Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14441
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Predicting Prospective Student Interests Using the C4.5 Algorithm and Naive Bayes

Annisa Amanda[1*)], Ali Akbar Ritonga[2)], Elysa Rohayani Hasibuan[3)],
[1,2,3)]Universitas Labuhanbatu, Indonesia
[1)]annisaamanda272@gmail.com, [2)] aliakbarritonga@gmail.com, [3)] elysa.hasby@gmail.com,

Submitted :Jan 7, 2025 | **Accepted** : Jan 25, 2025 | **Published** : Jan 29, 2025

Abstract: Students are individuals pursuing higher education at a university with the goal of enhancing their knowledge, skills, and character to succeed in the professional world and contribute to society. The purpose of this study is to analyze the factors that influence prospective students' interest in continuing their education using the C4.5 Algorithm and the Naïve Bayes Method. The importance of understanding prospective students' interest patterns is expected to help universities formulate more effective strategies. The purpose of this study is to determine how well the two methods classify data and understand the factors that most influence prospective students' decisions. The C4.5 Algorithm is known to be effective in building decision trees that are easy to interpret, while the Naïve Bayes Method has the advantage of handling datasets with independent attributes. This study uses the stages of data selection, data pre-processing, algorithm application, and model evaluation. The classification results obtained from the C4.5 Algorithm show that 132 data are included in the interest category and 8 data are not interested, while the Naïve Bayes Method produces 131 data of interest and 9 data are not interested. In conclusion, both methods have good accuracy levels, but the Naïve Bayes Method shows superiority in Recall value, while the C4.5 Algorithm excels in interpretation of results and clarity of classification patterns.

Keywords: C4.5 algorithm; Classification; Confusion Matrix; Machine Learning; Naïve Bayes Method;

INTRODUCTION

Students are individuals who are studying at a university with the aim of developing the knowledge, skills, and character needed in the professional world and social life (Rasela, 2022). As agents of change, students are expected to not only focus on academic achievement but also be active in organizational activities, research, and community service (Anam, Nurhakim, & Juliane, 2022). Through the educational process they undertake, students have an important role in bringing innovation and solutions to various problems in the surrounding environment (Budiman, 2021). Armed with the knowledge and experience gained during the study period, students are expected to become the next generation who contribute positively to the development of the nation and state.

Every prospective student certainly has their own goals in choosing the study program they want. These goals can be influenced by various factors, such as personal interests, future career prospects, the desire to develop certain skills, or family expectations (Nas, 2021). Some prospective students may choose a study program based on their passion and talent, while others consider job opportunities and financial stability in the future. Apart from that, the quality of teaching, campus facilities, and the reputation of the study program are also often important considerations. With clear goals, prospective students are expected to be more focused and motivated in completing their education and achieving success in their chosen field.

Universities are institutions designed to meet the needs of education and the world of work. Faculties in these institutions typically cover fields such as education, law, science, technology, economics, and management, all aimed at producing graduates who are ready to compete in the professional world. With a variety of study program choices, universities are committed to providing quality education and preparing students to contribute meaningfully to society.

However, despite offering quality education and diverse study programs, some prospective students remain reluctant to continue their education at certain campuses. Factors such as tuition fees considered expensive, the status of the campus as private, and facilities deemed incomplete by some students often become obstacles. Additionally, issues such as lecturer attendance, quality of knowledge transfer, administrative services perceived

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14441
e-ISSN : 2541-2019
p-ISSN : 2541-044X

as unfriendly, and difficult access to campus also influence perceptions. These conditions create a gap between the academic quality offered by universities and prospective students' perceptions.

From the explanation of the problem above, this research aims to analyze the factors that influence prospective students' interest or disinterest in continuing their education at a university. Through a classification approach, this study will examine interest patterns based on the data obtained and validate previously mentioned problems, such as cost factors, facilities, lecturer quality, and administrative services. It is hoped that the research results will provide a clearer picture of prospective students' perceptions and serve as a basis for universities to make improvements and enhance quality in various aspects.

This research will be carried out by applying machine learning using the C4.5 algorithm and the Naive Bayes method to analyze and classify prospective students' interests(Maizura, Sihombing, & Dar, 2023). The C4.5 algorithm is used to build a decision tree that can identify the main factors influencing interest, while the Naive Bayes method will be used to predict interest tendencies based on the probability of each factor analyzed (Nasution, Dar, & Nasution, 2023). The combination of these two methods is expected to provide accurate and reliable classification results, offering clear insight into interest patterns and helping universities develop more effective strategies to increase campus attractiveness.

LITERATURE REVIEW

Machine Learning

Machine learning is a branch of artificial intelligence that allows systems to learn from data and improve their performance without needing to be explicitly programmed (Abbas et al., 2020). Through certain algorithms, machines can analyze data, recognize patterns, and make predictions or decisions automatically. Machine learning is divided into three main types, namely supervised learning, unsupervised learning, and reinforcement learning, each of which has a different approach to processing data (Lakhdari, Soldevila, Rezgui, & Renault, 2023). The application of machine learning is found in various fields, such as spam email classification, disease detection, customer behavior analysis, and market trend prediction (Study, 2021). With the ability to process large amounts of data and generate deep insights, machine learning has become a very important tool in data-based decision making.

C4.5 Algorithm

The C4.5 algorithm is an algorithm used to build decision trees in machine learning. Developed by Ross Quinlan, this algorithm functions to solve classification problems by dividing data into subsets based on attributes that have the highest information value (Riansyah, Suwilo, & Zarlis, 2023). C4.5 uses the concepts of Information Gain and Entropy to determine which attributes are most effective in separating data into different categories. One of the advantages of this algorithm is its ability to handle data that has numeric and categorical attributes as well as handle data that is incomplete or has missing values (Alam, Alana, & Juliane, 2023). The C4.5 algorithm is widely used in various applications, such as risk analysis, disease prediction, and customer segmentation.

Naïve Bayes Method

The Naive Bayes method is a classification algorithm based on Bayes' Theorem, which predicts the probability of a class based on the attributes in the dataset (Rahman & Fauzi Abdulloh, 2023) (Apriyani, Maskuri, Ratsanjani, Pramudhita, & Rawansyah, 2023). This method is called "naive" because it assumes that each attribute is independent of each other, although in reality this is not always the case (Supendar, Rusdiansyah, Suharyanti, & Tuslaela, 2023) (Tanjung, Tampubolon, Panggabean, & Nandrawan, 2023) (Madjid, Ratnawati, & Rahayudi, 2023). Naive Bayes is very effective for handling large datasets and works well even if the data has independent attributes. This algorithm is often used in text classification, such as sentiment analysis, email spam filters, and document classification. The main advantage of the Naive Bayes method lies in its high computing speed and ability to provide accurate results even with limited or imbalanced data.

METHOD

This research will be carried out by utilizing machine learning technology using the Orange application, an intuitive and effective data visualization and analysis tool (Ahmed, Barczak, Rashid, & Susnjak, 2021) (Esteban, Zafra, & Ventura, 2022). This research was conducted at Labuhanbatu University using the C4.5 and Naive Bayes algorithms to analyze the factors that influence prospective students' interest in choosing the university as a place to study. With the support of the Orange application, the data pre-processing process, feature selection, and algorithm implementation can be carried out more systematically and efficiently. The expected results of the analysis are accurate and relevant insights into prospective students' interest patterns.

The C4.5 algorithm is used in this research to build a decision tree that classifies prospective students' interest in Labuhanbatu University. This algorithm selects attributes with the highest information to separate data, such as tuition and facilities, which helps map the factors that influence prospective students' decisions.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14441
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The Naive Bayes method is used to classify prospective students' interests based on attribute probabilities in the dataset (Hasibuan, Dar, & Yanris, 2023) (Siregar, Irmayani, & Sari, 2023) . Even though it assumes independence between attributes, this method is effective in providing accurate and fast predictions, thereby strengthening the classification results and providing a comparable alternative to the C4.5 Algorithm (Lubis & Chandra, 2023) (Saleh, Dharshinni, Perangin-Angin, Azmi, & Sarif, 2023).

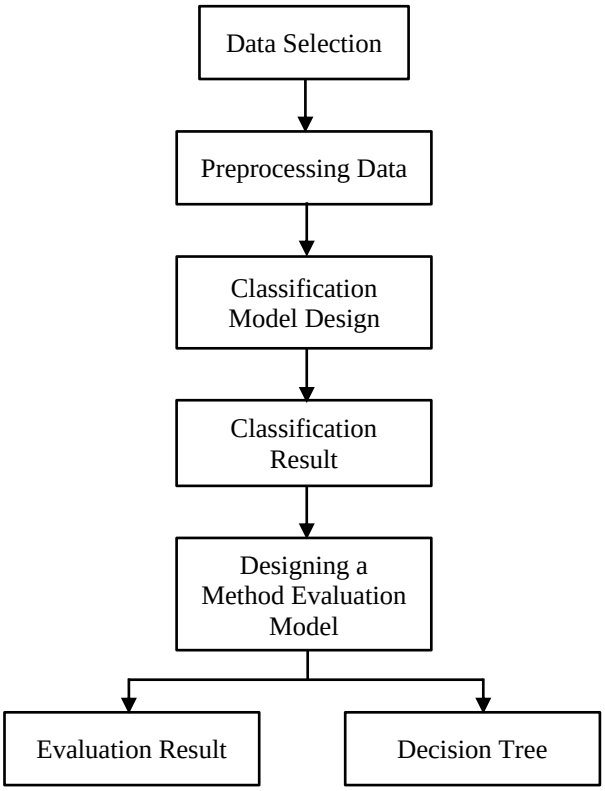

Fig. 1. Research Stages Flow

In the picture above is the process that will be carried out in this research to classify the interests of prospective students and female students. The explanation is as follows.

The process begins with data selection, where relevant and significant data is extracted from a large dataset to enhance the efficiency and quality of model training. This is followed by data preprocessing, which involves cleaning, transforming, and formatting the data to address issues like missing values, normalization, and encoding categorical variables. Next, a classification model design is developed by selecting and applying a suitable algorithm to build a model capable of categorizing data into specific classes based on its features. The classification results represent the predicted categories or labels generated by the model based on learned patterns. To assess the model's performance, a method evaluation model design is implemented, utilizing metrics such as confusion matrices, accuracy, precision, recall, and F1-score. The evaluation results then indicate the model's effectiveness in classifying data according to these metrics. Finally, a decision tree may be employed as a machine learning model in the form of a branching diagram, used to make decisions or classifications by iteratively splitting data based on key features until final results are achieved.

Confusion Matrix

Confusion Matrix is an evaluation tool used to measure the performance of a classification model by comparing the predictions produced by the model with actual data. This matrix consists of four main components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Using these values, various metrics such as accuracy, precision, recall, and F1-score can be calculated to assess how well the model is at classifying data correctly or incorrectly. The Confusion Matrix provides a more detailed picture of model error than using accuracy alone, making it very useful in analyzing classification model performance.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14441
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1. Confusion Matrix

Confusion Matrix		Prediction	
		P	N
Actual	P	True Positive (TP)	False Positive (FP)
	N	False Negative (FN)	True Negative (TN)

To determine the calculation of the confusion matrix, researchers can do it by calculating accuracy, precision and recall.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad \text{(Violita, Yanris, \& Hasibuan, 2023)}$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad \text{(Mawaddah, Dar, \& Yanris, 2023)}$$

$$\text{Accuracy} = \frac{TP}{TP+FN} \times 100\% \quad \text{(Irmayani, Sinaga, \& Masrizal, 2023)}$$

RESULT

Data Selection

Data selection in this research was carried out by collecting two different sets of data, namely training data and testing data. Training data is used to train the model, allowing the model to learn patterns and relationships between features and labels in the data. Meanwhile, testing data is used to test the performance of the model after it has been trained, ensuring that the model can predict accurately on data that has never been seen before. This selection process is important to ensure that the data used is relevant and of good quality, so that the model built can produce more accurate and reliable predictions.

Table 2. Data Training

Name	Access to Location	Teacher Quality	Campus Facilities	Quality of Administrative Services	Cost	Category
Aini Kurniasih	Difficult	Not good	Not good	Not good	Expensive	Not Interested
Akbar Pranata	Easy	Not good	Good	Good	Cheap	Interest
Aldi Firmansyah	Easy	Good	Not good	Not good	Expensive	Not Interested
Aldi Harahap	Easy	Good	Good	Good	Cheap	Interest
Alfi Harahap	Easy	Good	Good	Good	Cheap	Interest
Amelia Nasution	Easy	Good	Good	Good	Cheap	Interest
Andien Siregar	Easy	Good	Good	Good	Cheap	Interest
Aprilia Nasution	Easy	Good	Good	Good	Cheap	Interest
Arsyad Wibawa	Easy	Good	Not good	Good	Cheap	Interest
Arsyila Rambe	Easy	Good	Good	Good	Cheap	Interest
Bagas Siregar	Easy	Good	Good	Good	Cheap	Interest
Good Prasetyo	Easy	Good	Good	Good	Cheap	Interest
Bela Fitriani	Difficult	Good	Not good	Not good	Cheap	Not Interested
Cahyo Wibowo	Easy	Good	Good	Good	Cheap	Interest

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14441
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Daffa Nasution	Easy	Good	Good	Good	Cheap	Interest
Dedi Susanto	Difficult	Not good	Not good	Good	Cheap	Not Interested
Devi Harahap	Easy	Good	Good	Good	Cheap	Interest
Dewi Maharani	Easy	Good	Good	Good	Cheap	Interest
Dewi Triana	Difficult	Not good	Not good	Not good	Expen sive	Not Interested
Dina Ramadhani	Difficult	Good	Good	Not good	Cheap	Interest

In the table above is training data used to assist the process of classifying the interests of prospective students and female students. With the training data above, classification can be done well. The training data used in this research was 67 data from prospective students and female students. The data above is only a sample for the training data used.

Table 3. Data Testing

Name	Access to Location	Teacher Quality	Campus Facilities	Quality of Administrative Services	Cost
Abimanyu Nasution	Difficult	Not good	Good	Good	Cheap
Adelia Permatasari	Easy	Good	Good	Good	Cheap
Adit Prakoso	Easy	Good	Good	Good	Cheap
Aditya Harahap	Easy	Good	Good	Good	Cheap
Adnan Harahap	Easy	Good	Not good	Good	Cheap
Afifah Siregar	Easy	Good	Good	Good	Cheap
Afnan Harahap	Easy	Good	Good	Good	Cheap
Agung Purnomo	Easy	Good	Good	Good	Cheap
Aldianto Hasibuan	Easy	Good	Good	Good	Cheap
Aldo Saputra	Easy	Good	Good	Good	Cheap
Alfariz Nasution	Difficult	Not good	Not good	Good	Cheap
Alvina Nasution	Easy	Good	Good	Good	Cheap
Amelia Cahaya	Easy	Good	Good	Good	Cheap
Amira Harahap	Easy	Good	Good	Good	Cheap
Ananda Nasution	Easy	Not good	Not good	Not good	Expens ive
Andi Kurnia	Easy	Good	Good	Good	Cheap
Andika Nasution	Easy	Good	Good	Good	Cheap
Arif Wibisono	Easy	Good	Good	Good	Cheap
Arina Putri	Easy	Good	Good	Good	Cheap

*name of corresponding author

Arini Hasibuan	Easy	Good	Good	Good	Cheap

In the table above is the testing data used in this research. The testing data used was 140 prospective students and female students. The data above will later be used to classify using the orange application. The data above is only a sample for this research.

Preprocessing Data

In this research, the data preprocessing stage was carried out more simply because the data used was of good quality and neatly structured. The preprocessing process aims to ensure that the data is ready to be analyzed using the C4.5 and Naive Bayes algorithms in the Orange application. Even though the data is in good condition, several basic steps such as checking data completeness, normalization, and removing duplicate data are still carried out to ensure the consistency and accuracy of the analysis results. With optimal data, the analysis process can run more efficiently and provide more accurate results in identifying factors that influence prospective students' interest in studying.

Classification Model Design

The classification model design stage is the process carried out to form and design the model used in this research. The model used was designed using the orange application.

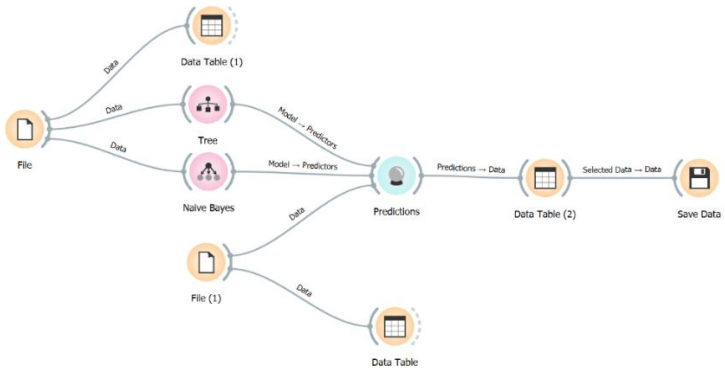

Fig. 2. Classification Model

In the image above is the classification model that will be used for this research. With the model above, classification results will be obtained using the c4.5 algorithm and the Naive Bayes method. The method can be seen in the red widget image above.

Classification Results

In the classification results of this researcher, the classification results obtained will be presented in the form of a table below, namely as follows.

Table 4 Classification Results

Name	Access to Location	Teacher Quality	Campus Facilities	Quality of Administrative Services	Cost	C4.5 Algorithm	Naïve Bayes
Abimanyu Nasution	Difficult	Not good	Good	Good	Cheap	Interest	Interest
Adelia Permatasari	Easy	Good	Good	Good	Cheap	Interest	Interest
Adit Prakoso	Easy	Good	Good	Good	Cheap	Interest	Interest
Aditya Harahap	Easy	Good	Good	Good	Cheap	Interest	Interest
Adnan Harahap	Easy	Good	Not good	Good	Cheap	Interest	Interest
Afifah Siregar	Easy	Good	Good	Good	Cheap	Interest	Interest

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14441
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Afnan Harahap	Easy	Good	Good	Good	Cheap	Interest	Interest
Agung Purnomo	Easy	Good	Good	Good	Cheap	Interest	Interest
Aldianto Hasibuan	Easy	Good	Good	Good	Cheap	Interest	Interest
Aldo Saputra	Easy	Good	Good	Good	Cheap	Interest	Interest
Alfariz Nasution	Difficult	Not good	Not good	Good	Cheap	Interest	Not Interested
Alvina Nasution	Easy	Good	Good	Good	Cheap	Interest	Interest
Amelia Cahaya	Easy	Good	Good	Good	Cheap	Interest	Interest
Amira Harahap	Easy	Good	Good	Good	Cheap	Interest	Interest
Ananda Nasution	Easy	Not good	Not good	Not good	Expensive	Not Interested	Not Interested
Andi Kurnia	Easy	Good	Good	Good	Cheap	Interest	Interest
Andika Nasution	Easy	Good	Good	Good	Cheap	Interest	Interest
Arif Wibisono	Easy	Good	Good	Good	Cheap	Interest	Interest
Arina Putri	Easy	Good	Good	Good	Cheap	Interest	Interest
Arini Hasibuan	Easy	Good	Good	Good	Cheap	Interest	Interest

In the table above are the results obtained from this research using the C4.5 Algorithm and the Naive Bayes method. The classification results obtained from the C4.5 Algorithm were 132 data of interest and 8 data of no interest. The classification results obtained from the Naive Bayes Method were 131 data of interest and 9 data of no interest.

Designing an Evaluation Model

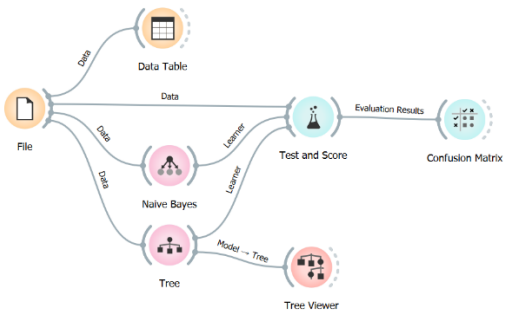

Fig. 3. Evaluation Model

In the picture above is the evaluation model used to determine the evaluation method used in this research. Evaluation of the C4.5 Algorithm method and the Naive Bayes Method was carried out to measure the performance of the two algorithms in classifying data related to the interests of prospective students. This evaluation process involves measuring accuracy, precision, recall, and f1-score to understand the extent to which both methods can predict results correctly. The C4.5 algorithm is expected to be able to form a decision tree that is clear and easy to interpret, while Naive Bayes is expected to be able to handle probability well in determining the interest patterns of prospective students. The results of this evaluation will later be compared to determine which method is superior and more effective in analyzing the data used in this research.

Evaluation Results

For the evaluation results carried out by the author, the author used the test and score widget and confusion matrix to obtain accurate results. The evaluation results are as follows.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14441
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Test and Score

Table 5
Test and Score Results

Model	AUC	CA	F1	Precision	Recall	MCC
Algoritma C4.5	0.867	0.971	0.971	0.971	0.971	0.735
Naïve Bayes	0.999	0.979	0.980	0.984	0.979	0.840

In the table above are the Test and Score results from the method used. For the results obtained, the accuracy of the Naïve Bayes method is better than the accuracy results of the C4.5 Algorithm. This means that the Naïve Bayes method is better than the C4.5 algorithm.

Confusion Matrix Results

Table 1Confusion Matrix Results in C4.5 Algorithm

		Predicted		
		Interest	Not Interested	∑
Actual	Interest	130	2	132
	Not Interested	2	6	8
	∑	132	8	140

In the table above is the composition of the confusion matrix results obtained from evaluating the KNN method in data mining. For the results, there are True Positive (TP) results which are 130 data, for True Negative (TN) results which are 6 data, for False Positive (FP) results which are 2 and for False Negative (FN) results which are 2 data. For these results, it is not possible to directly measure the accuracy value, the data above must be calculated first using the formula in the confusion matrix, which is as follows.

$$\boldsymbol{Accuracy} = \frac{130+6}{130+6+2+2} + \mathbf{100\%}$$ Then the Accuracy value = 97%

$$\boldsymbol{Precision} = \frac{130}{130+2} + \mathbf{100\%}$$ Then the Precision value = 98%

$$\boldsymbol{Recall} = \frac{130}{130+2} + \mathbf{100\%}$$ Then the Recall value = 98%

Table 7. Confusion Matrix Results in the Naïve Bayes Method

		Predicted		
		Interest	Not Interested	∑
Actual	Interest	129	3	132
	Not Interested	0	8	8
	∑	129	11	140

In the table above is the composition of the confusion matrix results obtained from evaluating the KNN method in data mining. For the results, there are True Positive (TP) results which are 129 data, for True Negative (TN) results which are 8 data, for False Positive (FP) results which are 3 and for False Negative (FN) results which are 0 data. For these results, it is not possible to directly measure the accuracy value, the data above must be calculated first using the formula in the confusion matrix, which is as follows.

$$\boldsymbol{Accuracy} = \frac{129+8}{129+8+3+0} + \mathbf{100\%}$$ Then the Accuracy value = 97%

$$\boldsymbol{Precision} = \frac{129}{129+3} + \mathbf{100\%}$$ Then the Precision value = 97%

$$\boldsymbol{Recall} = \frac{129}{129+0} + \mathbf{100\%}$$ Then the Recall value = 100%

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14441
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Based on the results of the Confusion Matrix evaluation, the C4.5 Algorithm and the Naïve Bayes Method both achieved an accuracy level of 97%, indicating excellent performance in classifying data. However, there are differences in the Precision and Recall values. The C4.5 Algorithm has a Precision and Recall value of 98% each, indicating good consistency in predicting the data of interest. On the other hand, the Naïve Bayes Method has a Precision value of *97% and a higher Recall value of 100%, indicating its better ability to detect positive data perfectly. Thus, although the accuracy of both is the same, the Naïve Bayes Method is superior in terms of sensitivity (Recall), while the C4.5 Algorithm is more stable in general classification accuracy.

Decision Tree

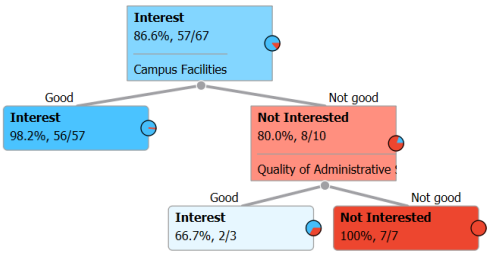

Fig. 4. C4.5 Algorithm Decision Tree Results

Based on the results of the C4.5 Algorithm decision tree model in the image above, the main factor influencing prospective students' interest in studying is the quality of administrative services. If the quality of administrative services is considered good, then all prospective students (100%, 130/130) tend to have an interest in continuing their studies at that campus. However, if the quality of administrative services is considered to be poor, the next influencing factor is campus facilities. From this data, 80% of prospective students (8/10) who rated campus facilities as poor tended not to be interested in studying. Meanwhile, if campus facilities are considered good, prospective students again show interest (100%, 2/2). On the other hand, if campus facilities are considered poor, then all prospective students in this category (100%, 8/8) have no interest in continuing their studies. This shows that the quality of administrative services and campus facilities plays an important role in shaping prospective students' decisions.

DISCUSSIONS

The application of the C4.5 Algorithm and the naïve bayes method provides good classification results, each method has its own advantages. The classification results of the C4.5 Algorithm obtained results of 132 interest data and 8 data without interest. The classification results obtained from the Naive Bayes Method were 131 interest data and 9 no interest data. These results indicate that many prospective students are interested in studying at the Campus. For the accuracy results obtained, both methods used have fairly good accuracy results.

The accuracy results obtained by the c4.5 algorithm are 97% and the accuracy results obtained by the Naïve Bayes method are also 97%. Both accuracy results are equally good because they provide almost perfect accuracy results. The comparison of the two methods is 1:1, which gives equally good results. This is also the result obtained is not free from the arubut used from access to the location that provides suitable variables for this study and not only that such as attributes of teacher quality, campus facilities, costs and quality of bureau services also provide good attributes for this study, so that this study can provide good results. The method used to provide good results. This means that this method used in this study is very suitable for use as a data classification method.

CONCLUSION

Based on the research results, a comparison between the C4.5 Algorithm and the Naïve Bayes Method shows that both methods have the same level of accuracy, namely 97%. However, there are differences in other metric values, where the C4.5 Algorithm has a Precision and Recall value of 98% each, while the Naïve Bayes Method has a Precision value of 97% and a higher Recall value, namely 100%. In addition, the AUC value for the C4.5 Algorithm is 0.867, while the Naïve Bayes Method reaches 0.999, which shows superiority in the ability to distinguish positive and negative classes. The results from the Confusion Matrix also confirm that the Naïve Bayes Method is superior in minimizing the prediction error for the negative class (False Negative). Thus, even though both methods show good performance, the Naïve Bayes Method has a slight advantage in classifying data related to prospective students' interest in studying.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

REFERENCES

Abbas, S. A., Aslam, A., Rehman, A. U., Abbasi, W. A., Arif, S., & Kazmi, S. Z. H. (2020). K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir. *IEEE Access*, *8*, 151847–151855. https://doi.org/10.1109/ACCESS.2020.3014021

Ahmed, N., Barczak, A. L. C., Rashid, M. A., & Susnjak, T. (2021). A parallelization model for performance characterization of Spark Big Data jobs on Hadoop clusters. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00499-7

Alam, A., Alana, D. A. F., & Juliane, C. (2023). Comparison Of The C.45 And Naive Bayes Algorithms To Predict Diabetes. *Sinkron*, *8*(4), 2641–2650. https://doi.org/10.33395/sinkron.v8i4.12998

Anam, K., Nurhakim, B., & Juliane, C. (2022). Komparasi Algoritma Klasifikasi Data Mining Menggunakan Optimize Selection untuk PeInterestan Program Studi. *Building of Informatics, Technology and Science (BITS)*, *4*(2), 606–613. https://doi.org/10.47065/bits.v4i2.2160

Apriyani, M. E., Maskuri, R. A., Ratsanjani, M. H., Pramudhita, A. N., & Rawansyah, R. (2023). Digital Forensic Investigates Sexual Harassment on Telegram using Naïve Bayes. *Sinkron*, *8*(3), 1409–1417. https://doi.org/10.33395/sinkron.v8i3.12514

Budiman, B. (2021). Perbandingan Algoritma Klasifikasi Data Mining untuk Penelusuran Interest Calon Mahasiswa Baru. *Nuansa Informatika*, *15*(2), 37–52. https://doi.org/10.25134/nuansa.v15i2.4162

Esteban, A., Zafra, A., & Ventura, S. (2022). Data mining in predictive maintenance systems: A taxonomy and systematic review. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1471

Hasibuan, F. F., Dar, M. H., & Yanris, G. J. (2023). Implementation of the Naïve Bayes Method to determine the Level of Consumer Satisfaction. *SinkrOn*, *8*(2), 1000–1011. https://doi.org/10.33395/sinkron.v8i2.12349

Irmayani, D., Sinaga, F. A., & Masrizal, M. (2023). Analysis of the Level of Public Satisfaction on the Tiktok Application as an E-Commerce. *Sinkron*, *8*(4), 2579–2591. https://doi.org/10.33395/sinkron.v8i4.13040

Lakhdari, Y., Soldevila, E., Rezgui, J., & Renault, É. (2023). Detection of Plant Diseases in an Industrial Greenhouse: Development, Validation & Exploitation. *2023 International Symposium on Networks, Computers and Communications, ISNCC 2023*. https://doi.org/10.1109/ISNCC58260.2023.10323932

Lubis, A. I., & Chandra, R. (2023). Forward Selection Attribute Reduction Technique for Optimizing Naïve Bayes Performance in Sperm Fertility Prediction. *Sinkron*, *8*(1), 275–285. https://doi.org/10.33395/sinkron.v8i1.11967

Madjid, F. M., Ratnawati, D. E., & Rahayudi, B. (2023). Sentiment Analysis on App Reviews Using Support Vector Machine and Naïve Bayes Classification. *Jurnal Dan Penelitian Teknik Informatika*, *8*(1), 556–562. Retrieved from https://doi.org/10.33395/sinkron.v8i1.12161

Maizura, S., Sihombing, V., & Dar, M. H. (2023). Analysis of the Decision Tree Method for Determining Interest in Prospective Student College. *SinkrOn*, *8*(2), 956–979. https://doi.org/10.33395/sinkron.v8i2.12258

Mawaddah, A., Dar, M. H., & Yanris, G. J. (2023). Analysis of the SVM Method to Determine the Level of Online Shopping Satisfaction in the Community. *SinkrOn*, *8*(2), 838–855. https://doi.org/10.33395/sinkron.v8i2.12261

Nas, C. (2021). Data Mining Prediksi Interest Calon Mahasiswa Memilih Perguruan Tinggi Menggunakan Algoritma C4.5. *Jurnal Manajemen Informatika (JAMIKA)*, *11*(2), 131–145. https://doi.org/10.34010/jamika.v11i2.5506

Nasution, R. F., Dar, M. H., & Nasution, F. A. (2023). Implementation of the Naïve Bayes Method to Determine Student Interest in Gaming Laptops. *Sinkron*, *8*(3), 1709–1723. https://doi.org/10.33395/sinkron.v8i3.12562

Rahman, R., & Fauzi Abdulloh, F. (2023). Performance of Various Naïve Bayes Using GridSearch Approach In Phishing Email Dataset. *Sinkron*, *8*(4), 2336–2344. https://doi.org/10.33395/sinkron.v8i4.12958

Rasela, F. (2022). Pengaruh Literasi Wakaf terhadap Interest Mahasiswa Berwakaf pada Forum Wakaf Mahasiswa Indonesia. *Jurnal Riset Perbankan Syariah*, 69–76. Retrieved from https://journals.unisba.ac.id/index.php/JRPS/article/view/969

Riansyah, M., Suwilo, S., & Zarlis, M. (2023). Improved Accuracy In Data Mining Decision Tree Classification Using Adaptive Boosting (Adaboost). *SinkrOn*, *8*(2), 617–622. https://doi.org/10.33395/sinkron.v8i2.12055

Saleh, A., Dharshinni, N., Perangin-Angin, D., Azmi, F., & Sarif, M. I. (2023). Implementation of Recommendation Systems in Determining Learning Strategies Using the Naïve Bayes Classifier Algorithm. *Sinkron*, *8*(1), 256–267. https://doi.org/10.33395/sinkron.v8i1.11954

Siregar, A. P., Irmayani, D., & Sari, M. N. (2023). Analysis of the Naïve Bayes Method for Determining Social Assistance Eligibility Public. *SinkrOn*, *8*(2), 805–817. https://doi.org/10.33395/sinkron.v8i2.12259

Study, A. C. (2021). *applied sciences Data Mining Techniques for Early Diagnosis of Diabetes :* 1–12.

Supendar, H., Rusdiansyah, R., Suharyanti, N., & Tuslaela, T. (2023). Application of the Naïve Bayes Algorithm in Determining Sales Of The Month. *SinkrOn*, *8*(2), 873–879. https://doi.org/10.33395/sinkron.v8i2.12293

Tanjung, J. P., Tampubolon, F. C., Panggabean, A. W., & Nandrawan, M. A. A. (2023). Customer Classification

*name of corresponding author

Using Naive Bayes Classifier With Genetic Algorithm Feature Selection. *Sinkron*, *8*(1), 584–589. https://doi.org/10.33395/sinkron.v8i1.12182

Violita, P., Yanris, G. J., & Hasibuan, M. N. S. (2023). Analysis of Visitor Satisfaction Levels Using the K-Nearest Neighbor Method. *SinkrOn*, *8*(2), 898–914. https://doi.org/10.33395/sinkron.v8i2.12257

*name of corresponding author