International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 867–882
https://doi.org/10.21449/ijate.1208809
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

# An investigation of data mining classification methods in classifying students according to 2018 PISA reading scores

**Emrah Buyukatak**[1,*], **Duygu Anil**[2]

[1]Independent Researcher
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey

**ARTICLE HISTORY**

*Received: June 10, 2022*
*Revised: Sep. 13, 2022*
*Accepted: Nov. 22, 2022*

**Keywords:**
Data Mining,
Artificial Neural Networks,
Decision Trees,
Cluster Analysis,
Classification.

**Abstract:** The purpose of this research was to determine classification accuracy of the factors affecting the success of students' reading skills based on PISA 2018 data by using Artificial Neural Networks, Decision Trees, K-Nearest Neighbor, and Naive Bayes data mining classification methods and to examine the general characteristics of success groups. In the research, 6890 student surveys of PISA 2018 were used. Firstly, missing data were examined and completed. Secondly, 24 index variables thought to affect the success of students' reading skills were determined by examining the related literature, PISA 2018 Technical Report, and PISA 2018 data. Thirdly, considering the sub-classification problem, the students were scaled in two categories as "Successful" and "Unsuccessful" according to the scores of PISA 2018 reading skills achievement test. Statistical analysis was conducted with SPSS MODELER program. At the end of the research, it was determined that Decision Trees C5.0 algorithm had the highest classification rate with 89.6%, the QUEST algorithm had the lowest classification rate with 75%, and four clusters were obtained proportionally close to each other in Two-Step Clustering analysis method to examine the general characteristics according to the success scores. It can be said that the data sets are suitable for clustering since the Silhouette Coefficient, which is calculated as 0.1 in clustering analyses, is greater than 0. It can be concluded that according to achievement scores, all data mining methods can be used to classify students since these models make accurate classification beyond chance.

## 1. INTRODUCTION

One of the most important criteria for the success of educational policies of countries is to be able to train qualified and successful individuals in accordance with the information and data era. Success is determined by evaluating the performances at the national and international levels whether the planned targets in the education systems have been achieved in the recent period. Today for this purpose, education systems are evaluated by using large-scale exams which are applied to large groups covering the specified knowledge and skills for more than one course to monitor what students learn in the school environment. In addition, the learning skills of students of a certain age and school group in different countries are regularly monitored and compared.

*CONTACT: Emrah Buyukatak ✉ ebuyukatak@hotmail.com

e-ISSN: 2148-7456 /

In large-scale exams, it has become important to use open-ended questions or open-ended and multiple-choice questions together, which allows measuring high-level cognitive skills and allows students to give their own answers since open-ended questions give students the opportunity to think and create their own answers.

PISA measures students' high-level cognitive skills by investigating not only whether the basic knowledge learned at school is re-used, but also whether students can guess about what they do not know using knowledge that they have learned and whether they can apply what they know inside and outside of school. In PISA, not only knowledge and skills in Turkish, mathematics and science, but also attitudes towards Turkish, mathematics, and science are discussed, and also whether they are aware of what opportunities the scientific competencies they gain at school will create for them is evaluated (Anıl, 2008). Large-scale achievement tests such as PISA are achievement tests that mostly consist of multiple subtests or dimensions in different grade levels and courses. PISA is applied to large student groups and a huge amount of data are obtained from this exam.

PISA is carried out regularly and information on many variables is collected. Since there is a large amount of information about students in such a large-scale exam, data in this application are also defined as big data. This information in different formats, which emerges from both test scores and questionnaires and is also obtained from more than half a million students, constitutes a large pile of data. The important thing here is to determine which is meaningful and which is meaningless from such a large amount of data in PISA in the decision process. As a result, decisions can be made as to whether this data can be used in data mining since large amount of information obtained from students in recent years is big data (Nisbet, Elder, & Miner, 2009). With these methods, behavioral patterns of individuals are analyzed and predictions are made for future behaviors.

The amount of information produced and stored at the global level is unimaginably large and on the increase every day. However, data in these areas should be stored and managed securely in a magnetic environment using database systems. As a result of such needs, powerful systems and tools are needed to systematically reveal efficient information from large amounts of data and to transform them into organized data and then knowledge. Data mining emerged in the 1980s when computers began to be used to solve data analysis problems. Data mining is called an interdisciplinary field of study that combines various techniques such as machine learning, pattern recognition, statistics, databases, and visualization to solve the problem of obtaining information from large data sets (Cabena et al., 1998). Data mining is also expressed as the process of applying one or more computer learning techniques to automatically extract and analyze information from the data in the database (Roiger, 2017). In addition, this process is the use of multiple data analysis tools to reveal patterns in the data and the relationship between the data in order to make valid predictions. In this direction, data mining techniques make it possible to reveal the relationship between the parameters of large amounts of data in largescale exams such as PISA and TIMMS.

Data such as the most important element of the education process, students' personal information, grade status, absenteeism, and successful and unsuccessful courses are obtained by Educational Data Mining (EDM), which examines data mining in terms of education. By applying different models to these data, it is possible to determine the reasons for success, to increase their success, to prevent their absenteeism, to choose the courses they will take, and to make recommendations regarding their career goals (Rizvi, Rienties, & Khoja, 2019). In this way, the discovery of patterns based on these data and the use of discovered patterns in the improvement of the learning process and in instructional design have emerged as important issues today. By this means, data mining techniques are used in education in forming groups according to students' personal characteristics and individual learning similarities, predicting

undesirable student behaviors such as low motivation, absenteeism, dropping out of school, and not following school rules, and taking necessary precautions (Aksoy, 2014).

Educational Data Mining (EDM) is the creation, research, and application of analysis methods in digital environments to detect patterns in multi-volume educational data, which is very difficult to analyze due to large data (Romero & Ventura, 2013). Data in EDM are not limited to interactions of students, and data from students, administrative data, demographic data, and emotional characteristics of trainees together constitute the EDM data (Witten & Frank, 2000). To make determinations about student success, to make inferences about failure in the education environment and its causes, and to create educational environments that meet the needs, educational data mining, which uses many different disciplines such as psychometry, learning analytics, and statistics can be benefitted (Özbay, 2015).

Nowadays, it has been thought that data mining will be useful especially in the selection and classification made taking into account the measurement results in the field of education. In this way, it will be possible to understand the learning level and behavior of students better by determining which variable may be effective in which cluster or class. As a result, the number of prediction studies conducted to determine the factors affecting student success and the shaping of this success has increased significantly (Anıl, 2008; Gelbal, 2008; Erdil, 2010; Özer & Anıl, 2011). In addition, it is very difficult to make prediction and classifications in groups that are similar to each other, which makes it necessary to carefully select the methods used in research and ensure the classification with the most accurate prediction.

When the related studies are evaluated as a whole, data mining methods can be seen to have been used intensively on a sectoral basis, especially in industry and banking. Although such methods offer a wide field of study in the field of education and the number of studies in education related to the concept of data mining has increased nationally and internationally, it is observed that very few studies have been carried out and specifically domestic studies and resources have been scarce. However, using the data collected in education is of central importance for achieving success and increasing student achievement in this field. As a result of collecting more data in the field of education along with technological developments, this research is important in terms of examining data mining methods in educational fields other than the usual sectoral basis.

Different methods and algorithms have been used in the literature as to recent prediction and classification studies in data mining, and the models used in these applications have a unique algorithm. Evaluating the algorithms by comparison or revealing which algorithm is successful in situations is important in terms of increasing classification performances. Research in data mining has generally been limited to Artificial Neural Networks and Decision Trees. However, this study is important in terms of using and comparing Artificial Neural Networks, Decision Trees, K-Nearest Neighbor, and Naive Bayes methods for classification models that will allow predicting the future success of students. In this study, apart from the most used classification methods, other data mining classification methods that are considered to make significant contributions to the literature are examined. In addition, using the Two-Step Clustering analysis, different groups gathered in the same cluster according to the similar characteristics of the students' data in the large-volume PISA 2018 data set were determined and the importance of the variables on these groups was examined.

The purpose of this research was to determine classification accuracy of the factors affecting the success of students' reading skills and the success scores of reading skills, based on PISA 2018 data by using data mining classification methods such as Artificial Neural Networks, Decision Trees, K-Nearest Neighborhood, and Naive Bayes and to examine the general characteristics of success groups. For this purpose, the following sub-problems were examined in this study.

1.Considering the factors affecting the students' success in 2018 PISA reading skills and their success scores, at what accuracy rate do Artificial Neural Networks, Decision Trees, K-Nearest Neighborhood, and Naive Bayes analyses classify students according to their success?

2.What are the general characteristics of the achievement groups according to the factors affecting the 2018 PISA reading skills success of the students and their success scores in reading skills?

3.What are the results regarding the comparison of the general classification rates of the students of Artificial Neural Networks, Decision Trees, K-Nearest Neighborhood, and Naive Bayes methods according to their success?

## 2. METHOD

This study was conducted to examine different classification models. In this respect, the research is a descriptive study.

### 2.1. Study Group

186 schools and 6890 students represented Turkey in the PISA 2018 application. Since the items that were mostly not answered or not entered any responses in the study were excluded from the data set, the sample of the study consisted of 6431 students.

### 2.2. Data Collection Tools

Each PISA application focuses on one of the fields of mathematics, reading, and science. PISA 2018 focused on predominantly reading skills and also mathematical literacy and science literacy.

PISA 2018 included cognitive tests aiming to measure the academic performance of students and questionnaires of student and school were prepared to evaluate the student as a whole. Students were expected to answer the questionnaire, which consisted of questions about oneself, family and home, language learning at school, the Turkish / Turkish Language and Literature Lesson learned at school, thoughts about life, school, school program, and learning periods. The main student questionnaire, computer-based, consisted of 79 questions and lasted 35 minutes. The data used in the study consist of student questionnaire and cognitive test results and were downloaded from the OECD website.

### 2.3. Data Analysis

In the literature there are not any assumptions that need to be tested before these techniques can be applied. However, missing data analysis was done by considering the mechanisms of missing data patterns and amounts. As a result of the examination carried out before the analysis in the study, 459 data were removed from the data set due to responses either mostly not answered or not entered at all, and the analysis was carried out with 6431 data. In addition, after the missing data analysis, it was determined that 1678 data were missing. Since the exclusion of the data of 1678 students from the analysis would not give correct results, missing data were completed with the EM logarithm.

In the study, importance was given to the selection of variables that affect the success of students' reading skills in the selection of variables based on PISA 2018 data. Within the scope of variable selection, literature, PISA 2018 Technical Report, and PISA 2018 data were examined. As a result, 24 indices that were considered to affect the success of students' reading skills were determined in this study. The variables used in the study were "Index of Economic, Social and Cultural Status (ESCS)", "Family Wealth (WEALTH)", "Understanding and Remembering (UNDREM)", "Summarizing (METASUM)", "Reading and Using Strategies (METASPAM)", "Joy/Like Reading (JOYREAD)", "Disciplinary Climate (DISCLIMA)", "Home Educational Resources (HEDRES)", "Home Possessions (HOMEPOS)", "Information

and Communication Technologies Resources (ICTRES)", "Cultural Possessions at Home (CULTPOSS)", "Teacher Support (TEACHSUP)", "Teacher's Stimulation of Reading Engagement Perceived by Student (STIMREAD)", "Self-Concept of Reading: Perception of Competence (SCREADCOMP)", "Self-Concept of Reading: Perception of Difficulty (SCREADDIFF)", "Perception of Difficulty of the PISA Test (PISADIFF)", "Parents' Emotional Support Perceived by Student (EMOSUPS)", "Perceived Feedback (PERFEED)", "Subjective Well-Being: Positive Affect (SWBP)", "Perception of Cooperation at School (PERCOOP)", "Subjective Well-Being: Sense of Belonging to School (BELONG)", "Use of ICT in Leisure Activities out of School (ENTUSE)", "Use of ICT for School Work Outside of School (HOMESCH)" and "Use of ICT at School (USESCH)".

Students were scaled in two categories as "Successful-Unsuccessful" according to the scores in PISA 2018 reading skills achievement test. First, "average reading achievement score" variable was formed by taking the average mean of the 10 reading achievement scores (Plausible Value: PV1READ, PV2READ … PV10READ) of every student. Then the mean of this variable was calculated as 470. If any students' "average reading success score" is below 469.9, it is called "unsuccessful-0", and if it is above 469.9, it is called "successful-1". The "success status" variable was created in such a way. In the light of these regulations, the number of "successful1" students was 3212 with 49.9%. The number of "unsuccessful-0" students was 3219 with 50.1% in the PISA 2018 Turkey application, in which 6431 students participated.

During the model evaluation process, both Cross Validation and Bootstrap methods were used to ensure that many models were created and tested. In order to increase the accuracy of the methods and algorithms, the analysis was run with Boosting, and the 10-fold Cross Validation technique was used in the development of the models. Before the analysis, the data set was divided into 70% training and 30% test data. In the literature, some studies split the data into three parts as training, test, and validation, while some research splits the data into training and test sets. In this study, data was split into two parts; namely, data as training and test set because in the study Cross Validation method was used to ensure that many models were created and tested. When using a method such as cross validation, two partitions may be sufficient and effective, thereby averaged after repeated rounds of model training and testing to help reduce bias and variability (Xu & Goodacre, 2018). The seed value of analysis to reproducibility is 2695748.

In the study, "success status" variable is a dependent variable and 24 index variables are independent ones. Artificial Neural Networks, Decision Tree algorithms, K-Nearest Neighborhood, and Naive Bayes analyzes were made using SPSS Modeler 18.0 program. As a result of the analyses, the correct classification rates of each model and algorithm's training and test data were calculated. The overall correct classification rate of all data set was calculated using the following equation:

$$\text{The Overall Correct Classification Rate} = \frac{\text{Number Of Correctly Classified Samples}}{\text{Total Number Of Samples}}$$

To test the accuracy of the classification of a model, the relative and maximum chance criteria need to be calculated and compared. According to the success status of the sample, the maximum chance criterion of "successful" and "unsuccessful" students is 0.51 The relative chance criterion is 0.49 In this study, the percentages of classification accuracy determined were evaluated by comparing them with the maximum and relative chance criteria.

In this specific research, clustering analysis was performed in order to group the ungrouped data according to their similarities. Two-Step Clustering algorithm is preferred for large and high-dimensional data consisting of both categorical and continuous data. In the study,

Two-Step Clustering Analysis was carried out through the SPSS program in order to determine the different groups by collecting the data of the students in the same cluster according to their similar characteristics (variables) and to examine the importance of the variables on these groups. In this analysis "success status" variable is a dependent variable and 24 index variables are the independent variables.

## 3. RESULT

This section presents the research findings obtained from the analyses carried out in parallel with the research questions and makes brief interpretations of these findings as well. To predict the success of students' reading skills, artificial neural networks "Multilayer Perceptron (MLP)" model was used. One dependent and 24 independent index variables were included in this model. Using the total data set, 70.8% (n=4556) of the data were allocated to the training set and 29.2% (n=1875) to the test set. While predicting the success of reading skills, 50 trials were made to find the architecture of the network that gives the best performance. Artificial neural network is three layers, and there are 24 artificial nerve cells (neurons) in the first layer (input layer) and seven artificial nerve cells in the hidden layer, which is the second layer. In the last layer (output), there are two nerve cells representing each level of the dependent variable. “Hyperbolic Tangent Function” is applied as activation function in hidden layer and “Softmax Function” is applied in output layer. The results of the analysis regarding the success of reading skills with the artificial neural network are shown in Table 1. The seed value of analysis to reproducibility is 2695748.

**Table 1.** *Analysis Results on Artificial Neural Networks Reading Skills Achievement.*

| Sample | Observed | Estimated | | |
|---|---|---|---|---|
| | | Unsuccessful | Successful | Classification Rate |
| Training | Unsuccessful | 1776 | 507 | 77.8% |
| | Successful | 553 | 1691 | 75.4% |
| | Total | 51.4% | 48.6% | 76.6% |
| Test | Unsuccessful | 688 | 248 | 73.5% |
| | Successful | 276 | 692 | 71.5% |
| | Total | 50.6% | 49.4% | 72.5% |

When Table 1 is examined, it is seen that the artificial neural network correctly predicted the reading skills success of the students in the training sample with a performance of 76.6% and the reading skills success of the students in the test sample with a performance of 72.5%. In addition, it correctly classified 75.4% of the successful students in the training dataset and 71.5% of the successful students in the test dataset. While 77.8% of the unsuccessful students in the training dataset were classified correctly, 73.5% of the unsuccessful students in the test dataset were classified correctly. The overall correct classification rate of the training and test data sets was calculated as 75.4%. The maximum chance criterion of the sample is 0.51 and the relative chance criterion is 0.49 The value of 75.4% is above the maximum and relative chance criterion. This result shows that artificial neural networks can be used successfully in classification in this model.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, the most important input variables related to the success of reading skills can be seen as "Home Possessions (100%)" and "Family Wealth (82%)" as the most important determinants of success regarding reading skills. The independent variables that have the least effect on the success of reading skills include “Teacher's Stimulation of Reading Engagement Perceived by Student (14.8%)” and “Subjective Well-Being: Positive Affect (9.9%)”.

ROC analyzes are performed in the analysis of artificial neural network models. The area under the ROC curve is called the "AUROC or AUC" (Area under the ROC curve) and is a measurement that helps determine the reliability of the model. The AUROC value takes values between 0.5 and 1.0. The closer the probabilities of the AUROC index are to one, the more successful the result will be. In the relevant literature, it is stated that the discrimination ability of the prediction model can be classified as follows:

'AUROC' =0.5 No prediction probability, so no discrimination.

0.7≤ 'AUROC'≤ 0.8 statistically acceptable discrimination.

0.8≤ 'AUROC' ≤0.9 statistically perfect discrimination.

'AUROC'>0.9 is statistically outstanding.

As can be seen in Table 2, a statistically perfect discrimination ability with 0.837 value was presented by the model. With this analysis, the performance of the model was also tested.

**Table 2.** *Areas under the Curve as a Result of ROC Analysis.*

| | | Areas Under the Curve |
|---|---|---|
| Success Status | Unsuccessful | 0.837 |
| | Successful | 0.837 |

To predict students' reading skills success with decision tree, the results of analysis of four decision tree algorithms were examined. One dependent and 24 independent index variables were included in the analysis of the decision trees algorithm, and using the total data set, 69.6% (n=4476) of the data were determined for training and 30.4% (n=1955) for testing.

In the study, the C5.0 algorithm was run with "Boosting" to increase the accuracy rate. In this model, a 10-fold cross-validation test was used as a validation test. The standard deviation of the model determined by the cross-validation method was 0.7% and the depth of the decision tree was 21. The analysis results regarding the success of reading skills with the C5.0 algorithm are shown in Table 3.

**Table 3.** *Analysis Results of C5.0 Algorithm.*

| Sample | Observed | Estimated | | |
|---|---|---|---|---|
| | | Unsuccessful | Successful | Classification Rate |
| Training | Unsuccessful | 2014 | 219 | 91% |
| | Successful | 259 | 1984 | 88.5% |
| | Total | 50.7% | 49.2% | 89. 3% |
| Test | Unsuccessful | 895 | 91 | 90.8% |
| | Successful | 96 | 873 | 90% |
| | Total | 50.7% | 49.3% | 90.4% |

When Table 3 is examined, it is seen that the C5.0 algorithm correctly predicted the success in reading skills of the students in the training sample with a performance of 89.3%, and the success in reading skills of the students in the test sample with a performance of 90.4%. In addition, it correctly classified 88.5% of the successful students in the training dataset and 90% of the successful students in the test dataset. While 91% of the unsuccessful students in the training dataset were classified correctly, 90.8% of the unsuccessful students in the test dataset were classified correctly. According to this result, the overall correct classification rate of the C5.0 algorithm was calculated as 89.6%. The maximum chance criterion of the sample is 0.51 and the relative chance criterion is 0.49 Since overall correct classification rate is above these values, it can be concluded that the C5.0 algorithm can be used successfully in classification.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, it is seen that the “Self-Concept of Reading: Perception of Difficulty (0.047)” variable is the most important determinant of the success of reading skills. It is seen that the variability in the " Self-Concept of Reading: Perception of Difficulty " greatly affects the success of reading skills. The variable “Index of Economic, Social and Cultural Status (0.037)”, that is, the socio-economic status of students, is the most ineffective independent variable on the success of reading skills. This situation shows that the socio-economic development and wealth of the student are not important on the success of reading and do not contribute to their success.

In the analysis of the CHAID algorithm, the largest tree depth was 10, the chi-square calculation method is Pearson, the stopping criteria were calculated as 2% for the root node, 1% for the child node, and the largest iteration was 100. The analysis results regarding the success of reading skills with the CHAID algorithm are shown in Table 4.

**Table 4.** *Analysis Results of CHAID Algorithm.*

| Sample | Observed | Estimated | | |
|---|---|---|---|---|
| | | Unsuccessful | Successful | Classification Rate |
| Training | Unsuccessful | 1830 | 403 | 82% |
| | Successful | 387 | 1856 | 82.7% |
| | Total | 49.5% | 50.4% | 82.3% |
| Test | Unsuccessful | 669 | 317 | 68% |
| | Successful | 291 | 678 | 70% |
| | Total | 49.1% | 50.9% | 69% |

When Table 4 is examined, it is seen that the CHAID algorithm correctly predicted the reading skills success of the students in the training sample with a performance of 82.3%, and the reading skills success of the students in the test sample with a performance of 69%. In addition, it correctly classified 82.7% of the successful students in the training dataset and 70% of the successful students in the test dataset. While 82% of unsuccessful students in the training dataset were classified correctly, 68% of unsuccessful students in the test dataset were classified correctly. Based on this result, it is possible to say that the CHAID algorithm gives good results in predicting successful students. The overall correct classification rate of the CHAID algorithm was calculated as 78.2%. It can be concluded that the CHAID algorithm can be used successfully in classification because the classification rate of the CHAID algorithm, which is 78.2%, is above the maximum and relative chance criterion values.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, it is seen that the most important input variables are “Reading and Using Strategies (0.18)" and "Summarizing (0.11)". It is seen that “Teacher Support (0.004)” is the most ineffective one on the success of reading skills. This situation shows that teachers' help to students in learning and their support in understanding a subject are not much important on the success of reading, and teacher support on the success of reading does not contribute to their success in learning and comprehension.

In the C&RT algorithm analysis, the largest tree depth is five, the largest number of proxies is zero (indicating that there is no missing value in the data set), impurity measurement is Gini for the categorical target area, stopping criteria is 2% for the root node and 1% for the child node. The analysis results obtained for the C&RT algorithm are shown in Table 5.

**Table 5.** *Analysis Results of C&RT Algorithm.*

| Sample | Observed | Estimated | | |
|---|---|---|---|---|
| | | Unsuccessful | Successful | Classification Rate |
| Training | Unsuccessful | 1734 | 799 | 68.4% |
| | Successful | 488 | 1755 | 78.2% |
| | Total | 46.5% | 53.4% | 77.9% |
| Test | Unsuccessful | 736 | 250 | 74.6% |
| | Successful | 264 | 705 | 72.7% |
| | Total | 51.1% | 48.8% | 73.7% |

Table 5 shows that the C&RT algorithm correctly predicted the reading skills success of the students in the training sample with a performance of 77.9%, and the reading skills success of the students in the test sample with a performance of 73.7%. In addition, it correctly classified 78.2% of the successful students in the training dataset and 72.7% of the successful students in the test dataset. While 68.4% of the unsuccessful students in the training dataset were classified correctly, 74.6% of the unsuccessful students in the test dataset were classified correctly. According to these results, it is possible to say that the C&RT algorithm gives good results in predicting especially successful students in the same way as the CHAID algorithm does. The overall correct classification rate of the C&RT algorithm was calculated as 76.6%. It can be concluded that the C&RT algorithm can be used successfully in classification because the value of 76.6% is above the maximum and relative chance criteria of the sample.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, it is seen that as in the CHAID algorithm analysis the most important input variables are "Reading and Using Strategies (0.18)" and "Summarizing (0.10)", while “Teacher Support (0.005)” is the most ineffective one on the success of reading skills. However, values of the degree of importance are different from those in the CHAID algorithm analysis. This situation shows that the teachers' help to students in learning and their support in understanding a subject are not much important on the success of reading, and teacher support on the success of reading does not contribute to their success in learning and comprehension.

Quadratic separation analysis was used in the QUEST algorithm, and each node was divided into two subgroups. Analysis parameters are maximum tree depth 10, maximum number of proxies 0, Alpha (for splitting) 0.05, stopping criteria 2% for root node, and 1% for child node. Analysis results of the QUEST algorithm are presented in Table 6.

**Table 6.** *Analysis Results of QUEST Algorithm.*

| Sample | Observed | Estimated | | |
|---|---|---|---|---|
| | | Unsuccessful | Successful | Classification Rate |
| Training | Unsuccessful | 1715 | 518 | 76.8% |
| | Successful | 555 | 1688 | 75.2% |
| | Total | 50.7% | 49.2% | 76.3% |
| Test | Unsuccessful | 732 | 254 | 74.2% |
| | Successful | 268 | 701 | 72.3% |
| | Total | 51.1% | 48.8% | 73.3% |

When Table 6 is examined, it can be seen that the QUEST algorithm correctly predicted the reading skills success of the students in the training sample with a performance of 76.3% and the reading skills success of the students in the test sample with a performance of 73.3%. In addition, it correctly classified 75.2% of the successful students in the training dataset and 72.3% of the successful students in the test dataset. While 76.8% of the unsuccessful students

in the training dataset were classified correctly, 74.2% of the unsuccessful students in the test dataset were classified correctly. According to these results, it is possible to say that the QUEST algorithm gives good results in predicting unsuccessful students. The overall correct classification rate of the QUEST algorithm was calculated as 75%. The maximum and relative chance criteria of the sample are 0.51 and 0.49, respectively. The results of the QUEST algorithm analysis show the classification rate of 75% above these values. This result shows that the QUEST algorithm can be used successfully in classification in this model.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, it is seen that the most important input variables are "Reading and Using Strategies (0.196)" and "Summarizing (0.115)". The independent variable that has the least effect on the success of reading skills is Subjective Well-Being: Positive Affect (0.0001)". "Subjective Well-Being: Positive Affect " variable, that is, different emotions that students may have when they evaluate themselves (joyful, cheerful, and happy), has little effect on the success of their reading skills. It shows that the positive effects and emotions of the students are not important on their success of reading, and the happiness of the students does not contribute to their success in reading.

In the K-Nearest Neighbor analysis, Manhattan Distance Measure was chosen as the distance measure since it increases the accuracy rate. For the validity test, the k value, which gives the lowest error rate as a result of the 10-fold cross-validation test, was calculated as five. K-Nearest Neighbor method analysis results are presented in Table 7.

When Table 7 is examined, it can be seen that the K-Nearest Neighbor method correctly predicted the reading skills success of the students in the training sample with a performance of 81.2%, and the reading skills success of the students in the test sample with a performance of 82.1%. In addition, it correctly classified 81.9% of the successful students in the training dataset and 83% of the successful students in the test dataset. While 80.6% of the unsuccessful students in the training dataset were classified correctly, 81.2% of the unsuccessful students in the test dataset were classified correctly. According to this result, it is possible to say that the K-Nearest Neighbor method gives good results, especially in predicting successful students. The overall correct classification rate of the K-Nearest Neighbor was calculated as 81.5%. This result of analysis shows that K-Nearest Neighbor method can be used successfully in classification in this model, since the classification rate is above the maximum and relative chance criterion values.

**Table 7.** *Analysis Results of K-Nearest Neighbor.*

| Sample | Observed | Estimated | | |
|---|---|---|---|---|
| | | Unsuccessful | Successful | Classification Rate |
| Training | Unsuccessful | 1801 | 432 | 80.6% |
| | Successful | 406 | 1837 | 81.9% |
| | Total | 49.3% | 50.6% | 81.2% |
| Test | Unsuccessful | 801 | 185 | 81.2% |
| | Successful | 164 | 805 | 83% |
| | Total | 49.3% | 50.6% | 82.1% |

When the degree of importance of the independent variables on the success of reading skills is examined, it is seen that the most important input variables for the K-Nearest Neighbor method are "Reading and Using Strategies (0.0438)" and "Summarizing (0.0433)". The effects of the variables are very close to each other, but the variability in "Reading and Using Strategies" greatly affects the success of reading skills. In addition, the variable "Perception of Cooperation at School (0.0409)", that is, cooperation among students in learning, is the most ineffective

independent variable on the success of reading skills. This situation that cooperation between students, or giving importance to cooperation, is not important on the success of reading and does not contribute to their reading success.

The findings regarding the success of reading skills with Naive Bayes analysis are shown in Table 8.

**Table 8.** *Analysis Results of Naive Bayes.*

| Sample | Observed | Estimated | | |
|---|---|---|---|---|
| | | Unsuccessful | Successful | Classification Rate |
| Training | Unsuccessful | 1716 | 517 | 76.8% |
| | Successful | 517 | 1726 | 76.9% |
| | Total | 49.8% | 50.1% | 76.9% |
| Test | Unsuccessful | 757 | 229 | 76.7% |
| | Successful | 225 | 744 | 76.7% |
| | Total | 50.2% | 49.7% | 76.78% |

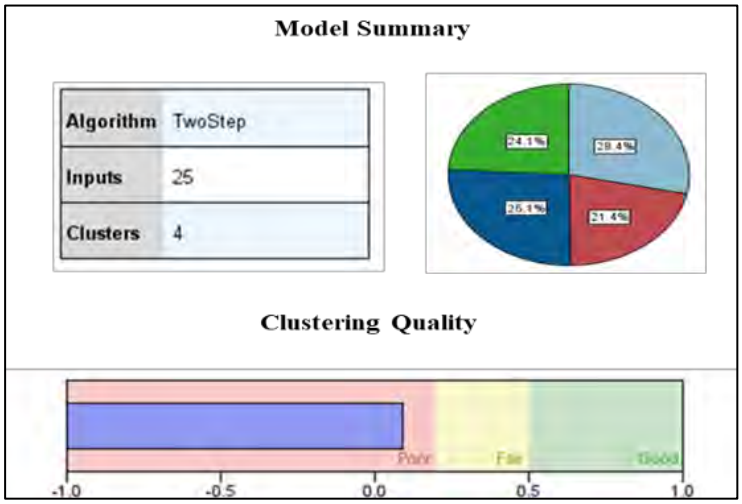
When Table 8 is examined, it can be seen that the Naive Bayes method correctly predicted the reading skills success of the students in the training sample with a performance of 76.9% and the reading skills success of the students in the test sample with a performance of 76.78%. In addition, it correctly classified 76.9% of the successful students in the training dataset and 76.7% of the successful students in the test dataset. While 76.8% of the unsuccessful students in the training dataset were classified correctly, 76.7% of the unsuccessful students in the test dataset were classified correctly. The overall correct classification rate of the Naive Bayes method was calculated as 76.8%. The overall classification rate as a result of analysis is above the maximum and relative chance criteria of the sample. According to this result, it can be concluded that the Naive Bayes method can be successfully used in classification in this model.

When the degree of importance of the independent variables on the success of reading skills is examined, it is seen that the most important input variables are “Disciplinary Climate (0.667)” and “Perception of Difficulty of the PISA Test (0.623)”. The independent variable that has the least effect on the success of reading skills is “Use of ICT at School (0.367)”. This situation that students' use of information and communication technologies at school is not important on the success of reading and does not contribute to their reading success.

In the Two-Step Clustering Analysis using one dependent and 24 independent index input variables, the Silhouette Coefficient was calculated as 0.1 and the clustering quality indexed to the Silhouette coefficient is shown in Figure 1. In the literature, a precise threshold value is not defined in the evaluations regarding the Silhouette coefficient. However, it is stated that a coefficient value greater than 0 is sufficient for clusters, and the larger the coefficient, the better the quality of the cluster. In this context, it can be concluded that although the Silhouette coefficient value (0.1) in the Two-Step Clustering Analysis is small, it is sufficient for clustering.

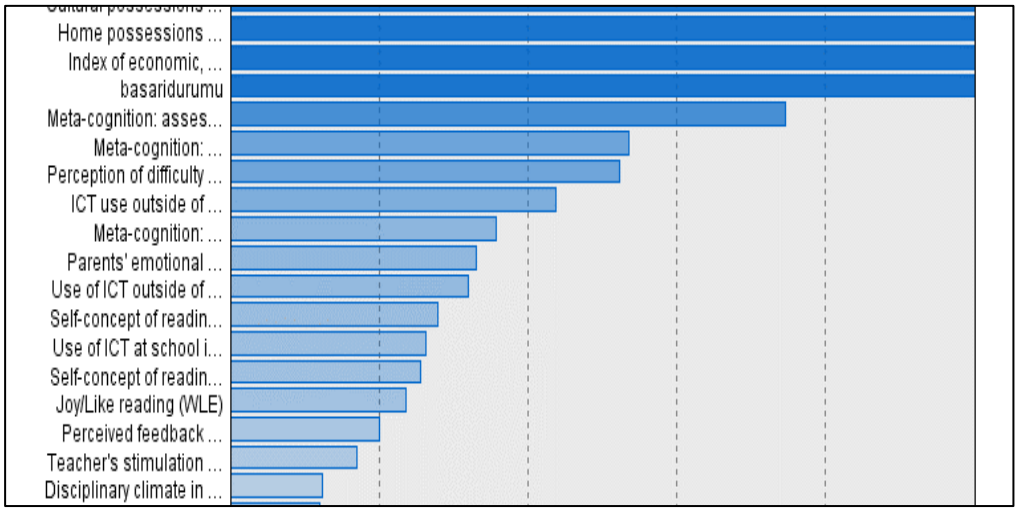
As a result of the clustering analysis, Silhouette coefficient four clusters were obtained, and it was determined that the distributions of these clusters were proportionally close to each other. The ratio from the largest to the smallest cluster was found to be 1.33. This ratio should be less than 2. In this context, it is seen that the size of the clusters and the ratio from the largest to the smallest cluster are appropriate. Variables according to their importance in cluster analysis are shown in Figure 2.

**Figure 1.** *Clustering Quality Indexed to Silhouette Coefficient.*




According to the findings, successful students gathered in the First and Second Clusters. It is seen that "Use of ICT for School Work Outside of School ", "Perceived Feedback", " Perception of Cooperation at School", " Perception of Difficulty of The PISA Test", and " Self-Concept of Reading: Perception of Difficulty" did not have a significant effect on successful students, while "Success Status", "Reading and Using Strategies", "Summarizing" "Family Wealth", "Information and Communication Technologies (ICT) Resources", and "Home Educational Resources " variables had a significant effect by performing well. On the other hand, unsuccessful students were gathered in the Third and Fourth Clusters. In terms of variables, it is revealed that "Reading and Using Strategies", "Summarizing", "Meta-Cognition: Understanding and Remembering" and "Joy/Like Reading" did not have a significant effect on unsuccessful students, while "Use of ICT For School Work Outside of School" of “Using” and “Use of ICT at School” hag a significant effect by performing well.

**Figure 2.** *The degree of importance of cluster analysis independent variables.*




The results of the comparison of correct classification rates according to the analysis of Artificial Neural Networks, Decision Trees, K-Nearest Neighbor, and Naive Bayes methods are shown in Table 9.

**Table 9.** *Analysis Results of Naive Bayes.*

| Method | | Classification Rate (%) |
|---|---|---|
| Artificial Neural Networks | | 75.4 |
| Decision Trees | C5.0 | 89.6 |
| | CHAID | 78.2 |
| | C&RT | 76.6 |
| | QUEST | 75 |
| K-Nearest Neighbor | | 81.5 |
| Naive Bayes | | 76.8 |

As is seen in Table 9, Decision Trees C5.0 algorithm has the highest classification rate with 89.6%. The second highest rate is the K-Nearest Neighbor method with 81.5%. QUEST algorithm has the lowest classification rate with 75%. However, the classification rates of other methods and algorithms are close to each other. The results of the analysis made according to the success status of the students participated in the PISA 2018 Turkey application are above the maximum chance criterion and the relative chance criterion of the samples. According to the results, Artificial neural networks, Decision Tree algorithms, K-Nearest Neighborhood, and Naive Bayes methods can be used successfully in classifying students according to their success since these models make accurate classification beyond chance.

These results are in parallel with the study of Calis, Kayapınar, and Çetinyokuş (2014), who used decision trees for classification in data mining and tested the accuracy of classification according to demographic structures of individuals in four decision tree algorithms and revealed that C5.0 had a higher correct classification rate than that of other algorithms. Similarly, credit scores were calculated by comparing neural networks, M5, logistic regression, and K-Nearest Neighborhood (KNN) algorithms in the study by Liu and Schumann (2005), and the highest classification accuracy was obtained with the K-Nearest Neighbor (KNN) method. In the study, where the models obtained by Artificial Neural Networks and Decision Trees methods to compare the insurance risk estimation performances, the prediction success of the Decision Trees method was found to be higher, although both methods are at an acceptable level (Şahin, 2018). These results also show that there is a parallelism between the studies.

## 4. DISCUSSION and CONCLUSION

Statistical results and inferences are revealed with the analysis of data types that occur for the solution of research problems in the field of education with Educational Data Mining. In this study, based on PISA 2018 data, those factors affecting the success of students' reading skills and the success scores of their reading skills were examined with data mining classification methods and classification accuracies.

When the findings are evaluated, it is seen that Artificial Neural Networks classify with 75.4%, Decision Tree algorithms C5.0 89.6%, CHAID 78.2%, CART 76.6%, QUEST 75%, K-Nearest Neighborhood 81.5%, and Naive Bayes method 76.8%. In the study Decision Trees C5.0 algorithm has the highest classification rate with 89.6%, and the QUEST algorithm has the lowest classification rate with 75%. It is seen that the classification rates of other methods and algorithms are close to each other. The K-Nearest Neighbor method has the second highest rate of classification by having a higher classification rate than that of other methods. These findings coincide with the study in which the C5.0 decision tree algorithm makes the best prediction based on the analysis of Logistic Regression, Artificial Neural Networks, Decision Tree algorithms using student credentials, previous success status, and electronic learning data (Aydın, 2007). In addition, it is possible to say that there is a parallelism with the study in which the success rate of the K-Nearest Neighbor (KNN) analysis was found to be much higher than

that of other data mining algorithms, as a result of examining the success rate of the model developed for estimating the cellular location of proteins in the field of biotechnology (Cai & Chou, 2003).

In order to examine the general characteristics of the success groups according to the 2018 PISA reading achievement scores of the students, four clusters were obtained as a result of the Two-Step Cluster Analysis method, and it was determined that the distributions of these clusters were proportionally close to each other. In the Two-Step Cluster Analysis, the Silhouette Coefficient was calculated as 0.1. Since this coefficient is greater than 0.1, it can be said that the data set is suitable for clustering. The ratio of largest cluster to the smallest cluster, which should be less than 2, is 1.33. According to these findings, it is revealed that clustering is appropriate. When the variables are examined according to their importance, it is seen that the degree of importance of "Information and Communication Technologies (ICT) Resources", "Family Wealth", "Home Educational Resources", "Cultural Possessions at Home", "Home Possessions", "Index of Economic, Social and Cultural Status", and "Success Status" is 1 and they are effective in clustering and the most distinguishing variables by making a significant difference. It was found that the variables of "Disciplinary Climate", "Subjective Well-Being: Positive Affect", "Subjective Well-Being: Sense of Belonging to School", "Perception of Cooperation at School", and "Teacher Support" are not effective in distinguishing those with the lowest discrimination and do not make a significant difference.

The maximum chance criterion calculated within the scope of the ratios of the "successful" and "unsuccessful" students in the sample is 0.51, and the relative chance criterion is 0.49.9. It is evaluated that these Artificial Neural Networks, Decision Trees, K-Nearest Neighbor (KNN), and Naive Bayes methods can be used to classify students according to their success status and the produced models can correctly classify beyond chance, since the classification rates are above the sample's maximum and relative chance criterion values. It has been revealed that reading achievement scores are effective in separating students according to their success status and make a significant difference.

There are different methods and algorithms for prediction and classification, which has been studied extensively in data mining. However, when the studies are examined, it is revealed that Artificial Neural Networks and Decision Trees are the most studied methods. In other studies, on the same or similar sample, success can be estimated or predicted by means of such other classification methods as Regression Support Vector Machines, K-Means, and Time Series Analysis.

In the first stage of the study, loss and missing data were completed, and then the analyses were made. However, some analyses of classification methods in data mining can also be performed with missing data. In this context, how the analyses of the same or similar data sets and other classification methods perform in missing data can be examined.

SPSS Modeler program was used for the analysis. There are many data mining analysis Üprograms. In order to compare the programs, data mining methods and analysis programs can be compared using the same data sets. However, similar studies can be conducted on exams with different data such as TIMMS and PIRLS.

Studies in the field of education are carried out with different sample sizes, including different variables and different methods to divide students as successful and unsuccessful such as upper-lower 27% groups. Therefore, it is necessary to use a large number of methods and algorithms for classification and comparison purposes in order to determine which method or algorithm performs better in the sample used.

## Acknowledgments

This study is based on a summary of the doctoral dissertation entitle "Examination of Reading Literacy Levels in PISA 2018 Turkey Sample with Different Data Mining Classification Methods".

## Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

## Authorship Contribution Statement

**Emrah Büyükatak:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Duygu Anıl:** Methodology, Supervision, and Validation. Authors may edit this part based on their case.

## Orcid

Emrah Büyükatak https://orcid.org/0000-0002-5341-5053
Duygu Anıl https://orcid.org/0000-0002-1745-4071

## REFERENCES

Aksoy, E. (2014). *Determination of the mathematically gifted and talented students using data mining in terms of some variables* [Master Thesis] Dokuz Eylül University Department of Educational Sciences.

Anıl, D. (2008). The analysis of factors affecting the mathematical success of Turkish students in the PISA 2006 evaluation program with structural equation modeling. *American-Eurasian Journal of Scientific Research*, *3*(2), 222-227.

Aydın, S. (2015). *Data mining and an application on Anadolu University distance education system* [Doctoral dissertation]. Anadolu University.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Cai, Y.D., & Chou, K.C. (2003). Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochemical and Biophysical Research Communications*, *305*(2), 407-411. https://doi.org/10.1016/S0006-291X(03)00775-7

Çalış, A., Kayapınar, S., & Çetinyokuş, T. (2014). An application on computer and internet security with decision tree algorithms in data mining. *Journal of Industrial Engineering*, *25*(3), 2-19. https://dergipark.org.tr/en/pub/endustrimuhendisligi/issue/46771/586362

Erdil, Z. (2010). Relationship of academic achievement and early intervention programs for children who are at socio-economical risk. Journal of *Hacettepe University Faculty of Nursing*, *17*(1), 72-78. https://dergipark.org.tr/en/pub/hunhemsire/issue/7840/103271

Gelbal, S. (2010). The effect of socio-economic status of eighth grade students on their achievement in Turkish. *Education and Science*, *33*(150). http://eb.ted.org.tr/index.php/EB/ article/view/626

Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, *56*(9), 1099-1108. https://doi.org/10.1057/palgrave.jors.2601976

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

Özbay, Ö. (2015). The current status of distance education in the world and Turkey. *The Journal of International Educational Sciences*, *2*(5), 376-394.

Özer, Y., & Anıl, D. (2011). Examining the factors affecting students' science and mathematics achievement with the structural equation modeling. *Hacettepe University - Journal of Education*, 41, 313-324. https://app.trdizin.gov.tr/makale/TVRJMU1qa3lNZz09

Rizvi, S., Rienties, B., & Khoja, S.A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, 137, 32-47. https://doi.org/10.1016/j.compedu.2019.04.001

Roiger, R.J. (2017). *Data mining: a tutorial-based primer*. Chapman and Hall/CRC.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

Şahin, M. (2018). *Risk assessment in car insurance using decision trees and artificial neural networks* [Doctoral dissertation]. Yıldız Technical University Department of Statistics.

Witten, I.H. & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*(3), 249-262.