International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 410–429
https://doi.org/10.21449/ijate.904456
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Estimation of the Academic Performance of Students in Distance Education Using Data Mining Methods

Resul Butuner[1,*], **M. Hanefi Calp**[2]

[1]Ankara Beypazarı Fatih Vocational and Technical Anatolian High School University, Faculty of Education, Department of Computer, Ankara, Türkiye
[2]Ankara Hacı Bayram Veli University, Faculty of Economics & Administrative Sciences, Department of Management Information Systems, Ankara, Türkiye

ARTICLE HISTORY
Received: Mar. 27, 2021
Revised: Feb. 03, 2022
Accepted: Mar. 15, 2022

Keywords:
Distance Education,
Academic Performance,
Estimation,
Data Mining,
Artificial Intelligence.

Abstract: Many institutions in the field of education have been involved in distance education with the learning management system. In this context, there has been a rapid increase in data in the e-learning process as a result of the development of technology and the widespread use of the internet. This increase is in the size of large data. Today, big data can be primarily processed, the relationships between data can be discovered, a meaningful conclusion can be drawn, and predictions about the future using big data can be made. However, these data are generally not used in a way to contribute to the people and institutions (educators, education administrators, ministries, etc.) involved in the education process. Therefore, this study aims to estimate the academic success of students who receive education in the distance education process using data mining methods. The reason why data mining is used is that these methods are particularly effective and powerful tools in classification and prediction processes. The methods used in the study are Random Forest, Artificial Neural Networks, Naive Bayes, Support Vector Machines, Logistic Regression, and Deep Learning algorithms, respectively. The dataset includes primary, secondary, and high school students' data, which were obtained from the learning management system used in the distance education process. As a result, the study findings showed that Deep Learning, Random Forest, and Support Vector Machines algorithms provide prediction success at higher performance than others.

1. INTRODUCTION

Coronavirus (COVID-19) is an epidemic disease that spreads all over the world in a very short time and has fatal consequences. This epidemic has adversely affected many institutions. Therefore, working methods in transportation, industry, health, and education have started to be carried out remotely (Karakaya et al., 2020; Savas et al., 2021). In the field of education, educators continue this process by providing students with education through the method of distance education (Yamamoto & Altun, 2020; Yilmaz & Buyrukoglu, 2021). Some tools are used in distance education. One of these tools is the Learning Management System (LMS). LMS is software created to manage e-learning processes more efficiently and effectively. Using

*CONTACT: M. Hanefi CALP ✉ hanefi.calp@hbv.edu.tr Ankara Hacı Bayram Veli University, Faculty of Economics & Administrative Sciences, Department of Management Information Systems, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

this web tool, an organization can systematically carry out, manage, and measure training activities. LMS provides teachers and students with an online classroom that reinforces learning processes. In online classroom environments, LMS reinforces teachers and students in the learning process. A standard LMS supports an inclusive learning environment for academic progress with interceding structures that promote online collaborative groupings, professional training, discussions, and communication among other LMS users (Dias & Dinis, 2014; Jung & Huh, 2019; Oakes, 2002). It is important to understand the function and responses of the human brain, which is one of the current research topics, and the events of solving the working principle of the brain. It is because it will help to understand the level of learning and changes, especially in individuals (Dogan, 2012). It is considered an academic achievement to state the status of learning with various measurement methods and values such as grades or points (Turgut & Baykul, 2013). It is an important issue in increasing success to know the academic achievement levels of students in advance as it will greatly benefit both students and educational institutions (Luan, 2002). Data such as grades and scores are transformed into information by processing with computer software for a specific purpose (Kurt & Erdem, 2012; Savas, 2021). Therefore, techniques that process data and make them usable have become of great importance today. In this context, the conversion process of raw data into information or meaningful results is realized by data mining methods (Kiray et al., 2015). Statistical methods may not always yield meaningful results in analyzing the data collected in the LMS and revealing meaningful information. Data mining methods are used to process and analyze data on these issues (Beitel, 2005; Luan, 2002; Siemens & Baker, 2012).

Data mining is defined as the manual or automatic processing of large amounts of data to obtain meaningful results by reaching meaningful data. The field of data mining is gradually developing as a result of the processing of data warehouses with the electronic storage and development of data and the development of various analysis tools. It can be seen as obtaining new information by processing large-scale data belonging to students through data mining in the field of education. It is necessary to store the data more effectively and efficiently, and the analysis to be made on the data should work effectively in the background since a lot of data is kept for students, managers, and parents in education. In this process, data mining produces descriptive and predictive models for processing recorded data and generating information. The interest in this method is increasing, and new methods are emerging for more successful decisions due to the high benefit of the prediction method in the decision-making process (Akpinar, 2000; Uzut & Buyrukoglu, 2020; Yurtoglu, 2005).

The studies using data mining in traditional and distance education, the prediction of academic success for students, and the determination of the reasons that affect their failures have started to be seen frequently in the field of education (Ozbay, 2015). Data mining for education is an understanding of a discipline that develops methods to examine large-scale data and it uses these methods to understand students' performance and learning environments more comprehensively (Algami, 2016). Data acquisition from LMS, social network analysis, and visualization are used in this area. The process of transforming raw data collected from students, teachers, and parents into meaningful results by educators, researchers, and software in the education system is defined as educational data mining (García et al., 2011). The meaningful results obtained from educational data mining can be used as information that management will need to increase the competence and productivity of educational institutions. It is beneficial to identify problems in educational institutions and to create a more efficient and productive educational environment in terms of analyzing the data of students with the method of data mining in education, predicting students' success, and determining the reasons for their failure (Ozbay, 2015). Predictions can be created, and models for students' academic performance, and these predictive models can be used to advise students for their academic studies by using data mining methods in the field of education. In addition, new approaches to learning analytics can

be created, and students' profiles can be modeled through data mining methods in education. Individualized education environment, curriculum, and new learning styles can be created by classifying similar students (Bienkowski et al., 2012). Currently, the methods used in the field of education cause a lot of time and great effort. However, less time will be spent, which will lead business processes to be made automatically using the models developed with data mining (Lopez et al., 2012).

Many studies using different methods, techniques, and applications are encountered when the field is searched on the subject. Subbanarasimha et al. (2000) compared two different datasets by using artificial neural networks (ANN) and regression techniques to predict the academic performance of MBA students. As a result, a high success rate was obtained with the ANN model (Subbanarasimha et al., 2000). Guneri and Apaydin (2004) classified student achievement status using ANN and Logistic Regression (LR) methods. The classification success rate was found to be 95% in the methods used in the classification of success situations. Ibrahim and Rusli (2007) compared algorithms by predicting student success using ANN, Decision Trees (DT), and linear regression methods in data mining. It was found that ANN analysis gave better results in the prediction of overall academic achievement.

Bresfelean et al. (2008) used classification through learning, and data clustering methods to determine the academic success or failure of students in the study on the "Farthest First" algorithm and "Weka J48". They achieved quite successful results and stated that the methods used could be used effectively in education. Sembiring et al. (2011) developed a model using data mining methods to analyze student behavior and achievements and predict student performance. They revealed that the model they developed was quite effective. Sengur and Tekin (2013) estimated the graduation grades of students at the Department of Computer and Instructional Technologies in the Faculty of Education at Fırat University by using ANN and DT data mining methods. According to the results, it was observed that ANN provided better prediction performance compared to DT. Buyrukoglu and Yilmaz (2021) proposed a semi-automated data mining method for answering questions asked by students for online learning. This case-based reasoning model provided 84% time saving for instructors.

Akcapinar (2014) developed a model by using data mining methods on the data from an online learning environment for 76 students. He clustered the academic achievements of the course as “passed-fail” and categorized similar student profiles. Aydin (2015) predicted students' success in courses using variables such as course name, time passed through the e-service, the number of times students took the course, an average of exams taken, and students' age, using different classification algorithms. Kiray et al. (2015) analyzed the success of Turkish students in science and mathematics with the data mining method based on the data obtained from the international TIMSS and PISA exam results of Turkish students in their study. At the end of the study, they revealed the variables that affect science and mathematics achievement. Amrieh et al. (2016) developed a model to estimate the academic performance of the students with ANN, Naive Bayes (NB), and DT data mining methods using the data from 480 students through E-Learning. The reliability of the model was proven by providing 80% accuracy in the study.

Ozbay and Ersoy (2016) examined the relationship between the mobility of undergraduate students on the LMS and their academic achievement using data mining methods. It has been revealed that there is a significant relationship between mobility on LMS and academic achievement levels using log records containing the mobility of 40 students on LMS and year-end academic achievement scores. In another study, a model was developed by estimating 80% of when the students will complete the course based on the first day of the online course (Cunningham, 2017). Alsuwaiket (2018) proposed a model that predicts the academic performance of students in mathematics lessons that completed the 4th grade for students, teachers, and school principals. Altun et al. (2019) created models to estimate the academic

graduation average of the students, using the data such as gender, marital status, age of enrollment, and midterm exam scores in the 1st semester of the 1st grade. In the study, 94.30% success was obtained from the model made with regression analysis and 94.43% from the ANN model. Aydemir (2019) used 3794 students' data taking the Foreign Language-II course in the study on data mining at a university in Turkey. In the study, prediction models developed by ANN, M5P, Decision Stump, M5Rules, Decision Table, and Bagging methods were created and compared with each other. It was concluded that the model established with the bagging method produced estimates with the best result, 1.22 mean absolute error and 0.80 correlation coefficient. It was concluded that the students would learn passing grades of the course in advance and take precautions.

In summary, data mining methods can enable us to determine the field in which the students are successful, to obtain their level of success, and the factors that affect their success and cause failure. In this context, the aim of the study is to predict the academic success of students with data mining methods using the data obtained from the distance education system. For this purpose, all the details of the material and method were given in the second part of the study. The third part contains the experimental results and the discussion of these results. Finally, the conclusion and recommendations from the study were included in the fourth part.

2. MATERIAL and METHOD

In this section, all the details about the method and technique of the study were given. First of all, the data of the study were obtained from the official website of Kaggle (Aljarah, 2016), which is an open platform "https://www.kaggle.com/aljarah/xAPI-Edu-Data", and it includes fields such as gender, place of birth, nationality, education stages. Data from 304 students were used which have different country codes (from the student questionnaire in the CSV data file format) as a data source for the analysis. Data mining methods were used to extract meaningful information from the used dataset. E-Learning data and personal data obtained from LMS belonging to preschool, secondary school, and high school students were used in the study. These data produce 3 class outputs using data mining methods. These are Low Level (0-69 point range-L), Medium Level (70-89 point range-M), and High Level (90-100 point range-H). In this framework, the study was explained in detail with the process of the study, the dataset, data definition, the algorithms used, and the creation of the model.

2.1. The Process of the Study

The process of the study consisted of three stages. First, the data were organized by pre-processing and distinguishing from the noise before being analyzed. Later, the data were analyzed and visualized. Rapid Miner Studio and Orange applications, which are platforms for data science, were used to analyze data and determine validity. A comparison of the results obtained from these two platforms was made. The purpose of using this software was to analyze data, use various data mining techniques with pre-processing on data and evaluate these models by creating new models. Finally, the prediction results were obtained by training the data utilizing the seven algorithms in the study, and the success levels of these prediction results were listed in detail. The steps of data cleaning, data integration, data selection, data transformation, data mining, and presentation of results were carried out in the study, in general, to obtain meaningful information from the data.

2.2. Dataset

The dataset is an educational dataset collected from the LMS called Kalboard 360. Kalboard 360 is a multi-agent LMS designed to facilitate learning using the latest technology. This system provides users with simultaneous access to educational resources from any device with an internet connection. The data were collected using a student activity-tracking tool called the

experience API (xAPI). When the data set for the academic performance of the students is examined, it has a multivariate structure with 115 female students, 189 male students, 16 features, and classifications. It was determined that there were noisy data for the features used in the study when the data were analyzed, and these data were removed from the dataset. As a result of the extracted dataset, there were data belonging to 304 students in total, and this number represents 63.33% of the whole dataset. The dataset included students from different countries including Kuwait, Jordan, Palestine, Iraq, Lebanon, Tunisia, Saudi Arabia, Egypt, Syria, USA, Iran, and Libya. The properties of the dataset were given in Table 1 in detail (Aljarah, 2016).

Table 1. *Properties of the dataset(Aljarah, 2016).*

Nu	Attribute Name (Input)	Type	Value Range
1	Gender	Binominal	Male (1), Female (0)
2	Country	Polynomial	Kuwait, Jordan, Palestine, Iraq, Lebanon, Tunisia, Saudi Arabia, Egypt, Syria, USA, Iran, Libya
3	Place of birth	Polynomial	Kuwait, Jordan, Palestine, Iraq, Lebanon, Tunisia, Saudi Arabia, Egypt, Syria, USA, Iran, Libya
4	School Level	Polynomial	Primary School, Secondary School, High School
5	Grade Level	Polynomial	G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12
6	Section ID	Polynomial	A, B, C
7	Lesson Topic	Polynomial	English, Spanish, French, Arabic, IT, Mathematics, Chemistry, Biology, Science, History, Quran, Geology
8	Semester	Binominal	First, Second
9	Parent of Student	Binominal	Mother, Father
10	Number of Hand Raises	Number	0-100
11	Number of Resources Visited	Number	0-100
12	Number of Views Announcements	Number	0-100
13	Discussion Numbers	Number	0-100
14	Parent Responding Questionnaire	Number	0-100
15	Parent School Satisfaction	Binominal	Yes, No
16	Student Attendance Day	Binominal	Above 7, Below 7
17	Class	Polynomial	L (Low Level), M (Medium Level), H (High Level)

2.3. Data Cleaning

The following processes were carried out in the data cleaning: completing missing data, removing inconsistencies, detecting outliers, and removing noise. Data cleaning must be realized before data analysis. The following methods can be used to complete missing data (Oguzlar, 2003).

- Records with missing values can be deleted.
- Average value can be used instead of missing values.
- Median can be used instead of missing values.
- Instead of missing values, the average of the class in which it is located can be taken.
- The most appropriate value can be used instead of missing values by using methods such as regression.

2.4. Algorithms Used in the Study

In this section, each of the algorithms used in the study was briefly explained.

2.4.1. *Random forest*

The random forest model is a tree-based learning algorithm. The algorithm aims to combine tree decisions trained in different training sets and present them to the user instead of a single decision tree. While determining the attributes of each level, some calculations are made in all trees, and the attribute is determined. Then, the attributes in other trees are combined, and the most used attribute is selected. The selected attribute is included in the tree, and the process is repeated at all levels. In order to start the algorithm, the number of variables and the number of trees to be used in each node should be determined by the user (Breiman, 2001; Resende,2018).

2.4.2. *Naive bayes*

Naive Bayes is a kind of probabilistic classification mechanism based on Bayes' theorem. It aims to create a simple and effective statistical forecast showing a high level of success rate in application areas. The basic logic of the algorithm is based on the classification process according to the most appropriate label and the dependence of the attributes in the data only on a certain class (Bayes, 1763; Yildiz et al., 2007).

2.4.3. *Support vector machine*

Support vector machines are machine learning algorithms based on convex optimization aimed at minimizing risk. It can learn independently since it does not need distribution information. It aims to obtain the most suitable hyperplane to separate the classes in the support vector machines technique. That is, it predicts maximizing the distance between different classes (Ayhan & Erdogmus, 2014; Cortes & Vapnik, 1995).

2.4.4. *K-Nearest neighborhood*

It is a supervised learning algorithm that performs learning based on the data in the training set. It is used for both classification and regression problems. It performs the classification process by comparing the new data in the group with the data in the training set. Each sample in the training set is kept to represent a point in space. The k samples in training are set closest to the new sample, and the class of the new sample is determined when a new sample joins space (Fix, & Hodges, 1951; Han et al., 2011; Kilinc et al., 2016).

2.4.5. *Logistic regression*

It is used to match an item of data to a real-valued prediction variable. The main goal is to fit the data to a known type of function. In the regression technique, it is tried to find the function that best models the data given in the process. It is used to determine which function is the best by determining the difference between actual values and predictions (Tolles et al., 2003).

2.4.6. *Artificial neural networks*

ANNs are an information processing system used in solving nonlinear and complex problems, performing the learning function, which is the most basic feature of the human brain, and they can use experiments, to generate, create, discover, and predict new information without any help. ANN has a structure that builds a different calculation technique compared to the traditional calculation techniques, and it can adapt to its environment. It can also make decisions in uncertain situations. ANN is used effectively in many different areas such as forecasting, optical character handling, fingerprint recognition, pattern recognition, robot technology, job scheduling and quality control, power systems, system modeling, finance applications, image processing, industrial applications, and defense applications (Butuner & Yuksel, 2021; Calp, 2019; Calp & Kose, 2020).

2.4.7. *Deep learning*

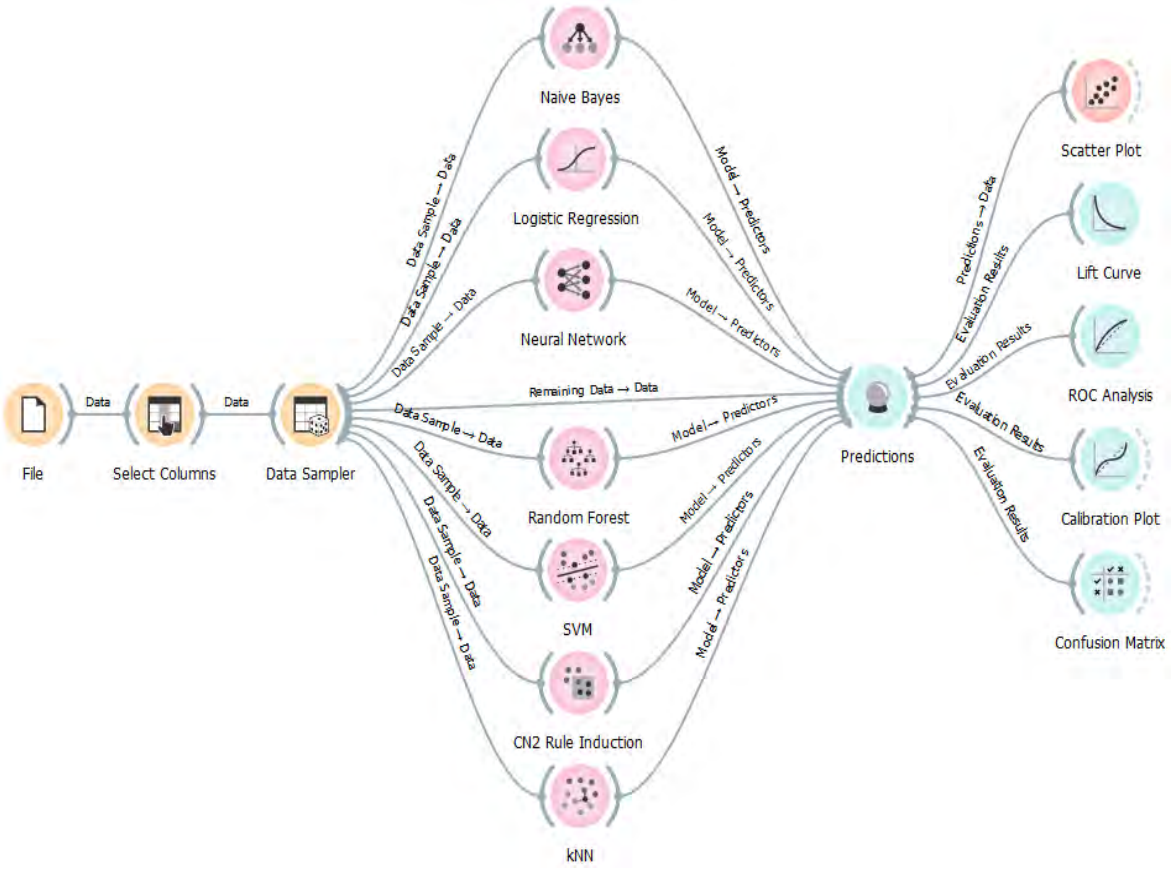
It is a technique that is used for machines to perceive and understand the world and provides solutions for nonlinear problems by using more than one layer. Generally, it is seen that it is

used more in areas such as data analysis, image classification, video analysis, speech recognition, information retrieval, object recognition, and natural language learning (Salman et al., 2020). In the deep learning technique, data is based on learning from the representation of data by learning more than one feature level. Deep learning methods are generally developed on ANNs, but they have more hidden neurons and layers. Deep learning methods have yielded very successful results in processing many types of data, such as video, audio, and text (Butuner, 2020; Calp, 2021).

2.5. Creating the Models

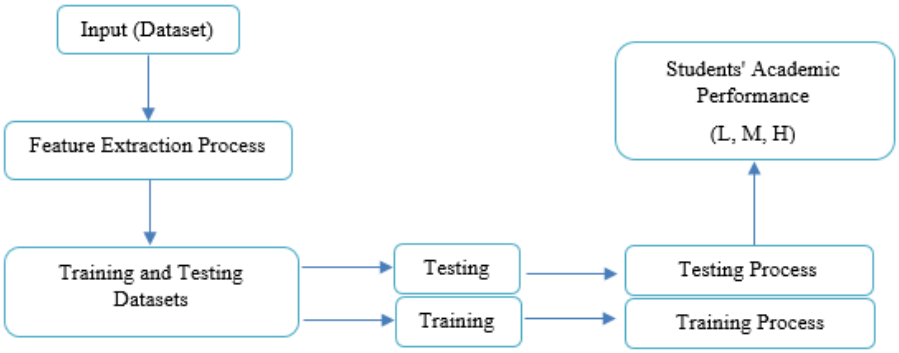
Rapid Miner and Orange programs were used to create the models. First of all, the methods to be used and their properties were determined. Then, the dataset containing the inputs and outputs of the study was loaded into the system. The models were created. Finally, the results and graphics of the created models were obtained (Figure 1).

Figure 1. *The process of creating models and obtaining results.*



There are two separate processes in the created models training and testing. First of all, the training of the model was carried out, then the testing process was applied. The performance of the models created was measured in the test part, and then the academic performance of the students was determined (Figure 2).

Figure 2. *Training, Testing, and Output Process.*



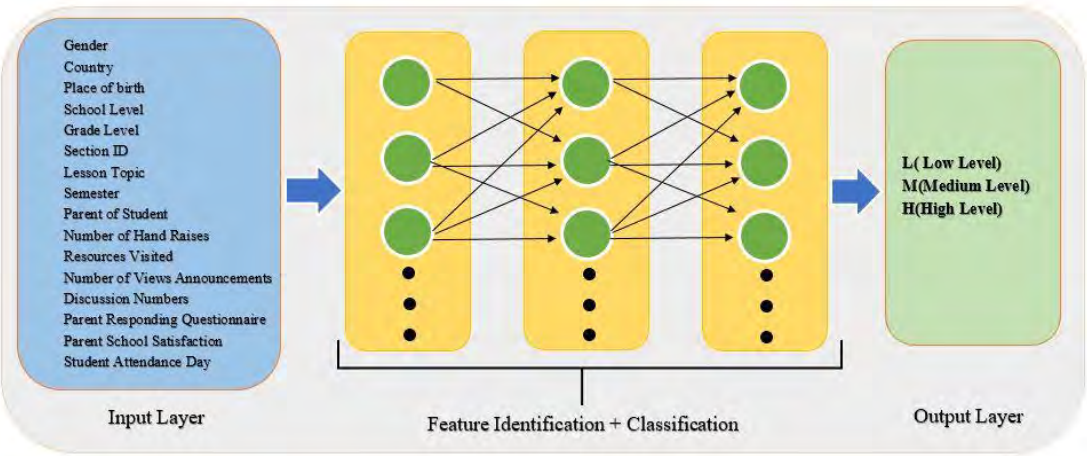
In the creation process of the models, the data were divided into two parts 75% to be used in the training of models, and 25% to be used in the testing process (Table 2). At this point, 228 of a total of 304 datasets belonging to preschool, secondary school, and high school students contained the training set, and 76 of them contained the testing set. These ratios and numbers may change to obtain the best model based on trial and error. The factor in determining these ratios and numbers used in the study was to obtain the training level of the model that gave the best result with these values.

Table 2. *The dataset used in the creation of the model.*

Purpose of usage	Number	%
Training	228	75
Testing	76	25

The mixed sampling "Shuffle Sampling" technique was used in the running of the models. RMSE (Root Mean Squared Error), CE (Classification Error), R2 (SC) Squared Correlation, and RE (Relative Error) performance measurement tools were used to evaluate the models. The general structure of the system used in all models was given in Figure 3. According to the general structure, 16 data received from the input layer were processed by using algorithms in feature determination and classification blocks, and meaningful results were produced for the output layer. These results were grouped as L (Low Level), M (Medium Level), and H (High Level).

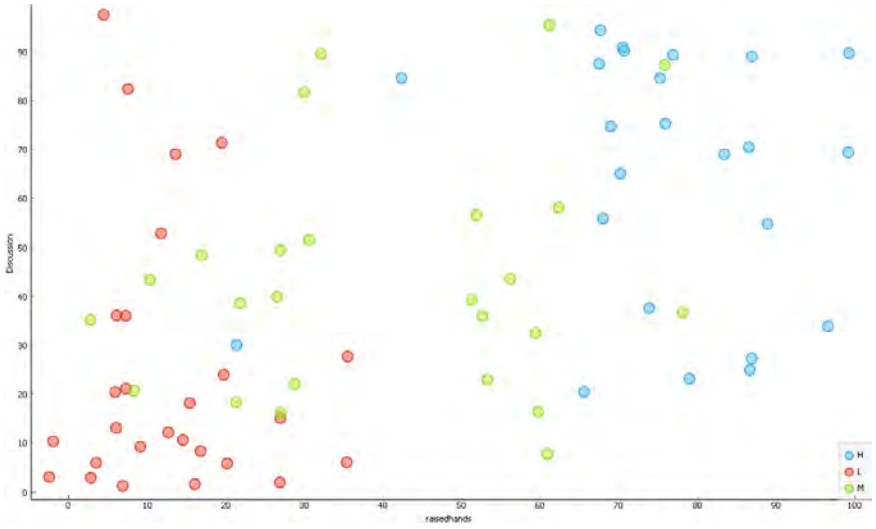
Figure 3. *The general structure of the system.*



3. EXPERIMENTAL RESULTS and DISCUSSION

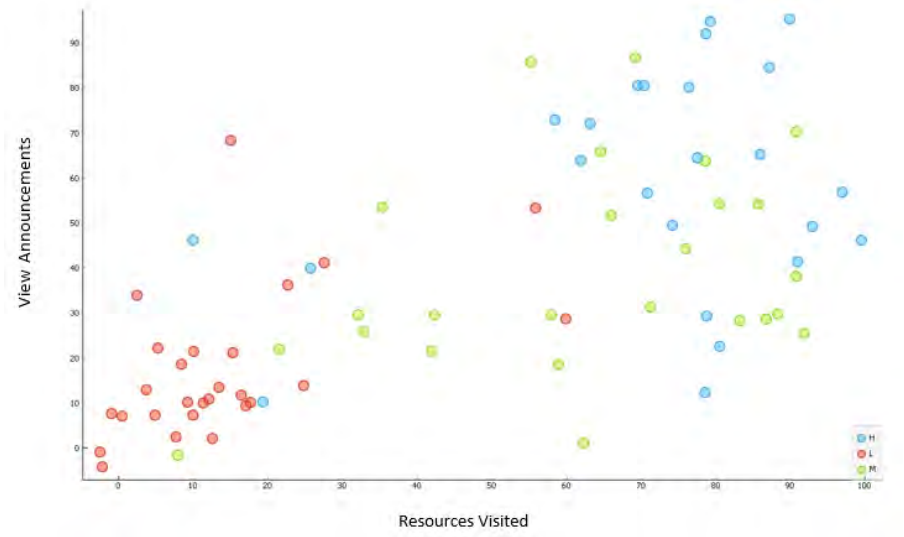
In this section, the experimental results obtained from the study were given, and these results were discussed in detail. In this context, data mining methods that consist of 16 column data entries belonging to students and produce meaningful results at three levels Low (L), Medium (M), and High (H) level were compared. In Figure 4, the correlation graph of the success levels in the Raisedhands and Discussion areas was shown. According to the density of the students in the Raisedhands and Discussion groups, it was seen that the success levels are High Level (H), but only the student density in the discussion groups was at Low Level (L). The students who gathered in the Raisedhands and Discussion groups in medium intensity were gathered at the Medium Level (M).

Figure 4. *The correlation graph between Raisedhands and Discussion.*



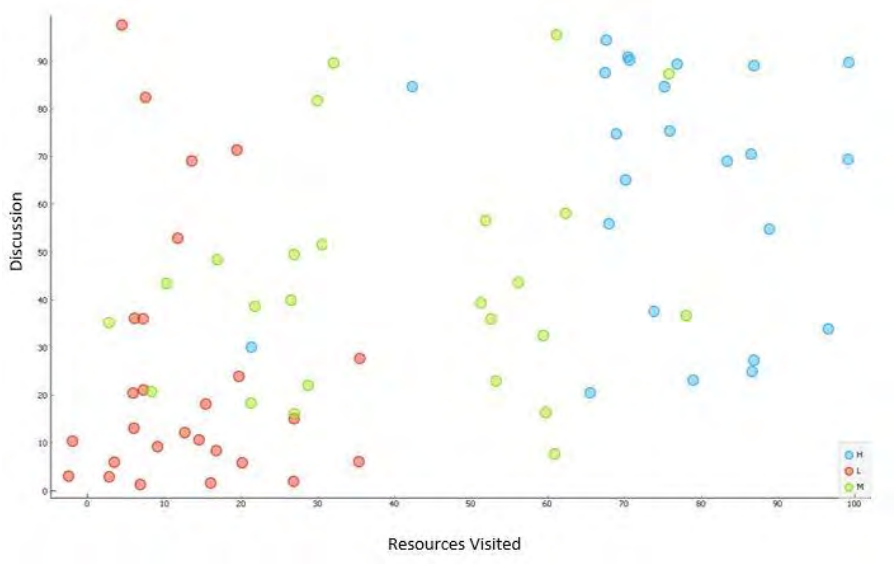
In Figure 5, the correlation graph between the Resources Visited and View Announcements by the students on the distance education platform was presented. According to the graph, student levels were also at (H) level and (M) level in direct proportion to the intensity of both elements (Resources Visited by students and View Announcements). It was observed that there was a student density at the (L) level in the region where the density of the resources visited by the students was low.

Figure 5. *The correlation graph between Resources Visited and View Announcements.*



In Figure 6, the correlation graph between the Resources Visited and Discussion areas by the students on the distance education platform was given. It was seen that the success levels were High Level (H) according to the density of the students in the discussion groups, but only the student density in the discussion groups were collected at the Low Level (L). It was observed that there was a student density at the (M) level in the region where the density of the resources visited by the students was medium in the Resources Visited and Discussion groups.

Figure 6. *The correlation graph between the Resources Visited and Discussion.*



The comparison of the academic achievement performances of the students according to the models used was given in Table 3. When Table 3 was examined, the success rates of over 96% were achieved with Deep Learning (DL), Support Vector Machines (SVM), LR, and Random Forest (RF) algorithms. In addition, the validation percentages of these algorithms obtained a 99% success rate.

Table 3. *Comparison of the models.*

Model	Validation	F1 Score	Accuracy	Recall
NB	0.965	0.896	0.904	0.895
LR	0.998	0.961	0.961	0,961
ANN	0.975	0.934	0.935	0.934
RF	0.994	0.960	0.961	0.961
SVM	0.996	0.974	0.960	0.974
DL	0.988	0.969	0.978	0.962
K-NN	0.929	0.839	0.850	0.842

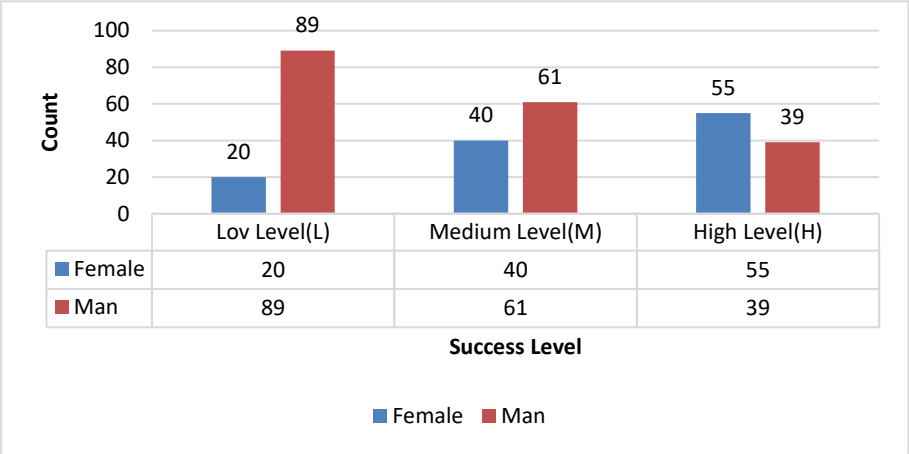
The information about the performance of the models prepared for the use of parameters according to the RF and DL algorithms was compared and given in Table 4.

Table 4. *Comparison of the results according to the models created.*

Model	Dataset	Testing	CE	RMSE	R^2 (SC)	RE
RF	304	0.961	0.049 +/- 0.084	0.222	5.33%	16.40% +/- 14.96%
DL		0.978	2.20%	0.141 +/- 0.000	0.969	6.53% +/- 12.46%

Figure 7 shows the success levels in the dataset according to gender distribution. It was observed that the number of male students in the (L) level exceeds four times the number of female students, and the number of female students in the (M) level and (H) level is 1.5 times the number of male students. As a result, it was seen that the success level of female students from online platforms was higher than male students in the distance education process.

Figure 7. *Success graph according to gender.*



In Figure 8, the success classification in the dataset according to school types was given. It was observed that the density of middle school students whose success classification was (H) level was high, and the density of high school students was lower. At the (L) level, it was understood that the density of both primary school and secondary school students was high and close to each other. At the (M) level, it was seen that the density of secondary school students was high, and the density of high school students was low.

Figure 8. *Student success intensity graph according to school types.*

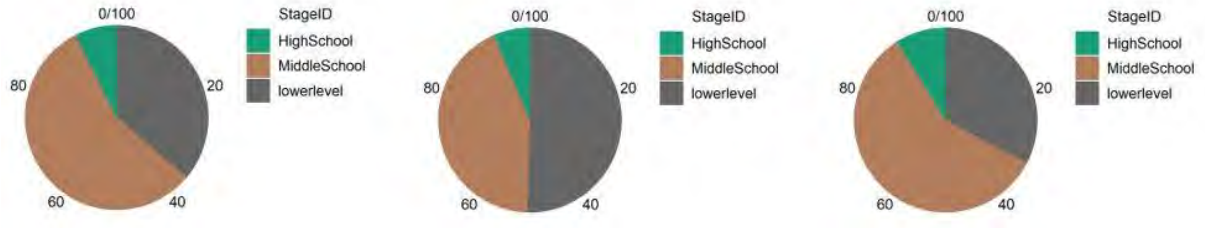


Figure 9 shows the success classification in the dataset according to absenteeism. Absenteeism was grouped as less than 7 days and more. It was seen that absenteeism of students with high levels of success was less than 7 days. On the other hand, it was seen that the students with low achievement levels have more than 7 days of absenteeism and the number of absenteeism of the students was concentrated in this number. In terms of success, the absenteeism of middle-level students under 7 days was approximately 1.5 times those above 7 days. In addition, it was understood that success classification according to the graph was inversely proportional to absenteeism and student numbers.

Figure 9. *Student success graph according to absenteeism.*

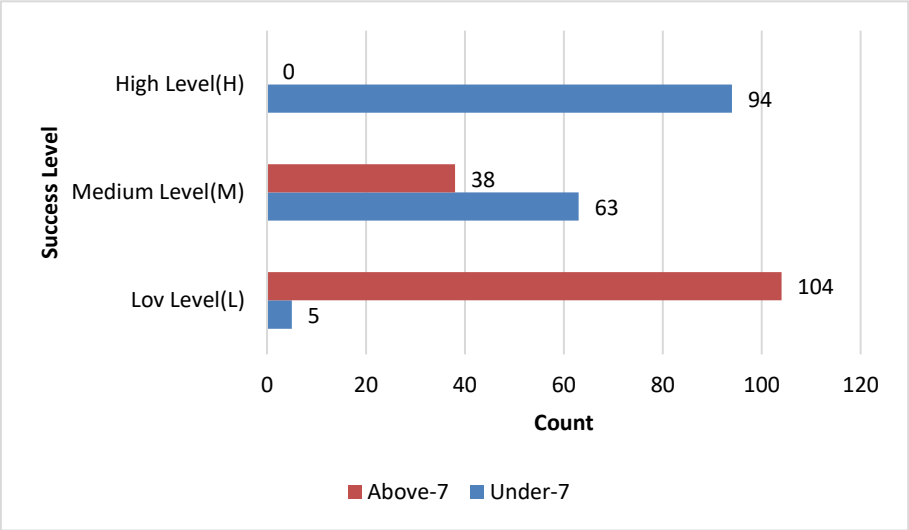
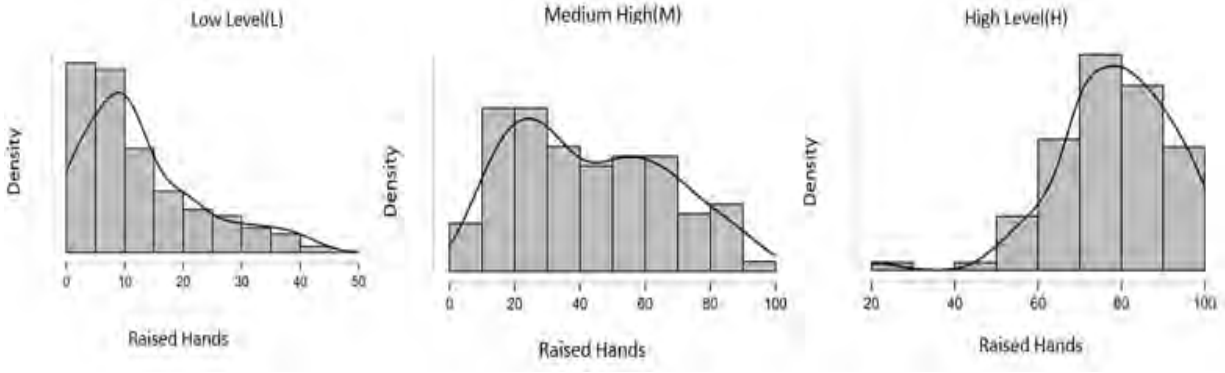


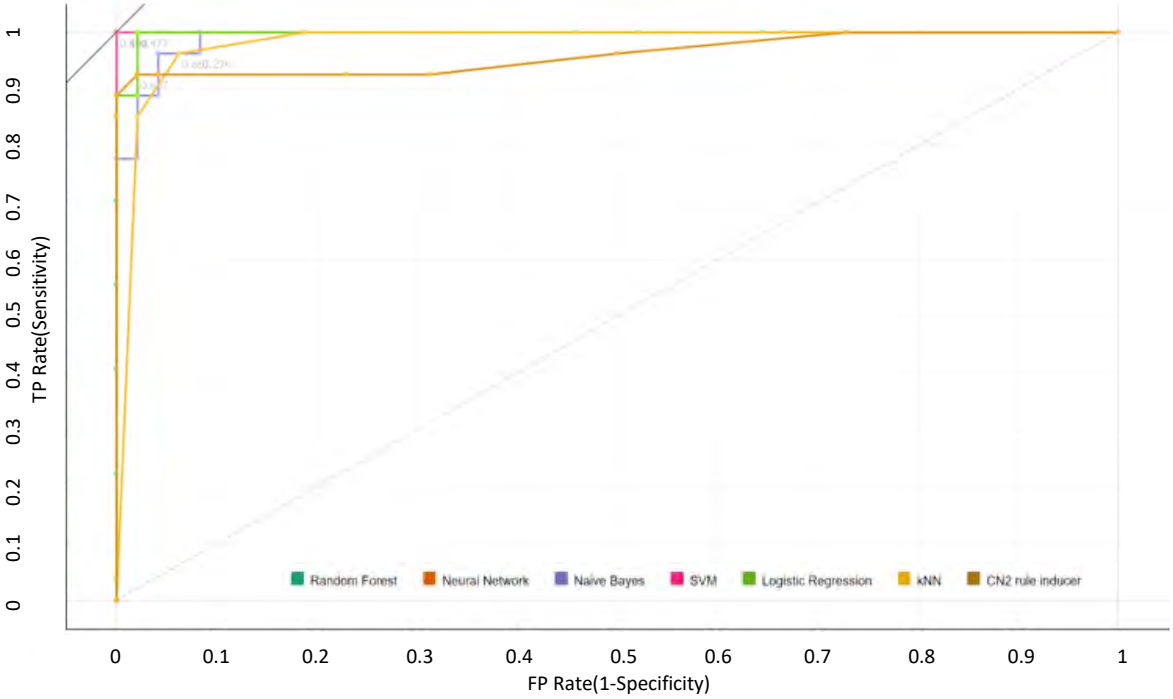
Figure 10 shows the graphs showing the success levels of students for asking questions by raising their fingers during the distance education process. As the number of Raisedhands of the students in the (L) level decreased in online education, the intensity of student achievement decreased. It revealed that while the Raisedhands intensity of the students was 20%, the (M) level had a peak value, and this value gradually decreased. It was understood that when the level of Raisedhands at (H) level made the peak value, the student success density was the highest.

Figure 10. *Relationship graph between Raised Hands and student achievement.*



In Figure 11, true positive and false positive ROC graphs of students with low achievement levels (L) according to the estimated percentages of the algorithms were given. According to the graph, the best accuracy rate was obtained from RF and DL algorithms.

Figure 11. *The accuracy graphs of the models (Low Level) (ROC Analysis).*



In Figure 12, the correct positive and false-positive ROC graphs of students whose achievement level was medium level (M) according to the estimation percentages of algorithms was given. According to the graph, the best accuracy rate was obtained from RF and DL algorithms.

Figure 12. *The accuracy graphs of the models (Medium Level) (ROC Analysis).*

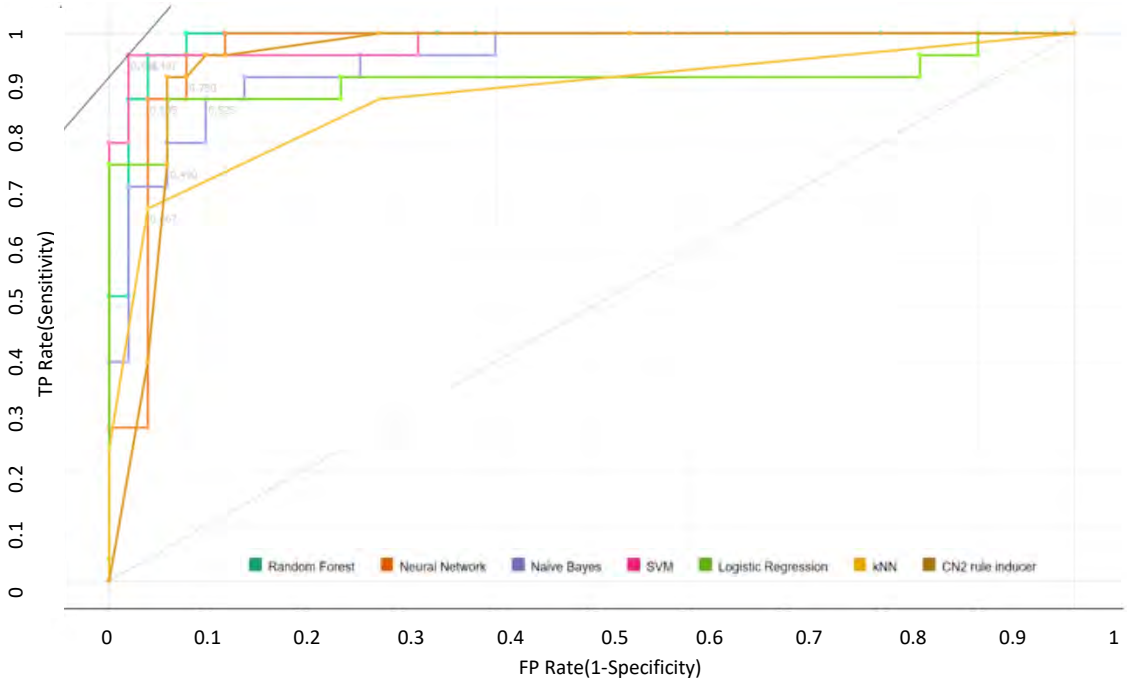


Figure 13 shows the correct positive and false-positive ROC graphs of the students with a high level of success (H) according to the estimated percentages of the algorithms. According to the graph, the best accuracy rate was again obtained from RF and DL algorithms.

Figure 13. *The accuracy graphs of the models (High Level) (ROC Analysis).*

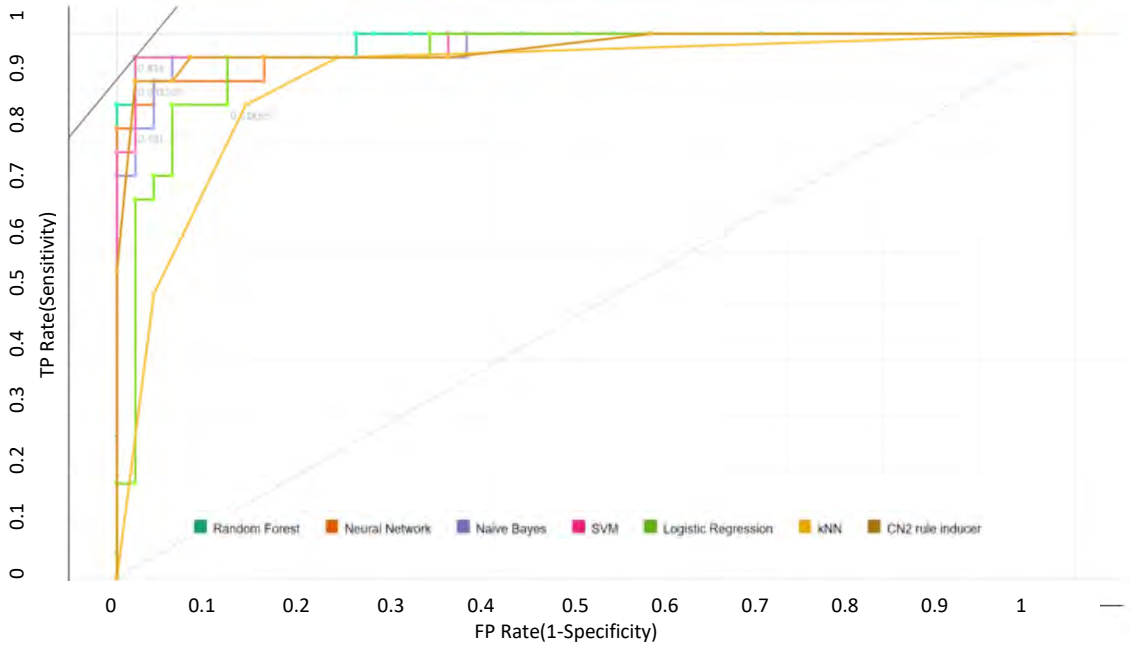


Table 5. *The confusion matrix results of the algorithms used in the study (respectively).*

Naive Bayes				
	H	L	M	∑
H	91.67%	0.00%	8.33%	24
L	7.41%	88.89%	3.70%	27
M	8.00%	0.00%	92.00%	25
∑	24	26	26	76

Artificial Neural Networks				
	H	L	M	∑
H	91.67%	4.17%	4.17%	24
L	3.70%	96.30%	0.00%	27
M	8.00%	0.00%	92.00%	25
∑	24	26	26	76

Logistic Regression				
	H	L	M	∑
H	95.83%	0.00%	4.17%	24
L	3.70%	96.30%	0.00%	27
M	4.00%	0.00%	96.00%	25
∑	24	26	26	76

Random Forest				
	H	L	M	∑
H	100.00%	0.00%	0.00%	24
L	3.70%	96.30%	0.00%	27
M	0.00%	8.00%	92.00%	25
∑	24	26	26	76

Support Vector Machines				
	H	L	M	∑
H	100.00%	0.00%	0.00%	24
L	3.70%	96.30%	0.00%	27
M	4.00%	4.00%	92.00%	25
∑	24	26	26	76

Deep Learning				
	H	L	M	∑
H	100.00%	0.00%	0.00%	24
L	7.41%	92.59%	0.00%	27
M	0.00%	0.00%	100.00%	25
∑	24	26	26	76

K-Nearest Neighbor				
	H	L	M	∑
H	87.50%	8.33%	4.17%	24
L	7.41%	81.48%	11.11%	27
M	8.00%	4.00%	88.00%	25
∑	24	26	26	76

The Confusion Matrix values for the algorithms used in the study were given in Table 5. It was seen that the Deep Learning model gave the best results according to (L) level, (M) level, and (H) level categories. In this DL model, the ratio of (H) level was 100%, (L) level was 92.59%, and (M) level was 100%. On the other hand, the KNN model was obtained with (L) level 81.48%, (M) level 88%, and (H) level 87.50%, with the lowest results in the study.

Finally, 15 (fifteen) different tests were carried out using real data with the proposed RF, ANN, KNN, NB, SVM, DL, and LR models. Experimental results obtained from these tests were given in Table 6.

Table 6. *Experimental results obtained from the models.*

Nu	RF	RF-Error	ANN	ANN-Error	KNN	KNN-Error	NB	NB- Error	SVM	SVM- Error	DL	DL- Error	LR	LR- Error	Output
1	0.86	0.14	0.86	0.14	0.75	0.25	0.92	0.08	0.86	0.14	0.90	0.10	0.88	0.12	M
2	0.94	0.09	0.99	0.01	0.58	0.42	0.93	0.07	0.98	0.02	0.99	0.01	0.94	0.06	H
3	0.94	0.06	0.94	0.06	0.83	0.17	0.90	0.10	0.97	0.03	0.99	0.01	0.93	0.07	H
4	0.95	0.05	0.99	0.01	0.75	0.25	0.94	0.06	0.95	0.05	0.97	0.03	0.97	0.03	M
5	0.95	0.06	0.94	0.06	0.50	0.50	0.98	0.02	0.96	0.04	0.98	0.02	0.95	0.05	M
6	0.97	0.05	0.93	0.07	0.83	0.17	0.92	0.08	0.98	0.02	0.97	0.03	0.97	0.03	M
7	0.97	0.03	0.93	0.07	0.92	0.08	0.91	0.09	0.94	0.06	0.98	0.02	0.99	0.01	H
8	0.96	0.08	0.94	0.06	0.42	0.58	0.85	0.15	0.97	0.03	0.93	0.07	0.96	0.04	H
9	0.97	0.03	0.97	0.03	0.92	0.08	0.92	0.08	0.99	0.01	0.98	0.02	0.94	0.06	L
10	0.96	0.06	0.90	0.10	0.92	0.08	0.91	0.09	0.97	0.03	0.96	0.04	0.97	0.03	L
11	0.97	0.03	0.90	0.10	0.92	0.08	0.95	0.05	0.96	0.04	0.99	0.01	0.96	0.04	M
12	0.99	0.06	0.99	0.01	0.92	0.08	0.89	0.11	0.98	0.02	0.97	0.03	0.96	0.04	L
13	0.96	0.06	0.89	0.11	0.83	0.17	0.76	0.24	0.99	0.01	0.99	0.01	0.98	0.02	L
14	0.98	0.02	0.95	0.05	0.75	0.25	0.95	0.05	0.98	0.02	0.99	0.01	0.97	0.03	L
15	0.98	0.02	0.95	0.05	0.58	0.42	0.88	0.12	0.97	0.03	0.98	0.02	0.98	0.02	M
Ort.	0.96	0.06	0.94	0.06	0.76	0.24	0.91	0.09	0.96	0.04	0.97	0.03	0.96	0.04	

To understand the success of the study more clearly, the obtained findings and test results were compared with other studies in the literature (Table 7). When Table 7 was examined, it was seen that the performance of the study was quite successful compared to other studies.

Table 7. *Comparison of the proposed study with the studies in the literature.*

Nu	Year	Authors	Number of Data (Student)	Method	Success Rate
1	2004	Guneri, N., & Apaydin, A. (2004).	352	LR	%95.17
				ANN	%97.14
2	2013	Sengur, D., & Tekin, A. (2013).	127	ANN	%93
				DT	%76
3	2013	Cokluk, O. T. D., & Cırak, G. Y. (2012).	419	LR	%66.10
				ANN	%70.16
4	2013	Turhan, K., Kurt, B., & Engin, Y. Z. (2013).	111	LR	%85
				ANN	%93
5	2016	Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016).	500	DT	%77.8
				NB	%72.4
				RF	%76.6
6	2016	Ozbay, O., & Ersoy, H. (2017).	40	CART	%85
				C5.0	%82.5
				CHAID	%65
				QUEST	%65
7	2019	Altun, M., Kayikci, K., & Irmak, S. (2019).	578	ANN	%94.43
				LR	%94.30
8	2019	Aydemir, E. (2019).	3794	M5P	%12
9	2018	Aydogan, I., & Zırhlıoglu, G. (2018).	1049	ANN	%97,2
10	2022	Butuner, R. & Calp, M. H. (2022) -The results of the models used in this study-	304	ANN	%97
				SVM	%99
				DL	%98
				NB	%96.5
				LR	%99.8
				RF	%99.4
				K-NN	%92.9

4. CONCLUSION and RECOMMENDATIONS

Education is the most important element for the future of society. It is inevitable that data mining methods, which have successful applications in many fields in the 21st century, will also be applied in the field of education and create new concrete outputs. This study proposed a model that predicts the academic achievements of primary, secondary, and high school students who receive education with the learning management system in the distance education process using data mining methods. 7 different models were created with data mining methods. Predictions were made in 3 classes Low (L), Medium (M), and High (H) according to the students' grades using the created models.

It can be said that predicting the academic performance of students is meaningful in increasing academic success when the models created in the study are evaluated. In addition, it was seen that results with high accuracy were obtained in predicting the academic success of students when the study was compared with similar studies using data mining techniques.

Finally, today, big and various data are stored in the systems of many institutions in Turkey such as e-government services, MEBSIS, and e-school which depend on the Ministry of National Education, OSYM, Universities, and Ministry of Health. However, it is seen that data mining or artificial intelligence techniques have not been used enough to draw meaningful

results from these data. It is obvious that the quality and production success of the institutions will increase by processing these data by each institution and obtaining meaningful and concrete outputs. As a result, it is expected that this study will set an example for other studies to increase student success.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Resul BUTUNER: Investigation, Resources, Visualization, Software, - original draf original draft. **M. Hanefi CALP**: Introduction, Methodology and Validation, Formal Analysis, and Writing -original draf.

Orcid

Resul BUTUNER https://orcid.org/0000-0002-9778-2349
M. Hanefi CALP https://orcid.org/0000-0001-7991-438X

REFERENCES

Akcapinar, G., Altun, A., & Aşkar, P. (2015). Modeling students' academic performance based on their interactions in an online learning environment. *Primary education Online*, *14*(3), 815-824.

Akpinar, H. (2000). Information discovery and data mining in databases. *Istanbul University Journal of the School of Business*, 1-22.

Aljarah, I. (2017). *Students' academic performance dataset*. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/aljarah/xAPI-Edu-Data

Alsuwaiket, M. (2018). *Measuring academic performance of students in higher education using data mining techniques* [Doctoral dissertation, Loughborough University].

Altun, M., Kayikci, K., & Irmak, S. (2019). Estimation of Graduation Grades of Primary Education Students by Using Regression Analysis and Artificial Neural Networks. *E-International Journal of Educational Research*, *10*(3), 29-43. https://doi.org/10.19160/ijer.624839

Amrieh, E.A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), 119-136. https://doi.org/10.14257/ijdta.2016.9.8.13

Aydemir, E. (2019). Forecasting of the Course Learning Notes by Data Mining Methods. *European Journal of Science and Technology*, 70-76. https://doi.org/10.31590/ejosat.518899

Aydin, S. (2015). Data Mining and an Application in Anadolu University Open Education System. *Journal of Research in Education and Teaching*, *4*(3), 36-44.

Aydogan, I., & Zirhlioglu, G. (2018). Estimation of Student Successes by Artificial Neural Networks. *YYU Journal of Education Faculty*, *15*(1), 577-610. http://dx.doi.org/10.23891/efdyyu.2018.80

Ayhan, S., & Erdogmus, S. (2014). Kernel Function Selection for the Solution of Classification Problems via Support Vector Machines. *Eskisehir Osmangazi University Journal of Economics and Administrative Sciences*, *9*(1), 175-201.

Bayes, T. (1763). LII. *An essay towards solving a problem in the doctrine of chances*. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. Philosophical transactions of the Royal Society of London, 370-418. https://doi.org/10.1098/rstl.1763.0053

Beitel, S. (2005). *Applying Artificial Intelligence Data Mining Tools to the Challenges of Program Evaluation*. Connecticut.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief*. Office of Educational Technology, US Department of Education.

Breiman, L. (2001). Random Forests, *Machine Learning, 45(1)*, 5–32.

Bresfelean, V.P., Bresfelean, M., Ghisoiu, N., & Comes, C.A. (2008, June 23-26). *Determining students' academic failure profile founded on data mining methods*. Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces, 317-322, https://doi.org/10.1109/ITI.2008.4588366

Butuner, R. (2020). *Sentiment Analysis with Deep Learning Methods and Its Use in School Guidance Services*, [Master's Thesis, Necmettin Erbakan University]. Coincil of Higher Education Libraries: https://tez.yok.gov.tr/UlusalTezMerkezi/TezGoster?key=fl0Kw4p1rmMDotyKRdYv1BKdBnLg10dCC3PJQ2laOIvx6m-b832uTqLlcfv5bVHP

Butuner, R., & Yuksel, H. (2021). Diagnosis and Severity of Depression Disease in Individuals with Artificial Neural Networks Method. *International Journal of Intelligent Systems and Applications in Engineering*, *9*(2), 55-63. https://doi.org/10.18201/ijisae.2021.234

Buyrukoglu, S., & Yilmaz, Y., (2021). A Novel Semi-Automated Chatbot Model: Providing Consistent Response of Students' Email in Higher Education based on Case-Based Reasoning and Latent Semantic Analysis, *International Journal of Multidisciplinary Studies and Innovative Technologies, 5*(1), 6-12.

Calp, M.H. (2019). An estimation of personnel food demand quantity for businesses by using artificial neural networks. *Journal of Polytechnic*, *22*(3), 675-686.

Calp, M.H. (2021). Use of Deep Learning Approaches in Cancer Diagnosis. In: Kose U., Alzubi J. (eds) *Deep Learning for Cancer Diagnosis. Studies in Computational Intelligence, vol 908.* Springer. https://doi.org/10.1007/978-981-15-6321-8_15

Calp, M.H., & Kose, U. (2020). Estimation of burned areas in forest fires using artificial neural networks. *Ingeniería Solidaria*, *16*(3), 1-22.

Cokluk, O.T.D., & Cirak, G.Y. (2013). The Usage of Artifical Neural Network and Logistic Regresssion Methods in the Classification of Student Achievement in Higher Education. *Mediterranean Journal of Humanities, 3*(2), 71-79. https://doi.org/10.13114/MJH/201322471

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20 (3),* 273-297. https://doi.org/10.1007/BF00994018. S2CID 206787478.

Cunningham, J. (2017). *Predicting student success in a self-paced mathematics MOOC* (Order No. 10272808). Available from *Pro Quest Dissertations & Theses Global*, (1900990574).

Dias, S.B., & Dinis, J.A. (2014). Towards an enhanced learning in higher education incorporating distinct learner's profiles. *Educational Technology & Society*, *17*(1), 307–319.

Dogan, A. (2012). *Yapay Zeka [Artificial intelligence]*. Kariyer Publishing.

Dunham, M.H. (2003). *Data mining introductory and advanced topics*. Prentice Hall.

Fix, E., & Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. USAF School of Aviation Medicine, Randolph Field, Texas.

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.

Guneri, N., & Apaydin, A. (2004). Logistic Regression Analysis and Neural Networks Approach in the Classification of Students Achievement. *Gazi University Journal of Commerce & Tourism Education Faculty*, *1*, 170-188.

Han, J., Pei J. & Kamber, M., (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Ibrahim, Z., & Rusli, D. (September, 2007). *Predicting Students' Academic Performance: Comparıng Artificial Neural Network*, *Decision Tree and Linear Regression*. 21st Annual SAS Malaysia Forum, (s. 5). Shangri-La Hotel, Kuala Lumpur.

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

Jung, S., & Huh, J.H. (2019). An Efficient LMS Platform and Its Test Bed. *Electronics*, *8*(2), 154.

Karakaya, F., Arik, S., Cimen, O., & Yilmaz, M. (2020). Investigation of the views of biology teachers on distance education during the COVID-19 pandemic. *Journal of Education in Science, Environment and Health (JESEH), 6*(4), 246-258. https://doi.org/10.21891/jeseh.792984

Kilinc, D., Borandag, E., Yucalar, F., Tunali, V., Simsek, M., & Ozcift, A. (2016), Classification of Scientific Articles Using Text Mining with KNN Algorithm and R Language. *Marmara Journal of Pure and Applied Sciences, 3*, 89-94. https://doi.org/10.7240/mufbed.69674.

Kiray, S.A., Gok, B., & Bozkir, A.S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *Journal of Education in Science, Environment and Health (JESEH), 1*(1), 28-48.

Kurt, C., & Erdem, O. (2012). Discovering the Factors Effect Student Success via Data Mining Techniques. *Journal of Polytechnic, 15*(2), 111-116.

Lopez, M.I., Luna, J.M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. International Educational Data Mining Society.

Luan, J. (2002). Data Mining and Knowledge Management in Higher Education-Potential Applications. *42nd Associate of Institutional Research International Conference (s. 1-20). Toronto*, Canada: ERIC.

Oakes, K. (2002). E-learning: LCMS, LMS—They're not just acronyms but powerful systems for learning. Training & Development, *56*(3), 73–75.

Oguzlar, A. (2003). Data Preprocessing. *Erciyes University Journal of Faculty of Economics and Administrative Sciences*, *21*, 67-76.

Ozbay, O. (2015). Data Mining Concept and Data Mining Applications in Education. *The Journal of International Education Science, 5,* 262-272.

Ozbay, O., & Ersoy, H. (2017). Analysis of Student Dynamism into Learning Management.

Resende, P.A.A., & Drummond, A.C. (2018). A survey of random forest based methods for intrusion detection systems. ACM Computing Surveys (CSUR), *51*(3), 1-36.

Salman, F.M., Abu-Naser, S.S., Alajrami, E., Abu-Nasser, B.S., & Ashqar, B.A. (2020). COVID-19 Detection using Artificial Intelligence, *International Journal of Academic Engineering Research*, 18-25.

Savas, S., (2021). Artificial Intelligence and Innovative Applications in Education: The Case of Turkey, *Journal of Information Systems and Management Research, 3*(1), 14-26.

Savas, S., Guler, O., Kaya, K., Coban, G., & Guzel, M.S., (2021). Digital Games in Education and Learning through Games, *International Journal of Active Learning, 6*(2), 117-140.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). Prediction of Student Academic Performance by an Application of Data Mining Techniques. *International Conference on Management and Artificial Intelligence, 6*(1), 110-114). IACSIT Press.

Sengur, D., & Tekin, A. (2013). Prediction of Student's Grade Point Average by Using the Data Mining Methods. *Journal of Information Technologies, 6*(3), 7-16.

Siemens, G., & Baker, R. (2012). Prediction of student academic performance by an application of k-means clustering algorithm. *Towards Communication and Collaboration. 2nd international conference on learning analytics and knowledge*. Vancouver, Canada.

Subbanarasimha, P., Arinzeb, B., & Anandarajanb, M. (2000). The Predictive Accuracy of Artificial Neural Networks and Multiple Regression in the Case of Skewed Data. Exploration of Some Issues. *Expert Systems with Applications*, 117-123.

System through Data Mining Methods. *Journal of Gazi University, Faculty of Education, 37*(2), 523-558.

Tolles, J., & Meurer, W.J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama, 316*(5), 533-534.

Turgut, M., & Baykul, Y. (2013). *Eğitimde Ölçme ve Değerlendirme [Measurement and Evaluation in Education]*. Pegem Yayıncılık.

Turhan, K., Kurt, B., & Engin, Y.Z. (2013). Estimation of Student Success with Artificial Neural Networks. Education and Science, *38(170)*, 112-120.

Uzut, O.G., & Buyrukoglu, S., (2020). Prediction of real estate prices with data mining algorithms, *Euroasia Journal of Mathematics, Engineering, Natural and Medical Sciences*, *8*(9), 77-84.

Yamamoto, G. T., & Altun, D. (2020). The Coronavirus and the Rising of Online Education. *Journal of University Research*, 25-34. https://doi.org/10.32329/uad.711110

Yildiz, H.K., Genctav, M., Usta, N., Diri, B., & Amasyalı, M.F. (2007). *A New Feature Extraction Method for Text Classification*. 2007 IEEE 15th Signal Processing and Communications Applications.

Yurtoglu, H. (2005). *Yapay Sinir Ağları Modellemesi ile Öngörü Modellemesi: Bazı Makroekonomik Değişkenler için Türkiye Örneği [Predictive Modeling with Artificial Neural Network Modeling: The Case of Turkey for Some Macroeconomic Variables]*. [Expertise Thesis, DPT]. https://www.sbb.gov.tr/wp-content/uploads/2018/11/HasanYurtoglu.pdf