Comparison of K-Nearest Neighbour Method with Naive Bayes Classifier and Support Vector Machines for Student Graduation Classification

Rubangi
University Technology Yogyakarta
Yogyakarta, Indonesia

Sutarman
University Technology Yogyakarta
Yogyakarta, Indonesia

ABSTRACT

Timely student graduation is a hallmark of student success in obtaining a bachelor's degree. During the lecture period, students are not necessarily able to complete the lecture period on time because many factors influence student graduation to be late. One of the factors that determine the quality of higher education in lectures is the presentation of students' ability to complete college studies on time. The length of time students study affects the quality of the study programme because student study time is used as one of the criteria in determining the assessment by BAN PT (National Accreditation Board of Higher Education). With the existence of these problems that occur, it can be overcome in research, namely the classification of student graduation on time by using the K-Nearest Neighbor Algorithm, Naive Bayes Classifier and Support Vector Machines algorithm methods to classify the accuracy of student graduation. The implementation of the three algorithms was carried out with Rapid miner software. After training and testing with 543 datasets, the classification of student graduation on time with the best accuracy of the three methods is the K-Nearest Neighbor method, the accuracy obtained is 100%, then for the classification of on-time graduation, the Naive Bayes Classifier method obtained an accuracy of 97.11%, and the Support Vector Machines method of classifying student graduation on time, the accuracy obtained is 84.56%.

General Terms

Student Recovery, Graduation, University

Keywords

Classification, Student Graduation, KNN, NBC, SVM

1. INTRODUCTION

Timely graduation is a characteristic of student success in obtaining a bachelor's degree [1]. The student graduation rate is an indicator of the success of higher education to achieve the right graduation rate, universities must plan the learning process so that students can graduate on time [2]. Student graduation is a very important thing for achievement in the world of education so that it affects the value of accreditation in education [3]. One of the factors that determine the quality of higher education in lectures is the presentation of students' ability to complete college studies on time [4]. In addition, the challenge in higher education is to improve the quality of educational programmes, in this case the success or failure of students to be able to complete lectures on time which can be done by evaluating to improve the quality of higher education in learning [5]. During the lecture period, students may not necessarily be able to complete the study period on time because many factors influence late student graduation such as the level of student understanding in understanding lecture material, working and non-working student status, and student marital status [6]. This is the importance of the classification of student graduation to determine the status of student graduation is the prediction of student graduation. However, the question is how to know whether a student will graduate on time and not on time [7]. There are several methods that can be used to overcome the prediction of student graduation, namely by using a classification algorithm approach. Algorithmic methods that can be used to complete the classification are K-Nearest Neighbour, Naive Bayes Classifier and Support Vector Machines. In the process of classifying student graduation predictions, there are many factors and criteria for measuring student graduation and determining whether or not students are eligible to complete their studies [8]. Classifying students based on academic achievement is an effective strategy to reduce the failure of resource management in higher education [9]. Prediction of student graduation can be done by the process of student graduation classification. So as to overcome the problem of predicting student study time with the K-Nearest Neighbor method being able to predict study time with cross validation accuracy of 70, 28%[10]. To solve the problem of predicting student study time, the Naive Bayes method has an accuracy of 69.33% [11]. while to solve the problem of predicting student study time with the Support Vector Machine method is able to predict study time with an accuracy of 86.36%[12]. Based on the background, this research compares the K-Nearest Neighbor method with Naive Bayes Classifier and Support Vector Machines algorithms for student graduation. This research is expected to provide benefits for universities to formulate policies so that students can graduate on time.

2. RESEARCH METHOD

In this study, a comparison of the K-Nearest Neighbor Method with Naive Bayes Classifier and Support Vector Machines was carried out for the classification of student graduation. This research begins with data collection. After data collection, several data preprocessing processes will be carried out. Then use the Rapid miner tool to manage the data by training and testing data using the K-Nearest Neighbor algorithm, Naive Bayes Classifier and Support Vector Machines. Training and testing aims to produce classification and accuracy values. There are several stages of this research method, namely: (1) collecting student graduation data; (2) preprocessing student graduation data; (3) algorithm design; (4) training & testing; (5) algorithm comparison. The following stages of the research methodology are illustrated in Figure 1.

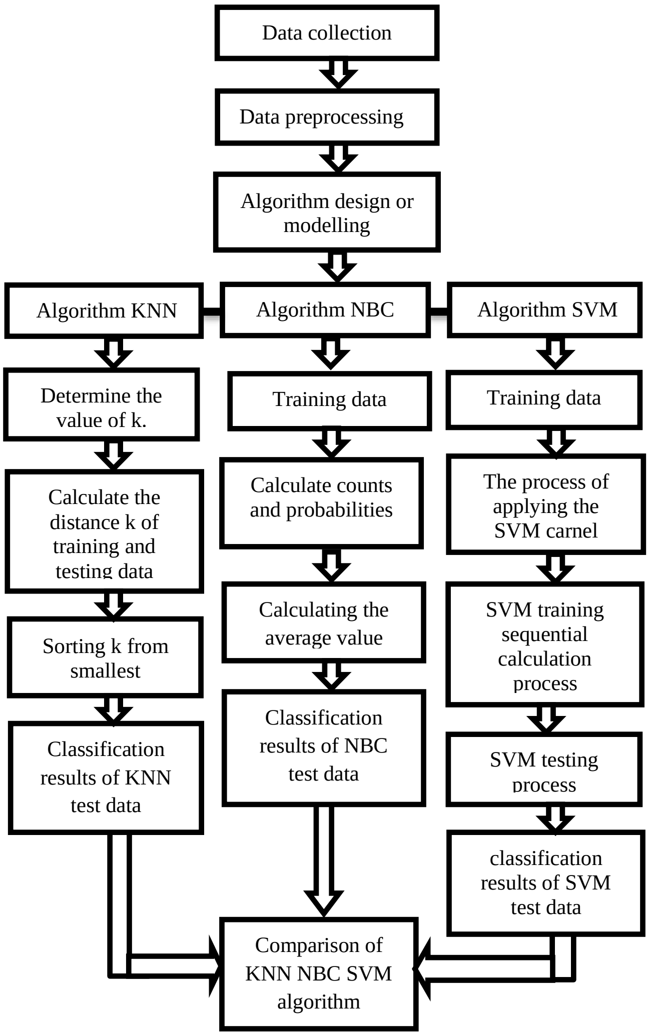

Data collection
Data preprocessing
Algorithm design or modelling
Algorithm KNN
Algorithm NBC
Algorithm SVM
Determine the value of k.
Calculate the distance k of training and testing data
Sorting k from smallest
Classification results of KNN test data
Training data
Calculate counts and probabilities
Calculating the average value
Classification results of NBC test data
Training data
The process of applying the SVM carnel
SVM training sequential calculation process
SVM testing process
classification results of SVM test data
Comparison of KNN NBC SVM algorithm

Fig 1: Research Methods

2.1 Data Collection

The data sampling process obtained in the form of Yogyakarta University of Technology student graduation dataset. The research took data from students of the Informatics Engineering study programme at Yogyakarta University of Technology class 2014 and 2015. The amount of data used is 543 student data. Variables used for student graduation classification consist of gender, major, SKS, GPA, age, and graduation status (graduated on time and not on time).

2.2 Data Preprocessing

Data Preprocessing Stage: Before using data through data mining methods or techniques, there are several steps that must be taken, namely preprocessing the data. The preprocessing stage looks for data whose values are missing, meaning that the data obtained is empty data, this stage creates tags/classes, converts string attributes into numbers. Preprocessing is the process of identifying incomplete, incorrect, irrelevant, inaccurate, or missing pieces of data, then correcting, replacing, or deleting them as necessary. In this research, data conversion of 2 (three) attributes (gender, major) in the form of text into numeric.

2.3 Algorithm Design

After the preprocessing stage is completed, the next stage is to classify the students' graduation process on time and not on time using various algorithm methods used. One of the algorithms that can be used to complete classifying student graduation data for on-time graduates is the K-Nearest Neighbour algorithm, Naive Bayes Classifier, and Support Vector Machine. The K-nearest neighbour algorithm is an object classification method based on the nearest neighbour to the object. The K-Nearest Neighbor algorithm performs classification to find the shortest distance between the data to be evaluated and the K nearest neighbours in the training data. Then the Naive Bayes algorithm method is a classification algorithm used in machine learning that processes training data to produce a classification model that can be used to predict the appropriate class label for previously unseen data. In testing, the Naive Bayes algorithm is used to predict the target class from unknown data by taking into account the probabilities calculated during training [13]. While the Support Vector Machine algorithm method is a classification method. SVM uses a kernel function to map low-dimensional sample data from the original nonlinear space to a high-dimensional feature space, and constructs an optimal classification hyperplane in the search space[14]. The graduation data is then processed by each method from KNN, NBC and SVM on the Rapidminer tool.

2.4 Training And Testing Data

In this research, the training data in the dataset is used to train the KNN, NBC and SVM algorithms. The training dataset is data that is processed and pre-processed using the previous KNN, NBC and SVM algorithms. The data used totalled 543 data sets, then the training data was divided into two parts, namely the training data set with a proportion of 70% and the test data set with a proportion of 30%. Training and testing can also be training sets with 80% and test data sets with a proportion of 20%. The results of data testing are carried out to determine the performance of each algorithm used. This research tests the use of KNN, NBC, and SVM algorithms to classify on-time and off-time graduates. This testing stage illustrates how the algorithm is used to determine the accuracy value. The performance of the resulting method will show the accuracy of the algorithm model has been used before.

2.5 Training And Testing Data

This research compares the K-Nearest Neighbour (KNN) method with Naive Bayes classifier (NBC) and support vector machine (SVM) to classify student graduation on time. To get the best results in determining whether students graduate from college on time or not, a comparison is made to produce higher accuracy. A classification technique is said to be good if it is able to produce an estimate of the model with a high accuracy value [15].

3. RESULTS AND DISCUSSION

3.1 Results

After analysing the data to compare the K-Nearest Neighbor, Naive Bayes Classifier and Support Vector Machines methods, the results achieved by researchers are to get the right or best method for graduating students on time at the Yogyakarta University of Technology Informatics study programme. Researchers use the Rapid Miner application to simplify the testing and training process which results in student graduation classification accuracy. The results of the classification of student graduation using the K-Nearest Neighbor method resulted in a graduation classification with an accuracy of 100%, the Naive Bayes Classifier method with a graduation classification accuracy of 97.11% while the Support Vector Machines method resulted in a graduation classification with an accuracy of 84.56%.

3.2 Classification Process Model of KNN, NBC and SVM Algorithms

3.2.1 KNN algorithm method process

Classification of student graduation using the K-Nearest Neighbour method there are several testing and training processes carried out. The K-Nearest Neighbour model process can be seen in Figure 2.

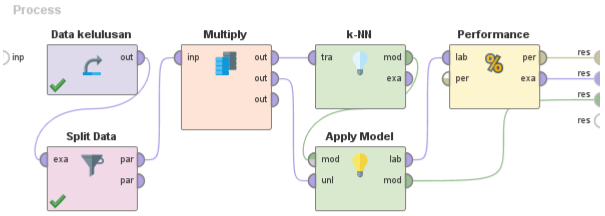


Fig 2: Model K-Nearest Neighbour Method

In classifying student graduation, the KNN algorithm model is used by performing several operators, namely: this student graduate data set contains student graduation data in Excel format, split data to separate training data from test data with parameters 70% to 30%, Multiply is used to divide the data, K-NN algorithm is used for classification, Apply model is used for testing the data seen from the KNN operator, and Performance serves to measure the performance results of the accuracy value of the KNN model.

3.2.2 NBC algorithm method process

In the process of classifying student graduation with the Naive Bayes Classifier method by performing several operators, namely graduation data, split data, multiply, Naive Bayes, apply model and performance. The Naive Bayes model process can be seen in the figure 3.

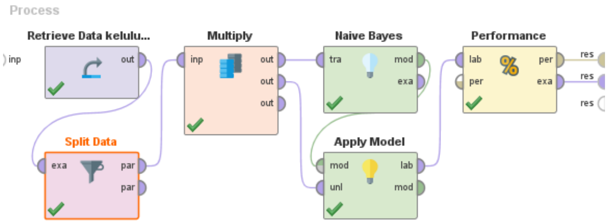


Fig 3: Model Naive Bayes Classifier Method

From the Naive Bayes classifier algorithm method, there are several steps, namely the graduation data operator with excel format containing student graduates. Split data operator used to separate training data from test data with parameters 70% to 30% and parameters 80% to 20%. Multiply is used to divide data from split to Naive Bayes operator and applay model. Naive Bayes operator is an algorithm used for student graduation classification. The apply model operator for the model has been done training and testing classification. While the performance operator measures the classification performance results of the accuracy value of the Naive Bayes method model.

3.2.3 SVM algorithm method process

Classification of student graduation using the Support Vector Machines algorithm method is carried out using the RapidMiner tool by performing several operators, namely graduation data, split data, Replace missing Value, set role, and testing data using cross validation, SVM, apply model and performance, after that connect all panels. The display of connected panels can be seen in Figure 4 and Figure 5 below.

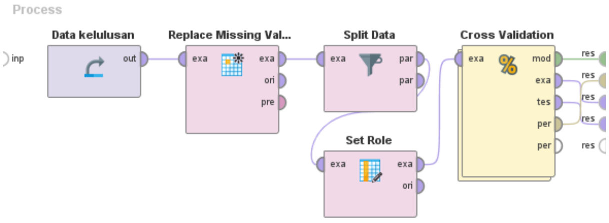


Fig 4: Model Support Vector Machines Method

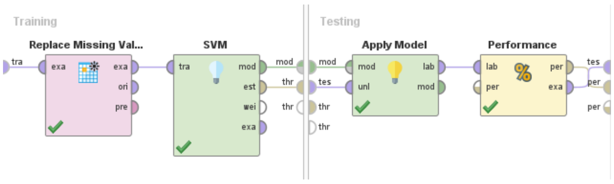


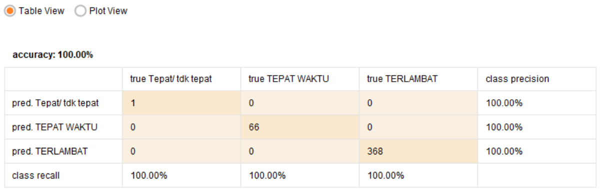
Fig 5: Testing with Cross Validation

In Figure 4 above the SVM model process by performing several operators, namely: graduate data set containing student graduation data in excel form, split data to separate training data from test data with parameters 70% by 30%. In the training process on the data set there is an error so it requires Replace Missing Values which is to fix the data there are missing values or empty data that needs to be corrected. SVM model validation process can be seen in Figure 5. The validation process involves two different sub-processes: training and testing. The training sub-process is used to instruct the model, which can then be applied in the model application and performance testing sub-processes. The Apply Model operator serves as a link to the SVM method. On the other hand, Performance is responsible for measuring the accuracy of the SVM model.

3.3 Analysis Results of KNN, NBC and SVM methods

3.3.1 KNN method student graduation classification results

The results of student graduation classification with the performance of the KNN algorithm method can be seen in Figure 6.



● Table View ○ Plot View

accuracy: 100.00%

	true Tepat/ tdk tepat	true TEPAT WAKTU	true TERLAMBAT	class precision
pred. Tepat/ tdk tepat	1	0	0	100.00%
pred. TEPAT WAKTU	0	66	0	100.00%
pred. TERLAMBAT	0	0	368	100.00%
class recall	100.00%	100.00%	100.00%	

Fig 6: Classification Results of K Nearest Neighbour Method

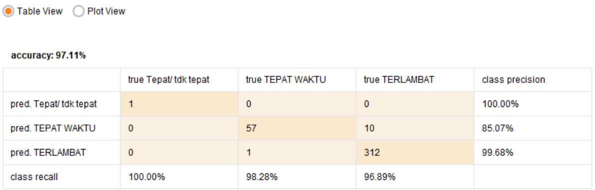
Based on the results of testing student graduation data using the KNN method of student graduation classification with parameters 70% training data and 30% test data, there are 66 on-time graduates and 368 graduates who are not on time, which shows an accuracy rate of 100.00%. In the K-Nearest Neighbors method to get the best accuracy results, testing with the value of k in the algorithm, namely the parameter value K = 1 K = 3, K = 7, K = 9, K = 11 as table 1 below.

Table 1. Comparison Results of Parameter K of KNN method

K Value Testing	KNN Accuracy Performance
K - 1	100,00%
K - 3	98,43%
K - 5	85,83%
K - 7	84,78%
K - 9	84,51%
K - 11	84,25%

In table 1 above that the performance results of the K-Nearest Neighbors method in the classification of student graduation in testing the best accuracy value is generated at k1 100.00%, the lowest accuracy at k11 is 84.37%. Based on the results of the accuracy data, the value of the parameter k greatly affects the accuracy results where the smaller k, the better the accuracy results.

3.3.2 NBC method student graduation classification results
The classification results of student graduation with the Naive Bayes classifier algorithm method can be seen in Figure 7.



accuracy: 97.11%

	true Tepat/ tdk tepat	true TEPAT WAKTU	true TERLAMBAT	class precision
pred. Tepat/ tdk tepat	1	0	0	100.00%
pred. TEPAT WAKTU	0	57	10	85.07%
pred. TERLAMBAT	0	1	312	99.68%
class recall	100.00%	98.28%	96.89%	

Fig 7: Classification Results of Naive Bayes Method

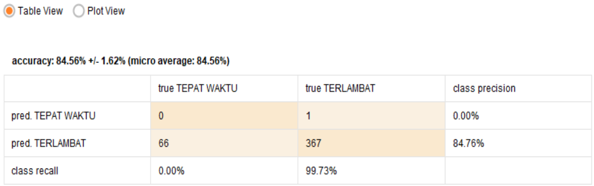
Based on the results of testing student graduation using the Naive Bayes classifier method of student graduation classification with parameters 70% training data and 30% test data, the results of on-time graduates are 57 people and 312 graduates are not on time, which shows an accuracy rate of 97.11%. In the Naive Bayes method to get optimal results, test data with parameters such as table 2.

Table 2. Comparison Of Naive Bayes Method Testing Parameters

Parameter Testing	Parameters 70% To 30%	Parameter 80% To 20%
Accuracy	97.11%	95.40%

In table 2 above that the results of parameter testing performance to get maximum accuracy using the Naive Bayes method in parameter testing 70% training data and 30% test data, based on the results of these tests the accuracy is 97.11%. while testing data with parameters 80% training data and 20% test data get an accuracy of 95.40%.

3.3.3 SVM method student graduation classification results
The classification results of student graduation using the Support Vector Machines algorithm can be seen in Figure 8.



accuracy: 84.56% +/- 1.62% (micro average: 84.56%)

	true TEPAT WAKTU	true TERLAMBAT	class precision
pred. TEPAT WAKTU	0	1	0.00%
pred. TERLAMBAT	66	367	84.76%
class recall	0.00%	99.73%	

Fig 8: Testing with Cross Validation

Based on the accuracy results with Support Vector Machines in Figure 8, the classification of student graduation shows that the Support Vector Machines method in classifying student graduation with the best parameters, namely 80% training data and 20% test data with the results of student graduation classification. Based on testing the Support Vector Machines method with an accuracy result of 84.56%, the classification of student graduates on time is 66 people and 367 graduates are not on time. In the Support Vector Machines method to get optimal results, test data with parameters such as table 3.

Table 3. Comparison Of Support Vector Machines Method Testing Parameters

Parameter Testing	Parameters 70% To 30%	Parameter 80% To 20%
Accuracy	84.47%	84.56%

In table 3. above that the results of parameter testing performance to get maximum accuracy using the Support Vector Machines method in parameter testing 70% training data and 30% test data, based on the results of these tests the accuracy is 84.47%. while testing data with parameters 80% training data and 20% test data get an accuracy of 84.56%.

3.4 Discussion of Result

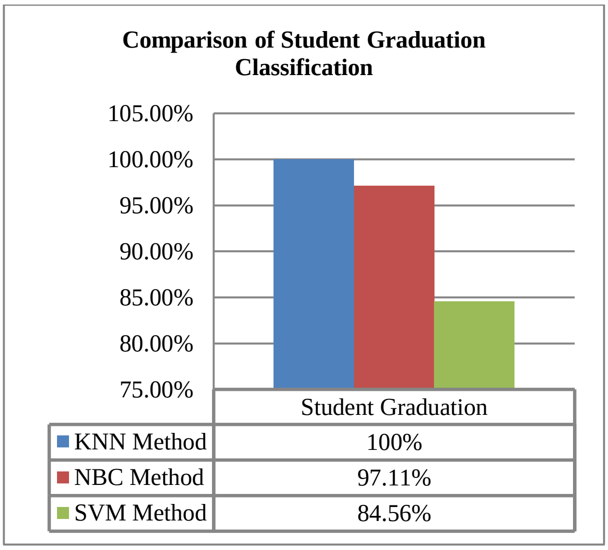


Fig 9: Student Graduation Classification Comparison Results

As shown in Figure 9, the comparison of the classification of on-time graduation of students between the K-Nearest Neighbor, Naive Bayes and Support Vector Machines methods. Based on the classification results, the best performance performance on the test is the K-Nearest Neighbor (KNN) algorithm method with 100% accuracy and followed by the Naive Bayes (NB) method which obtained 97.11% accuracy. While based on the results of the lowest classification performance is not good in processing student graduation data using the Support Vector Machines (SVM) algorithm method with an accuracy of 84.56%. Based on the performance results of Figure 9, the comparison of accuracy can make decisions for the classification of student graduation students on time and late with the K-Nearest Neighbor algorithm method, this is influenced by the smaller k, the better the accuracy obtained.

4. CONCLUSION

In this study using the K-Nearest Neighbor (KNN), Naive Bayes Classfier (NBC) and Support Vector Machines (SVM) methods of student graduation classification, the conclusions regarding the comparison of the KNN, NBC, and SVM methods are as follows:

1. Based on the results of testing the classification of student graduation using the KNN method produces a graduation classification with an accuracy of 100%, while the NBC method with a graduation classification accuracy of 97.11% and the SVM method produces a graduation classification with an accuracy of 84.56%.

2. Based on the results of the accuracy level produced by the K-Nearest Neighbor algorithm for the classification of student graduation with 66 on-time graduates and 368 untimely graduates showing an accuracy level of 100.00% on the K-1 value parameter.

3. Based on the results of the accuracy rate generated by the Naive Bayes algorithm for the classification of student graduation with 57 on-time graduates and 312 untimely graduates, which shows an accuracy rate of 97.11%.

4. Based on the results of the accuracy level generated by the SVM algorithm for the classification of student graduation with on-time graduates totalling 66 people and 367 graduates not on time, resulting in an accuracy of 84.56%.

5. Based on the results of the study Get the best method for graduating students on time so that it helps universities in making policies to be able to increase student graduation on time.

5. REFERENCES

[1] T. H. Hasibuan and D. Mahdiana, "Prediction of Timely Student Graduation Using the C4.5 Algorithm at Uin Syarif Hidayatullah Jakarta," SKANIKA Sist. Comput. and Tech. Inform., vol. 6, pp. 61-74, 2023.

[2] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," J. Eng. Appl. Technol., vol. 2, no. 2, pp. 86-91, 2021, doi: 10.21831/jeatech.v2i2.42777.

[3] S. R. Rani, S. R. Andani, and D. Suhendro, "Application of K-Nearest Neighbor Algorithm for Student Graduation Prediction at SMK Anak Bangsa," Pros. Semin. Nas. Ris. Inf. Sci., no. September, pp. 670-676, 2019.

[4] O. W. Yuda, D. Tuti, L. S. Yee, and Susanti, "Application of Data Mining for Classification of On-Time Student Graduation Using Random Method," SATIN Sains dan Teknol. Inf., vol. 8, no. 2, pp. 122-131, 2022, doi: 10.33372/stn.v8i2.885.

[5] R. H. Sukarna and Yulian Ansor, "Implementation of Data Mining Using the Naive Bayes Method with Feature Selection to Predict Student Graduation on Time," J. Ilm. Science and Technol., vol. 6, no. 1, pp. 50-61, 2022, doi: 10.47080/saintek.v6i1.1467.

[6] N. Khasanah, A. Salim, N. Afni, R. Komarudin, and Y. I. Maulana, "Prediction of Student Graduation with the Naive Bayes Method," Technol. J. Ilm., vol. 13, no. 3, pp. 207, 2022, doi: 10.31602/tji.v13i3.7312.

[7] K. R. Diska and K. Budayawan, "Graduation Prediction Information System Using the Naive Bayes Classifer Method (Case Study: Informatics Engineering Education Study Program)," J. Educ. Tambusai, vol. 7, no. 1, pp. 936-943, 2023, doi: 10.31004/jptam.v7i1.5375.

[8] M. R. Qisthiano, P. A. Prayesy, and I. Ruswita, "Application of Decision Tree Algorithm in Classification of Student Graduation Prediction Data," G-Tech J. Technol. Applied, vol. 7, no. 1, pp. 21-28, 2023, doi: 10.33379/gtech.v7i1.1850.

[9] N. Rijati and E. A. Hakim, "Model Development with Machine Learning Approach for Student Academic Performance Prediction," Semin. Engineering 2023, pp. 784-789, 2023.

[10] H. Manarul, A. Faqih, and T. Suprapti, "Implementation of K-Nearest Neighbour Algorithm for Graduation Accuracy Prediction," JURSIMA J. Sist. Inf. and Manaj., vol. 10, no. 2, 2022.

[11] T. M. Rahayu, B. A. Ningsi, Isnurani, and I. Arofah, "Classification of Student Graduation Timeliness with the Naïve Bayes Method," Media Bina Ilm., vol. 15, no. 8, pp. 4993-5000, 2021.

[12] M. F. Abdullah, Kusrini, and M. R. Arief, "Prediction of Student Grades and Graduation Time Using the Support Vector Machine Method," SAINTEKBU J. Science and Technol., vol. 14, no. 1, 2022.

[13] A. D. Cahyo, "Naive Bayes Method for Classification of Undergraduate Study Period," J. Technol. Pint., vol. 3, no. 4, 2023, [Online]. Available at: http://teknologipintar.org/index.php/teknologipintar/article/view/385%0Ahttp://teknologipintar.org/index.php/teknologipintar/article/download/385/370

[14] S. Bumbungan, Kusrini, and Kusnawi, "Application of Particle Swarm Optimisation (PSO) in Automatic Parameter Selection in Support Vector Machine (SVM) for Prediction of Graduation of Amamapare Timika Polytechnic Students," J. Tek. AMATA, vol. 04, no. 1, pp. 81-93, 2023.

[15] D. Indahsari, I. Maulana, and A. Primajaya, "Classification of Students' Level of Understanding of Gustav Jung's Personality Theory Using the C4.5 Algorithm," JUSTINDO (Journal of Systems and Technol. Inf. Indones., vol. 7, no. 1, pp. 31-41, 2022, doi: 10.32528/justindo.v7i1.546.