Eduvest - Journal of Universal Studies
Volume 4 Number 06, June, 2024
p- ISSN 2775-3735- e-ISSN 2775-3727

APPROPRIATENESS OF STUDENT MAJOR SELECTION USING NAIVE BAYES AND K-NEAREST NEIGHBOR ALGORITHMS AT SMK PLUS AL MUSYARROFAH

Kamaluddin Mustofa[1], Tyan Tasa[2], Denni Kurniawan[3]
[1,2,3] Universitas Budi Luhur, Indonesia
Email: 2111601627@student.budiluhur.ac.id, 2111601767@student.budiluhur.ac.id, denni.kurniawan@budiluhur.ac.id

ABSTRACT

The process of selecting a major is a critical stage for students because it can influence their motivation and learning outcomes while attending school, especially at Vocational High Schools (SMK). This challenge is becoming more significant with the emergence of many new schools in various cities and districts in Indonesia, especially in DKI Jakarta Province. Prospective students often choose majors not based on personal interests, which can then result in lower grades, especially in productive subjects or certain competencies. To overcome this problem, a major suitability system is needed that can provide recommendations based on student abilities through certain attributes. In this research, a department suitability classification process was carried out using the Naive Bayes and k-Nearest Neighbor methods using data from 238 tenth grade (X) students for the 2023/2024 academic year, which included 9 relevant attributes. The testing process was carried out with a composition of training data and test data in five comparisons, namely 90:10, 80:20, 70:30, 60:40, and 50:50. The research results show that the 80:20 composition provides the best results, with k-Nearest Neighbor achieving recall, accuracy and precision levels of 100%. On the other hand, the Naive Bayes Classifier produces a recall rate of 61%, with an accuracy of 73%. These results indicate that k-Nearest Neighbor is superior in predicting major suitability compared to Naive Bayes under these conditions.

KEYWORDS *Department Selection, Naive Bayes, K-Nearest Neighbor*

INTRODUCTION

Vocational High Schools (SMK) in Indonesia, parallel to Senior High Schools (SMA), have different roles. Whereas SMA is more geared towards continuing education to the university level, SMK is designed to prepare students for the world of work after graduation or provide the option of continuing education

How to cite:	Kamaluddin Mustofa, Tyan Tasa, Denni Kurniawan (2024). Appropriateness of Student Major Selection Using Naive Bayes and K-Nearest Neighbor Algorithms at SMK Plus Al Musyarrofah. *Journal Eduvest. 4* (6): 5436-5456
E-ISSN:	2775-3727
Published by:	https://greenpublisher.id/

to a higher level. As such, the focus of learning in SMK is more related to the development of practical skills as opposed to the understanding of general knowledge as in SMA. The process of choosing a major in SMK starts from the time of enrollment, in contrast to SMA which often starts in semester 3 or during grade XI. This condition is influenced by the increasing number of public and private schools in each city and district, especially in DKI Province, which creates fierce competition in SMK admissions. Information from various educational sources and researchers' experiences show that a mismatch in the selection of majors in SMK can have an impact on the fluctuation of students' grades in each semester. This emphasizes the importance of supporting students in the selection of majors that match their interests and abilities to minimize mismatches and improve their academic performance.(Kusumadewi et al., 2020)

Selection of appropriate majors will increase interest and provide a person's comfort in learning. With the same basic ability, it is expected that learning activities can run smoothly and have no difficulties and can increase students' interest and learning achievement. Conversely, the lack of interest in learning due to errors in choosing a major (Istighfar et al., 2023).

Based on this information, the student's own ability can make decisions that are contrary to the student's ability. Various information about SMK majors has been widely available in print media and on the internet, making it easy to get this information. However, the information available only provides a general explanation, such as: profile, cost, location, and other general information. And this information has not fully helped provide input regarding the majors desired by prospective students according to the abilities, interests and preferences of prospective students.(Mohamad Andri Rasyid et al., 2023)

Meanwhile, data mining, which is often used in handling large amounts of data like this, is a process that uses statistics, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases. (Widiastuti et al., 2023) Data mining on the other hand is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets. The output of this data mining can be used to improve future decision making. (Brilliant et al., n.d.) From these problems, a data mining technique is needed that can assist in the classification of the suitability of student majors at SMK Plus Al Musyarrofah in accordance with student interests so that it can increase the enthusiasm for learning and improve student achievement in each semester. The criteria that will be used as a reference in this study are the scores of Indonesian Language, English, Mathematics, Science, Entrance Selection Test and the overall average score. Data mining is the mining or discovery of new information by looking for certain patterns or rules from a very large amount of data. (Maulana, 2023) Data mining consists of various techniques that can be used to predict and classify, where this technique estimates the possibility that will occur in the future by looking at some existing information and data patterns. (Sayhidin et al., 2023)

The algorithms that will be used in the classification of majors at SMK Plus Al Musyarrofah are k-Naerest Neighbors (K-NN) and Naive Bayes. Because this research uses classification which will compare the level of accuracy, precision and recall of the attributes owned, namely the attributes that will be used for the classification of student major suitability, namely, NISN, Name, Indonesian

Language Score, English Language, Mathematics, Science, Entrance selection test and overall average value and major. Precision is the level of accuracy between the information requested by the user and the answer given by the system. While Recall is the success rate of the system in retrieving information.(Fakhri et al., n.d.) 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. (Lestari et al., 2019) As for determining the suitability, the results of the Indonesian Language, English, Mathematics, Science, Entrance Selection Test and the overall average value will be seen, it is possible that the student already feels that the major he chose is not in accordance with what he wanted and expected at the beginning of registration. (Yudhi Putra & Ismiyana Putri, n.d.)

The current conditions faced at the research location every time they enter semester 3 (three) or 4 (four) there are students whose grades go up and down and cause these students to tend to brood and feel that these students have chosen the wrong major when registering first so that their grades are not in accordance with their wishes and learning or learning methods do not focus on the material being taught (Nuraeni et al., 2023), The school has tried various methods during learning and specifically these students are often activated in class, but because the student already feels that it is not in accordance with his wishes or the wrong major which has an impact on not being enthusiastic about studying the subjects of his major which results in his grades going up and down in every semester or every time there is an assignment from his subject teacher.(Sayhidin et al., 2023).

A number of studies have been conducted to overcome various problems using methods such as Naïve Bayes and K-Nearest Neighbor (KNN). For example, Manurung and Putri (2020) used Naïve Bayes to determine high school students' talent interest in choosing a major, with the result of increasing accuracy by 1.69% compared to previous research. On the other hand, Arifin and Ariesta (2019) applied Naïve Bayes based on Particle Swarm Optimization for chronic kidney disease prediction, with confusion matrix accuracy reaching 98.75% and AUC 99%. Then, Ramadhani et al. (2022) used Naïve Bayes and Gaussian functions to determine the majoring of grade X students, with the development of special potential such as visual-spatial intelligence which has a high certainty of 99.60%. These studies show the various applications and performance improvements of these methods in the context of data-driven decision-making.

This research identifies the problem that prospective students often choose majors without considering their personal interests, which can result in a decrease in grades, especially in productive lessons or certain competencies. Problem restrictions include the object of research focusing on SMK Plus Al Musyarrofah, using data from 238 tenth grade students in the 2023/2024 academic year, and applying the Naïve Bayes and K-Nearest Neighbor (K-NN) methods. The problem formulation includes questions regarding the more optimal method between Naïve Bayes and K-NN in determining the suitability of majors as well as a comparison of accuracy, precision, and recall between the two. The purpose of this research is to produce a performance comparison between Naïve Bayes and K-NN algorithms in the context of student major suitability, and apply the results to provide solutions for SMK Plus Al Musyarrofah. The benefits include assisting the school in overcoming the problem of selecting unsuitable majors and providing a contribution to the development of science in similar research in the future.

RESEARCH METHODS

This research has prepared a research schedule that aims to ensure the achievement of targets in completing this research. This scheduling plan will be implemented after the thesis proposal has been approved or has passed the proposal hearing. Based on the detailed research schedule, in September 2023, from the first to the third week, data collection was conducted, followed by the preprocessing stage from the fourth to the second week of October. Data transformation was carried out from the third to the fifth week of October 2023. Implementation of the K-Nearest Neighbors (KNN) and Naïve Bayes algorithms focused on the first to second weeks of November 2023, followed by analysis in the third and fourth weeks. In December 2023, the analysis using KNN and Naïve Bayes was carried out in the first to third weeks, with the preparation of the conclusion and final report in the fourth week, including a comparison of the precision and accuracy of the two algorithms.

RESULTS AND DISCUSSION

Business Understanding

Currently, the challenge faced is the lack of a suitable method to predict the suitability of major selection in vocational high schools each year. Therefore, this research will focus on exploration using a dataset covering 238 students in the 2023/2024 academic year at Al Musyarrofah Plus Vocational High School (SMK). The aim is for the school to evaluate the effectiveness of learning, teaching methods, and academic support from the students themselves, especially in the context of the suitability of major selection, in the hope of improving the quality of education. In addition, this research aims to help students achieve optimal grades and learning outcomes. Furthermore, this research is geared towards improving the efficiency of institutional resource management by planning facilities, instructors, and student support needs based on the projected suitability of majors during the school years.

Choosing a major while in school is an important decision that affects a student's educational and career journey. This decision is influenced by a variety of factors, including personal interests, skills, career aspirations, as well as influences from the social and educational environment. This study aims to explore the dynamics and factors that motivate students in choosing a major at the secondary school level. The research method involved analyzing data from 238 students in the 2023/2024 academic year at Al Musyarrofah Plus Vocational High School (SMK). This data was used to evaluate the suitability of major selection, the effectiveness of learning methods, and students' academic support. This research also aims to provide insights that can help students achieve optimal learning outcomes and support school resource planning. The results of this study are expected to provide valuable information for schools, students, and parents in understanding and improving the major selection process at the secondary school level.

By addressing these barriers, schools can strengthen their readiness to predict students' major selection suitability and implement proactive measures for more effective education planning and management.

Data Understanding

In this phase, data cleaning is done by ensuring that there are no missing values and that the data has consistent values (not outside the Min-Max range). To get quality data, there are several preprocessing techniques used. Data Validation, Used to identify odd data (outer/noise), inconsistent data, and incomplete data (missing values) or lack of appropriate attributes. From 238 sample data and 13

	NO	NISN	NAMA	B_INDO	B_INGG	MATEMATIKA	IPA	TES_SELEKSI	RATA_RATA	M
0	1	******9742	DPL	80	76	72	84	80	78.4	TI
1	2	******1586	MNA	82	68	70	80	82	76.4	B
2	3	******8103	SPP	78	82	76	78	88	80.4	TI
3	4	******0975	MRS	84	80	72	78	78	78.4	TI
4	5	******7299	MAA	84	82	76	72	74	77.6	B
5	6	******9800	ALS	86	78	74	68	70	75.2	B
6	7	******0121	RFR	82	80	74	78	80	78.8	TI
7	8	******0412	ARP	82	74	76	78	80	78	TI
8	9	******5150	ARM	74	78	84	76	78	78	TI
9	10	******9166	REP	86	74	72	76	78	77.2	B

index: Index([1, 0], dtype='int32', name='JURUSAN')

Jumlah: [136 102]

Total Seluruh Data: 238

attributes, a cleaning process will be carried out which aims to eliminate attributes whose conditions are not suitable so that it becomes 238 sample data and 9 attributes only.

Figure 4 1 Student data for academic year 2023/2024 into 9 attributes

Understanding data related to the suitability of graduates' majors in SMK (Vocational High School) involves various aspects to help schools, policy makers, and industries understand the preferences, skills, and trends of graduates. Some elements that can be covered in this data understanding involve:

1. Suitability and Vocational Program: Data regarding the vocational programs that best match students' interests. This includes information on student participation rates in specific programs and trends in student suitability from year to year.
2. Academic Success: Data on the academic success of students in specific vocational programs. This includes graduation rates, grade point averages, and other academic achievements.
3. Major Suitability Survey: The result of a survey or research conducted to identify the suitability of a student's major. This survey may include career preferences, hobby interests, or other aspects that may influence vocational choice.

Understanding this data is important for shaping education policy, optimizing vocational programs, and ensuring students' readiness to succeed in education and

get the best grades. Careful analysis of the data can help schools adjust their vocational programs to suit the needs of students and the suitability of majors while attending school.

Data preparation

The steps in data preparation are critical stages in research that involve processing and cleaning raw data so that it can be used effectively in analysis. Here are some common steps and techniques that are often used in data preparation:

1. Data Understanding
   Data Description Start by understanding the structure of the dataset, looking at the number of rows and columns, and exploring basic descriptive statistics. Identify the data type of each column numeric, categorical, text and others.
2. Data Cleaning
   Handling Missing Values Identify and handle missing values. This could involve replacing missing values, deleting them, or using imputation techniques. Handling Outliers Detect and handle outliers or extreme values that may affect the analysis results and Handling Duplicates Identify and remove duplicate data if any.
3. Data Transformation:
   a. Scaling and Normalization: Scale or normalize numerical values to have similar ranges.
   b. Encoding Categorical Data: Convert categorical variables into a format that can be used by algorithms, such as one-hot encoding.
   c. Feature Engineering: Create new features that might improve model performance or provide additional insights.
   d. Text Data Processing: If the data includes text, perform tokenization, stemming, and vectorization in preparation for text analysis or modeling.
4. Dimensional Reduction:
   a. Principal Component Analysis (PCA): Dimensionality reduction for datasets with many features
   b. Feature Selection: Select the most informative and relevant features for the purpose of analysis.
   c. Train-Validation-Test Split: Split the dataset into train, validation, and test subsets for model evaluation.

Modeling

At the modeling stage, this is done with 2 (two) steps, namely the Lexicon Based process where the labeling results will be used as training data and the classification process with the Naïve Bayes and K-Nearest Neighbors (KNN) methods which will be described in the following steps:

1. Data Preparation
   In this step, data preparation for classification analysis is carried out by loading the dataset into Python using libraries such as Pandas, NumPy and others. In this step, data cleaning, converting data into a suitable format, and separating data into features and labels (targets) are also done.
2. Algorithm Selection

In this step, the classification algorithm to be used is selected, which includes Naïve Bayes and K-Nearest Neighbors (KNN) by using a library such as Scikit-learn to implement the algorithm.

3. Data Sharing
   In this step, the dataset is divided into training data and testing data. The training data is used to train the classification model, while the test data is used to test the performance of the model. The function used in this step is the train_test_split function from Scikit-learn to split the dataset. The ratio of training data and test data to be used is 90:10, 80:20, 70:30, 60:40 and 50:50.
4. Model Evaluation
   In this step, the model performance is evaluated using test data. Model testing is done by comparing the accuracy, precision, recall, and f1-score values. The functions used to calculate these metrics in scikit-learn are accuracy_score, precision_score, recall_score, and f1_score.
5. Prediction
   In this step, once the model is trained and configured properly, it is used to perform prediction on new data. The predict method is used to predict labels based on features of the data that are not yet known.
6. Cross-validation
   To ensure the reliability of the model in this step, cross-validation is performed by dividing the dataset into multiple folds and training and testing the model on each fold. This is done to help avoid overfitting and provide a more realistic estimate of the model's performance on new data.
7. Interpretation of Results
   After making predictions the next step is to interpret the results and take appropriate action based on the needs.

In this modeling stage, the two classification algorithms, Naïve Bayes and K-Nearest Neighbors (KNN), are initiated. The dataset used was originally 238 and 13 attributes to 238 and 9 attributes after going through the text preprocessing stage. Furthermore, the dataset is separated between training data and test data with a trial comparison of 90:10, 80:20, 70:30, 60:40 and 50:50 for each algorithm. The classification model for each algorithm is as shown below:

Evaluation

It is important to remember that these evaluations should be holistic and consider the individual needs of students. In addition, effective communication with all stakeholders, such as students, teachers, and parents, can enhance the success of the evaluation and implementation of interest and talent development programs.

The result of the test is to produce an accuracy value. The confusion matrix model will form a matrix consisting of true positive or positive tuples and true negative or negative tuples, then enter the prepared testing data into the confusion matrix. From this naïve bayes method, the results obtained still have to be optimized in order to get a maximum and high value. Deployment

At this deployment stage, the development of application prototypes is carried out using the python programming language and streamlit as an interface display, in order to make it easier to read the results of data that has been processed by

python programming. To run the application prototype that has been developed, is to run the command "streamlit run kamal.py" on the terminal and the main page will appear on the browser.

Figure 4.5 Prototype main menu display

On the main page of the prototype there are several menus including File Upload, Naïve Bayes algorithm classification, K-Nearest Neighbors (KNN) algorithm classification. Furthermore, the display of the Upload File menu is as follows.

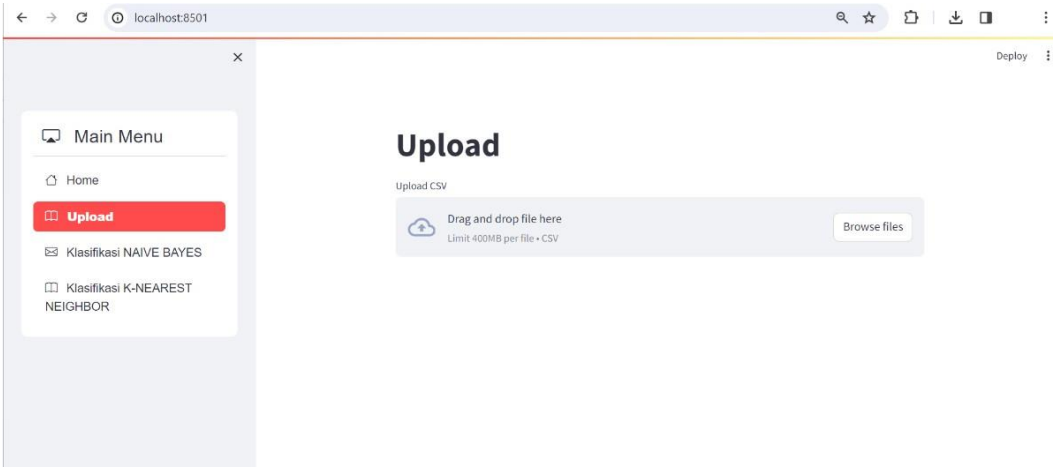

Figure 4.6 Dataset Upload

Figure 4.6 above is a form for uploading a dataset file in the form of a csv file and then a description of the dataset file will be displayed along with the data content into the dataframe. After the data is uploaded, the next process is to select the Naïve Bayes Classification menu which is displayed as follows:

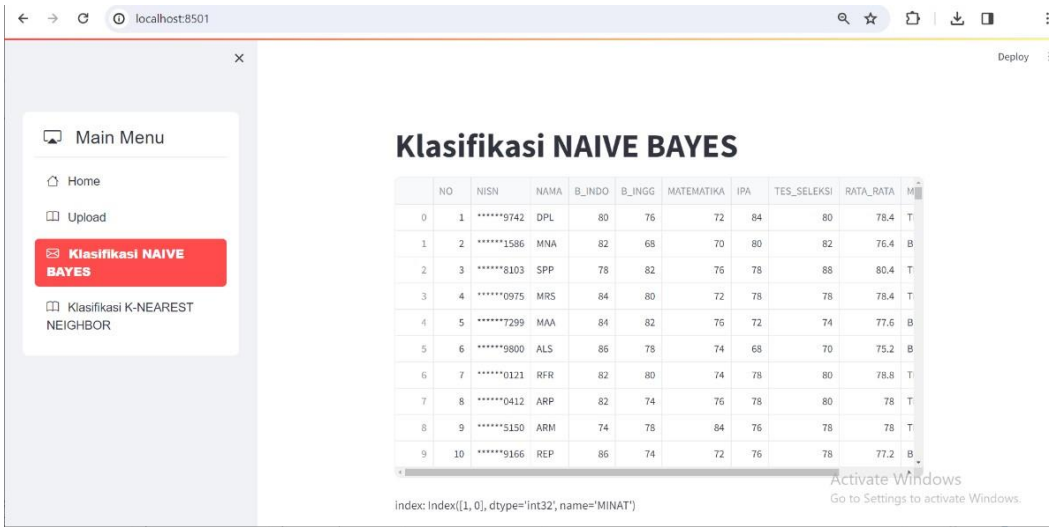

Figure 4.7 Display of Naive Bayes classification

By applying the Naive Bayes algorithm to the data of Xth grade vocational students, we analyzed academic grades and proficiency test results as key features. The Naive Bayes process helps in identifying class probability patterns based on these attributes, allowing us to predict major labels. Model evaluation involves measuring prediction accuracy and adjusting parameters to improve model precision.

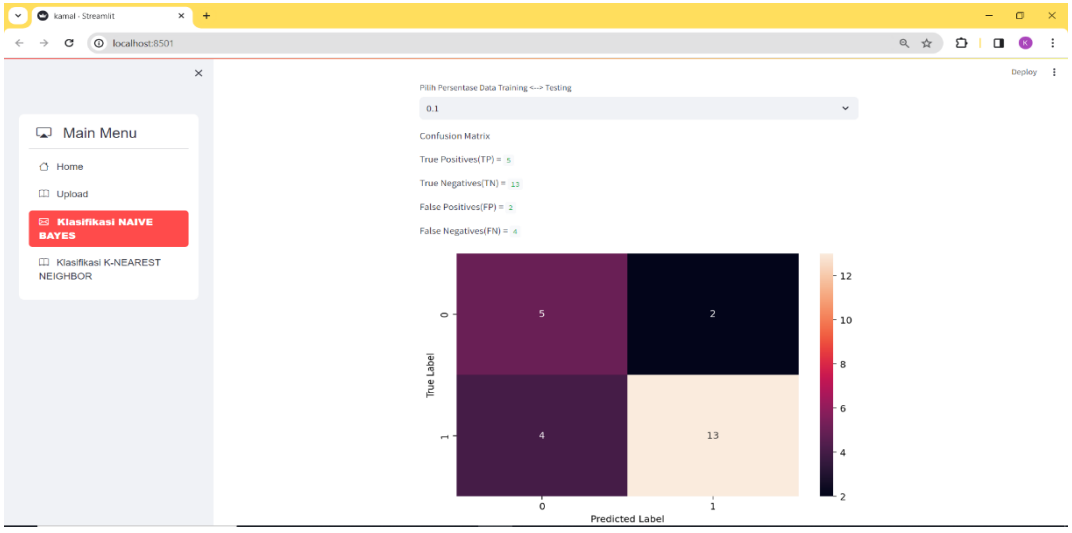

Figure 4 8 Naïve Bayes Confusion Matrix

The results of the initial experiment with a total of 338 data samples and a training and test data composition of 90:10, consisting of 304 training data and 34 test data showed that the Naïve Bayes algorithm successfully achieved an accuracy rate of 75%. In addition, there is an achievement of 71% precision, 74% recall, and 72% f1-score, as illustrated in the following figure:

Classification report:

precision recall f1-score support

0	0.56	0.71	0.63	7
1	0.87	0.76	0.81	17
accuracy			0.75	24

macro avg 0.71 0.74 0.72 24 weighted avg 0.78 0.75 0.76 24

Accuracy Model : 75 %

Figure 4.9 90:10 naive bayes accuracy results

By applying the K-Nearest Neighbors (KNN) algorithm to the data of grade X vocational students, we analyzed academic grades, and proficiency test results as key features. The K-Nearest Neighbors (KNN) process helps identify similar students based on these attributes, allowing us to predict major labels. Model evaluation involves measuring prediction accuracy and adjusting the K parameter to improve model precision.

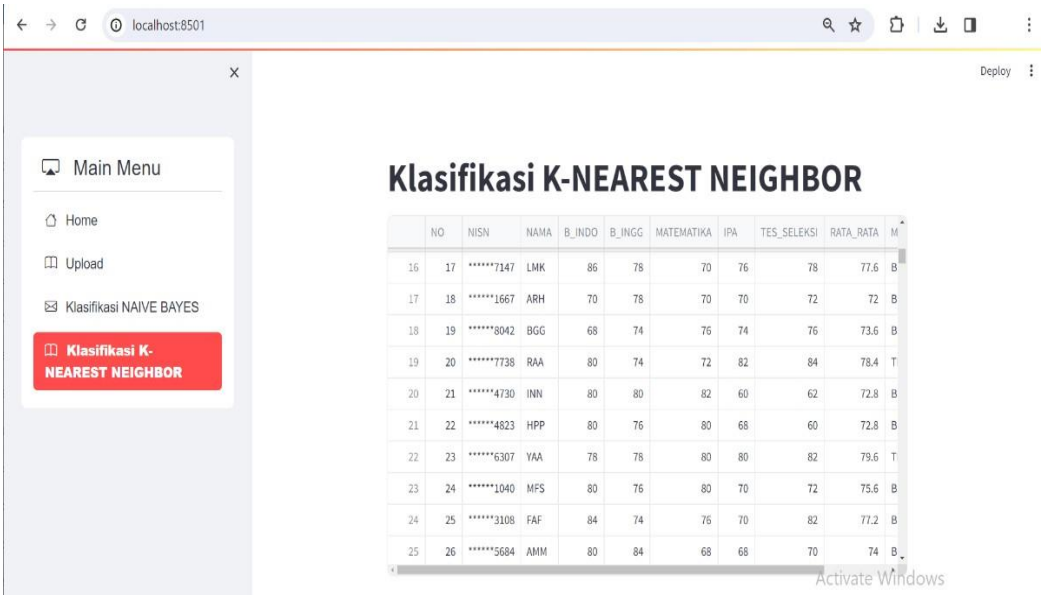

Figure 4.10 KNN classification data display

Figure 4.11 Confusion Matrix K-Nearest Neighbors (KNN) 90:10

The results of the initial experiment with a total of 338 data samples and a training and test data composition of 90:10, consisting of 304 training data and 34 test data showed that the K-Nearest Neighbors (KNN) algorithm successfully achieved an accuracy rate of 75%. In addition, there is an achievement of 87% precision, 57% recall, and 55% f1-score, as illustrated in the following figure:

Model accuracy is : 75.00 %

Classification Report:

precision recall f1-score support

0	1.00	0.14	0.25	7
1	0.74	1.00	0.85	17
accuracy			0.75	24

macro avg 0.87 0.57 0.55 24 weighted avg 0.82 0.75 0.67 24

Figure 4.12 Accuracy results of K-Nearest Neighbors (KNN) 90:10

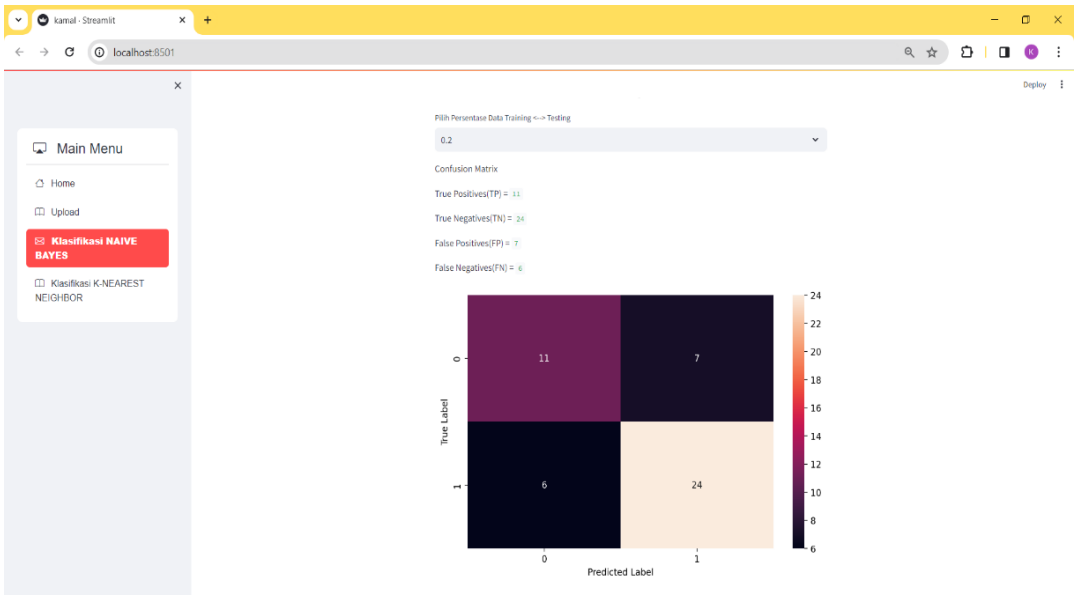

Figure 4.13 Confusion Matrix Naïve Bayes 80:20

The results of the second experiment with a total of 338 data samples and a training and test data composition of 80:20, consisting of 270 training data and 68 test data showed that the Naïve Bayes algorithm successfully achieved an accuracy rate of 73%. In addition, there is an achievement of 71% precision value, 71% recall, and 71% f1-score value, as illustrated in the following figure:

Classification report:

precision recall f1-score support

```
0       0.65      0.61      0.63        18
        1       0.77      0.80      0.79        30

accuracy                            0.73        48
```

macro avg 0.71 0.71 0.71 48 weighted avg 0.73 0.73 0.73 48

Accuracy Model : 73 %

Figure 4.14 Naive Bayes 80:20 accuracy results

Figure 4.15 Confusion Matrix K-Nearest Neighbors (KNN) 80:20

The results of the second experiment with a total of 338 data samples and a training and test data composition of 80:20, consisting of 270 training data and 68 test data showed that the K-Nearest Neighbors (KNN) algorithm successfully achieved an accuracy rate of 100%. In addition, there is an achievement of 100% precision, 100% recall, and 100% f1-score, as illustrated in the following figure:

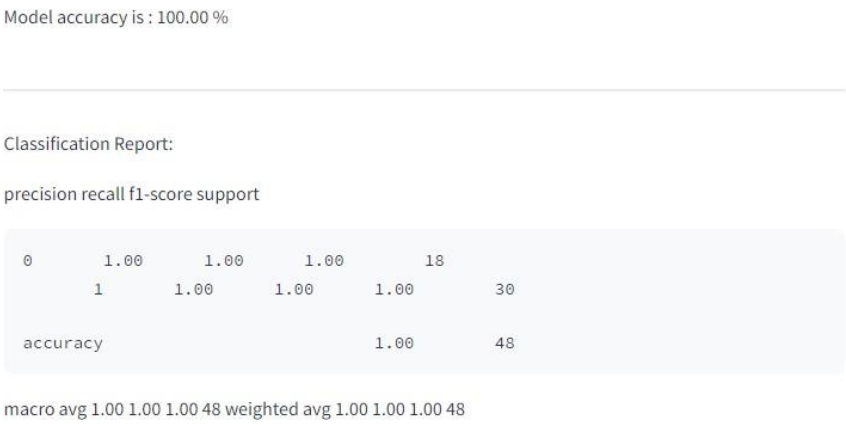

Figure 4.16 K-Nearest Neighbors (KNN) 80:20 accuracy results

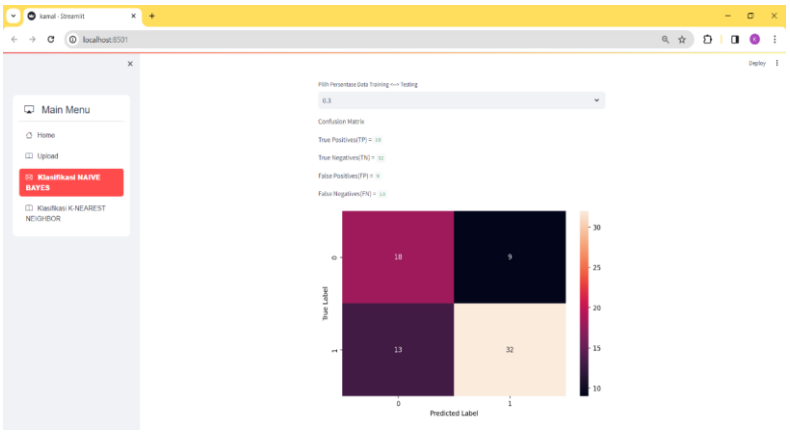

Figure 4.17 Confusion Matrix Naïve Bayes 70:30

The results of the third experiment with a total of 338 data samples and a training and test data composition of 70:30, consisting of 237 training data and 101 test data showed that the Naïve Bayes algorithm successfully achieved an accuracy rate of 69%. In addition, there is an achievement of precision value of 68%, recall of 69%, and f1-score value reaching 68%, as illustrated in the following figure:

Classification report:

precision recall f1-score support

0	0.58	0.67	0.62	27
1	0.78	0.71	0.74	45
accuracy			0.69	72

macro avg 0.68 0.69 0.68 72 weighted avg 0.71 0.69 0.70 72

Accuracy Model : 69 %

Figure 4.18 Naive Bayes 70:30 accuracy results

Figure 4.19 Confusion Matrix K-Nearest Neighbors (KNN) 70:30

The results of the third experiment with a total of 338 data samples and a training and test data composition of 70:30, consisting of 237 training data and 101 test data showed that the K-Nearest Neighbors (KNN) algorithm successfully achieved an accuracy rate of 97%. In addition, there is an achievement of 97% precision, 97% recall, and 97% f1-score, as illustrated in the following figure:

Model accuracy is : 97.22 %

Classification Report:

precision recall f1-score support

0	0.96	0.96	0.96	27
1	0.98	0.98	0.98	45
accuracy			0.97	72

macro avg 0.97 0.97 0.97 72 weighted avg 0.97 0.97 0.97 72

Figure 4.20 K-Nearest Neighbors (KNN) 70:30 accuracy results

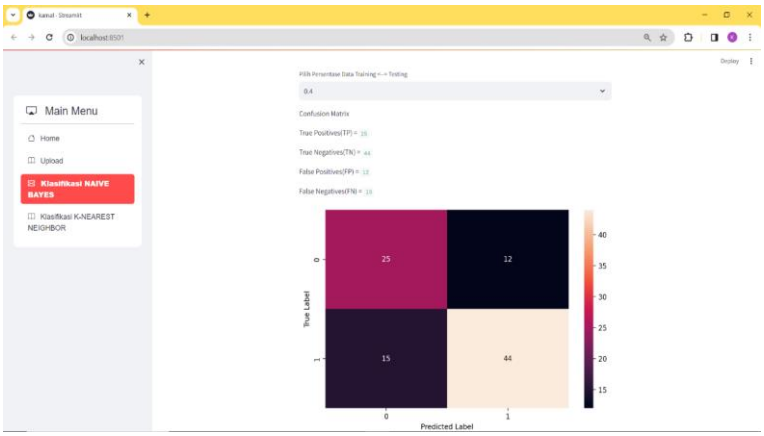

Figure 4.21 60:40 Naive Bayes Confusion Matrix

The results of the fourth experiment with a total of 338 data samples and a training and test data composition of 60:40, consisting of 203 training data and 135 test data show that the Naïve Bayes algorithm successfully achieved an accuracy rate of 72%. In addition, there is an achievement of 71% precision value, 71% recall, and 71% f1-score value, as illustrated in the following figure:

Classification report:

precision recall f1-score support

0	0.62	0.68	0.65	37
1	0.79	0.75	0.77	59
accuracy			0.72	96

macro avg 0.71 0.71 0.71 96 weighted avg 0.72 0.72 0.72 96

Accuracy Model : 72 %

Figure 4.22 Naive Bayes 60:40 accuracy result

Figure 4 23 Confusion Matrix K-Nearest Neighbors (KNN) 60:40

The results of the fourth experiment with a total of 338 data samples and a training and test data composition of 60:40, consisting of 203 training data and 135 test data showed that the K-Nearest Neighbors (KNN) algorithm successfully achieved an accuracy rate of 95%. In addition, there is an achievement of a precision value of 94%, a recall of 95%, and an f1-score value of 95%, as illustrated in the following figure:

Model accuracy is : 94.79 %

Classification Report:

precision recall f1-score support

0	0.90	0.97	0.94	37
1	0.98	0.93	0.96	59
accuracy			0.95	96

macro avg 0.94 0.95 0.95 96 weighted avg 0.95 0.95 0.95 96

Figure 4.24 K-Nearest Neighbors (KNN) 60:40 accuracy results

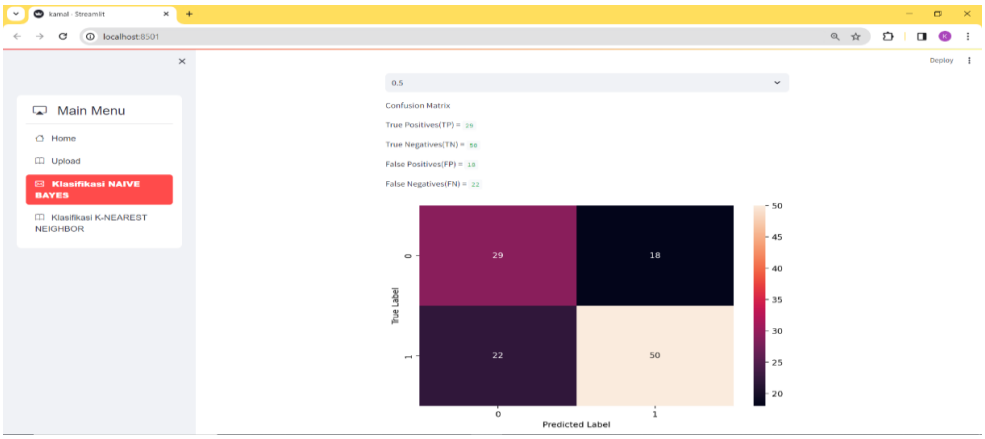

Figure 4.25 Confusion Matrix Naive Bayes 50:50

The results of the fifth experiment with a total of 338 data samples and a training and test data composition of 50:50, consisting of 169 training data and 169 test data showed that the Naïve Bayes algorithm successfully achieved an accuracy rate of 66%. In addition, there is an achievement of 65% precision, 66% recall, and 65% f1-score, as illustrated in the following figure:

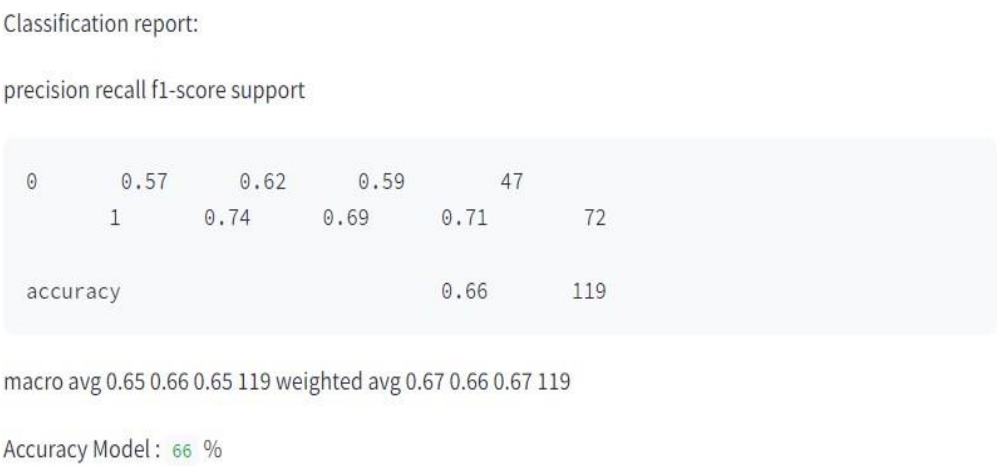

Figure 4.26 Naive Bayes 50:50 accuracy results

Figure 4.27 Confusion Matrix K-Nearest Neighbors (KNN) 50:50

The results of the fifth experiment with a total of 338 data samples and a training and test data composition of 50:50, consisting of 169 training data and 169 test data showed that the K-Nearest Neighbors (KNN) algorithm successfully achieved an accuracy rate of 93%. In addition, there is an achievement of precision value of 93%, recall of 94%, and f1-score value reaching 93%, as illustrated in the following

figure:

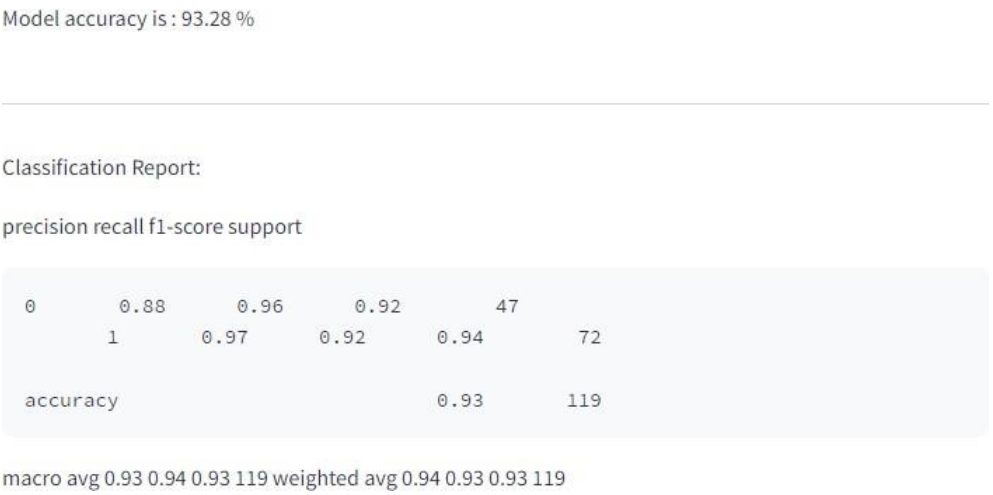

Figure 4.28 K-Nearest Neighbors (KNN) 50:50 accuracy results

4.7. Research Implications

Based on the research that has been done, it can provide research implications for several aspects including system aspects, managerial aspects, and aspects of further research.

System Aspects

To implement the research results, a good support system is needed, so that interested parties can use the research results. Therefore, adequate facilities and infrastructure are needed both in terms of hardware and software used. This research has at least the following specifications:

Table 4.1 Software Requirements

Software	Minimum Requirement
Operating System	Microsoft Windows 2010
Visual studio code	Version 1.86
Python programming	Streamlit

The minimum hardware specifications needed to support the data mining system can be seen in the following table:

Table 4. 2 Hardware Requirements

Hardware	Minimum Requirement
Processor	Core i7
Memory	16 GB
Harddisk	256 GB

Managerial Aspects

Based on the measurement and evaluation results, it shows that the K-Nearest Neighbors (KNN) algorithm is accurate in classifying student majors so that this

algorithm can provide solutions for schools in determining student majors. the contribution to this research is the addition of 5 attributes which include: average entrance test scores, Indonesian, English, Mathematics and Science scores.

Aspects of Future Research

This research still has shortcomings and limitations, so that future research references can consider the following points:

1) Linking aspects that have not been discussed in this research so that deeper research is obtained.
2) The results of this research need to be applied using other algorithmic methods and developed with other optimization algorithms.
3) This research can also be developed every three to four years in order to adjust to existing conditions and situations.

CONCLUSIONS

From the problem identification in the background, it can be concluded that the main problem is the tendency of prospective students to choose majors without considering personal interests, which has a negative impact on lowering grades, especially in productive subjects or certain competencies. This study found that using data from 238 tenth grade students in the 2023/2024 academic year, a model with a test data ratio of 80:20 produced the best performance. k-Nearest Neighbor (KNN) showed superior performance with recall, accuracy, and precision reaching 100%, while Naïve Bayes Classifier only achieved 61% recall and 73% accuracy. This finding confirms that KNN is more effective in identifying and classifying data than Naïve Bayes in the context of this study. Based on the results of the study, it is concluded that the application of the KNN model has the potential as a solution to help prospective students choose majors based on their personal interests, which is expected to improve academic performance and provide more precise directions in future career development. Suggestions made based on this research include conducting annual research to monitor the suitability of majors and students' interests and implementing a committee-guided major selection stage to ensure the suitability between students' interests and their choice of majors, especially in the early stages of selection.

REFERENCES

Help, P., Bpjs, I., Fajar, K., Putro, S., Utami, E., Hartanto, A. D., Yogyakarta, A., Approved, D. D., Key, K., Pbi, :, Bayes, N., & Pso, D. (2022). Particle Swarm Optimization-based Neive Bayes Classification for Prediction. Journal Computer Science, 1(1).

Brilliant, M., Nurhasanah, I. A., & Rahmadaniah, D. (n.d.). Comparison of Naïve Bayes and K-Nearest Neighbor Algorithms for Classification of Alumni Waiting Time in Obtaining Employment (Case Study Smks Pgri 2 Pringsewu).

Fakhri, J., Sunge, A. S., Zy, A. T., & Pelita Bangsa, U. (n.d.). Naive Bayes Algorithm Classification Design on Student Major Selection Data (Vol. 11, Issue 2).

Hardoni, A., Rini, D. P., & Sukemi, S. (2021). SMOTE Integration of Naive Bayes

and Logistic Regression Based on Particle Swarm Optimization for Software Defect Prediction. JOURNAL OF BUDIDARMA INFORMATICS MEDIA, 5(1), 233.

Homepage, J., A'yuniyah, Q., & Reza, M. (n.d.). IJIRSE: Indonesian Journal of Informatic Research and Software Engineering Application of The K-Nearest Neighbor Algorithm For Student Department Classification At 15 Pekanbaru State High School Application of K-Nearest Neighbor Algorithm For Student Department Classification At 15 Pekanbaru State High School.

Indriyani, S., Fatchan, M., & Firmansyah, A. (2023). Precious Metal Price Prediction with Naïve Bayes and Pso Algorithm Approach. In JINTEKS (Vol. 5, Issue 1).

Istighfar, F., Negara, A. B. P., & Tursina, T. (2023). Classification of Student Field of Expertise Using Naive Bayes Algorithm. Journal of Information Systems and Technology (JustIN), 11(1), 77.

Kusumadewi, V. A., Cholissodin, I., & Adikara, P. P. (2020). Classification of Student Majors using K-Nearest Neighbor and Optimization with Genetic Algorithm (Case Study: SMAN 1 Wringinanom Gresik) (Vol. 4, Issue 4).

Lestari, B. A., Hasbi, M., & Susyanto, T. (2019). Selection of the Best School Using the K-Nearest Neighbors Method and Taxonomic Matcher. Journal of Information and Communication Technology (TIKomSiN), 6(2).

Maulana, G. (2023). Application of Machine Learning Algorithm for Vocational High School Student Majoring Based on Report Card and Psychotest Score. Journal of Engineering and Computer Science, 07(01), 56.

Merawati, D. (2019). Application of Data Mining to Determine the Interests and Talents of Smk Students with the C4.5 Method. ALGOR JOURNAL, 1(1).

Mohamad Andri Rasyid, R. K., Riyanto, A., & Widyawati, R. (2023). Implementation Of The Naïve Bayes Algorithm For The Faculty Selection Recommendation System At Amikom University Yogyakarta. In JIKOM: Journal of Informatics and Computers (Vol. 13, Issue 1).

Muhabatin, H., Prabowo, C., Ali, I., Lukman Rohmat, C., Rizki Amalia, D., citation, C., & Rizki, D. (2021). Hoax News Classification Using PSO-Based Naïve Bayes Algorithm. Informatics for Educators and Professionals, 5(2), 156-165.

Muhidin, A., & Casdi, M. (2019). SIGMA-Journal of Technology Pelita Bangsa Optimization of Naïve Bayes Algorithm Based on Particle Swarm Optimization (Pso) and Stratified to Improve the Accuracy of Diabetes Disease Prediction (Vol. 10).

Novaldy, F., & Herliana, A. (2021). Application of Pso to Naïve Bayes for Prediction of Life Expectancy of Heart Failure Patients. RESPONSIVE JOURNAL, 3(1), 37-43.

Nuraeni, S., Syam, S. P. A., Wajdi, M. F., Firmansyah, B., & Malkan, M. (2023). Implementation of K-NN Method to Determine Student Majors at SMAN 02 Manokwari. G-Tech: Journal of Applied Technology, 7(1), 89-95.

Pambudi, A., & Abidin, Z. (2023). Application of Crisp-Dm Using Mlr K-Fold on Stock Data Pt. Telkom Indonesia (Persero) Tbk (Tlkm) (Case Study: Indonesia Stock Exchange 2015-2022). JDMSI, 4(1), 1-14.

Rani, H. A. D. (2021). Particle Swarm Optimization on Naïve Bayes for Baby Birth

Condition Prediction. Journal of Informatics Dialectics (Detika), 2(1), 28-33.

Rifai, A., Aulianita, R., Stmik,), Jakarta, N. M., & Jakartai, N. M. (n.d.). Comparison of C4.5 and Naïve Bayes Classification Algorithms Based on Particle Swarm Optimization for Credit Risk Determination. In Journal Speed - Center for Engineering and Education Research (Vol. 10). CDROM.

Saepudin, S., Muslih, M., Information Systems Studies, P., Nusa Putra, U., Raya Cibolang Kaler No, J., & Sukabumi, K. (2019). Major Selection with K-Nearest Neighbor Method for Prospective New Students. In Jurnal Rekayasa Teknologi Nusa Putra (Vol. 5, Issue 2).

Sayhidin, D., Haris, G., & Juliane, C. (2023). Implementation of Data Mining Student Leadership Level with K-Nearest Neighbor, Decision Tree, and Naïve Bayes. 7(1), 199-206.

Widaningsih, S. (2019). Comparison of Data Mining Methods for Predicting Grades and Graduation Times of Informatics Engineering Study Program Students with C4.5, Naïve Bayes, Knn and Svm Algorithms. Incentive Techno Journal, 13 (1), 16-25.

Widiastuti, N. A., Azhar, M., & Mulyo, H. (2023). Implementation of K- Nearest Neighbor Algorithm for Major Classification in New Learners. SIMETRIS Journal, 14(2).

Wulandari, N., & Etikasari, P. (2019). Analysis of Student Learning Interests at the Indonesian Education Institute Perumnas 3 Bekasi with the C4.5 Method. In Journal of Information Engineering (Vol. 8, Issue 1).

Yudhi Putra, M., & Ismiyana Putri, D. (n.d.). Utilization of Naïve Bayes and K-Nearest Neighbor Algorithms for Class XI Students' Major Classification (Vol. 16, Issue 2).

Yusuf, D., Mubarak, Y., Pangesti, A. R., Wulansari, N., & Zulqornain, R. (2023). Comparison of Naive Bayes Classifier and Decision Tree C4.5 Methods in Finding Interest Patterns for Major Selection in Madrasah Aliyah (Case Study: MA El-Bayan Majenang). In Journal of Information Systems, and Information Technology (Vol. 2, Issue 1).