e-ISSN: 2614-8404
p-ISSN: 2776-3234

COMPATIBILITY OF SELECTION OF STUDENT DEPARTMENTS USINGk-NEAREST NEIGHBOR AND NAÏVE BAYES CLASSIFIER IN INFORMATICS PRIVATE VOCATIONAL SCHOOL, SERANG CITY

Budi Pangestu
Master Program in Computer Science, Faculty of Information Technology, Budi Luhur University Jl. Raya Ciledug,Petukangan Utara, Kebayoran Lama, South Jakarta 12260 Tel. (021) 5853753, Fax. (021) 5869225
E-mail: pangesturj45@gmail.com

Abstract- **Selection of majors by prospective students when registering at a school, especially a Vocational High School, is very vulnerable because prospective students usually choose a major not because of their individual wishes. And because of the increasing emergence of new schools in cities and districts in each province in Indonesia, especially in the province of Banten. Problems experienced by prospective students when choosing the wrong department or not because of their desire, so that it has an unsatisfactory value or value in each semester fluctuates, especially in their Productive Lessons or Competencies. To provide a solution, a departmental suitability system is needed that can provide recommendations for specialization or major suitability based on students' abilities through attributes that can later assist students in the suitability of majors. The process of classifying the suitability of majors in data mining uses the k-Nearest Neighbor and Naive Bayes Classifier methods by entering 16 (sixteen) criteria or attributes which can later provide an assessment of students through this test when determining the majors for themselves, and there is no interference from people. another when choosing a major later. Research that has been carried out successfully using the k-Nearest Neighbors method has a higher recall of 99%, 81% accuracy and 82% precision compared to the Naïve Bayes Classifier whose recall only yields 98% while the accuracy and precision is the same as the k- Nearest Neighbors.**

Keywords: Data Mining, Department Suitability, K-NN, NBC, Classification

I. INTRODUCTION

Selection of appropriate majors will increase interest and provide comfort for someone in learning. On the basis of the same abilities, it is expected that learning activities can run smoothly and do not experience difficulties and can increase students' interest and learning achievement. Conversely, a lack of interest in learning is due to mistakes in choosing a major. The algorithms that will be used in the classification of majors at the SMKS Informatics at Serang City are k-Naerest Neighbors (K-NN) and Naive Bayes Classifier (NBC). Because this study uses a classification which will compare the level of accuracy, precision and recall of the attributes that are owned, namely the attributes that will be used to classify the suitability of student majors, namely, Academic Values for Semester I and II which consist of the values of Religious Education and Character, Pancasila and Citizenship Education, Indonesian, English, Mathematics, Basic Expertise, Expertise Program Basics, Interview Results, Al-Quran Reading and Writing Results, Interests, Counseling Guidance Teacher Recommendations. Where Precision is the level of accuracy between the information requested by the user with the answer given by the system. Meanwhile, Recall is the success rate of the system in recovering information. Accuracy is defined as the level of closeness between the predicted value and the actual value. The following illustration illustrates the difference between accuracy and precision. As for determining the suitability, the results of Semester 1 and Semester 2 scores will be seen. If there is a drop in semester 2 scores, there is a possibility that the student feels that the department he chose is not in accordance with what he wants and is expected at the beginning of registration. The current condition, which is faced at the location of the researcher, every time he enters semester 3 (three) or 4 (four) there are students whose grades fluctuate and cause these students to tend to reflect and feel that the student chose the wrong major when registering first so that the score does not match desire and learning or how to learn does not focus on the material being taught. The school has tried various methods when learning and specifically for these students, it is often activated in class, but because the students feel they are not in accordance with their wishes or have a wrong direction which results in not being enthusiastic about studying subjects from their majors which results in their grades fluctuating every semester or every there are assignments from the subject teacher.

Research that has been conducted by [1] conducted research on the Implementation of the Fuzzy K-Nearest Neighbor (FK-NN) Classification Method for Fingerprint Access Points in Indoor Positioning, which uses Fuzzy K-

Nearest Neighbor (FK-NN) and K -Nearest Neighbor (K-NN) with the results of research carried out from the initial stage to testing, and the results of testing the accuracy of client positions that have been carried out from the K-NN and FK-NN methods, resulting in a percentage index (%) on the K-NN method for k = 1 the value reaches 96%, k = 2 to k = 7 the value reaches 76%, and k = 8 to k = 10 the value reaches 73%. Meanwhile, the FK-NN method for k = 1 and k = 2 the value reaches 96%, k = 3 to k = 8 the value reaches 76%, k = 9 the value reaches 73%, and k = 10 the value reaches 76%. Based on these results, it shows that the system is running well and the accuracy results from the implementation of the FK-NN classification method for Fingerprint Access points in Indoor Positioning have a fairly good level of accuracy than the KNN method.

Subsequent research has also been carried out by [2] who conducted research on the Comparison of the Performance of the Naive Bayes and K-Nearest Neighbor Methods for Indonesian Language Article Classification, which uses Naive Bayes and K-Nearest Neighbor with the results of the Naive Bayes method having better performance good with an accuracy rate of 70%, while the K-Nearest Neighbor method has a fairly low level of accuracy, namely 40%.

Subsequent research has also been carried out by [3], who conducted Determination Analysis of High School Departments based on the Fuzzy Tsukamoto Method and the K-Nearest Neighbor (K-NN) Algorithm, using the Logical Fuzzy method, Tsukamoto Method, Data Maining, K-Nearest. National Algorithm with the results of the research that has been done The Tsukamotodan K-NN method can be used as decision support for majoring high school students based on the students' abilities, interests and talents. The Tsukamoto method performs majors calculating the percentage of department recommendations based on Fuzzy logic. The K-NN method determines the pointing by calculating the distance between the data stored as training data and new data as testing data

Subsequent research has also been carried out by [4], who applied the Naïve Bayes Classifier Method to Determine Final Project Topics on the STIKOM Binaniaga Library Website. Using the Naïve Bayesian Classification (NBC) with the research results that have been described, conclusions can be drawn: 1. The Naive Classifier method can be used to determine and display the topic of the IT department in the proposed title. 2. The accuracy and finding in determining the topic in the data of the new IT department thesis title is influenced by the learning data or training data in each category. This training data contains words that often appear in each category or words that can represent certain categories. Further research has also been carried out by Mafakhir [5] who conducted research on the Application of the Naïve Bayes Classifier Method for Student Designation at Madrasah Aliyah Al-Falah Jakarta, with the Naïve Bayes Classifier method, with the results of testing and testing of the methods and prototypes that have been developed, it is concluded that the Naïve Bayes method can be used to classify majors. students. However, the process of converting numeric values into categorical values results in low accuracy, precision, and recall results. Simplification causes detailed information to be lost.

Further research has also been carried out by [6],[7] who conducted research on the comparative analysis of the naïve Bayes Classifier and K-Nearest Neighbor methods against data classification, with the Naive Bayes method, the k-nn algorithm, confusion matrix, with the results of the comparison of the two methods. It can be concluded that the k-NN method has better accuracy than the NBC method. This is evidenced by an accuracy rate of 80% for the k-NN method and 73% for the nbc calculated using the Confusion matrix method.

This study will compare 2 methods using the k-Nearest Neighbor[8] and Naïve Bayes Classifier methods[9],[10],[11],[12] to see the suitability of the Department to students at the Informatics SMKS Serang City which uses the attributes that are owned, namely Semester I and II Academic Values which consist of the value of Religious Education and Character, Pancasila and Citizenship Education, Indonesian, English, Mathematics, Basic Expertise, Expertise Program Basics, Interview Results, Al-Quran Reading and Writing Results, Interests, Counseling Guidance Teacher Recommendations.

II. RESEARCH METHODOLOGY

The object of research is sourced from Basic Education Data (DAPODIK) and Data from Semester 1 and 2 e-report cards for SMKS Informatics at Serang City from 2014-2019, so it can be seen what kind of students study the subject in their majors. Where in the development of this research using software with the PHP programming language (Hypertext Processor), so it takes some software that supports it, namely as follows:

a. Web Browser
b. Web Server (XAMPP)
c. Editor (Sublime / Notepad ++ / Visual

Studio)

d. Output (Microsoft Office Word and Excel)

A. Technique of Analysis

The analysis technique that will be used is data mining which compares the k-Nearest Neighbor method with the Naïve Bayes Classifier using simple nonprobablity sampling. Which is used to process the data that will be made for Testing data and Test data on the suitability of the Department can be seen in Figure .

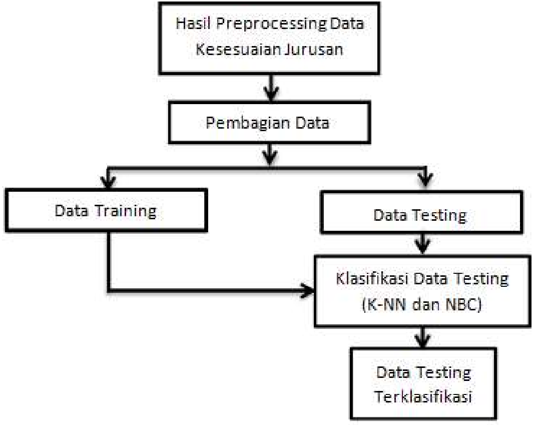

Hasil Preprocessing Data Kesesuaian Jurusan
Pembagian Data
Data Training
Data Testing
Klasifikasi Data Testing (K-NN dan NBC)
Data Testing Terklasifikasi

Figure 1. Data Analysis Techniques

B. Design

The system that will be created is a system that will compare the k-Nearest Neighbor[13] and Naïve Bayes Classifier methods[14],[15] to determine the suitability of the department in the selection of the department later.

The process to be designed in making a prototype or system is:

1) Import excel data

Import data is done to see the data that has been processed which will later process the data in the system with the data format in the form of .cvs or .xls

2) Preprocessing

Data that has been imported will be checked for eligibility in the data by the system, and later processed by the system.

3) Comparison results

After the data is in accordance with the needs, it will be processed directly by the system with a comparison stage in the research analysis process. The results of the comparison will be in the form of a description of Appropriate or Unsuitable, which will see the results of data processing in the selection of majors by students.

The steps in calculating the NBC value in the system can be described in the NBC Flowchart which can be seen in Figure 2.

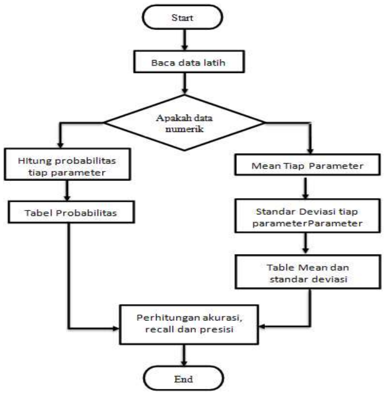

Start
Baca data latih
Apakah data numerik
HItung probabilitas tiap parameter
Tabel Probabilitas
Mean Tiap Parameter
Standar Deviasi tiap parameterParameter
Table Mean dan standar deviasi
Perhitungan akurasi, recall dan presisi
End

Figure 2. NBC Flowchart

III. RESULT AND DISCUSSION

The data analysis activity in this study aims to provide a detailed picture of what is examined in this study. In analyzing the data in this study the researcher used the method of observation on the attributes used in the calculation based on 997 data taken from 2014-2019 shown in table There are 4 (four) majors that will be of interest, namely Accounting, Office Administration, Software Engineering and Computer and Network Engineering.

Table 1 Total Data Amount

Id	Minat	Smstr Ganjil	Smstr Genap	Hsl Intrview	Hsl BTQ	Rek BK	Hasil
1	AK	83	82	80	90	AK	SESUAI
2	AK	82	86	90	80	AK	SESUAI
3	AK	78	74	80	90	AK	SESUAI
4	AK	83	85	80	80	AK	SESUAI
5	AK	82	86	80	70	AK	SESUAI
6	AK	79	84	90	80	AK	SESUAI
7	AK	77	82	80	80	APK	TIDAK SESUAI
8	AK	81	80	70	90	AK	SESUAI
9	AK	83	89	50	70	AK	SESUAI
….	…..	…..	…..	……	…..	…..	….
990	APK	84	83	80	70	APK	SESUAI
991	APK	80	78	70	80	APK	SESUAI
992	APK	84	80	50	80	APK	TIDAK SESUAI

Table 2 Test Data

Id	Minat	Smstr Ganjil	Smstr Genap	Hsl Intrview	Hsl BTQ	Rek BK	Hasil
1	AK	83	82	80	90	AK	SESUAI
2	AK	82	86	90	80	AK	SESUAI
3	AK	78	74	80	90	AK	SESUAI
4	AK	83	85	80	80	AK	SESUAI
5	AK	82	86	80	70	AK	SESUAI
6	AK	79	84	90	80	AK	SESUAI
7	AK	77	82	80	80	APK	TIDAK SESUAI
8	AK	81	80	70	90	AK	SESUAI
9	AK	83	89	50	70	AK	SESUAI
….	…..	…..	…..	……	…..	…..	….
990	APK	84	83	80	70	APK	SESUAI
991	APK	80	78	70	80	APK	SESUAI
992	APK	84	80	50	80	APK	TIDAK SESUAI

A. Research Results

In the results of this study, it is explained where to start taking the training data and test data that is owned and then processing the data by entering the data in the calculation of the k-NN and NBC methods which will later calculate in each method and confusion matrix will be carried out. The results of the k-NN and NBC calculations will produce the suitability of the majors that students choose. For the results of the research and steps in the system to determine the suitability of the Koran itself, it will be discussed in prototype testing to show that the results of the application made are as expected. The flowchart for comparison of the k-Nearest Neighbor and Naïve Bayes Classifier methods is shown in Figure 3.

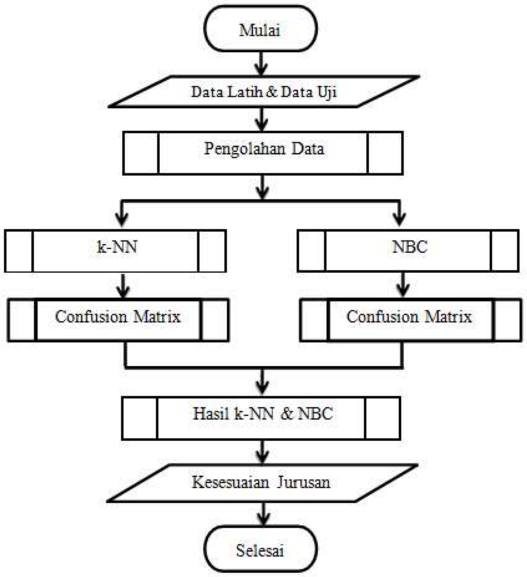


Figure 3. Classification Process Flow

B. System Design
System design determines how the system will meet the objectives of this study in the form of hardware, software, interface displays, databases and files that will be needed in designing this system.

In this design, it consists of a Class Diagram which will display several descriptions of the system, namely the attributes and operations in a class. The class diagram itself can be seen in Figure 4B

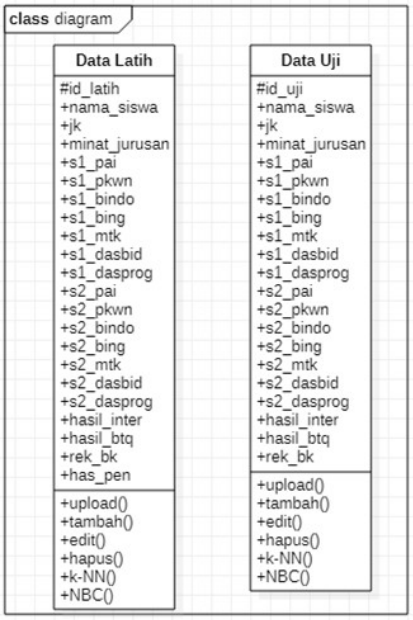
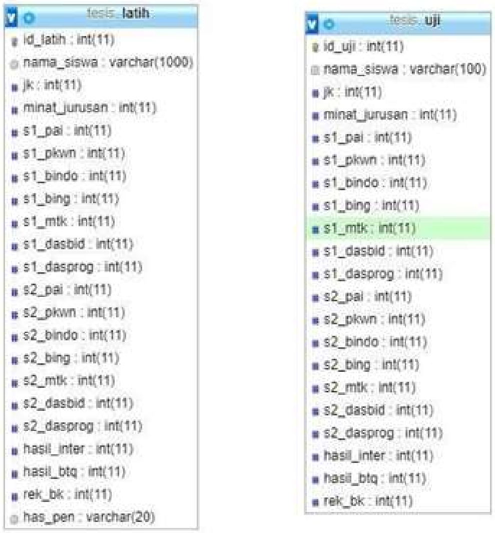


Figure 4. Class Diagram

Consists of 2 tables in the database for the suitability of this direction, and can be seen in Figure 5.

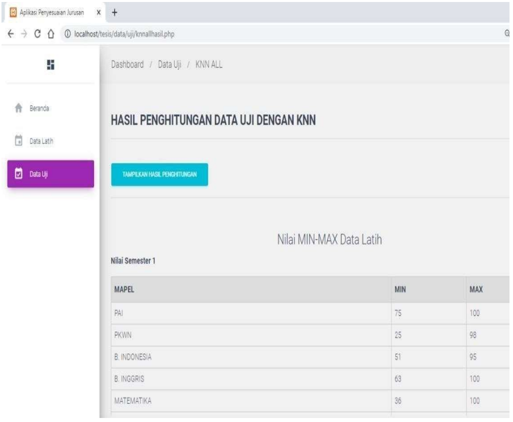


tesis latih
id_latih : int(11)
nama_siswa : varchar(1000)
jk : int(11)
minat_jurusan : int(11)
s1_pai : int(11)
s1_pkwn : int(11)
s1_bindo : int(11)
s1_bing : int(11)
s1_mtk : int(11)
s1_dasbid : int(11)
s1_dasprog : int(11)
s2_pai : int(11)
s2_pkwn : int(11)
s2_bindo : int(11)
s2_bing : int(11)
s2_mtk : int(11)
s2_dasbid : int(11)
s2_dasprog : int(11)
hasil_inter : int(11)
hasil_btq : int(11)
rek_bk : int(11)
has_pen : varchar(20)

tesis uji
id_uji : int(11)
nama_siswa : varchar(100)
jk : int(11)
minat_jurusan : int(11)
s1_pai : int(11)
s1_pkwn : int(11)
s1_bindo : int(11)
s1_bing : int(11)
s1_mtk : int(11)
s1_dasbid : int(11)
s1_dasprog : int(11)
s2_pai : int(11)
s2_pkwn : int(11)
s2_bindo : int(11)
s2_bing : int(11)
s2_mtk : int(11)
s2_dasbid : int(11)
s2_dasprog : int(11)
hasil_inter : int(11)
hasil_btq : int(11)
rek_bk : int(11)

Figure 5. Database Design

C. Display Design
The Min-Max value for odd semesters is taken from data obtained in research conducted by researchers which can be seen in Figure 6.


Aplikasi Penyesuaian Jurusan
localhost/tesis/data/uji/knnallhasil.php
Beranda
Data Latih
Data Uji
Dashboard / Data Uji / KNN ALL
TAMPILKAN HASIL PENGHITUNGAN

HASIL PENGHITUNGAN DATA UJI DENGAN KNN

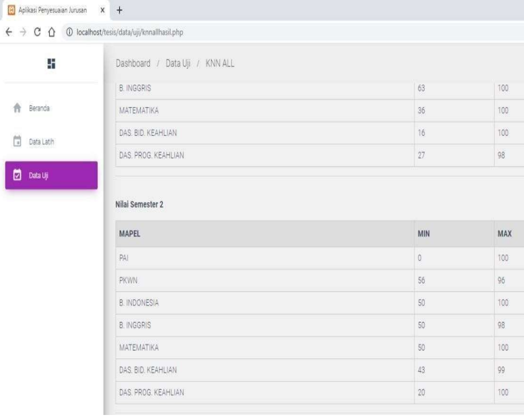
Nilai MIN-MAX Data Latih

Nilai Semester 1

MAPEL	MIN	MAX
PAI	75	100
PKWN	25	98
B. INDONESIA	51	95
B. INGGRIS	63	100
MATEMATIKA	36	100

Figure 6. Display of Odd Semester Min Max Value

The Min-Max value for the even semester is not much different from the odd semester value of data taken from the research place which can be seen in Figure 7


Aplikasi Penyesuaian Jurusan
localhost/tesis/data/uji/knnallhasil.php
Beranda
Data Latih
Data Uji
Dashboard / Data Uji / KNN ALL

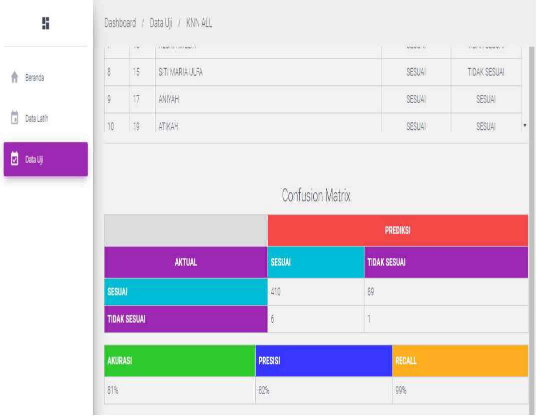
B. INGGRIS	63	100
MATEMATIKA	36	100
DAS. BID. KEAHLIAN	16	100
DAS. PROG. KEAHLIAN	27	98

Nilai Semester 2

MAPEL	MIN	MAX
PAI	0	100
PKWN	56	96
B. INDONESIA	50	100
B. INGGRIS	50	98
MATEMATIKA	50	100
DAS. BID. KEAHLIAN	43	99
DAS. PROG. KEAHLIAN	20	100

Figure 7. Display of Min Max Even Semester Value

The calculation of the Confusion Matrix on the test data can be seen in Figure 8.



8	15	SITI MARIA ULFA	SESUAI	TIDAK SESUAI
9	17	ANIYAH	SESUAI	SESUAI
10	19	ATIKAH	SESUAI	SESUAI

Confusion Matrix

	PREDIKSI	
AKTUAL	SESUAI	TIDAK SESUAI
SESUAI	410	89
TIDAK SESUAI	6	1

AKURASI	PRESISI	RECALL
81%	82%	99%

Figure 8. Display Confusion Matrix Value

D. White Box

The tester in white box testing is knowledgeable about coding and writing test cases with the appropriate parameters. This mainly concerns the control flow and data flow of a program. White Box itself has several techniques in testing, such as: Data Flow Testing, Control Flow Testing, Basic Path / Path Testing, and Loop Testing can be seen in Figure 9

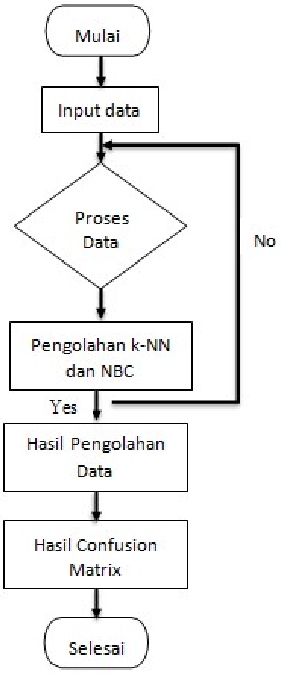


Figure 9. White Box Testing

E. Implementation Plan

1. Implementation of Development
   In the system development stage itself, the system will be refined to be more tailored to the needs, with an estimated time of around 2 months to adjust and be able to complete future developments.
2. System Socialization
   After the development of the system that has been socialized at the beginning of the meal, it will be socialized again for the system with the latest developments to the SMKS Informatics of Serang City.
3. Application of the system
   For the implementation of the system itself, it will be carried out with existing standards in the location, by preparing the necessary equipment, starting from hardware and software, which will be applied to the system, with a series ranging from 3 weeks - 1 month.
4. System Trial
   After it is finished, it will be tested on the whole system and look for whether there are still errors or deficiencies in the system.
5. Evaluation of System Trials
   If there are still errors or shortcomings that

result, the system will be refined and repaired again according to the needs of the Serang City Informatics SMKS.

6. Refinement of Systems and Procedures
   If the evaluation of the trial has been carried out and no more errors appear in the system, improvements will be made and included in standard operating procedures at the Serang City Informatics SMKS.

IV. CONCLUSION

This conclusion is drawn from the results of research that has been carried out starting from, problems, hypotheses and discussions that have been tested, the following conclusions are obtained:

a. Based on the results of the accuracy, recall and precision test in both methods, it was found that k-Nearest Neighbor has a higher recall of 99%, 81% accuracy and 82% precision compared to the Naïve Bayes Classifier whose recall only produces 98% while for accuracy and precision the same as k. -Nearest Neighbors.

b. So by using the k-Nearest Neighbor method it can be used for calculations in the Adjustment of Majors to Students at Informatics SMKS Serang City

After the results are obtained from this study, it is advisable to carry out further research with the following additions:

a. The data obtained from the research site must be better and try to include all data, not just certain items,

b. Research is carried out annually, in order to see and take into account the suitability of the majors in students later.

c. The stages of implementation for students and the committee for admitting new students are very much needed in the first stage of selecting a department directed by the committee to prospective new students.

REFERENCES

[1] Billyan, B. F., Bhawiyuga, A., & Primananda, R. (2017). Implementasi Metode Klasifikasi Fuzzy K-Nearest Neighbor (FK-NN) Untuk Fingerprint Access Point Pada Indoor Positioning. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (JPTIK)*, *1*(11).

[2] Devita, R. N., Herwanto, H. W., & Wibawa, A. P. (2018). Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(4), 427. https://doi.org/10.25126/jtiik.201854773

[3] Ariani, F., & Endra, R. Y. (2013). Implementation of Fuzzy Inference System with Tsukamoto Method for Study Progamme Selection. *2nd International Conference on Engineering and Technology Development (ICETD)*, *Icetd*, 189–200.

[4] Ghaniy, R., & Sihotang, K. (2019). Penerapan Metode Naïve Bayes Classifier Untuk Penentuan Topik Tugas Akhir. *Teknois : Jurnal Ilmiah Teknologi Informasi Dan Sains*, *9*(1), 63–72. https://doi.org/10.36350/jbs.v9i1.7

[5]. Mafakhir, A. Z., & Solichin, A. (2020). Penerapan Metode Naïve Bayes Classifier Untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta. *Fountain of Informatics Journal*, *5*(1), 21. https://doi.org/10.21111/fij.v5i1.4007

[6]. Handayani, I., & Ikrimach, I. (2020). Accuracy Analysis of K-Nearest Neighbor and Naïve Bayes Algorithm in the Diagnosis of Breast Cancer. *Jurnal Infotel*, *12*(4), 151–159. https://doi.org/10.20895/infotel.v12i4.547

[7]. Antaristi, M., & Kurniawan, Y. I. (2017). Aplikasi Klasifikasi Penentuan Pengajuan Kartu Kredit Menggunakan Metode Naive Bayes di Bank BNI Syariah Surabaya. *Jurnal Teknik Elektro*, *9*(2). https://doi.org/10.15294/jte.v9i2.12496

[8]. Indrayuni, E. (2017). Text Mining dalam Analisis Sentimen Review Restoran Menggunakan Algoritma K-Nearest-Neighbor (KNN). *JURNAL TEKNIK INFORMATIKA STMIK ANTAR BANGSA*, *3*(2).

[9] Sipayung, E. M., Maharani, H., & Zefanya, I. (2016). Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier. *Jurnal Sistem Informasi*.

[10] Zuhri, F. N., & Alamsyah, A. (2017). Analisis sentimen masyarakat terhadap brand smartfren menggunakan naive bayes classifier di forum kaskus. *E-Proceeding of Management*.

[11] Indriani, A. (2014). Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) Yogyakarta*.

[12] Oktasari, L., Chrisnanto, Y. H., & Yuniarti, R. (2016). Text Mining Dalam Analisis Sentimen Asuransi Menggunakan Metode Niave Bayes Classifier. *Prosiding SNST*.

[13] Rivki, M., & Bachtiar, A. M. (2017). IMPLEMENTASI ALGORITMA K-

NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN FOLLOWER TWITTER YANG MENGGUNAKAN BAHASA INDONESIA. *Jurnal Sistem Informasi*. https://doi.org/10.21609/jsi.v13i1.500

[14] Rinawati, R. (2017). Penentuan Penilaian Kredit Menggunakan Metode Naive Bayes Berbasis Particle Swarm Optimization. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*. https://doi.org/10.30645/j-sakti.v1i1.28

[15] Saleh, A. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Creative Information Technology Journal*.