Received: 14 September 2023 | Revised: 23 April 2024 | Accepted: 25 April 2024

DOI: 10.1111/hequ.12539

RESEARCH ARTICLE

Higher Education Quarterly WILEY

Development of an early warning system for higher education institutions by predicting first-year student academic performance

Cem Recai Çırak[1,2] | Hakan Akıllı[3] | Yeliz Ekinci[4]

[1]Department of Control and Automation Engineering, Istanbul Technical University, Istanbul, Türkiye

[2]Department of Electrical and Electronics Engineering, Istanbul University - Cerrahpaşa, Istanbul, Türkiye

[3]Business Intelligence Department, Anadolubank, Istanbul, Türkiye

[4]Department of Management Information Systems, Istanbul Bilgi University, Istanbul, Türkiye

Correspondence
Cem Recai Çırak, Department of Control and Automation Engineering, Istanbul Technical University, Istanbul, Türkiye.
Email: ccirak@itu.edu.tr

Abstract

In this study, an early warning system predicting first-year undergraduate student academic performance is developed for higher education institutions. The significant factors that affect first-year student success are derived and discussed such that they can be used for policy developments by related bodies. The dataset used in experimental analyses includes 11,698 freshman students' data. The problem is constructed as classification models predicting whether a student will be successful or unsuccessful at the end of the first year. A total of 69 input variables are utilized in the models. Naive Bayes, decision tree and random forest algorithms are compared over model prediction performances. Random forest models outperformed others and reached 90.2% accuracy. Findings show that the models including the fall semester CGPA variable performed dramatically better. Moreover, the student's programme name and university placement exam score are identified as the other most significant variables. A critical discussion based on the findings is provided. The developed model may be used as an early warning system, such that necessary actions can be taken after the second week of the spring semester for students predicted to be unsuccessful to increase their success and prevent attrition.

1 | INTRODUCTION

Universities need to develop early warning systems for students who are at risk in terms of academic performance and attrition. These early warning systems will help the university to take action for these students to increase their success and retain them. On the other hand, predicting students' and graduates' behaviours and standings is one of the biggest challenges that higher education faces today. Fortunately, in this century, institutions collect a vast of data and these data can be analysed by data mining techniques and meaningful results can be derived, which can be used for strategic and tactical decisions (Wang et al., 2022). Data mining techniques may be utilized in various higher education applications such as student loss prediction (Bowman et al., 2020; Márquez-Vera et al., 2016; Mason et al., 2018; Perez et al., 2018; Vafeiadis et al., 2015), elective course suggestion system development (Al-Badarenah & Alsakran, 2016; Dwivedi & Roshni, 2017; Kardan & Sadeghi, 2013), course schedule planning (Goyal & Vohra, 2012), curriculum development (Buitrago-Florez et al., 2022; Gupta et al., 2015) and student academic performance prediction that is reviewed in Section 2 in detail.

The literature review shows that studies on the prediction of student academic performance for freshmen based on data analytics in the last decades are few (Anderton et al., 2016; Garcia & Mora, 2011; van Zyl et al., 2016). Another gap in the current studies is that a limited number of input variables are analysed and taken into account in the prediction models. Therefore, the literature review motivates us to study student academic performance prediction for freshmen at universities while providing a detailed list of variables that affect the success of the prediction. Another motivation to study first-year students—freshman—is the idea that the first year is the most important period for students to get used to university life and for mitigating risks of attrition as this is the time when students develop a sense of belonging and academic and personal connections (Shcheglova et al., 2020).

The novelty of the study comes from the fact that there are very few studies that predict student success for freshmen (Beattie et al., 2018) and the current studies use limited number of input variables. The early warning system suggested in this study will make the predictions after the first-year spring semester registration and add–drop period which is the second week of the 14-week semester; so that, there are still 12 more weeks to take action for the students who are classified as unsuccessful. Early prediction of the students who will be unsuccessful will help the universities provide the necessary motivation and support for these students. Moreover, this early warning system will also list the significant variables and their effect on the predictions. Based on these results, the university will be able to make future plans for the future.

In this study, the dataset used in the proposed models consists of 11,698 undergraduate students' data retrieved from an anonymous university in Türkiye. The prediction models are constructed based on well-known machine learning classification algorithms; naive Bayes, decision tree and random forest, and the performances of these models are compared.

The paper is structured in five more sections. Section 2 presents an extensive literature review on the prediction of student academic performance in higher education. Section 3 explains the machine learning methods that are used in this study. Section 4 presents the methodology and application that is used to develop the early warning system. Section 5 summarizes the results of the applied methodology. Section 6 finalizes the paper with conclusions and future research directions.

2 | PREDICTION OF STUDENT ACADEMIC PERFORMANCE IN HIGHER EDUCATION

Student success has been studied in the literature in various aspects since the definition of it may vary among people. For instance, (Weatherton & Schussler, 2021) states that students define success as gaining leadership skills, building career networks, employability and being happy. Undoubtedly, an important part/factor of student

success is academic performance, as well. From the viewpoint of higher education institutions, student success/academic performance can be defined as meeting the conditions and achieving the objectives required to complete the registered programme. The main component of student academic performance is the grades received from the courses offered by the academic programme. Averages of the grades obtained from these courses are defined as student academic performance in the programme.

Studies revealed that the most important source of stress for students is the problems related to their academic success (Beiter et al., 2015). Moreover, various factors affect the academic performance of the students. Among these factors, the school's effect on students' success is evaluated as between 5% and 15% (Fong et al., 2015). The remaining factors affecting academic success are students' personal and social conditions, such as previous academic achievements and educational background, natural abilities and tendencies, socioeconomic status and living spaces (Cano & Leonard, 2019; Ruegg et al., 2021). Even under similar conditions and in the same class, there may be significant differences among students' academic performance. Academic performance can be increased by determining the reasons for these differences and preventing the factors that cause failure. These factors can be determined by statistical methods and data mining techniques (Polyzou & Karypis, 2019). Data mining techniques provide a great benefit in inferring these relationships. While each variable can affect success directly, these effects can also be measured by analysing groups of variables together (Beattie et al., 2018). For example, while the high school diploma grade may directly affect a student's academic performance, all students' success may be affected by the lecturer's academic knowledge and teaching skills.

According to existing studies in the literature, it has been seen that university students' academic successes and their successes in non-academic and non-university activities have no direct relationship (Beattie et al., 2018; Žuljević & Buljan, 2022). Having families with high socioeconomic and educational levels positively affects students' academic performance at varying levels based on students' gender (Aina et al., 2022; Glaesser & Cooper, 2012). To have a better understanding of the recent studies in the literature, Table 1 is constructed.

Table 1 summarizes recent studies that predict student academic performance. The datasets used in these studies and output and input variables with the utilized methodology are given in the table together with the findings; which may be seen in Table 1, only (Beattie et al., 2018) predicts first-year student success. Additionally, the current studies use limited number of input variables. The literature review also shows that machine learning techniques are used widely, especially in random forests.

The reviewed literature in Table 1 motivates us to study student academic performance prediction for freshmen at universities while providing a detailed list of variables that affect the success of the prediction. At universities, besides passing all the courses taken, the cumulative grade point average (CGPA) calculated based on the letter grades taken from these courses must be above the minimum level determined by regulations to graduate. Even though the conditions heavily depend on the universities and the countries where these universities are located; the international universities using a 4.0 grading scale generally accept a 2.00 CGPA of 4.00 as the minimum grade required to graduate from an undergraduate programme and pass semesters without facing negative status (such as on probation, suspended, etc.) as well. In this paper, the minimum CGPA required for academic success (and for graduation) is also 2.00 of 4.00, and the output variable is defined via categorical values (successful or unsuccessful) according to the CGPA.

The problem in this study is a classification problem and classification algorithms are supervised machine learning techniques. Since the dataset in this study includes both categorical and numeric variables, classification algorithms such as random forest, decision tree and naive Bayes which allow to use both types of variables are selected. These models performed successfully in similar studies as well (Beaulac & Rosenthal, 2019; Daud et al., 2017; Gray & Perkins, 2018; Hasan et al., 2018; Mengash, 2020; Rivas et al., 2021; Vinoth Kumar et al., 2021).

TABLE 1 Literature review.

Authors	Dataset	Output variable	Input variables	Methodology	Findings
Daud et al. (2017)	100 students	The degree is completed or not	23 variables including student personal information, family expenditures, income and assets	Support vector machine (SVM), decision tree, Bayes network and naive Bayes	SVM performs best with ~86% accuracy; family expenditure variables are more significant
Lu et al. (2018)	59 students	Course final performance is lower than 60 points or not	21 variables including homework and quiz scores, attendance and online learning and practice logs	Principal component regression	~82% accuracy is reached; quiz and online practice scores and online activity number variables show higher significance
Hasan et al. (2018)	22 students	Semester GPA (fail, pass, average, good or excellent)	11 variables including CGPA, previous failures, coursework and learning activities	Decision tree, random forest, naive Bayes and SMO classifier	Random forest and SMO classifiers outperform others with 100% accuracy
Beattie et al. (2018)	1317 students	Standardized first-year CGPA	7 variables including proneness to study for exams, expected GPA study and paid work hours, exercise start time, impatience and conscientiousness	Least angle regression and survey	If high school GPA is used, it becomes the best predictor, but non-academic variables are good at predicting unexplained extreme outcomes
Beaulac and Rosenthal (2019)	38,842 students	The degree is completed or not	142 variables including the first year's taken credits from 71 different departments, and average grades of all taken courses from each department	Random forest and logistic regression	Random forest performs better and ~79% accuracy is achieved; grades in math, finance and economics are important variables
Jokhan et al. (2019)	1403 students	Coursework score	2 variables including average login per week and task completion rate in the online learning module	Linear regression	Accuracy is 60.8%; the average task completion rate is highly correlated with coursework scores
Gray and Perkins (2018)	21,000 students	Semester GPA (pass, fail, conditional fail, repeat semester or repeat year)	5 variables including school or programme, year and online learning metrics of Weeks 1, 2 and 3	1 nearest neighbour, decision tree, random forest, naive Bayes and multilayer perceptron	Random forest outperforms other methods with ~88% overall accuracy

TABLE 1 (Continued)

Authors	Dataset	Output variable	Input variables	Methodology	Findings
Waheed et al. (2020)	32,593 students	Pass or fail the course	30 variables including students' demographics and online learning system analytics	Deep artificial neural network (ANN), SVM and logistic regression	Deep ANN is the best-performing method with ~88.6 overall accuracy
Mengash (2020)	2039 students	CGPA (poor, acceptable, good, very good and excellent)	3 variables including high school GPA, general aptitude test (GAT) and scholastic achievement test (SAT)	ANN, SVM, decision tree and naive Bayes	ANN performs better and reaches 79.2% accuracy; SAT has the highest correlation with CGPA
Rivas et al. (2021)	32,593 students	Pass or fail the course	39 variables including student demographics, previous academic recordings and online learning activities	Multilayer perceptron, decision tree, random forest and extreme gradient boosting	Multilayer performs slightly better than other methods with 78.2% accuracy; total click is the most important variable
Vinoth Kumar et al. (2021)	649 students	CGPA (fail or pass) and CGPA (average, good or outstanding)	13 variables including student demographics, past examinations, class activities, sports activities, health situation, free time and Internet availability	Naive Bayes, random forest, stacking, ADA boosting, decision tree, ANN, SVM and K nearest neighbours	The proposed multi-tier algorithm reaches up to 90.8% accuracy by employing naive Bayes on the first tier to predict 'fail' or 'pass' and random forest, stacking and ADA boosting methods are combined
Yakubu and Abubakar (2022)	747 students	CGPA (pass or fail)	5 variables including entry age, gender, region, study year and entry exam score (JAMB)	Logistic regression	83.5% prediction accuracy is reached. Study year and region are effective variables

3 | MACHINE LEARNING ALGORITHMS

3.1 | Naive Bayes

Naive Bayes is an efficient learning method that is commonly used as a probabilistic classification algorithm. The naive Bayes classifier algorithm is based on the Bayes theorem, named after Thomas Bayes and the assumption of independent variables (Tang et al., 2016). The desired information to be predicted is obtained by constructing proportional relations between dependent and independent variables. Proportional calculations are expressed simply in terms of the number of repetitions of categorical values of the output variable and all independent variables.

3.2 | Decision tree

Decision tree is a well-known learning method, which can be used for classification and regression, quite popular in machine learning and data mining applications (Al-Barrak & Al-Razgan, 2016). Decision trees may be split into two main groups: classification trees where the output variables are categorical and regression trees where the output variables are numerical. A decision tree is branched off to nodes starting from the root node according to the independent variables in the dataset. There are two types of nodes, which are the decision node and the leaf node. Decision nodes are used to make any decision and have multiple branches, whereas leaf nodes are the probable outcomes for those decisions and do not contain any further branches. The branching process continues until no new determinative variable is left (Loh, 2011). The decision at the root node gives the decision tree model output. A generic decision tree diagram is shown in Figure 1.

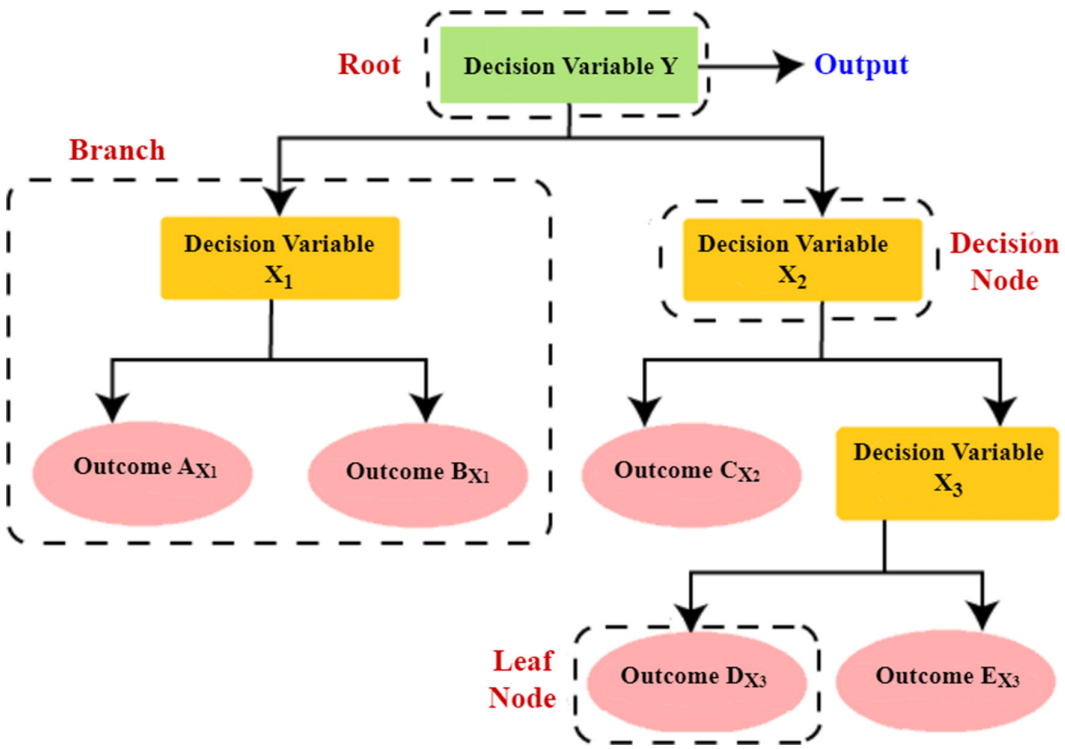

FIGURE 1 A generic decision tree diagram.

The advantages of the usage of the decision tree method may be listed as follows. (1) Training and testing of decision trees are convenient and fast. (2) Decision trees are easy to interpret and relatively more understandable than other algorithms. (3) Many algorithms give output but not its reason, whereas it is possible to capture the cause-and-effect relationship in decision trees. (4) Decision trees can generate rules (Kotsiantis, 2013).

3.3 | Random forest

Random forest is a learning method that can be used for classification and regression. A random forest structure consists of a collection of many decision trees. Each tree in the forest makes predictions based on a random subset of independent variables and uses a random sample of values from the dataset for training (Paul et al., 2018). Random forest classifier output is the category selected by the majority of all individual trees. Random forest regressor output is the average of the predicted outputs by all individual trees. Therefore, random forest models are expected to make more accurate and less sensitive predictions than decision trees in general (Beaulac & Rosenthal, 2019; Pranckevičius & Marcinkevičius, 2017). Random forest models are also able to identify and provide variable importance scores. Thus, alongside the model predictions, the most significant variables in the model may be obtained (Langan et al., 2018).

4 | METHODOLOGY

This study employs data mining and machine learning techniques for the early prediction of first-year higher education students' academic success.

4.1 | Problem definition

It can be stated that the primary and most important goal of all educational institutions is to graduate successful students. To achieve this, it is necessary to identify the factors that cause success and failure and make studies and improvements on these issues. In this paper, data mining techniques are used to predict academic success and to determine the relevant factors.

The dataset used for experimental analyses in this study is retrieved from an anonymous non-profit private (foundation) university from Türkiye. The university has more than 20,000 students, and approximately 10% of them are international students from all over the World (but mainly from countries in the Balkans, Eurasia, Middle East and Africa). The medium of instruction in the university is English, except for the faculty of health sciences where Turkish is used. An academic year in the university consists of two 14-week semesters, fall and spring. The programmes with a foreign language (English) medium of instruction have a 1-year English preparatory school (prep school) prior. Therefore, the first-year students who are in the prep school do not exist in the dataset. Thus, only freshman-year students are considered. The dataset includes 11,698 freshman-year undergraduate students' data.

In this study, academic performance for freshmen is predicted via machine learning. The determined output variable is the student's first-year academic performance which is either successful or unsuccessful. If a student's CGPA is greater than or equal to 2.00 over 4.00 at the end of the first year, the student is called successful and unsuccessful otherwise. The CGPA at the end of the first academic year is obtained as the weighted average of the end-term grades of all courses taken in the first year.

Once the output variable is determined and clarified, the next decision becomes when the predictions will be made. It is decided to make the predictions after the spring semester registration and add–drop period so that there

are still 12 more weeks to take action for the students who are classified as unsuccessful. The earlier the predictions are made, the more time there will be to take action. The accuracy rate of the prediction model is also related to the time of emergence of the independent variables included in the model. For instance, at the time of enrolment, only the students' demographic data are available. In order to use academic information as well as demographic information, it is required to wait until the data about students' academic behaviour emerge. Therefore, to predict the first-year academic performance of a student (end of spring semester), the data from the fall semester and the data which belong to the period when the spring semester course registration and add–drop period finish are needed. Add–drop is a period after the registration period—the second week of the 14-week semester—in the academic calendar that allows students to change or drop the elective courses out of the courses they registered for or to add new elective courses. Academic information such as students' first-term GPA, behaviours and course selections for both the fall and spring semesters will be available after this period. Additionally, prep school information is also available for this period.

4.2 | Data analysis

In this section, the data provided by an anonymous university will be examined and analysed. The available data, which may be included in the model as variables, are listed in Table 2.

A critical point to be considered in selecting the training dataset is how much of the historical data should be used. One of the main factors that will affect this decision is the quality and continuity of the data over time. In this study, as a result of the data analysis, it is decided to use the last five academic years' data.

4.3 | Data preparation

As the first step of data preparation, a transformation was applied. ETL (extract, transform and load) tools are frequently used in data mining and business intelligence and they provide convenience in data conversion processes. They fundamentally facilitate the management of data transformation, merging and moving processes by connecting to different data sources. In this paper, the Data Services (https://www.sap.com/products/technology-platform/data-services.html), which is an ETL tool and an SAP product, is used for the data transformation process.

In this study, to obtain single-row formatted data per student, the next preparation step should be transforming the transactions including multiple rows into the single-row format without loss of meaning. For instance, the course instructor info variable data include information for each course instructor. Accordingly, instead, the fall semester full-time instructor ratio and the spring semester full-time instructor ratio variables are introduced as the ratios of the weekly class hours taken from full-time instructors to the total weekly class hours for the fall and spring semesters.

Using ratios for some numerical variables instead of units may increase the model's success. For instance, students' academic behaviours may differ even for the same educational institution according to the programme and special conditions. Therefore, using the fall semester attempted credits and the fall semester gained credits variables as units will not be a balanced approach for the students taking different numbers of courses. Hence, these variables are combined in the fall semester gained credits ratio variable, as a relative value to balance the data. Similarly, the fall semester weekly online class hours and the spring semester weekly online class hours variables are transformed into the fall semester online class hour ratio and the spring semester online class hour ratio variables.

Data analysis revealed that some students were included in the target set for more than 1 year due to grade repetition. Thus, it is decided to include only the last academic year's record of students who repeat their freshman

TABLE 2 Available variables for the models.

Variable name	Explanation	Data type
Academic year	The academic year that student is in freshman year	Categorical
Gender	Student's gender	Categorical
TR citizenship	Whether the student is a Republic of Türkiye citizen or not	Categorical
Start academic year	The academic year that the student started the major programme	Categorical
Registration academic year	The academic year that the student got registered for the university first time	Categorical
Birthdate	Student's date of birth	Date
Father educational status	Student's father's educational level	Categorical
Mother educational status	Student's mother's educational level	Categorical
Scholarship rate	The rate of tuition fee waiver scholarship granted according to the student's university placement exam score as 100%, 50% or not applicable	Categorical
Course registration status	Whether the student made course registration or not	Categorical
Secondary education score	Student's secondary education composite score	Numerical
Prep school status	Whether the student studied in the 1-year English preparatory school before starting the major programme or not	Categorical
Prep school leave type	Whether the student passed the prep school based on the average grade during the prep school or via a proficiency exam	Categorical
Prep school grade	Student's prep school grade	Numerical
Prep school absence	Student's total number of classes of absence during the prep school	Numerical
Prep school starting level	Student's prep school starting level determined based on the proficiency exam score	Categorical
Prep school level repeat	Whether the student repeated a level due to unsuccessful performance in the level-up exams or not in the prep school	Categorical
Prep school repeat	Whether the student repeated the prep school year due to being unsuccessful at the end of the year or not	Categorical
Lateral transfer	Whether the student changed the programme via lateral transfer or not	Categorical
SIS login activities	The number of student's login activities for the student information system (SIS)	Numerical
Disciplinary punishment	Whether the student was subject to disciplinary action or not	Categorical
Disciplinary status report requests	Whether the student requested the disciplinary status report, which may be needed for various reasons such as external lateral transfer applications, or not	Categorical
Transcript document	Whether the student requested the transcript document, which may be needed for external lateral transfer applications, or not	Categorical
Fall semester CGPA	Student's CGPA based at the end of the fall semester	Numerical
Last spring semester GPA	Freshman-year repeating student's GPA based on the courses taken in the last repeated spring semester	Numerical
Fall semester gained credits	The number of total credits achieved by the student in the fall semester	Numerical
Fall semester attempted credits	The number of total credits attempted by the student in the fall semester	Numerical

(Continues)

TABLE 2 (Continued)

Variable name	Explanation	Data type
Transferred credits	The number of transferred credits into the student's transcript from another university	Numerical
Period of matriculation	The number of semesters passed since the student started the programme	Numerical
Course registration date	The date that the student completed the course registration	Date
Fall semester selected courses	The number of courses selected by the student in the fall semester	Numerical
Spring semester selected courses	The number of courses selected by the student in the spring semester	Numerical
Conflicting classes info	Whether the student selected courses with conflicting hours or not	Categorical
Summer school info	Whether the student enrolled in a course in summer school (an irregular semester with limited number of offered courses) or not	Categorical
Placement preference order	In which preference order, the student got placed in the university and the programme according to the student's university entrance exam result	Numerical
Number of students in the programme	The total number of students in the programme that the student got placed in	Numerical
Number of students in the faculty/school	The total number of students in the faculty/school that the student got placed in	Numerical
KYK scholarship	Whether the student gains a scholarship from KYK (credits and dormitories institution) or not	Categorical
Placement score	Student's university placement exam score	Numerical
Placement score category	The category of the student's university placement exam score (quantitative, verbal, equally weighted and language) that is used to be placed in the programme	Categorical
High school type	Whether the student studied in a public or private high school	Categorical
High school category	The category of the high school that the student studied in	Categorical
Region of the high school	The geographical region in which the high school was located	Categorical
City of the high school	The city in which the high school was located	Categorical
High school name	The name of the high school that the student studied in	Categorical
High school graduation year	The year that the student graduated from high school	Categorical
High school diploma grade	Student's high school diploma grade	Numerical
Course instructor info	Whether the instructor of the course that the student takes is a part-time or full-time instructor in the university	Categorical
Programme campus info	The university campus that is assigned for the student's programme	Categorical
Planned campus info	Whether the student takes courses on different campuses of the university or not	Categorical
Fall semester weekly class hours	The total number of weekly class hours of the courses that the student takes in the fall semester	Numerical
Spring semester weekly class hours	The total number of weekly class hours of the courses that the student takes in the spring semester	Numerical

TABLE 2 (Continued)

Variable name	Explanation	Data type
Fall semester weekly online class hours	The total number of weekly online class hours of the courses that the student takes in the fall semester	Numerical
Spring semester weekly online class hours	The total number of weekly online class hours of the courses that the student takes in the spring semester	Numerical
Number of frozen semesters	The number of semesters that the student freezes	Numerical
University registration type	Student's university registration type based on the placement	Categorical
Registration semester	Whether the student first registered for the programme in the fall (regular) or the spring (irregular) semester	Categorical
Programme name	The name of the programme that the student got placed in	Categorical
Department name	The name of the department that the student got placed in	Categorical
Faculty/school name	The name of the faculty/school that the student got placed in	Categorical
Medium of instruction	The medium of instruction in the student's programme	Categorical
Faculty/school type	Whether the major programme is under a vocational school or not	Categorical

year for more than one academic year in the dataset to guarantee that the dataset used for the model is created in the single-row per student format.

According to the abovementioned issues, a transformation is performed on the data, and new and transformed additional variables are defined and presented in Table 3. Accordingly, birthdate, prep school absence, disciplinary punishment, disciplinary status report request and fall semester gained credits variables are excluded. As a result, a total of 69 input variables provided in Table 2 (except for the excluded ones) and Table 3 are used in modelling.

In order to provide summary statistics of the dataset, the min, max and mean values of the most important numeric variables are given in Table 4. Additionally, there are 63 unique values for the programme name variable, 52.61% of the students are female and 47.39% are male, and 65.22% of the students studied in the English preparatory school. For the students' placement score category variable: 34.63% of the students are in the quantitative, 20.31% in verbal, 42.38% in equally weighted and 2.67% in the language category.

4.4 | Modelling

In this study, the machine learning algorithms discussed in Section 2 are used for modelling. Hence, three different models are run using naive Bayes, decision tree and random forest algorithms, and the performances of these models are compared. In this study, KNIME (https://www.knime.com), a practical open-source data mining software that includes commonly used machine learning algorithms, is used as software.

There are many parameter values that can be selected while running the models. In order to select the best parameter values, many combinations of different values are tried and the ones that resulted in higher performance (in terms of accuracy and precision) are selected as the final parameter values (split criterion: information gain; number of models: 500 for random forests; quality measure: gain ratio; pruning method: minimal description length for decision trees; and maximum number of unique nominal values per attribute: 20 for naive Bayes).

After analysing the dataset, it has been seen that there is a strong positive correlation between the fall semester CGPAs and the end of the first-year CGPAs; hence, it affects the student class (successful or unsuccessful) that is predicted by models. Additionally, the fall semester CGPA is seen as the variable with the highest importance in the classification models run in this study. Therefore, the models are run with and without the fall semester CGPA

TABLE 3 Additionally included and transformed variables for the models.

Variable name	Explanation	Data type
Student's age	A variable derived from the birthdate variable to convert the variable format from date to numerical that is more suitable for machine learning models	Numerical
New student	A variable added to make the model identify whether the student studied the freshman year the first time or repeated	Categorical
Prep school absence ratio	A variable derived from the prep school absence variable as the ratio to the total number of class hours in prep school to obtain a balanced variable for students who take different weekly class hours	Numerical
Number of disciplinary punishments	A variable that summarizes the disciplinary action data and converts it into the number of disciplinary punishments taken by the student	Numerical
Number of disciplinary status report requests	A variable that summarizes the disciplinary status reports as the number of student requests	Numerical
Fall semester gained credits ratio	A variable derived from the fall semester gained credits variable as the ratio to attempted credits variable to obtain a balanced variable for students who attempted different numbers of credits	Numerical
Is student registered in time	A variable showing whether the student completed the course registration during the registration week or as late registration	Categorical
Gap years after high school	A variable added to measure the effect of the number of gap years between the student's high school graduation and starting the university programme on the student's success	Numerical
Fall semester full-time instructor ratio	A variable that shows the ratio of the number of weekly class hours that the student takes from full-time instructors in the fall semester to the fall semester weekly class hours variable	Numerical
Spring semester full-time instructor ratio	A variable that shows the ratio of the number of weekly class hours that the student takes from full-time instructors in the spring semester to the spring semester weekly class hours variable	Numerical
Fall semester online class hour ratio	A variable that shows the ratio of the fall semester weekly online class hours variable to the fall semester weekly online class hours variable	Numerical
Spring semester online class hour ratio	A variable that shows the ratio of the fall semester weekly online class hours variable to the fall semester weekly online class hours variable	Numerical
Is student successful (output variable)	The output variable states whether the student is successful or not where it becomes 'successful' if the student's freshman year CGPA is greater than or equal to 2.00 over 4.00 and 'unsuccessful' otherwise	Categorical

variable, and the performances of the models and the importance scores of the other variables are calculated accordingly.

5 | RESULTS

The dataset consists of 11,698 freshman-year undergraduate students' data. It is divided into two subsets, train and test sets, before using them in the proposed models. Eighty per cent of the data are used to train the proposed models, and the remaining 20% are used to test the model predictions.

In machine learning-based binary classification problems, confusion matrices are constructed to measure the performance of the created models. The confusion matrix is a table that presents the frequencies of four possible outcomes that are the combinations of the predicted and actual classes. A standard confusion matrix format is given in Table 5.

In Table 4, actual class values are the realized results given in the test dataset, whereas predicted class values are the predicted results generated by the model. TP and TN represent the values correctly classified by the model, whereas FP and FN show the incorrectly performed classifications.

The output variable is labelled 'successful' if the student's CGPA at the end of the first academic year is 2.00 or higher out of 4.00, and 'unsuccessful' otherwise. In this paper, 80% of the data are used for training and 20% are used for testing. Thus, of 11,698 undergraduate students' data, 2340 students are selected randomly for the test dataset. In this test dataset, 1269 are successful and labelled as 'successful', whereas 1071 are unsuccessful and labelled as 'unsuccessful'. Therefore, the dataset is regarded to be balanced. Since predicting an 'unsuccessful' labelled student incorrectly as 'successful' may prevent taking the necessary precautionary actions for the student, the false-positive misclassification type is considered more important than the false-negative misclassification type.

The selection of the performance measure is another point in the evaluation of the model performances. Accuracy and precision are common confusion matrix-based metrics for binary classification models. Accuracy is the measure of the overall correct classification ratio. The precision represents the ratio of the number of correct classifications among all classes predicted as positive (Sokolova & Lapalme, 2009). The precision metric makes sense where the importance of FP and FN differs (Cui et al., 2008). Therefore, accuracy and precision metrics are used together to measure model performances. Accuracy and precision are calculated per Equations (1) and (2) respectively (Duman et al., 2012).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

TABLE 4 Summary statistics of the dataset.

Variable name	Minimum	Maximum	Average
Fall semester CGPA	0.00	4.00	1.98
Placement score	144	512	319
Fall semester selected courses	1	10	6.2
Spring semester selected courses	2	10	6.3
High school diploma grade	43.44	99.45	74.64
Fall semester weekly class hours	3	37	21.1
Spring semester weekly class hours	0	36	21.7
Prep school grade	60	100	71

TABLE 5 Standard confusion matrix.

	Predicted class	
	Positive	Negative
Actual class		
Positive	TP	FN
Negative	FP	TN

Abbreviations: FN, false negative; TP, true positive.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

5.1 | Model predictions

The confusion matrices for the models created by the naive Bayes algorithm with and without adding the fall semester CGPA as an input variable are provided in Tables 6 and 7, for test sets.

For the decision tree-based models including and excluding the fall semester CGPA variable, the confusion matrices are shown in Tables 8 and 9 respectively.

Tables 10 and 11 present the confusion matrices for the random forest models where the fall semester CGPA variable is added and removed in order.

5.2 | Comparison of models

Comparing three models that are run by naive Bayes, decision tree and random forest algorithms, it is seen that random forests show the best performance. The random forest models have the highest accuracy among

TABLE 6 Naive Bayes model (including fall semester CGPA variable) confusion matrix.

	Predicted result	
	Successful	**Unsuccessful**
Actual result		
Successful	TP: 1230 (52.56%)	FN: 39 (1.67%)
Unsuccessful	FP: 314 (13.42%)	TN: 757 (32.35%)

Abbreviations: FN, false negative; TP, true positive.

TABLE 7 Naive Bayes model (excluding fall semester CGPA variable) confusion matrix.

	Predicted result	
	Successful	**Unsuccessful**
Actual result		
Successful	TP: 1263 (53.97%)	FN: 6 (0.26%)
Unsuccessful	FP: 966 (41.28%)	TN: 105 (4.49%)

Abbreviations: FN, false negative; TP, true positive.

TABLE 8 Decision tree model (including fall semester CGPA variable) confusion matrix.

	Predicted result	
	Successful	**Unsuccessful**
Actual result		
Successful	TP: 1136 (48.55%)	FN: 133 (5.68%)
Unsuccessful	FP: 125 (5.34%)	TN: 946 (40.43%)

Abbreviations: FN, false negative; TP, true positive.

TABLE 9 Decision tree model (excluding fall semester CGPA variable) confusion matrix.

	Predicted result	
	Successful	**Unsuccessful**
Actual result		
Successful	TP: 963 (41.15%)	FN: 306 (13.08%)
Unsuccessful	FP: 609 (26.03%)	TN: 462 (19.74%)

Abbreviations: FN, false negative; TP, true positive.

TABLE 10 Random forest model (including fall semester CGPA variable) confusion matrix.

	Predicted result	
	Successful	**Unsuccessful**
Actual result		
Successful	TP: 1128 (48.21%)	FN: 141 (6.02%)
Unsuccessful	FP: 89 (3.80%)	TN: 982 (41.97%)

Abbreviations: FN, false negative; TP, true positive.

TABLE 11 Random forest model (excluding fall semester CGPA variable) confusion matrix.

	Predicted result	
	Successful	**Unsuccessful**
Actual result		
Successful	TP: 957 (40.90%)	FN: 312 (13.33%)
Unsuccessful	FP: 229 (9.79%)	TN: 842 (35.98%)

Abbreviations: FN, false negative; TP, true positive.

the used machine learning methods. Additionally, in random forest predictions, precision is higher than other methods.

The models have substantially higher prediction performances in terms of both accuracy and precision when the fall semester CGPA variable is included. Table 12 also shows the inclusion of the fall semester CGPA variable increases all model performances by 13%–28%. This means the fall semester CGPA is the most significant variable for the freshmen success prediction.

Additionally, the 10 most significant variables for the best-performing models that use the random forest algorithm are given with their importance scores in Table 13 for both cases that the fall semester CGPA variable included and excluded. The importance score I_j for each feature j is obtained via Equation (3) as the sum of the ratios of the number splits s_{jl} to the number candidates c_{jl} for each layer l of the generated trees during the random forest model training.

$$I_j = \sum_{\forall l \in \text{Layers}} \frac{s_{jl}}{c_{jl}}, \forall j \in \text{Features} \tag{3}$$

Table 13 shows that among the significant variables in freshman student success predictions for both random forest models (including and excluding the fall semester CGPA variable), fall semester CGPA becomes the most

significant variable when it is included. Programme name, placement score, spring semester selected courses and gender variables are seen as significant variables in both prediction models. Another finding is that prep school variables have a major cumulative effect on the prediction of student success. The intuitive assumption that the high school diploma grade provides a hint for the success of the student at the university is also supported by the results in Table 13.

Table 14 gives the details about how the values of important variables relate to the final prediction. As can be seen in Table 14, female students are more successful than male students. Additionally, students who do not repeat prep school are more successful. Moreover, students who have higher placement exam scores, high school diploma grades and prep school grades are more successful.

6 | CONCLUSION

In higher education, models created by machine learning techniques play a crucial role in improving the quality of education by identifying the factors affecting or indicating students' academic success and establishing early

TABLE 12 Summary of the model performances.

	Fall semester CGPA variable included		Fall semester CGPA variable excluded	
Model algorithm	**Accuracy**	**Precision**	**Accuracy**	**Precision**
Random forest	**90.2%**	**92.7%**	**76.9%**	**80.7%**
Decision tree	89.0%	60.9%	60.9%	61.3%
Naive Bayes	84.9%	79.7%	58.5%	56.7%

TABLE 13 The significant variables according to relative importance scores.

Fall semester CGPA variable included		Fall semester CGPA variable excluded	
Variable name	**Importance score**	**Variable name**	**Importance score**
Fall semester CGPA	2.851	Programme name	2.613
Fall semester gained credits ratio	2.125	Placement score	2.006
Programme name	1.816	Spring semester selected courses	1.562
Department name	1.643	Gender	1.390
Placement score	1.172	Prep school level repeat	1.259
Faculty/school name	1.016	Prep school absence ratio	1.240
Placement score category	0.966	Placement score category	1.190
Spring semester selected courses	0.899	High school diploma grade	1.040
Gender	0.812	Spring semester weekly class hours	0.954
Prep school repeat	0.781	Prep school grade	0.950

TABLE 14 Effects of important variables on prediction.

Is student successful	Male (%)	Female (%)	Prep school level repeat (%)	Placement score (average)	High school diploma grade (average)	Prep school grade (average)
Yes	41.14	65.30	28.73	788.823	77	73
No	58.86	34.70	71.27	384.659	72	68

warning systems. However, the literature review shows that studies on student academic performance prediction based on data analytics for freshmen have been scarce in the last decade. Moreover, in most studies, a limited number of input variables are included in the prediction models. These gaps in the literature motivate us to study student academic performance (successful/unsuccessful) prediction for freshmen at an international university while providing a detailed list of all input variables and the significant ones that affect the success of the predictions.

This study employs commonly used binary classification algorithms for academic performance prediction in higher education. The early warning system predictions are made after the spring semester registration and add–drop period, so there are still 12 more weeks to take action for the students classified as unsuccessful. The dataset used in the proposed models consists of 11,698 undergraduate first-year students' data retrieved from an anonymous international non-profit private university. The dataset is analysed, and finally, 69 input variables are used in the models. Models based on naive Bayes, decision tree and random forest algorithms are run, and their performances are compared. After analysing the dataset, a strong positive correlation is seen between the fall semester CGPA and the CGPA at the end of the first academic year. Accordingly, it highly affects the prediction of student classes (successful or unsuccessful) by models. Therefore, model performances are measured under the cases whether the fall semester CGPA data are fed into the models or not.

The results show that the models perform significantly better when the fall semester CGPA data are included. The random forest algorithm outperforms the naive Bayes and decision tree methods. Random forest models performed with 90.2% accuracy if the fall semester CGPA variable is included and with 76.9% accuracy if the fall semester CGPA variable is excluded. The most significant variables in model development are also analysed and discussed. The fall semester CGPA is identified as the most significant individual variable. Programme name (major), previous academic track variables, prep school performance variables and course load variables are seen as the other important variables in both cases where the fall semester CGPA variable is included and excluded. This finding shows convenience with the intuitive assumptions and the finding in (van Zyl et al., 2016) that the previous academic track is the most effective indicator of freshman-year students' success. Thus, these variables can be considered as direct indicators of future success.

Programme name, placement score and placement score category variables have strong individual effects on predictions in both cases where the fall semester CGPA variable is included or not. These variables together may imply two things about the programmes that accept students with placement scores in specific categories: (1) Some of these programmes are chosen by students who have better academic success. (2) Some of these programmes apply a curriculum consisting of courses that majority of them are relatively easier or harder to obtain better grades. These implications may point to the heterogeneous distribution of students in terms of academic performance among programmes and the heterogeneity of the programmes' curriculum levels. Even though these variables are practical for academic success and risk predictions, there may be a need for the development of well-balanced curriculums in the university.

In the case of the fall semester CGPA variable excluded, it is seen that prep school level repeat, prep school absence ratio and prep school grade variables become significant. They have high individual and significant cumulative effects on student success predictions. Moreover, Table 14 shows that the students tend to be more successful in the first year if the success in prep school is higher. This may imply that prep school performance, which is the indicator of proficiency level in the medium of instruction in the university, affects students' academic success and probably their learning performances. This is not surprising where students are predominantly non-native speakers of the medium of instruction since there is a strong relationship between academic success and the level of proficiency in the medium of instruction (Wait & Gressel, 2009). These variables may clue in the reason behind low academic performance for some students.

The spring semester selected course variable is one of the most important predictors in both cases, and the spring semester weekly class hours variable is also found significant in the case of the fall semester CGPA

excluded. These variables are related to students' course loads. Based on these findings, two intuitive implications can be made: (1) An excessive course load may be overwhelming for students and decrease their academic performance. (2) A course load far less than the regular load determined in the curriculum may lead to disengagement for students. Therefore, course load data should also be regarded while identifying potential risk factors for the students who are predicted to be unsuccessful.

Findings imply that demographic variables (except gender) show no significant effect on student success predictions. Even though (Windham et al., 2014) also found gender significant, it should not be considered a sole or direct indicator of student success because the effect of gender on academic success may vary from programme to programme and university to university (Cerdeira et al., 2018).

Although it already feels intuitive that universities and higher education institutes should have targeted interventions for students with low grades in the first semester of their course, this study suggests predicting the success of the student at the end of the first year. Hence, the students who do not have low grades in the first semester but are predicted to be unsuccessful at the end of the second semester may also be captured. Using the proposed early warning system, universities may take action to motivate and help the students at risk, who are predicted to be unsuccessful, to increase their academic performances. Moreover, the provided list of the most significant variables and the discussion about them may also help universities where they have difficulties in retrieving all the input variables.

The volume of the used dataset and the model performances obtained in this study are highly competitive when compared to the existing studies. However, the dataset is limited to a single university. In future research, the proposed early warning system may be extended to multiple institutes and students from all years rather than only freshmen. Additionally, further machine learning algorithms may be employed to broaden the performance comparisons for the models.

AUTHOR CONTRIBUTIONS

Cem Recai Çırak: Conceptualization; methodology; supervision; validation; visualization; writing – original draft; writing – review and editing. **Hakan Akıllı:** Data curation; formal analysis; resources; software; validation; writing – original draft. **Yeliz Ekinci:** Conceptualization; methodology; project administration; supervision; validation; writing – review and editing.

CONFLICT OF INTEREST STATEMENT

Authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from Istanbul Bilgi University. Restrictions apply to the availability of these data, which were used under license for this study. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Cem Recai Çırak https://orcid.org/0000-0001-6380-3669

REFERENCES

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, *79*, 101102. https://doi.org/10.1016/j.seps.2021.101102

Al-Badarenah, A., & Alsakran, J. (2016). An automated recommender system for course selection. *International Journal of Advanced Computer Science and Applications*, *7*(3), 166–175. https://doi.org/10.14569/IJACSA.2016.070323

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: A case study. *International Journal of Information and Education Technology*, *6*(7), 528–533. https://doi.org/10.7763/IJIET.2016.V6.745

Anderton, R. S., Evans, T., & Chivers, P. T. (2016). Predicting academic success of health science students for first year anatomy and physiology. *International Journal of Higher Education*, *5*(1), 250. https://doi.org/10.5430/ijhe.v5n1p250

Beattie, G., Laliberté, J.-W. P., & Oreopoulos, P. (2018). Thrivers and divers: Using non-academic measures to predict college success and failure. *Economics of Education Review*, *62*, 170–182. https://doi.org/10.1016/j.econedurev.2017.09.008

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, *60*(7), 1048–1064. https://doi.org/10.1007/s11162-019-09546-y

Beiter, R., Nash, R., McCrady, M., Rhoades, D., Linscomb, M., Clarahan, M., & Sammut, S. (2015). The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of Affective Disorders*, *173*, 90–96. https://doi.org/10.1016/j.jad.2014.10.054

Bowman, N. A., Jang, N., Kivlighan, D. M., Schneider, N., & Ye, X. (2020). The impact of a goal-setting intervention for engineering students on academic probation. *Research in Higher Education*, *61*(1), 142–166. https://doi.org/10.1007/s11162-019-09555-x

Buitrago-Florez, F., Sanchez, M., Pérez Romanello, V., Hernandez, C., & Hernández Hoyos, M. (2022). A systematic approach for curriculum redesign of introductory courses in engineering: A programming course case study. *Kybernetes*, *52*, 3904–3917. https://doi.org/10.1108/K-10-2021-0957

Cano, A., & Leonard, J. D. (2019). Interpretable Multiview early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies*, *12*(2), 198–211. https://doi.org/10.1109/TLT.2019.2911079

Cerdeira, J. M., Nunes, L. C., Reis, A. B., & Seabra, C. (2018). Predictors of student success in higher education: Secondary school internal scores versus national exams. *Higher Education Quarterly*, *72*(4), 304–313. https://doi.org/10.1111/hequ.12158

Cui, G., Leung, M. W., Zhang, G., & Li, L. (2008). Model selection for direct marketing: Performance criteria and validation methods. *Marketing Intelligence & Planning*, *26*(3), 275–292. https://doi.org/10.1108/02634500810871339

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In Proceedings of the 26th International Conference on World Wide Web Companion—WWW '17 Companion (pp. 415–421). https://doi.org/10.1145/3041021.3054164

Duman, E., Ekinci, Y., & Tanriverdi, A. (2012). Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Systems with Applications*, *39*(1), 48–53. https://doi.org/10.1016/j.eswa.2011.06.048

Dwivedi, S., & Roshni, V. S. K. (2017). Recommender system for big data in education. In 2017 5th National Conference on E-learning & E-learning technologies (ELELTECH) (pp. 1–6). https://doi.org/10.1109/ELELTECH.2017.8074993

Fong, K. E., Melguizo, T., & Prather, G. (2015). Increasing success rates in developmental math: The complementary role of individual and institutional characteristics. *Research in Higher Education*, *56*(7), 719–749. https://doi.org/10.1007/s11162-015-9368-9

Garcia, E. P. I., & Mora, P. M. (2011). Model prediction of academic performance for first year students. In 2011 10th Mexican International Conference on Artificial Intelligence (pp: 169–174). https://doi.org/10.1109/MICAI.2011.28

Glaesser, J., & Cooper, B. (2012). Gender, parental education, and ability: Their interacting roles in predicting GCSE success. *Cambridge Journal of Education*, *42*(4), 463–480. https://doi.org/10.1080/0305764X.2012.733346

Goyal, M., & Vohra, R. (2012). Applications of data mining in higher education. *International Journal of Computer Science*, *9*(2), 113–120.

Gray, C. C., & Perkins, D. (2018). Utilizing early engagement and machine learning to predict student outcomes. *Computers in Education*, *131*, 22–32. https://doi.org/10.1016/j.compedu.2018.12.006

Gupta, B., Goul, M., & Dinter, B. (2015). Business intelligence and big data in higher education: Status of a multi-year model curriculum development effort for business school undergraduates, MS graduates, and MBAs. *Communications of the Association for Information Systems*, *36*, 449–476. https://doi.org/10.17705/1CAIS.03623

Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S., & Sarker, K. U. (2018). Student academic performance prediction by using decision tree algorithm. In 2018 4th International Conference on Computer and Information Sciences (ICCOINS) (pp. 1–5). https://doi.org/10.1109/ICCOINS.2018.8510600

Jokhan, A., Sharma, B., & Singh, S. (2019). Early warning system as a predictor for student performance in higher education blended courses. *Studies in Higher Education*, *44*(11), 1900–1911. https://doi.org/10.1080/03075079.2018.1466872

Kardan, A. A., & Sadeghi, H. (2013). A decision support system for course offering in online higher education institutes. *International Journal of Computational Intelligence Systems*, *6*(5), 928. https://doi.org/10.1080/18756891.2013.808428

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Langan, A. M., Harris, W. E., Barrett, N., Hamshire, C., & Wibberley, C. (2018). Benchmarking factor selection and sensitivity: A case study with nursing courses. *Studies in Higher Education*, *43*(9), 1586–1596. https://doi.org/10.1080/03075079.2016.1266613

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

Lu, O., Huang, A., Huang, J., Lin, A., Ogata, H., & Yang, S. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology & Society*, *21*(2), 220–232.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Mason, C., Twomey, J., Wright, D., & Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. *Research in Higher Education*, *59*(3), 382–400. https://doi.org/10.1007/s11162-017-9473-z

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, 55462–55470. https://doi.org/10.1109/ACCESS.2020.2981905

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved random forest for classification. *IEEE Transactions on Image Processing*, *27*(8), 4012–4024. https://doi.org/10.1109/TIP.2018.2834830

Perez, B., Castellanos, C., & Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study. In 2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI) (pp: 1–6). https://doi.org/10.1109/ColCACI.2018.8484847

Polyzou, A., & Karypis, G. (2019). Feature extraction for next-term prediction of poor student performance. *IEEE Transactions on Learning Technologies*, *12*(2), 237–248. https://doi.org/10.1109/TLT.2019.2913358

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221–232. https://doi.org/10.22364/bjmc.2017.5.2.05

Rivas, A., González-Briones, A., Hernández, G., Prieto, J., & Chamoso, P. (2021). Artificial neural network analysis of the academic performance of students in virtual learning environments. *Neurocomputing*, *423*, 713–720. https://doi.org/10.1016/j.neucom.2020.02.125

Ruegg, R., Ruegg, R., Petersen, N., Hoang, H., & Marianne, M. (2021). Effects of pathways into university on the academic success of international undergraduate students. *Higher Education Research and Development*, *40*(6), 1283–1297. https://doi.org/10.1080/07294360.2020.1804336

Shcheglova, I., Gorbunova, E., & Chirikov, I. (2020). The role of the first-year experience in student attrition. *Quality in Higher Education*, *26*(3), 307–322. https://doi.org/10.1080/13538322.2020.1815285

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tang, B., Kay, S., & He, H. (2016). Toward optimal feature selection in naive Bayes for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2508–2521. https://doi.org/10.1109/TKDE.2016.2563436

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

van Zyl, A., Gravett, S., & de Bruin, G. (2016). To what extent do pre-entry attributes predict first year student academic performance in the South African context? *South African Journal of Higher Education*, *26*(5), 1095–1111. https://doi.org/10.20853/26-5-210

Vinoth Kumar, E. S., Balamurugan, S., & Sasi Kala, S. (2021). Multi-tier student performance evaluation model (MTSPEM) with integrated classification techniques for educational decision making. *International Journal of Computational Intelligence Systems*, *14*(1), 1796. https://doi.org/10.2991/ijcis.d.210609.001

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*, 106189. https://doi.org/10.1016/j.chb.2019.106189

Wait, I. W., & Gressel, J. W. (2009). Relationship between TOEFL score and academic success for international engineering students. *Journal of Engineering Education*, *98*(4), 389–398. https://doi.org/10.1002/j.2168-9830.2009.tb01035.x

Wang, T., Xiao, B., & Ma, W. (2022). Student behavior data analysis based on association rule mining. *International Journal of Computational Intelligence Systems*, *15*(1), 32. https://doi.org/10.1007/s44196-022-00087-4

Weatherton, M., & Schussler, E. E. (2021). Success for all? A call to re-examine how student success is defined in higher education. *CBE Life Sciences Education*, *20*(1), es3. https://doi.org/10.1187/cbe.20-09-0223

Windham, M. H., Rehfuss, M. C., Williams, C. R., Pugh, J. V., & Tincher-Ladner, L. (2014). Retention of first-year community college students. *Community College Journal of Research & Practice*, *38*(5), 466–477. https://doi.org/10.1080/10668926.2012.743867

Yakubu, M. N., & Abubakar, A. M. (2022). Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes*, *51*(2), 916–934. https://doi.org/10.1108/K-12-2020-0865

Žuljević, M. F., & Buljan, I. (2022). Academic and non-academic predictors of academic performance in medical school: An exploratory cohort study. *BMC Medical Education*, *22*(1), 366. https://doi.org/10.1186/s12909-022-03436-1

How to cite this article: Çırak, C. R., Akıllı, H., & Ekinci, Y. (2024). Development of an early warning system for higher education institutions by predicting first-year student academic performance. *Higher Education Quarterly*, *00*, e12539. https://doi.org/10.1111/hequ.12539