

International Journal of Business Intelligence and Data Mining

ISSN online: 1743-8195 - ISSN print: 1743-8187
<https://www.inderscience.com/ijbidm>

Presenting a model to reduce students' academic drop by using analytical comparison of machine learning algorithms in data mining (case study of Shahed University)

Mozhdeh Salari, Reza Radfar, Mahdi Faghihi

DOI: [10.1504/IJBIDM.2025.10067362](https://doi.org/10.1504/IJBIDM.2025.10067362)

Article History:

Received:	10 October 2023
Last revised:	31 July 2024
Accepted:	06 August 2024
Published online:	14 July 2025

Presenting a model to reduce students' academic drop by using analytical comparison of machine learning algorithms in data mining (case study of Shahed University)

Mozhdeh Salari, Reza Radfar* and Mahdi Faghihi

Department of Information Technology Management,
Science and Research Branch,
Islamic Azad University,
Tehran, Iran

Email: msalari@shahed.ac.ir

Email: r.radfar@srbiau.ac.ir

Email: mahdifaghihi59@gmail.com

*Corresponding author

Abstract: This research aims to find factors that predict undergraduate student educational performance. To achieve this goal, the study follows the CRISP-DM method. This study used various classification algorithms to predict the total GPA. The data used in this research are records of undergraduate students from 2012 in Shahed University. We used 1,468 data records in data mining. We used the Rapidminer 9.9 tool for modelling. This study also considers four feature selection techniques. This study used K-fold cross-validation to split the data. This study introduced the best model for predicting students' academic performance. In two-class modelling, we get better results and higher accuracy than four-class modelling. This research found the random forest algorithm best for predicting students' performance. It achieved 94.17% accuracy with two classes. The random forest results show a higher chance of success in students with a higher 1st semester GPA.

Keywords: student performance prediction; data mining; machine learning; data science applications in education.

Reference to this paper should be made as follows: Salari, M., Radfar, R. and Faghihi, M. (2025) 'Presenting a model to reduce students' academic drop by using analytical comparison of machine learning algorithms in data mining (case study of Shahed University)', *Int. J. Business Intelligence and Data Mining*, Vol. 27, No. 1, pp.1–39.

Biographical notes: Mozhdeh Salari is a PhD student in Information Technology Management, Islamic Azad University, Science and Research Branch, Tehran, Iran. He received a Bachelor's degree in Computer Engineering (Software) from Islamic Azad University, Meybod branch in 2014 and a Master's degree in Information Technology Management from Payam Noor University, Tehran branch in 2014. He has articles in the field of information technology management in journals. He is more active in information technology, innovation, education and machine learning.

Reza Radfar is a Full Professor of Islamic Azad University of Science and Research in Tehran, Iran. He is the Editorial Board of *Innovation Management Journal* and has 62 conference articles and 145 journal articles in domestic Journals. In addition, he has also participated in the publication of 11 ISI articles. During 18 years, he has had direct scientific cooperation with 51 different researchers. His published articles are mostly on innovation, research and development, technology and information technology.

Mahdi Faghihi is an Associate Professor at Islamic Azad University, Science and Research Unit, Tehran, Iran. His education is in the Doctoral level of Management – Decision-Making and Policy-Making, and the subject of his Doctoral Thesis is the presentation of the electronic government model in Iran in 2025. His scientific records include authoring three books, more than 10 scientific and research articles and ISI in the field of information and communication technology, authoring 15 scientific and promotional and conference articles (presented in international and domestic journals and conferences), authoring and editing more 100 research and policy reports have been published.

1 Introduction

Educational data mining (EDM) has greatly improved educational outcomes and processes (Stinar et al., 2024). EDM can help. It can find insights in academic data. This can predict future performance and highlight areas needing improvement (Alamgir et al., 2024). EDM aims to find patterns in educational data. It uses machine learning to do this. Therefore, using data mining to find factors that affect students' performance is useful (Alboaneen et al., 2022). EDM can analyse student learning. It can predict academic performance and identify students at risk of failure (Waheed et al., 2020). Many existing articles do not explicitly define performance (Hu et al., 2017). Lee, for example, defines performance as "the measure of a student's competence for future courses" (Lei and Li, 2015). Muthukrishnan et al. (2017) noted that most articles predicted final course grades. Universities must now improve their ability to use data. They need to predict and ensure student success (Bernacki et al., 2020). Educational big data is a strategic asset. It is vital for advancing educational reform (Bai et al., 2021). The field of data mining in education has changed in the past decade (Romero and Ventura, 2020). A major challenge for higher education admins is predicting a student's success (Arcinas et al., 2021). EDM can predict students' final results (El Aissaoui et al., 2020). A key issue in this research is predicting students' future GPA before their final exams. By making academic predictions, we can help students. It can improve their performance and reduce dropouts (Batool et al., 2023). To accurately predict a student's performance, we must understand many factors. They all affect student results and progress (Alshanhqiti and Namoun, 2020). Increasing the graduation rate, identifying at-risk students, and managing university resources. Also, optimising performance and renewing the curriculum (Ampadu, 2023), creating a behavioural model for students, and making predictions based on it, working with large datasets, and learning about algorithms (Garcia and Skrita, 2019), learning (Garcia and Skrita, 2019), evaluating managers in adopting

educational measures, and monitoring the learning process (Debang and Hassan, 2023) are among the applications of EDM. Classification is the best data mining technique. It can help students and faculty improve education by identifying strong and weak individuals at the start of the semester/year (Saa, 2016). The classification shows a connection between the internal exam and final grades (Sarker et al., 2024). In the future, big data and the internet of things will make a student's info available from birth (Al-Emran et al., 2020; Romero and Ventura, 2020). We can combine data from different sources to personalise and improve students' learning (Ding et al., 2019). We ought to make sure that the data we have are as useful as possible while still upholding student privacy (Stinar et al., 2024).

1.1 Statement of the problem

A major issue in academics is predicting students' performance. It is also vital to quickly counsel those at risk of failing (Asif et al., 2017). Early detection enables interventions that improve student performance (Bellaj et al., 2024). EDM uses data mining methods to analyse educational datasets (Kaur and Dahiya, 2023). Academic institutions have databases of student demographics and records (Rostami et al., 2015). Choosing data types, models, and optimal criteria will increase transparency, interpretability, and accessibility for teachers and students (Gunasekara and Saarela, 2024). Aldowah et al. (2019) show that some EDM and learning analytics (LA) techniques can help with teaching challenges (Aldowah et al., 2019). With many processes now electronic, schools have databases. They hold a lot of data on people's educational backgrounds (Romero and Ventura, 2013). From a knowledge management view, data mining aims to find hidden knowledge in raw data (Romero and Ventura, 2007). EDM develops methods to analyse educational data. It aims to improve education and decision-making (Romero and Ventura, 2013; Hasani and Bazrafshan, 2018). EDM can personalise education, boost efficiency, and cut costs (Zhang et al., 2010). Also, in education, we must evaluate programs, predict results, and create a plan for future changes (Batirovna, 2023). The recent decline in education in most developing countries has led to research for solutions to some of the problems (Ogunde and Ajibade, 2014). Researchers in the academic world have conducted various analyses on the entrance exam and the performance of students at the normal level, but they have not given much attention to the relationship between them and the final graduation grades (Ogunde and Ajibade, 2014). Educational administrators can then install support initiatives for low-performing students. They should also encourage high-performing students to continue excelling (Sarker et al., 2024). The lack of information structure and their large volume are not very effective for students and professors. Through EDM, we can discover existing patterns. This can raise educational issues during students' studies. It can improve the educational system and course delivery. This study's main problem is to find the factors that predict students' academic performance at Shahed University. In this regard, researchers discover the best machine learning algorithm for predicting students' academic performance by validating and evaluating various modelling methods. There has been little research on factors affecting

students' performance in Iran. So, we must research this field to fill the gap with other countries. This research aims to find factors that predict students' academic performance at Shahed University. It will use data mining with classification models.

1.2 Research questions

1.2.1 Main question

What factors predict undergraduate student performance, using two- and four-class models?

1.2.2 Sub-questions

- 1 Which classification algorithm, two-class or four-class, is better at predicting students' performance?
- 2 What methods have been used to predict students' performance?
- 3 Which prediction models do data mining classification techniques identify using the predicted variables?
- 4 How is the validation of the developed model for Shahed University students?

2 Research literature

Between 2018 and 2021, researchers identified 'academic performance prediction', 'machine learning', 'process mining', 'participation', 'KNN', 'J48', and 'Bayesian network' as engine themes. Researchers discovered that the themes of 'deep learning (DL)' and 'emotions' are emerging themes for future studies to shape the future of education based on sustainable goals (Boztaş et al., 2024). An extensive review by Batool et al. (2023), this review found that students' academic and demographic factors are the best predictors of their academic performance (Batool et al., 2023). Triayudi et al. (2024) study uses decision trees, neural networks, and simple Bayes. The results showed that the framework accurately predicts student performance. Xiaowei He's article shows that the proposed method has higher accuracy and a shorter detection time in detecting students' online learning status (He, 2024). Jayaprakash et al. (2020) found that their 'improved random forest (RF) classification' method outperformed RF and Naive Bayes (NB) algorithms in both classification and prediction tasks (Jayaprakash et al., 2020). The research findings of Mishra et al. (2014) showed that RF trees are better than the j48 decision tree algorithm (Mishra et al., 2014). Hussain et al. (2018) found that the RF algorithm outperformed other classifiers in accuracy and error rates. Marjan's (2023) research found that RF and support vector machine (SVM) models are better at predicting students than other models. In the article by Alboaneen et al. (2022) Academic factors had a greater effect on students' performance than demographic factors.

Table 1 The results of the research literature review and present research on classification algorithms

Data mining algorithm	DT	RF	NB	KNN	SVM	ANN	PART	Logit boost	Line RL	LR	Correlation laws	Clustering and association rules	Group method	Accuracy
Hoyos and Caicedo-Castro (2024)					*	*								
Batool et al. (2023)		*				*								
Marjan et al. (2023)	*	*	*	*	*	*				*				
Abdelmagid and Qahmash (2023)		*		*	*	*				*		*		
Manoharan et al. (2023)	*	*		*		*				*				
Alghamdi and Rahman (2023)	*	*	*											99.34%
Alboaneen et al. (2022)		*		*	*	*			*					
Yağcı (2022)	*		*	*	*									
Dabbade et al. (2021)	*				*	*			*					70–75%
Cruz-Jesus et al. (2020)		*		*	*	*			*					83.44%
Sokkhey and Okazaki (2020)	*	*	*	*	*	*				*				50–81%
Hung et al. (2020)	*	*												
Rebai et al. (2020)	*	*								*				
Jayaprakash et al. (2020)	*	*	*								*			
Zulfiker et al. (2020)	*	*				*								
Musso et al. (2020)						*								
Waheed et al. (2020)						*								85%
Salal and Abdullaev (2019)	*		*	*	*	*								
Turabieh (2019)	*		*	*		*								
Xu et al. (2019)	*			*	*	*								
Ajibade et al. (2019)	*		*	*	*								*	91.5%
Ahmad and Shahzadi (2018)						*								85%
Hasani and Bazrafshan (2018)	*		*				*							

Table 1 The results of the research literature review and present research on classification algorithms (continued)

<i>Data mining algorithm</i>	<i>DT</i>	<i>RF</i>	<i>NB</i>	<i>KNN</i>	<i>SVM</i>	<i>ANN</i>	<i>PART</i>	<i>Logit boost</i>	<i>Line RL</i>	<i>LR</i>	<i>Correlation laws</i>	<i>Clustering and association rules</i>	<i>Group method</i>	<i>Accuracy</i>
Hussain et al. (2018)	*	*	*						*					
Umer et al. (2017)	*	*	*	*						*				
Khasanah (2017)	*		*											
Fernandes et al. (2017)					*									
Rahman et al. (2017)	*		*			*								
Asif et al. (2017)	*													
Hoffait and Schlyns (2017)						*				*				92.34%
Mueen et al. (2016)	*		*			*								86%
Amrieh et al. (2015)	*		*	*										92.34%
Yehuala (2015)	*		*											
Punlunjek and Rachburee (2015)	*													
Osmanbegović et al. (2014)	*	*										*		71%
Harwati (2014)														
Mishra et al. (2014)	*	*												
Kabakchieva (2013)	*		*	*										
Oskoueï and Askari (2014)	*	*	*			*		*			*			60–75%
Pandey and Pal (2011)			*											96%
Nghe et al. (2007)	*		*											
Han and Kamber (2006)	*										*			
Present research	*	*	*	*	*	*								94.17%

The midterm exam score was the best predictor (Alboaneen et al., 2022). Hussain et al. (2024) said their model had an 88.6% accuracy. This was better than existing studies. This study accurately estimates students' final grades using current academic records. The study by Asif et al. (2017) emphasised the need for timely help for weaker students. It also recommended supporting high-achieving students (Asif et al., 2017). Sökkhey and Okazaki (2020) proposed a hybrid approach. It combined PCA with four ML algorithms. The study showed that hybrid techniques can improve classification model accuracy (Sökkhey and Okazaki, 2020). In Hoyos and Caicedo-Castro's (2024) article, the Algorithms showed an average performance of more than 80%. It is noteworthy that ANN performed better than SVM in the analysed dataset. Ajibade et al. (2019) studied ways to boost academic performance. They used group-based classification methods. Their approach achieved an impressive 91.5% accuracy. This shows its potential (Ajibade et al., 2019). Rahman et al. (2017) proposed using feature selection to improve the accuracy of classifying students' academic performance (Rahman et al., 2017). Abdelmagid and Qahmash's (2023) study revealed some insights. It found that linear regression was best for predicting academic performance. The SVM algorithm was less effective (Abdelmagid and Qahmash, 2023).

The research of Debhadeh et al. (2021) found that the decision tree algorithm had the best fit, with an 83.44% accuracy rate. These findings suggest that larger datasets could improve model training (Dabhade et al., 2021). shows that the RF model is effective at predicting students' performance in programming courses (Hung et al., 2020). Punlumjeak and Rachburee (2015) found a way to predict students' total academic success based on their scores in different subjects (Punlumjeak and Rachburee, 2015). In the research of Manoharan et al. (2023). A mix of statistics and machine learning shows the innovative ways to classify students (Manoharan et al., 2023). Qi (2024) generated a list of online cultural education resources in their research. Based on predicted interest, it ranked recommendations from most to least. The experiments show this method works better for recommending online Chinese teaching resources. Waheed's research (2020) showed that student participation in education does not affect academic performance. Also, the researchers concluded that DL models can be effective in predicting student performance early (Waheed et al., 2020). In the study of Amrieh et al. (2015). The findings stress a link between learners' behaviours and their academic success. They also emphasise that behaviours shape educational outcomes (Amrieh et al., 2015). In Alghamdi and Rahman's (2023) study, the NB model had a 99.34% accuracy rate. The RF model was next at 98.7% (Alghamdi and Rahman, 2023). The use of larger datasets leads to better training of advanced models (Dabhade et al., 2021). In Xiao (2024) research, he selects results based on features and uses the K-nearest neighbour classification (KNN) algorithm to classify educational digital resources. The simulation results show that the proposed method is better. It is more accurate and faster at classifying. Rabaani et al. (2023) found that a neural network model has the highest accuracy, 96%, in predicting customer interaction levels (Rabaani et al., 2023). The review of previous studies used in this research showed that the decision tree algorithm was the most common method. In this research, the second most-used technique highlighted is the Naïve Bayes algorithm, following this trend. The next ranks are for RF, artificial neural network (ANN), SVM, and KNN algorithms. Table 1's highlighted cells show the best algorithms. They are the most accurate and efficient at predicting students' academic performance. In a recent paper by Batool et al. (2023) analysis found that the most used data mining algorithms in educational research are ANN and RF (Batool et al.,

2023). Also, Alshantiti and Namoun (2020) found that, in recent years, the use of RF algorithms in EDM research has grown (Alshantiti and Namoun, 2020). EDM research uses these features. They include: demographic info before university, previous and current educational status, internet usage, personality, and psychological and behavioural traits of students (Yağcı, 2022; Cruz-Jesus et al., 2020; Fernandes et al., 2017; Rebai et al., 2020; Musso et al., 2020; Harwati, 2014; Salal and Abdullaev, 2019; Khasanah, 2017; Punlumjeak and Rachburee, 2015; Zulfiker et al., 2020; Nghe et al., 2007; Kabakchieva, 2013; Oskouei and Askari, 2014; Waheed et al., 2020; Yehuala, 2015; Amrieh et al., 2015; Umer et al., 2017; Xu et al., 2019; Hasani and Bazrafshan, 2018; Pandey and Pal, 2011; Galla et al., 2014; Chamorro-Premuzic and Furnham, 2009; Hellas et al., 2018).

3 Research method

3.1 Introduction

This research aims to find factors that predict undergraduate student performance. It will compare two-class and four-class classification models. To achieve this, the study follows CRISP-DM. The current method collects data from student databases. It also investigates the causes of changes in academic status. Previous research and experts have found good indicators of students' performance. In modelling, classification is used to check students' performance. This study used data from 2012 on undergraduate students at Shahed University. This research used RapidMiner 9.9 data mining software. We assessed prediction accuracy using 10-fold cross-validation. Diploma grade point average, semester 1 GPA, semester 2 GPA, and graduation GPA are all from 0 to 20. After determining the model, we classified these variables according to the discretisation model. Table 2 shows the grade point average format in the two-class mode. Also, Table 3 shows the GPA format in four classes.

Table 2 Classification of features in the two-class mode

<i>Feature</i>	<i>Class</i>	<i>Classification</i>
GPA ≥ 14	Excellent	1
14 > GPA	weak	2

Table 3 Classification of features in the four-class mode

<i>Feature</i>	<i>Class</i>	<i>Classification</i>
GPA ≥ 17	Excellent	1
17 > GPA ≥ 15	Good	2
GPA = 14	medium	3
14 > GPA	weak	4

3.2 CRISP-DM process

The data mining process must be reliable and repeatable by people with little background in it. To this end, standards have been developed: KDD, SEMMA, and

CRISP-DM (Oreski et al., 2017). In this research, CRISP-DM, which stands for cross-industry standard process for data mining (Wirth and Hipp, 2000), is used.

3.2.1 Business understanding

A look at the poor performance of undergraduate students at Shahed University, The students' academic results and talks with the administrators show a problem with the students' education. A model that predicts student performance can set standards. These must attract accepted students who can graduate with high grades. This can affect the quality and grades of their future students.

3.2.2 Data understanding

The first phase of this research is to gather data. It aims to find the causes of low GPA and high dropout rates at Shahed University. Shahed University has 7,000 undergraduate students, with 5,631 student records available for analysis. Raosoft sample size calculation enabled the research to sample from this target number (Raosoft, 2004). Using this sample size calculator, a 95% confidence level, and a 5% margin of error, a sample size of 365 or more was sufficient. This study is important. Many students, especially those with a GPA below 15, have poor academic performance.

3.2.2.1 Data collection

The researcher contacted the heads of education and experts at Shahed University. He sought help to collect data from the Nad educational student information system.

3.2.2.2 Removal of outliers

This article uses Euclidean distance to remove outliers. In the modelling process using predictive algorithms in RapidMiner 9.9, the 'detect outlier' operator relies on distance-based techniques. Parameters such as 'number of neighbours' determine the neighbours considered, 'number of outliers' controls the outlier count, and the 'distance function' parameter specifies the distance type, including Euclidean distance.

3.2.2.3 Data normalisation

The Normalise operator performs data normalisation in this process. Data normalisation is a key pre-processing step. It scales or transforms data so each feature contributes equally. Machine learning algorithms need high-quality data. It helps them make a good predictive model for classification problems (Singh and Singh, 2020).

3.2.3 Data preparation process

Data cleaning is a key pre-processing step in data mining. It removes inconsistencies in the data to improve quality. We removed additional features and finally, our number of features reached 19. Next, we used two-class and four-class classification algorithms to model the data in the mining process (Rahm and Do, 2000).

Table 4 Features affecting the prediction of students’ academic performance in past and present research

<i>Qualitative – nominal</i>	<i>Feature</i>		<i>Encoding</i>
Qualitative	Male	1	Gender
	Female	2	
Numerical	In numbers		Age
Qualitative – nominal	Computer science	1	Field – orientation
	Business management	2	
	Government management	3	
	Industrial management	4	
	Electrical engineering	5	
	Electrical-control engineering	5	
	Electrical-electronic engineering	5	
	Electrical-power engineering	5	
	Electrical-communication engineering	5	
	Medical engineering	6	
	Medical-clinical engineering	6	
	Civil engineering	7	
	Civil engineering	7	
	Computer engineering	8	
	Computer engineering (hardware)	8	
	Computer engineering (computer systems architecture)	8	
Qualitative – nominal	Years entry 2021	1	Year of entering the university
	Years entry 2020	2	
	Years entry 2019	3	
	Years entry 2012 to 2018	4	

Table 4 Features affecting the prediction of students' academic performance in past and present research (continued)

Qualitative – nominal		Feature	Encoding
Qualitative – nominal	Sanjesh organisation test	1	Type of admission (entering the university)
	Transfer from other universities with the change of field	2	
	Transfer from other universities without change of field	2	
	A transfer within the university with a change of major	3	
	A change of major within the university	3	
	Full semester guests from other universities	3	
	One-course guests from other universities	3	
	The second course inside the university	4	
Qualitative – nominal	Guest base	4	Place of living
	Native	1	
Qualitative – nominal	Non-native	2	Military status
	Students in the final year of secondary or pre-university courses	1	
	Have an educational permit from the duty system	2	
	Exempted for girls (not included)	3	
	Graduated from an expert degree up to one year ago	4	
	Permanently exempted	5	
	Exempted under supervision	5	
	Other cases	6	
Qualitative – nominal			6
Qualitative – nominal			7

Table 4 Features affecting the prediction of students’ academic performance in past and present research (continued)

<i>Qualitative – nominal</i>	<i>Feature</i>	<i>Encoding</i>
Qualitative – nominal	Region one	1
	Sacrificers 5% capacity	2
	Regions 3	3
	Azad	4
	Region 2	5
	Isargan 25% capacity	6
Qualitative – nominal	Children of veterans	7
	Martyrs	8
	Witnesses	9
	Children of Azadeh veterans	10
	Children of academic staff members	11
	Combatants	12
	Azadeh's children	10
	Children of deceased veterans	7
	Regions 3	3
	Quota status	8
	Technical and engineering	1
	Computer science	2
Qualitative – nominal	Management	3
	Faculty of engineering	1
	Faculty of basic sciences	2
	Faculty of humanities	3
Qualitative – nominal	Educational group	9
	Faculty	10

Table 4 Features affecting the prediction of students' academic performance in past and present research (continued)

<i>Qualitative – nominal</i>	<i>Feature</i>		<i>Encoding</i>
Qualitative – nominal	Single	1	Marital status 11
	Married	2	
	Unknown	3	
Qualitative – nominal	Daily	1	Course 12
	Second shift	2	
	At night	3	
Qualitative	In numbers		Diploma average 13
Numerical	In numbers		The number of failed units 14
Qualitative – nominal	0 semester	1	The number of honors semesters 15
	1 semester	2	
	2 semester	3	
	More than 3 semesters	4	
Qualitative – nominal	0 semester		The number of conditional semesters 16
	1 semester: 2	2	
	2 semester	3	
	More than 3 semesters	4	
Qualitative – nominal	In numbers		GPA for semester 1 17
Qualitative – nominal	In numbers		GPA for semester 2 18
Qualitative – nominal	In numbers		FGPA (Final grade point average) 19

3.2.3.1 Feature selection

Feature selection aims to find a small set of relevant features. They should solve the problem without making the model more complex (Mwadulo, 2016). The data prep process selected for extraction 19 numbers, including the GPA class feature. Table 4 shows these selected attribute fields and corresponding values.

3.2.3.2 Principal component analysis algorithm

In this research, principal component analysis (PCA) played a crucial role in data cleansing. PCA is a math technique. It transforms correlated observations into uncorrelated ones, called principal components, using orthogonal transformations. These components are selected based on their variance, with the first component having the highest variance and representing uncorrelated information. PCA is valuable for feature selection, optimising data subsets, and enhancing data visualisation (Jolliffe and Cadima, 2016). It also makes machine learning algorithms more efficient. It reduces data dimensionality without losing valuable information (Dabhade et al., 2021).

3.2.3.3 Features encoding

In Machin Learning models, all data must be numerical variables. Therefore, before using category data in the model, you need to code it. There are various techniques for encoding features. We encoded the features into numeric ones. We used different techniques, including LabelEncoder and one-hot encoding. This was to find the best method for our dataset. LabelEncoder gave us a lower MAPE than we used in our dataset. Table 4 shows the variables used in the research and the possible values.

3.2.4 Modelling

For modelling, use classification techniques. They are a type of prediction. In classification, a class (label or discrete value) is predicted. A classifier creates a classification model using the training data. This model should fit the training data. It must predict the test data well (Han and Kamber, 2006).

3.2.4.1 Decision trees

Decision tree algorithms are a powerful tool in EDM. They classify students by their learning status (Chen and Lin, 2023). Data mining relies on decision trees to build accurate classification models. This technique assigns weights to dataset features by computing the increased ratio and, based on the resulting rules, predicts effectiveness (Alsalman et al., 2019). Decision trees offer simplicity and interpretability, making them suitable for interval numeric variables. In this structure, each branch node is a choice among options. Each leaf node is a decision (Quinlan, 1986). In a decision tree, internal nodes contain partitions that test attribute expressions, and each leaf node carries a corresponding class label. Users recursively divide nodes using a decision tree algorithm. Each branch shows a possible decision scenario and its outcome (Ogunde and Ajibade, 2014). The decision tree algorithm aims to find the best tree. It does this by checking the impurity of child nodes. Common impurity criteria include entropy, Gini index, and

classification error (Tan et al., 2013). Higher entropy values indicate a more homogeneous class distribution (Yulianto et al., 2020).

3.2.4.1.1 ID3 decision tree

ID3 is a simple decision tree developed by Ross Quinlan (Surjeet and Saurabh, 2012; Yadav et al., 2012; Ogunde and Ajibade, 2014). The ID3 algorithm searches the features of the training samples and extracts a feature that separates the best samples. This algorithm uses greedy search, i.e., it does not look back and selects the best feature (Ogunde and Ajibade, 2014).

3.2.4.1.2 Decision tree C4.5

The C4.5 algorithm is an improvement of the IDE3 algorithm developed (Ross Quinlan, 1993). C4.5 is also called J48 algorithm and it is better than ID3, such as:

- Can handle categorical and discrete data.
- C 4.5 decision tree algorithm can handle missing values. Missing feature values are simply not used in gain and entropy calculations.
- C4.5 prunes trees by returning to the tree. It removes non-contributing branches by replacing internal nodes with leaf nodes (Nijhawan et al., 2017).

The ID3 algorithm has a limitation in that it is sensitive to attributes with large values. This sensitivity is overcome by the C4.5 algorithm, which employs 'information gain' to measure a gain ratio (Hssina et al., 2014).

3.2.4.2 Random forest algorithm

RF is a supervised learning algorithm introduced by Leo Breiman in 2001 (Breiman, 2001). It leverages an ensemble of decision trees for regression and classification tasks. A larger number of trees usually improves its predictions. RF is now a standard tool for data analysis. It is effective for prediction (Akar and Güngör, 2012). In this ensemble approach, each regression decision tree predicts a number for a dataset. RF averages these predictions to get the final output (Kulkarni, 2014). RF makes its ultimate decision through a voting mechanism. Common feature selection criteria for decision tree induction include the Information gain ratio (Ross Quinlan, 1993) and the Gini index (Breiman et al., 1984). Like entropy, the Gini index reaches its maximum value when all classes in the dataset have equal probabilities (Pal, 2005; Ahmed and Sadiq, 2018).

3.2.4.3 Neural network algorithm

The ANN method began with McCulloch's 1940s work. He tried to simulate biological nervous systems. ANNs have interconnected units. Their traits depend on the network's topology and the neurons' attributes. This methodology stands as one of the most prevalent and efficacious learning systems. Notably, ANNs can learn, train, and adapt like the human brain (Russell and Norvig, 2016). ANNs analyse events. They find patterns in related ones. Then, they gather information. Lastly, they make decisions based on this knowledge when faced with new situations (Bahadir, 2016).

3.2.4.4 K nearest neighbour algorithm

KNN is a supervised machine learning technique. It solves classification and regression problems. Notably, KNN defers the entirety of its inductive generalisation endeavours until the regression phase (Han et al., 2011). The algorithm saves the feature vectors and classifies the training data during training. Next, during the classification step, the same features are computed for the unknown testing data. The algorithm calculates the distance between this new vector and all training sample vectors. Subsequently, it selects the nearest ‘k’ vectors based on this distance metric (Yulianto et al., 2020).

3.2.4.5 Gradient boosted tree algorithm

GBT leverages a collection of decision tree weak prediction models to formulate predictions. In this method, predictions are generated through the integration of a robust learner. This robust learner is a classifier. It can closely match a true classifier, made by combining various cases. Conversely, weak learners are classifiers that exhibit only marginal correlation with the true classifier. GBT employs an iterative process to iteratively refine predictions (Machado et al., 2019).

3.2.4.6 Support vector machines

SVM aims to find the closest data vectors, known as support vectors, to the decision boundary in the training dataset. It then uses these support vectors to classify new test vectors (Vladimir and Vapnik, 1998). A margin is the distance from the decision boundary to the closest points of the two classes (Vapnik, 1995). This classifier categorises objects as either high or low based on their inherent characteristics. Also, kernel techniques can turn non-linear relationships into linear ones. This helps with partitioning (Adejo and Connolly, 2018).

3.2.4.7 Naïve Bayes classifier

This algorithm is a simple probabilistic classifier. It computes probabilities by aggregating frequencies and merging dataset values. It also predicts membership probabilities for each class. It determines the likelihood of a record belonging to a specific class. The class boasting the highest probability is identified as the most probable. The classifier has a unique feature. The presence or absence of one feature does not affect any other feature (Karo et al., 2022).

3.2.4.8 Naïve Bayes kernel classifier

In the realm of classification, the initial step involves constructing a machine-learning model through a dataset known as the training dataset. Next, the model’s efficacy is tested on an unfamiliar dataset. Notably, the Naïve Bayes Kernel classifier (KNB) is a simple Bayesian method. It uses kernel density estimation (KDE) (Bernardo and Smith, 2009).

3.2.5 Test and evaluation stage

During the evaluation stage, the goal is to find a model with optimal attributes. It should generalise information well. The application of 10-fold cross-validation serves to verify and validate the outcomes of the research algorithms, ensuring accuracy, precision, and meticulous measurement (Arlot and Celisse, 2010; Anguita et al., 2009). Cross-validation serves as a method to assess and validate the accuracy of a constructed model (Ting, 2017). Accuracy is a very important criterion, but it is not enough to evaluate the performance of the model. When evaluating, we often use error rate and accuracy. What matters is our model's reliability, its generalisability, and its flexibility (Romero and Ventura, 2010; Shaikh et al., 2015). So, each criterion considers different aspects of the classification's performance. We must evaluate the classification by considering all of them at once (Ballabio et al., 2018).

3.2.5.1 Confusion matrix

Confusion matrix is a commonly used method to calculate accuracy in data mining concepts (Yulianto et al., 2020). The confusion matrix shows the current state of the dataset and the number of correct/incorrect predictions of the model.

The model's performance is based on the count of correctly and incorrectly classified samples (Costa et al., 2007). Table 5 shows the confusion matrices.

Table 5 Confusion matrix

<i>Data class</i>	<i>Classified as Pos</i>	<i>Classified as neg</i>
Positive	True positive (tp)	False negative(fn)
Negative	False positive(fp)	True negative(tn)

3.2.5.2 Accuracy

The accuracy measure (*Acc*) evaluates the effectiveness of the classifier with its percentage of correct predictions (Costa et al., 2007). CA is the ratio of correct predictions ($TP + TN$) to the total number of samples ($TP + TN + FP + FN$). The average per-class effectiveness of a classifier (Sokolova and Lapalme, 2009).

$$Acc = \frac{|TN| + |TP|}{|FN| + |FP| + |TN| + |TP|} \text{ Costa et al. (2007)} \quad (1)$$

3.2.5.3 Precision

Precision (P) is a significant measure that assesses the likelihood of a positive prediction being accurate (Costa et al., 2007). Precision is calculated as the ratio of true positives correctly predicted to all instances predicted as positive by the model. It lies within the range of [0, 1] and is a key indicator of the model's ability to minimise false positive errors (Nisbet et al., 2009).

$$P = \frac{|TP|}{|TP| + |FP|} \quad (2)$$

$$F - measure = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R} \text{ Costa et al. (2007)} \quad (3)$$

3.2.5.4 Recall or sensitivity

The recall (R) measure evaluates the effectiveness of a classifier for each class in a binary problem. Recall is the proportion of samples belonging to the positive class that are correctly predicted as positive (Costa et al., 2007). Recall, also called sensitivity, measures the ability of a model to classify the positives efficiently (Sokolova and Lapalme, 2009). The specificity is the percentage of negative examples that are correctly predicted to be negative (Costa et al., 2007).

$$R = \frac{|TP|}{|TP| + |FN|} \quad (4)$$

$$Spc = \frac{|TN|}{|FP| + |TN|} \text{ Sokolova and Lapalme (2009), Costa et al. (2007)} \quad (5)$$

3.2.5.5 Specificity

Feature measures the ability of a model to efficiently classify negatives. A feature is the ratio of correctly identified true negative records to true negative records in the data. Specificity fights against false positives.

$$\frac{|TN|}{|TN| + |FP|} \text{ Sokolova and Lapalme (2009)} \quad (6)$$

3.2.5.6 F1-measure

The F1-measure comes from the harmonic mean of precision and recall. It accounts for their conflicting relationship. This calculation aims to achieve a balanced outcome that prioritises both precision and sensitivity in the results (Dixit et al., 2021; Shaikh et al., 2015).

$$F - measure = \frac{2}{1 / precision + 1 / recall} \text{ Shaikh et al. (2015)} \quad (7)$$

3.2.5.7 ROC area (receiver operating characteristics)

The ROC metric assesses a model's ability to avoid misclassification by plotting sensitivity against specificity on a graph. AUC, or area under the curve, represents a single point on the ROC curve and quantifies the model's predictive performance. A larger AUC means better class recognition. An AUC of 1 is ideal (Sarlis and Christopoulos, 2014). The area under the ROC curve is a composite index. It shows how likely the model is to classify positive and negative instances correctly. This index ranges from a minimum of 0.5 to a maximum of 1 and reflects the algorithm's efficiency (Hand et al., 2001).

3.2.5.8 Kappa

The kappa statistic, introduced by Jacob Cohen in 1960, is frequently used to assess inter-rater reliability. Cohen's kappa considers the likelihood that raters may make random guesses due to uncertainty. Similar to most correlation statistics, kappa's range spans from -1 to $+1$ (McHugh, 2012). Cohen's Kappa assesses the reliability of the classification of objects. Interrater reliability, on the other hand, measures the agreement level between two binary raters (Wood, 2007). The kappa criterion value typically stands at 0.60 (Simundic, 2008).

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (8)$$

3.2.5.9 Error classification (error rate)

The most common measures of performance in classification problems are error rate and accuracy. This includes both binary and multi-class classification. Error rate evaluates a classifier by its percentage of incorrect predictions. The error rate is the ratio of misclassified samples to all samples. While accuracy is the proportion of samples that are correctly classified. error rate as follows: error rate as follows:

$$Err = \frac{|FN| + |FP|}{|FN| + |FP| + |TN| + |TP|} = 1 - Acc \text{ Costa et al. (2007)} \quad (9)$$

3.2.6 Deployment stage

The study utilised RapidMiner 9.9 and excel for all its data mining implementation and processing tasks. The outcomes of the research design process have contributed to the generation of new knowledge presented in the article.

4 Data analysis (findings)

4.1 Introduction

The Education data mining results use the RapidMiner 9.9 tool. Among the methods to assess model accuracy, the 10-fold cross-validation is used. This method involves randomly dividing the dataset into 10 segments. The model is then trained and tested 10 times (Ting, 2017). The best model is based on the chosen performance criteria (Lever et al., 2016). In this section, the research questions are addressed and answered.

4.2 Modelling

Creating an effective, user-friendly model to predict student performance is hard. It must consider many factors (Xing et al., 2015). The research explores several classification models: C4.5 decision tree, ID3, KNN, NB, Naive Bayes Kernel (NBK), neural network, RF, SVM, and gradient boosted trees. We assess these models using the parameter configurations outlined in Table 6.

Table 6 Parameter settings in classification algorithms

<i>Parameters</i>	<i>Algorithm</i>
Number of tree = 20	RF
Criterion = gain ratio	
maximal depth = 10	
Voting strategy = confidence vote	
Splitting criterion = information gain ratio	DT C4.5
Minimal size of split = minimal leaf size =1	
Minimal gain = 0.01	
Maximal depth = 20	
Confidence = 0.1	DT ID3
Split criterion = the ratio of obtaining information	
The minimum split size = 4	
Minimum leaf size = 2	
Minimum gain = 0.01	ANN
Training cycles = 500	
Learning rate = 0.3	
Momentum = 0.2	
Error epsilon = 1.0E-5	
The neural network has two hidden layers with 5 nodes	
Parameter settings of Weight by Gini index operator in modelling with neural network	
Weight relation = top p%	
p = 0.5	KNN
The number of K = 5	
Measure types = mixed measures	
Mixed measure = mixed Euclidean distance	
Value c = 1	SVM
Kernel = linear	
Epsilon = 0.1	
Cost = 1.0	
Naive Bayes operator was set based on Laplace correlation	NB
Weight relation= top p%	
P = 0.5	
Number of trees = 50	
Maximal depth = 5	GBT
Min rows =10	
min split improvement = 1.0E-5	
Number of bins = 20	
Learning rate = 0.01	
Sample rate = 1.0	

4.2.1 Modelling results

Figure 1 is a graph that compares the accuracy of the algorithms used in this research in two classes. Figure 2 is a graph that compares the accuracy of the algorithms used in this research in four classes.

Figure 1 Comparative chart of the accuracy of the algorithms used in the two-class mode (see online version for colours)

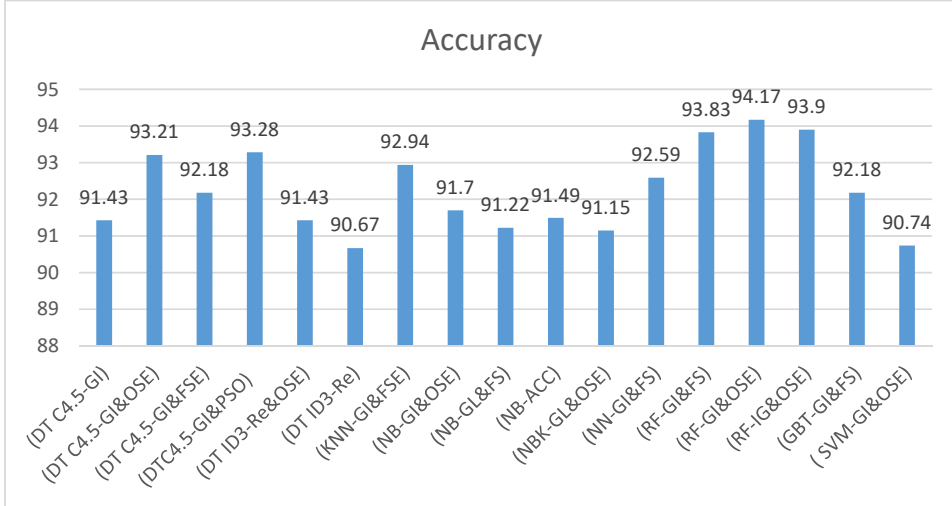


Figure 3 is a table that shows the confusion matrix in the case of 2 classes. Figure 4 is a table that shows the confusion matrix in the case of 4 classes.

Figure 2 Comparative chart of the accuracy of the algorithms used in the four-class mode (see online version for colours)

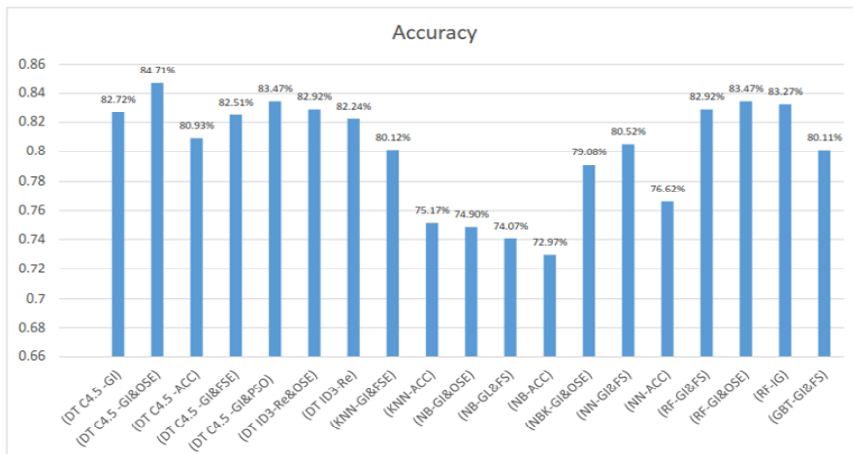


Table 7 Comparison of evaluation criteria of models in two classes

	<i>Predictor</i>	<i>Accuracy</i>	<i>Classification_error</i>	<i>Kappa</i>	<i>Precision</i>	<i>Recall</i>	<i>F-measure</i>	<i>Specificity</i>	<i>AUC ROC</i>
The best model	Random forest with Gini index and optimise selection (evolutionary) (RF-GI and OSE)	94.17%	5.83%	0.862	95.42%	96.23%	95.81%	89.56%	0.984
Model 2	Random forest trees with information gain and optimise selection (evolutionary) (RF-IG&OSE)	93.90%	6.10%	0.856	95.20%	96.03%	95.61%	89.11%	0.982
Model 3	Random forest with Gini index and forward selection (RF-GI&FS)	93.83%	6.17%	0.853	94.59%	96.63%	95.59%	87.56%	0.983
Model 4	Decision tree C4.5 with Gini index and with optimise selection (evolutionary) (DT C4.5-GI&OSE)	93.21%	6.79%	0.841	95.23%	95.04%	95.08%	89.11%	0.906

Table 8 Comparison of evaluation criteria of models in four classes

	Predictor	Accuracy	Kappa	Classification_error	Recall	Precision	Kendall_Tau	Spearman_rho	Correlation
The best model	Decision tree C4.5 with Gini index and with optimise selection (evolutionary)(DT C4.5 – GI&OSE)	84.71%	0.780	15.29%	83.42%	84.17%	0.858	0.888	0.891
Model 2	Decision tree C4.5 with Gini index and optimise weights (PSO) (DT C4.5 – GI&PSO)	83.47%	0.761	16.53%	82.60%	83.14%	0.841	0.872	0.880
Model 3	Random forest with Gini index and optimise selection (evolutionary)(RF-GI&OSE)	83.47%	0.762	16.53%	82.19%	83.61%	0.845	0.876	0.882

The RF algorithm predicted the class of 1,373 out of 1,458 objects. Its accuracy is 94.17%. Each matrix contains all correct predictions in the diagonals. In the first column (class 1), i.e., successful students, the classifier predicted 403 of 450 students as successful. The recall of this class is 89.56% (ie., 403/450). The accuracy for this class is 91.38%. In the second column (class 2 of 2), i.e., unsuccessful students, the classifier correctly predicted 970 of the 1,008 students as unsuccessful. The recall is 96.23% (ie., 970/1,008). The accuracy for this class is 95.38%. The RF-GI&OSE model in two-class mode has better predicted poor students' performance than successful students. The confusion matrix in Figure 3 shows the RF-GI&OSE RF method has an 89.56% success detection rate and a 96.23% failure detection rate.

Figure 3 Confusion matrix of the best classifier (RF-GI and OSE) in 2-class mode (see online version for colours)

accuracy: 94.17% +/- 2.22% (micro average: 94.17%)

	true 1	true 2	class precision
pred. 1	403	38	91.38%
pred. 2	47	970	95.38%
class recall	89.56%	96.23%	

Figure 4 Confusion matrix of the best classifier (DT C4.5-GI and OSE) in 4-class mode (see online version for colours)

accuracy: 84.71% +/- 2.36% (micro average: 84.71%)

	true 1	true 2	true 3	true 4	class precision
pred. 1	81	22	0	0	78.64%
pred. 2	22	295	49	9	78.67%
pred. 3	1	27	498	50	86.46%
pred. 4	0	2	41	361	89.36%
class recall	77.88%	85.26%	84.69%	85.95%	

The C4.5 algorithm can predict the class of 1,235 objects out of 1,458, which gives it an accuracy value of 84.71%. Each matrix contains correct predictions in its diagonals. In the first column (class 1 of 4), i.e., excellent students, the classifier correctly predicted 81 of 104 students as excellent in achievement. The recall is 77.88% (ie., 81/104) and the precision for this class is 78.64%. In the 4th column (class 4 of 4), i.e., weak students, the classifier correctly predicted 361 of 420 students as weak. The recall is 85.95% (that is, 361/420) and the accuracy for this class is 89.36%. Thus, the model (DT C4.5-GI&OSE) has predicted poor students' performance better than that of successful students in four classes. The confusion matrix in Figure 4 shows the detection rates for the DTC4.5-GI&OSE method. They are: excellent 77.88%, good 85.26%, average 84.59%, and poor 85.95%. Decision trees are popular. They make easy-to-interpret classification rules. We study popular decision tree classifiers. We test them to find the best one for predicting student GPA from student data. The research shows that the C4.5 decision tree is better than ID3 for data classification.

Table 7 shows the evaluation criteria for modelling two classes. Table 8 shows the criteria for modelling four classes.

4.2.2 Results of predicting academic performance using classifiers in two-class mode

Table 8 shows that, in the two-class mode, the RF-GI&OSE algorithm has the highest accuracy of 94.17%. No other algorithm is as accurate. The RF-IG&OSE model also has an acceptable accuracy of 93.90%. Thus, the classifiers 'RF with Gini index with optimise selection (evolutionary) feature selection algorithm' and 'RF with information gain with optimise selection (evolutionary) feature selection algorithm' are key models for forecasting in this research. Considering the high accuracy of both of these models, we will check other items to choose the best model. The (RF-GI&OSE) model has more precision (95.42%) than the (RF-IG&OSE) model with 95.20% accuracy. Also, Table 6 shows the (RF-GI&OSE) model has a higher recall of 96.23%. The (RF-IG&OSE) model's recall is 96.03%. The (RF-GI&OSE) model has a higher f-measure (95.81%) than the (RF-IG&OSE) model with an f-measure of 95.61%. The (RF-GI&OSE) model has a higher specificity (89.56%) than the (RF-IG&OSE) model with a specificity of 89.11%. The model (RF-GI&OSE) has a lower error of 5.83% than the model (RF-IG&OSE) with a 6.1% error. The (RF-GI&OSE) model has a higher Kappa of 0.862% compared to the Kappa value of 0.856% in the (RF-IG&OSE) model. The comparison shows that the model (RF-GI&OSE) is better than (RF-IG&OSE) at predicting students' performance in two classes. Thus, the model (RF-GI&OSE) is chosen as the main model of this research. It is the best algorithm to predict students' performance in two classes. The RF with Gini index and forward selection (RF-GI&FS) and the decision tree C4.5 with Gini index and optimise weights (PSO) (DTC4.5-GI&PSO) algorithms, and the decision tree C4.5 with Gini index and optimise selection (evolutionary) (DT C4.5-GI&OSE) also have high accuracy. They are among the best algorithms in 2-class mode. The comparison of the algorithms shows that the best accuracy is with the RF-GI&OSE decision tree. The RF-GI&OSE algorithm has a better recognition rate for predicting excellent students (89.56%) than RF-IG&OSE (89.11%). The RF-GI&FS algorithm has the best detection for predicting unsuccessful students (96.63 %) against all algorithms. Also, the NBK-GL&OSE algorithm better predicts excellent students (95.11%) than RF-GI&OSE (89.56%) and RF-IG&OSE (89.11%) or other algorithms.

Figure 5 The area under the roc chart (AUC) (see online version for colours)

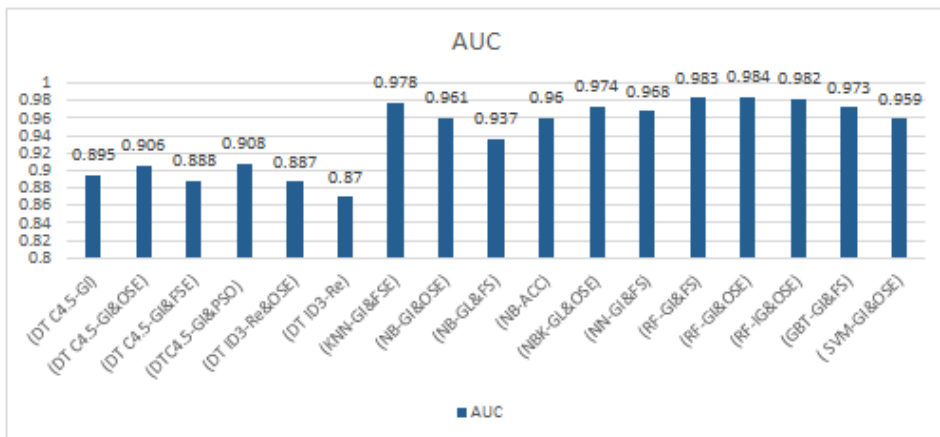
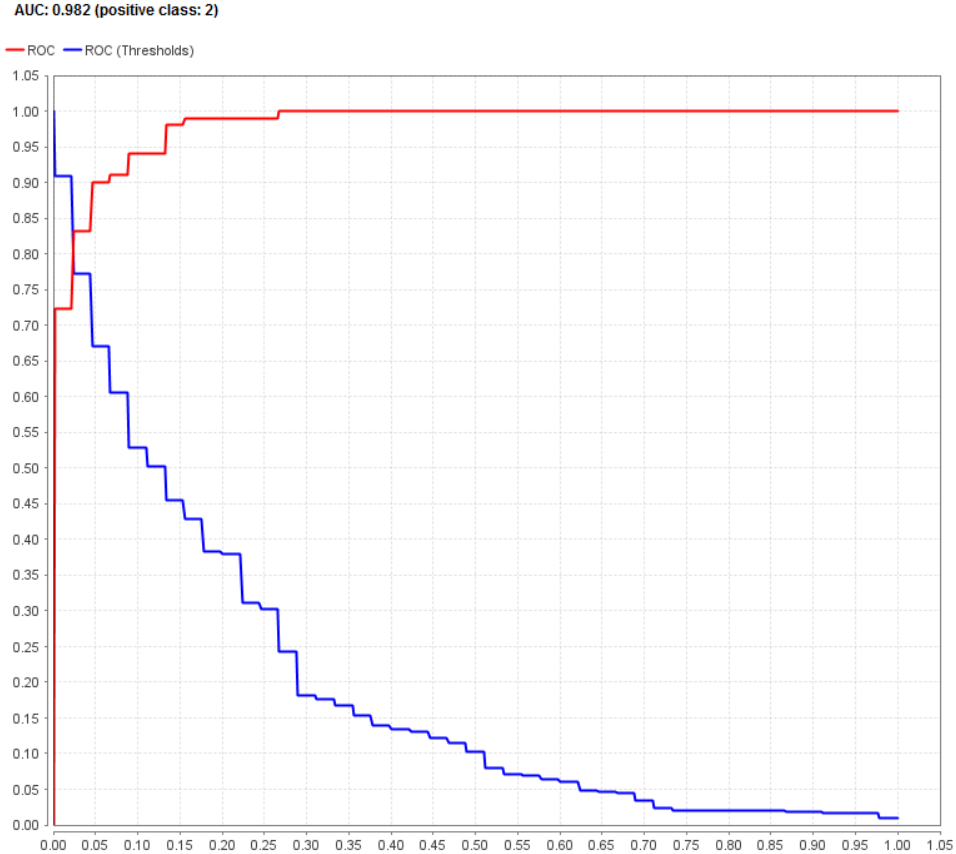


Figure 5 shows the amount of AUC in different classification algorithms.

One of the criteria for assessing the fit of the category models is the ROC diagram. If the model works completely randomly, the area under the graph will be equal to 0.5. If the model is very accurate in finding the right cases, the area under the graph will approach 1. According to Figure 6, the best AUC in the ROC chart is RF-GI&OSE (0.984).

Figure 6 Roc diagram of AUC value in a two-class mode in the model (RF-GI&OSE) (see online version for colours)



4.2.2.1 Important features in two-class mode

Table 9 shows the prioritisation of predictive variables. It is based on their weight in a two-class mode in a RF with Gini Index and optimise selection (evolutionary) (RF-GI&OSE).

In the main research method, RF-GI&OSE, the number of failed units has the greatest impact on predictions. The 1st-semester GPA feature is next. Age, diploma average, and number of honours semesters are in the middle of the graph. Several variables have been removed due to the Gini index feature selection's low impact. The semester 2 GPA and a

few other factors had little effect on the total GPA. In this model, the age variable affects students' academic performance among demographic characteristics.

Table 9 Prioritising important variables in predicting performance in two classes in random forest with Gini index and optimise selection (evolutionary)(RF-GI&OSE)

<i>Number of conditional</i>	<i>GPA for semester 2</i>	<i>Number of honours semester</i>	<i>Diploma average</i>	<i>Age</i>	<i>GPA for semester 1</i>	<i>Number of fallen units</i>
0.062	0.108	0.145	0.150	0.160	0.162	0.173

4.2.2.2 The results of the student performance prediction model in two-class mode

Table 10 Results of the student performance prediction model in two-class mode random forest with Gini index and optimise selection (evolutionary) (see online version for colours)

<i>Row no.</i>	<i>Classified GPA</i>	<i>Prediction</i>
1	1	1
2	2	2
3	2	2
4	2	2
5	2	2
6	1	1
7	1	1
8	2	2

4.2.3 Results of predicting academic performance using classifiers in four-class mode

Table 9 shows that, for four classes, the DT C4.5-GI&OSE algorithm has the highest accuracy of 84.71% among the algorithms. Thus, the classifier 'decision tree with Gini index with Optimise selection (Evolutionary) feature selection algorithm' is a key model in this research's forecasts. This model is more accurate than other models. It also ranks higher on other criteria, such as precision, recall, kappa, error classification, correlation, Kendal, and Spearman. The comparison shows that the model (DT C4.5-GI&OSE) is the best at predicting students' performance in 4 classes. The Decision Tree C4.5 with Gini index and optimise weights (PSO) and RF with Gini index and optimise selection (Evolutionary) (RF-GI&OSE) also had high accuracy. They are important algorithms in 4-class mode. The comparison of the algorithms shows that the best accuracy is from the decision tree, DT C4.5-IG&ose. The DT C4.5-IG&ose algorithm better predicts excellent students (77.88%) than RF-OS&GI (75.96%). The RF-OS&GI algorithm has the best detection for predicting weak and unsuccessful students (90%) against all algorithms. Also, DT ID3-IG&ose and NBK-GI&OSE algorithms predict high school students' scores better (82.69%). They outperform RF-OS&GI (75.96%) and DT C4.5-IG&ose (77.88%).

4.2.3.1 Important features in four-class mode

Table 11 shows the priority of predictive variables. It is based on their weight in the four-class mode for the decision tree with Gini index and optimise selection (evolutionary) (RF-GI&OSE).

In the main method of research, DT C4.5-GI&OSE, it can be seen that the diploma average has the greatest effect on predictions. Next is the average for the 1st semester. Semester 2 GPA and the number of honours semesters are in the middle of the graph. We removed several variables. The Gini Index feature selection technique had a very low impact. The input and the number of selected features had little effect on the average. In this model, all demographic features had little effect.

Table 11 Prioritising important variables in predicting performance in four classes in decision tree with Gini index and optimise selection (evolutionary)(RF-GI&OSE)

<i>Input</i>	<i>Number of conditional</i>	<i>Number of honours semesters</i>	<i>GPA for semester 2</i>	<i>GPA for semester 1</i>	<i>Diploma average</i>
0.104	0.114	0.122	0.197	0.201	0.262

4.2.3.2 The results of using the student performance prediction model in the four-class mode

Table 12 Model implementation results (DT C4.5-GI&OSE) to predict student performance decision tree C4.5 with Gini index and with optimise selection (evolutionary)

<i>Row no.</i>	<i>Classified GPA</i>	<i>Prediction</i>
1	3	3
2	3	4
3	1	1
4	3	3
5	2	3
6	3	3
7	2	2
8	3	3

5 Discussion

In response to research sub-question 2, what methods predict student performance? We used several models to analyse and categorise students. They are: C4.5 and ID3 decision tree classifiers, RF, Bayesian networks, k-nearest neighbours, ANNs, SVMs, and gradient boosted trees. This study introduced the best model. It was found by comparing algorithms in 2-class and 4-class tests. This study also considers the use of four feature selection techniques.

Which classification algorithm, two-class or four-class, predicts student performance better? The RF-GI&OSE algorithm has the highest accuracy of 94.17% compared to other algorithms. The RF-IG&OSE model also has an acceptable accuracy of 93.90%. Thus, the classifiers ‘RF with Gini index with optimise selection (Evolutionary) feature

selection algorithm' and 'RF with information gain with optimise selection (evolutionary) feature selection algorithm' are key models for forecasting in this research. Considering the high accuracy of both of these models, we will check other items to choose the best model. The (RF-GI&OSE) model has more precision (95.42%) than the (RF-IG&OSE) model with 95.20% accuracy. Table 6 shows that the (RF-GI&OSE) model has a higher recall (sensitivity) of 96.23%. The (RF-IG&OSE) model's sensitivity is 96.03%. The (RF-GI&OSE) model has a higher f-measure (95.81%) than the (RF-IG&OSE) model with an f-measure of 95.61%. The (RF-GI&OSE) model has a higher specificity (89.56%) than the (RF-IG&OSE) model with a specificity of 89.11%. The model (RF-GI&OSE) has a lower classification error of 5.83% compared to the model (RF-IG&OSE) with a classification error of 6.1%. The (RF-GI&OSE) model has a higher Kappa of 0.862% compared to the Kappa value of 0.856% in the (RF-IG&OSE) model. The comparison shows that the model (RF-GI&OSE) is better than the model (RF-IG&OSE) at predicting students' performance in two classes. Thus, the model (RF-GI&OSE) is chosen as the main one. It is the best for predicting students' performance in two classes. The best algorithm in the four-class mode is decision tree C4.5. It uses the Gini index and an evolutionary optimiser. Its prediction accuracy is 84.71%. According to these results, using NB algorithms, the weakest prediction accuracy of about 70% occurs. In other words, in four-class mode, the best algorithm for predicting academic performance. In this study, the best algorithms are C4.5 and RF. They have the highest accuracy and rank well on other criteria. The four-class mode decision tree C4.5-GI & OSE model showed 84.71% accuracy, 84.17% precision, 83.42% sensitivity, and a 0.780 kappa.

What factors predict undergraduate student performance? This is the main question of the research. It compares two-class and four-class classification models. This study used RF-GI&OSE in 2 classes. It found that, of the features, the number of failed units had the greatest impact on the prediction process. The next most important feature was the 1st-semester GPA. 2st-semester GPA and conditional number among the selected characteristics had little effect on the total grade point average. In this model, the age variable affects students' academic performance among demographic characteristics. In the main method of research, DT C4.5-GI&OSE in four classes, the average of the diploma has the greatest effect on the prediction process. The 1st-semester average is next. The input and the conditional number of the selected features had little effect on the overall average. In this model, all demographic characteristics had little effect.

What prediction models are found using data mining classification techniques on the predicted variables, in response to sub-question 3? The comparison shows that the model (RF-GI&OSE) is better than the model (RF-IG&OSE) at predicting students' performance in two classes. Thus, the model (RF-GI&OSE) is chosen as the main one for this research. It is the best at predicting students' performance in two classes. The RF-GI&FS the Decision Tree C4.5 with Gini index and optimise weights (PSO) (DTC4.5-GI&PSO), and the DTC4.5-GI&OSE are important 2-class algorithms. They have high accuracy. The model (DT C4.5-GI&OSE) is better than other models at predicting students' performance in 4 classes, according to the comparison. In 4-class mode, the algorithms C4.5 with Gini index and optimise weights (PSO) and RF with Gini index and optimise selection (evolutionary) (RF-GI&OSE) had high accuracy. They are among the most important algorithms.

In response to sub-question 4, how is the validation of the developed model for the students of Shahed University done? This research uses K-fold cross-validation to split

the data into educational and experimental sets. Estimating the accuracy of algorithms in this method is much more than other methods. In this study, we consider the value of k to be equal to 10. This study ranked the key Attributes that impact students' academic performance. The academic factor had a greater impact than the demographic factors. This result is consistent with the results of studies by Bhutto et al. (2020). Waheed et al. (2020) predicted the failure or success of students with 85% accuracy (Waheed et al., 2020). Costa Mendes et al. (2021) and Cruz-Jesus et al. (2020) predicted student performance based on income, age, employment, cultural level, residence, and socioeconomic information. Hoffait and Schyns (2017) ranked the algorithms by accuracy: LR, ANN, then RF. But, Waheed et al. (2020) found that the SVM algorithm performs better than the LR algorithm. In the research of Xu et al. (2019), the most efficient algorithm is SVM. The NN algorithm is next. The decision tree has the lowest performance (Xu et al., 2019). Meanwhile, this research found that the RF and DT algorithms are best at predicting students' academic progress scores. The NB algorithm has the lowest accuracy. Bernacki et al. (2020) Their algorithm identified 75% of students who failed to get a B or better to advance to the next course (Bernacki et al., 2020). Ahmad and Shahzadi (2018) predicted academic performance with 85% accuracy (Ahmad and Shahzadi, 2018).

Academic performance was predicted with many diverse variables, including students' demographic features (Bernacki et al., 2020; Rebai et al., 2020; Garcia and Skrita, 2019; Rizvi et al., 2019), learning skills and study habits (Ahmad and Shahzadi, 2018) learning strategies, social support, perception, motivation, social demographics and academic performance features (Musso et al., 2020; Costa-Mendes et al., 2021), etc. Cruz et al. (2020) predicted students' final grades with 16 independent variables. The prediction was made only with the parameters of the mid-semester score, faculty, and educational group (Cruz-Jesus et al., 2020). In most models in EDM studies, the prediction accuracy is above 70%. However, collecting and processing data is time-consuming. Hoffait and Schyns (2017) found data collection too hard. They said socio-economic data was unnecessary (Hoffait and Schyns, 2017). In addition, these data do not always provide a correct idea of failure prevention (Bernacki et al., 2020; Yağcı, 2022). The research found that academic features are better than demographics at predicting students' performance.

6 Conclusions

This study finds models that best predict students' performance in education. They are based on individual, social, and educational factors. This research aims to find factors that predict undergraduate student performance. This study used various algorithms to predict the total GPA. The data used in this research are records of undergraduate students from 2012. Researchers used 1,468 data records in data mining. We used the CRISP-DM research method for the data mining aspect of this research. We used the RapidMiner 9.9 tool for modelling. This paper's results are in the case of two and four classes. The study compared different algorithms using some criteria. It then introduced the best models. This study also considers four attribute selection techniques. The best algorithm in two-class mode was the RF. It used the Gini index and optimise selection (evolutionary). It had an accuracy of 94.17%. This model has shown 95.42% accuracy, 96.23% sensitivity, and 89.56% specificity. The RF results show that a higher 1st

semester GPA leads to a higher chance of academic success. The proposed model in the four-class mode Decision Tree C4.5-GI and OSE showed an accuracy of 84.71%. It had a precision of 84.17%, a sensitivity of 83.42%, and a kappa of 0.780. This research shows that machine learning algorithms can predict students' academic achievement. These algorithms are decision tree C4.5-GI and OSE and RF-GI and OSE. C4.5's decision tree shows that higher first-semester and diploma GPAs predict academic success. The comparison of results in two-class and 4-class classification showed that two-class modelling is better. It has higher accuracy than 4-class modelling. Thus, the current research shows that the RF algorithm is best for predicting students' performance. It uses the Gini index and evolutionary optimal selection. It has a 94.17% accuracy for two classes. This research used K-fold cross-validation and the confusion matrix to evaluate the data. We must use DM methods to solve students' learning problems. This will optimise the educational environment and enable data-based decision-making. Our research compared the predicted and actual results. They showed a significant improvement. The prediction helped identify successful and unsuccessful students. It also helped them get better grades.

The proposed research method may encounter certain limitations, which include:

- 1 Data collection: we need to analyse two types of data. First, we have some records from the education system. Second, we need to collect and add some other data for effective data mining. Also, the data's cross-sectional nature suggests a need for long-term studies.
- 2 Outlier data: outlier data: A high number of outliers or missing data can mislead results. However, this issue can be partly addressed through appropriate data pre-processing techniques.
- 3 Generalisability of findings: the sample is limited to undergraduate students at Shahed University. So, be cautious when applying the findings to students at other schools or universities.

Here are the suggestions provided in the text:

- 1 Real-world implementation: it is recommended to use the proposed research method in Iranian schools and universities. This real-world application would confirm the methodology's effectiveness in the local educational context.
- 2 Graduate-level research: future studies can extend the research to the graduate level. They can test the predictive model in higher education settings.
- 3 Faculty evaluation: researchers should analyse faculty evaluations. They may reveal factors like competency, qualifications, and employment criteria.
- 4 Exploration of various factors: future studies can explore more factors that affect student performance. These may include students' interests, skills, work history, and career goals. Also, their cultural background, family income, social media use, and extracurricular. Lastly, their academic engagement, interactions with professors, and various learner behaviours.
- 5 Model enhancement: to improve accuracy, add new parameters and expand the ML models.

In summary, these recommendations stress real-world implementation. They call for research at different academic levels, faculty evaluations, and a look at diverse factors. Also, they urge improving machine learning models for more reliable results in future EDM work.

References

- Abdelmagid, A. and Qahmash, A. (2023) 'Utilizing the educational data mining techniques' orange technology' for detecting patterns and predicting the academic performance of university students', *Inf. Sci. Lett.*, Vol. 12, No. 3, pp.1415–1431.
- Adejo, O.W. and Connolly, T. (2018) 'Predicting student academic performance using a multi-model heterogeneous ensemble approach', *Journal of Applied Research in Higher Education*, Vol. 10, No. 1, pp.61–75.
- Ahmad, Z. and Shahzadi, E. (2018) 'Prediction of students' academic performance using artificial neural network', *Bulletin of Education and Research*, Vol. 40, No. 3, pp.157–164.
- Ahmed, N.S. and Sadiq, M.H. (2018) 'Clarify of the random forest algorithm in an educational field', *2018 International Conference on Advanced Science and Engineering (ICOASE)*, IEEE, pp.179–184.
- Ajibade, S-S.M., Ahmad, N.B. and Shamsuddin, S.M. (2019) 'An heuristic feature selection algorithm to evaluate academic performance of students', *2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC)*, EE, pp.110–114.
- Akar, Ö. and Güngör, O. (2012) 'Classification of multispectral images using random forest algorithm', *Journal of Geodesy and Geoinformation*, Vol. 1, No. 2, pp.105–112.
- Alamgir, Z., Akram, H., Karim, S. and Wali, A. (2024) 'Enhancing student performance prediction via educational data mining on academic data', *Informatics in Education*, Vol. 23, No. 1, pp.1–24.
- Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L. and Alharthi, R. (2022) 'Development of a web-based prediction system for students' academic performance', *Data*, Vol. 7, No. 2, p.21.
- Aldowah, H., Al-Samarraie, H. and Fauzy, W.M. (2019) 'Educational data mining and learning analytics for 21st century higher education: a review and synthesis', *Telematics and Informatics*, Vol. 37, No. 2019, pp.13–49.
- Al-Emran, M., Malik, S.I. and Al-Kabi, M.N. (2020) 'A survey of internet of things (IoT) in education: opportunities and challenges', in Hassanien, A., Bhatnagar, R., Khalifa, N.E.M. and Taha, M.H.N. (Eds.): *Toward Social Internet of Things (SIoT): Enabling Technologies, Architectures and Applications: Emerging Technologies for Connected and Smart Social Objects*, Springer International Publishing, Cham.
- Alghamdi, A. S. and Rahman, A. (2023) 'Data mining approach to predict success of secondary school students: a Saudi Arabian case study', *Education Sciences*, Vol. 13, No. 3, p.293.
- Alsaman, Y.S., Halemah, N.K.A., Alnagi, E.S. and Salameh, W. (2019) 'Using decision tree and artificial neural network to predict students academic performance', *2019 10th International Conference on Information and Communication Systems (ICICS)*, IEEE, pp.104–109.
- Alshanqiti, A. and Namoun, A. (2020) 'Predicting student performance and its influential factors using hybrid regression and multi-label classification', *IEEE Access*, Vol. 8, No. 2020, pp.203827–203844.
- Ampadu, Y.B. (2023) 'Handling big data in education: a review of educational data mining techniques for specific educational problems', *AI, Computer Science and Robotics Technology*, Vol. 13, No. 2, pp.1–16.
- Amrieh, E.A., Hamtini, T. and Aljarah, I. (2015) 'Preprocessing and analyzing educational data set using X-API for improving student's performance', *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, IEEE, pp.1–5.

- Anguita, D., Ghio, A., Ridella, S. and Sterpi, D. (2009) 'K-fold cross validation for error rate estimate in support vector machines', *International Conference on Data Mining (DMIN)*, pp.291–297.
- Arcinas, M.M., Sajja, G.S., Asif, S., Gour, S., Okoronkwo, E. and Naved, M. (2021) 'Role of data mining in education for improving students performance for social change', *Turkish Journal of Physiotherapy and Rehabilitation*, Vol. 32, No. 3, pp.6519–6526.
- Arlot, S. and Celisse, A. (2010) 'A survey of cross-validation procedures for model selection', *Statistics Surveys*, Vol. 4, No. 2010, pp.40–79.
- Asif, R., Merceron, A., Ali, S.A. and Haider, N.G. (2017) 'Analyzing undergraduate students' performance using educational data mining', *Computers and Education*, Vol. 113, No. 2017, pp.177–194.
- Bahadir, E. (2016) 'Using neural network and logistic regression analysis to predict prospective mathematics teachers' academic success upon entering graduate education', *Educational Sciences: Theory and Practice*, Vol. 16, No. 3, pp.943–964.
- Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A. and Xia, F. (2021) 'Educational big data: predictions, applications and challenges', *Big Data Research*, Vol. 26, No. 2021, p.100270.
- Ballabio, D., Grisoni, F. and Todeschini, R. (2018) 'Multivariate comparison of classification performance measures', *Chemometrics and Intelligent Laboratory Systems*, Vol. 174, No. 2018, pp.33–44.
- Batirovna, S.B. (2023) *Educational Data Mining and Learning Analytics*, Vol. 4, No. 6, pp.61–64.
- Batool, S., Rashid, J., Nisar, M.W., KIM, J., Kwon, H-Y. and Hussain, A. (2023) 'Educational data mining to predict students' academic performance: a survey study', *Education and Information Technologies*, Vol. 28, No. 1, pp.905–971.
- Bellaj, M., Dahmane, A.B., Boudra, S. and Sefian, M.L. (2024) 'Educational data mining: employing machine learning techniques and hyperparameter optimization to improve students' academic performance', *International Journal of Online and Biomedical Engineering*, Vol. 20, No. 3, p.55.
- Bernacki, M.L., Chavez, M.M. and Uesbeck, P.M. (2020) 'Predicting achievement and providing support before STEM majors begin to fail', *Computers and Education*, Vol. 158, No. 2020, p.103999.
- Bernardo, J.M. and Smith, A.F. (2009) *Bayesian Theory*, Vol. 405, Book of John Wiley & Sons, Wiley, Available from the British Library, England.
- Bhutto, E.S., Siddiqui, I.F., Arain, Q.A. and Anwar, M. (2020) 'Predicting students' academic performance through supervised machine learning', *2020 International Conference on Information Science and Communication Technology (ICISCT)*, IEEE, Pakistan, pp.1–6.
- Boztaş, G.D., Berigel, M. and Altınay, F. (2024) 'A bibliometric analysis of educational data mining studies in a global perspective', *Education and Information Technologies*, Vol. 29, No. 7, pp.8961–8985.
- Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) *Classification and Regression Trees*, Google Scholar There is no corresponding record for this reference, Wadsworth, Monterey, CA.
- Chamorro-Premuzic, T. and Furnham, A. (2009) 'Mainly openness: the relationship between the big five personality traits and learning approaches', *Learning and Individual Differences*, Vol. 19, No. 4, pp.524–529.
- Chen, S. and Lin, X. (2023) 'Application of decision tree algorithm in educational data mining', *Curriculum and Teaching Methodology*, Vol. 6, No. 8, pp.120–127.
- Costa, E., Lorena, A., Carvalho, A. and Freitas, A. (2007) 'A review of performance evaluation measures for hierarchical classifiers. Evaluation methods for machine learning II', *Papers from the AAAI-2007 workshop*, pp.1–6.

- Costa-mendes, R., Oliveira, T., Castelli, M. And Cruz-Jesus, F. (2021) 'A machine learning approximation of the 2015 Portuguese high school student grades: a hybrid approach', *Education and Information Technologies*, Vol. 26, No. 2, pp.1527–1547.
- Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M. and Rosa-Louro, A. (2020) 'Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country', *Heliyon*, Vol. 6, No. 6, p.e04081.
- Dabhade, P., Agarwal, R., Alameen, K., Fathima, A., Sridharan, R. and Gopakumar, G. (2021) 'Educational data mining for predicting students' academic performance using machine learning algorithms', *Materials Today: Proceedings*, Vol. 47, No. 2021, pp.5260–5267.
- Debang, M. and Hassan, B.U. (2023) *Educational Data Mining: Prospects and Applications*, Authorea Preprints, Institute of Engineering 2, Nepal.
- Ding, W., Jing, X., Yan, Z. and Yang, L.T. (2019) 'A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion', *Information Fusion*, Vol. 51, No. 2019, pp.129–144.
- Dixit, P., Nagar, H. and Dixit, S. (2021) 'The role of multi-agent system based educational data mining techniques for student performance prediction', *Design Engineering*, Vol. 10, No. 7, pp.2752–2767.
- El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A. and El Alloui, Y. (2020) 'A multiple linear regression-based approach to predict student performance', in Ezziyyani, M. (Ed.): *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)*, pp.9–23, Springer International Publishing, Cham.
- Fernandes, E., Carvalho, R., Holanda, M. and Van Erven, G. (2017) 'Educational data mining: Discovery standards of academic performance by students in public high schools in the federal district of brazil', *World Conference on Information Systems and Technologies*, Springer, pp.287–296.
- Galla, B.M., Wood, J.J., Tsukayama, E., Har, K., Chiu, A.W. and Langer, D.A. (2014) 'A longitudinal multilevel model analysis of the within-person and between-person effect of effortful engagement and academic self-efficacy on academic performance', *Journal of School Psychology*, Vol. 52, No. 3, pp.295–308.
- Garcia, J.D. and Skrita, A. (2019) 'Predicting academic performance based on students' family environment: evidence for Colombia using classification trees', *Psychology, Society and Education*, Vol. 11, No. 3, pp.299–311.
- Gunasekara, S. and Saarela, M. (2024) 'Explainability in educational data mining and learning analytics: an umbrella review', *International Conference on Educational Data Mining*, pp.887–892.
- Han, J. and Kamber, M. (2006) *Data Mining: Concepts and Techniques*, 2nd, Morgan Kaufmann, University of Illinois at Urbana Champaign.
- Han, J., Kamber, M. and Pei, J. (2011) *Data Mining Concepts and Techniques*, 3d edition, Vol. 5, pp.83–124, The Morgan Kaufmann Series in Data Management Systems, Università degli Studi di Milano, Chapter 2.
- Hand, D., Mannila, H. and Smyth, P. (2001) *Principles of Data Mining*, MIT Press, Cambridge MA.
- Harwati, A.A.P. and Wulandari, F.A. (2014) 'Mapping student's performance based on data mining approach', *The 2014 International Conference on Agro-industry (ICoA): Competitive and Sustainable Agroindustry for Human Welfare*, pp.173–177.
- Hasani, A.A. and Bazrafshan, M. (2018) 'Analyzing students' educational information to evaluate their success via using data mining method (case study: faculty of management and industrial engineering, Shahrood university of technology)', *Journal of Management and Planning In Educational System*, Vol. 11, No. 2, pp.187–208.
- He, X. (2024) 'Detection method of students' online learning state based on posture recognition', *International Journal of Business Intelligence and Data Mining*, Vol. 24, Nos. 3–4, pp.278–292.

- Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C. and Liao, S.N. (2018) 'Predicting academic performance: a systematic literature review', *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, Association for Computing Machinery, Larnaca, Cyprus.
- Hoffait, A-S. and Schyns, M. (2017) 'Early detection of university students with potential difficulties', *Decision Support Systems*, September, Vol. 101, No. 9, pp.1–11.
- Hoyos, W. and Caicedo-Castro, I. (2024) 'Tuning data mining models to predict secondary school academic performance', *Data*, Vol. 9, No. 7, p.86.
- Hssina, B., Merbouha, A., Ezzikouri, H. and Erritali, M. (2014) 'A comparative study of decision tree ID3 and C4. 5', *International Journal of Advanced Computer Science and Applications*, Vol. 4, No. 2, pp.13–19.
- Hu, X., Cheong, C.W.L., Ding, W. and Woo, M. (2017) 'A systematic review of studies on predicting student learning outcomes using learning analytics', *Proceedings of the Seventh International Learning Analytics andamp; Knowledge Conference*, Association for Computing Machinery, Vancouver, British Columbia, Canada.
- Hung, H-C., Liu, I.F., Liang, C-T. and SU, Y-S. (2020) 'Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education', *Symmetry*, Vol. 12, No. 2, p.213.
- Hussain, M.M., Akbar, S., Hassan, S.A., Aziz, M.W. and Urooj, F. (2024) 'Prediction of student's academic performance through data mining approach', *Journal of Informatics and Web Engineering*, Vol. 3, No. 1, pp.241–251.
- Hussain, S., Dahan, N.A., Ba-Alwib, F.M. and Ribata, N. (2018) 'Educational data mining and analysis of students' academic performance using WEKA', *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 9, No. 2, pp.447–459.
- Jayaprakash, S., Krishnan, S. and Jaiganesh, V. (2020) 'Predicting students academic performance using an improved random forest classifier', *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, pp.238–243.
- Jolliffe, I.T. and Cadima, J. (2016) 'Principal component analysis: a review and recent development', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 374, No. 2065, p.20150202.
- Kabakchieva, D. (2013) 'Predicting student performance by using data mining methods for classification', *Cybernetics and Information Technologies*, Vol. 13, No. 1, pp.61–72.
- Karo, I., Fajari, M., Fadhillah, N. and Wardani, W. (2022) 'Benchmarking Naïve Bayes and ID3 algorithm for prediction student scholarship', *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, Vol. 1232, No. 1, p.12002.
- Kaur, K. and Dahiya, O. (2023) 'Role of educational data mining and learning analytics techniques used for predictive modeling', *2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, IEEE, pp.1–6.
- Khasanah, A.U. (2017) 'A comparative study to predict student's performance using educational data mining techniques', *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, p.12036.
- Kulkarni, V.Y. (2014) *Effective Learning and Classification Using Random Forest Algorithm*, Savitribai Phule Pune University, College of Engineering Pune.
- Lei, C. and Li, K.F. (2015) 'Academic performance predictors', *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*, IEEE, pp.577–581.
- Lever, J., Krzywinski, M. and Altman, N. (2016) 'Points of significance: model selection and overfitting', *Nature Methods*, Vol. 13, No. 9, pp.703–705.

- Machado, M.R., Karray, S. and De Sousa, I.T. (2019) 'LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry', *2019 14th International Conference on Computer Science and Education (ICCSE)*, IEEE, pp.1111–1116.
- Manoharan, R., Stalin, M.S. and Loganathan, G.B. (2023) *Ensemble Model for Educational Data Mining Based on Synthetic Minority Oversampling Technique*, Apollo Engineering College.
- Marjan, M.A., Uddin, M.P. and IBN Afjal, M. (2023) 'An educational data mining system for predicting and enhancing tertiary students' programming skill', *The Computer Journal*, Vol. 66, No. 5, pp.1083–1101.
- Mchugh, M.L. (2012) 'Interrater reliability: the Kappa statistic', *Biochemia Medica*, Vol. 22, No. 3, pp.276–282.
- Mishra, T., Kumar, D. and Gupta, S. (2014) 'Mining students' data for prediction performance', *2014 Fourth International Conference on Advanced Computing and Communication Technologies*, IEEE, pp.255–262.
- Mueen, A., Zafar, B. and Manzoor, U. (2016) 'Modeling and predicting students' academic performance using data mining techniques', *International Journal of Modern Education and Computer Science*, Vol. 8, No. 11, pp.36–42.
- Musso, M.F., Hernández, C.F.R. and Cascallar, E.C. (2020) 'Predicting key educational outcomes in academic trajectories: a machine-learning approach', *Higher Education*, Vol. 80, No. 5, pp.875–894.
- Muthukrishnan, S., Govindasamy, M. and Mustapha, M. (2017) 'Systematic mapping review on student's performance analysis using big data predictive model', *Journal of Fundamental and Applied Sciences*, Vol. 9, No. 4, pp.730–758.
- Mwadulo, M.W. (2016) *A Review on Feature Selection Methods for Classification Tasks*, Vol. 5, No. 6, pp.395–402.
- Nghe, N.T., Janecek, P. and Haddawy, P.A. (2007) 'Comparative analysis of techniques for predicting academic performance', *2007 37th Annual Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, IEEE, T2G-7-T2G-12.
- Nijhawan, V.K., Madan, M. and Dave, M. (2017) 'The analytical comparison of ID3 and C4. 5 using WEKA', *International Journal of Computer Applications*, Vol. 167, No. 11, pp.1–4.
- Nisbet, R., Elder, J. and Miner, G. (2009) *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press, Elsevier.
- Ogunde, A.O. and Ajibade, D.A. (2014) 'A data mining system for predicting university students' graduation grades using ID3 decision tree algorithm', *Journal of Computer Science and Information Technology*, Vol. 2, No. 1, pp.21–46.
- Oreski, D., Pihir, I. and Konecki, M. (2017) 'Crisp-DM process model in educational setting', *Economic and Social Development: Book of Proceedings*, pp.19–28.
- Oskouei, R.J. and Askari, M. (2014) 'Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies)', *Computer Engineering and Applications Journal*, Vol. 3, No. 2, pp.79–88.
- Osmanbegović, E., Suljić, M. and Agić, H. (2014) 'Determining dominant factor for students performance prediction by using data mining classification algorithms', *Tranzicija*, Vol. 16, No. 34, pp.147–158.
- Pal, M. (2005) 'Random forest classifier for remote sensing classification', *International Journal of Remote Sensing*, Vol. 26, No. 1, pp.217–222.
- Pandey, U.K. and Pal, S. (2011) *Data Mining: a Prediction of Performer or Underperformer Using Classification*, arXiv preprint arXiv:1104.4163.
- Punlumjeak, W. and Rachburee, N. (2015) 'A comparative study of feature selection techniques for classify student performance', *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 29–30 October, pp.425–429.

- Qi, H. (2024) 'Knowledge mapping-based online teaching resource recommendation method for Chinese education', *International Journal of Business Intelligence and Data Mining*, Vol. 24, Nos. 3–4, pp.252–265.
- Quinlan, J.R. (1986) 'Induction of decision trees', *Machine Learning*, Vol. 1, No. 1986, pp.81–106.
- Rabaani, F.S.A., Said, A.M., Fageeri, S.O. and Alabdulsalam, A.K. (2023) 'Using the data mining technique to predict successful customer engagement of marketing campaigns in social media', *International Journal of Business Intelligence and Data Mining*, Vol. 23, No. 2, pp.166–183.
- Rahm, E. and Do, H.H. (2000) 'Data cleaning: problems and current approaches', *IEEE Data Eng. Bull.*, Vol. 23, No. 4, pp.3–13.
- Rahman, L., Setiawan, N.A. and Permanasari, A.E. (2017) 'Feature selection methods in improving accuracy of classifying students' academic performance', *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, pp.267–271.
- Raosoftware (2004) *Sample Size Calculator* [online] <http://www.raosoftware.com/samplesize.html>.
- Rebai, S., Yahia, F.B. and Essid, H. (2020) 'A graphically based machine learning approach to predict secondary schools performance in Tunisia', *Socio-Economic Planning Sciences*, Vol. 70, No. 2020, p.100724.
- Rizvi, S., Rienties, B. and Khoja, S.A. (2019) 'The role of demographics in online learning; a decision tree based approach', *Computers and Education*, Vol. 137, No. 10, pp.32–47.
- Romero, C. and Ventura, S. (2007) 'Educational data mining: a survey from 1995 to 2005', *Expert Systems with Applications*, Vol. 33, No. 1, pp.135–146.
- Romero, C. and 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)*, Vol. 40, No. 6, pp.601–618.
- Romero, C. and Ventura, S. (2013) 'Data mining in education', *WIREs Data Mining and Knowledge Discovery*, Vol. 3, No. 1, pp.12–27.
- Romero, C. and Ventura, S. (2020) 'Educational data mining and learning analytics: an updated survey', *WIREs Data Mining and Knowledge Discovery*, Vol. 10, No. 3, p.e1355.
- Ross Quinlan, J. (1993) 'Book Of C4. 5: programs for machine learning', *Mach. Learn.*, Morgan Kaufmann Publishers.
- Rostami, M., Ayat, S., Saghari, F. and Yaghoobi, F. (2015) 'Applying fuzzy clustering to assess and anticipate students' educational progress in learning environments', *Technology of Education*, Vol. 9, No. 4, pp.245–258.
- Russell, S.J. and Norvig, P. (2016) *Artificial Intelligence: a Modern Approach*, Malaysia Pearson Education Limited London, UK.
- Saa, A.A. (2016) 'Educational data mining and students' performance prediction', *International Journal of Advanced Computer Science and Applications*, Vol. 7, No. 5, pp.212–220.
- Salal, Y. and Abdullaev, S. (2019) 'Optimization of classifiers ensemble construction: case study of educational data mining', *Вестник Южно-Уральского государственного университета, Компьютерные технологии, управление, радиоэлектроника, Серия*, Vol. 19, No. 4, pp.139–143.
- Sarker, S., Paul, M.K., Thasin, S.T.H. and Hasan, M.A.M. (2024) 'Analyzing students' academic performance using educational data mining', *Computers and Education: Artificial Intelligence*, Vol. 7, No. 2, pp.1–16.
- Sarlis, N.V. and Christopoulos, S-R.G. (2014) 'Visualization of the significance of receiver operating characteristics based on confidence ellipses', *Computer Physics Communications*, Vol. 185, No. 3, pp.1172–1176.
- Shaikh, A., Mahoto, N., Khuhawar, F. and Memon, M. (2015) 'Performance evaluation of classification methods for heart disease dataset', *Sindh University Research Journal-SURJ (Science Series)*, Vol. 47, No. 3, pp.1172–1176.

- Simundic, A-M. (2008) 'Confidence interval', *Biochemia Medica*, Vol. 18, No. 2, pp.154–161.
- Singh, D. and Singh, B. (2020) 'Investigating the impact of data normalization on classification performance', *Applied Soft Computing*, Longowal, Punjab, India, Vol. 97, No. 13 (Part B), p.105524.
- Sokkhey, P. and Okazaki, T. (2020) 'Hybrid machine learning algorithms for predicting academic performance', *Int. J. Adv. Comput. Sci. Appl.*, Vol. 11, No. 1, pp.32–41.
- Sokolova, M. and Lapalme, G. (2009) 'A systematic analysis of performance measures for classification tasks', *Information Processing and Management*, Vol. 45, No. 4, pp.427–437.
- Stinar, F., Xiong, Z. and Bosch, N. (2024) 'An approach to improve k-anonymization practices in educational data mining', *Journal of Educational Data Mining*, Vol. 16, No. 1, pp.61–83.
- Surjeet, K. and Saurabh, P. (2012) 'data mining: a prediction for performance improvement of engineering students using classification', *WCSIT*, Vol. 2, No. 2, pp.51–56, ISSN.
- Tan, P., Steinbach, M. and Kumar, V. (2013) *Introduction to Data Mining: Pearson New International Edition (English Edition)*, Pearson Education Limited, Harlow, Essex, UK.
- Ting, K.M. (2017) 'Confusion matrix', *Encyclopedia of Machine Learning and Data Mining*, Vol. 260 No. 2017, p.260.
- Triayudi, A., Aldisa, R.T. and Sumiati, S. (2024) *New Framework of Educational Data Mining to Predict Student Learning Performance*, March, Vol. 15, No. 1, pp.115–132.
- Turabieh, H. (2019) 'Hybrid machine learning classifiers to predict student performance', *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, 9–11 October, pp.1–6.
- Umer, R., Susnjak, T., Mathrani, A. and Suriadi, S. (2017) 'On predicting academic performance with process mining in learning analytics', *Journal of Research in Innovative Teaching and Learning*, Vol. 10, No. 2, pp.160–176.
- Vapnik, V. (1995) *The Nature of Statistical Learning*, Theo~. Springer Verlag, New York.
- Vladimir, N.V. and Vapnik, V. (1998) 'Statistical learning theory', in Xu, J.H. and Zhang, X.G. (Eds.): *Translation*, 2004, Publishing House of Electronics Industry, Beijing.
- Waheed, H., Hassan, S-U., Aljohani, N.R., Hardman, J., Alelyani, S. and Nawaz, R. (2020) 'Predicting academic performance of students from VLE big data using deep learning models', *Computers In Human Behavior*, Vol. 104, No. 2019, p.106189.
- Wirth, R. and Hipp, J. (2000) 'CRISP-DM: towards a standard process model for data mining', *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Manchester, pp.29–40.
- Wood, J.M. (2007) 'Understanding and computing Cohen's Kappa: a tutorial', *WebPsychEmpiricist*, October, Vol. 3, pp.145–160.
- Xiao, Z. (2024) 'Research on classification of educational digital resources based on KNN algorithm', *International Journal of Business Intelligence and Data Mining*, Vol. 24, Nos. 3–4, pp.352–363.
- Xing, W., Guo, R., Petakovic, E. and Goggins, S. (2015) 'Participation-based student final performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory', *Computers in Human Behavior*, Vol. 47, No. 6, pp.168–181.
- Xu, X., Wang, J., Peng, H. and Wu, R. (2019) 'Prediction of academic performance associated with internet usage behaviors using machine learning algorithms', *Computers in Human Behavior*, Vol. 98, No. 9, pp.166–173.
- Yadav, S.K., Bharadwaj, B. and Pal, S. (2012) 'Data mining applications: a comparative study for predicting student's performance', arXiv preprint arXiv:1202.4815.
- Yağci, M. (2022) 'Educational data mining: prediction of students' academic performance using machine learning algorithms', *Smart Learning Environments*, Vol. 9, No. 11, pp.1–19.

- Yehuala, M.A. (2015) 'Application of data mining techniques for student success and failure prediction (The case of Debre Markos university)', *International Journal of Scientific and Technology Research*, Vol. 4, No. 4, pp.91–94.
- Yulianto, L.D., Triayudi, A. and Sholihati, I.D. (2020) 'Implementation educational data mining for analysis of student performance prediction with comparison of K-nearest neighbor data mining method and decision tree C4. 5', *Jurnal Mantik*, Vol. 4, No. 1, pp.441–451.
- Zhang, Y., Oussena, S., Clark, T. and Kim, H. (2010) 'Use data mining to improve student retention in He-A case study', in *12th International Conference on Enterprise Information Systems (ICEIS 2010)*, Funchal, Portugal, 8–12 June.
- Zulfiker, M.S., Kabir, N., Biswas, A., Chakraborty, P. and Rahman, M.M. (2020) 'Predicting students' performance of the private universities of Bangladesh using machine learning approaches', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 3, pp.672–679.