Journal of Educational Technology & Online Learning
Volume 3 | Issue 3 | 2020
https://dergipark.org.tr/tr/pub/jetol

Predicting Academic Achievement with Machine Learning Algorithms

Muhammed Berke YILDIZ[a]
Caner BÖREKCİ[b]

[a] kaplanke10@hotmail.com; Şehit Prof. Dr. İlhan Varank Science and Art Center, Balıkesir, Turkey; ORCID: 0000-0002-3586-6723
[b] canerborekci@hotmail.com, Şehit Prof. Dr. İlhan Varank Science and Art Center, Balıkesir, Turkey; ORCID: 0000-0001-5749-2294

Doi: 10.31681/jetol.773206
Suggested citation: Yıldız, M.B., Börekçi, C. (2020). Predicting Academic Achievement with Machine Learning Algorithms. *Journal of Educational Technology & Online Learning, 3*(3), 372-392.

Article Info

Received : 08.05.2020
Revised : 03.06.2020
Accepted : 23.06.2020

Research Article

Abstract

Education systems produce a large number of valuable data for all stakeholders. The processing of these educational data and making studies on the future of education based on the data reveal highly meaningful results. In this study, an insight was tried to be developed on the educational data collected from ninth-grade students by using data mining methods. The data contains demographic information about students and their families, studying routines, behaviours of attending learning activities, and their epistemological beliefs about science. Thus, this research aimed to solve a classification problem, two-class (successful or unsuccessful according to the exam result) was tried to be estimated from the collected data. In the study, the prediction accuracy of the supervised classification algorithms were compared and it was defined which variables were effective in the formation of classes. When the prediction accuracy of machine learning algorithms was compared, the findings indicated that the Neural Network algorithm (98.6%) had the highest score. The accuracy rate of the other algorithms are kNN (86.2%), Logistic Regression (78.4%), SVM (90.3%), Decision Tree (91.9%), Random Forest (90.0%), and Naive Bayes (81.7%). The information gain coefficient of the variables was examined to determine the factors affecting the prediction accuracy. It was revealed that demographic variables of the family, scientific epistemological beliefs of the student, study routines and attitudes towards some courses affected the classification. It can be concluded that there was a relationship between these variables and academic success. Studies on these variables will support students' academic success.

Keywords: Educational Data Mining, Machine Learning, Academic Achievement

1. INTRODUCTION

Data mining is a multidisciplinary field that acts as a bridge between many technical fields such as database technology, statistics, artificial intelligence, machine learning, pattern definition and data visualization (Özekes, 2003). With data mining, relationships and rules can be obtained from large amounts of data and future predictions can be made (Norton, 1999; Romero & Ventura, 2007). With the relations and rules have obtained, complex data patterns have become clear, important information has discovered and future predictions have made (Koh & Tan, 2011; Savaş, Topaloğlu & Yılmaz, 2012). Today, data mining is widely used in many fields

such as business, insurance, marketing, medicine, biology, telecommunications, education, etc. (İnan,2003; Romero & Ventura, 2013).

The implementation of data mining in education is called educational data mining and it is an interdisciplinary field (Romero and Ventura, 2007). Educational data mining is based on general data mining which is used to examine and the analysis of data obtained from educational environments. The initial aim is to improve the educational outcomes of the students, to understand the educational environments they learn, to understand better how they learn, how they define learning, and how they learn and explain educational phenomena (Romero & Ventura, 2007, 2013; Baradwaj & Pal, 2012, Osmanoğlu et all., 2020). The data obtained with educational data mining from any learning environment, which is conducive and promotes learning, can be analyzed (Baker, 2010).

Educational data mining methods have different sources such as data mining, machine learning, psychometrics and computational modeling, statistics and information visualization (Algarni, 2016). Romero and Ventura (2013, 2020) state that educational data mining studies can be used for:

- estimating student performance,
- testing or developing technology-supported learning theories,
- to give teachers feedback on how to improve the educational outcomes of students,
- personalization of the service provided to students,
- to give students suggestions about learning activities, tasks, problems to be solved or lessons to be completed,
- to inform the stakeholders of education about the undesired behaviours of the students during the learning process,
- developing and organizing cognitive models for students to present their knowledge and skills,
- grouping or defining students profile ,
- to help teachers to develop and conduct the content of a lesson and planning lessons for the future.

Baker (2010) classified educational data mining studies into five main categories; prediction, clustering, relationship mining, distilation of data for human judgement, and discovery with models. These categories and some studies are exemplified here.

(1) Prediction models: In prediction, the aim is to develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Prediction models can be used for identify students at risk or predicting

students' achievements. Cha, Kim, Park, Yoon, Jung, and Lee (2006) preferred decision trees and hidden Markov models from classification algorithms to classify students' learning styles and preferences. The data, collected through a smart learning system involving 600 higher education students, were analyzed. Dekker, Pechenizkiy and Vleeshouwers (2009) used decision trees to predict whether students studying in the electrical and electronic engineering department would drop out of school at the end of the first year. Also, they determined the success criteria for the electrical and electronic engineering department. Cano and Leonard (2019) utilized genetic programming classification algorithms in their studies and they developed an early warning system to predict the academic performance of university students. They stated that the results obtained from the study would help teachers and policymakers to improve education.

(2) Clustering models are used for clustering the data which have same attribute in the data set. Determining students' behaviours and learning styles can be given as an example. Vale, Madeira and Antunes (2014) conducted a block cluster analysis using the decision trees algorithms to determine the study areas of 443 graduate students who got a degree between 1997 and 2012. D'Mello (2017) used a learning analytics tool to collect interaction patterns and body language data and analyzed this data for the effects of emotions on learning. Şahin, Keskin, Özgür and Yurdugül (2017) collected data about students' interactions and self-regulation skills from an online course named “computer networks and communication”. The data were analyzed with grey interaction analysis and optimal scaling analysis and they clustered students depending on their interactions. Milliecamp, Broos, De Laet, and Verbert (2019) collected data online and used visualization tools to identify users' behaviour and improve their learning.

(3) Relationship mining is used for revealing the relationship between two or more variables. One of the areas of relationship mining is revealing the relationship between students' academic achievement and parents' attitudes. Another area is determining the effect of a learning method on academic success. The four sub-categories are (a) association rule mining, (b) correlation mining, (c) sequential pattern mining, (d) causal data mining.. Rashid, Asif, Butt, and Ashraf (2013) analyzed collaboration and sequential pattern mining by processing feedback on students' own performance in the student information system. They stated that GPS (generalized sequential pattern mining) gave more successful results than Apriori algorithm in terms of finding the relationship between data in text format. Dalkılıç and Aydın (2017) set out the association rules to reveal the causes of university students' absenteeism by using the apriori algorithm.

(4) Distillation of data for human judgement is the fourth type of categories. This model aims to find new ways to define or classify the characteristics of the data and can be used to identify patterns in students' behaviour, learning styles, collaborative work. In their research, Scheuer and McLaren (2012) analyzed students' scientific research performances, research design skills and experimental design skills using data mining methods on the data they obtained from the online research teaching system. Hernández-García, Acquila-Natale, Chaparro-Peláez, and Conde (2018) analyzed data in a learning analytics system to predict students' performance in their collaborative work on a project-based learning task.

(5) Another categort is Discovery with models and in this models, a model of a phenomenon is developed via prediction, clustering and then used as a component in another analysis, such as prediction or relationship mining. Determining the relationship between students 'characteristics to determine students' behaviors can be an example of such studies. de Carvalho and Zarate (2019) conducted a causality investigation on the features of the training dataset that was created during a 20-week online course on Algorithm and Data Structures. They determined what kind of behaviour patterns had what kind of results. Wong et al. (2019) created a model in their analysis of how learning theories and learning analytics tools are used in educational research. Wong et al (2019) created a model for how learning theories and learning analytics tools are used in educational research. As a result of their research, they revealed that learning theories in learning analytics tools are used in two ways. First, learning theories guide what kind of data will be collected and the data collection approach. The second one helps in explaining students behaviors' to achieve success. Bravo-Agapito, Frances, and Seaone (2019) analyzed studies that using educational data mining methods to improve foreign language learning. They concluded that studies were usually done to estimate student performance, control students' motivation and provide feedback to teachers. Botelho, Baker and Heffernan (2019) aimed to create a new model that perceives students' behaviour and emotions by using machine learning techniques. They stated that the use of the feature selection method is a more consistent way to develop high-performance models

It has been observed that educational data mining studies have increased in recent years (Romero and Ventura, 2020). Based on the results of these studies, we can concluded that educational data mining can be used in different fields to solve problems, and make predictions. In this study, a model was tried to be developed on the educational data collected from high school students by using data mining methods. Demographic information about themselves and their families, HSE (High School Exam) study routines, class engagement behaviours and

scientific epistemological beliefs were collected from 9th-grade students. The research considered as a classification problem, and two classes (being successful in the exam or not) tried to be estimated. Prediction accuracy of the supervised classification algorithms was compared and at the same time, the variables that affect the formation of classes have determined. Based on the studies in the literature, it was understood that the participants of the studies conducted in the field of educational data mining were generally university students and the data collected from the learning analytics or student information systems of universities. Many phenomena such as academic success, learning styles, and behavioural styles of students were analyzed in the studies. In this current research, a model was created by analyzing various data collected from high school students with classification algorithms. With the model created, it is aimed to determine the academic success of students who are preparing for HSE and to reveal the inputs that will affect their academic success.

2. METHODOLOGY

Data mining methods are used in this research to create foresight from data and develop strategies accordingly. In this study, the classification method, which takes place under data mining techniques, used for collecting educational data, determining prediction models, classifying and testing the classification accuracy. The CRISP-DM (Cross Industry Standard Process Model - Data Mining) methodology applied in the research process for machine learning. The CRISP methodology proposes step-by-step procedures to make the process reliable and standard. This methodology; follows a cyclical process that includes understanding the problem, understanding the data, data preparation, modelling, implementation, evaluation, and reporting (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, and Wirth, 2000; Almahadeen, Alkaya and Sarı, 2017; Demirkol, Kartal, Şeneler, and Gülseçen, 2019). The path followed in the research process has shown in Figure 1.

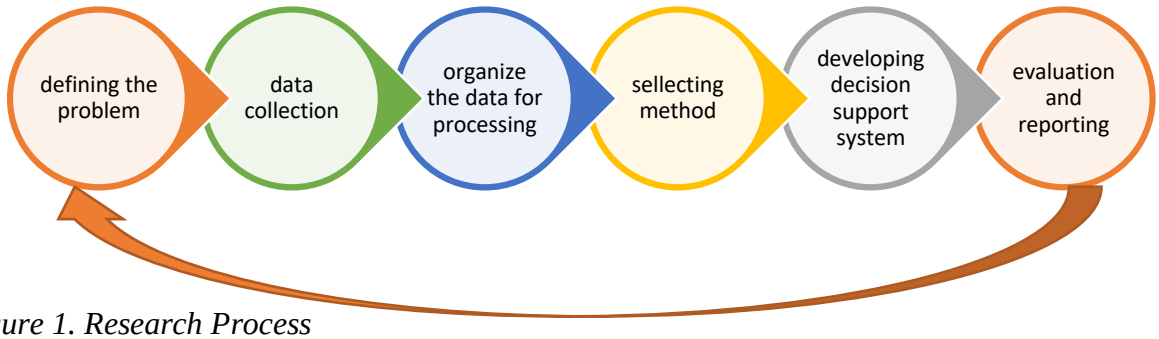


Figure 1. Research Process

421 9th-grade students participated in this study. 160 of them were successful in HSE and the other 261 were not. Participants of the study are 421 9th-grade students. 160 of them were successful in HSE and the other 261 were not. Students were studying at 14 different schools and data were collected from approximately 30 students from each school. Types of the schools are science high schools, Anatolian high schools, vocational and technical schools and imam hatip high schools.

The data collection tool created by the researchers consists of an information form and a scientific epistemological belief scale. The education level of the parents of the students, the income level of their families, the interest level of the families to students' school life, the existence of their study environment, the study routines, the general attitudes and behaviours in the lessons, the support from other sources during the preparation process for HSE (private lessons, private teaching institution, extra school courses), daily study hours, number of questions they solved in a day, and the attitudes to the lesson which they liked (Turkish, mathematics, science, history of the Turkish revolution, foreign language, religious culture and moral knowledge) took place in the information form. The parameters of the form based on the results of various researches in the literature which studied the factors affecting academic success (Savaş, Taş & Duru, 2010; Özer & Korkmaz, 2016; Aslanargun, Bozkurt, & Sarıoğlu, 2016; Bıyıklı, 2017; Gök, 2017; İncirci, İlğan, Sirem & Bozkurt, 2017; Sari, Arikan & Yıldızlı, 2017; Uzun & Bökeoğlu, 2017; Yenice, Hiğde & Özden, 2017; Börekci & Uyangör, 2018).

The scientific epistemological beliefs scale was developed by Elder (1999) and have been used for collecting data from students' scientific epistemological beliefs. Adaptation study of the scale has been carried out by Acat, Tüken and Karadağ (2010). The scale is a 5-point Likert type scale and consists of 25 items. It consists of five (5) sub-dimensions called (i) Authority and Integrity, (ii) Information Generation Process, (iii) Source of Information, (iv) Reasoning and (v) Variability of Information. According to Acat, Tüken and Karadağ (2010) explanations of the sub-dimensions were listed below;

(i)Authority and Truth: Scientific knowledge is precise and comes from authority. The sub-dimension includes undeveloped beliefs about the origin and precision of scientific knowledge. It is stated in the items that absolute truth exists, information comes from a source other than individuals and authority have the information.

(ii)Information Generation Process: Scientific knowledge is of empirical origin. Observation and experimentation play an important role in the creation of scientific knowledge. Items in the

subdimension state that the role of the experiment in the creation of scientific knowledge and the student beliefs about the questioning of evidence and decision-making period.

(iii)Source of Information: Information obtained from books and teachers is always correct. Items in the sub-dimension state that searching the source of scientific knowledge in books/teachers indicates students' undeveloped/immature beliefs.

(iv)Reasoning: The scientist is curious and creates scientific information based on his initial knowledge, observations and logic. Items in the subdimension state that the scientist uses reasoning and logic in the process of creating scientific information.

(v)Variability of Knowledge: Scientific knowledge is not certain. Items in the subdimension state that scientists create and test explanations about nature using observation, experiment, theoretical and mathematical models. They change their views when they encounter new experimental evidence that does not fit the existing explanations.

In this research, the main purpose of collecting data about students 'scientific epistemological beliefs is to determine students' views on the structure, source, limits and development of scientific knowledge. Students' epistemological beliefs positively affect their use of data processing strategies, learning and controlling learning materials from a metacognitive perspective, showing academic success, showing positive attitudes towards school and forming deep, complex thoughts (Deryakulu & Büyüköztürk, 2005). The scientific epistemological beliefs of the students are the predictors of their academic success (Evcim, 2010; Yenice, Hiğde & Özden, 2017; Kanadlı & Akay, 2019). Feature information regarding the collected data is given in Table 1. It consists of 29 independent 1 dependent variables.

Table 1.

Variables of The Study

Feature	**Type**
Gender	Categorical (female / male)
Mother' educational status	Categorical (be literate, primary, secondary, high school, university, postgraduate)
Mother' educational status	Categorical (be literate, primary, secondary, high school, university, postgraduate)
Family Income	Categorical (< 2000 TL, between 2000-7000 TL, > 7000 Tl üstü)
Do you have own studying room?	Categorical (yes / no)
Do you take notes during the lesson and make homeworks?	Categorical (yes / sometimes / no)
Are lessons hard for you?	Categorical (yes / sometimes / no)
Do you think your exam scores are below your potential?	Categorical (yes / sometimes / no)
Do you like going to the blackboard and lecture a lesson?	Categorical (yes / sometimes / no)
Did your family supports you for being successful in your lessons?	Categorical (yes / sometimes / no)
Did your family provide a suitable working environment for you to be successful in your lessons?	Categorical (yes / sometimes / no)
Did your family attend parent-teacher meetings and school events?	Categorical (yes / sometimes / no)
Did your family communicating with you about school?	Categorical (always / sometimes / never)
How many hours in a day you were studying?	Categorical (< 1 hour / 1-2 hour / 2-3 hour / 3 or more hour)
Did you continue your courses regularly?	Categorical (yes / no)
Have you taken lessons from an institution or person other than school?	Categorical (private teaching enstitute / Private tutor / no)
On average, how many test questions did you solve per day?	Categorical (0-50 / 50-100 / 100-200 / > 200)
Lesson Attitudes (Turkish, mathematics, science, history of the Turkish revolution, foreign language, religious culture and moral knowledge)	Categorical (1 to 5)
The scientific epistemological beliefs scale (a) Authority and Integrity	Scale (1 to 5)
(b) Information Generation Process	Scale (1 to 5)
(c) Source of Information	Scale (1 to 5)
(d) Reasoning	Scale (1 to 5)
(e) Variability of Information	Scale (1 to 5)
Succesful in HSE (target varible)	Categorical (1, 0)

The collected data were analyzed using the Orange 3 Data Mining tool. Orange was developed by Ljubljana University, Faculty of Computer and Information Science, Bioinformatics Laboratory using Python programming language (Demsar et al, 2013). In the analysis process, 7 different classification algorithms were used. These are kNN (k = 3), Logistic Regression, SVM, Neural Networks, Naive Bayes, Decision Tree and Random Forest algorithms. k-fold

cross-validation method used for all algorithms. In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k−1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. In this study, the value of k was taken as 10. This method aims to determine the best classifier for this dataset. In the analyze process accuracy rates of classification algorithms were calculated, confusion matrices were prepared and ROC analysis was performed. At the same time, in order to increase the success of classification algorithms, the information gain value of the features in the data set was calculated. Histogram graphiscs of features with high values were drawn. The system architecture is given in Figure 2.

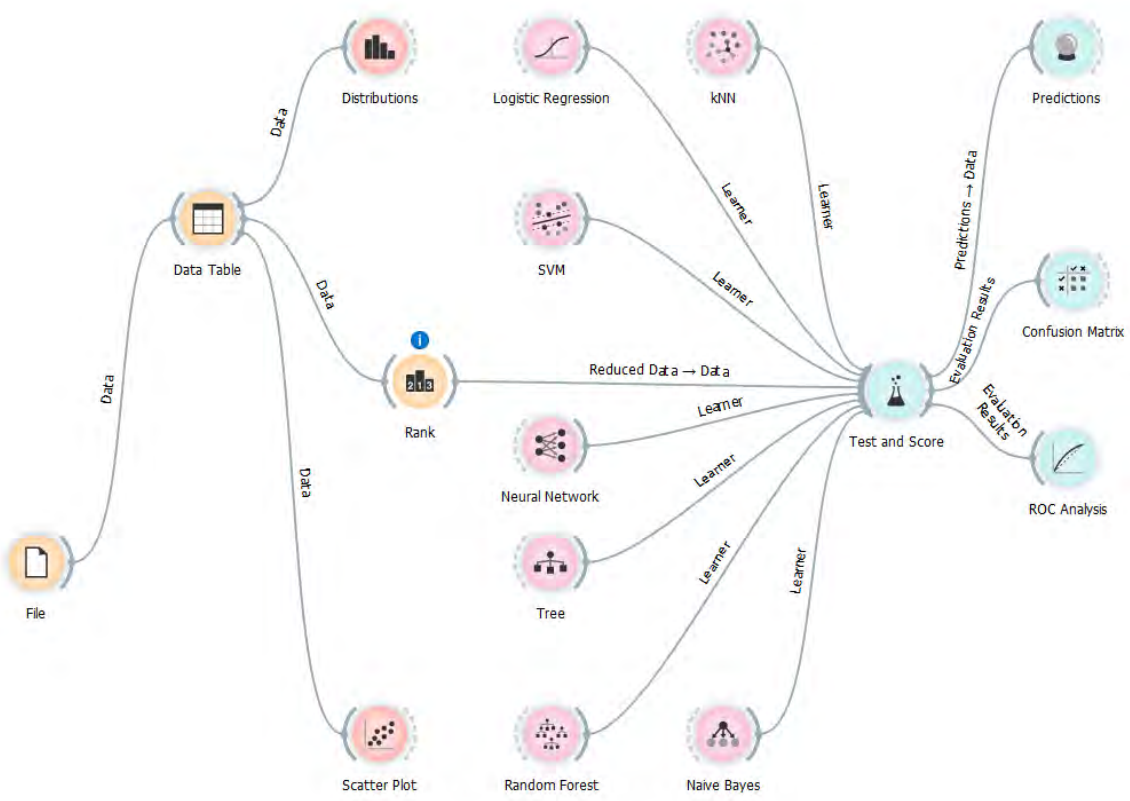


Figure 2. Model Architecture

3. FINDINGS

After the pre-processing of the data set, 7 different classification algorithms were used to determine the success of the students. For classification kNN (k = 3), Logistic Regression,

SVM, Neural Network, Decision Tree, Random Forest, and Naive Bayes algorithms were used then the results were examined. To increase the success of classification algorithms, information gain values of the independent variables in the data set were calculated and the results are given in Figure 3.

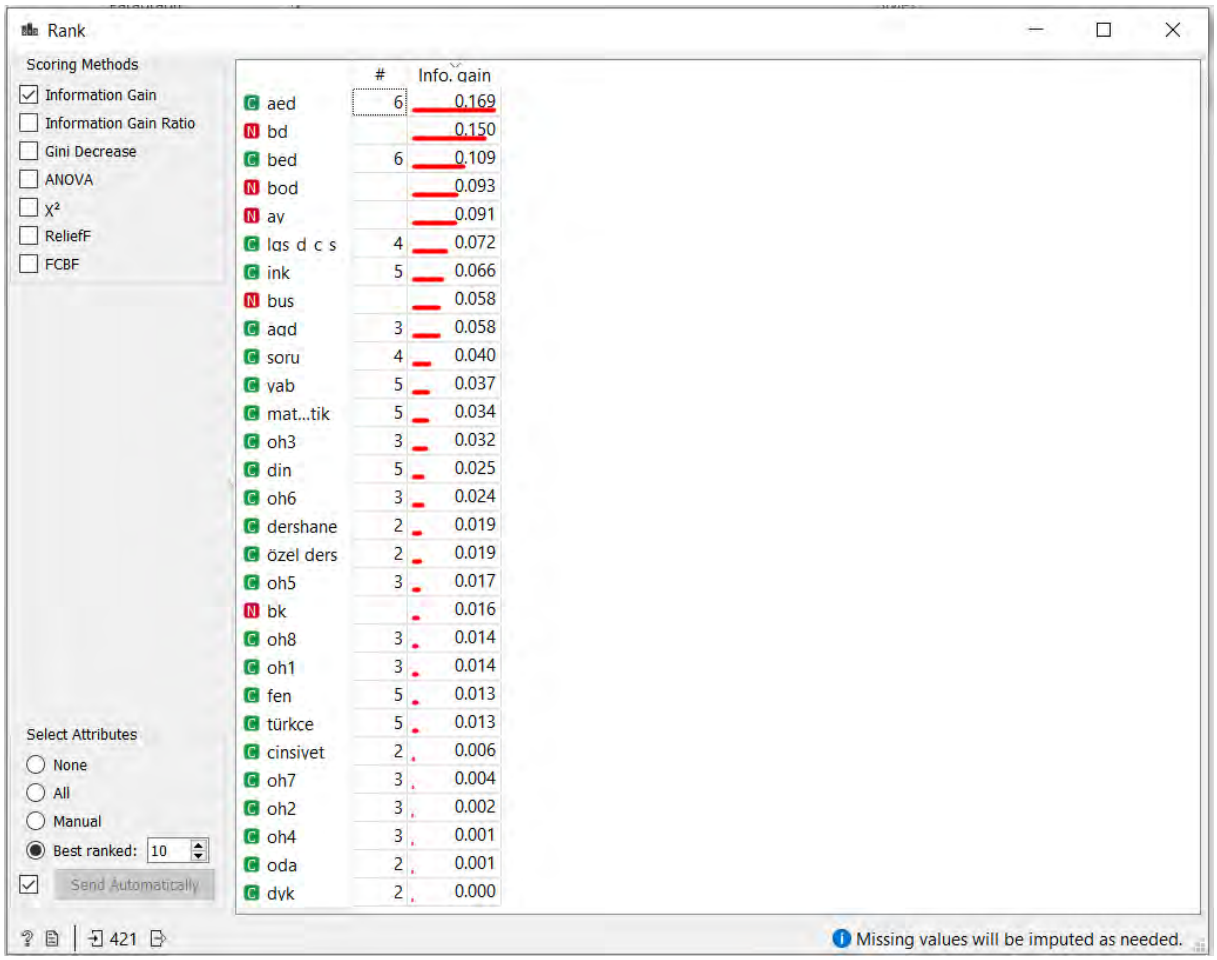


Figure 3. Information Gain values of features

When the results were analyzed, the order of 10 independent variables, in which the greatest value is determined, is (1) education level of the mother, (2) variability of ınformation, (3) education level of the father, (4) authority and integrity, (5) reasoning, (6) study duration per day, (7) history of the Turkish revolution, (8) information generation process, (9) family income, (10) solved problem per day. The size effect of the information gain coefficient is in these variables indicates that it plays an important role in the formation of classes. The histogram plots of these variables and the average values of the distributions are given in Figure 4. The red columns represent high school students who were successful in the HSE (1), and the blue columns represent high school students who were not successful in the HSE (0).

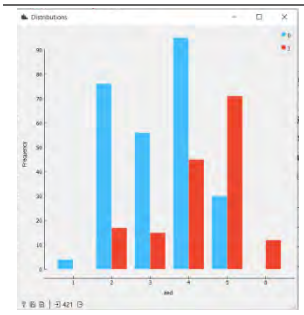
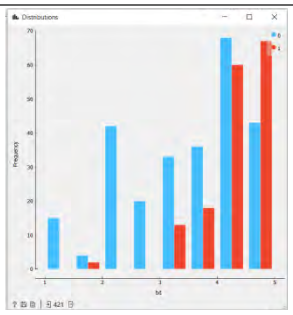
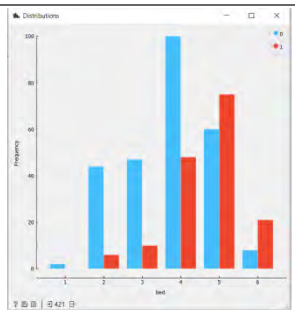
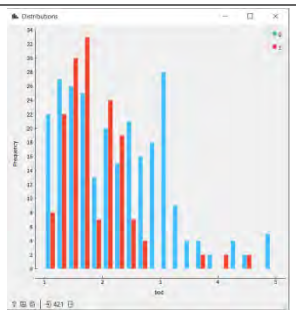
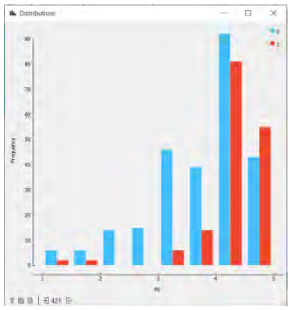
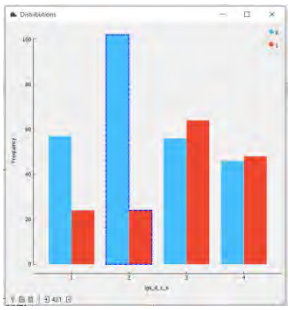
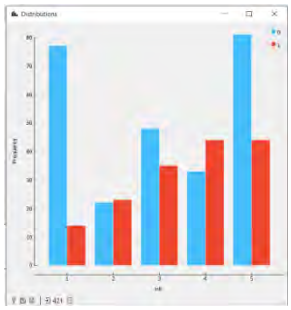
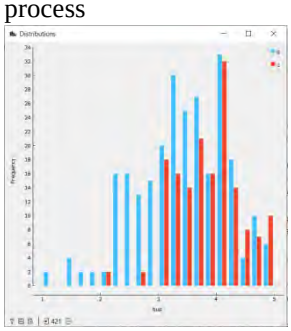
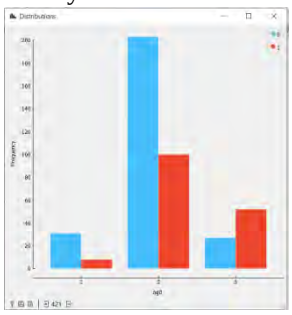
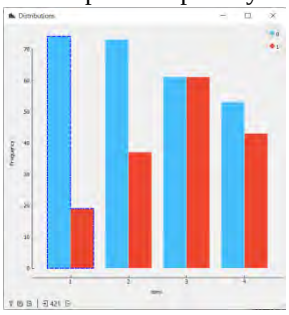
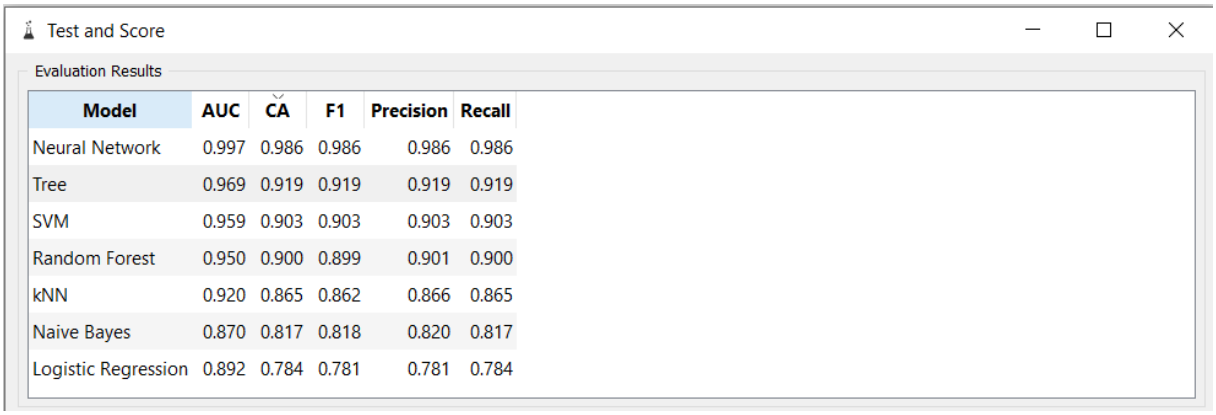
Education level of the mother	Variability of information	Education level of the father	Authority and integrity
			
0 = 3.27 1 = 4.29	0 = 3.43 1 = 4.25	0 = 3.75 1 = 4.59	0 = 2.26 1 = 1.80
Reasoning	Study duration per day	History of the Turkish revolution	Information generation process
			
0 = 3.71 1 = 4.25	0 = 2.35 1 = 2.85	0 = 3.07 1 = 3.51	0 = 3.40 1 = 3.83
Family income	Solved problem per day		
			
0 = 1.98 1 = 2.28	0 = 2.36 1 = 2.80		

Figure 4. Histogram graphics and mean scores of variables

Figure 4 shows the most effective variables in the data set on being successful in HSE. Average values of these variables were calculated for both classes. It is understood from the distribution of the graphs and the mean values that the scores of the students who succeed in HSE higher than the others (except for the Authority and Accuracy factor; the low score here is the accepted perspective of the scientific epistemological belief). The presence of a similar distribution was also observed in the graphs of other variables. When the results of the prediction accuracy of the algorithms are examined (Figure 5) it is seen that the Neural Network algorithm has the

highest prediction success rate. The accuracy rate of the algorithm is 98.6%. In other words, the Neural Network algorithm accurately predicted 98.6% of 421 entries in the dataset. The lowest success rate is the Logistic Regression algorithm (78.4%).



Test and Score

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.997	0.986	0.986	0.986	0.986
Tree	0.969	0.919	0.919	0.919	0.919
SVM	0.959	0.903	0.903	0.903	0.903
Random Forest	0.950	0.900	0.899	0.901	0.900
kNN	0.920	0.865	0.862	0.866	0.865
Naive Bayes	0.870	0.817	0.818	0.820	0.817
Logistic Regression	0.892	0.784	0.781	0.781	0.784

AUC: Area under ROC is the area under the receiver-operating curve.

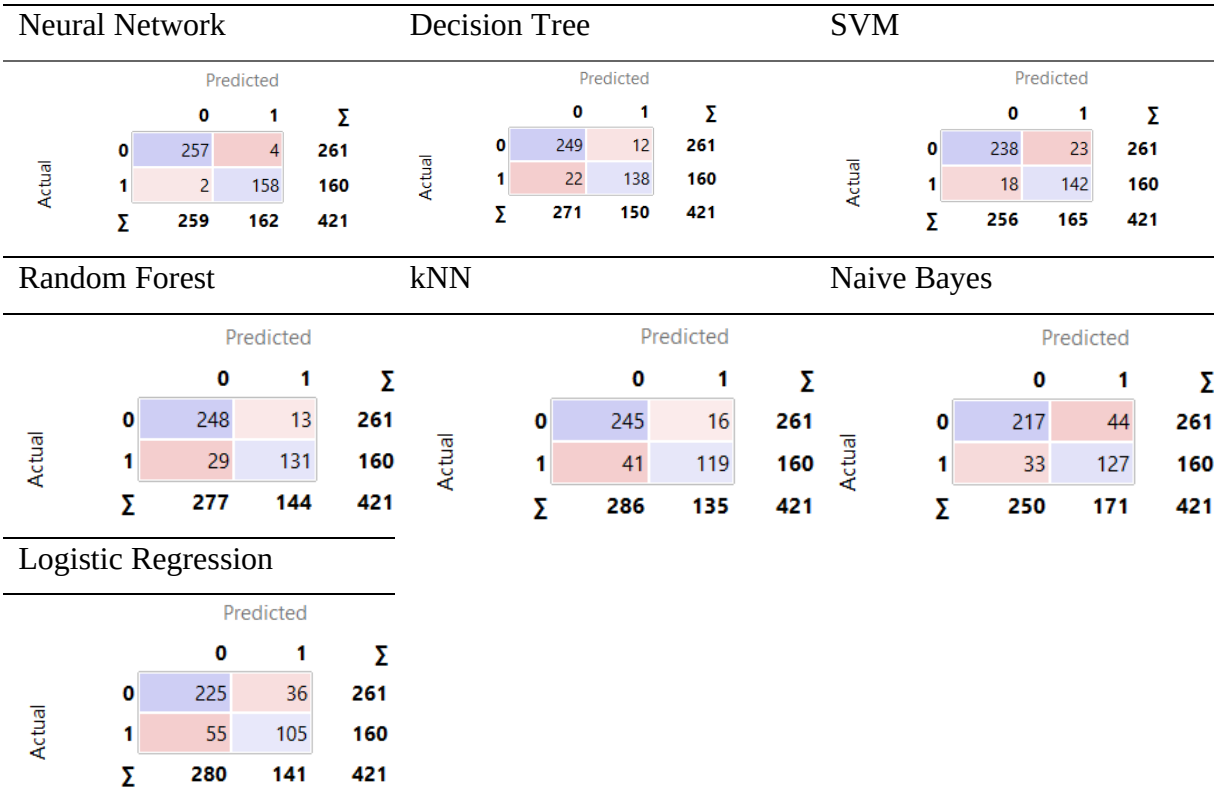
CA: Classification accuracy is the proportion of correctly classified examples.

F-1is a weighted harmonic mean of precision and recall.

Precision is the proportion of true positives among instances classified as positive.

Recall is the proportion of true positives among all positive instances in the data.

Figure 5. Classification Accuracy of algorithms



Neural Network

Actual \ Predicted	0	1	Σ
0	257	4	261
1	2	158	160
Σ	259	162	421

Decision Tree

Actual \ Predicted	0	1	Σ
0	249	12	261
1	22	138	160
Σ	271	150	421

SVM

Actual \ Predicted	0	1	Σ
0	238	23	261
1	18	142	160
Σ	256	165	421

Random Forest

Actual \ Predicted	0	1	Σ
0	248	13	261
1	29	131	160
Σ	277	144	421

kNN

Actual \ Predicted	0	1	Σ
0	245	16	261
1	41	119	160
Σ	286	135	421

Naive Bayes

Actual \ Predicted	0	1	Σ
0	217	44	261
1	33	127	160
Σ	250	171	421

Logistic Regression

Actual \ Predicted	0	1	Σ
0	225	36	261
1	55	105	160
Σ	280	141	421

Figure 6. Confusion Matrices

To compare the classification algorithms, the area under the ROC (Receiver Operating Characteristic) curve was plotted for each (Figure 7). ROC curve allows to determine the discriminative power of the test and to compare the effectiveness of various tests. The size of the area under the curve (AUC) indicates the success of the algorithm used (Tomak & Yüksel, 2009). As, it can be concluded from Figure 5 and Figure 7, the algorithm with the largest area under the ROC is the Neural Network algorithm (0.997), followed by the Decision Tree algorithm (0.969). The AUC of the Logistic Regression algorithm has the lowest area (.892).

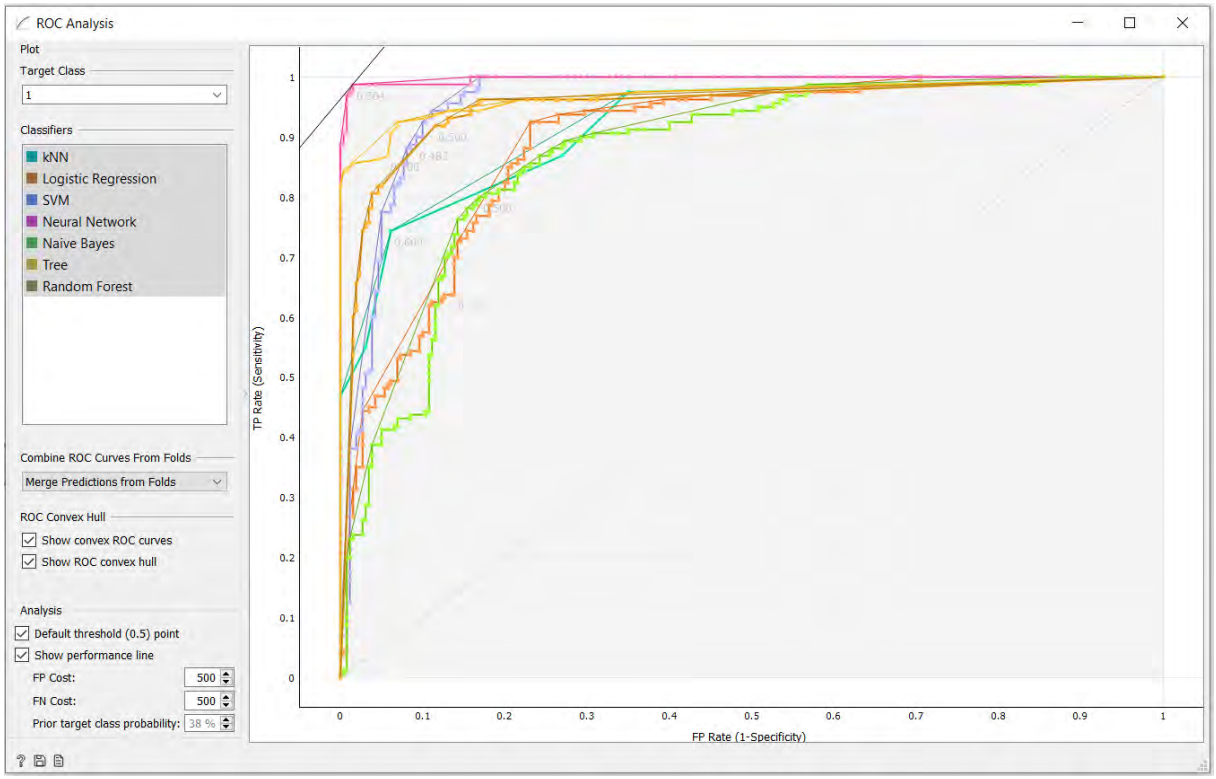


Figure 7. ROC Analysis

4. DISCUSSION AND CONCLUSION

In this study, a data set created with students' scientific epistemological beliefs, study routines, behaviours in a particular course, demographic information about families, and attitudes to courses. The data set used for the solution of a classification problem (being successful in HSE). a model was developed with classification algorithms and these algorithms accuracy rates were calculated. Also, the variables that play a role in the formation of classes was examined. According to the findings, the Neural Network (98.6%) algorithm has the highest classification accuracy rate. The education levels of mother and father, the income level of the family, the

students' scientific epistemological beliefs, study routines and attitudes towards the lessons are effective on HSE success. Numerous studies using classification algorithms for the prediction of academic success have been conducted with different data sets. Some of these studies conducted with high school students were given in this section. Márquez-Vera, Cano, Romero and Ventura (2013) states that ADTree algorithm (99.7%) is the most successful algorithm for estimating high schools students overall performance and the most effective factors in the 77 factors were students ages, motivation levels, physical wellbeing's, smoking habits and their working groups. Márquez - Vera et al. (2016) in another study examined the school dropout reasons of high school students, they find that the JRip algorithm (96%) has highest prediction success and stated that the variables that caused school dropout were students success in secondary school, alcohol smoking habits, working environments, motivations for learning, and success in mathematics, social and humanities courses. When studies with university students examined; Ha, Loan, Giap and Huong (2020) used eight different classification algorithms for the prediction of student achievement and compared their achievements for five grades (A, B, C, D and F). They concluded that the Naive Bayes algorithm (86.19%) had the highest success rate. Saa, Al-Emran, and Shaalan (2019) used seven different algorithms to predict the academic success of university students. While the random forest algorithm (75.52%) yielded the most successful results, they revealed that the factors affecting the success are information about the high school, university entrance exam, and the performance of the student in the previous courses. Similarly, while working on the data of university students, Roy and Garg (2017) concluded that the J48 algorithm (73.92%) was more successful, they concluded that the students' health status, education of their families, alcohol use and friend relationship were among the factors affecting their success. Al-Saleem, Al-Kathiry, Al-Osimi, and Badr (2015) used the J48 and ID3 algorithms to create a model for predicting students' achievements and tried to determine their success in various elective courses with these algorithms. As a result, they concluded that the J48 (83.75%) algorithm gives more successful results than ID3 (69.27%). Strecht, Cruz, Soares, and Mendes-Moreira (2015) compared the success of classification and regression algorithms to predict academic success. They concluded that between these two approaches, classification algorithms are more successful. In the study, they used the university' academic database and they reveal that the SVM algorithm has the highest prediction success. Guo, Zhang, Xu, Shi, and Yang (2015) created a deep learning architecture and used 4 different algorithms to compare students' final grades and their achievements. They stated that the most successful algorithm for that problem was the neural network algorithm,

which they named SPPN (Student Performance Prediction Network) with 77.2% success rate. In the other study conducted by Affendey, Paris, Mustapha, Sulaiman, and Muda (2010), the effects of the courses that taken by students are examined and the predictive success of the algorithms are compared, and the Naive Bayes algorithm has a 95.29% success rate and 5 courses (Computer Programing 2, Multimedia Technology, Computer Organization, Assembly Language and Programming Language) have shown that have a great effect on success.

5. SUGGESTIONS AND LIMITATIONS

When the research findings and the studies in the field are evaluated together, it was concluded that different algorithms were successful on different datasets, in other words, different solutions were depending on the nature of the problem to be solved. It is understood that the success of the algorithms used in the researches is directly correlated with the collected data which affective on the problem result. In this study, it was revealed that the education level of the family, the thoughts of the students about the nature of scientific knowledge, students' study routines, and the income level of their family play an important role on the success of the algorithm which was used for predicting academic achievement.

Also, there were some limitations to this research. A two-class educational data mining problem was discussed in this article. The data collected were collected from 421 9th-grade students at 14 different schools in the city centre. 29 features used for predicting exam success. It is thought that level of algorithms prediction success is about these limitations. A similar problem can be applied to different settlement levels and locations or larger participant groups. Standard models can be developed by including different variables that affect academic success. Another important issue that stands out here is, the intervention areas can be determined by examining this research and other researches to increase the academic success of students.

Makine Öğrenimi Algoritmalarıyla Akademik Başarı Tahmini

Özet

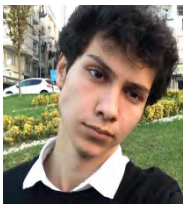
Eğitim sistemleri eğitimin bütün paydaşları için, çok sayıda ve değerli veriler üretir. Bu eğitsel verilerin işlenmesi ve veriye dayanarak eğitimin geleceği ile ilgili çalışmalar yapmak son derece anlamlı sonuçlar ortaya çıkarmaktadır. Bu çalışmada dokuzuncu sınıf öğrencilerinden toplanan eğitsel veriler üzerinde veri madenciliği yöntemleri kullanılarak bir görüş geliştirilmeye çalışılmıştır. Sınavla öğrenci alan ve sınavsız öğrenci alan ortaöğretim okullarında eğitime devam eden öğrencilerden kendileri ve aileleri ile ilgili demografik veriler, bir önceki yılda LGS (Liselere Geçiş Sınavı) öncesi ders çalışma ve derse katılma

davranışları ve bilimsel epistemolojik inançlarına ilişikin veriler toplanmıştır. Araştırma bir sınıflandırma problemi olarak ele alınmış ve iki sınıf (sınavla öğrenci alan okulda okuyan ve sınavsız öğrenci alan okulda okuyan) veriler üzerinden tahmin edilmeye çalışılmıştır. Araştırmada gözetimli sınıflandırma algoritmalarının tahmin başarıları karşılaştırılmış ve sınıfların oluşmasında hangi değişkenlerin etkili olduğu tespit edilmiştir. Makine öğrenmesi algoritmalarının tahmin başarıları karşılaştırıldığında en yüksek başarıyı Neural Network algoritmasının (%98.6) gösterdiği bulunmuştur. Tahmin başarısını etkileyen faktörleri belirlemek için değişkenlere ait Information Gain katsayısına bakılmış. Ailenin demografik bilgileri, öğrencinin bilimsel epistemolojik anlayışı, ders çalışma alışkanlıkları ve bazı derslere olan tutumlarının en çok etki eden değişkenler olduğu ortaya çıkarılmıştır. Tespit edilen değişkenlere yönelik yapılacak iyileştirme çalışmalarının öğrencilerin akademik başarıları üzerinde olumlu etkisinin olacağı düşünülmektedir.

Anahtar kelimeler: Eğitsel Veri Madenciliği, Makine Öğrenmesi, Akademik Başarı

About the Author(s)

Muhammed Berke YILDIZ



Muhammet Berke Yıldız is a computer science student at Şehit Prof. Dr. İlhan Varank Science and Art Center (BILSEM). He is continueing his education at Şehit Turgut Solak Science Highschool. His interest areas are: Programming, machine learning, data mining and text mining.

Mailing Address: Şehit Prof. Dr. İlhan Varank Bilim ve Sanat Merkezi,Balıkesir, TÜRKİYE, 10100
E-mail: kaplanke10@hotmail.com

Caner BÖREKCİ

Caner Börekci is a computer science teacher at Şehit Prof. Dr. İlhan Varank Science and Art Center (BİLSEM). He gained a Ph.D. in Curriculum and Instruction department at Balıkesir University in 2018. His interest areas are curriculum design, program evaluation, teacher education, philosophical, historical and social foundations of education, instructional design, and programming.

Mailing Address: Şehit Prof. Dr. İlhan Varank Bilim ve Sanat Merkezi,Balıkesir, TÜRKİYE, 10100
E-mail: canerborekci@hotmail.com

REFERENCES

Acat, M. B., Tüken, G., & Karadağ, E. (2010). Bilimsel epistemolojik inançlar ölçeği: Türk kültürüne uyarlama, dil geçerliği ve faktör yapısının incelenmesi. *Türk Fen Eğitimi Dergisi*, 7(4), 67-89.

Affendey, L. S., Paris, I. H. M., Mustapha, N., Sulaiman, M. N., & Muda, Z. (2010). Ranking of influencing factors in predicting students' academic performance. *Information Technology Journal*, 9(4), 832-837.

Algarni, A. (2016). Data mining in education. *International Journal of Advanced Computer Science and Applications*, 7(6), 456-461.

Almahadeen, L., Akkaya, M., & Sari, A. (2017). Mining student data using CRISP-DM model. *International Journal of Computer Science and Information Security*, 15(2), 305.

Al-Saleem, M., Al-Kathiry, N., Al-Osimi, S., & Badr, G. (2015). Mining educational data to predict students' academic performance. In International Workshop on Machine Learning and Data Mining in Pattern Recognition (pp. 403-414). Springer, Cham.

Aslanargun, E., Bozkurt, S., & Sarıoğlu, S. (2016). Sosyo ekonomik değişkenlerin öğrencilerin akademik başarısı üzerine etkileri. *Uşak Üniversitesi Sosyal Bilimler Dergisi*, 9(27/3), 201-234.

Baker, R. (2010). Data mining for education. *International encyclopedia of education*, 7(3), 112-118

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.

Bıyıklı, C. (2017). Ortaokul öğrencilerinin Türkçe dersi akademik başarıları ile ders çalışma alışkanlıkları arasındaki ilişki. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 42*(42), 59-73.

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2019). Machine-learned or expert-engineered features? Exploring feature engineering methods in detectors of student behavior and affect. In The twelfth international conference on educational data mining, Montréal, Canada.

Börekci, C., & Uyangör, N. (2018). Family attitude, academic procrastination and test anxiety as predictors of academic achievement. *International Journal of Educational Methodology*, 4(4), 219-226. doi: 10.12973/ijem.4.4.219

Bravo-Agapito, J., Frances, C., & Seaone, I. (2019). Data mining in foreign language learning. *WIREs: Data Mining and Knowledge Discovery*, 10(1), e1287.

Cano, A., & Leonard, J. (2019). Interpretable multi-view early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies, 12*, 198–211.

Cha, H. J., Y. S. Kim, S. H. Park, T. B. Yoon, Y. M. Jung, and J.-H. Lee (2006). Learning styles diagnosis based on user interface behaviors for the customization of learning interfaces in an intelligent tutoring system. In Proceedings of the 8th International Conference on Intelligent Tutoring Systems, ITS 2006, volume 4053 of Lecture Notes in Computer Science , 513-524, Springer.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 9, 13.

Dalkılıç, F., & Aydın, Ö. (2017). Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi öğrencilerinin devamsızlık davranışlarını etkileyen faktörler. *Journal of Higher Education & Science/Yükseköğretim ve Bilim Dergisi*, 7(3), 546-553.

de Carvalho, W. F., & Zarate, L. E. (2019). Causality relationship among attributes applied in an educational data set. In Proceedings of the 34th ACM/SIGAPP symposium on applied computing (pp. 1271–1277). Limassol, Cyprus: ACM.

Dekker. G., Pechenizkiy, M., & Vleeshouwers, J. (2009). Predicting students drop out: A case study. In Proceedings of the 2nd International Conference on Educational Data Mining, EDM'09, pages 41-50.

Demirkol, D., Kartal, E., Şeneler, Ç., & Gülseçen, S. (2019). Bir öğrenci bilgi sisteminin kullanılabilirliğinin makine öğrenmesi teknikleriyle tahmin edilmesi. *Veri Bilimi*, 2(1), 10-18.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in Python, *Journal of Machine Learning Research*, 14(Aug), 2349−2353.

Deryakulu, D. ve Büyüköztürk, Ş. (2005). Epistemolojik inanç ölçeğinin faktör yapısının yeniden incelenmesi: Cinsiyet ve öğrenim görülen program türüne göre epistemolojik inançların karşılaştırılması. *Eğitim Araştırmaları*, 18, 57-70.

D'Mello, S. (2017). Emotional learning analytics. In Handbook of learning analytics (p. 115). New York, NY: SOLAR.

Evcim, İ. (2010). İlköğretim 8. Sınıf öğrencilerinin epistemolojik inanışlarıyla, fen kazanımlarını günlük yaşamlarında kullanabilme düzeyleri ve akademik başarıları arasındaki ilişki.Yayımlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi Eğitim Bilimleri Enstitüsü, İstanbul.

Gök, M. (2017). Makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, 5 (3), 139-148.

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting students performance in educational data mining. In 2015 International Symposium on Educational Technology (ISET) (pp. 125-128). IEEE.

Ha, D. T., Loan, P. T. T., Giap, C. N., & Huong, N. T. L. (2020). An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques. *International Journal of Computer Science and Information Security (IJCSIS)*, 18(3).

Hernández-García, Á., Acquila-Natale, E., Chaparro-Peláez, J., ve Conde, M. Á. (2018). Predicting teamwork group assessment using log data-based learning analytics. *Computers in Human Behavior*. doi:10.1016/j.chb.2018.07.016

İnan, O. (2003). Öğrenci işleri veri tabanı üzerinde veri madenciliği uygulamaları. Yayınlanmamış Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

İncirci, A., İlğan, A., Sirem, Ö., & Bozkurt, S. (2017). Ortaöğretim destekleme ve yetiştirme kurslarına ilişkin öğrenci görüşleri. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, (42), 50-68.

Kanadlı, S., & Akay, C. (2019). Schommer'in epistemolojik inançlar modelinin cinsiyet ve akademik başarı açısından incelenmesi: Bir meta-analizi çalışması. *Eğitim ve Bilim*, 44(198), 389-411.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2), 65.

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, 38(3), 315-330.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1), 107-124.

Millecamp, M., Broos, T., De Laet, T., & Verbert, K. (2019). DIY: learning analytics dashboards. In Companion Proceeding of the 9th International Conference on Learning Analytics & Knowledge (LAK'19) (pp. 947-954). Solar.

Norton, M. J. (1999). Knowledge discovery in databases. Library Trends, 48(1), 9.

Osmanoglu,U.O., Atak,O.N., Caglar,K., Kayhan, H. &Can, T.C. (2020). Sentiment Analysis for Distance Education Course Materials: A Machine Learning Approach. Journal of Educational Technology & Online Learning, 3(1), 31-48.

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

Özer, B., & Korkmaz, C. (2016). Yabancı dil öğretiminde öğrenci başarısını etkileyen unsurlar. *Ekev Akademi Dergisi*, 20, 59-84.

Rashid, A., Asif, S., Butt, N. A., & Ashraf, I. (2013). Feature level opinion mining of educational student feedback data using sequential pattern mining and association rule mining. *International Journal of Computer Applications*, 81(10).

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. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355

Roy, S., & Garg, A. (2017). Predicting academic performance of student using classification techniques. In 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) (pp. 568-572). IEEE.

Saa, A. A., Al-Emran, M., & Shaalan, K. (2019). Mining student information system records to predict students' academic performance. In International conference on advanced machine learning technologies and applications (pp. 229-239). Springer, Cham.

Sarı, M. H., Arıkan, S., & Yıldızlı, H. (2017). Factors predicting mathematics achievement of 8th graders in TIMSS 2015. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 8(3), 246-265.

Savaş, E. , Taş, S., & Duru, A. (2010). Matematikte Öğrenci Başarısını Etkileyen Faktörler. *İnönü Üniversitesi Eğitim Fakültesi Dergisi*, 11 (1) , 113-132 .

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

Scheuer, O., & McLaren, B. M. (2012). Educational data mining. *Encyclopedia of the Sciences of Learning*, 1075-1079.

Strecht, P., Cruz, L., Soares, C., & Mendes-Moreira, J. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance. *In International Conference on Educational Data Mining (EDM)*;392-395.

Şahin, M., Keskin, S., Özgür, A., & Yurdugül, H. (2017).E-öğrenme ortamlarında öğrenen özelliklerine dayalı etkileşim profillerinin belirlenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, 7(2), 172 – 192. DOI: 10.17943/etku.297075

Tomak, L., & Yüksel, B., E., K. (2009). İşlem karakteristik eğrisi analizi ve eğri altında kalan alanların karşılaştırılması. *Journal of Experimental and Clinical Medicine*, 27(2), 58-65.

Uzun, G., & Bökeoğlu, Ö. Ç. (2017). Akademik başarının okul, aile ve öğrenci özellikleri ile ilişkisinin çok düzeyli yapısal eşitlik modellemesi ile incelenmesi. *Ankara University Journal of Faculty of Educational Sciences (JFES)*, 52 (3), 655-684. DOI: 10.30964/auebfd.525770

Vale, A., Madeira, S. C., & Antunes, C. (2014). Mining coherent evolution patterns in education through biclustering. In 7th International Conference on Educational Data Mining 2014.

Wong J. et al. (2019) Educational Theories and Learning Analytics: From Data to Knowledge. In: Ifenthaler D., Mah DK., Yau JK. (eds) Utilizing Learning Analytics to Support Study Success. Springer, Cham

Yenice, N., Hiğde, E., & Özden, B. (2017). Ortaokul öğrencilerinin üstbiliş farkındalıklarının ve bilimin doğasına yönelik görüşlerinin cinsiyet ve akademik başarılarına göre incelenmesi. *Ondokuz Mayis University Journal of Education*, 36(2), 1-18.