EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI : 10.20885/EKSAKTA.vol7.iss1.art5

Orange Classification using Naïve Bayes and K-Nearest Neighbor Algorithms based on Its Physical Properties

Fadli Hafizulhaq[1] **, Andasuryani Andasuryani**[2]*

[1,2] Department of Agricultural and Biosystem Engineering, Faculty of Agricultural Technology, Universitas Andalas, Limau Manis, Padang, Sumatra Barat 25163, Indonesia

*Corresponding author: andasuryani@ae.unand.ac.id

Abstract: Oranges are among the most widely consumed fruits globally. While many farmers possess extensive knowledge of orange cultivation, they often lack expertise in post-harvest handling and processing. Classification or grading is a crucial step after harvest to ensure quality. Machine learning offers an efficient solution for automating this process and decreasing the time consumed. This study implements two machine learning algorithms, Naïve Bayes and K-Nearest Neighbor, to classify Gerga oranges based on different training-to-test data ratios (75:25, 50:50, and 25:75). The results indicate that as the training data decreases, the accuracy of Naïve Bayes improves, but its precision declines, whereas K-Nearest Neighbor exhibits the opposite trend. The best accuracy (90% accuracy) was produced by NB-25 and KNN-75. Meanwhile, precision and recall value were more important in order to reduce economic losses and buyer dissatisfaction, so that users can profit more. In this case, the KNN-75 model is the best to classify Gerga oranges into the right groups (85% precision, 91% recall). Despite the differences in class importance, KNN offers a steadier and more balanced outcome for both sides of the dataset. KNN is also more reliable to handle many number of samples in real practice when the model is used to design sorting or grading machines for oranges.

Keywords: Orange, Classification, Naïve bayes, K-neareast neighbor, Machine learning

Introduction

As an agrarian country, Indonesia produces tons of fruits every year. This archipelagic country exported 1.05 million metric tons of fruits to many countries in 2023 [1]. One type of fruit that is exported is orange. Indonesia has many orange-producing areas, including Kerinci, Jambi, Indonesia. The famous orange from Kerinci is called the Gerga orange.

Gerga orange (*Citrus reticulate*) is a hybrid orange between *Citrus sinensis* (L.) Osbeck with *Citrus reticulata* (L.) Blanco [2]. Gerga orange is officially catalogued and designated as the Keprok Rimau Gerga Lebong (RGL), which was first introduced to the Lebong Regency of Bengkulu Province in the year 2009 [3]. Following the introduction, local farmers in Kerinci Regency extensively expanded the agribusiness and established the Gerga orange as a horticultural focal point not only in Bengkulu but also in Kerinci Regency, Jambi Province. In 2024, Badan Pusat Statistik Kerinci reported that Kerinci produced 403,589.14 quintals of Siam/Keprok oranges [4].

The Gerga orange is mainly distributed in a local network where farmers are still selling their product directly to final consumers. Besides that, Gerga orange farmers in Kerinci also make their land a tourist destination. This condition prevents the farmers from increasing the economic value of the orange. In order to increase the revenue of the farmers, the export scenario must be considered. However, various post-harvest processes are needed to make sure that the orange meets the export standards.

One of the important post-harvest processes in export preparation is grading. Grading or classification is a technique for grouping fruits based on their quality. This process can provide benefits for both producers and consumers [5]. In the traditional method, grading is done by relying on observations from producers which rely on special abilities and are subjective. The grading process of the Gerga orange is still using that traditional method, which consumes a lot of time. However, in this digital and artificial intelligence (AI) era, machine learning has been proven to help the grading process become faster and more efficient. [6], [7]. As far as we know, there has been no research that has developed a machine learning method for grading Gerga oranges.

Among various machine learning methods, Naïve Bayes (NB) and K-Nearest Neighbor (KNN) are two common and easy-to-use classification methods. NB is a classification algorithm based on the Bayesian hypothesis and naive presumption. However, it has proven to be useful in complex real-world conditions [8]. Based on previous research, the application of the NB Classifier to tomatoes obtained classification results with an accuracy of 76% [9].

The KNN classifier determines the class of a test sample by majority voting among its k closest neighbors. These neighbors are identified using a distance metric, with Euclidean distance being one of the most frequently used options [10]. A previous study reported that the KNN classifier classified tangerine fruit for ripeness level with 93% accuracy [11]. Neither NB nor KNN has ever been used to classify Gerga oranges.

The aim of this study is to apply the NB and KNN machine learning algorithms to the classification of Gerga oranges based on their physical properties. The amount of training data will be varied to see changes in the accuracy and precision of the model built. This research will contribute to the development of an AI-based sorting and grading machine for Gerga oranges in the future.

Materials and Methods

Data Collecting

Gerga oranges were randomly selected from a field in Kerinci, Jambi, Indonesia. The fruits were harvested in July 2024 and shipped to Padang, West Sumatra. Data collection was carried out at room temperature with the number of specimens being 120 Gerga oranges with a medium level of ripeness. Geometric measurements of the fruit were carried out using a vernier calliper accuracy (0.01 mm accuracy). Three fruit diameters that represent length (L), width (W), and thickness (T) were taken to calculate the size and shape of the fruit samples. An illustration of the measurement parameters can be seen in Figure 1. Meanwhile, the individual fruit mass was measured with an analytical balance (0.01 g accuracy).

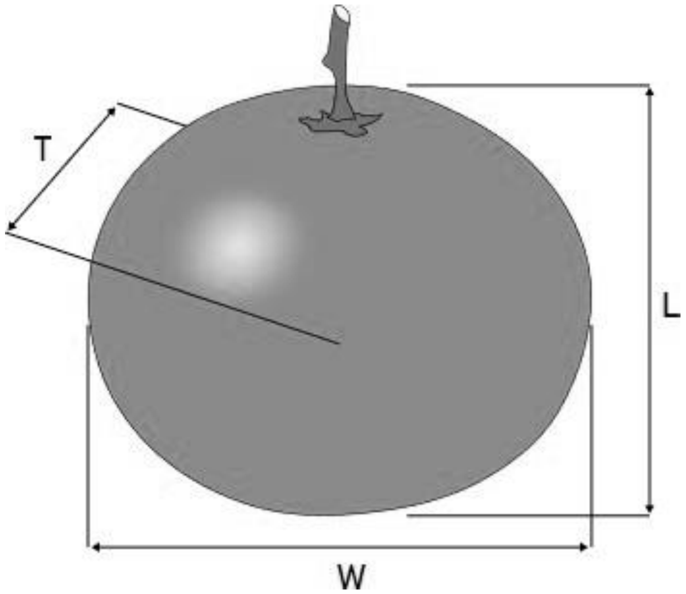


Figure 1. Illustration of measuring the dimensions of an orange fruit

The geometric mean diameter (GMD) was calculated using the results of measuring length (L), width (W), and thickness (T) as written in equation 1. After that, grading was carried out manually based on the GMD and mass of the Gerga orange fruit with the number of groups being 3 (Grade A, B, and C).

$$GMD = \sqrt[3]{LxWxT} \qquad (1)$$

Due to the lack of a standardized grading system specific to the Gerga orange variety, a custom composite scoring method was developed. The grading criteria were based on two primary physical parameters: fruit mass and GMD. First, individual scores ranging from 1 to 3 were assigned to both the mass (M_range) and the diameter (GMD_range) based on predetermined physical thresholds (see Table 1).

Table 1. Scoring threshold for mass and GMD

Parameter	Score 1 (Low)	Score 2 (Medium)	Score 3 (High)
Mass	< 160 g	160-189.9 g	≥ 190 g
GMD	< 65 mm	65 – 67.9 mm	≥ 68 mm

To determine the final classification, a composite index (MxGMD) was calculated for each fruit by taking the arithmetic mean of its mass score and diameter score as written in equation 2.

$$MxGMD = \frac{M\_range + GMD\_range}{2} \qquad (2)$$

The final ground-truth labels for the machine learning algorithms were then assigned based on this MxGMD index: fruits with an index of 2.5 or higher were classified as Grade A, an index of exactly 2.0 as Grade B, and an index of 1.5 or lower as Grade C.

Data Preprocessing and Splitting

Before applying the machine learning algorithms (NB and KNN), three individual dimensional measurements (length, width, and thickness) were excluded from the final dataset to reduce noise. Only mass, GMD, and orange level are used to run the classification process. Next, the dataset was split into two groups (training and test data). The composition of training data and test data is varied into three as shown in Table 2.

The classification process is carried out using the Scikit-learn library in the Python programming language. Both models are created with random_state=True to avoid different results each time the program is executed.

Following the data split, feature scaling was applied. The data was standardized using a standard scaler procedure, which centered the variables to a mean of zero and scaled them to a standard deviation of one. The standardization parameters were fitted exclusively on the training data and subsequently applied to transform the testing data, thereby preventing data leakage. While standardizing features is strictly required to ensure equal weight distribution in the KNN distance calculations, applying this identical scaled dataset to the Gaussian Naïve Bayes algorithm ensured a uniform baseline for an objective comparison of their predictive performance. The prediction, as the result of the models, was created using test data unseen during the training phase. Overall performance was summarized with confusion matrix metrics.

Table 2. Dataset splitting for machine learning algorithms

	Training data (%)	Testing data (%)
NB-75	75	25
NB-50	50	50
NB-25	25	75
KNN-75	75	25
KNN-50	50	50
KNN-25	25	75

Confusion Matrix

A confusion matrix is a table used in classification that summarizes information about the overall performance of the model's predictions. This table contains a recapitulation of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) generated by the classifier. Accuracy, precision, recall, and F1 score are determined from the prediction outcomes as shown in equations 2, 3, 4 and 5.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

(3)

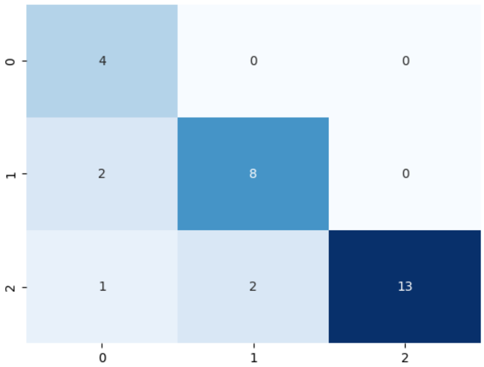
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

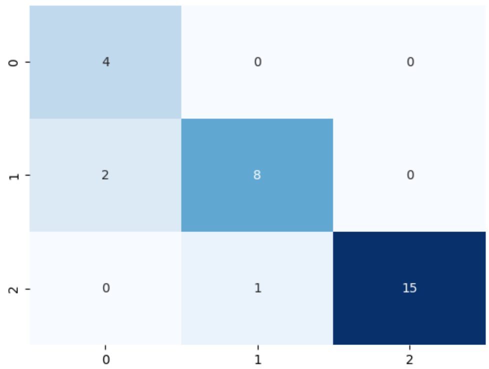
$$F1\ score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \tag{5}$$

Results and Discussion

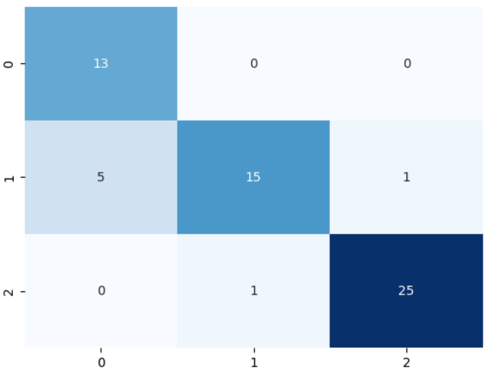
Classification of Gerga oranges using 2 classifiers has been completed and the accuracy, precision, recall, and F1 score values for both have been obtained. These values are calculated based on predictions summarized with the confusion matrix. The accuracy, precision, recall, and F1 score on NB models were increased as the amount of training data decreased. Meanwhile, KNN models showed the opposite results. All confusion matrices can be seen in Figure 2, the matrices were created based on unseen testing data. There were 25% testing data for NB-75 and KNN-75, 50% testing data for NB-50 and KNN-50, and 75% testing data for NB-25 and KNN-25. By evaluating models on this unseen testing data, the results can reflect the model's generalization capabilities for real-world classification tasks. The matrices can explain that values change when the amount of testing data changes. The changes in all values are detailed in Table 2.



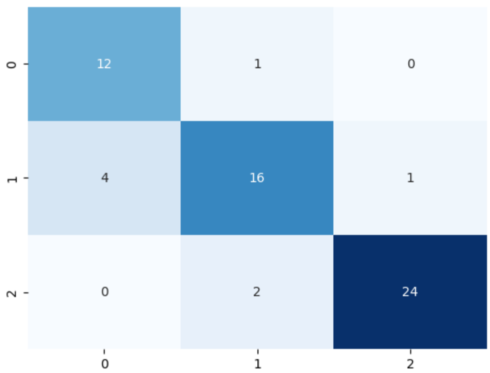
(a)



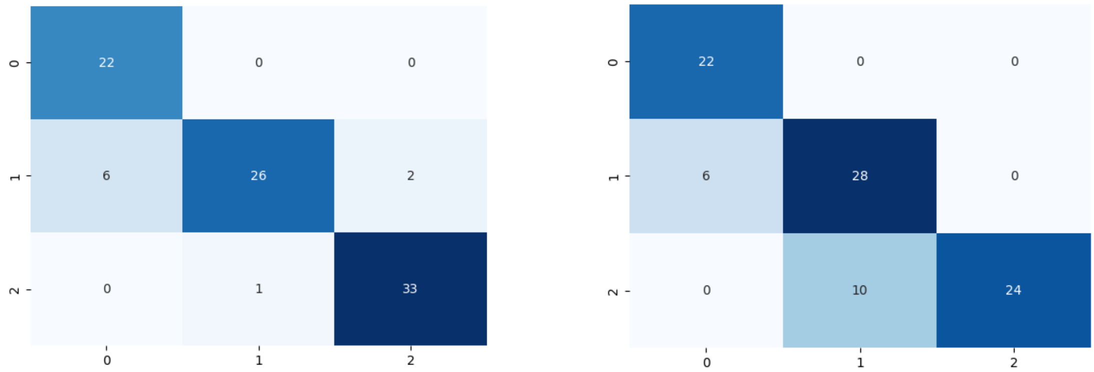
(b)



(c)



(d)



(e) **(f)**

Figure 2. Confusion matrices of all tests for unseen testing data (a) NB-75 (b) KNN-75 (c) NB-50 (d) KNN-50 (e) NB-25 (f) KNN-25

Table 3. Accuracy and precision of models

	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
NB-75	83	79	87	81
NB-50	88	87	89	87
NB-25	90	90	91	90
KNN-75	90	85	91	87
KNN-50	87	85	87	86
KNN-25	82	84	84	83

As can be seen in Figure 2, FP and FN tend to change with the change in the amount of training data. There are only 3 prediction errors in KNN-75 (Figure 2b) while there are 5 prediction errors in NB-75 (Figure 2a). This shows that the accuracy of KNN on a large amount of training data is better than NB. On the same amount of training data and test data, the accuracy of the two models is almost the same, they are about 88% and 87%. However, this is reversed when the training data is only 25%, the accuracy of NB increased to 90% while KNN decreased to 82%.

This phenomenon can occur because when the amount of training data decreases, NB lacks data to build presumptions, thus increasing its accuracy. This result is in line with the findings in the implementation of the Naïve Bayes algorithm to forecast the study duration of students, NB accuracy increases when the amount of training data decreases [12]. On the other hand, the accuracy of KNN is highly dependent on the number of *k* and data points. The decrease in KNN accuracy probably occurred because of the decrease in the number of data points, so the model lacked data to calculate the distances. Other factors are the influence of outliers and unbalanced datasets [13].

The precision of NB models was in line with their accuracy. NB-50 and NB-25 have almost the same number in both accuracy and precision. The big difference is spotted on NB-75 where its accuracy 83% but the precision 79%. The precision of the KNN model tends to be stable with the increase in training data. Small decrease is only spotted from KNN-50 (85%) to KNN-25 (82%). In terms of precision value consistency, KNN is considered better than NB. Precision includes false positives in the calculation, so the lower the precision value, the greater the possibility of incorrectly estimating high-grade orange as low-grade orange. This will have an impact on economic losses for users if they sell fruits classified by the system. From the simulation results, the KNN model is the right model if the user wants to avoid economic losses.

Moreover, the recall value increases when the amount of training data decreases in the classification model using the NB algorithm. The opposite is seen in the experiment with KNN. The increase in recall value is consistent in the NB-75, NB-50, and NB-25 models, which is 2%. While the decrease in recall value in the KNN model is irregular, which is 4% from KNN-75 to KNN-50 and 3% from KNN-50 to KNN-25. The best model based on recall value is KNN-75, which is 91%. This is based on the greater amount of training data that makes the model able to predict better. Recall includes false negatives in the calculation, this indicates that the lower the recall value, the greater the chance of low-grade oranges being included in the grouping. This can potentially cause dissatisfaction from buyers who buy products classified by the model.

If the user wants metrics that support both things related to precision and recall, then the F1 score value can be considered. In this case, the most reliable model is KNN-75 with an F1 score of 87%. This is because the large amount of training data is more reliable, and the value obtained is quite high. These results were in line with a previous study that compared NB and KNN in apple classification [14]. The KNN model showed better results in terms of both accuracy and precision. It demonstrates more equitable and stable performance across both classes, even after adjusting for the varying significance of each class. KKN, as a distance-based algorithm, can achieve good classification accuracy without requiring a complex neural network.

The results of this study are relevant for the post-harvest handling and processes of Gerga oranges. They can be used to modernize the fruit grading method in producing regions like Kerinci, which is highly subjective and prone to human error. By establishing a quantifiable grading index approach (MxGMD) and validating a robust KNN classification model, this research provides a foundational algorithm for automated sorting and grading systems.

Conclusion

Both machine learning models have been successfully applied in classifying Gerga oranges based on their physical properties. Both models show opposite results when the number of samples for training and testing was varied. The accuracy and precision of the Naïve Bayes model improved as the training data size decreased, likely due to changes in its naive presumptions. In contrast, the accuracy of the K-Nearest Neighbor model declined with less training data, possibly because there were fewer reference points for making predictions. The highest accuracy model in both algorithms is NB-25 and KNN-75 with a 90% accuracy value. Thus, the KNN-75 model can be said to be the most reliable model in this Gerga orange classification. This is because in real practice, the number of oranges to be classified will certainly be very large, so a suitable model is needed that can predict well on such a large sample. A larger dataset is required to assess how changes in the data ratio impact the classification metrics of both Naïve Bayes and K-Nearest Neighbor. These results can be used as a consideration in designing the Gerga orange automated sorting or grading machine in the future. This automation can minimize post-harvest time and increase the economic value of the Gerga oranges.

References

[1] B. P. Statitik, "Annual Fruit Exports by Major Destination Countries, 2012-2023."

[2] W. D. Wahyu, D. Syarif, and A. Yani, "Analysis of Microeconomic , Macroeconomic , and Maqashid Sharia Feasibility of Gerga Orange Agribusiness as a Source of New Economic Growth in Jambi Province , Indonesia," *Ekon. J. Econ. Bus.*, vol. 8, no. 2, pp. 1281–1293, 2024, doi: 10.33087/ekonomis.v8i2.1740.

[3] A. Irawan and R. Yuristia, "The economic feasibility of Gerga ' s orange (Citrus sp .) agribusiness as a new source of economic growth in Bengkulu , Indonesia," *Econ. Agrar. y Recur. Nat.*, vol. 21, no. October, pp. 29–45, 2021.

[4] Badan Pusat Statistik Kerinci, "Annual Fruit and Vegetable Crop Production by Crop Type in Kerinci Regency, 2024." Accessed: Feb. 24, 2026. [Online]. Available: https://kerincikab.bps.go.id/id/statistics-table/3/WXpSVU5uUTBOSEl5WVhGQmVESTVSVnBSVlhWeVVUMDkjMw==/produksi-tanaman-buah---buahan-dan-sayuran-tahunan-menurut-jenis-tanaman-di-kabupaten-kerinci--2024.html?year=2024

[5] N. S. M. Elkaoud and A. M. M. Elglaly, "Development of Grading Machine for Citrus Fruits (Valencia Orange)," *J. Soil Sci. Agric. Eng.*, vol. 10, no. 11, pp. 671–677, 2019, doi: 10.21608/jssae.2019.62864.

[6] S. K. Behera, A. K. Rath, A. Mahapatra, and P. K. Sethy, "Identification, classification & grading of fruits using machine learning & computer intelligence: a review," *J. Ambient Intell. Humaniz. Comput.*, no. March, 2020, doi: 10.1007/s12652-020-01865-8.

[7] P. U. Patil, S. B. Lande, V. J. Nagalkar, S. B. Nikam, and G. C. Wakchaure, "Grading and sorting technique of dragon fruits using machine learning algorithms," *J. Agric. Food Res.*, vol. 4, no. January, p. 100118, 2021, doi: 10.1016/j.jafr.2021.100118.

[8] G. Guillergan, "Naive Bayes Classifier in Grading Carabao Mangoes," *Technium*, vol. 22, pp. 14–32, 2024.

[9] A. Kusuma, D. R. Ignatius, and M. Setiadi, "Tomato Maturity Classification using Naive Bayes Algorithm and Histogram Feature Extraction," *J. Appl. Intell. Syst.*, vol. 3, no. 1, pp. 39–48, 2018.

[10] S. Ghazal, W. S. Qureshi, U. S. Khan, J. Iqbal, N. Rashid, and M. I. Tiwana, "Analysis of visual features and classifiers for Fruit classification problem," *Comput. Electron. Agric.*, vol. 187, no. February, p. 106267, 2021, doi: 10.1016/j.compag.2021.106267.

[11] I. Rasyid, I. Saputra, R. K. S. Suryanegara, M. R. A. Yudianto, and Maimunah, "Classification of Tangerines on Fruit Ripening Levels Using K-Nearest Neighbor Algorithm," in *Prosiding University Research Colloquium*, 2022, pp. 403–409.

[12] I. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, "Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm," *Int. J. Eng. Emerg. Technol.*, vol. 2, no. 1, p. 53, 2017, doi: 10.24843/ijeet.2017.v02.i01.p11.

[13] K. U. Syaliman, E. B. Nababan, and O. S. Sitompul, "Improving the accuracy of k-nearest neighbor using local mean based and distance weight," *J. Phys. Conf. Ser.*, vol. 978, no. 1, pp. 0–6, 2018, doi: 10.1088/1742-6596/978/1/012047.

[14] L. M. C. Padasas, J. C. A. Jr, E. I. Balaba, and J. A. Olaybar, "Performance Assessment of K-Nearest Neighbors and Naive Bayes Classifiers in Apple Quality Prediction for Data-Driven Agriculture," *Insight Comput. Sci.*, vol. 1, pp. 1–9, 2025.