www.bircu-journal.com/index.php/birci
Budapest International Research and Critics Institute-Journal (BIRCI-Journal)
Humanities and Social Sciences
ISSN 2615-3076 Online)
ISSN 2615-1715 (Print)
Budapest Institute

Classification of the Nave Bayes Method in Determining the Concentration of Information Programs

Mia Miskiatul Atiroh
Universitas Telkom, Indonesia
miamiskiatulatiroh@student.telkomuniversity.ac.id

Abstract

The Faculty of Information Technology is one of the faculties at Serang Raya University which has three study programs, namely Information Systems (IS), Computer Engineering and Informatics. In the informatics study program there are two concentrations, namely programming and multimedia concentrations. Determination of concentration cannot be separated from student supervision, on the other hand, academics must have their own policies in determining student concentration. The method used in this research is Nave Bayes which has 10 variables, namely Algorithm, Calculus I, Internet Html, Commerce Package Program, Graphic Design, Hardware & Software, Calculus II, Data Communication, Introduction to Object Oriented I, Introduction to Information Technology. In this study, there were 248 student data from the informatics study program where 100 student data consisted of students who had taken concentration and 148 data were students who had not taken concentration. In the calculation of 100 student data who have taken concentration, it is known that the number of data with programming concentration is 87 students while for multimedia concentration there are 13 students. Furthermore, the 100 data is divided into 80 training data and 20 testing data using random sampling technique. Based on student academic data which was used as testing data, the Naive Bayes method was successful in classifying 20 student data from 100 student data. Thus, the Naive Bayes method is successful in classifying concentrations with an accuracy success rate of 0.80 (80%) and an accuracy of 0.565 Kappa statistics so that the concentration selection using the Nave Bayes classifier method is accurate.

Keywords
Concentration; naïve bayes; success rate; kappa statistics

I. Introduction

The Faculty of Information Technology is one of the faculties at Serang Raya University which has three study programs, namely Information Systems (IS), Computer Engineering and Informatics. In the informatics study program there are two concentrations, namely programming and multimedia concentrations. Determination of concentration cannot be separated from student supervision, on the other hand, academics must have their own policies in determining student concentration. In addition, determining the concentration will help students focus more on what they are interested in and see from the academic value of a student himself. According to the Head of Informatics Engineering Study Program, Mr. Akip Suhendar., M.Kom "while the selection of concentration was chosen directly by the student". The problem in this research is that there are no conditional courses for each programming and multimedia concentration,

DOI: https://doi.org/10.33258/birci.v5i2.5655

which makes it difficult for students to choose a concentration that matches the value of the course. In addition, there is no system for the process of determining the concentration.

Several researches in the field of computing have been carried out to contribute to the world of education. The researchers used the concept of data mining to classify student data (Dimitoglou et al., 2016., Ayu & Saryanti, 2019., Supriyanti et al, 2016). Several approaches to determine the concentration of student study programs are the nave Bayes algorithm (Pramata & Yulmaini., 2018), the use of unsupervised discretization techniques (Saleh et al., 2018).

Naive 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. The algorithm uses Bayes theorem and assumes all attributes are independent or not interdependent given by the value of the class variable. Based on these problems, this research uses the Naïve Bayes classification method in determining the concentration of students in the informatics engineering study program. With the technique of classifying or grouping student score data according to the similarity of the weights of certain courses related to the concentration of the existing majors.

The results of this study are expected to help students not to choose the wrong concentration and to be able to implement the Naïve Bayes method in determining concentration.

II. Review of Literature

The type of research used in writing this thesis is to use the type of applied research that is intended to find solutions to problems as a continuation of pure research. In addition, applied research is more about computational methods in comparison "tools".

The following are the stages of research by explaining how to apply computational methods to research. Can be seen in Figure 1.

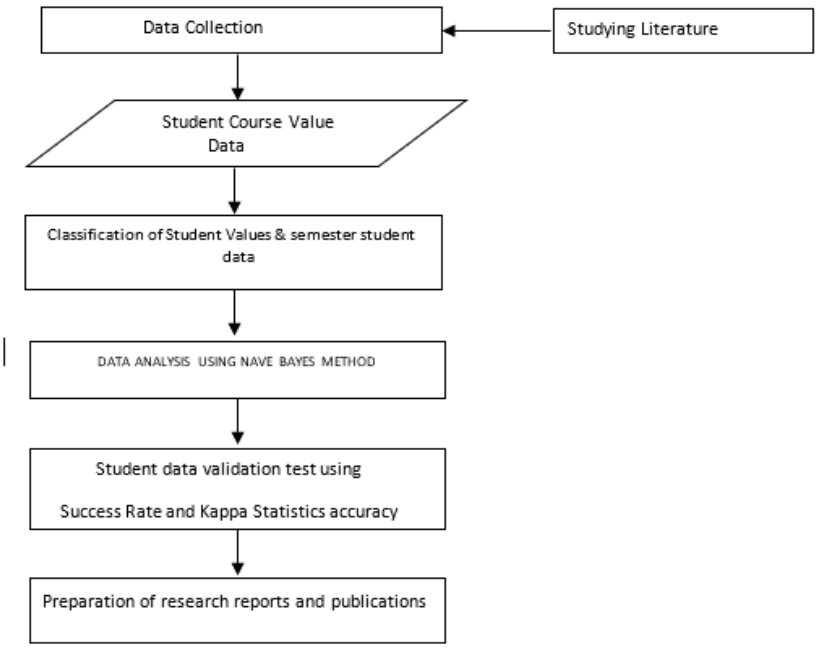

Figure 1. *Research Stage*

The method used in this research is Nave Bayes which has 10 variables, namely Algorithm, Calculus I, Internet Html, Commerce Package Program, Graphic Design, Hardware & Software, Calculus II, Data Communication, Introduction to Object Oriented I, Introduction to Information Technology. Based on the 10 data used from research data, it is divided into 8 data for training data and 2 data for testing data.

***Figure 2**. Results of data accuracy testing on 2 student data*

It can be seen from Figure 2 explaining the results of the accuracy test between the observed student data and the predicted student data. It can be seen in the first column table, the first true positive (TP) correct data detected correctly for the results of the observed multimedia concentration data accuracy test with the predicted multimedia concentration data as many as 0 student data. Furthermore, in the first column table, the first row of false positive (FP) data is incorrect but detected as correct data explaining the results of the observed multimedia concentration data accuracy test with programming concentration data that is predicted to be 0 student data in the second column table, the second row is false negative (FN) the data is correct but incorrectly detected explains the results of the observed multimedia concentration data accuracy test with programming concentration data that is predicted to be 1 student data. And for the second column table, true negative (TN) incorrect data detected correctly for the results of the observed multimedia concentration data accuracy test with the predicted multimedia concentration data as much as 1 student data.

So the results for class precision or the level of accuracy between the information requested by the user and the answer given by the system (TP/(TP+FP))*100% on programming prediction data is 0.00% and multimedia prediction data is 50.00% . As for class recall or the success rate of the system in retrieving an information (TP/(TP+FN))*100%, the observed data for programming is 0.00% and the observed data for multimedia is 100%.

The prediction results will be calculated using the success rate and kappa. The following is the calculation of accuracy for student classification results in programming concentration and multimedia concentration:

1) Succes Rate

True Positive rate	= TP/(TP+FN)
	= 0 / (0 + 1)
	= 0/1
	= 0
False Positive Rate	= FP/ (FP+TN)
	= 0 / (0 + 1)
	= 0/1
	= 0
Success rate	= (TP+TN)/(TP+TN+FP+FN)
	= (0 + 1)/ (0+1+0+1)
	= 1/2
	= 0,5
Error rate	=1 – Success Rate.
	= 1- 0,5
	= 0,5

2) Kappa Statistic

If the data matrix is known as follows:

Table 1. Data matrix

	X	Y
X	A	B
Y	C	D

Table 2. Data matrix

	X	Y
X	0	0
Y	1	1

Then the value of the accuracy of the observed data can be calculated as follows:

$$p_0 = \frac{a+d}{a+b+c+d}$$

$$p_0 = \frac{0+1}{0+0+1+1}$$

$$p_0 = \frac{1}{2}$$

$$p_0 = 0{,}5$$

Meanwhile, to find the expected accuracy value (Px) it can be calculated by the following equation:

$$p_X = \frac{a+b}{a+b+c+d} x \frac{a+c}{a+b+c+d}$$

$$p_X = \frac{0+0}{0+0+1+1} x \frac{0+1}{0+0+1+1}$$

$$p_X = \frac{0}{2} x \frac{1}{2}$$

$$p_X = 0 \, x \, 0{,}5$$

$$p_X = 0$$

Meanwhile, to find the expected accuracy value P (y) can be calculated by the following equation:

$$p_Y = \frac{c+d}{a+b+c+d} x \frac{b+d}{a+b+c+d}$$
$$p_Y = \frac{1+1}{0+0+1+1} x \frac{0+1}{0+0+1+1}$$
$$p_Y = \frac{2}{2} x \frac{1}{2}$$
$$p_Y = 1 \; x \; 0{,}5$$
$$p_Y = 0{,}5$$

Thus, obtained:

$$expected\ accuracy = p_X + p_Y$$
$$expected\ accuracy = 0 + 0{,}5$$
$$expected\ accuracy = 0{,}5$$

Here is the equation to calculate the Kappa . value.

$$Kappa = \frac{observed\ accuracy - expected\ accuracy}{(1 - expected\ accuracy)}$$
$$Kappa = \frac{(0{,}5 - 0{,}5)}{(1 - 0{,}5)}$$
$$Kappa = \frac{0}{0{,}5}$$
$$Kappa = 0$$

Based on the results of the classification calculation using an accuracy test in the form of a success rate of 0.50 and kappa statistics, namely that the Nave Bayes method for selecting student concentrations is accurate. The following is table 4.5 the results of the calculation of the accuracy test using the success rate and kappa statistics methods.

Table 3. Accuracy test results of SR and Kappa methods

Success rate	ba statistic
0,50	0

III. Result and Discussion

At this stage, we will present the results of student data that have been calculated using the Naïve Bayes method. In this study, there were 248 data on students from the informatics study program where 100 student data consisted of students who had taken concentration and 148 data of students who had not taken concentration which can be seen in Appendix 1.

In the calculation of 100 student data who have taken concentration, it is known that the number of data with programming concentration is 87 students while for multimedia concentration there are 13 students. Furthermore, the 100 data is divided into 80 training data and 20 testing data using random sampling technique. The following is a sample of training data which can be seen in table 4.

Table 4. Sample of student training data

NIM	Algorithm	Calculus I	HTML	PPN	Graphic design	Hardware Software	Calculus II	Data communication	PBO I	PTI	con-center
11216104	B+	B-	C	A	B	A-	C+	A	A-	A	PROGRAMMING
11216065	A-	B-	B-	C+	A	A-	B	B+	A-	A	MULTIMEDIA

The following is a sample of testing data which can be seen in table 5

Table 5. Sample of student testing data

NIM	Algorithm	Calculus I	HTML	PPN	Graphic design	Hardware Software	Algorithm II	Data communication	PBO I	PTI	con-cente
11217025	A-	A	A-	B+	C	E	B+	A	E	A	PROGRAMMING

From the data that has been collected, the next step is to calculate using Naive Bayes. The first calculation is to calculate the probability for each programming and multimedia concentration. Based on the results of the trainer, the probability for each concentration is as follows which can be seen in table 6.

Table 6. Probability output results for concentration

PROBABILITY	PROGRAMMING	MULTIMEDIA
	0.86	*0.14*

In table 6 it can be seen that the probability results for programming concentration are 0.86 and the multimedia probability is 0.14 from 80 student training data, by calculating the concentration probability value. $P(konsentrasi) = \frac{\text{n(pemrograman/multimedia)}}{\text{N}}$ where n is the number of sample data for programming or multimedia concentration to which the probability value will be calculated, while N is the total sample data.

Then the next step is to calculate the probability concentration of the conditions on the hypothesis. Table 5.2 shows a sample of student data with 11217025 entering concentration. Programming to generate probability values for each variable by testing data

based on data that has been trained. After performing calculations using nave Bayes by finding the same variable value for programming or multimedia concentration, the results of the probability for each variable in the concentration of programming and multimedia for students with the number 11217025 can be seen in table 7/

Table 7. Probability results of each data testing variable

con-center	Algorithm	Calculus I	HTML	PPN	Graphic design	Hardware Software	Algorithm II	Data communication	PBO I	PTI
P(programming)	0,07	0,07	0,01	0,04	0,04	0,01	0,13	0,61	0,09	0,28
P(multimedia)	0,09	0,09	0	0,09	0	0	0,09	0,45	0,18	0,27

Table 7 is the result of calculating the probability value for each variable in the testing data (see Table 7) against the target class for programming and multimedia concentration. The first line of the target class of programming concentration for the calculation results on the algorithm variable is 0.07 where the algorithm variable with an A- value to the programming concentration is 5 students with a total programming concentration of 69 student data, so the probability for the algorithm variable with an A- value is 0.07 . Likewise with other variables such as the calculus I variable with an A value of 5 students with a probability of 0.07, the internet html variable with an A- value of 1 student with a 0.01 probability, the commercial package program variable (VAT) with a B+ value of 3 students. with a probability of 0.04, graphic design variable with a value of C as many as 3 students with a probability of 0.04, hardware & software variable with a value of E as many as 1 student with a probability of 0.01, calculus II variable with a value of B+ as many as 9 students with a probability of 0.13, data communication variable with A value of 42 students with a probability of 0.61, PBO I variable with an E value of 6 students with a probability of 0.09, PTI variable with an A value of 19 students with a probability of 0.28. Next is the second line, the target class for the multimedia concentration for the calculation results on the algorithm variable is 0.09 where the algorithm variable with an A- value to the programming concentration is 1 student with a total programming concentration of 11 student data, so the probability for the algorithm variable with an A- value is 0 ,09. Likewise with other variables such as the calculus I variable with an A value of 1 student with a probability of 0.09, the internet html variable with an A- value of 0 students with a probability of 0, the commercial package program variable (VAT) with a B+ value of 1 student with a probability 0.09, graphic design variable with a C value of 0 students with probability 0, hardware&software variable with an E value of 0 students with a probability of 0, a calculus II variable with a B+ value of 1 student with a probability of 0.09, a data communication variable with an A value as many as 5 students with a probability of 0.45, the PBO I variable with an E value of 2 students with a probability of 0.18, the PTI variable with an A value of 3 students with a probability of 0.27.

After calculating the probability value of each Data Testing variable. The next step is to calculate the probability value of Data Testing for each target class by calculating the probability of programming and multimedia target classes against the probability of variables in the target class both in programming and in multimedia. The following are the results of the probability calculation for each target class as seen in table 8.

Table 8. Results of probability values in the target class

NIM	CLASS PREDICTION	PROGRAMMING	MULTIMEDIA
11218004	PROGRAMMING	0.00000016193	-

It can be seen in table 8 the results of the calculation that the probability value in the target class for programming concentration is greater than the probability for the target class for multimedia concentration. Because at the time of calculating the probability for each variable in the target class by multiplying all the variables in the target class, the concentration of programming or multimedia with the target class of concentration itself.

Based on the results of calculations using Naive Bayess, it can be seen that Data TestingP = (Algorithm = A-, Calculus I = A, Internet Html = A-, Commerce Package Program = B+, Graphic Design = C, Hardware&Software = E, Calculus II = B+, Data Communication = A, Object-Oriented Introduction I = E, Introduction to Information Technology = A) entered into the target class concentration = programming. This can be seen from the probability value for the concentration target class = programming, which is higher than the probability value for the concentration target class = multimedia.

Thus, based on the comparison of programming concentration with multimedia concentration,

1) If programming concentration > multimedia concentration, then students enter programming concentration
2) If the multimedia concentration > programming concentration, the student enters the multimedia concentration

3.1 Discussion

Based on Table 8, it can be seen the accuracy results of the class prediction between the data students who fall into the classification of programming and multimedia concentrations in the Study Program Informatics such as Figure 1.

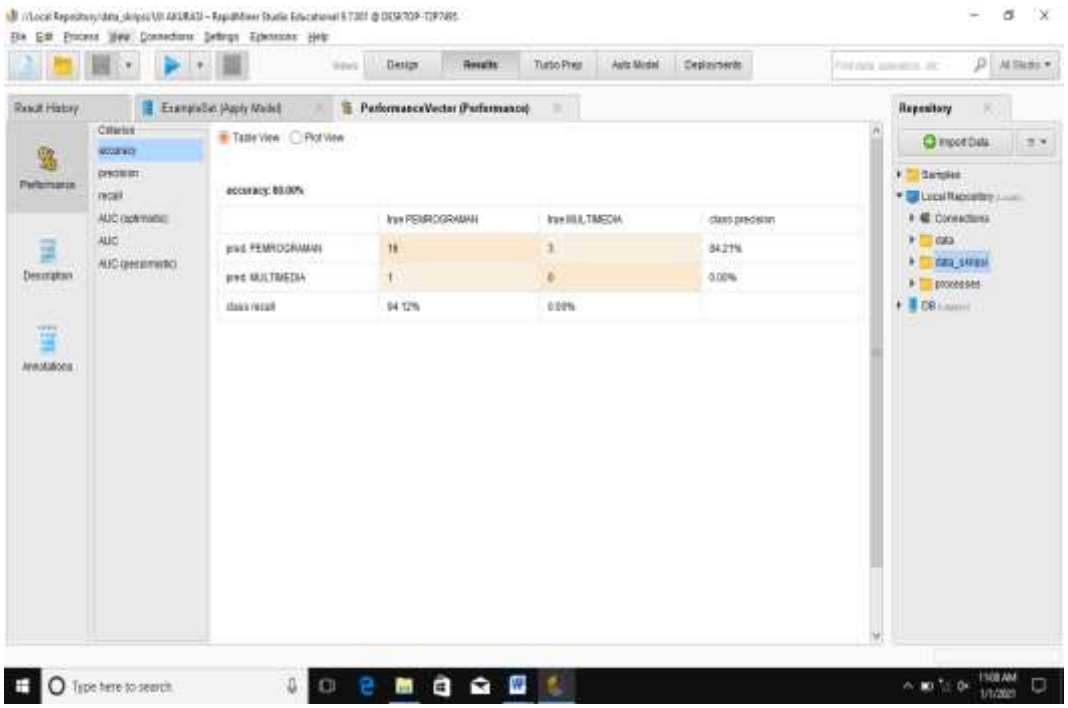

Figure 3. *Results of data accuracy testing on 20 student data*

1. Succes Rate

True Positive rate	= TP/(TP+FN)
	= 16 / (16+1)
	= 16/17
	= 94,12
False Positive Rate	= FP/ (FP+TN
	= 3 / (3+0)
	= 3/3
	= 1
Success rate	= (TP+TN)/(TP+TN+FP+FN)
	= (16+0)/ (16+0+3+1)
	= 16/20
	= 0,80
Error rate	=1 – Success Rate.
	= 1- 0,80
	= 0,2

1) Kappa Statistic

If the data matrix is known as follows:

Table 9. Data matrix

	X	Y
X	A	B
Y	C	D

Table 10. Data matrix

	X	Y
X	16	3
Y	1	0

Then the value of the accuracy of the observed data can be calculated as follows:

$$p_0 = \frac{a+d}{a+b+c+d}$$

$$p_0 = \frac{16+0}{16+3+1+0}$$

$$p_0 = \frac{16}{20}$$

$$p_0 = 0{,}80$$

Meanwhile, to find the expected accuracy value (Px) it can be calculated by the following equation:

$$p_X = \frac{a+b}{a+b+c+d} x \frac{a+c}{a+b+c+d}$$

$$p_X = \frac{16+3}{16+3+1+0} x \frac{16+1}{16+3+1+0}$$

$$p_X = \frac{19}{20} x \frac{17}{20}$$

$$p_X = 0{,}95x\ 0{,}85$$

$$p_X = 0{,}81$$

Meanwhile, to find the expected accuracy value P (y) can be calculated by the following equation:

$$p_Y = \frac{c+d}{a+b+c+d} x \frac{b+d}{a+b+c+d}$$

$$p_Y = \frac{1+0}{16+3+1+0} x \frac{3+0}{16+3+1+0}$$

$$p_Y = \frac{1}{20} x \frac{3}{20}$$

$$p_Y = 0{,}05\ x\ 0{,}15$$

$$p_Y = 0{,}01$$

Thus, obtained:

$$expected\ accuracy = p_X + p_Y$$

$$expected\ accuracy = 0{,}81 + 0{,}01$$

$$expected\ accuracy = 0{,}82$$

Here is the equation to calculate the Kappa . value.

$$Kappa = \frac{observed\ accuracy - expected\ accuracy}{(1 - expected\ accuracy)}$$

$$Kappa = \frac{(0{,}80 - 0{,}82)}{(1 - 0{,}82)}$$

$$Kappa = \frac{0{,}02}{0{,}18}$$

$$Kappa = 0{,}1$$

Based on the results of the classification calculation using an accuracy test in the form of a success rate of 0.80 and kappa statistics, namely that the Nave Bayes method for selecting student concentrations is accurate.

IV. Conclusion

Data Mining helps in the application of the Naive Bayes method in obtaining information by classifying student data into programming and multimedia concentrations. The Naive Bayes method utilizes training data to generate the probability of each criterion for a different class, so that the probability values of these criteria can be optimized for the classification carried out by the Naive Bayes method itself. Based on student academic data which was used as testing data, the Naive Bayes method was successful in classifying 20 student data from 100 student data. Thus, the Naive Bayes method is successful in classifying concentrations with an accuracy success rate of 0.80 (80%) and an accuracy of 0.565 Kappa statistics so that the concentration selection using the Nave Bayes classifier method is accurate.

References

Ayu, I. G., & Saryanti, D. (2019). "Penerapan Teknik Clustering Untuk Pengelompokan Konsetrasi Mahasiswa Dengan Metode K-Means," Universitas Dakyana Pura, Tahun 2017, Hal.519-526.

Dimitoglou, G., Adams, J. A., & Jim, C. M. (2016). "Perbandingan Kinerja Algoritma C4.5 Dan Naive Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa. 1(2012)," Jurnal Informa Politeknik Indonusa Surakarta Vol. 1 Nomor 3 Tahun 2016, Hal. 61-67.

Antony Anwari Rahman Dan Agus Suryanto. (2017). "Implementasi Sistem Informasi Seleksi Penerima Beasiswa Dengan Metode Naive Bayes Classifier," Jurnal Penelitian Pendidikan Indonesia (Jppi) Vol. 2, No. 3, Tahun 2017, Hal. 1-8.

Fitria, A., & Azis, H. (2018). "Analisis Kinerja Sistem Klasifikasi Skripsi Menggunakan Metode Naïve Bayes Classifier. Prosiding Seminar Nasional Ilmu Komputer Dan Teknologi Informasi," Prosiding Seminar Nasional Ilmu Komputer Dan Teknologi Informasi Vol. 3, No. 2, Tahun 2018, Hal. 102-106.

Luh, N., Sri, W., Ginantra, R., & Wardani, N. W. (2019). "Implementasi Metoda Naïve Bayes Dan Vector Space Model Dalam Deteksi Kesamaan Artikel Jurnal Berbahasa," Jurnal Infomedia Vol. 4 No. Tahun 2019, Hal. 94-100.

Mahfudh, A. A., Mustofa, H., Islam, U., Walisongo, N., & Indonesia, S. (2019). "Klasifikasi Pemahaman Santri Dalam Pembelajaran Kitab Kuning Menggunakan Algoritma Naive Bayes Berbasis Forward Selection," Walisongo Journal Of Information Technology, Vol. 1 No. 2, Tahun 2019, Hal. 101-110.

Manga, A. R. (2018). "Penerapan Metode Naïve Bayes Pada Klasifikasi Judul Jurnal," Prosiding Seminar Nasional Ilmu Komputer Dan Teknologi Informasi Vol. 3, No. 2, Tahun 2018, Hal. 92-101

Munandar T. A., (2019) " Data Mining Dengan Bahasa R Revisi-3" Tahun 2019

Maulida, L. (2018). "Penerapan Data Mining Dalam Mengelompokkan Kunjungan Wisatawan Ke Objek Wisata Unggulan Di Prov. Dki Jakarta Dengan K-Means," Jiska (Jurnal Informatika Sunan Kalijaga), Vol. 2, No. 3, Tahun 2018, Hal.. 167–174.

Mustofa M. S., & Rahmadhan M. R., (2017). "implementasi data mining untuk evaluasi kinerja akademik mahasiswa menggunakan alogoritma naïve bayes" Citi c Journal Vol.4 No.2, Tahun 2017, Hal. 151-162

Muqorobin, M., Kusrini, K., & Luthfi, E. T. (2019). "Optimasi Metode Naive Bayes Dengan Feature Selection Information Gain Untuk Prediksi Keterlambatan Pembayaran Spp Sekolah," Jurnal Ilmiah Sinus (Jis) Vol : 17, No. 01, Tahun 2019, Hal. 1-14.

Nugroho, M. F., & Wibowo, S. (2017). "Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer Unaki Semarang Menggunakan Algoritma Naive Bayes," Jurnal Informatika Upgris Vol. 3, No. 1, Tahun 2017, Hal. 63-70

Olivita, D., & Vitriani, Y. (2016). "Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier Dan K-Nearest Neighbor," Jurnal Sains, Teknologi dan Industri, Vol. 14, No. 1, Tahun 2016, Hal. 79 - 85

Pratama, T., & Komputer, F. I. (2018). "Implementasi Algoritma Naive Bayes Dalam. Menentukan Konsentrasi Skripsi Dan Rekomendasi Bahasa Pemrograman," Jurnal Informatika, Vol. 18, No.1, Tahun 2018, Hal. 1-13

Saleh, A., & Mulia, T.,. (2015). "Penerapan Data Mining Dalam Menentukan. Jurusan

Siswa," Seminar Nasional Informatika, Tahun 2015, Hal. 351-355
Saleh, A., Nasari, F., Utama, U. P., & Siswa, K. J. (2018). "Penggunaan Teknik Unsupervised Discretization Pada Metode Naive Bayes Dalam Menentukan Jurusan Siswa," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) Vol. 5, No. 3, Tahun 2018, Hal. 353-360
Setiawan, A., Astuti, I. F., & Kridalaksana, A. H. (2015). "Klasifikasi Dan Pencarian Buku Referensi Akademik Menggunakan Metode Naïve Bayes Classifier (Nbc)," Jurnal Informatika Mulawarman Vol. 10 No. 1, Tahun 2015, Hal. 1-10.
Supriyanti, W., Kusrini, & Amborowati, A., (2016) "Perbandingan Kinerja Algoritma C4.5 dan Naïve Bayes untuk Ketetapan pemilihan konsentrasi mahasiswa," Jurnal Informa Politeknik Indonesia Surakarta, Vol. 1 NO.3, Tahun 2016, Hal.61-67.
Wibisono, A. B., & Fahrurozi, A., (2019) " Perbandingan algoritma klasifikasi dalam pengklasifikasikan data penyakit jantung koroner," jurnal ilmiah teknologi dan rekayasa Vol. 24 No.3, Tahun 2019, Hal.161-170.