e-ISSN : 2597-3673 (Online) , p-ISSN : 2579-5201 (Printed)
Vol.6 No.1, June 2022

Journal of Information System, Informatics and Computing
Website/URL: http://journal.stmikjayakarta.ac.id/index.php/jisicom
Email: jisicom@stmikjayakarta.ac.id , jisicom2017@gmail.com

OPTIMIZATION OF NAÏVE BAYES ALGORITHM PARAMETERS FOR STUDENT GRADUATION PREDICTION AT UNIVERSITAS DIRGANTARA MARSEKAL SURYADARMA

Muryan Awaludin [1], Verdi Yasin [2], Mega Wahyuningsih [3*]

[1] Faculty of Industrial Technology, Universitas Dirgantara Marsekal Suryadarma, Indonesia

[2] Informatics Engineering study program, STMIK Jayakarta, Jakarta - Indonesia

[2] Faculty of Industrial Technology, Universitas Dirgantara Marsekal Suryadarma, Indonesia

*Corresponding E-mail: muryanawaludin1@gmail.com, verdiyasin29@gmail.com, megawahyuningsih97@gmail.com

Received: February 27, 2022. **Revised:** March 25, 2022. **Accepted:** May 10, 2022.
Issue Period: Vol.6 No.1 (2022), **Pp.** 91-106

Abstract. The Information Systems Study Program at Unsurya is a new department and only a few graduate students. Based on data obtained from graduates of the 2018/2019 academic year, 41 students graduated, including 26 students who experienced delays in taking their studies. A system that can predict student graduation is needed so that the Information Systems department can produce more student graduations than before. By optimizing the parameters of the Naïve Bayes algorithm, it can be applied in predicting graduation by utilizing previous student graduation data, the attributes used are gender, age, sks, gpa, and student status. The results of research testing using Rapid Miner 9.8 with 41 training data and 25 testing data, yielding 96% accuracy, 90.91% recall, and 100% precision.

Keywords: Information Systems, Naïve Bayes Algorithm, optimizing the parameters, predicting graduation, rapid mine

1. Introduction

The more students who graduate on time, the better the college's performance will be, so that the student graduation rate on time is one of the performance assessments of accreditation for a university or Study Program [1]. Ensuring adequate quality in the education system is critical to student performance and the overall value of the knowledge provided, there are many benefits to detecting student problems and learning difficulties early on, as this provides a unique opportunity to address causal factors in a timely manner to prevent student failure and trends. drop out of university [2].

DOI: 10.52362/jisicom.v6i1.785

The prediction system for student graduation uses Naïve Bayes data mining, the use of historical data on student graduation is more optimal for predicting student graduation [3]. The Naive Bayes algorithm has the advantage of being simple, fast and highly accurate, with successful research having an accuracy of 88.2% [4]. System accuracy testing is carried out by matching the predicted results with real data with the Naïve Bayes algorithm using the confusion matrix testing method, the test results show an accuracy of 80% [5]. Naïve Bayes often performs much better in predicting most complex real-world situations than might be expected [6]. The Naïve Bayes classifier often performs very well in practice, and excellent classification results can be obtained even when the probability estimate contains large errors [7].

Results of training data testing and test data with semester 1 to 4 credit attributes and semester 1 to 4 IP and a comparison scale of 60% training data and 40% test data obtained 91.86% accuracy using the Naïve Bayes algorithm [8]. The Naïve Bayes classifier often performs very well in practice, and excellent classification results can be obtained even when the probability estimate contains large errors [9]. The Naïve Bayes data mining method can make a prediction regarding student graduation on time by taking into account the attributes of the GPA (Cumulative Achievement Index), so that the accuracy of literature results in accuracy above 90% [10]. The results prove that the Naïve Bayes algorithm has been successfully implemented to predict student graduation and is able to produce 73.725% accuracy from 4000 instance data obtained [11]. Studies comparing classification algorithms have found the Naïve Bayes Classifiers to be comparable in performance with certain decision trees and Neural Networks, the Naïve Bayes Classifiers also show high accuracy and speed when applied to large databases [12].

There are various classification algorithms, but only a few are often used to classify data in data mining, namely Naïve Bayes, Neural Network, Decision Tree, K-Nearest Neighbor, and Logistic Regrees. The algorithm that is often used to predict student graduation is the Naïve Bayes algorithm, because it has high accuracy, is fast and space efficient, only requires a small amount of training data to estimate the parameters required for classification, and document classification in terms of the process that takes its class based on data- pre-existing data is easy to understand, however Naïve Bayes has the disadvantage of assuming the independent variable, and if the conditional probability is zero, the prediction probability will be zero.

It is calculated that the Information Systems department has graduated several students based on the data obtained for graduates of the 2018/2019 academic year, 41 students graduated, among the students who graduated there were 26 students who graduated late and 15 students who graduated on time. A total of 22 employee class students with working status experienced late graduation. This is a problem that must be faced by the Department of Information Systems because many students experience delays in taking their studies, especially from the employee class. Students of the Information Systems Study Program have small student graduations so that it will affect accreditation of majors, and the unavailability of student graduation prediction applications in Unsurya makes the Head of Study Program unable to predict students who are worried that their study period will be late. So that there will be problems in the future for students who are late who cannot graduate on time, because there is no preparation for these students to catch up. The more students majoring in Information Systems who successfully graduate, the more graduates can continue to the world of work, so that the Head of Study Program can bring the Information System to increase points for major accreditation. This makes a measure of the excellence of the quality and potential of students, and related agencies can receive and provide assistance to students and graduates.

DOI: 10.52362/jisicom.v6i1.785

2. Literature Review

Research conducted by [13] which applies the Naïve Bayes Classifier Algorithm to determine the topic of the final assignment for classification where existing titles will be classified with related topic statements, thus the title data previously obtained were only IT department students, because of the data used as a learning trial is only the title of the alumni of the STikom IT department students in 2015-2017.

The research conducted [14] uses the Naive Bayes algorithm to classify the welfare status of poor households, namely very poor households (RTSM) and poor households (RTM). The test results obtained are that the Naive Bayes algorithm produces an accuracy of 85.80% and an AUC value of 0.930, the stages in the experiment can be seen in Figure 1.

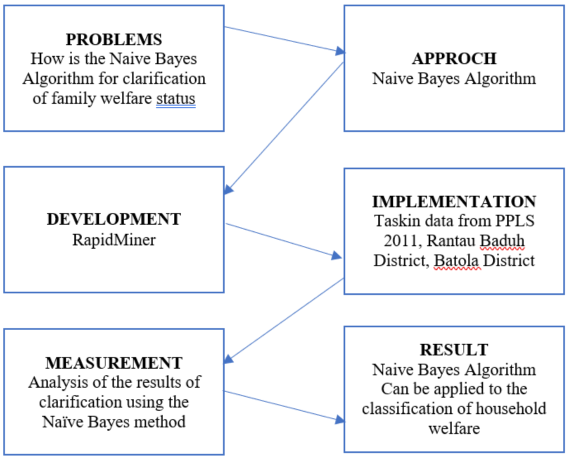

PROBLEMS
How is the Naive Bayes Algorithm for clarification of family welfare status
APPROCH
Naive Bayes Algorithm
DEVELOPMENT
RapidMiner
IMPLEMENTATION
Taskin data from PPLS 2011, Rantau Baduh District, Batola District
MEASUREMENT
Analysis of the results of clarification using the Naïve Bayes method
RESULT
Naive Bayes Algorithm Can be applied to the classification of household welfare

Figure 1. The stages in the experiment.

In the research conducted by [15] they used the cross-validation method in which the research divided 158 images into five groups to train and train the model. The Naïve Bayes model achieves 94.31% accuracy, 94.33% recall, 94.11% precision, and an average F1 score of 0.93, which is the highest among the other three algorithms. Based on the results, it is concluded that Optical Character Recognition (OCR) and Naïve Bayes Algorithm are quite suitable for this problem, the algorithmic flow is as shown in Figure 2.

DOI: 10.52362/jisicom.v6i1.785

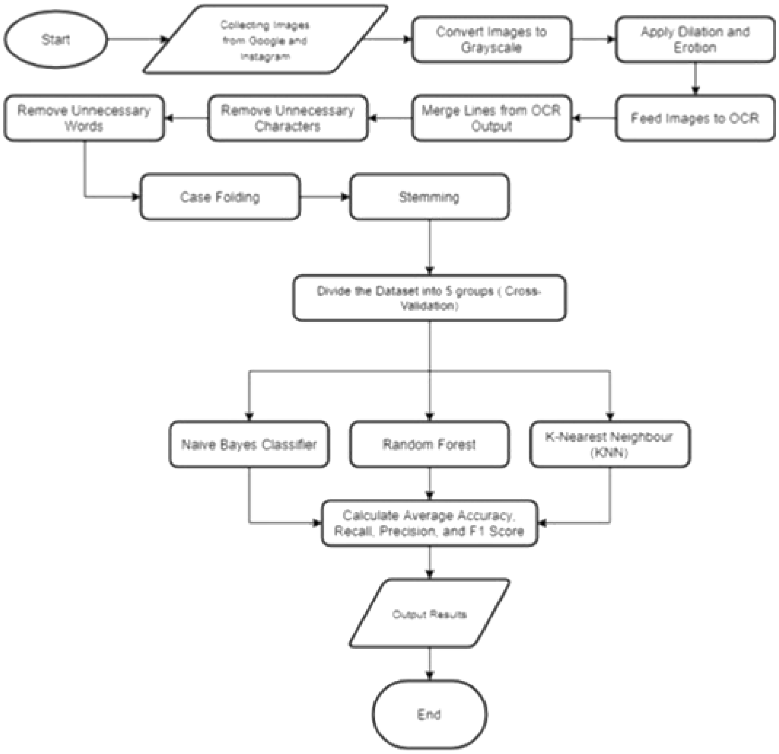

Start
Collecting Images from Google and Instagram
Convert Images to Grayscale
Apply Dilation and Erotion
Feed Images to OCR
Merge Lines from OCR Output
Remove Unnecessary Characters
Remove Unnecessary Words
Case Folding
Stemming
Divide the Dataset into 5 groups (Cross-Validation)
Naive Bayes Classifier
Random Forest
K-Nearest Neighbour (KNN)
Calculate Average Accuracy, Recall, Precision, and F1 Score
Output Results
End

Figure 2. Flowchart

Research conducted by [16] to determine emotions among online diabetes communities using the string-based Multinomial Naïve Bayes algorithm. Facebook posts from official diabetes support groups were crawled, with a total of 15,000 posts pre-processed. Of these, 800 were described manually by human experts. Posts were first classified according to Plutchik's emotion wheel, which consists of eight dominant emotions: anger, sadness, fear, joy, surprise, trust, anticipation, and disgust using the NRC Emotion Lexicon (Emolex).

Emotion classification was then refined using the string-based Naïve Bayes Multinomial algorithm, with results showing a 6.3% increase (i.e. 82% vs. 75.7% for mean F score) when compared to the Emolex approach, and other machine learning algorithms, namely, Naïve Bayes and Multinomial Naïve Bayes. The higher accuracy in emotion classification reflects the feasibility of our approach. Further analysis also revealed emotions such as joy, fear and sadness to be the highest frequency in the diabetes community.

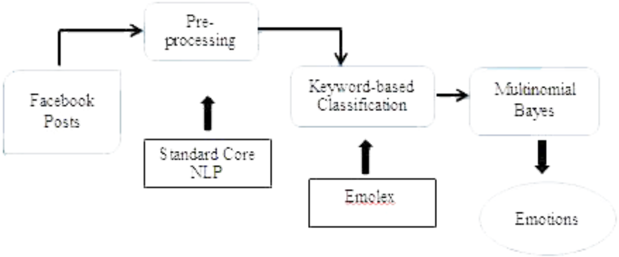


Figure 3. Emotion Detection Pipeline

Analysis of the feature selection for automatic patent categorization carried out by [17] for the corpus of 7,309 patent applications from the World Patent Information Test Collection (WPI) [18], I assign the International Code of the Patent Classification (IPC) section using the Naïve Bayes classifier which. Then the authors compare the precision, gain, and f-measure for various meta parameter settings including data refinement and acceptance threshold. Finally, we found a model optimized to IPC class and group codes and compared the results of categorization of patents in academic literature.

3. Discussion

Data mining is a technology for finding structures and patterns in large data sets. Data mining refers to extracting important knowledge / information from a number of datasets, and presenting it in a form that is easily understood by users [19].

Decision assistance is the goal of data mining and statistics; we now expect this technique to do so rather than just providing a reality model to help us understand it [20]. According to [21] data mining is a combination of a number of computer sciences, which is defined as the process of discovering new patterns from very large data sets, which includes methods that are slices of artificial intelligence, machine learning, statistics, and database systems.

Classification is a technique by looking at the behavior and attributes of a group whose class / label has been defined [22]. Classification is a model formation technique from unclassified data, to be used to classify new data [23]. Classification is defined as a form of data analysis to extract a model that will be used to predict class / label [19]. The purpose of classification is to find a model from training data that differentiates each attribute into the Appropriate class / label, the model is then used to classify attributes whose class / label is not yet known [24]. There are many other methods for building classification models, one of which is Naïve Bayes, which can predict which includes distributed identification based on available data.

Of course, what is of interest is likely future performance on new data, not previous performance on old data [25]. The data classification consists of two stages, namely the learning stage and the classification stage. The learning stage is the stage in the formation of a classification model, while the classification stage is the stage of using a classification model to predict the class / label of data [23].

The method in which the calculation is in the form of dividing the differences into classes, then the calculation used by statistics will be calculated by means of probability and will be divided into classes. Naïve Bayes is one of the most effective and efficient inductive learning algorithms for machine learning and data mining [26].

The Naïve Bayes algorithm has stages in the process, namely:

1. Count the number of classes / labels in the dataset.

DOI: 10.52362/jisicom.v6i1.785

2. Calculating the number of cases per class from the existing data.
3. Multiplying all the class variables.
4. Comparing the results per class, in order to get a conclusion from these comparisons.

Bayes' Theorem simplifies understanding of the Naïve Bayes algorithm, simplified by the following equation:

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \qquad (1)$$

Information:
X : Data with unknown class.
C : The data hypothesis X is a specific class.
P (C | X) : Probability of hypothesis C based on condition X.
P (X | C) : Probability of X based on the conditions in hypothesis C.
P (C) : Probability of hypothesis C (prior probability).
P (X) : Probability X.

Naïve Bayes is a simple classification that calculates probability by adding up frequencies and values from existing data [11]. Naïve Bayes also shows high accuracy and speed when applied to large databases [12].

The following is the process of calculating the Naïve Bayes algorithm starting with data collection. The data obtained to calculate the probability of a condition is called training data, then to perform test data is called testing data. After the data has been successfully obtained, calculations can be carried out from calculating the probability of the condition to calculating the predicted probability.

The training data contained in the table below is an example of data taken to be tested for success.

Table 1. Example Training Data

NO	Gender	Age	Credits	GPA	Status	Graduation
1	male	24-29	average>=19	3.02	working	late
2	male	24-29	average>=19	2.98	working	late
3	male	>29	average>=19	2.91	working	late
4	male	>29	average>=19	3.05	working	late
5	male	18-23	average>=19	3.32	students	late
6	female	18-23	average>=19	3.82	students	late
7	male	24-29	average>=19	3.17	working	late
8	male	24-29	average>=19	2.98	working	late
9	male	24-29	average>=19	2.84	working	late
10	male	24-29	average>=19	2.76	working	late
11	male	24-29	average>=19	3.53	working	Appropriate
12	female	18-23	average>=19	3.13	working	Appropriate
13	female	18-23	average>=19	3.69	students	Appropriate
14	male	18-23	average>=19	3.31	students	Appropriate
15	male	18-23	average>=19	3.17	working	Appropriate

DOI: 10.52362/jisicom.v6i1.785

16	male	18-23	average>=19	2.76	students	Appropriate
17	male	18-23	average>=19	3.51	students	Appropriate
18	female	18-23	average>=19	3.58	working	Appropriate
19	male	24-29	average>=19	3.31	working	Appropriate
20	female	18-23	average>=19	3.38	working	Appropriate

The following is a calculation of the probability of conditions from the existing training data.

Table 2. Probability of Sample Conditions for Training Data

Probability (Late/Appropriate)	C (Appropriate)= 10/20 = 0.5	C (Late)= 10/20 = 0.5
Probability (Gender)	**Appropriate**	**Late**
L	X | C (Male | Appropriate) = 6/10 = 0.6	X | C (Male | Late) = 9/10 = 0.9
P	X | C (Female | Appropriate) = 4/10=0.4	X | C (Female | Late) = 1/10=0.1
Probability (Age)	**Appropriate**	**Late**
18-23	X | C (18-23 | Appropriate) = 8/10=0.8	X | C (18-23 | Late) = 2/10=0.2
24-29	X | C (24-29 | Appropriate) = 2/10=0.2	X | C (24-29 | Late) = 6/10=0.6
> 29	X | C (> 29 | Appropriate) = 0/10=0	X | C (> 29 | Late) = 2/10=0.2
Probability (Credits)	**Appropriate**	**Late**
Average Credits <19	X | C (Average Credits < 19 | Appropriate) = 1/10=0.1	X | C (Average Credits < 19 | Late) = 0/10=0
Average Credits >=19	X | C (Average Credits >= 19 | Appropriate) = 9/10=0.9	X | C (Average Credits >= 19 | Late) = 10/10=1
Probability (Gpa)	**Appropriate**	**Late**
>= 3.00	X | C (GPA >= 3.00 | Appropriate) = 9/10=0.9	X | C (GPA >= 3.00 | Late) = 5/10=0.5
< 3.00	X | C (GPA < 3.00 | Appropriate) = 1/10=0.1	X | C (GPA < 3.00 | Late) = 5/10=0.5
Probability (Status)	**Appropriate**	**Late**
Students	X | C (Students | Appropriate) = 4/10=0.4	X | C (Students | Late) = 2/10=0.2
Working	X | C (Working | Appropriate) = 6/10=0.6	X | C (Working | Late) = 8/10=0.8

DOI: 10.52362/jisicom.v6i1.785

After knowing the probability of conditions from the training data above, then the data you want to test can be entered. As an example of testing data in the table below.

Table 3. Examples of Data Testing

Gender	Age	Credits	GPA	Status	Graduation
Male	24-29	Average >=19	< 3.00	Working	Late
Male	18-23	Average >=19	>= 3.00	Students	Appropriate
Female	24-29	Average >=19	>= 3.00	Students	Appropriate
Male	24-29	Average >=19	>= 3.00	Working	Late
Male	24-29	Average >=19	>= 3.00	Students	Appropriate
Female	24-29	Average >=19	>= 3.00	Working	Appropriate
Male	>29	Average >=19	>= 3.00	Working	Late

Based on the calculation of the probability of conditions, the testing data can be tested, as in the table below.

Table 4. Probability of Prediction Example Data Testing

Appropriate	Late	Result	Prediction
= 0.5 x 0.6 x 0.2 x 0.9 x 0.1 x 0.6 = 0.0032	= 0.5 x 0.9 x 0.6 x 1 x 0.5 x 0.8 = 0.108	P (Late) > P (Appropriate).	Late
= 0.5 x 0.6 x 0.8 x 0.9 x 0.9 x 0.4 = 0.0778	= 0.5 x 0.9 x 0.2 x 1 x 0.5 x 0.2 = 0.009	P (Appropriate) > P (Late)	Appropriate
= 0.5 x 0.6 x 0.2 x 0.9 x 0.9 x 0.4 = 0.013	= 0.5 x 0.1 x 0.6 x 1 x 0.5 x 0.2 = 0.003	P (Appropriate) > P (Late)	Appropriate
= 0.5 x 0.6 x 0.2 x 0.9 x 0.9 x 0.6 = 0.0292	= 0.5 x 0.9 x 0.6 x 1 x 0.5 x 0.8 = 0.108	P (Late) > P (Appropriate)	Late
= 0.5 x 0.6 x 0.2 x 0.9 x 0.9 x 0.4 = 0.0194	= 0.5 x 0.9 x 0.6 x 1 x 0.5 x 0.2 = 0.027	P (Late) > P (Appropriate)	Late
= 0.5 x 0.4 x 0.2 x 0.9 x 0.9 x 0.6 = 0.0194	= 0.5 x 0.1 x 0.6 x 1 x 0.5 x 0.8 = 0.012	P (Appropriate) > P (Late)	Appropriate
= 0.5 x 0.6 x 0 x 0.9 x 0.9 x 0.6 = 0	= 0.5 x 0.9 x 0.2 x 1 x 0.5 x 0.8 = 0.036	P (Late) > P (Appropriate)	Late

DOI: 10.52362/jisicom.v6i1.785

After calculating the Naïve Bayes algorithm, the results will be seen in the confusion table. Calculating the level of data accuracy aims to find out how the system works. The results of this calculation are then tested for the accuracy of the data using confusion matrix. Confusion matrix is a method commonly used to perform calculations in data accuracy in data mining.

According to [19] confusion matrix is a useful tool for analyzing how well a classifier can recognize tuples from various classes. A confusion matrix is a table used to describe the performance of the classification model on a test dataset whose true value is known.

Confusion matrix visualizes classifier accuracy by comparing actual classes with predictions [11]. A confusion matrix is a method used to calculate the accuracy of the data mining concept. Confusion matrix aims to calculate the accuracy of each test [8].

Tabel 5. Confusion matrix

		Predictions	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FN	TN

Information:
• TP (True Positive): the number of data whose actual class is a positive class with the prediction class being a positive class.
• FN (False Negative): the amount of data whose actual class is a positive class and the prediction class is a negative class.
• FP (False Positive): the number of data whose actual class is a negative class with the prediction class being a positive class.
• TN (True Negative): the amount of data whose actual class is a negative class with the prediction class being a negative class.

From the confusion matrix above, a measurement matrix can be made to obtain accuracy, precision, and recall values.

a. Accuracy is a test method based on the level of closeness between predictions and actual values, to find out the amount of data classified correctly, the accuracy of the prediction results can be seen.

$$Accuracy = \frac{\mathbf{TP+TN}}{\mathbf{TP+FN+FP+TN}}\, x\ \mathbf{100\%} \tag{2}$$

b. Precision is a test method by comparing the amount of relevant information obtained by the system with the total amount of information taken by the system, both relevant and not.

$$Precision\ = \frac{\mathbf{TP}}{\mathbf{TP+FP}}\, x\ \mathbf{100\%} \tag{3}$$

DOI: 10.52362/jisicom.v6i1.785

c. Recall is a test method that compares the amount of relevant information obtained from the system with the total number of relevant information in the information collection (whether taken or not by the system).

$$\boldsymbol{Recall} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \, \boldsymbol{x} \ \mathbf{100\%} \quad (4)$$

After getting the predictions from the data contained in Table 5, a confusion table can be created.

Tabel 5. Confusion Matrix

	Class	
Predicted ↓	Appropriate	Late
Appropriate	3	0
Late	1	3

From the confusion table above, it can be calculated using the confusion matrix method to determine the accuracy, precision, and recall of the sample data.

1. Accuracy = $\frac{3+3}{7}$ x 100% = 86 %
2. Precision = $\frac{3}{3+0}$ x 100% = 100%
3. Recall = $\frac{3}{3+1}$ x 100% = 75%

From this calculation, the accuracy result is 86%, precision is 100%, and the recall is 75%.

4. Experiment Report

The experiments are conducted using a computing platform based on Intel Core i5 2.2 GHz CPU, 8 GB RAM, and Microsoft Windows 10 Professional 64-bit with SP1 operating system. The development environment is Sublime Text 3 and RapidMiner 9.8 library.

Table 6. Calculation Results of Training Data

Probability (Appropriate/Late)	C (Appropriate)= 15/41 = 0.366	C (Late)= 26/41 = 0.634
Probability (Gender)	**Appropriate**	**Late**
Male	X | C (Male | Appropriate) = 9/15 = 0.6	X | C (Male | Late) = 21/26 = 0.808
Female	X | C (Female | Appropriate) = 6/15 = 0.4	X | C (Female | Late) = 5/26 = 0.192
Probability (Age)	**Appropriate**	**Late**

DOI: 10.52362/jisicom.v6i1.785

18-23	X | C (18-23 | Appropriate) = 8/15 = 0.533	X | C (18-23 | Late) = 4/26 = 0.154
24-29	X | C (24-29 | Appropriate) = 7/15 = 0.467	X | C (24-29 | Late) = 16/26 = 0.615
> 29	X | C (> 29 | Appropriate) = 0/15 = 0	X | C (> 29 | Late) = 6/26 = 0.231
Probability (Creadits)	**Appropriate**	**Late**
Average Creadits <19	X | C (Average Creadits < 19 | Appropriate) = 2/15 = 0.133	X | C (Average Creadits < 19 | Late) = 1/26 = 0.038
Average Creadits >=19	X | C (Average Creadits >= 19 | Appropriate) = 13/15 = 0.867	X | C (Average Creadits >= 19 | Late) = 25/26 = 0.962
Probability (Gpa)	**Appropriate**	**Late**
>= 3.00	X | C (GPA >= 3.00 | Appropriate) = 15/15 = 1	X | C (GPA >= 3.00 | Late) = 17/26 = 0.654

Based on the results of the calculation of the training data above, it will produce a probability condition for all of the correct passing of 0.366 and being late in the amount of 0.634. The results of the probability of passing the Male Gender Correct condition are 0.6 and Late is 0.808, for the Correct Female passing the number is 0.4 and Late being 0.192.

The results of the probability of passing the correct age 18-23 age are 0.533 and 0.154 late, for the correct 24-29 passing the number of 0.467 and 0.615 for the correct pass over 29, 0 and 0.231 late. The results of the probability of passing conditions for the correct SKS mean <19 are 0.133 and the overdue is 0.038, for the correct passing average SKS> = 19 are 0.867 and overdue is 0.962.

The results of the probability of passing the GPA> = 3.00, which are correct in the amount of 1.00 and late in the amount of 0.654, for the passing of <3.00, the correct number is 0 and the number is 0.346. The results of the probability of passing the Right Student Student Status Status are 0.133 and 0.154 late, for the right graduation is 0.867 and 0.846 is late.

The results of the calculation of data testing using the Naïve Bayes algorithm, are in Table 7

Table 7. Calculation Results of Testing Data

No	Appropriate	Late	Result	Prediction
1	= 0.366 x 0.6 x 0.467 x 0.867 x 1 x 0.867	= 0.634 x 0.808 x 0.615 x 0.962 x 0.654 x 0.846	P (Late) > P (Appropriate)	Late

	= 0.0769	= 0.1677		
2	= 0.366 x 0.6 x 0.467 x 0.867 x 0 x 0.867 = 0	= 0.634 x 0.808 x 0.615 x 0.962 x 0.346 x 0.846 = 0.0888	P (Late) > P (Appropriate)	Late
3	= 0.366 x 0.6 x 0 x 0.867 x 0 x 0.867 = 0	= 0.634 x 0.808 x 0.231 x 0.962 x 0.346 x 0.846 = 0.0333	P (Late) > P (Appropriate)	Late
4	= 0.366 x 0.6 x 0 x 0.867 x 1 x 0.867 = 0	= 0.634 x 0.808 x 0.231 x 0.654 x 0.846 = 0.0629	P (Late) > P (Appropriate)	Late
5	= 0.366 x 0.6 x 0.533 x 0.867 x 1 x 0.133 = 0.0135	= 0.634 x 0.808 x 0.154 x 0.962 x 0.654 x 0.154 = 0.0076	P (Appropriate) > P (Late)	Appropriate
6	= 0.366 x 0.4 x 0.533 x 0.867 x 1 x 0.133 = 0.009	= 0.634 x 0.192 x 0.154 x 0.962 x 0.654 x 0.154 = 0.0018	P (Appropriate) > P (Late)	Appropriate
7	= 0.366 x 0.6 x 0.467 x 0.867 x 1 x 0.867 = 0.0769	= 0.634 x 0.808 x 0.615 x 0.962 x 0.654 x 0.846 = 0.1677	P (Late) > P (Appropriate)	Late
8	= 0.366 x 0.6 x 0.467 x 0.867 x 0 x 0.867 = 0	= 0.634 x 0.808 x 0.615 x 0.962 x 0.346 x 0.846 = 0.0888	P (Late) > P (Appropriate)	Late
9	= 0.366 x 0.6 x 0.467 x 0.867 x 0 x 0.867 = 0	= 0.634 x 0.808 x 0.615 x 0.962 x 0.346 x 0.846 = 0.0888	P (Late) > P (Appropriate)	Late
10	= 0.366 x 0.6 x 0.467 x 0.867 x 0 x 0.867 = 0	= 0.634 x 0.808 x 0.615 x 0.962 x 0.346 x 0.846 = 0.0888	P (Late) > P (Appropriate)	Late
25	……………..	…………………	………………..	……………………

The results of the prediction probability of passing Data 1 that are Right are 0.0769 and Late are 0.1677, then Probability (Late) < Probability (Exact) so Data 1 is predicted to pass Late. The results of the prediction probability of passing Data 2 that are Right are 0 and Late are 0.0888, then Probability (Late) < Probability (Correct) so Data 2 is predicted to pass Late. The results of the Probability of Passing Data 3 that are Right are 0 and Late are 0.0333, then Probability (Late) < Probability (Correct) so Data 3 is predicted to pass Late.

The results of the prediction probability of passing Data 4 that are Right are 0 and Late are 0.0629, then Probability (Late) < Probability (Exact) so Data 4 is predicted to pass Late. The results of the prediction probability of passing Data 5 that are Right are 0.0135 and Late are 0.0076, then Probability (Correct) < Probability (Late) so Data 5 is predicted to pass Exactly. The results of the prediction probability of passing Data 6 that are Right are 0.009 and too late are 0.0018, then Probability (Correct) < Probability (Late) so Data 6 is predicted to pass Exactly.

The results of the Probability Prediction of Passing Data 7 that are Right are 0.0769 and Late are 0.1677, then Probability (Late) < Probability (Correct) so Data 7 is Predicted to Pass Late. The results of the prediction probabilities of passing Data 8, Data 9, and Data 10 that are correct are 0 and late are 0.0888, so Probability (Late) < Probability (Correct) so Data 8, Data 9, and Data 10 are predicted to pass Late.

Based on the calculation of the testing data above, there are 10 student data that are predicted to pass right and 15 student data are predicted to pass late, based on the 25 student data tested.

Based on the results of the calculation of training data and testing data, the results of calculations using a confusion matrix are obtained to determine the accuracy, recall, and precision of the calculated data, which can be seen in table 8.

Tabel 8. Confusion Matrix

	Class	
Predicted ↓	**Appropriate**	**Late**
Appropriate	10	0
Late	1	14

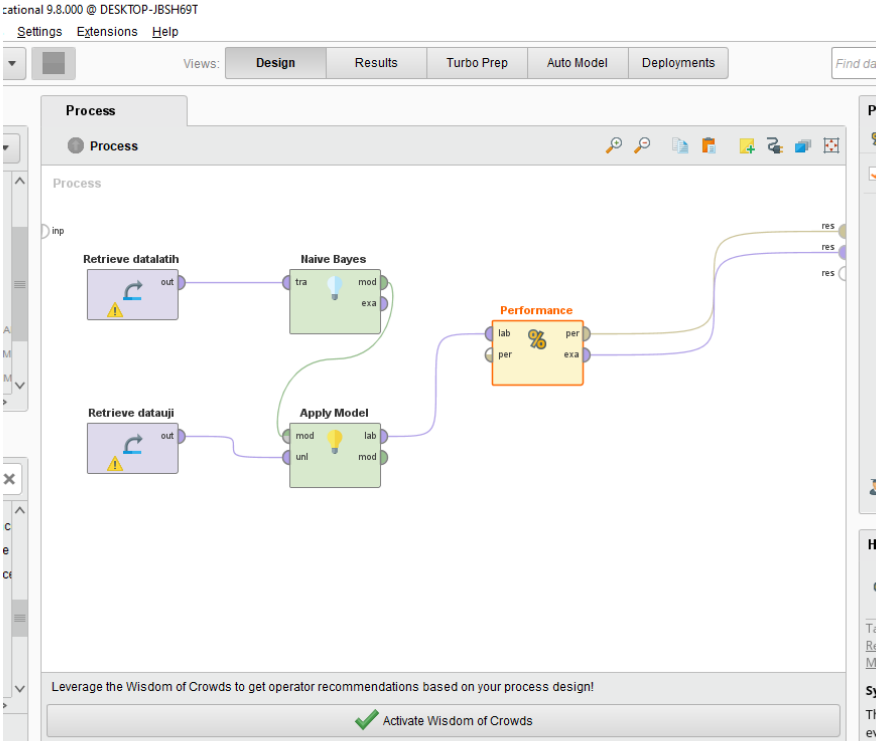
The known amount is then calculated for accuracy, recall and precision in the calculations below:

Tabel 9. Confusion Matrix Calculation

	Calculation	**Result**
Accuracy	$= \frac{10+14}{10+1+0+14}$ x 100%	96%
Recall	$= \frac{10}{10+1}$ x 100%	90,91%
Precision	$= \frac{10}{10+0}$ x 100%	100%

This research was conducted using RapidMiner data mining software. The assessment will be done by looking at the accuracy produced, the higher the percentage of the classification results, the higher the accuracy of the method used.

DOI: 10.52362/jisicom.v6i1.785



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 3. RapidMiner Main Process Settings

After testing the accuracy, recall, and precision values of the Naïve Bayes algorithm using RapidMiner software, it produces 96% accuracy values, 90.91% recall, and 100% precision. This result is the same value as the calculation performed manually. This proves that the Naïve Bayes algorithm can be applied to predict the graduation rate of Information Systems students at The Universitas Dirgantara Marsekal Suryadarma.

5. Conclusion

Based on the results of calculations using the Naïve Bayes Algorithm and proven using the RapidMiner application using testing data taken as much as 60% of the training data managed to get 96% accuracy, 90.91% recall, and 100% precision. These results are obtained from testing data taken as much as 60% of the training data. So it can be concluded that the application of the Naïve Bayes Algorithm can be used to determine the prediction of the punctuality of graduation rates for Information Systems students at The Universitas Dirgantara Marsekal Suryadarma. Show us that the proposed method achieved higher classification accuracy. Therefore, we can conclude that proposed method makes an improvement in neural network prediction performance.

References

[1] M. Broto Legowo and B. Indiarto, "Model Sistem Penjaminan Mutu Berbasis Integrasi Standar Akreditasi BAN-PT dan ISO 9001:2008," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 2, p. 90, 2017.

[2] A. I. Adekitan and O. Salau, “The impact of engineering students’ performance in the first three years on their graduation result using educational data mining,” *Heliyon*, vol. 5, no. 2, p. e01250, 2019.
[3] M. K. Diqy Fakhrun Shiddieq, S.T. and Patricia, “Implementasi Algoritma Data Mining Naive Bayes untuk Prediksi Kelulusan Mahasiswa,” no. 456, pp. 8–13, 2020.
[4] A. Syarifah and M. A. Muslim, “Pemanfaatan Naïve Bayes Untuk Merespon Emosi Dari Kalimat Berbahasa Indonesia,” *Unnes J. Math.*, vol. 4, no. 2, 2015.
[5] I. . K. S. Putu Sainanda Cahyani Moonallika, Ketut Queena Fredlina, “Penerapan Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus STMIK Primakara),” *J. Ilm. Komput.*, vol. 6, no. 1, pp. 47–56, 2020.
[6] A. Saleh, “Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga,” *Creat. Inf. Technol. J.*, vol. 2, no. 3, pp. 207–217, 2015.
[7] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*, vol. 53, no. 9. 2008.
[8] Firman Azhar Riyadi, “Implementasi Metode naive Bayes Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu Prodi Informatika (Studi Kasus : Universitas Teknologi Yogyakarta),” pp. 1–9, 2020.
[9] T. Daniel, *Uncovering Patterns in Student Work*. 2015.
[10] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, “Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review,” *Fakt. Exacta*, vol. 13, no. 1, pp. 38–47, 2020.
[11] E. Sutoyo and A. Almaarif, “Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 95–101, 2020.
[12] J. Han and M. Kamber, *Data mining: Data mining concepts and techniques*. 2006.
[13] K. Sihotang and R. Ghaniy, “Penerapan Metode Naïve Bayes Classifier Untuk Penentuan Topik Tugas Akhir Pada Website Perpustakaan STIKOM Binaniaga,” vol. 9, no. 2009, pp. 63–72, 2019.
[14] E. Karyadiputra, S. Kom, and M. Kom, “ANALISIS ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI STATUS KESEJAHTERAAN RUMAH TANGGA KELUARGA BINAAN SOSIAL,” vol. 7, no. 4, pp. 199–208, 2016.
[15] P. Phoenix, R. Sudaryono, and D. Suhartono, “ScienceDirect ScienceDirect Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier,” *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 498–506, 2021.
[16] V. Balakrishnan and W. Kaur, “ScienceDirect ScienceDirect String-based Multinomial Naïve Bayes for Emotion Detection String-based Multinomial Naïve Bayes for Emotion Detection among Facebook Diabetes Community among Facebook Diabetes Community,” *Procedia Comput. Sci.*, vol. 159, pp. 30–37, 2019.
[17] C. Cassidy, “Parameter tuning Naïve Bayes for automatic patent classification ☆,” *World Pat. Inf.*, vol. 61, no. June 2019, p. 101968, 2020.
[18] M. Awaludin, “Penerapan Sistem Piranti Lunak Personal Finance Berbasis Android untuk Peningkatkan Kualitas Ekonomi Individu,” *J. Sist. Inf. Univ. Suryadarma*, vol. 3, no. 2, pp. 107–114, 2018.
[19] J. Han, M. Kamber, and J. Pei, *Data mining: Data mining concepts and techniques (Third Edition)*. 2012.
[20] S. Tufféry, *Data Mining and Statistics for Decision Making*. 2011.
[21] A. R. Febie Elfaladonna, “Analisa Metode Classification-Decission Tree Dan Algoritma C.45 Untuk Memprediksi Penyakit Diabetes dengan Menggunakan Aplikasi Rapid Miner,” *Sci. Inf.*

DOI: 10.52362/jisicom.v6i1.785

Technol., vol. 2, no. 1, pp. 10–17, 2019.

[22] D. Iskandar and Y. K. Suprapto, “Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan Antara Algoritma C 4.5 dan Naive Bayes,” *J. Ilm. NERO*, vol. 2, no. 1, pp. 37–43, 2015.

[23] D. Sartika and D. Indra, “Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian,” *J. Tek. Inform. Dan Sist. Inf.*, vol. 1, no. 2, pp. 151–161, 2017.

[24] P. Harmianty, “Aplikasi Prediksi Kelulusan Tepat Waktu Mahasiswa Menggunakan Algoritma C4.5,” pp. 1–9, 2017.

[25] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Third Edition*. Elsevier Inc., 2011.

[26] Syarli and A. A. Muin, “Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi),” *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 22–26, 2016.

DOI: 10.52362/jisicom.v6i1.785