

Educational Data Mining: Comparison of Models for Predicting Non-Academic Students

Aldin Febriansyah¹, Ihsan Fauzi²

¹Universitas Negeri Surabaya, Surabaya, Indonesia

²Department Research and Development, SEAMOLEC, Indonesia

*aldin.20005@mhs.unesa.ac.id

Abstract

Passing the national assessment test in Bahasa Indonesia for Paket C students is still a challenge, especially for non-academic students. This is caused by various factors, such as educational background, economic conditions, and learning motivation. This research aims to develop a prediction model for Package C students passing the national assessment exam in Indonesian using four machine learning algorithms, namely neural network, logistic regression, support vector machine, and naive bayes. The data used in this study are data on equivalency education exam tryout scores, equivalency education exam scores, schools, and regional information from 1,240 non-academic students who took the Equivalency Education Exam in the Indonesian language subject in the 2020-2021 academic year. This research uses machine learning methods in the context of Educational Data Mining. The preliminary analysis results show that the four machine learning algorithms can be used to predict the graduation of Paket C students with a fairly high accuracy. The neural network algorithm shows the best performance with 57.5% accuracy. SVM, LR, and NB algorithms achieved 56.2%, 54.8%, and 48.4% accuracy, respectively. The results of this study have the potential to increase the pass rate of Package C students on the Indonesian language national assessment test. The prediction model developed can be used to identify students at risk of not graduating, so that appropriate educational interventions can be provided.

Keywords: Educational Data Mining, Machine Learning, Accuracy, Paket C students, Equivalency Education Examination.

Background

Education is an important aspect of human existence as it improves the quality of life and helps individuals reach their full potential. The National Education System Law No.20 of 2003 emphasises that education is conducted through three pathways emphasises that education is conducted through three channels, namely: formal education, non-formal education, and informal education. Formal education is conducted in schools, non-formal education is carried out in the community, and informal education is mainly carried out in the family. Equivalency education provides access to non-formal levels of education for those who cannot attend formal education.

The low quality of education is believed to be the cause of the low quality of human resources. The demand for improving the quality of education lies not only in improving the quality of inputs and outputs, but also the quality of the process driven by the managerial power of the school principal (Rochayati et al., 2023). Management is a necessity to facilitate the achievement of human goals in organisations, as well as to manage various organisational resources, such as facilities and infrastructure, time, human resources, methods and others effectively, innovatively, creatively, solutively and efficiently.

The online seTARA application was developed in 2016 and inaugurated in 2018 by the Directorate of Literacy and Equality Education of the Ministry of Education and Culture (Fibrianti & Suhanadji, 2020). The online

seTARA application is a learning media for equivalency programmes in the form of e-learning (Wica & Marzuki, 2021). E-learning is expected to be able to solve problems in the world of education because e-learning can be adjusted to the needs as a complement or substitute for learning activities in the classroom (Alshehri et al., 2019). Distance learning helps students who cannot participate in face-to-face learning activities (Eddy & Suryono, 2019).

The problem faced by non-academic students in the equivalency education exam is the low percentage of passes. This is caused by various factors, such as educational background, economic conditions and learning motivation. However, the main problem faced is the low quality of education. The low quality of education can cause students to be unable to follow the lessons properly, causing the pass percentage to decrease.

To increase the percentage of non-academic students passing the equivalency education exam, one possible solution is to utilise data mining techniques. Data mining can be used to identify factors that influence the pass rate among non-academic students. By finding patterns and trends in large data sets, it is possible to create accurate prediction models that can forecast pass rates. This will greatly improve the effectiveness of equivalency education in Indonesia.

Literature Review

Many studies have focused on analyzing e-learning systems in the field of Educational Data Mining (EDM). Some studies have categorized educational information, and others have attempted to forecast student performance.

According to Cruz-Jesús et al. (2020), it is possible to predict student academic performance by evaluating 16 demographic factors including age, gender, attendance, internet access, computer ownership, and number of courses taken. The study used various machine learning techniques, such as Random Forest, Logistic Regression, k-Nearest Neighbors, and Support Vector. The results showed that these methods were able to predict student performance with accuracy between 50% and 81%.

Bernacki et al. (2020) investigated the relationship between student achievement and LMS logs. Their findings showed that 75% of students who were required to retake the course were correctly predicted by the behavior-based prediction model. The study also demonstrated how the approach can identify students who are going to be at-risk for the next semester and offer them the necessary help.

Waheed et al. (2020) modeled an artificial neural network using student records of how they navigated a Learning Management System (LMS). The study revealed that students' demographics and clickstream activity were significant predictors of their academic performance. These findings suggest that students who actively explore the subject matter tend to achieve better academic results. Although there was no significant correlation between students' participation in the learning environment and their academic performance.

Fernandes et al. (2019) developed a model that combines demographic characteristics and student performance scores collected from semester activities to predict academic performance. This study used a classification model based on Gradient Boosting Machine (GBM). The results showed that the best predictors for achievement were the previous year's achievement and absence rate. The researchers also found that demographic factors such as neighborhood, school and age can be potential indicators for academic success or failure. They argue that this model can help formulate new policies aimed at preventing academic failure.

Xu et al. (2019) demonstrated a deep learning model has the ability to accurately predict students' academic achievement right from the start. This study used machine learning techniques for predictive modeling to look at the connection between academic achievement and internet usage habits among college students. With a high degree of accuracy, the research model accurately predicted the academic achievement of the pupils. The results showed that academic performance and internet connection frequency were positively correlated, but academic performance and internet traffic volume were negatively correlated. Furthermore, the study came to the conclusion that a student's academic performance is significantly influenced by their internet usage habits.

Table 1 presents a comparative comparison of research that used machine learning techniques to predict academic achievement grades. Previous studies have demonstrated the remarkable accuracy with which machine learning algorithms can forecast students' academic achievement. Random Forest, Genetic Programming, and Naive Bayes are the best machine learning methods for predicting academic success.

Students' academic achievement is influenced by a variety of aspects and traits, including course performance, student involvement, motivation, habits, social and financial difficulties, lack of advancement, and job choices. It is possible to predict pupils' academic achievement using these indicators, but gathering and evaluating the data might be difficult.

The goal of this research is to develop a novel model that can forecast academic achievement of students based just on their grades, without taking socioeconomic background or demographic data into account. Using machine learning algorithms, this model will identify students who are at danger of falling short of their academic potential and assist in the creation of new educational interventions and policies aimed at raising academic performance.

Research Method

This section describes in detail the dataset, pre-processing techniques, and machine learning algorithms used to analyze the data.

Dataset

The dataset used in this study is the data of equivalency education exam tryout scores and Indonesian language equivalency education exam scores of Paket C students. The data is obtained from the online seTARA website dataset. The data contains the participant id, the name of the parent school that organizes the exam, the tryout score of the Indonesian equivalency education exam, the score of the Indonesian equivalency education exam, the regional origin of the parent school, and the province of the parent school. Table 2 presents the distribution of non-academic students by parent school.

The collected data will be cleaned to remove outliers and incomplete data. Outliers are data that are far away from other values in the data. Incomplete data is data that has empty values. The cleaned data will be processed to remove irrelevant features. Relevant features are features that can be used to predict the performance of the Indonesian final exam.

Data identification and collection

Pre-processing technique is a process done to clean and prepare the data before it is analyzed. This technique is important to do to ensure that the data used for analysis is valid and reliable.

Some of the pre-processing techniques that will be carried out on this research dataset include Data cleaning is the process of removing outliers and incomplete data, Outliers are data that are far from other values in the data (incomplete data is data that has empty values), Data transformation is the process of changing the format of the data so that it can be analyzed more easily (for example, nominal data can be converted into numeric data), Feature selection is the process of selecting relevant features from the data, Relevant features are features that can be used to predict Indonesian final exam performance.

Building the DM Model and Implementing the Algorithm

The pass rate of Paket C students on the Indonesian national assessment test will be predicted using four machine learning techniques in this study: neural networks, logistic regression, support vector machines, and naive bayes. This choice was made after taking into account a number of criteria, including comparing the accuracy of each prediction method, increasing our knowledge of the variables influencing students' academic success, and creating a more thorough and accurate prediction model.

Table 1
Comparison Analysis

Reference	Variable	Object	Degree	Dataset	Algoritma	Accuracy	
						Min	Max
Cruz-Jesús dkk. (2020)	Year of study cycle, gender, age, number of years enrolled in secondary school, scholarships, internet access, class size, school size, economic level, population density, number of course units attended	Predicting student performance	high school students	110627	ANN, DT, ET, RF, SVM, KNN, LR	LR (81.1 %)	SVM (51.2 %)
Fernandes dkk. (2019)	Classes with people with special needs, Environment of use classroom, Gender, age (average), Student advantage, city, neighborhood, Students with special needs, Class (average), Absenteeism (average)	Predicting academic outcomes from student performance	high school students	Dataset1: 19000 Dataset2: 19834	Gradient Boosting Machine	89.5%	91.9%
Waheed dkk. (2020)	Student demographics, clickstream events	Pass-fail, withdrawn-pass, distinction-fail, distinction-pass	Undergraduate students	32593	ANN, SVM, LR	84%	93%
Xu dkk. (2019)	Internet usage consists of time online, frequency of connection between networks, volume of internet traffic, and duration online	Predicting student performance	Undergraduate students	4000	DT, NN, SVM	71%	76%
Bernacki dkk. (2020)	Records in learning management system	Predicting performance	Undergraduate students	337	LR, NB, J-48 DT, J-Rip DT	J-48 (53.71 %)	LR (67.36 %)

Logistic regression, support vector machine, neural network, and naive bayes are different prediction techniques, each with its own advantages and disadvantages. Logistic regression is a prediction technique that is simple and easy to implement, but its accuracy may not be as good as other prediction techniques. Support vector machine is a more complex prediction technique that requires more resources, but generally has higher accuracy than logistic regression. Neural networks are highly accurate prediction techniques, but require significant resources due to their complexity. On the other hand, Naive Bayes is a simple and fast prediction technique, but its accuracy may not be as good as other methods.

It is possible to identify which prediction method is more successful in projecting kids' academic achievement based on their grades by comparing the accuracy of each method. To create a prediction model that identifies children who are at danger of underachieving, the most accurate method will be employed.

The two primary goals of data mining (DM) are descriptive modeling and predictive modeling. The goal of predictive modeling is to forecast via the examination of past data. Predictive modeling is utilized in this study to forecast the pass rate of Paket C students on the national assessment test in Indonesia. In order to get understanding and make decisions, descriptive modeling looks for patterns and trends in the data that already exists.

Models that can precisely predict output values based on available input data are created using statistical approaches. Conversely, machine learning techniques solve supervised optimization problems to automatically construct models that link input data to predicted goal values.

Table 2
the Dataset

Parent School Name	Number of Students
Direktorat Pendidikan Masyarakat dan Pendidikan Khusus	155
PKBM Al-Madinah Kota Kediri	56
PKBM Ar-Rahman	35
PKBM Bhakti Mandiri	54
PKBM Cahaya Iman	2
PKBM Cerdas	19
PKBM Cipta Karya Intelektual	25
PKBM Citra Bangsa	30
PKBM Citra Ilmu	9
PKBM Faradika	1
PKBM Laily	23
PKBM Lentera Hati	1
PKBM Mandiri Utama	68
PKBM Mashaghi	3
PKBM Mulia	20
PKBM Mutiara Kampar	80
PKBM Nosarara Nosabatutu	21
PKBM Nurhidayah	10
PKBM Nusantara	25
PKBM Padamu Negeri	81
PKBM Pandu Generasi Indonesia	45
PKBM Pandu Pelajar Mandiri	26
PKBM Pelita Riau	6
PKBM Permata	22
PKBM Permata Buana	13
PKBM Pioneer Palur	90
PKBM Pratama	38
PKBM Pratama Mandiri	15
PKBM Prospektif	17
PKBM Wahana Ilmu	36
SKB Blora	20
SKB Kab. Batang	27
SKB Kab. Jeneponto	106
SKB Kab. Bekasi	61

The confusion matrix indicator gauges the outcome of this model study. It is well established from the literature that no one classifier produces the best prediction outcomes. Investigating whether categorization is better suited for the data under review is therefore essential.

Result and Discussion

Throughout the whole trial, Orange machine learning software is a potent and user-friendly component-based programming tool designed for both experienced and inexperienced data scientists was utilized. Orange streamlines data analysis by combining workflows made up of widgets that manage different tasks including gathering data, pre-processing, visualizing, modeling, and assessing it. A workflow is, in this sense, a set of actions performed on the platform in order to accomplish a particular objective. As illustrated by the design workflow flowchart in Figure 1.

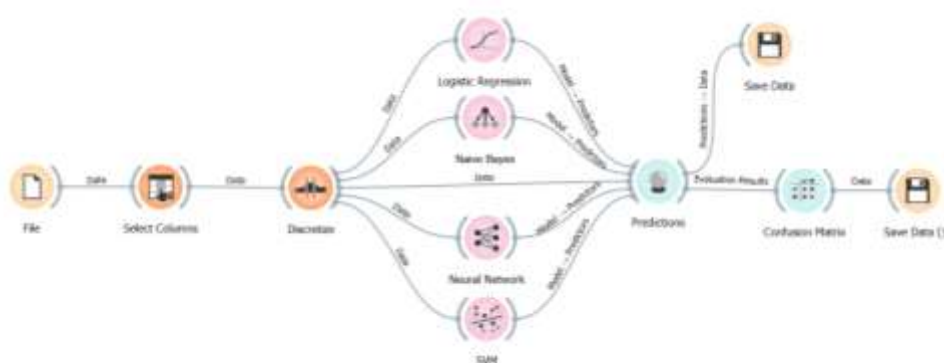


Image 1. Design of workflow model

The datasets used comprise the participant ID, the name of the parent school that is in charge of administering the exam, the results of the Indonesian language equivalency education exam, the results of the exam tryout, the results of the exam, the parent school province, and the origin of the parent school region of the 1240 non-academic students who took the exam in the 2020–2021 academic year. A subset of the dataset is displayed in Table 3.

The elements of the dataset include school and geographical information, equivalency education exam scores, and tryout test results. Every metric includes student-related data. The header "dataset" describes the variables equivalency education test scores and equivalency education tryout test scores. In the online seTARA website, the province, district/city, and school variables stand in for the school's geographic location, while the school variable represents the parent school. The equivalency education test score is designated as the dependent variable in the model formulation, whereas the parent school, district/city, province, and tryout score are indicated as independent factors. Table 4 displays the model that varies.

The discretization model with the same width is used to classify the equivalency education test scores and tryout results once the variable model has been established. The conversion parameters for equivalency education test results and tryout scores into a category format are displayed in Table 5.

The values in Table 6's first column correspond to the real values. The values that the suggested model predicts are those found in the SVM, LR, NB, and NN columns. For instance, Table 6 shows that the actual end result, 5c33768ba2f3844256cba896, is between 55 and 77. SVM, LR, and NN have projected values bigger than 77, however the NB model's values are in the range less than 32.

Model performance evaluation

The performance of the model is assessed using a number of measures, such as the confusion matrix, area under the ROC curve (AUC), recall, f-score (F1), precision, and classification accuracy (CA).

Confusion matrix

The confusion matrix, which indicates the proportion of accurate and inaccurate predictions the model made, is shown in Table 7. The number of cases that are successfully and wrongly identified indicates how well the model performs. The number of samples in the test set is shown by the lines, and the model estimates are indicated by the columns.

Table 3
Part of the dataset consists of 1240 rows

User ID	Parent School Name	Tryout	UPK	City	Province
5c33768ba2f3844256cba896	Direktorat Pendidikan Masyarakat dan Pendidikan Khusus	33	65	Central Jakarta City	DKI Jakarta
5d2d8aeea2f3847137fc8197	PKBM Citra Ilmu	88	63	Reg. Semarang	Central Java
5d2d8bcb2f384c942fc80ef	PKBM Citra Ilmu	43	53	Reg. Semarang	Central Java
5d2d8bd9a2f384c942fc80f1	PKBM Citra Ilmu	38	58	Reg. Semarang	Central Java
5d440987a2f384f91adbbcd0	PKBM Citra Ilmu	80	53	Reg. Semarang	Central Java
...	...	...	...	...	...
6412bf49a2f3849b37d3e0be	PKBM Mutiara Kampar	88	55	Reg. Kampar	Riau
64153baca2f384deded3e0df	PKBM Mutiara Kampar	88	50	Reg. Kampar	Riau

Table 4
Variable Models

Feature	Target variabel	Meta Attributes
Tryout	UPK	ID user
Parent School Name		
City		
Province		

Table 5
Category criteria

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

The number of successfully categorized cases in the True Negative (TN) and True Positive (TP) categories is shown in Table 6. When events are expected to be negative (0) when they should be positive (1), they are classified as False Positive (FP) cases. When instances are projected to be negative (0) when they should be positive (1), they are classified as False Negative (FN) cases.

The confusion matrix for the NN algorithm is displayed in Table 8. The proportion of properly predicted cases is displayed on the main diagonal of the 4×4 matrix, while the percentage of anticipated mistakes is displayed on the other components.

According to the table, 30.3% of those with less than 32.5, 34.3% of those in the range 32.5-55, 53.8% of those in the range 55-77.5, and 63.3% of those with real end scores more than 77.5 were all accurately predicted. For the other techniques, the confusion matrix is displayed in Tables 9, 10, and 11.

										UP K	Ta ble 6 Pr ob abi lity an d fin al de cisi on of pre dic tio n mo del s (L R, SV M, NB , an d NN)
$\geq$ 77.5	32.5 - 55	$\geq$ 77.5	32.5 - 55	55 - 77.5	32.5 - 55	55 - 77.5	32.5 - 55	55 - 77.5	55 - 77.5		
5f0	5ec	5e1	5df4	5d6	5d4			5d2	5c3	Us er ID	
ed7	35b	433	7e1	3a3	409	5d2	5d2	d8a	376		
93a	33a	99a	1a2f	a7a	87a	d8b	d8b	eea	8ba		
2f3	2f38	2f38	384	2f38	2f38	d9a	cba	2f38	2f38		
846	4da	444	382	451	4f91	2f38	2f38	471	442		
747	317	6dff	30c	137	adb	4c9	4c9	37fc	56c	NB	
944	151	3be	dcc	cc3	bed	42fc	42fc	819	ba8		
3d1	32	c	6	32	0	80f1	80ef	7	96		
55 - 77.5	32.5 - 55	32.5 - 55	32.5 - 55	32.5 - 55	32.5 - 55	32.5 - 55	32.5 - 55	32.5 - 55	< 32.5	LR	
$\geq$ 77.5	32.5 - 55	32.5 - 55	55 - 77.5	32.5 - 55	32.5 - 55	32.5 - 55	32.5 - 55	32.5 - 55	$\geq$ 77.5		
$\geq$ 77.5	55 - 77.5	55 - 77.5	32.5 - 55	55 - 77.5	55 - 77.5	32.5 - 55	32.5 - 55	55 - 77.5	$\geq$ 77.5		
$\geq$ 77.5	55 - 77.5	55 - 77.5	32.5 - 55	55 - 77.5	55 - 77.5	55 - 77.5	55 - 77.5	55 - 77.5	$\geq$ 77.5	SV M	
$\geq$ 77.5	55 - 77.5	55 - 77.5	32.5 - 55	55 - 77.5	55 - 77.5	55 - 77.5	55 - 77.5	55 - 77.5	$\geq$ 77.5		

PK BM Al- Ma dina h Kot a Ked iri	PK BM Citr a Ilmu	PK BM Citr a Ilmu	PK BM Citr a Ilmu	PK BM Citr a Ilmu	PK BM Citr a Ilmu	PK BM Citr a Ilmu	PK BM Citr a Ilmu	PK BM Citr a Ilmu	Dire ktor at Pen didi kan Mas yara kat Dan Pen didi kan Khu sus	Par ent Sc ho ol	
≥ 77. 5	≥ 77. 5 Reg	≥ 77. 5 Reg	< 32.5 Reg	≥ 77. 5 Reg	≥ 77. 5 Reg	32.5 - 55 Reg	32.5 - 55 Reg	≥ 77. 5 Reg	32.5 - 55 Cen tral Jak arta City	Try out	
Ked iri City	Se mar ang	Se mar ang	Se mar ang	Se mar ang	Se mar ang	Se mar ang	Se mar ang	Se mar ang		Cit y	
Eas t Jav a	Cen tral Jav a	Cen tral Jav a	Cen tral Jav a	Cen tral Jav a	Cen tral Jav a	Cen tral Jav a	Cen tral Jav a	Cen tral Jav a	DKI Jak arta	Pro vin ce	

Table 7
The Confusion matrix

		Prediction	
		Positive (1)	Negative (0)
Aktual	Positive (1)	TP	FP
	Negative (0)	FN	FN

Table 8
Confusion matrix from the NN algorithm

		Prediction				
		< 32.5	32.5 – 55	55 – 77.5	≥ 77.5	Total
Actual	< 32.5	33.3%	21.2%	0.3%	8.2%	89
	32.5 – 55	19.0%	34.3%	14.6%	11.6%	180
	55 – 77.5	14.3%	20.2%	53.8%	16.8%	356
	≥ 77.5	33.3%	24.2%	31.3%	63.3%	615
Sum		21	99	390	730	1240

Table 9
Confusion matrix from the SVM algorithm

		Prediction				
		< 32.5	32.5 – 55	55 – 77.5	≥ 77.5	Total
Actual	< 32.5	NA	0.0%	0.0%	9.9%	89
	32.5 – 55	NA	100.0%	15.8%	13.8%	180
	55 – 77.5	NA	0.0%	54.0%	19.4%	356
	≥ 77.5	NA	0.0%	30.1%	56.9%	615
	Sum	0	2	335	903	1240

Table 10
Confusion matrix from the LR algorithm

		Prediction				
		< 32.5	32.5 – 55	55 – 77.5	≥ 77.5	Total
Actual	< 32.5	31.6%	0.0%	0.0%	8.8%	89
	32.5 – 55	21.1%	44.4%	14.8%	14.0%	180
	55 – 77.5	5.3%	44.4%	54.4%	21.7%	356
	≥ 77.5	42.1%	11.1%	30.7%	55.5%	615
	Sum	19	9	270	942	1240

Table 11
Confusion matrix from the NB algorithm

		Prediction				
		< 32.5	32.5 – 55	55 – 77.5	≥ 77.5	Total
Actual	< 32.5	23.3%	5.4%	0.0%	1.6%	89
	32.5 – 55	22.4%	27.0%	14.1%	6.9%	180
	55 – 77.5	18.8%	29.7%	47.4%	19.9%	356
	≥ 77.5	35.5%	37.8%	38.5%	71.6%	615
	Sum	335	74	384	447	1240

Classification accuracy: The ratio of accurate predictions (TP+TN) to all instances (TP+TN+FP+FN) is known as the CA.

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

Precision: Precision is defined as the ratio of the total number of projected positive occurrences to the number of accurately categorized positive instances. We get values in the interval [0,1].

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

Recall: The ratio of properly identified positive examples to all occurrences whose real class is positive is known as recall. Another name for recall is the true positive rate. We get values in the interval [0,1].

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

Criterion F (F1): Recall and precision have the inverse connection. Therefore, for more precise and sensitive answers, the harmonic mean of both criteria is computed. The F criteria is the name given to this standard.

$$F - \text{Criterion} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Receiver operating characteristics (ROC) curve

AUC-ROC curves are used to assess the classification problem's performance. AUC-ROC, a metric that gauges a model's predictive accuracy, is frequently used to evaluate the efficacy of machine learning algorithms, particularly in cases when datasets are imbalanced.

AUC: Area under the ROC curve

Larger coverage makes machine learning algorithms more effective at differentiating between different classes. One is the optimal AUC value. Table 12 displays the models' AUC, recall value, accuracy, F-Criteria (F1), and classification accuracy (CA).

The NN, SVM, LR, and NB algorithms have AUC values of 0.793, 0.739, 0.761, and 0.735, in that order. The NN, SVM, LR, and NB algorithms had classification accuracies of 0.575, 0.562, 0.548, and 0.484, in that order. These results showed that the NN algorithm's accuracy was 57.5%. This shows that the expected and actual data have a significant connection, leading to the accurate categorization of 57.5% of the samples.

Conclusion and Suggestion

Using the mid-exam results as raw data, this study suggests a novel model based on machine learning algorithms to forecast the final test scores of undergraduate students. To predict students' final exam scores, the performance metrics of machine learning methods, such as closest neighbor, support vector machines, logistic regression, and naïve bayes, are computed and compared. Two parameters are the main focus of this study. Predicting academic success based on past accomplishment rankings is the first parameter. The second is a comparison of machine learning algorithms' performance metrics.

The findings demonstrate that the suggested model attains a 48–58% classification accuracy. Based on these findings, it can be concluded that midterm test scores of students are significant indicators that have to be taken into consideration when estimating their final exam scores. Algorithms with sufficient accuracy to forecast students' final test scores include NN, SVM, LR, and NB. Furthermore, just three categories of parameters were used to make the predictions: school data, region data, and tryout test results.

Table 12
Model AUC, CA, F1, precision, and recall values

Model	AUC	Classification Accuracy	F1	Precision	Recall
Naïve Bayes	0.735	0.484	0.489	0.547	0.484
Logistic Regression	0.761	0.548	0.482	0.519	0.548
Neural Network	0.793	0.575	0.547	0.542	0.575
SVM	0.739	0.562	0.489	0.583	0.562

The findings of this study are in line with earlier studies that demonstrate how well students' midterm test scores predict their future academic success. For instance, Waheed et al.'s research from 2020 indicated that there was an 85% accuracy rate in forecasting students' success or failure based on their midterm test results. Xu et al.'s (2019) research also shown that it is possible to predict and distinguish students' academic success based on their performance on midterm exams.

The study's findings also demonstrate the reasonable degree of accuracy with which student academic success may be predicted using machine learning algorithms. The accuracy of classification is higher for the LR, NN, and SVM algorithms compared to the NB method. This is in line with other studies that shown the efficacy of the LR, NN, and SVM algorithms in predicting students' academic success.

It may be inferred from the study's findings that machine learning models have a reasonable degree of accuracy when it comes to predicting students' academic success. Tryout test results, school data, and region data are the only three categories of factors included in the model presented in this study. This demonstrates that the model may reasonably be used to predict students' academic achievement.

By adding more factors as input variables and machine learning techniques to the modeling process, further research may be done. Utilizing the efficacy of DM techniques is also essential for addressing students' issues, examining their learning behaviors, improving the learning environment, and facilitating data-driven decision-making.

Reference

- Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. *Bulletin of Education and Research*, 40(3), 157–164.
- Alshehri, A., Rutter, M., & Smith, S. (2019). Assessing the Relative Importance of an Elearning System's Usability Design Characteristics Based on Students' Preferences. *European Journal of Educational Research*, 8(3), 839–855.
- Babić, I. D. (2017). Machine learning methods in predicting the student academic motivation. *Croatian Operational Research Review*, 8(2), 443–461. doi:<https://doi.org/10.17535/corr.2017.0028>
- Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education*, 158, 103999. doi:<https://doi.org/10.1016/j.compedu.2020.103999>
- Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon*, 6(6). doi:<https://doi.org/10.1016/j.heliyon.2020.e04081>
- Eddy, G. S., & Suryono, Y. (2019). Implementation of Online Learning Lessons in the Package C Program. *Journal of Nonformal Education*, 5(2), 117–124.
- Fibrianti, S., & Suhanadji. (2020). Analisis Penyelenggaraan 'Setara Daring' Sebagai Media Pembelajaran Pendidikan Luar Sekolah di Satuan Pendidikan Nonformal (SPNF) SKB Gudo Kabupaten Jombang. *Jurnal Pendidikan Untuk Semua*, 4(2), 36–45.
- Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of business research*, 94, 335–343. doi:<https://doi.org/10.1016/j.jbusres.2018.02.012>
- Hoffait, A.-S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, 101, 1–11. doi:<https://doi.org/10.1016/j.dss.2017.05.003>
- Sumpena, J., & Kurnia, N. (2019). Analisis prediksi kelulusan siswa pkbm paket c dengan metoda algoritma naïve bayes. *Jurnal TEDC*, 13(2), 127–133.
- Undang-Undang No.20 tahun 2003 tentang Sistem Pendidikan Nasional. Accessed on December 23, 2023 from <https://pusdiklat.perpusnas.go.id/regulasi/download/6>

- Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior*, 104, 106189. doi:<https://doi.org/10.1016/j.chb.2019.106189>
- Wica, P. Y., & Marzuki, M. (2021). Character Education with Setara Daring Application in Non-Formal Education. *JPI (Jurnal Pendidikan Indonesia)*, 10(4), 656–666.
- Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), 11. doi:<https://doi.org/10.1186/s40561-022-00192-z>
- Yuningsih, L., Setiawan, I. R., & Sunarto, A. A. (2020). Rancangan Aplikasi Prediksi Kelulusan Siswa Menggunakan Algoritma C4. 5. *Progresif: Jurnal Ilmiah Komputer*, 16(2), 121-132. doi:<https://doi.org/10.35889/progresif.v16i2.517>