Contents lists available at https://csi.unisi.my.id

Computer Science Insights

Journal Page is available to https://csi.unisi.my.id

Research article

Comparison of K-Nearest Neighbor and Naïve Bayes *Classification Methods for Coconut Maturity Data*

M. Dhanul Paridzhi [1 ,*] *,Dwi Yuli Prasety* [2] *, Samsudin* [3]

[1234] *Faculty of Engineering and Computer Science, Indragiri Islamic University, Tembilahan, Riau*
email: [1,*] *mparidzhi@gmail.com*, [2] *dwiyuliprasetyo@gmail.com, samsudinsadek@gmai.com*
** Correspondence*

ARTICLE INFO

Article history:
Received June 18, 2025
Revised June 21, 2025
Accepted June 23, 2025
Available online June 30, 2025

Keywords:
Classification
K-Nearest Neighbors
Naïve Bayes
Coconut Maturity
Data Mining

Please cite this article in IEEE style as:
F. Author, S. Author, T. Author and F. Author, "Article Title," *Data Science Insights*, vol. X, no. X, pp. XX-XXX, 202X.

ABSTRACT

Classification of coconut maturity level is an important aspect in the coconut industry to ensure product quality and optimal selling value. This study aims to compare the performance of two data mining classification methods, namely K-Nearest Neighbors (KNN) and Naïve Bayes, in classifying coconut maturity levels in Tembilahan, Indragiri Hilir Regency. The dataset used consists of 300 data with parameters such as skin color, fruit weight, diameter, and fruit texture condition. Evaluation was conducted using accuracy, precision, recall, F1-score, and Cohen's Kappa metrics with 80:20 data division for training and testing. The results showed that the Naïve Bayes method provided superior performance with an accuracy of 86.67% and a Cohen's Kappa value of 0.795 compared to KNN which only achieved an accuracy of 43.33% and a Cohen's Kappa value of 0.144. Naïve Bayes also showed better consistency in classifying the three categories of coconut maturity (young, semi-old, and old) with an average F1-score of 0.8455 versus 0.4119 for KNN. This study provides a recommendation that the Naïve Bayes method is more effective for coconut maturity classification, especially in data conditions that are not fully optimized in terms of preprocessing. The results show that the application of the Naïve Bayes-based automatic classification system has great potential in helping coconut farmers and processors improve production efficiency and selling value through accurate and consistent maturity classification, although improvements are still needed in the recognition of the “Half-Old Coconut” class which has low recall due to the complexity of its characteristics.

1. Introduction

Indonesia as an agricultural country has great potential in the agricultural sector, especially coconut commodities. The Tembilahan area in Indragiri Hilir Regency is one of the largest coconut production centers in Indonesia, with significant contributions to the local economy. Coconut has a strategic role not only economically but also socially, as its various derivative products such as coconut oil, copra, and coconut milk have become an integral part of the lives of Indonesian people.

The main problem faced by coconut farmers and processors is determining the maturity level of coconut accurately and consistently. Errors in maturity classification can result in products that do not meet market standards, potentially reducing their selling value and causing significant economic losses. Traditional methods that rely on farmers' experience and intuition often result in subjective and inconsistent assessments, especially when large production volumes are involved.

The development of artificial intelligence and data mining technologies provide innovative solutions to overcome these problems. These technologies enable the development of automated classification systems that can provide objective, consistent and accurate results. Classification is one of the techniques in data mining that aims to group data into certain categories or classes based on the attributes possessed by the data.

In the context of coconut maturity classification, several methods have been developed and used, but two of the most popular and effective methods are K-Nearest Neighbors (KNN) and Naïve Bayes. The KNN method works on the principle of proximity, where classification is done by finding the k nearest neighbors of the data to be classified, then determining the class based on the majority of the classes of these neighbors. Meanwhile, Naïve Bayes is a probabilistic classification method based on Bayes' Theorem with the assumption that each attribute is independent of each other.

Although these two methods have been widely used in various classification applications, there is no research that specifically compares their performance in the context of coconut maturity classification. Therefore, this study aims to analyze and compare the effectiveness of the two methods in classifying coconut maturity data, so as to provide recommendations for the most suitable method to be implemented in the coconut industry.

2. Research Methods

This research uses a quantitative approach with an experimental design to compare the performance of two data mining classification methods. The research location was focused on Tembilahan, Indragiri Hilir Regency, which was chosen because it is one of the largest coconut production centers in Indonesia with representative coconut characteristics for the Indonesian region.

2.1. Data Collection

The data used in this study is primary data collected directly from the field through a combination of observation and structured interview methods. The data collection process involved direct observation of the physical characteristics of the coconuts and interviews with experienced farmers to validate the maturity categories. The final dataset consists of 300 coconut samples that have been categorized into three maturity levels: young coconut, semi-old coconut, and old coconut.

Each coconut sample was characterized based on four main attributes: coconut skin color, fruit weight in grams, fruit ring diameter in centimeters, and fruit texture condition. The selection of these attributes was based on consultations with agricultural experts and experienced farmers who identified these parameters as the most relevant indicators for determining coconut maturity level.

2.2. Data Preprocessing

The data preprocessing stage is a critical step in preparing the dataset for the classification process. In this stage, data cleaning is performed to remove missing or irrelevant values, as well as data normalization to ensure that all attributes have a balanced scale. This process is especially important for the KNN method which is sensitive to scale differences between attributes.

The data was then divided into two subsets using a stratified sampling technique with a ratio of 80:20, where 80% of the data (240 samples) were used as training set and 20% of the data (60 samples) were used as testing set. This division ensures that the proportion of each maturity class is proportionally represented in both subsets.

3. Implementation of Classification Methods

3.1.1. K-Nearest Neighbors (KNN)

The KNN method is implemented based on the principle of Euclidean distance calculation to determine the closeness between data. The distance calculation formula used is:

$$d_{(xy)} = \sqrt{\sum_{j=1}^{m}(X_i - Y_i)^2} \qquad (1)$$

where d is the closeness distance value, x is the training data, y is the test data, and m is the total attributes. The KNN algorithm works by calculating the distance between the test data and all data in the training set, then selecting the k nearest neighbors to determine the class based on majority voting.

3.1.2. Naïve Bayes

The implementation of Naïve Bayes is based on Bayes' Theorem with the assumption of conditional independence between attributes. The basic formula used is:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \qquad (2)$$

where P(H|X) is the probability of hypothesis H based on condition X, P(X|H) is the probability of X based on condition hypothesis H, P(H) is the prior probability of hypothesis H, and P(X) is the probability of evidence X. This method calculates the posterior probability for each class and selects the class with the highest probability as the classification result.

3.2. Evaluation Metrics

Evaluation of the performance of both methods was done using various standard metrics in classification. Accuracy was calculated using the formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

where TP (True Positive) is the number of correct positive predictions, TN (True Negative) is the number of correct negative predictions, FP (False Positive) is the number of incorrect positive predictions, and FN (False Negative) is the number of incorrect negative predictions.

Precision and recall are calculated using the formulas:

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad (4)$$

$$Recall = \frac{TP}{TP+FN} x\ 100\% \quad (5)$$

F1-score as the harmonic mean of precision and recall is calculated with:

$$F_1 = 2 . \frac{Precision .recall}{Precision+recall} \quad (6)$$

Cohen's Kappa is used to measure the level of agreement between predicted and true values with correction for agreement that occurs by chance:

$$k = \frac{P_o - P_e}{1 - P_e} \quad (7)$$

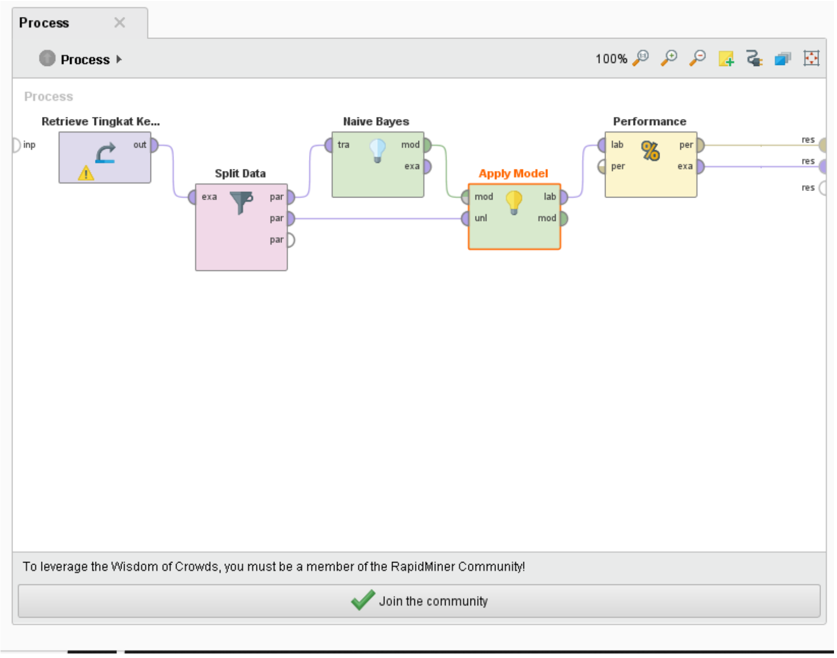
3.3. Tools and Devices

The implementation and experiments were conducted using RapidMiner software which provides an intuitive visual interface for the data mining process. The hardware used is a computer with a minimum specification of Intel Core i5 processor, 4GB RAM, and SSD storage to ensure the smoothness of the analysis process. Visualization of the analysis results was done using Tableau to provide an easy-to-understand graphical representation.

4. Results and Discussion

This section presents the results of the classification process of coconut maturity level data using two methods, namely k-Nearest Neighbor (k-NN) and Naive Bayes. The research was conducted using RapidMiner software with a dataset of 300 data, which was divided into 80% training data and 20% test data. Evaluation is done based on accuracy, precision, recall, F1-score, and Cohen's Kappa value metrics.

4.1. Implementation of Classification Process



To leverage the Wisdom of Crowds, you must be a member of the RapidMiner Community!

Figure 1. Implementation of Classification Process

In Figure 1, the implementation of the two classification methods is done using RapidMiner with a systematic process flow. The process begins with reading the coconut maturity dataset from the local repository, followed by data division using the Split Data operator with a ratio of 80:20. The training data is then processed using each classification algorithm (KNN and Naïve Bayes) to build a model, which is then applied to the testing data using the Apply Model operator. Performance evaluation is performed using the Performance (Classification) operator which generates various evaluation metrics including confusion matrix, accuracy, precision, recall, and F1-score.

4.2. Classification Results of the K-Nearest Neighbors (KNN) Method

Table 1. Classification Results of the K-Nearest Neighbots (KNN) Method

	True Young Coconut	True Half Old Coconut	True Old Coconut	Total Pred.
Pred. Young Coconut	16 (TP)	4	2	22
Pred. Half Old Coconut	9	4 (TP)	9	22
Pred. Old Coconut	1	9	6 (TP)	16
Total Actual	26	17	17	60

The results of the KNN method implementation show a relatively low performance in classifying coconut maturity levels. The confusion matrix analysis shows that out of 60 test data, only 26 data were classified correctly, resulting in an accuracy of 43.33%. The misclassification distribution showed that the KNN model had significant difficulty in distinguishing between the three coconut maturity classes.

The calculation of the evaluation metrics per class showed that the "Young Coconut" class performed best with a precision of 72.73% and a recall of 61.54%, resulting in an F1-score of 0.6664. In contrast, the "Half-Old Coconut" class showed the worst performance with a precision of only 18.18% and a recall of 23.53%, resulting in an F1-score of 0.2051. The "Old Coconut" class is in the middle position with a precision of 37.50% and a recall of 35.29%, resulting in an F1-score of 0.3632.

The obtained Cohen's Kappa value of 0.144 indicates a very weak level of agreement between model predictions and actual values. This indicates that the performance of the KNN model is barely better than random prediction. The low performance of KNN can be caused by several factors, including sensitivity to scale differences between attributes, dependence on the quality and distribution of training data, and difficulty in handling data with overlapping characteristics between classes.

4.3. Classification Results of Naïve Bayes Method

Table 2. Classification Results of Naïve Bayes Method

	True Young Coconut	True Half Old Coconut	True Old Coconut	Total Pred.
Pred. Young Coconut	25 (TP)	1	0	26
Pred. Half Old Coconut	1	11 (TP)	1	13
Pred. Old Coconut	0	5	16 (TP)	21
Total Actual	26	17	17	60

In contrast to KNN, the implementation of the Naïve Bayes method shows a very satisfactory performance in classifying the maturity level of coconut. Out of 60 test data, 52 data were classified correctly, resulting in an accuracy of 86.67%. This result shows a significant improvement of 43.34% compared to the KNN method.

Per-class analysis showed that Naïve Bayes provided consistently high performance for all classes. The "Young Coconut" class achieved perfect performance with a precision and recall of 96.15% each, resulting in an F1-score of 0.9615. The "Old Coconut" class also performed very well with 76.19% precision and 94.12% recall, resulting in an F1-score of 0.8413. Although the "Half-Old Coconut" class has a relatively lower performance with 84.62% precision and 64.71% recall, it still produces a good F1-score of 0.7336.

The Cohen's Kappa value of 0.795 indicates a strong level of agreement between the model prediction and the actual value. This value indicates that the performance of the Naïve Bayes model is much better than random prediction and has high reliability. The superiority of Naïve Bayes can be explained by its ability to handle data with complex distributions through a probabilistic approach, as well as its resistance to noise and outliers in the dataset.

4.4. Performance Comparison Analysis

Metrik	k-NN	Naive Bayes
Akurasi	43,33%	86,67%
Kappa	0,144	0,795
F1 Score	0,4119	0,8455
Recall rata2	40,12%	85,00%
Precision rata2	42,80%	85,65%

A comprehensive comparison between the two methods shows the clear superiority of Naïve Bayes. In terms of accuracy, Naïve Bayes outperforms KNN by a very significant margin (86.67% vs. 43.33%). This difference is also consistent across all other evaluation metrics, including average precision (85.65% vs. 42.80%), average recall (85.00% vs. 40.12%), and average F1-score (0.8455 vs. 0.4119).

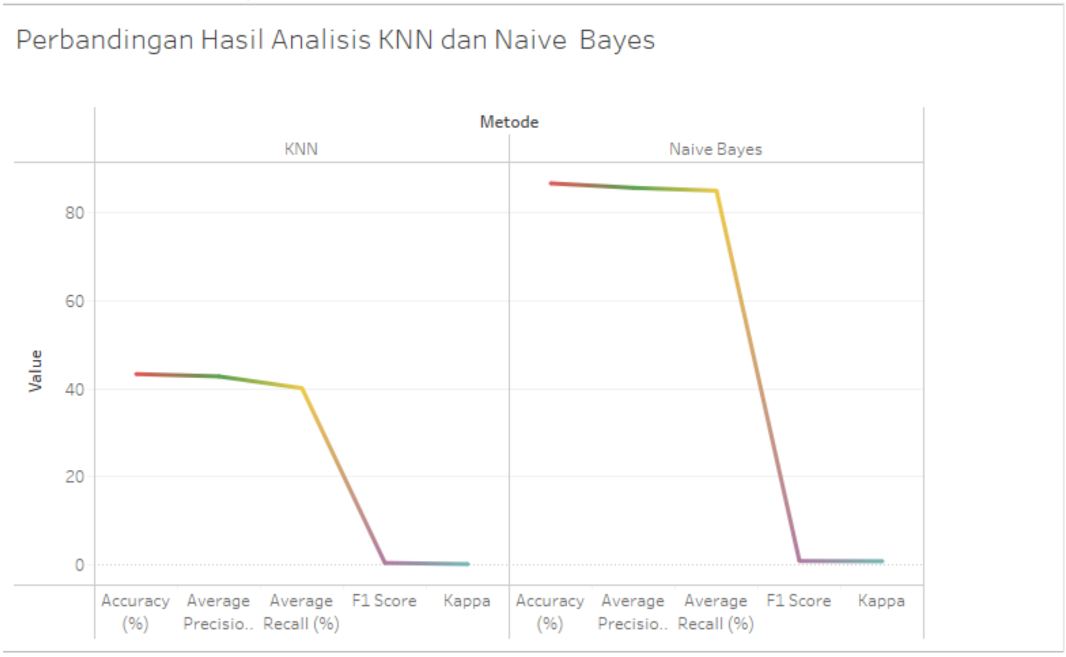


Figure 2. Comparison of KNN and Naive Bayes analysis results

The graph in Figure 2. displays the performance comparison of the two classification methods, KNN and Naive Bayes, based on several metrics such as accuracy, average precision, average recall, F1 Score, and Kappa. Each metric gives an idea of how well the model works in classifying the data in general. It can be seen that the metric values of Naive Bayes tend to be higher than KNN in all aspects of evaluation.

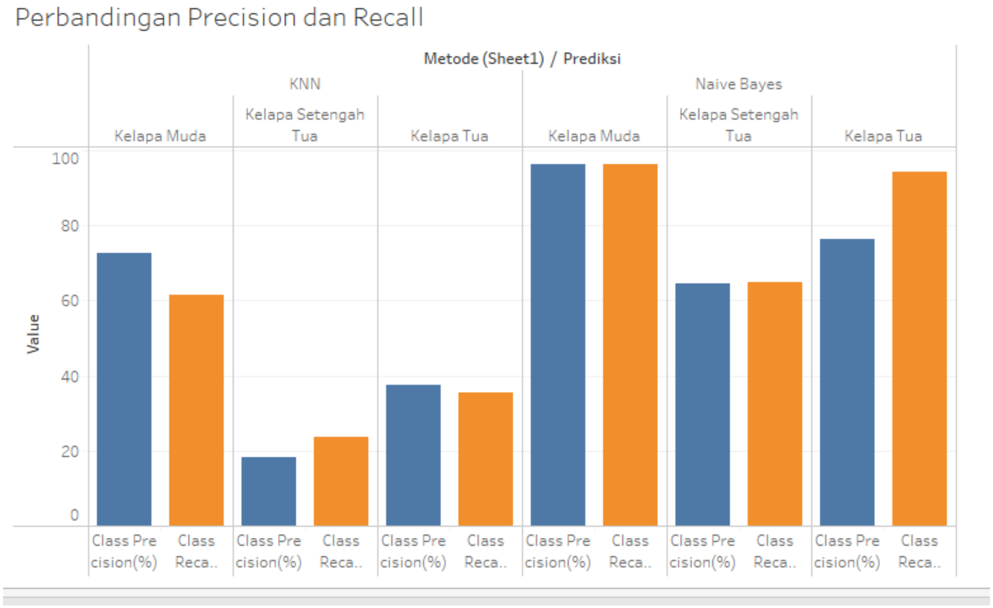


Figure 3. Precision and Recall Comparison

The second graph in Figure 3. presents the comparison of precision and recall values for each coconut category class (Young Coconut, Half-Old Coconut, and Old Coconut) in both classification methods, namely KNN and Naive Bayes. The graph shows that Naive Bayes consistently outperforms in each class, both in terms of precision and recall.

For the Young Coconut class, Naive Bayes performed almost perfectly, with very high precision and recall and almost the same. The same can be seen for the Old Coconut class, where the recall is very high, indicating the ability of Naive Bayes to recognize the class very well. As for the Half-Old Coconut class, although not as high as the other classes, Naive Bayes still performed much better than KNN.

In contrast, KNN performed poorly, especially for the Half-Old Coconut class, where its precision and recall were very low. This shows that KNN has difficulty in distinguishing and classifying this class. Even in other classes, the precision and recall values of KNN are still lower than Naive Bayes. This graph confirms that Naive Bayes is not only superior overall, but also more stable in recognizing each class.

From the conclusions of Figure 2 and Figure 3, overall, the performance of k-NN is very low due to the difficulty of the model in recognizing close patterns between classes, especially if the features between classes are not significantly separated. Meanwhile, Naive Bayes is able to capture broader feature distribution patterns through a probabilistic approach, resulting in a more stable and accurate classification.

The main factors that lead to this difference include:

- KNN is highly dependent on the quality and distribution of training data and is sensitive to outlier or non-standardized data.

- Naive Bayes is able to provide better predictions because it takes into account the probability distribution between features even under the assumption that features are independent of each other.

- The "Half-Old" class presents its own classification challenges due to the similarity of its attributes to the other two classes, causing a decrease in recall in both models, but the impact is greater for k-NN.

Thus, based on the evaluation results, the Naive Bayes model is recommended as a more effective classification method for classifying coconut maturity levels in Tembilahan compared to the k-NN method.

4.5. Practical Implications

The results of this study have significant practical implications for the coconut industry. The implementation of a Naïve Bayes-based automatic classification system can assist coconut farmers and processors in determining the maturity level with high and consistent accuracy. This can improve the efficiency of the production process, reduce losses due to misclassification, and ultimately increase the selling value of coconut products.

However, it should be noted that although Naïve Bayes shows superior performance, there is still room for improvement, especially in classifying the "Half-Old Coconut" class which shows a relatively lower recall. This indicates that the class has more complex characteristics and may require additional attributes or more sophisticated preprocessing techniques to improve classification accuracy.

5. Conclusion

Based on the analysis of coconut maturity classification using K-Nearest Neighbors (KNN) and Naive Bayes algorithms, it is concluded that data quality greatly affects the performance of each method. Naive Bayes showed superior performance with an accuracy of 86.67% compared to KNN which only reached 43.33%, mainly due to its ability to handle numerical and categorical data without normalization and its resistance to outliers and noise. In contrast, KNN proved sensitive to scale differences between features and required well-processed data, such as normalization of numerical attributes and encoding of categorical features. Fruit weight and diameter features play an important role in classification, while color and texture features are particularly challenging for KNN due to their complex and non-uniform formats.

The findings of this study provide a significant practical contribution to the coconut industry, particularly in the development of an automated classification system that can improve efficiency and accuracy in determining the maturity level of coconut. The implementation of the Naïve Bayes method is recommended as an optimal solution for coconut maturity classification applications, with the potential to improve product quality and selling points in the Indonesian coconut industry.

For future research, it is recommended to explore more advanced classification methods such as Support Vector Machine, Random Forest, or Deep Learning, as well as perform more comprehensive data preprocessing optimization to further improve classification performance. In addition, expansion of the dataset with a larger sample and wider geographical variation can help validate the generalizability of the developed model.

References

[1] A. Muni *et al.*, “Deteksi dan Klasifikasi Hama dan Penyuakit Tanaman Kelapa Menggunakan Nearest Mean Classifier Di Kabupaten Indragiri Hilir,” *J. Perangkat Lunak*, vol. 6, pp. 414–427, 2024.doi: https://doi.org/10.32520/jupel.v6i3.3655

[2] U. Masparudin, Abdullah, “Sistem Prediksi Kualitas Santan Kelapa Menggunakan Nearest Mean Classifier (NMC),” *Sist. J. Sist. Inf.*, vol. 9, no. 3, pp. 645–655, 2020. doi: 10.32520/stmsi.v9i3.1015

[3] A. Muni and M. Jibril, “Penerapan Metode Naive Bayes Classifier Dalam Pemilihan Kualitas Bibit Kelapa Untuk Masyarakat Petani Kelapa Di Indragiri Hilir,” *J. Perangkat Lunak*, vol. 5, no. 3, pp. 313–322, 2023, doi: 10.32520/jupel.v5i3.2780.

[4] R. R. Marlis, Abdullah, and F. Yunita, “Sistem Prediksi Kualitas Kopra Putih Menggunakan k-Nearest Neighbor (k-NN),” *Sist. J. Sist. Inf.*, vol. 10, no. 2, pp. 290–299, 2021, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[5] O. D. Kurnia, E. T. A. Fena, D. Yuliana, A. Ningrum, E. Daniati, and A. Ristyawan, “Analisis Perbandingan Algoritma Naïve Bayes dengan K-Nearest Neighbor (KNN) Pada Dataset Mobile Price Classification,” vol. 8, pp. 1174–1183, 2024.

[6] S. Sahar, “Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung,” *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 79–86, 2020, doi: 10.33096/ijodas.v1i3.20.

[7] I. Yuswandi, D. Y. Prasetyo, “Sistem Pakar Diagnosa Penyakit Refraksi Menggunakan Metode Forward Chaining Berbasis Web,” *J. Perangkat Lunak*, vol. 14, no. 1, pp. 1–10, 2022, doi: 10.56708/progres.v14i1.300.

[8] H. Mardesci, M. Maryam, and K. Ihwan, “Forecasting Coconut Production in Indragiri Hilir with Autoregressive Integrated Moving Average Model,” *Sistemasi*, vol. 12, no. 1, p. 219, 2023, doi: 10.32520/stmsi.v12i1.2531.

[9] Samsudin, Z. Ilyas, and F. Tanjung, “Sistem Pakar Diagnosa Penyakit Sindrom Metabolik pada Rumah Sakit Umum Daerah Tembilahan.,” *Sist. J. Sist. Inf.*, vol. 12, no. September, pp. 1007–1018, 2023. doi: 10.32520/stmsi.v12i3.3516

[10] A. Ikhlas, A. Abdullah, and D. Y. Prasetyo, “Mesin Pembelajaran Ensemble Untuk Identifikasi Varietas Padi,” *Inform. Pertan.*, vol. 29, no. 2, p. 123, 2020, doi: 10.21082/ip.v29n2.2020.p123-130.

[11] F. M. Delta Maharani, A. Lia Hananto, S. Shofia Hilabi, F. Nur Apriani, A. Hananto, and B. Huda, “Perbandingan Metode Klasifikasi Sentimen Analisis Penggunaan E-Wallet Menggunakan Algoritma Naïve Bayes dan K-Nearest Neighbor,” *Metik J.*, vol. 6, no. 2, pp. 97–103, 2022, doi: 10.47002/metik.v6i2.372.

[12] C. Kurniawan and H. Irsyad, “Perbandingan Metode K-Nearest Neighbor Dan Naïve Bayes Untuk Klasifikasi Gender Berdasarkan Mata,” *J. Algoritm.*, vol. 2, no. 2, pp. 82–91, 2022, doi: 10.35957/algoritme.v2i2.2358.

[13] V. Nadya, “Perbandingan Metode Klasifikasi Naïve Bayes dan K-Nearest Neighbor pada Data Status Pembayaran Pajak Pertambahan Nilai di Kantor Pelayanan Pajak Pratama Surakarta,” pp. 231–236, 2024., doi: 10.31284/p.snestik.2024.5880

[14] S. I. Guslianto and S. ‘Uyun, “Klasifikasi Kematangan Buah Sawit Berdasarkan Fitur Warna, Bentuk dan Tekstur Menggunakan Algoritma K-NN,” *J. Edukasi dan Penelit. Inform.*, vol. 9, no. 3, p. 407, 2023, doi: 10.26418/jp.v9i3.64877.

[15] W. D. Prasetya and B. Sujatmiko, “Rancang Bangun Aplikasi dengan Perbandingan Metode K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penderita Penyakit Diabetes,” *J. Informatics Comput.*

Sci., vol. 3, no. 04, pp. 515–525, 2022, doi: 10.26740/jinacs.v3n04.p515-525.

[16] Z. Endesah, “Sistem Informasi Agronomi tanaman kelapa pada industri hulu dan hilir di Kabupaten Indragiri Hilir,” 2021.

[17] S. N. sari Muslim, “Perbandingan Algoritma Naive Bayes dan KNN dalam Analisis Sentimen Ulasan Pengguna Aplikasi Capcut,” *J. Inform. dan Teknol.*, vol. 2, no. 3, pp. 61–69, 2019. doi: 10.23960/jitet.v12i3S1.5156

[18] M. Amin and A. Bindas, “Pengklasifikasi Bibit Kelapa Menggunakan Algoritma Deep Learning Convolutional Neural Network,” *J. Perangkat Lunak*, vol. 6, pp. 405–413, 2024. doi: 10.32520/jupel.v6i3.3654

[19] M. R. Ridha and F. Yunita, “Pemilihan Bibit Kelapa Menggunakan Metode Nearest Mean Classifier Untuk Masyarakat Petani Kelapa Di Kabupaten Indragiri Hilir,” *J. Perangkat Lunak*, vol. 2, no. 3, pp. 101–114, 2020, doi: 10.32520/jupel.v2i3.1306.