The 2nd International Conference on Multidisciplinary Engineering (2nd ICOMDEN 2024)

The 2nd International Conference on Multidisciplinary Engineering: 00013 (2024) https://doi.org/10.29103/icomden.v2.xxxx
Volume 2, 2024
eISSN: xxxx-xxxx
Research Original Article

Comparative Analysis of K-Nearest Neighbor and Support Vector Machine Methods for Assessing Quality Standards of Palm Oil Bunches

Siti Hajar✉[1]**Rozzi Kesuma Dinata**[2]**Maryana**[3]
[1]Department of Informatics Engineering, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24353, Indonesia, siti.200170103@mhs.unimal.ac.id
[2]Department of Informatics Engineering, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24353, Indonesia, rozzi@unimal.ac.id
[3]Department of Informatics Engineering, Universitas Malikussaleh, Bukit Indah, 24353, Indonesia, maryana@unimal.ac.id
✉Corresponding Author: siti.200170103@mhs.unimal.ac.id | **Phone: +6285767408433**

Abstract
Oil palm (Elaeis guineensis Jacq) is a crucial crop in the agricultural sector, particularly in Indonesia, as it produces various economically valuable products. The quality of oil palm fruit bunches (TBS) significantly influences the production process of crude palm oil (CPO), making accurate quality assessments essential for maintaining industry standards. This study aims to compare the effectiveness of two machine learning methods, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM), in determining the acceptable quality of TBS. Using TBS data from the years 2019 to 2023, the research analyzes several variables, including maturity level and yield percentage, to develop a web-based system for classifying TBS. The classification process involves preprocessing the data, applying the algorithms, and evaluating their performance based on key metrics such as accuracy, recall, and precision. The results indicate that the K-NN method outperforms SVM, achieving an accuracy of 100%, a recall of 100%, and a precision of 100%. In contrast, the SVM method demonstrates an accuracy of 91%, a recall of 100%, and a precision of 91%. These findings highlight the effectiveness of K-NN in classifying TBS quality while also demonstrating the reliability of SVM. This research is expected to provide valuable insights and effective solutions for decision-making regarding the acceptance of TBS quality, ultimately benefiting stakeholders in the palm oil industry and serving as a reference for future studies in data mining classification.

Keywords: K-Nearest Neighbors, Support Vector Machine, Quality of Palm Oil Fruit Bunches, Data Mining, Classification

Introduction
Oil palm (Elaeis guineensis Jacq) is a perennial palm species that thrives in tropical regions. Currently, oil palm is one of the most important plantation crops in the agricultural sector, particularly due to its high economic value per hectare compared to other oil or fat-producing plants worldwide [1]. The oil palm industry significantly influences Indonesia's economic growth. The distribution of oil palm in Indonesia spans across the islands of Sumatra, Kalimantan, Java, Sulawesi, Papua, and several other islands. This sector plays a crucial role, as oil palm fruit is used as raw material for cooking oil, margarine, soap, cosmetics, and pharmaceuticals.

Quality is a critical factor related to company performance, especially at Inti Mitra Sawit Lestari. The Crude Palm Oil (CPO) industry is highly dependent on the quality of oil palm fruit, where the quality of the fruit bunches used in the production process determines the final yield of palm oil. This study compares the results of two classification algorithms, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM), in determining the acceptable quality of oil palm fruit.

The comparison aims to identify which classification method is more effective for determining the quality of oil palm bunches suitable for acceptance at Inti Mitra Sawit Lestari. A previous study titled "Comparison of Support Vector Machine and K-Nearest Neighbor Algorithms in Oil Palm Maturity" by Syawaluddin Kadafi Parinduri, Rika Rosnelly, Anton Purnama, Ameliana Sihotang, and Mimi Chintya Adelina found that from 34 image data of 2 categories, extracted features using Inception-V3, both SVM and K-NN achieved the same accuracy, with values of 0.500 each. The models obtained an Accuracy of 100%, F1 Score of 100%, Recall of 100%, and Precision of 100%. The classification of 6 image data resulted in all correct classifications.

Based on the discussion above, an appropriate approach to address these issues is to leverage information technology

advancements data mining technology is essential for effectively and efficiently processing data. The data mining algorithms used in this study are K-Nearest Neighbor and Support Vector Machine.

Literature Review

1. Data Mining

Data mining is the process of extracting data into information that allows users to quickly access large volumes of data. With the appropriate techniques, the data mining process can yield optimal results. Each data point in data mining consists of a specific class along with variables and determinants of that class. Through data mining, researchers can identify a class from the data variables they possess [2]. The general functions of data mining include association, sequence analysis, clustering, classification, regression, forecasting, and providing solutions.

2. Classification

Classification is a stage in discovering a pattern that differentiates or separates a concept or class of data. Its purpose is to estimate the class of an object whose label is unknown, while the model is derived from data analysis with known class labels. Classification is one of the data processing techniques that divides objects into several classes according to the desired number of classes. The classification technique works by grouping data based on training data and classification attribute values. Grouping rules are used to classify new data into existing groups, making the goal of classification the accuracy in predicting a value. In this study, the classification systems used are the K-Nearest Neighbor (K-NN) algorithm and the Support Vector Machine (SVM) algorithm, which are commonly employed in data classification.

3. K-Nearest Neighbor

The K-Nearest Neighbor (K-NN) algorithm is a method used for classifying objects based on the learning data closest in proximity to the object. K-Nearest Neighbor works by identifying groups of objects in the training data that are closest (most similar) to the new or testing data. K-NN is one of the algorithms used in classification. The working principle of K-Nearest Neighbor (K-NN) is to find the nearest distance between the data to be evaluated and its closest neighbors in the training data [2]. The Euclidean Distance is used as a measure for interpreting proximity between two objects, and it is commonly applied in distance calculations. In this study, the approach method used is the Euclidean Distance method. The details of the K-Nearest Neighbor algorithm process can be illustrated in the following flowchart :

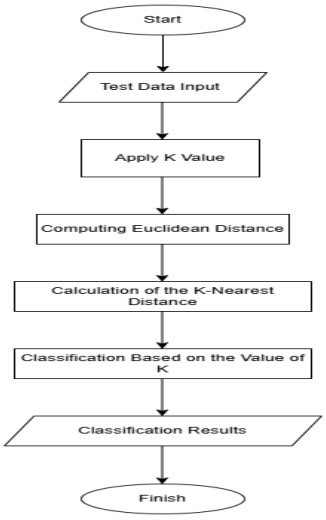

Figure 1. K-Nearest Neighbor algorithm

Based on the previous definition, the K-NN method can be interpreted as a technique that performs classification based on the nearest neighbor data or previously stored samples to reach a collective decision. The formula for calculating distance using Euclidean Distance is explained below.

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \qquad (1)$$

Description:
x1 = Sample data
x2 = Test data or testing data
i = Data variable
d = Distance
p = Number of training data

4. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning technique for data classification. It compares a selection of standard discrete parameter values known as the candidate set. The SVM method defines a boundary between two classes at the maximum distance from the nearest data points [3]. Introduced by Vapnik in 1992 with colleagues Bernhard Boser and Isabelle Guyon, SVM is prominent in pattern recognition. The algorithm uses nonlinear mapping to transform original training data into a higher dimension, aiming to find the best hyperplane that separates the two classes [4]. The details of the Support Vector Machine algorithm process can be illustrated in the following flowchart:

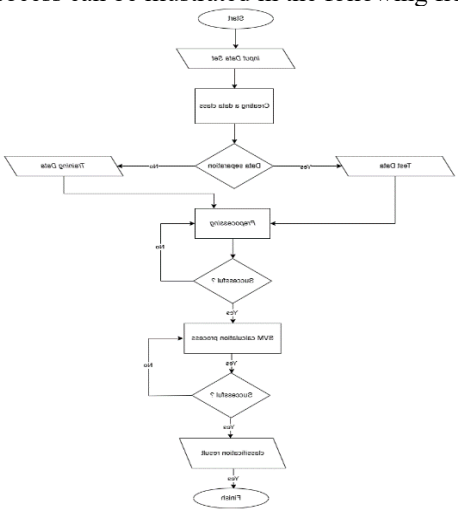

Figure 2. Support Vector Machine algorithm

$$f(x) = w_1 * x_1 + \cdots + wn * xn + b \qquad (2)$$

Description:
$f(x)$ = Predicted value
w = Weight value
x = Input value
b = Bias value

Here is the formula for Support Vector Machine using the Gradient Descent algorithm:

$$w_j = w_j + \eta \times \left(y_i - f(x_i)\right) \times x_{ij} \qquad (3)$$

$$b = b + \eta \times \left(y_i - f(x_i)\right) \qquad (4)$$

Description:
η = Learning rate (e.g., $\eta = 0{,}01$)
y_i = Actual target class value
$f(x_i)$ = Predicted result
x_{ij} = j-th feature of the i-th data point

5. Confusion Matrix

A confusion matrix performs testing to estimate correct and incorrect objects. The test order is tabulated in the confusion matrix, where the predicted class is displayed at the top of the matrix, and the observed class is on the left side. Each cell contains a number indicating how many cases actually belong to the observed class that were predicted [5].

6. Types of Oil Palm

In general, based on the thickness of the fruit flesh, kernel size, or shell thickness, there are three types of oil palm fruit: Dura, Pisifera, and Tenera. Although there are many types of oil palm, both long-established varieties and new developments through crossbreeding, these aim to achieve superior seed quality.

Materials & Methods

1. Data Collection

The sample data used in this research was obtained from PT Inti Mitra Sawit Lestari and consists of 600 TBS samples. The data is organized based on weighted assessments of several criteria established in this study. The calculation will yield outputs indicating whether the quality is "good" or "not good," as well as the accuracy rate of the two methods employed.

Table 1. Data for Classification Process

NO	SUPPLIER NAME	BUDGET	FRUIT RIPENESS COLOR LEVEL	OIL YIELD PERCENTAGE	OIL PALM FRUIT BUNCH TYPE
1	Usman	1850	4	26	Tenera
2	Ida	1750	3	23	Psifera

3	Gultom	1550	1	18	Dura
4	Sultan	1850	4	26	Tenera
5	Zendi	1855	4	26	Tenera
6	Raju	1855	4	26	Tenera
7	Rijal	1755	3	23	Psifera
8	Udin	1600	1	18	Dura
9	Jamal	1560	1	18	Dura
10	Aldo	1860	4	26	Tenera
…...	…...	…...	…...	…...	…...
600	Sodrik	1600	2	18	Dura

Table 1 presents the complete dataset that will be used in the classification process with the K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) methods. A total of 480 records will be used for training, while 120 records will be used for testing. This data will undergo a weighting process to be used in the calculations for the K-NN and SVM methods.

2. METHODS

Below is the system scheme from the research conducted.

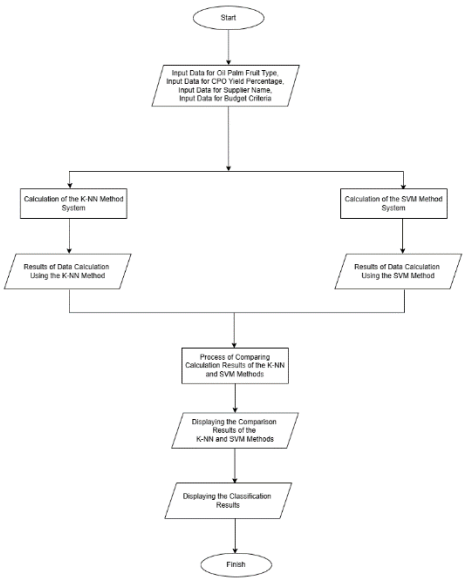

Figure 3. System Scheme

Figure 3 illustrates the assessment scheme for oil palm fruit bunches. The first step is to start. Starting is the initial process to begin the system. Next, input the data for the type of oil palm fruit, CPO yield percentage, supplier name, and budget criteria. The next stage involves the system calculation using the K-Nearest Neighbor (K-NN) method. Following this, the system calculation will be performed using the Support Vector Machine (SVM) method. Then, display the results of the data calculation using the K-NN method. Next, display the results of the data calculation using the SVM method. The subsequent step is the process of comparing the results of the calculations from the K-NN and SVM methods. Then, display the comparison results of the K-NN and SVM calculations. After that, perform the data classification and display the classification results. The classification results are the output that will be presented by the system, indicating the quality of oil palm fruit bunches that are eligible for acceptance at PT Inti Mitra Sawit Lestari. Finish serves as the final marker of the processes that have been carried out by the system.

3. Evaluation

To calculate the accuracy between the actual results and the predicted outcomes, the evaluation stage can use a Confusion Matrix. Table 2 presents the Confusion Matrix for calculating the accuracy of the research results.

Table 2. Confusion Matrixting Rules

	Positive Prediction	Negative Prediction
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Results and Discussion

The criteria used in this research are the supplier name, budget, ripeness level of the oil palm fruit bunches, yield

percentage, and type of oil palm fruit bunch. The values for each criterion can be seen in Table 2.

Table 3. Criteria Data

Criteria Code	Criteria
X1	Budget
X2	Fruit Ripeness Color Level
X3	Yield Percentage
X4	Type of Oil Palm Fruit Bunch

Table 4. Training Data For K-NN And SVM

No	Supplier Name	X1	X2	X3	X4	Classification
1	Usman	6	4	3	3	High Quality
2	Ida	4	3	2	2	High Quality
3	Gultom	1	1	1	1	Low Quality
4	Sultan	6	4	3	3	High Quality
5	Zendi	6	4	3	3	High Quality
…...	…...	…...	…...	…...	…...	…...
480	Sodrik	1	2	1	1	Low Quality

Table 5. Testing Data For K-NN And SVM

No	Supplier Name	X1	X2	X3	X4
1	Usman	5	3	2	2
2	Ida	8	4	3	3
3	Gultom	8	4	3	3
4	Sultan	5	3	2	2
5	Zendi	7	4	3	3
…...	…...	…...	…...	…...	…...
120	Sodrik	1	2	1	1

In the available training data, there are only four parameters related to determining the quality of oil palm fruit bunches: budget, ripeness level of the oil palm fruit bunches, percentage of yield, and type of oil palm fruit bunches. In the implementation of the K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) methods, we need to convert the target classes. In this study, for K-NN, a conversion value of 1 is assigned to the high quality category and a value of 2 to the low quality category. For SVM, a conversion value of 1 is assigned to the high quality category and -1 to the low quality category. Therefore, the data used for training can be seen in Table 5.

Table 6. Conversion Data Labels For K-NN

No	Supplier Name	X1	X2	X3	X4	Classification'
1	Usman	6	4	3	3	1
2	Ida	4	3	2	2	1
3	Gultom	1	1	1	1	2
4	Sultan	6	4	3	3	1
5	Zendi	6	4	3	3	1

…...	…...	…...	…...	…...	…...	…...
480	Sodrik	1	2	1	1	2

Table 7. Conversion Data Labels For SVM

No	Supplier Name	X1	X2	X3	X4	Classification'
1	Usman	6	4	3	3	1
2	Ida	4	3	2	2	1
3	Gultom	1	1	1	1	-1
4	Sultan	6	4	3	3	1
5	Zendi	6	4	3	3	1
…...	…...	…...	…...	…...	…...	…...
480	Sodrik	1	2	1	1	-1

Table 8. Results of the data arranged from smallest to largest distance and majority for K=5.

NO	Distance 1	Classification	NO	Distance 2	Classification	….	NO	Distance 120	Classification
7	0	High Quality	39	0	High Quality	….	63	0	Low Quaity
25	0	High Quality	41	0	High Quality	….	68	0	Low Quaity
31	0	High Quality	55	0	High Quality	….	78	0	Low Quaity
40	0	High Quality	59	0	High Quality	….	90	0	Low Quaity
48	0	High Quality	61	0	High Quality	….	102	0	Low Quaity

Table 8 shows the results of distance calculations from distance 1 to distance 120. These distance values will then be sorted from the smallest to the largest, and the value of K=5 will be determined.

Table 9. K-NN Calculation Results.

No	Supplier Name	X1	X2	X3	X4	Classification'	Classification results.
1	Usman	5	3	2	2	1	1
2	Ida	8	4	3	3	1	1
3	Gultom	8	4	3	3	1	1
4	Sultan	5	3	2	2	1	1
5	Zendi	7	4	3	3	1	1
…...	…...	…...	…...	…...	…...	…...	…...
120	Sodrik	1	2	1	1	2	2

Table 10. Confusion Matrix Results for K-NN

	Positive Prediction	Negative Prediction
Actual Positive	True Positive (TP) = 109	False Negative (FN) = 0
Actual Negative	False Positive (FP) = 0	True Negative (TN) = 11

To view the number of predicted data points categorized as high quality and low quality compared to the actual data, refer to Figure 4.

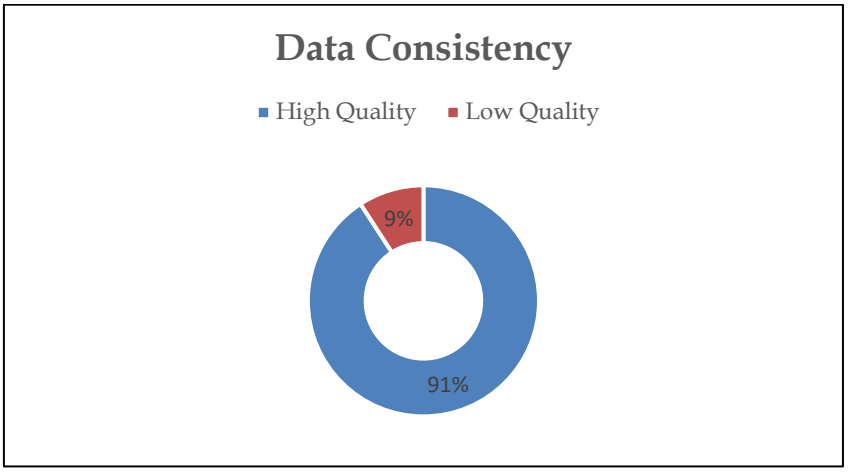

Figure 4. Data Consistency K-NN

It can be seen from Table 10 that there are 109 classified data points using the K-NN method that fall into the high-quality category. However, there are 11 classified data points that fall into the low-quality category. Referring to Figure 4, 91% are in the high-quality category, while 9% are in the low-quality category.

After performing the K-NN calculations, the next step is to conduct SVM calculations to allow for a comparison of the results between the two methods.

Table 11. SVM Calculation Results

X1	X2	X3	X4	Classification Labels	F(X)	Prediction
0,571428571	0,666666667	0,5	0,5	1	0,759282	1
1	1	1	1	1	0,694194	1
1	1	1	1	1	0,694194	1
0,571428571	0,666666667	0,5	0,5	1	0,759282	1
0,857142857	1	1	1	1	0,695344	1
…...	…...	…...	…...	…...	…...	…...
0	0,333333333	0	0	-1	0,825519	1

Table 12. Confusion Matrix Results for SVM

	Positive Prediction	Negative Prediction
Actual Positive	True Positive (TP) = 109	False Negative (FN) = 0
Actual Negative	False Positive (FP) = 11	True Negative (TN) = 0

To view the number of predicted data points categorized as high quality and low quality compared to the actual data, refer to Figure 4.

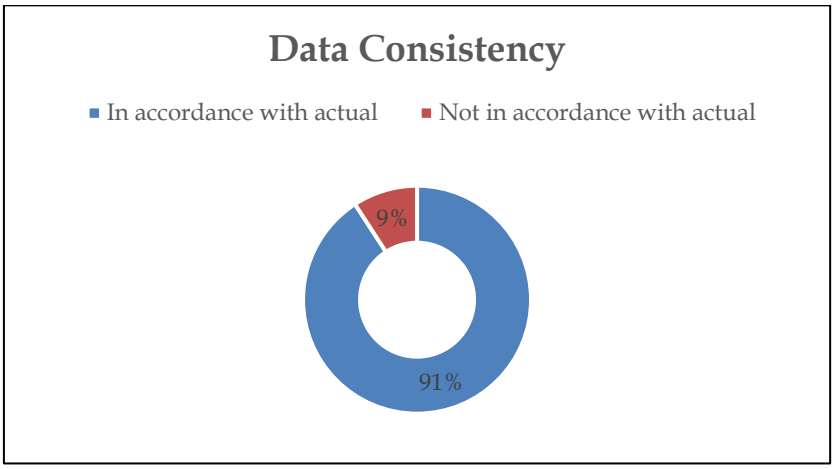

Figure 5. Data Consistency SVM

It can be seen from Table 12 that there are 109 predicted data points using the SVM method that match the actual data. However, there are 11 predicted data points that do not match the actual data. Referring to Figure 5, 91% of the predicted data aligns with the actual data, while 9% does not.

Therefore, the accuracy, precision, and recall obtained from the comparison of the K-NN and SVM methods for determining the quality of oil palm bunches that are suitable for acceptance can be seen in Figure 6 below.

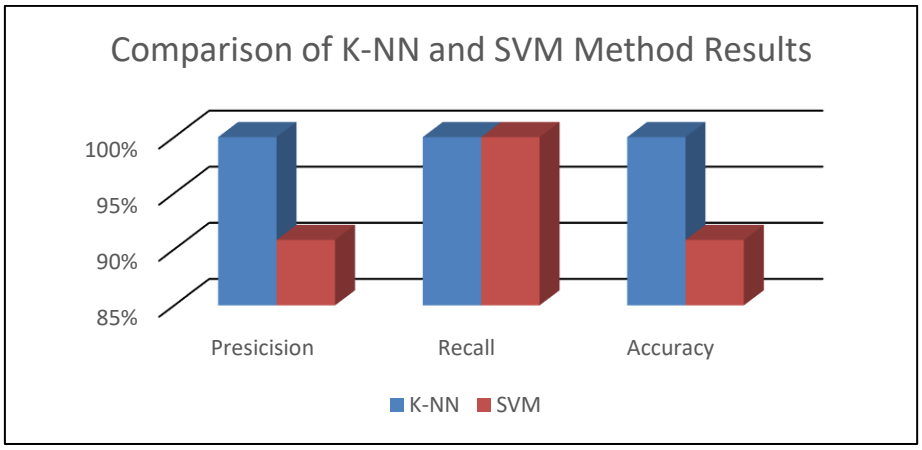

Figure 6. Evaluation of the Comparison of K-NN and SVM Results Using the Confusion Matrix

The accuracy results from the K-Nearest Neighbor method are as follows: accuracy 100%, precision 100%, and recall 100%. The accuracy results from the Support Vector Machine (SVM) method are: accuracy 91%, precision 91%, and recall 100%. The final results of the implementation of the system for determining the quality of oil palm bunches that are suitable for acceptance, comparing the calculations of the K-Nearest Neighbor method and the Support Vector Machine method, can be seen in Figure 7.

Figure 7. Comparison Results of K-NN and SVM

Conclusions

With the classification system for determining the quality of oil palm bunches that are suitable for acceptance, a solution can be provided for PT in assessing the quality of high and low-quality oil palm bunches for acceptance. The implementation of the K-Nearest Neighbor and Support Vector Machine algorithms allows for high-accuracy classification of oil palm bunches. By utilizing features such as supplier name, budget, ripeness level of the oil palm bunches, yield percentage, and type of oil palm bunch, the K-NN and SVM algorithm models can automatically classify the quality of oil palm bunches into two target classes: high quality and low quality. The accuracy results from the K-NN method are 100%, with a precision of 100% and a recall of 100%. In contrast, the SVM method has an accuracy of 91%, with a precision of 91% and a recall of 100%. From the comparison of the accuracy results obtained, it can be concluded that both methods are reliable, as they both exhibit high accuracy values.

Acknowledgments

The author expresses heartfelt gratitude to both parents and to Mr. Rozzi Kesuma Dinata, S.T., M.Eng, as the main supervisor, for his invaluable guidance during the completion of this research. The author also extends thanks to Mrs. Maryana, S.Si., M.Si., as the second supervisor, for her patient direction and support throughout the research process. Special appreciation is given to PT. Inti Mitra Sawit Lestari for their assistance and guidance during this research. Additionally, the author is grateful to friends who have provided encouragement and support throughout the completion of this research.

References

[1] H. Stephanie, N. Tinaprilla, and D. A. Rifin, "EFISIENSI PABRIK KELAPA SAWIT DI INDONESIA," *Jurnal Agribisnis Indonesia*, vol. 6, no. 1, pp. 27–36, 2018, [Online]. Available: http://journal.ipb.ac.id/index.php/jagbi

[2] R. K. Dinata, H. Akbar, and N. Hasdyna, "Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus," *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 104–111, Aug. 2020, doi: 10.33096/ilkom.v12i2.539.104-111.

[3] R. Umar, I. Riadi, D. Astria Faroek, A. Dahlan, S. Sistem Informasi, and U. Ahmad Dahlan, "Komparasi Image Matching Menggunakan Metode K-Nearest Neighbor (KNN) dan Metode Support Vector Machine (SVM)," 2020. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[4] W. R. U. Fadilah, D. Agfiannisa, and Y. Azhar, "Analisis Prediksi Harga Saham PT. Telekomunikasi Indonesia Menggunakan Metode Support Vector Machine," *Fountain of Informatics Journal*, vol. 5, no. 2, p. 45, Sep. 2020, doi: 10.21111/fij.v5i2.4449.

[5] R. Kesuma Dinata and N. Hasdyna, "KLASIFIKASI SEKOLAH MENENGAH PERTAMA/SEDERAJAT WILAYAH BIREUEN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS BERBASIS WEB," 2020.

Document Length Restrictions

Papers accepted for publication are strictly limited to **6-12 pages** in a one-column format following this template. (one single space, 9pt font). **PLEASE USE THIS TEMPLATE FORMAT FOR PREPARE YOUR MANUSCRIPT**