Journal of Applied Engineering and Technological Science
Vol 3(2) 2022 : 126-132

ANALYSIS K-NEAREST NEIGHBOR METHOD IN CLASSIFICATION OF VEGETABLE QUALITY BASED ON COLOR

Purwa Hasan Putra[1*], Muhammad Syahputra Novelan[2], Muhammad Rizki[3]
Faculty of Science and Technology, Universitas Pembangunan Pancabudi[1,2]
Faculty of computer science, Universitas Sumatera Utara[3]
Email: purwahasanputra@dosen.pancabudi.ac.id[1*]

Received : 21 May 2022, Revised: 20 June 2022, Accepted : 24 June 2022
**Corresponding Author*

ABSTRACT
In this research, the process of applying the K-Nearst Neighbor (KNN) method will be carried out, which is a classification method for a collection of data based on the majority of categories and the goal is to classify new objects based on attributes and sample samples from training data. So that the desired output target is close to the accuracy in conducting learning testing. The results of the test of the K-Nearest Neighbor method. It can be seen that from the K values of 1 to 10, the percentage of the results of the analysis of the K-NN method is higher than the results of the analysis of the K-NN method. And from the K value that has been tested, the K 2 value and the K 9 value have the largest percentage so that the accuracy is also more precise. As for the results of testing the K-Nearest Neighbor method in data classification. As for the author's test using a variation of the K value of K-Nearest Neighbor 3,4,5,6,7,8,9. Has a very good percentage of accuracy compared to only K-NN. The test results show the K-Nearest Neighbor method in data classification has a good percentage accuracy when using random data. The percentage of variation in the value of K K-Nearest Neighbor 3,4,5,6,7,8,9 has a percentage of 100%.

Keywords: *Machine Learning, Data Classification, Vegetable Quality*

1. Introduction

This image classification process refers to an artificial intelligence method that focuses on machine learning. Many other methods in machine learning which is used to process classifications include K-Nearest Neighbor and Naïve Bayes Classifier. Classification is the grouping of an object into classes based on the characteristics similarities and differences(Safri, et al., 2018; Bayhaqy, et al., 2018; Yang, et al., 2018; Findawati, et al., 2019).

Models for pattern recognition in domains ranging machine learning has shown great success in building from computer vision over speech recognition and text understanding to Game AI(Bhatt, et al., 2021). In addition to these classical domains, machine learning and in particular deep learning are increasingly important and successful in engineering and the science. These success stories are grounded in the data-based nature of the approach of learning from a tremendous number of examples(Von Rueden, et al., 2019).

In his research on Indoor Localization using the K-Nearest Neighbor (K-NN) and Backpropagation methods get the result that the k-NN method produces better accuracy compared to the Backpropagation method. In his research uses the k-NN method to classify image databases hand-based biometric which is a fingerprint database finger print and finger vein, as well as optimizing the kNN method to get a better percentage(Jain & Lella, 2020; Triguero, et al., 2019).

Using the K-Nearst Neighbor (KNN) method is a method of classifying a set of data based on the majority of categories and the goal is to classify new objects based on attributes and sample samples from training data. So that the desired output target is close to the accuracy in conducting learning testing(Diah, et al., 2019; Abu Alfeilat, et al., 2019).

The X-Means algorithm is an algorithm used for grouping data. The x means algorithm is the development of k-means(Mughnyanti, et al., 2020). X-means clustering is used to solve one of the main weaknesses of Kmeans clustering, namely the need for prior knowledge about the number of clusters (K). In this method, the true value of K is estimated in an unsupervised way and only based on the data set itself(Putra & Novelan, 2020; Ahmed, et al., 2020).

This fact inspired us to work on the traffic flow forecast problem build on the traffic data and models.It is cumbersome to forecast the traffic flow accurately because the data available for the transportation system is insanely huge. In this work, we planned to use machine learning, genetic, soft computing, and deep learning algorithms to analyse the big-data for the transportation system with much-reduced complexity. Also, Image Processing algorithms are involved in traffic sign recognition, which eventually helps for the right training of autonomous vehicles(Gempita, et al., 2019; Tournier, et al., 2019).

Machine learning allows in data classification, this application recognizes patterns in data either with training or without training. The classification of data is called clustering in machine learning. Some examples of clustering algorithms include K-Means, Farthest-First Maximization-Expectation (EM), and others(Farhat, et al., 2020; Naranjo-Torres, et al., 2020).

Machine Learning is a branch of artificial intelligence that has the concept that computers as machines have the ability to adapt to new environments and are able to detect patterns from existing facts. The definition of machine learning is when the machine from the experience of E to the task T and measures the increase in the performance of P, if the performance of the task T is measured by the performance of P, improve the experience of E(Binti Jaafar, et al., 2016).

3. Research Methods

Machine learning allows in data classification, this application recognizes patterns in data either with training or without training. The classification of data is called clustering in machine learning. Some examples of clustering algorithms include K-Means, Farthest-First Maximization-Expectation (EM), and others(Putra, 2021).

The data collected for this research are image files with the format Portable Network Graphics (PNG) obtained using the camera digital. The image that becomes the input is the image of the papaya fruit. The sample data used are 3 data in each sample image, with each having 3 attributes, namely red, green, blue. The dataset is the result of image extraction which will be the data source for the classification of papaya fruit images using the K-Nearest Neighbor method(Yahaya, et al., 2014).

Image processing will be carried out in 2 stages, namely resizing the image and the extraction process which is also described in pseudo-code. The following is a sample image of papaya that will identify the level of maturity. The image used in this sample is the original papaya image which will be resized to a size of 100 x 100 pixels, the goal is to make the classification process easier due to uniformity in image size. The original image and the resized image are shown in Figure 1.

Fig. 1. Image of Papaya

Data classification is the process of sorting and grouping data into various types, forms or other different classes. Data classification enables the separation and classification of data according to data set requirements for various business or personal purposes. It is primarily a data management process.

Group analysis as a method for classifying data into several groups using the method of measuring the size of the association, so that the same data is in one group and data with large differences are placed in other data groups. The input to the group analysis system is a data set and the size similarity between the two data. While the results of the group analysis are a number of groups that form a partition or partition structure of the data set and a general description of each group, which is very important for a deeper analysis of the characteristics contained in the data.

Grouping data must use an approach to find similarities in the data so as to be able to place the data into the right groups. Data grouping will divide the data set into several groups where the similarity in a group is greater when compared to other groups.

There are two learning methods available in the classification model, namely:

a. Eager learning is a learning process on training data intensively so that the model can correctly predict the output class label. Several methods are eager learning, including Neural Network, Bayesian, decision tree, Support Vector Machine.
b. Lazy learning is a learning process without training and only storing the value of the training data for use in the prediction process. Some methods are lazy learning, including: K-Nearest Neighbor, Linear Regression, Fuzzy K-Nearest Neighbor.

The classification process in machine learning has four components, namely:

1. Class
   The guest dependent variable must be in a form that represents the label held by the object.
2. Energy
   The independent variables are represented by the characteristic attributes of the data. For example, salary, attendance, smoking, blood pressure.
3. Training Dataset
   A data set that has both of the above component values used to determine the appropriate class based on energy.
4. Testing the dataset
   A new data set that will be classified with the model that has been created and will be evaluated in the classification accuracy process.

In the classification process, before making predictions, it is necessary to carry out a learning process first. The learning process requires data.

The data required during the classification process consists of two types, namely:

a. Training data or training data is data used in the learning process in the classification process.
b. Test data or testing data is data used in the prediction process in the classification process.

The Euclidean distance is simply the sum of the intensity differences pixel-wise and, consequently, small deformations can result in large Euclidean distances. This paper proposes a new Euclidean distance, which we call the Euclidean Distance Image (IMED). Unlike traditional ones, IMED takes into account the spatial relationships of pixels. Based on three properties that (arguably) an image metric that can intuitively satisfy, we show that IMED is the only Euclidean distance that has this property. Euclidean distance is the distance between points in a straight line. This distance method uses the Pythagorean theorem. And is the distance calculation that is most often used in the machine learning process. The Euclidean Distance formula is the result of the square root of the difference between two vectors.

4. Results and Discussions

In this study, the process of determining the type of data will be carried out on the Bougenville papaya fruit or commonly referred to as the paper papaya, which is an ornamental plant whose existence is quite popular among the public and is widespread in various regions in Indonesia. The data collected for this research are image files in Portable Network Graphics (PNG) format which were obtained using a digital camera. The image that is input is the image of the quality of the papaya fruit. The sample data used are 3 data on each image sample, with

each having 3 attributes, namely red, green, blue. The dataset is the result of image extraction which will be a data source for fruit image classification using the K-Nearest Neighbor method.

The designed application can be implemented in papaya fruit image classification based on the level of maturity with the method that has been included in the algorithm. The analysis of the K-Nearest Neighbor method in vegetable image classification that has been implemented into the application can be described as follows: Before carrying out the test, training data must be prepared in the form of an image extraction dataset that is stored in the database. The following figure shows an application that was developed to extract training images and save the extraction results into a database.

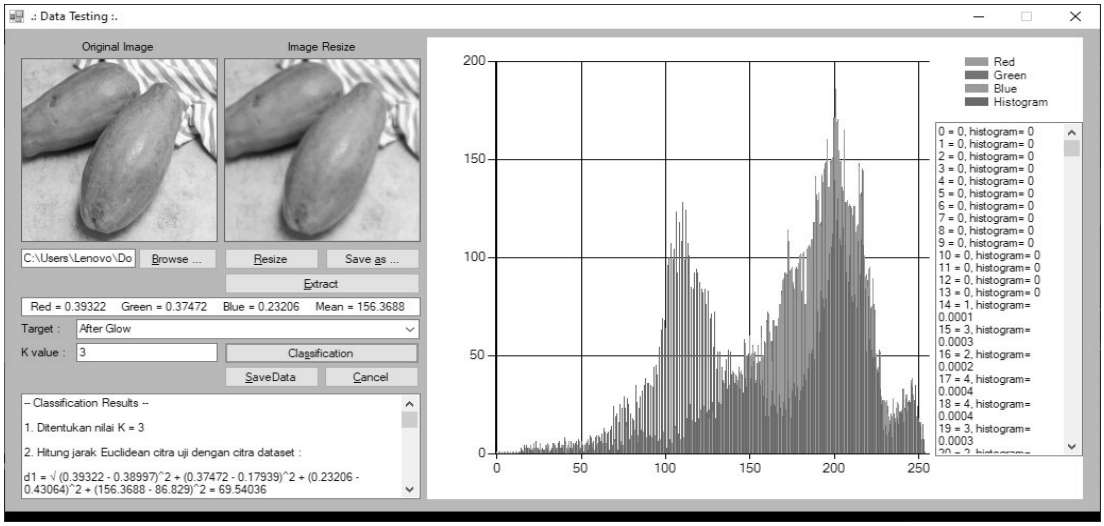

Fig. 2. Image Extraction Train

In the picture above, you can see the image processing process to get the dataset. There is a process of resizing, extracting, and saving data. The resize process is used to change the size or resolution of the image to be uniform, ie in this case 100 x 100 pixels. After the image is resized, it is then extracted to get the attribute values of red, green, blue, hue, saturation, value, mean, variance, skewness, kurtosis, and energy.

The application also displays a graph of the red, green, blue, and histogram values in the image which later these values are processed to get the specified attribute values. The results of the extraction will then be stored in a database by providing a target label according to the color level of the papaya fruit which will later be used as test data.

In this study, the authors analyzed the test with variations in the value of (K) K-NN. From the analysis results show that in this test the author also analyzes the variation in the K value of papaya fruit. It is shown as follows. In this test using 30 test data with 4 attributes and 3 species in the data classification.

Table 1. Results of Variations in K Values K-NN Method

	Total Value (K) K-NN	K-NN Method Analysis Results
Papaya Fruit Data	3, 5, 7, 8, 9	85%
	3, 5, 7, 8, 9	86%
	3, 5, 7, 8, 9	77.3%
	3, 5, 7, 8, 9	77%
	3, 5, 7, 8, 9	81%
	3, 5, 7, 8, 9	73.6%

	3, 5, 7, 8, 9	68%
	3, 5, 7, 8, 9	73%
	3, 5, 7, 8, 9	83%
	3, 5, 7, 8, 9	92%

The analysis from Table 1 presents information on the accuracy level of the K-Nearest Neighbor algorithm specitivity. The analysis was carried out by calculating the number of correct / total data * 100%.

Accuracy is the percentage of the total number of correct predictions in the classification process (Deng et al, 2016). This is done based on the table of Confusion for each class in the Confusion Matrix obtained from the results of training and testing.

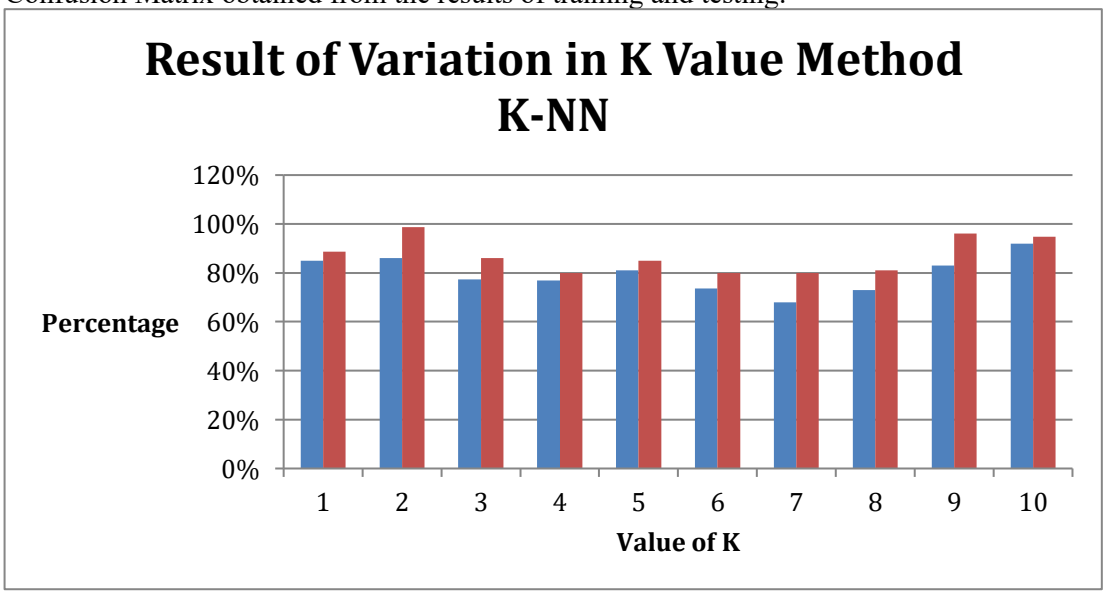

Fig. 3. Testing Results of K-Value Variations in the K-NN Method with 30 Test Data

In Figure 3 above, it can be seen that from the K values of 1 to 10 tested the percentage of the results of the analysis of the K-NN method is higher than the results of the analysis of the K-NN method. And from the K value that has been tested, the K 2 value and the K 9 value have the largest percentage so that the accuracy is also more precise. As for the results of testing the K-Nearest Neighbor method in data classification. As for the author's test using a variation of the K value of K-Nearest Neighbor 3,4,5,6,7,8,9. Has a very good percentage of accuracy compared to only K-NN. The test results show the K-Nearest Neighbor method in data classification has a good percentage accuracy when using random data. The percentage of variation in the value of K K-Nearest Neighbor 3,4,5,6,7,8,9 has a percentage of 100%.

5. Conclusion

The system built can facilitate the K-Nearest Neighbor process to determine performance and increase accuracy in image classification. The results of the test of the K-Nearest Neighbor method. It can be seen that from the K values of 1 to 10, the percentage of the results of the analysis of the K-NN method is higher than the results of the analysis of the K-NN method. And from the K value that has been tested, the K 2 value and the K 9 value have the largest percentage so that the accuracy is also more precise. As for the results of testing the K-Nearest Neighbor method in data classification. The author's test uses variations in the value of K K-Nearest Neighbor 3,4,5,6,7,8,9. Has a very good percentage of accuracy compared to only K-NN. The test results show the K-Nearest Neighbor method in data classification has a good percentage accuracy when using random data. The percentage of variation in the value of K K-Nearest Neighbor 3,4,5,6,7,8,9 has a percentage of 100%.

References

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data, 7*(4), 221-248.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics, 9*(8), 1295.

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018, October). Sentiment analysis about E-commerce from tweets using decision tree, K-nearest neighbor, and naïve bayes. In *2018 international conference on orange technologies (ICOT)* (pp. 1-6). IEEE.

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., ... & Ghayvat, H. (2021). CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics, 10*(20), 2470.

Binti Jaafar, H., binti Mukahar, N., & Ramli, D. A. B. (2016, December). A methodology of nearest neighbor: Design and comparison of biometric image database. In *2016 IEEE Student Conference on Research and Development (SCOReD)* (pp. 1-6). IEEE.

Diah, K. T., Faqih, A., & Kusumoputro, B. (2019, November). Exploring the feature selection of the EEG signal time and frequency domain features for k-NN and weighted k-NN. *In 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129)* (pp. 196-199). IEEE.

Farhat, R., Mourali, Y., Jemni, M., & Ezzedine, H. (2020, February). An overview of Machine Learning Technologies and their use in E-learning. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)* (pp. 1-4). IEEE.

Findawati, Y., Astutik, I. I., Fitroni, A. S., Indrawati, I., & Yuniasih, N. (2019, December). Comparative analysis of Naïve Bayes, K Nearest Neighbor and C. 45 method in weather forecast. In *Journal of Physics: Conference Series (Vol. 1402, No. 6, p. 066046)*. IOP Publishing.

Gempita, G. P., Wilasari, D., Kristalina, P., & Sukaridhoto, S. (2019, September). Implementation of K-NN fingerprint method on receiving server for indoor mobile object tracking. In *2019 International Electronics Symposium (IES)* (pp. 411-416). IEEE.

Jain, A., & Lella, R. L. (2020, December). Pearson Correlation Coefficient Based Attribute Weighted k-NN for Air Pollution Prediction. In *2020 IEEE 17th India Council International Conference (INDICON)* (pp. 1-8). IEEE.

Mughnyanti, M., Efendi, S., & Zarlis, M. (2020). Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. In *IOP Conference Series: Materials Science and Engineering (Vol. 725, No. 1, p. 012128)*. IOP Publishing.

Naranjo-Torres, J., Mora, M., Hernández-García, R., Barrientos, R. J., Fredes, C., & Valenzuela, A. (2020). A review of convolutional neural network applied to fruit image processing. *Applied Sciences, 10*(10), 3443.

Safri, Y. F., Arifudin, R., & Muslim, M. A. (2018). K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor. *Sci. J. Informatics,* 5(1), 18.

Tournier, J. D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., ... & Connelly, A. (2019). MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage, 202*, 116137.

Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to

obtain quality data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(2), e1289.

Putra, P. H., & Novelan, M. S. (2020). Analysis Of The Use Of X-Means Method In Grouping Interest And Talent Data Students. *Jurnal Ipteks Terapan*, *14*(2), 152-159.

Putra, P. H. (2021). Application Of The K-Means Algorithm In Identifying Types Of Skin Disease1. *Infokum*, *9*(2, June), 281-286.

Von Rueden, L., Mayer, S., Garcke, J., Bauckhage, C., & Schuecker, J. (2019). Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning. *Learning*, *18*, 19-20.

Yahaya, O. K. M., MatJafri, M. Z., Aziz, A. A., & Omar, A. F. (2014, August). Non-destructive quality evaluation of fruit by color based on RGB LEDs system. In *2014 2nd International Conference on Electronic Design (ICED)* (pp. 230-233). IEEE.

Yang, C. C., Soh, C. S., & Yap, V. V. (2018). A systematic approach in appliance disaggregation using k-nearest neighbours and naive Bayes classifiers for energy efficiency. Energy Efficiency, 11(1), 239-259.