HUMANIS
(Humanities,Management and Science Proceedings)

Vol. 3 • No. 1 • Desember 2022

Pege (*Hal.*) : **170 – 179**

ISSN (*online*) : 2746 - 4482
ISSN (*print*) : 2746 - 2250

© LPPM Universitas Pamulang
JL.Surya Kencana No.1 Pamulang, Tangerang Selatan – Banten
Telp. (021) 7412566, Fax (021) 7412491
Email : humanisproccedings@gmail.com

Special Issue :
ICoMS2022
The 3rd International Conference on Management and Science
Website. :
http://www.openjournal.unpam.ac.id/index.php/SNH

Majors Determination for High School Students Using the Naïve Bayes Algorithm, C4.5 and the K-Nearest Neighbor Algorithm (Case Study: SMA 1 Barunawati Jakarta)

Yudisti Prayigo Permana[1]; Taswanda Taryo[2]; Makhsun[3]

Pamulang University, Indonesia

E-mail:[a]mystr190.id@gmail.com
[b]dosen02234@unpam.ac.id
[c]makhsun_toha@yahoo.com

Abstract: Education is the most important function in life to form a good mindset and also to help develop the potential in students to become better individuals and the knowledge gained can be useful for many people. The majors process is the most important aspect in determining the interests and talents of students to facilitate students in carrying out learning. The majors must be done carefully and be seen from various aspects so that there is no mistake in determining the majors because it will have an impact on students' academic scores. In the majoring process, there are several aspects that are used as material for consideration, namely, by looking at the academic scores of students obtained from academic tests and then comparing them with the results of psychological tests and questionnaires regarding the majors of interest, so it takes quite a long time to get the results of majors. The difficulty in the process of classifying majors is an obstacle for the school to calculate from each criterion because there is no major system capable of producing majors classification with a high degree of accuracy so that the results obtained are in accordance with the abilities and interests of students. This study aims to get the best results from three algorithms, namely, Naïve Bayes, C4.5 and the K-Nearest Neighbor algorithm to determine the classification of majors in order to create more interesting, active learning because the learning that students get is in according to their interests and talents. The classification method using Naïve Bayes is a classification method based on probability which is used to predict with the assumption that between one class and another are not interdependent. In addition, the method using the C4.5 algorithm functions to classify data that has numeric and categorical attributes and the K-Nearest Neighbor algorithm works based on the assumption that a data will have the same class or category as the surrounding data. From the results of the tests carried out in this research of 214 datasets, the Naïve Bayes algorithm method has a better accuracy rate than the C4.5 and K-Nearest Neighbor algorithms from the amount of data processed resulting in an accuracy value of 98.13%. The comparisons have been made using data random data with real data of 50, 100, 214, 300, 400 and 428 data and it can be concluded that the nave Bayes algorithm is suitable to be applied in this case because it has the highest level of accuracy and is stable and not affected by the amount of data being tested.

Keywords: Data Mining, Classification, Major, Naïve Bayes, C4.5, K-Nearest Neighbor

INTRODUCTION

Education is the most important function in life to form a good mindset and also help develop the potential in students to become better individuals and the knowledge gained can be useful for many people. Majoring is a process that will determine the success of students, both while studying in high school and after college. Major is one of the most important aspects in determining student interests and talents so that students can more easily adjust learning according to their abilities. In determining a major, many new students do not know their talents and interests and feel confused in choosing what major is more suitable for their abilities and mistakes in choosing majors that are not in accordance with abilities.

The process of determining majors is one of the efforts to direct new students based on their academic abilities and interests. In the majors, there are several aspects that are used as material for consideration, namely, by looking at the academic scores of students obtained from academic tests and then comparing them with the results of psychological tests and questionnaires regarding the majors of interest, so it takes quite a long time to get the results of majors. The difficulty in the process of classifying majors is an obstacle for the school to calculate from each criterion because there is no major system capable of producing majors classification with a high degree of accuracy so that the results obtained are in accordance with the abilities and interests of students. In this study the authors chose SMA 1 Barunawati Jakarta as the object of research. The attributes used in determining the majors are student interest, the results of interviews and psychological tests as well as the results of the assessment of the matriculation test which consists of several types of core subjects from each major and the average value of the previous school. The method used to determine student majors is the classification method. In this study, researchers used three classification methods to compare the results of the two algorithms. The algorithm used in the majors classification is the Naïve Bayes algorithm, C4.5 and K-Nearest Neighbor.

The Naïve Bayes method is a classification method using probability and statistical methods which aims to classify data in certain classes then from the patterns obtained can be used to estimate what majors are suitable for the new students. Meanwhile, the C4.5 algorithm is applied to classify data that has numeric and categorical attributes and can then be used to predict the value of the discrete type attribute of the new record. Finally, the K-Nearest Neighbor method is a classification of objects based on the learning data closest to the object (Bramer; 2007:93). Based on the three classification methods, one of the advantages of the Naïve Bayes classification method is that it does not require a large amount of data and can also be utilized for quantitative and qualitative data. Meanwhile, C4.5 is used for predictive classification or segmentation and the K-Nearest Neighbor classification method is tranquil to implement and the data represented is adjusted to the structure. With this classification, students' majors that have been carried out are expected to make students not wrong in determining majors and it is easier to follow learning according to their abilities.

LITERATURE REVIEW AND HYPOTHESIS DEVELOPMENT

Research conducted firstly by Christiandita, Rahayuningtyas and Dedy Satrio Winarso (2017) uses the K-Nearest Neighbor algorithm for the majors of high school students. The difference from this study is that the data applied uses sample data where the results of the data are taken as much as 25% and then implemented into an application design for majors. The research is more on application design and does not lead to data analysis. While what the authors do is to analyze the data with various methods. For the equation lies in the method used, namely, using the K-Nearest Neighbor algorithm.

Research conducted secondly by Aditya Maulana Habibi and Reva Ragam Santika (2020) uses the K-Nearest Neighbor algorithm in determining majors using the Web-Based Euclidean Distance Method at Setia Gama Middle School. The difference from this research is that this research only utilizes one algorithm, namely K-Nearest Neighbor as a majoring method and does not experiment with other methods and divides the data into 2 parts, namely, training data and test data into 100 training data and 64 test data. Thirdly, Zuleha (2020) then majored in high school using the K-Nearest Neighbor Classifier method at SMAN 2 Singingi. The research is more on application design, does not lead to data analysis and there is no more detailed explanation related to testing the data. Fourthly, Sumarni Adi (2018) made predictions in the majors of new high school students with the Naïve Bayes

Classifier algorithm. In this study, starting from the initial stage carried out until the final process, the author did the same, except that the data applied used sample data and carried out sample testing with only 2 data.

Research conducted fifthly by Ahmad Zainul Mafakhir and Achmad Solichin (2020) on the Application of the Naïve Bayes Classifier Method for Majoring Students at Madrasah Aliyah Al-Falah Jakarta. In this study, the level of accuracy obtained is quite low, around 33.34% because there are obstacles when converting numeric values to categorical so as to produce a low level of accuracy. Research conducted sixthly by Endang Etriyanti, Dedy Syamsuar and Yesi Novaria Kunang (2020) on the Implementation of Data Mining Using the Naïve Bayes Classifier Algorithm and C4.5 to Predict Student Graduation. In this study, we compared the level of accuracy with 2 methods, namely the nave Bayes algorithm and C4.5 with a difference in the level of accuracy obtained by 0.62% which method uses the C4.5 algorithm.

When compared to other studies used by the author in references to previous research from some of the literature that the author took, almost entirely focused on the methods taken and applied to a system and there were no studies that used a combination of the three methods, namely the nave Bayes algorithm method, C4.5 and K-nearet neighbor where the three combinations of methods are used to analyze the data to produce a high level of accuracy then the method that has a high level of accuracy will be used as a reference for majoring. So that is where the originality value of the research conducted in this study lies.

Naïve Bayes Algorithm

Naïve Bayes is one of the algorithms contained in the classification technique and the algorithm is a classification with probability and statistical methods discovered by British scientists Bayes, which predicts future opportunities based on previous experience. It is, therefore, known as Bayes' Theorem. The following is Bayes' theorem (Supriyanto, 2013):

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

where,
X : data with unknown class
H : data hypothesis X with a specific class
P(H|X) : probability of hypothesis H based on condition X (posteriori probability)
P(H) : hypothesis probability H (prior probability)
P(X|H) : probability X based on condition on hypothesis H
P(X) : probability of X

C4.5 . Algorithm

The C4.5 algorithm is an algorithm that is widely known and used for data classification that has numeric and categorical attributes. The results of the classification process in the form of rules can be used to forcast the attribute values of discrete types and new records. The C4.5 algorithm itself is a development of the ID3 algorithm, where development is carried out in terms of being able to overcome missing data, being able to overcome continuous data and pruning. In general, the C4.5 algorithm for building a decision tree is as follows:

1. To select attribute as root.
2. To create a branch for each value
3. To split cases in branches.
4. To repeat the process for each branch until all cases on the branch have the same class.

To select the root attribute, it is based on the highest gain value of the existing attributes. To determine the gain, the formula as shown in the following equation is used:

Gain (sa) = Entropy (s) – Zi; =I Entropy (Si) (2)

where,
S : case set
A : attribute
N : number of partition attribute A
Si\ : number of cases on partition i
|*Si*| : number of cases in S

K-Nearest Neighbor Algorithm

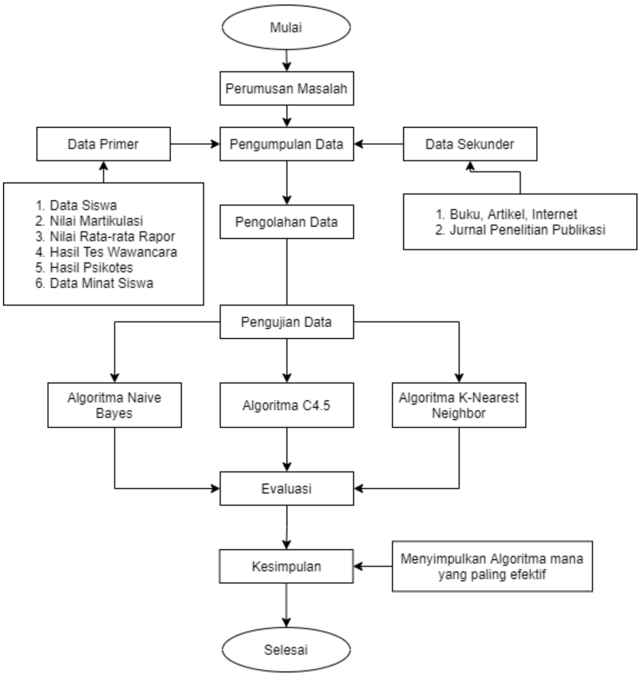
In general, data mining has techniques used in classifying and the approach of these techniques consists of 2 techniques, namely, supervised learning and unsupervised learning techniques. Supervised learning is an approach where the K-Nearest Neighbor Algorithm is a method that observes a supervised algorithm. The difference between supervised learning and unsupervised learning is that supervised learning aims to find new patterns in the data by connecting existing data patterns with new data. While in unsupervised learning the data does not yet have any pattern, and the purpose of unsupervised learning is to find patterns in a data. In this new student admissions study, a supervised learning algorithm is used. The K-Nearest Neighbor method works based on the assumption that a data will have the same class or category as the data around it. This concept is known as the neighboring concept (Adinugroho and Sari, 2018). In this research, RapidMiner Studio is software or software that is open source. Rapidminer is a solution for analyzing data mining, text mining and predictive analysis. RapidMiner also uses a variety of descriptive and predictive techniques to provide users with insights so they can make the best decisions.

METHODS

Research Design, Data collection and Processing

RapidMiner Studio is software or software that is open source and the Rapidminer is a solution for analyzing data mining, text mining and predictive analysis. RapidMiner uses a variety of descriptive and predictive Techniques to provide users with insights so they can make the best decisions. This research design is structured can be seen in Figure 1.

At this stage the thing that needs to be done to analyze a major is to collect data. The data collection is carried out directly in the field, namely by collecting new student data for the 2021/2022 academic year obtained from the academic section. Data collection aims to obtain real data that will be processed for analysis. The data obtained were 218 new student data sets with 11 attributes or variables. The attributes used are Name, NISN, Gender, Average Score Report, School Origin, Student Interests, Interview Results, IQ, Psychological Test Results, Science Marticulation, Social Studies Marticulation. The results of data collection obtained a record of 218 new student data sets based on the number of students who entered SMA 1 Barunawati in the 2021/2022 school year with 11 attributes. Students who are categorized as science or social studies are by comparing the data on student interest with some data and test results that have been carried out. However, from the results of data collection, data records and attributes cannot all be used because it is necessary to preprocess data or initial data processing to get good data. Based on the data before preprocessing the data, the writer needs to pre-process the data to get good quality data. The pre-processing that the author uses includes data cleaning, which is removing empty and incomplete data. For example, records of students who have discontinued or during the testing process resign or withdraw the registration file. So the data that originally amounted to 218 becomes only 214 data sets, so as much as 1.83% of data are blank and incomplete are cleaned at the pre-data review stage which aims to avoid missing values in the data set.



Source : research 2022

Figure 1. Flow Chart of Research.

Data Test

At this stage, data testing is carried out using 214 data that have been cleaned of incomplete data, testing is then carried out to find and produce the best method from the method that is analyzed by comparing the values of precision, recall and accuracy of each algorithm. Here are formulas for determining accuracy, precision and recall.

1. Accuracy

   Accuracy is the ratio of Correct predictions (positive and negative) to the overall data. The accuracy formula is:

$$\text{Accuracy}=\left(\frac{\text{True Positif} + \text{True Negatif}}{\text{Total data}}\right) \text{ x } 100\% \qquad (3)$$

2. Precision

   Precisionis the ratio of positive correct predictions to the overall positive predicted outcome. The precision formula is:

$$\text{Precision} =\left(\frac{\text{True Positif}}{\text{True Positif} + \text{False positif}}\right) \text{ x } 100\% \qquad (4)$$

3. Recall

   Recall is the ratio of true positive predictions to the total number of true positive data. The recall formula is:

$$\text{Recall} =\left(\frac{\text{True Positif}}{\text{False Positif} + \text{True Negatif}}\right) \text{ x } 100\% \qquad (5)$$

Analysis Techniques

This research technique uses the Naïve Bayes Algorithm, C4.5 Algorithm, and K-Nearest Neighbor Algorithm. Naïve Bayes is a classification model that uses probabilities that can be used for quantitative and qualitative data and does not require a large amount of data to run by taking into account speed and time efficiency. While C4.5 is one of the algorithms used to perform classification or segmentation or grouping and is predictive, the K-Nearest Neighbor algorithm is a classification method for a set of data based on previously classified data learning.

RESULT AND DISCUSSION

Research data

Research data of 240 students that have been obtained through the observation and documentation phase during July - September 2021 at SMA 1 Barunawati Jakarta. The student data taken were 240 students for the 2021/2022 academic year.

Tabel 1. Dataset of Students

Name	Gender	School Origin	IQ	Average grades	Major interest	Interview result	Psychotest result	Matriculation of Science	Matriculation of Social	Final Result
SISWA 1	M	Negeri	112	84.3	1	1	1	57	54	IPA
SISWA 2	M	Swasta	100	77	2	2	2	39	35	IPS
SISWA 3	M	Negeri	106	79.5	1	1	1	59	64	IPA
SISWA 4	F	Swasta	100	57.9	1	1	2	59	54	IPA
SISWA 5	M	Negeri	112	85	1	1	1	75	57	IPA
SISWA 6	M	Negeri	100	81.67	1	1	2	56	50	IPA
SISWA 7	M	Negeri	100	79.27	2	2	2	23	54	IPS
SISWA 8	F	Negeri	106	81.57	1	1	1	70	43	IPA
SISWA 9	M	Swasta	90	77.4	2	2	2	20	27	IPS
SISWA 10	M	Negeri	90	74	2	2	2	34	14	IPS
.										
.										
.										
SISWA 240										

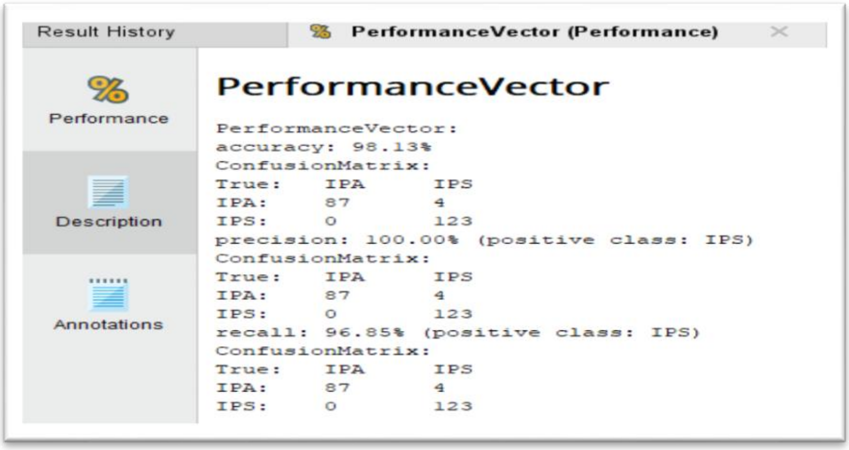
Source : Research 2022

Algorithm Performance Measurement Using Confusion Matrix

Measuring the performance of an algorithm is important to find out how well a system organizes data. Confusion Matrix is a comparison of the classification results performed by the system (model) with the actual classification results. The Confusion Matrix is in the form of a matrix table that describes the performance of the classification model on a series of test data whose actual values are known. Majoring data processing with classification method uses 3 algorithm models, the results are as follows:

Results of the Naïve Bayes Algorithm

The measurement of classification performance from the results of the Naïve Bayes algorithm uses a confusion matrix, namely, by calculating the values of accuracy, precision and recall. The value of the prediction results as follows:



PerformanceVector

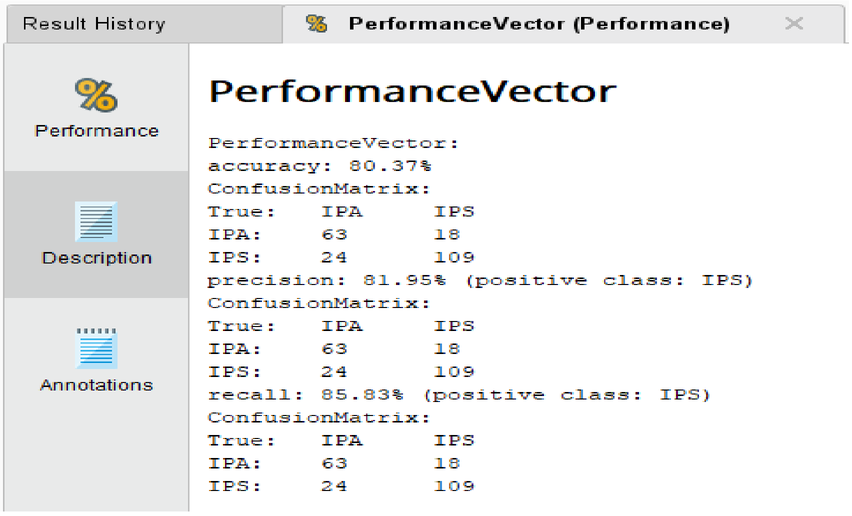
```
PerformanceVector:
accuracy: 98.13%
ConfusionMatrix:
True:   IPA     IPS
IPA:    87      4
IPS:    0       123
precision: 100.00% (positive class: IPS)
ConfusionMatrix:
True:   IPA     IPS
IPA:    87      4
IPS:    0       123
recall: 96.85% (positive class: IPS)
ConfusionMatrix:
True:   IPA     IPS
IPA:    87      4
IPS:    0       123
```

Source : Research 2022

Figure 2. Performance Vector of Algorithm Naïve Bayes.

C4.5 Algorithm Results

The classification performance measurement from the results of the C4.5 algorithm applies a confusion matrix, namely, by calculating the values of accuracy, precision and recall. The value of the extrapolation results as follows:



PerformanceVector

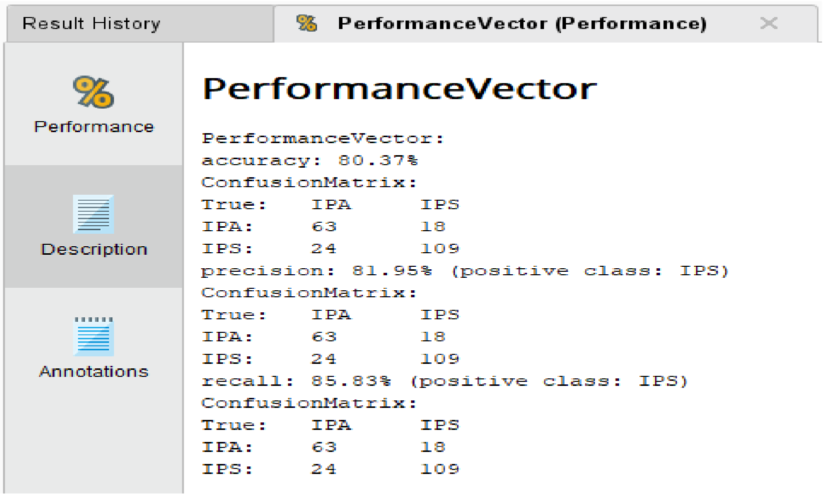
```
PerformanceVector:
accuracy: 80.37%
ConfusionMatrix:
True:   IPA     IPS
IPA:    63      18
IPS:    24      109
precision: 81.95% (positive class: IPS)
ConfusionMatrix:
True:   IPA     IPS
IPA:    63      18
IPS:    24      109
recall: 85.83% (positive class: IPS)
ConfusionMatrix:
True:   IPA     IPS
IPA:    63      18
IPS:    24      109
```

Source : Research 2022

Figure 3. Performance Vector of Algorithm C4.5.

K-Nearest Neighbor Algorithm Results

Measurement of classification performance from the results of the K-Nearest Neighbor algorithm using a confusion matrix, namely by calculating the values of accuracy, precision and recall. The value of the calculation results as the followings:



PerformanceVector

```
PerformanceVector:
accuracy: 80.37%
ConfusionMatrix:
True:   IPA     IPS
IPA:    63      18
IPS:    24      109
precision: 81.95% (positive class: IPS)
ConfusionMatrix:
True:   IPA     IPS
IPA:    63      18
IPS:    24      109
recall: 85.83% (positive class: IPS)
ConfusionMatrix:
True:   IPA     IPS
IPA:    63      18
IPS:    24      109
```

Source : Research 2022

Figure 4. Performance Vector of Algorithm K-Nearest Neighbor.

Results Comparison of the Three Methods

After implementing majors using the three methods, namely the Naïve Bayes algorithm, C4.5 and K-Nearest Neighbor using the RapidMiner application, as seen in the Figures 2, 3 and 4, the best method is using the Naïve Bayes algorithm. The results of data processing classification algorithm Naïve Bayes, C4.5 and K-Nearest Neighbor comparison based on the level of accuracy. The higher the value of accuracy obtained from the test results, the higher the value of determining the resulting direction. The comparison of the accuracy values of the results of data processing for determining majors using the classification method among the nave Bayes algorithm, C4.5 and K-Nearest neighbor can be seen in Table 2.

Table 2. Comparation of Accuracy Values for the Three Methods.

Algorithm	# Samples	Results of Interests		***Values of Confusion Matrix*** **(%)**		
		True	***False***	***Accuracy***	***Precision***	***Recall***
Naïve Bayes	214	210	4	**98.13**	100.00	96.85
C4.5	214	196	18	**91.59**	97.39	88.19
K-Nearest Neighbor	214	172	42	**80.37**	81.95	85.83

Source : Research 2022

As seen in Table 2 and based on the comparison of the accuracy level of determining the direction using the classification method between the nave Bayes algorithm, C4.5 and K-nearest neighbor, it is shown that the classification method using the Naive Bayes algorithm is a classification method with the highest level of confusion matrix compared to the C4 algorithm C4.5 and K-nearest neighbor which get values of accuracy, precision and recall are respectively 98.13%, 100% and 96.85%. It can be finally concluded that to determine students' majors, the most appropriate algorithm is to use the Naïve Bayes algorithm.

Comparison of Random Data Results

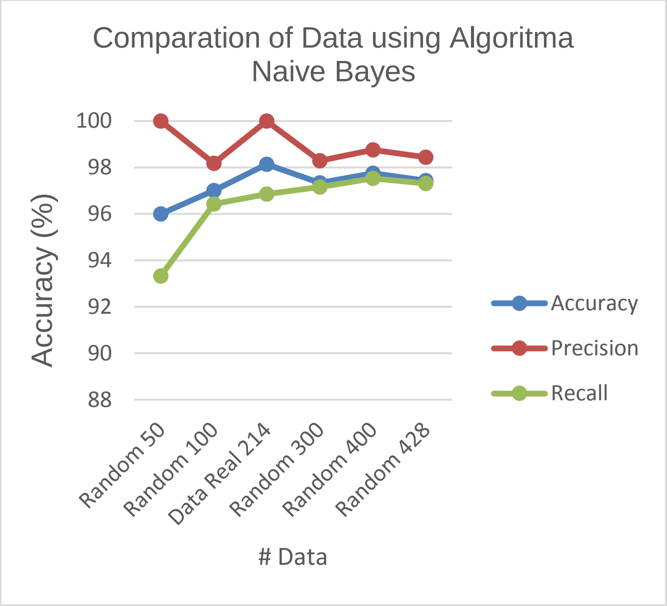
After implementing the majors using the three methods, namely, the Naïve Bayes algorithm, C4.5 and K-Nearest Neighbor, the best method is to use the Naïve Bayes algorithm with the highest level of accuracy. To prove that the Naive Bayes algorithm is indeed the most suitable algorithm in this case, it is necessary to prove it by testing with random data that will be compared with the results of real data processing. The followings are the results of testing with random data using the nave Bayes algorithm which were taken and reproduced by doubling the data randomly which can be seen in Table 3.

Table 3. Confusion Matrix with Random Data Random using Algorithm Naïve Bayes.

Classification of Sample	**# Sample**	**Results of Interests**		**Confusion Matrix Values (%)**		
		True	**False**	**Accuracy**	**Precision**	**Recall**
Random	50	48	2	96.00	100.00	93.33
Random	100	97	3	97.00	98.18	96.43
Data Real	214	210	4	98.13	100.00	96.85
Random	300	292	8	97.33	98.28	97.16
Random	400	391	9	97.75	98.75	97.53
Random	428	417	11	97.43	98.44	97.30

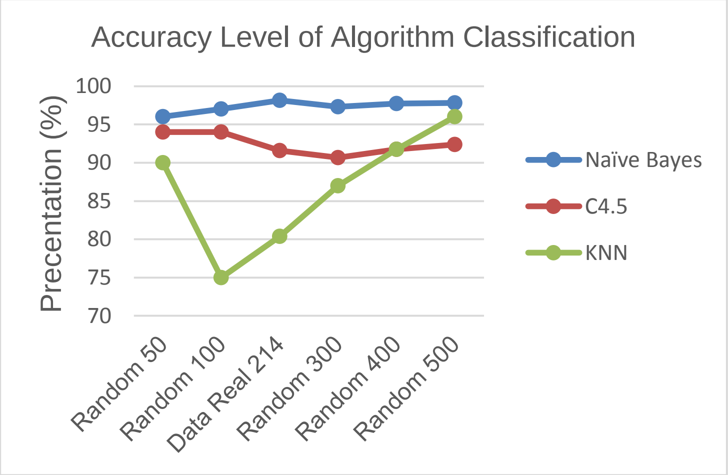
Source : Research 2022

The results of the comparison of the confusion matrix data with real data and random data can be displayed in graphical form as shown in Figure 5.



Source : Research 2022

Figure 5. Comparation of *Random Data using* Algoritma *Naïve Bayes*



Source : Research 2022

Figure 6. Accuracy Level of Algorithm Classification.

CONCLUSIONS

Based on the results of research conducted at SMA 1 Barunawati Jakarta using the classification method from the amount of data tested, 214 datasets, the Naïve Bayes algorithm have the highest level of accuracy compared to the C4.5 and K-Nearest Neighbor algorithms with the accuracy value of 98.13%. It is also proved by comparing the data by testing for each algorithm using real and random data with 50, 100, 214, 300, 400 and 500 datasets. Finally, it can concluded the Naïve Bayes algorithm is the best method for the research regarding students intersts for science and social in SMA 1 Barunawati Jakarta.

ACKNOWLEDGEMENT

The author would like to thank the entire academic community of the Informatics Engineering Magister Program of the University of Pamulang (UNPAM) and also the entire ranks of SMA 1 Barunawati Jakarta so that this research has been implemented properly and met with the author's expectations.

REFERENCES

Adinugroho, Sigit & Arum Sari, Yuita. (2018). *Implementasi Data Mining Menggunakan Weka.* Universitas Brawijaya Press, ISBN 978-602-43-2445-2

Aditya Maulana Habibi, Reva Ragam Santika (2020). Implementasi Algoritma K-Nearest Neighbor dalam Menentukan Jurusan Menggunakan Metode Euclidean Distance Berbasis Web Pada SMP Setia Gama. SKANIKA, Vol. 3, No. 4, Juli 2020, 7-14, E-ISSN: 2721-4788

Ahmad Zainul Mafakhir, Achmad Solichin (2020). Penerapan Metode Naïve Bayes Classifier untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta. Fountain of Informatics Journal, Volume 5, No. 1, Mei 2020, ISSN: 3652-4313 (print) / 25485113 (online)

Bramer, M. 2007. *Principles Of Data Mining*. London: Springer-Verlag London Limited

Christiandita Rahayuningtyas, Dedy Satrio Winarso (2017). Implementasi Algoritma k-Nearest Neighbor untuk Penjurusan Siswa SMA. Cahayatech Vol.6, No. 02, September 2017 ISSN : 2302 – 2426

Endang Etriyanti, Dedy Syamsuar dan Yesi Novaria Kunang (2020). Implementasi Data Mining Menggunakan Algoritme Naïve Bayes Classifier dan C4.5 untuk Memprediksi Kelulusan Mahasiswa. Telematika. Vol. 13 No. 1, Februari 2020 pp. 56-67, e-ISSN 2242-4528, p-ISSN 1979-925X

Gani, Ruslan A. (1986). *Bimbingan Penjurusan*. Bandung : Angkasa, ISBN : 979-404-149-1

Sumarni Adi, Jurnal Mantik Penusa (2018). Prediksi dalam Penjurusan Siswa Baru Tingkat SMA Menggunakan Algoritma Naïve Bayes Classifier. Vol. 2, No. 2, Desember 2018, e-ISSN 2580-9741, p-ISSN 2088-3943

Supriyanto, Catur. Purnama Parida. (2013). Deteksi Penyakit Diabetes Type II Dengan Naïve Bayes Berbasis Particle Swarm Optimization. Jurnal Teknologi Informasi, Volume 9 Nomor 2, Oktober 2013, ISSN 1414-9999

Zuleha (2020). Penentuan Jurusan Sekolah Menengah Atas Menggunakan Metode K-Nearest Neighbor Classifier Pada SMAN 2 Singingi. JuPerSaTeK, Vol. 3 No. 1, Juli 2020, hal. 199-206, ISSN : 2622-108X.