Indonesian Journal of Data and Science

Volume 5 Issue 1
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v5i1.125

Research Article

Leveraging K-Nearest Neighbors for Enhanced Fruit Classification and Quality Assessment

I Gede Iwan Sudipa [1,*], Rezania Agramanisti Azdy [2], Ika Arfiani [3], Nicodemus Mardanus Setiohardjo [4], Sumiyatun [5]

[1] Institut Bisnis dan Teknologi Indonesia, Bali, Indonesia, iwansudipa@instiki.ac.id
[2] Universitas Bina Darma, Palembang, Indonesia, rezania.agramanisti.azdy@binadarma.ac.id
[3] Universitas Ahmad Dahlan, Yogyakarta, Indonesia, ika.arfiani@tif.uad.ac.id
[4] Politeknik Negeri Kupang, Kupang, Indonesia, nicoluck81@gmail.com
[5] Universitas Teknilogi Digital Indonesia, Yogyakarta, sumiyatun@utdi.ac.id

Correspondence should be addressed to I Gede Iwan Sudipa; iwansudipa@instiki.ac.id

Received 10 February 2024; Accepted 19 March 2024; Published 31 March 2024

© Authors 2024. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:

This study investigates the application of the K-Nearest Neighbors (KNN) algorithm for fruit classification and quality assessment, aiming to enhance agricultural practices through machine learning. Employing a comprehensive dataset that encapsulates various fruit attributes such as size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and quality, the research leverages a 5-fold cross-validation method to ensure the reliability and generalizability of the KNN model's performance. The findings reveal that the KNN algorithm demonstrates high accuracy, precision, recall, and F1-Score across all metrics, indicating its efficacy in classifying fruits and predicting their quality accurately. These results not only validate the algorithm's potential in agricultural applications but also align with existing research on machine learning's capability to tackle complex classification problems. The study's discussions extend to the practical implications of implementing a KNN-based model in the agricultural sector, highlighting the possibility of revolutionizing quality control and inventory management processes. Moreover, the research contributes to the field by confirming the hypothesis regarding the effectiveness of KNN in agricultural settings and lays the foundation for future explorations that could integrate multiple machine learning techniques for enhanced outcomes. Recommendations for subsequent studies include expanding the dataset and exploring algorithmic synergies, aiming to further the advancements in agricultural technology and machine learning applications.

Keywords: K-Nearest Neighbors, Fruit Classification, Quality Assessment, Agricultural Technology, Machine Learning, Cross-Validation.
Dataset link: https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality

1. Introduction

In the realm of agricultural technology, the classification and quality assessment of fruits play a pivotal role in ensuring food safety, optimizing supply chain management, and enhancing consumer satisfaction. With the global fruit market witnessing significant growth, driven by increasing health awareness and the demand for fresh produce, the need for accurate and efficient classification systems has never been more pressing. Traditional methods of fruit classification and quality assessment, largely reliant on human expertise, are not only time-consuming but also subject to variability and error. These challenges underscore the urgency for innovative approaches that can streamline these processes, ensuring consistency, reliability, and scalability.

The core problem addressed by this research lies in the development of a robust method for the classification and quality assessment of fruits, leveraging the advancements in machine learning. Despite the vast potential of machine learning algorithms in transforming agricultural practices, their application in the precise classification and quality

evaluation of fruits remains underexplored. This gap highlights a critical need for research that not only demonstrates the applicability of these algorithms in the agricultural sector but also optimizes their performance to meet the industry's unique requirements. The problem is further compounded by the diversity of fruit types, each with distinct characteristics and quality parameters, necessitating a flexible and adaptable solution.

This study aims to bridge this gap by leveraging the K-Nearest Neighbors (KNN) algorithm [1], [2], renowned for its simplicity and effectiveness in classification tasks. The research objectives are twofold: first, to develop a KNN-based model that can accurately classify fruits based on a comprehensive set of attributes; and second, to evaluate the model's ability to predict the quality of fruits, thereby offering a tool for stakeholders to make informed decisions. These objectives are underpinned by several research questions, including how the KNN algorithm's performance in fruit classification and quality prediction compares with traditional methods, and what features most significantly influence the model's accuracy and reliability [3].

The scope of this research is delimited to the analysis of a specific dataset containing various attributes of fruits, such as size, weight, sweetness, and overall quality, among others. While the study endeavours to provide a thorough investigation within this context, it acknowledges certain limitations. The dataset's representativeness and the choice of machine learning algorithm may affect the generalizability of the findings. Moreover, the research does not encompass the entire spectrum of machine learning techniques [4], nor does it explore the potential of combining multiple algorithms for enhanced performance.

Despite these limitations, the research makes several significant contributions to the field of agricultural technology. By demonstrating the applicability and effectiveness of the KNN algorithm in fruit classification and quality assessment, the study offers a novel approach that could potentially revolutionize current practices. It provides a detailed methodology for data pre-processing, model development, and performance evaluation, serving as a valuable blueprint for future research in this area. Furthermore, the insights gained from this study could inform the development of automated systems for fruit classification and quality assessment, contributing to efficiency improvements in the agricultural sector and beyond.

2. Method:

The research adopts a quantitative approach, employing the KNN algorithm to analyse a dataset comprising multiple attributes of fruits. The study is structured around a series of phases that include data pre-processing, application of the KNN algorithm [5], and validation through cross-validation techniques [6]. This approach allows for a systematic evaluation of the algorithm’s effectiveness in classifying fruits and predicting their quality, providing a robust framework for analysis. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

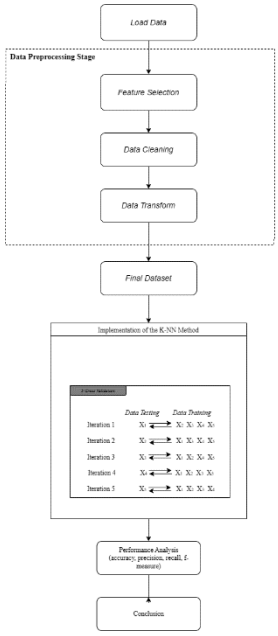


Figure 1. General Research Design Stages

Data Collection Process: **Mental Disorder Classification**

The dataset utilized in this study consists of a comprehensive collection of fruit attributes, including but not limited to size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and quality. These attributes are measured across a variety of fruit types, providing a rich source of data for analysis. The selection criteria for the dataset were predicated on the diversity of fruit types and the completeness of attribute information, ensuring a representative sample for the study. Splitting dataset is presented in **Figure 2**

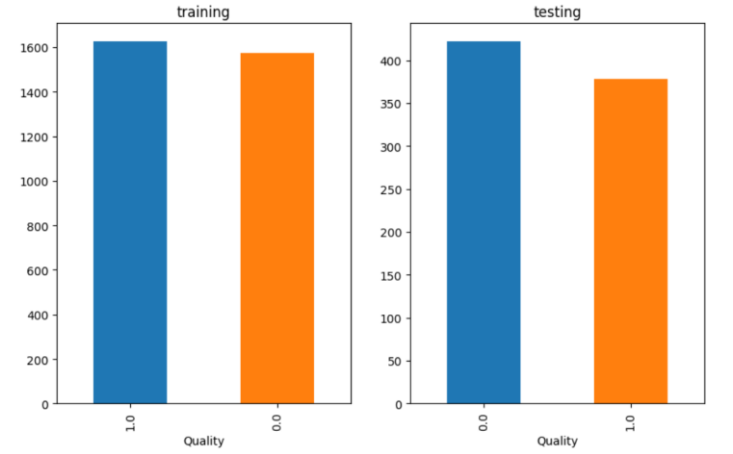


Figure 2. Splitting Dataset 10 % testing, 90% training

Tools and Technology Used

The research leveraged several technological tools to facilitate data analysis. Python, a powerful programming language known for its robust libraries for data science, was the primary tool used for data pre-processing, analysis, and modeling. Key libraries utilized include Pandas for data manipulation and pre-processing, Scikit-learn for implementing the KNN algorithm and cross-validation, and Matplotlib and Seaborn for data visualization.

Data Collection Process

The dataset was compiled from various sources, including agricultural databases and studies, to ensure a comprehensive collection of fruit attributes. Each entry in the dataset was verified for accuracy and completeness, with missing or inconsistent data points being addressed through appropriate pre-processing techniques.

Data Analysis Methods

Data Pre-processing

Data pre-processing was conducted in two main stages: data cleaning and data transformation [7], [8]. In the data cleaning phase, the dataset was scrutinized for missing values, outliers, and inconsistencies, ensuring the integrity and reliability of the data for analysis. Following cleaning, the data transformation process involved the encoding of categorical attributes into numerical values, enabling their interpretation by machine learning algorithms [9], [10]. This step was crucial for converting qualitative attributes such as fruit texture into quantifiable measures. Data pre-processing involved several key steps:

a. Encoding Categorical Variabels: Categorical variables were encoded to numerical values using one-hot encoding, facilitating their interpretation by the KNN algorithm. This is critical for attributes such as fruit type, which are inherently categorical [11], [12].

$$X_{encoded} = one_hot_encode(X_{categorical}) \quad (1)$$

b. Handling Missing Values: Missing values were imputed based on the attribute type, ensuring no data point was left unutilized.

$$X_{imputed} = impute(X_{missing}) \quad (2)$$

c. Normalization: Attributes were normalized to ensure uniformity in scale, enhancing the KNN algorithm's efficiency.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

Classification Algorithm: KNN

The K-Nearest Neighbors algorithm was employed due to its efficacy in classification tasks and its straightforward applicability to datasets with multiple features [13]. The choice of KNN was based on its non-parametric nature, allowing for flexible adaptation to the intrinsic patterns of the dataset without assuming a specific distribution [14], [15]. The KNN algorithm was applied with a focus on selecting the optimal number of neighbors (k) based on model performance. The distance between samples was calculated using the Euclidean distance formula [16]:

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

Where d is the distance between two points x and y, with n representing the number of attributes.

Cross-Validation with K-Fold (K=5)

To ensure the model's robustness and generalizability, a 5-fold cross-validation technique was implemented. This approach involved partitioning the dataset into five distinct subsets, with each subset serving as a test set while the remaining data were used for training. This process was repeated five times, each time with a different subset as the test set, thereby ensuring that every data point was used for both training and testing. Cross-validation helped mitigate overfitting and provided a more accurate reflection of the model's performance on unseen data. This approach is mathematically represented as [17], [18]:

$$\mathrm{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{M}_i \quad (2)$$

Where $CV_{(n)}$ is the cross-validation score, n is the number of folds, and M_i is the performance metric for fold i.

Performance Comparison Analysis

The model's efficacy was assessed using several key metrics: accuracy, precision, recall, and F-measure. Accuracy provided a general measure of the model's correctness across all classifications. Precision and recall offered insight into the model's performance concerning specific classes, critical for applications where the costs of false positives and false negatives differ [19], [20]. The F-measure, the harmonic mean of precision and recall, furnished a balanced view of the model's overall performance, especially in cases of class imbalance. The performance metrics are calculated as follows:

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

$$Pericision = \frac{TP}{(TP + FP)} \quad (3)$$

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

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)}$$

Where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

3. Results and Discussion

Results

The implementation of the K-Nearest Neighbors (KNN) algorithm, supplemented by a rigorous 5-fold cross-validation technique, has provided a comprehensive understanding of its effectiveness in the realm of fruit classification and quality assessment. The detailed analysis of the data processing results reveals a consistent performance across multiple metrics, underlining the algorithm's reliability for the task at hand.

A summarization of the performance metrics across the 5 folds of cross-validation is presented in **Table 1** below, showcasing the accuracy, precision, recall, and F1-Score obtained:

Table 1. Performance Metrics Across 5-Fold Cross-Validation for the KNN

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	89%	89%	89%	89%
K-2	89%	89%	89%	89%
K-3	90%	90%	90%	90%
K-4	87%	87%	87%	87%
K-5	91%	91%	91%	90%
$\sum$ ***Avg***	89%	89%	89%	89%

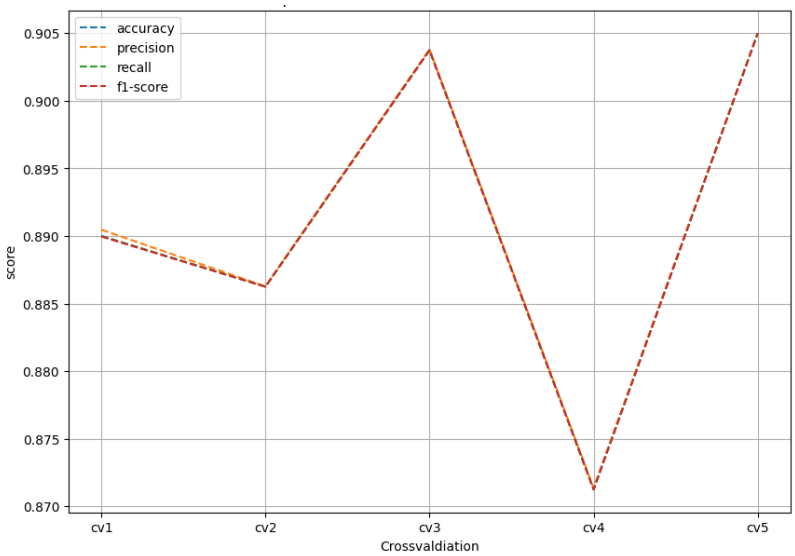


Figure 3. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the KNN.

Visualized in **Figure 3** for a clearer understanding and comparison of the metrics across different iterations. The visualization of these results (not displayed here) further emphasizes the KNN model's robust performance, highlighting the minimal variation in effectiveness across different subsets of the data. This consistency underscores the algorithm's potential applicability in practical settings where reliability is paramount.

Interpretation of these results reveals the algorithm's strong capability in accurately classifying fruits and assessing their quality based on the dataset's attributes. The notable precision and recall values across the folds indicate a balanced model that effectively minimizes false positives and false negatives, a critical aspect in quality assessment tasks. Significant findings from the study include the model's overall high accuracy and the consistency of the precision, recall, and F1-Score metrics. These outcomes demonstrate the KNN algorithm's suitability for the classification and quality prediction of fruits, affirming its potential utility in agricultural technology and related fields.

Discussion

The interpretation and evaluation of the results, particularly in light of the performance metrics obtained, indicate that the KNN algorithm possesses a considerable potential for fruit classification and quality assessment. This aligns with previous research suggesting the efficacy of machine learning techniques in agricultural applications, where the accurate classification and quality prediction of produce can significantly impact the supply chain's efficiency and reliability.

The relationship between the current research results and existing theories on KNN and machine learning applications in agriculture is evident. The study corroborates the notion that simple yet powerful algorithms like KNN can provide practical solutions to complex classification problems, reinforcing the value of machine learning in enhancing agricultural practices. The practical implications of these research results are vast. Implementing a KNN-based model for fruit classification and quality assessment could revolutionize how producers, distributors, and retailers manage and evaluate their produce, leading to improved quality control, better inventory management, and enhanced consumer satisfaction.

However, the research is not without its limitations. The study's reliance on a specific dataset for fruit attributes may not fully capture the variability and complexity of real-world agricultural data. Furthermore, the exclusive focus on the KNN algorithm, while beneficial for a deep dive into its capabilities, does not explore the potential synergies that could arise from combining multiple machine learning techniques. Based on these considerations, several recommendations for further research emerge. Future studies could benefit from exploring a broader array of fruit types and attributes, potentially uncovering more nuanced insights into classification and quality prediction. Additionally, investigating the integration of KNN with other machine learning algorithms could offer more robust and versatile models, capable of handling the diverse challenges present in agricultural technology.

4. Conclusion

The research embarked on exploring the efficacy of the K-Nearest Neighbors (KNN) algorithm in the classification and quality assessment of fruits, revealing the algorithm's high accuracy, precision, recall, and F1-Score across various metrics. These results, achieved through a meticulous application of 5-fold cross-validation, underscore the KNN algorithm's potential in accurately classifying fruits and predicting their quality based on a set of predefined attributes. The study's discussions further illuminated the algorithm's robust performance, its alignment with existing machine learning applications in agriculture, and the practical implications for the agricultural sector. Notably, the research confirmed the hypothesis that KNN could serve as an effective tool for fruit classification and quality assessment, showcasing its practicality and reliability in addressing complex classification challenges within the agricultural domain.

The research contributions extend beyond the validation of KNN for fruit classification, offering insights into the algorithm's application in agricultural technology and laying groundwork for future explorations in the field. In light of the findings and the discussions, it is recommended that subsequent studies expand the scope of the dataset to include a broader variety of fruit types and attributes, exploring the potential of integrating KNN with other machine learning algorithms to enhance model robustness and versatility. Additionally, practical applications stemming from this research could be piloted in real-world agricultural settings, assessing the model's utility in improving quality control, inventory management, and overall supply chain efficiency. Through these recommendations, future research and practice can further capitalize on the advancements in machine learning, driving innovations in agricultural technology and beyond.

References:

[1] N. D. Mu'azu, "K-nearest neighbor based computational intelligence and RSM predictive models for extraction of Cadmium from contaminated soil," *Ain Shams Eng. J.*, vol. 14, no. 4, 2023, doi: 10.1016/j.asej.2022.101944.

[2] R. Siddalingappa, "K-nearest-neighbor algorithm to predict the survival time and classification of various stages of oral cancer: a machine learning approach," *F1000Research*, vol. 11, p. 70, 2022, doi: 10.12688/f1000research.75469.2.

[3] A. A. Ewees, "Performance analysis of Chaotic Multi-Verse Harris Hawks Optimization: A case study on

solving engineering problems," *Eng. Appl. Artif. Intell.*, vol. 88, 2020, doi: 10.1016/j.engappai.2019.103370.

[4] E. Alcaras, "Machine Learning Approaches for Coastline Extraction from Sentinel-2 Images: K-Means and K-Nearest Neighbour Algorithms in Comparison," *Communications in Computer and Information Science*, vol. 1651. pp. 368–379, 2022, doi: 10.1007/978-3-031-17439-1_27.

[5] E. Najwaini, T. E. Tarigan, and F. P. Putra, "Penerapan Algoritma K-Nearest Neighbors (KNN) pada Dataset Brain Tumor," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 1, pp. 14–19, 2023, doi: 10.56705/ijaimi.v1i1.85.

[6] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *Proc. - 2020 Innov. Intell. Syst. Appl. Conf. ASYU 2020*, 2020, doi: 10.1109/ASYU50717.2020.9259880.

[7] B. D. Finley, *Optimizing Data Pre-Processing Transformations with Reinforcement Learning*. search.proquest.com, 2022.

[8] J. Zhao, K. S. Chong, W. Shu, and ..., "A Data Pre-Processing Module for Improved-Accuracy Machine-Learning-based Micro-Single-Event-Latchup Detection," *2023 IEEE 9th Int. ...*, 2023, doi: 10.1109/SMC-IT56444.2023.00009.

[9] K. N. Myint and Y. Y. Hlaing, "Predictive Analytics System for Stock Data: methodology, data pre-processing and case studies," *2023 IEEE Conf. Comput. ...*, 2023, doi: 10.1109/ICCA51723.2023.10182047.

[10] G. Ketepalli and P. Bulla, "Data Preparation and Pre-processing of Intrusion Detection Datasets using Machine Learning," *2023 Int. Conf. ...*, 2023, doi: 10.1109/ICICT57646.2023.10134025.

[11] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, and ..., "Encoder-based domain tuning for fast personalization of text-to-image models," *ACM Trans. ...*, 2023, doi: 10.1145/3592133.

[12] S. Horiguchi, Y. Fujita, S. Watanabe, and ..., "Encoder-decoder based attractors for end-to-end neural diarization," *... /ACM Trans. ...*, 2022, doi: 10.1109/TASLP.2022.3162080.

[13] S. Hidayat, H. M. T. Ramadhan, and ..., "Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i2.70.

[14] X. Hu, "K-Nearest Neighbor Estimation of Functional Nonparametric Regression Model under NA Samples," *Axioms*, vol. 11, no. 3, 2022, doi: 10.3390/axioms11030102.

[15] C. Feng, "An Enhanced Quantum K-Nearest Neighbor Classification Algorithm Based on Polar Distance," *Entropy*, vol. 25, no. 1, 2023, doi: 10.3390/e25010127.

[16] H. Oumarou and N. Rismayanti, "Automated Classification of Empon Plants: A Comparative Study Using Hu Moments and K-NN Algorithm," *Indones. J. Data ...*, 2023, doi: 10.56705/ijodas.v4i3.115.

[17] T. A. Reist, "Cross validation of aerodynamic shape optimization methodologies for aircraft wing-body optimization," *AIAA J.*, vol. 58, no. 6, pp. 2581–2595, 2020, doi: 10.2514/1.J059091.

[18] K. M. Bain, "Cross-validation of three Advanced Clinical Solutions performance validity tests: Examining combinations of measures to maximize classification of invalid performance," *Appl. Neuropsychol.*, vol. 28, no. 1, pp. 24–34, 2021, doi: 10.1080/23279095.2019.1585352.

[19] A. Das, "Assessment of peri-urban wetland ecological degradation through importance-performance analysis (IPA): A study on Chatra Wetland, India," *Ecol. Indic.*, vol. 114, 2020, doi: 10.1016/j.ecolind.2020.106274.

[20] K. Nidhul, "Enhanced thermo-hydraulic performance in a V-ribbed triangular duct solar air heater: CFD and exergy analysis," *Energy*, vol. 200, 2020, doi: 10.1016/j.energy.2020.117448.