International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Prediction of the Waiting Time Period for Getting a Job Using the Naive Bayes Algorithm

Riski Amalia[1], Arief Wibowo[2]
[1,2] Study Program of Informatics Engineering, Faculty of Information Technology, Universitas Budi Luhur Jakarta, Indonesia
Email address: [1]1611500438 @ student.budiluhur.ac.id, [2]arief.wibowo @ budiluhur.ac.id

Abstract**—** *Various attempts were made by universities to prepare graduates to be ready to face the world of work. One of the readiness is to predict the waiting period for first time employment for prospective graduates. With this prediction, universities can evaluate the quality of education always to be superior and better. This study aims to make a prediction model of students' waiting periods when getting their first job. The problem is solved using classification data mining techniques, namely the Naïve Bayes algorithm. The analysis results using 199 training data and 22 testing data obtained an accuracy level of 90.90%, recall of 90.48%, and 100% precision. The prediction model is implemented in a prediction prototype application that is utilized by the head of the college's study program.*

Keywords**—** *Data Mining, Naïve Bayes, Predictions, Job waiting times.*

I. INTRODUCTION

One of the roles of tertiary institutions is to be committed to producing quality graduates to win workplace competition. To achieve the success of competition in the world of work requires the best educational process in Higher Education. The higher the level of education that has been taken, the more qualified graduates should be. One indicator of graduates' quality is how these graduates can compete in the world of work with a short waiting time to get a job. Various attempts were made by the University to prepare this output to be ready in facing the world of work, one of which is by predicting the waiting period to get a job. With this prediction, Higher Education can improve the quality of education of students so that students are superior and faster in getting jobs. Based on these problems, this study analyzed and predicted students' waiting time to get their first job. The analysis process uses Data Mining Techniques and will be analyzed using the Naïve Bayes algorithm.

Several previous studies discussing data mining using the Naïve Bayes algorithm have been carried out, for example, the prediction of new student admissions, this study aims to predict the admission of new students, by analyzing using some junior high school student data analyzed to determine the level of student acceptance in vocational schools which aims to determine Naïve Bayes algorithm test results in the prediction of new student admissions at the Vocational High School [1].

Previous research related to the Naïve Bayes algorithm, which discusses the prediction of student failure. This study aims to increase the success of students in studies so that students do not fail again. Based on testing, the Naïve Bayes algorithm produces an accuracy rate of 77.97% from 395 datasets to predict student failure [2].

Other research related to the Naïve Bayes algorithm which discusses the classification of scholarship recipients. This research was conducted to analyze the eligibility of scholarship recipients. In this research, it is proven by the ability of Naive Bayes Classifier to classify data on PPA scholarship applicants. It is resulting in a classification probability model for class determination on the next scholarship registrant. Testing the accuracy of the model of the system developed. It produces the smallest accuracy value of 64% in testing with a sample of 100 data and produces the highest accuracy value of 97.66% [3]. This study aims to make a prediction model of students' waiting periods when getting their first job. Modeling is done by analyzing past data about waiting times for college graduates to get prediction rules. The rules formed are then implemented in the form of an application prototype.

II. LITERATURE REVIEW

Data Mining is a process of finding meaningful relationships of trends by examining an extensive collection of data stored in storage with pattern recognition techniques such as statistical and mathematical techniques. Abundant data availability, the need for information or knowledge to support decision making to create business solutions, and infrastructure support in the field of information technology are the support of data mining processing. Data Mining is used to detect strange events such as various applications of data mining, namely market and management analysis, company analysis and risk management, telecommunications, finance, insurance, sports, and astronomy. As a technology that can produce knowledge, data mining processes in several stages [4].

Naïve Bayesian method is a data mining technique used to predict an event in the future, by comparing it with data or evidence in the past. The use of word probabilities or tokens is used as input - the probability of events. Naïve Bayesian classification will look at old data in determining the similarity value of new data. There must be old data used as comparative data in the Bayes process [5]. The equation of the Bayes theorem is:

$$P(X|H) = \frac{P(X|H)P(H)}{PX} \qquad (1)$$

Remarks:

X	: Data with unknown classes
H	: The data hypothesis is a class specific
P (H|X)	: Probability of hypothesis H based condition X
P (H)	: Hypothesis probability H

Riski Amalia and Arief Wibowo, "Prediction of the Waiting Time Period for Getting a Job Using the Naive Bayes Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 2, pp. 225-229, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

P (X|H) : Probability of X based on the conditions at hypothesis H

P (X) : Probability of X

Generally, Bayes Theorem is easily calculated for categorical type features. If the feature has numeric (continuous) data, special treatment is needed before it is included in Naive Bayes. The plot of Naïve Bayes can be seen from the following picture [6]:

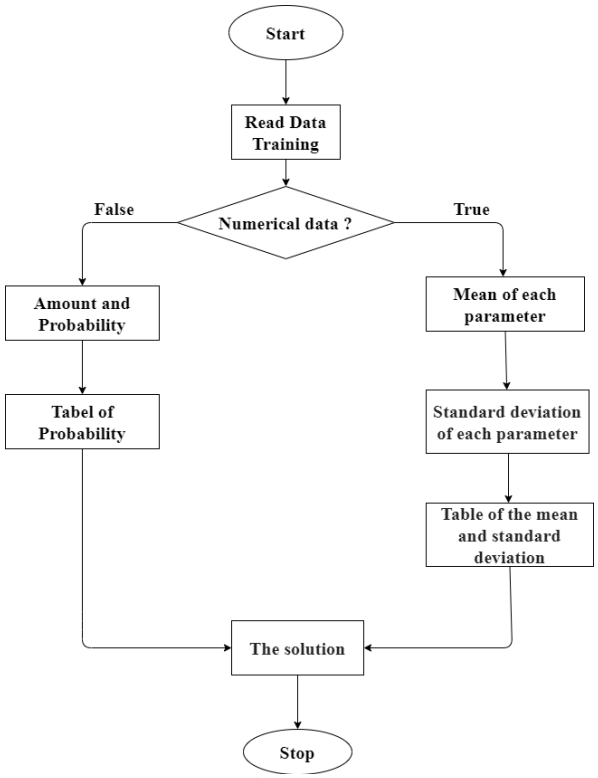

Fig. 1. The Plot of the Naive Bayes Algorithm

Based on Figure 1, it can be seen that Bayes naïve is known as a flexible algorithm that is capable of processing numerical and non-numeric data for classification of data mining. In numerical data, the steps to be taken are calculating the average value of each parameter, calculating the standard deviation of each parameter, and getting the value in the mean, the table of standard deviations and probabilities, to get a solution. Non-numeric data will be processed by counting numbers and probabilities, getting probability tables, and getting solutions.

III. Research Metodology

This research consists of several stages, starting with the determination of the problem formulation and the literature study to find alternative solutions and continue with data mining methods consisting of data processing, training, and data testing. The final stage of this study is to build a prototype application for waiting period prediction to understand the pattern or knowledge gained from data mining methods. The flow of this research can be seen in Figure 2.

The first thing to do in this research is to conduct a literature study relating to previous research on the use of the Naïve bayes method. In this case, the researcher wants to make predictions about the waiting time of students in getting a job using data mining techniques. After conducting a literature study, the next stage is determining the right method. In this study, the researcher chose the Naïve bayes algorithm because it can be used to calculate a set of probabilities from a combination of values in a dataset.

In this study, the object of research was the alumni of the noble Budi Secretariat Academy in Jakarta. The steps in data processing include data collection, data selection, data cleaning, and data transformation. After processing the data, it will get a dataset that can be used for further processing.

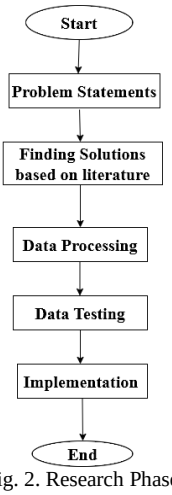

Fig. 2. Research Phase

In the next step, the data is tested by the Naïve Bayes method. Testing is done by manual calculation and comparing the results with prototype applications. From the resulting prediction model, the accuracy, precision, and recall are calculated, using the confusion matrix.

IV. Result and Discussion

In this research, the object of research is historical data of the Budi Luhur Secretary Academy alumni who have gotten a job. The steps or stages in data processing in this study include data collection, data selection, data cleaning, data transformation.

At the data collection stage, 314 primary data were obtained for data selection. Data selection is a process where the selection of the many attributes that exist, then selected based on the needs in this study, the authors use several attributes to be used as parameters in predicting waiting time for work, the attributes that will be used can be seen in Table I.

TABLE I. Data Attributes.

Attribute	Description
NIM	Student ID number
High School	Previous High School
Town	High School location
HS Distance	The distance of High School from the college
School Category	School Category (Private/Public)
Age	Age of student in the admission period
GPA1	Semester Performance Index1
GPA2	Semester Performance Index2
GPA3	Semester Performance Index3
Label	Status of waiting time to get a job

Riski Amalia and Arief Wibowo, "Prediction of the Waiting Time Period for Getting a Job Using the Naive Bayes Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 2, pp. 225-229, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Data cleaning is performed at the next stage to eliminate data that is still null to facilitate the calculation process. After cleaning the data, 221 data were obtained, which will be processed into training data and testing data. the data split into 90% for training data and 10% for test data.

The data transformation process is done so that the data do not have too many variations in values. This transformation phase is done on two attributes, namely, the school's origin and the school's city, which will be transformed into school distance. Data changes were also made to the attributes that will be labeled in this study, namely, the waiting time whose initialized value can be seen in Table II.

TABLE III. Prediction Class.

Remark	Class Label
Getting a job before graduation	Negative
Getting a job after graduation	Positive

After the data has been transformed, the analysis is carried out using the Naïve Bayes formula. The workings of the Naïve Bayes calculation process are started by taking training data. Examples of training data used can be seen in Figure 3.

Traning Data (199 rows)
File name: training.xlsx

No.	Student Id	High School	Town	HS Distance	School Category	Age	GPA 1	GPA 2	GPA 3	Label
1	1221300021	Cyber Media	Jakarta	0	Private	19	3.00	3.42	3.38	positive
2	1221300039	SMK Muhammadiy	Tangerang	2	Private	19	2.52	2.63	3.26	positive
3	1221300047	SMK Yadika 4	Tangerang	2	Private	20	2.96	2.68	2.95	positive
4	1221300096	SMAN 12	Tangerang	2	Public	20	3.61	3.67	3.48	positive
5	1221300104	SMKN 23 Jakarta	Jakarta	0	Public	18	3.61	3.58	4.00	positive
6	1221300138	SMK Yadika 4	Tangerang	2	Private	18	2.61	3.05	2.76	negative
7	1221300153	Public 1 Pedes	Karawang	3	Public	20	2.43	3.13	3.33	negative
8	1221300179	SMK Jakarta 1	Jakarta	0	Private	18	3.43	3.25	3.33	positive
9	1221300187	SMK Yadika 4	Tangerang	2	Private	18	3.09	3.63	3.52	positive
10	1221300195	SMK Budi Murni	Jakarta	0	Private	18	2.83	3.50	2.95	positive
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
199	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Fig. 3. Training Data

Figure 3 has shown that 199 data are used in training. Thus the remaining 22 data are used for the testing phase.

Testing Data (22 rows)
File name: testing.xlsx

No.	Student Id	High School	Town	HS Distance	School Category	Age	GPA 1	GPA 2	GPA 3	Label
1	1521300093	SMK An-Nurmaniyah	Tangerang	2	Private	18	3.73	3.71	3.63	Positive
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
22	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Fig. 4. Testing Data

Next, an example of naïve bayes calculation is performed with the steps as:

a. Count the number of Class/Label
   P (Y= Positive) = 161/199 = 0.80905
   (The amount of positive data divided by the total amount of data).
   P (Y= Negative) = 38/199= 0.19095
   (the amount of Negative data divided by the total amount of data).
b. Counting the same number of cases with the same class, in this study there are 2 types of attributes, namely categorical data and numerical data. as can be seen in Figure 1, for the categorical attribute, the probability can be directly calculated, but for the numerical attribute to calculate the probability, the first value and the standard deviation must be found afterwards, then calculate the probability using the dentus gauss formula which can be seen in Formula 2 [6].

$$P(Xi = xi|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi=\mu ij)^2}{2\sigma^2 ij}} \quad (2)$$

Probability of categorical data types:

P (School category = Public | Y=Positive)
= 70/161=0.43478261
P (School category = Public | Y=Negative)
=16/38=0.4210526
P (School category = Private | Y=Positive)
= 91/161 = 0.5652173
P (School category = Private | Y=Negative)
= 22/38 = 0.5789473

Probability of numeric data types:

The first step is calculating the mean value, while the formula for calculating the mean (mean) [1], with the following example:

$$\mu = \frac{x1+x2+x3+\ldots\ldots+xn}{n} \quad (3)$$

Remarks:
μ = mean
xi = sample value of i
n = count of sample

Distance of school in positive class
$= \frac{0+2+2+2+0+0+2+0+2+3+0+\cdots+0}{161} = 1.075$

Distance of school, in negative class
$= \frac{2+3+0+0+0+2+1+0+3+0+0+\cdots+3}{38} = 1.079$

Age, in positive class
$= \frac{19+19+20+20+18+18+18+18+\ldots\ldots+18}{161} = 18.429$

Age, in negative class
$= \frac{18+20+18+17+19+20+18+18+18+\ldots\ldots+20}{38} = 18.921$

GPA1, in positive class
$= \frac{3.00+2.52+2.96+3.61+3.61+3.43+3.09++\cdots+4.00}{161} = 3.326$

GPA1, in negative class
$= \frac{2.61+2.43+2.91+3.22+2.35+2.78+2.43+\ldots.+3.26}{38} = 3.141$

GPA2, in positive class
$= \frac{3.42+2.63+2.68+3.67+3.58+3.25+3.63+\ldots\ldots\ldots+2.60}{161} = 3.487$

GPA2, in negative class
$= \frac{3.05+3.13+3.05+3.50+2.93+3.05+2.93+\ldots.+3.65}{38} = 3.256$

GPA3, in positive class
$= \frac{3.38+3.26+2.95+3.48+4.00+3.33+3.52+\ldots.+2.00}{161} = 3.519$

GPA3, in negative class
$= \frac{2.76+3.33+3.14+3.48+2.90+3.00+2.76+\cdots+3.68}{38} = 3.331$

The second stage is looking for the standard deviation, the formula for calculating the standard deviation is as follows [1]:

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n} (xi - \mu)^2}{n}} \quad (4)$$

Remarks:
σ = standar deviation
x_i = value of x_i

Riski Amalia and Arief Wibowo, "Prediction of the Waiting Time Period for Getting a Job Using the Naive Bayes Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 2, pp. 225-229, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

μ = average count
n = count of sample

School distance, in positive class

$$= \frac{(0-1.075)^2+(2-1.075)^2+(2-1.075)^2+\ldots.+(0-1.075)^2}{161}$$

= 1.081

School distance, in negative class

$$= \frac{(2-1.079)^2+(3-1.079)^2+(0-1.079)^2+\ldots.+(3-1.079)^2}{38}$$

= 1.148

Age when entering, in positive class

$$= \frac{(19-18.429)^2+(19-18.429)^2+(20-18.429)^2+\ldots.+(19-18.429)^2}{161}$$

= 1.345

Age when entering, in negative class

$$= \frac{(18-18.921)^2+(20-18.921)^2+(18-18.921)^2+\ldots.+(20-18.921)^2}{38}$$

= 2.813

GPA1, in positive class

$$= \frac{(3.00-3.326)^2+(2.52-3.326)^2+(2.96-3.326)^2+\ldots.+(4.00-3.326)^2}{161}$$

= 0.411

GPA1, in negative class

$$= \frac{(2.61-3.141)^2+(2.43-3.141)^2+(2.91-3.141)^2+\ldots.+(3.26-3.141)^2}{38}$$

= 0.428

GPA2, in positive class

$$= \frac{(3.42-3.487)^2+(2.63-3.487)^2+(2.68-3.487)^2+\ldots.+(2.60-3.487)^2}{161}$$

= 0.353

GPA2, in negative class

$$= \frac{(3.05-3.256)^2+(3.13-3.256)^2+(3.05-3.256)^2+\ldots.+(3.65-3.256)^2}{38}$$

= 0.372

GPA3, in positive class

$$= \frac{(3.38-3.519)^2+(3.26-3.519)^2+(2.95-3.519)^2+\ldots+(2.00-3.519)^2}{161}$$

= 0.332

GPA3, in negative class

$$= \frac{(2.76-3.331)^2+(3.331-3.331)^2+(3.14\ 3.331)^2+\ldots+(3.68-3.331)^2}{38}$$

= 0.338

The third step is calculating the probability using the Gausian approach that can be seen in formula 2.

Distance of school, in positive class

$$= \frac{1}{\sqrt{2\pi(1.081)}} e^{-\frac{(2-1.075)^2}{2(1.081)^2}} = 0.55348187$$

Distance of school, in negative class

$$= \frac{1}{\sqrt{2\pi(1.148)}} e^{-\frac{(2-1.079)^2}{2(1.148)^2}} = 0.51382054$$

Age, in positive class

$$= \frac{1}{\sqrt{2\pi(1.345)}} e^{-\frac{(18-18.429)^2}{2(1.1.345)^2}} = 0.36203494$$

Age, in negative class

$$= \frac{1}{\sqrt{2\pi(2.813)}} e^{-\frac{(18-18.921)^2}{2(2.813)^2}} = 0.251022$$

GPA1, in positive class

$$= \frac{1}{\sqrt{2\pi(0.411)}} e^{-\frac{(3.73-3.326)^2}{2(0.411)^2}} = 1.0090501$$

GPA1, in negative class

$$= \frac{1}{\sqrt{2\pi(0.428)}} e^{-\frac{(3.73-3.141)^2}{2(0.428)^2}} = 1.572317$$

GPA2, in positive class

$$= \frac{1}{\sqrt{2\pi(0.353)}} e^{-\frac{(3.71-3.487)^2}{2(0.353)^2}} = 0.819958$$

GPA2, in negative class

$$= \frac{1}{\sqrt{2\pi(0.372)}} e^{-\frac{(3.71-3.256)^2}{2(0.372)^2}} = 1.3777777$$

GPA3, in positive class

$$= \frac{1}{\sqrt{2\pi(0.332)}} e^{-\frac{(3.63-3.519)^2}{2(0.332)^2}} = 0.7323592$$

GPA3, in negative clas

$$= \frac{1}{\sqrt{2\pi(0.338)}} e^{-\frac{(3.63-3.331)^2}{2(0.338)^2}} = 1.0150540$$

Multiplication for each class
(P|Positive) = 0.05552310
(P|Negative) = 0.03135385

The final stage is to compare the results of positive and negative classes. From the above results, it appears that the highest probability value is in the class (P|Positive). It can be concluded that the student is predicted to have a "Positive" waiting period, which means the student is predicted to get a job after graduating.

After the prediction model in the form of rules is obtained, the rules can be implemented into a prediction application for students to wait for a job. The screen display is shown in Figure 5 and Figure 6.

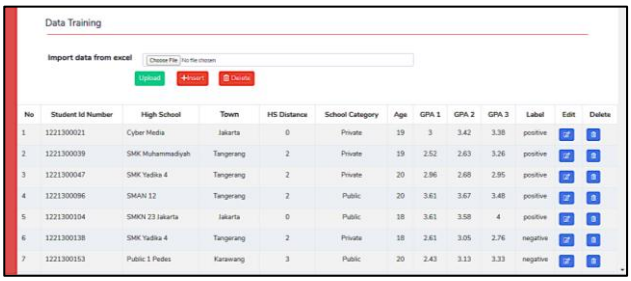

Fig. 5. Data Training Process

The training data page is used to import training data. After the data import process, all the data is processed for Naïve Bayes calculation, the calculation can be seen in Figure 6.

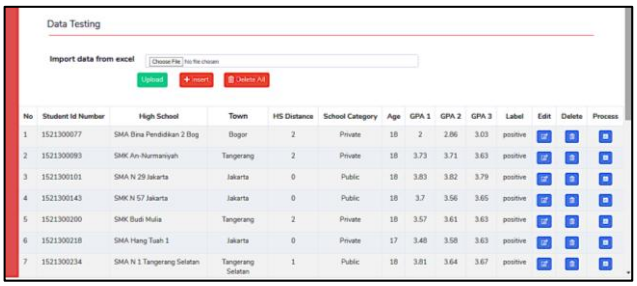

Fig. 6. Data Testing Process

On the next screen page, the testing process is carried out. The test was conducted on 22 testing data, which

Riski Amalia and Arief Wibowo, "Prediction of the Waiting Time Period for Getting a Job Using the Naive Bayes Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 2, pp. 225-229, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

constituted 10% of all data collected for this study. The test results are shown in Figure 7.

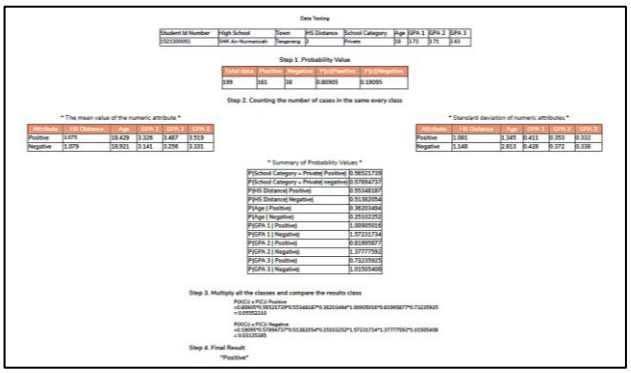

Fig. 7. Data Testing Result

In this section, the test data are completed using Naïve Bayes calculations based on the rules of modeling results using training data. The test results are shown in Figure 8.

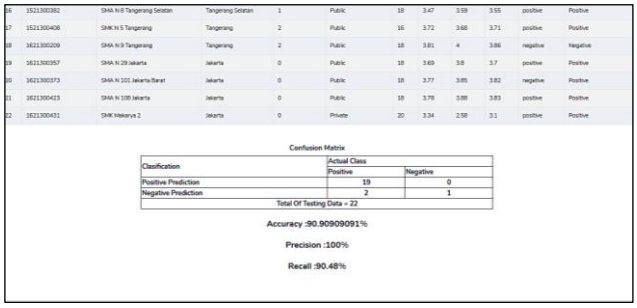

Fig. 8. Testing Result Parameters

After processing the training data, the accuracy of the data is obtained. Accuracy is calculated using Confusion matrix. The following describes the confusion matrix in Table III.

TABLE IIIII. Confusion Matrix.

Prediction Result	**Actual**	
	Positive	**Negative**
Positive	TP = 19	FP = 0
Negative	FN = 2	TN = 1

Based on the results of the confusion matrix seen in class positive, 19 students are predicted to be precisely in positive class, while found two students in the negative class are predicted to be incorrect class and found one student who is predicted to be correct class.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{19+1}{19+1+0+2} = \frac{20}{22} = 90.90\%$$

$$\text{Precision} = \frac{TP}{FP+TP} = \frac{19}{0+19} = \frac{19}{19} = 100\%$$

$$\text{Recall} = \frac{TP}{FN+TP} = \frac{19}{2+19} = \frac{19}{21} = 90.48\%$$

V. CONCLUSION

Based on the results of the study, it can be concluded that the Naïve Bayes method can be used to predict the waiting time of students to obtain work using 199 training data and 22 testing data, consisting of 1 categorical and five numerical attributes.

The confusion matrix was obtained from the testing process with an accuracy value of 90.90%, a precision value of 100%, and a recall value of 90.48%. The conclusions of this study show that more students are expected to have a positive waiting period or get a job after graduation.

REFERENCES

[1] S. Rizal dan M. Lutfi, "Penerapan Algoritma Naïve Bayes Untuk Prediksi Penerimaan Siswa Baru di SMK Al-Amien Wonorejo," *Jurnal Keilmuan dan Aplikasi Teknik Informatika*, vol. 10, no. 1, hal. 14–21, 2018. (Published in Bahasa).

[2] R. Rumini dan N. Norhikmah, "Prediksi Kegagalan Siswa Dalam Data Mining Dengan," *Jurnal Mantik Penusa*, vol. 3, no. 1, hal. 42–46, 2019. (Published in Bahasa).

[3] S. Adi, "Implementasi Algoritma Naive Bayes Classifier Untuk Klasifikasi Penerima Beasiswa PPA Di Universitas Amikom Yogyakarta," *Jurnal Mantik Penusa*, vol. 22, no. 1, hal. 11–16, 2018. (Published in Bahasa).

[4] M. F. Rifai, H. Jatnika, dan B. Valentino, "Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Tingkat Kelulusan Peserta Sertifikasi Microsoft Office Specialist (MOS)," *Petir*, vol. 12, no. 2, hal. 131–144, 2019. (Published in Bahasa).

[5] R. D. Pambudi, A. A. Supianto, dan N. Y. Setiawan, "Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Pendekatan Data Mining Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Brawijaya," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 3, hal. 2194–2200, 2019. (Published in Bahasa).

[6] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," *Creative Information Technology Journal*, vol. 2, no. 3, hal. 207–217, 2015. (Published in Bahasa).