



Early prediction of Student academic performance based on Machine Learning algorithms: A case study of bachelor's degree students in KSA

Mouna Ben Said^{1,2} · Yessine Hadj Kacem² · Abdulmohsen Algarni³ · Atef Masmoudi^{3,4}

Received: 8 June 2023 / Accepted: 13 November 2023 / Published online: 13 December 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

In the current educational landscape, where large amounts of data are being produced by institutions, Educational Data Mining (EDM) emerges as a critical discipline that plays a crucial role in extracting knowledge from this data to help academic policy-makers make decisions. EDM has a primary focus on predicting students' academic performance. Numerous studies have been conducted for this purpose, but they are plagued by challenges including limited dataset size, disparities in grade distributions, and feature selection issues. This paper introduces a Machine Learning (ML) based method for the early prediction of bachelor students' final academic grade as well as drop-out cases. It focuses on identifying, from the first semester of study, the students requiring specific attention because of their academic weaknesses. The research employs nine classification models on students' data from a Saudi university, subsequently implementing a majority voting algorithm. The experimental outcomes are noteworthy, with the Extra Trees (ET) algorithm achieving a promising accuracy of 82.8% and the Majority Voting (MV) model outperforming all existing models by an accuracy reaching 92.7%. Moreover, the study identifies the factors exerting the greatest impact on students' academic performance, which belong to the three considered feature types: demographic, pre-admission, and academic.

Keywords Educational data mining · Machine learning · Predictive model · Student academic performance · student drop-out

Yessine Hadj Kacem, Abdulmohsen Algarni and Atef Masmoudi contributed equally to this work.

✉ Yessine Hadj Kacem
yessine.hadjkacem@enis.tn

Extended author information available on the last page of the article

1 Introduction

Educational institutions are witnessing a crucial growth of the amount of data they generate from several sources including registration platforms, learning management systems, and educational applications. This large amount of data, if properly harnessed, can serve as a valuable resource to improve the quality of education. Educational Data Mining (EDM) (Romero & Ventura, 2020) emerges as a promising field that permits the exploration of large educational datasets in order to extract valuable knowledge and uncover meaningful patterns to inform decision-making and enhance learning outcomes. One of the main goals of EDM is the students' performance prediction (Kumar & Pal, 2011). A powerful approach in this field is the application of Machine learning (ML) techniques (Suthaharan, 2016). By leveraging ML algorithms, patterns and relationships can be identified within educational data which empowers the prediction of individual students' achievements, the identification of students requiring special attention or who are at risk of drop-out, and the understanding of factors affecting students' learning outcomes (Batool et al., 2022; Khan & Ghosh, 2020). Such predictive models provide valuable information to help educators and policy-makers make timely interventions and provide targeted support to students at risk of falling behind.

Multitude of studies on the prediction of students academic performance using ML algorithms exist in the literature. However, they often face challenges such as small dataset size and grade distribution imbalance which may potentially skew the accuracy of predictions. They also struggle with the choice of features to be used for prediction which are of a great impact on the model performance. While demographic and academic-related attributes are predominant in the literature, there exist other captivating features that have the potential to enhance the model accuracy which is still sub-optimal. In addition, the use of multi-semester features may hinder the early prediction and thus the early intervention efforts.

The present work proposes an approach to the early prediction of students' academic performance that addresses the aforementioned issues. We specifically consider a case of bachelor degree students in Saudi Arabia. The novelty of this work lies in working on a dataset that comprises students' data from two different programs offered in twelve different branches in distant geographic locations. Our prediction model operates from the first semester, allowing for the early identification of the students' grade in the final year as well as the students who are at risk of drop-out. The timely prediction of students' grade may help the instructors better assess the students' capabilities and then tailor the learning tasks to their needs in order to help them reach their full academic potential. It may also assist the university administrators in prioritizing the departments and allocating resources among them. Furthermore, identifying the students who are at risk of drop-out at an early stage permits the administrators and policymakers provide timely interventions in order to save them from delinquency and joblessness. Such prediction may equally help the departments predict the number of students that will register to the rest of courses in the second, third and fourth year and thus be prepared in terms of resources. To make such predictions, ten different ML models have been developed and evaluated to achieve more accurate and reliable predictive models than existing related works. The main outcomes of this study have

proved the efficiency of using the SMOTE technique and the ensemble approach in dealing with data imbalance and model generalizability.

The remainder of this paper is organised as follows. Section 2 gives an overview of the related works tackling the prediction of students performance using ML algorithms. The developed methodology is presented in Section 3. The acquired dataset is described and the steps of its pre-processing are explained. The developed ML models are then exposed along with the evaluation metrics. Section 4 shows and discusses the results of predictive models. Finally, Section 5 concludes the paper.

2 Literature review

2.1 EDM and ML

In recent years, Educational Data Mining (EDM) (Romero & Ventura, 2010; na-Ayala, 2014) and Machine Learning have gained significant attention for their potential to improve educational outcomes. EDM utilizes data mining techniques to extract valuable information and insights from educational datasets, while ML involves the use of algorithms and statistical models to enable machines to learn from data and improve their performance over time. These two fields have been combined to address a range of educational challenges, including predicting student performance, identifying at-risk students, and enhancing personalized learning experiences (Shafiq et al., 2022).

2.2 ML-based prediction of student performance

Prior researches have delved into predicting students' academic performance using machine learning algorithms. One example is a recent study that introduced a hybrid classification model utilizing Decision Tree and Support Vector Machine algorithms (Hussain et al., 2022). This study made use of a dataset containing 520 records of Bachelor students which is collected through a questionnaire about 29 different features. It identified factors that may affect students' academic performance which are basically related to social networks and mobile games. This study has some limitations. Firstly, the relatively small size of the dataset may impact the generalizability of the study's findings. Secondly, the issue of imbalanced dataset was not tackled, which could result in biased model performance. Thirdly, the high number of features used in their models could introduce noise and potentially reduce prediction accuracy. Lastly, the models' accuracy, which ranged from 69% to 78% for different splits, was relatively low, indicating the need for further improvement.

In (Chen & Zhai, 2023), the authors considered three separate educational datasets and validated seven ML models. The first dataset consisted of 400 graduates from the National Institute of Engineering in India, with the objective of predicting graduate exam admission likelihood based on student-related characteristics. The second dataset was obtained from a university's campus placement records for their engineering programs in 2013-2014, containing 2966 sample records. The third dataset focused on student performance classification in two Portuguese secondary schools,

with 649 samples and 32 various student-related factors. The authors reported that the Random Forest algorithm demonstrated highly effective performance in predicting student performance across all three datasets, with accuracies of 87%, 89%, and 80% for the first, second, and third datasets, respectively. However, the study did not compare its findings with those of previous studies using the same datasets, which limits the ability to evaluate the novelty and significance of its results. Additionally, it did not address the issue of imbalanced datasets, which has the potential to impact the models' performance.

S. and N. Alturki conducted in (Alturki & Alturki, 2021) a research on predicting the academic achievement of undergraduate students at Princess Nourah Bint Abdulrahman University in Saudi Arabia. They collected records of 300 female students from three departments within the Computer and Information Science College. The collected dataset consists in demographic, pre-enrollment and post-enrollment features related to the first four semesters of study. The authors compared the performance of six data mining methods and determined that Naïve Bayes achieved the best accuracy of 67% in predicting the students' final academic grade, while RF performed significantly better in honorary students prediction with an accuracy of 90%. The dataset used in this study had limitations including the small sample size, the absence of male students, and the imbalance in grades' distribution. The authors acknowledged that the imbalance issue was not tackled in their study, which could affect the results' generalizability.

In (Beaulac & Rosenthal, 2019), the dataset used was obtained from the University of Toronto and consisted of seven dimensions of observations, such as the student ID, course title, department of the course, semester, credit value of the course, and the numerical grade obtained by the student. The dataset contained information on 38,842 students, with 26,488 completing their undergraduate program and 12,294 dropping out. Two classifiers were built, one to predict if a student would complete his undergraduate program and the second to predict the student's major. The RF algorithm was used to predict the major and achieved an accuracy of 47.41%. Such a relatively low accuracy, along with the use of an imbalanced dataset, represent the major limitations of this work. It is suggested that an increase in the number of features could improve the performance of the model.

On the other hand, in (Olabanjo et al., 2022), the main objective was to develop a Radial Basis Function Neural Network for predicting the academic performance of secondary school students. The study analysed the results of 1927 students from a Nigerian secondary school located in Lagos State. The dataset included subjects such as Mathematics, English, and local language subjects depending on the students' major which can be either science, commercial, or arts. The proposed model achieved an accuracy of 86.59% in predicting the students' majors. However, the major limitation of this work is that it is based on the students' graduation scores of all six years of study which disallows the early prediction or early intervention. Indeed, it did not provide a justification for the features selection.

In another research work (Uliyan et al., 2021), predictive deep learning techniques, particularly the Bidirectional Long Short Term Model, were used to identify students who were at risk of leaving. The research collected data from the Saudian Univer-

sity of Ha'il, focusing on 2,000 first-year students from the preparatory dataset and 949 second-year from the College of Computer Science and Engineering dataset. The features used in the study were the course grades from the first four semesters. The model achieved a 90% accuracy in predicting students at risk of retention using Bidirectional Long and Short Term Memory (BLSTM)-Conditional Random Field (CRF). However, the study could be improved by including the student branch feature and addressing the issue of imbalanced dataset.

Sujan et al. (Brdesee et al., 2022) developed a hybrid 2D CNN model to predict students' academic performance by combining two different 2D CNN models. The study used the Open University learning analytics dataset (Kuzilek et al. Kuzilek et al., 2017) comprising observations of 32,593 students studying 22 courses in the Open University during 2013 and 2014. The considered factors included student demographic information, daily interaction with the university's Virtual Learning Environments (VLE), and student assessment as well as final results. The proposed model achieved an accuracy of 88%. The value of this study would be enhanced by solving the problem of imbalanced dataset and adding a comparison with other studies working on the same dataset.

Hani et al. (Poudyal et al., 2022) analyzed a dataset, acquired from the student information system of a Saudian government university, to predict academic performance, infer student behavior, predict the time needed for the student to graduate, and analyze the capacity of the campuses present in the institute. The dataset comprises data on course registration and the academic performance of over 230,000 students from different programs for the years 2006 to 2015. Demographic and academic-related features were considered. The Random Forest model reached an accuracy of 86% and it was suggested that including the program name as a feature could enhance the model performance since the courses during the first two years vary depending on the program.

One more recent research work (Alghamdi & Rahman, 2023) is concerned by the prediction of higher school students' academic performance using ML classifiers. The dataset, comprising 526 records with 26 features, was collected through an electronic questionnaire tool. The research identified key factors that significantly influence students' success. They consist of a set of demographic features, such as accommodation place and type, family income, and father's and mother's job, and a set of academic factors, which are the grades of all semesters (from S1 to S6). The Naïve Bayes classifier achieved the highest accuracy at 99.34% which is a promising result. Nevertheless, this study presents some limitations regarding the dataset, namely its relatively small size, big number of features, as well as concerns about the reliability of the questionnaire used to collect it. Furthermore, the inclusion of grades from all semesters might potentially limit the timeliness of academic success predictions in a student's educational journey.

In summary, multitude of studies have delved into predicting students' academic performance using ML algorithms, each proposing a different approach to solve the problem. However, most of them share the same shortcomings. First, they often grapple with the constraint of a limited dataset which may affect the results' generalizability. Additionally, the issue of imbalance in grades' distribution, which may result in

biased model performance, is largely unaddressed. Another critical issue is the choice of features used for prediction which significantly influences the model performance, with demographic and academic-related attributes being predominant in the literature. However, the student branch, which may differ geographically from one student to the other, may be a relevant feature, but is, to the best of our knowledge, not tackled in literature. Furthermore, the use of features from different semesters of study may hinder the prediction and thus the intervention at an early stage. Lastly, the existing models' accuracy is often relatively low underscoring the need for further improvement. In the present work, we aim to address all of the above limitations.

3 Methodology

As shown in Fig. 1, our methodology consists of a multi-stage process that commences with the step of collecting a real-world dataset from a higher education institute, which is a challenging and time-consuming task. Subsequently, a data pre-processing procedure is carried out to ensure data readiness. After this preparatory stage is finished, we go on to the pivotal phase of model training and testing. It involves the use of nine ML algorithms and is conducted through the cross-validation technique (Refaeilzadeh et al., 2009) to generate nine predictors of students' academic performance. A majority voting approach is then applied to determine the final value of the prediction as well as the evaluation metrics. The details of each of these steps are elaborated upon in subsequent subsections.

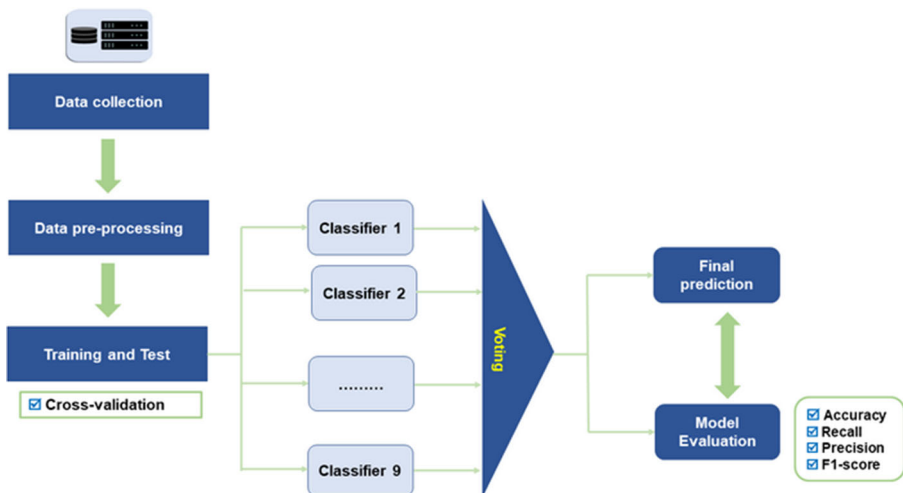


Fig. 1 Proposed approach

3.1 Data collection

The step of data collection constitutes a crucial foundation for such research. The dataset used in this study was acquired from the registration deanship of a university in KSA. A tailored data collection script was developed to generate this dataset, since the specific requisite features for our study are not inherently available within the registration platform. For anonymity purpose, we will not reveal its ethnicity. It originally contains 2444 records of undergraduate students majoring in two different programs, Information System (IS) and Computer Science (CS). These programs are offered for 10 semesters. The repository comprises 27 factors including demographic, pre-admission, and academic information. This dataset covers 5 batches; It has been collected over a period of 5 years (from 2013 to 2017, included). It is also important to mention that the participant students in the dataset are spread over twelve different branches of the same university which are geographically distant (between 40 and 200Km of distance).

3.2 Data pre-processing

We explain in the subsequent subsections the different steps of feature engineering and data pre-processing that led to an understandable and balanced dataset.

3.2.1 Feature engineering

The initial dataset contains raw data with a big number of features of which many are redundant (such as date of birth and age, which are the same), irrelevant (such as student identifier which had the values std1, std2, etc ...), private (such as the branch name) or empty (such as place of birth). Moreover, multiple records are almost empty. Therefore, we were led to clean and annotate the dataset in order to make it

Table 1 Description and labelling of the selected features

Feature	Description	Labelling
Gender	The student's gender	0 for female, 1 for male
Age	The student's age	from 20 to 44 years
Branch	The branch to which the student belongs	from 1 to 12
Program	The program that the student follows	0 for IS, 1 for CS
Semester	The semester of student enrollment in the program	
HSAA	High School Accumulative Averages	0-100%
GAT	General Aptitude Test scores	0-100%
STAAT	Standard Achievement Admission Test scores	0-100%
Score	Final admission score	0-100%
English	Grade of English course obtained in the first semester	0-100%
Math	Grade of Math course obtained in the first semester	0-100%
Computing	Grade of Computing course obtained in the first semester	0-100%

understandable and interpretable by the predictive models. We finally retained 2125 records and 12 features presented in Table 1. These features are of 3 types:

- 3 demographic features which are the age of the student, ranging from 20 to 44, his gender, and the branch to which he belongs. The branch name has been transformed into an integer value ranging from 1 to 12 for privacy reasons.
- 4 pre-admission information: the first two attributes are the STAAT (Standard Achievement Admission Test) and GAT (General Aptitude Test) which are the scores of the exams used for students' admission into public Saudi universities. These exams are mastered by the National Center for Assessment and Evaluation (QIYAS). Each of them is dedicated to assess specific students' skills. The STAAT measures the students' overall comprehension of basic subjects which are English, mathematics, physics, biology, and chemistry. The GAT evaluates numerical, verbal, logical reasoning, and deductive /inductive skills. The third pre-admission attribute is the HSAA (High School Accumulative Averages) which is the accumulative average score of the secondary school. Using these three exam scores, a final admission score is calculated as follows:

$$\text{score} = \text{GAT} * 0.3 + \text{HSAA} * 0.3 + \text{SAAT} * 0.4$$

The exam weights used in this formula are specified by the university based on its own admissions criteria.

- 5 academic factors which are the program that the student follows, the semester of his enrollment, and finally the grades he obtained in 3 core courses from the first semester of study which are English, Math, and computing. These courses are shared by both of the studied programs, thus giving a facility of switch between the majors, in case it is recommended upon an early prediction of the final grade.

As it is shown in Table 1, an integer encoding was used for categorical values which are the student's gender (0 for female and 1 for male), branch (from 1 to 12) and program (0 for IS and 1 for CS). Figure 2 presents the distribution of students based on their branch and program.

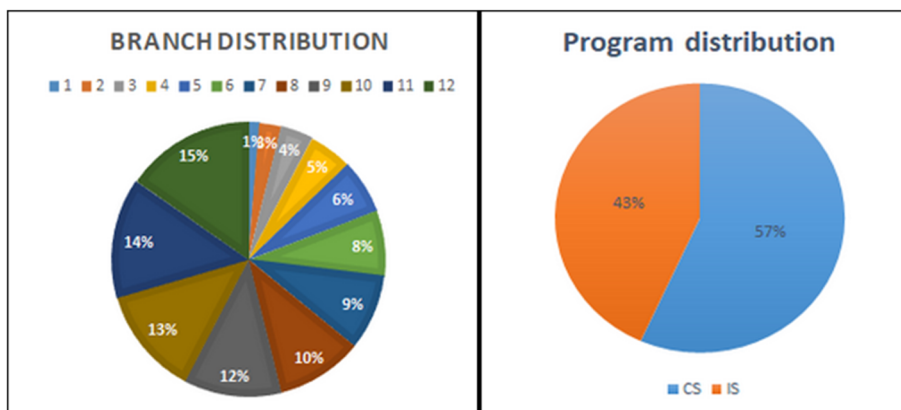


Fig. 2 Distribution of students in branches and in CS/IS programs

Table 2 Student performance labelling

Grade	Criteria	Class
Not graduated	Drop-out student (transferred to another program, left because $GPA < 2$, or got expelled)	0
Acceptable	$2 < GPA \leq 2,75$	1
Good	$2,75 < GPA \leq 3,75$	2
Very good	$3,75 < GPA \leq 4,75$	3
Excellent	$4,75 < GPA \leq 5$	4

As for the target class, which is the final academic grade, it is initially given by the student's GPA in the last semester of study. The GPA is a floating point number ranging from 0 to 5. However, since we deal with a classification problem of the students' grades, we were led to transform the continuous GPA data into a nominal one. As it is presented in Table 2, five classes of grade are defined: integers from 1 to 4 represent the different levels of grade from acceptable to excellent. A fifth class labelled 0 is defined to represent the students who did not complete the program (drop-out students) either because they transferred to another program, left the program because their GPA is under 2, or got expelled from university.

3.2.2 Dataset resampling

We analyzed the distribution of our target classes, which are five classes representing the students' final grade and ranging from 0 to 4. We noticed that there is a significant disparity between the classes; class 2 (grade "Good") and class 3 (grade "Very good") are the majority, having 40% and 27.81% of the total representations, respectively. However, the remaining classes have notably lower observations in the dataset. This is a case of imbalanced dataset which is among the key challenges in classification as it may result in biased models which perform poorly on the underrepresented classes. To address the imbalance and improve the model's performance, several resampling techniques can be employed such as oversampling, undersampling, and generating synthetic samples (Batista et al., 2004). In the present work, we used the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). This technique generates synthetic samples of the minority class through the interpolation between one or more existing minority class samples and their nearest neighbors. Figure 3 illustrates the distribution of the target classes before and after using SMOTE. After resampling, each class counts 850 students records. The resulting dataset thus comprises 4250 rows x 12 columns.

3.3 Modeling and prediction

In EDM, classification techniques are used for prediction tasks. In the present work, we developed nine models for the prediction of student academic performance using



Fig. 3 Distribution of classes before and after using SMOTE

ML classification algorithms which are the K Nearest Neighbors Classifier (KNN), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Gaussian Naive Bayes (GNB), Random Forest (RF), Gradient Boosting (GB), Multilayer Perceptron (MLP) and Extra Trees (ET). The Majority Voting (MV) algorithm is then used to combine the predictions of the above individual models to provide the majority vote as the final prediction. The classifiers were trained on 80% then tested on 20% of the dataset. The cross-validation technique (Refaeilzadeh et al., 2009) was used to provide a more reliable evaluation of models' performance and help in hyperparameter tuning by identifying the optimal combination that averagely yields the best performance across multiple folds. The experiments are performed using Python 3.11.5. The ML models are developed using the scikit-learn library version 1.2.2.

In the subsections below, the principle of each of the used algorithms is briefly described and the different metrics used to evaluate their performance are presented.

3.3.1 ML algorithms

KNN (Cover & Hart, 1967) is a non-parametric algorithm for regression and classification tasks. It assigns the class of a new observation by looking at their k-nearest neighbors. The value of k usually depends on the dataset size and dimensionality and is usually chosen using cross-validation to minimize bias. KNN can be computationally expensive for large datasets and high-dimensional feature spaces.

SVM (svm) is a powerful algorithm that has been used successfully to solve several classification problems with both linearly and non-linearly separable data. It attempts to find the best hyperplane that maximizes the boundary between the different classes of labeled data points in order to minimize the classification error of the model on the unknown data set. SVM is equally able to handle multi-class classification tasks.

DT (Breiman et al., 1984) is a simple yet effective ML algorithm that is commonly used to solve classification problems with both categorical and continuous data. Based

on the features' values, DT performs recursive partition of the data into subsets to construct a tree structure that is used to classify unknown data points.

LR (Hastie et al., 2009) is a linear supervised learning classifier that is mainly used for binary classification problems, but is also able to handle multi-class ones. It is a statistical method for analysing a dataset with one or more independent variables. It uses a logistic function to model the probability of a data point belonging to a particular class. LR can be sensitive to outliers and needs attentive feature selection to avoid overfitting.

GNB (Duda et al., 2001) is a computationally efficient probabilistic classifier that can be used to solve both binary and multi-class classification problems. It performs predictions based on Bayes' theorem and assumes that parameters are independent of each other and distributed according to Gaussian (Normal) Distribution. GNB is sensitive to correlated parameters and requires careful data pre-processing.

RF (Breiman, 2001) is a supervised ML algorithm that is widely used in Classification and Regression problems. It is an ensemble method that builds multiple DT over bootstrapped subsets of data with random subset of features. Its final prediction results from voting or averaging the predictions of the individual trees. The use of random feature selection permits to improve the model performance and cope with overfitting. Indeed, by using multiple trees, RF is able to efficiently handle large datasets and improve the model's accuracy and stability.

GB (Friedman, 2001) is also an ensemble learning method that iteratively builds weak prediction models, typically Decision Trees, working towards a stronger model by minimizing a loss function at every stage. It is computationally very expensive since it often requires many trees. An attentive tuning of its hyperparameter is also required to avoid overfitting.

MLP (Bishop, 1995) is a multilayer feedforward Artificial Neural Network. It is composed of three types of layers of nodes; input, hidden, and output layers. It uses backpropagation learning technique that enables it to learn complex non-linear relationships between inputs and outputs. MLP can be used for both classification and regression problems.

ET (Geurts et al., 2003) (also called Extremely Randomized Trees) is an ensemble learning method that is similar to RF in that it builds multiple DT to perform its task. However, unlike RF, it constructs the trees over the entire dataset and not only over bootstrapped subsets of the data, which helps to obtain lower variance. Indeed, in addition to the random selection of features and thresholds while constructing each tree, ET uses a randomised node split instead of searching for the best node to split on, which makes it faster. This additional randomness leads to better generalization of the model and enhanced resistance to noise. Similarly to RF, ET algorithm is computationally efficient and is able to effectively handle large datasets.

MV (Smith & Johnson, 2022) is also an ensemble learning method where multiple base models are independently trained and then their predictions are combined. In a classification problem, MV is used to determine the final prediction. When the base models have different strengths and weaknesses, this method helps improve the accuracy of the predictions by leveraging the collective knowledge of multiple models.

3.3.2 Evaluation metrics

In order to assess the performance of our predictive models, we considered 4 of the most used model evaluation metrics which are as follows:

- Accuracy: it represents the proportion of the correctly classified samples across the entire dataset¹. It is calculated using (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- Recall: also called Sensitivity. It represents the fraction of positive samples that were retrieved. A low recall score means that the model struggles to identify positive instances. A high recall indicates that the model retrieves most of the relevant samples. Equation 2 represents the recall formula.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

- Precision: also called positive predictive value. It represents the fraction of the truly predicted positive instances among the total instances that are predicted as positive (TP+FP) as shown in (3). A Higher precision indicates that the model predicts more relevant instances than irrelevant ones.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

- F1-score: combines the precision and recall scores of a classification model. The higher the precision and recall, the higher the F1-score. The closer the F1-score to 1, the better is the model.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

4 Results and discussion

4.1 Prediction results analysis

The performance evaluation of the ten developed models is performed twice: before SMOTE evaluation is shown in Table 3, and after SMOTE evaluation is summarised in Table 4. Figure 4 represents a visualisation of these outcomes using bar charts which facilitates the comparison between the different models and the interpretation of results.

We deduce from Table 3 that most of the models struggled with accurately classifying minority classes before applying SMOTE, as proved by low values of recall,

¹ TP is the true positive, FP is the false positive, TN is the true negative and FN is the false negative.

Table 3 Model Performance before SMOTE application

Model	Accuracy	Recall	Precision	F1-score
Gaussian Naive Bayes	55,80%	59,70%	59,30%	54,70%
Logistic Regression	62,40%	62,40%	60,20%	59,30%
SVM	63,00%	63,00%	57,60%	58,80%
Decision tree	47,80%	47,80%	51,30%	47,40%
Gradient Boosting	57,10%	57,10%	59,80%	55,60%
KNN	53,30%	53,30%	58,20%	50,90%
Multilayer Perceptron	57,00%	57,00%	54,40%	56,10%
Random Forest	57,20%	57,20%	60,10%	55,30%
Extra Trees	56,60%	56,60%	61,50%	54,70%
Majority voting	80,70%	80,70%	82,04%	80,05%

precision, and F1-scores. However, it is apparent from Fig. 4 that the use of SMOTE had positive effect on the performance of almost all the models with different degrees, indicating the ability of SMOTE to address class imbalance issue. The majority of the models, including DT, GB, KNN, MLP, RF, and ET, show significant performance improvement of about 14 to 24% for all metrics. The final achieved accuracy is between nearly 71% for both DT and GB, and 80% for RF and ET, making these two latter strong performers in this classification problem.

LR and SVM classifiers exhibited a slight improvement after SMOTE, ranging from nearly 2 to 10% only. This can be attributed to the ability of these algorithms to perform relatively well on unbalanced data. This can be seen from their relatively high accuracy of 62.40% for LR and 63% for SVM before SMOTE application. However, the precision and F1-score were relatively low for these models, showing consistent challenge to correctly predict the minority classes. As for GNB, SMOTE

Table 4 Model Performance after SMOTE application

Model	Accuracy	Recall	Precision	F1-score
Gaussian Naive Bayes	59,70%	59,70%	59,30%	59,00%
Logistic Regression	65,20%	65,20%	67,40%	64,50%
SVM	66,10%	66,10%	68,50%	65,10%
Decision tree	71,20%	71,20%	69,10%	70,50%
Gradient Boosting	71,30%	71,30%	73,70%	70,80%
KNN	72,10%	72,10%	72,40%	69,30%
Multilayer Perceptron	75,70%	75,70%	75,10%	74,70%
Random Forest	80,10%	80,10%	80,10%	79,20%
Extra Trees	80,80%	80,80%	81,80%	79,90%
Majority voting	92,70%	92,70%	92,68%	92,58%

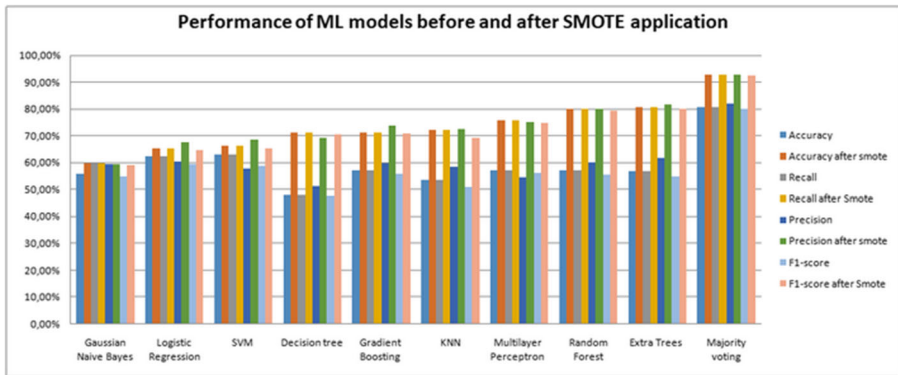


Fig. 4 Visualisation of classification performance of ML models before and after SMOTE application

had no effect on its performance. This model has the worst overall performance in this work, reaching moderate final value of only 59% for all four metrics.

Finally, the MV model consistently achieved the highest classification performance among all models before and after SMOTE. This indicates that the synthetic samples generated by SMOTE had no significant impact on the performance of the model. Before SMOTE, MV reached an accuracy of 80.70%. The F1-score, recall, and precision were also high, indicating good performance to accurately classify all classes. After applying SMOTE, it maintained its high performance, with an accuracy, F1-

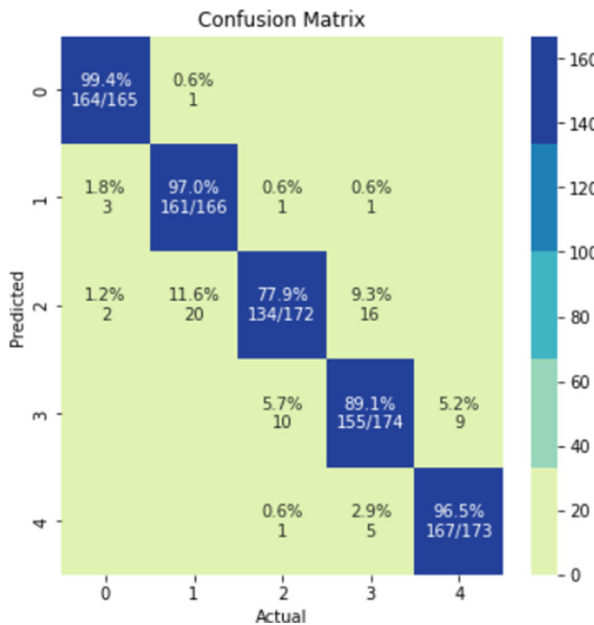


Fig. 5 Confusion matrix of the Majority Voting Model| The proportions of model predictions vs the actual results

score, recall, and precision of 92.7%. The insignificant improvement of the MV ensemble model after SMOTE application and its consistently high performance can be attributed to the strong ability of ensemble methods to capture the collective knowledge of the individual models and make robust and reliable predictions.

The above results demonstrate the advantage of using the SMOTE technique as well as the ensemble approach in providing more accurate and reliable predictions. We additionally visualise the performance of the best model (the MV Model) using its confusion matrix as illustrated by Fig. 5. The matrix shows that the ensemble model performs a successful prediction of the different classes that reached high values of 99.4% for class 0, 97% for class 1 and 96.5% for class 4, and a bit inferior values of 89.1% for class 3 and 77.9% for class 2. Such significant percentages further prove the effectiveness of the proposed model in predicting the students’ academic performance.

Figure 6 illustrates the stability curve of the MV model which permits to assess the performance and generalization capabilities of the model across different training dataset sizes. This curve includes the curves of training score and validation score which can help determining the optimal size of training dataset, providing insights into the model’s performance and guiding decisions on data collection, model architecture, and regularization techniques. The training score curve represents the performance (accuracy) of the ensemble model on the training data as a function of the training dataset size. It shows how well the model is fitting the training data as more data is added. We observe that the model achieves a high training score. This could indicate that the model is able to capture the underlying patterns and complexities in the data adequately. The validation score curve represents the model’s performance on a separate validation dataset (not used during training) as a function of the training dataset size. It evaluates how well the model generalizes and is able to perform on

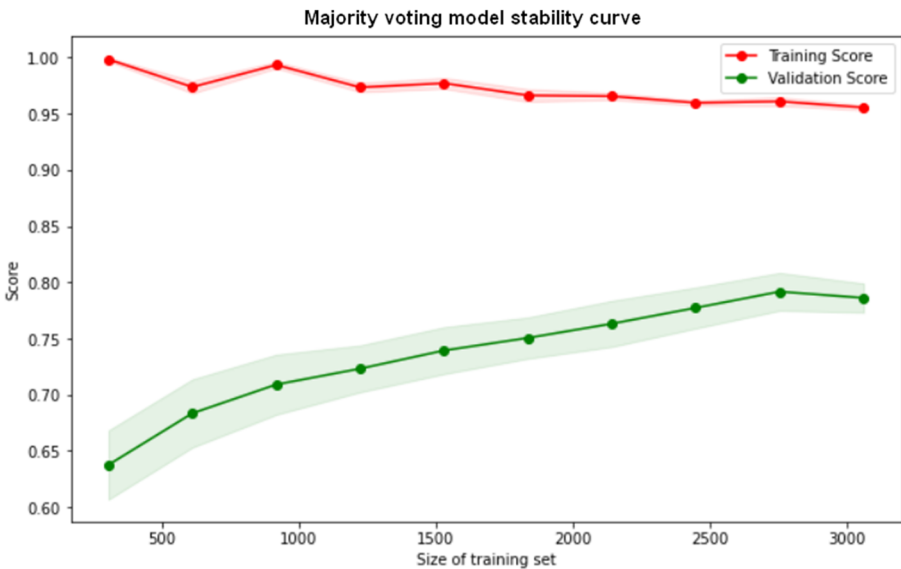


Fig. 6 Majority voting model stability curve

unseen data. We observe that, with a small training dataset size, the validation score may be low (65%). This indicates that the model is not able to generalize well to unseen data, as it might be overfitting to the limited training examples. As the training dataset size increases, the validation score typically improves. This suggests that with more diverse training data, the model becomes better at capturing the underlying patterns in the data and can generalize well to unseen examples.

4.2 Top-10 factors affecting student academic performance

One important task in EDM is the identification of the factors that have a significant influence on the student performance. Such task has several benefits including the enhancement of the classification models' performance by directing the efforts towards the understanding of the most influential factors and their incorporation into predictive models. It serves also as a support for decision-makers by allowing more targeted interventions and informed strategies. In this study, we determined the importance score of the top-10 factors that affect the prediction of students' performance, using the ET model, which is the best single model in terms of accuracy. Table 5 shows the resulting scores.

It does appear that the student gender is the factor that most influences the predictive decision of the ET model, with a substantially higher importance score (0.34) compared to the other factors. This is consistent with the findings of (Parajuli Thapa, 2017) where authors concluded that female students are more likely to outperform their male counterparts. The next nine features exhibit importance scores that are relatively close to each other, ranging from 0.089 to 0.064. The second most affecting factor is the branch to which the student belongs. This can be returned to the difference between the branches in terms of geographic location and quality of the offered educational resources and facilities, such as laboratories, libraries, and online materials. Students who have access to better resources may have better learning experience and thus better

Table 5 Top-ten attributes on the prediction of student performance using Extra Trees model

Order	Factor	Importance Score
1	Gender	0,34
2	Branch	0.089
3	Semester	0.081
4	HSAA	0.075
5	Program	0.074
6	STAAT	0.074
7	Compting	0.068
8	GAT	0.065
9	English	0.065
10	Math	0.064

academic performance. This finding further proves the contribution of this study which incorporated the branch feature in the predictive models.

In addition to the above demographic factors, two academic features were determined among the top-5 high-impact factors which are the semester of the student enrollment and the program type. This finding is also in line with the recent literature review by (Alyahyan & Dustegor, 2020) which confirmed that the learning environment (Mueen et al., 2016) including semester period and program type are among the five factors that are the most frequently studied in students' performance prediction. In fact, there may be differences between the batches in terms of academic background: the students educational background may differ from one year to the other due to variations in the primary and secondary education quality as well as the academic programs' rigor, which can lead to varying students levels. Moreover, the admission criteria may change from one year to the other thus causing variable admissible students level. Furthermore, students entering college in fall semester may have access to more academic support resources, such as orientation programs and mentoring, than those admitted in spring semester. As for the program feature, it is obvious that it is a high impact factor since the difficulty level, requirements, and teaching methods may vary from one program to the other.

Past-academic performance (Adekitan & Salau, 2019) is also among the most influencing factors determined by (Alyahyan & Dustegor, 2020). This fits in with our results recognising the HSAA and the STAAT attributes as the fourth and sixth factors affecting student academic performance, with adjacent importance scores of 0.075 and 0.074, respectively. The GAT attribute is determined as the eighth factor with a minor difference. These factors reflect the general level of the student in the secondary school which will substantially affect his level in college. Ultimately, the course grades are equally identified among the top-10 influencing factors with relatively similar importance scores of 0.068, 0.065, and 0.064 for Computing, English, and Math, respectively. This result is in line with the findings of multiple literature researches such as the comparative study proposed in (Tatar & Dustegor, 2020). This study demonstrated that individual course grades should be used for earlier predictions of graduation GPA (before the third term) to avoid model over-simplification, whereas semester GPAs are recommended for later terms to mitigate model over-fitting.

4.3 Comparison with related works on academic performance prediction

Table 6 shows a comparison between our proposed approach and various existing studies that are highly relevant to our specific context. The comparison demonstrates that our approach distinguishes itself in terms of dataset size, the use of data balancing techniques, the types of input data, and the achieved accuracy. The combination of these criteria, taken as a whole, constitutes the distinctive contribution of this work. A large dataset comprising a total of 2125 records is employed. The issue of imbalanced dataset is addressed through the use of SMOTE technique. Furthermore, pertinent features, in addition to the academic and demographic factors, are utilized and they significantly improved our model's predictive accuracy, surpassing the performance of existing models. Moreover, multiple ML classifiers were explored and a better

Table 6 Comparison with related works on academic performance prediction

Reference	Year	Dataset source	Number of records	Data Balancing	Data types	Best model	Accuracy
(Hussain et al., 2022)	2022	City University of Science and Information Technology in Pakistan	520	No	Academic, personal (social networks, games...), Family, & institution	SVM & DT	78%
(Chen & Zhai, 2023)	2023	3 datasets from an Indian university and two Portuguese secondary schools	400+2966+649=4015	No	academic	RF	87%, 89%, 80%
(Alturki & Alturki, 2021)	2021	Princess Nourah Bint Abdulrahman University in Saudi Arabia	300	No	Pre-enrollment, academic, & demographic of first 4 semesters	RF	90%
(Tatar & Dustegor, 2020)	2020	College of Computer Science and Information Technology from a Saudi university	357	No	Academic for first semester	LR	65.6%
(Olabanjo et al., 2022)	2022	Nigerian secondary school	1927	No	Academic of 6 years of study	RBFNN	86.59%
(Uliyan et al., 2021)	2021	University of Ha'il in Saudi Arabia	2949	No	Academic of first 4 semesters	BSTML-CRF	90%

Table 6 continued

Reference	Year	Dataset source	Number of records	Data Balancing	Data types	Best model	Accuracy
(Brdsee et al., 2022)	2022	Open University Learning Analytics dataset	32593	No	Demographic, academic, & daily interaction with VLE	2D CNN	88%
(Poudyal et al., 2022)	2022	Student information system of a Saudi government university	230000	Yes	Demographic & academic/ various programs	RF	86%
(Alghamdi & Rahman, 2023)	2023	A Saudi government secondary school via electronic questionnaire	526	Yes	Demographic & academic for all 6 semesters	Native Bayes	99.34%
Our approach	2023	Registration deanship of a government university in Saudi Arabia	2125	Yes	Pre-admission, demographic, & academic for first semester/ different branches	Majority voting	92.7%

prediction performance is achieved using the MV Model. This model reached the highest accuracy of 92.7% and then its performance can be attributed to our innovative use of majority voting, wherein multiple models influence the final decision. Indeed, our proposal enabled us to early capture a wider range of factors that influence the performance of the students.

The aforementioned findings provide a valuable contribution to the existing studies on predicting students' academic performance at an early stage and thus present a beneficial support for educational institutions and policymakers.

5 Conclusion

Educational data has a great importance in bettering teaching pedagogy and enabling informed decision-making. The worth of data generated within academic environments has valuable insights and knowledge which can be used for prediction goals. Our study focuses on predicting the graduation performance of undergraduate students from the first semester of study based on ML predictive models. The employed dataset was collected from the registration deanship of a Saudi university over a period of 5 years. It comprises data of Bachelor students majoring from 2 different programs and belonging to 12 different branches in distant geographic locations. After a cleaning and annotation step, the dataset includes 2125 records and 12 features of 3 types: demographic, pre-admission, and academic. The study employed SMOTE technique to solve the problem of imbalanced data. Nine individual models were harvested using ML algorithms for performance comparison. The training and test step was carried out using the cross-validation technique. The performance of our predictive models was evaluated and compared using 4 evaluation metrics which are the accuracy, recall, precision, and F1-score. The ET algorithm achieved the highest accuracy among individual models reaching 80.8%. The MV algorithm was subsequently implemented to combine the individual predictions outperforming all the existing models by an accuracy of 92.7%. A study of the top-10 factors affecting the student academic performance was equally performed. Our findings showed that The prediction decision is most affected by the student gender, whereas the relevance scores of the following nine factors, including branch localisation, admission semester, program, past-academic performance, and course grades, are too close to one another.

The proposed predictive model is utilised upon the conclusion of the first semester. It anticipates students' graduation result, which is denoted by an integer grade class ranging from 0 to 4. Class 4 signifies an excellent level of performance, while class 0 represents drop-out students. In case where the model predicts a lower grade class for a student, we possess the ability to identify the key factors that have influenced his result. These insights offer valuable guidance for educators and policymakers to better understand and support students in reaching their utmost academic potential from the early stage. This counselling can be directed towards formulating strategies, whether that entail strengthening the student's areas of weakness, or, in cases where it's appropriate, switching to a more compatible academic program.

Looking ahead, there are several potential paths for further research. One prospective direction involves applying the developed prediction model to other academic

programs in order to validate its effectiveness in various contexts. Furthermore, exploring the use of deep learning techniques could offer further insights and improvements in predicting students' academic performance.

Acknowledgements The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number (R.G.P.2/549/44).

Author Contributions All authors contributed to the article and approved the submitted version.

Funding This research was financially supported by the Deanship of Scientific Research at King Khalid University under research grant number (R.G.P.2/549/44).

Availability of data and materials The dataset analysed during the current study are not publicly available due to institutional policies but are available from the corresponding author on reasonable request.

Code availability The code is available from the corresponding author on reasonable request.

Declarations

Conflict of interest/Competing interests None

Ethics approval None

Consent for publication All authors have given their consent for publication.

References

- Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5(2), e01250. <https://doi.org/10.1016/j.heliyon.2019.e01250>
- Alghamdi, A. S., & Rahman, A. (2023). Data mining approach to predict success of secondary school students: A Saudi Arabian case study. *Education Sciences*, 13(3). <https://doi.org/10.3390/educsci13030293>
- Alturki, S., & Alturki, N. (2021). Using educational data mining to predict students' academic performance for applying early interventions. *Journal of Information Technology Education: Innovations in Practice*, 20, 121–137. <https://doi.org/10.28945/4835>
- Alyahyan, E., & Düşteğör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1). <https://doi.org/10.1186/s41239-020-0177-7>
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>
- Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2022). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1), 905–971. <https://doi.org/10.1007/s10639-022-11152-y>
- 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>
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.
- Brdese, H. S., Alsaggaf, W., Aljohani, N., & Hassan, S.-U. (2022). Predictive model using a machine learning approach for enhancing the retention rate of students at-risk. *International Journal on Semantic Web and Information Systems*, 18(1), 1–21. <https://doi.org/10.4018/ijswis.299859>

- Breiman, L. (2001). *Random forests*. *Machine learning*, 45(1), 5–32. <https://doi.org/10.1023/a:1010933404324>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC Press.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, Y., & Zhai, L. (2023). A comparative study on student performance prediction using machine learning. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-023-11672-1>
- Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. *Machine learning*, 20(3), 273–297. <https://doi.org/10.1007/bf00994018>
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. <https://doi.org/10.1109/tit.1967.1053964>
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. Wiley.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
- He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/tkde.2008.239>
- Hussain, A., Khan, M., & Ullah, K. (2022). Student's performance prediction model and affecting factors using classification techniques. *Education and Information Technologies*, 27(6), 8841–8858. <https://doi.org/10.1007/s10639-022-10988-8>
- Ioannis, B., & Maria, K. (2018). Gender and student course preferences and course performance in computer science departments: A case study. *Education and Information Technologies*, 24(2), 1269–1291. <https://doi.org/10.1007/s10639-018-9828-x>
- Khan, A., & Ghosh, S. K. (2020). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, 26(1), 205–240. <https://doi.org/10.1007/s10639-020-10230-3>
- Kumar, B., & Pal, S. (2011). Mining educational data to analyze students performance. *textitInternational Journal of Advanced Computer Science and Applications*, textit2(6). <https://doi.org/10.14569/ijacsa.2011.020609>
- Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, 4(1). <https://doi.org/10.1038/sdata.2017.171>
- Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, 8(11), 36–42. <https://doi.org/10.5815/ijmecs.2016.11.05>
- Olabanjo, O. A., Wusu, A. S., & Manuel, M. (2022). A machine learning prediction of academic performance of secondary school students using radial basis function neural network. *Trends in Neuroscience and Education*, 29, 100190. <https://doi.org/10.1016/j.tine.2022.100190>
- Parajuli, M., & Thapa, A. (2017). Gender differences in the academic performance of students. *Journal of Development and Social Engineering*, 3(1), 39–47. <https://doi.org/10.3126/jdse.v3i1.27958>
- Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4), 1432–1462. <https://doi.org/10.1016/j.eswa.2013.08.042>
- Poudyal, S., Mohammadi-Aragh, M. J., & Ball, J. E. (2022). Prediction of student academic performance using a hybrid 2d CNN model. *Electronics*, 11(7), 1005. <https://doi.org/10.3390/electronics11071005>
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems* (pp. 532–538). https://doi.org/10.1007/978-0-387-39940-9_565
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3). <https://doi.org/10.1002/widm.1355>
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, textit40 (6), 601–618. <https://doi.org/10.1109/tsmcc.2010.2053532>

- Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access*, *10*, 72480–72503. <https://doi.org/10.1109/access.2022.3188767>
- Smith, J., & Johnson, M. (2022). Majority voting in ensemble classifiers. *Journal of Machine Learning*, *10*(3), 123–145. <https://doi.org/10.1234/jml.2022.12345>
- Suthaharan, S. (2016). *Machine learning models and algorithms for big data classification*. <https://doi.org/10.1007/978-1-4899-7641-3>
- Tatar, A. E., & Düşteğör, D. (2020). Prediction of academic performance at undergraduate graduation: Course grades or grade point average? *Applied Sciences*, *10*(14), 4967. <https://doi.org/10.3390/app10144967>
- Uliyan, D., Aljaloud, A. S., Alkhalil, A., Amer, H. S. A., Mohamed, M. A. E. A., & Alogali, A. F. M. (2021). Deep learning model to predict students retention using BLSTM and CRF. *IEEE Access*, *9*, 135550–135558. <https://doi.org/10.1109/access.2021.3117117>
- Wang, X., Zhao, Y., Li, C., & Ren, P. (2023). ProbsAP: A comprehensive and high-performance system for student academic performance prediction. *Pattern Recognition*, *137*, 109309. <https://doi.org/10.1016/j.patcog.2023.109309>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Mouna Ben Said^{1,2} · Yessine Hadj Kacem²  · Abdulmohsen Algarni³  ·
Atef Masmoudi^{3,4} 

Mouna Ben Said
mouna.bensaid@crns.tn

Abdulmohsen Algarni
a.algarni@kku.edu.sa

Atef Masmoudi
atef.masmoudi@ipeis.usf.tn

- ¹ Digital Research Center of Sfax, Sfax 3021, Tunisia
- ² CES Laboratory, National Engineering School of Sfax, University of Sfax, Sfax 3038, Tunisia
- ³ Department of Computer Science, King Khalid University, Alfara, Abha 61421, Saudi Arabia
- ⁴ Laboratory of Electronics and Technology of Information, National Engineering School of Sfax, University of Sfax, 3038 Sfax, Tunisia