Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 203–212
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46659

Predictive Modeling of Student Dropout Using Academic Data and Machine Learning Techniques

Qurrotul Aini[1*], Elsy Rahajeng[2], Mufadha Tiohandra[3], Hamzah Aji Pratama[4], Jehad Hammad[5]

***Abstract*—This study's objective is to investigate the performance of a predictive model for students at risk of dropout (DO) by considering several internal criteria of an academic program. This research uses academic information from UIN Syarif Hidayatullah Jakarta and applies the C4.5, Naive Bayes Classification (NBC), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) to forecast which students might drop out. The data used consists of 714 student records from Department of Information Systems for the academic year 2010–2015 as training and 2018 as testing data. The research method refers to the SEMMA framework (Sample, Explore, Modify, Model, and Assess) to ensure systematic and accurate data processing. Meanwhile, the internal criteria used are the completed courses, the status of the internship report, and the final project proposal. According to the study's findings, the C4.5 and SVM algorithms get the best accuracy rates of 94.44%, while KNN and NBC come in second and third, respectively, with 93%. The results show that the C4.5 and SVM algorithms work well with academic data. This study provides a substantial contribution to the development of a prediction system for students at risk of dropping out, which can be integrated into data-based applications or dashboards. This solution is expected to help higher education institutions identify students who need further academic support. In addition, this research also opens up opportunities for the progress of more accurate forecasting models through the integration of additional variables such as behavioral or psychological data. With this data-driven approach, higher education institutions can enhance their efficiency in monitoring and preventing student dropouts, thereby supporting a vision of quality and sustainable education.**

***Index Terms*—C4.5, naive bayes classification, support vector machine, k-nearest neighbor, predictive analysis, student dropout.**

Received: 11 May 2025; Revised: 6 June 2025; Accepted: 19 June 2025.
*Corresponding Author

[1*]Qurrotul Aini, UIN Syarif Hidayatullah Jakarta Indonesia (e-mail: qurrotul.aini@uinjkt.ac.id).
[2]Elsy Rahajeng, UIN Syarif Hidayatullah Jakarta Indonesia (e-mail: qurrotul.aini@uinjkt.ac.id).
[3]Mufadha Tiohandra, UIN Syarif Hidayatullah Jakarta Indonesia (e-mail: mufadha.tiohandra21@mhs.uinjkt.ac.id).
[4]Hamzah Aji Pratama, PT. Pionirbeton Industri, Jakarta Timur, Indonesia (e-mail: hamzahaji1999@gmail.com).
[5]Jehad Hammad, Department of Computer Information Systems, Al-Quds Open University (QOU), PalestineAl-Quds Open University, Palestinian Territory, Occupied (e-mail: jhammad@qou.edu).

I. INTRODUCTION

UIN Syarif Hidayatullah (UIN Jakarta), established with a vision to provide quality higher education, has set strict academic standards to ensure optimal student graduation. One of the faculties at UIN Jakarta is the Faculty of Science and Technology. According to the 2022 academic guidelines, the regular Undergraduate Program (S1) at this university has a study load of between 144 and 150 credits planned for 8 semesters, or 4 years. However, some students may take longer to complete their study program and even experience dropout or termination of status as a student [1]. As a concern, policies from universities can minimize the high percentage of student dropouts (DO). Consequently, identifying students at risk of dropping out can reduce the actual dropout [2].

Based on data from the UIN Jakarta Academic Office, it can be seen that the number of students of the Department of Information Systems, Faculty of Science and Technology, who dropped out increased by 7.95% from the 2012–2014 student entry year. Table 1 documents the data and illustrates the trend in student graduation.

Table 1.
Student's Graduation Status

Enty Year	No. of Students	Grad.	DO	Resigned	DO (%)
2010	89	62	18	9	20,22
2011	116	90	21	5	18,10
2012	154	125	24	5	15,58
2013	138	98	28	12	20,29
2014	170	120	40	10	23,53
2015	88	82	6	0	6,82

Grad. = graduated

Nevertheless, early identification of students at risk of dropout remains a challenge. This research aims to explore and discuss the high dropout rate of students in higher education, particularly in the Dept. of Information Systems, which impacts academic efficiency and the utilization of institutional resources.The phenomenon of student dropout is a serious

challenge for higher education institutions. In addition to causing financial losses for students and institutions, dropout also reflects a failure in the learning and mentoring process. Student dropout risk is often identified reactively, after the problem worsens or the student quits. The lack of a proactive system to early identify students at risk of dropping out causes preventive interventions to be ineffective or delayed.This research aims to address this issue by developing a machine learning-based predictive model capable of proactively identifying students at risk of dropping out based on internal academic data. By identifying key factors from academic data that correlate with dropout, and hope to provide an effective tool for institutions to:

- Predicting students at risk of dropping out long before the problem becomes critical.
- Enabling early and targeted interventions, such as academic guidance, counseling, or curriculum adjustments, to prevent dropout.
- Enhancing the effectiveness of managing academic resources and boosting the graduation rates of students is the main objective.

Thus, the main substance of the problem is the need for an accurate and proactive prediction system to identify students at risk of dropping out to minimize its impact and improve student study success.

Information about student graduation rates is essential for the department to monitor and improve the quality of education. Measuring the accuracy of student graduation rates also reflects the effectiveness of the implemented curriculum. These variables serve as valuable inputs for making predictions. Prediction is an estimation action based on learning data patterns or historical data [3]. Prediction as an initial step provides an overview ofstudents at risk of attrition in the Department of Information Systems. The results of this prediction can be further interpreted as a prescriptive analysis aimed at providing strategic recommendations to prevent student dropouts.

Machine learning (ML) algorithms can perform predictions. One of the algorithms that can be used is the classification algorithm. In the process of classifying student academic data, many algorithm models can be used, including the C4.5 algorithm, NBC, SVM, and KNN [4], [5]. These machine learning algorithms can “learn” from given data, so they can adapt to emerging patterns and trends. This allows for improved prediction performance over time with more data or changing information.

This study examines students at risk of attrition by employing a machine learning algorithm on student data from the Department of Information Systems at UIN Jakarta. While the contribution of the research is to identify the criteria for dropout students, obtain a prediction model of four ML algorithms, and determine the performance of the application of the ML algorithm.

II. Literature Review

There are two types of predictive analysis models: classification models and regression models. Classification models predict if a value belongs to a specific class, whereas regression models predict a number. Some important techniques that are popularly used in developing predictive models are described below.

A. Decision Tree C4.5

C4.5 is the most frequently used and influential algorithm today. Compared to ID3, there are some improvements in C4.5. Firstly, C4.5 allow for pruning to prevent overfitting during construction. Second, C4.5 can handle incomplete data and discrete data [6]. Figure 1 illustrates a decision tree that is the outcome of the C4.5 algorithm..

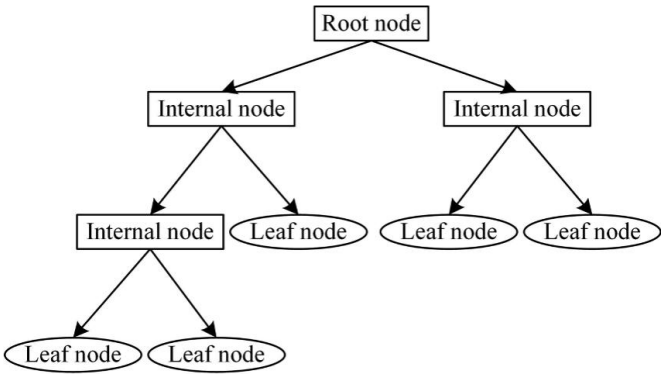

Fig. 1. Example of C4.5 decision tree [7].

The C4.5 algorithm is explained as follows: [8]

- Prepare the training dataset.
- Determine the root of the decision tree.
- Determine the gain. The selection of a root attribute is determined by the largest gain value among the current attributes. To calculate the gain value, apply (1):

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{N} \frac{|S_i|}{|s|} \times Entropy(S_i) \quad (1)$$

- The procedure for each branch is fulfilled. Repeat step 2. On the other hand, to determine the value of entropy using (2):

$$Entropy(S) = \sum_{i=1}^{N} -\pi \times log2\pi \quad (2)$$

- The decision tree participation process finishes when every node *N* branch receives the same class..

B. Naive Bayes Classification

Naive Bayes classification is a straightforward probability classification technique based on Bayes' theorem. Bayes' theorem is paired with "naive," which implies that each attribute or variable is independent [9], [10].The advantages of Naive Bayes are simplicity, speed, and high accuracy.NBC is a classification technique in data mining that employs fundamental probability and statistics [11]. Naive Bayes is a classification method that forecasts probability derived on historical data [12]. The disadvantage of Naive Bayes is that the attributes of data are independent and have no relationship with each other[13].Mathematically, the Naive Bayes formula is as follows:

$$P(C|X) = \frac{P(X_1|C)PP(X_2|C)....PP(X_n|C)P(C)}{P(X_1)P(X_2)....P(X_n)} \quad (3)$$

Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 203–212
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46659

where X is data with an unknown class and C is a class in the dataset. Posterior is the probability of data X in class C or $P(C|X)$, which is the result of multiplying likelihood and prior divided by evidence. Likelihood is the probability of the attribute of data X in class C or $(X_n|C)$, prior is the probability of class C from the total dataset or (C), and evidence is the probability of the attribute of data X from the entire total dataset or $P(x)$.

C. K-Nearest Neighbors

KNN has various advantages, including tolerance to training data with a lot of noise and a large quantity. In addition, it has simple steps, which affect its computational speed. The weakness of KNN is that it requires determining the number of nearest neighbors of the target data, symbolized by the parameter K value. The training data based on distance calculations is less accurate because it requires selecting, trying, and determining the type of distance used and which attributes to use to obtain the best distance calculation results. Additionally, it has high computational costs because it requires distance calculations for each query instance across the entire training data set [14].

The value of K can be found by taking the square root of the available sample or usually starting from 3 and continuing with odd numbers to avoid ties in voting [15]. Here is the KNN algorithm:

- Determine the value of K.
- Calculate the distance between test and training data.
- Find training data (k) that is closest to test data.
- Determine the frequency of each class.
- Assign test data to the class with the highest frequency.
- Repeat steps (b) to (e).

To measure the proximity or distance between two objects (new data to old data/neighbors), you can use the Manhattan distance and Euclidean distance formulas [16]. Manhattan distance, or city block distance, is a measure of distance inspired by the distance between two locations in a city.

$$D(a,b) = \sum_{i=1}^{N} |X_{an} - X_{bn}| \quad (4)$$

where $D(a,b)$ is the distance between two points, n is the number of attributes/dimensions, X_{an} is the value of the n-th attribute on object a, while X_{bn} is the value of the n-th attribute on object b.

D. Support Vector Machine

The principle of SVM is to identify the optimal hyperplane that acts as a delimiter between two data classes. Figure 2 shows two classes, +1 representing agree and -1 representing disagree, with 2 input attributes, A1 and A2.

The two-dimensional data can be linearly split, as a straight line can be drawn to separate all the +1 class tuples from all the -1 class tuples, the black line in the middle is the hyperplane (Fig. 3) [17], [18], [19]. The hyperplane is a line formed as the best separator between the two classes, -1 and +1, and can be maximized by measuring the margin and finding its maximum point. Margin (the dashed black line) is the distance between the hyperplane and the nearest tuple from each class (Fig. 3) [20]. The closest tuple is called the support vector.

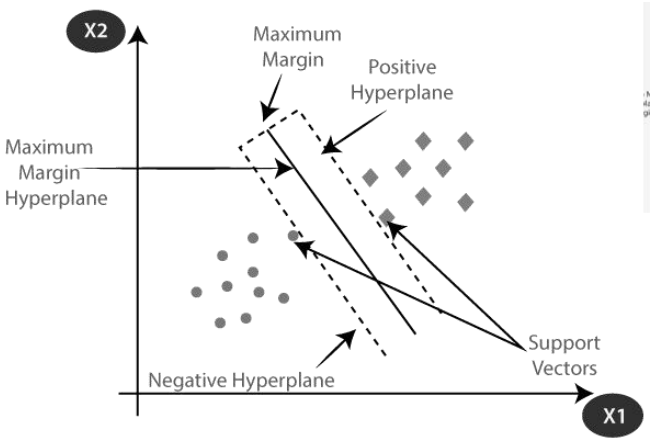

Fig. 2. SVM that separates two data with a hyperplane.

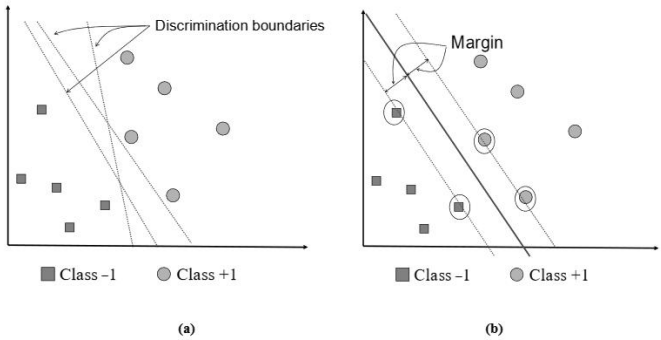

Fig. 3. Two hyperplanes in one data.

E. Performance Metrics of Machine Learning Models

The performance metrics commonly used in predictions are:

1) The confusion matrix

It is a table composed of the scores of correct and incorrect test data (Table 2.1). A confusion matrix is a metric for evaluating performance in machine learning classification where the results can be in the form of two or more classes. This table consists of four different combinations of predicted values and actual values. This matrix contains four terms that denote the outcomes of the categorization process, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [21].

Tabel 2.
Confusion Matrix

Confusion Matrix		Nilai Aktual	
		TRUE	FALSE
Predicted value	TRUE	TP (True Positive Correct result)	FP (False Positive Unexpected result)
	FALSE	FN (False Negative Missing result)	TN (True Negative Correct absence of result)

Meanwhile, a false negative is defined as a condition where the predicted output label score is incorrect but the actual score is correct [22]. Mathematically, the calculation of accuracy, precision, and recall scores are as follows: (5)–(8):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (5)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (6)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (7)$$

$$\text{F}_1\ \text{Score} = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad (8)$$

2) ROC and AUC

A receiver-operating characteristic curve (ROC) is a visual representation of a model's performance at all threshold levels. ROC is a kind of performance measurement tool for classification problems in determining the threshold of a model. The ROC curve is created by calculating the true positive rate (TPR) and false positive rate (FPR) at each possible threshold (in practice, at selected intervals) and then plotting the TPR against the FPR. A perfect model, which at a certain threshold has a TPR of 1.0 and an FPR of 0.0, can be represented by a point at (0, 1) if all other thresholds are ignored, as shown in Fig. 4 [23]. AUC denotes the area under the receiver operating characteristic (ROC) curve. A higher AUC score often indicates the superior performance of a classifier for the specified task.

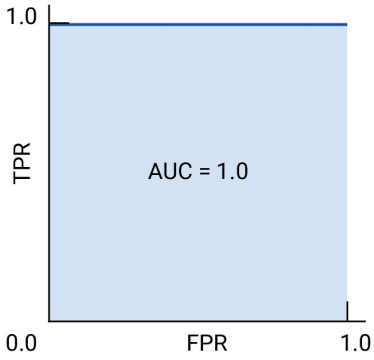

Fig. 4. ROC and AUC of the perfect hypothesis model.

F. Related Work

In recent years, various studies have focused on predicting student dropout rates using machine learning techniques. Each offers valuable insights into the factors contributing to student retention or attrition while employing different methodologies and datasets.

The C4.5 algorithm has been widely used for classification tasks related to educational outcomes. As demonstrated by [24], [25], [26], this algorithm successfully identified key predictors of student success, achieving an accuracy rate of 90–96.45%. Nevertheless, the study limited its analysis to demographic factors, thus overlooking potentially influential variables like academic stressors and support systems. This gap highlights the need for more holistic models that take into account a wider array of factors affecting student retention. However, there is a particular case where [27] applied C4.5 to the prediction of graduating students, and the results indicated that the model is less than optimal in the overall prediction.

Research by [28] and [29] applied the Naive Bayes classifier to examine dropout risks, reporting significant findings related to the socio-economic background of students. Accuracy in both studies reached 93–96%. While the results underscore the importance of demographic data, they also suggest a limited view of dropout causation. The reliance on probabilistic assumptions inherent in Naive Bayes may have constrained the depth of analysis regarding behavioral and situational influences that could affect student persistence.

The KNN approach was explored by [29] and [30], where it was used to predict dropout rates based on historical academic data. This study noted an accuracy of 96–97.68%, but it raised concerns about scalability and the model's sensitivity to noise in the datasets. However, there are cases where the accuracy only reaches 83% due to the limited dataset [31]. Additionally, KNN's dependency on the local neighborhood of data points may not adequately capture the broader trends influencing student retention.

A significant body of work has employed SVM due to its capacity to handle high-dimensional data effectively. For instance, [32] and [33] utilized SVM to analyze academic performance indicators, achieving an accuracy of 83–92%. However, this research primarily centered on quantitative metrics and did not delve into qualitative aspects such as student engagement and motivation, which are critical in understanding dropout patterns.

While these studies collectively contribute to understanding dropout prediction, they often focus on limited factors or use specific algorithms that may not represent an optimal choice for all contexts. Notably, the integration of behavioral data and additional demographic variables is frequently missing, which could enhance prediction accuracy and enable more effective interventions.

This study aims to bridge these gaps by employing multiple machine learning algorithms, including C4.5, SVM, Naive Bayes, and KNN, alongside a more comprehensive dataset that incorporates both academic performance and behavioral indicators. This approach not only seeks to improve predictive accuracy but also aspires to provide actionable insights for educational institutions striving to minimize dropout rates.

III. Research Method

This study adopts a quantitative approach with a predictive research design based on machine learning techniques. The primary goal is to develop a predictive model capable of identifying patterns and relationships among variables using historical data. Rather than relying solely on descriptive analysis, this approach emphasizes both exploratory and inferential insights to uncover key factors that influence the observed outcomes.Machine learning is employed as the core analytical tool due to its ability to process large-scale, complex datasets that traditional statistical methods may not adequately handle. Various machine learning algorithms—such as decision trees, support vector machines, NBC, and KNN—are implemented and compared to determine the most effective model. The performance of each algorithm is evaluated using

Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 203–212
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46659

standard metrics such as accuracy, precision, recall, and the area under the ROC curve (AUC-ROC). To structure the modeling process, this study utilizes the SEMMA framework (Sample, Explore, Modify, Model, and Assess), developed by the SAS Institute. This framework provides a systematic and iterative approach for developing predictive models, beginning with data sampling, followed by data exploration and preparation, modify stage, model construction using machine learning algorithms, and ending with performance evaluation and validation. The flow of this research is more comprehensively visualized in Fig. 5.

A. Sample

Sample is the first part of SEMMA, which collects significant data and information. The source of the dataset for this research was obtained from the Center for Information Technology and Database (Pustipanda) UIN Jakarta via email to the relevant parties. This dataset is presented in the Microsoft Excel Open XML Spreadsheet (.xlsx) format, which contains information such as student names. Student ID, student status, semester, courses, and course grades for students admitted in 2010 to 2015 (714 records). Whereas for students who entered in 2018, the data includes the student's name, student ID, semester, courses, and student grades (150 records). Additionally, interviews were conducted with the Secretary of Department of Information Systems to obtain the criteria for student dropouts, which include courses not completed by the 8th semester, not finishing the internship report, and not submitting a final project proposal. From the students who entered in 2010, 2011, 2012, 2013, 2014, and 2015, there were 136 students recorded as having dropped out and 578 students who successfully completed their studies.

B. Explore

The preceding data collection (sample) included six columns of variables for analysis. Furthermore, this study focuses on three major elements that have the potential to drive students to DO: the number of incomplete courses, the status of internship reports, and the production of thesis proposals. Table 3 shows the outcomes of the student data exploration for the years 2010–2015.

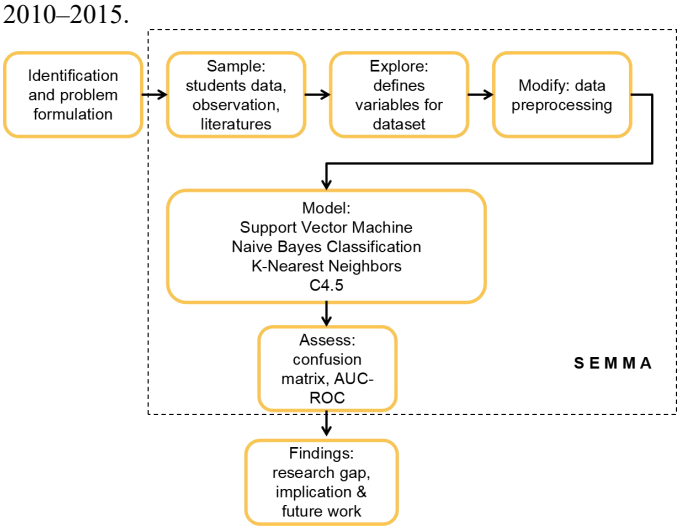

Fig. 5. Flow of research on predicting students at risk of dropping out.

According to this data, 136 students have dropped out, while 578 have successfully completed their education.

C. Modify

The modify stage is a data preprocessing stage that includes several important steps, one of which is data cleansing to ensure better data quality [34]. In the modify stage, this research performs a cleansing process by deleting data on students in the 2010, 2011, 2012, 2013, 2014, and 2015 entry years who have the status resigned. This is done to avoid affecting the analysis results, as the research focuses on students who graduated or dropped out. After the cleansing process is complete, the next step is to convert the data into numeric form. The goal of this conversion is to transform categorical attributes into a format that classification algorithms can process. In this study, we carry out the modification process using the Python programming language and the Pandas library for data manipulation.

D. Model

Next stage, it is applied four different machine learning algorithms: C4.5, Naive Bayes Classification (NBC), Support

Table 3.
Explore Student's Data

Criteria		2010	2011	2012	2013	2014	2015
	Number of Students (grad. & DO)	80	111	149	126	160	88
Status	Graduated	62	90	125	98	120	82
	DO	18	21	24	28	40	6
All subjects completed by 8th semester?	Yes	67	94	127	87	101	61
	No	13	17	22	39	59	27
Completed internship report?	Yes	66	91	124	79	104	42
	No	14	20	25	47	56	46

Vector Machine (SVM), and K-Nearest Neighbor (KNN). The selection of these algorithms is based on previous literature studies, their proven effectiveness in classification tasks, and their different characteristics in handling data.

1) *Splitting data,* which begins with a dataset that has undergone the stages of sampling, exploring, and modifying (i.e., clean and ready-to-process data), is divided into training data and testing data. Authors conducted tests with three different splitting ratios: 90:10, 80:20, and 70:30 (train:test) to evaluate the stability and generalization of the model under varying data conditions.
2) Cross-validation, this process to minimize bias and improve the reliability of model evaluation, with applying the *k*-fold cross-validation technique on the training data. With $k = 10$, the training data is divided into 10 equally sized subsets. The iteration is performed 10 times, where each time one subset is used as validation data and the other nine subsets as training data. This process ensures that every part of the dataset is used for both training and testing, providing a more robust estimation of the model's performance and reducing the risk of overfitting.
3) Model training, where each algorithm (C4.5, NBC, SVM, KNN) is trained using the training data from each fold of cross-validation. The default parameters of each algorithm are used unless significant optimizations are found through preliminary testing (for example, adjusting the K value in KNN).
4) Prediction, which fter training, authors used the model to predict class labels (pass or dropout) on separate test data.

E. Assess

In the assess stage, the algorithm's results are tested and evaluated. Testing is done using two metrics, namely the confusion matrix and the receiver operating characteristic (ROC) and area under the curve (AUC), or commonly known as the ROC-AUC score. In the confusion matrix, testing the accuracy of the search results will be evaluated by the value of recall, precision, accuracy, and F1-score. Additionally, the scikit-learn library in Python performs cross-validation using stratified validation techniques. This process is carried out using 10-fold as the default of split validation, which aims to produce maximum accuracy.

IV. Result

A. Cross-Validation

This study employs 10-fold validation to objectively assess model performance by partitioning the training data into folds. Table 4 shows the performance of stratified *k*-fold cross-validation.

Table 4.
Performance of Stratified *K*-Fold Cross-Validation

K-fold	C4.5	NBC	SVM	KNN
1	0.93	0.93	0.93	0.93
2	0.96	0.99	0.96	0.99
3	0.96	0.96	0.96	0.96
4	0.92	0.94	0.92	0.94
5	0.97	0.97	0.97	0.97
6	0.89	0.89	0.90	0.89
7	0.85	0.83	0.86	0.83
8	0.96	0.90	0.96	0.90
9	0.89	0.87	0.89	0.87
10	0.85	0.72	0.86	0.72
Mean	0.918	0.90	0.921	0.90

The SVM has the best overall performance. It is a suitable model for predicting student graduation or dropout. C4.5 is also very good and almost on par with SVM; it is stable and accurate, and it has the advantage of result interpretation (due to the easily readable decision tree format). Meanwhile, NBC and KNN are less consistent, experiencing a significant drop in accuracy on certain folds (as low as 0.72 on the 10th fold), indicating a lack of stability. It can be highlighted that model stability is important in the final selection. Models that have high accuracy and stability across folds (such as SVM and C4.5) are more reliable for real-world predictions, as their performance is less affected by data variations.

B. Prediction of Students at Risk of DO

Table 5 presents the percentage outcomes of estimating the graduated students and the risk of DO in the Department of Information Systems, using four algorithms on a dataset of 150 records from the class of 2018, while the training data consisted of 714 records.

Tabel 5.
Percentage of Graduates and Students at Risk of DO

Ratio	C4.5		NBC		SVM		KNN	
(%)	DO	Grad.	DO	Grad.	DO	Grad.	DO	Grad.
90:10	20.7	79.3	33.3	66.7	20.7	79.3	32.7	67.3
80:20	21.3	78.7	33.3	66.7	20.7	79.3	21.3	78.7
70:30	21.3	78.7	33.3	66.7	20.7	79.3	58.7	41.3

As seen in Table 5, SVM demonstrates the most stable and optimal performance in classifying students who are likely to graduate versus those at risk of dropping out. The results indicate that SVM possesses strong generalization capabilities across different training and testing datasets. The performance of C4.5 is nearly equivalent to that of SVM, exhibiting a slight decrease in accuracy at the 80:20 and 70:30 ratios. These results demonstrate that C4.5 excels at classification, although it appears to be somewhat more sensitive to variations in data ratio. NBC shows lower performance than other algorithms. This decrease may be due to the assumption of independence among features in NBC, which does not align with the actual characteristics of the data, leading to an overestimation of the DO prediction. KNN shows very high variability, particularly at the 70:30 ratio, where the DO jumps dramatically to 58.7%. This evidence shows that KNN is very sensitive to changes in the training-testing data and can experience overfitting or underfitting if not properly managed. The detailed results of the number of predictions of class 2018 students using the four models are shown in Table 6.

Tabel 6.
Predicted Number of Students Graduating and at Risk of DO Class of 2018

Ratio	C4.5		NBC		SVM		KNN	
(%)	DO	Grad.	DO	Grad.	DO	Grad.	DO	Grad.
90:10	31	119	50	100	31	119	49	101
80:20	32	118	50	100	31	119	32	118
70:30	32	118	50	100	31	119	88	62

Table 6 shows that the NBC model is not adaptive to the ratio data, as it is unresponsive or too rigid to adjust to changes in the data distribution. In contrast, KNN is very sensitive to the data ratio, meaning that its performance can be unstable due to

Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 203–212
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46659

the amount of training data. However, C4.5 and SVM provide stable and consistent predictions, which are suitable for use in variable data situations. In the case of KNN, the ratio of 70:30 may indicate overfitting or improper parameter selection (e.g., *k* value) for that ratio.

C. *Confusion Matrix*

The confusion matrix offers an overview of accurate and erroneous predictions while also delineating the types of errors, which is a critical element in an educational situation. The confusion matrix comprises four primary components:

- True positive (TP) indicates that the data categorized as a graduate accurately represents a graduate.
- False positive (FP) refers to data incorrectly labeled as graduated when, in reality, the individual did not graduate, resulting in a critical error that may prompt unwarranted intervention.
- False negative (FN) indicates that data categorized as DO is, in fact, graduating, representing a significant risk due to the failure to implement early intervention.
- True negative (TN) refers to instances where the data categorized as DO are indeed dropouts.

Figure 6 presents the confusion matrix of the C4.5 model, which uses a training-to-testing data ratio of 70:30. This figure implies that the model is balanced in handling both classes, where the number of FPs and FNs are equal (8 each), meaning that the model is not biased towards one particular class (does not overpredict or underpredict DOs). The model is suitable for initial implementation, but it requires enhancements for effective early intervention. Threshold adjustments or additional features are needed to increase sensitivity to dropout risk. The findings of the four models' confusion matrices are displayed in Table 7.

Table 7 implies that although the SVM model generally shows the highest performance in terms of accuracy, precision, recall, and F1 score, there is an indication that its effectiveness highly depends on the proportion of sufficient training data. The decrease in the F1 score from 0.910 (90:10 ratio) to 0.870 (70:30 ratio) indicates that SVM tends to be less stable on smaller datasets. On the other hand, NBC model performs quite stably and has the highest recall value across all data ratios, reflecting its ability to detect students who are truly at risk of dropping out.

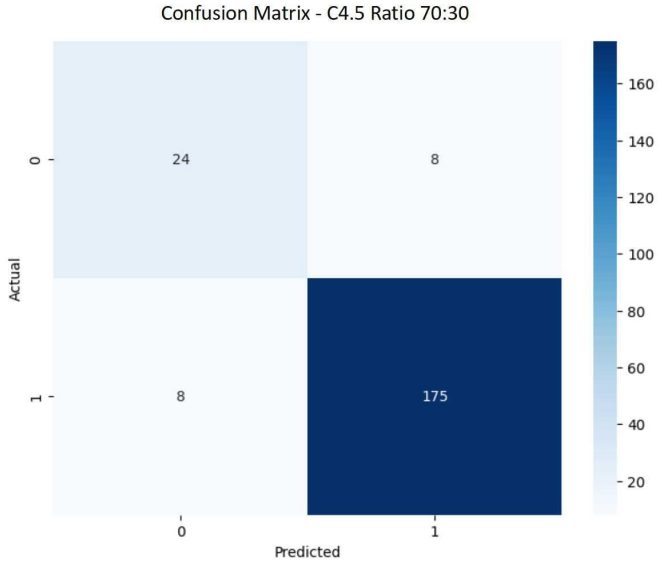

Fig. 6. Confusion matrix of C4.5

This method can be interpreted as a defensive approach useful for institutions to avoid negligence in providing early interventions, even though it sacrifices the level of prediction accuracy (precision). Model C4.5, although it does not always record the highest metric values, offers the advantage of high interpretability through an easily understandable decision tree structure. This advantage becomes important if the prediction results will be used as the basis for administrative or academic policies that need to be explained transparently. On the other hand, KNN shows a strong dependence on the size and distribution of the training data. The sharp decline in the F1 score to 0.808 at the 70:30 ratio indicates that KNN is less suitable for real-world scenarios with limited or uneven data. Interestingly, all models achieved an F1 score above 0.80 at all ratios, which implies that the features used in the models (such as course completion, internship reports, and final project status) are quite representative in distinguishing between graduating students and those at risk of DO. Therefore, the selection of the ideal model should not only consider the highest numerical performance (accuracy) but also the stability across data conditions, sensitivity to DO cases, and the model's transparency level in supporting decision-making at the program study level.

Table 7.
Confusion Matrices of Models

		Accuracy (%)			Precision (%)			Recall (%)			F1 Score		
	Ratio	90:10	80:20	70:30	90:10	80:20	70:30	90:10	80:20	70:30	90:10	80:20	70:30
Model	C4.5	94.44	92.30	92.55	91	86	85	91	84	85	0.874	0.850	0.850
	NBC	90.27	93	92.55	82	85	84	91	92	92	0.863	0.884	0.878
	SVM	94.44	94.40	93	91	92	88	91	86	86	0.910	0.889	0.870
	KNN	91.66	93	88.37	84	88	77	92	85	85	0.878	0.865	0.808

D. ROC and AUC

This study's model performance evaluation also includes ROC and AUC values. The ROC curve shows how well a model can tell the difference between two groups (students who graduate and those who drop out) by looking at the trade-off at different levels between the true positive rate (TPR) and the false positive rate (FPR). The AUC number represents the area beneath the ROC curve; thus, a value closer to 1 indicates superior classification accuracy of the model. The ROC curve showing the AUC value of the C4.5 model with a 90:10 ratio is illustrated in Fig. 7. The complete AUC value is shown in Table 8 and Fig. 8 (ratio 70:30).

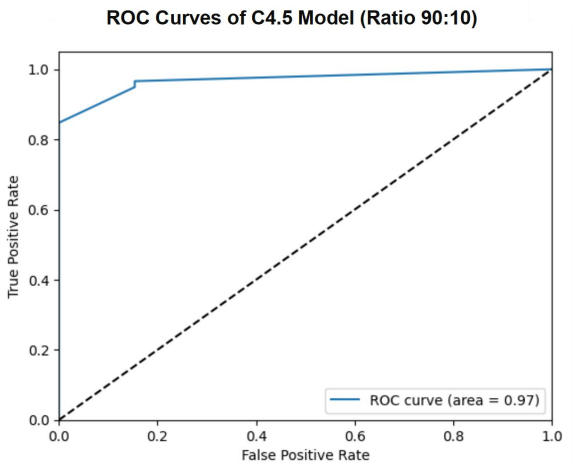

Fig. 7. ROC curve and AUC of C4.5

The results in Table 8 show that the NBC and KNN models perform reliably well, with high AUC values for all training/testing data ratios. NBC achieved an AUC of 0.97 (90:10), 0.98 (80:20), and 0.96 (70:30), while KNN attained 0.97, 0.97, and 0.95, respectively. This ratio signifies that both models possess robust and consistent capacities to differentiate between drop-out students and graduates, notwithstanding fluctuations in the training data proportions.The C4.5 model has commendable performance with an AUC of 0.97 at a 90:10 ratio, which marginally diminishes to 0.93 and 0.92 for 80:20 and 70:30 ratios, respectively. This drop can be attributed to the fact that decision tree-based models typically demand ample training data to construct an appropriate tree structure. Simultaneously, SVM exhibits inconsistency, with commendable AUC values of 0.91 at the 90:10 ratio and 0.95 at the 70:30 ratio, yet experiencing a significant decline to 0.73 at the 80:20 ratio. This unusual behavior shows that SVM performance is very sensitive to changes in data distribution or hyperparameter settings, requiring careful attention during validation and parameter tuning.

Table 8.
AUC Scores of Models

Ratio	AUC		
	90:10	80:20	70:30
C4.5	0.97	0.93	0.92
NBC	0.97	0.98	0.96
SVM	0.91	0.73	0.95
KNN	0.97	0.97	0.95

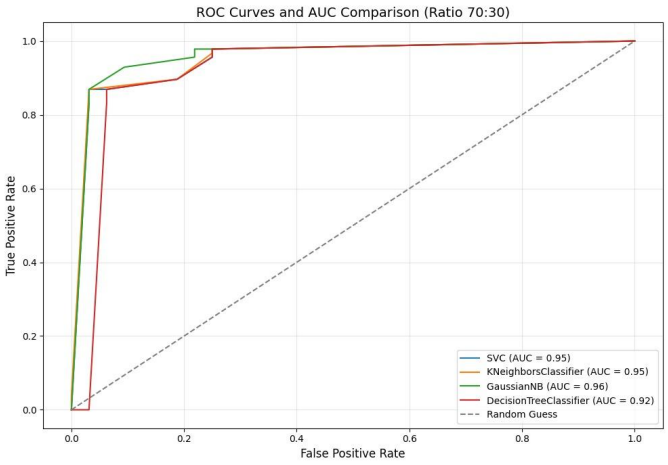

Fig. 8. ROC curve and AUC all models

E. Interpretation

The C4.5 and SVM algorithms showed the highest accuracy rate of 94.44%, indicating their effectiveness in identifying patterns in academic data related to student dropout tendencies. The KNN and NBC algorithms also showed high accuracy at 93%, indicating strong predictive capabilities. The small percentage difference between the four algorithms (1.44% between the highest and lowest) suggests that all tested algorithms have good relevance and application for this dropout prediction problem.

Internal academic criteria such as the number of completed courses, the progress of internship reports, and the status of thesis proposals are excellent and predictive indicators of student dropout risk. The high accuracy of the model, particularly C4.5 and SVM, confirms that changes or delays in any of these criteria can serve as an early signal for institutions to intervene. For example, students who show a decline in course completion, significant delays in internship reports, or lack of progress in thesis proposals can be accurately identified as a high-risk group. This demonstrates the model's performance and highlights the importance of proactively monitoring these academic indicators. Furthermore, these findings suggest that educational institutions, can utilize this data-driven predictive model for the following purposes:

1) *Early intervention* with identifying students at risk of dropping out early to provide guidance, academic support, or counseling.
2) *Resource allocation* with support resources more efficiently to students who need them the most.
3) *Improving student retention* with contributing to strategies for enhancing student retention by taking preventive actions based on accurate predictions. This interpretation provides a strong foundation for the practical implementation of this model in the university's academic management system.

F. Discussion

This study aims to fill an important gap in the literature regarding the prediction of student dropout (DO) risk, which

Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 203–212
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46659

has not been fully addressed in previous research. Most earlier studies, like those by [24] and [25], have shown that the C4.5 algorithm is good at predicting dropout risk, but they mainly looked at demographic factors and didn't fully consider important academic details. Other studies that used the NBC or KNN had high accuracy, but they struggled with data noise and weren't very good at dealing with uneven data distributions. Previous studies have primarily focused on a single algorithm type, neglecting the performance of predictive models in the same context and overlooking issues of interpretability and model performance stability across validation scenarios. This research addresses that issue by providing a detailed study of four machine learning algorithms—C4.5, Naive Bayes, SVM, and KNN [28], [29], and [30] —carefully tested using the SEMMA framework and cross-validation. This research not only assesses accuracy, but also takes into account precision, recall, F1-score, and AUC at different training and testing data ratios. In addition, this research emphasizes internal academic variables that are operational and can be directly actionable, such as course completion status, internship reports, and thesis proposals. The finding that SVM and C4.5 show the most stable and accurate performance indicates the importance of using a data-driven approach to support an effective early warning system in the context of higher education. Thus, the main contribution of this study lies not only in the technical performance of the predictive model but also in mapping the methodological and conceptual gaps present in the literature, as well as providing a stronger foundation for evidence-based academic decision-making.

V. Conclusion

This study investigates how well prediction models work for students who are at risk of dropout (DO) by taking into account a number of internal program characteristics at UIN Syarif Hidayatullah Jakarta. Predictive models can be developed and tested by utilizing machine learning techniques such as C4.5, NBC, SVM, and KNN in conjunction with academic data from 714 student records in the Information Systems Department. According to this study, the best methods for identifying which students could drop out based on their academic achievement were the C4.5 and SVM. The best accuracy level of 94.44% was attained by these two methods. At 93% apiece, the KNN and NBC algorithms likewise showed outstanding accuracy. The primary conclusion drawn from this study is that C4.5 or SVM machine learning algorithms perform incredibly well at predicting which students may drop out when combined with internal academic data (such as finished courses, internship report status, and thesis proposal). These results confirm that C4.5 and SVM algorithms are the best options for this purpose and immediately answer our study topic about predictive model performance. It is implied that educational institutions possess strong instruments to detect kids in need of intervention at an early stage. Universities can enhance student retention rates and academic success by taking preemptive steps like offering academic advice or psychological support, made possible by the strong predictive ability.

This research only uses internal academic criteria (completed courses, internship report status, and thesis proposal) for dropout prediction. This may not cover all the complex factors contributing to student dropout (e.g., financial, personal, social issues, etc.). Also, data was only gathered from one department/university. This restriction may limit the generalization of the findings to other departments or institutions. Moreover, the limited data time-frame, with the training data from 2010–2015 and testing in 2018, may not reflect newer trends or academic policies.

As for future work, external or non-academic criteria (such as demographic, economic, student activity participation, and satisfaction survey) can be combined to build a more comprehensive model. Next, explore advanced algorithms such as ensemble learning techniques or deep learning. Finally, the implementation of a real-time predictive system integrated with the university's academic information system for early intervention.

Acknowledgment

The authors gratefully acknowledge the financial support for the research and publication of this article from the Center for Research and Publication (Puslitpen) of the Institute for Research and Community Service (LP2M) UIN Syarif Hidayatullah Jakarta.

References

[1] A. Lubis *et al*., *Pedoman Akademik UIN Syarif Hidayatullah Jakarta 2022*. [Online]. Available: https://asset.uinjkt.ac.id/uploads/assets/pedoman/pedoman-akademik-uin-2022-2023.pdf

[2] M. Sharma and M. Yadav, "Predicting students' drop-out rate using machine learning models: A comparative study," *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, pp. 1166–1171, Aug. 2022, doi: 10.1109/icicict54557.2022.9917841.

[3] M. Wach and I. Chomiak-Orsa, "The application of predictive analysis in decision-making processes on the example of mining company's investment projects," *Procedia Computer Science*, vol. 192, pp. 5058–5066, Jan. 2021, doi: 10.1016/j.procs.2021.09.284.

[4] Q. Aini, H. A. Pratama, and R. H. Kusumaningtyas, "Potential drop out students: utilizing C4.5 algorithm and naive Bayes classification," *2024 12th International Conference on Cyber and IT Service Management (CITSM),* pp. 1–7, Oct. 2024, doi: 10.1109/citsm64103.2024.10775424.

[5] G. Gunawan, H. Hanes, and C. Catherine, "C4.5, k-nearest neighbor, naïve bayes, and random forest algorithms comparison to predict students' on time graduation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 4, no. 2, pp. 62–71, Nov. 2021, doi: 10.24014/ijaidm.v4i2.10833.

[6] X. Wang, C. Zhou, and X. Xu, "Application of C4.5 decision tree for scholarship evaluations," *Procedia Computer Science*, vol. 151, pp. 179–184, Jan. 2019, doi: 10.1016/j.procs.2019.04.027.

[7] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on C4.5 algorithm for online voltage stability assessment,"

International Journal of Electrical Power & Energy Systems, vol. 118, Art. no. 105793, Dec. 2019, doi: 10.1016/j.ijepes.2019.105793.

[8] F. Akbar, H. W. Saputra, A. K. Maulaya, M. F. Hidayat, and R. Rahmaddeni, "Implementasi algoritma decision tree C4.5 dan support vector regression untuk prediksi penyakit stroke," *MALCOM Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 2, pp. 61–67, Sep. 2022, doi: 10.57152/malcom.v2i2.426.

[9] G. I. Webb, "Naïve Bayes," in *Encyclopedia of Machine Learning*, 2010, pp. 713–714. doi: 10.1007/978-0-387-30164-8_576.

[10] S. Sumanto, L. S. Marita, L. Mazia, and T. W. Ratnasari, "Analisis kelayakan kredit rumah menggunakan metode naïve bayes untuk mengurangi kredit macet," *Applied Information System and Management (AISM)*, vol. 4, no. 1, pp. 17–22, Apr. 2021, doi: 10.15408/aism.v4i1.20274.

[11] M. Muhathir, M. H. Santoso, and R. Muliono, "Analysis naive bayes in classifying fruit by utilizing hog feature extraction," *Journal of Informatics and Telecommunication Engineering*, vol. 4, no. 1, pp. 151–160, Jul. 2020, doi: 10.31289/jite.v4i1.3860.

[12] O. Peretz, M. Koren, and O. Koren, "Naive bayes classifier – An ensemble procedure for recall and precision enrichment," *Engineering Applications of Artificial Intelligence*, vol. 136, Art. no. 108972, Jul. 2024, doi: 10.1016/j.engappai.2024.108972.

[13] A. L. Fairuz, R. D. Ramadhani, and N. A. F. Tanjung, "Analisis sentimen masyarakat terhadap COVID-19 pada media sosial Twitter," *Journal of Dinda Data Science Information Technology and Data Analytics*, vol. 1, no. 1, pp. 42–51, Feb. 2021, doi: 10.20895/dinda.v1i1.180.

[14] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "Implementasi algoritma klasifikasi k-nearest neighbor (KNN) untuk klasifikasi seleksi penerima beasiswa," *IJCIT (Indonesian Journal on Computer and Information Technology)*, vol. 6, no. 2, pp. 118–127, Dec. 2021, doi: 10.31294/ijcit.v6i2.10438.

[15] B. Lantz, *Machine Learning with R: Expert techniques for predictive modeling*. Packt Publishing Ltd, 2019.

[16] Y. Heryadi and T. Wahyono, *Machine Learning: Konsep dan Implementasi*. Penerbit Gava Media, 2020.

[17] A. A. Ajhari, "The comparison of sentiment analysis of moon knight movie reviews between multinomial naive bayes and support vector machine," *Applied Information System and Management (AISM)*, vol. 6, no. 1, pp. 13–20, Apr. 2023, doi: 10.15408/aism.v6i1.26045.

[18] N. Hasanati, Q. Aini, and A. Nuri, "Implementation of support vector machine with lexicon based for sentiment analysis on twitter," *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–4, Sep. 2022, doi: 10.1109/citsm56380.2022.9935887.

[19] B. Waspodo, Q. Aini, F. R. Singgih, R. H. Kusumaningtyas, and E. Fetrina, "Support vector machine and lexicon based sentiment analysis on kartu prakerja (indonesia pre-employment cards government initiatives)," *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, pp. 01–06, Sep. 2022, doi: 10.1109/citsm56380.2022.9935990.

[20] J. Han, M. Kamber, and J. Pei, *Data mining: Data mining Concepts and Techniques*. Morgan Kaufmann, 2014.

[21] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Information Sciences*, vol. 507, pp. 772–794, Jul. 2019, doi: 10.1016/j.ins.2019.06.064.

[22] O. Arifin and T. B. Sasongko, "Analisa perbandingan tingkat performansi metode support vector machine dan naïve bayes classifier," *Seminar Nasional Teknologi Informasi dan Multimedia*, vol. 6, no. 1, pp. 67–72, 2018.

[23] Google Developers, "Classification: ROC and AUC," *Google for Developers*. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=AUC%20stands%20for%20%22Area%20under,across%20all%20possible%20classification%20thresholds.

[24] L. R. Haidar, E. Sediyono, and A. Iriani, "Analisa prediksi mahasiswa drop out menggunakan metode decision tree dengan algoritma ID3 dan C4.5," *Jurnal Transformatika*, vol. 17, no. 2, p. 97–106, Jan. 2020, doi: 10.26623/transformatika.v17i2.1609.

[25] D. Sinaga, E. J. Solaiman, and F. J. Kaunang, "Penerapan Algoritma Decision Tree C4.5 Untuk Klasifikasi Mahasiswa Berpotensi Drop out Di Universitas Advent Indonesia," *TeIKa Jurnal Teknologi Informasi Dan Komunikasi*, vol. 11, no. 2, pp. 167–173, Oct. 2021, doi: 10.36342/teika.v11i2.2613.

[26] M. T. Anwar, L. Heriyanto, and F. Fanini, "Model prediksi dropout mahasiswa menggunakan teknik data mining," *Jurnal Informatika Upgris*, vol. 7, no. 1,pp. 56–60, Jun. 2021, doi: 10.26877/jiu.v7i1.8023.

[27] A. F. A. Rahman, S. Sorikhi, and S. Wartulas, "Prediksi kelulusan mahasiswa menggunakan algoritma C4.5 (studi kasus di universitas peradaban)", ijir, vol. 1, no. 2, pp. 70–77, Dec. 2020.S. D. Purba, L. Harahap, and J. F. R. Panggabean, "Prediction of students drop out with support vector machine algorithm," *Jurnal Mantik*, vol. 6, no. 1, pp. 582–586, 2022.

[28] Q. Aini, H. A. Pratama, and R. H. Kusumaningtyas, "Potential drop out students: utilizing C4.5 algorithm and naive Bayes classification," *2024 12th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–7, Oct. 2024, doi: 10.1109/citsm64103.2024.10775424.

[29] A. Anwarudin, W. Andriyani, B. P. Dp, and D. Kristomo, "The prediction on the students' graduation timeliness using naive Bayes classification and K-Nearest neighbor," *Journal of Intelligent Software Systems*, vol. 1, no. 1, pp. 75–88, Jul. 2022, doi: 10.26798/jiss.v1i1.597.

[30] T. Triase, S. Sriani, and K. Khairuna, "Usability algoritma supervised learning untuk prediksi kelulusan mahasiswa pada sistem bimbingan akademik," *Indonesian Journal of Computer Science*, vol. 11, no. 3, pp. 1015–1022, 2022.

[31] N. Nurwati, N. Azizah, and Y. Santoso, "Penerapan algoritma K-Nearest Neighbor untuk prediksi mahasiswa berpotensi dropout," *Journal Sensi*, vol. 9, no. 1, pp. 74–83, Jan. 2023, doi: 10.33050/sensi.v9i1.2624.

[32] S. D. Purba, L. Harahap, and J. F. R. Panggabean, "Prediction of students drop out with support vector machine algorithm," *Mantik*, vol. 6, no. 1, pp. 582-586, May 2022.

[33] H. Hairani, "Peningkatan konerja metode SVM menggunakan metode KNN imputasi dan k-means-smote untuk klasifikasi kelulusan mahasiswa universitas bumigora," *Jurnal Teknologi Informasi Dan Ilmu Komputer*, vol. 8, no. 4, pp. 713–718, Jul. 2021, doi: 10.25126/jtiik.2021843428.

[34] M. Hosseinzadeh *et al*., "Data cleansing mechanisms and approaches for big data analytics: a systematic study," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 1, pp. 99–111, Nov. 2021, doi: 10.1007/s12652-021-03590-2.