TISER
TECHNOLOGY, INNOVATION AND SPECIAL EDUCATION RESEARCH

www.tiserjournal.com

Modeling the Mathematical Performance of Secondary School Students Using E-Learning Environment Data

Koray AKDENİZ*, Ministry of National Education, Türkiye,
ORCID ID: 0000-0002-6343-4514, korayakdeniz@hacettepe.edu.tr
Doç. Dr. Bahadır YILDIZ, Hacettepe University, Faculty of Education, Türkiye,
ORCID ID: 0000-0003-4816-3071, bahadir@bahadiryildiz.net

Abstract: Research on the applicability and effectiveness of data obtained from educational environments using data mining techniques has a guiding feature for educational stakeholders and policy makers. This study aims to model students' academic performance in a mathematics course by using interaction data in an e-learning environment, survey data containing student information, and data obtained from attitude and anxiety scales towards mathematics. The participants in the study were 112 students in the 5th, 6th, 7th and 8th grades of secondary school. The students used the e-learning environment for extra-curricular activities and face-to-face classes throughout the semester. In addition, the socio-demographic status of the students was determined through a questionnaire, and their attitudes and anxieties towards mathematics were determined through scales. The data obtained were analysed using data mining techniques. According to the results of the study, the regression type with the best parameter values in predicting students' final scores compared to other regression types is the elastic net regression. In the regression model, the placement test score, the average academic performance of the previous year and the percentage of completion of the content sent to students online are the variables that make the highest positive contribution to the model. In the classification problem, logistic regression has better parameter values than other classification algorithms. In the model built with the logistic regression algorithm, 85% of the students were correctly classified at the end of the semester.

Keywords: *Educational Data Mining, Mathematics Achievement Prediction, E-learning Environment*

1. INTRODUCTION

With the acceleration of technological development, access to data has become easier than ever. The proliferation of the Internet, the growth of mobile devices, and the digitalization process have enabled individuals and institutions to produce and access vast amounts of data effortlessly. This development has led to the emergence of the concept of data mining. Over time, the importance and application of data mining have significantly increased. Data mining (DM) is defined as the process of discovering hidden patterns and structures within raw data obtained from various sources and transforming it into meaningful and actionable information for research and applications (Baker & Siemens, 2014; Romero & Ventura, 2010).

The data mining approach is used in a wide range of fields, from financial markets to marketing, from engineering to health sciences. In recent years, rapid advances in information technology

*** Corresponding Author**
Ministery of National Education, Türkiye

have enabled schools to digitally store student demographic and academic data through information systems. This provides educators and school administrators with the opportunity to easily access, analyse and use student data in information processes (Polat, 2021). In particular, the analysis of collected educational data for predicting student achievement has recently gained significant importance (Koyuncu, 2018). Educational data mining (EDM) stands out as a discipline that aims to improve student behaviour, learning processes and educational outcomes by analysing this data.

Romero and Ventura (2007) stated that EDM is an iterative cyclical process and defined the components of the process as generating, testing and developing educational hypotheses. Educational stakeholders are responsible for designing, planning and developing educational systems. As students interact with these systems, educational data mining techniques can be used to analyse data such as student personal information, grades, absenteeism and achievement status. These analyses can be used for purposes such as determining factors that affect student achievement, improving achievement levels, reducing absenteeism, providing guidance in course selection, and providing suggestions for career planning (Rizvi, Rienties, & Khoja, 2019). In addition, data mining techniques are an effective tool for forming groups according to students' personal characteristics and learning styles, identifying problems such as low motivation, absenteeism, dropping out of school or not following school rules, and predicting these problems and taking preventive measures (Aksoy, 2014). Identifying problems that may arise in educational processes in advance and developing these processes is one of the important benefits of these techniques (Can, 2017).

The Education Information Network (EBA), which is provided free of charge by the Ministry of National Education in Türkiye, plays an important role in providing digital educational materials and teaching resources to students, teachers and parents. It can be said that there are many studies on EBA in the literature. Some of these studies examine the effects on students' academic achievement, attitudes, interests and opinions by using the content on EBA in the teaching process. Quasi-experimental studies have been conducted in almost every discipline to determine the effectiveness of EBA-supported teaching. In addition, there are studies that look at the views of teachers, pupils and parents on the frequency of use, purposes, benefits and challenges of EBA. In this context, the views of primary, secondary and higher education students and teachers from different sectors on EBA have been investigated. However, there is no study in the literature on modelling student achievement using EBA as an online learning environment with EDM techniques. There are studies in the literature that aim to model student achievement with e-learning environments other than EBA. In the study conducted by Koca (2019), an attempt was made to model student achievement with data from an e-learning system used in a private school. As a result of the study, it was found that some of the variables created with the system data were effective in modelling student achievement in Turkish and mathematics. EBA provides a rich resource for educational data mining studies by creating a large data pool thanks to its extensive educational content and user interactions. Analysing the data on EBA has great potential for increasing student achievement, improving the learning process and developing educational policies.

The e-learning environment used in this study, which aims to model the mathematics performance of middle school students, is the EBA platform. Interaction data related to the students' learning processes were collected from this environment. In addition to the EBA data, a readiness test, a socio-demographic questionnaire, maths attitudes and anxiety scales were used as data collection tools. With the data obtained from the questionnaire, the scales and the e-learning environment, mathematics performance was predicted using classification and prediction regressions, which are commonly used in educational data mining. The model was created by optimising the regression algorithms with the best parameter values. At this point, the research sought to answer the question: What is the success of different predictive and classification models created using student data from the questionnaire, the scale and the e-learning environment in predicting mathematics course performance?

2. LITERATURE REVIEW

2.1. Data Mining

Data mining (DM) can be defined as the process of making future predictions by extracting meaningful relationships from large data sets (Baker & Siemens, 2014; Romero & Ventura, 2010; Şentürk, 2006). The research underlying the concept of data mining emerged in the 1960s with the computerisation of data (Ledley, 1960). Initially, simple data models were developed; over time, entity-relational, relational and object-oriented data models were developed as the needs increased. In the 1990s, the growth of databases with the widespread use of the Internet led to the need to analyse this data, and in this context the concept of data mining was introduced by computer scientists (Efron & Tibshirani, 1991). Today, the applications of data mining are becoming more and more widespread. It has become inevitable that DM, which is used in many different fields from health to banking, from law to commerce, will also be used in education. The use of DM in education has led to the concept of Educational Data Mining (EDM). The diagram in Figure 1 shows the relationship between data mining and educational data mining.

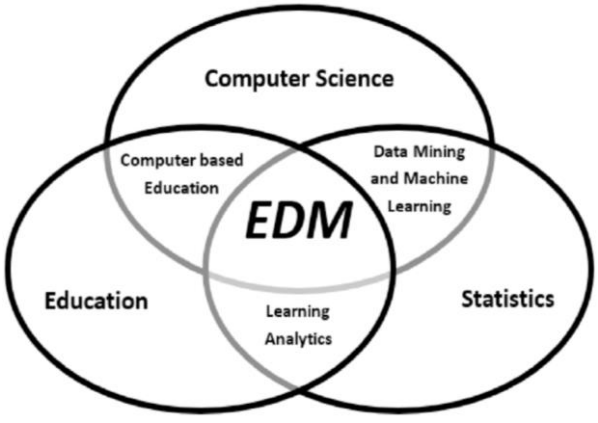

Figure 1. *Interdisciplinary Diagram of EDM (Romero&Ventura, 2013)*

2.1.1. Educational Data Mining

In e-learning environments, the discovery of patterns from the traces left by learners, such as video viewing and the number of clicks on content, has become possible with the use of EDM (Şahin & Yurdugül, 2020). García et al. (2011) define EDM as the process of organising the raw data obtained in educational environments and transforming it into information that can be utilised by educational software, developers, teachers and researchers. Bousbia & Belamri (2014) posit that the EDM process comprises seven stages, as illustrated in Figure 2.

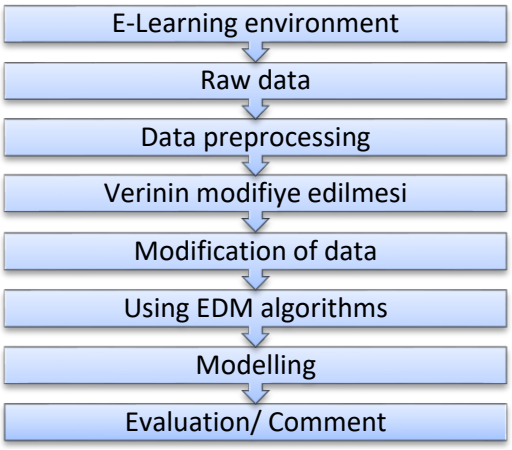

Figure 2. *EDM Process*

The components of this process are explained by Şahin & Yurdugül (2020) as follows;

- *E-learning environment*: Online environments where learners undergo learning processes. Examples of these environments are massive open online course environments (MOOCs), learning management systems, adaptive hypermedia environments.
- *Raw data:* the initial, unprocessed form of all interaction data collected from learners within an e-learning environment. This data can also be defined as "unstructured data."
- *Data preprocessing***:** It is the process of making the data obtained from e-learning environments ready for analysis by conducting structuring studies such as missing data, noise removal and dimension reduction.
- *Modifying the data***:** This is the process of making the data ready for analysis so that it can be used in EDM algorithms. At this stage, operations such as boxing, categorization, etc. can be performed.
- *Use of EDM algorithms***:** The use of classification, clustering and association rules algorithms on the data prepared for analysis in a purposeful way.
- *Modeling***:** It is the process of modeling the data at the end of the analysis and revealing patterns.
- *Evaluation / Interpretation***:** These are the evaluations presented to the researchers for how and for what purpose the obtained models and patterns will be used.

2.2. Education Information Network (EBA)

The Ministry of National Education has made significant progress in recent years with regard to the establishment of technological infrastructure in schools. Nevertheless, the mere establishment of technological infrastructure is insufficient for the comprehensive utilisation of technology in education. Accordingly, the EBA was developed in accordance with the necessity for educational materials to facilitate the utilisation of this infrastructure (Tüysüz & Çümen, 2016). EBA is a social education platform that provides a secure and personalised learning environment for students at all levels, from kindergarten to 12th grade. The platform

provides users with curriculum-compliant course content and personal and professional development guidance services. EBA offers a plethora of resources, including interactive books, textbooks, applications and tests, video or interactive lectures on a multitude of topics and learning outcomes, practice questions, infographics, summaries and project documents at each course level. Furthermore, there are numerous instructor-specific content types. In addition to course content, there is a library area that contributes to the personal development of students, teachers, and parents, and is designed to be engaging and enjoyable to use (MoNE, 2020).

The EBA is an educational platform that is accessible to all students, teachers and parents. The number of users and shares is on a steady upward trajectory. The principal objective of EBA is to furnish users with efficacious materials and a plethora of rich content, irrespective of location (Sarıkaya & Aydın, 2020).

The content on the EBA platform is presented to students through an intelligent recommendation system. The personalisation approach adopted by EBA ensures that the system is designed according to the specific interests and content preferences of each student. The social education platform feature enables students and teachers to interact, create posts on their own pages, participate in discussions and polls, and exchange messages. Teachers can utilise the platform to send homework assignments to their students and analyse reports to determine the extent to which students are completing those assignments successfully (MoNE, 2020).

3. METHODOLOGY

3.1. Research Method

In this study, the steps of the CRISP-DM process model were used to predict student achievement. CRISP-DM is a methodology used in data mining applications and consists of sub-processes such as understanding business processes, understanding data, data preparation, modeling, evaluation, and dissemination (Marinez Plumed, 2019; Schröer, Kruse & Gómez 2021). This model was preferred in this study as it allows the EDM model to be used in independent studies and includes EDM implementation steps. Figure 3 shows the stages of the applied business model.

Figure 3. *The Process Followed in the Research*

This section is intended to provide a comprehensive account of the methodologies, materials, collaborative partners, and participants involved in the study. The protocols utilized for data acquisition, techniques and procedures, investigated parameters, methods of measurement, and apparatus must be described in sufficient detail to allow other scientists to comprehend, analyze, and compare the results. The number, age and sex of the study subjects and participants should be provided. A detailed description of the statistical methods employed should be provided to facilitate verification of the reported results. This section may include a separate subsection that provides an explanation of the abbreviations used in the study.

During the research process, data were collected from the e-learning environment, questionnaires, and lecture notes. These data underwent preprocessing, where variables anticipated to predict end-of-semester course success were selected. Subsequently, the data were formatted for Educational Data Mining (EDM) analysis and analyzed using classification and prediction regressions commonly employed in data mining. The obtained results were then reported.

3.2. Participants

The participants of the study were 5th, 6th, 7th and 8th grade students studying in the middle school where the researcher works. Before starting the application, a meeting was held with the students and information about the study and its scope was given. Then, 157 students were selected on the basis of volunteerism and permission was obtained from their parents with parental consent forms. Although the participant group initially consisted of 157 students, it was determined that 45 students did not log in to EBA during the research process due to technical limitations and student preferences. Since the main focus of the study was to model student achievement with EBA data, these students' data were not included in the analysis process. Therefore, the participant group of the study consisted of 112 middle school students. Information about the study group is given in Table 1.

Table 1. *Information about the Participants*

Class Level	***f***	***%***
5 th grade	*27*	*24.1*
6 th grade	*21*	*18.8*
7 th grade	*33*	*29.4*
8 thgrade	*31*	*27.7*
Total	*112*	*100*
Gender	***f***	***%***
Male	*66*	*58.9*
Female	*46*	*41.1*
Total	*112*	*100*

When Table 1 is examined, it is seen that most of the students in the grade levels included in the study are in the seventh grade (f:33, 29.4%). In addition, the majority of the research participants were male students (f:66, 58.9%). In general, the sample size required for regression analysis should be at least 5 observations for each independent variable (Sümbüloğlu & Sümbüloğlu, 2005). In this study, 10 independent variables were used in regression algorithms and 18 independent variables were used in classification regressions. The participant group of 112 students was suitable for data analysis.

3.3. Data Collection

Data were collected from three different categories of data sources.

3.3.1. E-Learning Environment Data

The first data source is the EBA environment designed by the General Directorate of Innovation and Educational Technologies (YEĞİTEK);

- EBA Score (the score given to students by the platform as a result of the interactions made by students in the EBA environment),
- Number of students' discussion, voting and message interactions,
- Percentage of completion of the content sent by the teacher-researcher,
- Percentage of overall performance in mathematics,
- Data on students' individual content completion percentages in mathematics course.

These data were used in digital environment and in line with the permission of YEĞİTEK, in order to perform data analysis in accordance with the Law on the Protection of Personal Data.

3.3.2. Scales

The second category of data sources comprises the student information questionnaire and the scales utilized to ascertain students' attitudes and concerns. Ethical permissions for the use of these data sources were obtained by the developers. Furthermore, the readiness examination administered at the outset of the academic term represents an additional data source within this category.

The student information survey is a questionnaire prepared by the researcher that includes a range of student-related information, such as study time, the number of siblings, and the status of academic support. The survey is designed to determine the impact of students' socio-demographic status on academic achievement.

The attitude scale is comprised of the following items: In order to ascertain the attitudes of students towards mathematics, the five-point Likert-type Mathematics Attitude Scale, developed by Aşkar (1986), comprising 20 items, was employed. The scale was evaluated based on the following response options: "Strongly Agree," "Agree," "Undecided," "Disagree," and "Strongly Disagree." Responses were assigned a value of 5 for "Totally Agree," 4 for "Agree," 3 for "Undecided," 2 for "Disagree," and 1 for "Strongly Disagree." The scale was found to have a unidimensional structure, with a Cronbach alpha reliability coefficient of 0.96 in the development phase (Aşkar, 1986).

The Anxiety Scale is a psychometric instrument designed to assess the extent to which an individual experiences anxiety. The Mathematics Anxiety Scale (MAAS), developed by Erktin, Dönmez, and Özel (2006), is a four-point Likert-type assessment tool comprising 45 questions. The factor analysis yielded the identification of the sub-dimensions of the scale and the delineation of four distinct dimensions of mathematics anxiety: fear of examinations and assessments, avoidance of mathematics lessons, utilization of mathematics in daily life, and self-confidence in mathematics. The items comprising the scale were assigned a value of 1, 2, 3, or 4, respectively, depending on the frequency with which the respondent indicated the occurrence of the behavior in question. The Cronbach alpha reliability coefficient for the scale was determined to be 0.92.

The readiness exams consisted of valid and reliable questions from the questions of the scholarship exams and the placement test (Seviye Belirleme Sınavı [SBS]) exams conducted by the Ministry of National Education in previous years. The distribution of the questions was determined by taking into account the time allocated to learning areas in the mathematics curriculum. Then, three mathematics education experts, one measurement and evaluation expert and one Turkish language expert were consulted. For each question, the field experts stated whether the question was appropriate for the learning outcome, the language expert stated whether the question was appropriate in terms of language expression, and the measurement and evaluation expert stated whether the question was appropriate in terms of measurement and evaluation principles. The content validity ratio (CVR) was calculated for each item to quantitatively present the evaluation results from the experts.

$$CVR = \frac{NA}{N/2} - 1$$

NA: Number of experts who approved the article

N: Total number of experts

Items with a CSR > 0 are items that are considered appropriate by at least half of the test experts (Yurdugül, 2005). As a result of the expert feedbacks, the CSR ratio was calculated more than 0 for all items and the achievement test was applied to the students after it was finalized.

3.3.3 E-School Data

E-School is a digital system that facilitates the management of student data and school processes in institutions affiliated with the Ministry of National Education. It enables the digital recording of information as outlined in the Ministry of National Education Regulation on Primary Education Institutions (2007). At this juncture, the end-of-term mathematics course grades, number of absences, and academic grade point average data from the previous year, obtained from the participants' E-School environment, were utilized as the third data source in the study.

3.4. Data Analysis

The data obtained from the data collection tools within the scope of the study were initially collected as raw data and variables were identified within the context of educational data mining processes. The determination of variables is of significant importance for the model to yield accurate results. The utilization of incorrect or irrelevant variables may result in the generation of erroneous results. Consequently, the following variables were determined from the measurement tools, and their characteristics are presented in Table 2.

Table 2. *Variables Used in the Study and Their Characteristics*

Data Source	Variable	Variable Name	Variable Type
E-School	Y	Student end of semester score	Continuous
	Y1	End of semester passing status	Categorical
Achievement Test	X1	Readiness score	Continuous
Attitude Scale Towards Mathematics (Aşkar, 1986)	X2	Attitude scale score	Continuous
Mathematics Anxiety Scale (Erktin, Dönmez & Özel ,2006)	X3	Anxiety scale score	Continuous
E-School	X4	Previous year success score	Continuous
	X5	Number of absences	Continuous
E-Learning Environment	X6	EBA point	Continuous
	X7	Total number of Discussion-Vote-Message interactions	Continuous
	X8	Mathematics course general exam performance percentage	Continuous
	X9	Submitted content completion percentage	Continuous
	X10	Percentage of non-homework (individual work) content	Continuous
Demographic Scale	X11	Number of people living in the same house	Continuous
	X12	Financial situation	Categorical
	X13	Having a room of one's own	Categorical
	X14	Study time	Categorical
	X15	Status of receiving academic support	Categorical
	X16	Mother's education leve	Categorical
	X17	Father’s education leve	Categorical
	X18	Family unity status	Categorical

The RStudio program was employed for the purpose of data analysis. In the initial phase of data preprocessing, an analysis of missing data was conducted, which revealed that no instances of missing data were present. An outlier analysis was conducted on each variable in the dataset using the boxplot method. Outliers were observed between 5-12 for each variable. In case of possible data loss, outliers were not deleted due to the inability to reach a sufficient sample for regression analysis and to obtain reliable results. With the suppression method, first quartile values were assigned instead of values smaller than the mean of the variables and third quartile values were assigned instead of values larger than the mean. The reason for choosing this method is that the closest values to the actual values of the variables with outliers are the

average values of the quartiles to which they are connected. In order to improve the performance of the model, the "standardscaler" method was used to scale the data by converting the data into a data format with a mean of zero and a standard deviation of one. This method was chosen because the research variables are on different scales. While some data measured academic achievement between 0-100, variables such as the number of siblings, daily study time, etc. took values in very different ranges from this range. This causes the model to behave biased in algorithms such as support vector machines, logistic regression, etc. The StandardScaler method allows for a fairer comparison of attributes with different units of measurement. This prevents unnecessary dominance of certain features in the decision mechanism of the model. After the data scaling process, the data analysis phase started and regression algorithms such as multiple linear regression, ridge regression, lasso regression and elastic-net regression, and classification algorithms such as K-nearest neighbor, support vector machine, artificial neural network, classification and regression trees, random forest and logistic regression were used. Students' end-of-semester achievement scores were used as the dependent variable for predicting achievement status. In the classification algorithms, if the end-of-semester achievement score was less than 50 points, it was coded as "Failed" and if not, it was coded as "Passed". The reason for this is that the passing grade for the Turkish course is at least 70 points and the passing grade for other courses is at least 50 points (MoNE, 2023).

3.5.Ethics Committee Approval

Approval for this study was granted by the Ethics Committee of Hacettepe University on 14/04/2022 under registration number 2136407.

4. FINDINGS

4.1. Descriptive Statistics

Table 3 below displays the fundamental descriptive statistical information pertaining to the dependent and independent continuous variables employed in this investigation. The aim of this study is to predict students' end-of-semester achievement scores based on the data obtained from the online environment and scales.

Table 3. *Descriptive Statistics of Variables*

Variable	n	Mean	Median	Min	Max	Standard Deviation	Skewness	Kurtosis
Y	112	68.02	66.85	25.7	100	19.41	-0.08	1.92
X1	112	40.98	40	10	100	17.17	0.89	3.71
X2	112	74.72	79	25	100	18.34	-0.82	2.74
X3	112	85.94	81.5	45	154	23.19	0.72	3.46
X4	112	93.5	93.5	58.1	100	7.55	-1,68	6.65
X5	112	2.32	2	0	14	2.57	1,59	6.15
X6	112	130.01	70.5	1	787	163.95	2.17	7.84
X7	112	1.87	0	0	13	3.16	1.61	4.63
X8	112	16.53	0	0	100	28.39	1.51	3.99
X9	112	7.25	0	0	100	31.17	3.15	14.37
X10	112	4.85	2	0	51	8.15	3.3	15.99
X11	112	4.58	5	2	9	1.18	1.08	2.79

As indicated in the table above, the mean score for the students' end-of-semester assessments (Y), which serves as the dependent variable in this study, is 68.02. The lowest end-of-semester score observed among the participants in the study was 25.7 points, while the highest score was 100 points. Moreover, the skewness and kurtosis values of the student scores align with the characteristics of a normal distribution. It can be stated that the independent variables X4, X5, X6, X7, X8, X9, and X10 deviate from the normal distribution, as evidenced by their skewness values exceeding -1 or 1 and kurtosis values exceeding 3. To ensure that the variables in the table deviate from the normal distribution and that the variables are on the same scale, data scaling was performed on the independent variables using the "scale" function in the Rstüdio program. Subsequently, regression analyses were conducted on the new values of the variables. Table 4 presents the numerical code values and frequencies of categorical variables.

Table 4. *Frequencies of Categorical Variables*

Variable	Variable Value	f	%	Total %
YS	Failed	25	22.3	22.3
	Passed	87	76.7	100
X12	0-3000 TL	42	37.5	37.5
	3000-6000TL	44	39.2	76.7
	6000-10000 TL	15	13.3	90
	+10000 TL	11	10	100
X13	None	40	35.7	35.7
	Yes	72	64.3	100
X14	0-1 hour	35	31.3	31.3
	1-3 hour	64	57.1	88.4
	3-6 hour	13	11.6	100
	6-10 hour	0	0	100
X15	No support	90	80.3	80.3
	Private classroom	17	15.2	95.5
	Privete tutoring	5	4.5	100
X16	Primary School	24	21.4	21.4
	Middle School	23	20.5	41.9
	High School	45	40.2	82.1
	License	19	17	99.1
	Postgraduate	1	0.9	100
X17	Primary School	11	10	10
	Middle School	21	18.7	28.7
	High School	41	36.6	65.3
	License	37	32.9	98.2
	Postgraduate	2	1.8	100
X18	Together	99	88.4	88.4
	Seperate	13	11.6	100

As is evident from the data presented in the above table, the majority of the students who participated in the study (n = 87; 76.7%) successfully completed the mathematics course at the conclusion of the semester. It can be posited that the income of the student families in question typically falls within the range of 3,000 to 6,000 Turkish Lira. The majority of students (64.3%) had their own room, while the majority of students (80.3%) did not receive any academic support. Additionally, the data indicates that the majority of students (57.3%) study between one and three hours per day. Furthermore, the data on the parents of the students reveals that the majority of parents had completed high school, and a significant proportion of students resided with their parents (88.4%).

4.2. Regression Results

Correlation analysis was conducted with the scaled data in order to determine the relationships between the variables. The data obtained are shown in Figure 4.

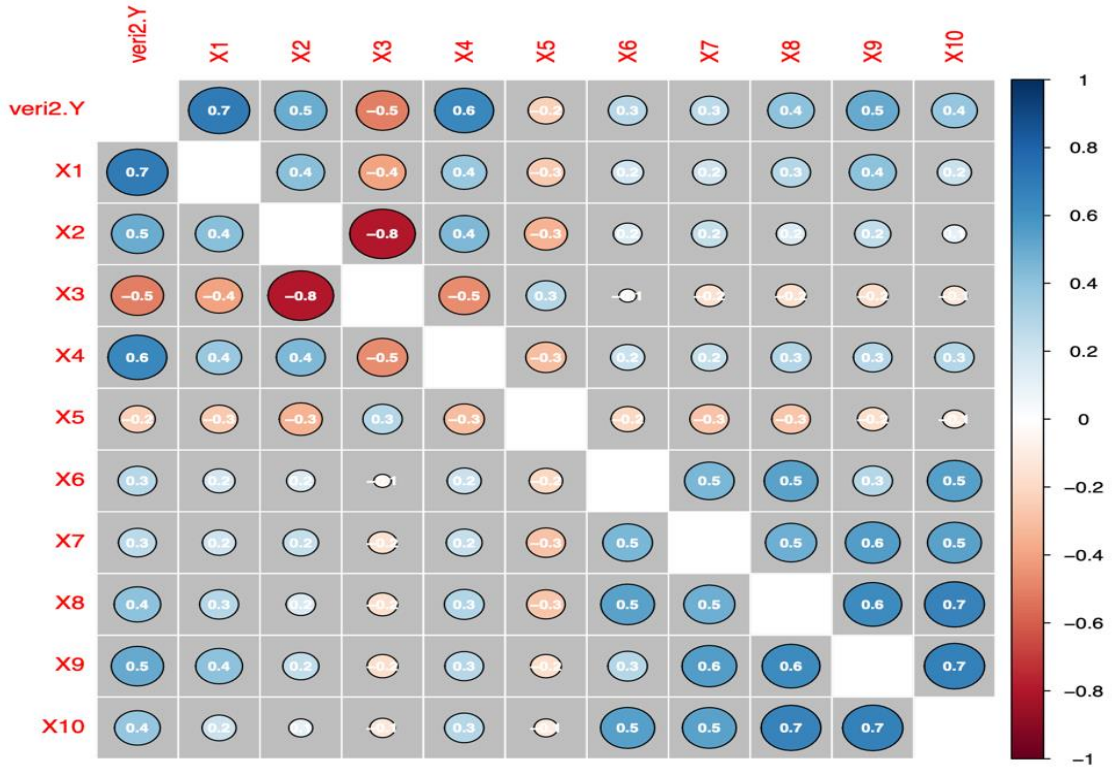

Figure 4. *Correlation Matrix of Independent Variables*

Figure 4 illustrates a strong negative correlation between the X2 and X3 variables (-0.8) and a robust positive correlation between the X8 and X10, as well as the X9 and X10 variables (0.7). Subsequent to this, a multiple linear regression analysis was conducted, and variance inflation factors (VIF) were examined to ascertain the presence of multicollinearity between the variables. Moreover, the optimal multiple linear regression model was identified by partitioning the data into training and test sets at varying proportions.

Table 5. *Statistics of Multiple Linear Regression Models according to Different Training Data*

Regression Type	Percantage of traning-test data	RSE	RMSE	MSE	MAE	Adjusted-R^2
Linear Regression	100	10.92	10.36	107.52	8.17	0.683
	95	10.92	11.10	123.26	9.31	0.681
	90	11.11	8.54	72.87	6.41	0.676
	85	11.04	10.84	117.51	7.80	0.682
	80	**10.37**	**14.88**	**221.41**	**8.45**	**0.731**
	75	11.06	11.15	124.52	8.79	0.674
	70	10.44	12.56	157.78	10.32	0.702
	65	10.38	12.71	161.46	10.46	0.729

Table 5 illustrates the process of dividing the data set into distinct proportions for training and testing, with the objective of identifying the optimal regression model. Following the division of the data set into 80% training data and 20% test data the model was tuned using the 10-fold cross-validation method, resulting in an Adj. R2 value of 0.731. The adjusted R^2 value indicates the proportion of the change in the dependent variable that can be explained by the independent variables. Accordingly, 73% of the changes in the end-of-semester score can be explained by the independent variables of the study. Furthermore, the standard error of residuals (RSE) value, which assesses the model's fit to the data, was calculated as 10.37, and the root mean square error squares (RMSE) value was calculated as 14.88. Table 6 illustrates the regression model's optimal coefficients.

Table 6. *Coefficients of the Best Regression Model*

Variable	B	Std. Error	T	VIF	P
Fixed	68.638	1.103	62.239	-	2.e-16***
X1	7.281	1.291	5.641	1.462	2.40e-07***
X2	1.187	1.940	0.612	3.512	0.542
X3	-3.011	2.066	-1.458	3.855	0.148
X4	5.863	1.386	4.230	1.632	6.12e-05***
X5	1.510	1.211	1.247	1.338	0.216
X6	1.260	1.586	0.794	1.956	0.429
X7	-2.568	1.440	-1.784	1.782	0.078
X8	0.715	1.640	-0.436	2.315	0.664
X9	8.396	2.177	3.857	2.980	0.002***
X10	0.087	1.747	0.05	2.742	0.960
R^2=0.775	***Ad. R^2= 0.731***		***F= 23.7***		***p < 2e-16***

Table 6 reveals that when VIF values are considered, all variables exhibit VIF values less than 5, respectively. It can thus be concluded that there is no evidence of multicollinearity among the variables. Given that the p-value is less than 0.05, it can be concluded that at least one variable in the model has the capacity to explain the model. Furthermore, the p-values of X1, X4, and X9 variables are less than 0.05, indicating that these variables significantly contribute to the model. Furthermore, increases in the X3 and X7 variables have a negative impact on the end-of-semester achievement score. Increases in other variables have a positive effect on the dependent variable. In light of the high correlation coefficients of 0.7 and 0.8 observed in Figure 4, additional analyses were conducted, including ridge regression, lasso regression, and elastic-net regression, in addition to the initial multiple linear regression analysis. These supplementary analyses were performed to ensure robustness against outliers and multicollinearity issues. The results of the aforementioned additional analysis, which yielded the most optimal multiple linear regression model data, are presented in Table 7.

Table 7. *Statistics for Different Regression Types*

Regression Type	Percantage of traning	RMSE	MSE	MAE	Adjusted-R^2
Rigde regression	80	10.17	103.41	8.12	0.705
Lasso regression	80	10.73	115.28	8.734	0.690
Elastic-net regression	**80**	**9.96**	**99.34**	**7.84**	**0.733**

When Table 7 is analyzed, Elastic-net regression has the best parameters. The Adj R2 value, which has the greatest importance for regression parameters, was calculated as 0.733. This means that 73% of the changes in the end-of-semester score can be explained by the independent variables of the study. In addition, since values such as RMSE (9.96) and MSE (99.34) are lower than the parameter values of other algorithms, it is seen that the fit of the model to the data is better in Elastic-net regression. Figure 5 shows the coefficients of Elastic-net regression.

```
10 x 1 sparse Matrix of class "dgCMatrix"
                s0
X1   0.390103019
X2   0.075375628
X3  -0.109217406
X4   0.699873550
X5   0.282254251
X6   0.001215253
X7  -0.294639463
X8   .
X9   0.384423134
X10  0.043557059
>
```

Figure 5. *Elastic-Net Regression Coefficients*

As illustrated in Figure 5, the independent variables X3 and X7 exert a negative influence on the dependent variable, whereas the remaining variables contribute a positive effect to the regression model. In the regression model, the effect of the X8 variable on the model is zero. The X1, X4, and X9 variables made the highest positive contribution to the dependent variable. The results obtained show that the increase in anxiety score and the total number of discussions and messages in the e-learning environment have a negative effect on academic achievement. However, the increase or decrease in students' overall exam percentage in EBA has no significant effect on academic achievement. On the other hand, it was found that the increase in the readiness score, the previous year's grade point average and the percentage of completion of the content sent to the students by the teacher positively affected academic achievement.

4.3. Classification Results

After the predictive regression analyses, students' end-of-semester achievement scores were classified as "Failed" and "Passed" and logistic regression analysis was performed first. Then, using the training and test data with the best parameter values in logistic regression, the model with the best parameters was created using K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Classification and Regression Trees (CART), Random Forest, Neural networks (NN) classification regressions. Table 8 shows the results of logistic regression analysis.

Table 8. *Statistics of Logistic Regression Models according to Different Training Data Faces*

Regression Type	Percantage of traning-test data	Accuracy	%95 Confidence interval	Kappa	Sensivity	Specificity	AIC
Logistic regression	100	0.893	0.82 – 0.94	0.691	0.93	0.76	82.385
	95*	1	0.48 - 1	1	1	1	82.378
	90	0.9	0.55 – 0.99	0.615	1	0.50	78.35
	85	0.9	0.55 – 0.99	0.615	1	0.50	78.35
	80	0.773	0.55 – 0.92	0.396	0.82	0.60	562.61
	75	**0.852**	**0.66 – 0.96**	**0.617**	**0.86**	**0.83**	**58**
	70*	1	0.95 - 1	1	1	1	56
	65*	1	0.95 - 1	1	1	1	54

*Overfitting status

An analysis of Table 8 reveals that the logistic regression model developed with 75% of the training data exhibits optimal results when evaluated based on the specified performance indicators. The overall accuracy of the model is indicated by the Accuracy value, which is 0.852. This indicates that the model is accurate in 85% of instances. Moreover, the logistic regression model demonstrates satisfactory performance with high sensitivity (0.86) and specificity (0.83). The Kappa value of 0.617 indicates that the model is relatively effective in comparison to random forecasts, yet there is potential for further enhancement. The AIC value (58) suggests that the model offers a balanced fit and is not excessively complex. Overall, it can be stated that this model is capable of effectively discriminating between positive and negative classes. The coefficients of the model are illustrated in Figure 6.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.785e+01  3.549e+01  -2.194  0.02826 *
X1           5.583e-02  7.520e-02   0.742  0.45782
X2           2.568e-01  1.550e-01   1.657  0.09754 .
X3          -3.836e-02  1.034e-01  -0.371  0.71066
X4           6.091e-01  2.363e-01   2.578  0.00994 **
X5          -4.040e-01  4.355e-01  -0.928  0.35361
X6          -7.352e-02  3.289e-02  -2.235  0.02542 *
X7           4.862e-01  6.142e-01   0.792  0.42856
X8           1.675e-01  1.942e-01   0.862  0.38847
X9           1.239e+00  6.284e-01   1.971  0.04873 *
X10          2.475e+00  1.282e+00   1.930  0.05364 .
X11          7.851e-02  5.608e-01   0.140  0.88866
X122         1.255e+01  5.153e+00   2.435  0.01488 *
X123        -2.227e+00  5.043e+00  -0.442  0.65874
X124         2.919e+01  6.053e+03   0.005  0.99615
X131        -7.907e+00  3.841e+00  -2.059  0.03954 *
X142        -4.805e+00  2.920e+00  -1.645  0.09988 .
X143         1.331e+01  3.662e+03   0.004  0.99710
X152        -2.636e+00  6.000e+00  -0.439  0.66049
X153         1.120e+01  1.038e+04   0.001  0.99914
X162         3.858e+00  2.748e+00   1.404  0.16025
X163         5.049e+00  2.715e+00   1.859  0.06297 .
X164         5.136e+00  6.109e+00   0.841  0.40049
X165         4.272e+01  4.305e+04   0.001  0.99921
X172         9.754e+00  4.684e+00   2.082  0.03730 *
X173         5.814e+00  3.427e+00   1.697  0.08978 .
X174         1.519e+01  7.754e+00   1.959  0.05014 .
X175        -1.147e+01  3.161e+04   0.000  0.99971
X182         3.990e+00  2.430e+00   1.642  0.10063
```

Figure 6. *Coefficients of the Best Regression Model*

A detailed examination of Figure 6 reveals that the students' success or failure in mathematics at the conclusion of the semester was markedly influenced by a number of factors. These include the students' previous year's achievement score (X4), their EBA score for the semester (X6), and the extent to which they had completed the content provided to them by the teacher. (X9), family income (X12_2) between 3000-6000 TL, the lack of a room belonging to the student (X13_1), and the educational status of the students' fathers being secondary school (X17_2). The complexity matrix, formed when the model created using 75% training data was tested on test data, is shown in Table 9.

Table 9. *Complexity Matrix of the Model*

		Real Value	
		Failed	Passed
Estimated Value	Failed	5	3
	Passed	1	18

Upon analysis of Table 9, the model demonstrated a 85% accuracy rate, correctly predicting the failure of five out of eight students and the success of 18 out of 19 students. This indicates that the model is highly reliable in predicting academic performance. Subsequent to the logistic

regression analysis, the optimal model parameters were identified through the application of diverse classification regression algorithms. The outcomes are illustrated in Table 10.

Table 10. *Model Parameters Generated from Different Classification Regressions*

Regression Type	**Accuracy**	**%95 Confidence interval**	**Kappa**	**Sensivity**	**Specificity**
KNN	0.78	0.58 – 0.91	0	1	0
SVM	**0.78**	**0.58 – 0.91**	**0.36**	**0.86**	**0.50**
CART	**0.78**	**0.58 – 0.91**	**0.36**	**0.86**	**0.50**
Random Forest	0.78	0.58 – 0.91	0.16	0.95	0.16
YSA	0.78	0.58 – 0.91	0.27	0.90	0.33

Upon examination of Table 10, it can be observed that the accuracy value is 0.78 in models belonging to different regression types. This indicates that the correct classification rate is 78% across all regression types. The table also indicates that the regression types with the optimal parameter values are the SVM and CART algorithms. The sensitivity value of the models is 0.86, indicating that the models are highly successful in detecting positive classes. However, the specificity value of 0.50 indicates that the model has low success in detecting negative classes and tends to misclassify negative classes as positive. Furthermore, the Kappa value of 0.36 indicates that the model has a moderate fit and that its performance requires improvement.

5. RESULTS AND DISCUSSION

Unlike traditional classroom environments, online learning environments make it possible to record traces of all student activities. These traces may include various activities such as following course announcements, responding to questions on discussion platforms or simply logging in and out of the system. Research has shown that many educational problems can be solved by recording such data in databases about learning processes and analyzing them with the EVM approach. In this study, one of these problems is focused on predicting student performance.

Within the scope of the study, 19 variables were produced with the data obtained from the scales at the beginning of the semester and the data obtained from the e-learning environment during the semester. These variables were used to predict students' end-of-semester mathematics course scores and to classify the passing and failing status of the mathematics course. Prior to the analysis of the data, the data underwent a series of pre-processing stages. The initial step involved the identification of missing data. Once it was established that no data was missing, outlier analysis was conducted on the variables. The outliers identified were not removed to prevent data loss; instead, new values were assigned using the suppression method. Subsequently, data scaling was conducted to guarantee that disparate variables were aligned

on an equivalent scale and to enhance the efficacy of the model. Following data preprocessing, different regression and classification algorithms were divided into different percentages of training and test data, and models that best predict academic achievement were created. While selecting the best model, algorithms with the best values in more than one regression parameter were taken into consideration. The final version of the model was obtained by organizing the new models using 10-k cross-validation. All these procedures are important for ensuring the accuracy and generalizability of the research results.

In line with the research problem, the data prepared for analysis were first analyzed by multiple linear regression analysis using different proportions of training and test data for the prediction of the end-of-semester achievement score. In the results obtained, the regression model with the best parameters was obtained when the data was divided into 80% training data and 20% test data. In this model; students' readiness test scores, academic achievement averages of the previous year and the percentage of completion of the content sent to students in the online learning environment made significant contributions. Determining the readiness level of an individual enables guidance in accordance with his/her individual and characteristic features, determining his/her needs and making plans, programs and preparations according to these needs (Yapıcı, 2004). According to Yenilmez and Kakmacı (2008), individuals with a high level of readiness can comprehend subjects faster and make comments on the subject. These individuals complete their homework more easily because they have learned the concepts and become more ready for the next subject because they understand the previous subject well. In parallel with this, the results of the study revealed that readiness test scores made a positive and significant contribution to the regression model. According to Ülker (2021), interactive assessment tools in the e-learning environment have a high level of impact on student achievement. The percentage of completion of interactive exams, exercises, etc. sent to students in the EBA environment also made a positive and significant contribution to the regression model. In the study conducted by Koca (2019), it was aimed to model mathematics course achievement with the data obtained from the learning management system used by students. As a result of the study, it was concluded that students' performance in the test they solved on the system and the number of questions had a significant effect on predicting academic achievement. This result is consistent with the findings of the research.

In multiple linear regression, VIF values below 5 may indicate that there is no multicollinearity problem between variables. However, in the correlation analysis between variables, algorithms that are resistant to multicollinearity can be used due to high correlation values between some variables. In this study, in addition to multicollinear regression, Ridge regression, Lasso regression and Elastic-net regression algorithms were used to analyze the data. In the results obtained, Elastic-net regression has the best parameter values. In addition, elastic-net regression has the best parameter values such as RMSE, Adj. R^2, MSE have better values than multiple linear regression. In the model obtained from the elastic-net regression, the Anxiety score and the total number of students' votes, discussions and messages contributed negatively to the model. In the model, the effect of students' overall exam performance percentages on the model was zero. When the literature is examined, in studies aiming to predict success in online environments, the increase in the number of messages and posts written by students in the

environment positively affects academic achievement (Akçapınar, 2014; Lopez et al. 2011). The reason why parallel results could not be obtained at this point may be that the interaction of discussion, message and voting in the EBA environment increases the EBA score and students make interactions outside the focus of the question asked. According to Bindak (2005), mathematics anxiety is one of the most important affective factors that negatively affect individuals' mathematics achievement, regardless of their intelligence level. Parallel results were obtained in this study. The overall exam average is the percentage of correct answers in all exams taken by the student. However, since some students took very few exams and had a high level of overall performance, the effect of this variable may be zero. The variables that made the highest contribution to the model were students' readiness test scores, academic achievement averages of the previous year, and the percentage of completion of the content sent to students in the online learning environment. These results are in parallel with the results obtained from multiple linear regression analysis.

In alignment with the research question, students' final semester grades were classified as unsuccessful if they fell below 50 points, and as successful if they reached or exceeded 50 points. Logistic regression, random forest, support vector machines, decision tree, artificial neural networks and k-nearest neighbor algorithms, which are frequently used in the literature, were used on the classified data. Among the classification algorithms, logistic regression is the algorithm that best predicts the pass/fail status of students at the end of the semester using 75% training data. The logistic regression algorithm demonstrated an 85% accuracy rate in correctly predicting the end-of-semester achievement classification. The variables that were found to significantly affect the pass or fail status of students in the mathematics course at the end of the semester were the achievement score of the previous year, the EBA score of the students for the semester, the percentage of content completed by students, the income of the family being between 3000-6000 TL, the availability of a room belonging to the student, and the educational status of the students' fathers being secondary school. The regression model established by Akhan and Bindak (2017) to predict students' achievement in mathematics revealed a significant relationship between variables such as the grade point average of the previous year and the presence of a room belonging to the student and achievement. Furthermore, the study conducted by Anıl (2010) aimed to predict students' achievement in the science course. The variable of father's education level was identified as the variable that made the highest significant contribution to achievement in the regression model. These results are in alignment with the findings of the aforementioned study.

School success is greatly influenced by many different factors besides mental factors. These factors include achievement motivation, anxiety level, family characteristics, socio-economic status, inadequate conditions of school and education, environmental factors, nutrition and health conditions. These variables affecting the learning process are closely related to the physiological, psychological and social conditions of the individual (Güleç & Alkış, 2003). In parallel with this in the study, variables such as financial status, parental education level, and working environment at home, which are factors outside the student, have a significant effect on the regression model.

In existing studies which have aimed to model student mathematics achievement in the literature, it was observed that the factors affecting student achievement were not addressed using a sufficiently multidimensional approach. In the study conducted by Akhan and Bindak (2017), mathematical achievement was modelled using variables obtained from the personal information form, the attitude and school culture scale, and other relevant data. In the study conducted by Çalışkan (2014), mathematical achievement was modelled through the variables of students' cognitive entry behaviours and time allocated to mathematics. In the study conducted by Koca (2019), achievement in the Turkish and mathematics courses was modelled through the variables obtained from the learning management system used by the students. The effect of mathematics anxiety on mathematics achievement was investigated by İlhan and Sünkür (2013). These studies concentrated on a limited number of variables for modelling mathematics achievement. The modelling of each variable separately affects the error variance. In this study, mathematics achievement was modelled with 18 different independent variables, including students' traces left in the online learning environment, attitude and anxiety scales, students' demographic data, and cognitive input behaviours. The model provided a more comprehensive prediction of the multidimensional structure of mathematics achievement.

As a result, in this study, which aims to predict middle school students' mathematics achievement with EDM methods, variables such as students' cognitive input behaviors, structured interactions with online learning environment, family income and father's education level have a positive effect on academic achievement. Variables such as attitude towards mathematics, anxiety and number of absenteeism did not show a significant effect on academic achievement.

6. RECOMMENDATIONS

This section includes recommendations for researchers and teachers.

6.1. Recommendations for Researchers

The online learning environment employed in the present study is the EBA platform. It should be noted that videos watched, exercises or exams completed in the EBA programme may appear as incomplete even though students have completed these contents. This may result in issues with data collection. Therefore, it would be beneficial to utilise e-learning environments that provide access to more reliable data in new studies.

In the present study, data were analysed from 112 students. In future studies, the application of data mining algorithms may yield more robust results and enhance the generalisability of findings when studies are conducted with a larger sample size, including students from multiple educational institutions and those from diverse socioeconomic backgrounds.

6.2. Recommendations for Teachers

The study demonstrated that the studies sent to students in the EBA environment had a positive and significant impact on academic achievement. In contrast, the unstructured interactions of students with the system did not have a significant effect on achievement. It is recommended that teachers structure the content in the EBA environment and send it to students, as well as

provide follow-up. Additionally, students' cognitive input behaviors have been found to have a significant effect on the prediction of academic achievement. Therefore, it is advised that teachers determine the cognitive input behaviors of the students they teach at the beginning of the semester and take measures accordingly.

6.3. Recommendations for Policy Making Institutions in Education

In the study, some of the content completed by learners on the EBA platform was not completed due to internet infrastructure or systemic failures. It is recommended that policy-making institutions in education reinforce the infrastructure of online learning environments such as EBA. Furthermore, some students were unable to log in to EBA due to a lack of necessary facilities. The MoNE provided complimentary internet and device support for EBA, particularly during the pandemic period. It would be beneficial to increase the number of such supports to enable students with financial limitations to access EBA.

7. ABOUT THE AUTHORS

Koray AKDENİZ: He continues his doctoral studies in the field of mathematics education at Hacettepe University. He works on the use of technology in mathematics education, machine learning and artificial intelligence. He works as a mathematics teacher at a school affiliated with the Ministry of National Education.

Bahadır YILDIZ: He is a faculty member of Hacettepe University Faculty of Education, Department of Mathematics Teaching. He graduated from Hacettepe University, Elementary Education Undergraduate Program and completed her master's and doctoral education in the Department of Computer Education and Instructional Technology at the same university. She is working on computational thinking, design thinking, interdisciplinary approach and STEAM, machine learning and artificial intelligence in mathematics education, with the main focus on the integration of information and communication technologies into mathematics education. He produces national and international publications and projects on related topics.

8. References

Akçapınar, G. (2014). *Çevrimiçi öğrenme ortamındaki etkileşim verilerine göre öğrencilerin akademik performanslarının veri madenciliği yaklaşımı ile modellenmesi [A data mining approach to students' academic performance modeling in online learning environment based on their interaction data]*. (Doktoral Dissertation). Hacettepe University, Ankara, Türkiye.

Akhan, Ş. & Bindak, R. (2017). Bazı kişisel değişkenlerin ortaokul öğrencilerinin matematik başarısı üzerindeki etkisi: Bir regresyon modeli [The effect of some personal variables on secondary school students' mathematics achievement: A regression model]. *Ihlara Journal of Educational Research, 2(*2), 5-17.

Aksoy, E. (2014). *Matematik alanında üstün yetenekli ve zekalı öğrencilerin bazı değişkenler açısından veri madenciliği ile belirlenmesi [Determination of gifted and intelligent students in the field of mathematics by data mining in terms of some variables]*. (Master Thesis). Dokuz Eylül University, İzmir, Türkiye.

Anıl, D. (2010). Uluslararası öğrenci başarılarını değerlendirme programı (PISA)'nda Türkiye'deki öğrencilerin fen bilimleri başarılarını etkileyen faktörler [Factors affecting the science achievements of students in Turkey in the International Student Achievement Assessment Program (PISA)]. *Education and Science*, *34*(152).

Aşkar, P. (1986). Matematik dersine yönelik tutumu ölçen likert-tipi bir ölçeğin geliştirilmesi [Development of a likert-type scale to measure attitudes towards mathematics course]. *Education and Science, 62*, 31-36.

Baker, R. & Siemens, G. (2014). Learning analytics and educational data mining. *Cambridge handbook of the leaning sciences*, 253-272. Retrieved from https://www.researchgate.net/profile/Ryan-Baker-2/publication/316628053_Educational_data_mining_and_learning_analytics/links/611682a60c2bfa282a41e893/Educational-data-mining-and-learning-analytics.pdf

Bindak, R. (2005). İlköğretim öğrencileri için matematik kaygı ölçeği [Mathematics anxiety scale for primary school students]. *Fırat University Journal of Science and Engineering Sciences*, *17*(2), 442-448.

Bousbia, N. & Belamri, I. (2014). Which contribution does EDM provide to computer-based learning environments?. In *Educational data mining* (pp. 3-28). Springer, Cham.

Can, E. (2017). *Temel eğitimden ortaöğretime geçiş sınavı kazanımlarının veri madenciliği yöntemleri ile değerlendirilmesi [Evaluation of basic education to secondary education transition exam achievements using data mining methods]*. (Master Thesis). Afyon Kocatepe University, Afyon, Türkiye.

Çalışkan, M. (2014). Bilişsel giriş davranışları, matematik özkavramı, çalışmaya ayrılan zaman ve matematik başarısı arasındaki ilişkiler [Relationships between cognitive entry behaviors, mathematics self-concept, time allocated to studying, and mathematics achievement]. *Turkish Journal of Social Research, 181*(181), 345-358.

Doğan, O. (2017). Türkiye'de veri madenciliği konusunda yapılan lisansüstü tezler üzerine bir araştırma [A Research on postgraduate theses on data mining in Turkey]. *Journal of Gazi University Faculty of Economics and Administrative Sciences, 19*(3), 929-951.

Dunham, M.H. (2003). *Data mining introductory and advanced topics*. Upper Saddle

Efron, B. & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science, 253*(5018), 390-395.

Erktin, E., Dönmez, G. & Özel, S. (2006). Matematik kaygısı ölçeği'nin psikometrik özellikleri [Psychometric properties of the mathematics anxiety scale]. *Education and Science, 31*(140).

García, E., Romero, C., Ventura, S. & de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education, 14*(2), 77-88. https://doi.org/10.1016/j.iheduc.2010.07.006

Güleç, S. & Alkış, S. (2003). İlkögretim birinci kademe öğrencilerinin derslerdeki başarı düzeylerinin birbiri ile ilişkisi [The relationship between the success levels of primary school students in courses]. *Primary education Online, 2*(2).

İlhan, M. & Sünkür, M. Ö. (2013). Matematik kaygısının matematik başarısını yordama gücünün cinsiyet ve sınıf değişkenleri açısından incelenmesi [Examining the power of mathematics anxiety in predicting mathematics achievement in terms of gender and grade variables]. *Gaziantep University Journal of Social Sciences, 12*(3).

Koca, M. H. (2019). *Ortaokul öğrencilerinin ders başarı düzeylerinin öğrenme analitiği ile tahmini [Prediction of secondary school students' course success levels with learning analytics]*. (Master Thesis), Yıldız Technical University, İstanbul, Türkiye.

Koyuncu, İ. (2018). *Öğrencilerin PISA matematik başarılarının yordanmasında veri madenciliği yöntemlerinin karşılaştırılması [Comparıson of data mining methods in predicting PISA mathematical achievements of students]*. (Doktoral Dissertation). Hacettepe University, Ankara, Türkiye.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., ... & Flach, P. (2019). CRISP-DM twenty years later: From data mining processes

to data science trajectories. *IEEE transactions on knowledge and data engineering*, *33*(8), 3048-3061. https://doi.org/10.1109/TKDE.2019.2962680

Ministry of National Education (MoNE) (2020). Eğitimde FATİH projesi [FATİH project in education]. Ministry of National Education Primary Education Institutions Regulation, 2007. Retrieved from http://fatihprojesi.meb.gov.tr

Ministry of National Education (MoNE) (2023). *Okul öncesi eğitim ve ilköğretim kurumları yönetmeliği [Preschool education and primary education institutions regulation]*. *Retrieved from*[*https://www.mevzuat.gov.tr/File/GeneratePdf?mevzuatNo=19942&mevzuatTur=KurumVeKurulusYonetmeligi&mevzuatTertip=5*](https://www.mevzuat.gov.tr/File/GeneratePdf?mevzuatNo=19942&mevzuatTur=KurumVeKurulusYonetmeligi&mevzuatTertip=5).

Ledley, R. S. (1960). A pedagogical aspect of the development of the real numbers. *The American Mathematical Monthly*, *67*(3), 280-281.

Lopez, M. I., Luna, J. M., Romero, C. & Ventura, S. (2012). *Classification via clustering forpredicting final marks based on student participation in forums.* Paper presented at the 5th International Conference on Educational Data Mining, EDM 2012, Chania, Greece.

Polat, A. (2021). *Açık öğretim liseleri öğrencilerinin okul terki ve mezuniyet durumlarının eğitsel veri madenciliği ile incelenmesi [Examining the school dropout and graduation status of open education high school students using educational data mining]*. (Doktoral Dissertation). Sakarya University, Sakarya, Türkiye.

Rizvi, S., Rienties, B. & Khoja, S. A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education, 137*, 32-47.

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

Romero, C. & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618. http://dx.doi.org/10.1109/TSMCC.2010.2053532

Romero, C. & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, *3*(1), 12-27. https://doi.org/10.1002/widm.1075

Sarıkaya, D., & Aydın, A. (2020). Eğitim Bilişim Ağı EBA ve deney destekli etkinliklerin 7. sınıf elektrik devreleri ünitesinin öğretimine etkisi [The effect of Educational Information Network EBA and experiment supported activities on teaching the 7th grade electrical circuits unit]. *Journal of Science Teaching, 9*(2), 265-310.

Schröer, C., Kruse, F. & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science, 181*, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

Sümbüloğlu V, Sümbüloğlu K, (2005). *Klinik ve saha araştırmalarında örnekleme yöntemleri ve örneklem büyüklüğü [Sampling methods and sample size in clinical and field research]* (1nd ed). Seçkin.

Şahin, M. & Yurdugül, H. (2020). Educational data mining and learning analytics. (T. Güyer, H. Yurdugül, & S. Yıldırım, Ed.), In *Educational data mining and learning analytics* (pp. 5-32). Anı Publishing.

Şentürk, A. (2006). *Veri madenciliği: Kavram ve teknikler [Data mining: Concepts and techniques]*. Ekin Publishing House

Tüysüz, C. & Çümen, V. (2016). EBA ders web sitesine ilişkin ortaokul öğrencilerinin görüşleri [Opinions of secondary school students regarding the EBA course website]. *Uşak University Journal of Social Sciences, 9*(27/3), 278-296.

Ülker, Ü. (2021). *E-öğrenme ortamlarında etkileşimli değerlendirme araçları kullanımının başarı kaygısına, motivasyona, başarıya etkisi ve öğrenen görüşleri [The effect of using interactive assessment tools on achievement anxiety, motivation, achievement in elearning environments and learners' views]* (Doktoral Dissertation), Gazi University, Ankara, Türkiye.

Yapıcı, M. (2004). İlköğretim dilbilgisi konularının çocuğun bilişsel düzeyine uygunluğu [Suitability of primary school grammar subjects to the child's cognitive level]. *Primary education-Online, 3*(2), 35-41.

Yenilmez, K. & Kakmacı, Ö. (2008). İlköğretim yedinci sınıf öğrencilerinin matematikteki hazır bulunuşluk düzeyi [Readiness level of primary school seventh grade students in mathematics]. *Kastamonu Education Journal, 16*(2), 529-542.

Yurdugül, H. (2005). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması [Using content validity indices for content validity in scale development studies]*. XIV. National Educational Sciences Congress (28-30 September), Pamukkale University Faculty of Education.