Eurasian Journal of Educational Research 84 (2019) 115-134

Eurasian Journal of Educational Research
www.ejer.com.tr

Comparison of Results Obtained from Logistic Regression, CHAID Analysis and Decision Tree Methods*

Gokhan AKSU[1], Cigdem REYHANLIOGLU KECEOGLU[2]

ARTICLE INFO

Article History:
Received: 31 Jul. 2018
Received in revised form: 01 May 2019
Accepted: 20 Aug. 2019
DOI: 10.14689/ejer.2019.84.6

Keywords
CHAID Analysis, Logistic Regression Analysis, Data Mining, PISA

ABSTRACT

Purpose: In this study, Logistic Regression (LR), CHAID (Chi-squared Automatic Interaction Detection) analysis and data mining methods are used to investigate the variables that predict the mathematics success of the students.

Research Methods: In this study, a quantitative research design was employed during the data collection and the analysis phases.

Findings: The findings obtained in this study showed that the variables, which had significant effects on the mathematical success of the students in each method, differ from each other. Although the independent variables with a significant effect on the dependent variable were different according to different methods, the findings indicated that the importance order of the variables did not change according to the method used. In this study, the correct classification ratios obtained by the class concerning PISA mathematics literacy differed by different methods.

Implications for Research and Practice: CHAID analysis and REPTree algorithm may be an alternative for one another in the studies that aimed to classify individuals concerning their success. However, LR analysis should not be considered as an alternative method since it will provide significantly different results compared to the other two methods.

*This study was partly presented at the 5th International Eurasian Educational Research Congress in Antalya, 2 – 5 May, 2018

[1] Aydın Adnan Menderes University, TURKEY, ORCID: 0000-0003-2563-6112
[2] Gaziantep College Foundation Private Schools, TURKEY, ORCID: 0000-0002-6168-6662

Introduction

Many contexts are considered to direct the education policies of the countries. Policymakers around the world use the results of international practices to compare the knowledge and skill levels of the pupils in their own countries with the knowledge and skill levels of the pupils in other countries to set standards to raise the level of education (such as average scores achieved by countries, countries' educational outcomes and their capacity to achieve equality in education opportunities at the highest level) and to determine the strengths and weaknesses of education systems (International Student Assessment Program [PISA], 2015). The data obtained from international examinations, such as the International Student Assessment Program (PISA) applied by the Organization for Economic Co-operation and Development (OECD) to the 15-year-old students, and the International Mathematics and Science Trends Survey (TIMSS) applied to students at the fourth and eighth grade by the International Education Achievement Assessment Organization (IEA), are extremely important for the direction of the education policies of the countries (International Mathematics and Science Trends Survey [TIMSS], 2015). In these applications, which are realized with the participation of different countries, achievement tests, and various questionnaires, are used to gather information about students' performances in science and mathematics, education systems, curriculum, student characteristics, characteristics of teachers and schools (TIMSS, 2015). Thanks to this information, countries are able to evaluate their educational processes according to an international perspective.

The findings obtained from international examinations, which play an important role in shaping the countries' educational policies, are derived from a large-scale database where variables in different areas are measured. Very large-scale data and large-scale databases in different areas can be considered as a data mine, including valuable data. From this data mine, which has a complex structure, to generate meaningful information that is not known beforehand, process management with different operations is required. This process management takes place with data mining. Data mining performs this process using a computer, machine learning, database or data warehouse management, mathematical algorithms and statistical techniques (Albayrak & Koltan-Yılmaz, 2009). Data mining is basically defined as the use of software techniques for accessing useful information through the relationships or patterns within the large data sets (Can, Özdiland Yılmaz, 2018). Thanks to data mining software and techniques, large scale data can be decomposed, and useful information can be revealed.

There are many processing steps that must be performed in the data mining process. Larose and Larose (2014) emphasize that the data mining process takes place in five stages as follows: definition of the task, recognition of data, preparation of data, modelling and evaluation. The most troublesome of these stages are the stages of recognition and preparation of the data (Özkekeş, 2003). The data obtained from international applications, such as PISA, can be considered as data that are complex, and therefore, suitable for arrangement and modeling with data mining stages. Using data mining methods, maximum information can be obtained about the independent variables predicting the dependent variable on the complex data obtained from the applications that direct the educational policies of the countries.

Determining independent (predictive) variables that influence the dependent variable is one of the main focuses of scientific research. In these studies conducted for this purpose, various methods are used to determine predictor variables that have relationships with the dependent variable. The common feature of these methods is to test the significance of the effects of independent variables on the dependent variable. The methods used have different characteristics, as well as their common characteristics. The most important of these are the assumptions required by the methods.

The methods are generally divided into parametric and non-parametric methods concerning assumptions that must be met to be applied. In cases where parametric conditions do not occur (such as quantitative variables come from a multivariate normal distribution assumption is not established, homogeneity of variance/covariance matrices is not established), it has been a great convenience for researchers to develop and use nonparametric methods, which can be used for the same purpose with a parametric method (Baştürk, 2016). However, if the need to make a choice among these alternative nonparametric methods arises, it is necessary to compare the methods to decide which method to be used concerning the accuracy of the results. The comparison studies performed for this purpose are important concerning ensuring the validity of the decision. Thus, it has great importance to determine the method which is based on regression analysis is used for estimating the dependent variable with the help of independent variables. In this way, researchers will contribute to science using the method that produces the most accurate and most consistent results.

Another feature that separates the methods used is the type of data that the method can be applied to. Some of the statistical methods can only be applied to continuous data, while some can also be applied to categorical data. Categorical data analysis is a method commonly used in educational applications (Azen & Walker, 2011). Although the results of the measurement are obtained as continuous scores by accepting measurement tools used to determine the academic achievement of the students as equal intervals on the scale level, the success scores in the decision-making process are converted into categorical data in the form successful/unsuccessful according to a certain criterion score (Baştürk, 2016). Thus, the students are classified as successful/unsuccessful according to their achievement score.

Although a criterion score is often used in determining student achievement, there are many factors affecting student achievement. At this point, statistics is concerned with identifying those who have a significant impact on these factors, and these factors need to be considered in the process of assessing student achievement.

Statistical methods can be classified as descriptive and predictive methods according to the intended purpose (Tütek & Gümüşoğlu, 2008). In general, predictive methods are used to determine the factors affecting success score methods (Zuckerman & Albrecht, 2001). One of these methods is the logistic regression analysis. Logistic regression analysis can be considered as a special case of regression analysis methods (Peng, Lee & Ingersoll, 2002). The regression analysis is a strong statistical method which aims to explain the relationship between two variables, one of the two

or more variables are taken as dependent and the other as independent variables using a mathematical equation (Çokluk, 2012).

The strong assumptions of linear regression analysis do not allow to be implemented when its parametric conditions are not available. In such cases, regression-based, non-parametric multivariate statistical methods are used. One of the methods that can be used in cases where the dependent variable is categorical or classified and an assumption is not required for the distribution of independent variables is Logistic Regression analysis (Mertler & Vannatta, 2005; Tabachnick & Fidell, 2001). Logistic regression analysis has an important place in categorical data analysis concerning requiring fewer assumptions than methods that have regression logic and used to determine predictive variables (Kılıç, 2000). In the application phase of the method, users important have advantages because it requires fewer assumptions (Park, 2013). Besides, the results of the model can be interpreted easily. One of the methods that can give similar results with regression analysis and do not take into account the assumptions of regression analysis is the CHAID analysis method. The method can use algorithms, user-defined rules, criteria specified via an interactive graphical user interface, or a combination of these methods. This enables users to try various predictors and splitting criteria in combination with almost all the functions of automatic tree building (Nisbet, Miner & Yale, 2017). Thanks to the tree diagrams, the independent variables predicting the dependent variable and the importance levels of these variables can be seen (DíazPérez & Bethencourt-Cejas, 2016).

As a result, the results of two methods, which are nonparametric, logistic regression analysis and CHAID analysis methods, can be used to determine the factors that have a significant effect on student achievement as a dependent variable. The common feature of all three methods is to focus on determining the independent variables that have a significant effect on the dependent variable. However, the most basic feature that separates these three methods is the learning algorithm that runs in the background. The CHAID algorithm, proposed by Kaas in 1980, was formed by combining the predictive variable in the category pairs with no significant difference (SPSS, 1999). REPTree algorithm is used in data mining. In data mining, there is the C4.5 algorithm, which is called the statistical classifier (Witten and Fransk, 2005). An extension of the C4.5 algorithm is REPTree algorithm) is used to construct a decision tree (Quinlan, 1993). This study aims to determine the independent variables which are thought to have a significant effect on mathematics achievement and to reveal the order of importance of these variables. Moreover, another focus of the study is that whether the order of importance of variables differs according to the methods used. In this study, it was also investigated how the students were categorized according to variables of their interest, attitude, motivation, perception, self-efficacy, anxiety and working discipline towards mathematics course. In this way, it is thought that researchers working in national and international fields will obtain more precise and more consistent measurement results by working with the best data analysis method they need in the analysis stage. It is of great importance to determine how similar or different results logistic regression analysis, which is frequently used in estimating the categorical dependent variable, and CHAID analysis and data mining methods, which have been used more recently, will show from the same data set. There are studies comparing these methods in the literature (Antipov & Pokryshevskaya, 2009; Şata & Çakan, 2018; Rudd &

Priestley, 2017). Some researchers have proposed CHAID as an aid for better specifying and interpreting a logit model. In this study, the CHAID, data mining approach and logistic regression were used for finding whether independent variables significantly have a lower or higher effect on the dependent variable according to the used analysis method. This approach is employed for diagnostic purposes, as well as for improving the initial model. We demonstrated that the proposed method could be used for splitting the dataset into several segments, followed by building separate models for each segment, which led to a significant increase in classification accuracy both on training and test datasets and, therefore, enhanced logistic regression.

The problem statement of the research within the framework of specified purposes is as follows: "Does statistical significance of the features, such as the students' interest, attitude, motivation, perception, self-efficacy, anxiety and work discipline differ according to the method used?" In accordance with the determined basic problem statement, the following questions were raised within the scope of this research:

1. Do the significance levels of the independent variables differ according to the method used?
2. Do the accurate classification rates of independent variables differ according to the method used?
3. Is the order of importance of the independent variables in classifying the students concerning mathematics achievement differ according to the method used?

Method

Research Design

In this study, logistic regression (LR), CHAID analysis and data mining methods were used to investigate the predictors of student achievement. This study aimed that the relationship between two or more variables is examined in any way without any interference from these variables. Due to examining the relationship between variables, this study is correlational research (Büyüköztürk, Çakmak, Akgün, Karadeniz and Demirel, 2016).

Research Sample

The data used in this study were obtained with the help of the responses, which were about the subscales of interests, attitudes, self-efficacy, perception, motivation, anxiety and study discipline of students who took part in the PISA 2012 Student Questionnaire. The data file used in the study was obtained from the official OECD website, http://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm. The data file in the format of the text document was converted to the appropriate format for analysis in SPSS program using the syntax.

The universe of the study consisted of 4818 students participating in PISA 2012 student survey and was determined by a stratified random sampling method. However, it was decided that the missing data should be excluded from the analysis because the loss data rate was high for the seven different affective features used in this study and the missing data were not randomly distributed, which may lead to bias in the statistical analysis results (Garson, 2015; Groves, 2006; Tabachnick & Fidell,

2014). After the missing data analysis, the results of this study were obtained from the data collected from 1000 participants using systematic sampling from the universe. In the study, a systematic sampling method was used, which is one of the probabilistic sampling methods given that the boundaries of the universe are certain and the universe is relatively large (Cohen, Manion & Morrison, 2007). The population can be represented with a high degree by a systematic sampling method (Koç Başaran, 2017).

Data Analysis

The independent variables of this study consisted of the variables which are investigated whether they have a significant effect on mathematics achievement. In the variable selection stage, the findings of the study conducted by Aksu and Güzeller (2016) were used. In the study, it was stated that PISA 2012 dataset consisted of seven different sub-scales to determine the affective qualities of the students. These are the variables of the students' interest in mathematics, mathematics motivation, attitude towards mathematics, self-efficacy in mathematics, math anxiety, mathematics study discipline and student's math perception. The mathematical achievement in which the effects of independent variables are investigated is the dependent variable of the study.

For the analysis of the data, dependent and independent variables were first analyzed according to Binary LR analysis, and consequently, the correct classification ratio was determined according to the mathematics achievement of the students with independent variables, which had a significant effect on mathematics achievement. However, in the first stage, the assumptions, which are required for LR analysis, were tested. According to Tabacknick and Fidell (2001), there are four assumptions that are to be tested for LR analysis. First, binary logistic regression requires the dependent variable to be binary, and ordinal logistic regression requires the dependent variable to be ordinal. In this study, dependent variables had an ordinal scale. Second, logistic regression requires observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data. In this study, the observations, which come from repeated measurements or matched data, were not determined. Third, logistic regression requires there to be little or no multicollinearity among the independent variables, which means that the independent variables should not be too highly correlated with each other. In this study, the Pearson Correlation between independent variables was under the critic level (<0,70) and was not statistically significant. Fourth, logistic regression assumes the linearity of independent variables and log odds. Although this analysis does not require the dependent and independent variables to be related linearly, in this study, the linear relationship between independent variables was specified clearly via a scatter chart. Finally, logistic regression typically requires a large sample size. A general guideline is that you need a minimum of 10 cases with the least frequent outcome for each independent variable in your model (Bush, 2015). In this study, more than 10 cases were used for each independent variable. After the assumptions were tested, dependent and independent variables were analyzed according to Binary LR analysis.

In the second stage, the same variables and the correct classification ratio were analyzed using CHAID analysis. As a result of CHAID analysis, a decision tree was

obtained from the dependent variable and independent variables, which had a significant effect on the dependent variable. In the third stage, the analysis was carried out using the REPTree algorithm in the weka program, and independent variables that had a significant effect on mathematics achievement were determined. In addition, as in the other two methods, the classification result was obtained according to the success of the students. In the final stage, the common variables predicting mathematics achievement were determined according to the results obtained from each method. Moreover, it was tested whether there was a significant difference between the correct classification rates for each method. T and z statistics can be used to test this difference between ratios and the significance of this difference. If n> 30 for the calculation, the z statistic is calculated; if n <30 for the calculation, t statistic is calculated. In this study, the z test was used because the sample size was greater than 30 (Lehmann, 2006). The z test was performed with the help of the equation given below.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

This results in the standardized statistic, which, when both n1p1 and n2p2 are greater than 5, can be shown to approximately follow the standard normal distribution (Massey & Miller, 2006).

Results

LR was carried out by the comparison of CHAID analysis and REPTree algorithm methods, determining the variables that had a significant effect on the students' success level and by comparing the correct classification rates as successful and unsuccessful (1-0) concerning PISA mathematics achievement. In this study, the results obtained by LR analysis are given first.

Findings with LRA

As a result of logistic regression analysis, the variables that have a significant effect on the classification of the students regarding their success levels are shown in Table 1.

Table 1.

Logistic Regression Analysis Results

		B	S.H	Wald	sd	p.	Exp(B)
1. Stage	interest	.058	.032	3.289	1	.070	1.060
	motivation	.008	.030	.070	1	.791	1.008
	attitude	.107	.019	31.283	1	.000	1.113
	self-efficacy	.241	.018	177.447	1	.000	.786
	anxiety	-.096	.020	22.352	1	.000	1.101
	perception	.012	.032	.143	1	.705	1.012
	discipline	.057	.015	14.922	1	.000	1.059
	constant	-.806	.470	2.944	1	.086	.447

a. 1. Variables included in the analysis: interest, motivation, attitude, self-efficacy, anxiety, perception, discipline.

In the evaluation of students remained left or pass an average of the obtained success scores was taken ($\bar{x}$=449.00), and this value was determined as the cutting value. The students over the average were categorized as 1, while those below the average were categorized as 0. When the Table 1 is examined, it is observed that the variables of attitude (β = 0,107, p <.01), self-efficacy (β = -0,241, p <.01), anxiety (β = 0,096, p <.01) and study discipline (β = 0.057, p <.01) had a significant effect in the classifying student performance as successful and unsuccessful concerning total scores. β values given in the table are the values to be used in an equation where the probability of taking a sample in a given category is calculated. According to this, if students have high self-efficacy, attitudes and discipline scores, they would be more likely to be successful. According to the Exp (β), also known as Odds Ratio, self-efficacy (0.79), attitude (1.11), anxiety (1.10) and working discipline (1.06) show the values that are likely to be successful concerning science literacy in PISA. The significance levels of the independent variables on the dependent variables were self-efficacy (177,44), attitude (31,28), anxiety (22,35) and working discipline (14,92), respectively.

In the logistic regression analysis, while the SPSS package program classifies individuals as passed/failed as a result of an achievement test, it creates a classification percentage for a predicted variable by accepting all students as 'passed' or 'failed' in the initial model. Then, the predictive variables are added to the initial model, and the actual classification percentage is obtained concerning the success of the individuals. According to this, while the correct classification rate of the students was 53,50% in the initial model, this ratio was determined as 71,20% by including the variables in the model. Regarding the total variance explained of the model, Cox & Snell R2 value was calculated as..20, and Nagelkerke R2 was calculated as .,27. Accordingly, 27% of the variability in mathematics literacy is explained by the variables added to the model. As a result of the Omnibus test, it was determined that the first model obtained by the inclusion of both the initial model and the variables in the model was statistically significant (χ2=31028, sd=7, p<.01).

Findings by CHAID Analysis

As a result of logistic regression analysis, the results which were dependent on important statistics for independent variables that had a significant effect on the dependent variable were obtained. In CHAID analysis, the predictive variables that are significant can be seen through the nodes in the decision tree branching process. The variables that appear in the nodes of the tree from top to bottom provide information about the order of importance of the variable, respectively. Accordingly, the decision tree obtained by CHAID analysis is shown in Figure 1.

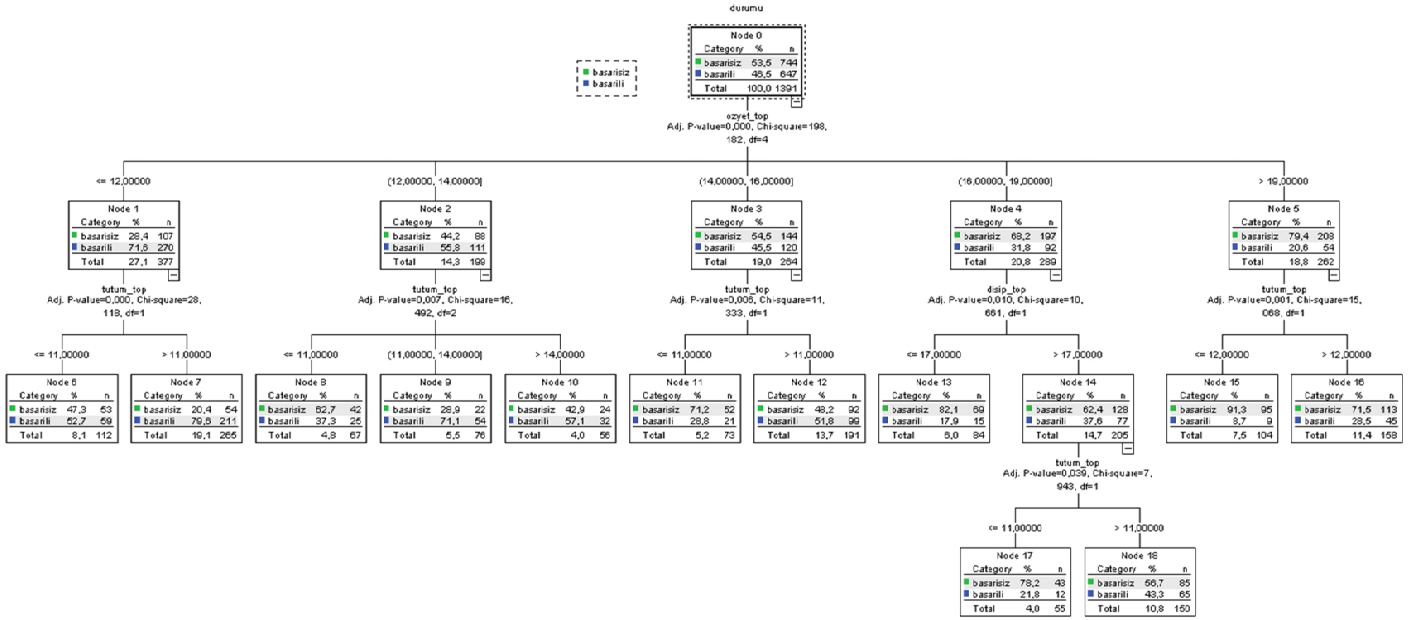

Figure 1. *Decision Tree for the Classification of PISA Mathematics Performance*

When Figure 1 is examined, the findings showed that mathematics self-efficacy was the best predictor of PISA literacy ($\chi 2=198.00$, df=4; $p<.01$). The most effective variable in a sub-branch of the tree was the attitude for 4 nodes while the most effective variable for the remaining node number 4 was the study discipline ($\chi 2=10.00$, df=1; $p<.05$). On the branching obtained in the third stage of the decision tree, the most effective variable was the attitude ($\chi 2=7.00$, df=1; $p<.05$). In addition to this, the amount of information gain (gain) in each node is shown in Table 2.

Table 2.

Knowledge Gain Quantities on Nodes in CHAID Analysis

Node	Node		Amount of Gain		Response ratio (%)	Indices (%)
	N	Percentage	N	Percentage		
7	265	19,10	211	32,60	79,60	171,20
9	76	5,50	54	8,30	71,10	152,80
10	56	4,00	32	4,90	57,10	122,90
6	112	8,10	59	9,10	52,70	113,30
12	191	13,70	99	15,30	51,80	111,40
18	150	10,80	65	10,00	43,30	93,20
8	67	4,80	25	3,90	37,30	80,20
11	73	5,20	21	3,20	28,80	61,80
16	158	11,40	45	7,00	28,50	61,20
17	55	4,00	12	1,90	21,80	46,90
13	84	6,00	15	2,30	17,90	38,40
15	104	7,50	9	1,40	8,70	18,60

When Table 2 is examined, it was seen that most information was obtained from node 7. In this node where 211 students were successful, and 54 of them were classified as unsuccessful, while the correct classification rate was 7960%, the overall success rate was determined as 19,0%. Based on these values, the amount of information gain obtained from node 7 was calculated as 32,60%, and it was observed that self-efficacy and attitude variables are effective in making classification, respectively. It was seen that the second node that provided the most information was the node number 12, which had the correct classification rate of 51,80%, by classifying 99 of 191 students correctly. Plus, the self-efficacy and attitude variables were respectively effective in the classification of this node, which had an information gain of 15.30% throughout the tree. The third most common node was nodes number 18, which had a correct classification rate of 43.30% by classifying 65 of the 150 students as successful. While

the categorization of this node with an information gain amount of 10.00% throughout the tree, it was observed that self-efficacy, work discipline and attitude variables were effective. It was seen that the fourth node giving the most information is node 6 with the correct classification rate by classifying 59 of 112 students as successful, and it had a 52,70% correct classification rate. Throughout the tree, it was seen that during classification, the variables of self-efficacy and attitude were effective in the tree, respectively. While the categorization of this node with 9.10% knowledge gain, when the results obtained by CHAID analysis were evaluated as a whole, the severity of the variables predicting success is respectively self-efficacy, attitude and study discipline. Accordingly, it was determined that the anxiety variable, which had a significant effect on logistic regression, had no significant effect on the three-level decision tree.

While the accurate classification rate of classifying the real successful students as successful was 70.30%, the rate of real failure students as failure was determined as 67.10%. Accordingly, the correct classification rate for all students in the decision tree obtained by CHAID analysis was determined as 68.60% with 0,314 risk value and with 0,012 standard error. The risk ratio shows that 31.40% of the classification can be misclassified. This rate was determined as 71.20% in LR.

Findings Obtained by REPTree Algorithm

The decision tree obtained with the aim of determining the variables which had a significant effect on classification as successful and unsuccessful in mathematics literacy with the help of REPTree classification algorithm from data mining methods is shown in Figure 2.

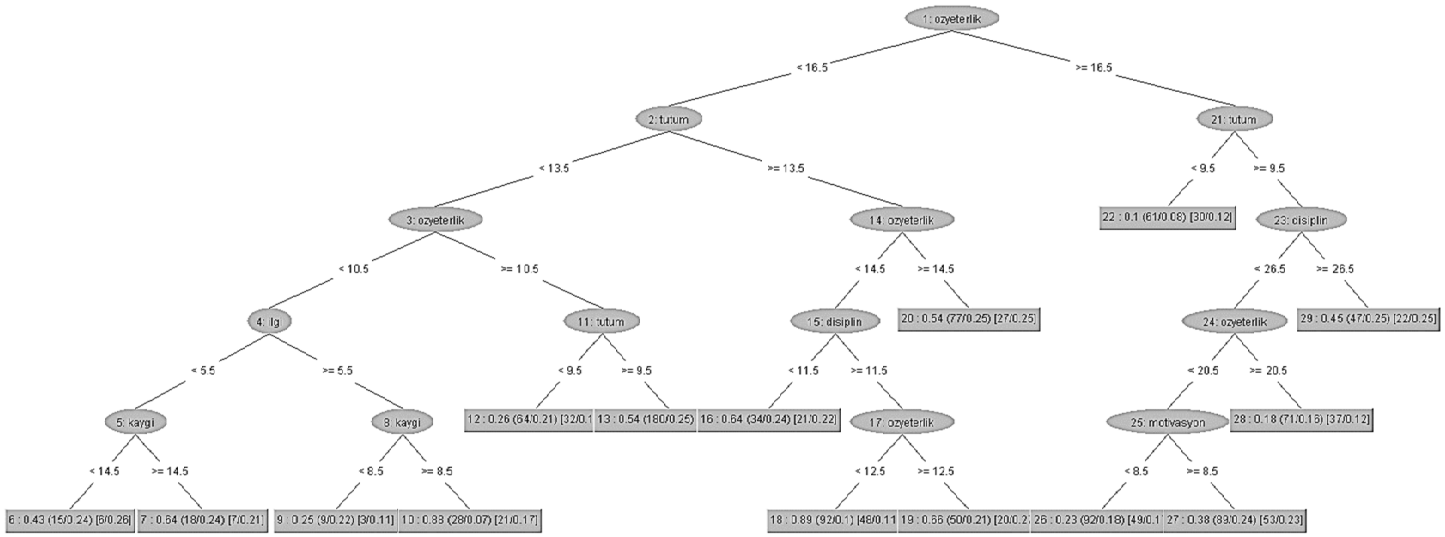

Figure 2. *Decision Tree Obtained with the help of the REPTree Algorithm*

When Figure 2 is examined, it was determined that the best predictor of PISA literacy was mathematics self-efficacy. In the decision tree obtained by the REPTree algorithm, the cut-off point concerning the self-efficacy level was determined as 16.50. This cut-off value was obtained due to the default settings of the program. If you change the properties of tree-like maximum tree dept or batch size the cut values, there will be minor changes in the cut point. It was determined that the best predictor variable was the self-efficacy for the students who scored equal to the cut-off point and who scored below/above this value. It was seen that regarding mathematics literacy, the most effective variable in tree branching at the third level was perception for 2 nodes, study discipline for 2 nodes and anxiety and motivation variables for the remaining nodes. It was also seen that in the fourth step of the decision tree, the most effective variables were anxiety for 2 nodes, interest for 2 nodes, motivation for 2 nodes, and attitude, self-efficacy and motivation for the remaining nodes. When the results obtained with the help of the REPTree algorithm were evaluated as a whole, the order of importance of independent variables was obtained in the following order: self-efficacy, attitude, working discipline, interest, motivation, and anxiety. According to the results obtained by the REPTree algorithm, the first three variables that affected the success were self-efficacy, attitude and working discipline, respectively.

When the confusion matrix of the classification process is examined by the REPTree algorithm, it was seen that the number of the correctly classified students was determined as 372 students were successful, and 523 students failed. The model made a mistake by classifying 275 students as successful who were unsuccessful in reality; and 221 students as unsuccessful who were successful in reality. According to this, 895 of 1391 students were assigned to the right classes, and the correct classification rate of the REPTree algorithm was determined as 64.34%. The mean square root of the errors was 0,513, and the kappa statistic was 0,279. There are no standard assessment criteria for the level of estimation of each learning method in data mining (Sokolova &Lapalme, 2009). In the analysis, the desired error and kappa statistics are as low as possible. In addition, one of the validity criteria obtained in data mining is the area under the ROC curve. When this value is close to 1, it is stated that an excellent classification is made. According to the results of the analysis, it was determined that the area under the ROC curve was .63, and the sensitivity value of the model was .642.

As a result of the analysis, the order of importance of independent variables of LR, CHAID analysis and REPTree algorithm and classification results related to all three methods according to the student success were obtained. When deciding which method to choose, the classification results obtained from the methods can be considered. After determining whether there was a significant difference between the classification results obtained, it can be decided which method to choose according to the size of the correct classification ratio. As a result of the analysis, the difference between the classification rates obtained from the three methods was tested in binary groups. In other words, the classification ratios obtained from LR and CHAID analysis were compared first and then the results of the LR and REPTree algorithm and finally, the classification results obtained from CHAID analysis and REPTree algorithm were compared.

The Z-statistic calculated for the comparison of the correct classification rates obtained as a result of the LR and CHAID analysis was smaller than the critical value of the Z-statistic

at the significance level of 0.05 (Zd calculated = 1.26 <Zcritical = 1.96). From this point of view, it was revealed that the difference between the two sizes was statistically significant. Considering the magnitude of the classification rates obtained from both methods, it can be concluded that this difference is in favor of LR analysis. In other words, LR analysis was more accurate than CHAID analysis. According to this, it is concluded that LR analysis gives a more accurate result than CHAID analysis in determining the independent variables that have a significant effect on the dependent variable.

Another comparison of the obtained classification ratios was made between LR analysis and REPTree algorithm. The Z statistic calculated for the comparison of the correct classification ratios obtained from LR analysis and REPTree algorithm was greater than the critical value of the Z-statistic at the level of 0.05 (Z calculated = 3.29> Zcritical= 1.96). According to this result, there was no statistically significant difference between LR analysis and correct classification rates obtained from the REPTree algorithm. Therefore, considering the correct classification results, it can be thought that both methods are the alternatives of each other. However, it should be kept in mind that the independent variables that have a significant effect on the dependent variable differ according to the two methods.

Finally, the Z statistic calculated for the correct classification ratios obtained according to the CHAID analysis and REPTree algorithm was higher than the critical value of the Z-statistic at the level of 0.05 (Zcalculated = 2.02> Zcritical = 1.96). As in the LR analysis, there was no statistically significant difference between the correct classification ratios obtained from the CHAID analysis and the REPTree algorithm. This shows that the comments made upon the LR and REPTree algorithm can be made for the CHAID analysis and REPTree algorithm, too. In other words, any of the two methods may be preferred according to the correct classification results. However, two methods had different results in determining independent variables, which have a significant effect on the dependent variable.

Discussion, Conclusion and Recommendations

In this study, Logistic Regression, CHAID and data mining methods are used to determine the variables that predict the students' mathematics success in PISA. This study also aimed to investigate whether the significance level and order of importance of the independent variables in classifying the students concerning mathematics achievement differ according to the method used. As a result of this study, in accordance with the findings related to the first subproblem, it was concluded that the significance of the independent variables of interest, attitude, motivation, perception, self-efficacy, anxiety and study discipline on mathematics achievement differed according to the method used. According to LR analysis, while independent variables having a significant effect on the dependent variable are listed as self-efficacy, attitude, anxiety and working discipline, whereas predictive variables having a significant effect on the dependent variable according to CHAID analysis and importance of these variables are self-efficacy, attitude and study discipline. The predictive variables determined according to the REPTree algorithm used in data mining and the order of importance of these variables were determined as self-efficacy, attitude and anxiety.

According to the findings obtained from the analysis of the second sub-problem, classification rate results concerning mathematics achievement according to the independent variables of interest, attitude, motivation, perception, self-efficacy, anxiety and study discipline related to the mathematics course differed in size according to the method used. The highest correct classification rate belongs to the LR analysis, the second is CHAID analysis and the smallest classification result belongs to the REPTree algorithm. This result is different from the findings of the study conducted by Abessi and Yazdi (2015). Similarly, in the study conducted by Baran-Kılıçalan (2018), it was observed that there were very close correct classification ratios like the C5.0 method was 75.90%, the CHAID analysis was 75.40%, and the LR analysis was 75.10%. In the relevant studies, the success sequence concerning correct classification rates is in the form of a learning method based on data mining, CHAID analysis and LR analysis.

It is thought that the use of C4.5 and C5.0 algorithms, which are considered to be the previous version of the REPTree algorithm, is the cause of this difference. There are studies about different results obtained by different algorithms in the literature. However, the findings of the study conducted by the researchers in the relevant literature showed that the most successful method was LR and then CHAID analysis, and this is similar to by Şata and Çakan's (2018) study. When the results of similar studies in the literature are evaluated as a whole, it is seen that the logistic regression analysis has a better classification rate compared to CHAID analysis (Duran, Pamukçu and Bozkurt, 2014; Heckerd and Gondolf 2005; Kurt, Türe and Kurum, 2008).

However, it is thought to be one of the reasons for the low rate of classification obtained by the REPTree algorithm is that the decision tree is not limited to three levels as in the CHAID analysis in the SPSS program, and the release of the number of levels to be obtained for the tree. It was determined that the classification rate obtained by the REPTree method was lower as more precise classification was performed by increasing the number of levels. In this respect, the findings suggest that the sensitivity of the measurement results obtained has increased in a sense. This difference between the classification results obtained for LR and CHAID analysis was statistically significant, while there was no statistically significant difference between LR analysis and REPTree algorithm and CHAID analysis and REPTree algorithm. However, this result differs from the McCarty and Hastak's (2007) findings. They found that CHAID tends to be superior to RFM (recency, frequency, and monetary value) and logistic regression. The difference between the classification rates obtained from CHAID analysis and LR analysis in each of four different sample sizes is not significant.

In a similar study, Güldal and Çakıcı (2017) compared 70 students with Naive Bayes, Decision Tree (C4.5) and k-nearest neighbor method for k = 1, 3 and 5 according to different evaluation criteria. In their study, the highest accuracy rates were obtained by the nearest neighbor for k=3, J48, k01 and the nearest neighborhood for k=5, respectively and Naive Bayes methods. The accuracy values of the classification algorithms discussed in the classification of students as successful and unsuccessful with the help of different algorithms varied between 55.7% and 64.3%. In a similar study conducted by Mehdiyev, Enke, Fettke and Loos (2016), the findings showed that the most accurate predictions were performed by artificial neural networks, Random forest, Logistic regression, Radial based

networks and C4.5 methods, respectively. Accordingly, it can be said that the findings of the research are similar to the studies conducted in the field (Almuniri & Said, 2017).

Finally, in line with the findings obtained for the solution of the third sub-problem, the importance of independent variables in classifying students concerning mathematics achievement showed similarity depending on the method used. According to the obtained results, although the independent variables having a significant effect on the dependent variable are different according to the different methods, the order of importance of the variables did not change according to the method used. The self-efficacy variable, regardless of the method used in the study, is the variable that describes the dependent variable best. Similarly, no matter which method is used, the attitude variable is the variable with the best predictive power after the self-efficacy variable. Anxiety variable was not determined as a significant predictor variable in CHAID analysis, but according to the results of the LR analysis and REPTree algorithm, the predictive power ranking was followed by the attitude variable. Finally, the study discipline variable, which is not significant according to the REPTree algorithm, is the last predictor among the significant variables. This result is similar to the findings of Vale (2012). In this study, it is seen that the variables that have a significant effect on both methods, according to both methods, are the same in estimating the rates of automobile insurance using the CHAID analysis and Logistics model.

Based on the findings of this study, it is recommended that because of the CHAID, LR analysis and the REPTree algorithm give different results in determining the independent variables, which have a significant effect on the dependent variable; it is recommended that the studies which aim to determine the predictor variable should not be limited to one method. Combining findings from different methods serving the same purpose may be stronger evidence for research results. In addition, in studies where it is aimed to reveal the importance of the independent variables that have a significant effect on the dependent variable, LR and CHAID analysis and any of the REPTree algorithm may be preferred. This study is limited to Turkey's sample with 1000 students. Similar studies can be carried out on the data sets of different countries to provide new insights. This study is limited to REPTree, one of the data mining classification methods. In addition to this, it is suggested that comparative studies should be conducted to determine the level of similarity between CHAID and LR rather than a single data mining method.

References

Abessi, M., & Yazdi, E. H. (2015). Marketing data mining classifiers: Criteria selection issues in customer segmentation. *International Journal of Computer Applications, 106*(10), 5-10.

Almuniri, I., & Said, A. M. (2017). School's performance evaluation based on data mining. *International Journal of Engineering and Information Systems*, *1*(9), 56-62.

Albayrak, A.S., & Koltan-Yılmaz, Ş. (2009). Veri madenciliği: Karar ağacı algoritmaları ve imkb verileri üzerine bir uygulama. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14*(1), 31-52.

Antipov, E., & Pokryshevskaya, E. (2010). Applying CHAID for logistic regression diagnostics and classification accuracy improvement, *Journal of Targeting, Measurement and Analysis for Marketing*, *18*(2), 109 – 117. doi: 10.1057/jt.2010.3

Azen, R., &, Walker, C.M. (2011). *Categorical data analysis for the behavioral and social sciences* (1st Ed.), New York: Routledge.

Baran-Kılıçalan, M. (2018). *Hanehalkı işgücü araştirma verileri ile veri madenciliği yöntemlerinin uygulanmasi ve modellerin karşilaştirilmasi*. Yayımlanmamış Yüksek Lisans Tezi. Hacettepe Üniversitesi İstatistik Ana bilim Dalı, Ankara.

Baştürk, R. (2016). *Bütün yönleriyle SPSS örnekli nonparametrik yöntemler* (3. Baskı). Ankara: Anı Yayıncılık.

Bush S. (2015). Sample size determination for logistic regression: A simulation study. *Communications in Statistics - Simulation and Computation*, *44*, 360–373.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2016*). Bilimsel araştırma yöntemleri* (20. Baskı). Ankara: Pegem Akademi.

Can, Ş., Özdil, T., & Yılmaz, C. (2018). Üniversite öğrencilerinin ders başarısını etkileyen faktörlerin lojistik regresyon analizi ile tahminlenmesi. *International Review of Economics and Menagement*, *6*(1), 28-49.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education (6. Baskı)*. London: Routledge.

Çokluk Ö., Şekercioğlu G., & Büyüköztürk Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları*. Pegem yayınları, Ankara.

Díaz-Pérez, F. M., & Bethencourt-Cejas, M. (2016). CHAID algorithm as an appropriate analytical method for tourism market segmentation. *Journal of Destination Marketing & Management*, 5, 275-282. Doi: http://dx.doi.org/10.1016/j.jdmm.2016.01.006

Duran, A. E., Pamukçu, A., & Bozkurt, H. (2014). Comparison og data mining techniques for direct marketing campaings. *Sigma*, 32, 142-152.

Garson, D. (2015). *Missing values analysis and imputation methods*. USA: Statistical Publishing Associates.

Guldal, H., & Çakıcı, Y. (2017) Ders yönetim sistemi yazılımı kullanıcı etkileşimlerinin sınıflandırma algoritmaları ile analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 21*(4),1355-1367.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in house hold surveys. *Public Opinion Quartely*, *7*(5), 646-675

Heckert, A., & Gondolf, E. (2004). Battered women's perceptions of risk versus risk factors and instruments in predicting repeat reassault. *Journal of Interpersonal Violence*, 19, 778–800.

Kilic, S. (2000). *Lojistik regresyon analizi ve pazarlama araştırmalarında bir uygulama*. Yayınlanmamış Yüksek Lisans Tezi. İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Koc Basaran Y. (2017) Sosyal bilimlerde Örnekleme Kuramı, *Akademik Sosyal Araştırmalar Dergisi*, *5*(47), 480-495.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems xith Applications*, 34, 366-374.

Lehmann, E. L. (2006). On likelihood ratio tests, In IMS Lecture Notes- 2nd Lehmann Symposium, 49, 1–8.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. 2nd Edition. NewJersey, USA: John and Wiley Sons Incorporated,.

Massey, A., & Miller, S. J. (2006). *Tests of hypotheses using statistics*. Mathematics Department, Brown University, Providence, RI, 2912.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic Regression. *Journal of Business Research*, 60, 656–662.

Mertler, C.A., & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd Edition). Glendale, CA: Pyrczak Publishing.

Nisbet, R., Miner, G., & Yale, K. (2017). *Handbook of statistical analysis and data mining applications* (2nd Edition). Elsevier Science Inc., ISBN: 0124166326,9780124166325

Ozekes, S. (2003). Veri madenciliği modelleri ve uygulama alanları. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 2(3)*, 65-82.

Park, H. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, *43*(2), 154–164.

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

Rudd, J. M., & Priestley, J. L. (2017). A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Nonfinancial Payment Status in Commercial Credit, *Grey Literature from PhD Candidates*. 5. http://digitalcommons.kennesaw.edu/dataphdgreylit/5

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks, *Information Processing and Management*, 45, 427-437.

SPSS White Paper Inc. (1999). *Answer tree algorithm summary.* ATALGWP-0599, USA.

Sata, M., & Cakan, M. (2018). Comparison of results of CHAID analysis and logistic regression analysis. *Dicle University Journal of Ziya Gökalp Faculty of Education*, *33*, 48-56.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn& Bacon.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics.* USA: Pearson Education Limited.

Tutek, H., & Gumusoglu, Ş. (2008). *İşletme istatistiği,* İstanbul: Beta Basım Yayım Dağıtım A.Ş.

Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating forecasting methods by considering different accuracy measures. *Procedia Computer Science, 95*, 264 – 271.

Milli Eğitim Bakanlığı-Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, (2015). Uluslararası Öğrenci Değerlendirme Programı PISA 2015 Ulusal Raporu, Ankara.

Milli Eğitim Bakanlığı-Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, (2016). *Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS) 2015 Ulusal Matematik Ve Fen Ön Raporu*. Ankara.

Vale, J. B. (2012). *Using Data mining to predict automobile insurance fraud*. Dissertation of the degree of MSc in Business Administration. Universidade Católica Portuguesa.

Zuckerman, I., & Albrecht, D.W. (2001). Predictive statistical models for user modeling. *User Modeling and User Adapted Interaction*, *11*(1-2), 5-18.

Yordayıcı Değişkenlerin Belirlenmesinde Kullanılan Yöntemler: Lojistik Regresyon, Veri Madenciliği Yöntemleri ve CHAID Analizi

Atıf:

Aksu, G., & Reyhanlioglu Keceoglu, C. (2019). Comparison of results obtained from logistic regression, chaid analysis and decision tree methods. *Eurasian Journal of Educational Research*, *84, 115-134*, DOI: 10.14689/ejer.2019.84.6

Özet

Problem Durumu: Ülkelerin eğitim politikalarına yön vermek amacıyla göz önünde bulundurulan birçok durum vardır. Dünya genelinde politika belirleyicileri, kendi ülkelerindeki öğrencilerin bilgi ve beceri düzeylerini araştırmaya katılan diğer ülkelerdeki öğrencilerin bilgi ve beceri düzeyleriyle karşılaştırmak, eğitim düzeyinin yükseltilmesi amacıyla standartlar oluşturmak ve eğitim sistemlerinin güçlü ve zayıf yönlerini belirlemek amacıyla uygulanan uluslararası uygulamaların sonuçlarından yararlanılmaktadır. Ülkeler bu bilgiler sayesinde eğitim süreçlerini uluslararası bir perspektife göre değerlendirebilmektedir. Ülkelerin eğitim politikalarının şekillendirilmesinde önemli rol oynayan uluslararası sınavlardan elde edilen bulgular, farklı alanlarda değişkenlerin ölçüldüğü büyük ölçekli bir veri tabanından elde edilmektedir. Çok büyük ölçekli veriler, farklı alanlardaki büyük ölçekli veri tabanları içinde değerli verileri bulunduran bir veri madeni gibi düşünülebilir. Veri madenciliği yöntemleri sayesinde ülkelerin eğitim politikalarına yön veren uygulamalardan elde edilen karmaşık veriler üzerinden bağımlı değişkeni yordayan bağımsız değişkenlere dair maksimum bilgi elde edilebilir. Bağımlı (yordanan) değişkenin üzerinde etkili olan bağımsız (yordayıcı) değişkenlerin belirlenmesi bilimsel araştırmaların temel odağında yer alan konulardan bir tanesidir. Bu amaçla gerçekleştirilmiş çalışmalarda yordayıcı değişkenlerin belirlenmesinde çeşitli yöntemlerden yararlanılır. Bu yöntemlerin ortak özelliği bağımsız değişkenlerin bağımlı değişkenler üzerindeki etkilerinin anlamlılığını test etmesidir. Kullanılan yöntemlerin ortak özellikleri kadar birbirinden farklılaşan özellikleri de bulunmaktadır. Kullanılan yöntemleri birbirinden ayıran temel özelliklerden biri uygulanabildiği veri türüdür. İstatistiksel yöntemlerin bazıları sadece sürekli verilere uygulanabilirken, bazıları kategorik verilere de uygulanabilmektedir. Kategorik veri analizi eğitim uygulamalarında sıklıkla kullanılan bir yöntemdir. Her ne kadar öğrencilerin akademik başarılarını belirlemek için kullanılan ölçme araçları eşit aralık ölçek düzeyinde kabul edilerek, ölçme sonuçları sürekli puanlar olarak elde edilse de, öğrenciler hakkında karar verme sürecinde başarı puanları belli bir ölçüt puana göre başarılı/başarısız şeklinde kategorik verilere dönüştürülmektedir. Sonuç olarak bir bağımlı değişken olarak öğrenci başarıları üzerinde anlamlı etkiye sahip olan faktörlerin belirlenmesi için veri madenciliği ile parametrik olmayan iki yöntem olan Lojistik Regresyon analizi ve CHAID analizi yöntemlerinin sonuçlarından yararlanılabilir. Her üç yöntemin de ortak özelliği bağımlı değişken üzerinde anlamlı etkiye sahip olan bağımsız değişkenleri belirlemeyi hedeflemesidir. Bununla birlikte üç yöntemi birbirinden ayıran en temel özellik arka planda çalıştırdığı öğrenme algoritmasıdır. Tüm bunlara bağlı olarak başarı üzerinde anlamlı bir etkiye sahip olduğu düşünülen bağımsız değişkenlerin belirlenmesi ve bu değişkenlerin önem sırasının ortaya konulması birçok bilimsel çalışmanın ortak amaçlarından biridir. Ayrıca değişkenlerin önem sırasının kullanılan yöntemlere göre değişmesi çalışmalarda hangi yöntemin kullanılması gerektiği konusunda karışıklık yaratacaktır.

Araştırmanın Amacı: Çalışma kapsamında ele alınan üç farklı yönteme göre bağımsız değişken olarak kabul edilen matematik dersine ilişkin ilgi, tutum, motivasyon, algı, öz yeterlik, kaygı ve çalışma disiplini değişkenlerine göre öğrencilerin başarı durumları bakımından nasıl sınıflandıkları araştırılmıştır. Bu çalışmada öğrencilerin matematik

başarısını yordayan değişkenlerin belirlenmesi amacıyla Lojistik Regresyon (LR) ve CHAID analizi ile veri madenciliği yöntemlerinden yararlanılmaktadır. Mevcut bir durumun sonuçlarının belirlenmesi sebebiyle çalışma ilişkisel (korelasyonel) bir araştırma niteliğindedir. Çalışmada kullanılan veriler PISA 2012 öğrenci anketinde yer alan ve uygulamaya katılan öğrencilerin ilgi, tutum, özyeterlik, algı, motivasyon, kaygı ve çalışma disiplini alt ölçeklerine verdikleri yanıtlar yardımıyla elde edilmiştir. Çalışmanın evreni PISA 2012 öğrenci anketine katılan ve tabakalı seçkisiz örnekleme yöntemiyle belirlenen 4818 öğrenciden oluşmaktadır. Ancak analizler sistematik örnekleme yöntemi ile seçilmiş 1000 öğrenci üzerinden gerçekleştirilmiştir. Verilerin analizi LR ve CHAID analizi ile veri madenciliği yöntemlerinden REPTree algoritmasına göre gerçekleştirilmiştir. Böylece her üç yönteme göre öğrencilerin matematik başarısı üzerinde anlamlı etkisi olan bağımsız değişkenler belirlenmiştir. LR, CHAID analizi ve REPTree algoritması yöntemlerinin karşılaştırılması öğrencilerin başarı durumuna göre anlamlı etkisi olan değişkenlerin ve her bir yönteme ilişkin öğrencilerin matematik başarılarına göre doğru sınıflandırma oranlarının belirlenmesi ile gerçekleştirilmiştir.

Araştırmanın Bulguları: Elde edilen sonuçlara göre her bir yönteme ilişkin öğrencilerin matematik başarısı üzerinde anlamlı etkisi olan değişkenler birbirinden farklı çıkmıştır. Bunun yanı sıra her ne kadar farklı yöntemlere göre bağımlı değişken üzerinde anlamlı etkiye sahip olan bağımsız değişkenler farklı olsa da, değişkenlerin önem sırasının kullanılan yönteme göre değişmediği belirlenmiştir. Çalışmada ayrıca farklı yöntemler tarafından öğrencileri PISA matematik okuryazarlığı bakımından sınıflamada elde edilen doğru sınıflama oranlarının farklılık gösterdiği belirlenmiştir.

Araştırmanın Sonuçları ve Öneriler: LR analizine göre bağımlı değişken üzerinde anlamlı etkiye sahip olan bağımsız değişkenler özyeterlik, tutum, kaygı ve çalışma disiplini şeklinde sıralanırken, CHAID analizine göre bağımlı değişken üzerinde anlamlı etkisi olan yordayıcı değişkenler ve bu değişkenlerin önem sırası özyeterlik, tutum ve çalışma disiplini şeklindedir. Veri madenciliğinde kullanılan REPTree algortimasına göre belirlenen yordayıcı değişkenler ve bu değişkenlerin önem sırası ise özyeterlik, tutum ve kaygı şeklinde belirlenmiştir. En büyük sınıflandırma oranı LR analizi, ikinci olarak CHAID analizi ve en küçük sınıflandırma sonucu ise REPTree algoritmasına aittir. REPTree algoritması ile elde edilen sınıflama oranının düşük çıkma sebeplerinden bir tanesi karar ağacının SPSS programında gerçekleşen CHAID analizinde olduğu gibi 3 düzeyle sınırlandırmayarak ağaç için elde edilecek düzey sayısının serbest bırakılmasından kaynaklanabileceği düşünülmektedir. Çalışmada matematik başarısı bakımından öğrencileri sınıflandırmada bağımsız değişkenlerin önem sırası kullanılan yönteme göre benzerlik göstermiştir. Bağımlı değişken üzerinde anlamlı etkiye sahip olan bağımsız değişkenlerin önem sırasının ortaya konmasının amaçlandığı çalışmalarda LR ve CHAID analizi ile REPTree algoritmasından herhangi biri tercih edilebilir. Bireylerin başarı durumları açısından sınıflandırılmasının amaçlandığı çalışmalarda CHAID analizi ile REPTree algoritması birbirinin alternatifi olabilir. Ancak LR analizi diğer iki yönteme göre anlamlı derecede farklı sonuçlar vereceği için alternatif bir yöntem olarak düşünülmemelidir.

Anahtar Kelimeler: CHAID Analizi, Lojistik Regresyon Analizi, Veri Madenciliği, PISA