

# A comparative study of machine learning algorithms for virtual learning environment performance prediction

Edi Ismanto<sup>1</sup>, Hadhrami Ab. Ghani<sup>2</sup>, Nurul Izrin Binti Md Saleh<sup>2</sup>

<sup>1</sup>Department of Informatics Education, Universitas Muhammadiyah Riau, Pekanbaru, Indonesia

<sup>2</sup>Department of Data Science, Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Kota Bharu, Malaysia

## Article Info

### Article history:

Received Jun 11, 2022

Revised Jan 10, 2023

Accepted Mar 10, 2023

### Keywords:

Classification techniques  
Exploratory data analysis  
Machine learning  
Performance evaluation  
Virtual learning environment

## ABSTRACT

Virtual learning environment is becoming an increasingly popular study option for students from diverse cultural and socioeconomic backgrounds around the world. Although this learning environment is quite adaptable, improving student performance is difficult due to the online-only learning method. Therefore, it is essential to investigate students' participation and performance in virtual learning in order to improve their performance. Using a publicly available Open University learning analytics dataset, this study examines a variety of machine learning-based prediction algorithms to determine the best method for predicting students' academic success, hence providing additional alternatives for enhancing their academic achievement. Support vector machine, random forest, Nave Bayes, logical regression, and decision trees are employed for the purpose of prediction using machine learning methods. It is noticed that the random forest and logistic regression approach predict student performance with the highest average accuracy values compared to the alternatives. In a number of instances, the support vector machine has been seen to outperform the other methods.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Edi Ismanto

Department of Informatics Education, Universitas Muhammadiyah Riau

Jalan Tuanku Tambusai, Kota Pekanbaru, Provinsi Riau - Indonesia

Email: edi.ismanto@umri.ac.id

## 1. INTRODUCTION

The emergence of more and more online and open courses in various countries has shown a significant impact on the progress of the development of distance education. This enables multiple course delivery formats to be developed by higher education institutions and organizations. Researchers are interested in analyzing student activity patterns in taking online courses because of the growing number of higher education institutions that use distance learning with the use of a virtual learning environment (VLE) such as massive online open courses (MOOCs). Large datasets from VLEs may be evaluated and utilized to provide recommendations for enhancing the online learning experience.

Learning analytics' major goal is to extract students' study patterns in order to improve the quality of learning and instruction. Learning analysis data not only provides instructional references for instructors to improve the quality of their teaching but also ideas for instructors to assist students in changing their learning practices. The success of students taking online courses is one of the indicators of the success of higher education.

The advancement of machine learning (ML), which is now widely utilized to tackle data problems, is causing ML research to expand [1]–[3]. In order to increase model performance, ML algorithm capabilities are constantly upgraded [4]–[6]. Several high-performance classification algorithms, such as support vector machine (SVM), random forest (RF), Nave Bayes (NB), logistic regression (LR), and decision trees (DT), have

been investigated in the literature and will be utilized to develop prediction models in the Open University learning analytics dataset to solve VLE data challenges (OULAD). OULAD is a database that contains information on courses, students, and their interactions with the virtual learning environment (VLE), which currently has 32,593 registered students [7]. Preprocessing will be performed on the OULAD dataset before it is separated into training and testing data. A confusion matrix will be used to measure and evaluate each model of the ML algorithm. Grid search and random search procedures are used to find model hyperparameters. The measurement model's outputs will be compared to assess how well it can classify VLE data.

## 2. METHOD

ML approaches have been applied in this research [8] to investigate student participation in various VLE activities. Both domain and category educational qualities were covered by the strategies chosen. Figure 1 depicts the essential stages involved in the current research. The dataset, preprocessing approaches, and machine learning algorithms used in this work are all described in this section.

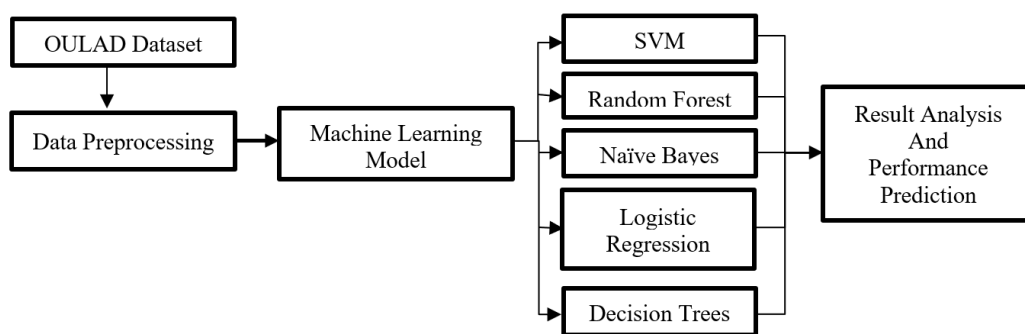


Figure 1. Research methodology

### 2.1. Exploratory data analysis (EDA)

The dataset OULAD comprises information on the courses, students, and their interactions with the virtual learning environment (VLE). Presentations are the term for class meetings. Course presentations begin in February and October, and are designated by the letters "B" and "J". In the OULAD data, each course is referred to as a module. The module includes two main disciplines: social sciences and science, technology, engineering and mathematics (STEM). The course module codes found in the OULAD data are the AAA module codes, BBB module codes, and GGG module codes for the social sciences and the CCC, DDD, EEE, and FFF Module Codes for the STEM sciences. The full class meeting module are identified by the course module codes BBB, DDD, and FFF. As displayed in Table 1. Therefore, the study is restricted to these three courses since they represent the highest number of students across the longest period of time.

Table 1. Course domain

Module	Domain	Students	Presentations
AAA	Social Sciences	748	2
BBB	Social Sciences	7,909	4
CCC	STEM	4,434	2
DDD	STEM	6,272	4
EEE	STEM	2,934	3
FFF	STEM	7,762	4
GGG	Social Sciences	2,534	3

There are no final exam results in the dataset, and there is no test date for BBB and FFF courses. Based on the following citation in the official documentation for the OULAD dataset, it is believed that the final test will take place on the last day of the course presentation. Only three variables were considered in order to avoid adding excessive complexity to the model. The first is the number of times per day that students click the VLE. The second factor is the number of assessments that students submit each day. The third is that the average value of student assessments is updated every day. To account for structural differences between

course presentations, such as varying durations and test structures, the three variables are rescaled to fall between 0 and 1 for each course presentation independently.

The first characteristic, as previously indicated, is the number of times students click something in the VLE every day. Because the Open University's learning is mainly done online, the daily number of clicks is seen to be a good proxy for student effort. The second feature is the number of assignments students submit on a given day. The number of assignments a student submits and the date on which they are completed may have an impact on their overall grade. The third and final feature is the student assignment average. Student assignment scores are updated on a daily basis and are computed as the average of all the tasks students have submitted so far in the course. In addition to the attributes stated, the dataset holds information about each student's final course outcomes. Students are classified as failing the course, passing it, passing it with an excellent predicate, or withdrawing from it. This research focuses on binary target variables with the labels "pass" and "fail" as the two class labels. Intuitively, a target variable with two labels will provide more accurate forecasts than one with three labels.

## 2.2. ML classification algorithm

The goal of machine learning is to create a learning tool that can acquire knowledge on its own, without the help of people. Predicting the class of the provided data is the process of classification. Several machine-learning techniques were used to build the classification model, and the results were compared. We surveyed a number of articles that used ML methods including support vector machine, random forest, Naive Bayes, logistic regression, and decision tree to predict student performance as seen in Table 2.

Table 2. Literature survey

No	Algorithm	References
1	Support Vector Machine (SVM)	[9]–[16]
2	Random Forest (RF)	[17]–[22]
3	Naive Bayes (NB)	[23]–[25]
4	Logistic Regression (LR)	[26], [27]
5	Decision Tree (DT)	[28]–[33]

### 2.2.1. Support vector machine (SVM)

SVM is a supervised technique for solving classification issues [11]. SVM may be applied to both linear and nonlinear models. The goal of this approach is to discover a hyperplane in an N-dimensional space that clusters the data points sympathetically [13]. There are numerous different hyperplanes that may be used to split data points into classes. According to studies [9] that have used SVM to model student academic performance, the greatest accuracy attained by SVM employing a linear kernel is 73.68%. Using a radial basis kernel, SVM was able to predict academic achievement with 90% accuracy [12]. When compared to ten category machine learning algorithms, linear support vector machines outperformed them 90% of the time in predicting student performance [10].

### 2.2.2. Random forest (RF)

RF is a basic yet adaptable ML algorithm that gives excellent results in the vast majority of cases and is widely used in a variety of problem statements due to its ease of usage and ability to conduct both classification and regression [22]. Forest is a collection of Decision Trees that are generally taught using the "bagging" approach, which combines several learning methods to improve accuracy. The results show that the improvised random forest outperforms the other classifiers, predicting students' academic success with a 93% accuracy rate [17]. The results show that RF can accurately classify numerous courses based on a variety of differentiating characteristics and predict student performance with a 96.88 accuracy [19]. The accuracy of the approach for predicting student achievement based on random forest classification was 81% [20].

### 2.2.3. Naive Bayes (NB)

A Naive Bayes classifier is a quantifiable clear classifier based on Bayes' theorem with a strong independent assumption [25]. Nave Bayes' probability distribution contributed challenging classification efficiency to statistics. Discreteness, on the other hand, has a high level of accuracy after its individual features of it have been gathered. Due to its ability to deal with large data sets and ease of implementation, this form of classifier has gained a lot of traction in recent years. The results suggest that Naive Bayes may be used to predict students' academic achievement early in the first year with a 72.46% accuracy [24]. The results demonstrate that by employing the Nave Bayes method, students' performance accuracy is above 90%, which is quite high [23].

### 2.2.4. Logistic regression (LR)

LR is one of the most well-known algorithms, LR is mostly used to tackle classification issues. The logistic or sigmoid function is the source of its name. It's a curve that transfers a real-valued number to a number between 0 and 1. Binomial, multinomial, and ordinal logistic regression are the three types of logistic regression. The experiment indicated an accuracy of 85.71% in predicting educationists' success using a regression model [26]. Early detection of at-risk students using iterative logistics regression yielded findings with a 98% accuracy rate [27].

### 2.2.5. Decision trees (DT)

Decision trees are widely used to gather information for the purpose of making decisions. The end result is a tree-like structure that decides on a condition at each level, with the preceding level's decision determining the next course of action [28]. Algorithms reduced error pruning tree (REPTree) is a decision tree method whose initial concept was from enhancing the C45 algorithm by extending the pruning phase, so that the rules formed are more minimal and useful. A 91.9% accuracy rate in predicting student achievement was achieved by the REPTree method used in the study [30]. All mining models exhibited a predictive probability of 70.25% to 95.1% in predicting student behaviors and performance in online learning experiments, demonstrating that all mining models were very reliable and accurate [31]. The J48 outperformed REPTree and random tree in forecasting students' success with a decent accuracy of 69.3% [32]. Based on the receiver operating characteristics (ROC) curve, the decision tree model predicted 85.31% accuracy for Pass, 79.41% accuracy for conditional, and 91.67% accuracy for Failed [33].

### 2.3. Performance evaluation of the model

A confusion matrix, classification accuracy (CA), precision, recall, and f-score (F1), were used to assess the model's performance [6]. The confusion matrix depicts the present state of the dataset as well as the number of accurate and wrong model predictions. Accuracy is an important and intuitive metric as it measures the proportion of correct predictions to the total number of predictions. Precision is the ratio of positive correct predictions compared to the overall positive predicted results. The recall is the ratio of true positive predictions compared to the total number of true positive data. F-score (F1) is a weighted comparison of the average precision and recall. Formally,

$$Accuracy = \frac{\sum True\ positives + \sum True\ negatives}{\sum Total\ population} \quad (1)$$

$$Recall = \frac{\sum True\ positives}{\sum True\ positives + \sum False\ negatives} \quad (2)$$

$$Precision = \frac{\sum True\ positives}{\sum False\ positives + \sum True\ positives} \quad (3)$$

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

Positives denote students who really fail, whereas negatives denote students who actually pass, while true denotes a valid prediction and false denotes an incorrect forecast. Table 3 illustrates the confusion matrix associated with various combinations of actual and predicted.

Table 3. The confusion matrix

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

## 3. RESULTS AND DISCUSSION

A model must be evaluated in order to create machine learning software that is effective. Evaluation criteria are used to explain model output, which frequently helps to distinguish between different model results. The study employed a confusion matrix along with the four evaluation metrics accuracy score, precision score, recall score, and F1-score, which are the most used evaluation metrics for machine learning models. This part will assess the results of the constructed model on a course-by-course basis, led by the two research questions that were defined,

- How well can the ML methods predict student performance in the virtual learning environment (VLE)?

- Which five ML algorithms (SVM, RF, NB, LR, and DT) have the best performance in predicting student performance in the virtual learning environment (VLE)?

To assess the SVM, RF, NB, LR, and DT algorithm models, the BBB, DDD, and FFF confluence module codes were chosen, due to the fact that the BBB, DDD, and FFF module codes have the most classroom sessions. The length of each course varies, so the course data is divided into ten deciles. A parameter grid search with cross-validation is used to determine parameter settings. The cross-entropy and Gini index are used to establish parameters in DT and RF algorithms [34]. In addition, the SVM algorithm uses a linear kernel or a radial basis function [35]. The final setting of these parameters is determined by their performance on the training set as measured by 5-fold cross-validation. Table 4 displays the data for training and testing the models.

Table 4. Total training and testing data

Training data	Total data	Testing data	Total data
BBB data course	3,858	BBB data course	1,520
DDD data course	2,830	DDD data course	1,149
FFF data course	3,818	FFF data course	1,503

### 3.1. Model performance

This work analyzes and presents the accuracy and recall values of the SVM, RF, NB, LR, and DT algorithms based on the outcomes of model training and testing. Table 5 shows the results of utilizing the SVM, RF, NB, LR, and DT algorithms to measure accuracy and recall for the BBB course. The LR algorithm model has almost the best average accuracy; however, the SVM algorithm outperforms the LR, RF, NB, and DT algorithms only in decile 1 testing namely 64%.

Table 5. Course BBB model performance

Decile	Accuracy					Recall				
	SVM	RF	NB	LR	DT	SVM	RF	NB	LR	DT
0	0.75	0.75	0.24	0.75	0.73	0.00	1.00	1.00	0.00	0.04
1	0.64	0.56	0.26	0.56	0.55	0.41	0.57	0.99	0.52	0.54
2	0.59	0.56	0.26	0.60	0.54	0.49	0.58	0.99	0.47	0.52
3	0.57	0.54	0.25	0.59	0.57	0.49	0.53	0.99	0.48	0.50
4	0.56	0.56	0.39	0.59	0.55	0.78	0.50	0.96	0.68	0.73
5	0.62	0.61	0.50	0.63	0.61	0.71	0.58	0.88	0.71	0.71
6	0.65	0.61	0.58	0.84	0.49	0.78	0.58	0.79	0.67	0.89
7	0.64	0.64	0.61	0.80	0.70	0.79	0.60	0.80	0.77	0.77
8	0.70	0.68	0.63	0.91	0.68	0.82	0.63	0.83	0.70	0.83
9	0.70	0.69	0.65	0.92	0.62	0.83	0.64	0.83	0.70	0.85
10	0.67	0.72	0.66	0.92	0.62	0.84	0.69	0.84	0.70	0.86

The accuracy and recall performance curves for the BBB course model are shown in Figure 2. For deciles 8, 9, and 10, Algorithm LR has a 91% and 92% accuracy rate, respectively. When compared to other algorithms, this BBB course's LR Algorithm's accuracy is far superior.

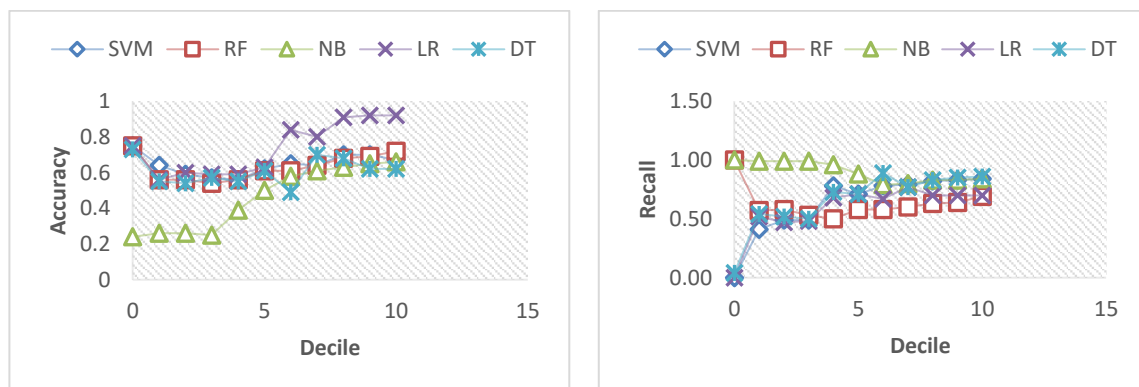


Figure 2. The performance curve of the BBB course model

The accuracy and recall for the DDD course were measured using the SVM, RF, NB, LR, and DT algorithms, as shown in Table 6. The RF algorithm model has almost the best average accuracy, while the LR algorithm delivers superior accuracy in decile 5, decile 6, and decile 7 tests, namely 82%, 83%, and 84%, respectively, when compared to SVM, RF, NB, and DT. The RF model almost has the best average value for recall value.

Table 6. Course DDD model performance

Decile	Accuracy					Recall				
	SVM	RF	NB	LR	DT	SVM	RF	NB	LR	DT
0	0.68	0.71	0.31	0.70	0.60	0.44	0.89	1.00	0.25	0.47
1	0.74	0.74	0.31	0.74	0.72	0.24	0.92	1.00	0.33	0.33
2	0.75	0.76	0.38	0.75	0.70	0.32	0.96	0.99	0.32	0.44
3	0.76	0.80	0.48	0.76	0.75	0.28	0.97	0.95	0.30	0.49
4	0.78	0.79	0.74	0.79	0.64	0.38	0.92	0.70	0.52	0.61
5	0.81	0.80	0.79	0.82	0.69	0.46	0.95	0.63	0.58	0.57
6	0.81	0.81	0.80	0.83	0.67	0.44	0.96	0.61	0.50	0.68
7	0.82	0.83	0.82	0.84	0.82	0.47	0.98	0.61	0.60	0.49
8	0.82	0.84	0.82	0.80	0.81	0.48	0.98	0.58	0.39	0.58
9	0.84	0.85	0.82	0.84	0.82	0.55	0.96	0.57	0.56	0.63
10	0.87	0.88	0.83	0.86	0.81	0.63	0.94	0.64	0.64	0.79

The accuracy and recall performance curves for the DDD course model are shown in Figure 3. The RF algorithm's accuracy and recall in this DDD course are, on the whole, far superior to those of the others. At the 10th decile, the RF algorithm yields a maximum accuracy of 88%.

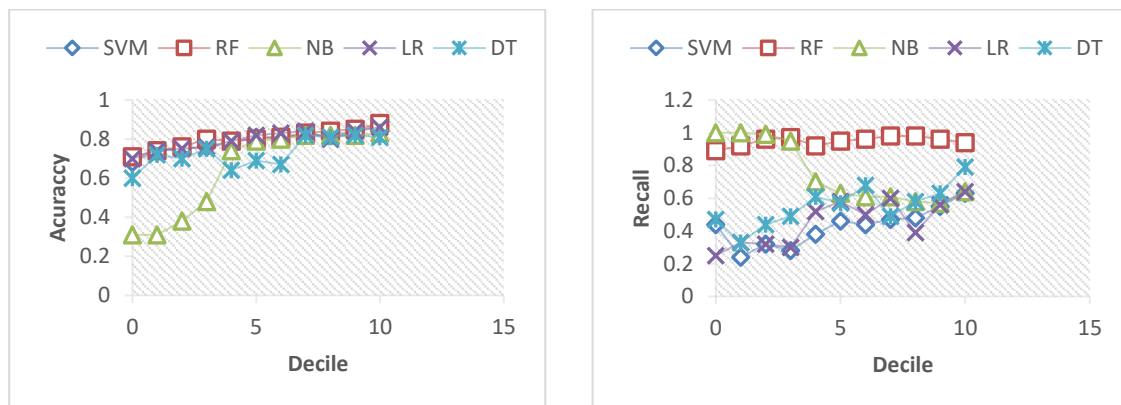


Figure 3. The performance curve of the DDD course model

The results of accuracy and recall measures for the FFF course utilizing the SVM, RF, NB, LR, and DT algorithms can be shown in Table 7. The RF algorithm model has almost the best average accuracy; in the 8-decile test, the SVM algorithm outperforms the LR, RF, NB, and DT algorithms, with an accuracy of 87%. SVM has a 75% accuracy in the 0 decile test, but its recall is still low when compared to RF and NB.

Table 7. Course FFF model performance

Decile	Accuracy					Recall				
	SVM	RF	NB	LR	DT	SVM	RF	NB	LR	DT
0	0.75	0.74	0.34	0.75	0.66	0.09	0.92	0.93	0.07	0.42
1	0.75	0.75	0.45	0.73	0.68	0.46	0.85	0.88	0.53	0.52
2	0.78	0.79	0.67	0.79	0.70	0.58	0.87	0.74	0.55	0.60
3	0.81	0.83	0.77	0.81	0.81	0.57	0.93	0.67	0.59	0.58
4	0.83	0.85	0.78	0.84	0.80	0.67	0.94	0.65	0.61	0.60
5	0.86	0.87	0.80	0.85	0.83	0.65	0.95	0.66	0.65	0.65
6	0.87	0.89	0.83	0.86	0.84	0.73	0.95	0.70	0.74	0.72
7	0.88	0.89	0.83	0.87	0.85	0.75	0.95	0.71	0.78	0.69
8	0.87	0.84	0.83	0.85	0.79	0.80	0.85	0.72	0.83	0.76
9	0.89	0.90	0.85	0.87	0.87	0.81	0.93	0.75	0.86	0.84
10	0.91	0.91	0.88	0.88	0.86	0.84	0.94	0.76	0.88	0.87

The accuracy and recall performance curves for the FFF course model are shown in Figure 4. The RF algorithm yields the maximum accuracy, with 9th and 10th decile values of 90% and 91% respectively. When compared to other algorithms, the RF algorithm produces accuracy and recall numbers that are nearly universally superior.

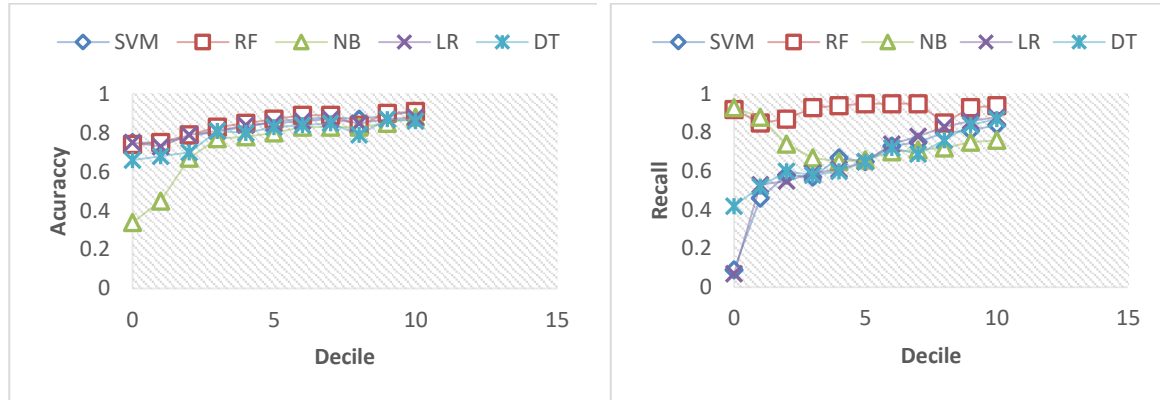


Figure 4. The performance curve of the FFF course model

Based on the comparison of the SVM, RF, NB, LR, and DT performance measures of the ML method in the BBB course, the LR model was found to be the best. For the 8th, 9th, and 10th deciles, the LR algorithm achieves accuracy rates of 91% and 92%. The results of the LR test between actual conditions, prediction results, and accuracy are shown in Table 8.

Table 8. Best ML algorithm model on BBB course

ML Algorithms	Course BBB	Predicted		Accuracy	
Logistic Regression (LR)	Decile 0	Fail	Pass		
	Actual	Fail	0	369	75%
		Pass	0	1152	
	Decile 1	Fail	Pass		
	Actual	Fail	192	177	56%
		Pass	482	670	
	Decile 2	Fail	Pass		
	Actual	Fail	172	197	60%
		Pass	403	749	
	Decile 3	Fail	Pass		
	Actual	Fail	178	191	59%
		Pass	431	721	
	Decile 4	Fail	Pass		
	Actual	Fail	251	118	59%
		Pass	496	656	
	Decile 5	Fail	Pass		
	Actual	Fail	262	107	63%
		Pass	443	709	
	Decile 6	Fail	Pass		
	Actual	Fail	247	122	84%
	Pass	121	1031		
Decile 7	Fail	Pass			
Actual	Fail	283	86	80%	
	Pass	207	945		
Decile 8	Fail	Pass			
Actual	Fail	260	109	91%	
	Pass	20	1132		
Decile 9	Fail	Pass			
Actual	Fail	260	109	92%	
	Pass	10	1142		
Decile 10	Fail	Pass			
Actual	Fail	260	109	92%	
	Pass	11	1141		

While the findings of the DDD course's comparison of the performance assessment of the ML method, namely SVM, RF, NB, LR, and DT, show that the RF model is the best. A maximum accuracy of 88% is attained by the RF algorithm at the 10th decile. The results of the RF test between actual conditions, prediction results, and accuracy are shown in Table 9.

The best model chosen in the FFF course is the RF model, based on the comparative findings of the performance assessment of the ML method, namely SVM, RF, NB, LR, and DT. The RF algorithm yields the highest accuracy, with 9th and 10th decile values of 90% and 91%, respectively. The results of the RF test between actual conditions, prediction results, and accuracy are shown in Table 10.

Table 9. Best ML algorithm model on DDD course

ML Algorithms	Course DDD	Predicted		Accuracy	
		Fail	Pass		
Random Forest (RF)	Decile 0	Actual	Fail 707	Pass 85	71%
		Pass	246	112	
	Decile 1	Actual	Fail 732	Pass 60	74%
		Pass	229	129	
	Decile 2	Actual	Fail 757	Pass 35	76%
		Pass	238	120	
	Decile 3	Actual	Fail 767	Pass 25	80%
		Pass	205	153	
	Decile 4	Actual	Fail 729	Pass 63	79%
		Pass	178	180	
	Decile 5	Actual	Fail 756	Pass 36	80%
		Pass	188	170	
	Decile 6	Actual	Fail 763	Pass 29	81%
		Pass	187	171	
	Decile 7	Actual	Fail 777	Pass 15	83%
		Pass	180	178	
	Decile 8	Actual	Fail 779	Pass 13	84%
		Pass	171	187	
	Decile 9	Actual	Fail	Pass	85%
		Pass			
	Decile 10	Actual	Fail	Pass	88%
		Pass			

Table 10. Best ML algorithm model on FFF course

ML Algorithms	Course FFF	Predicted		Accuracy	
		Fail	Pass		
Random Forest (RF)	Decile 0	Actual	Fail 1025	Pass 92	74%
		Pass	297	90	
	Decile 1	Actual	Fail 947	Pass 170	75%
		Pass	198	189	
	Decile 2	Actual	Fail 972	Pass 145	79%
		Pass	169	218	
	Decile 3	Actual	Fail 1036	Pass 81	83%
		Pass	162	225	
	Decile 4	Actual	Fail 1048	Pass 69	85%
		Pass	145	242	
	Decile 5	Actual	Fail 1066	Pass 51	87%
		Pass	135	252	
	Decile 6	Actual	Fail 1065	Pass 52	89%
		Pass	112	275	
	Decile 7	Actual	Fail 1064	Pass 53	89%
		Pass	105	282	
	Decile 8	Actual	Fail 948	Pass 169	84%
		Pass	66	321	
	Decile 9	Actual	Fail 1034	Pass 83	90%
		Pass	60	327	
	Decile 10	Actual	Fail 1052	Pass 65	91%
		Pass	57	330	

#### 4. CONCLUSION

In conclusion, the purpose of this article was to carry out and report the results of a comparative study on the performance of several machine learning-based algorithms in terms of making predictions. Training and testing operations have been carried out by making use of the publicly available OULAD dataset in order to evaluate and observe the performance of each of the prediction algorithms that are being considered. It has been discovered that certain algorithms that are based on machine learning perform better than the other algorithms in a number of different scenarios. According to the findings, the methods of logistic regression and random forest have the highest average accuracy achievement when compared to the other approaches. It has been discovered that the support vector machine method performs superiorly to the other options in certain specific instances.




#### REFERENCES

- [1] N. A. K. Rosili, N. H. Zakaria, R. Hassan, S. Kasim, F. Z. C. Rose, and T. Sutikno, "A systematic literature review of machine learning methods in predicting court decisions," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 4, pp. 1091–1102, 2021, doi: 10.11591/IJAI.V10.I4.PP1091-1102.
- [2] V. Siddesh Padala, K. Gandhi, and D. V. Pushpalatha, "Machine learning: The new language for applications," *IAES International Journal of Artificial Intelligence*, vol. 8, no. 4, pp. 411–421, 2019, doi: 10.11591/ijai.v8.i4.pp411-421.






- [3] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data-ai integration perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2021, doi: 10.1109/TKDE.2019.2946162.
- [4] Aized Amin Soofi and Arshad Awan, "Classification techniques in machine learning: Applications and issues," *Journal of Basic & Applied Sciences*, vol. 13, pp. 459–465, 2017, doi: 10.6000/1927-5129.2017.13.76.
- [5] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015, doi: 10.1126/science.aaa8415.
- [6] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," pp. 01–14, 2020, doi: 10.5121/csit.2020.100801.
- [7] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Data descriptor: Open University learning analytics dataset," *Scientific Data*, vol. 4, 2017, doi: 10.1038/sdata.2017.171.
- [8] J. W. & Sons, "The machine-learning approach," 2020.
- [9] N. A. M. S. Et. al., "Modeling Student's Academic Performance During Covid-19 Based on Classification in Support Vector Machine," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 5, pp. 1798–1804, 2021, doi: 10.17762/turcomat.v12i5.2190.
- [10] N. Naicker, T. Adelyi, and J. Wing, "Linear support vector machines for prediction of student performance in school-based education," *Mathematical Problems in Engineering*, vol. 2020, 2020, doi: 10.1155/2020/4761468.
- [11] E. A. Mahareek, A. S. Desuky, and H. A. El-Zhni, "Simulated annealing for svm parameters optimization in student's performance prediction," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1211–1219, 2021, doi: 10.11591/eei.v10i3.2855.
- [12] L. H. Alamri, R. S. Almuslim, M. S. Alotibi, D. K. Alkadi, I. Ullah Khan, and N. Aslam, "Predicting student academic performance using support vector machine and random forest," *ACM International Conference Proceeding Series*, vol. PartF16898, pp. 100–107, 2020, doi: 10.1145/3446590.3446607.
- [13] S. A. Oloruntuba and J. L. Akinode, "Student performance prediction using support vector machine," *International Journal of Engineering Sciences & Research Technology*, pp. 380–385, 2019, doi: 10.48175/ijarsct-3939.
- [14] I. Burman and S. Som, "Predicting students academic performance using support vector machine," *Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019*, pp. 756–759, 2019, doi: 10.1109/AICAI.2019.8701260.
- [15] H. Al-Shehri et al., "Student performance prediction using support vector machine and K-nearest neighbor," *Canadian Conference on Electrical and Computer Engineering*, 2017, doi: 10.1109/CCECE.2017.7946847.
- [16] S. Bhutto, I. F. Siddiqui, Q. A. Arain, and M. Anwar, "Predicting students' academic performance through supervised machine learning," *ICISCT 2020 - 2nd International Conference on Information Science and Communication Technology*, 2020, doi: 10.1109/ICISCT49550.2020.9080033.
- [17] S. K. Ghosh and F. Janan, "Prediction of student's performance using random forest classifier," *Proceedings of the International Conference on Industrial Engineering and Operations Management*, pp. 7089–7100, 2021.
- [18] M. Nachouki and M. A. Naaj, "Predicting student performance to improve academic advising using the random forest algorithm," *International Journal of Distance Education Technologies*, vol. 20, no. 1, 2022, doi: 10.4018/IJDET.296702.
- [19] S. Jayaprakash, S. Krishnan, and J. Jaiganesh, "Predicting students academic performance using an improved random forest classifier," *2020 International Conference on Emerging Smart Computing and Informatics, ESCI 2020*, pp. 238–243, 2020, doi: 10.1109/ESCI48226.2020.9167547.
- [20] A. Behr, M. Giese, H. D. Teguium K., and K. Theune, "Early prediction of university dropouts - A random forest approach," *Jahrbucher fur Nationalokonomie und Statistik*, vol. 240, no. 6, pp. 743–789, 2020, doi: 10.1515/jbnst-2019-0006.
- [21] P. Ajay, M. Pranati, M. Ajay, P. Reena, T. Balakrishna, and U. G. Scholar, "Prediction of student performance using random forest classification technique," *International Research Journal of Engineering and Technology*, pp. 405–408, 2020.
- [22] C. Beaulac and J. S. Rosenthal, "Predicting university students' academic success and major using random forests," *Research in Higher Education*, vol. 60, no. 7, pp. 1048–1064, 2019, doi: 10.1007/s11162-019-09546-y.
- [23] K. S. Thant, E. T. T. Thu, M. M. Khaing, K. L. Myint, and H. H. K. Tin, "Evaluation of student academic performance using Naïve Bayes classifier," *Advances in Computer and Communication*, vol. 1, no. 1, pp. 46–52, 2021, doi: 10.26855/acc.2020.12.005.
- [24] A. Baz, F. Alshareef, E. Alshareef, H. Alhakami, and T. Alsuibat, "Predicting students' academic performance using Naïve Bayes," *IJCSNS International Journal of Computer Science and Network Security*, vol. 20, no. 4, p. 182, 2020.
- [25] Havaluddin, N. Dengen, E. Budiman, M. Wati, and U. Hairah, "Student academic evaluation using Naïve Bayes classifier algorithm," *Proceedings - 2nd East Indonesia Conference on Computer and Information Technology: Internet of Things for Industry, EIConCIT 2018*, pp. 104–107, 2018, doi: 10.1109/EIConCIT.2018.8878626.
- [26] L. Zhang and H. Rangwala, "Early identification of at-risk students using iterative logistic regression," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10947 LNAI, pp. 613–626, 2018, doi: 10.1007/978-3-319-93843-1\_45.
- [27] S. Arora, M. Agarwal, and R. Kawatra, "Prediction of educationist's performance using regression model," *Proceedings of the 7th International Conference on Computing for Sustainable Global Development, INDIACom 2020*, pp. 88–93, 2020, doi: 10.23919/INDIACom49435.2020.9083708.
- [28] G. H. Wang, J. Zhang, and G. S. Fu, "Predicting student behaviors and performance in online learning using decision tree," *Proceedings - 2018 7th International Conference of Educational Innovation through Technology, EITT 2018*, pp. 214–219, 2018, doi: 10.1109/EITT.2018.00050.
- [29] M. Al Karim, M. Y. Ara, M. M. Masnad, M. Rasel, and D. Nandi, "Student performance classification and prediction in fully online environment using Decision tree," *AIUB Journal of Science and Engineering*, vol. 20, no. 3, pp. 70–76, 2021, doi: 10.53799/AJSE.V20I3.173.
- [30] M. A. Gotardo, "Using decision tree algorithm to predict student performance," *Indian Journal of Science and Technology*, vol. 12, no. 8, pp. 1–8, 2019, doi: 10.17485/ijst/2019/v12i5/140987.
- [31] A. Shanthini, G. Vinodhini, and R. M. Chandrasekaran, "Predicting students' academic performance in the University using meta decision tree classifiers," *Journal of Computer Science*, vol. 14, no. 5, pp. 654–662, 2018, doi: 10.3844/jcssp.2018.654.662.
- [32] E. Alyahyan and D. Dusteaor, "Decision trees for very early prediction of student's achievement," *2020 2nd International Conference on Computer and Information Sciences, ICCIS 2020*, 2020, doi: 10.1109/ICCIS49240.2020.9257646.
- [33] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood, and K. U. Sarker, "Student academic performance prediction by using decision tree algorithm," *2018 4th International Conference on Computer and Information Sciences: Revolutionising Digital Landscape for Sustainable Smart Society, ICCOINS 2018 - Proceedings*, 2018, doi: 10.1109/ICCOINS.2018.8510600.
- [34] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, 2019, doi: 10.1002/widm.1301.
- [35] J. Liu and E. Zio, "SVM hyperparameters tuning for recursive multi-step-ahead prediction," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3749–3763, 2017, doi: 10.1007/s00521-016-2272-1.




**BIOGRAPHIES OF AUTHORS**

**Edi Ismanto**    completed education bachelor's degree in the Informatics Engineering Department, State Islamic University of Sultan Syarif Kasim Riau. And master's degree in Master of Computer Science at Putra Indonesia University Padang. Now working as a lecturer in the Department of Informatics, University Muhammadiyah of Riau. With research interests in the field of Machine learning algorithms and AI. He can be contacted at email: [edi.ismanto@umri.ac.id](mailto:edi.ismanto@umri.ac.id)



**Hadhrami Bin Ab Ghani**    received his bachelor degree in electronics engineering from Multimedia University Malaysia (MMU) in 2002. In 2004, he completed his masters degree in Telecommunication Engineering at The University of Melbourne. He then pursued his Ph.D. at Imperial College London in intelligent network systems and completed his Ph.D in 2011. He can be contacted at email: [hadhrami.ag@umk.edu.my](mailto:hadhrami.ag@umk.edu.my).



**Nurul Izrin Binti MD Saleh**    obtained a bachelor's degree in information technology from Multimedia University Malaysia (MMU). He completed his master's degree in computer science at The University of Putra Malaysia. Then complete a Ph.D. at The University of Brunel London in the same field of study. She can be contacted at email: [nurulizrin.ms@umk.edu.my](mailto:nurulizrin.ms@umk.edu.my).