ISSN: 2581-1274

Comparison of K-Nearest Neighbors Method and Naïve Bayes Method in Classifying the Quality of Oil Palm Seed Varieties

Suhaiba Nasyira Hariono [a,1], Nurdin [b,2,*], Lidya Rosnita [c,3]

[a,c] *Departement of Informatics, Universitas Malikussaleh, Aceh, Indonesia*
[b] *Departement of Information Technology, Universitas Malikussaleh, Aceh, Indonesia*
[1] *suhaiba.210170253@mhs.unimal.ac.id;* [2] *nurdin@unimal.ac.id*;* [3] *lidyarosnita@unimal.ac.id*
** corresponding author*

ARTICLE INFO

Article history:
Published
December 12, 2025

Keywords:
Oil Palm Seeds
Classification
Quality
K-Nearest Neighbors
Naïve Bayes

ABSTRACT

Rapid advances in technology and science have driven significant transformations in various sectors, including the palm oil industry. As one of the nation's strategic commodities, the success of the palm oil industry is greatly influenced by the quality of the seeds used. The selection of high-quality seeds from the initial nursery stage pre-nursery to the main nursery stage is a critical factor in supporting productivity, harvest quality, and resistance to pests and diseases. However, the seed selection process often faces challenges such as genetic variation among individuals and differences in adaptive capacity to the environment, leading to inconsistent growth performance. This study aims to classify the quality of oil palm seedlings using the K-Nearest Neighbors and Naïve Bayes classification algorithms based on seedling growth observation data at PTPN IV Pabatu Plantation, Serdang Bedagai District. The criteria used in the classification process include seedling age, seedling height, stem diameter, number of leaves, leaf length, and pest infestation per seedling. Based on the test results, the K-Nearest Neighbors method achieved an accuracy of 94%, precision of 89.47%, and recall of 94.44%. Meanwhile, the Naïve Bayes method achieved an accuracy of 82%, precision of 92.85%, and recall of 78.78%. These results indicate that the K-Nearest Neighbors algorithm performs better in classifying the quality of oil palm seedlings compared to the Naïve Bayes algorithm. Thus, this data mining-based classification approach can serve as a strategic solution to enhance the accuracy of seedling selection in an objective and efficient manner.

I. Introduction

Rapid advances in technology and science have had a significant impact on various sectors, including agriculture, particularly palm oil [1]. Palm oil is a strategic commodity that plays a major role in the national economy, with Indonesia having been the world's largest producer since 2006. Data from the Central Statistics Agency (BPS) shows that palm oil production from large plantations increased from 19,912 tons in 2016 to 30,996 tons in 2024 [2]. This increase demonstrates the importance of the palm oil sector in supporting the country's economy. The success of the palm oil industry is largely determined by the quality of the seeds used. High-quality seeds are a key factor in achieving high productivity, good harvest quality, and resistance to pests and diseases. Conversely, the use of unsuitable seeds can reduce crop yields and the quality of fresh fruit bunches. PTPN IV Pabatu, as a state-owned enterprise, plays a strategic role in oil palm seedling production through the pre-nursery (1–3 months) and main nursery (4–12 months) stages. The quality of seeds at these two stages greatly determines the success of plant growth in the field [3]. However, the seed selection process is not easy due to genetic variations between individuals and differences in their ability to adapt to the environment. This results in varying seed growth performance [4]. Therefore, a technological approach is needed to improve the accuracy and efficiency of the selection process. One

solution that can be applied is the use of data mining techniques, particularly the K-Nearest Neighbors and Naïve Bayes classification algorithms, which are capable of providing more objective assessments than conventional methods. Previous studies have demonstrated the effectiveness of both algorithms in various classification cases. [5] compared the K-Nearest Neighbors and Naïve Bayes methods in classifying potential blood donors, with K-Nearest Neighbors achieving 86% accuracy and Naïve Bayes achieving 76%. [6] applied the Naïve Bayes Classifier to classify palm oil quality and obtained an accuracy of 64.25%. Meanwhile, [7] related research predicting low-income communities eligible for government assistance using the K-Nearest Neighbors method yielded an accuracy rate of 86.02%. Based on this background, this study aims to compare the effectiveness of the K-Nearest Neighbors and Naïve Bayes methods in classifying the quality of oil palm seedlings at PTPN IV Pabatu. The results of this study are expected to support a more accurate and efficient seedling selection process, thereby contributing to increased productivity and sustainability of the oil palm industry in Indonesia.

II. Method

The research method is a series of steps taken by researchers in collecting, analyzing, and interpreting data to answer research questions or solve existing problems. Figure 1 shows a research process diagram that illustrates each part and stage of the overall research procedure. This illustration provides a comprehensive understanding of the research design and makes it easier to see the relationships between interrelated components in order to achieve the research objectives.

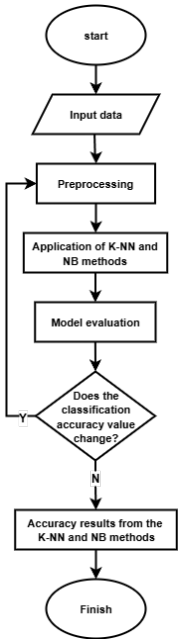

Fig. 1. Flowchart of the System Work Process Using K-NN and NB Method

A. Data Collecting

In the initial stage of the research, the author collected a dataset relevant to the topic. The dataset used in this study is primary data obtained directly from the PTPN IV Pabatu Oil Palm Nursery. The data contained information on the number of seedlings in March 2025 that was relevant to the research objectives. The dataset consisted of six variables, namely age, plant height, stem diameter, number of leaves, leaf length, and number of plants affected by pests per seedling. The data was then classified into two categories, namely superior seedlings and non-superior seedlings, with a total of 170 samples.

B. Preprocessing

Data preprocessing is the initial stage in data analysis that involves cleaning, transforming, and preparing raw data so that it is ready for use in data mining, machine learning, and statistical analysis. This stage aims to improve data quality by reducing noise and irrelevant data, as well as adjusting the data format to suit the needs of the modeling algorithm, thereby producing more accurate and efficient analysis. The preprocessing stages discussed include [8]:

1. Data Cleaning

Data Cleaning at this stage, missing values are identified and handled through imputation (e.g., using the mean, median, or other methods) or deletion of incomplete data to prevent bias in modeling. Duplicate removal is also important to maintain the integrity of the dataset so that the model is not trained on the same data repeatedly. In addition, inconsistencies in data formats, such as differences in date formats or inconsistent categorical values, are corrected so that the k-nearest neighbors and naïve bayes algorithms can read and process the data correctly without errors.

2. Data Transformation

Data transformation aims to prepare data in a scale and format suitable for algorithms. In the k-nearest neighbors' method, normalization or standardization of numerical data is very important because this algorithm uses the distance between data to determine the nearest neighbor; without a uniform scale, features with larger values can dominate the distance calculation. Meanwhile, for naïve bayes, categorical features must be converted using encoding techniques such as one-hot encoding or label encoding so that they can be processed probabilistically. If the data is too granular, aggregation such as summarizing daily data into monthly averages can also be done to reduce noise and data complexity.

3. Data Distribution

The dataset is then divided into training, validation, and testing subsets. This division allows for objective evaluation of model performance, where training data is used to build the model, validation data is used for parameter setting (e.g., selecting k in k-nearest neighbors), and testing data is used to measure the final performance of the model on data it has never seen before. Proper data division is crucial so that k-nearest neighbors and naïve bayes models do not experience overfitting and the prediction results are reliable.

C. K-Nearest Neighbors Method

K-nearest neighbors is a method used to classify objects based on the training data closest to the object [9]. K-Nearest Neighbors is a method in supervised learning that determines the classification of new data based on existing data. This algorithm works by grouping objects into specific categories based on the training data closest to the object [10]. K-Nearest Neighbors uses an approach in which new data is classified into the group most commonly found among its closest neighbors. The class that appears most frequently in that environment will be the classification result [11]. The K-Nearest Neighbors method in classification has shown good performance according to various previous studies. The success of this method is achieved through structured and systematic stages, as described in the literature. The K-Nearest Neighbors calculation process can be seen in Figure 2.

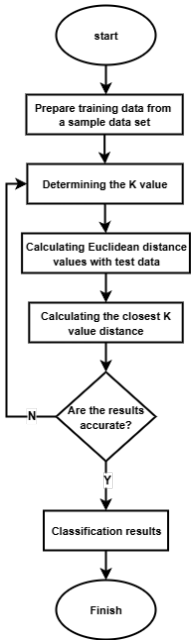

Fig. 2. Flowchart of Oil Palm Seedling Quality Classification Using the K-NN Method

The steps for calculating the K-Nearest Neighbors algorithm include:

1. Determining the value of K, which is the number of nearest neighbors used as a reference in the classification process [12].
2. Calculate the distance between the new data (to be predicted) and all data in the dataset. In this study, the metric used for measurement is theEuclidean distance (query instance) of each object to the given sample data.

$$d(x,y)= \sqrt{\sum_{i=1}^{n} (x_i-y_i)^2} \quad (1)$$

Description:
x_i : Data sampel
y_i : Data testing
i : Data Variables/Criteria
d(x,y) : Distance
n : Data Dimensions

3. Identify the nearest neighbors by selecting the data points that are closest to the new data point. These values will be used to determine the majority class as the prediction result.

4. Determine the category of the new data by calculating the frequency of occurrence of each class from the nearest neighbors. The class that appears most frequently is selected as the classification.

D. Naïve Bayes Method

Naïve Bayes is a method for clustering or classification based on Bayes' theorem [13]. This approach uses the concepts of probability and statistics developed by Thomas Bayes, a British scientist. Naïve Bayes uses probability theory to determine the highest probability of a classification by analyzing the frequency of each classification in the training data [14]. The class label with the highest probability will be selected as the classification result for the [15]. For continuous data classification, the Gaussian probability density function is used to calculate the probability of each feature value [16]. The Naive Bayes calculation process is illustrated in Figure 3.

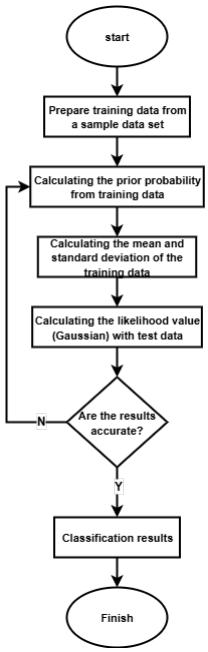

Fig. 3. Flowchart of Oil Palm Seedling Quality Classification Using the NB Method

The steps for calculating the Naïve Bayes algorithm include:

1. Calculating Prior Probability is the initial probability that a piece of data belongs to a particular class [17].

$$P(C) = \frac{Jumlah\ class\ yang\ muncul}{Total\ jumlah\ data} \quad (2)$$

2. Calculate the mean (μ) and standard deviation (σ) for each feature in each class.
   a. Calculating the mean (μ) is the average value of a feature.

$$\mu = \frac{\sum X}{N} \quad (3)$$

Description:
μ: average of features in class
X: number of feature data values X
N: number of data in class C

   b. Calculate the standard deviation (σ) obtained by taking the square root of the variance value, where the variance itself is the average of the squares of the variance values, wherethe variance itself is the average of the squares of the differences between each data value and the overall average.

$$\sigma = \sqrt{\frac{\Sigma(X-(\mu))^2}{N}} \quad (4)$$

Explanation:
σ: standard deviation
X: feature values in class C
μ: average of these features
N: Number of data in class C

   c. Calculating likelihood using a Gaussian (normal) distribution.

$$P(X|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right) \quad (5)$$

Explanation:
P($X|C$) : likelihood used to determine how likely a data value x is to appear from class C.
$\frac{1}{\sqrt{2\pi\sigma^2}}$: part of the Gaussian distribution that functions as a normalization factor.
π : mathematical constant = 3.14
σ : standard deviation of features in class C.
μ : mean of the features in class C.
$(x-\mu)^2$: the square of the difference between the feature value x and the mean μ.
$\frac{(x-\mu)^2}{2\sigma^2}$: measures the distance of the value x from the mean in units of variance to determin the probability of occurrence based on the normal distribution.

E. Model Evaluation

Confusion Matrix is a method often used to calculate accuracy in data mining. This method is presented in the form of a table that describes the amount of test data that has been classified correctly and incorrectly [18]. Confusion Matrix is arranged by displaying the predicted class at the top of the table and the observed class on the left side. This table compares the model's predictions with the actual data to identify whether the model made correct or incorrect predictions [19].The formula for calculating the Confusion Matrix is written as follows: Precision is useful for measuring the accuracy between the information requested by the user and the answer provided by the system.

$$Precision = \frac{TP}{(TP+FP)} \quad (6)$$

Recall is useful for measuring the success rate of a system in retrieving information in an equation.

$$Recall = \frac{TP}{(TP+FN)} \quad (7)$$

Accuracy is useful for measuring the performance of a method.

$$Acuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (8)$$

III. Results and Discussion

In this study, the author will test the K-Nearest Neighbors and Naïve Bayes methods to predict the results of comparing the two methods in classifying the quality of 10 palm oil seedling varieties at ages ranging from 1 to 12 months (pre-nursery to main nursery) using data from observations of seedling growth in March 2025 at PTPN IV Pabatu. Both algorithms are applied to classify 120 training data with 50 test data of seedling varieties based on certain predefined variables. This study shows the classification results of each method in determining the quality of seedling varieties.

Table 1. Seeds Criteria

Kode Kriteria	Kriteria
X1	Age (months)
X2	Plant Height (cm)
X3	Stem Diameter (cm)
X4	Number of Leaves
X5	Leaf Length (cm)
X6	Attacked by pests (per plant)

Table 2. Data Training

No	Varietas	X1	X2	X3	X4	X5	X6	Class
1	Ppks 50 ng	1	18	0,5	1	12	5	Superior
2	Ppks 50 ng	2	19	1,3	3	13	7	Superior
3	Ppks 50 ng	3	20	1,5	4	14	6	Superior
4	Ppks 50 ng	4	25	1,8	5	15,2	4	Superior
5	Ppks 50 ng	5	27	2,0	6	16,5	5	Superior
…	…	…	…	…	…	…	…	…
115	Socfin MTG	7	45	4,7	11	19	8	Superior
116	Socfin MTG	8	60	6,0	12	20,5	9	Superior
117	Socfin MTG	9	74	8,0	14	22,7	7	Superior
118	Socfin MTG	10	90	10,0	15	24,5	10	Superior
119	Socfin MTG	11	108	12,0	17	26,5	13	Non-Superior
120	Socfin mtg	12	140	16,0	25	29	10	Superior

The above data is training data consisting of 120 data points that will be classified with all 50 test data points. The test data to be classified is as follows:

Table 3. Data Testing

No	Varietas	X1	X2	X3	X4	X5	X6	Prediction
1	Ppks 50 ng	1	18	0,5	1	11	6	Superior
2	Ppks 50 ng	3	20	1,4	3	13	8	Superior
3	Ppks 50 ng	4	25	1,7	5	15,5	5	Superior

Table 3. Data Testing

No	Varietas	X1	X2	X3	X4	X5	X6	Prediction
4	Ppks 50 ng	9	70	6,0	14	20,8	13	Non-Superior
5	Ppks 50 ng	12	139	10,0	19	30	10	Superior
…	…	…	…	…	…	…	…	…
45	PPKS SP	11	90	6,0	15	22,5	15	Non-Superior
46	Socfin MTG	1	10	0,3	3	12	3	Superior
47	Socfin MTG	3	25	0,9	5	14,5	5	Superior
48	Socfin MTG	10	90	10,0	15	24,5	7	Superior
49	Socfin MTG	11	108	12,0	17	26,5	11	Non-Superior
50	Socfin mtg	12	140	16,0	25	30	10	Superior

A. K-Nearest Neighbors Method Calculation

Table 3 above contains test data that will be classified in the training data in Table 2, To obtain accurate classification results, this process begins with systematic data normalization. Normalization is performed using the min-max normalization method, which aims to equalize the scale between variables before the classification stage is carried out. The normalization results for 120 training data are as follows:

Table 4. Results of Training Data Normalization Calculations

No	X1	X2	X3	X4	X5	X6
1	0	0,061	0,018	0	0,052	0,090
2	0,090	0,069	0,069	0,083	0,105	0,272
3	0,181	0,076	0,082	0,125	0,157	0,181
4	0,272	0,115	0,101	0,166	0,221	0
5	0,363	0,130	0,113	0,208	0,289	0,090
6	0,454	0,223	0,145	0,291	0,357	0,181
…	…	…	…	…	…	…
115	0,545	0,269	0,284	0,416	0,4216	0,363
116	0,636	0,384	0,367	0,458	0,5	0,454
117	0,727	0,492	0,493	0,541	0,615	0,272
118	0,818	0,615	0,6202	0,583	0,710	0,545
119	0,909	0,753	0,746	0,666	0,815	0,818
120	1	1	1	1	0,947	0,545

Table 4 above presents the results of the normalization process for 120 training data, which was then also applied to 50 test data. The normalization process was carried out using the min-max normalization method with the aim of standardizing the scale for each variable so that there were no differences in the value ranges between attributes before entering the classification stage. The results of the normalization for 50 test data are shown as follows:

Table 5. Results of Testing Data Normalization Calculations

No	X1	X2	X3	X4	X5	X6
1	0	0,061	0,018	0	0	0,181
2	0,181	0,076	0,075	0,083	0,105	0,363
3	0,272	0,115	0,094	0,166	0,236	0,090
4	0,727	0,461	0,367	0,541	0,515	0,818
5	1	0,992	0,620	0,75	1	0,545
…	…	…	…	…	…	…

Table 5. Results of Testing Data Normalization Calculations

No	X1	X2	X3	X4	X5	X6
45	0,909	0,615	0,367	0,583	0,605	1
46	0	0	0,003	0,083	0,052	-0,090
47	0,181	0,115	0,041	0,166	0,184	0,090
48	0,818	0,615	0,620	0,583	0,710	0,272
49	0,909	0,753	0,746	0,666	0,815	0,636
50	1	1	1	1	1	0,545

After the normalization process on the test data is done, the next step is to calculate using the K-Nearest Neighbor (K-NN) method. The calculation process is as follows:

1. Determine the K parameter value. In solving this case, we use the parameter K=3.
2. Calculate the square of the Euclidean distance of each object to the available sample data. The results of the distance calculation are presented in the following table:

Table 6. Distance Calculation Results Data 1

No	Distance	Class
1	0,105	Superior
2	0,193	Superior
3	0,279	Superior
4	0,440	Non-Superior
5	0,530	Superior
…	…	…
115	0,891	Superior
116	1,079	Superior
117	1,273	Superior
118	1,521	Superior
119	1,830	Non-Superior
120	2,207	Superior

Table 7. Distance Calculation Results Data 2

No	Distance	Class
1	0,347	Superior
2	0,129	Superior
3	0,194	Superior
4	0,404	Non-Superior
5	0,402	Superior
…	…	…
115	0,650	Superior
116	0,831	Superior
117	1,509	Superior
118	1,280	Superior
119	1,577	Non-Superior
120	1,989	Superior

Table 8. Distance Calculation Results Data 3

No	Distance	Class
1	0,380	Superior
2	0,305	Superior
3	0,162	Superior

Table 8. Distance Calculation Results Data 3

No	Distance	Class
4	0,092	Non-Superior
5	0,116	Superior
…	…	…
115	0,552	Superior
116	0,751	Superior
117	0,908	Superior
118	1,195	Superior
119	1,533	Non-Superior
120	1,880	Superior

3. Next, the objects are sorted based on the smallest distance, then selected as many as K=3. This sorting process is done by arranging the data starting from the smallest distance to the largest distance, so that the three closest objects are obtained.

Table 9. Minimum Distance Calculation Results Data 1

No	Distance	Description	K
13	0,057	Superior	1
61	0,098	Superior	2
97	0,100	Superior	3

Table 10. Minimum Distance Calculation Results Data 2

No	Distance	Description	K
63	0,058	Superior	1
87	0,043	Superior	2
99	0,049	Superior	3

Table 11. Minimum Distance Calculation Results Data 3

No	Distance	Description	K
4	0,092	Superior	1
5	0,115	Superior	2
111	0,118	Superior	3

Using equation (1), the author calculated the distance value of test data 1 against 120 training data. The calculation results show that the smallest distance values were obtained in data 13, 62, 97, and so on. For the calculation of distances 4 to 50, the steps taken are the same as for the calculation of distance 1. That is, determine the value of k = 3, then calculate the distance using the Euclidean Distance matrix, sort the data based on the distance value, and select the data with the smallest distance value for each distance n, which corresponds to the number of k that has been determined in first step.

B. Naïve Bayes Method Calculation

Calculations were performed using the Naïve Bayes method, with the following calculation steps:

1. Calculating Prior Probability

In the training dataset, the data is classified into two classes, namely superior seeds and non-superior seeds. The probability (P) for each class can be calculated by dividing the number of data occurrences in each class by the total data available. The prior probability calculation is performed using formula (2), with the following results:

Table 12. Prior Probability of Training Data

Class	Value
Non-Superior	0,35
Superior	0,65

2. Calculating the Mean (μ) and Standard Deviation (σ) (μ)
   a. Non-Superior Seed Class
   The following is a training data table from the non-superior seed class with a total of 42 data points:

Table 13. Training Data for Non-Superior Seedlings

Class Label = Non-Superior						
No	X1	X2	X3	X4	X5	X6
1	9	70	6,5	14	20,8	11
2	11	103	9,5	18	25	14
3	5	25	2,0	4	15	13
…	…	…	…	…	…	…
40	12	110	7,0	17	24	14
41	6	39	3,3	10	17,7	11
42	11	108	12,0	17	26,5	13

The following are the mean (μ) and standard deviation (σ) values for each variable (X1, X2, X3, X4, X5, X6) in the non-superior seed class label, using the calculation equation in formula (3) for the mean and formula (4) for the standard deviation:

Table 14. Mean (μ) and Standard Deviation (σ) of Non-Superior Seed Class

Non- Superior Class						
	X1	X2	X3	X4	X5	X6
Mean (μ)	8,55	68,21	4,53	10,93	19,89	12,60
Standard Deviasi (σ)	2,62	32,28	2,64	5,33	3,53	1,50

   b. High-Quality Seed Class
   The following is a training data table from the high-quality seed class with a total of 78 data points:

Table 15. Training Data Superior Seedlings

Class Label = Superior						
No	X1	X2	X3	X4	X5	X6
1	1	18	0,5	1	12	5
2	2	19	1,3	3	13	7
3	3	20	1,5	4	14	6
…	…	…	…	…	…	…
76	9	74	8,0	14	22,7	7
77	10	90	10,0	15	24,5	10
78	12	140	16,0	25	29	10

The following are the mean (μ) and standard deviation (σ) values for each variable (X1, X2, X3, X4, X5, X6) in the superior class label, using the calculation equation in formula (3) for the mean and formula (4) for the standard deviation:

Table 16. Mean (μ) and Standard Deviation (σ) of Superior Seed Class

Superior Class						
	X1	X2	X3	X4	X5	X6

Mean (μ)	5,40	42,99	2,75	6,59	16,44	8,21
Standard Deviasi (σ)	3,37	31,40	2,82	5,33	4,32	1,65

3. Calculating Likelihood Using Gaussian Distribution

After obtaining the mean (μ) and standard deviation (σ) values for each class the next step is to calculate the Likelihood value using Gaussian Distribution through formula (5) with test data in Table 3.

C. Model Evaluation

Based on manual calculations in Excel, classification results were obtained for each method for two classes, namely non-superior seeds and superior seeds. The results of the K-Nearest Neighbors (K-NN) and Naïve Bayes (NB) methods were then evaluated using a confusion matrix model, as shown in the following table:

Table 17. Comparison of K-Nearest Neighbors and Naïve Bayes Methods

No	Varieties	Class	*K-Nearest Neighbors Result*	*Naïve Bayes Result*
1	Ppks 50 NG	Superior	Superior	Superior
2	Ppks 50 NG	Superior	Superior	Superior
3	Ppks 50 NG	Superior	Superior	Superior
4	Ppks 50 NG	Non-Superior	Non-Superior	Non-Superior
5	Ppks 50 NG	Superior	Superior	Non-Superior
…	…	…	…	…
45	Ppks sungai pancur	Non-Superior	Non-Superior	Non-Superior
46	Socfin MTG	Superior	Superior	Superior
47	Socfin MTG	Superior	Superior	Superior
48	Socfin MTG	Superior	Superior	Superior
49	Socfin MTG	Non-Superior	Non-Superior	Non-Superior
50	Socfin MTG	Superior	Superior	Non-Superior

Based on the results of the K-Nearest Neighbors (K-NN) method calculations obtained in the table above, the Confusion Matrix can be calculated as follows:

Table 18. Confusion Matrix K-Nearest Neighbors Method

	Prediction	
Actual	Non-Superior Seeds	Superior Seeds
Non-Superior Seeds	17	1
Superior Seeds	2	30

Based on the results of the Confusion Matrix analysis using the K-Nearest Neighbors method, the accuracy value obtained was 94%, precision was 89.47%, and recall was 94.44%. Then, for the results of the Naïve Bayes method calculation obtained in Table 4.17 above, the Confusion Matrix can be calculated as follows:

Table 19. Confusion Matrix Naïve Bayes Method

	Prediction	
Actual	Non-Superior Seeds	Superior Seeds
Non-Superior Seeds	15	2
Superior Seeds	7	26

Based on the results of the Confusion Matrix calculation using the Naive Bayes method, the accuracy value obtained was 82%, precision was 92.85%, and recall was 78.78%.

Thus, the accuracy, precision, and recall values obtained from the comparison between the K-Nearest Neighbors (K-NN) and Naïve Bayes (NB) methods in classifying the quality of oil palm seed varieties can be seen in the following figure:

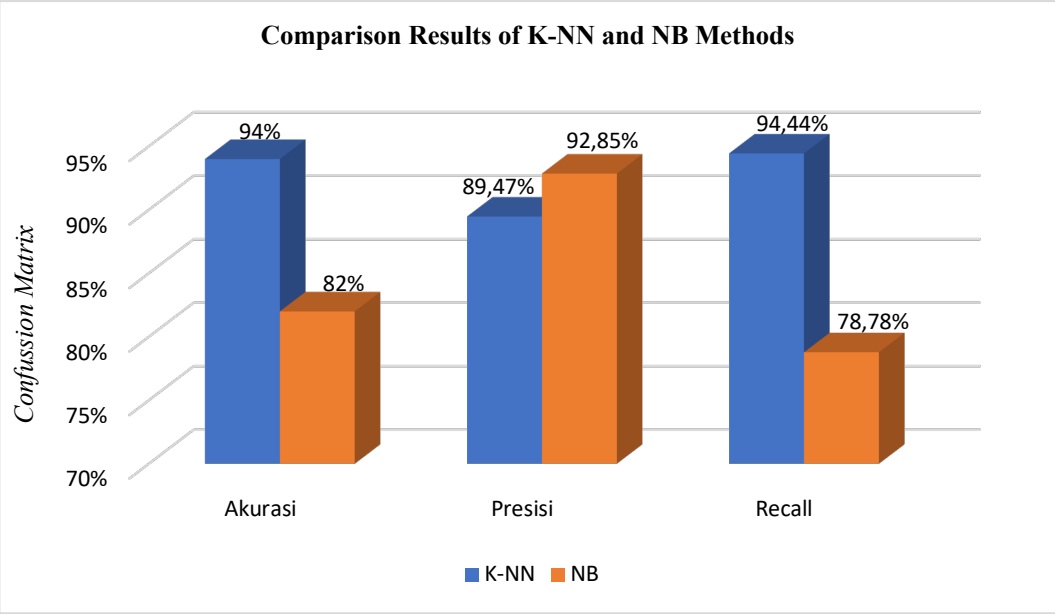

Fig. 4. Chart of Evaluation Results Comparing the K-Nearest Neighbors and Naïve Bayes Methods Using a Confusion Matrix

IV. Conclusion

Based on the results of the research that has been conducted, it can be concluded that this study has successfully developed a web-based oil palm seedling variety quality classification system that utilizes six assessment criteria, namely seedling age (months), plant height (cm), stem diameter (cm), number of leaves, leaf length (cm), and number of pest attacks per seedling. This system was designed by implementing two classification algorithm methods, namely K-Nearest Neighbors and Naïve Bayes. The results of the classification process using these two algorithms were used to determine the quality of the seedling varieties, whether they were classified as Superior or Non-Superior from the ten varieties observed. Based on the evaluation of the system's performance using the Confusion Matrix model, it was found that the K-Nearest Neighbors method produced an accuracy rate of 94%, a precision rate of 92.85%, and a recall rate of 78.78%. Meanwhile, the Naïve Bayes method showed an accuracy rate of 82%, a precision rate of 89.47%, and a recall rate of 94.44%. From the comparison of the two methods, it can be concluded that the K-Nearest Neighbors method has better performance in classifying seed variety quality than the Naïve Bayes method, in terms of the accuracy obtained.

References

[1] Sunarko, *Budi Daya dan Pengelolaan Kebun Kelapa Sawit dengan Sistem Kemitraan (Partnership System in Managing Palm Oil Plantation)*. AgroMedia Pustaka, 2009.

[2] Badan Pusat Statistik, "Produksi Perkebunan Besar Tahunan Menurut Jenis Tanaman (Ribu Ton)," *BADAN PUSAT STATISTIK*, 2024. [Online]. Available: https://www.bps.go.id/id/ statistics-table/2/OTQjMg==/produksi-perkebunan-besar-tahunan-menurut-jenis-tanaman--ribu-ton-.html.

[3] S. Thakur, S. Ratnam, and A. Singh, "Perkebunan Kelapa Sawit Indonesia Dalam Perspektif Pembangunan Berkelanjutan," *Palm Oil Agribus. Strateg. Policy Inst.*, vol. 4, no. 12, pp. 50–55, 2024.

[4] E. Hambali and M. Rivai, "The Potential of Palm Oil Waste Biomass in Indonesia in 2020 and 2030," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 65, no. 1, 2017.

[5] N. Nur, Asmawati, and N. Syahra, "Perbandingan Metode k-NN dan Naïve Bayes dalam Klasifikasi Penentuan Calon Pendonor Darah," *J. Comput. Inf. Syst. (J-CIS)*, vol. 1, no. 1, pp. 21–28, Aug. 2021.

[6] N. Puspitasari, R. Rosmasari, F. W. Pratama, and H. Sulastri, "Quality Classification of Palm Oil Varieties Using Naive Bayes Classifier," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 13, no. 1, pp. 11–23, May 2022.

[7] N. Nurdin, "Analisa Data Mining Dalam Memprediksi Masyarakat Kurang Mampu Menggunakan Metode K-Nearest Neighbor," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 2, pp. 1090–1098, 2024.

[8] Zunan Setiawan, Muhammad Fajar, and Arif Mudi Priyatno, *Buku Ajar Data Mining*. Jakarta, Indonesia: SonPedia Publishing, 2023.

[9] A. Muzakir, A. Desiani, and A. Amran, "Klasifikasi Penyakit Kanker Prostat Menggunakan Algoritma Naïve Bayes dan K-Nearest Neighbor," *Komputika J. Sist. Komput.*, vol. 12, no. 1, pp. 73–79, 2023.

[10] A. Garavand, C. Salehnasab, A. Behmanesh, and Aslani, "Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms," *J. Healthc. Eng.*, vol. 22, no. 1, p. 9, 2022.

[11] S. Saumina, Y. Afrillia, and N. Nurdin, "Klasifikasi Usaha Mikro Kecil dan Menengah Menggunakan Metode K-Nearest Neighbors di Kota Lhokseumawe," Jurnal Sistem Informasi Kaputama (JSIK), vol. 9, no. 2, 2025.

[12] A. A. Fikhri, and N. Nurdin, "Implementasi Algoritma K-Nearest Neighbor Pada Sistem Pemantau Suhu Dan Kelembapan Ruang Server Menggunakan Protokol Mqtt Berbasis Iot," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, 2024.

[13] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Eng. Appl. Artif. Intell.*, vol. 136, Oct. 2024.

[14] N. Nurdin, S. Abadi, and Y. Afrillia "Classification System for Soil Types Suitable for Food Crops using Naïve Bayes Method," Sist. J. Sist. Inf., vol. 13, no. 3, hal. 1102–1113, 2025.

[15] Moch. Rizky Yuliansyah, M. B, and A. Franz, "Perbandingan Metode K-Nearest Neighbors dan Naïve Bayes Classifier Pada Klasifikasi Status Gizi Balita di Puskesmas Muara Jawa Kota Samarinda," *Adopsi Teknol. dan Sist. Inf.*, vol. 1, no. 1, pp. 08–20, 2022.

[16] M. Qamal, I. Sahputra, N. Nurdin, M. Maryana, and M. Mukarramah, Sistem Pendukung Keputusan Penentuan Penerimaan Bantuan PKH Menggunakan Metode Naïve Bayes, TECHSI: J. Tek. Inform., vol. 14, no. 1, p. 21, 2023, doi: 10.29103/techsi.v14i1.6960.

[17] Nurdin and D. Astika, "Penerapan Data Mining Untuk Menganalisis Penjualan Barang Dengan Menggunakan Metode Apriori Pada Supermarket Sejahtera Lhokseumawe," *TECHSI - J. Tek. Inform.*, vol. 7, no. 1, pp. 132–155, 2019.

[18] Isman, Andani Ahmad, and Abdul Latief, "Perbandingan Metode KNN Dan LBPH Pada Klasifikasi Daun Herbal," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 557–564, 2021.

[19] B. Valentino Jayadi, T. Handhayani, and M. Dolok Lauro, "Perbandingan Knn Dan Svm Untuk Klasifikasi Kualitas Udara Di Jakarta," *J. Ilmu Komput. dan Sist. Inf.*, vol. 11, no. 2, pp. 11–17, 2023.