Smart Learning Environments

RESEARCH Open Access

Educational data mining: prediction of students' academic performance using machine learning algorithms

Mustafa Yağcı*

*Correspondence:
mustafayagci06@gmail.com
Kırşehir Ahi Evran University,
Faculty of Engineering
and Architecture,
40100 Kırşehir, Turkey

Abstract

Educational data mining has become an effective tool for exploring the hidden relationships in educational data and predicting students' academic achievements. This study proposes a new model based on machine learning algorithms to predict the final exam grades of undergraduate students, taking their midterm exam grades as the source data. The performances of the random forests, nearest neighbour, support vector machines, logistic regression, Naïve Bayes, and k-nearest neighbour algorithms, which are among the machine learning algorithms, were calculated and compared to predict the final exam grades of the students. The dataset consisted of the academic achievement grades of 1854 students who took the Turkish Language-I course in a state University in Turkey during the fall semester of 2019–2020. The results show that the proposed model achieved a classification accuracy of 70–75%. The predictions were made using only three types of parameters; midterm exam grades, Department data and Faculty data. Such data-driven studies are very important in terms of establishing a learning analysis framework in higher education and contributing to the decision-making processes. Finally, this study presents a contribution to the early prediction of students at high risk of failure and determines the most effective machine learning methods.

Keywords: Machine learning, Educational data mining, Predicting achievement, Learning analytics, Early warning systems

Introduction

The application of data mining methods in the field of education has attracted great attention in recent years. Data Mining (DM) is the discovery of data. It is the field of discovering new and potentially useful information or meaningful results from big data (Witten et al., 2011). It also aims to obtain new trends and new patterns from large datasets by using different classification algorithms (Baker & Inventado, 2014).

Educational data mining (EDM) is the use of traditional DM methods to solve problems related to education (Baker & Yacef, 2009; cited in Fernandes et al., 2019). EDM is the use of DM methods on educational data such as student information, educational records, exam results, student participation in class, and the frequency of students' asking questions. In recent years, EDM has become an effective tool used to identify hidden patterns in educational data, predict academic achievement, and improve the learning/teaching environment.

Learning analytics has gained a new dimension through the use of EDM (Waheed et al., 2020). Learning analytics covers the various aspects of collecting student information together, better understanding the learning environment by examining and analysing it, and revealing the best student/teacher performance (Long & Siemens, 2011). Learning analytics is the compilation, measurement and reporting of data about students and their contexts in order to understand and optimize learning and the environments in which it takes place. It also deals with the institutions developing new strategies.

Another dimension of learning analytics is predicting student academic performance, uncovering patterns of system access and navigational actions, and determining students who are potentially at risk of failing (Waheed et al., 2020). Learning management systems (LMS), student information systems (SIS), intelligent teaching systems (ITS), MOOCs, and other web-based education systems leave digital data that can be examined to evaluate students' possible behavior. Using EDM method, these data can be employed to analyse the activities of successful students and those who are at risk of failure, to develop corrective strategies based on student academic performance, and therefore to assist educators in the development of pedagogical methods (Casquero et al., 2016; Fidalgo-Blanco et al., 2015).

The data collected on educational processes offer new opportunities to improve the learning experience and to optimize users' interaction with technological platforms (Shorfuzzaman et al., 2019). The processing of educational data yields improvements in many areas such as predicting student behaviour, analytical learning, and new approaches to education policies (Capuano & Toti, 2019; Viberg et al., 2018). This comprehensive collection of data will not only allow education authorities to make data-based policies, but also form the basis of software to be developed with artificial intelligence on the learning process.

EDM enables educators to predict situations such as dropping out of school or less interest in the course, analyse internal factors affecting their performance, and make statistical techniques to predict students' academic performance. A variety of DM methods are employed to predict student performance, identify slow learners, and dropouts (Hardman et al., 2013; Kaur et al., 2015). Early prediction is a new phenomenon that includes assessment methods to support students by proposing appropriate corrective strategies and policies in this field (Waheed et al., 2020).

Especially during the pandemic period, learning management systems, quickly put into practice, have become an indispensable part of higher education. While students use these systems, the log records produced have become ever more accessible. (Macfadyen & Dawson, 2010; Kotsiantis et al., 2013; Saqr et al., 2017). Universities now should improve the capacity of using these data to predict academic success and ensure student progress (Bernacki et al., 2020).

As a result, EDM provides the educators with new information by discovering hidden patterns in educational data. Using this model, some aspects of the education system can be evaluated and improved to ensure the quality of education.

Literature

In various studies on EDM, e-learning systems have been successfully analysed (Lara et al., 2014). Some studies have also classified educational data (Chakraborty et al., 2016), while some have tried to predict student performance (Fernandes et al., 2019).

Asif et al. (2017) focused on two aspects of the performance of undergraduate students using DM methods. The first aspect is to predict the academic achievements of students at the end of a four-year study program. The second one is to examine the development of students and combine them with predictive results. He divided the students into two parts as low achievement and high achievement groups. He have found that it is important for the educators to focus on a small number of courses indicating particularly good or poor performance in order to offer timely warnings, support underperforming students and offer advice and opportunities to high-performing students. Cruz-Jesus et al. (2020) predicted student academic performance with 16 demographics such as age, gender, class attendance, internet access, computer possession, and the number of courses taken. Random forest, logistic regression, k-nearest neighbours and support vector machines, which are among the machine learning methods, were able to predict students' performance with accuracy ranging from 50 to 81%.

Fernandes et al. (2019) developed a model with the demographic characteristics of the students and the achievement grades obtained from the in-term activities. In that study, students' academic achievement was predicted with classification models based on Gradient Boosting Machine (GBM). The results showed that the best qualities for estimating achievement scores were the previous year's achievement scores and unattendance. The authors found that demographic characteristics such as neighbourhood, school and age information were also potential indicators of success or failure. In addition, he argued that this model could guide the development of new policies to prevent failure. Similarly, by using the student data requested during registration and environmental factors, Hoffait and Schyns (2017) determined the students with the potential to fail. He found that students with potential difficulties could be classified more precisely by using DM methods. Moreover, their approach makes it possible to rank the students by levels of risk. Rebai et al. (2020) proposed a machine learning-based model to identify the key factors affecting academic performance of schools and to determine the relationship between these factors. He concluded that the regression trees showed that the most important factors associated with higher performance were school size, competition, class size, parental pressure, and gender proportions. In addition, according to the random forest algorithm results, the school size and the percentage of girls had a powerful impact on the predictive accuracy of the model.

Ahmad and Shahzadi, (2018) proposed a machine learning-based model to find an answer to the question whether students were at risk regarding their academic performance. Using the students' learning skills, study habits, and academic interaction features, they made a prediction with a classification accuracy of 85%. The researchers concluded that the model they proposed could be used to determine academically unsuccessful student. Musso et al., (2020) proposed a machine learning model based on learning strategies, perception of social support, motivation, socio-demographics, health condition, and academic performance characteristics. With this model, he predicted the academic performance and dropouts. He concluded that the predictive variable with the highest effect on predicting GPA was learning strategies while the variable with the greatest effect on determining dropouts was background information.

Waheed et al., (2020) designed a model with artificial neural networks on students' records related to their navigation through the LMS. The results showed that demographics and student clickstream activities had a significant impact on student performance. Students who navigated through courses performed higher. Students' participation in the learning environment had nothing to do with their performance. However, he concluded that the deep learning model could be an important tool in the early prediction of student performance. Xu et al. (2019) determined the relationship between the internet usage behaviors of university students and their academic performance and he predicted students' performance with machine learning methods. The model he proposed predicted students' academic performance at a high level of accuracy. The results suggested that Internet connection frequency features were positively correlated with academic performance, whereas Internet traffic volume features were negatively correlated with academic performance. In addition, he concluded that internet usage features had an important role on students' academic performance. Bernacki et al. (2020) tried to find out whether the log records in the learning management system alone would be sufficient to predict achievement. He concluded that the behaviour-based prediction model successfully predicted 75% of those who would need to repeat a course. He also stated that, with this model, students who might be unsuccessful in the subsequent semesters could be identified and supported. Burgos et al. (2018) predicted the achievement grades that the students might get in the subsequent semesters and designed a tool for students who were likely to fail. He found that the number of unsuccessful students decreased by 14% compared to previous years. A comparative analysis of studies predicting the academic achievement grades using machine learning methods is given in Table 1.

A review of previous research that aimed to predict academic achievement indicates that researchers have applied a range of machine learning algorithms, including multiple, probit and logistic regression, neural networks, and C4.5 and J48 decision trees. However, random forests (Zabriskie et al., 2019), genetic programming (Xing et al., 2015), and Naive Bayes algorithms (Ornelas & Ordonez, 2017) were used in recent studies. The prediction accuracy of these models reaches very high levels.

Prediction accuracy of student academic performance requires an deep understanding of the factors and features that impact student results and the achievement of student (Alshanqiti & Namoun, 2020). For this purpose, Hellas et al. (2018) reviewed 357 articles on student performance detailing the impact of 29 features. These features were mainly related to psychomotor skills such as course and pre-course performance, student participation, student demographics such as gender, high school performance, and self-regulation. However, the dropout rates were mainly influenced by student motivation, habits, social and financial issues, lack of progress, and career transitions.

The literature review suggests that, it is a necessity to improve the quality of education by predicting the academic performance of the students and supporting those who are in the risk group. In the literature, the prediction of academic performance was made with many and various variables, various digital traces left by students on the internet (browsing, lesson time, percentage of participation) (Fernandes et al., 2019; Rubin et al., 2010; Waheed et al., 2020; Xu et al., 2019) and students demographic characteristics

Table 1 Comparative analysis

References	Variables	Objectives	Level	Dataset	Algorithms	Accuracy	
						Min	Max
Asif et al. (2017)	The marks for all the courses that are taught in the four years of the degree programme	Predicting students' performance	Undergraduate students	210	DT, 1-NN, NB, NN, RF	NN (62.50%)	NB (83.65%)
Cruz-Jesus et al. (2020)	Year of the study cycle, gender, age, number of enrolled years in high school, scholarship, internet access, class size, school size, economic level, population density, number of unit courses attended	Predicting students' performance	High schools students	110627	ANN, DT, ET, RF, SVM, kNN, LR	LR (81.1%)	SVM (51.2%)
Fernandes et al. (2019)	Class with persons with special needs, Classroom usage environment, Gender, age (mean), Student benefit, city, neighbourhood, Student with special needs, Grade (mean), Absence (mean)	Predict academic outcomes of student performance	High schools students	Dataset1:19000 Dataset2:19834	Gradient Boosting Machine	89.5%	91.9%
Hoffait and Schyns (2017)	Gender, Nationality, Studies, Prior schooling, math, scholarship, success	Predicting students at high risk of failure	secondary school students	2244	RF, LR, ANN	ANN (70.4%)	RF (90%)
Rebai et al. (2020)	Socioeconomic status, school type, school location, competition, teacher characteristic (experience, salary), class size, school size, gender, parental education, political context, parental pressure	to identify the key factors that impact schools' academic performance and to explore their relationships	Secondary schools	105 schools	RT, RF		

Table 1 (continued)

References	Variables	Objectives	Level	Dataset	Algorithms	Accuracy	
						Min	Max
Ahmad and Shahzadi (2018)	Previous degree marks, Home environment, Study habits Learning skills, Hardworking and Academic interaction	Identification of students in the risk group	Undergraduate students	300	MPNN		95%
Musso et al., (2020)	Learning strategies, coping strategies, cognitive factors, social support, background, self-concept, self-satisfaction, use of IT and reading	Grade point average, academic retention, and degree completion	Undergraduate students	655	ANN	60.5%	80.7%
Waheed et al., (2020)	Students' demographics, clickstream events	Pass-fail, withdrawn-pass, distinction-fail, distinction-pass	Undergraduate students	32593	ANN, SVM, LR	84%	93%
Xu et al. (2019)	Internet usage behaviours comprise online time, internet connection frequency, internet traffic volume, and online time	Predicting students' performance	Undergraduate students	4000	DT, NN, SVM	71%	76%
Bernacki et al. (2020)	Log records in the learning management system	Predict achievement	Undergradeate students	337	LR, NB, J-48 DT, J-Rip DT	J-48 (53.71%)	LR (67.36%)
Burgos et al. (2018)	Historical student course grade data	Drop out of a course	Undergradeate students	100	SVM, FFNN, PESFAM, LOGIT_Act	SVM (62.50)	LOGIT_Act(97.13%)

(gender, age, economic status, number of courses attended, internet access, etc.) (Bernacki et al., 2020; Rizvi et al., 2019; García-González & Skrita, 2019; Rebai et al., 2020; Cruz-Jesus et al., 2020; Aydemir, 2017), learning skills, study approaches, study habits (Ahmad & Shahzadi, 2018), learning strategies, social support perception, motivation, socio-demography, health form, academic performance characteristics (Costa-Mendes et al., 2020; Gök, 2017; Kılınç, 2015; Musso et al., 2020), homework, projects, quizzes (Kardaş & Güvenir, 2020), etc. In almost all models developed in such studies, prediction accuracy is ranging from 70 to 95%. Hovewer, collecting and processing such a variety of data both takes a lot of time and requires expert knowledge. Similarly, Hoffait and Schyns (2017) suggested that collecting so many data is difficult and socio-economic data are unnecessary. Moreover, these demographic or socio-economic data may not always give the right idea of preventing failure (Bernacki et al., 2020).

The study concerns predicting students' academic achievement using grades only, no demographic characteristics and no socio-economic data. This study aimed to develop a new model based on machine learning algorithms to predict the final exam grades of undergraduate students taking their midterm exam grades, Faculty and Department of the students.

For this purpose, classification algorithms with the highest performance in predicting students' academic achievement were determined by using machine learning classification algorithms. The reason for choosing the Turkish Language-I course was that it is a compulsory course that all students enrolled in the university must take. Using this model, students' final exam grades were predicted. These models will enable the development of pedagogical interventions and new policies to improve students' academic performance. In this way, the number of potentially unsuccessful students can be reduced following the assessments made after each midterm.

Method

This section describes the details of the dataset, pre-processing techniques, and machine learning algorithms employed in this study.

Dataset

Educational institutions regularly store all data that are available about students in electronic medium. Data are stored in databases for processing. These data can be of many types and volumes, from students' demographics to their academic achievements. In this study, the data were taken from the Student Information System (SIS), where all student records are stored at a State University in Turkey. In these records, the midterm exam grades, final exam grades, Faculty, and Department of 1854 students who have taken the Turkish Language-I course in the 2019–2020 fall semester were selected as the dataset. Table 2 shows the distribution of students according to the academic unit. Moreover, as a additional file 1 the dataset are presented.

Midterm and final exam grades are ranging from 0 to 100. In this system, the end-of-semester achievement grade is calculated by taking 40% of the midterm exam and 60% of the final exam. Students with achievement grade below 60 are unsuccessful and those above 60 are successful. The midterm exam is usually held in the middle of the academic semester and the final exam is held at the end of the semester. There are approximately

Table 2 The dataset

Academic unit	Number of Students
Faculty of Education	404
Faculty of Arts and Sciences	319
Faculty of Health Sciences	296
Faculty of Economics and Administrative Sciences	221
School of Physical Education and Sports	192
Faculty of Engineering and Architecture	116
School of Physical Therapy and Rehabilitation	92
Faculty of Islamic Sciences	88
Faculty of Agriculture	68
Faculty of Fine Arts	30
Vocational School of Applied Sciences	28
Total Number of Students	1854

9 weeks (2.5 months) from the midterm exam to the final exam. In other words, there is a two and a half month period for corrective actions for students who are at risk of failing thanks to the final exam predictions made. In other words, the answer to the question of how effective the student's performance in the middle of the semester is on his performance at the end of the semester was investigated.

Data identification and collection

At this phase, it is determined from which source the data will be stored, which features of the data will be used, and whether the collected data is suitable for the purpose. Feature selection involves decreasing the number of variables used to predict a particular outcome. The goal; to facilitate the interpretability of the model, reduce complexity, increase the computational efficiency of algorithms, and avoid overfitting.

Establishing DM model and implementation of algorithm

RF, NN, LR, SVM, NB and kNN were employed to predict students' academic performance. The prediction accuracy was evaluated using tenfold cross validation. The DM process serves two main purposes. The first purpose is to make predictions by analyzing the data in the database (predictive model). The second one is to describe behaviors (descriptive model). In predictive models, a model is created by using data with known results. Then, using this model, the result values are predicted for datasets whose results are unknown. In descriptive models, the patterns in the existing data are defined to make decisions.

When the focus is on analysing the causes of success or failure, statistical methods such as logistic regression and time series can be employed (Ortiz & Dehon, 2008; Arias Ortiz & Dehon, 2013). However, when the focus is on forecasting, neural networks (Delen, 2010; Vandamme et al., 2007), support vector machines (Huang & Fang, 2013), decision trees (Delen, 2011; Nandeshwar et al., 2011) and random forests (Delen, 2010; Vandamme et al., 2007) is more efficient and give more accurate results. Statistical techniques are to create a model that can successfully predict output values based on available input data. On the other hand, machine learning methods automatically create a model that matches the input data with the expected target values when a supervised optimization problem is given.

The performance of the model was measured by confusion matrix indicators. It is understood from the literature that there is no single classifier that works best for prediction results. Therefore, it is necessary to investigate which classifiers are more studied for the analysed data (Asif et al., 2017).

Experiments and results

The entire experimental phase was performed with Orange machine learning software. Orange is a powerful and easy-to-use component-based DM programming tool for expert data scientists as well as for data science beginners. In Orange, data analysis is done by stacking widgets into workflows. Each widget includes some data retrieval, data pre-processing, visualization, modelling, or evaluation task. A workflow is a series of actions or actions that will be performed on the platform to perform a specific task. Comprehensive data analysis charts can be created by combining different components in a workflow. Figure 1 shows the workflow diagram designed.

The dataset included midterm exam grades, final exam grades, Faculty, and Department of 1854 students taking the Turkish Language-I course in the 2019–2020 Fall Semester. The entire dataset is provided as Additional file 1. Table 3 shows part of the dataset.

In the dataset, students' midterm exam grades, final exam grades, faculty, and department information were determined as features. Each measure contains data associated with a student. Midterm exam and final exam grade variables were explained under the heading "dataset". The faculty variable represents Faculties in Kırşehir Ahi Evran University and the department variable represents departments in faculties. In the development of the model, the midterm, the faculty, and the department information were determined as the independent variable and the final was determined as the dependent variable. Table 4 shows the variable model.

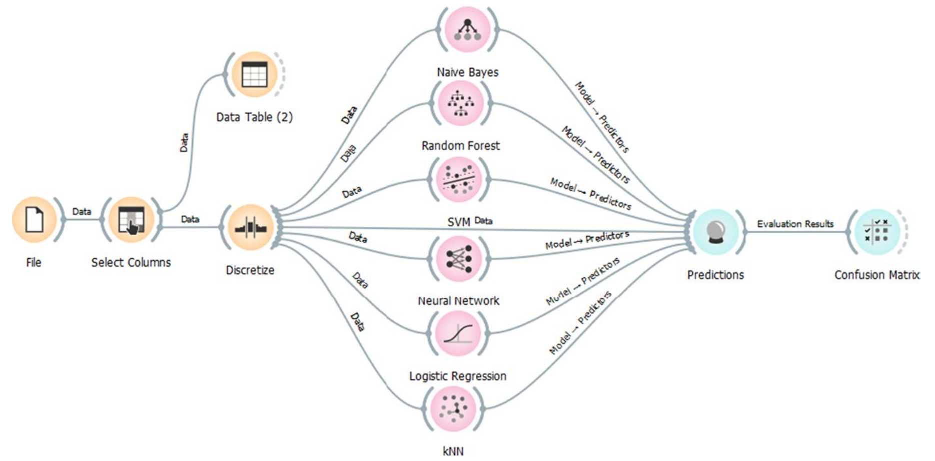

Fig. 1 The workflow of the designed model

Table 3 Part of the dataset consist of 1854 rows

stdID	Midterm	Final	Faculty	Department
std1	60	68	Faculty of Economics and Administrative Sciences	Political Science and Public Administration
std2	34	67	School of Physical Education and Sports	Coaching Education
std3	25	75	Faculty of Education	Computer Education and Instructional Technology
std4	50	66	Faculty of Education	Social Sciences Teaching
std5	50	66	Faculty of Education	Early Childhood Education
std6	88	72	Faculty of Education	Garden Plants
std7	45	37	School of Physical Education and Sports	Physical Education and Sports Teaching
std8	52	50	School of Physical Education and Sports	Coaching Education
…	…	…	…	…
std1853	88	88	School of Physical Therapy and Rehabilitation	Physiotherapy and Rehabilitation
std1854	84	96	School of Physical Therapy and Rehabilitation	Physiotherapy and Rehabilitation

After the variable model was determined, the midterm exam grades and final exam grades were categorized according to the equal-width discretization model. Table 5 shows the criteria used in converting midterm exam grades and final exam grades into the categorical format.

In Table 6, the values in the final column are the actual values. The values in the RF, SVM, LR, KNN, NB, and NN columns are the values predicted by the proposed model. For example, according to Table 5, std1's actual final grade was in the range 55 to 77. While the predicted value of the RF, SVM, LR, NB, and NN models were in the range of, the predicted value of the kNN model was greater than 77.

Evaluation of the model performance

The performance of model was evaluated with confusion matrix, classification accuracy (CA), precision, recall, f-score (F1), and area under roc curve (AUC) metrics.

Table 4 The model of variables

Features	Target variable	Meta Attributes
Midterm	Final	stdID
Faculty		
Department		

Table 5 Categorical criteria

Category	Criteria
1	grade < 32.5
2	32.5 < = grade < 55
3	55 < = grade < 77.5
4	grade > = 77.5

Table 6 Probabilities and final decisions of predictive models (RF, LR, SVM, kNN, NB, NN)

stdID	RF	SVM	LR	kNN	NB	NN	final	midterm	faculty	Department
st1	55–77.5	55–77.5	55–77.5	≥ 77.5	55–77.5	55–77.5	**55–77.5**	< 32.5	Faculty of Education	Computer Education and Instructional Technology
st2	55–77.5	≥ 77.5	≥ 77.5	55–77.5	55–77.5	55–77.5	**55–77.5**	32.5–55	Faculty of Education	Social Sciences Teaching
st3	55–77.5	55–77.5	55–77.5	55–77.5	55–77.5	55–77.5	**55–77.5**	32.5–55	Faculty of Education	Early Childhood Education
st4	55–77.5	55–77.5	55–77.5	55–77.5	55–77.5	55–77.5	**55–77.5**	≥ 77.5	Faculty of Agriculture	Garden Plants
st5	55–77.5	55–77.5	55–77.5	< 32.5	55–77.5	55–77.5	**32.5–55**	32.5–55	School of Physical Education and Sports	Physical Education and Sports Teaching
st6	32.5–55	55–77.5	55–77.5	55–77.5	32.5–55	32.5–55	**32.5–55**	32.5–55	School of Physical Education and Sports	Coaching Education
st7	55–77.5	55–77.5	≥ 77.5	55–77.5	< 32.5	< 32.5	**< 32.5**	< 32.5	Faculty of Education	Social Sciences Teaching
st8	55–77.5	55–77.5	55–77.5	≥ 77.5	55–77.5	55–77.5	**55–77.5**	55–77.5	Faculty of Education	Psychological Counseling and Guidance
st9	< 32.5	32.5–55	≥ 77.5	< 32.5	< 32.5	32.5–55	**≥ 77.5**	< 32.5	Faculty of Education	Primary Education
st10	55–77.5	55–77.5	55–77.5	55–77.5	55–77.5	55–77.5	**55–77.5**	32.5–55	Faculty of Arts and Sciences	Archaeology

Table 7 The Confusion matrix

		Predicted	
		Positive (1)	**Negative (0)**
Actual	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 8 Confusion matrix of the RF algorithm

		Predicted				
		< 32.5	**32.5–55**	**55–77.5**	**≥ 77.5**	**Sum**
Actual	< 32.5	**60%**	3.8%	1.2%	0.6%	38
	32.5–55	26.7%	**65.4%**	9.5%	0.8%	154
	55–77.5	10.0%	30.8%	**71.2%**	13.6%	1016
	≥ 77.5	3.3%	0.0%	18.1%	**84.9%**	646
	Sum	30	26	1320	478	1854

Confusion matrix

The confusion matrix shows the current situation in the dataset and the number of correct/incorrect predictions of the model. Table 7 shows the confusion matrix. The performance of the model is calculated by the number of correctly classified instances and incorrectly classified instances. The rows show the real numbers of the samples in the test set, and the columns represent the estimation of the model.

In Table 6, true positive (TP) and true negative (TN) show the number of correctly classified instances. False positive (FP) shows the number of instances predicted as 1 (positive) while it should be in the 0 (negative) class. False negative (FN) shows the number of instances predicted as 0 (negative) while it should be in class 1 (positive).

Table 8 shows the confusion matrix for the RF algorithm. In the confusion matrix of 4×4 dimensions, the main diagonal shows the percentage of correctly predicted instances, and the matrix elements other than the main diagonal shows the percentage of errors predicted.

Table 8 shows that 84.9% of those with the actual final grade greater than 77.5, 71.2% of those with range 55–77.5, 65.4% of those with range 32.5–55, and 60% of those with less than 32.5 were predicted correctly. Confusion matrixs of other algorithms are shown in Tables 9, 10, 11, 12, and 13.

Classification accuracy: CA is the ratio of the correct predictions (TP + TN) to the total number of instances (TP + TN + FP + FN).

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

Table 9 Confusion matrix of the NN algorithm

		Predicted				
		< 32.5	**32.5–55**	**55–77.5**	**≥ 77.5**	**Sum**
Actual	< 32.5	**64%**	9.7%	1.2%	0.6%	38
	32.5–55	24%	**61.3%**	9.6%	1.0%	154
	55–77.5	12.0%	25.8%	**71.8%**	14.9%	1016
	≥ 77.5	0.0%	3.2%	17.4%	**83.5%**	646
	Sum	25	31	1296	502	1854

Table 10 Confusion matrix of the SVM algorithm

		Predicted				
		< 32.5	**32.5–55**	**55–77.5**	**≥ 77.5**	**Sum**
Actual	< 32.5	**68.8%**	14.3%	1.6%	0.6%	38
	32.5–55	31.2%	**52.4%**	9.9%	0.9%	154
	55–77.5	0.0%	14.3%	**70.1%**	14.3%	1016
	≥ 77.5	0.0%	19.0%	18.4%	**84.2%**	646
	Sum	16	21	1349	468	1854

Table 11 Confusion matrix of the LR algorithm

		Predicted				
		< 32.5	**32.5–55**	**55–77.5**	**≥ 77.5**	**Sum**
Actual	< 32.5	**56.0%**	8.3%	1.5%	0.8%	38
	32.5–55	24.0%	**41.7%**	10.3%	1.7%	154
	55–77.5	4.0%	25.0%	**70.0%**	20.1%	1016
	≥ 77.5	16.0%	25.0%	18.1%	**77.4%**	646
	Sum	25	12	1295	522	1854

Table 12 Confusion matrix of the NB algorithm

		Predicted				
		< 32.5	**32.5–55**	**55–77.5**	**≥ 77.5**	**Sum**
Actual	< 32.5	**40.0%**	9.5%	0.9%	0.0%	38
	32.5–55	18.2%	**42.9%**	9.4%	1.2%	154
	55–77.5	18.2%	42.9%	**70.4%**	19.3%	1016
	≥ 77.5	23.6%	4.8%	19.2%	**79.5%**	646
	Sum	55	42	1270	487	1854

Table 13 Confusion matrix of the kNN algorithm

		Predicted				
		< 32.5	**32.5–55**	**55–77.5**	**≥ 77.5**	**Sum**
Actual	< 32.5	**50.0%**	2.6%	1.1%	0.5%	38
	32.5–55	30.0%	**31.3%**	8.9%	1.5%	154
	55–77.5	15.0%	55.7%	**72.9%**	24.9%	1016
	≥ 77.5	5.0%	10.4%	17.1%	**73.1%**	646
	Sum	40	115	1089	610	1854

Precision: Precision is the ratio of the number of positive instances that are correctly classified to the total number of instances that are predicted positive. Gets a value in the range [0.1].

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall i**s** the ratio of the correctly classified number of positive instances to the number of all instances whose actual class is positive. The Recall is also called the true positive rate. Gets a value in the range [0.1].

$$Recall = \frac{TP}{TP + FN}$$

F-Criterion (F1): There is an opposite relationship between precision and recall. Therefore, the harmonic mean of both criteria is calculated for more accurate and sensitive results. This is called the F-criterion.

$$F\text{-}Criterion = \frac{2 \times Duyarlilik \times Kesinlik}{Duyarlilik + Kesinlik}$$

Receiver operating characteristics (ROC) curve

The AUC-ROC curve is used to evaluate the performance of a classification problem. AUC-ROC is a widely used metric to evaluate the performance of machine learning algorithms, especially in cases where there are unbalanced datasets, and explains how well the model is at predicting.

AUC: Area under the ROC curve

The larger the area covered, the better the machine learning algorithms at distinguishing given classes. AUC for the ideal value is 1. The AUC, Classification Accuracy (CA), F-Criterion (F1), precision, and recall values of the models are shown in Table 14.

The AUC value of RF, NN, SVM, LR, NB, and kNN algorithms were 0.860, 0.863, 0.804, 0.826, 0.810, and 0.810 respectively. The classification accuracy of the RF, NN, SVM, LR, NB, and kNN algorithms were also 0.746, 0.746, 0.735, 0.717, 0.713, and 0,699 respectively. According to these findings, for example, the RF algorithm was able to achieve 74.6% accuracy. In other words, there was a very high-level correlation between the

Table 14 AUC, CA, F1, precision and recall values of the models

Model	(AUC)	Classification accuracy (CA)	F1	Precision	Recall
Random Forest	0.860	0.746	0.721	0.752	0.746
Neural Network	0.863	0.746	0.723	0.748	0.746
SVM	0.804	0.735	0.704	0.735	0.735
Logistic Regression	0.826	0.717	0.685	0.700	0.717
Naïve Bayes	0.810	0.713	0.692	0.706	0.713
kNN	0.810	0.699	0.694	0.691	0.699

data predicted and the actual data. As a result, 74.6% of the samples were been classified correctly.

Discussion and conclusion

This study proposes a new model based on machine learning algorithms to predict the final exam grades of undergraduate students, taking their midterm exam grades as the source data. The performances of the Random Forests, nearest neighbour, support vector machines, Logistic Regression, Naïve Bayes, and k-nearest neighbour algorithms, which are among the machine learning algorithms, were calculated and compared to predict the final exam grades of the students. This study focused on two parameters. The first parameter was the prediction of academic performance based on previous achievement grades. The second one was the comparison of performance indicators of machine learning algorithms.

The results show that the proposed model achieved a classification accuracy of 70–75%. According to this result, it can be said that students' midterm exam grades are an important predictor to be used in predicting their final exam grades. RF, NN, SVM, LR, NB, and kNN are algorithms with a very high accuracy rate that can be used to predict students' final exam grades. Furthermore, the predictions were made using only three types of parameters; midterm exam grades, Department data and Faculty data. The results of this study were compared with the studies that predicted the academic achievement grades of the students with various demographic and socio-economic variables. Hoffait and Schyns (2017) proposed a model that uses the academic achievement of students in previous years. With this model, they predicted students' performance to be successful in the courses they will take in the new semester. They found that 12.2% of the students had a very high risk of failure, with a 90% confidence rate. Waheed et al. (2020) predicted the achievement of the students with demographic and geographic characteristics. He found that it has a significant effect on students' academic performance. He predicted the failure or success of the students by 85% accuracy. Xu et al. (2019) found that internet usage data can distinguish and predict students' academic performance. Costa-Mendes et al. (2020), Cruz-Jesus et al. (2020), Costa-Mendes et al. (2020) predicted the academic achievement of students in the light of income, age, employment, cultural level indicators, place of residence, and socio-economic information. Similarly, Babić (2017) predicted students' performance with an accuracy of 65% to 100% with artificial neural networks, classification tree, and support vector machines methods.

Another result of this study was RF, NN and SVM algorithms have the highest classification accuracy, while kNN has the lowest classification accuracy. According to this result, it can be said that RF, NN and SVM algorithms perform with more accurate results in predicting the academic achievement grades of students with machine learning algorithms. The results were compared with the results of the research in which machine learning algorithms were employed to predict academic performance according to various variables. For example, Hoffait and Schyns (2017) compared the performances of LR, ANN and RF algorithms to identify students at high risk of academic failure on their various demographic characteristics. They ranked the algorithms from those with the highest accuracy to the ones with the lowest accuracy as LR, ANN, and RF. On the other hand, Waheed et al. (2020) found that the SVM algorithm performed higher than the LR algorithm. According to Xu et al. (2019), the algorithm with the highest performance is SVM, followed by the NN algorithm, and the decision tree is the algorithm with the lowest performance.

The proposed model predicted the final exam grades of students with 73% accuracy. According to this result, it can be said that academic achievement can be predicted with this model in the future. By predicting students' achievement grades in future, students can be allowed to review their working methods and improve their performance. The importance of the proposed method can be better understood, considering that there is approximately 2.5 months between the midterm exams and the final exams in higher education. Similarly, Bernacki et al. (2020) work on the early warning model. He proposed a model to predict the academic achievements of students using their behavior data in the learning management system before the first exam. His algorithm correctly identified 75% of students who failed to earn the grade of B or better needed to advance to the next course. Ahmad and Shahzadi (2018) predicted students at risk for academic performance with 85% accuracy evaluating their study habits, learning skills, and academic interaction features. Cruz-Jesus et al. (2020) predicted students' end-of-semester grades with 16 independent variables. He concluded that students could be given the opportunity of early intervention.

As a result, students' academic performances were predicted using different predictors, different algorithms and different approaches. The results confirm that machine learning algorithms can be used to predict students' academic performance. More importantly, the prediction was made only with the parameters of midterm grade, faculty and department. Teaching staff can benefit from the results of this research in the early recognition of students who have below or above average academic motivation. Later, for example, as Babić (2017) points out, they can match students with below-average academic motivation by students with above-average academic motivation and encourage them to work in groups or project work. In this way, the students' motivation can be improved, and their active participation in learning can be ensured. In addition, such data-driven studies should assist higher education in establishing a learning analytics framework and contribute to decision-making processes.

Future research can be conducted by including other parameters as input variables and adding other machine learning algorithms to the modelling process. In addition, it is necessary to harness the effectiveness of DM methods to investigate students' learning behaviors, address their problems, optimize the educational environment, and enable data-driven decision making.

Abbreviations

EDM: Educational data mining; RF: Random forests; NN: Neural networks; SVM: Support vector machines; LR: Logistic regression; NB: Naïve Bayes; kNN: K-nearest neighbour; DT: Decision trees; ANN: Artificial neural networks; ERT: Extremely randomized trees; RT: Regression trees; MPNN: Multilayer perceptron neural network; FFNN: Feed-forward neural network; PESFAM: Adaptive resonance theory mapping; LMS: Learning management systems; SIS: Student information systems; ITS: Intelligent teaching systems; CA: Classification accuracy; F1: F-score; AUC: Area under roc curve; TP: True positive; TN: True negative; FP: False positive; FN: False negative; ROC: Receiver operating characteristics.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40561-022-00192-z.

Additional file 1: Dataset.

Acknowledgements

Not applicable.

Authors' contributions

All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 15 November 2021 Accepted: 15 February 2022
Published online: 03 March 2022

References

Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. *Bulletin of Education and Research, 40*(3), 157–164.

Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. *IEEE Access, 8*, 203827–203844. https://doi.org/10.1109/access.2020.3036572

Arias Ortiz, E., & Dehon, C. (2013). Roads to success in the Belgian French Community's higher education system: predictors of dropout and degree completion at the Université Libre de Bruxelles. *Research in Higher Education, 54*(6), 693–723. https://doi.org/10.1007/s11162-013-9290-y

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education, 113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Aydemir, B. (2017). *Predicting academic success of vocational high school students using data mining methods graduate*. [Unpublished master's thesis]. Pamukkale University Institute of Science.

Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. *Croatian Operational Research Review, 8*(2), 443–461. https://doi.org/10.17535/crorr.2017.0028

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. *Learning analytics* (pp. 61–75). Springer.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education, 158*(August), 103999. https://doi.org/10.1016/j.compedu.2020.103999

Burgos, C., Campanario, M. L., De, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering, 66*(2018), 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005

Capuano, N., & Toti, D. (2019). Experimentation of a smart learning system for law based on knowledge discovery and cognitive computing. *Computers in Human Behavior, 92*, 459–467. https://doi.org/10.1016/j.chb.2018.03.034

Casquero, O., Ovelar, R., Romo, J., Benito, M., & Alberdi, M. (2016). Students' personal networks in virtual and personal learning environments: A case study in higher education using learning analytics approach. *Interactive Learning Environments, 24*(1), 49–67. https://doi.org/10.1080/10494820.2013.817441

Chakraborty, B., Chakma, K., & Mukherjee, A. (2016). A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory. In *Proceedings of 2nd IEEE international conference on engineering and technology, ICETECH 2016, March* (pp. 431–436). https://doi.org/10.1109/ICETECH.2016.7569290

Costa-Mendes, R., Oliveira, T., Castelli, M., & Cruz-Jesus, F. (2020). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Education and Information Technologies, 26*, 1527–1547. https://doi.org/10.1007/s10639-020-10316-y

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon*. https://doi.org/10.1016/j.heliyon.2020.e04081

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory and Practice, 13*(1), 17–35. https://doi.org/10.2190/CS.13.1.b

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining : Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94*(February 2018), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Fidalgo-Blanco, Á., Sein-Echaluce, M. L., García-Peñalvo, F. J., & Conde, M. Á. (2015). Using Learning Analytics to improve teamwork assessment. *Computers in Human Behavior, 47*, 149–156. https://doi.org/10.1016/j.chb.2014.11.050

García-González, J. D., & Skrita, A. (2019). Predicting academic performance based on students' family environment: Evidence for Colombia using classification trees. *Psychology, Society and Education, 11*(3), 299–311. https://doi.org/10.25115/psye.v11i3.2056

Gök, M. (2017). Predicting academic achievement with machine learning methods. *Gazi University Journal of Science Part c: Design and Technology, 5*(3), 139–148.

Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm. *Systems Research and Behavioral Science, 30*(2), 194–203. https://doi.org/10.1002/sres.2130

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S.N. (2018). Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175–199).

Hoffait, A., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems, 101*(2017), 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education, 61*(1), 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

Kardaş, K., & Güvenir, A. (2020). Analysis of the effects of Quizzes, homeworks and projects on final exam with different machine learning techniques. *EMO Journal of Scientific, 10*(1), 22–29.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science, 57*, 500–508. https://doi.org/10.1016/j.procs.2015.07.372

Kılınç, Ç. (2015). *Examining the effects on university student success by data mining techniques.* [Unpublished master's thesis]. Eskişehir Osmangazi University Institute of Science.

Kotsiantis, S., Tselios, N., Filippidi, A., & Komis, V. (2013). Using learning analytics to identify successful learners in a blended learning course. *International Journal of Technology Enhanced Learning, 5*(2), 133–150. https://doi.org/10.1504/IJTEL.2013.059088

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European Higher Education Area—Application to student data from Open University of Madrid, UDIMA. *Computers and Education, 72*, 23–36. https://doi.org/10.1016/j.compedu.2013.10.009

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 31–40.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. *Higher Education, 80*(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications, 38*(12), 14984–14996. https://doi.org/10.1016/j.eswa.2011.05.048

Ornelas, F., & Ordonez, C. (2017). Predicting student success: A naïve bayesian application to community college data. *Technology, Knowledge and Learning, 22*(3), 299–315. https://doi.org/10.1007/s10758-017-9334-z

Ortiz, E. A., & Dehon, C. (2008). What are the factors of success at University? A case study in Belgium. *Cesifo Economic Studies, 54*(2), 121–148. https://doi.org/10.1093/cesifo/ifn012

Rebai, S., Ben Yahia, F., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences, 70*(August 2018), 100724. https://doi.org/10.1016/j.seps.2019.06.009

Rizvi, S., Rienties, B., & Ahmed, S. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education, 137*(August 2018), 32–47. https://doi.org/10.1016/j.compedu.2019.04.001

Rubin, B., Fernandes, R., Avgerinou, M. D., & Moore, J. (2010). The effect of learning management systems on student and faculty outcomes. *The Internet and Higher Education, 13*(1–2), 82–83. https://doi.org/10.1016/j.iheduc.2009.10.008

Saqr, M., Fors, U., & Tedre, M. (2017). How learning analytics can early predict under-achieving students in a blended medical education course. *Medical Teacher, 39*(7), 757–767. https://doi.org/10.1080/0142159X.2017.1309376

Shorfuzzaman, M., Hossain, M. S., Nazir, A., Muhammad, G., & Alamri, A. (2019). Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. *Computers in Human Behavior, 92*(February 2017), 578–588. https://doi.org/10.1016/j.chb.2018.07.002

Vandamme, J.-P., Meskens, N., & Superby, J.-F. (2007). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405–419. https://doi.org/10.1080/09645290701409939

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior, 89*(July), 98–110. https://doi.org/10.1016/j.chb.2018.07.027

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*(October 2019), 106189. https://doi.org/10.1016/j.chb.2019.106189

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47*, 168–181.

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior, 98*(January), 166–173. https://doi.org/10.1016/j.chb.2019.04.015

Zabriskie, C., Yang, J., DeVore, S., & Stewart, J. (2019). Using machine learning to predict physics course outcomes. *Physical Review Physics Education Research, 15*(2), 020120. https://doi.org/10.1103/PhysRevPhysEducRes.15.020120

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.