Naive Bayes Classifier Method Analysis and Support Vector Machine (SVM) Student Graduation Prediction

Siti Mukodimah[1] , Muhamad Muslihudin[2] , Dwi Rohmadi Mustofa[3], Didi Susianto[4], Suyono[5]
[1,2,4,5] Institut Bakti Nusantara, Lampung, Indonesia
[3] STIT Pringsewu, Lampung, Indonesia
Corresponding author e-mail:

Abstract:

Graduating on time is the dream of every student, but the reality is not as expected. Many students graduate after more than four years. Based on data, the Lampung region has a total of 84 institutions with 24,216 new students per year and 19,486 graduates per year. The number of college graduates only reached 54.96 % of the number of new students enrolled each year in universities. This means that the number of students entering and leaving the university has not been balanced every year so that there is an accumulation of students in each graduation period. With this research can identify the causes of delays in graduation students. Prediction of student graduation based on predetermined parameters can help classify student data, thus helping students who are indicated to be not on time to complete their studies. In this study, two models were used to predict student graduation, namely the Naïve Bayes Classifier and SVM. Both methods are classification methods. The Naïve Bayes method is an algorithm that is simple, fast and has a high level of accuracy, while the SVM method is a method that is able to identify separate Hyperplanes that maximize the margin between two different classes. Based on observations and research studies that have been carried out at STMIK Pringsewu have not led to the identification of the causes of student delays in completing studies, so it is necessary to predict student graduation using the right classification method to describe the class on time of graduation. The results of this study are taken into consideration by the leadership in making internal higher education policies in an effort to balance the ratio of student enrollment and student graduation at STMIK Pringsewu.

Keywords: Prediction, Punctuality, Graduation, Classification Method

DOI Number: 10.14704/NQ.2022.20.12.NQ77360 **NeuroQuantology2022;20(12): 3522-3533**

1. INTRODUCTION

Higher Education is an institution providing the final stage of academic education in formal education. Universities that provide higher education can take the form of academies, polytechnics, high schools, institutes, or universities. Universities must have quality and governance capabilities according to the standards set by BAN-PT so that they can be publicly recognized or accredited. Based on data from the Central Statistics Agency (BPS) in 2019, the Lampung Region has a total of 84 institutions with 24,216 new students per year and 19,486 graduates annually. The number of college graduates only reached 54.96 % of the number of new students enrolled each year in universities. This means that the number of students entering and leaving the university has not been balanced every year so that there is an accumulation of students in each graduation period.

Research that has been carried out by Abdul Rohman (2019) by comparing data mining, namely neural network algorithms, k-nearest neighbors and decision trees to predict student graduation by using student graduation data as a dataset. The results of the test by measuring the three algorithms using cross validation, confusion matrix, and ROC curve, it is known that the neural network has the highest accuracy value of 87.32%, then the decision tree algorithm method is 83.57%, then the k-algorithm method, nearest neighbor with 83.66% accuracy. The AUC value for the neural network method shows the highest value, which is 0.917 and the lowest is the Decision Tree method, which is 0.844 [1] . Ratna Puspita (2018) in this study applied the C4.5 algorithm to the student graduation prediction application. This study utilizes student graduate data and processes it using data mining to obtain information in the form of student graduation predictions. The method that will be used is the decision tree method built with the C4.5 algorithm accompanied by an error-based pruning algorithm for the decision tree cutting process. The criteria that will be used are gender, regional origin, GPA, and TOEFL. In its application, the C4.5 algorithm can be used to generate graduation predictions with an average precision of 63.93%, recall of 60.73%, and accuracy of 60.52%. After the decision tree is cut using the error-based pruning method, better results are obtained. Trees cut using a confidence value of 0.4 resulted in a precision of 70.70%, recall of 50.65%, and accuracy of 61.57%. Meanwhile, trees that were cut using a confidence value of 0.25 resulted in a precision of 73.77%, recall of 48.84%, and accuracy of 62.44% [2] .

Based on two studies that have been done previously, predictions of student graduation are carried out using comparative data mining methods, namely neural network, decision tree, C4.5 algorithm, and k-nearest neighbor, while in this study, predictions of student graduation will be carried out using two data mining methods. Namely the Bayesian classifier and SVM using data from STMIK Pringsewu graduate students as a dataset that has never been done before.

The punctuality of student graduation is an interesting topic discussed in every university. This also happens at STMIK Pringsewu, where there is an imbalance between the number of graduates each year and the number of new students enrolled each year. Therefore, the authors are interested in raising the topic of the punctuality of graduates at STMIK Pringsewu by applying the *Bayes classifier* method and the *SVM method* which will be used as methods to predict the timeliness of graduates in completing studies. With this research, it is hoped that it can help students who are indicated to be not on time to complete their studies, and help improve the quality of STMIK Pringsewu tertiary institutions.

2. LITERATURE REVIEW

2.1 Previous Research

To support research and strengthen literature reviews and interesting state of the art In this study, the results of previous studies relevant to this study were compiled, namely Lila Setiyani. This study aims to provide a systematic literature review on data mining methods in predicting the number of graduations on time. The results of this study obtained the level of accuracy of the method used to predict student graduation on time by taking into account the attributes of the database used, while the accuracy of the three literatures produced an accuracy above 90% with the number of attributes and different applications for testing. There is one attribute in common from the three literatures, namely the GPA attribute. Christin Nandari Dengen [3] the results of this expert system can display the level of accuracy of the use of decission trees on student graduation predictions, the accuracy obtained is 60%. Ryan Dwi Pambudi [3] The results of this study are a pattern based on the probability of the attribute used to predict graduation. In addition, the development of a website-based dashboard visualization to predict graduation is carried out by using Weka CLI as an API. Based on several references, the research studies that were carried out were all focused and oriented on the classification method to predict the punctuality of graduation, so research was needed to examine the identification of factors causing the imprecise time of graduation for students. This research was conducted by focusing on predicting student graduation times by combining the Naïve Bayes Algorithm Model, and the SVM algorithm as a classification algorithm which will help make it easier for universities to identify and classify students who are indicated to graduate not on time.

Theoretical Framework

2.2.1 Data Mining Concept

According to Daniel T. Larose (2004) Data Mining is the process of finding meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technology as well as statistical and mathematical techniques [4] [5] . According to Kusrini (2009) the terms data mining and *knowledge discovery in databases* are often used interchangeably to describe the process of extracting hidden information in a large database. Understanding the two terms have different concepts, but are related to each other. One of the stages in the whole process of knowledge discovery in databases is data mining . *Knowledge discovery in the database* can be broadly described as follows.

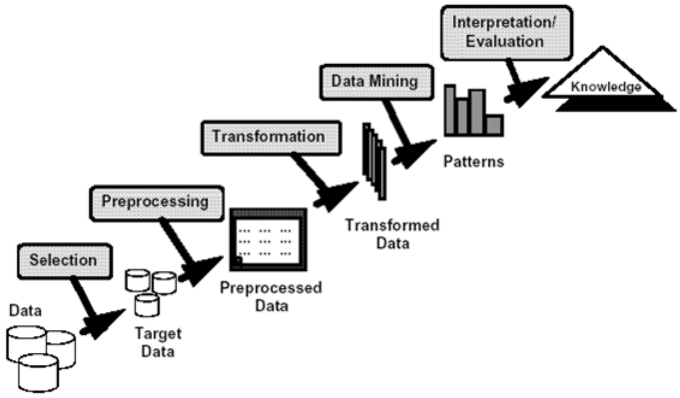


Figure 1. Stages *of Knowledge Discovery in Databases* [6]

Data mining is divided into several groups based on the tasks that can be done, namely [4] :

1. Description
   Sometimes analytical research simply wants to try to find a way to describe the patterns and trends contained in the data. Descriptions of patterns and trends often provide possible explanations for a pattern or trend.
2. Estimate
   Estimation is almost the same as classification, except that the estimation target variable is more numerical than categorical. The model is built using a complete record that provides the value of the target variable as the predicted value. Furthermore, in the next review, the estimated value of the target variable is made based on the value of the predictive variable.
3. Prediction
   Prediction has similarities with estimation and classification. It's just that, the prediction of the result shows something that hasn't happened yet (may happen in the future).
4. Classification
   In the classification of variables, objectives are categorical. For example, we will classify income into three classes, namely high income, medium income, and low income.
5. *Clustering*

Clustering is a grouping of records, observations, or attention and form a class of objects that have similarities. *Cluster* is a collection of *records* that have similarities with one another and have dissimilarities with *records* in other *clusters*. Clustering is different from classification in that there is no target variable in the clustering.

6. Association
   Identify the relationship between various events that occur at one time

2.2.2 Student Graduation

Based on the Regulation of the National Accreditation Board for Higher Education (BAN-PT) Number 3 of 2019 concerning Higher Education Accreditation Instruments, it is stated that one of the indicators of higher education performance is the presentation of student graduation on time. Based on data obtained from BPS, the number of university graduates in the Lapung Region only reached 54.96 % of the number of new students enrolled each year. The data shows that the number of students graduating on time for college every year in the Lampung area has not been balanced with the number of new students enrolled each year. Graduation of students at STMIK Pringsewu itself is not balanced with the number of registered students, the presence of students who do not graduate on time each period causes the accumulation of students with unclear status. There are two factors that hinder student graduation, namely internal factors such as laziness, self-motivation, IQ and external factors such as parental income, and GPA scores.

2.2.3 STMIK Pringsewu

Pringsewu College of Informatics and Computer Management (STMIK) was established on March 17, 2008 by the STARTECH foundation based on the Decree of the Minister of Education and Culture No. 41/D/O/2008. STMIK Pringsewu has won the trust of the government and the community in the implementation of higher education because it is able to produce quality human resources and are ready to work. Since its establishment in 2008, STMIK Pringsewu has grown rapidly to date. This is indicated by the number of students who continue to increase significantly every year. Judging from the strategic location of Pringsewu district on the West Cross route which is one of the busiest routes in Lampung province to a number of provinces on the west coast of Sumatra, Pringsewu district has a very potential position for the development of the education sector. STMIK Pringsewu has considerable potential to develop because it is located close to the city of Bandar Lampung and is the first computer college in the district. All planning and activities of STMIK Pringsewu are based on core objectives related to education, research, and community service concerns, with the hope that STMIK Pringsewu will become one of the leading universities in Indonesia.

3. RESEARCH METHODS

3.1. Method of collecting data

The data collection method is an important thing in research and is a strategy or method used by researchers in collecting the data needed in their research. Data collection methods used in this study are:

a. Observation
   The data and information collection technique carried out during the field study was direct *observation*, namely data collection carried out by the author during direct observation of the data of STMIK Pringsewu students.

b. Literature review
   The literature review is carried out by reading, citing and making notes sourced from library materials that support and relate to research in

this case regarding data mining Naïve Baye's classifier and SVM.

c. Documentation
Data collection is done by looking at the campus database in the Academic section, the data used is 258 student data from the 2019 batch.

3.2. Naïve Bayes Classifier Method

Naïve Bayes is a simple probabilistic classifier that calculates a set of probabilities by adding up the frequencies and combinations of values from a given dataset. [7] [8]

Naïve Bayes is based on the simplifying assumption that attribute values are conditionally independent if given an output value. In other words, given the output value, the probability of observing together is the product of individual probabilities (Ridwan 2013) [9] [10] . The advantage of using Naive Bayes is that this method only requires a small amount of training data to determine the parameter estimates needed in the classification process. Naive Bayes often performs much better in most complex real-world situations than expected (Pattekari and Parveen 2012) [11] .

The Naive Bayes Classifier is considered to work very well compared to other classifier models, namely the Naive Bayes Classifier has a better accuracy rate than other classifier models (Xhemali 2009). *Naive Bayes Classifier is* included in *supervised learning*, *Naive Bayes estimates conditional* class opportunities by assuming that the attribute is conditionally independent given the label y. The conditional independent assumption can be expressed in the following form (Suyanto , 2017). The calculation of *Naive Bayes with the following* equation .

$$\mathrm{P(H|X)} = \frac{\mathrm{P(H|X)}P(H)}{\mathrm{P(X)}} \quad (3\ .1\ )$$

Information:

X = Data with unknown class (proof)
H = Hypothesis data X is a class specification
P(H|X) = Probability of hypothesis H true for condition X (posterior prob.)
P(H) = Hypothesis probability H (prior prob.)
PX = Probability prior to proof X

Classification with Naive Bayes works based on probability theory which views all data as evidence in probability. This gives the characteristics of Naive Bayes as follows . (Indrawan, 2017).

1. The Naïve *Bayes method is robust* to isolated data which is usually data with different characteristics (*outliner*). Naive Bayes can also handle incorrect attribute values by ignoring training data during the model building and prediction process.
2. Resistant to irrelevant attributes.
3. A bias correlation degrade the performance of the *Naïve Bayes classification* because the assumption of the independence of the attribute does not exist.

The Naive Bayes Algorithm has the advantage that it is relatively easy to implement because it does not use numerical optimization, matrix calculations and others, Efficient in training and use, Can use *binary* or *polynomial data*, because it is assumed to be independent, it allows this method to be implemented in various ways. Dataset, relatively high accuracy. The Naive Bayes algorithm also has drawbacks, namely the estimation of the possibility of an inaccurate class and the threshold or *threshold* must be determined manually and not analytically.

3.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) was developed by Boser, Guyon, Vapnik. SVM is one of the best methods that can be used in pattern classification problems. The concept of SVM stems from the problem of classifying two classes such as positive and negative training sets. SVM tries to find the best hyperplane (separator) to separate into two classes and maximize the margin between the two classes [12].

The SVM concept can be explained simply as an attempt to find the best hyperplane that functions as a separator of two classes in the input space. Patterns that are members of two classes: +1 and -1 and share alternative discrimination boundaries. Margin is the distance between the hyperplane and the closest pattern from each class. This closest pattern is called a support vector. Efforts to find the location of this hyperplane is the core of the learning process in SVM [13] .

3.3.1 Linear SVM

Linear SVM is a data classification method where the data separation process is carried out linearly. In the case of classification with linear SVM, let $X_i \in \{x_1, ..., x_n\}$ is the dataset and $y_i \in \{+1, -1\}$ is the class label of the data $X_{i.}$ the thing to do is to look for a dividing line or hyperplane from the second group of classes where there are various alternative hyperplanes. The function used to find the hyperplane is f (w, b) = $x_i . w + b$ were separated into positive and negative groups (vijayakumar 1999). [14]

$X_i . w + b\ 1 \textit{ for } y_i = 1$ (3.1)

$X_i . w + b\ {-1} \textit{ for } y_i = -1$ (3.2)

Information:

X_i : data to i

W : weight of support vector or vector that is perpendicular to the hyperplane

B : bias value

Y_i : data class to i

which is equivalent to:

$y_i (x_i . w + b) – 1\ 0 \textit{ for } I = 1,…,n$ *(3.3)*

Description :

n : number of data

Hyperplane is the one that lies midway between two sets of objects of two classes. The best hyperplane can be found by measuring the largest margin of several alternative hyperplanes. Margin is the shortest distance between two sets of objects of two classes. The margin can be calculated by finding the distance between the two hyperplanes supporting the two classes.

An illustration of SVM with an optimal hyperplane that separates the positive and negative groups can be seen in Figure 2.

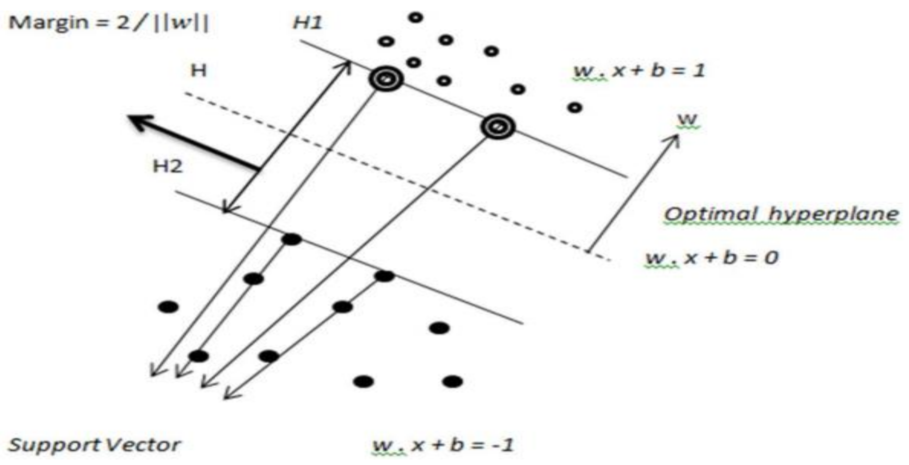


Figure 2. hyperplane optimal illustration

Figure 3.1 shows that H1 is $_a$ supporting hyperplane of class +1 which has the function wx $_1$ + b = +1. While H2 is a supporting hyperplane from class -1 $_{\text{which}}$ has a function wx $_2$ + b = -1.

$$margin = |d_{H1} - d_{H2}| = \frac{2}{\|w\|} \qquad (3.4)$$

Information:

d_{H_1} = class support hyperplane distance +1

d_{H_2} = class support hyperplane distance -1

Then for optimal hyperplane both classes use the following equation:

$$minimize\, J_1\,[w] = \frac{1}{2}\,\|w\|^2 \qquad (3.5)$$

With y $_i$ (x $_i$. w + b)– 1 *0 for I = 1,…,n* (3.6)

3.3.2 Non-Linear SVM

In some cases, the data cannot be classified using the SVM linear method. In classifying data in non-linear form, SVM was developed with kernel functions. Several kinds of SVM kernel functions can be seen in table 3.1. [15] [16]

No	Kernel Name	Function Definition
1	*Linear*	*K (x, y) = x . y*
2	*Polynomial of degree*	*K (x, y) = (x . y)* ^{d}
3	*Polynomial of degree 2*	*K (x, y) = (x . y + c)* ^{d}
4	*Gaussian RBF*	*K (x, y) = exp*$\left(\frac{-\|x-y\|^2}{2_{\sigma^2}}\right)$
5	*Sigmoid (Hyperbolic Tangent)*	*K (x, y) = tanh* $(\sigma(x\,.\,y) + c)$
6	*Inverse Multi Quadratic*	*K (x,y) =*$\frac{1}{\sqrt{\|x-y\|^2+c^2}}$
7	*Additive*	*K (x, y) =* $\sum_{i=1}^{n} K_i\,(x_i,y_i)$

3.4. Research Framework

To provide an overview of the chart of the research process to be carried out, it is necessary to explain the stages of the research as shown in Figure 3 below :

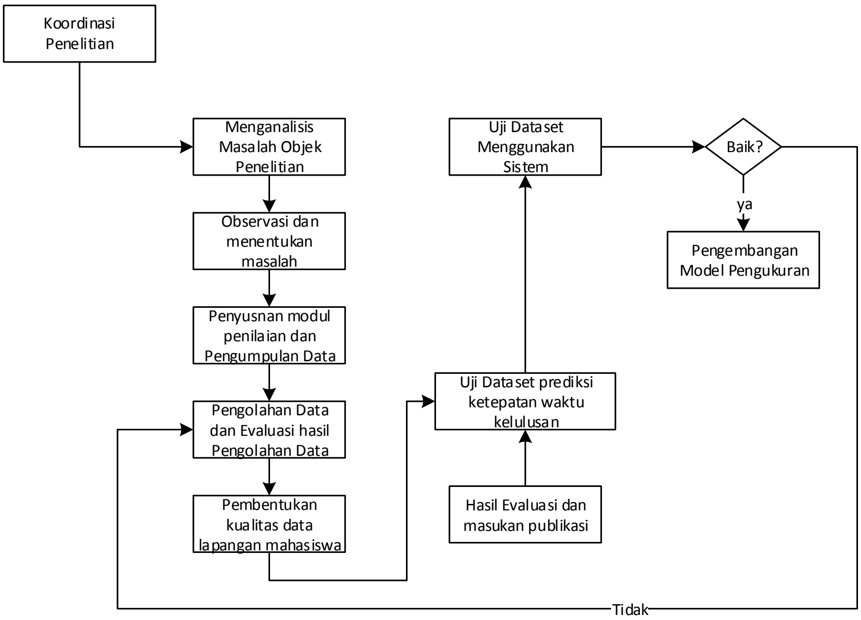

Koordinasi Penelitian
Menganalisis Masalah Objek Penelitian
Observasi dan menentukan masalah
Penyusnan modul penilaian dan Pengumpulan Data
Pengolahan Data dan Evaluasi hasil Pengolahan Data
Pembentukan kualitas data lapangan mahasiswa
Uji Dataset prediksi ketepatan waktu kelulusan
Hasil Evaluasi dan masukan publikasi
Uji Dataset Menggunakan Sistem
Baik?
ya
Pengembangan Model Pengukuran
Tidak

Figure 3. Flow of research stages

4. RESULTS AND DISCUSSION

4.1 Discussion

In this study, the data used amounted to 259 data divided into two for training data and testing data. For training data used 77 data. And the rest of the data is used for testing. For testing, the system will take a number of data randomly from the dataset according to the number of input testing data entered in the form by the user. The training data is calculated using one of the SVM training data completion methods, namely the RBF method and compared with the naive Bayes classifier method. The result of the training is a learning stored in the system which will be a reference for the system to determine an input test data for graduates on time and not on time.

Table 4.1 Test data table

Student name	landfill	IQ TEST	GPA	NEM	Income
Ages Riski Amalia	71.70	85.00	3.41	71.70	3,000,000
Agus Asngari	70.90	75.00	3.75	70.90	2,700,000
Agusta Hendri Yanti	78.50	65.00	3.91	78.50	1,800,000
Angga Rio Sudrajat	70.50	85.00	2.98	70.50	1,600,000
Ariska Arrystantia	68.00	74.00	3.33	68.00	2,500,000
Dedi Suhermanto	74.25	72.00	3.44	74.25	5,000,000
Deni Hidayati	74.25	80.00	3.24	74.25	4,000,000
Destia Mufkholifah	70.50	82.00	3.22	70.50	4,500,000
Eka Lina Sari	83.25	78.00	3.32	83.25	5,000,000
Endang Dewi is sustainable	75.00	75.00	3.65	75.00	2,200,000
Evi Sulismawati	80.25	74.00	3.13	80.25	4,000,000
Farida Hasna	70.50	78.00	3.43	70.50	2,800,000
Febrian Setiawan	80.10	75.00	3.49	80.10	5,000,000
Ferdi Aprianto	0.00	76.00	3.31	0.00	5,000,000
Furusim Marfu'ah	71.25	80.00	3.46	71.25	4,500,000

4.2 Results

After the training and testing data is determined, then the data is processed using the orange application as shown in Figure 4.

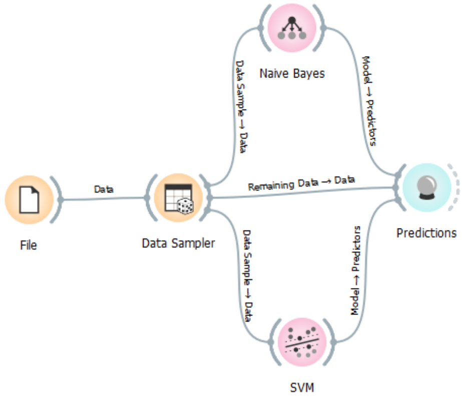
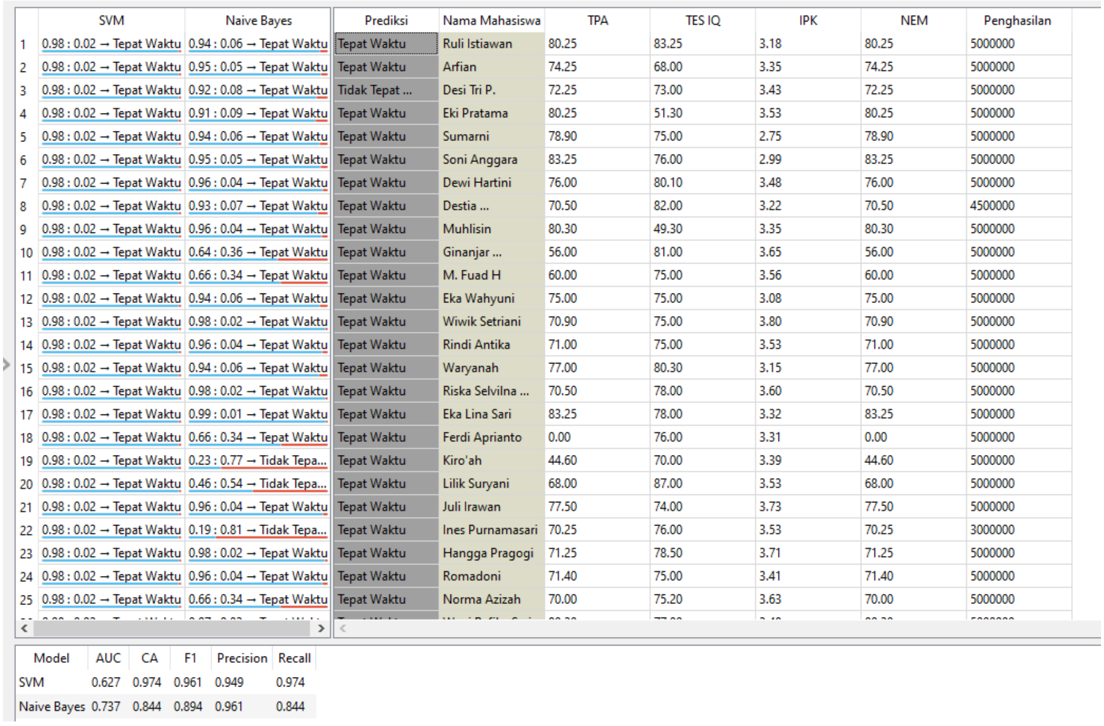


Figure 4. testing data processing using orange

After the prediction process using the model and SVM can be seen in Figure 5. it can be seen that the accuracy of the naive Bayes method is higher than the SVM method. The accuracy level of the naive Bayes method is 0.961 while the SVM is 0.949.



	SVM	Naive Bayes	Prediksi	Nama Mahasiswa	TPA	TES IQ	IPK	NEM	Penghasilan
1	0.98 : 0.02 → Tepat Waktu	0.94 : 0.06 → Tepat Waktu	Tepat Waktu	Ruli Istiawan	80.25	83.25	3.18	80.25	5000000
2	0.98 : 0.02 → Tepat Waktu	0.95 : 0.05 → Tepat Waktu	Tepat Waktu	Arfian	74.25	68.00	3.35	74.25	5000000
3	0.98 : 0.02 → Tepat Waktu	0.92 : 0.08 → Tepat Waktu	Tidak Tepat ...	Desi Tri P.	72.25	73.00	3.43	72.25	5000000
4	0.98 : 0.02 → Tepat Waktu	0.91 : 0.09 → Tepat Waktu	Tepat Waktu	Eki Pratama	80.25	51.30	3.53	80.25	5000000
5	0.98 : 0.02 → Tepat Waktu	0.94 : 0.06 → Tepat Waktu	Tepat Waktu	Sumarni	78.90	75.00	2.75	78.90	5000000
6	0.98 : 0.02 → Tepat Waktu	0.95 : 0.05 → Tepat Waktu	Tepat Waktu	Soni Anggara	83.25	76.00	2.99	83.25	5000000
7	0.98 : 0.02 → Tepat Waktu	0.96 : 0.04 → Tepat Waktu	Tepat Waktu	Dewi Hartini	76.00	80.10	3.48	76.00	5000000
8	0.98 : 0.02 → Tepat Waktu	0.93 : 0.07 → Tepat Waktu	Tepat Waktu	Destia ...	70.50	82.00	3.22	70.50	4500000
9	0.98 : 0.02 → Tepat Waktu	0.96 : 0.04 → Tepat Waktu	Tepat Waktu	Muhlisin	80.30	49.30	3.35	80.30	5000000
10	0.98 : 0.02 → Tepat Waktu	0.64 : 0.36 → Tepat Waktu	Tepat Waktu	Ginanjar ...	56.00	81.00	3.65	56.00	5000000
11	0.98 : 0.02 → Tepat Waktu	0.66 : 0.34 → Tepat Waktu	Tepat Waktu	M. Fuad H	60.00	75.00	3.56	60.00	5000000
12	0.98 : 0.02 → Tepat Waktu	0.94 : 0.06 → Tepat Waktu	Tepat Waktu	Eka Wahyuni	75.00	75.00	3.08	75.00	5000000
13	0.98 : 0.02 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Wiwik Setriani	70.90	75.00	3.80	70.90	5000000
14	0.98 : 0.02 → Tepat Waktu	0.96 : 0.04 → Tepat Waktu	Tepat Waktu	Rindi Antika	71.00	75.00	3.53	71.00	5000000
15	0.98 : 0.02 → Tepat Waktu	0.94 : 0.06 → Tepat Waktu	Tepat Waktu	Waryanah	77.00	80.30	3.15	77.00	5000000
16	0.98 : 0.02 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Riska Selvilna ...	70.50	78.00	3.60	70.50	5000000
17	0.98 : 0.02 → Tepat Waktu	0.99 : 0.01 → Tepat Waktu	Tepat Waktu	Eka Lina Sari	83.25	78.00	3.32	83.25	5000000
18	0.98 : 0.02 → Tepat Waktu	0.66 : 0.34 → Tepat Waktu	Tepat Waktu	Ferdi Aprianto	0.00	76.00	3.31	0.00	5000000
19	0.98 : 0.02 → Tepat Waktu	0.23 : 0.77 → Tidak Tepa...	Tepat Waktu	Kiro'ah	44.60	70.00	3.39	44.60	5000000
20	0.98 : 0.02 → Tepat Waktu	0.46 : 0.54 → Tidak Tepa...	Tepat Waktu	Lilik Suryani	68.00	87.00	3.53	68.00	5000000
21	0.98 : 0.02 → Tepat Waktu	0.96 : 0.04 → Tepat Waktu	Tepat Waktu	Juli Irawan	77.50	74.00	3.73	77.50	5000000
22	0.98 : 0.02 → Tepat Waktu	0.19 : 0.81 → Tidak Tepa...	Tepat Waktu	Ines Purnamasari	70.25	76.00	3.53	70.25	3000000
23	0.98 : 0.02 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Hangga Pragogi	71.25	78.50	3.71	71.25	5000000
24	0.98 : 0.02 → Tepat Waktu	0.96 : 0.04 → Tepat Waktu	Tepat Waktu	Romadoni	71.40	75.00	3.41	71.40	5000000
25	0.98 : 0.02 → Tepat Waktu	0.66 : 0.34 → Tepat Waktu	Tepat Waktu	Norma Azizah	70.00	75.20	3.63	70.00	5000000

Model	AUC	CA	F1	Precision	Recall
SVM	0.627	0.974	0.961	0.949	0.974
Naive Bayes	0.737	0.844	0.894	0.961	0.844

Figure 5. Prediction table

After the training data, the profit testing of the testing data is then carried out as shown in Figure 6.

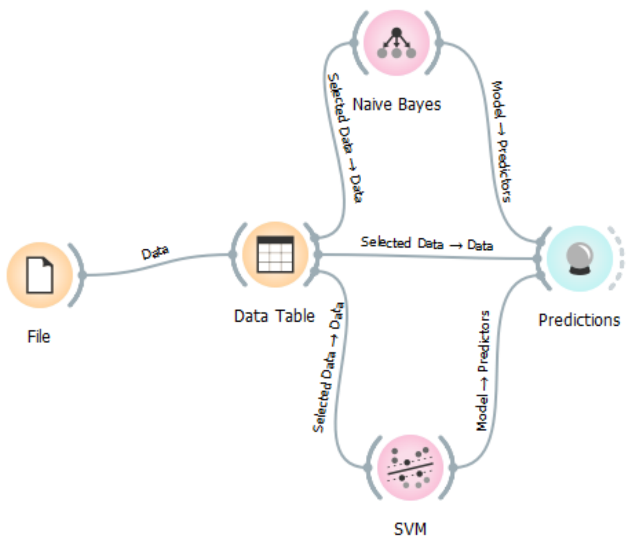
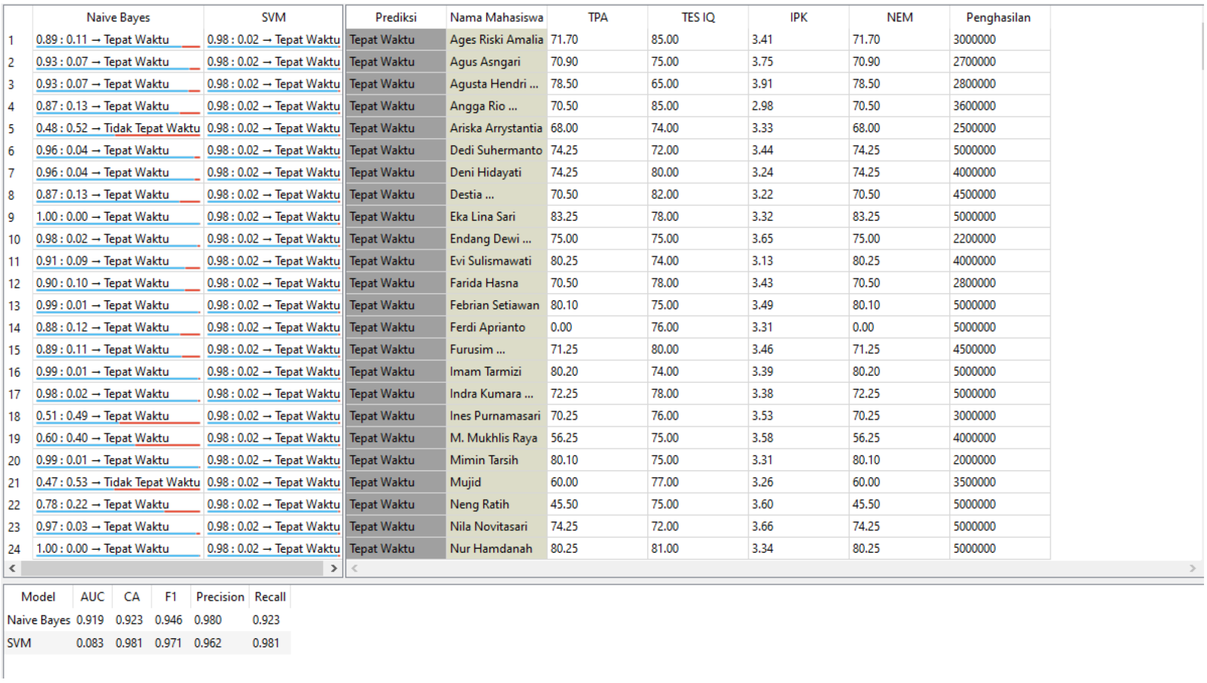


Figure 6. Testing data

After the data testing prediction process has been carried out using the model and SVM, it can be seen in table 4.4. It can be seen that the accuracy of the Naïve Bayes Classifier method is higher than the SVM method. The accuracy of the naive Bayes method reached 0.980 while the SVM was 0.962.



	Naive Bayes	SVM	Prediksi	Nama Mahasiswa	TPA	TES IQ	IPK	NEM	Penghasilan
1	0.89 : 0.11 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Ages Riski Amalia	71.70	85.00	3.41	71.70	3000000
2	0.93 : 0.07 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Agus Asngari	70.90	75.00	3.75	70.90	2700000
3	0.93 : 0.07 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Agusta Hendri ...	78.50	65.00	3.91	78.50	2800000
4	0.87 : 0.13 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Angga Rio ...	70.50	85.00	2.98	70.50	3600000
5	0.48 : 0.52 → Tidak Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Ariska Arrystantia	68.00	74.00	3.33	68.00	2500000
6	0.96 : 0.04 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Dedi Suhermanto	74.25	72.00	3.44	74.25	5000000
7	0.96 : 0.04 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Deni Hidayati	74.25	80.00	3.24	74.25	4000000
8	0.87 : 0.13 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Destia ...	70.50	82.00	3.22	70.50	4500000
9	1.00 : 0.00 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Eka Lina Sari	83.25	78.00	3.32	83.25	5000000
10	0.98 : 0.02 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Endang Dewi ...	75.00	75.00	3.65	75.00	2200000
11	0.91 : 0.09 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Evi Sulismawati	80.25	74.00	3.13	80.25	4000000
12	0.90 : 0.10 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Farida Hasna	70.50	78.00	3.43	70.50	2800000
13	0.99 : 0.01 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Febrian Setiawan	80.10	75.00	3.49	80.10	5000000
14	0.88 : 0.12 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Ferdi Aprianto	0.00	76.00	3.31	0.00	5000000
15	0.89 : 0.11 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Furusim ...	71.25	80.00	3.46	71.25	4500000
16	0.99 : 0.01 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Imam Tarmizi	80.20	74.00	3.39	80.20	5000000
17	0.98 : 0.02 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Indra Kumara ...	72.25	78.00	3.38	72.25	5000000
18	0.51 : 0.49 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Ines Purnamasari	70.25	76.00	3.53	70.25	3000000
19	0.60 : 0.40 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	M. Mukhlis Raya	56.25	75.00	3.58	56.25	4000000
20	0.99 : 0.01 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Mimin Tarsih	80.10	75.00	3.31	80.10	2000000
21	0.47 : 0.53 → Tidak Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Mujid	60.00	77.00	3.26	60.00	3500000
22	0.78 : 0.22 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Neng Ratih	45.50	75.00	3.60	45.50	5000000
23	0.97 : 0.03 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Nila Novitasari	74.25	72.00	3.66	74.25	5000000
24	1.00 : 0.00 → Tepat Waktu	0.98 : 0.02 → Tepat Waktu	Tepat Waktu	Nur Hamdanah	80.25	81.00	3.34	80.25	5000000

Model	AUC	CA	F1	Precision	Recall
Naive Bayes	0.919	0.923	0.946	0.980	0.923
SVM	0.083	0.981	0.971	0.962	0.981

Figure 7. Prediction of testing data

Figure 7 shows a table with AUC (Area Under Curve) values with Naïve Bayes Clasiier of 0.919 and for the SVM model of 0.083. AUC accuracy is said to be perfect if the AUC value reaches 1,000 and poor accuracy if the AUC value is below 0.500. From the data test that has been carried out by training and testing, it can be seen that the level of accuracy of testing of the two models where the naive Bayes classifier model is superior to SVM. The level of accuracy itself is influenced by how much data is used, it is proven that it can be seen from the level of accuracy of the training and testing data test. Where the accuracy of the testing data is higher than the training data.

5 CONCLUSION

The conclusion obtained from this research is that the Naïve Bayes Classifier and Support Vector Machine methods can be used to predict the timeliness of graduate students by learning from the learning generated from training data training. The resulting learning comes from dataset training for graduate students. Where the level of accuracy is higher for the naive Bayes classifier method than the SVM method, from the test data that has been carried out by training and testing it can be seen that the level of accuracy of testing of the two models where the naive Bayes clasifier model is superior to SVM. The level of accuracy itself is influenced by how much data is used, it is proven that it can be seen from the level of accuracy of the training and testing data test. Where the accuracy of the testing data is higher than the training data. while the AUC (Area Under Curve) value with Naïve Bayes Clasifier is 0.919 and for the SVM model is 0.083. AUC accuracy is said to be perfect if the AUC value reaches 1,000 and poor accuracy if the AUC value is below 0.500.

ACKNOWLEDGMENT

This paper has been presented in National Seminar on Technology, Business and Multidisciplinary Research in Yogyakarta, Indonesia, 23 – 24 August 2022. This work is supported by Institut Bakti Nusantara, Lampung, Indonesia. We gratefully appreciate this support.

6 REFERENCES

[1] A. Rohman and M. Rochcham, “Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Mahasiswa,” *Neo Tek.*, vol. 5, no. 1, pp. 23–29, 2019, doi: 10.37760/neoteknika.v5i1.1379.

[2] R. P. S. Putri and I. Waspada, “Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika,” *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 1, p. 1, 2018, doi: 10.23917/khif.v4i1.5975.

[3] R. Dwi, Pambudi, A. Afif, Supianto, and N. Y. Setiawan, “Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Pendekatan Data Mining Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Brawijaya,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. 2196*, vol. 3, no. 3, pp. 2194–2200, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/4655/2154.

[4] D. T. Larose, *Discovering Knowledge in Data An Introduction to Data Mining*. Canada: Published simultaneously in Canada, 2005.

[5] N. Purwati, R. Nurlistiani, N. Purwati, R. Nurlistiani, and O. Devinsen, “DATA MINING DENGAN ALGORITMA NEURAL NETWORK DAN VISUALISASI DATA,” vol. 20, no. 2, 2020.

[6] K. E. T. Lutfi, *Algoritma Data Mining*. Yogyakarta: Penerbit Andi Yogyakarta, 2009.

[7] E. Fitriani, “Perbandingan Algoritma C4.5 Dan Naïve Bayes Untuk Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan,” *Sistemasi*, vol. 9, no. 1, p. 103, 2020, doi: 10.32520/stmsi.v9i1.596.

[8] M. F. Rifai, H. Jatnika, and B. Valentino, “Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Tingkat Kelulusan Peserta Sertifikasi Microsoft Office Specialist (MOS),” *PETIR*, vol. 12, no. 2, pp. 131–144, 2019.

[9] D. P. Pertiwi and R. Anggrainingsih, “Evaluation of Campaign Categories on Kitabisa . Com By Naive Bayes Classifier Method,” *ITSMART J. Ilm. Teknol. dan Inf.*, vol. 8, no. 1, 2019.

[10] H. N. Ahmad, V. Suhartono, and I. N. Dewi, “Penentuan Tingkat Kelulusan Tepat Waktu Mahasiswa Stmik Subang Menggunakan Algoritma C4.5,” *J. Teknol. Inf.*, vol. 13, no. 1, pp. 46–56, 2017.

[11] H. Willa Dhany and F. Izhari, “Analisis Algorithms Support Vector Machine Dengan Naive Bayes Kernel Pada Klasifikasi Data,” vol. 6, pp. 595–598, 2019.

[12] A. Pratama, R. C. Wihandika, and D. E. Ratnawati, *Implementasi Algoritme Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa*, vol. 2, no. March. 2018.

[13] L. Istyfaiyah, M. Wati, D. Mining, and S. V. Machine, “Algoritma Support Vector Machine (SVM) Untuk Klasifikasi Status Kelayakan Keluarga Penerima Bantuan,” vol. 1, no. 1, 2017.

[14] S. Widaningsih, “Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm,” *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.

[15] F. Gorunescu, *Data Mining: Concepts, Model and Techniques*. Verlag Berlin Heidelberg: Springer, 2011.

[16] Y. Kurnia and K. Kusuma, “Comparison of C4.5 Algorithm, Naive Bayes and Support Vector Machine (SVM) in Predicting Customers that Potentially Open Deposits,” *bit-Tech*, vol. 1, no. 2, pp. 40–47, 2018, doi:

10.32877/bt.v1i2.46.

[17] Munawarah, Raudlatul. 2016. Implementasi Metode Support Vector Machine untuk Mendiagnosa Penyakit Hepatitis. Program S-1 Ilmu Komputer, Universitas Lambung Mangkurat: Banjarbaru.

[18] BAN-PT, Buku Matriks Penilaian Instrumen Akreditasi Program Studi Badan Akreditasi Nasional Perguruan Tinggi, Jakarta, 2019.

[19] M. Ridwan, H. Suyono dan M. Sarosa, "Penerapan Data mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," Jurnal EECCIS, vol. 7 No.1, 2013.

[20] H. Romadhona, Agus; suprapedi; himawan, "Prediksi Kelulusan Tepat Waktu Mahasiswa Stmik-Ymi," J. Teknol. Inf., vol. 13, no. 1, pp. 69–83, 2017.

[21] Fadilah, Arif Rahman. 2014. "Analisis dan Perbandingan Metode Support Vector Machine (SVM) dan Naïve Bayes untuk E-mail Spam Filtering". Universitas Telkom Bandung.