Journal of Educational Technology & Online Learning
Volume 7 | Issue 2 | 2024
http://dergipark.org.tr/jetol

Data mining approach for prediction of academic success in open and distance education

Selma Tosun [a] *, Dilara Bakan Kalaycıoğlu [a]

[a] Gazi University, Türkiye; 0000-0001-7444-7903

Suggested citation: Tosun, S. & Bakan Kalaycıoğlu, D. (2024). Data Mining Approach for Prediction of Academic Success in Open and Distance Education. *Journal of Educational Technology & Online Learning*, 7(2), 168-176.

Highlights

- This paper focuses on the classification of academic success of students in open and distance education faculty by educational data mining.
- Classification algorithms demonstrate notable effectiveness in categorizing students as either successful or unsuccessful, achieving model accuracy rates that range from 0.78 to 0.92.
- The C&RT algorithm performs best when the accuracy and specificity criteria are evaluated together in determining the data mining model.
- Grades in the compulsory courses taught in the first semester in higher education programs are among the most critical criteria for classifying ultimate academic success.

Article Info: Research Article

Keywords: *Data mining in education, academic success, open and distance education, C&RT, RF*

Abstract

Predicting and improving the academic achievement of university students is a multifactorial problem. Considering the low success rates and high dropout rates, particularly in open education programs characterized by mass enrollment, academic success is an important research area with its causes and consequences.

This study aimed to solve a classification problem (successful or unsuccessful), predict students' academic success, and identify those at risk. The primary objective was to predict the academic success status with 26,708 students enrolled in Istanbul University open and distance education programs between 2011 and 2017. Predictions were based demographic data and success grades in Turkish, Atatürk's Principles and History of Revolution, English, and Disaster Culture courses. The study utilized classification models from supervised learning algorithms and was conducted using the SPSS Modeler 18 program. Initially, the data was divided into 70% training and 30% test data. Then, models were constructed by using Random Forest, Tree-AS, C&RT, C5.0, CHAID, QUEST, Naive Bayes, Logistic Regression, NeuralNet, and SVM algorithms. Model performances were compared according to accuracy, sensitivity, specificity, F1 score, positive predictive value, negative predictive value, and Matthews Correlation Coefficient criteria. The C&RT model demonstrated the best performance, achieving the highest specificity value of 0.915.

1. Introduction

Open and distance education is currently undergoing a remarkable expansion. Each year, there is a consistent increase in the number of students enrolling open and distance higher education programs. A staggering 23 million higher education students are engaged in open and distance education courses offered by institutions in the twelve nations comprising Australia, Brazil, Canada, China, Germany, India, Russia, South Africa, South Korea, Turkiye, the United Kingdom, and the United States, collectively representing 51% of the world's population. Notably, emerging economies like Brazil, China, and Turkiye have

*Corresponding author: Assessment and Evaluation in Education, Gazi University, Ankara.
e-mail address: selma.tosun@istanbul.edu.tr
This study was partly presented as a proceeding at the 3rd International Conference on Educational Technology and Online Learning held between 20-23 June 2023.

doi: http://doi.org/10.31681/jetol.1334687
Received 30 Jul 2023; Revised 15 Mar 2024; Accepted 19 Mar 2024
ISSN: 2618-6586.

witnessed substantial increase in open and distance education participation (Zawacki-Richter & Qayyum, 2019). According to the statistics of higher education in Turkiye in 2022, a total of 2,835,686 students in open and distance education programmes distributed as follows: 1,560,050 in associate degree programs and 1,275,636 in undergraduate programmes (YÖKSİS, 2023).

The number of students who enrolled in open and distance learning at universities is very high, but these students often stand out for having lower success rates compared to their peers in traditional education, and they tend to have a higher tendency to drop out (Bağrıacık Yılmaz & Karataş, 2022; Radovan, 2019).

Open and distance learning accommodates diverse groups of learners with different backgrounds and learning needs. Many factors effective in predicting students' academic achievement have been investigated in the literature (Okur et al., 2019). Students' previous academic achievement and student demographics were presented as an important variable in 69% of the research articles (Alyahyan & Düştegör, 2020). For this reason, student demographics and external assessments, which include the grades of other courses, are the most frequently used student attributes. While student demographics include variables such as gender, age, family background, etc., external assessment includes the grade obtained in the final exam of a particular course or the success grade.

The migration of student data to digital platforms has led to the development of specialized methods to prevent academic failure (de Oliveira, et al., 2021). During this phase, data mining techniques, which are increasingly prominent in contemporary educational research, play a pivotal role (Bilici & Özdemir, 2021; Bonde & Kirange, 2018; Tekin & Öztekin, 2018). Educational data mining seeks correlations and rules that will enable us to make predictions about the present and future from the large amount of data obtained from different sources regarding the educational process using a computer program (Hotaman, 2020). In other words, it transforms raw data from educational institutions into usable patterns (Tan et al., 2018). Educational data mining, which uses many different disciplines such as psychometrics, learning analytics, and statistics together (Türel & Baz, 2016), is used to make determinations about student success, to make inferences about the problems in the educational environment and their causes, and to create correct and need-meeting educational environments (Özbay, 2015). Using all the students' data to determine the student's academic achievement reliably will positively affect the success of the process evaluation approach.

Data mining methods, widely employed across various fields in higher education, offer valuable capabilities such as predicting student performance, analyzing student preferences, pinpointing weaknesses in educational programs, and optimizing resource allocation (Bhise et al., 2013). Consequently, educators must be adept at identifying students who might be at risk of underperforming, taking preventive measures, and equipping themselves to handle such scenarios effectively (Kotsiantis et al., 2004). Modern educational institutions leverage these techniques, such as data mining, to provide decision-makers with accurate information, enabling them to make informed choices crucial for developing and enhancing the educational system, as emphasized in research by Veeramuthu and Periasamy (2014). An also in higher education, the initial semester or year is particularly critical, as failures during this period can serve as early indicators of academic risk (Apaydın et al., 2020).

Thus, the ability to predict student performance in advance and intervene with those who may be at risk of failure early on can play a critical role in improving the educational experience for students. Effectively predicting student success and failure from the outset of their educational journey allows for a better understanding of their specific needs and the implementation of proactive measures to support them. This, in turn, can lead to improved academic outcomes for students and a subsequent decrease in dropout rates (Albreiki et al., 2021).

2. Literature

Researchers have conducted numerous studies to achieve a more comprehensive understanding of student success and academic performance. These studies have focused on factors influencing student achievement at the university level (Taşdemir, 2012), comparing academic success in open education (Tosun, 2016),

predicting students at risk within higher education (Apaydın et al., 2020), and identifying the key factors impacting academic performance alongside demographic variables, including the identification of the most influential variables for predicting mathematics performance (Bulut et al., 2022).

When the literature is examined, it is seen that the academic achievement and dropout rates of open and distance education students are the subject of many studies (Batool et al., 2023; Dabhade et al., 2021; Durairaj and Vijitha, 2014; Khasanah and Harwati, 2017; Kotsiantis et al., 2004;2005; Nahar et al., 2021; Sembiring et al., 2011). In addition, there are many studies that use data mining methods to predict or determine academic success and performance (Alan & Temiz, 2019; Dabhade et al., 2021; Elakia et al., 2014; Natek & Zwilling, 2014; Osmanbegović and Suljic, 2008; Ramesh et al., 2013; Saheed et al., 2018; Shahiri et al., 2015).

Kotsiantis et al. (2005) tried to predict student performance with regression models based on demographic data and grades obtained from written assignments of some courses in their research conducted in Hellenic Open University. Finally, a prototype version of software support tool for tutors has been constructed implementing algorithm, which proved to be the most appropriate among the tested algorithms. Another study conducted at the Hellenic Open University proposed a method for predicting university distance education students' course success and performance using a supervised machine learning algorithm with essential demographic characteristics and grades on several written assignments to predict underperforming students. According to the research findings, the Naive Bayes algorithm was the most appropriate and easiest to implement, with highly satisfactory accuracy and overall precision (Kotsiantis et al., 2004).

Khasanah & Harwati (2017), in their research at the Islamic University of Indonesia, showed that among the factors affecting student performance, attendance and grade point average in the first semester are among the variables that will provide the best prediction and Bayesian Network has a higher accuracy rate than decision trees. Saheed et al. (2018), who investigated various machine learning algorithms for predicting and classifying student performance, developed models to predict student performance using ID3, C4.5, and CART; C4.5 performed better than other classifiers. In addition, educational factors, parental factors and sociodemographic factors (age, gender, religion, marriage status, etc.) in data mining analyses were found to be effective in students' academic performance. Bulut et al. (2022) built prediction models using Random Forest and LogitBoost algorithms to identify at-risk students for low mathematics performance. Nahar et al. (2021) compared six classification algorithms (random forest, DT(J48), Naive Bayes, PART bagging, boosting) in their study on improving academic achievement and created two final models based on decision tree and Naive Bayes algorithms for two of the data sets. According to the data set collected through a survey, firstly, student performance is classified as good-medium-bad depending on a course, while in the second data set, final grades of a course are classified (A-B-C) and predicted. As a result, the proposed dercision tree and naive bayes models are compared not only in terms of accuracy but also in terms of many other performances (economy, complexity).

In addition, Issah et al. (2023) conducted a systematic literature review on machine learning applications to determine the features affecting academic performance. Of the 114 articles analysed within the scope of the research, 34.20% of the 114 articles were related to academic performance (GPA, grade level, high school score, attendance, number of courses per semester); 22.80% on demographic variables (gender, nationality, place of birth, age), 17.50% on behavioural characteristics (hands raised, resources visited, school satisfaction, discussion, class participation, answering questions), 8.80% on psychological characteristics (personality, motivation, learning strategies, learning approach, contextual influence), 8.80% family background (mother and father education, family income, parents' position) and 7.90% school environment characteristics (school size, educational environment, lecturer/teacher behaviour in the classroom) were used to predict performance.

Sembiring et al. (2011) developed models to predict student performance by analyzing student behaviors and achievements with data mining; He and Zhang (2011) introduced a decision support system based on data mining to support the complex decision-making process of universities by tracking students and making a comprehensive performance evaluation their study. Durairaj and Vijitha (2014) aimed to predict student performance from grade point averages and to develop a trust model using data mining techniques

that extract the necessary information for current educational management. There is a study comparing supervised (Naive Bayes and Logistic Regression) and unsupervised (K-Means and Hierarchical Clustering) learning algorithms to evaluate the academic performance of students at Panjab University (Rana & Garg, 2016), Yossy et al. (2019) investigated the classification performance of university students' achievement scores in mathematics; K-nearest neighbor 86.52%, classification and regression algorithms tree 86.08%, Naive Bayes 84.78%, AdaBoost 88.04%, extratree 81.30%, Bernoulli Naive Bayes 79.34%, random forest E of 87.82%, random forest G 89.78%; it was shown that the most appropriate classification method was random forest G with 89.78%. In their study, Orrego Granados et al. (2022) reveal the importance of machine learning and data mining approach in developing and targeting policies to support students with low academic performance or to encourage advanced students. In the study, recommendations were developed on the impact of counselling for students to develop their careers after graduation and revealed that the XGBoost Machine Learning model showed the best performance for obtaining information about students' academic performance.

3. Aim and Research Questions

In Turkiye, all undergraduate and associate degree programs in higher education's initial semesters include compulsory courses mandated by the Higher Education Law No. 2547, which cover subjects like the Turkish Course, Atatürk's Principles and History of the Turkish Revolution (Atatürk's Principles HTR) Course and English Course. Additionally, each university has the authority to establish additional compulsory courses in their programs, such as the Disaster Culture Course. Since 2013, these compulsory courses and the broader adoption of digital education technologies have been conducted through distance learning methods in Turkiye. No previous research has investigated the role of students' performance in these introductory courses at the end of the first semester in the classification of their grades at graduation.

Within the scope of this study, it is aimed to discover the factors affecting the academic success of students taking compulsory courses between 2011-2017 by using data mining methods. Based on the finding that academic measurements will be an influential variable in producing the most helpful information for the prediction of academic success (Bulut et al., 2022), the academic success status of students in different programs was tried to be predicted with the course success grades of their compulsory courses, the type of program they enrolled in and their demographic features. This approach is considered a valuable method for developing policies to support students with low academic performance or to encourage advanced students, especially from the first semester.

In this study, the effect of variables under the headings of prior academic achievement (university-data: WGPA (weighted grade point average), individual course letter marks and individual assessment grades), student demographics (gender, age, place of residence, marital status) and students' environment (class type, semester duration, type of programme etc.) were investigated for the prediction of students' academic achievement (successful or unsuccessful).

Considering all these, this study will seek answers to two main questions:

1) Which data mining algorithm shows the highest performance for the classification of students' academic achievement (success-unsuccess) in open and distance education?

2) What are the effective factors in classifying the academic achievement of open and distance education students?

4. Methodology

4.1. Research Model

In the research, a model will be developed with data mining algorithms to predict academic success and to identify students at risk of failure early by taking the demographic data of open and distance education students in higher education and their achievement scores in compulsory courses as variables.

This research is a correlational survey design, which is one of the quantitative research methods. Correlational surveys investigate the relationship between one or more quantitative variables and one or more variables (Fraenkel et al., 2012).

Data mining methods were used to extract meaningful information from the data set used in the study. Data mining can be defined as the determination of rules that will enable prediction for the future in the light of available data (Şengür & Tekin, 2013).

4.2. Research Procedures

After obtaining the necessary legal permissions and database access rights, an appropriate data analysis method was developed to comprehensively explore and analyse the data set.

The subsequent steps were followed: organising the data into multiple data files, converting the data into meaningful wholes, examining missing and extreme values and determining the missing data strategy, merging student demographic data with the related course success notes data, cleaning unnecessary data columns and repetitive data, and then recoding the data. In the recoding stage, continuous variables are transformed into categorical variables.

After completing these data preparation stages, classification algorithms were selected based on the nature of the prediction variable, which was nominal in this case. Random Forest (RF), Tree-AS (Tr-AS), Classification and Regression Trees (C&RT), C5.0, CHAID, QUEST, Naive Bayes (NB), Logistic Regression (LR), Artificial Neural Network (NeuralNet) and Support Vector Machine (SVM) algorithms were used in the research.

Figure 1 represents the overall research process and the specific data mining classification algorithms.

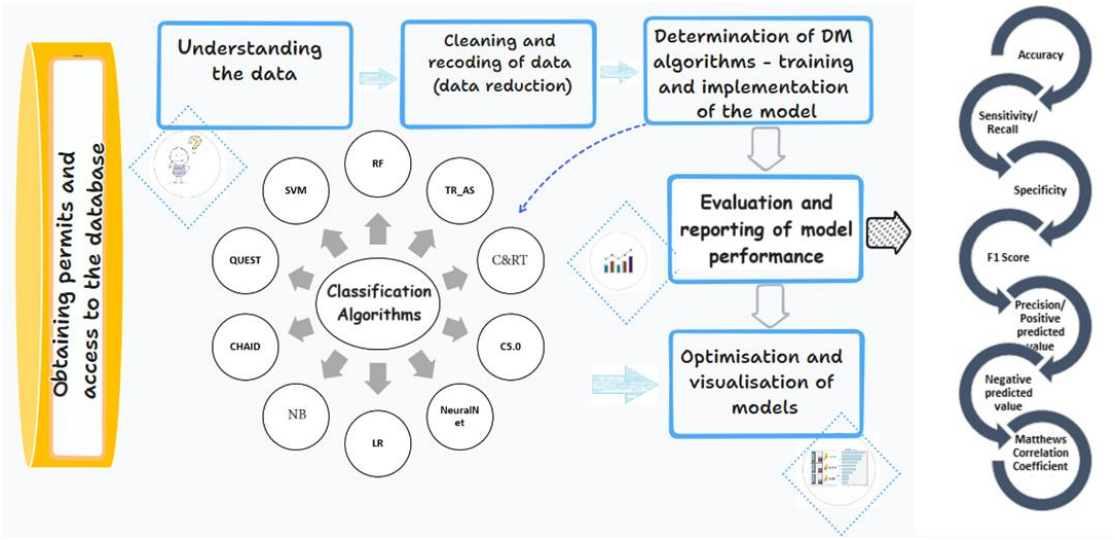

Figure 1: Research Process and Data Mining Algorithms

4.3. Study Group

The study utilized a dataset comprising demographic information from 26,708 students enrolled in Istanbul University's Open and Distance Education Faculty programs from 2011 to 2017. This dataset that was extracted from the university's academic database also included the the academic performance grades of compulsory courses taken for the first time and demographic data of students enrolled in the first semester.

Among the students in the study, 55% are female, while 45% are male. Additionally, 44% reside in Istanbul, and 79% have been placed in education programs through with an examination. The remaining 21% have previously participated in a higher education program or are continuing their education (Second University without Exams). Furthermore, it is evident that the students predominantly fall within the 26-30 age range. Among the group, 2000 students are over 40 years old, according to the Weighted Grade Point Average

(WGPA), which indicates the graduation status of all students, 31% were classified as successful as their score was 2 and above, while the others were classified as unsuccessful (69%) as their score was below 2.

The variables used in the models and their possible values are listed in Table 1.

Table 1: Definition and Possible Values of Variables

Variable/Input	Definition Variable	Possible values
Gender	Male-female	0-1
Province of residence	Istanbul and others	0-1
Program type/ name	The higher education program in the relevant faculty	43 different programs; values between 1-43
Program level	Undergraduate and associate degree levels	1-2
Program teaching type	Open education and distance education	1-2
Form of enrolment	Variable indicating the admission conditions of students to programs	University Entrance Examination (UEE), Vertical Transfer Exam (VTE), Amnesty student (AS), Examination For Foreign Students (EFS), Second University without Exams, Exceptional Student Status (ESS-OZEL)
Student enrolment year	From 2011 to 2017	2011, 2012, 2013, 2014, 2015, 2016, 2017
Marital status	Marital status data of the student at the enrolment stage	Married, Single
Age	Age ranges of students	1= 25 years old and below; 2= 25 to 30 years old; 3=31 to 35 years old; 4=35 to 40 years old; 5= 4 years old and above
Course (Atatürk's Principles HTR, Turkish, English, Disaster Culture) letter grade	The 9-category evaluation result shows the end-of-term success status of the related course; AA is the top success grade, F is the low success grade, M is exempt	AA, BA, BB, CB, CC, DC, DD, F, M
Weighted Grade Point Average (WGPA)	Successful (point average 2 and above), Unsuccessful (point average below 2)	1-0

Figure 2 illustrates the distribution of letter scores in Compulsory Courses for students.

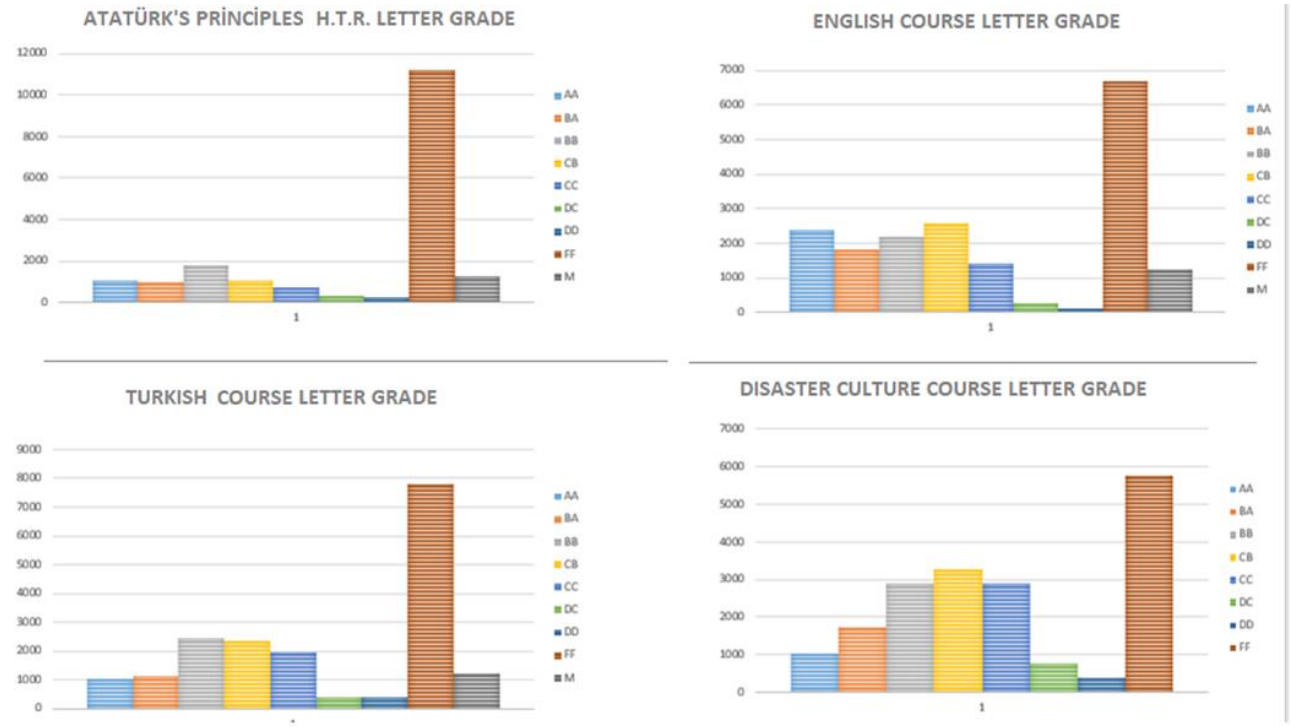

Figure 2: Distribution of Letter Grades

When Figure 2 is analyzed, it is seen that the highest frequency is in the FF grade in all letter grades of the four courses. Other letter grades have similar frequencies.

4.4. Data Analysis

In the context of this study, the execution of processes for data comprehension and preparation utilized an SQL database and Microsoft Excel 2016. Following this, IBM SPSS Statistics 25 took on tasks such as data recoding, cleaning, handling missing values, identifying outliers, and addressing noisy data. The development of data mining algorithms relevant to the research was achieved through the use of IBM SPSS Modeler 18, and the analysis incorporated data mining models available within this software. Which of the data obtained from the database could be included in the analysis was determined with the guidance of the literature. Then, in order to identify the effective variables and reduce the number of features, the "feature selection" method in SPSS Modeler was used and 13 variables were selected to be included in the models for further analysis. This approach aimed to identify the key factors that can effectively predict the risk of student failure.

In line with the study's objectives and the dataset's characteristics, the analysis involved the use of several classification algorithms, including Random Forest (RF), TR-AS, C&RT, C5.0, CHAID, QUEST, Naive Bayes (NB), Logistic Regression (LR), NeuralNet and SVM, with their respective performance evaluated. Initially, the dataset was split into two subsets: 70% for training and 30% for testing. This division assessed model performance on data outside the training set and mitigated overfitting. Literature suggests a training-to-test data ratio of approximately two to one is suitable (Özcan, 2013). Lastly the study examined the hierarchical impact of variables in the data mining model, especially those contributing to high classification accuracy in predicting academic success.

In Additionally Accuracy, Sensitivity/Recall, Specificity, F1 score, Precision/positive predictive value, Negative predictive value and Matthews Correlation Coefficient (MCC) were obtained from the confusion matrix criteria used to evaluate the classification performance.

The development process of data mining models has progressed for two purposes: to reveal the characteristics of large data groups (descriptive) or to make predictions based on these data (predictive). This study categorized the students' academic achievements into "Successful" and "Unsuccessful." The analysis employed various classification models, including Random Forest (RF), Tree-AS, C&RT, C5.0,

CHAID and QUEST algorithms. Additionally, Naive Bayes (NB), Logistic Regression (LR), Artificial Neural Network (NeuralNet) and Support Vector Machine (SVM) algorithms were utilized. The main properties of the data mining algorithms and classification trees in the IBM SPSS Modeler 18 program are as follows.

4.5. Data Mining Algorithms

The RF is a big data processing model that creates a single tree. The method uses recursive partitioning to split the training records into segments by minimizing the impurity at each step.

The Tree-AS model allows for decision trees using either a CHAID or Exhaustive CHAID model. It examines crosstabulations between input fields and outcomes and uses a chi-square independence test for significance. If more than one relation is statistically significant, CHAID selects the most significant input field. Exhaustive CHAID is a modified version of CHAID that examines all possible splits for each predictor but takes longer to compute. It can generate nonbinary trees with multiple branches, works for all input types, and accepts case weights and frequency variables.

The C&RT model generates a decision tree for predicting future observations using recursive partitioning and minimizing impurity. A model is considered "pure" if 100% of cases fall into a target field category. Target and input fields can be numeric or categorical, with binary splits.

The C5.0 model generates a decision tree or rule set by splitting the sample based on the field with the maximum information gain. It allows multiple splits into more than two subgroups. The Tree-AS model, similar to CHAID, processes big data. It uses chi-square statistics to identify optimal splits, and exhaustive CHAID is a modified version that examines all possible splits but takes longer to compute.

CHAID model generates decision trees using chi-square statistics to identify optimal splits, with nonbinary trees and numeric range or categorical input fields. Exhaustive CHAID is a more thorough approach but takes longer to compute.

The QUEST model provides a binary classification method for building decision trees, designed to reduce the processing time required for extensive C&RT analyses while reducing the tendency found in classification tree methods to favor inputs that allow more splits. Input fields can be numeric ranges (continuous), but the target field must be categorical. All splits are binary. The QUEST model offers binary classification for decision trees, reducing processing time and favoring categorical inputs with binary splits, reducing C&RT analysis bias (IBM Corporation, 2022).

The Bayesian Network is a probability model that uses evidence and real-world knowledge to predict outcomes. It uses Tree Augmented Naïve Bayes and Markov Blanket networks for classification, making predictions even in missing information. The network helps understand causal relationships, predict outcomes, and avoid overfitting.

LR is a statistical technique for classifying records based on input values using a categorical target field. It uses binomial and multinomial models to build equations relating input field values to output field probabilities. Logistic regression models can handle symbolic and numeric input fields and provide predicted probabilities for all target categories. They are most effective when group membership is a categorical field. Logistic models can used as a baseline for other modeling techniques.

A NeuralNet can approximate various predictive models with minimal structural and assumption requirements. The relationship form is established through learning, achieving close approximation for linear relationships. In cases where nonlinear relationships are suitable, the neural network autonomously adapts the correct model structure. Nevertheless, it lacks ease of interpretation, making traditional statistical models preferable for explaining underlying processes.

SVM classifies data into one of two groups without overfitting. The SVM classifies data using a support vector machine, making it ideal for comprehensive datasets with numerous predictor fields.

4.6. Model Performance Evaluation Criteria

In this study, Accuracy, Sensitivity/Recall, Specificity, F1 score, Precision/positive predictive value, Negative predictive value, and Matthews Correlation Coefficient (MCC) were obtained from the confusion matrix criteria used to evaluate the classification performance of the data mining models.

Accuracy is the ratio of the number of students correctly predicted by the model to the total number of students. Sensitivity is the rate at which the model correctly predicts positive situations; Specificity is the rate at which the model correctly predicts negative situations; the F1 score expresses the balance between Sensitivity and Specificity in the model. Precision is the ratio of the number of positive instances correctly classified by the model to the total number of positively classified instances. This measure measures the ability of the classification model to eliminate false positives. The negative predictive value is the ratio of actual negative cases to the cases the model classifies as negative. Another criterion, MCC, which is less common in the literature than other criteria, is more successful, especially in comparing algorithms in unbalanced classes (Bulut et al., 2022). As values other than F1 near 1, it indicates perfect classification. The higher the F1 criterion, the better. Unlike the other criteria, MCC can take a value between -1 and 1, where -1 represents perfect misclassification and 1 represents perfect correct classification.

Suppose the intersection values of observations and predictions in the classification (Confusion) matrix are expressed as True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP). In that case, the evaluation criteria are calculated as follows:

- Accuracy = $\frac{(TP+TN)}{(TP+TN+FP+FN)}$
- Sensitivity/ Recall = $\frac{TP}{TP+FN}$
- Specificity = $\frac{TN}{TN+FP}$
- F1 Score = $2\ x\ \frac{Sensitivity\ x\ Specificity}{Sensitivity\ +\ Specificity}$
- Precision/positive predicted value = $\frac{TP}{TP+FP}$
- Negative predicted value = $\frac{TN}{FN+TN}$

Matthews Correlation Coefficient = $\frac{(TPxTN)-(FPxFN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

4.7. Findings and Discussions

Table 1 presents the classification performance data for model training and test analyses conducted using RF, TR-AS, C&RT, C5.0, CHAID, QUEST, NB, LR, NeuralNet and SVM models to predict the academic success of open and distance learning students.

Journal of Educational Technology & Online Learning
Volume 7 | Issue 2 | 2024
http://dergipark.org.tr/jetol

Table 1: Classification Performances of Data Mining Models

	Accuracy		Sensitivity/ Recall		Specificity		F1 score		Precision/ positive predicted value		Negative predicted value		MCC	
Models	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.
RF	0.815	0.776	0.845	**0.784**	0.801	0.772	0.823	**0.778**	0.654	0.601	0.921	**0.891**	0.609	0.523
Tr-AS	0.803	0.791	0.562	0.543	0.909	0.899	0.695	0.677	0.734	0.703	0.824	0.818	0.513	0.480
C&RT	0.790	0.792	0.504	0.510	0.917	**0.915**	0.651	0.655	0.728	**0.725**	0.806	0.810	0.474	0.477
C5.0	0.815	0.789	0.573	0.529	0.922	0.902	0.707	0.667	0.766	0.704	0.830	0.814	0.543	0.473
NeuralNet	0.795	0.796	0.598	0.600	0.882	0.882	0.713	0.714	0.692	0.691	0.832	0.834	0.502	0.503
LR	0.809	**0**.813	0.602	0.612	0.901	0.901	0.722	0.729	0.730	0.721	0.836	0.841	0.533	0.542
NB	0.800	0.800	0.626	0.622	0.878	0.878	0.731	0.728	0.694	0.691	0.841	0.841	0.519	0.516
CHAID	0.784	0.780	0.530	0.526	0.897	0.891	0.666	0.662	0.696	0.679	0.811	0.811	0.465	0.452
QUEST	0.781	0.779	0.491	0.489	0.910	0.906	0.638	0.635	0.707	0.696	0.801	0.802	0.451	0.444
SVM	0.923	0.789	0.830	0.626	0.964	0.861	0.892	0.725	0.911	0.663	0.927	0.840	0.816	0.495

When reviewing the classification performances of the models in Table 1 for both training and test data, it becomes evident that the model accuracy values in the test data are as follows: RF (0.776), Tr-AS (0.791), C&RT (0.792), C5.0 (0.789), NeuralNet (0.796), LR (0.813), NB (0.800), CHAID (0.780), QUEST (0.779), SVM (0.789).

Based on the accuracy criterion, the most effective classification performances are achieved with the following algorithms: C&RT (0.792), NeuralNet (0.796), NB (0.800), and LR (0.813). However, especially in data mining research in the field of education, analyses for the prediction of academic success and performance, and in particular, the aims of this study include early prediction of failure and development of measures for this. Accordingly, the correct classification and precise identification of the unsuccessful student (hence low false positive value) may be a higher priority than the classification of the successful student as unsuccessful (false negative). Considering this perspective, the models' specificity, precision (positive predicted values), and overall model accuracy are crucial factors in determining the most suitable model for classification performance. While LR exhibits the highest accuracy rate at 0.813, the C&RT model has the highest specificity (0.915) and positive predicted value (0.725). Furthermore, the RF model boasts superior sensitivity (0.784), F1 score (0.778), and Matthews Correlation Coefficient value (0.523) compared to the other models. To achieve a more comprehensive understanding of the models' performances, Figure 3 graphically illustrates a comparative analysis.

doi: http://doi.org/10.31681/jetol.1334687
Received 30 Jul 2023; Revised 15 Mar 2024; Accepted 19 Mar 2024
ISSN: 2618-6586.

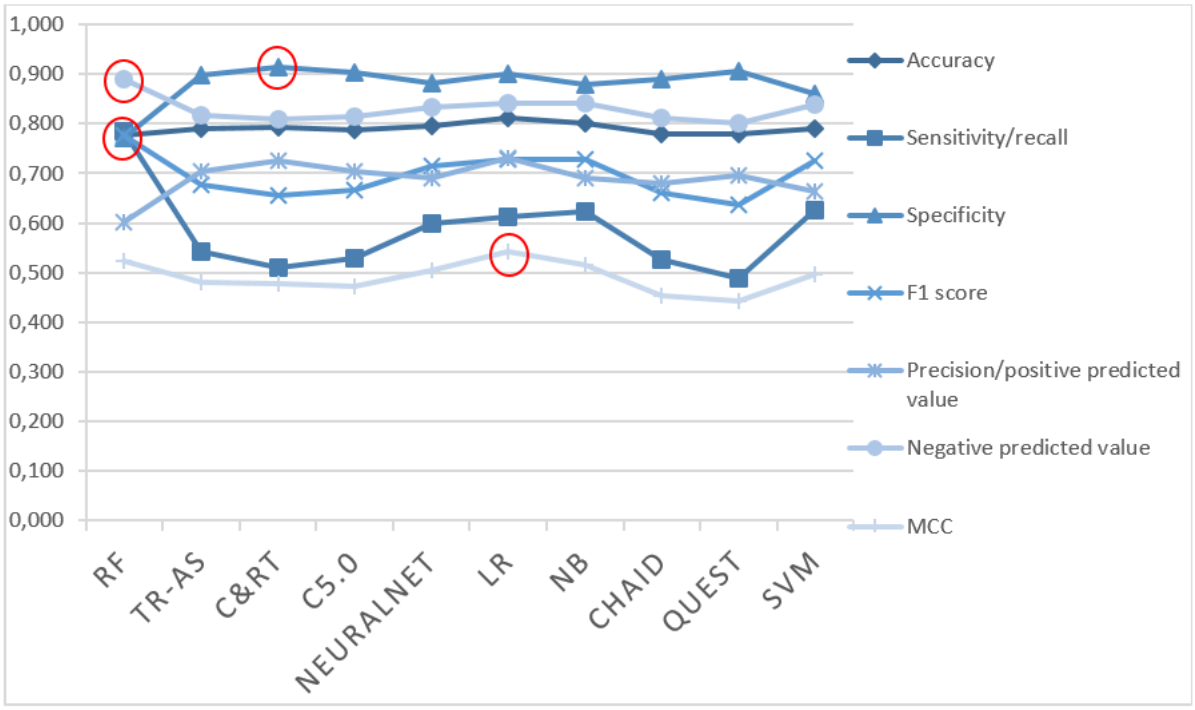

Figure 3: Model Performance Comparison Graph

At the Figure 3 is analysed, accuracy, sensitivity/recall, specificity, F1 score, precision/positive predictive value, negative predictive value, and Matthews Correlation Coefficient values are compared in data mining algorithms; accordingly, RF has the highest sensitivity and negative predictive value and F1 value; C&RT has the highest specificity value and the LR has the highest MCC value. Especially in cases where the target value (successful or unsuccessful) is not evenly distributed, model performance evaluation criteria other than accuracy are preferred.

Considering these results, the variables/features that will best classify academic success in RF and C&RT algorithms will be analysed. Figure 4 shows the importance of the variables in the Random Forest and C&RT models.

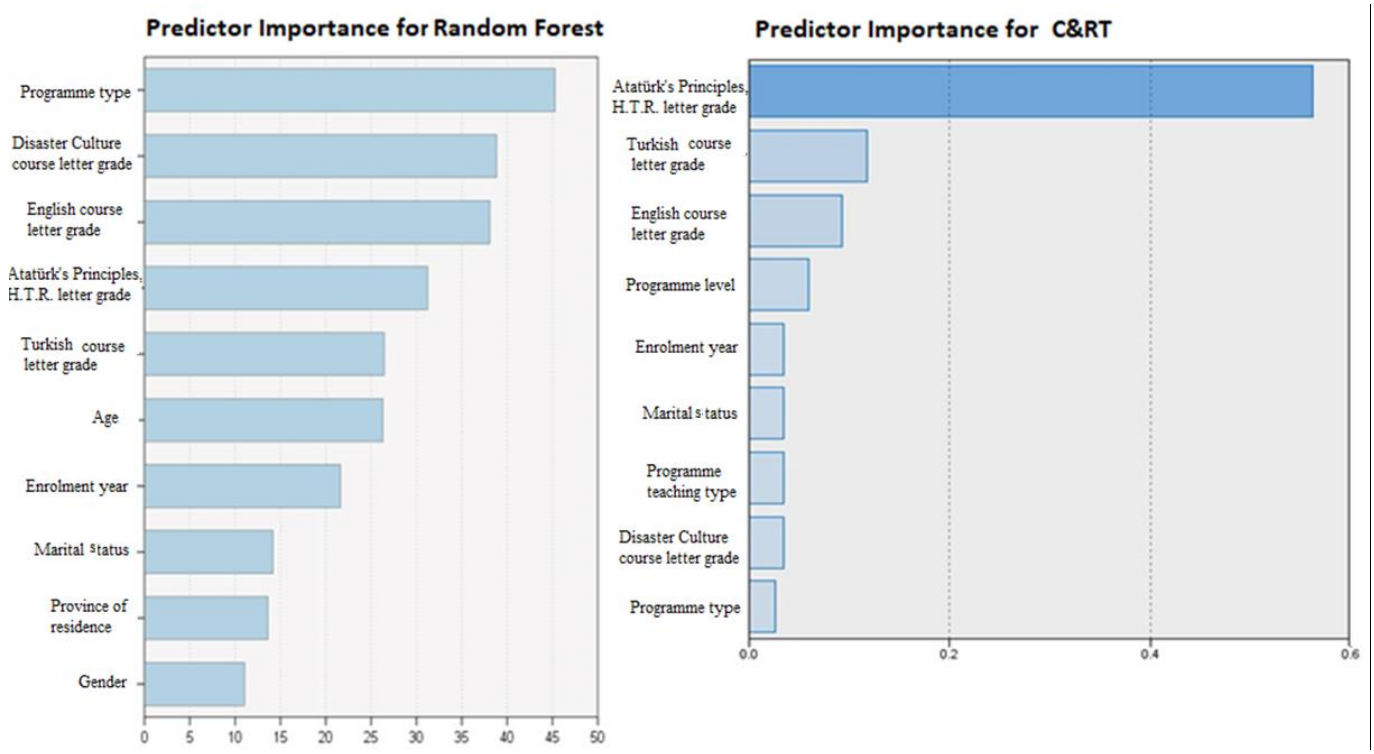

Figure 4: C&RT and RF Predictor Importance

In the C&RT model, Ataturk's Principles and History of Revolution Course letter grades are the most crucial prediction variable with a coefficient of 0.58, followed by Turkish Course letter grades, English Course letter grades, program level, year of enrolment, marital status, the program teaching type (open or distance), Disaster Culture Course letter grades and program type variables. In addition, unlike Orrego Granados et al. (2022) study, gender and residence province variables are not among the important prediction variables. In the Random Forest, branching starts with the Programme type variable (with a coefficient of 0.45). It continues with the following variables: Disaster Culture Course letter grade, English letter grade, Atatürk's Principles and History of Turkish Revolution Course letter grade, Turkish Course letter grade, age, year of enrolment, marital status, province of residence, and gender. In addition, 12 nodes were formed in the C&RT model, and the first eight are shown in Figure 5.

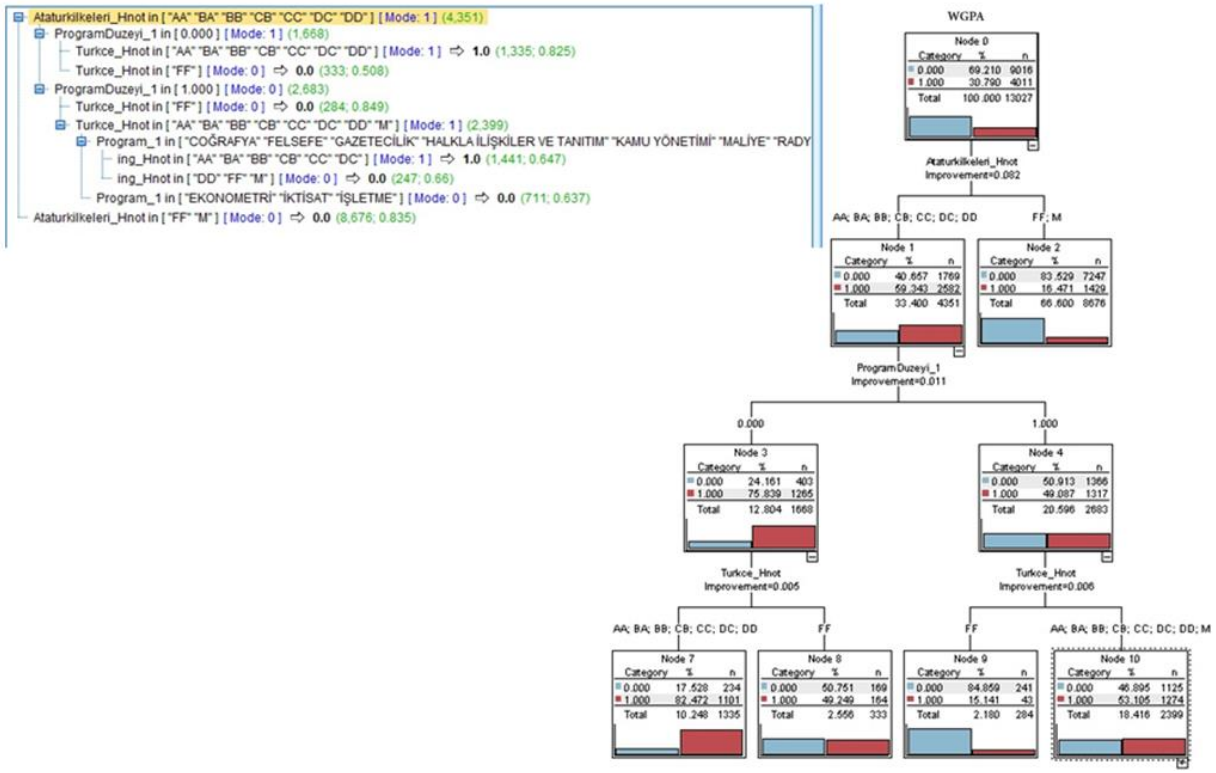

Figure 5: C&RT Nodes

According to the C&RT algorithm, the first node was constructed with the letter grade variable of Atatürk's Principles and History of Turkish Revolution Course; students with a letter grade of FF and M were classified as unsuccessful. The students with a letter grade of AA, BA, BB, CB, CB, DC, DD in Atatürk's Principles H.T.R. course and the students with a letter grade of AA, BA, BB, CB, DC, DD in the Turkish Course were classified as successful. The students with a letter grade of FF in the Turkish Course were classified as unsuccessful. Again, students with a letter grade of AA, BA, BB, CB, DC, DC, DD in Atatürk's Principles and History of Turkish Revolution and Principles and those with a letter grade of FF in the Turkish Course are classified as unsuccessful; students with a letter grade of AA, BA, BB, CB, DC, DD, M in Turkish Course and Geography, Philosophy, Public Relations and Publicity, Public Administration, Public Finance, Finance, Radio, Television and Cinema, Sociology, History, Labour Economics, and Industrial Relations, students with a letter grade of AA, BA, BB, CB, CB, DC in English Course are successful. Those with a letter grade of DD, FF, or M in English Course are classified as unsuccessful; those in Econometrics, Economics, and Business Administration programs are directly classified as unsuccessful.

The RF algorithm constructed its initial node based on program type, leading to the determination of the top five classification rules with the highest accuracy rates as follows:

1. Programme type, Emergency and Disaster Management, Department of Justice, Banking and Insurance, Foreign Trade, Econometrics, Philosophy, Journalism, Public Relations and Publicity, Law Office Management and Secretariat, Public Administration, Finance, Retail Sales and Store Management, Radio, Television and Cinema, Health Institutions Management, Social Services, Civil Air Transport Management, Medical Documentation and Secretariat, Labour Economics and Industrial Relations, Child Development, Economics, Occupational Health and Safety, Business Administration, Emergency and Disaster Management, Banking and Insurance, Geography, Geographical Information Systems, Foreign Trade, Econometrics, Cultural Heritage and Tourism, Finance, Media And Communication, Retail Sales and Store Management, Sociology, History, Economics, Business Administration and students with AA, CC, DC, DD, FF, M letter grade for Turkish Course and DC, DD, FF letter grade for Disaster Culture Course and FF, M letter grade for

Ataturk's Principles and History of Turkish Revolution Course are classified as with 0.975 Rule Accuracy and classified as unsuccessful.

2. Students whose program type is Geography, Geographical Information Systems, Foreign Trade, Philosophy, Journalism, Retail Sales and Store Management, Sociology, History, Labour Economics and Industrial Relations, Economics, Business Administration and who enrolled in the programs after 2013 and whose letter grade in Turkish Course is DC, FF and whose letter grade in Atatürk's Principles and History of Turkish Revolution is FF, M are classified as unsuccessful with 0.970 Rule Accuracy.
3. Students who are enrolled in the programs through the UEE exam and have a letter grade of BA, DC, DD, FF in English Course and Banking and Insurance, Geography, Foreign Trade, Econometrics, Philosophy, Journalism, Public Relations and Publicity, Law Office Management and Secretariat, Public Administration, Finance, Retail Sales and Store Management, History, Labour Economics and Industrial Relations, Child Development, Economics, Business Administration programs with a letter grade of DD, FF, M in Disaster Culture and FF, M in Turkish Course are classified as unsuccessful with 0.967 Rule Accuracy.
4. Undergraduate students whose gender variable is Male and whose letter grade in the Turkish Course is CC, DC, FF, M and whose letter grade in the Disaster Culture Course is AA, BA, BB, CB, CC, DD, DC, FF, M are classified as unsuccessful with 0.968 Rule Accuracy.
5. Students enrolled before 2013 who have a letter grade of FF in Turkish Course and who have a letter grade of DD, FF in Turkish Course and who are younger than 35 years old and who have a letter grade of FF, M in Atatürk's Principles and Revolution History Course are classified as unsuccessful with 0.964 Rule Accuracy.

5. Conclusion and Suggestions

As a result of the research, it was observed that data mining classification algorithms were successful in predicting academic success; model accuracy rates were between 0.78 and 0.92; student demographic variables and success grades of compulsory courses were essential variables for prediction. In line with Khasanah and Harwati (2017) showed that the courses' first-semester attendance and grade point averages were the primary prediction variables, and NB performed classification with the highest accuracy rate. This study concluded that Accuracy and Specificity criteria are also important in determining the data mining model that can be used in predicting the risky group. When the criteria are evaluated together, the C&RT algorithm performs best. For a similar purpose (determining the most critical factors in predicting student success and revealing the profiles of typical successful and unsuccessful students), Kovacic (2010) conducted a study with open education students, and again C&RT was found to be the most successful method with an overall correct classification percentage.

In addition, it was observed that the RF algorithm has a high precision value; therefore, it produces essential criteria for predicting unsuccessful groups according to the program type variable. The success grades in the compulsory courses of the first semester of the academic year are the most crucial predictor variable for subsequent academic success. In his study, Çırak (2012) found the artificial neural network algorithm successful in predicting students' academic achievement and showed that the most critical variable was the university entrance score. Based on this, one of this study's results is that academic scores are the primary predictors of academic achievement. In particular, according to the C&RT algorithm, the letter grade of Atatürk's Principles and History of Turkish Revolution Course and Turkish Course letter grades are among the prioritized variables in the classification rules.

Batool et al.(2023) state that previous academic achievement and demographic factors are the most important attributes in predicting student performance. The results of this study are partially consistent with the study of Batool et al., who concluded that Artificial Neural Networks and RF algorithms produce more effective results.

A university administration needs to be able to predict student performance to prevent student failure. Khasanah and Harwati (2017) applied the Bayesian Network and Decision Tree algorithms to predict student performance. They showed that the Bayesian Network has a higher accuracy rate, while students' attendance and grade point average in the first semester are the most critical variables. Therefore, similar to this study, it is once again confirmed that the success of the first semester courses is one of the prioritized variables for student performance.

Orrego Granados et al. (2022) found that student's academic achievement averages at the end of the first and second semesters had the highest correlation with final achievement in their data mining-based student performance prediction models. This result coincides with success in the first semester of compulsory courses of the universities in this study, which is the most important variable in predicting academic success at graduation. Similarly, in the C&RT algorithm, program types are of low importance compared to other factors. Again, Orrego Granados et al. (2022) reported that in addition to the first-semester end-of-year score, variables such as age, gender, time to graduation, and program types were also found to be essential for the prediction model.

Kotsiantis et al. (2004) found that the accuracy (0.724) and precision (0.78) of the Naive Bayes algorithm were relatively high in a data mining study using final exam scores of several courses and essential demographic variables to predict the academic performance of distance education students. However, considering that the algorithms compared in this study (Naïve Bayes, C4.5, LR, SMO, and 3-NN) are limited compared to the ones in this study, we can say that they do not contradict the results of this study.

Yossy et al. (2019), who aimed to find out which classification model has the best performance for student performance data, concluded that the best classification method is random forest G with 89.78%; demographic variables such as age, gender, and school and social characteristics of the student are important variables.

This study has certain limitations worth noting. Firstly, the socio-economic data of the students in the study group and information about their previous education history remained unavailable. Furthermore, while the dataset size and the processing time of data mining algorithms are typically essential performance indicators, in this study, the data size did not significantly impact algorithm processing times. As a result, processing time was not considered an evaluation criterion. It would be beneficial to conduct a similar study using a larger dataset and compare algorithm performance with processing time, especially when dealing with models that need to run concurrently in extensive databases.

The most accurate model for early prediction of academic failure can be achieved by utilizing data from the relevant academic unit. Since some variables in this study's data mining model (e.g., compulsory courses) pertain to the specific university where the research was conducted, it is advisable to conduct similar studies with additional variables at other open education universities. Furthermore, increasing the number of purpose-relevant variables can enhance the classification success of data mining models, warranting an investigation into its impact on model performance. A review of existing research in the field reveals variations in data mining methods and predictive variables used for student performance and academic achievement prediction. Therefore, it is recommended to conduct studies tailored to the data structures of individual educational institutions.

Moreover, data mining analyses performed after the first semester provide valuable insights to higher education administrators and educational planners, enabling them to anticipate student academic success. Consequently, this contributes to more effective educational planning tailored to at-risk student populations. Furthermore, acknowledging the pivotal role of program types as a significant variable in predicting academic success can inform the development of targeted support strategies for these students as they progress through their academic pursuits.

Ultimately, educational institution administrators have the opportunity to convert these and similar models into permanent tools for predicting academic success and implementing timely interventions for identified underperforming student groups.

References

Alan, M., & Temiz, M. (2019). A study on profiling students via data mining. *Alphanumeric Journal*, *7(2)*, 239-248. https://doi.org/10.17093/alphanumeric.630866

Albreiki, B., Habuza, T., Shuqfa, Z., Serhani, M.A., Zaki, N., & Harous, S. (2021). Customized rule-based model to identify at-risk students and propose rational remedial actions. *Big Data and Cognitive Computing, 5*(71), 1-17. https://doi.org/10.3390/bdcc5040071

Altıntaş, Ö., Başer, F., & Babadoğan, M. C. (2021). *Yükseköğretimde akademik riskli öğrencilerin kestiriminde makine öğrenmesi yöntemleri* (A. Apaydın & Ö. Kutlu, Eds.). Pegem Akademi.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 3. https://doi.org/10.1186/s41239-020-0177-7

Aydemir, E. (2019). Ders geçme notlarının veri madenciliği yöntemleriyle tahmin edilmesi. *European Journal of Science and Technology*, (15), 70-76. https://doi.org/10.31590/ejosat.518899

Bağrıacık Yılmaz, A.& Karataş, S. (2022). Why do open and distance education students drop out? Views from various stakeholders. *International Journal of Educational Technology in Higher Education*,19(28), 1-22. https://doi.org/10.1186/s41239-022-00333-x

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y

Bhise, R., Thorat, S.S. & Supekar, A.K. (2013). Importance of data mining in higher education system. *Journal of Humanities and Social Science (IOSR-JHSS)*, *6*(6), 18–21. https://doi.org/10.9790/0837-0661821

Bilici, Z., & Özdemir, D. (2021). Data mining studies in education: Literature review for the years 2014-2020. *Bayburt Eğitim Fakültesi Dergisi*. *17*(33), 342 - 376. https://doi.org/10.35675/befdergi.849973

Bonde, S. N., & Kirange, D. K. (2018). Educational data mining survey for predicting student's academic performance. A. P. Pandian, T. Senjyu, S. M. S. Islam, H. Wang (Eds), In Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018): Vol.31. (pp. 293-302). Springer International Publishing. https://doi.org/10.1007/978-3-030-24643-3

Bulut, O., Cormier, D. C., & Yildirim-Erbasli, S. N. (2022). Optimized screening for at-risk students in mathematics: A machine learning approach. *Information*, *13*(8), 400. https://www.mdpi.com/2078-2489/13/8/400

Çırak, G. (2012). *Yükseköğretimde öğrenci başarılarının sınıflandırılmasında yapay sinir ağları ve lojistik regresyon yöntemlerinin kullanılması* [Unpublished master thesis, Ankara Üniversitesi]. Ankara.

de Oliveira, C.F., Sobral, S.R., Ferreira, M.J., &Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing, 5*(64), 1-33. https:// doi.org/10.3390/bdcc5040064

Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, *pp. 47*, 5260–5267. https://doi.org/https://doi.org/10.1016/j.matpr.2021.05.646

Durairaj, M., & Vijitha, C. (2014). Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies*, *5*(4), 5987-5991.

Elakia, G., Aarthi, N. J. (2014). Application of data mining in educational database for predicting behavioural patterns of the students. International Journal of Computer Science and Information Technologies (IJCSIT), 5 (3), 4649–4652.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. McGraw-Hill.

He, Y., & Zhang, S. (2011, May 28-29). Application of data mining on students' quality evaluation. In *2011 3rd International Workshop on Intelligent Systems and Applications*. IEEE.

Hotaman, D. (2020). Öğrenci başarısının değerlendirilmesinde eğitsel veri madenciliğinin kullanılması. *Ulakbilge Dergisi*, *8*(48), 577–587. https://doi.org/10.7816/ulakbilge-08-48-08

IBM Corporation. (2022). *Decision Tree Nodes*. Retrieved 23.07.2023 from https://www.ibm.com/docs/en/cloud-paks/cp-data/4.7.x?topic=palette-modeling

Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision Analytics Journal*, *7*, 100204. https://doi.org/https://doi.org/10.1016/j.dajour.2023.100204

Khasanah, A. U., & Harwati. (2017). A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques. *IOP Conference Series: Materials Science and Engineering*, *215*, 012036. https://doi.org/10.1088/1757-899x/215/1/012036

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*(5), 411-426. https://doi.org/10.1080/08839510490442058

Kotsiantis, S. B., & Pintelas, P. E. (2005, 5-8 July 2005). Predicting students marks in Hellenic Open University. Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05),

Kovacic, Z. (2010). *Early prediction of student success: Mining students' enrolment data.* In Proceedings of Informing Science and IT Education Conference.

Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., & Islam, A. H. M. S. (2021). Mining educational data to predict students' performance. *Education and Information Technologies*, *26*(5), 6051–6067. https://doi.org/10.1007/s10639-021-10575-3

Natek,S., Zwilling, M. (2014. Student data mining solution–knowledge management system related to higher education institutions. *Expert systems with applications*, 41 (14), 6400–6407. https://doi.org/10.1016/j.eswa.2014.04.024

Okur M.R., Paşaoğlu Baş D., & Uça Güneş E.P., (2019). Açık ve uzaktan öğrenmede öğrenimi bırakma sebeplerinin incelenmesi. *Journal of Higher Education and Science*, 9(2), 225-235. https://doi.org/10.5961/jhes.2019.324

Orrego Granados, D., Ugalde, J., Salas, R., Torres, R., & López-Gonzales, J. L. (2022). Visual-predictive data analysis approach for the academic performance of students from a Peruvian university. *Applied Sciences*, *12*(21), 11251. https://doi.org/10.3390/app122111251

Osmanbegović, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Journal of Economics & Business/Economic Review*, 10, 3-12.

Özbay, Ö. (2015). Veri madenciliği kavramı ve eğitimde veri madenciliği uygulamaları. *Uluslararası Eğitim Bilimleri Dergisi*, (5), 262-272. https://dergipark.org.tr/tr/pub/inesj/issue/40015/475764

Özcan, T. (2013). *Veri Madenciliği*. İÜ AUZEF. https://cdn-acikogretim.istanbul.edu.tr/auzefcontent/21_22_Bahar/veri_madenciligi/1/index.html

Radovan, M. (2019). Should I stay, or Should I go? Revisiting Learner retention models in distance education. *Turkish Online Journal of Distance Education*, 20(3), 29–40. https:// doi. org/ 10. 17718/ tojde. 598211

Ramesh, V., P.Parkavi, & Ramar, K. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. *International Journal Of Computer Applications*, 63, 975-8887.

Rana, S., & Garg, R. (2016, February 12-13). *Application of hierarchical clustering algorithm to evaluate students' performance of an institute.* Second International Conference on Computational Intelligence & Communication Technology (CICT).

Saheed, Y. K., Oladele, T. O., Akanni, A. O., & Ibrahim, W. M. (2018). Student performance prediction based on data mining classification techniques. *Nigerian Journal of Technology*, *37*(4), 1087. https://doi.org/10.4314/njt.v37i4.31

Shahiri, A. M., Husain, W., & Rashid, N. a. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414-422. https://doi.org/https://doi.org/10.1016/j.procs.2015.12.157

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011, April). *Prediction of student academic performance by an application of data mining techniques* International Conference on Management and Artificial Intelligence IPEDR (Vol. 6, pp. 110-114).

Şengür, D., & Tekin, A. (2013). Öğrencilerin mezuniyet notlarının veri madenciliği metotları ile tahmini. *Bilişim Teknolojileri Dergisi,6*(3), 7-16.

Tan, S. S., Göktaş, Y., & Koçak, Ö. (2018, December 12-13). Veri madenciliği ile ÖSYM verileri kullanılarak akademik başarı tahmini [Conference presentation abstract] 2. Uluslararası Uzaktan Eğitim ve Yenilikçi Eğitim Teknolojileri Konferansı, Amasya.

Taşdemir, M. (2012). *Veri madenciliği (Öğrenci başarısına etki eden faktörlerin regresyon analizi ile tespiti).* (Publication Number 326726) [Master's thesis, Dicle Üniversitesi]. Diyarbakır. https://acikerisim.dicle.edu.tr/xmlui/handle/11468/789

Tekin, A., & Öztekin, Z. (2018). Eğitsel veri madenciliği ile ilgili 2006-2016 yılları arasında yapılan çalışmaların incelenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, *8*(2), 108-124. https://doi.org/10.17943/etku.351473

Tosun, M. (2016). *Açık öğretim öğrencilerinin akademik başarı düzeylerinin karşılaştırılması*. (Publication Number 427908) [Master's thesis, İstanbul Üniversitesi]. İstanbul.

Türel, Y. K., & Baz, E. (2016, 6-8 Ekim). Eğitsel veri madenciliği üzerine bir araştırma. 4. Uluslararası Öğretim Teknolojileri ve Öğretmen Eğitimi Konferansı Bildirileri, Fırat Üniversitesi, Elazığ.

Veeramuthu, P., & Periasamy, R. (2014). Application of higher education system for predicting student using data mining techniques. *International Journal of Innovative Research in Advanced Engineering*, *1*(5), 36-38.

Yossy, E. H., & Heryadi, Y. (2019, December). Comparison of data mining classification algorithms for student performance. In *2019 IEEE International Conference on Engineering, Technology, and Education (TALE)* (pp. 1-4). IEEE. doi:10.1109/TALE48000.2019.9225887

YÖKSİS. (2023). *Higher Education Information Management System; 2022-2023 Academic Year Higher Education Statistics*. Retrieved 12.11.2023 from https://istatistik.yok.gov.tr/

Zawacki-Richter, O., & Qayyum, A. (2019). *Open and distance education in Asia, Africa and the Middle East: National perspectives in a digital age*. Springer Nature.