mezun olma durumlarına göre yaptıkları sınıflandırmada farklı yıllarda verinin yapısına göre farklı yöntemlerin etkili olabileceğini göstermişlerdir. Bu nedenle, bu araştırmadan elde edilen bulgular, kullanılan verinin yapısı ve analiz koşulları çerçevesinde yorumlanmıştır.

Bu araştırmada, Naive Bayes yöntemi için koşullu bağımsızlık varsayımının sağlanmadığı bir veri ile analizler gerçekleştirilmiştir. Bu sonuç, Hamalainen ve Vinni (2011) tarafından belirtildiği gibi, Naive Bayes'in koşullu bağımsızlık varsayımı karşılanmamış olsa bile iyi performans gösterebileceğini göstermiştir. Yapılacak araştırmalarda bu varsayımın sağlandığı, k-en yakın komşuluk yöntemi için seçilecek k değerinin farklı şekillerde belirlendiği, yapay sinir ağları yönteminde, katman sayısının seçildiği, lojistik regresyon analizi ve yapay sinir ağları yöntemleri için varyans-kovaryans matrislerinin homojenliğinin sağlandığı koşullarda yöntemlerin sınıflandırma performansları değerlendirilip karşılaştırılabilir.

Benzer koşullarda yapılacak uygulamalarda Naive Bayes yönteminin kullanılması, zaman kaybı yaşanmadan geçerliği ve güvenirliği yüksek sonuçların elde edilmesini sağlayacaktır. benzer koşullar için yapılacak uygulamalarda daha yüksek sınıflandırma performansı sağlayabilmek için diğer yöntemler, k-en yakın komşuluk yöntemine tercih edilebilir. Örneklem büyüklüğü fazla olduğunda Naive Bayes ve lojistik regresyon yöntemlerinin YSA'ya tercih edilmesi daha yüksek performans ve zaman tasarrufu sağlayacaktır.