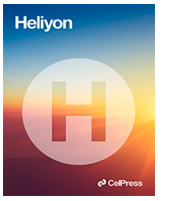
Heliyon 10 (2024) e32570

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Predicting performance of students by optimizing tree components of random forest using genetic algorithm

Mengyao Chen [a,b,*], Zhengqi Liu [c]

[a] School of Teacher Development, Shaanxi Normal University, Xi'an, 710000, Shaanxi, China
[b] Shaanxi Institute of Teacher Development, Shaanxi Normal University, Xi'an, 710000, Shaanxi, China
[c] Xi'an Aviation Computing Technology Research Institute of Aviation Industry, Xi'an, 710065, Shannxi, China

ARTICLE INFO

Keywords:
Academic performance
Optimizing random forest
Genetic algorithm
Feature selection

ABSTRACT

Prediction of student academic performance is still a problem because of the limitations of the existing methods specifically low generalizability and lack of interpretability. This study suggests a new approach that deals with the current problems and provides more reliable predictions. The proposed approach combines the information gain (IG) and Laplacian score (LS) for feature selection. In this feature selection scheme, combination of IG and LS is used for ranking features and then, Sequential Forward Selection mechanism is used for determining the most relevant indicators. Also, combination of random forest algorithm with a genetic algorithm for is introduced for multi-class classification. This approach strives to attain more accuracy and reliability than current techniques. The case study shows the proposed strategy can predict performance of students with average accuracy of 93.11 % which shows a minimum improvement of 2.25 % compared to the baseline methods. The findings were further confirmed by the analysis of different evaluation metrics (Accuracy, Precision, Recall, F-Measure) to prove the efficiency of the proposed mechanism.

1. Introduction

In today's world, education is one of the most important necessities of life. Since education requires a lot of money and budget, the goal of students' academic education is to increase their academic performance. Governments allocate huge sums of national income to education, and in addition, families bear a lot of expenses for their children's education. Academic achievement is correlated with improved academic performance. Academic success is the degree to which students have met the training course objectives [1,2]. The problem of studying academic performance is considered one of the important topics and attention of researchers in the field of educational management. One of the significant issues in academic performance is the correct prediction of students' academic performance and timely action and advice to students at risk of academic failure [3]. The meaning of academic performance is all the activities and efforts that a person makes in order to acquire sciences and pass various educational levels in educational centers. They have conducted significant researches in the field of academic failure and academic performance of students and the variables related to them and have presented several theories about improving the teaching-learning process; among them, learning strategies are considered important tools to enable students to achieve educational goals [4].

Presently, various societies are experiencing substantial advancements in the field of information and communication technology.

* Corresponding author. School of Teacher Development, Shaanxi Normal University Xi'an, 710000, Shaanxi, China.
E-mail address: chenmengyao0623@outlook.com (M. Chen).

https://doi.org/10.1016/j.heliyon.2024.e32570
Received 30 March 2024; Received in revised form 4 June 2024; Accepted 5 June 2024
Available online 6 June 2024

Nomenclature

Abbreviation	*Explanation*
AI	Artificial intelligence
ANN	Artificial neural network
AUC	Area under the curve
DL	Deep Learning
FPR	False Positive Rate
GA	Genetic algorithm
ML	Machine Learning
MLP	Multilayer perceptron
NN	Neural network
RF	Random forest
RFSPP	RF Students' Performance Prediction
ROC	Receiver Operating Characteristic
SVM	Support vector machine
SFS	Sequential Forward Selection
TPR	True Positive Rate

As a result of these developments, scientific leadership institutions such as colleges and universities have adopted electronic processes for education administration; additionally, databases containing vast quantities of data are prevalent in educational environments. Through the analysis of these enormous datasets pertaining to educational systems, feasible approaches to ameliorating the academic circumstances of pupils can be identified. The objective of educational data mining has been to extract knowledge from the educational system's data. Anticipating pupils' academic performance is one of the possible uses of educational data mining. The effectiveness of educational systems greatly depends on the ability to predict students' academic achievement and provide helpful solutions. This information may also assist management in making the best choices possible to improve student performance and system efficiency. The prediction of student learning outcomes is a significant and important use of prediction in educational data mining [3,5,6]. Today, in most schools, there is a large data bank of students' characteristics, which includes a large amount of information related to educational, educational records, etc. Finding patterns and knowledge hidden in this information can help educational managers to improve and improve educational processes such as evaluation, academic performance recognition, and counseling. The computer software used for this purpose is often only responsible for mechanizing the registration and recording of grades, running routine queries and short-term administrative planning. While in the depth of this volume of data, very interesting patterns and relationships between different parameters remain hidden. But by using data mining, it is possible to extract understandable, useful, unknown, valid and novel patterns from the training data of large databases. The patterns that have been found assist school education systems in making better choices and developing more sophisticated plans for student guidance. One of the primary objectives and strategies of the educational system that may profit from the outcomes of these models and the information gleaned from them is assisting students' academic advancement and improving their academic performance [7–9].

On the other hand, the issue that distinguishes the present era from the past is information and communication technology. The level of benefit and use of information and communication technology is directly related to the gap between countries and people. Therefore, it can be said that the most important indicator of progress is the level of development and application of information and communication technology in education [10–12]. Predictive studies were found to be the most prevalent in the field of educational data mining by some scholars who reviewed literature in this area. When doing internal forecasting research, they often use top-notch statistical techniques. Undoubtedly, the proliferation of novel data mining techniques in the domains of engineering and management raises the possibility that such techniques will find application in the realm of educational management as well, thereby bridging the divide between domestic and international research [13–15].

Presently, artificial intelligence (AI) is rapidly revolutionizing every facet of human existence. Not even education deviates from this principle. In recent years, artificial intelligence has advanced at an exceptionally rapid rate. Machine learning (ML), neural networks (NNs), and deep learning (DL) are among the AI capabilities that are developing at an exponential rate [16]. Teachers empowered by AI can provide individualized instruction, practice, and feedback to students in accordance with their individual strengths and limitations. Furthermore, implementation of precise academic performance forecasts for students can contribute to the enhancement of teaching quality. In light of this matter, the present study endeavored to propose a methodology for classifying and forecasting the academic standing of students by employing intelligent techniques [17].

A number of machine learning algorithms have been designed and are being used to predict student performance. Here, we discuss some prominent approaches grouped by type:

A) Artificial Neural Networks (ANNs):

Lau et al. [19] investigated ANNs (artificial neural networks) for predicting and classifying student performance. Although their model was correct 84.8 % of the time, ANNs are considered as black-box models, which is difficult to explain how the input variables

are linked to the prediction. Moreover, Aydoğdu [23] used ANNs to forecast the performance of students in online learning settings, reaching an accuracy of 80. 47 %. Their study established the presence of variables like attendance and time spent on the content as substantial predictors.

B) Ensemble Learning Methods:

Jain et al. [18] demonstrated a hybrid model that is a mixture of MLP (Multi-Layer Perceptron) and Random Forest. MLP outperformed in prediction of student grades whereas Random Forest was proficient in identifying the areas of improvement. This is an approach that is data-hungry and prone to overfitting if not properly tuned.

Kumar et al. [27] presented a multilevel ensemble learning algorithm that combined various algorithms: naïve bayes, random forest, and logistic regression (NB-RF-LR-SEMod) and obtained an accuracy of 88.3 % on a data set. Ensemble techniques can be more precise than single models but they are quite complex and require a careful selection and configuration of base learners.

Asselman et al. [29] decided to check the effectiveness of XGBoost, another ensemble learning method, for the prediction of student performance. XGBoost was found to be superior to other algorithms including Random Forest and AdaBoost, and this result justifies its application as an accurate predictor. Although XGBoost is computationally much more expensive than other simple algorithms, it provides a better model at the same time.

C) Decision Trees and Random Forests:

Jayaprakash et al. [21] have posited an enhanced Random Forest classifier that was developed for student performance prediction. Their model had a strong accuracy, but feature selection and hyperparameter adjustment are very important for getting the best performance.

Ghosh et al. [26] have also employed Random Forest to predict student performance, which was correct 96.88 % of the times. The study used fuzzy analysis to prepare the input data. Batool and et al. [25] designed a Random Forest Students' Performance Prediction (RFSPP) model, which is based on demographics of the students. Their approach got the accuracy from 81.20 % to 95.10 % on different datasets. Random Forest is known to be very accurate and deals well with the complex relationships between variables, but it has the downside of being less interpretable when compared to simpler models.

Hussain and Khan [30] had a system that used machine learning to predict the marks of the students at secondary and intermediate levels. The method they used had data from a Pakistan educational board and it involved data pre-processing, feature selection, training a regression model for marks prediction, and a decision tree classifier for grade classification. This research uses genetic algorithm for feature selection. However it leads to high computational complexity for big datasets.

D) Other Classification Algorithms:

Table 1
Summary of the literature.

Reference	Year	Research Goal	Method	Limitations
Jain et al. [18]	2019	Predict student grades	Hybrid model combining MLP (ANN) and Random Forest	Requires significant computational resources, prone to overfitting if not carefully tuned
Lau et al. [19]	2019	Predict and classify student performance	Artificial Neural Networks (ANNs)	Lacks interpretability ("black-box" nature of ANNs)
Alamri et al. [22]	2020	Predict student performance	Compared SVM and Random Forest	Optimal method depends on specific data and goals
Aydoğdu [23]	2020	Predict student performance in online learning	Artificial Neural Networks (ANNs)	Lacks interpretability ("black-box" nature)
Sekeroglu et al. [20]	2019	Predict student performance	Various machine learning algorithms (decision trees, Naive Bayes)	Importance of data pre-processing for effectiveness
Jayaprakash et al. [21]	2020	Predict student performance	Improved Random Forest classifier	Feature selection and hyperparameter tuning crucial
Ghosh et al. [26]	2021	Predict student performance	Random Forest with fuzzy ANFIS analysis for data preparation	Feature selection and hyperparameter tuning crucial
Batool et al. [25]	2021	Predict student performance	Random Forest Students' Performance Prediction (RFSPP) model	Random Forest can be challenging to interpret
Kumar et al. [27]	2022	Predict student performance	Multilevel ensemble learning model (NB-RF-LR-SEMod)	Requires careful selection and configuration of base learners
Asselman et al. [29]	2023	Predict student performance	XGBoost (ensemble learning method)	Computationally expensive compared to simpler algorithms
Ofori et al. [24]	2020	Identify best model for prediction and improvement	Literature review	Limited exploration of diverse techniques
Alam & Mohanty [28]	2022	Provide framework for educators to use data mining for prediction	Develop framework for educators	May not address interpretability or generalizability
Hussain & Khan [30]	2023	Predict student grades and marks	Decision Tree for regression and classification	High computational complexity for big datasets

Alamri et al. [22] perform a comparison between Support Vector Machine (SVM) and Random Forest models for the purpose of forecasting the students' performance. The two algorithms demonstrated high accuracy, but the choice of the optimum technique may be dependent on the particular dataset or the learning objectives. Sekeroglu et al. [20] employed decision trees and Naive Bayes, being the machine learning algorithms, to predict student performance. The report indicates that pre-processing raw data may be an important factor in better machine learning model performance.

Ofori et al. [24] have done a review of the machine learning for student performance prediction using literature. They underscored the critical role that early prediction plays in the enhancement of learning and the difficulties in deciding the best model for both prediction and learning improvement. Their findings imply that socioeconomic factors may be the reason of low accuracy of the prediction.

Alam and Mohanty [28] provided a structural framework which can be used by educators for predicting student performance by data mining techniques. They focused on the need for setting the student achievement indicators, selecting the key attributes, and choosing the most suitable machine learning models. Their objective is to simplify the use of data mining tools for teachers. Table 1, presents a summary of the literature.

The literature review showed that various methods have been presented in predicting students' performance using ML techniques, but the results presented in these studies are still far from the ideal method and achieving appropriate accuracy. This reason can be investigated from two ways; 1- The lack of clarity of the important indicators affecting the academic performance of students and 2- The models used are often single models, which usually cannot be used to have the necessary generality in the problem of predicting the academic performance of students. In other words, these models may not perform properly in real applications and based on real data. For this reason, the use of ensemble learning methods and the use of the capabilities of multiple models, each of which can make predictions independently, can lead to improving the accuracy of predictions. For this purpose, in this study, a combined strategy was used to select indicators related to academic performance. This strategy firstly uses two criteria of information gain and Laplacian score to evaluate the importance of each index. Then, by combining the values of these two criteria, it ranks the indicators. Finally, the Sequential Forward Selection (SFS) strategy is used to select the optimal features. Then, a combined strategy based on optimized RF is used to predict academic performance. This strategy, by using several decision tree classification models, improves the accuracy of the proposed model in solving the multi-class classification problem. The novel contribution of this research includes:

- Using the combination of Information Gain (IG) and Laplacian Score (LS) to identify the most relevant indicators related to students' academic performance.
- Optimizing decision tree components in order to increase the performance of multiple learning models in predicting students' academic performance.

Table 2
List of candidate indices for predicting academic performance.

ID	Title	Type
1	Type of school	Nominal
2	gender	Nominal
3	Age	Numerical
4	Housing	Nominal
5	Number of family members	Numerical
6	Parents' living situation (divorce status)	Nominal
7	Mother's level of education	Nominal
8	Father's level of education	Nominal
9	Mother's employment status	Nominal
10	Father's employment status	Nominal
11	The reason for choosing the place of study	Nominal
12	Legal guardian of the student	Nominal
13	Travel time from residence to school	Numerical
14	Duration of study lessons per week	Numerical
15	The number of previous failed courses	Numerical
16	Has a scholarship	Nominal
17	Financial support of parents for education	Nominal
18	Using extraordinary classes	Nominal
19	Has extracurricular activities	Nominal
20	History of attending online courses	Nominal
21	Willingness to continue studying	Nominal
22	Internet access at home	Nominal
23	The state of emotional relationships	Nominal
24	Quality level of communication with family members	Nominal
25	Amount of free time after school	Numerical
26	Fun status with friends	Nominal
27	Alcohol consumption during the week	Nominal
28	Alcohol consumption on weekends	Nominal
29	Current health status	Nominal
30	The number of absences in class	Numerical
–	Average final grades of students	A continuous number

- Improving the generalizability of the student academic performance prediction model using optimized ensemble learning systems.

The structure of the article is as follows: The first segment included the introduction. The research approach was provided in section 2. The study findings were reported in part 3, and the conclusion was reported in section 5.

2. Research methodology

This section presents a novel approach that uses a genetic algorithm (GA) to optimize numerous decision tree models in order to predict students' academic achievement. In this respect, the stages of the suggested technique are offered after the data gathering method is explained.

2.1. Data collection

In this research, the data were collected by distributing a questionnaire among high school students in Nanjing, China. This dataset contains 712 samples, of which 305 samples are related to male students and 407 samples are related to female students. At the beginning of the distribution of questionnaires, written consent was obtained from all the participants. All students studied in technical fields. Data collection has been done in two stages. In the first stage, students' information was collected through a questionnaire at the beginning of the academic year. In the second stage, the average grades of the students at the end of the academic semester have been obtained.

The data gathered for this study is listed in Table 2. This table indicates that the dataset comprises one dependent variable (student academic achievement) and thirty possible indicators (as independent variables). The aim of this study is to forecast students' academic achievement using a subset of the potential markers given in Table 2.

The student's academic performance is described as a numerical variable that shows the average of his final grades. In this research, this continuous numerical variable has been converted into a ranked variable. For this purpose, the score interval is divided into four heterogeneous intervals:

1. Bad: [0, 11)
2. Average: [11, 14)
3. Good: [14, 17)
4. Very good [17,20].

The value of this variable has been changed for each sample based on the resulting intervals. By doing this, the database samples are

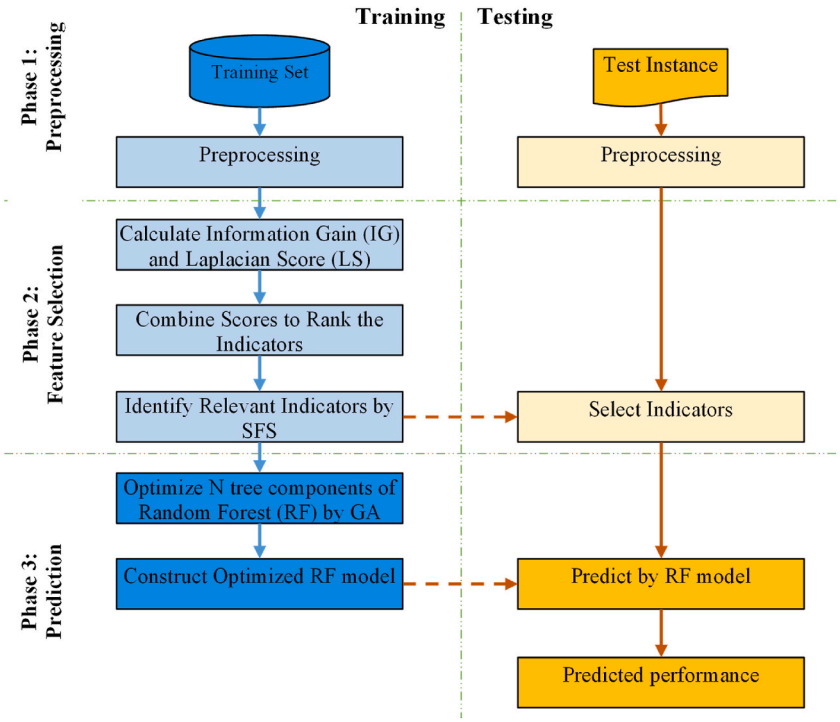


Fig. 1. Steps of the proposed method.

targeted in four categories:

- Bad: 105 samples
- Average: 214 samples
- Good: 180 samples
- Very good: 212 samples

In Table 2, time indicators are defined as minutes. The indicators related to the level of education of the parents are defined in the form of ranks with the values of 0- illiterate, 1- diploma and below, 2- post-graduate diploma and bachelor, 3- post-graduate and 4- doctorate and above. Also, the indicators related to parents' occupation are determined based on the values of education, treatment, service, technical and engineering, housewife or others. The legal guardian indicator can be set with one of its own values, mother, father or others. Qualitative indicators, such as the quality level of communication with family members or the state of recreation with friends, are rated in the form of ratings with values from very bad (0) to very good (5). Other collected discrete indicators are logical and have one of the values 0 = *False* and 1 = *True*.

2.2. The proposed method

The steps of the proposed method for predicting students' academic performance are shown in Fig. 1 as a diagram. The proposed method divides the problem of predicting students' academic performance into three phases:

1. Preprocessing
2. Identifying indicators related to academic performance
3. Prediction of academic performance based on identified indicators.

In the first step, the data collected from the students are cleaned and the nominal values are converted into numbers. Then, in the second stage, a combined strategy is used to select indicators related to academic performance. This strategy firstly uses two criteria of information gain and Laplacian score to evaluate the importance of each index. Then, by combining the values of these two criteria, it ranks the indicators. Finally, the SFS strategy is used to select the optimal features. To forecast academic achievement, a combination technique based on improved RF is used in the third stage. This approach increases the suggested model's accuracy in addressing the multi-class classification issue by using several decision tree classification models. GA optimizes every decision tree component in the suggested RF model. Every decision tree model is optimized by trimming the tree's leaves and establishing split points in each decision node. This combination classifier may be used to predict academic achievement in fresh data once all the components have been tuned and the optimal forest has been formed.

2.3. Pre-processing

The collection of characteristics indicating the state of the students is preprocessed as the first step in the proposed technique. All of the nominal characteristics are initially translated into numbers during the pre-processing stage. In this process, all True/Yes and False/No values are replaced with one and zero values, respectively. Then all the records with missing values are removed from the data set. At the end of the pre-processing process, each of the numerical features of the input are normalized. It should be noted that the values of each feature of the samples in the research problem are not uniformly distributed around the mean and the value of some features cannot be described linearly. For this reason, in order to normalize the features in the proposed method, the SoftMax scaling method has been used. This normalization operator maps each feature non-linearly in the interval $[0, 1]$ and is formulated as follows [31]:

$$F_x = \frac{1}{1 + e^{-\overrightarrow{y}}} \tag{1}$$

Which we have in the above relation [31]:

$$\overrightarrow{y} = \frac{\overrightarrow{x} - \overline{x}}{r\sigma} \tag{2}$$

In Equations (1) and (2); $\overline{x}$ represents the average feature vector $\overrightarrow{x}$, and σ represents its standard deviation. Also, r is an adjustable parameter that is set as $r = 1$ in this research. Using Equation (1), the values adjacent to the mean of the feature vector have a quasi-linear behavior, and as the feature value moves away from the mean, the magnitude of the difference is exponentially suppressed.

2.4. Feature selection

The second stage of the suggested approach involves eliminating unnecessary signs that might lead to predicting mistakes. This procedure not only raises the prediction model's accuracy but also speeds up processing. Ranking features is a common way to determine their importance. This can be done using different criteria. However, any feature ranking solution has limitations depending

on the method used. IG and LS methods are two popular techniques for feature ranking. However, both of these methods have limitations. For example, the IG method tends to select features with higher singular values. In contrast, the LS method requires multiple observations to reduce the risk of over-fitting. To solve these challenges, in the proposed method, the ranking combination determined by two methods, IG and LS, has been used.

Using the IG criterion, the amount of information entropy loss is measured for each feature under the conditions of the target variable. Thus, this criterion can be formulated as Equation (3) [32]:

$$I_f(x, T) = E[x] - E[x|T] \tag{3}$$

In the above equation, x and T represent the input and target features, respectively, and $E[x]$ represents the entropy of the x feature. Based on the above relationship, the imaginary interval for the information benefit of a feature is $[0, +\infty)$ where the higher value of IG indicates the higher importance of the feature; and lower values also indicate the less importance of the feature and its weaker relationship with the target variable.

On the contrary, in the LS feature ranking strategy, the Laplacian score values extracted through the nearest neighbor in the similarity matrix are used. This process for a dataset containing n samples include the following steps:

a) A neighborhood value is defined for each data point and the distance values of each pair of neighbors such as x and y are calculated as $d_{x,y}$.
b) Using the following relationship, the values of the distance $d_{x,y}$ are converted into the similarity $s_{x,y}$ [33]:

$$s_{x,y} = e^{-\left(\frac{d_{x,y}}{\delta}\right)^2} \tag{4}$$

In Equation (4), δ specifies the distance scale and it is set equal to 1 in the current research.

c) Each feature is normalized by Equation (5) and based on its average [33]:

$$\bar{f} = f - \frac{f^T GI}{I^T GI} \tag{5}$$

In the above equation, G is the degree matrix with dimensions $n \times n$, which is calculated as a diagonal matrix based on the sum of similarity values in the rows of S ($G_{i,i} = \sum_{j=1}^{n} S_{i,j}$) Also, I is a vector of length n and I^T describes the output of I.

d) The score of each feature such as f is calculated as Equation (6) [33]:

$$R_f = \frac{\bar{f}^T S \bar{f}}{\bar{f}^T G \bar{f}} \tag{6}$$

As a result of the implementation of the above steps, the score of each index such as f will be described using the criterion R_f. In the proposed method, in order to rank the indicators based on their importance, the combination of R_f and I_f scores are used. For this purpose, first, each of the score vectors R and I is normalized using Equation (1) to solve the problem of incompatibility of rating scales. In the following, the rank of each index such as f is calculated as Equation (7):

$$Rank_f = \frac{1}{2}\left(N_{R_f} + N_{I_f}\right) \tag{7}$$

In the above equation, N_{R_f} describes the normalization value of the Laplacian score of the feature f after normalization by Equation (1). SFS technique is used to eliminate characteristics that aren't important after the feature selection stage. This method involves taking the first two characteristics out of the sorted list and using them to train the classification model. The trained model's accuracy is then assessed for these features, and the training and accuracy assessment processes are then repeated by adding a feature with a higher rank. This process is repeated until adding a new feature does not affect the accuracy of the learning model. In this case, the feature selection process is completed and the feature set that has achieved the highest accuracy of the model is considered as optimal features. The third phase of the suggested technique uses the collection of chosen characteristics as its input.

2.5. Prediction of academic performance

After identifying the indicators related to academic performance, the tree component optimization strategy is used in the RF classification model to predict the target variable. The proposed RF consists of a set of CART decision trees, which can achieve a more powerful classification structure by combining the decision results of these components.

The RF model must be formed using a collection of optimum trees in order for it to function as best it can. Given that the tree construction algorithms are unable to ensure the attainment of the ideal learning model [34], the suggested approach employs the GA to refine and optimize the parameters of each CART. The remaining portion of this section describes how to improve each CART component in the suggested classifier. The first step in optimizing each decision tree component in the proposed combined classification model is to create each initial CART model in an RF structure using training data. An ML approach for classification or regression

that describes the learning model as a tree structure (or a collection of equivalent rules) is known as a CART model. Decision nodes and leaf nodes make up two categories of the tree components in this form. The input sample may be examined in accordance with the requirements of one of the decision variables based on the restrictions represented by each decision node, which together form the rules of the tree. Every decision node contains two sub-branches (of the decision node or leaf type), on the basis of which, depending on whether the sub-branch is a decision node or a leaf, the target variable's value may be defined or the decision rules can be followed. In order to design an appropriate tree, the CART construction procedure entails choosing input variables and figuring out split points for those variables. In order to minimize the cost function, a greedy algorithm is used in this procedure. The point with the lowest cost is therefore chosen at each division step. The suggested approach uses the CART model for classification, and the Gini impurity index is employed to calculate the cost function for tree building [35]:

$$Gini = \sum_{i=1}^{C} p_i(1 - p_i) \tag{8}$$

The number of target classes, denoted by C in Equation (8), and the percentage of training samples labeled I that are categorized by the decision node are shown by p_i. For nodes where every training sample belongs to the same class, the impurity Gini index will be 0. Following the formation of the tree structure, it should be pruned to reduce its complexity while preserving its intended purpose. In the suggested approach, the GA is used to establish the division points in the decision nodes and prune the tree's leaves after the creation of the initial CART model. Setting the dividing points in the decision nodes and pruning each current CART model in the RF constitute the optimization issue that is covered in the third stage of the suggested technique. The GA is therefore carried out individually for every CART model. The optimization variables in this issue may be separated into the following two groups for a CART tree with N decision nodes and M leaf nodes: The number N real variables, each of which represents a decision tree node and whose value designates the node's dividing line. The quantity of M binary variables, each of which represents a tree leaf node. The zero value in these variables denotes pruning the branch in the tree structure that corresponds to that leaf node, while the value one denotes maintaining that node in the tree structure. Each response vector in the GA to optimize each CART model will have a heterogeneous structure made up of N+ M areas based on the factors listed above. The first N elements of each response vector (chromosome) are used to identify the decision nodes' splitting points, while the last element M specifies whether to keep or remove leaf nodes from the tree structure. In Fig. (2), chromosomal structural examples are shown.

In Fig. 2-a, a basic CART model with five decision nodes $I_1 \sim I_5$ and six leaf nodes $A \sim F$ is introduced. For each chromosome corresponding to this CART model in the proposed GA, there are 11 genes, the first 5 of which are real and the next 6 are binary. Binary genes indicate whether leaf nodes are present or absent, whereas true genes correlate to the division point of the decision node. Fig. 2-b displays an example chromosome designed for this CART model. The initial element (0.2) on this chromosome displays the decision

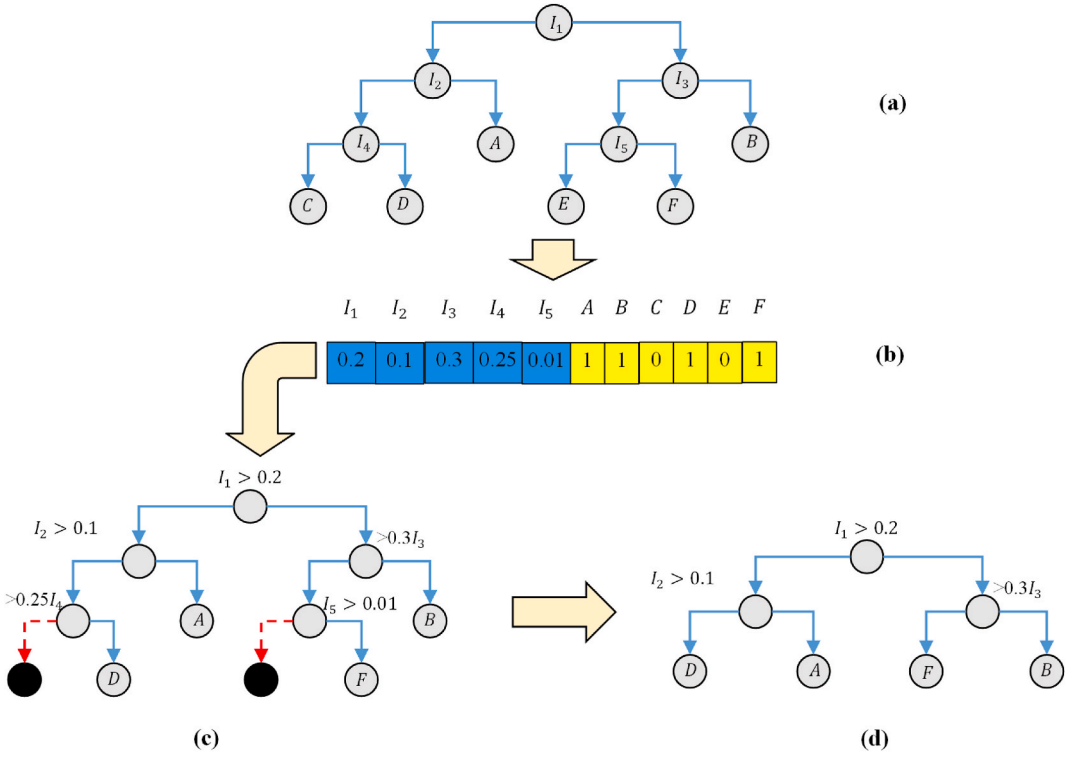


Fig. 2. A response sample and how to apply it in a CART model to optimize the learning model in the proposed method (a) the initial tree, (b) a sample solution vector, (c) the result of applying split points determined by the solution on the initial tree, and (d) the pruned tree.

node I_1's splitting point. Therefore, the root node of the tree is split with the condition $I_1 > 0.2$.

Because the CART structure is binary, eliminating a leaf node also eliminates all other nodes in the branch with that same rank. For instance, in Fig. 2-c, eliminating leaf node *C* will also eliminate leaf node *D*. In the CART model, this operation turns the decision node I_4 into a leaf node. Drawings of the tree's reduced structure after pruning are made using the response vector shown in Fig. 2-d. Based on the resulting tree, the chromosomal fit may be computed after applying each chromosome to the original tree structure. The training error criteria has been used in the suggested technique as a fitness assessment function in the GA. Training examples are applied to the tree that corresponds to the response vector for this purpose, and the output labels of the tree are compared with the actual labels for these cases. The response vector's fit may then be determined using the following formula:

$$Fitness = \frac{E}{N} \tag{9}$$

The number of training instances (E) for which the tree output differs from the example's actual label in the equation above is equal to N, the total number of training examples. The suggested approach optimizes each tree component in the RF structure using GA. In this process, we try to obtain a structure for each CART model that can minimize Equation (9). GA is an optimization strategy based on natural evolution that tries to discover the optimal answer to a problem by creating an initial population of candidate solutions and improving them during different cycles. During each GA cycle, a number of chromosomes are first selected as parents and merged using a combination operator to yield a set of new responses (offspring). Next, the mutation operator is applied to some chromosomes of the population so that the problem space can be searched more effectively. After creating each new chromosome in the population, its fitness is calculated based on Equation (9) to calculate the quality of the solution.

Chromosomes of superior quality are carried over to the next algorithmic cycle at the conclusion of each cycle, forming a new generation. Until one of the algorithm's termination requirements is not satisfied, this procedure is repeated. Considering the described optimization model, the GA performs the optimization process of each tree in the RF based on the following steps:

Step 1) an initial population of N chromosomes is created randomly.

Step 2) the fitness of each chromosome is calculated based on Equation (9).

Step 3) Based on the roulette wheel algorithm, $0.8 \times N$ number of parent chromosomes are selected.

Step 4) Parent chromosomes are merged together based on binary combination to create child chromosomes.

Step 5) each child chromosome is a mutation with probability N_m. During the mutation process, m random bits are selected in the chromosome and these bits are randomly replaced.

Step 6) the fitness of the offspring chromosomes created in the current population is calculated by Equation (9).

Step 7) the child chromosomes created in the current population are merged with the previous population and the resulting set is sorted in ascending order based on the fitness. Then, the first N chromosome in this collection is selected as the new generation population.

Step 8) the chromosome with the least fitness in the new generation is kept as the best discovered solution.

Step 9) If one of the following termination conditions is met; Step 10 is executed and in this case the algorithm is repeated from Step 3. The termination conditions of the GA in the proposed method are:

- If the minimum value of fitness does not change for N_t consecutive generations.
- If the number of iterations of the algorithm reaches a predetermined fixed value G.

Step 10) Add the optimized CART model based on the best chromosome to the RF classifier.

The proposed classification model includes 100 components of the CART tree, which uses the above steps to optimize each one. Each CART component in this classification model is formed and optimized based on a random subset of training data. Finally, after optimizing all CART models, the obtained forest is used to predict academic performance in new samples. In this case, the prediction of the target label for each test sample is done using the majority voting strategy among the optimized CART models.

3. Results and discussion

This section reviews and analyzes the research findings. MATLAB version 2020a is the coding program utilized for this purpose. For a broad evaluation on the performance of the proposed method, it was examined in three modes:

- Proposed: refers to the case that the method introduced in section 2 is used for performance prediction.
- Prop (All Features): is the same as the proposed method, with the difference that the feature selection method is not considered. In other words, all 30 features have been used for prediction. The goal of this comparison was to assess the effect of the features selection step on the performance of the proposed method.
- RF: Refers to an operational mode of the proposed method that the optimization of the decision tree components is omitted and a conventional RF model is used for prediction. By comparing the proposed method with this mode, the effect of optimizing learning components on the prediction performance can be evaluated.

In addition to above cases, the proposed method was compared with models in Lau et al. [19], Sekeroglu et al. [20], and Hussain & Khan [30]. We selected ANN in Ref. [19] as a comparative method because ANN is a well-established and powerful technique for student performance prediction, often achieving high accuracy in classification tasks. Our method can be compared against ANN in

Ref. [19] to determine if it achieves similar or potentially better classification accuracy while offering additional benefits, such as interpretability or efficiency. Also, the study by Sekeroglu et al. (2010) explored various machine learning algorithms, including decision trees (used for classification in this context) and Naive Bayes [20]. We focus on the decision tree aspect of their work for comparison. Our classification-based method can be evaluated against decision trees to see if it achieves similar or better results while potentially offering advantages in specific areas. Also, comparing the proposed method with the presented decision tree model in Ref. [30], can show the effect of utilizing multiple optimized learners, compared to using single learners on prediction efficiency.

To further generalize the findings, the k fold cross validation approach ($k = 10$) was used. 90 % of the data were utilized for training and 10 % for testing in each iteration. Additionally, the models' performance was examined using the various evaluation indicators, which include F-Measure, Accuracy, Precision, and Recall. As shown in the study methodology, the suggested technique identified markers of academic achievement by combining a strategy based on knowledge acquisition criterion and Laplacian score. Finally, the SFS strategy was used to select the optimal features. In Fig. 3, the results related to the selection of the indicators with the most relationship with the academic performance of the students are presented. Fig. 3-a specifies how to select each feature in different iterations (10 iterations). In this figure, each column corresponds to a feature and the selected features are displayed in yellow and the unselected features are displayed in dark blue. Finally, based on the score rate, the features that were selected in at least 50 % of repetitions, and in other words, their selection rate is greater than 0.5 (as shown in Fig. 3-b), were considered as features related to academic performance.

According to the results of Fig. 3, among the 30 reviewed features, 14 features have a selection rate of less than 0.5 and are therefore considered irrelevant. As a result, the remaining 16 features were considered as features related to academic performance and were used in the prediction process.

The suggested method's average accuracy values in comparison to other approaches are shown in Fig. 4. This chart shows that, in comparison to other models, the model suggested in this research has the greatest values of this index. Therefore, this model has been most accurate in predicting academic performance than others. After this model, Hussain & Khan [30], Lau et al. [19], prop (All Features), Sekeroglu et al. [20], and RF models are respectively in the next ranks of the model with the highest accuracy and the lowest classification error. Therefore, the proposed method has been able to improve 2.25 %, 2.63 %, 4.42 %, 5.41 %, and 6.61 % of average accuracy index compared to Hussain & Khan [30], Lau et al. [19], prop (All Features), Sekeroglu et al. [20], and RF models, respectively. Higher accuracy of the proposed method compared to the prop (All Features) mode, demonstrates the effectiveness of the proposed feature selection algorithm which can improve the prediction accuracy by 4.42 %. On the other hand, 6.61 % improvement in accuracy compared to conventional RF, shows the considerable effect of proposed classifier optimization mechanism on the accuracy of the model.

Fig. 5-a shows the confusion matrix for the proposed model. Fig. 5-b and 5-c, show the same results for the prop (All Features) and conventional RF, respectively. Additionally, Fig. 5-d to 5-e show the confusion matrices for the models presented by Seleroglu et al. [20], Lau et al. [19], and Hussain & Khan [30], respectively. In this figure, the columns represent the real classification and the rows represent the classification predicted by the models. According to this figure, it can be seen that the average value of the accuracy index

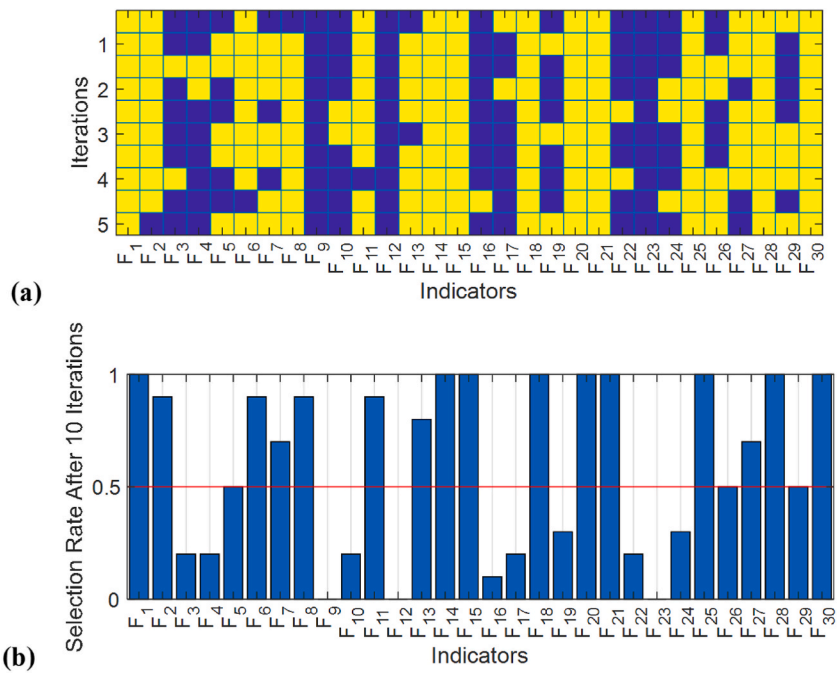


Fig. 3. Selecting the most relevant features with proposed approach (a) status of selecting each indicator during iterations, and (b) the selection rate after 10 iterations.

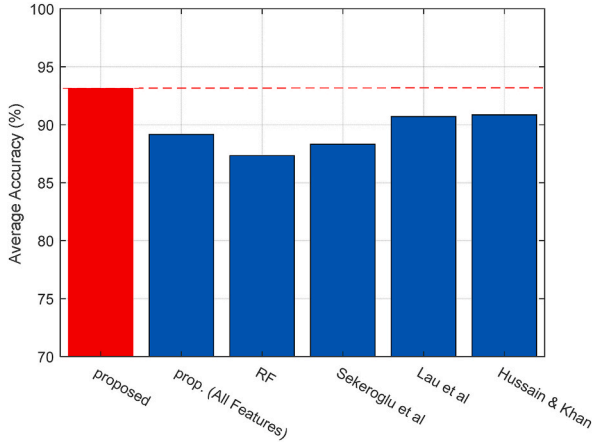


Fig. 4. Comparing the models based on the average *Accuracy* index.

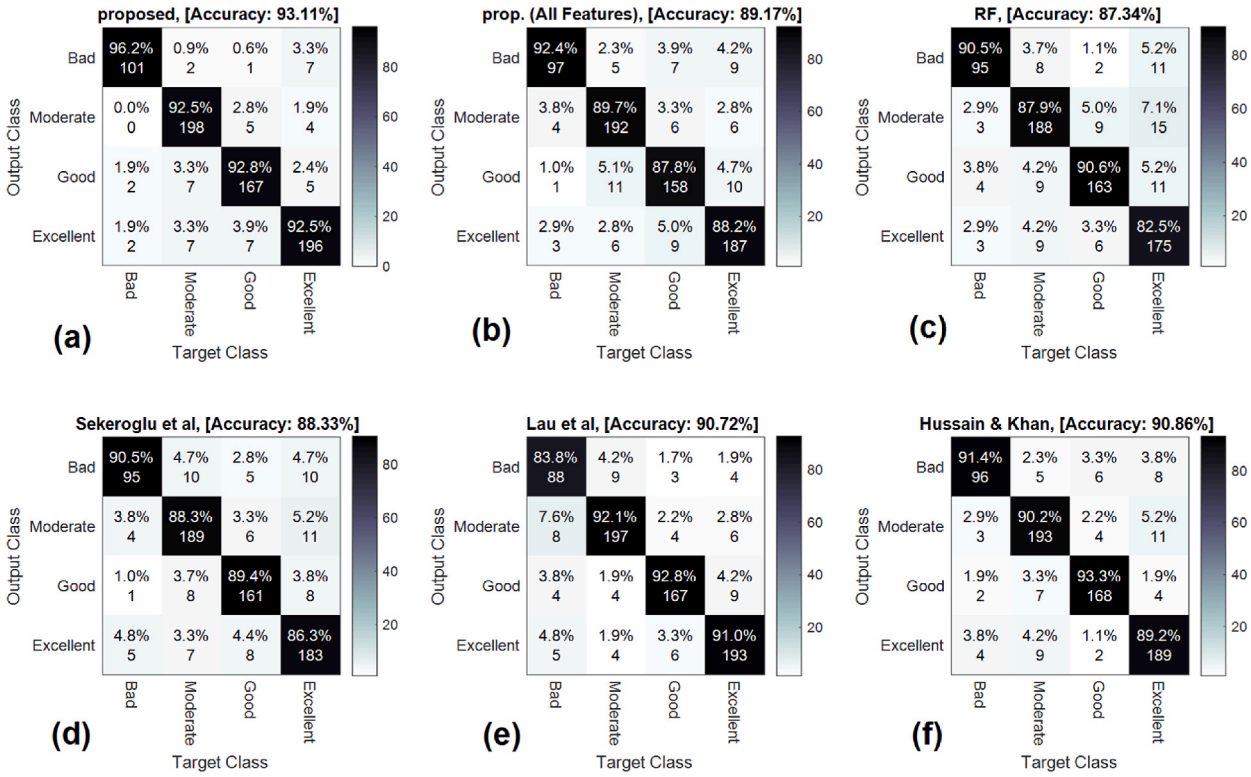


Fig. 5. The confusion matrix related to (a) the proposed method, (b) proposed method using all features, (c) the conventional RF, and the models proposed by (d) Seleroglu et al. [20], (e) Lau et al. [19], and (f) Hussain & Khan [30].

related to the proposed model is equal to 93.11 %, which is more than the corresponding values in other models. On the other hand, the percentage values on the main diameter indicate the performance of the models by categories. For example, for the bad category, the accuracy value of the proposed method is 96.2 %, which means that the proposed method was able to correctly predict 96.2 % of the data. The accuracy values related to the proposed method for Bad, Moderate, Good and Excellent categories are equal to 96.2 %, 92.5 %, 92.8 % and 92.5 %, respectively, which are all higher than the corresponding values in other models. Therefore, the proposed method is not only generally more accurate than others in predicting academic performance, but also has the highest accuracy in all categories.

The Precision, Recall, and F-Measure index values for each model broken down by class are shown in Fig. 6. This figure shows that, both per class (Fig. 6-a, 6-b, and 6-c) and overall (Fig. 6-d), the suggested technique has the greatest values for the Precision, Recall, and F-Measure indices. Therefore, in comparison to other approaches, these indices also validate the superiority of the suggested

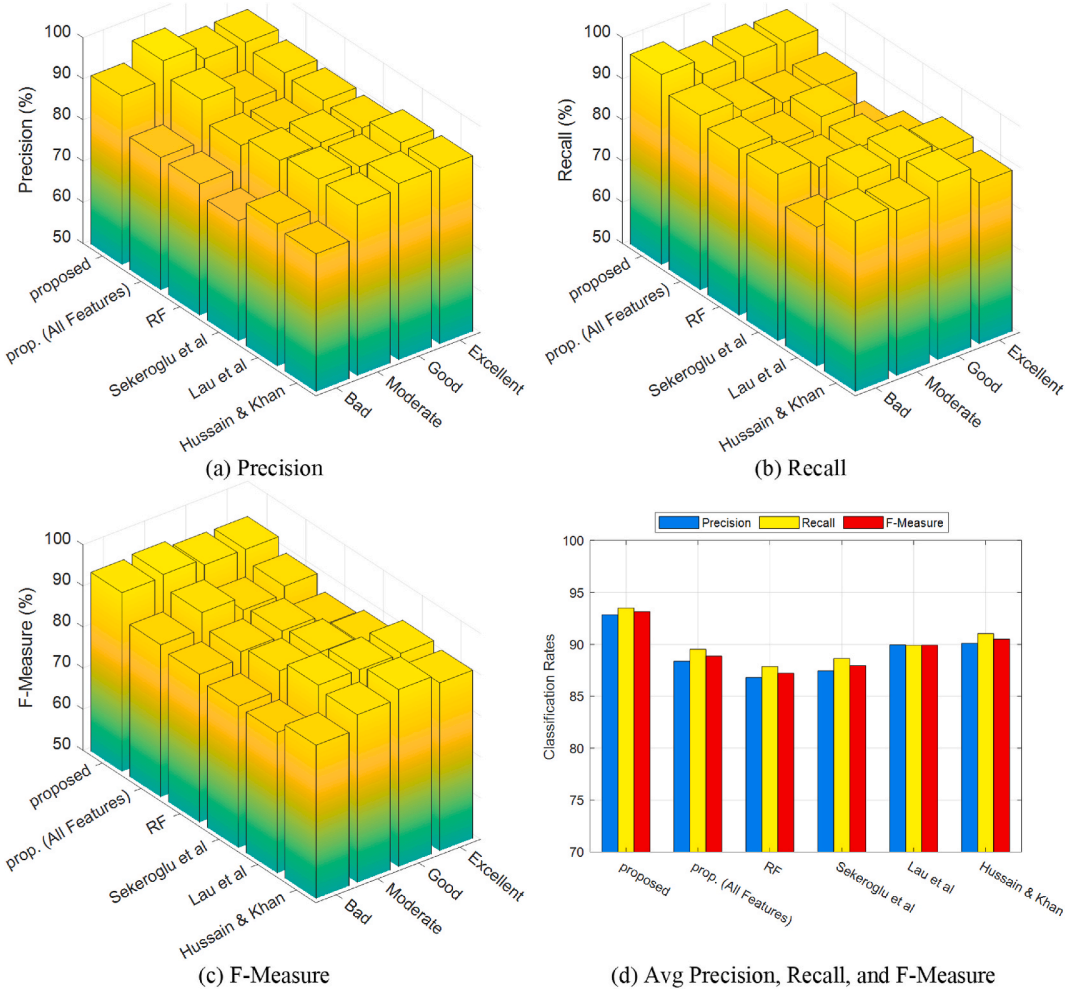


Fig. 6. The values of classification rates for each class, (a) Precision, (b) Recall, (c) F-Measure, and (d) the average of these criteria for all classes.

strategy in forecasting academic success. Conversely, the findings indicate that the RF model has the lowest accuracy level due to its low index values.

The receiver operating characteristics (ROCs) for each model are shown in Fig. 7. The True Positive Rate (TPR) is shown vertically, whereas the False Positive Rate (FPR) is displayed horizontally. Different criteria are used to draw these points. For each input item, the final output of the model calculates a score between 0 and 1. Given that it has the greatest TPR value and the lowest FPR value, the point on the graph with coordinates of (0, 1) in this figure will perform the best in terms of classification. The ideal classification is represented by this point. In this figure, AUC or area under the curve values for each model are also displayed.

This figure shows that the suggested approach has the lowest FPR value and the greatest TPR value. The suggested method's AUC value of 0.9716 is greater than the similar value in previous models. A minimum increase of 1.84 % compared to Hussain & Khan [30] and a maximum improvement of 4.15 % compared to the conventional RF are shown by ROC curve analysis. Therefore, the ROC analysis also confirms the superiority of the proposed method over other models. After this model, Hussain & Khan [30] is placed in the next rank of the best performance of models in predicting academic performance. Table 3 shows the values of evaluation indices including Precision, Recall, and F-Measure related to all models.

The conducted experiments, proved the effectiveness of the proposed feature selection and tree optimization techniques on the prediction efficiency. In the continuation of this section, we will examine the effect of two parameters on the performance of the proposed method. These parameters include "number of tree components" and "rate of optimized components" in the proposed random forest model. The number of tree components parameter refers to the number of decision trees that construct the proposed random forest classifier. We examined this parameter by considering values 5, 10, 25, 50, and 100; and then calculating the prediction accuracy of the model with each setting. On the other hand, the rate of optimized components parameter refers to the ratio of decision tree components in the random forest classifier that are optimized using GA. For this parameter, various rates of 0.2, 0.4, 0.6, 0.8 and 1 has been considered. For example, rate of 0.6 for optimizing components means that 60 % of decision trees in the random forest classifier are optimized using GA and the remaining 40 % are constructed based on the conventional tree formation approach. Also,

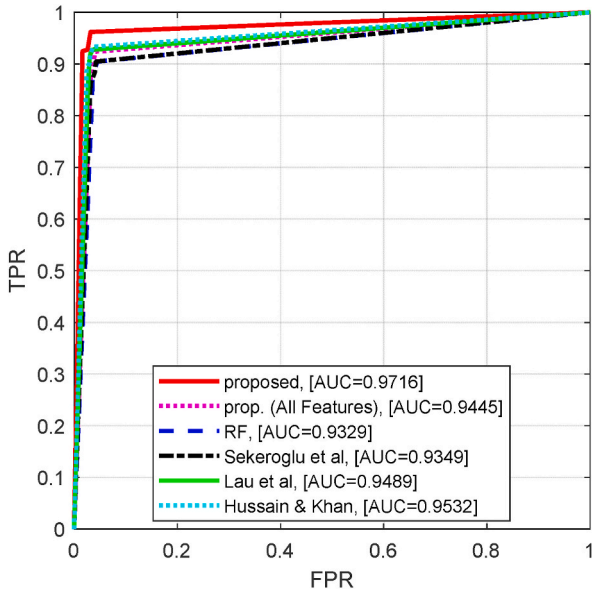


Fig. 7. Performance of all models based on ROC curves.

Table 3
Evaluation indices related to all models.

Model	Precision	Recall	F-Measure	Accuracy (%)	AUC
Proposed	92.8403	93.4861	93.1385	93.1083	0.9716
prop (All Features)	88.3771	89.5215	88.8642	89.1702	0.9445
RF	86.7944	87.8573	87.2167	87.3418	0.9329
Sekeroglu et al. [20]	87.4410	88.6398	87.9331	88.3263	0.9349
Lau et al. [19]	89.9482	89.9203	89.9288	90.7173	0.9489
Hussain & Khan [30]	90.1030	91.0249	90.5091	90.8579	0.9532

selection of tree components for optimization was performed randomly. Table 4 shows the effect of these parameters on the accuracy of the model.

It should be noted that in this experiment a similar 10-fold cross validation was considered. According to the results, the best parameter combination which leads to the highest accuracy in the proposed model is (25,1.0). This means that highest accuracy is achieved when the proposed RF model includes 30 decision tree classifiers and all of these trees are optimized using GA. The findings demonstrate that using more than 30 trees leads to decreasing in accuracy (and also, a slower classifier) which is the result of model overfitting. Also, using a smaller forest (less than 30 decision trees) show the evidences of under-fitting which means that the model cannot fit efficiently with the data.

4. Conclusion

Predicting students' academic performance is one of the most important and interesting topics for researchers in the field of educational management. One of the significant issues in academic performance is the correct prediction of students' academic performance and timely action and advice to students at risk of academic failure. For this purpose, in this study, a combined strategy was used to select the indicators related to academic performance. This strategy firstly uses two criteria of information gain and Laplacian score to evaluate the importance of each index. Then, by combining the values of these two criteria, it ranks the indicators. Finally, the SFS strategy is used to select the optimal features. Then, a combined strategy based on optimized RF is used to predict academic performance. This strategy, by using several decision tree classification models, improves the accuracy of the proposed model in solving the multi-class classification problem. The results showed that among the 30 reviewed features, 16 features have a selection rate greater than 0.5 and as a result, and they were used as features related to academic performance in the prediction process. According to the findings, when compared to other methods, the suggested approach has the greatest average values for the indices of accuracy, precision, recall, and F-measure. However, the examination of the categories revealed that the Bad, Moderate, Good, and Excellent categories all had higher indicator values associated with the suggested approach than the comparable values in other models. As a result, the suggested approach has the best accuracy both overall and across categories. ROC curve study further supported the suggested method's superiority. The suggested method's AUC value of 0.9716 is greater than the similar value in previous models. Analysis of ROC curves shows at least 2.39 % improvement compared to Lau et al. [19], and maximum 4.15 %

Table 4
The effect of the number of trees and optimization rate on the accuracy of the proposed classification model.

		Number of tree components				
		5	10	25	50	100
Rate of optimized components	**0.2**	82.49 %	83.90 %	89.66 %	90.03 %	87.12 %
	0.4	86.27 %	84.01 %	90.14 %	90.16 %	87.36 %
	0.6	87.17 %	84.11 %	91.85 %	90.89 %	88.81 %
	0.8	87.29 %	85.16 %	92.17 %	90.92 %	89.22 %
	1.0	88.01 %	88.56 %	**93.11 %**	90.15 %	89.15 %

improvement compared to RF. The results also showed that after the proposed method, the models of Lau et al. [19], prop (All Features), Sekeroglu et al. [20], and RF were ranked next in terms of classification accuracy.

Data availability

All data generated or analyzed during this study are included in this published article.

CRediT authorship contribution statement

Mengyao Chen: Investigation. **Zhengqi Liu:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e32570.

References

[1] E. Matusov, E. Matusov, Necessities-based society and technological education. Envisioning Education in a Post-Work Leisure-Based Society: A Dialogical Approach, 2020, pp. 21–46.
[2] E. Ortiz-Ospina, M. Roser, Government spending. Our World in Data, 2023.
[3] J.L. Rastrollo-Guerrero, J.A. Gómez-Pulido, A. Durán-Domínguez, Analyzing and predicting students' performance by means of machine learning: a review, Appl. Sci. 10 (2020) 1042.
[4] S. Iglesias-Pradas, Á. Hernández-García, J. Chaparro-Peláez, J.L. Prieto, Emergency remote teaching and students' academic performance in higher education during the COVID-19 pandemic: a case study, Comput. Hum. Behav. 119 (2021) 106713.
[5] H. Waheed, S.-U. Hassan, N.R. Aljohani, J. Hardman, S. Alelyani, R. Nawaz, Predicting academic performance of students from VLE big data using deep learning models, Comput. Hum. Behav. 104 (2020) 106189.
[6] A.Y. Huang, O.H. Lu, J.C. Huang, C. Yin, S.J. Yang, Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs, Interact. Learn. Environ. 28 (2020) 206–230.
[7] H.A. Mengash, Using data mining techniques to predict student performance to support decision making in university admission systems, IEEE Access 8 (2020) 55462–55470.
[8] H. Zeineddine, U. Braendle, A. Farah, Enhancing prediction of student success: automated machine learning approach, Comput. Electr. Eng. 89 (2021) 106903.
[9] M. Akour, H. Alsghaier, O. Al Qasem, The effectiveness of using deep learning algorithms in predicting students achievements, Indonesian Journal of Electrical Engineering and Computer Science 19 (2020) 387–393.
[10] N. Tomasevic, N. Gvozdenovic, S. Vranes, An overview and comparison of supervised data mining techniques for student exam performance prediction, Comput. Educ. 143 (2020) 103676.
[11] H. Khajavi, A. Rastgoo, Improving the prediction of heating energy consumed at residential buildings using a combination of support vector regression and meta-heuristic algorithms, Energy 272 (2023/06/01/2023) 127069.
[12] A. Rastgoo, H. Khajavi, A novel study on forecasting the airfoil self-noise, using a hybrid model based on the combination of CatBoost and Arithmetic Optimization Algorithm, Expert Syst. Appl. 229 (2023/11/01/2023) 120576.
[13] A. Alhadabi, A.C. Karpinski, Grit, self-efficacy, achievement orientation goals, and academic performance in University students, Int. J. Adolesc. Youth 25 (2020) 519–535.
[14] M. Ghadiri, A.A. Rassafi, B. Mirbaha, The effects of traffic zoning with regular geometric shapes on the precision of trip production models, J. Transport Geogr. 78 (2019/06/01/2019) 150–159.
[15] H. Khajavi, A. Rastgoo, Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms, Sustain. Cities Soc. 93 (2023/06/01/2023) 104503.
[16] T.T. Dien, S.H. Luu, N. Thanh-Hai, N. Thai-Nghe, Deep learning with data transformation and factor analysis for student performance prediction, Int. J. Adv. Comput. Sci. Appl. 11 (2020).
[17] R. Ghorbani, R. Ghousi, Comparing different resampling methods in predicting students' performance using machine learning techniques, IEEE Access 8 (2020) 67899–67911.
[18] A. Jain, P. Chaturvedi, A. Tambe, Prediction and analysis of student performance using hybrid model of multilayer perceptron and random forest, in: 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), 2018, pp. 1–7.

[19] E. Lau, L. Sun, Q. Yang, Modelling, prediction and classification of student academic performance using artificial neural networks, SN Appl. Sci. 1 (2019) 1–10.
[20] B. Sekeroglu, K. Dimililer, K. Tuncal, Student performance prediction and classification using machine learning algorithms, in: Proceedings of the 2019 8th International Conference on Educational and Information Technology, 2019, pp. 7–11.
[21] S. Jayaprakash, S. Krishnan, V. Jaiganesh, Predicting students academic performance using an improved random forest classifier, in: 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), 2020, pp. 238–243.
[22] L.H. Alamri, R.S. Almuslim, M.S. Alotibi, D.K. Alkadi, I. Ullah Khan, N. Aslam, Predicting student academic performance using support vector machine and random forest, in: Proceedings of the 2020 3rd International Conference on Education Technology Management, 2020, pp. 100–107.
[23] Ş. Aydoğdu, Predicting student final performance using artificial neural networks in online learning environments, Educ. Inf. Technol. 25 (2020) 1913–1927.
[24] F. Ofori, E. Maina, R. Gitonga, Using machine learning algorithms to predict students' performance and improve learning outcome: a literature based review, J. Inf. Technol. 4 (2020) 33–55.
[25] S. Batool, J. Rashid, M.W. Nisar, J. Kim, T. Mahmood, A. Hussain, A random forest students' performance prediction (rfspp) model based on students' demographic features, in: 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), 2021, pp. 1–4.
[26] S.K. Ghosh, F. Janan, Prediction of student's performance using random forest classifier, in: Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, 2021, pp. 7–11. Singapore.
[27] M. Kumar, K. Bajaj, B. Sharma, S. Narang, A comparative performance assessment of optimized multilevel ensemble learning model with existing classifier models, Big Data 10 (2022) 371–387.
[28] A. Alam, A. Mohanty, Predicting students' performance employing educational data mining techniques, machine learning, and learning analytics, in: International Conference on Communication, Networks and Computing, 2022, pp. 166–177.
[29] A. Asselman, M. Khaldi, S. Aammou, Enhancing the prediction of student performance based on the machine learning XGBoost algorithm, Interact. Learn. Environ. 31 (2023) 3360–3379.
[30] S. Hussain, M.Q. Khan, Student-performulator: predicting students' academic performance at secondary and intermediate level using machine learning, Annals of data science 10 (2023) 637–655.
[31] B.K. Singh, K. Verma, A.S. Thoke, Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification, Int. J. Comput. Appl. 116 (19) (2015).
[32] S. Jadhav, H. He, K. Jenkins, Information gain directed genetic algorithm wrapper feature selection for credit rating, Appl. Soft Comput. 69 (2018) 541–553.
[33] R. Huang, W. Jiang, G. Sun, Manifold-based constraint Laplacian score for multi-label feature selection, Pattern Recogn. Lett. 112 (2018) 346–352.
[34] Y. Cui, Optimizing decision trees for English teaching quality evaluation (ETQE) using artificial Bee Colony (ABC) optimization, Heliyon 9 (8) (2023) e19274.
[35] T. Daniya, M. Geetha, K.S. Kumar, Classification and regression trees with gini index, Adv. Math.: Scientific Journal 9 (10) (2020) 8237–8247.