Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 2023; 20(2), s. 459-483.

Van Yüzüncü Yıl University Journal of Education, 2023; 20(2), p. 459-483. DOI:10.33711/yyuefd.1221541

Investigation of Variables Affecting PISA Reading Comprehension Achievement Levels of Countries with Different Levels of Achievement with CRT and RF Methods

Yusuf KASAP[1], Nuri DOĞAN[2], Cem KOÇAK[3]

Abstract: The aim of this research is to determine the important variables that predict the PISA 2018 reading comprehension achievement score of countries with different achievement levels, using 34 independent variables obtained from the student questionnaire given to the students who participated in PISA in 2018. For this purpose, 79 countries that participated PISA were ranked according to their success percentages then, these countries were sorted into lower, middle and upper group countries. A sample of lower, middle and upper group countries was formed then, three countries were selected from each of the lower group, middle group and upper group countries and a sample of lower, middle and upper group countries was formed. Data mining analyzes were carried out on the samples obtained by using the Classification and Regression Tree and Random Forest methods. It has been observed that the number of important variables that predict reading comprehension success can be reduced from 34 to three to eight. Like this; Data mining classification prediction models, which can predict the success level of PISA, were obtained by using a small number of variables. It has been determined that the models obtained have an acceptable level of predictive performance in predicting success in three categories (low, medium-high). The most important predictor variables obtained from the models are information and communication technologies resources, perception of reading difficulty, professional status expected from the student, perception of difficulty in the PISA test, reading pleasure, weekly test language learning time, disciplinary climate, socio-economic status index.

Keywords: Machine learning, data mining, prediction performance, PISA, reading comprehension skill, decision trees

CRT ve RF Yöntemleri ile Farklı Başarı Düzeyine Sahip Ülkelerin PISA Okuduğunu Anlama Başarı Düzeylerini Etkileyen Değişkenlerin İncelenmesi

Öz: Bu araştırmanın amacı, 2018 yılında PISA'ya katılan öğrencilere sınavla birlikte verilen öğrenci anketinden elde edilen 34 bağımsız değişkeni kullanarak, farklı başarı düzeyine sahip ülkelerin PISA 2018 okuduğunu anlama başarı puanını yordayan önemli değişkenleri belirlemektir. Bu amaç için PISA'ya giren

Geliş tarihi/Received: 19.12.2022 Kabul Tarihi/Accepted: 01.06.2023 Makale Türü: Araştırma Makalesi

*Bu çalışma birinci yazarın doktora tezinden üretilmiştir.

[1] Dr., Hacettepe Üniversitesi Eğitim Bilimleri Bölümü., yusufkasap5806@hotmail.com, 0000 0002 5114 1175

[2] Prof. Dr., Hacettepe Üniversitesi Eğitim Bilimleri Bölümü., nuridogan2004@gmail.com, 0000 0001 6274 2016

[3] Prof. Dr., Hitit Üniversitesi Maliye Bölümü cemkocak@hotmail.com, 0000 0002 7339 7438

Atıf için/To cite: Kasap, Y., Doğan N. & Koçak, C. (2023). CRT ve RF yöntemleri ile farklı başarı düzeyine sahip ülkelerin PISA başarı düzeylerini etkileyen değişkenlerin incelenmesi. *Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, 20(2), 459-483. https://doi.org/10.33711/yyuefd.1221541

79 ülke başarı yüzdeliklerine göre sıralanmış ve bu sıralamaya göre bu ülkeler alt, orta ve üst grup ülkeler olarak ayrılmıştır. Daha sonra alt grup, orta grup ve üst grup ülkelerin her birinden üçer ülke seçilerek alt, orta ve üst grup ülkeler örneklemi oluşturulmuştur. Elde edilen örneklemler üzerinde Sınıflama ve Regresyon Ağacı ve Rastgele Orman yöntemleri kullanılarak veri madenciliği analizleri gerçekleştirilmiştir. Yapılan uygulamalarda okuduğunu anlama başarısını yordayan önemli değişkenlerin sayısının 34'ten üç ile sekiz arasında bir sayıya indirgenebildiği görülmüştür. Böylece; az sayıda değişken kullanılarak PISA başarı düzeyini yordayabilen veri madenciliği sınıflama tahmin modelleri elde edilmiştir. Elde edilen modellerin başarıyı üç kategorili (düşük, orta yüksek) yordama da kabul edilebilir düzeyde tahmin performansına sahip oldukları saptanmıştır. Modellerden elde edilen tahmin edici değişkenlerden en önemli olanları bilgi iletişim teknolojileri kaynakları, okuma zorluk algısı, öğrenciden beklenen mesleki statü, PISA testinin zorluk algısı, okuma keyfi, haftalık test dili öğrenme süresi, disiplin iklimi, sosyo-ekonomik durum indeksi biçimindedir.

Anahtar kelimeler: Makine öğrenme, veri madenciliği, tahmin performansı, PISA, okuduğunu anlama becerisi, karar ağaçları

Introduction

The amount of data obtained from individuals is increasing exponentially as information technologies find a wide place in many areas of our lives. Raw data from research alone is worthless. Thanks to the development of computer software, processing these raw data for a specific purpose, revealing patterns and relationships between the data obtained, predicting the current situation and making predictions for the future reveal meaningful results for both researchers and decision-makers. The increase in the data obtained and the existence of a wide range of variables related to different qualities of individuals necessitates the use of advanced statistical methods such as data mining in the analysis process. Data mining is a method that allows discovering important information hidden in the data and making predictions about the future by identifying relationships, patterns, correlations and rules in large amounts of data (Kayri, 2008).

More detailed statistical results can be obtained by using the techniques within the scope of data mining in the analysis of the data obtained from the researches. It is tried to predict the future behavior of individuals by determining the behavior patterns of individuals from a large amount of raw data obtained from individuals using data mining (Sieber, 2008). Data mining is used in many disciplines such as trade, medicine, banking, engineering, stock market, education (Savaş vd., 2012).

In recent years, the tendency to collect multivariate data based on the results obtained from both international and national exams has become widespread in education. One of these exams is the International Student Program (PISA). By using PISA data, with multivariate data mining models, it can be determined which independent variables predict success or are associated with success.

PISA is an application that includes Mathematics literacy, Science literacy and reading comprehension tests, which are held every 3 years. In each PISA application, weight is given to one of these three tests. The main area of focus in PISA 2018 is reading comprehension. The main purposes of this application: It was determined as measuring the skills of understanding, using, evaluating texts, thinking about texts and interacting with texts (OECD, 2019).

Reading comprehension in PISA; It is the individual's making real or figurative meanings from the text. It can be both understanding the words and expressing the main idea of a text written for narrative purposes. Using the information in the text in PISA; refers to the use of information obtained from a text for a purpose, to support a thought. Most readings are of this type. While reflecting individual's thoughts, the individual establishes a connection between the text individual reads and individual's own thoughts and experiences. At this stage, the individual brings a different approach to the individual's own life or comes to a decision regarding the accuracy and reality of the information in the text. Thus, the individual needs to read the text and determine whether it is suitable for the purpose of the text (MEB, 2010). The concept of literacy mentioned in PISA is expressed as high-level literacy, which provides the acquisition of basic knowledge and skills for participation in production relations in modern society by focusing on the knowledge and skills needed in real-life situations (Coulombe, Trembly ve Marchand, 2004).

Considering variables such as science literacy, mathematical literacy and reading comprehension measured by PISA, and data size, data mining may be an appropriate method for better interpretation of PISA data. Therefore, it is seen that data mining studies on PISA data have intensified in recent years (Abad and Gamazo, 2020; Aksu 2019; Bezek Güre, Kayri and Erdoğan, 2020).

In the literature, there are many national and international studies using data mining methods to predict academic success (Abad & Lopez, 2016; Gamazo & Abad, 2020; Abdous, He & Yen, 2012; Aksoy, 2014; Özarslan, 2014; Yu & et al., 2012; Yung & et al., 2012). Of these studies, Yu et al. (2012) and Gamazo & Abad (2020) studies are aimed at predicting the success level of PISA.

In these studies; using data mining methods, it is aimed to determine the variables that predict success or are related to success, some of these studies are aimed at determining the variables that predict the level of PISA success (Abad & Lopez, 2016; Aksu, 2016; Gamazo & Abad, 2020; Kiray vd., 2015; Yu & et al., 2012). E.g; Abad and Lopez (2016) used the C4.5 decision tree classification algorithm to determine the factors that predict academic success in their study. As another example, Yu et al. (2012) used the Logistic regression classification algorithm to predict students' PISA science literacy success in their study and determined the important factors predicting success according to the odds ratios in the model obtained.

In the studies on the determination of the important factors that predict success in the literature; In general, it is seen that R, Jamovi, Jasp, Orange, Matlab, Modeler and Weka programs are used.

With this; Bezek Güre et al. (2020) determined the important variables that can predict success at the highest level by using SPSS Modeler and Matlab programs. The study of Bezek Güre et al. (2020) is different from other studies in the literature. In almost all studies in the literature, while determining the factors affecting success, decisions are made according to the outputs obtained from data mining algorithms, Bezek Güre et al. In the approach of (2020), the variables that make the model performance the highest are seen as the best predictors. With this; Yue (2021) used a data mining performance-based approach in his study to identify important variables that can predict the performance of college students.

Bezek Güre et al. it can be said that the approaches of (2020) and Yue (2021) are more practical than the approach of determining important variables according to the outputs obtained from the data mining models used in the majority of the studies in the literature. Because; In determining the important variables according to the outputs of the models used, the performance of the data mining prediction model is not taken into account. Like this; Important variables can be determined by using models whose performance is not high enough. However, Bezek Güre et al. (2020) and Yue (2021)'s approach to detecting important variables requires making decisions according to a data mining prediction model with high performance.

On the other hand, when the studies in the literature are examined, studies with homogeneous groups limited to a single country are intense. There was no study on whether the variables predicting success in different success groups changed by using data mining methods. In fact, it is important to investigate whether the results obtained from data mining methods change as the data (sample) changes. Many methods can be used within the scope of data mining to determine the variables that predict success, but when the performance level of the used model changes and the success level of the studied sample changes, it is necessary to compare the results obtained with different data mining estimation methods to see what kind of result is encountered. In this study, Classification and Regression Tree and Random Forest methods were used.

In the literature, it is seen that the samples used in the studies of determining the important variables that predict the PISA success score are narrow. E.g; Aksu and Güzeller (2016) and Bezek Güre et al. (2020) used only the PISA Turkey sample in their study. Studies have focused on the variables that predict success or performance, but these studies are similar to each other, accepting science, mathematics or reading scores as dependent variables, and aiming to reveal the factors affecting success in these three areas. However, different levels of success were not taken into account in these studies. However, due to the idea that different variables are effective in different sub-success groups, it is possible to verify this claim by working in different achievement groups. In PISA, in which many countries participate, the success levels of countries are quite different from each other.

Therefore; In this study; By determining the variables affecting success for low, medium and high-achievement groups, it can be ensured that the countries in the relevant groups give more importance to the factors that positively affect success while directing their education policies. On the other hand, for countries with different success levels; By determining the same and different factors that affect success, it can be ensured that countries with low achievement levels can make more accurate decisions in increasing their success. In this respect, it is thought that the research will contribute to the literature.

The Aim of the Research

The main purpose of this research is to determine the important predictors that affect the reading comprehension success of countries with different achievement levels (low, medium and high achievement groups) by using the 34 independent variables found in the PISA 2018 student questionnaire.

The other aim of the study is; The aim is to show whether high-performance classification models can be obtained by using a small number of more important independent variables. Therefore, It is aimed that the study will be one of the few studies to reduce the number of variables by choosing the more important independent variables that make the performance of the data mining method the highest. In this study, answers to the following questions are looked for:

1. Using the student characteristics measured in the PISA 2018 student questionnaire, what are the important variables that predict the success of the countries in the samples of high, medium and low achievement countries by using the Classification and Regression Tree and Random Forest methods to predict three-category reading comprehension achievement?

2. What is the classification and estimation performance of the Classification and Regression Tree and Random Forest methods used to predict three-category reading comprehension achievement using the student characteristics measured in the PISA 2018 student questionnaire?

Method

Research Model

In the study, using scales measuring students' affective characteristics and questionnaires measuring socio-demographic characteristics; In line with the determination of the important variables that can predict the PISA reading comprehension achievement level in three categories (low achievement-medium achievement-high achievement) for the country samples selected from the low, medium and high achievement groups, the types of research are quantitative research designs descriptive design, comparative descriptive design, relational design. (Büyüköztçürk, Kılıç-Çakmak, Akgün, Karadeniz, & Demirel, 2018).

Population and Sample

For the purpose of the study, countries; Purposive sampling method was used as the PISA reading score was determined according to the lower (weak), medium, and upper (high) achievement groups, taking into account their order. Because purposeful sampling is the selection of a sample rich in information in line with the purpose of the study in order to conduct in-depth research (Büyüköztürk et al., 2018). In addition, while selecting countries from the lower, middle and upper groups for the study sample, care was taken to ensure that the rate of missing data is low and that it creates a pattern to be distributed in different percentiles. In addition, while selecting the countries to create the lower, middle and upper group samples, attention was paid to ensure that the missing data rate is low and that it creates a pattern to be distributed in different percentiles. Thus, the information provided by the selected samples for each group was enriched. According to the purpose of this research, it is aimed to represent countries with different success levels in 79 countries by choosing three countries for the lower, middle and upper groups from 79 countries that participated in the PISA 2018 exam. In line with this target, nine countries; It has been determined by taking into account the percentage rankings of 79 countries according to their success scores. The percentages of countries selected from the last third with 68% and 100% success rates are 91 for Indonesia, 82 for Saudi Arabia and 73 for Colombia. Like this; The sample consisting of Indonesia, Saudi Arabia and Colombia countries was named as the sample of countries in the sub-successful group, representing the most unsuccessful countries in the last 33%. The percentile order of the countries selected from the 33% and 67% success rate is; It is 57 for Serbia, 50 for Turkey and 41 for Hungary. Like this; The sample consisting of Serbia, Turkey and Hungary was named as the countries in the medium success group, representing the moderately successful countries in the 33% to 67% slice range. Finally; The percentages of the countries selected from the top third of the achievement tranches are 27 for Slovenia, 16 for the USA and nine for Finland. Like this; The sample of Slovenia, USA and Finland was named as the top achievement countries sample, representing the most successful countries in the top 33%. In this case, the sample size for low-achieving countries was 25378, for medium-achieving countries 18523, and for high-achieving countries 16567.

Obtaining Data

The data used in the research process were obtained from the database opened for sharing in 2020, using the address http://www.oecd.org/pisa/data/2018database/. A total of 34 independent variables belonging to 2018 PISA data and a dependent variable, which is the average of 10 reasonable values corresponding to the reading comprehension success level, were included in the research. When the data were examined, no missing data was found regarding the dependent variable. However, after excluding individuals who did not respond to 34 independent variables from the analysis, a small number of outliers were excluded from the data. In addition, missing data was found in each of all independent variables. Since these missing data were randomly distributed, the assignment was made using the multiple assignment method. In many applications of multiple assignment, accurate assignments are made (He, 2006). In this study, the multiple assignment process was performed by assigning five times based on the linear regression model for continuous variables.

Data Collection Tools

The data in the research were obtained from the test in the PISA reading comprehension exam and the student questionnaire applied in 2018. In order to obtain the dependent variable we will use, first of all, the reading comprehension success score was obtained as a continuous quantitative variable and named as PVREAD. Here, PVREAD refers to the reasonable reading comprehension achievement score. The PVREAD variable is the average of 10 reasonable point values (PV1,PV2,…,PV10) obtained from the PISA 2018 reading comprehension achievement test and expressing the range of abilities that students can have (Wu & Adams, 2002). Then, in the process of classifying the PVREAD score, the reading skills proficiency level table in the PISA 2018 Turkey preliminary report was used (MEB, 2019). Threshold values in this table in the preliminary report of MEB (2019) were revised and the number of PISA 2018 reading proficiency levels was reduced from 8 to 6, and the thresholds and category information of PISA 2018 reading comprehension proficiency levels were obtained (Table 1). This reduction was achieved by combining the 1a, 1b, 1c levels at the 1st level specified in the PISA 2018 report as a single level and determining the number of levels as 6.

Table 1
Categories of PISA 2018 Reading Comprehension Proficiency Levels

proficiency levels	Score (x)	Category
Level 1	0<X<407	Low
Level 2	407<X<480	Low
Level 3	480<X<553	Medium
Level 4	553<X<626	Medium
Level 5	626<X<698	High
Level 6	698<X<1000	High

According to Table 1, three-category achievement variable; The PVREAD value was obtained by classifying it as low success between 0-479,999, medium success between 480-625,999 and high success between 626-1000.

Analysis of Data

In the analysis of the data, SPSS Modeler and WEKA programs were used as a basis and in a systematic way, respectively. Excel and SPSS programs; SPSS Modeler and WEKA were used as utilities while transferring data and calculating some statistics. In the SPSS program, multiple assignment method was also used to fill in the missing data. W_FSTUWT (Final Trimmed Nonresponse Adjusted Student) weight variable, which shows the weight of each student in the data set, was included in the analysis because the weight variable should be used due to the nature of PISA data (Arıkan et al., 2020) while determining the important variables in the SPSS Modeler program. In this study, predictor importance values were taken into account in the determination of independent variables with high predictive importance in the analysis process of the data.
In the analyzes performed with CRT and RF methods with 34 independent variables, the first 10 variables that best-predicted success were determined by taking into account the predictive importance values calculated by each method. Then, in the Weka program using important independent variables; Performance criteria were calculated with all data, 10-fold cross-validation data and test data. It has been determined that these calculated performance criteria do not have a significant difference from the calculated performance criteria when 34 independent variables are used, and thus, it has been shown that high-performance models can be established with fewer variables in predicting success. These analyzes were made for the sample of low, medium and high-achievement countries, and the results were interpreted by making a detailed comparison.

Classification and Regression Tree Method

It is one of the most widely used algorithms when creating decision trees. This algorithm was developed by Breiman, Friedman, Olshen, and Stone in 1984. In the classification and regression tree (CRT) algorithm, a node is divided according to a certain criterion. In the division phase, first the values with all the features are taken into account, and after all the matches are completed, the selection process is carried out in the form of yes-no with binary branching (Özkan, 2008). The classification and regression tree algorithm can be used when the dependent variable is categorical or continuous data. In this respect, the CRT algorithm can be considered as an algorithm that includes multiple regression analysis when the dependent variable is continuous and logistic regression analysis when it is categorical (Güner, 2014). This algorithm should also calculate from which point the determined node will be divided into two, together with determining which node to be root or node (Silahtaroğlu, 2016).

In the process of using the CRT algorithm in continuous dependent variables, it uses convergence-based numerical methods to determine the partitioning point. When the dependent variable is categorical data, the class in which the object is located is estimated, while if it is continuous, the numerical value of the class in which the object is located is estimated (Köse, 2018). Contrary to other algorithms, in the CRT algorithm, a large tree is created initially, and then the tree is pruned down to minimize the misclassification estimation error (Loh, 2011). In this algorithm that produces a binary tree, the sub-options of each feature should be divided into two groups. The section to which each option will be directed is determined by the gini measure. In addition, when the dependent variable is categorical data, the method created using the gini measure, which is the branching criterion, is expressed as a classification tree, while the method created by using the sum of squares of error for the branching criterion for continuous data is expressed as a regression tree (Altunkaynak, 2019).

Branching based on the gini measure is made according to the gini separation index calculated for the variables (Loh, 2011). Since only two branches are made for each of the nodes in the Classification and Regression Tree, each of the variables should have two categories (Altunkaynak, 2019).

p(j/p) returns the relative probability of class j at node t. The formula for the Gini coefficient is as follows (Akar, Güngör, & Akar, 2010).

$$Gini = 1 - \sum_j [p(j/t)]^2$$

In the branching that occurs depending on the Twoing Criteria, all possible binary decompositions of each of the independent variables are considered. This criterion divides the relevant feature into left and right. Then, probability values are calculated for each category (repeating discrete value) from the values found in each branch (Köse, 2018). In the CRT algorithm, the records are divided into subclasses according to the formula below, taking into account all possible values (Larose, 2005).

$$\emptyset(\mathrm{D_i}) = 2\mathrm{P_L}\mathrm{P_R}\sum_{j=1}^{k}\left|\mathrm{P}\left(\mathrm{Y_j}\backslash \mathrm{L}\right) - \mathrm{P}(\mathrm{Y_j}\backslash \mathrm{R})\right|$$

According to the formula, the variable that makes the $\emptyset(\mathrm{D_i})$ value the largest at the $\mathrm{D_i}$ node and has the most impact on the category is selected for branching (Altunkaynak, 2019).

Random Forest Method

Random Forest (RF) algorithm, which is a simple classification algorithm, is an algorithm consisting of a combination of a certain number of trees with the highest accuracy and independence. Each of the decision trees branches according to the variables in the data set (Breiman, 2001). In this algorithm, it is aimed to make an effective prediction by combining the predictions made with each of the decision trees (Atasever, 2011).

The basic logic of decision forests is to create a model from a collection of algorithms that perform higher than the performance of an algorithm. The ensemble learning algorithm is like a decision made by a person who consults with the people around her and considers different opinions before making a critical decision (Polikar, 2006).

Since this method is based on CRT, it can be used in classification and regression processes. The main problem in models in which the decision tree is used is the overfitting due to insufficient data (Liao, Ju, & Zou, 2016). The RF algorithm is designed to learn from subsets of the dataset to avoid overfitting, taking into account a certain number of generated CRTs. Therefore, it is a resistant method against overfitting (Bhalla, 2014). The working process of the RF algorithm. During the operation of this algorithm, the training data is obtained from the studied data set using the repetitive sampling method (Breiman, 2004).

Sampling is taken as much as the number of decision trees that are intended to be created with the repetitive sampling method, and the training and test data set is divided for each sample (Akman, 2010). While creating the model with the RF algorithm, 2/3 of the data set is divided as training data (inBag) and 1/3 as test data (Out-Of-Bag), provided that the selected data is randomly selected by putting it back in the next step (Bhalla, 2014). Unpruned classification and regression trees are grown for each preloading examples. For this, instead of choosing the best division provider among all variables present in the learning data set (inBag), initially m pieces examples are chosen randomly and then the one that will provide the best division is determined (Breiman, 2001).

Each of the decision trees to be found in the forest is created with the CRT algorithm. For this, after determining the best branching criterion with the gini measure, the node is divided into two branches (Akman, 2010). When each tree is reduced based on regression, the test vector x is assigned to the mean value of y at the node it is located. The predicted value in the classification process is the class that receives the most votes (mod) in the forest (Breiman, 2004). In other words, predictions are made according to the new data set created by combining the prediction values of n trees. In this unification process, the mean for regression and the highest majority (mod) for classification are used (Liaw & Wiener, 2002).

To make an estimate at a new point x;

$\hat{f}(x) = \frac{1}{J}\sum_{j=1}^{J}\hat{h}_j(x)$, used for regression.

$\hat{f}(x) = \mathrm{argmax}_y \sum_{j=1}^{J} I(\hat{h}_j(x) = y)$, used for classification.

Here, $\hat{h}_j(x)$ j. is the estimation of the response of the predictor variable x in the tree.

Performance Criteria

Finally, absolute performance criteria for comparing results; Percentage of Correct Classification (PCC), Root Mean Square of Error (RMSE) and Mean Absolute Error (MAE) and relative performance criteria are; Kappa (κ) coefficient, Root Relative Square Error (RRSE) and Relative Absolute Error (RAE) (Tabachnick and Fidel, 2007; Field, 2009) were calculated separately for each data mining method and the results of the methods were compared with each other.

$$PCC = \frac{DP+DN}{DP+DN+YP+YN}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{N}}$$

$$MAE = \frac{\sum_{i=1}^{N}|P_i - O_i|}{N}$$

TP indicates the number of true positive, TN true negative, FP false positive, FN false negative. P_i represents the predicted values, O_i the observed values. RMSE and MAE should be close to zero. With relative performance criteria; Kappa (κ) coefficient, Root Relative Square Error (RRSE) and Relative Absolute Error (RAE) calculation formulas are given.

$$Kappa(\kappa) = \frac{Gözlenen\ doğruluk - Beklenen\ doğruluk}{1 - Beklenen\ doğruluk}$$

$$RRSE = \sqrt{\frac{\sum_{j=1}^{n}(P_{ij} - O_j)^2}{\sum_{j=1}^{n}(O_j - \bar{O})^2}}$$

$$RAE = \frac{\sum_{j=1}^{n}|P_j - O_j|}{\sum_{j=1}^{n}|O_j - \bar{O}|}$$

in their formulas; P_j and P_{ij} represent the predicted values, O_j the observed values, and $\bar{O}$ the mean of the observed values.

Results

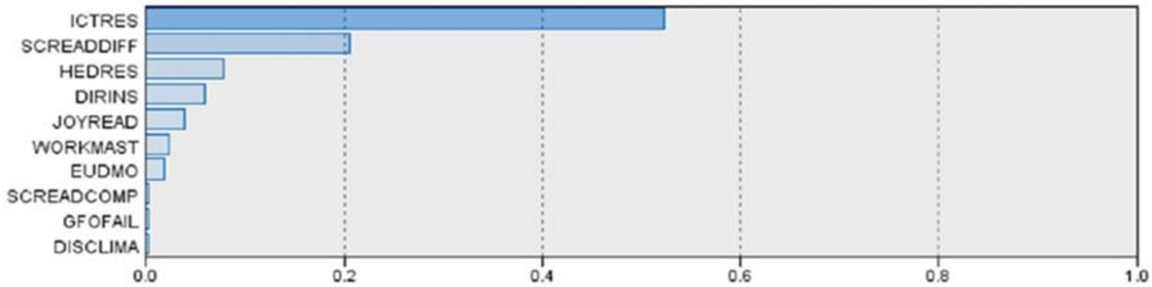
Findings Related to the First Sub-Problem

Classification and Regression Tree (CRT) Method

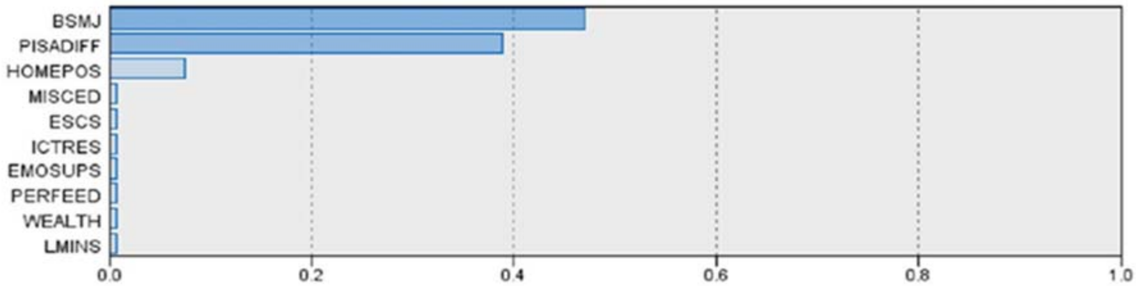
The graphs of the predictive significance percentages for the top 10 important variables obtained from the three-category CRT application are given in Figure 1.

Figure 1

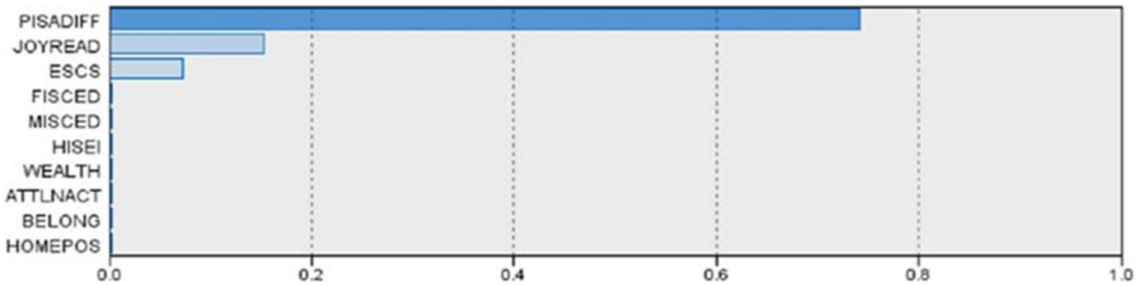
Graphs of 10 important variables selected with the CRT model in the case of three-category reading comprehension achievement



Sample of lower-group



Sample of middle-group



Sample of upper-group

Less than 10 independent variables, which are called "the best predictor of PISA reading comprehension achievement", obtained from the regression tree that emerged as a result of the analysis made in the modeler program, were determined and presented in Table 2. When Table 2 is examined, it is seen that for each sample group, the order of importance is given in the table and the important covariates are indicated in bold.

Table 2
Significant and Common Variables Selected by Classification and Regression Tree Model

Sample of lower-group	Sample of middle-group	Sample of upper-group
-	2. PISADIFF	1. PISADIFF
5. JOYREAD	-	2. JOYREAD
1. ICTRES	1. BSMJ	3. ESCS
2. SCREADDIFF	3. HOMEPOS	-
3. HEDRES	-	-
4. DIRINS	-	-
6. WORKMAST	-	-
7. EUDMO	-	-

1.2.3, order of importance of important variables

When Table 2 is examined, the variable PISADIFF (perception of the difficulty of the PISA test) was found to be an important covariate in predicting the three-category PISA reading comprehension achievement for each of the other sample groups, except for the lower-group. Here, in terms of predicting success, the PISADIFF variable has the first place in the sample of upper-group countries and the second place in the sample of middle-group countries; For these countries, it can be said that the difficulty perception variable of the PISA test is an important predictor that affects the success of PISA reading comprehension.

JOYREAD (reading pleasure) was determined as the important covariate predicting PISA reading comprehension achievement for the sample of lower and upper group countries. In terms of predicting success, this variable is in the fifth place in the sample of lower group countries and in the second place in the sample of upper group countries; For these countries, it can be said that the variable of reading pleasure is an effective predictor of PISA reading comprehension achievement.

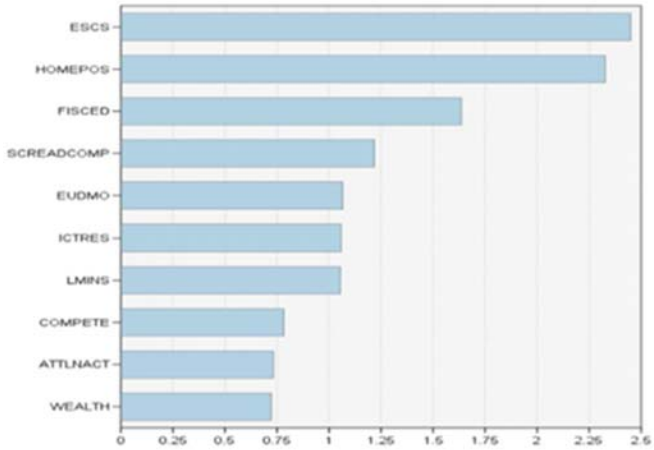
On the other hand, PISA predicts reading comprehension success; In the sample of lower group countries, ICTRES (information and communication technologies resources) first, SCREADDIFF (self-reading difficulty perception) second, HEDRES (home education resources) third, DIRINS (teacher's training orientation) fourth, WORKMAST (motivation for advanced tasks) sixth, EUDMO (meaning of life) are the seventh-ranked variables. In the sample of middle group countries, BSMJ (expected professional status of the student) is the first and HOMEPOS (educational items at home) is the third most important variable. In the sample of upper-group countries, ESCS (socio-economic status index) is the third most important variable.

Random Forest (RF) Method

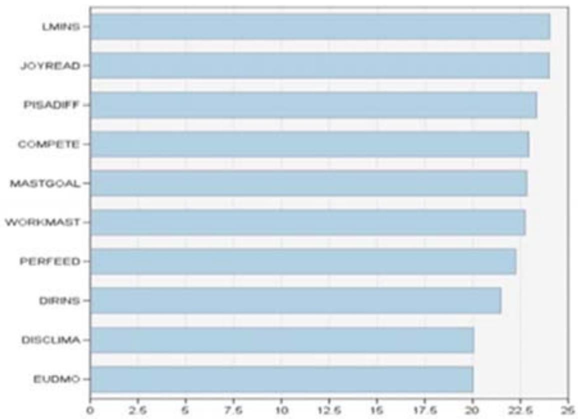
The graphs of the predictive significance percentages for the top 10 important variables obtained from the three-category RF application are given in Figure 2.

Figure 2

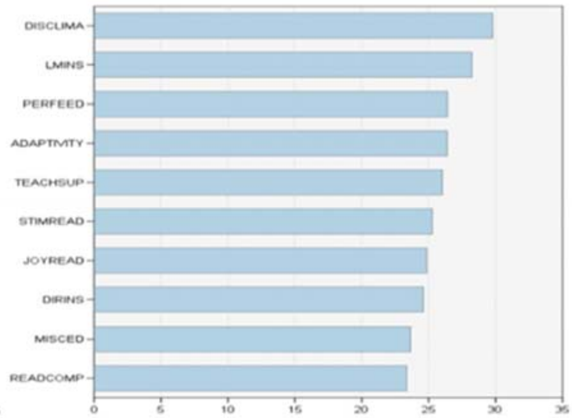
Graphs of 10 important variables selected with the RF model when reading comprehension achievement is in three categories



Sample of upper-group



Sample of lower-group



Sample of middle-group

In Figure 2, with the reduced model, by removing the variables with the lowest predictive value from the first 10 predictor independent variables that predict the PISA reading comprehension achievement variable selected for each group sample with the RF method, the classification performance in the WEKA program is reduced by decreasing the variable so that it is close to the performance obtained with 10 variables. After adjustment, less than 10 independent variables named as "the best predictor of PISA reading comprehension achievement" were determined and presented in Table 3. When Table 3 is examined, it is seen that the order of importance for each sample group is given in the table and the important common variables are indicated in bold.

Table 3

Significant and Common Variables Selected with the Random Forest Model

Sample of lower-group	Sample of middle-group	Sample of upper-group
1. LMINS	2. LMINS	7. LMINS
2. JOYREAD	7. JOYREAD	-
3. PISADIFF	1. DISCLIMA	1. ESCS
4.COMPETE	3. PERFEED	2. HOMEPOS
5. MASTGOAL	4. ADAPTIVITY	3. FISCED
6. WORKMAST	5. TEACHSUP	4.SCREADCOMP
-	6. STIMREAD	5. EUDMO
-	8.DIRINS	6. ICTRES
-	-	8.COMPETE

1.2.3, order of importance of important variables

When Table 3 is examined, the LMINS (weekly test language learning time) variable was found to be an important covariate in predicting the three-category PISA reading comprehension achievement in the sample of the lower group, middle group and upper group countries. Here, in terms of predicting success, the LMINS variable is in the first place in the lower group, in the second place in the middle group, and in the seventh place in the upper group; For these countries, it can be said that the weekly test language learning time variable is an important predictor affecting the PISA reading comprehension achievement.

JOYREAD (reading pleasure) was determined as the important covariate predicting PISA reading comprehension achievement for the lower group and middle group. In terms of predicting success, this variable is in the second place in the sample of lower group countries and in the seventh place in the sample of middle group countries; For these countries, it can be said that the variable of reading pleasure is an important predictor of PISA reading comprehension achievement.

In addition, PISA predicts reading comprehension success; In the sample of lower group countries, PISADIFF is third, COMPETE (attitudes to competition) fourth, MASTGOAL (learning goals) fifth, and WORKMAST (motivation for further tasks) sixth important predictors. In the sample of middle group countries, DISCLIMA (disciplinary climate) ranks first, PERFEED (teacher feedback) third, ADAPTIVITY (teaching adaptation) fourth, TEACHSUP (Teacher support) fifth, STIMREAD (Teacher's encouragement to read) sixth, and DIRINS (Teacher's direction of instruction) is the eighth-ranked important predictor. in the sample of upper group countries, ESCS (socio-economic status index) ranks first, HOMEPOS (educational items at home) ranks second, FISCED (paternal education level) ranks third, SCREADCOMP (self-reading efficacy perception) ranks fourth, EUDMO (meaning of life) ranks fifth, ICTRES (information and communication technologies resources) ranked sixth and COMPETE (attitudes to competition) ranked eighth important predictor.

Findings Regarding the Second Sub-Problem

Performance calculations were made in the WEKA program using CRT and RF methods, using 34 independent variables included in the study and the first 10 independent variables that predicted the three-category PISA reading comprehension success, and less than 10 important variables selected from these 10 independent variables. The results obtained are shown in Table 4 and Table 5.

Table 4

CRT Model Performances for Tree Category Success Condition

			Absolute Performances			Relative Performances		
	Variable	Data	PCC	MAE	RMSE	Kappa	RAE	RRSE
Lower Group Sample	All Variables (34)	All data	84.037	0.164	0.286	0.315	81.815	90.458
		Validity data	83.328	0.165	0.290	0.291	82.585	91.839
		Test data	83.277	0.165	0.290	0.272	82.126	91.559
	Selected variables (10)	All data	83.793	0.167	0.289	0.281	83.551	91.412
		Validity data	82.476	0.173	0.296	0.212	86.481	93.685
		Test data	82.570	0.175	0.298	0.208	87.414	93.855
	Important variables (7)	All data	82.693	0.176	0.296	0.199	87.886	93.754
		Validity data	82.465	0.174	0.296	0.200	86.821	93.759
		Test data	82.466	0.175	0.297	0.195	87.508	93.780
Middle Group Sample	All Variables (34)	All data	71.813	0.268	0.366	0.428	78.395	88.543
		Validity data	68.288	0.279	0.378	0.354	81.669	91.560
		Test data	68.561	0.277	0.377	0.355	81.134	91.306
	Selected variables (10)	All data	71.494	0.271	0.368	0.423	79.185	88.988
		Validity data	67.564	0.282	0.381	0.343	82.638	92.107
		Test data	67.497	0.282	0.382	0.340	82.390	92.366
	Important variables (3)	All data	68.660	0.286	0.378	0.363	83.681	91.480
		Validity data	67.543	0.288	0.381	0.342	84.243	92.256
		Test data	67.481	0.291	0.383	0.344	85.015	92.718
Upper Group Sample	All Variables (34)	All data	68.274	0.301	0.388	0.427	77.864	88.241
		Validity data	62.709	0.323	0.407	0.322	83.693	92.602
		Test data	62.595	0.328	0.408	0.322	84.844	92.850
	Selected variables (10)	All data	62.896	0.328	0.405	0.328	84.974	92.182
		Validity data	60.922	0.331	0.410	0.293	85.710	93.236
		Test data	61.299	0.335	0.410	0.297	86.738	93.446
	Important variables (3)	All data	62.515	0.328	0.405	0.319	84.943	92.166
		Validity data	61.447	0.330	0.408	0.300	85.281	92.891
		Test data	61.157	0.331	0.408	0.293	85.781	92.904

PCC: Percentage of correct classification, MAE: Mean absolute error, RMSE: Root of square mean squared error, RAE: Relative absolute error, RRSE: Root Relative Square Error

In Table 4, according to the analysis results for all data, cross-validation and test data for the sample of the lower group, middle group, and upper group countries, when the first 10 predictive variables and less than 10 important variables are used, the percentage of correct classification (PCC) for the lower group, it is seen that there is no significant change in the sample of middle group countries. For example, while the percentage of success is 71,813 for 34 variables in the sample of middle group countries, it is 71,494 for 10 variables and 68,660 for 3 variables. Here, the decrease in the number of variables does not cause a major change for PCC. The conclusion to be drawn from this is that, for example, for the sample of middle group countries, the PCC obtained from 34 variables can be approached with 3 variables. It is seen that similar results are the same for the sample of lower group countries. However, this situation is not the same for the upper group countries sample. The decrease in the number of variables in the sample of upper group countries caused a significant decrease in PCC. For example, in the sample of upper group countries, the success rate for 34 variables is 68,274, while the success rate for 10 variables is 62,896, and 61,157 for 3 variables. Here, the decrease in the number of variables causes a significant change for PCC.

On the other hand, the order of the reduced model established by CRT analysis with important variables, from the most successful to the most unsuccessful in terms of PCC, is listed as the lower group, the middle group, and the upper group. In addition, in the sample of lower group countries, the Kappa coefficient is noteworthy because it is between 0.20 and 0.40 when the number of variables is between 34 and 10, but low because it is below 0.20 when the number of variables is 7, RAE (relative absolute error) and RRSE (Root Relative Square Error) values were high. The reason for this can be thought that both the Lower group students are not evenly distributed to the success categories and the decrease in the number of variables weakens the model. On the other hand, it is seen that the Kappa coefficient is significant in the sample of middle group and upper group countries and it is higher than the sample of lower group countries.

Table 5
RF Model Performances for Tree Category Success Condition

			Absolute Performances			Relative Performances		
	Variable	Data	PCC	MAE	RMSE	Kappa	RAE	RRSE
Lower Group Sample	All Variables (34)	All data	100	0.057	0.101	1.000	28.766	32.046
		Validity data	84.336	0.157	0.271	0.316	78.494	85.856
		Test data	84.216	0.158	0.272	0.313	78.774	85.666
	Selected variables (10)	All data	99.976	0.061	0.108	1.000	30.883	34.146
		Validity data	82.973	0.168	0.287	0.242	83.765	90.726
		Test data	83.068	0.169	0.287	0.241	84.371	90.649
	Important variables (6)	All data	99.791	0.063	0.114	0.993	31.857	36.045
		Validity data	81.779	0.172	0.299	0.211	86.046	94.563
		Test data	81.724	0.172	0.297	0.202	86.005	93.683
Middle Group Sample	All Variables (34)	All data	100	0.097	0.132	1.000	28.423	32.082
		Validity data	72.051	0.265	0.354	0.432	77.444	85.761
		Test data	71.816	0.266	0.355	0.425	77.749	86.011
	Selected variables (10)	All data	99.967	0.107	0.142	0.999	31.287	34.475
		Validity data	66.479	0.291	0.381	0.311	85.156	92.224
		Test data	66.449	0.292	0.383	0.303	85.449	92.780
	Important variables (8)	All data	99.908	0.109	0.146	0.998	31.963	35.281
		Validity data	64.627	0.298	0.389	0.274	87.035	94.147
		Test data	65.115	0.298	0.389	0.282	87.104	94.129
Upper Group Sample	All Variables (34)	All data	100	0.114	0.144	1.000	29.611	32.883
		Validity data	66.686	0.312	0.386	0.395	80.721	87.947
		Test data	66.536	0.313	0.386	0.388	81.126	88.052
	Selected variables (10)	All data	99.987	0.123	0.154	0.999	31.773	35.047
		Validity data	59.684	0.334	0.412	0.267	86.406	93.734
		Test data	60.660	0.334	0.410	0.284	86.503	93.294
	Important variables (8)	All data	99.987	0.122	0.155	0.999	31.687	35.297
		Validity data	59.002	0.334	0.415	0.259	86.314	94.548
		Test data	59.186	0.333	0.414	0.261	86.273	94.407

PCC: Percentage of correct classification, MAE: Mean absolute error, RMSE: Root of square mean squared error, RAE: Relative absolute error, RRSE: Root Relative Square Error

In Table 5, when the first 10 predictive variables and less than 10 significant variables are used in the analyses for all data, cross-validation and test data for the sample of the lower group, middle group and upper group countries, PCC does not change for all data, but for cross-validation and test data. has changed very little. For example, while the success rate for 34 variables is 100 in the model created with all data in the sample of upper group countries, the success rate is very close to 100 for 10 and 8 variables. In this case, when it decreases from 34 variables to 8 variables, the PCC for all data almost does not change, it is seen that the PCC for cross-validation and test data decreases somewhat. Accordingly, the decrease in the number of variables does not cause a significant change in the PCC in the model created with all data. The conclusion to be drawn from this is that, for example, for the upper group countries sample, the PCC value obtained from 34 variables can be reached with 8 variables. It is seen that similar results are almost the same for the sample of lower group and middle group countries. However, the PCC value obtained from the cross-validation and test data in all of the groups was lower than the PCC value obtained from the whole data.

As a result, it is seen that the performance of the RF method in determining the relationship is better than the cross-validation and prediction performance. In addition, the reduced model established by RF analysis with significant variables is ranked as lower group, middle group and upper group in order of PCC from most successful to most unsuccessful in terms of all data cross-validity and test data. In addition, the Kappa coefficient is excellent for 34 variables in the sample of the lower group, middle group and upper group countries and for all data. On the other hand, cross-validity for 34 variables and test data is reasonable in the sample of lower and upper group countries, while it is moderate in the sample of middle group countries. When the number of variables is reduced, the Kappa coefficient calculated for all the data in the sample of lower group, middle group and upper group countries does not change almost, while a slight decrease is observed in the Kappa coefficient calculated for the cross-validation and test data. In this case, the decrease in the number of variables weakened the prediction performance of the model.

Discussion and Conclusion

In the analyzes made with the CRT method, the difficulty perception variable of the PISA test emerged as an important common variable in predicting the success of reading comprehension in the sample of middle group and upper group countries. Since the importance of this variable is more in the sample of upper group countries, the difficulty perception of the PISA test affects the reading comprehension achievement of the students in the upper group countries more.

Educational items at home variable was found to be a significant predictor only in the sample of middle group countries. Educational items at home variable was found to be a significant predictor only in the sample of middle group countries. The result obtained regarding the educational items at home variable; It is similar to the research findings of Güzle Kayır (2012) that educational items related to the study environment such as the study room and study desk at home increase the reading success of the student.

The reading pleasure variable was found to be an important common predictor in the sample of lower and upper group countries. The significance of this variable is higher in the sample of upper group countries. The findings of the study conducted by Tavşancıl et al. (2019) that the variable of reading pleasure significantly affects the success of PISA reading comprehension is consistent with the results of this research.

The variable of professional status expected from the student was determined as a significant predictor only in the sample of middle group countries. Since this variable has the first estimation importance in the sample of middle group countries, it has been determined that the professional status expected from the students for these countries affects the PISA reading comprehension achievement.

The socio-economic status variable was found to be an important predictor only in the sample of upper group countries. Since the predictor importance of this variable is in the first place in the sample of upper group countries, it can be said that the socio-economic status of these countries has a significant effect on success. In the study conducted by Arıcı and Altıntaş (2014), the finding that the socio-economic status variable is a significant predictor of students' PISA 2009 reading comprehension success is similar to the results of this study.

Information and communication technologies resources emerged as an important variable only in the sample of lower group countries. Since this variable has predictive significance in the first rank in the sample of lower group countries, it can be said that information and communication technologies resources for these countries have a significant impact on success. This result is in line with the finding of the research conducted by Urfalı Dadandı et al. (2018) that information and communication technologies resources significantly predict reading comprehension success. If the sample of lower group, middle group and upper group countries is compared in terms of CRT method estimating variables; All three groups differ from each other.

In the analysis made with the RF method, the weekly test language learning time variable emerged as an important covariate predicting the success of reading comprehension in the sample of lower group, middle group and upper group countries. The significance of this variable is higher in the sample of lower group countries. While the variable of reading pleasure was found to be a significant predictor in the sample of lower and middle group countries, it was not found as a significant predictor in the upper group. On the other hand, the RF method differed from the CRT method in terms of this variable, since the variable of reading pleasure was found to be an important estimator in the sample of upper group countries in the CRT method.

The variable of educational items at home was determined as the second-ranked predictor only in the sample of upper group countries. The result obtained for this variable; It is similar to the research findings of Güzle Kayır (2012) that educational items related to the study environment such as the study room and study desk at home increase the reading success of the student. The father's education level variable emerged as the third most important predictor only in the sample of upper group countries. This result was determined by Urfalı Dadandı et al. (2018), it was determined that the father's education level was consistent with the finding of a significant predictor of PISA reading comprehension achievement.

If the sample of lower group, middle group and upper group countries is compared in terms of RF method estimator variables; All three groups differ from each other. The group that differs most is the sample of upper group countries. Since the most difference in terms of predictive variables in the CRT method is in the sample of lower group countries, RF and CRT methods differ in this respect. As a result, it has been determined that the important variables that predict reading comprehension success in the three category PISA with CRT and RF methods differ for each method in the sample of the lower group, middle group and upper group countries.

In this study, the correct classification and prediction performance of each model for the accuracy of the results were determined by the important variables selected by that model. When the prediction performance for the validity and test data of the models created by reducing the number of variables and the Kappa fit coefficient of the created models are considered together, the estimation performances obtained from the CRT and RF models in the sample of lower group countries were at a good level and the Kappa fit coefficients were significant. In the sample of the middle group and upper group countries, the estimation performance for the validity data and test data of the CRT method was found to be at an acceptable level, and the Kappa coefficient of agreement was significant. On the other hand, the estimation performance of the RF method was found to be moderate and the Kappa coefficient of agreement was significant. In the sample of middle group and upper group countries, the estimation performance of the RF method for the validity data and test data was slightly lower than the CRT method.

Suggestions

As a result of the study, the following recommendations were developed.

- In the study, it was determined that the important variables predicting the success of CRT and RF methods and PISA reading comprehension differ from method to method. In line with these findings, it can be suggested to compare the results to be obtained with different data mining methods in research to predict the success of PISA reading comprehension.
- Based on the data obtained from the application of Abide's reading comprehension test in the Turkish sample using the methods used in the research, it is suggested that important variables that predict Abide's reading comprehension success are determined and compared with the important variables that predict reading comprehension success in PISA.
- The 34 variables used to predict PISA reading comprehension success and the number of important variables selected with a sample of lower group, middle group, and upper group countries using CRT and RF methods varied between three and eight. It is recommended that a similar study be conducted in the field of PISA science and mathematics.
- As a study similar to this study, it is recommended to conduct analyzes based on data mining methods to determine the important variables that predict success in exams such as TYT, AYT, DGS and ALES conducted by ÖSYM.

Ethics Committee Permission Information: This research was carried out with the permission of Hacettepe University Scientific Research and Publication Ethics Committee with the decision dated 14/07/2020 and numbered 35853172-300-E.00001160884

Author Conflict of Interest Information: The author declares that there is no conflict of interest with any institution or person within the scope of the study.

Statement of Contribution Rate: The authors of the study contributed equally to all processes of the study.

References

Abad, F. M., & Lopez, A. C. (2016). Data-mining techniques in detecting factors linked to academic achievement. *An International Journal of Research, Policy and Practice, 28* (1), 39-55. https://doi.org /10.1080/09243453.2016.1235591

Abdous, M., He, W., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Journal of Educational Technology ve Society, 15*(3), 77-88.

Akman, M. (2010). *Veri madenciliğine genel bakış ve random forests yönteminin incelenmesi: Sağlık alanında bir uygulama* [Yayınlanmamış yüksek lisans tezi]. Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Aksu, G., & Güzeller, C. O. (2016). Classification of PISA 2012 mathematical literacy scores using decision-tree method. *Journal of Education and Science, 41*(185), 101-122. https://doi.org 10.15390/EB.2016.4766

Altunkaynak, B. (2019). *Veri madenciliği yöntemleri ve r uygulamaları*. (2.Baskı). Seçkin yayıncılık.

Arıcı, Ö., & Altıntaş, Ö. (2014). An investigation of the PISA 2009 reading literacy in terms of socio-economical backgrounds and receiving pre-school education. *Journal of Faculty of Educational Sciences*, *47*(1), 423-448.

Arıkan, S., Özer, F., Şeker, V., & Ertaş, G. (2020). Geniş ölçekli testlerde örneklem ağırlıklarının ve olası değerlerin önemi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *11*(1), 43-60. https://doi.org 10.21031/epod.602765

Atasever, Ü. H. (2011). *Uydu görüntülerinin sınıflandırılmasında hızlandırma (boosting), destek vektör makineleri, rastgele orman (random forest) ve regresyon ağaçları yöntemlerinin kullanılması* [Yayınlanmamış yüksek lisans tezi]. Erciyes Üniversitesi Fen Bilimleri Enstitüsü, Kayseri.

Bezek Güre, Ö., Kayri, M., & Erdoğan, F. (2020). Analysis of factors effecting PISA 2015 mathematics literacy via educational data mining. *Journal of Education and Science 45*(202), 393-415. https://doi.org 10.15390/EB.2020.8477

Bhalla, D. (2014). Random forest in R: Step by Step Tutorial. http//www.listendata.com/2014/11/random-forest-with-r.html sayfasından erişilmiştir.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Breiman, L. (2004). *Consistency for a simple model of random forests.* Technical Report 670, Statistics Department, University Of California At Berkeley.

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö.E., Karadeniz, Ş. ve Demirel, F. (2018). *Bilimsel Araştırma Yöntemleri*. (25.Baskı). Pegem Yayıncılık.

Coulombe S., Tremblay, J. F., & Marchand, S. (2004). *Literacy scores, human capital and growth across fourteen OECD countries*, Statistics Canada, Ottawa.

Field, A. (2009). *Discovering Statistics using SPSS*. London: Sage.

Gamazo, A., & Abad, F. M. (2020). An exploration of factors linked to academic performance in PISA 2018 through data mining techniques. *Journal Frontiers in Psycholog, 11,* 575167. https://doi.org doi: 10.3389/fpsyg.2020.575167

Güner, Z. B. (2014). Veri madenciliğinde cart ve lojistik regresyon analizinin yeri: ilaç provizyon sistemi verileri üzerinde örnek bir uygulama. *Sosyal Güvence Dergisi, 6,* 53-99.

Güzle Kayır, Ç. (2012). *PISA 2009 Türkiye verilerine dayanarak okuma becerileri alanında başarılı okullar ile başarısız okulları ayırt eden okul içi etmenler ve sosyo-ekonomik faktörler* [Yayınlanmamış yüksek lisans tezi]. Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü, Antalya.

He, Y. (2006). *Missing data imputation for tree-based models.* [Yayınlanmamış Doktora Tezi]. California Üniversity.

Kayri, M. (2008). Elektronik portfolyo değerlendirmeleri için veri madenciliği yaklaşımı. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, 5(1), 98-110.

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, 1*(1), 28-48. https://doi.org/10.21891/jeseh.41216

Köse, İ. (2018). *Veri madenciliği teori uygulama ve felsefesi*, (1.Baskı), İstanbul: Papatya yayıncılık, 48-50.

Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining.* John Wiley ve Sons, New Jersey.

Liao, Z., Ju, Y., & Zou, Q. (2016). Prediction of G Protein-Coupled Receptors with SVM-Prot Features and Random Forest. *Journal of Scientifica*, 8309253. https://doi.org/10.1155/2016/8309253.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R news, 2*(3), 18-22.

Loh, W.Y. (2011). Classification and regression trees. *Wires Data Mining and Knowledge Discovery, 1,* 14-23. https://doi.org/10.1002/widm.8

MEB (2010). *PISA 2009 ulusal ön raporu*. Ankara: http://pisa.meb.gov.tr/wp-content/uploads/2013/07/PISA-2009-Ulusal-On-Rapor.pdf.

MEB (2019). *PISA 2018 Ulusal Ön Raporu*. Ankara: T.C. Milli Eğitim Bakanlığı

OECD (2019). *PISA 2018 Technical Report*, OECD Publishing, Paris.

Özarslan, S. (2014). *Öğrenci performansının veri madenciliği yöntemleri ile belirlenmesi* [Yüksek Lisans Tezi]. Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü, Kırıkkale.

Özkan, Y. (2008). *Veri madenciliği yöntemleri*. İstanbul: Papatya yayıncılık.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine, 6(3),*21–44.

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11*(21), 1-23.

Sieber, J. E. (2008). Knowledge discovery for human research ethics. *J Empir Res Hum Res Ethics, 3* (3), 1-2. https://doi.org/10.1525/jer.2008.3.1.1

Silahtaroğlu, G. (2016). *Veri madenciliği kavram ve algoritmaları*. (3.Baskı). İstanbul: Papatya yayıncılık. 12.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson/Allyn and Bacon.

Tavşancıl, E., Yıldırım, Ö., & Bilican Demir, S. (2019). Direct and indirect effects of learning strategies and reading enjoyment on PISA 2009 reading performance. *Eurasian Journal of Educational Research, 82,* 169-190.

Urfalı Dadandı, P., Dadandı, İ., & Koca, F. (2018). PISA 2015 türkiye sonuçlarına göre sosyoekonomik faktörler ile okuma becerileri arasındaki ilişkiler. *Uluslararası Türkçe Edebiyat Kültür Eğitim Dergisi, 7*(2), 1239-1252.

Wu, M. L., & Adams, R. J., (2002). Plausible Values: Why they are important. Paper presented at the International Objective Measurement workshop, New Orleans, LA.

Yu, C. H., Kaprolet C., Pennel, A. J., & Digangi, S. (2012). A data mining approach to comparing american and canadian grade 10 students' pisa science test performance. *Journal of Data Science 10,* 441-464. https://doi.org/10.6339/JDS.201207_10(3).0006

Yue, Y. (2021). The prediction model of college english performance based on data mining under the concept of obe. *Engineering Intelligent Systems, 2*, 89–96.

Yung, J. L., Hsu, Y. C., & Rice, K. (2012). Integrating data mining in program evaluation of k-12 online education. *Journal of Educational Technology ve Society, 15*(3), 27-41.

Geniş Türkçe Özet

Problem Durumu

Eğitimde de son yıllarda hem uluslararası düzeyde hem de ulusal düzeyde yapılan çalışmalarda çok değişkenli veri toplama eğilimi yaygınlaşmıştır. Bu sınavlardan birisi de Uluslararası Öğrenci Değerlendirme Programıdır (PISA). PISA verileri kullanılarak, çok değişkenli VM modelleri ile, başarıyı yordayan veya başarı ile ilişkili bağımsız değişkenlerin hangileri olduğu saptanabilmektedir.

Alanyazında yapılan araştırmalar incelendiğinde tek bir ülke ile sınırlı olan homojen gruplarla yapılan çalışmalar yoğunluktadır. Veri madenciliği yöntemleri kullanılarak farklı başarı gruplarında başarıyı yordayan değişkenlerin değişip değişmediğine yönelik çalışmaya rastlanmamıştır. Aslında veri (örneklem) değiştikçe veri madenciliği yöntemlerinden elde edilen sonuçların değişip değişmediğinin araştırılması önem teşkil etmektedir.

Başarıyı yordayan değişkenleri saptamak için veri madenciliği kapsamında birçok yöntem kullanılabilir ancak kullanılan modelin performans düzeyi değiştiğinde ve çalışılan örneklemin başarı düzeyi değiştiğinde nasıl bir sonuçla karşılaşıldığını görmek için farklı veri madenciliği kestirim yöntemleri ile elde edilen sonuçların karşılaştırılması gerekmektedir. Bu araştırmada Sınıflama ve Regresyon Ağacı ve Rasgele Orman yöntemi kullanılmıştır.

Alanyazında PISA başarı puanını yordayan önemli değişkenlerin belirlenmesi çalışmalarında kullanılan örneklemlerin dar kapsamlı olduğu görülmektedir. Örneğin; Aksu ve Güzeller (2016) ve Bezek Güre ve ark. (2020) çalışmalarında sadece PISA Türkiye örneklemini kullanmışlardır. Yapılan çalışmalar başarı veya performansı yordayan değişkenler üzerinde yoğunlaşmakta ancak bu çalışmalar birbirine benzer şekilde fen, matematik veya okuma puanını bağımlı değişken kabul ederek sözü edilen üç alandaki başarıyı etkileyen faktörlerin açığa çıkarılmasına yöneliktir. Ancak bu çalışmalarda farklı başarı düzeyleri dikkate alınmamıştır. Oysaki farklı alt başarı gruplarında farklı değişkenlerin etkili olması düşüncesinden dolayı bu iddiayı farklı başarı gruplarında çalışarak doğrulamak mümkündür.

Birçok ülkenin katıldığı PISA'da ülkelerin başarı düzeyleri birbirinden oldukça farklıdır. Bu nedenle; bu çalışmada, alt, orta ve üst başarı grubundaki ülkeler için başarıyı etkileyen değişkenlerin belirlenmesi ile ilgili gruplardaki ülkelerin eğitim politikalarına yön verirken başarıyı olumlu etkileyen faktörlere daha fazla önem vermeleri sağlanabilir. Diğer yandan farklı başarı düzeylerinde ülkeler için; başarıyı etkileyen aynı ve farklı faktörlerin neler olduğu saptanarak, düşük başarı düzeyindeki ülkelerin başarılarını artırmada daha doğru kararlar almaları sağlanabilir. Bu bakımdan araştırmanın alanyazına katkı sağlayacağı düşünülmektedir. Bu araştırmanın temel amacı, PISA öğrenci anketinde bulunan 34 bağımsız değişkeni kullanarak farklı başarı düzeyine sahip (alt, orta ve üst başarı grubundaki) ülkelerin okuduğunu anlama başarısını etkileyen önemli yordayıcıları belirlemektir. Çalışmanın diğer amacı ise; belirlenen az sayıda daha önemli bağımsız değişkenleri kullanılarak performansı yüksek sınıflama modellerinin elde edilip edilmeyeceğini göstermektir. Böylece; çalışmanın, kullanılan veri madenciliği yönteminin performansını en yüksek yapan daha önemli bağımsız değişkenleri seçerek değişken sayısını azaltmaya yönelik az sayıda çalışmadan biri olması hedeflenmektedir.

Yöntem

Bu araştırma nicel araştırma tasarımlarından tanımlayıcı tasarım, karşılaştırmalı tanımlayıcı tasarım ve ilişkisel tasarım olarak tasarlanmıştır. Çalışmanın amacı için ülkeler; PISA okuma puanı sıraları dikkate alınarak alt (zayıf), orta, üst (yüksek) başarı grubuna göre belirlendiği için amaçlı örnekleme yöntemi kullanılmıştır Ayrıca, alt, orta ve üst grup örneklemlerini oluşturmak için ülkeler seçilirken kayıp veri oranının düşük olmasına ve farklı yüzdelik dilimlere dağılacak şekilde örüntü oluşturmasına dikkat edilmiştir. Böylece her bir grup için seçilen örneklemlerin sağlayacağı bilgilerin zenginleşmesi sağlanmıştır. Bu araştırmada CRT ve RF yöntemleri ile 34 bağımsız değişken ile yapılan analizlerde her bir yöntemin hesapladığı tahmin edici önem (predictor importance) değerleri dikkate alınarak başarıyı en iyi yordayan ilk 10 değişken saptanmıştır. Daha sonra, önemli bağımsız değişkenler kullanılarak Weka programında; tüm veri, 10 katlı çapraz geçerlik verisi ve test verisi ile performans kriterleri hesaplanmıştır. Bu hesaplanan performans kriterlerinin 34 bağımsız değişken kullanıldığında hesaplanan performans kriterlerinden önemli düzeyde bir farklılığa sahip olmadığı belirlenmiş ve böylece başarıyı tahmin etmede, daha az sayıda değişken ile yüksek performansa sahip modeller kurulabildiği gösterilmiştir. Bu analizler alt, orta, üst başarı grubu ülkeler örneklemi için yapılmış ve elde edilen sonuçlar ayrıntılı bir karşılaştırma yapılarak yorumlanmıştır.

Bulgular

Araştırmanın sonuçları incelendiğinde sınıflama ve regresyon ağacı modeli ile seçilen önemli ve ortak değişkenler PISADIFF (PISA testinin zorluk algısı) değişkeni alt grup hariç diğer örneklem gruplarının her biri için, üç kategorili PISA okuduğunu anlama başarısını yordama da önemli ortak değişken olarak saptanmıştır. Burada, başarıyı yordama bakımından PISADIFF değişkeni üst grup ülkeler örnekleminde ilk sırada, orta grup ülkeler örnekleminde ikinci sırada öneme sahip olup; bu ülkeler için PISA testinin zorluk algısı değişkeninin, PISA okuduğunu anlama başarı durumunu etkileyen önemli tahmin edici olduğu söylenebilir. JOYREAD (okuma keyfi), alt grup ve üst grup ülkeler örneklemi için, PISA okuduğunu anlama başarı durumunu yordayan önemli ortak değişken olarak belirlenmiştir. Bu değişken başarıyı yordama bakımından alt grup ülkeler örnekleminde beşinci sırada ve üst grup ülkeler örnekleminde ise ikinci sırada öneme sahip olup; bu ülkeler için okuma keyfi değişkeninin, PISA okuduğunu anlama başarı durumunu yordayan etkili tahmin edici olduğu söylenebilir. Rastgele orman modeli ile seçilen önemli ve ortak değişkenler LMINS (haftalık test dili öğrenme süresi) değişkeni alt grup, orta grup, üst grup ülkeler örnekleminde üç kategorili PISA okuduğunu anlama başarısını yordama da önemli ortak değişken olarak saptanmıştır. Burada, başarıyı yordama bakımından LMINS değişkeni alt grupta birinci sırada, orta grupta ikinci sırada, üst grupta yedinci sırada öneme sahip olup; bu ülkeler için haftalık test dili öğrenme süresi değişkeninin, PISA okuduğunu anlama başarı durumunu etkileyen önemli tahmin edici olduğu söylenebilir. JOYREAD (okuma keyfi), alt grup ve orta grup için, PISA okuduğunu anlama başarı durumunu yordayan önemli ortak değişken olarak belirlenmiştir. Bu değişken başarıyı yordama bakımından alt grup ülkeler örnekleminde ikinci sırada, orta grup ülkeler örnekleminde yedinci sırada öneme sahip olup; bu ülkeler için okuma keyfi değişkeninin, PISA okuduğunu anlama başarı durumunu yordayan önemli tahmin edici olduğu söylenebilir.

Tartışma ve Sonuç

CRT yöntemi ile yapılan analizlerde PISA testinin zorluk algısı değişkeni orta grup ve üst grup ülkeler örnekleminde okuduğunu anlama başarısını tahmin etmede önemli ortak değişken olarak ortaya çıkmıştır. Bu değişkenin önemi üst grup ülkeler örnekleminde daha fazla olduğu için üst grup ülkelerdeki öğrencilerin okuduğunu anlama başarı durumlarını PISA testinin zorluk algısı daha fazla etkilemektedir.

Okuma keyfi değişkeni alt grup ve üst grup ülkeler örnekleminde önemli ortak tahmin edici olarak tespit edilmiştir. Bu değişkenin önem derecesi üst grup ülkeler örnekleminde daha fazladır. Tavşancıl ve arkadaşları (2019) tarafından yapılan araştırmada okuma keyfi değişkeninin PISA okuduğunu anlama başarısını önemli ölçüde etkilediğine yönelik bulguları bu araştırma sonuçları ile uyumludur. CRT yöntemi tahmin edici değişkenler bakımından alt grup, orta grup ve üst grup ülkeler örneklemi karşılaştırılacak olursa; üç grubun tamamı birbirinden farklılık göstermektedir.

RF yöntemi ile yapılan analizde Okuma keyfi değişkeni alt grup ve orta grup ülkeler örnekleminde önemli tahmin edici olarak bulunurken üst grupta önemli tahmin edici olarak bulunmamıştır. Diğer yandan CRT yönteminde üst grup ülkeler örnekleminde okuma keyfi değişkeninin önemli tahmin edici olduğu saptandığı için RF yöntemi bu değişken bakımından CRT yönteminden farklılaşmıştır.

RF yöntemi tahmin edici değişkenler bakımından alt grup, orta grup ve üst grup ülkeler örneklemi karşılaştırılacak olursa; üç grubun tamamı birbirinden farklılık göstermektedir. En çok farklılık gösteren grup ise üst grup ülkeler örneklemidir. CRT yönteminde tahmin edici değişkenler bakımından en çok farklılık alt grup ülkeler örnekleminde olduğu için bu bakımdan RF ve CRT yöntemi ayrışmıştır. Sonuç olarak CRT ve RF yöntemleri ile üç kategorili PISA okuduğunu anlama başarısını yordayan önemli değişkenlerin her bir yöntem için alt grup, orta grup, üst grup ülkeler örneklemi için farklılık gösterdiği tespit edilmiştir.