COMPARATIVE ANALYSIS OF ALGORITHMS NAÏVE BAYES AND C45 FOR STUDENT SATISFACTION WITH ADMINISTRATIVE SERVICES

1st Ramadani Ramadani
Computer of Science
Potensi Utama University
Medan, Indonesia
ramadans.ordinary@gmail.com

2nd B.Herawan Hayadi
Computer of Science
Potensi Utama University
Medan, Indonesia
b.herawan_hayadi@gmail.com

3rd Hartono Hartono
Computer of Science
Potensi Utama University
Medan, Indonesia
hartonoibbi@gmail.com

2023 International Conference of Computer Science and Information Technology (ICOSNIKOM) | 979-8-3503-6075-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICOSNIKOM60230.2023.10364373

***Abstract*— One of the reasons for increasing the number of people interested in schools is the service aspect. This is faced with the demands of the interests of various parties, such as leaders, shareholders (foundations), teachers and students at Siti Banun Vocational School. For this reason, good and effective management is needed so that the target of service recipients, namely the level of student satisfaction with administrative services, can be achieved. Data mining is an academic field focused on the systematic exploration and analysis of extensive datasets to discover valuable insights through the extraction of knowledge and detection of patterns. The best classification algorithm currently is the C45 algorithm and Naïve Bayes. The accuracy and AUC values will be compared. From 126 RPL Department Data from Class From the comparison of the amount of training data and test data, the best level of accuracy and AUC is by dividing test data: training data 10: 90. The accuracy and AUC values for the C45 algorithm are 85% and 88%. each. The accuracy and AUC values for the Naïve Bayes algorithm are 92% and 94% respectively.**

Keywords*— *Satisfaction Level, Data Mining, C45, Naïve Bayes

I. INTRODUCTION

Based on the data obtained from the Central Bureau of Statistics, the number of educational institutions in Indonesia exhibits a persistent upward trend on an annual basis. This includes a significant increase in the number of Vocational High Schools (SMK). North Sumatra has 270 public and 703 private SMKs with 137,972 students for public SMKs and 170,921 for private SMKs [1]. From the BPS data above, it can be concluded that this increase reflects the efforts of the government and society in providing wider and more diverse access to education for students across the country. The data also indicates that competition between SMKs is also getting higher and tighter, which requires each SMK to maintain its survival by competing for a large number of students. The number of enrollees in a private educational institution serves as a reliable metric for measuring its growth. It is imperative to acknowledge that the rise in student population is deeply interconnected with the level of student contentment within the institution. In addition to facilities and teachers, the factor that most supports the quality of schools is administrative services. Data Mining is a scientific field that focuses on uncovering knowledge and identifying patterns from extensive data. The act of extracting meaningful and previously unrecognized information from data is known as Data Mining[2].

Naïve Bayes is a classification algorithm that utilizes probabilities for predicting classes from data. The algorithm computes the likelihood of a class based on its features and selects the class with the highest likelihood as the prediction [3], [4]. This algorithm works under the assumption that all attributes are independent, although this is often not accurate in the real world. Nonetheless, Naive Bayes often gives good results and is used in many fields such as text analysis, email classification, and more [5].

Research with the title Detection of Samarinda Sarong Using Naïve Bayes Method Based on Image Processing [6] obtained an accuracy rate of 98.4%. There exist other investigations employing the naïve Bayes algorithm, specifically the research entitled Comparison of Feature Selection Optimization on Naïve Bayes for Airline Passenger Satisfaction Classification [7] achieved a precision rate of 86.13%.

He C4.5 algorithm is a method for generating a decision tree for data classification. The algorithm chooses the most valuable characteristics to divide the data into clusters using information gain as a criterion. This process is iterated until a fully grown tree or decision rules are obtained. The C4.5 algorithm is capable of handling heterogeneous data and extracting meaningful information from it[8].

In research [4] explains that the C45 algorithm is able to predict better than test data compared to sampling data. Research conducted on the C45 algorithm, which involved examining the C45 method for identifying factors contributing to student satisfaction in online learning, demonstrated the successful application of the C45 algorithm for analyzing these factors related to student satisfaction in online learning. As for the results obtained from questionnaire data and the application of the C45 algorithm, information can be obtained that the most important aspect of obtaining online learning satisfaction is the facilitation of student-lecturer and student-student interactions with students, the next most important aspect is online learning that can facilitate the completeness of learning objects, while the accuracy of predicting online learning satisfaction obtained from test data obtained an accuracy rate of 75%. Naïve Bayes has a unique ability to recognize patterns in data. Particularly in text analysis, such as in sentiment analysis on social media, Naïve Bayes can effectively identify deep patterns, aiding in the understanding of the broader context [9]. Moreover, based on the investigations conducted on the C45 algorithm, it can be deduced that the C45 algorithm is extensively employed in the process of data mining to derive multiple inferences, which are then presented in the form of a decision tree[10]. The research aims to evaluate the efficacy of the Naïve Bayes algorithm and the C45 algorithm in predicting student satisfaction with administrative services at Siti Banun Private Vocational High School (SMK). It seeks to determine the algorithm with superior performance and accuracy through a comparative analysis.

II. LITERTURE REVIEW

A. *Data Mining*

Data mining is a scientific method employed to extract valuable knowledge or concealed regularities from vast sets of data. The primary objective of data mining is to discern connections, regular patterns, and evolving tendencies within data, which can be utilized for making informed decisions or making predictions [11], [12]. Data mining is described as the process of employing computational techniques to extract valuable and practical insights from vast databases.[13].

B. Naïve Bayes

Naive Bayes is a machine learning technique that employs probability computations. This algorithm leverage the probability and statistical approach proposed by Thomas Bayes, a British scientist, to predict future probabilities based on prior experience. The Naïve Bayes classifier algorithm is a classification method that relies on the conditional probability of Bayes' theorem [3]. Naive Bayes is a prominent classification algorithm in the field of data mining, renowned for its superior classification efficacy, leading to its extensive application in practical classification tasks.

The workings of the Naïve Bayes classifier method can be sorted as the following steps [6] :

1. Consider a training dataset, D, which contains data rows and their corresponding class labels. Each row in D has attributes that are represented by n-dimensional vectors, X = (x1, x2, ... xn).

   This statement elucidates n observations conducted across n variables, denoted as A1, A2,...An.

2. Given m categories, C1, C2, ...Cm and a sample X, the classifier will determine that X belongs to a category with a conditional probability (posteriori probability), given X. X will be classified as belonging to category Ci only if:

   $$P(Ci \mid X) > P(Cj \mid X) \text{ untuk } 1 \le j \le m, j \ne I \quad (1)$$

   Therefore, the highest probability class P(Ci | X) was determined. The class Ci that corresponds to the maximum value of P(Ci | X) is referred to as the maximum a posteriori hypothesis. This concept is based on Bayes' theorem equation.

   $$P(C_i \mid X) = \frac{P(X \mid C_i)\, P(C_i)}{P(X)} \quad (2)$$

   With:

   $P(C_i \mid X)$ = Hypothesis probability of class Ci based on condition X

   $P(X \mid C_i)$ = Probability of X data based on condition in class C_i

   $P(C_i)$ = Initial class probability Ci

   $P(X)$ = Initial probability of data

   X

3. The probability distribution P(X) is invariant across all classes, with the objective being to maximize the product of the conditional probability P(X | Ci) and the prior class probability P(Ci). In cases where the prior class probability P(Ci) is unknown, it is commonly assumed that P(Ci) is equal for all classes (P(Ci) = P(C2) = ... = P(Cm)) in order to maximize P(X | Ci). However, the goal is to optimize P (X | Ci) P (Ci) instead. It is worth mentioning that the a priori probability of a class can be approximated by P(Ci)= |Ci, D| / |D|, where |Ci, D| refers to the count of duplicate instances of class Ci in D.
4. Datasets with a high number of attributes require significant computational resources to calculate the conditional probability P(X | Ci). Based on the simplistic assumption that classes exhibit conditional independence.

   It is postulated that attribute values exhibit conditional independence in relation to one another, given a particular sample class. This postulate can be expressed mathematically as:

   $$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$$

   $$= P(x_1 \mid C_i) \times P(X_2 \mid C_i) \times \ldots \times P(X_n \mid C_i) \quad (3)$$

5. In order to make a prediction about the class label of X, we need to evaluate P(X | Ci) P(Ci) for each class Ci. The classifier accurately estimates that the categorization of X belongs to class Ci if and only if.

   $$P(X \mid C_i)\ P(C_i) > P(X \mid C_j) \text{ untuk } 1 \le j \le m, j \ne i \quad (4)$$

C. *C4.5 Algorithm*

The C45 algorithm is a component of the algorithm utilized for creating structured partitions or clusters within the dataset. The underlying foundation of the C4.5 algorithm revolves around a variation of a decision tree structure [8].

A decision tree can be used to represent and make decisions, it can be seen as a series of points consisting of single nodes or spread to leaf parts [10], [14], [15]. The attribute chosen as the root is determined by selecting the attribute with the highest information gain among all available attributes. The information gain is calculated using the following formula:

$$\text{Gain S, A} = \text{Entropy S} - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \quad (5)$$

Explanation

S = set of cases
A = attribute
N = number of partitions of attribute A
|Si| = number of cases in partition i
|S| = Total number of cases in S

$$Entropy\ S = \sum_{i=1}^{n} - pi * log2\ pi$$

Explanation

S = set of cases
A = features
N = number of partitions
Pi = ratio of Si to S

III. METHODOLOGY

A. Research Stages

In this study, an assessment of the C4.5 and Naive Bayes algorithm classification techniques is conducted in order to address the research problem at hand. The following is the framework that the author makes for research as Figure 1 below

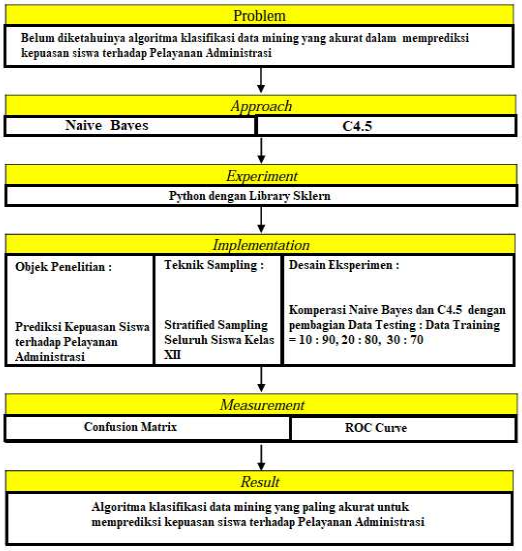

FIG 1. FRAMEWORK OF THOUGHT

In this study, the research method used is an experimental research method with stages as in Figure 2 below:

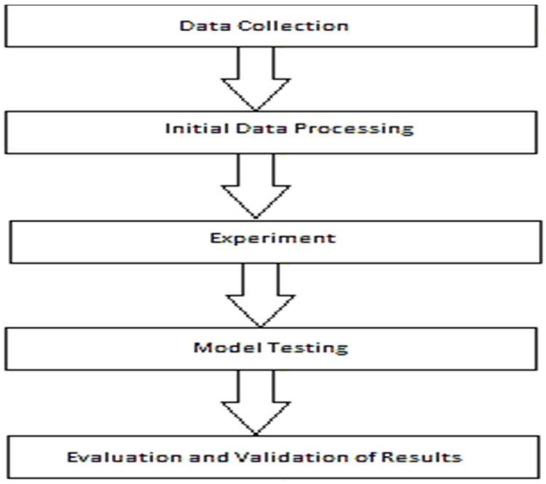

FIG 2. RESEARCH STAGES

From Figure 2 above, it is explained step by step :

1. Data Collection

In implementing the comparison of Algorithm C45 and Naïve Bayes requires data to train the model so as to get the best model. The data utilized is primary data acquired through the tabulation of questionnaires. The quantity of data derived from the questionnaire tabulation results is 126 data points comprising 11 variables. The categories include: title, division, mobile/WhatsApp number, inquiry 1, inquiry 2, inquiry 3, inquiry 4, inquiry 5, and inquiry 6. The dependent variable is attitude. Questions 1 to 6 will necessitate responses in the format of a Likert scale, consisting of the subsequent answer alternatives:

1. Not very good
2. Not Good
3. Average
4. Good
5. Very Good

While for the 9th attribute or 9th question the answers are only Satisfied (1) and Dissatisfied (0). In Figure 3 below is an example of Tabulation of Dataset:

TABLE I. DATASET OF STUDENT SATISFACTION SURVEY

Email	rizky071105@gmail.com
Name	Rizky Alwali Syahputra
Major	XII RPL 1
Phone	085265414328
Question 1	very efficient
Question 2	very good
Question 3	very easy to access
Question 4	very friendly
Question 5	very effective
Question 6	very good
Attitude	satisfied

2. Pre-processing

Prior to data processing, the data undergoes the process of data cleansing. Only a subset of fields from the primary data, specifically Question1, Question2, Question3, Question4, Question5, Question6, and Attitude, are utilized. After that, the answers to each question are converted into numeric with a predetermined value for the answer points for each question. The pre-processing outcomes of the primary data can be observed in Figure 3 depicted beneath:

TABLE II. DATASET AFTER PREPROCESSING

No	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	attitude
1	5	5	5	5	5	5	1
2	3	3	3	3	3	3	0
3	3	3	3	3	3	3	0
4	3	3	4	3	5	3	0
5	3	4	4	3	4	4	1
6	4	4	4	4	5	3	1
7	3	4	5	3	5	4	1
8	4	3	4	4	5	4	1
9	4	4	3	4	4	4	1
10	3	3	3	3	5	3	1
11	3	3	2	3	3	3	0
12	5	5	5	5	5	5	1
13	4	4	4	4	4	4	1
14	3	2	3	3	3	4	1
15	3	3	3	3	3	3	0
16	5	5	5	5	5	5	0
17	5	5	4	5	5	3	0
18	5	5	5	5	5	5	1
19	4	3	3	4	3	3	0
20	3	4	4	3	4	4	1
21	2	3	3	2	3	3	0
22	4	4	4	4	4	5	0
23							
126	5	4	4	5	4	3	1

The aforementioned dataset has been partitioned by dividing the dataset. Before splitting the dataset, the step taken is to split the variables from the dataset, namely predictor variables and target variables and the results after splitting the data variables are converted into numpy form so that computing is easier. The stages of splitting the data can be observed in Figure 3 presented underneath:

```
kolom_tertentu = df.iloc[:, [0,1,2,3,4,5]]
# kolom_tertentu

hasil = df.iloc[:,6]
```

```
arr = kolom_tertentu.to_numpy()
arrHasil = hasil.to_numpy()
arr
```

```
array([[5, 5, 5, 5, 5, 5],
       [3, 3, 3, 3, 3, 3],
       [3, 3, 3, 3, 3, 3],
       [3, 3, 4, 3, 5, 3],
       [3, 4, 4, 3, 4, 4],
       [4, 4, 4, 4, 5, 3],
       [3, 4, 5, 3, 5, 4],
       [4, 3, 4, 4, 5, 4],
       [4, 4, 3, 4, 4, 4],
       [3, 3, 3, 3, 5, 3],
       [3, 3, 2, 3, 3, 3],
       [5, 5, 5, 5, 5, 5],
       [4, 4, 4, 4, 4, 4],
       [3, 2, 3, 3, 3, 4],
       [3, 3, 3, 3, 3, 3],
       [5, 5, 5, 5, 5, 5],
```

FIG 3. SPLIT VARIABLE

In Figure 3 above, variable splitting is carried out, where the predictor variable is saved into a certain_column variable then converted to a numpy array which is saved to the arr variable, while the target variable is then converted to a numpy array then the results are saved into the arrResults variable. After splitting the variables, then split the data on the dataset. This stage involves partitioning the dataset into separate categories of training data and testing data, allowing the model to be assessed. The process of dividing the data can be visualized in Figure 4 presented.

```
X_train, X_test, y_train, y_test =
train_test_split(arr, arrResult, test_size=0.1,
random_state=0)
```

FIG 4. SPLITTING DATA

In Figure 4 above is a snippet of the script for splitting data. Meanwhile, the Target Variable Data is also separated into variables y_train and y_test. The percentage of data splitting is 90% for training data (X_train, y_train) and 10% for testing data (X_test, y_test).

3. Experiment

After the initial data processing stage, experiments were then carried out on testing data and training data using the Naïve Bayes and C4.5 algorithms. In this experiment, 70% Training Data was carried out, 30% Testing Data, then 80% Training Data, 20% Testing Data and also 90% Training Data, 10% Testing Data.

4. Testing Model

In order to validate the model, the Python programming language is employed along with its associated library. The evaluation is performed iteratively on the available dataset multiple times to yield optimal outcomes and demonstrate the suitability of the applied approach. After splitting the data, model training is carried out. The following is for training the model using the Python programming language using the Sklearn library as in Figure 5 below:

```
model = DecisionTreeClassifier(max_depth=4)
model.fit(X_train, y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=4)
```

FIG 5. CREATING A C45 MODEL WITH THE SKLEARN LIBRARY

In Figure 5 above is a script for creating a C45 model using the Sklearn Library. In the DecisionTreeClassifier function there is a max_depth function whose function helps prevent overfitting. In this research, the max_depth value is 4. The training model utilizes a fit function that accepts X_train and y_train parameters. On the other hand, the Naïve Bayes model is depicted in Figure 6.

```
### Model Naive Bayes
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score

model = GaussianNB()
model.fit(X_train, y_train)
```

```
GaussianNB
GaussianNB()
```

FIG 6. MODEL WITH NAIVE BAYES ALGORITHM

In Fig 6 above is a script for creating a Naïve Bayes model using the Sklearn Library. The training model utilizes a fit function that accepts 2 input parameters, namely X_train and y_train.

5. Assessment and authentication of findings

In the final phase of the investigation, the assessment and authentication of empirical findings and prototype experimentation are executed. Conclusions can be inferred from the evaluation outcomes regarding the conducted investigation and trials. The proposed technique depicted in Figure 7 will be utilized to process the current dataset.

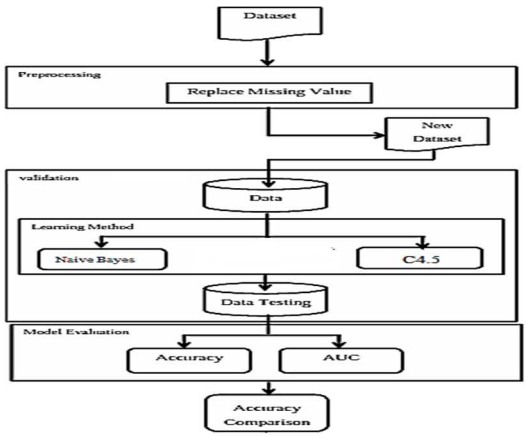

FIG 7. MODEL COMPARISON

In the model proposed in Figure 7 above, it is explained that this research is aimed at finding the best algorithm between Naïve Bayes and C45. The accuracy of the algorithm will be measured using a confusion matrix. Meanwhile, AUC will be measured using the ROC Curve. Thetest results with the highest accuracy are the methods that will be used to determine the classification of student satisfaction with administrative services.

IV. RESULT

A. *Accuracy*

Accuracy is a metric used to assess the satisfaction or dissatisfaction of a model in predicting a class or label. Accuracy is used to measure how well the C45 and Naïve Bayes algorithms predict the level of student satisfaction with administrative services at SMK Siti Banun. Accuracy is determined by evaluating the ratio of correct classifications made by the model to the overall number of test instances. A higher accuracy score signifies enhanced predictive capability of the model in assigning labels to data. The degree of precision in forecasting the degree of student satisfaction with administrative services by employing the C45 algorithm is depicted in Figure 8 provided.

	precision	recall	f1-score	support
0	1.00	0.75	0.86	8
1	0.71	1.00	0.83	5
accuracy			0.85	13
macro avg	0.86	0.88	0.85	13
weighted avg	0.89	0.85	0.85	13

FIG 8. PRECISION USING THE C4.5 ALGORITHM

From Figure 8 above, it can be seen that the accuracy level (f1-score) is 85%. For each prediction for the label Satisfied (1), the accuracy level is 86%, while for the prediction level for the label Not Satisfied (0), the accuracy level is 83%.

This data is obtained from the prediction results from testing data which is depicted in the confusion matrix below:

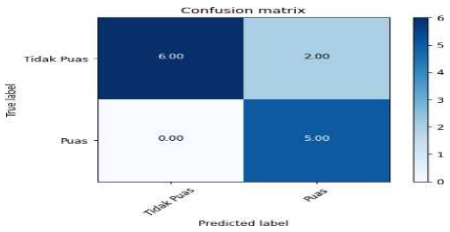

FIG 9. CONFUSION MATRIX C45 ALGORITHM

In Figure 9, the observed data that is both dissatisfied in reality and predicted dissatisfied by the model is 6, while the cases where the actual data is dissatisfied, but the model predicts satisfaction is 2. Meanwhile, the actual amount of data that is satisfied and predicted by the model is satisfied is 5 and vice versa is satisfied and predicted by the model is dissatisfied as much as 0. From the Confusion Matrix above obtained: TP = 6, TN = 5, FP = 2, FN = 0. From the Confusion Matrix above, the accuracy level is obtained:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

$$Acc = \frac{6 + 5}{6 + 5 + 2 + 0}$$

$$Acc = \frac{11}{13}$$

$$Acc = 0.846$$

Meanwhile, for the same thing, the level of accuracy in predicting the level of student satisfaction with administrative services using the Naïve Bayes algorithm is depicted in Figure 10 displayed underneath:

	precision	recall	f1-score	support
0	1.00	0.88	0.93	8
1	0.83	1.00	0.91	5
accuracy			0.92	13
macro avg	0.92	0.94	0.92	13
weighted avg	0.94	0.92	0.92	13

Fig 10. NAÏVE BAYES ALGORITHM ACCURACY

From Figure 10 above, it can be seen that the accuracy level (f1-score) is 92%. For each prediction for the label Satisfied (1), the accuracy level is 93%, while for the prediction level for the label Not Satisfied (0), the accuracy level is 91%. This data is obtained from the prediction results from testing data which is depicted in the confusion matrix below:

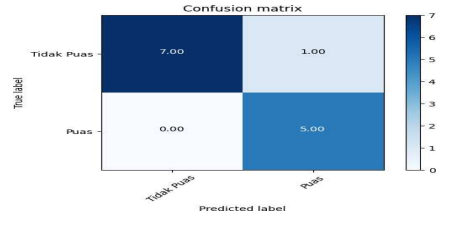

FIG 11. CONFUSION MATRIX NAÏVE BAYES ALGORITHM

In Figure 11 above, the observed count of dissatisfied data points that are correctly classified as dissatisfied by the model is 7, and the reverse scenario of dissatisfied data points being incorrectly classified as satisfied by the model is 1. Meanwhile, the actual amount of data that is satisfied and predicted by the model is satisfied is 5 and vice versa is satisfied and predicted by the model is dissatisfied as much as 0.

From the Confusion Matrix above obtained: TP = 7, TN = 5, FP = 1, FN = 0. From the Confusion Matrix above, the accuracy level is obtained:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

$$Acc = \frac{7 + 5}{7 + 5 + 1 + 0}$$

$$Acc = \frac{12}{13}$$

$$Acc = 0.923$$

B. *ROA AUC*

The AUC (Area Under the Curve) metric serves as an evaluation measure in classification tasks for assessing the model's capability to differentiate between positive and negative categories using different prediction thresholds. This is achieved by employing the ROC (Receiver Operating Characteristic) curve. The main function of AUC measures how well the model separates different classes. The area under the curve (AUC) for Algorithm C45 is greater than the accuracy value by a margin of 3%. The receiver operating characteristic (ROC) curve for Algorithm C45 is depicted in Figure 12 below.

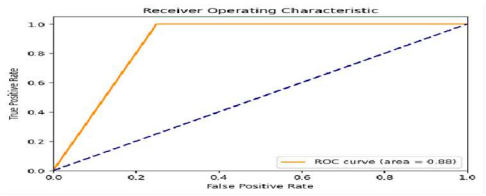

FIG 12. ROC CURVE FOR C45 ALGORITHM

The results depicted in Figure 12 demonstrate satisfactory performance with values exceeding 0.5, and the area under the curve (AUC) value is calculated to be 0.88 (88%). Conversely, the AUC value achieved using the Naïve Bayes Algorithm is 94.0%. Notably, this AUC value is 2% higher than the accuracy value. The ROC curve for the Naïve Bayes algorithm is visualized in Figure 13 below.

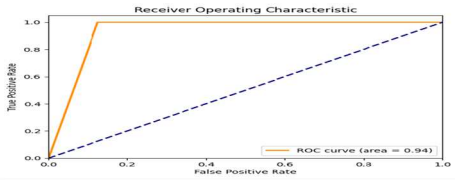

FIG 13. ROC CURVE IS PLOTTED FOR THE NAÏVE BAYES ALGORITHM

Figure 13 above shows good results where the value is above 0.5, and the AUC value is 0. (94%).

C. *Comparison of C45 and Naïve Bayes Models*

Table III displays the evaluation of each C45 and Naïve Bayes algorithm, showcasing the correlation between their accuracy value and AUC value.

TABLE III. RESULTS OF TESTING DATA AND TRAINING DATA

Algoritma	Data Testing (%)	Data Training (%)	Accuracy	AUC
C45	10	90	85%	88%
	20	80	62%	62%
	30	70	67%	66%
Naïve Bayes	10	90	92%	94%
	20	80	77%	77%
	30	70	74%	74%

The comparison outcomes of the utilized algorithms (C45 and Naive Bayes) in this study, based on the partitioning of testing data into training data as follows: 10% testing data and 90% training data, can be observed from Table III. The C4.5 algorithm demonstrated an accuracy value of 85%, while Naive Bayes achieved an accuracy of 92%. In the evaluation of the C4.5 algorithm and Naive Bayes, using a 20% subset of the data for testing purposes and an 80% subset for training, the accuracy rates obtained were 62% and 77%, respectively. In the interim, while contrasting the evaluation data consisting of 30% and the training data encompassing 70%, the precision metric of the C4.5 algorithm is 67% while the Naive Bayes algorithm achieves a precision of 74%. From the comparison of splitting data for each algorithm, the best composition is 10% for testing data and 30% for training data, and for the best accuracy, the Naïve Bayes algorithm is still used.

In the interim, the Area Under Curve (AUC) metric obtained from the C4.5 algorithm, using 10% of the data for testing and 90% for training, yielded a value of 88%. On the other hand, the Naive Bayes algorithm achieved an AUC value of 94% under the same experimental conditions. When evaluating the performance of the C4.5 algorithm and Naive Bayes, it was observed that when using a 20% testing data and 80% training data split, the C4.5 algorithm achieved an Area Under the Curve (AUC) value of 62%, while Naive Bayes achieved a higher AUC value of 77%. Conversely, when using a 30% testing data and 70% training data split, the C4.5 algorithm obtained a higher AUC value of 66%, whereas Naive Bayes yielded a slightly lower AUC value of 74%. Based on the evaluation of data partitioning for each algorithm, the optimal distribution is to allocate 10% of the data for testing purposes and 30% for training purposes. Additionally, the Naïve Bayes algorithm continues to yield the highest level of accuracy. The summary of the overall experiment can be observed in Table IV provided below.

TABLE IV. AVERAGE ALGORITHM COMPARISON RESULTS

Algorithm	Data Testing (%)	Data Training (%)	Accuracy	AUC
C45	10	90	85%	88%
	20	80	62%	62%
	30	70	67%	66%
	Average		71%	72%
Naïve Bayes	10	90	92%	94%
	20	80	77%	77%
	30	70	74%	74%
	Average		81%	82%

Based on the data presented in Table IV, it is evident that the mean precision of the C4.5 algorithm is 71%, while the Naive Bayes precision stands at 81%. Meanwhile for the average Area Unde Curve (AUC) value, the C45 algorithm has a value of 72%, while Naive Bayes has a value of 82%. From the data above, the accuracy and AUC values of the Naïve Bayes algorithm show the best results.

V. CONCOLUTION

From the comparison results of the C45 and Naive Bayes algorithms from experiments with testing data division: training data 10: 90, 20: 80, 30: 70. In all experiments involving the sharing of testing data, the accuracy values obtained using the Naïve Bayes classification algorithm outperformed those obtained using the C45 and Naïve Bayes algorithms, making it the most accurate algorithm. In the interim, the assessment employs the receiver operating characteristic (ROC) curve, which relies on the area under the curve (AUC) value. The naive Bayes algorithm consistently achieves the highest performance in all experiments, utilizing varying proportions of training and testing data (90:10, 80:20, and 70:30) with the same AUC value, representing the maximal result. Based on the comprehensive outcomes of model experimentation, it can be inferred that Naive Bayes exhibits the most optimal performance, as manifested by superior accuracy and AUC values.

REFERENCES

[1] B. P. Statistik, "Badan Pusat Statistik SMK," 2023. .

[2] N. A. Sinaga, B. H. Hayadi, and Z. Situmorang, "Perbandingan akurasi algoritma naïve bayes, k-nn dan svm dalam memprediksi penerimaan pegawai," *Tekinkom*, vol. 5, no. 1, pp. 27–34, 2022, doi: 10.37600/tekinkom.v5i1.446.

[3] N. Umar and M. Adnan Nur, "Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 4, pp. 585–590, 2022, doi: 10.29207/resti.v6i4.4179.

[4] H. Hairani, K. E. Saputro, and S. Fadli, "K-means- SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93,2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[5] Y. Narayan, "Comparative analysis of SVM and Naive Bayes classifier for the SEMG signal classification," *Mater. Today Proc.*, vol. 37, no. Part 2, pp. 3241–3245, 2020, doi: 10.1016/j.matpr.2020.09.093.

[6] A. Septiarini, Rizqi Saputra, Andi Tejawati, and Masna Wati, "Deteksi Sarung Samarinda Menggunakan Metode Naive Bayes Berbasis Pengolahan Citra," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 5, pp. 927–935, 2021, doi: 10.29207/resti.v5i5.3435.

[7] Yoga Religia and A. Amali, "Perbandingan Optimasi

Feature Selection pada Naïve Bayes untuk Klasifikasi Kepuasan Airline Passenger," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 527–533, 2021, doi: 10.29207/resti.v5i3.3086.

[8] F. Fersellia, E. Utami, and A. Yaqin, "Sentiment Analysis of Shopee Food Application User Satisfaction Using the C4.5 Decision Tree Method," *Sinkron*, vol. 8, no. 3, pp. 1554–1563, 2023, doi: 10.33395/sinkron.v8i3.12531.

[9] M. D. Muafa, "Pengembangan Aplikasi Berbasis Web dengan Rshiny untuk Data Klasifikasi Menggunakan Metode Naive Bayes," *Automata*, vol. 3, no. 1, p. 8, 2022, [Online]. Available: https://journal.uii.ac.id/AUTOMATA/article/view/21875.

[10] W. R. Sari Oktapia Ningse, S. Sumarno, and Z. M. Nasution, "C4.5 Algorithm Classification for Determining Smart Indonesia Program Recipients at MIS Al-Khoirot," *JOMLAI J. Mach. Learn. Artif. Intell.*, vol. 1, no. 1, pp. 65–76, 2022, doi: 10.55123/jomlai.v1i1.165.

[11] N. A. Sinaga, T. S. Gunawan, and Wanayumini, "Sentiment Analysis on Hotel Ratings Using Dynamic Convolution Neural Network," *2nd Int. Conf. Inf. Sci. Technol. Innov.*, pp. 1–6, 2023.

[12] N. Rochmawati *et al.*, "Covid Symptom Severity Using Decision Tree," *Proceeding - 2020 3rd Int. Conf. Vocat. Educ. Electr. Eng. Strength. Framew. Soc. 5.0 through Innov. Educ. Electr. Eng. Informatics Eng. ICVEE 2020*, 2020, doi: 10.1109/ICVEE50212.2020.9243246.

[13] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.

[14] W. D. Kusuma, Suhada, and Saifullah, "Klasifikasi Kepuasan Siswa Terhadap Kinerja Guru SMK Satrya Budi 2 Perdagangan Menggunakan Algoritma C4.5," *TIN Terap. Inform. Nusant.*, vol. 2, no. 7, pp. 460– 465, 2021, [Online]. Available: https://ejurnal.seminar-id.com/index.php/tin.

[15] D. Alita, S. Setiawansyah, and A. D. Putra, "C45 Algorithm for Motorcycle Sales Prediction On CV Mokas Rawajitu," *J. Sisfotek Glob.*, vol. 11, no. 2, p. 127, 2021, doi: 10.38101/sisfotek.v11i2.392.