SEMESTA TEKNIKA
Vol.23, No.1, 1-13, Mei 2020
DOI: 10.18196/st.231251

Classification of Student Majors with C4.5 and Naive Bayes Algorithms (Case Study: SMAN 2 Bekasi City)

ANTONIUS YADI KUNTORO, HERMANTO, TAUFIK ASRA , FERRY SYUKMANA,
HERMANTO WAHONO

ABSTRACT

The interest majors process carried out in high school or high school / MA is certainly by the interests of the students and is carried out to provide opportunities for students to develop competency attitudes and knowledge of the interests of students. The mismatch of student competencies towards the majors they take will affect the success of student learning. Students' competency skills are following their interests, talents, and academic abilities in a group of scientific subjects. In this study, researchers used two algorithm models, namely a comparison between the C4.5 algorithm and the Naive Bayes algorithm. This study also aims to build a system that can provide high school / MA majors recommendations based on the classification of students' abilities, talents, and interests. These classifications are needed by students in facing obstacles in choosing a major. The data used in this study are the results of the school entrance test data and also the data from the results of psychological assessments for students who have been declared to have passed the entrance test for the SMAN 2 Bekasi City academic year 2018/2019. The chosen research method is using CRISP-DM with six stages (North, 2012). By comparing the two data mining classification algorithms, For the highest accuracy, the NB algorithm = 76.43% and the AUC value = 0.846, while for the accuracy of the algorithm C4.5 = 70.29% and the AUC value = 0.738.

Keywords: Algorithm C4.5; Naïve Bayes; Student majors, AUC

INTRODUCTION

In the new curriculum, the majors are conducted at the beginning of the school entry, namely in class X. Changes in the curriculum are intended to enable the adjustment of educational programs to educational units with the potential conditions and peculiarities in the student area (Kemdikbud regulations). The possibility that will occur if students experience errors in majors is the low student achievement or can cause incompatibility with the direction chosen by the previous student or student (Education, Culture, & Indonesia, 2013).

In the process of education in school, differences in each student must be considered because it can determine the good and bad of student learning achievement. The basic purpose of the school is to develop all the talents and abilities of students during the education process. Individual differences between students in schools include differences in cognitive abilities, achievement motivation, interest, and creativity. And with these individual differences, the function of education is not only in the teaching and learning process, but also counseling, so that the selection and placement of majors for students must be based on the individual capacity as a student (Bahar, 2011).

Placement of students according to their capacity or often referred to as student majors in secondary school is determined by academic abilities supported by factors of interest. It is because of the characteristics of science require the same characteristics from those who learn it. Thus, students who study science that is in accordance with their personality characteristics will feel happy when studying science. Interest can affect the quality of student learning outcomes in a particular field of study. A student who is interested in Mathematics, for example, will focus more on the Mathematics field than others. Because of the concentration

of intensive attention to the material, students will study harder and achieve desired achievements (Bahar, 2011).
The incompatibility of student's competencies towards the majors they take is that the decision maker must really consider the criteria that have been set in the decision-making department. This will affect the success of student learning. Determination of majors is a problem experienced by students who want to continue their education to a higher level (Hertyana, 2018).

From the results of interviews with the school in determining the student's specialization department conducted by SMAN 2 Bekasi City, there are only two majors, namely majors in Natural Sciences and Social Sciences. Determination of these majors is considered based on the entrance test scores of students of class X who pass the entrance examination, as well as student's interests and talents seen from the psychological test results after they have passed the entrance examination. However, there are still lots of students who ignore their abilities or desires because they prefer following friends' choice since they want to be in the same class or other possible factors.

However, Classification and Clustering activities carried out by humans still have some limitations, especially in the ability of humans to accommodate the amount of data they need to process. In addition, errors can also occur due to inaccuracies carried out. One way to overcome this problem is to use data mining techniques to process data into strategic information sources using Classification and Clustering methods. Data mining can help an organization that has abundant data to provide information that can support decision making. (Nugroho, 2015).

The following are some of the literature that can be related to this study.

1. Data Mining Process

Data Mining is a process that must be in accordance with the procedures of the process itself, this process is called The Cross-Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM methodology is an effort to standardize the data mining process. The process model in CRISP-DM provides an overview of the life cycle of a data mining project. The process contains the project phase, each task and the relationship between tasks.
In CRISP-DM, there are six phases that are interconnected to describe the data mining process, namely: business understanding, understanding data, data preparation, modeling, evaluation, and deployment (Maimon & Rokach, 2010), as in the following Figure 1.

2. Algorithm C4.5

The C4.5 algorithm was introduced as an improved version of ID3. In ID3, induction of decision trees can only be done on features of the categorical type (nominal and ordinal), while the numeric type (interval or ratio) cannot be used. The improvement that distinguishes C4.5 algorithm from ID3 is that it can handle features with numeric types, pruning decision trees, and deriving rule sets. The C4.5 algorithm also uses gain criteria in determining the features that become node breakers in the induced tree (Asroni et al., 2018).

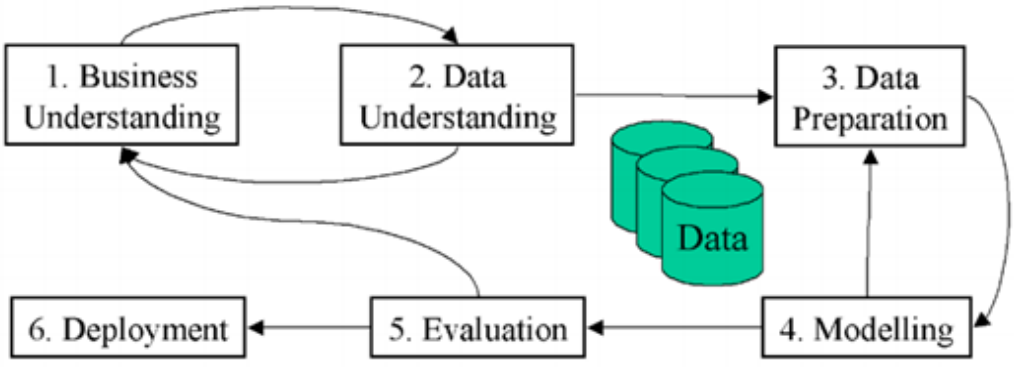

FIGURE 1. Process CRISP-DM
Source: (Maimon & Rokach, 2010)

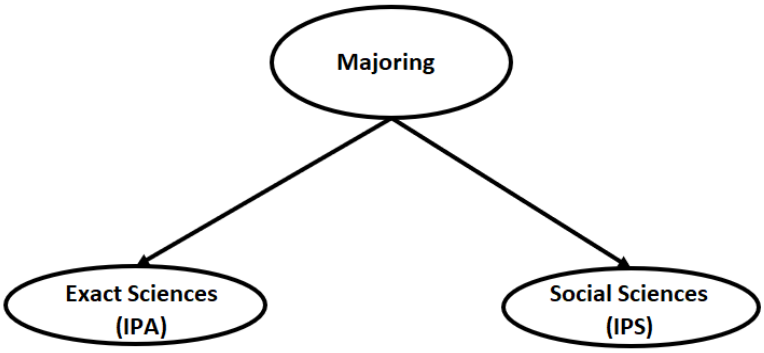

FIGURE 2. Terms of testing binary features (Asroni et al., 2018)

What is important in the induction of decision trees is how to declare the testing conditions on the node. There are 3 important groups in node testing requirements:

a. Binary Features
Features that only have two different values are called binary features. Testing conditions when this feature becomes a node (root or internal) only have two options. Examples of solutions are presented in Figure 2.

b. Categorical type features
For features that have categorical type values (nominal and ordinal) can have several different values. An example is the 'weather' feature has 3 different values, and this could have many combinations of test testing requirements. In general, there are two, namely binary splitting and multi splitting.

c. Numerical type features
For numerical features, the testing requirements in nodes (root or internal) are expressed by comparison testing (A <v) or (A ≥ v) with binary results, or for multi with results in the form of a range of values in the form $v_i \leq A < v_i + 1$, for i = 1, 2, ..., k. In the case of binary solving, the algorithm will check all possible solving positions v and choose the best position v. For multi-methods, the algorithm must check all possible continuous values.

2. Naïve Bayes

The Naive Bayes algorithm is a statistical classification that can be used to predict the probability of a class membership. According to Wu and Kumar, Naive Bayes is a popular classification method and is included in the ten best algorithms in data mining. Naive Bayes uses a branch of mathematics known as probability theory to find the greatest probability of possible classifications, by looking at the frequency of each classification in the training data. (Hermanto et al., 2019).

3. Rapid Miner

RapidMiner is a software that is open (open source). RapidMiner is a solution for analyzing data mining, text mining and prediction analysis. RapidMiner uses a variety of descriptive and predictive techniques to provide insight to users so that they can make the best decisions. RapidMiner has approximately 500 data mining operators, including operators for input, output, preprocessing data and visualization. RapidMiner is stand-alone software for data analysis and as a data mining machine that can be integrated into its own products. RapidMiner is written using Java language so that it can work on all operating systems (Aprilla Dennis, 2013).

4. Confusion Matrix

Confusion Matrix (Gorunescu, 2011) The Confusion Matrix is a very useful tool for analysing how well classifiers can recognize tuples from different classes. Evaluation using the confusion matrix function will produce accuracy, precision, and recall values. One of the advantages of The Confusion Matrix is that it is easy to tell if the data is between two classes (mis-labeling). The Confusion Matrix contains information about actual and predicted conditions in the classification system. System performance like this is usually evaluated using data in a matrix.

The evaluation of a classification model is based on tests to estimate true and false objects, the test sequence is tabulated in a confusion matrix where the predicted classes are shown at the top of the matrix and the observed classes on the left side. Each cell contains a number indicating how many actual cases of the class were observed to be predicted (Hastuti, 2012).

5. *ROC Curve*

ROC Curve (Gorunescu, 2011) is another way to test the performance of classifiers. A ROC chart is a plot with a false positive level (FP) on the X-axis and a true positive level (TP) on the Y-axis. Point (0.1) is a perfect classification that classifies all positive and negative cases correctly, because the positive rate is wrong (FP) is 0 (none), and the true positive level (TP) is 1. Point (0,0) is a classification that predicts each case to be negative, while point (1,1) corresponds to a classification that predicts each case becomes positive. Point (1.0) is an incorrect classification for all classifications.

6. *K-Fold Cross Validation*

K-Fold Cross Validation is a validation technique (Witten et al. In 2011, it was one of 10 00 00 who divided the data into k parts, and then each classification process was carried out. Using K-Fold Cross Validation, K Each experiment will use one testing data and the k-1 part will be the training data, then the testing data will be exchanged with one training data so that for each experiment different testing data will be obtained. For example in the case of 10 fold cross-validation, The data will be divided into 10 sets of parts, then 10 iterations will be carried out for testing and validation.

8. *T-Test*

The T-Test is a method of testing hypotheses using one individual (research object) using two different treatments. Despite using the same object, the sample is still divided into two, namely data with the first treatment and data with the second treatment. Performance can be known by comparing the conditions of the first research object and the condition of the object in the second study (Hastuti, 2012).

9. *Classification*

Classification is a process of finding a collection of patterns or functions that describe and separate data classes from one another to state that the object belongs to a certain predetermined category. Classification is one of the most common objectives in data mining. Classification is the process of grouping a variable into predetermined classes. A model in classification has the same meaning as a black box, where there is a model that accepts input and can then think about the input and provide answers as the output of the results of their thoughts (Asroni et al., 2018)..

RESEARCH METHOD

The research method used in this experiment using a standard methodology in data mining research is the Cross-Standard Industry for Data Mining (CRISP-DM) model. CRISP-DM is a collaboration of several companies, including Daimler-Benz, OHRA, NCR Corp, and SPSS Inc. which started since 1999 (North, 2012).

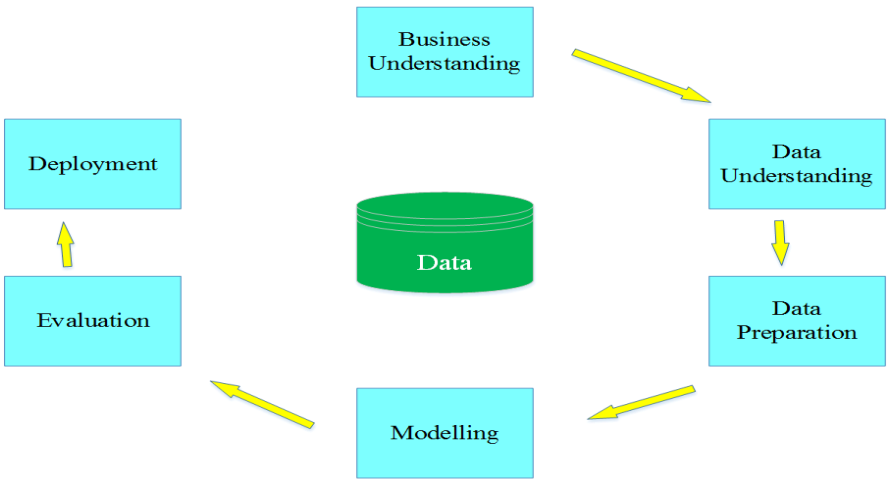

FIGURE 3. CRISP-DM (North, 2012)

Figure 3, will explain regarding CRISP-DM has six stages (North, 2012), namely:

a. Business Understanding
At the business understanding stage, an understanding of the object of research is carried out. In this study, the authors used data from prospective new students of SMAN 2 Bekasi City in knowing the 2018/2019 lessons.

b. Understanding Data
This stage is the process of understanding the data that will be used as the material to be researched so that it can be carried out to the next stage, namely processing.

c. Data Preparation
The next stage is to prepare the data before the data will be modeled or known as Data Preparation. For this second stage, namely preparing data to perform steps called preprocessing, using the GataFramework website can be used for free and easy to use because you don't have to create an account to use the service and continue preprocessing from rapidminer.

d. Modeling
This stage is also called the learning stage because at this stage the training data is classified by a model and then results in a decision. In this study, the modeling used the Rapid Miner application.

e. Evaluation
At this stage, model testing is carried out to obtain accurate model information. In this case, using confusion matrix and AUC

f. Deployment
After forming the model and analyzing and measuring in the previous stage, then an application is made using the best algorithm in this study.

1. CRISP-DM Stages

The research method used in this study is to use the experimental method. This study aims to compare and evaluate data mining classification algorithms in determining majors in at SMAN 2 Bekasi city. In this study using the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, which consists of 6 stages, namely: Business Understanding, Understanding Data, Data Preparation, Modeling, Evaluation, and Deployment.

2. Framework

In completing this study, the author makes a framework of thinking that is useful as a guideline or reference in this study so that this research can be done consistently. This study consists of several stages as seen in Figure 3.2 frame of mind. The problem in this study is whether the C4.5 and Naive Bayes algorithms can be applied to the determination of the science class majors and the IPS class at SMAN 2 Bekasi city and which algorithms will provide the best model for classifying the majors in SMAN 2 Bekasi city.

For this reason, a model is made using the C4.5 and NB algorithms to solve the problem, then test the two performance models. After testing the two models formed, 10 fold cross validation will be tested. The accuracy of the two models that have been formed will be measured by using a confusion matrix, while the Under Curve Area (AUC) will be measured using ROC Curve. To develop the application (deployment) based on the model that has been made, the Rapid Miner 8.0 tools are used. The following is a description of the framework that has been carried out as follows, in Figure 4:

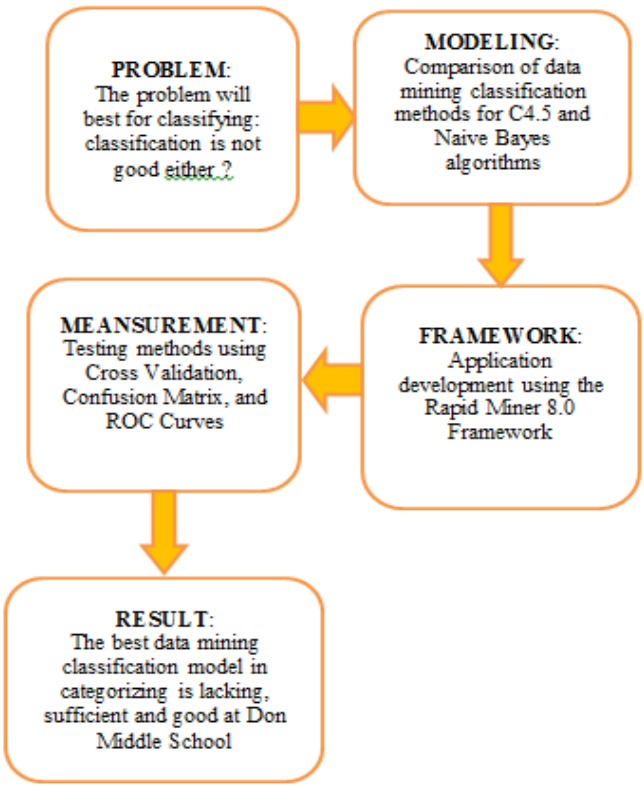

FIGURE 4. Framework

RESULT AND DISCUSSION

Based on existing research data as well as the literature, the following are the stages and results of data testing.

1. *Implementation Of The Methodology*

Based on the research methodology described in chapter III, the following methodology implementation was carried out in this study.

2. *Research Methods*

In this study, the authors conducted a study using the CRISP-DM research method (Cross-Standard Industry for Data Mining). The Stage of Crisp-Dm consists of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment (Brown, 2014).

3. *Business Understanding*

The first stage of CRISP-DM is Business Understanding which is to understand student data at SMAN 2 Bekasi City in determining its special majors new class X students entering the determinants of the majors only refer to the results of the psychological test only, hence researchers here propose a new attribute that is the result of test entrance tests for X grade students used with the results of psychological tests to determine the results of student majors. Which will be tested by comparing two algorithms namely C4.5 and NB.

4. *Data Understanding*

At this stage, the author examines the data of new prospective students of SMAN 2 Bekasi City in knowing the 2018/2019 lesson. The data taken is the results of the school entrance test data, and psycho-test of new students of SMAN 2 Bekasi City. The contents of entrance test result are subject values that were tested include Bahasa Indonesia, English, Mathematics, Physics, Biology, and Religion. The data used in this study is the data from the X class entrance test results at SMAN 2 Bekasi city and also the psychological test results of prospective class X students in the 2018-2019 academic year with a total of 148 students from two science classes and three social studies classes. Where there were as many as 60 female students and 88 female students. There were 83 students in the science class and 65 in the social studies class.

The following is the initial data before the value of the numbers is converted, see Table 1 below. Table 2 above is the amount of data for class X students in the 2018/2019 academic year received at SMAN 2 Bekasi City in the IPA and IPS classes which the data will be processed. Class X psychological test results for school year 2018/2019, as in the Table 3.

TABLE 1. Examples of Test Exam Values

No	Mathematics	Biology	English	Physics	Bahasa	Religion
1	15	14	14	17	17	23
2	4	13	9	13	17	18
3	14	17	5	17	17	21
4	13	13	16	11	18	17
5	6	17	10	17	17	20
6	9	12	7	9	13	14
7	15	16	9	16	17	23
8	7	11	16	16	19	19
9	11	14	8	10	15	21
10	5	16	13	16	17	21

TABLE 2. Table Number of Students

Class	The number of students
X IPA 1	33 students
X IPA 2	31 students
X IPS 1	30 students
X IPS 2	28 students
X IPS 3	26 students

TABLE 3. Sample Table of psychological test results

No	Gender	Majoring
1	F	IPS
2	F	IPA
3	F	IPA
4	M	IPS
5	M	IPS
6	M	IPA
7	M	IPA
8	F	IPA
9	F	IPS
10	F	IPS

5. Data Preparation

The third stage of CRISP-DM is Data Preparation. This stage prepares data so that it can be processed at the next stage. This stage adjusts the attributes of the data table that will be processed using the Rapid Miner 8.0 Framework. Several tables that have been described will be processed in this study and in this phase, the author creates a new table combining the tables into one table. The data in Table 4 below is the result of merging from the previous data tables that will be included in the process using the Rapid Miner 8.0 Framework.

In the student data Table 4 above will explain, the researcher reprocessed the data by converting redundant values or values that were too diverse into smaller groups to facilitate the formation of the model. For example, a value with a number of ≤ 2 is categorized sufficient and ≥ 17 is categorized as very good, Table 5 is a table of categorization of attributes can be seen below

6. Modeling

The fourth stage in CRISP-DM is Modeling. At this stage, the dataset that was made in the previous stage is used as input for the classification algorithm, which is used as a training dataset. In this study, two types of classification algorithms will be used, namely C4.5 and NB. The following is the design process that is used along with the description as in the Figure 5.

TABLE 4. Student data table

No	Mathematics	Biology	English	Physics	Bahasa	Religion	Gender	Majoring
1	9	9	10	12	16	20	F	IPS
2	15	16	9	15	17	23	F	IPA
3	5	16	13	16	17	21	F	IPA
4	7	13	7	13	13	20	M	IPS
5	9	6	13	15	20	16	M	IPS
6	14	17	5	17	17	21	M	IPA
7	7	11	16	16	19	19	F	IPA
8	15	14	14	17	17	23	F	IPA
9	9	12	9	12	18	14	F	IPS
10	10	14	11	12	18	21	M	IPS

TABLE 5. Attribute category

Score	Category
2 – 9	Enough
10 – 16	Good
17 – 24	Very Good

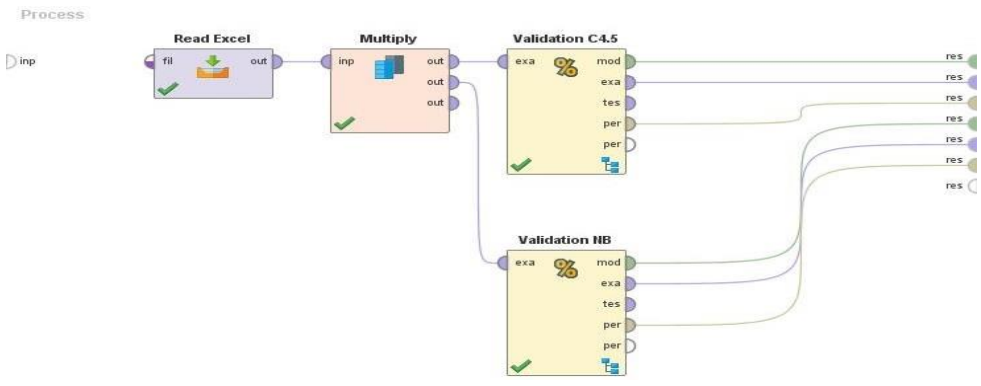

FIGURE 5. Rapidminer Algorithm Model

7. Evaluation

The fifth stage of the CRISP-DM method is Evaluation aiming to determine the usefulness of the model we have succeeded in making in the previous step. For evaluation, 10-fold cross-validation is used. The following is the design process used.

After testing in the Figure 6 with the model above, the results formed will look like in Figure 7 below

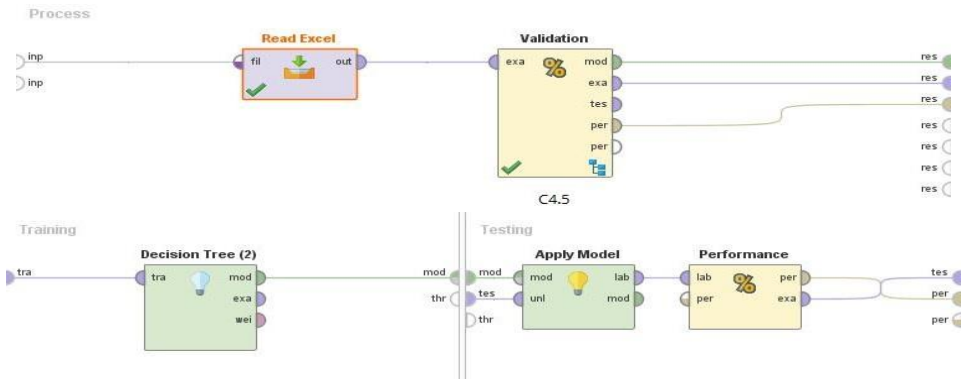

FIGURE 6. Testing the C4.5 Algorithm Model

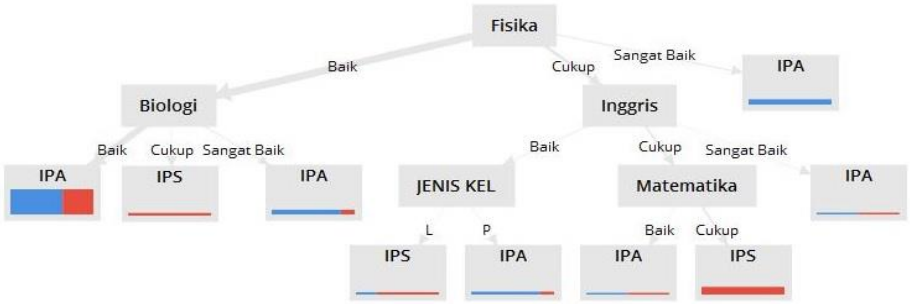

FIGURE 7. C4.5 Algorithm Decision Tree Model

The following are the rules of the C4.5 algorithm tree decision model, as follows:

1. R1: Physics = Good and Biology = Good then Science
2. R2: Physics = Good and Biology = Enough Social Studies
3. R3: Physics = Good and Biology = Very Good then Science
4. R4: Physics = Very Good then Science
5. R5: Physics = Enough and English = Good and Type Kel = L then IPS
6. R6: Physics = Enough and English = Good and Type Kel = P then Science
7. R7: Physics = Enough and English = Enough and Mathematics = Good then Science
8. R8: Physics = Enough and English = Enough and Mathematics = Enough then IPS
9. R9: Physics = Enough and English = Very Good then Science

Based on the results of testing using Rapid Miner, a decision tree and model rules are obtained as seen above with attributes such as Name, Bahasa, English, Mathematics, Physics, Biology, Religion, IQ, Gender, and Department. However, in the decision tree above, not all attributes appear because the attribute has a small gain value.

The results of the model testing that has been done aims to get the results of accuracy and Under Curve Area (AUC) and to get the results of the ROC graph with the value of Area Under Curve (AUC) of 0.738 with accuracy performance, namely Fair Classification, as seen in Figure 8.

FIGURE 8. AUC value in C4.5 Algorithm

From the results of the tests that have been done with the C4.5 Algorithm model in the Figure 8, the accuracy value is 70.29%, as shown in Table 6.

The number of True Positives (TP) is 74 classified records as selected IPA and False Negative (FN) as many as 9 records classified as selected IPS. 35 records for Positive False are classified as selected IPA, and 30 records for True Negative are classified as selected IPS. Based on Table 4.1 above, it shows that the level of accuracy using the C4.5 algorithm is 70.29%. Similar to the evaluation process in the C4.5 algorithm above, the NB algorithm is also evaluated as below in the Figure 9.

TABLE 6. Model for C4.5 Algorithm
Accuracy:70.29% +/- 9.72% (mikro: 70.27%)

	True IPA	True IPS	Class Precision
Pred. iPA	74	35	67.89%
Pred. IPS	9	30	76.92%
Class Recall	89.16%	46.15%	

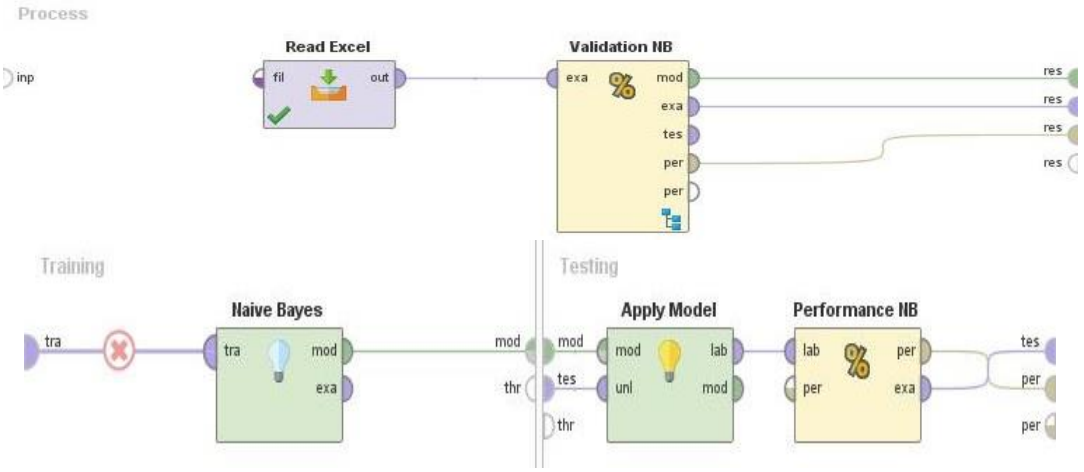

FIGURE 9. Testing the Naive Bayes Algorithm Model

The results of the model testing that has been done aims to get the results of accuracy and Under Curve Area (AUC) and to get the results of the ROC chart with the value of Area Under Curve (AUC) of 0.846 with accuracy performance, namely Good Classification, as seen in Figure 10. From the results of testing that has been done with the NB Algorithm model, the accuracy value is 76.43%, as shown in Table 7.

The number of True Positives (TP) is 69 records classified as selected IPA and False Negative (FN) as many as 21 records classified as selected IPS. In addition, 21 records for Positive False are classified as selected IPA, and 44 records for True Negative are classified as selected IPS. Based on Table 4.6 above, it shows that the level of accuracy using the Naive Bayes algorithm is 76.43%.

8. *Deployment*

This stage is the last stage in the standard modeling in data mining (CRISP-DM). In this stage, the report will be made in the form of writing the results of thesis and journal research from the introduction to conclusions, as well as creating a Graphical User Interface (GUI) so that later, users who use the results of this study can interact and apply it easily.

9. *Comparative Performance*

Based on the results of the analysis of each of the algorithm tests above, the results can be summarized as in Table 8.

FIGURE 10. The AUC value in the Naive Bayes Algorithm

TABLE 7. Model for Naive Bayes Algorithm
Accuracy:76.43% +/- 12.38% (mikro: 76.35%)

	True IPA	True IPS	Class precision
Pred. IPA	69	21	76.67%
Pred. IPS	14	44	75.86%
Class recall	83.13%	67.69%	

TABLE 8. Comparison of Performance Algorithms

	C4.5	**Naive Bayes**
ACCURACY	70,29%	76,43%
AUC	0,738	0,846

TABLE 9. Test results for C4.5 and Naive Bayes T-Tests

A	B	C
	0.729 +/- 0.089	0.764 +/- 0.124
0.729 +/- 0.089		0.469
0.764 +/- 0.124		

10. Analysis of Comparative Results

Based on the results of experiments conducted using data from prospective new students of SMAN 2 Bekasi City in knowing the 2018/2019 subject classification of student majors, wherein this experiment using the C4.5 Algorithm and NB using data from 148 students from two science classes and three study classes social. The test results with the results of the accuracy of the performance classification of the Area Under Curve (AUC), the accuracy of the good algorithm, C4.5, the AUC value = 0.738, while for the NB algorithm, the Good Classification category, the AUC value = 0.846.

After testing with 10Fold Cross Validation in this study, it was tested again using T-Test to test the truth and falseness of the two models. In this test, a comparison between two algorithms will be carried out, namely C4.5 and NB. To get the results of the T-Test statistics calculation on the C4.5 and NB algorithms, it will be seen in Table 9. Based on Table 9 above, it can be analyzed that the C4.5 and NB algorithms have insignificant differences in values, and have a probability of> 0.05, which is 0.469.

CONCLUSION

Based on the results of the research, it can be concluded that: from the results of the tests that have been produced, the accuracy value for determining student majors at SMAN 2 Bekasi City with a comparison of the two data mining classification algorithms, can be proven by the results of the accuracy and AUC value of each algorithm. The highest accuracy results are 76.43% and the AUC value = 0.846, namely the Naive Bayes algorithm, while for the C4.5 algorithm test results for accuracy is 70.29% and the AUC value = 0.738. The suggestion given by the author to improve the results of this study is the application of the Naive Bayes algorithm in determining high school majors so that they can be more useful for students in the majors. It is hoped that they can provide solutions for students of SMAN 2 Bekasi City in determining majors according to the abilities, interests, and talents of students.

REFERENCES

Aprilla Dennis. (2013). Belajar Data Mining dengan RapidMiner. *Innovation and Knowledge Management in Business Globalization: Theory & Practice, Vols 1 and 2*, *5*(4), 1–5. https://doi.org/10.1007/s13398-014-0173-7.2

Asroni, A., Fitri, H., & Prasetyo, E. (2018). Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik). *Semesta Teknika*, *21*(1), 60–64. https://doi.org/10.18196/st.211211

Bahar. (2011). *Penentuan Jurusan Sekolah Menengah Atas Dengan Algoritma Fuzzy C-Means*. Universitas Dian Nuswatoro Semarang.

Gorunescu, F. (2011). *Data Mining: Concepts, Model and Techniques*. https://doi.org/10.1007/978-3-642-19721-5

Hastuti, K. (2012). Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif. *Semantik*, 241–249.

Hermanto, Mustopa, A., & Kuntoro, A. Y. (2019). Hasil Akhir Penelitian Mandiri: Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa. Jakarta, Indonesia

Hertyana, H. (2018). Analisa Penentuan Jurusan Pada Sma. Kartika Viii-1 Menggunakan Metode Fuzzy Inference System Mamdani. *Jurnal Ilmu*

Pengetahuan Dan Teknologi Komputer, *3*(2), 119–126. Retrieved from http://ejournal.nusamandiri.ac.id/ejurnal/index.php/jitk/article/view/699/409

Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition. In *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition* (Vol. 9780470908). https://doi.org/10.1002/9781118874059

Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. In *Springer Handbook of Geographic Information* (Second). https://doi.org/10.1007/978-0-387-09823-4

North, M. (2012). Data Mining for the Masses. In *Computer*. Retrieved from http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/DataMiningForTheMasses.pdf%5Cnhttps://sites.google.com/site/dataminingforthemasses/

Nugroho, Y. S. (2015). Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *1*(1), 1. https://doi.org/10.23917/khif.v1i1.1175

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tool and techniques* (third Edit). Morgan Kaufmann.

PENULIS:

Antonius Yadi Kuntoro

STMIK Nusa Mandiri, Jalan Kramat Raya No.18 Jakarta

Email: antonius.aio@nusamandiri.ac

Hermanto

Universitas Bina Sarana Informatika, Jakarta.

Email: Hermanto.hmt@bsi.ac.id

Taufik Asra

Universitas Bina Sarana Informatika, Jakarta.

Email: Taufik.tas@bsi.ac.id

Ferry Syukmana

Universitas Bina Sarana Informatika, Jakarta.

Email: ferry.fsk@bsi.ac.id

Hermanto Wahono

STMIK Nusa Mandiri, Jakarta.

Email: mr.h3rm4n@gmail.com