Online Version Available at: https://jurnal.ubd.ac.id/

JOURNAL OF TECH-E

|2581-1916 (Online) | 2598-7585 (Printed) |

Article

Application Of Data Mining For Student Department Using Naive Bayes Classifier Algorithm

Yohana Tri Utami[1], *Debby Alita*[2*], *Ade Dwi Putra*[3]

[1] *Lampung University,Computer Science, Lampung, Indonesia*
[2] *Universitas Teknokrat Indonesia, Informatics, Lampung, Indonesia*
[3] *Universitas Teknokrat Indonesia, Information System, Lampung, Indonesia*

SUBMISSION TRACK

Received: Feb 27, 2022
Final Revision: Mar 07, 2022
Available Online: Mar 25, 2022

KEYWORD

Data Mining, Algorithm, Naïve Bayes Classifier, Student

CORRESPONDENCE

E-mail: debbyalita@teknokrat.ac.id

ABSTRACT

SMAN 02 Negeri Agung does not have a system that can assist schools in determining majors. The problem is that SMAN 02 Negeri Agung, when doing majors, it still uses existing data. For example, using a majoring interest questionnaire, there are questions about the interests that students want, and the values of their junior high school report cards, which consist of Indonesian, Mathematics, Science, Social Studies, and English. However, many students still choose majors not based on their interests or historical grades, such as following friends' choices. It can hinder student academic activities in the future, which will affect the value and development of student potential. This effective system hopes to help schools and students minimize errors in determining and choosing a major. Based on the problems described above, the authors want to apply the Naïve Bayes method, which will produce a high level of accuracy in determining new student majors more effectively and efficiently.The accuracy of the naive Bayes classifier can be stated quite well. It can be seen based on accuracy, 63.46%, error rate 0.3653%, false positive rate 0.2424%, sensitivity 0.6035%, specificity 0.7575%, and precision 0.944% Naive Bayes classifier method can It is recommended to predict student majors.

I. INTRODUCTION

Referring to the 2013 Curriculum regulations, the majors' process is carried out when students sit in class X (ten) in process. Students are allowed to choose a major, be it a science or social studies major, before being re-predicted based on a major decision by the school, taking into account the junior high school grades and the grade of the major test results. For example, grade of C is one of the academic scores that are not optimal, which can impact subsequent academic activities and affect the selection of fields of science or study for students who want to continue to higher education levels later [1].

SMAN 2 Negeri Agung is one of the senior high schools in Way right Regency, which has 2 (two) majors, namely Natural Sciences (MIPA) and Social Sciences (IPS). The major is carried out in grade 10. This student major aims to direct students to focus more on developing their abilities and interests. SMAN 2 Negeri Agung does not yet have a system that can assist the school in determining majors. The problem is that SMAN 2 Negeri Agung uses existing data when doing majors. For example, using a majors interest questionnaire, there are questions of interest that students want in the majors' interest questionnaire. Their junior high school report cards consist of Indonesian, Mathematics, Science, and Science scores. IPS, and English. However, many students still choose majors not based on their interests or value history, such as following the choice of friends and so on. That matter can hinder students' academic activities in the future, affecting the value and development of student potential. The majors' system hopes to help the school and students minimize errors in determining and choosing majors. Naïve Bayes is one method that can be used by SMAN 2 Negeri Agung to determine majors. The Naïve Bayes method is one method that can be used in terms of decision making to get better results on a prediction problem.

Based on the problems described above, the authors want to apply the Naïve Bayes method, which will produce a high accuracy in determining new students' majors more effectively and efficiently.

II. LITERATURES REVIEW

Syarli and Muin [2] conducted a study using the Naive Bayes method to predict student graduation. The evaluation results show that the accuracy percentage value indicates the Admissions dataset's effectiveness applied to the Naïve Bayes Classification method, which reaches 94%.

Furthermore, Peling et al. [3] research apply the Naive Bayes To Predict Period of Students Study Using Naive Bayes Algorithm. This study showed that Naïve Bayes was able to classify the proper data testing on average by 86.16% and 13.84% error. In addition, other information obtained from the data testing used that the students who entered from the PMDK Pass graduated on time as much as 40%, other paths graduated on time by 26.7% and passed filter exam on-time 13.3%.

Another study conducted by Naparin [4] used the naive Bayes method to classify the specialization of high school students. The result of the try out by using the Naïve Bayes method to assess high school students' specialization reached the assessment result that has the highest accuracy level 99.47% and AUC value 1,000.

In Putra and Wibowo's [5] research in predicting the decision of majoring in Yadika 5 SMA students using the Naïve Bayes Algorithm, the results showed an accuracy rate of 93.75%, a precision level of 83.33% and a recall rate of 100%.

Furthermore, in research conducted by Hozairi, et al [6] in applying data mining in determining student majors in this study, there were 100 student data used as data to see the accuracy of the Naive Bayes method in classifying student majors and the results from 100 student data tested, there were 90 student data obtained successfully classified with a success percentage of 90% while ten student data were not successfully classified.

Based on the research journal above, this designed system will use the Naïve Bayes method because it is a good algorithm for determining student majors. Results will be validated and measured the accuracy of the results achieved using 10 Fold Cross Validation [7].

III. FRAMEWORK

Here is the framework is shown in Figure 1:

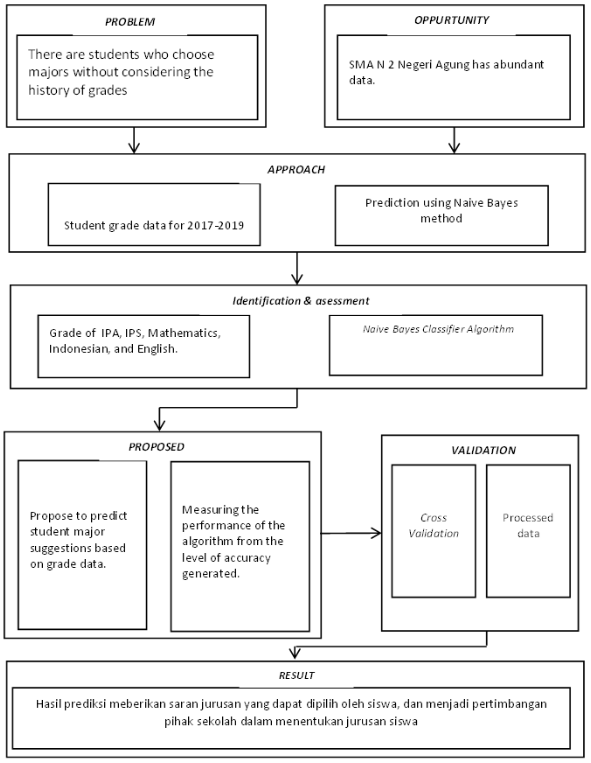

PROBLEM
There are students who choose majors without considering the history of grades
OPPURTUNITY
SMA N 2 Negeri Agung has abundant data.
APPROACH
Student grade data for 2017-2019
Prediction using Naive Bayes method
Identification & asessment
Grade of IPA, IPS, Mathematics, Indonesian, and English.
Naive Bayes Classifier Algorithm
PROPOSED
Propose to predict student major suggestions based on grade data.
Measuring the performance of the algorithm from the level of accuracy generated.
VALIDATION
Cross Validation
Processed data
RESULT
Hasil prediksi meberikan saran jurusan yang dapat dipilih oleh siswa, dan menjadi pertimbangan pihak sekolah dalam menentukan jurusan siswa

Fig 1: Framework

The following is a description of the framework is as follows:

1. Problems

   There are still many students who choose majors not based on their interests or history of grades. It can hinder students' academic activities in the future, which will affect the value and development of student potential. With the majors' system, it is hoped to help the school and students minimize errors in determining and choosing majors.
2. Opportunity

   The abundance of student value data can be processed using the application of data mining which can generate knowledge to provide advice on majors to students.
3. Approach

   The approach stages in this research are how the researchers approach existing problems to find solutions in this research, including using data mining techniques using the Naive Bayes Classifier method [8][9].
4. Identification and Assessment

   The identification stage and research assignments are matters relating to the attributes used in this research to produce information that is by the expected goals, namely to produce information on majors that students can choose.
5. Proposed

   The proposal that will be put forward in this study is to predict student majors based on junior high school report cards using the Naive Bayes Classifier method and knowing the performance of the method used.
6. Validation

   Tests are carried out using manual testing and WEKA tools [9].
7. Result

   The results obtained with this research provide information on major suggestions to students and can be taken into consideration by the school in determining the selection of majors by students. So that it can minimize errors/mistakes in the selection of majors by students.

IV. METHODS

The following are the parts of the research phase carried out:

1. Determine the dataset used, namely student grade data for 2018-2019, with attributes in Table 1 that include the value of junior high school report cards from semesters 1-5 with a total of 491 data. The dataset was obtained from direct observation at SMA N 02 Negeri Agung. In addition, it was also obtained from the results of interviews with the curriculum leader of SMA N 02 Negeri Agung.
2. Select the data to be selected and then clean the data and group the data to make

the prediction process easier. Based on the 491 data used, the features used are the grade of IPS, Bahasa Indonesia (B. Indo), IPA, Mathematics (MAT), and English (B.Ing),Where this feature will be used to classify "Is the student predictable in the science or social studies major?".

3. To make these predictions, the features needed must be categorical. If the feature is not as desired, the transformation process can occur. Feature transformation (FT) is another way to handle heterogeneous feature selection. The transformation method unifies the dataset format and allows conventional feature selection algorithms to handle heterogeneous data sets.
4. In the stage of determining and determining features, it is known that several factors that determine students' majors are subjects related to school majors, namely:

Table 1. Required attributes

Attributes	Description
MAT	It Is the grade of mathematical attributes in semesters 1-5 in the form of unconditional.
IPA	It is a natural science absolute grade in semesters 1-5.
IPS	Is the gradeof social science attributes in the form of unconditional in semesters 1-5?
B. Ing	Is the value of the English attribute in semesters 1-5 in the form of unconditional
B. Indo	It is the value of the Indonesian attribute in semesters 1-5, which is categorical

Table 2 Range Value

Range	Value	Description
89-100	A	Very Good
78-88	B	Good
67-77	C	Average
0-66	D	Bad

5. Based on the training data above, calculations can be made using the Naive Bayes Classifier algorithm, with the following working method:
   For the prediction problem, what is calculated is the probability that the hypothesis is valid (valid) for the observed sample B data, where B is the sample data with an unknown label. At the same time, A hypothesizes that B is the data with a label. P(A) is the probability of hypothesis A, and P(B) is the probability of the observed sample data. Is the probability of sample B data, if it is assumed that the hypothesis is valid. So the formula is as follows:

$$P\ (A|B) = \frac{p(A)p(B|A)}{P(B)} \qquad (1)$$

6. The test data is used to predict the majors of a student in Social Sciences or Science if it is known that the condition of the score:
   IPS = Good, B.indo = Good,
   B.ing = Average, Mat = very good,
   IPA = Good.

Stage 1 counts the number of Classes/labels
P (IPS) = 9/14 = 0.6428
P (IPA) = 5/14 = 0.3571

Counting the number of the same problem with the same class.
By Major:

a. Based on IPS :
   P(Good|IPS) = 4/9 = 0.4444
   P(Good|IPA) = 3/5 = 0.6

b. Based on B. Indo :
   P(Good|IPS) = 9/9 = 1
   P((GoodIPA) = 3/5 = 0.6

c. Based on B. Ing :
   P(Average|IPS) = 1/9 = 0.1111
   P(Average|IPA) = 0/5 = 0

d. Based on Mathematics :
   P(Very Good|IPS) = 1/9 = 0.1111
   P(Very Good|IPA) = 1/5 = 0.2

e. Based on IPA :
P(Good|IPS) = 8/9 = 0.8888
P(Good|IPA) = 4/5 = 0.8

Then calculate the probability or probability of each attribute and multiplied for each equal class:

IPS = P(Ips) x P(Good|IPS) x P(Good|IPS) x P(Average|IPS) x P(Very Good|IPS) x P(Good|Ips)
= 0.6428 x 0.4444 x 1 x 0.1111 x 0.1111 x 0.8888
= 0.003133878

IPA = P(IPA) x P(Good|IPA) x P(Good|IPA) x P(Average|IPA) x P(Very Good) x P(Good|IPA)
= 0.3571x 0.6 x 0.6 x 0 x 0.2 x 0.8
= 0

With the test data above, the prediction of the majors of the data using the Naive Bayes Classifier algorithm results in IPS.

7. The next process is to analyze the data obtained from the school through collecting data sourced from student grade data reports for 2017-2019. Four hundred and ninety-one (491) data were used as datasets based on these data.
8. Then, based on the data that has been obtained, the data is predicted using the Naive Bayes classifier method based on predetermined features. Based on the predicted data set, the majors are predicted for 2021.
9. After modelling, the next step is to validate and measure the accuracy of the results achieved using 10 Fold Cross Validation [10].
10. Evaluating the prediction method using the Confusion Matrix. Confusion Matrix is used to evaluate the prediction results such as the value of Accuracy, Error Rate, False Positive Rate, Recall, Specificity and Precision.

V. DISCUSSION AND RESULT

A. Preprocessing

The results contained in the preprocessing stage include data selection, data cleaning and data grouping. The results of the preprocessing stage are as follows:

1. **Data Selection**
The selection is made on the student's score data obtained. It needs to be done to group or divide the attributes according to the required information. Attributes selected or selected are student report cards from semester 1 (one) to semester 5 (five). The grade of all subjects selected was the grade of Mathematics, the grade of IPA (Natural Science), the grade of IPS (Social Sciences), the grade of English and the grade of Indonesian as the selected attribute.

2. **Data Cleansing**
Data can be clean if it does not contain impurities in the form of empty values and noise and outliers, and/or inconsistencies. Meanwhile, data can be unclean or dirty if it contains impurities in the form of empty values and/or noise and/or outliers and/or inconsistencies [11]. The level of data cleanliness is very influential on whether or not data mining results are good. So that dirty data can be cleaned by filling in empty values, smoothing noisy data, removing outliers, or correcting inconsistencies.
At the data cleaning stage, the researcher did not find ten dirty or unclean data, so the value data from all the selected attributes could be used because they had complete information.
3. **Data Grouping**
The purpose of grouping data is to simplify the prediction process.

Table 3. Ungrouped Grade Data

IPA	IPS	MTK	B.Indo	B.Ing	Label
80	83	77	83	79	IPA
85	83	74	83	82	IPA
80	83	84	84	84	IPS
91	84	77	84	80	IPS
76	86	72	86	81	IPA
80	79	77	82	77	IPS
80	80	82	83	80	IPA
80	83	79	85	81	IPA
70	83	74	87	80	IPS
83	83	74	84	78	IPA

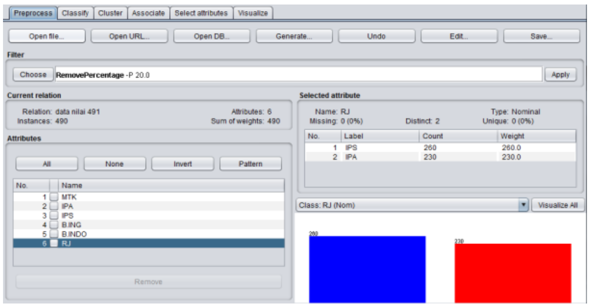
Table 4. Grouped Grade Data

IPA	IPS	MTK	B.Indo	B.Ing	Label
Average	Good	Good	Good	Good	IPA
Poor	Good	Good	Good	Good	IPA
Poor	Good	Good	Good	Good	IPS
Average	Very Good	Good	Good	Good	IPS
Poor	Good	Good	Good	Good	IPA
Average	Good	Good	Average	Good	IPS
Poor	Good	Good	Good	Good	IPA
Good	Good	Good	Good	Good	IPA
Poor	Poor	Good	Good	Good	IPS
Poor	Good	Good	Good	Good	IPA

The difference that occurs in the grade data after being grouped is that before being grouped, the grades of each attribute have varying or heterogeneous grades and turn into smoother or homogeneous grades. In addition, attribute grades are also changed from numerical type to categorical type.

B. Implementation of Naive Bayes on WEKA

Using the Naive Bayes method can be done easily because this algorithm has been embedded in weka. In addition, weka also provides several modifications of Naive Bayes, including Naive Bayes Multinominal, Naive Bayes Multinominal Text, and Naive Bayes Multinominal Updateable. In predicting, the evaluation method chosen is ten folds cross-validation.



Filter: RemovePercentage -P 20.0

Current relation
Relation: data nilai 491
Instances: 490
Attributes: 6
Sum of weights: 490

Attributes

No.	Name
1	MTK
2	IPA
3	IPS
4	B.ING
5	B.INDO
6	RJ

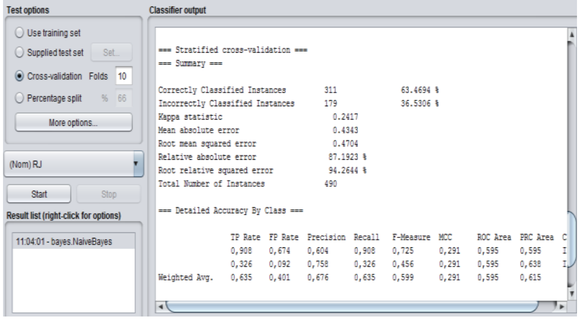
Selected attribute
Name: RJ
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	IPS	260	260.0
2	IPA	230	230.0

Class: RJ (Nom)

Fig 2: Detail of grades data.arff

The cross-validation method shows that 235 data were predicted incorrectly, so the error value is 47.8615%. Based on the confusion matrix reading, the error is because 100 IPA data are predicted as IPS class, and 135 data with IPS class are incorrectly predicted as IPA class.



○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) RJ

11:04:01 - bayes.NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         311               63.4694 %
Incorrectly Classified Instances       179               36.5306 %
Kappa statistic                          0.2417
Mean absolute error                      0.4343
Root mean squared error                  0.4704
Relative absolute error                 87.1923 %
Root relative squared error             94.2644 %
Total Number of Instances              490

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  C
                 0,908    0,674    0,604      0,908    0,725      0,291    0,595     0,595     I
                 0,326    0,092    0,758      0,326    0,456      0,291    0,595     0,638     I
Weighted Avg.    0,635    0,401    0,676      0,635    0,599      0,291    0,595     0,615
```

Fig 3: Output ofgrades data.arff

C. Classification Method Evaluation

The following table displays the confusion matrix for implementing the Naive Bayes Classifier on WEKA. Based on the table, an evaluation of the prediction method used is carried out to determine the level of accuracy of each method and compare which method has the highest level of accuracy.

Table 5. Confusion Matrix Display

		IPA	IPS
Classification Value	IPS	236	24
	IPA	155	75

Based on the table above, there are 490 total cases. From the case, the prediction results of the system stated that 311 students entered the science department based on predictions using the Naive Bayes Classifier method, and 179 students were predicted to be majoring in social studies.

Table 6. Classification Evaluation Display

Evaluasi	Hasil
Accuracy	*0,6319*
Error rate	*0,36817*
False-positive rate	*0.175*
Sensitivity	*0.5865102*
Specificity	*0. 8149*
Precision	*0.934579*

From the calculation results above, it can be seen that the overall Naive Bayes Classifier algorithm has a fairly good performance in terms of accuracy, Error Rate, False Positive Rate, Sensitivity (Recall), Specificity (True Negative Rate) and Precision. So that the Naive Bayes Classifier algorithm is recommended to provide advice on majors to students, where the school can use the prediction results as material and reference in evaluating student learning outcomes at school. The school can use the evaluation in determining and applying appropriate and effective learning techniques in improving students' understanding in the learning process because the level of student understanding is very influential in the grades of each subject. So that every potential possessed by students can be developed properly in supporting student achievement at school. While the benefits for students themselves are that they can see grade patterns that can be used as references to take majors to the next level.

VI. CONCLUSION

The results of applying the Naive Bayes Classifier algorithm in predicting student majors can be stated quite well. It can be seen from the results of the resulting accuracy, which reached 63.46 so that the Naive Bayes Classifier algorithm can be used to predict student majors.

The accuracy of the naive Bayes classifier can be stated quite well. This can be seen based on accuracy, 63.46%, error rate 0.3653%, false positive rate 0.2424%, sensitivity 0.6035%, specificity 0.7575%, and precision 0.944% Naive Bayes classifier method can It is recommended to predict student majors.

REFERENCES

[1] Y. S. Nugroho, T. D. Salma, and S. Rokhanuddin, “Implementasi Data Warehouse Dan Data Mining Untuk Pengembangan Sistem Rekomendasi Pemilihan SMA,” *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 2, no. 2, p. 49, 2016, doi: 10.23917/khif.v2i2.2333.

[2] S. Syarli and A. Muin, “Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi),” *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 22–26, 2016.

[3] I. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, “Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm,” *Int. J. Eng. Emerg. Technol.*, vol. 2, no. 1, p. 53, 2017, doi: 10.24843/ijeet.2017.v02.i01.p11.

[4] H. Naparin, “Klasifikasi Peminatan Siswa SMA Menggunakan Metode Naive Bayes,” *Syst. Inf. Syst. Informatics J.*, vol. 2, no. 1, pp. 25–32, 2016, doi: 10.29080/systemic.v2i1.104.

[5] D. Putra and A. Wibowo, “Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika

5 Menggunakan Algoritma Naïve Bayes," *Pros. Semin. Nas. Ris. ...*, vol. 2, pp. 84–92, 2020, [Online]. Available: http://tunasbangsa.ac.id/seminar/index.php/senaris/article/view/147.

[6] H. Hozairi, A. Anwari, and S. Alim, "Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes," *Netw. Eng. Res. Oper.*, vol. 6, no. 2, p. 133, 2021, doi: 10.21107/nero.v6i2.237.

[7] S. Adinugroho and Y. A. Sari, *Implementasi data mining menggunakan WEKA*. Universitas Brawijaya Press, 2018.

[8] D. T. Larose, *Discovering Knowledge in Data : An Introduction to Data Mining*. Wiley Interscience, 2005.

[9] J. Han and M. Kamber, *Data Mining Concepts and Techniques*. USA: Morgan Kaufmann Publishers, 2006.

[10] F. Alam and S. Pachauri, "Detection using weka," *Adv. Comput. Sci. Technol.*, vol. 10, no. 6, pp. 1731–1743, 2017.

[11] J. G. Moreno-Torres, J. A. Sáez, and F. Herrera, "Study on the impact of partition-induced dataset shift on $ k $-fold cross-validation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 23, no. 8, pp. 1304–1312, 2012.

[12] M. L. Lee, H. Lu, T. W. Ling, and Y. T. Ko, "Cleansing data for mining and warehousing," in *International Conference on Database and Expert Systems Applications*, 1999, pp. 751–760.

BIOGRAPHY

Yohana Tri Utami, Graduated from the Information System Study Program (S1) in 2012, she continued her Masters in Information Systems in 2014 and graduated in 2016. She is currently a Lecturer Computer Science Study Program at Lampung University.

Debby Alita, Graduated from the Informatics Study Program (S1) in 2012, she continued her Masters in Information Systems in 2014 and graduated in 2016. She is currently a Lecturer Informatics Study Program at Universitas Teknokrat Indonesia.

Ade Dwi Putra, Graduated from the Informatics Study Program (S1) in 2013, he continued his Masters in Information Systems in 2014 and graduated in 2018. He is currently a Lecturer Informatics Study Program at Universitas Teknokrat Indonesia.