Düzce University Journal of Science & Technology

Research Article

A Cloud Based Web-Tool to Predict the High School Entrance Exam Scores of the Students

Gökçen ALTUN[a,*], Ekrem GÜLCÜOĞLU [b]

[a] *Department of Econometrics, Hacı Bayram Veli University, Ankara, Turkey*
[b] *Ministry of National Education, Kastamonu, Turkey*
** Corresponding author's e-mail address: gokcenefendioglu@gmail.com*
DOI: 10.29130/dubited.1535345

ABSTRACT

Multivariate adaptive regression splines (MARS) model, one of the non-parametric regression methods, is used to predict the achievement scores of the 8th-grade students before the LGS (High School Entrance System) exam with the developed web-tool. The demographic information of the students and all the test results they took in the last year are used before the LGS exam. The significant variables on the LGS scores of the students are the number of siblings, mother's education level, revolution history and Kemalism, English, mathematics courses. A web-based machine learning-based application has been developed to predict the LGS scores of the students in line with these data. The web-tool is accessible with the following website https://beststat.shinyapps.io/lgs2/. R Shiny program is used in the development of the web-tool. The program is cloud-based and works independently of the operating system and web browsers. The developed application helps students prepare for the LGS exam to offer pre-exam advice to guide their studies.

Keywords: *MARS, Machine learning, LGS, Student achievement, Data mining, R Shiny.*

Öğrencilerin Lise Giriş Sınavı Puanlarını Tahmin Etmek için Bulut Tabanlı Bir Web Aracı

ÖZ

Çok değişkenli uyarlanabilir regresyon splinleri (MARS) modeli, parametrik olmayan regresyon yöntemlerinden biri olarak, geliştirilen web aracıyla LGS (Liselere Geçiş Sistemi) sınavı öncesinde 8. sınıf öğrencilerinin başarı puanlarını tahmin etmek amacıyla kullanılmaktadır. Bu süreçte, öğrencilerin demografik bilgileri ve son bir yıl içinde girdikleri tüm sınav sonuçları LGS sınavı öncesinde dikkate alınmaktadır. Öğrencilerin LGS puanları üzerinde etkili olan önemli değişkenler arasında kardeş sayısı, annenin eğitim seviyesi, inkılap tarihi ve Atatürkçülük, İngilizce, matematik dersleri yer almaktadır. Bu veriler doğrultusunda, öğrencilerin LGS puanlarını tahmin etmek için web tabanlı, makine öğrenimine dayalı bir uygulama geliştirilmiştir. Web aracı, https://beststat.shinyapps.io/lgs2/ adresinden erişilebilir durumdadır. Web aracının geliştirilmesinde R Shiny programı kullanılmıştır. Program bulut tabanlıdır ve işletim sistemi ile web tarayıcılarından bağımsız olarak çalışmaktadır. Geliştirilen uygulama, öğrencilerin LGS sınavına hazırlık sürecinde onlara rehberlik etmek amacıyla ön sınav önerileri sunmaktadır.

Anahtar Kelimeler: *MARS, Makine öğrenimi, LGS, Öğrenci başarısı, Veri madenciliği, R Shiny.*

Received: 18/08/2024, Revised: 10/09/2024, Accepted: 30/12/2024

I. INTRODUCTION

In the recent years, technological advancements and changes have exerted a significant influence on learning processes. E-assessment systems emerge as alternative systems to traditional assessment methods. The evaluation and assessment of these systems conducted online have been extensively studied for their advantages in various research works (Bayrak and Yurdugül, 2015). The goal of e-assessment is to evaluate and measure student performance based on feedback obtained from students (Simpson, 2016). Identifying meaningful data within the plethora of complex information in e-assessment systems and making predictions in the most suitable model pose a significant challenge.

In modeling relationships between variables within an e-assessment system, regression methods can be employed. Regression analysis is a statistical method that attempts to model the functional structure between a dependent variable and one or more independent variables. The aim of regression analysis is to find the model that best explains the change in the dependent variable using the independent variables. In other words, regression analysis seeks to increase the correlation or reduce the difference between the predicted value and the actual value of the dependent variable (Kılıç, 2013). In regression analysis, parameter estimates are made using the Least Squares Method (LSM). The LSM estimation method is based on minimizing the difference between the observed dependent variable value and the estimated independent variable value.

One of the most important assumptions in regression analysis is the problem of multi-collinearity, which arises when there is high correlation among independent variables. In such cases, LSM estimates are not reliable, and various regression models have been developed to address the issue of multi-collinearity (Şahinler, 2000). Additionally, data mining models such as CHAID and CART, which could serve as alternatives to multiple regression models, can also be employed (Argüden, 2008). Studies utilizing data mining with the MARS model have been conducted in various fields, including economics (Albayrak and Yilmaz, 2009; Tunay, 2010; Tunay, 2011), health (Ekrem et al., 2020; Zakeri, 2010), engineering (Eyduran et al., 2019), and education (Şevgin, 2020). The MARS model, like the well-known multiple linear regression model, is based on maximizing the explained variance of errors.

Changes in computer and software technologies have facilitated the analysis of a large number of variables using complex algorithms such as the MARS model. Leathwick et al. (2006) examined environmental factors influencing the species distribution of freshwater fish using the MARS model. Significant environmental factors were processed onto a geographic information system to determine suitable habitats for freshwater fish. Nacar et al. (2020) aimed to predict dissolved oxygen concentration using the MARS and regression analysis (RA) based on temperature (T), specific conductivity (SC) data calculated from specific conductance (EC), pH, and flow rate (Q). They identified basis functions and equations producing the best prediction values with the MARS model. Orhan et al. (2018) modeled milk yield using the MARS model with control day, milking time, conductivity, and mobility as independent variables based on daily lactation records of 80 cows from 2006 to 2011. Sevimli (2009) combined data collected using the split-mouth design method in dentistry and, statistically, demonstrated an effective prediction model using the MARS model. Tosun (2021) conducted a study on 25 animals on a local farm in Şanlıurfa, including variables such as breed, age, along with monthly milk yield, lactation period, and recorded live weight data for June-September 2020. In their study on the relationship between macroeconomic indicators and economic growth in Turkey, Bağcı and Hoş (2021) ranked variables affecting economic growth in their MARS model according to their importance, including imports, unemployment rate, credit volume, interest rate, exchange rate, and inflation. Kartal et al. (2018) conducted a study on factors affecting the USD/TRY and EUR/TRY exchange rates in Turkey using monthly data from January 2006 to June 2017 and 12 independent variables with the MARS model. The developed MARS model was found to be within acceptable limits of explanatory power. As observed in the aforementioned studies, the MARS model has been successfully applied in diverse fields such as medicine, economics, agriculture, and animal husbandry.

MARS model is a frequently used model in the field of educational sciences. Depren (2018) examined the factors affecting the achievement of Turkish students in science with the MARS model using PISA 2015 data. Ahmed et al. (2022) estimated the Introductory Engineering Mathematics score values of students with the MARS model. Addini et al. (2023) examined the quality of education in Indonesia with the MARS model. Zurimi (2020) examined the factors affecting the study periods of students with the MARS model and found that the economic status of the family is one of the most important variables. Apart from the MARS model, machine learning methods are also used to predict student achievement. Bujang et al. (2021) obtained grade predictions of students with random forest, support vector machine and decision trees methods.

The rest of the study is organized as follows. In Section 2, the MARS model is summarized. In Section 3, the data set is analyzed by the MARS model and the results of the model are discussed. In Section 4, the developed web-tool is introduced. Section 5 contains the concluding remarks.

II. MULTIVARIATE ADAPTIVE REGRESSION SPLINES

The MARS model is one of the non-parametric regression methods. Unlike traditional regression models, the MARS model does not assume a functional form for the relationship between variables. The primary aim of the MARS model is to reveal the relationship between the dependent variable and independent variables using basis functions. The MARS model is a regression model that utilizes basis functions corresponding to knot points obtained from independent variables. The relationship between the dependent variable and independent variables is explained functionally based on the obtained knot points. Thus, the MARS model stands out as a suitable statistical model for datasets with a multivariate structure (Temel et al., 2005).

Compared to traditional methods, the MARS model offers the advantage of optimal data transformation and the ability to determine interactions in complex datasets. Developing a robust regression model, even for small datasets, requires considerable time and effort. However, in conjunction with the MARS model, regression models can be easily developed for large databases and highly complex data structures (Tunay, 2001). The MARS model is a flexible and fast model that can be used both as a regression model and a binary classification model.

In the MARS model, unlike the known regression model, subsets formed by input variables are evaluated (Xu et al., 2006). Eventually, in the universe formed by predictive variables, extension functions are created, divided into many coherent regions, and these regional regression curves are expressed as basis functions (Put et al., 2004). In the MARS model, variables are divided into regions, and appropriate basis functions and their coefficients are determined through transformations. The created basis functions are in a linear relationship with the dependent variable. The MARS model can easily evaluate non-monotonic relationships between the dependent and independent variables, providing robust models (Olecka, 2007). The process of creating the equation within the MARS model can assess complex relationships in large datasets and achieve better results than other linear and parametric methods (Xu et al., 2006).

The setup of the MARS model is carried out in 2 steps (Kolyshkina et al., 2004). In the first step, all possible basis functions are created until the largest model is found. Basis functions can consist of a single variable or the interaction of multiple variables. Functions evaluating interactions and non-linear transformations of independent variables in the model constitute the basis function. While creating the basis function, all dependent and independent variables, as well as combinations of variables, are individually evaluated. Each basis function also generates a mirror image function. However, the mirror image function has no effect on the model as its slope at the relevant knot point is zero (Çelik et al., 2018). In the second step, the largest model obtained in the first step is reduced to the optimal model through a pruning process. The algorithm used in this process is the backward stepwise algorithm.

In the MARS model, rather than assuming a linear relationship between variables, modeling is carried out by considering the non-linear functions of variables individually and their linear combinations. Basis functions are employed for this purpose. Extensions made with basis functions are used to predict an appropriate model. The knot points within the defined intervals of the dataset signify the beginning and end of a data point. Therefore, knot points are where the function's behavior changes (Everingham and Sexton, 2011). When constructing the MARS model, the last value where the slope of the line does not change in the interval where the independent variable value is located is taken as the knot point. In defining the MARS model, basis functions defined at knot points are utilized. Figure 2.1 illustrates the correlation distribution between the independent variable X and the dependent variable Y. Considering the distribution of the independent variable X at the knot points, it is observed that the slope of the function changes at the x_1 variable's knot point. Similarly, at the x_2 knot point, the slope of the function has also changed.

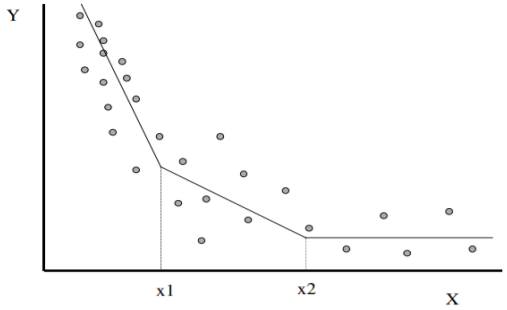

Figure 2.1. *Knot points for the function y=f(x) (Briand et al., 2004)*

In the basis function, the variable x is derived in a piecewise linear structure. Below, the piecewise linear basis functions $(x-t)_+$and $(t-x)_+$ are provided. Note that the symbol "+" denotes the positive side.

$$(\mathrm{x}-\mathrm{t})_+ = \begin{cases} x-t, & x > t \\ 0 \end{cases} \tag{2.1}$$

$$(\mathrm{t}-\mathrm{x})_+ = \begin{cases} t-x, & x < t \\ 0 \end{cases} \tag{2.2}$$

The sum of the basis functions is given in Equation 2.3.

$$C = \{(X_j - t), (t - X_j)\} \qquad t \in \{x_1, x_1, \dots, x_{Nj}\} \qquad j = 1, 2, \dots, p. \tag{2.3}$$

The knot value indicates the end of any region of data and the beginning of another. The MARS model is determined by using the number of knot values that yield the minimum prediction error (Sabancı, 2019).

In Figure 2.2, the relationship between the dependent and independent variables for the MARS model is depicted.

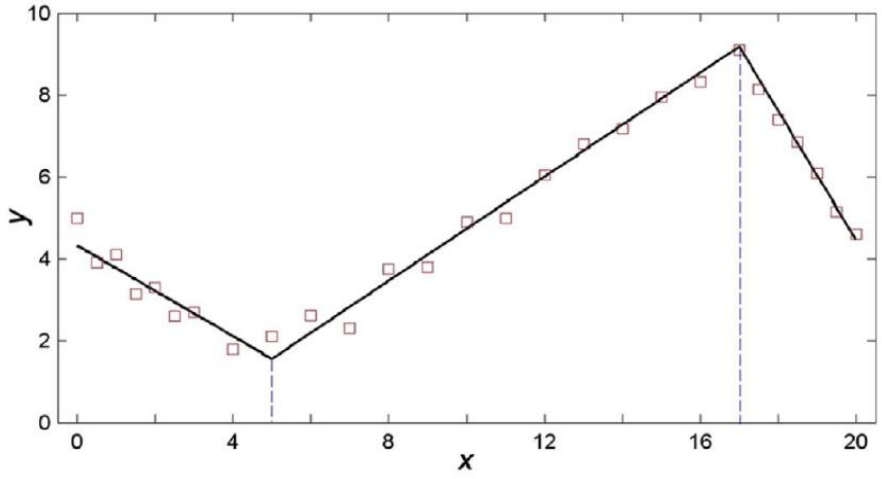

Figure 2.2. *Function with two knot points belonging to the MARS model (Zhang and Goh, 2016)*

Based on the Figure 2.2, the knot values for x=5 and x=17 are depicted. Corresponding to these knot values, there exist three basis functions, which are provided in Equation 2.4.

$$Max(0,5-x), \ 0<x<5 \qquad Max(0,x-5), 5<x<17 \qquad Max(0,x-17), x>17 \tag{2.4}$$

A. FORWARD SELECTION

The first stage of the MARS model is forward selection. This stage resembles the first stage of a stepwise regression model. The distinction lies in the use of basis functions as input variables (Hill and Lewicki, 2006). The Greedy algorithm is employed to find the pair of basis functions with the smallest sum of squared errors. In MARS forward selection, to add a new basis function, it evaluates existing main terms; then, it evaluates all variables to choose the new basis function, and finally, it assesses combinations of all observed values for each variable to determine the last knot value (Strickland, 2015).

It employs an intuitive technique in each forward selection to minimize the number of main terms to be evaluated. The number of terms added in the forward selection process is up to the maximum number of terms entering the model, and this number is determined by the user. Therefore, it is crucial for the user to accurately determine this number (Friedman, 1993). The MARS model, as formulated by Friedman (1991), is expressed as follows.

$$f(x) = a_0 + \sum_{m=1}^{M} a_m B_m(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{m=1}^{K_m} [S_{km}(x_{v(k,m)} - t_{km}) \tag{2.5}$$

According to the above function, the number of basis functions is denoted as M, and the MARS model can be formulated as $m = 1,2,\dots,M$. The value K_m represents the number of interactions. The value S_{km} takes the values ± 1 and a_0 is the constant term. Regression coefficients are represented by a_m. $B_m(x)$ denotes the basis function (Friedman, 1991).

B. BACKWARD ELIMINATION

The second stage of the MARS model is constituted by the backward elimination stage. The aim in this stage is to reduce the complexity of the model created in the first step. In the backward elimination process, the largest model obtained in the first stage is considered. This model, due to harboring the problem of overfitting, does not yield statistically reliable results in the test set (Hill and Lewicki, 2006). The backward elimination algorithm is a continuation of the forward selection algorithm.

Through the backward elimination algorithm, the overfitting model undergoes a pruning process. The best subset is obtained using the GCV criterion. The GCV value takes into account the model complexity and the HKT value. Basis functions minimizing the GCV value are identified. A model with a smaller GCV value is considered the best model (Briand et al., 2004; Hastie et al., 2009; Hill and Lewicki, 2006). GCV is calculated using the following formula (Friedman, 1991).

$$GCV = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}(x_i)\right]^2 / [1 - \frac{C(M)}{N}]^2 \tag{2.6}$$

The value N, as presented in Equation 2.6, denotes the number of observations in the dataset. $C(M)$ is the cost-complexity measure of a model containing M basis functions, excluding the constant basis function (the a_0 coefficient in the MARS model), and is defined as follows (Orhan et al., 2018).

$$C(M) = M + dM \tag{2.7}$$

Here, d represents a smoothing parameter in the creation process of the MARS model, accompanied by a cost optimization for each basis function (Chou et al., 2004; Leblanch and Tibshirani, 1994).

III. APPLICATION

The study was conducted with the participation of all 68 students at Mehmet Akif Ersoy Middle School in Tosya district, Kastamonu province, at the 8th-grade level. Three different sources of data were utilized within the scope of the research, namely the exam evaluation system, student information form, and the results of the students' High School Entrance Exam (LGS). The data are available from the authors upon request. The exam evaluation system, during the academic year 2020-2021, read the optical form information of the students from the trial exams, recording the number of correct and incorrect answers for each student in subjects such as mathematics, Turkish, religious culture and ethics, history of revolution and Kemalism, English, and science. Throughout the semester, students took a total of 34 trial exams.

A. DESCRIPTIVE STATISTICS FOR DEMOGRAPHIC VARIABLES

Descriptive statistics for demographic variables such as gender, number of siblings, and parents' educational levels are provided in Table 3.1.

Table 3.1. *Descriptive Statistics for Demographic Variables*

Variables	Level	n	%
Gender	**Male**	37	54.4
	Female	31	45.6
The number of siblings	**none**	4	5.9
	1	32	47.1
	2	19	27.9
	Three or more	13	19.1
Mother's education level	**Elementary school**	19	27.9
	Middle school	14	20.6
	High school	20	29.4
	University	12	17.6
	Master's degree	3	4.4
Father's education level	**Elementary school**	11	16.2
	Middle school	13	19.1
	High school	23	33.8
	University	18	26.5
	Master's degree	3	4.4
Social activity participation	**Yes**	36	52.9
	No	32	47.1
Reading books	**Yes**	60	88.2
	No	8	11.8
Private room	**Yes**	58	85.3
	No	10	14.7
Computer/Tablet	**Yes**	61	89.7
	No	7	10.3
Mobile phone	**Yes**	53	77.9
	No	15	22.1
Grade repetition	**Yes**	12	17.6
	No	56	82.4
Academic support	**Yes**	26	38.2
	No	42	61.8

As indicated in Table 3.1, 54.4% of the students are male, while 45.6% are female. Four levels have been used for the information on the number of siblings. In terms of the number of siblings, 4 students (5.9%) have no siblings. Nearly half of the students (47.1%) have only one sibling. In terms of the mothers' education level, 29.4% of student mothers graduated from high school, 27.9% from elementary school, 20.6% from middle school, 17.6% from university, and 4.4% from master's degree programs. Regarding the fathers' education level, 33.8% of students' fathers graduated from high school, 26.5% from university, 19.1% from middle school, 16.2% from elementary school, and 4.4% from master's degree programs. More than half of the students (52.6%) participate in social activities. The majority of students (88.2%) have a habit of reading books. More than half of the students (85.3%) have their own private rooms. Only 7 students (10.3%) do not have a computer and tablet, while 15 students (22.1%) do not have a mobile phone. The rate of students repeating a grade is 17.6%, while 56 students (82.4%) have never repeated a grade up to the 8th grade. In terms of academic support, more than half of the students (61.8%) take private lessons outside of school, while 38.2% of students have not taken any private lessons.

B. DESCRIPTIVE STATISTICS FOR STUDENTS' ACADEMIC ACHIEVEMENTS

Students took a total of 34 trial exams from September 2020 to June 2021. Descriptive statistics for the calculated average net values related to students' academic achievements and the descriptive statistics values for their scores in the High School Entrance Exam (LGS) are presented in Table 3.2.

Table 3.2. *Descriptive statistics for students' academic achievements*

Variables	Minimum	Maximum	Mean	Median	Standard deviation	Variance
Turkish course average score	-0.144	17.344	10.168	11.472	5.134	26.355
Mathematics course average score	-0.709	15.647	5.533	4.353	4.988	24.875
Revolution history and Kemalism course average score	-0.367	9.093	5.038	6.104	2.831	8.013
Religious Culture and Ethics course average score	-0.332	9.286	5.774	6.531	2.777	7.714
English course average score	-1.000	9.587	4.913	5.209	3.506	12.292
Science course average score	-0.259	17.656	8.683	8.653	5.875	34.516
LGS Score	190.855	451.915	318.771	190.855	78.877	6221.643

As indicated in Table 3.2, the subject with the smallest variance is the course of Religious Culture and Ethics, while the subject with the largest variance is the Science course. It can be inferred that the achievement variable in the Science course has a higher level of heterogeneity compared to other subjects. In comparison to the Science course, it can be stated that the achievement variable in the Religious Culture and Ethics course is more homogeneous. Since the number of questions in Turkish, Science, and Mathematics courses is equal (20 questions each), it is determined that students are more successful in the Turkish course based on the average values. As the number of questions in the History of Revolution and Kemalism, Religious Culture and Ethics, and English courses is also equal (10 questions each), it is found that students are more successful in the Religious Culture and Ethics course among these four subjects.

C. NORMALITY ANALYSIS OF STUDENTS' ACADEMIC ACHIEVEMENTS

The normality analysis results of the average values of the trial exam results taken by students throughout the academic year are presented in Table 3.3. As seen in Table 3.3, the relevant variables do not satisfy the assumption of normal distribution (($p < 0.05$). When the skewness values are analyzed, it is seen that all variables are left skewed except mathematics course average score. All kurtosis values are smaller than 0, so the distributions of variables are platykurtic.

Table 3.3. *Skewness, kurtosis coefficients, and normality test results for variables related to students' academic achievements*

Variables	**Skewness**	**Kurtosis**	**Shapiro-Wilk**
Turkish course average score	-0.507	-0.978	0.922 (< 0.001)
Mathematics course average score	0.452	-1.124	0.897 (< 0.001)
Revolution history and Kemalism course average score	-0.447	-1.196	0.910 *(<0.001)*
Religious Culture and Ethics course average score	-0.841	-0.417	0.884 *(<0.001)*
English course average score	-0.182	-1.538	0.889 *(<0.001)*
Science course average score	-0.061	-1.547	0.903 *(<0.001)*
LGS Score	-0.013	-1.449	0.923 *(<0.001)*

D. RELATIONSHIP BETWEEN LGS SCORES AND TRIAL EXAMS

The correlation between students' LGS scores and the trial exams they took in the 8th grade is presented in Figure 3.2. Trial exams refer to the exams that students take at school in preparation for the LGS exam. Since the variables are non-normal, we use the spearman correlation coefficient.

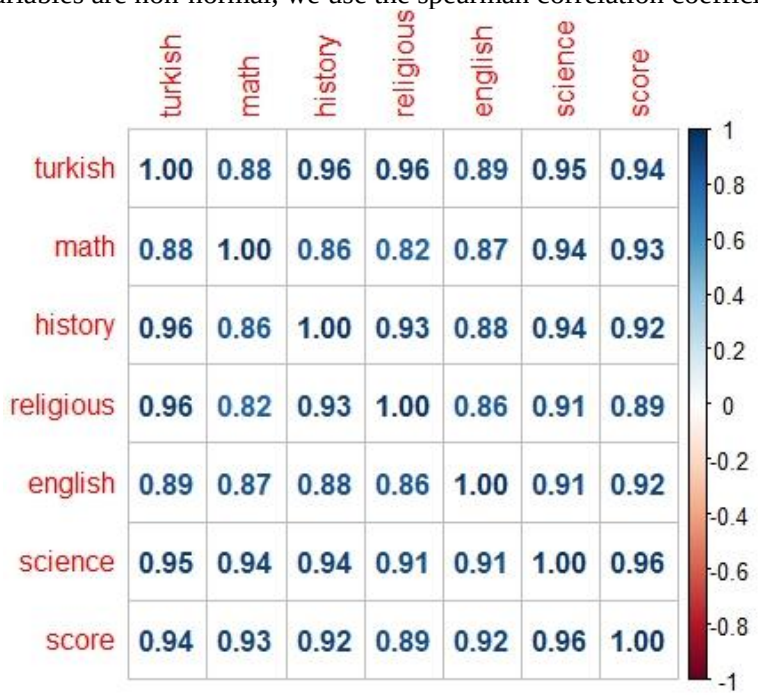

Figure 3.2. *Relationship between students' LGS scores and trial exam achievements*

As indicated in Figure 3.2, when the relationship between successful subjects in the trial exams and the LGS score is assessed, the correlation between the achievement in the science course and the LGS score is higher compared to other subjects (0.96). The success in the science course is followed by Turkish (0.94) and mathematics (0.93) courses, respectively. Mathematics is followed by equal-weighted History of Revolution and Kemalism and English courses. In the trial exams, the relationship between the success in the Religious Culture and Ethics course and the LGS score is the lowest compared to other subjects, and the correlation coefficient is calculated as 0.89.

E. MARS MODEL RESULTS

The parameter estimates for the MARS model were obtained using the R program, specifically utilizing the 'earth' package in the program. Initially, it is crucial to estimate the tuning parameters of the model, which include the degree of interaction and the number of terms. Therefore, a grid search algorithm and k-fold cross-validation were employed to determine the optimal values for these two parameters. During this process, the prediction error was minimized, and the value of k was set to 10. The 'caret' package was utilized for this procedure. The obtained results are presented in Figure 3.3. According to the results provided in Figure 3.3, the optimal interaction degree was determined to be 1, and the number of terms was found to be 7.

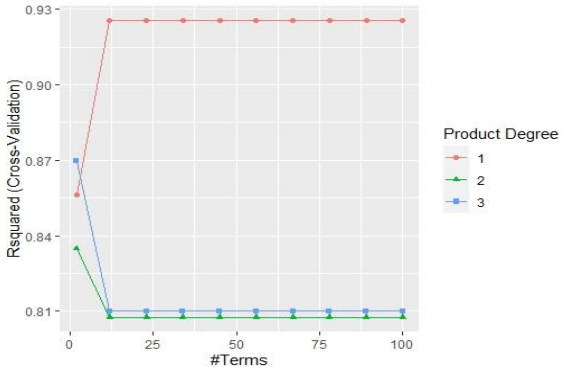

Figure 3.3. *Tuning parameters obtained through cross-validation and grid search algorithm for the MARS model.*

Once the tuning parameters of the MARS model were determined, the model parameters were obtained using the 'earth' package. The results are presented in Table 3.5.

Table 3.5. *Parameter estimates and basis functions of the MARS model.*

Parameter Estimation	Basis Function
285.597	Constant term
-16.241	The number of siblings (3 and more)
13.728	Mother's Education Level (Middle School)
-11.965	max(0, 5.936 - Average Net Number in Mathematics Course)
4.060	max(0, Average Net Number in Mathematics Course - 5.936)
9.610	max(0, Average Net Number in Revolution history and Kemalism Course - 1.730)
5.504	max(0, Average Net Number in English Course - 1.466)
$GRSq$	0.933
R^2	0.954
GCV	423.003
RSS	18817.430

When examining Table 3.5, according to the obtained results, the number of basis functions is 6. The MARS model automatically performs the variable selection process. According to the model results, the significant variables are: number of siblings, mother's education, average score in mathematics course, average score in history and Kemalism course, and average score in English course. The constant term value is determined as 285.597. When the number of siblings is 3 or more, the parameter estimate value is -16.241. It is observed that student achievement decreases when the number of siblings is 3 or more. When the mother's education is at the middle school level, the parameter estimate value is 13.728. Middle school graduates positively influence student achievement. The inflection point for the

mathematics course is found to be 5.936. If the student's average net score in the mathematics course is less than 5.936, the parameter estimate coefficient is determined to be -11.965. The situation identified here is that having an average net score in the mathematics course less than 5.936 negatively affects student achievement. However, when the average net score in the mathematics course is greater than 5.936, the parameter estimate coefficient is 4.060. Student achievement is positively influenced when the average net score in the mathematics course exceeds 5.936. The inflection point for the history and Kemalism course is found to be 1.730. When the average net score for this course exceeds 1.73, the parameter estimate value is 9.610. For the English course, the inflection point is found to be 1.466. When the average net score in English exceeds 1.466, the parameter estimate value is 5.504. The GCV value is calculated as 423.003. The GRSq and R2 values are obtained as 0.933 and 0.954, respectively. The RSS value is determined to be 18817.430. Figure 3.4 presents the residual analysis of the MARS model. Note that the definition of the GCV can be found in Çanga (2022).

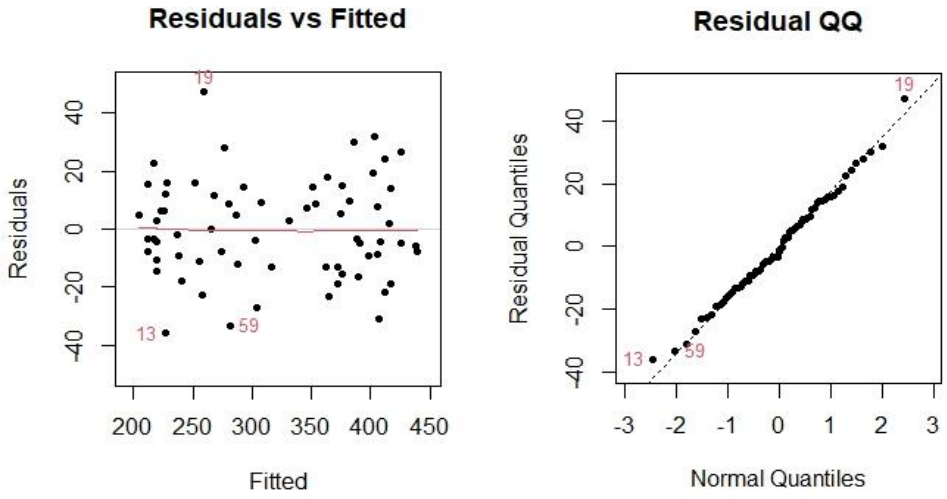

Figure 3.4. *Examination of model fit for the MARS model*

As indicated in Figure 3.4, a residual analysis has been conducted. Three observations can be considered as potential outlier observations. These observations are removed from the model and model success is re-evaluated. Since no significant improvement in model performance is achieved by removing the relevant observations from the data, it is decided to include these observations in the data set. Figure 3.5 provides graphs of the basis functions.

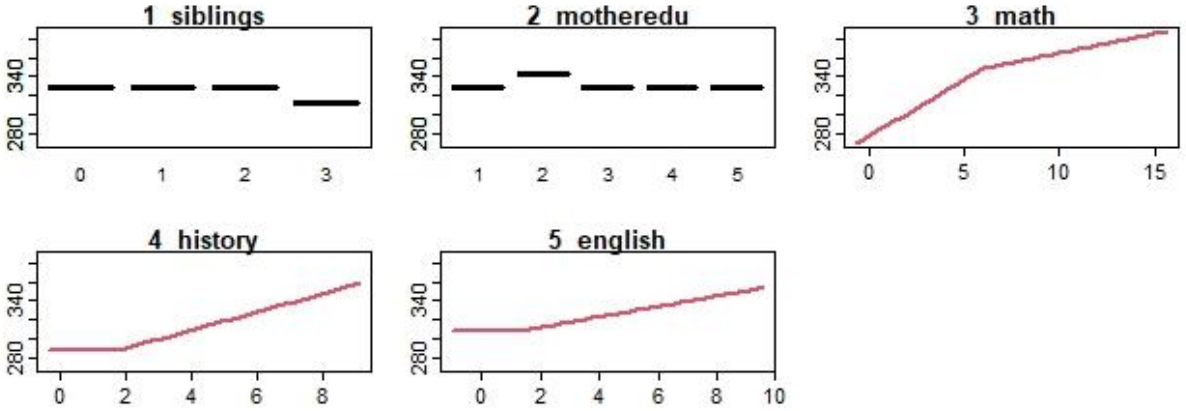

Figure 3.5. *Graphical examination of basis functions*

The knot points for the 5 significant variables are summarized in the graphs in Figure 3.5. As indicated in Figure 3.5, when the number of siblings is 0, 1, and 2, it does not affect the student's LGS score. However, when the number of siblings exceeds 2, the student's LGS score decreases. Therefore, having 3 or more siblings has a negative impact on the LGS score. Additionally, when the mother's education

level is at the middle school, it has a positive effect on the LGS score compared to other education levels. The average net score of the mathematics course has two breakpoints on the LGS score. The slope of the regression line changes at the value of 5.936. Having a mathematics net score below 5.936 negatively affects the LGS score, while having a net score above this value positively influences the LGS score. The average net score in the History of Revolution and Kemalism course does not affect the LGS exam until 1.73, but beyond this value, it positively influences the LGS exam. The regression graph for the English course is similar to the graph of the History of Revolution and Kemalism course. The average net score in the English course does not affect the LGS exam until 1.466, but beyond this value, it positively influences the LGS exam. In Figure 3.6, the importance levels of the significant variables are given according to the GCV and RSS criteria.

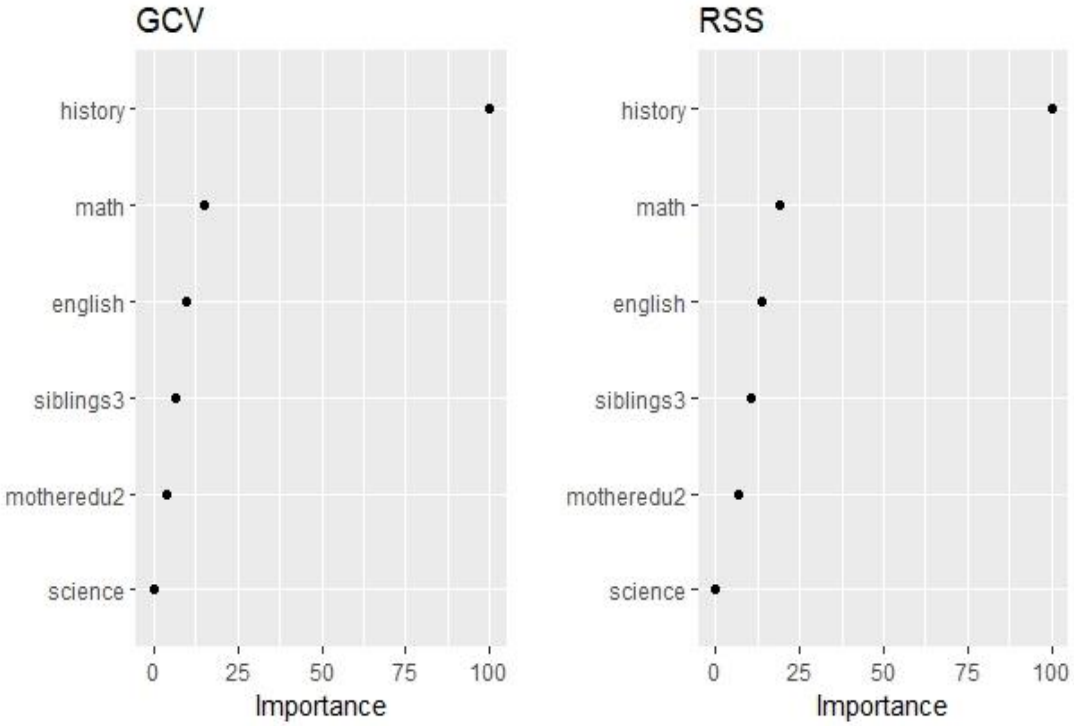

Figure 3.6. *Importance Levels of Variables Based on GCV and RSS*

In Figure 3.6, the significance levels of variables, along with their importance degrees, have been determined. This was calculated based on RSS or GCV values. Ultimately, the average net score in the History of Revolution and Kemalism course was found to be the most important variable. The ranking of variables in terms of their importance on the LGS score is as follows:

- Average net score in the History of Revolution and Kemalism course
- Average net score in the Mathematics course
- Average net score in the English course
- Number of siblings
- Mother's education level
-

In Figure 3.7, the predicted values of the MARS model are compared with the actual values. As depicted in Figure 3.7, the red squares represent the values predicted by the MARS model, while the black dots indicate the actual values. The predicted values by the MARS model closely align with the actual values, demonstrating the model's success in modeling the relevant data.

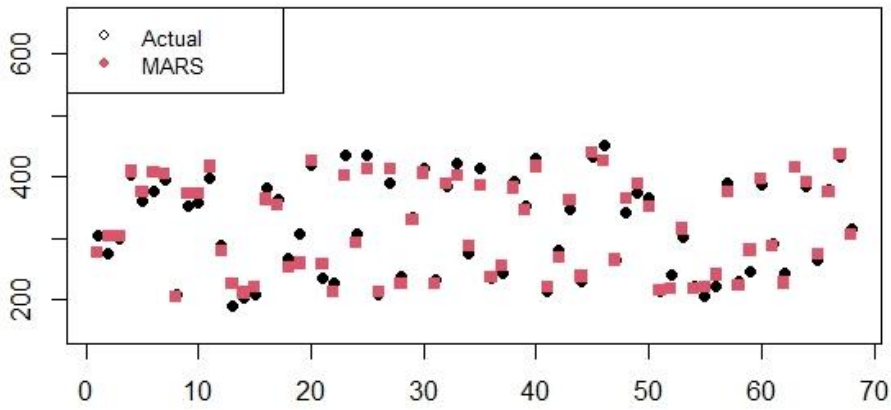

Figure 3.7. *Comparison of Prediction Values of the MARS Model with Actual Values*

IV. DEVELOPED CLOUD-BASED APPLICATION

In this section, we introduce the developed cloud based web-tool. The web-tool is developed using the shiny package of the R software. We use the the MARS model to produce the estimated LGS score based on the input data for students. The user interface of the web-tool is displayed in Figure 4.1. The web-tool is accessible with the following website https://beststat.shinyapps.io/lgs2/.

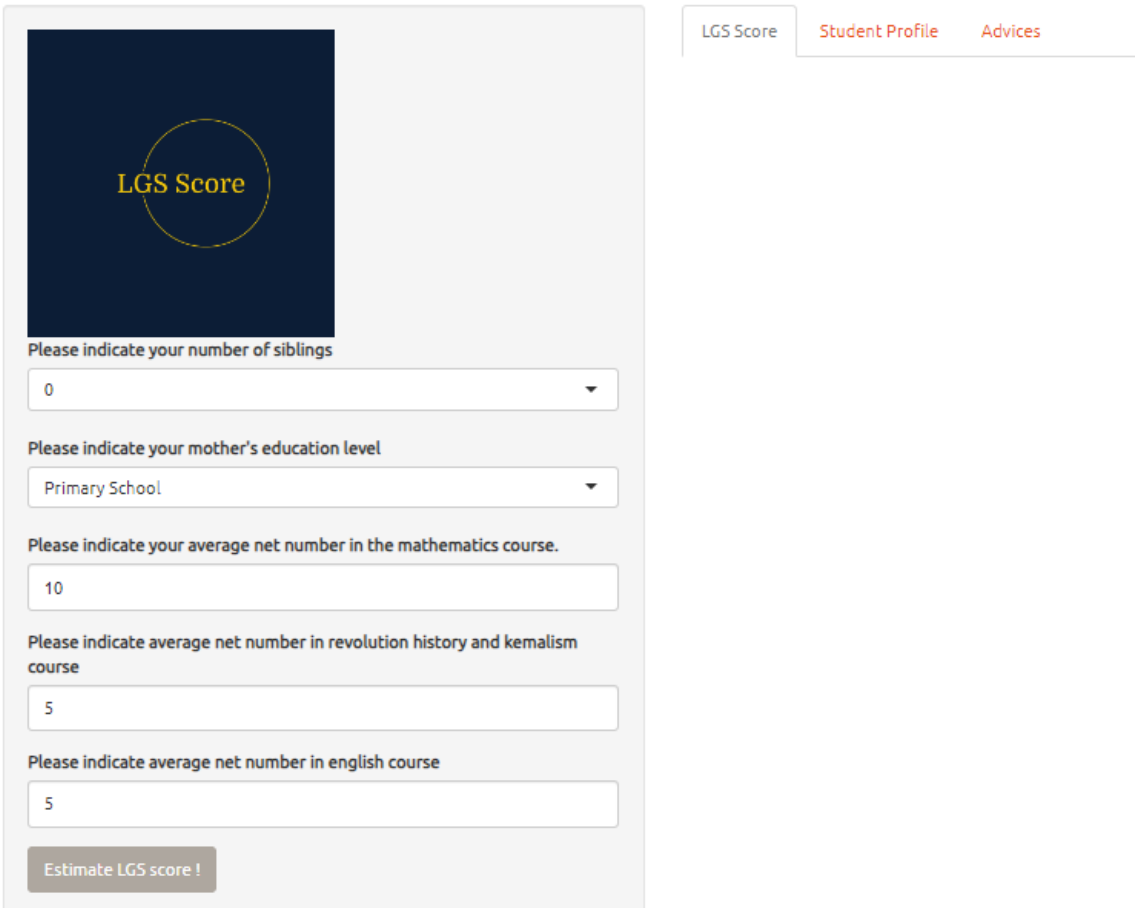

Figure 4.1. *The user interface of the developed web-tool.*

According to the results of the MARS model, the five variables are found statistically significant on the performance of the LGS scores of the students. Therefore, the web-tool considers only these variables

to estimate the LGS score based on the MARS model. In the right panel of the web-tool, there are three tabs. These are LGS Score, Student Profile and Advices. After all the necessary input variables are provided, the user runs the model by clicking the "Estimate LGS score! ". After that, the estimated LGS score, student profile and advises are shown at the right panel of the web-tool.
The advises are provided based on the obtained knot values. For instance, the knot value for the mathematics variable is 5.936. When the net value of the student for the mathematics course is less than 5.936, it affects the LGS score badly. So, the system generates an advice such as *"You need to spend more time in math class!*". The other advises are also generated by using the knot values of the predictor variables. The advises are listed below.

Conditions	**Advices**
Math<5.936, revolution history>1.730 and english>1.466	You need to spend more time in math class!
Math>5.936, Revolution History<1.730, English>1.466	You need to study more in the revolution course!
Math>5.936, Revolution History>1.730, English<1.466	You need to study more in English class!
Math<5.936, Revolution History<1.730, English>1.466	Study mathematics and revolution more!
Math<5.936, Revolution History>1.730, English<1.466	Study math and English more!
Math>5.936, Revolution History<1.730, English<1.466	Study revolution and English more!
Math<5.936, Revolution History<1.730, English<1.466	You need to increase your score by working more efficiently!
In other cases	Keep going like this!

V. RESULTS AND RECOMMENDATIONS

Considering the non-linear relationship between datasets obtained in the field of education, it has been deemed appropriate to utilize the MARS model, a non-linear and non-parametric prediction method. The MARS model attempts to explain the dependent variable by evaluating both the explanatory variables individually and their interactions. Moreover, it seeks to describe the created model with the basis function it generates. A web application based on the MARS model, relying on machine learning, has been developed to predict students' scores in the LGS exam.

To assess the prediction performance of the MARS model, goodness-of-fit criteria ($GRSq$, R^2, GCV, RSS) were employed. The constructed MARS prediction model utilized 7 terms, including the constant term, number of siblings, mother's education level, and average net scores in trial exams for Mathematics, English, and the History of Revolution and Kemalism course.

The calculated goodness-of-fit criteria, based on the lowest GCV (423.003)and RSS (18817.430), resulted in satisfactory values for GRSq (0.933) and cross-validation R^2 (0.954). It was observed that when the mother's education level is at the middle school level, student achievement increases (parameter estimate value was found to be 13.728 when the mother's education level is at the middle school level). This positively influences the student's LGS success. However, when the mother's education level is at the primary school, high school, college, or postgraduate level, the student's LGS success is not affected. Sociologically, this can be interpreted as follows: when the mother is a primary school graduate, she is likely unable to participate in the relevant educational activities, with a very low literacy rate in her family, or insufficient importance given to education within the family. For middle school graduate mothers, the situation can be interpreted as follows: a mother who has not achieved her

own goals can positively contribute to her child's education by providing support, thereby positively influencing the child's goals. This support can be evaluated as following the child's assignments or courses, motivating the child, and so on. On the other hand, mothers who are graduates of high school, college, or postgraduate programs are heavily involved in professional life (Keskin, 2018). Due to their extensive participation in professional life, they may have difficulty following their students. Therefore, they may not be able to show the necessary importance to their child's education. This is because they allocate more of their working hours to professional life. The time they spend with their child is less. This situation does not significantly positively affect the student's success.
In the MARS model developed within the scope of this study, when the number of siblings is 3 or more, the parameter estimate value was -16.241. Student achievement is negatively affected when the number of siblings is 3 or more. This is an expected result, as a large number of siblings may lead to a crowded environment, where there is not enough suitable space for studying, resulting in a decrease in the student's success.

It was found that success in the History of Revolution and Kemalism course positively affects the student's LGS success (parameter estimate value was 9.610 for the function max(0, History of Revolution and Kemalism average - 1.730)). This result supports the study conducted by Gençtürk (2001) on the factors affecting the diploma grades of high school students. According to the findings obtained, to positively influence LGS success in trial exams, the average net score in the History of Revolution and Kemalism course should be above 1.73.

Success in the English course was found to positively affect the student's LGS success (parameter estimate value was 5.504 for the function max(0, English average - 1.466)). This result supports the studies of Baş and Beyhan (2012) and Kazazoğlu (2013), indicating that success in the English course positively affects students' academic achievements. According to the findings, to positively influence LGS success in trial exams, the average net score in the English course should be above 5.504.
The Mathematics course was found to affect LGS success in two ways. When the net score in the trial exams for Mathematics is below 5.936, it negatively affects the LGS score (parameter estimate value was -11.965 for the function max(0, 5.936 - Mathematics average)). When the net score in the Mathematics course is above 5.936, the LGS success is positively affected (parameter estimate value was 4.060 for the function max(0, Mathematics average - 5.936)). This result supports studies indicating a significant and positive effect of mathematical success on students' problem-solving skills and success in other courses (Özsoy, 2005; Şentürk, 2010). This result suggests that not solving any questions in the mathematics course can negatively affect LGS success.

In contrast to studies in the literature, in this study, a new prediction equation was realized using the MARS model to predict the LGS scores of students based on machine learning. The developed machine learning-based web application has been published in an interactive format. Since the study is limited to a single school, the obtained findings are limited to the school where the study was conducted. For students, the recommendations in the advice table of the web application can be improved by working with subject matter experts. The results obtained in the study have shown that the MARS model could be an important option for predicting the LGS scores of middle school students.

Funding The authors state that this work has not received any funding.
Data availability The authors do not have permission to share data.
Declarations
Conflict of interest The authors of this manuscript declare no conflicts of interest.
Ethical approval Not applicable.
Informed consent Not applicable.

VI. REFERENCES

[1] S. Albayrak, ve Ş. Koltan Yılmaz. "Veri madenciliği: Karar ağaci algoritmalari ve İMKB verileri üzerine bir uygulama", *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi* , vol. 14, no. 1, pp. 31-52, 2009.

[2] Addini, P. F., Hadi, W., & Harahap, P. M. R. "Application of the multivariate adaptive regression spline (Mars) method in analyzing misclassification of elementary school accreditation data in the city of Tebing Tinggi", *Journal Scientia*, vol. 12, no. 1, pp. 617-620, 2023.

[3] Ahmed, A. A. M., Deo, R. C., Ghimire, S., Downs, N. J., Devi, A., Barua, P. D., & Yaseen, Z. M. "Introductory engineering mathematics students' weighted score predictions utilising a novel multivariate adaptive regression spline model", *Sustainability*, vol. 14, no. 17, pp. 11070, 2022.

[4] A. Yılmaz., *Veri madenciliği: Veriden bilgiye, masraftan değere*, ARGE danışmanlık, 2008.

[5] Bağcı, B., ve Hoş, S. "Türkiye'de Ekonomik Büyümenin Makroekonomik Göstergeler İle İlişkisi: MARS Modeli", *Ekonomi İşletme ve Maliye Araştırmaları Dergisi*, vol. 3, no. 2, pp. 193-202, 2021.

[6] Baş, G., ve Beyhan, Ö. "İngilizce Dersinde Yansıtıcı Düşünme Etkinliklerinin Öğrencilerin Akademik Başarılarına ve Derse Yönelik Tutumlarına Etkisi", *Amasya Üniversitesi Eğitim Fakültesi Dergisi*, vol. 1, no. 2, pp. 128-142, 2012.

[7] Bayrak, F., ve Yurdugül, H., "E-Değerlendirme ve Dönüt", *The Turkish Online Journal of Educational Techno Bujang*, 2025

[8] S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M., "Multiclass prediction model for student grade prediction using machine learning", *Ieee Access*, vol. 9, pp. 449-465, 2021.

[9] Briand, L., Freimut, B., ve Vollei, F., " Using multiple adaptive regression splines to support decision making in code inspections", *Journal of Systems and Software*, vol. 73, no. 2, pp. 205-217, 2004.

[10] Çanga, D., "Use of MARS Data Mining algorithm based on training and test sets in determining carcass weight of cattle in different breeds", *Journal of Agricultural Sciences*, vol. 28, no. 2, pp. 259-268, 2022.

[11] Chou, Shieu-Ming, et al. "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines." *Expert systems with applications*, vol. 27, no. 1, pp. 133-142 2004.

[12] Çelik, Ş., Şengül, T., Şengül, A. Y., ve Hakan, İ., "Tüketici fiyat indeksini etkileyen bitkisel ve hayvansal üretim değerlerinin çok değişkenli uyarlanabilir regresyon uzanımları ile incelenmesi: Türkiye örneği", *Journal of Awareness*, vol. 3, no. 3, pp. 399–408, 2018.

[13] Depren, S. K. (2018). "Prediction of Students'science Achievement: An Application of Multivariate Adaptive Regression Splines and Regression Trees", *Journal Of Baltic Science Education*, vol.17, no. 5, pp. 887-903, 2018.

[14] Ekrem, Ö., Salman, O. K., Aksoy, B., ve İnan, S. A., "Yapay zeka yöntemleri kullanılarak kalp hastalığının tespiti", *Mühendislik Bilimleri ve Tasarım Dergisi*, vol. 8, no. 5, pp. 241–254, 2020.

[15] Everingham, Y., ve Sexton, J., "An introduction to Multivariate Adaptive Regression Splines for the cane industry" 33rd Annual Conference of the Australian Society of Sugar Cane Technologists. Mackay, 2011.

[16] Eyduran, E., M. Akin, and S. P. Eyduran., *Application of multivariate adaptive regression splines through R software*, Ankara Turkey: Nobel Academic Publishing (2019).

[17] Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines", *The Annals of Statistics*. Vol. 19, no. 1, pp. 1-67, 1991.

[18] Friedman, J. H., " Fast MARS", Technical Report., Department of Statistics. Stanford University, Stanford, 1993

[19] Gençtürk, Ö., "Meslek ve Anadolu Meslek Liselerinde Öğrenci Başarısını Etkileyen Faktörler", Yüksek Lisans Tezi, Marmara Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2001.

[20] Hastie, T., Tibshirani, R., ve Friedman, J., *The Elements of Statistical Learning: Data mining, inference, and prediction*, Springer, New York, 2009.

[21] Hill, Thomas, Pawel Lewicki, and Paweł Lewicki. *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, Inc., 2006.

[22] Kartal, M., Depren, S. K., ve Depren, Ö., "Türkiye'de Döviz Kurlarını Etkileyen Makroekonomik Göstergelerin Belirlenmesi: MARS Yöntemi İle Bir İnceleme", *MANAS Sosyal Araştırmalar Dergisi*, vol. 7, no. 1, pp. 209-229, 2018.

[23] Kazazoğlu, S., "Türkçe ve İngilizce derslerine yönelik tutumun akademik başarıya etkisi", *Eğitim ve Bilim*, vo. 38, no. 170, pp. 295-306, 2013.

[24] Keskin, S., "Türkiye'de Eğitim Düzeyine Göre Kadınların İş Hayatındaki Yeri" *Kadın Araştırmaları Dergisi* , vol. 17, pp. 1-30, 2018.

[25] Kılıç, S., "Doğrusal regresyon analizi" *Journal of Mood Disorders*, vol. 3, pp. 90–92, 2013.

[26] Kolyshkina, Inna, Sylvia Wong, and Steven Lim. "Enhancing generalized linear models with data mining." *Casualty Actuarial Society*. 2004.

[27] Leathwick, J., Elith, J., ve Hastie, T., "Comparative performance of generalized additive models and multi-variate adaptive regression splines for statistical modelling of species distributions", *Ecological modelling*, vol. 199, no. 2, pp. 188-196, 2006.

[28] Leblanch, M., ve Tibshirani, R., "Adaptive Principle Surfaces", *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 53-64, 1994.

[29] Nacar, S., Betül, M. E., ve Bayram, A., " Günlük Çözünmüş Oksijen Konsantrasyonunun Çok Değişkenli Uyarlanabilir Regresyon Eğrileri İle Tahmin Edilmesi", *Uludağ University Journal of The Faculty of Engineering*, vol. 25, no. 3, pp. 1479-1498, 2020.

[30] Olecka, Anna. "Beyond classification: Challenges of data mining for credit scoring." *Knowledge Discovery and Data Mining: Challenges and Realities*. IGI Global, pp. 139-161, 2007.

[31] Orhan, H., Teke, E. Ç., ve Karcı, Z., "Laktasyon Eğrileri Modellemesinde Çok Değişkenli Uyarlanabilir Regresyon Eğrileri (Mars) Yönteminin Uygulanması", *Kahramanmaraş Sütçü İmam Üniversitesi Tarım ve Doğa Dergisi*, vol. 21, no. 3, pp. 363-373, 2018.

[32] Özsoy, G., "Problem Çözme Becerisi İle Matematik Başarısı Arasındaki İlişki", *Gazi Eğitim Fakültesi Dergisi,* vol. 25, no. 3, pp.179-190, 2005.

[33] Put, R., Xu, Q. S., Massart, D. L., ve Vander Heyden, Y., "Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies" *Journal of Chromatography A,* vol. 1055 pp. 11–19, 2004.

[34] Sabancı, D., "Rastgele Orman Yaklaşımı Kullanılarak Çok Değişkenli Uyumlu Regresyon Şeritlerinde Model Seçimi," Doktora Tezi, İstatistik Anabilim Dalı, Ondokuz Mayıs Üniversitesi, , Samsun 2019.

[35] Sevimli, Y., " Çok Değişkenli Uyarlanabilir Regresyon Uzanımlarının Bir Split-Mouth Çalışmasında Uygulaması," Yüksek Lisans Tezi, Biyoistatistik Anabilim Dalı, Marmara Üniversitesi, İstanbul, 2009.

[36] Simpson, L. P., "Perception of examsoft feedback reports as autonomy-support for learners," Doktora Tezi, Morehead State University, Eğitim Bilimleri, Kentucky, 2016.

[37] Strickland, J., *Predictive Analytics Using R.*, Morrisville, North Carolina, ABD: Lulu Press, 2015.

[38] Şahinler, S., "En küçük kareler yöntemi ile doğrusal regresyon modeli oluşturmanın temel prensipleri", *Mustafa Kemal Üniversitesi Ziraat Fakültesi Dergisi*, vol. 5, pp. 57–73, 2000.

[39] Şentürk, B., "İlköğretim Beşinci Sınıf Öğrencilerinin Genel Başarıları, Matematik Başarıları, Matematik Dersine Yönelik Tutumları ve Matematik Kaygıları Arasındaki İlişki." Yüksek Lisans Tezi, Sosyal Bilimler Enstitüsü, Afyon Kocatepe Üniversitesi, Afyon, 2010.

[40] Şevgin, H., "ABİDE 2016 fen başarısının yordanmasında MARS ve BRT veri madenciliği yöntemlerinin karşılaştırılması", Doktora Tezi, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, 2020.

[41] Temel, G. O., Ankarali, H., & YAZICI, A. C, "Regresyon Modellerine Alternatif Bir Yaklaşım MARS", *Turkiye Klinikleri Journal of Biostatistics*, vol. 2, no. 2, pp. 58-66, 2010.

[42] Tosun, F., "Veri Madenciliği İle Çok Değişkenli Uyarlanabilir Regresyon Eğrileri (MARS Modellemesi) Yönteminin Uygulanması", Yüksek Lisans Tezi, Zootekni Anabilim Dalı, Harran Üniversitesi, , Şanlıurfa, 2021.

[43] Tunay, K. Batu. "Türkiye'de paranin gelir dolasim hizlarinin MARS yöntemiyle tahmini." *METU Studies in Development* vol. 28 , no.2, pp.175, 2001.

[44] Tunay, K. B., "Bankacılık Krizleri ve Erken Uyarı Sistemleri: Türk Bankacılık Sektörü İçin Bir Model Önerisi", *BDDK Bankacılık ve Finansal Piyasalar Dergisi*, vol. 4, pp. 9–46, 2010.

[45] Tunay, K. B., "Türkiye'de Durgunluklarin MARS Yöntemi ile Tahmini ve Kestirimi", *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, vol. 30, no. 1, pp. 71–91, 2011.

[46] Xu, Q. S., Daeyaert, F., Lewi, P. J., ve Massart, D. L., "Studies of relationship between biological activities and HIV Reverse Transcriptase Inhibitors by Multivariate Adaptive Regression Splines with Curds and Whey", *Chemometrics and Intelligent Laboratory Systems*, vol. 82, pp. 24-30, 2006.

[47] Zakeri, F. A., "Multivariate adaptive regression splines models for the prediction of energy expenditure in children and adolescents", *Journal of Applied Physiology*, vol. 108, no. 1, pp. 128-136, 2010.

[48] Zhang, W., ve Goh, A. T., "Multivariate adaptive regression splines and neural network models for prediction of pile drivability", *Geoscience Frontiers*, vol. 7, no.1, pp. 45-52, 2016.

[49] Zurimi, S. , Analysis of Multivariate Adaptive Regression Spline (MARS) Model in Classifying factors affecting on Student the Study Period at FKIP Darussalam University of Ambon" In *Journal of Physics: Conference Series*, Vol. 1463, no. 1, pp. 012005, 2020.