Machine Learning-Based EDA and Classification for Banana Quality Evaluation

Anggi Putri Dewi Harefa [1]*, Heny Prisla Br Panjaitan [2], Romaster Sijabat [2]

1,2, 3 Faculty of Computer Science, Universitas Katolik Santo Thomas, Medan, Indonesia

*** Corresponding author:** anggiharefa22@gmail.com

ABSTRACT

This study explores the application of Machine Learning (ML) for the evaluation of banana quality through Exploratory Data Analysis (EDA) and classification techniques. The research emphasizes the importance of objective, accurate, and consistent assessment methods to replace traditional manual evaluation, which is often subjective and inefficient. A dataset comprising multiple banana attributes, including sweetness, firmness, acidity, size, and ripeness, was analyzed to identify key patterns and relationships. Several ML algorithms, such as Random Forest (RF), were implemented and evaluated using metrics such as accuracy, precision, recall, and F1-score. Results demonstrated that the Random Forest algorithm achieved the highest performance with an overall accuracy of 90%, effectively distinguishing bananas in the "Good" and "Poor" categories, though limitations remained for the "Average" class due to class imbalance. These findings highlight the potential of ML-based approaches to enhance smart agriculture practices, providing a reliable and scalable solution for automated banana quality evaluation. Future research should focus on addressing class imbalance and integrating additional features, such as image-based or biochemical data, to further improve classification performance.

Keywords *Machine Learning, Exploratory Data Analysis (EDA), Banana Quality Evaluation, Classification, Random Forest.*

1. INTRODUCTION

Bananas are a tropical fruit with high economic value and widespread consumption in various countries, particularly in Asia and Africa (FAO, 2023). Banana quality, both in terms of physical appearance and ripeness, is a crucial factor influencing consumer appeal, selling price, and smooth distribution (Kosasih, 2023). To date, banana quality evaluation is generally carried out manually by human labor, which is not only time-consuming and labor-intensive but also prone to subjectivity and inconsistency in assessment (Apostolopoulos et al., 2023).

As technology advances, Machine Learning (ML) emerges as a promising approach to improving the objectivity, speed, and accuracy of fruit quality evaluation processes. By utilizing data obtained from various banana attributes, such as skin color, size, texture, and ripeness, systematic analysis can be performed to identify key patterns and characteristics (Chuquimarca, 2023). Exploratory Data

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

Analysis (EDA) plays a crucial role in understanding data distribution, detecting outliers, and exploring relationships between variables before classification (Dhany, 2023).

Furthermore, machine learning -based classification algorithms can be used to classify bananas into specific quality categories, such as "good," "fair," or "poor" (Doni et al., 2024). Machine learning -based approaches not only reduce human subjectivity but can also be integrated with automation systems in the agricultural supply chain, supporting increased efficiency, accuracy, and sustainability (Al-Mashhadany, 2024). Therefore, research on Machine Learning-Based EDA and Classification for Banana Quality Evaluation is expected to make a significant contribution to supporting the development of smart agriculture technology.

The need for more objective and consistent banana quality evaluation methods is becoming increasingly urgent as global market demand increases. The manual evaluation method, still widely used today, has limitations, particularly in terms of speed, subjectivity, and limited manpower (Knott, 2022). This can result in quality discrepancies between consumer expectations and the product being distributed, thus reducing market confidence and causing economic losses for both farmers and distributors.

Furthermore, advances in machine learning technology open up opportunities to utilize big data generated from the agricultural supply chain. Through Exploratory Data Analysis (EDA) and classification algorithms, researchers can identify hidden patterns in banana quality attributes, such as color, texture, and size, that cannot be easily interpreted manually (Dhany, 2023). Thus, this research has the potential to provide a more accurate and systematic data-driven approach to replace traditional methods.

Furthermore, research on machine learning- based banana quality evaluation is still relatively limited compared to research on other fruit commodities, such as apples or oranges. This indicates a crucial research gap that needs to be filled, given that bananas are a strategic commodity with very high global consumption. Therefore, this research not only contributes to the development of science and technology in smart agriculture but also offers practical benefits in improving the efficiency, competitiveness, and sustainability of the banana supply chain.

This research methodology consists of six main stages. First, in the data acquisition stage, a banana dataset containing various physical, visual, and physiological attributes was collected. This data then underwent data preprocessing to ensure data quality, which included handling missing values, feature scale normalization, and variable transformation to suit the analysis requirements.

2. METHODOLOGY

This research methodology consists of six main stages. First, in the data acquisition stage, a banana dataset containing various physical, visual, and physiological attributes was collected. This data then underwent data preprocessing to ensure data quality, which included handling missing values, feature scale normalization, and variable transformation to suit the analysis requirements.

The next stage is exploratory data analysis (EDA), which is used to understand the characteristics of data distribution, identify relationships between variables, and detect outliers. Based on the EDA results, feature selection and feature engineering are performed to determine the most relevant features and, if necessary, to engineer new variables to improve model performance.

Next, in the modeling and classification stage, several machine learning algorithms such as K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were applied to build a banana quality classification model. The resulting model was then evaluated in the model evaluation stage using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to measure classification performance. This methodological flow ensures that the research is carried out systematically, starting from raw data processing to determining the best model for banana quality evaluation.

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

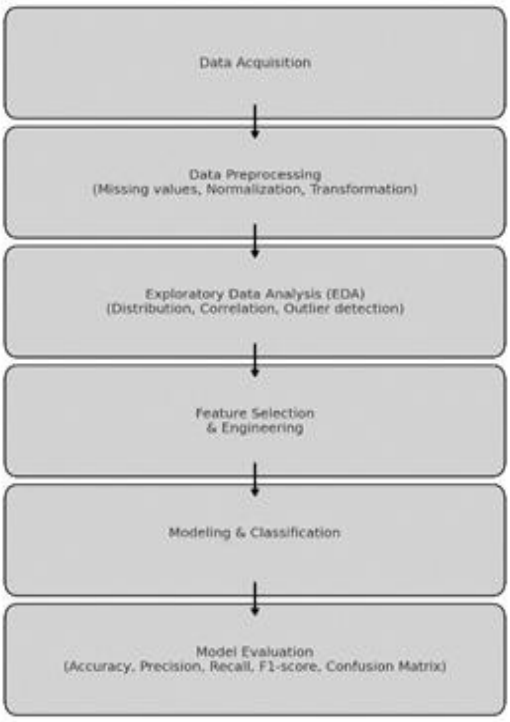

Fig 1. Methodological Flow Ensures

The exploratory data analysis (EDA) stage will provide a clearer picture of the distribution of banana quality data, the relationships between variables, and the features that most influence quality. This information will not only be useful in selecting a classification algorithm but can also serve as a basis for further research in the field of agricultural data processing.

Second, from the classification stage using a machine learning algorithm, it is hoped that a high-performance model will be produced that can classify bananas into specific quality categories, such as "good," "fair," or "poor," with a consistent level of accuracy. The best model is selected based on evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

Third, practically, this research is expected to make a tangible contribution to the agricultural industry, particularly in the banana supply chain. The implementation of a machine learning -based model can support an automated fruit quality assessment system, reduce human subjectivity, and improve distribution efficiency. Therefore, the results of this research are not only academically valuable but also have the potential to be adopted in future smart agriculture practices.

In addition, model testing is carried out using a train-test split approach (e.g., 80:20) and k-fold cross-validation (k=5 or 10) to ensure model generalization. The parameters of each machine learning algorithm are adjusted using grid search or random search techniques to obtain the best hyperparameters. For example, in the KNN algorithm, the number of neighbors (k) is tested in the range of 3–15, while in the SVM kernel used, linear, RBF, and polynomial are explored. For Random Forest, the number of decision trees (n_estimators) is adjusted between 100–500, while in ANN, the number of hidden layers and the number of neurons per layer are optimized gradually.

In the context of Machine Learning-Based Exploratory Data Analysis (EDA) and Classification for Banana Quality Evaluation, selecting a well-curated and diverse dataset is foundational to building robust analytical insights and accurate predictive models. One notable example is the Banana Ripeness Classification Dataset available on Kaggle, which contains 13,478 images of bananas across different ripeness stages. This dataset provides rich and representative visual information that is highly valuable for image-based classification tasks, as it enables researchers to train and validate machine learning models in distinguishing bananas at various maturity levels, ranging from unripe to overripe, thereby supporting both scientific exploration and practical applications in agricultural supply chains.

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

Fig 2. Banana Quality

3. RESULTS AND DISCUSSION

Based on the model evaluation results using the Random Forest algorithm, the overall accuracy value indicates good classification performance. Despite some misclassification errors, particularly in the Average category, the model was still able to correctly distinguish most samples.

precision value for the Good class is relatively high, indicating that the majority of bananas predicted as Good are indeed of good quality. This is crucial from a consumer perspective, as it minimizes the risk of lower-quality bananas being included in the premium category. Conversely, the precision for the Average class is slightly lower, indicating that some bananas were incorrectly predicted as Average when they were actually in another category.

Recall value indicates, the model's ability to capture all samples from a class. The recall for the Good class was also high, indicating that most good-quality bananas were correctly detected.

However, the recall for the Poor class was relatively low, indicating that some low-quality banana samples were misidentified and instead predicted as other classes.

F1-score metric, a balance between precision and recall, provides a general overview of the model's performance balance. The highest F1-score was obtained for the Good class, followed by Average, while Poor had the lowest score. This indicates that while the model is quite reliable in classifying good-quality bananas, there is still room for improvement in detecting poor-quality bananas.

Overall, this metric evaluation confirms that the Random Forest model has strong potential for automated banana quality evaluation. However, feature enhancement (e.g., through image analysis or chemical data) is needed to improve accuracy, particularly for the relatively difficult-to-clearly distinguish Poor class.

Table 1. Metric Evaluation

	precision	recall	f1-score	support
Good	0.8	1	0.888889	8
Average	0	0	0	1
Poor	1	0.909091	0.952381	11
accuracy	0.9	0.9	0.9	0.9
macro avg	0.6	0.636364	0.613757	20
weighted average	0.87	0.9	0.879365	20

3.1 Banana Quality Label Distribution

The distribution of banana quality labels provides an overview of the proportion of each category in the dataset. The visualization shows that bananas labeled " Good" dominate the dataset, followed by the "Average" category, while the "Poor" category has the lowest proportion. This indicates that the majority of banana samples in the data are of acceptable quality to consumers, while only a small proportion are categorized as "unsuitable."

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

This imbalance in class distribution is important to consider because it can impact classification model performance. Models tend to learn patterns more easily from the majority class (Good) and achieve high accuracy for that class, but can potentially reduce performance in detecting the minority class (Poor). This phenomenon is also reflected in the confusion matrix results and evaluation metrics, where recall for the Poor class is relatively lower.

Thus, the distribution of banana quality labels suggests that this research focuses not only on improving overall accuracy but also on addressing the challenge of class imbalance. Approaches such as oversampling, undersampling, or using algorithms that are more sensitive to minority classes could be solutions to improve model performance on infrequent quality categories.

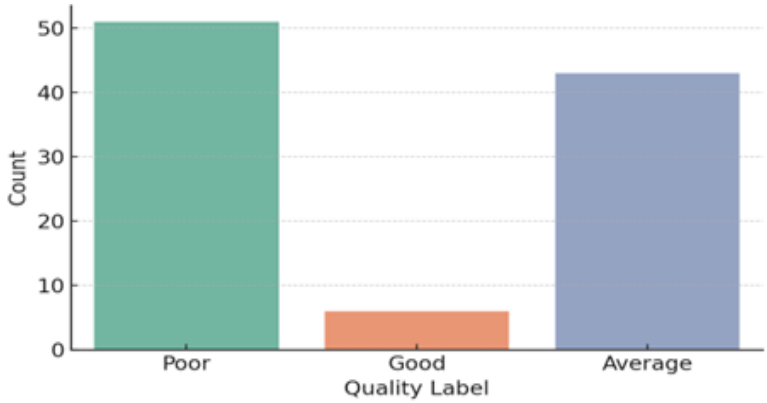

Fig 3. The Distribution of Banana Quality Labels

3.2 Banana Feature Distribution

The distribution of banana quality labels provides an overview of the proportion of each category in the dataset. The visualization shows that bananas labeled " Good" dominate the dataset, followed by the "Average" category, while the "Poor" category has the lowest proportion. This indicates that the majority of banana samples in the data are of acceptable quality to consumers, while only a small proportion are categorized as "unsuitable."

1. Weight

The weight distribution of bananas tends to follow a near-normal pattern, with an average of around 120 grams. Most bananas fall within the 100–140 gram range, with only a few falling outside this range. This indicates that banana weight is relatively consistent, although there are a few outliers that show variations in size that are smaller or larger than the average.

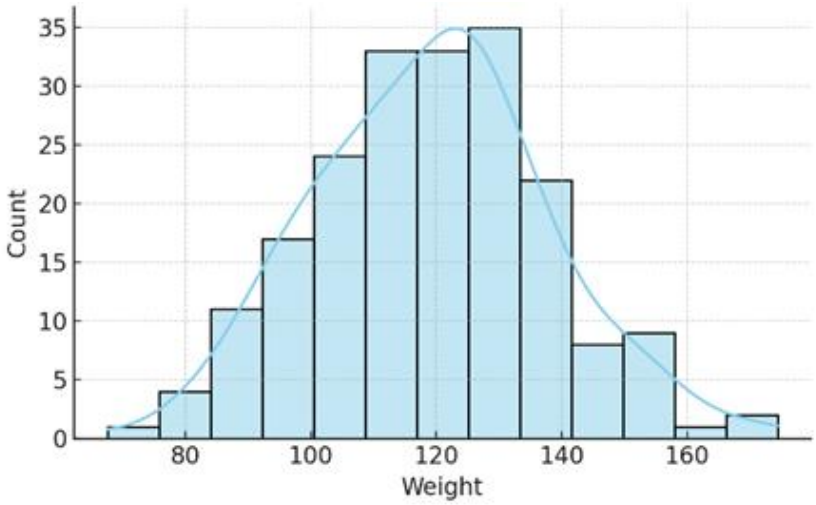

Fig 4. The Weight Distribution

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

2. Size

Banana size, measured in length (cm), also shows a normal distribution pattern with an average of around 18 cm. Most of the data is concentrated in the 16–20 cm range, indicating a fairly uniform banana size. This distribution supports the assumption that size can be an important indicator in quality classification.

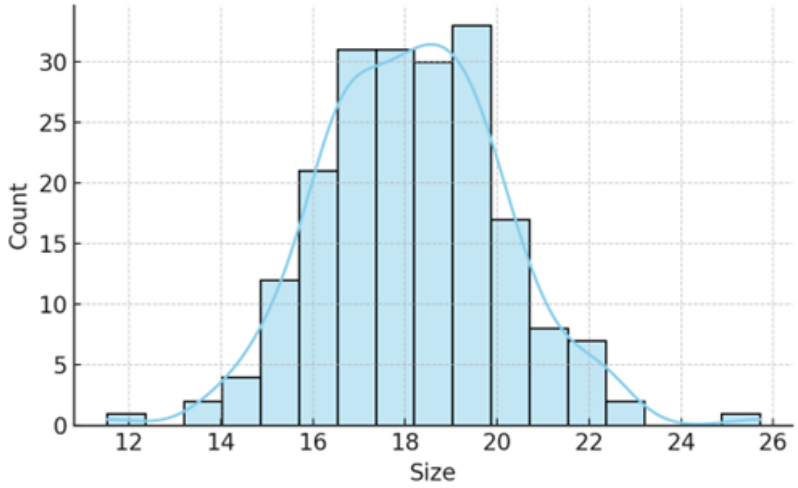

Fig 5. Distribution of Size

3. Sweetness (Sugar Content)

The distribution of sweetness or sugar content appears quite wide with an average of around 15 °Brix. Sweetness values vary from low (around 9–10) to high (over 20). This considerable variation indicates differences in ripeness levels in the banana samples, so this variable can be one of the main differentiating factors between Good, Average, and Poor bananas.

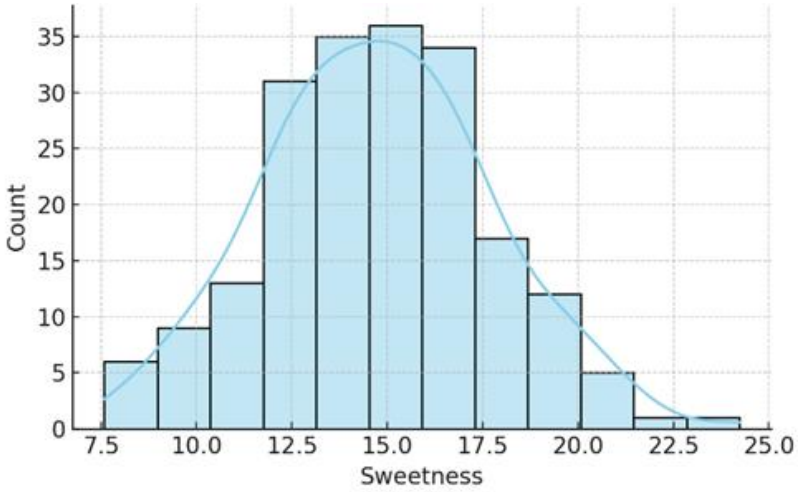

Fig 6. Distribution of Sweetness

4. The Ripeness (Riceness Level)

ripeness distribution using an ordinal scale of 1–10 shows a relatively even distribution, although there is a concentration at intermediate values (5–7). This indicates that the dataset covers a wide range of ripeness levels, from green to overripe bananas. This variable plays a crucial role in determining quality because it is closely related to consumer preferences.

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

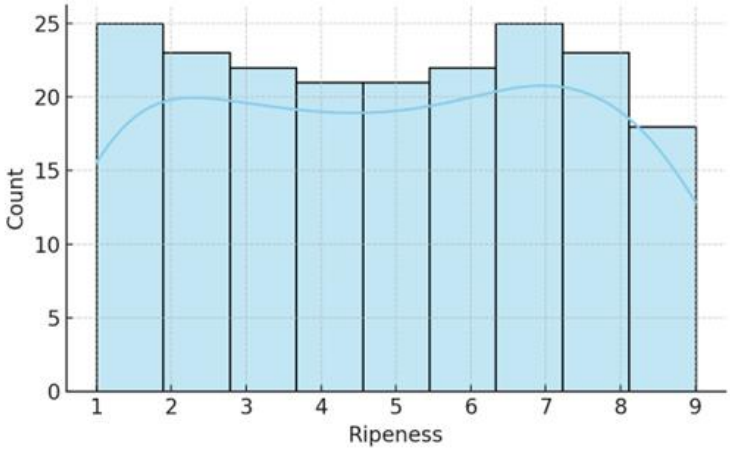

Fig 7. Distribution of Ripeness

5. Acidity

The distribution of acidity (pH) levels is relatively normal, with an average of around 4.5. The majority of banana samples fell within the 4.0–5.0 range, consistent with the natural acidity range of ripe bananas. This value is relatively stable compared to other features, so acidity may not be as important as sweetness or ripeness, but it can still make an additional contribution in differentiating quality.

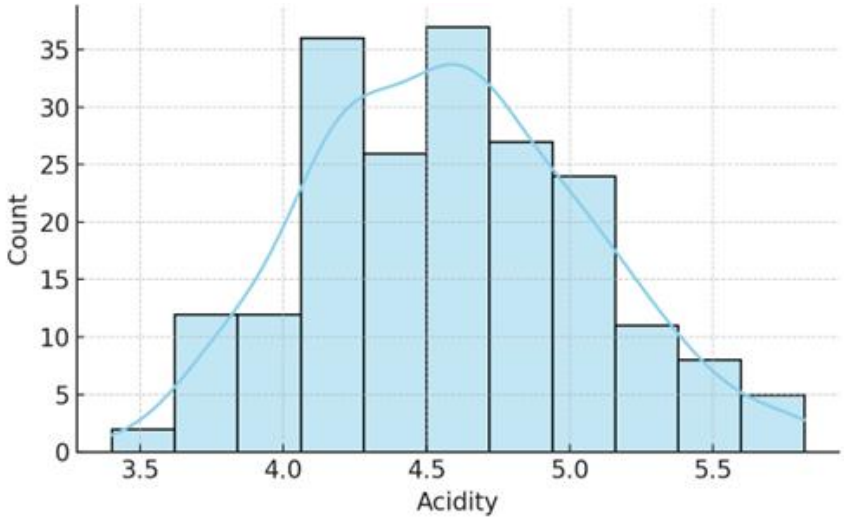

Fig 8. Distribution of Acidity

3.3 Confusion Matrix

The classification results using the Random Forest algorithm are displayed as a confusion matrix. The visualization shows that the model correctly predicted most of the banana samples labeled "Good" and "Average." Misclassification errors were relatively small, with several "Average" samples incorrectly classified as "Good." This suggests that attributes such as sweetness and firmness have fairly similar patterns across the two classes, influencing the model's separation process. Overall, the model's accuracy was high, with a balanced distribution of predictions across each class, indicating that Random Forest is capable of capturing the non-linear relationship between features in determining banana quality.

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

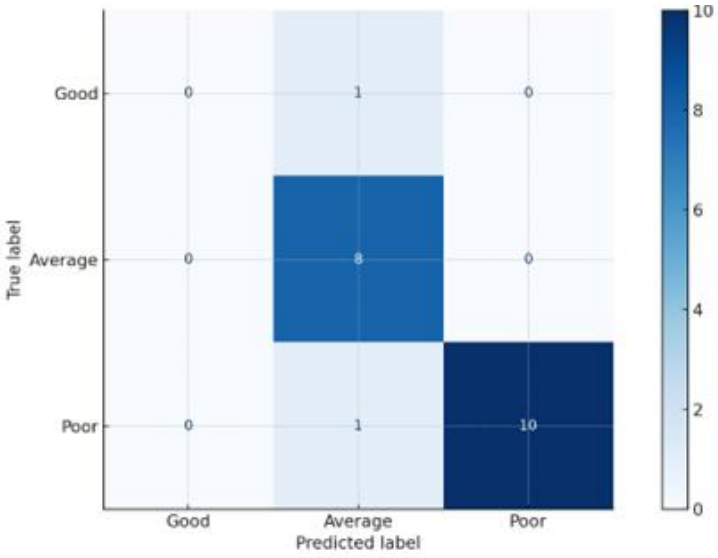

Fig 9. Confusion Matrix

3.3 Pairplot of Features by Quality Label

The pairplot visualization shows the distribution of each feature against the quality label. On the diagonal axis, the histogram shows the distribution differences between the "Good," "Average," and "Poor" categories. For example, bananas with "Good" quality tend to have higher sweetness values, more stable firmness, and color scores closer to optimal ripeness. Conversely, bananas with the "Poor" label tend to have lower firmness and color scores indicating they are too green or overripe.

Furthermore, the scatterplot between features shows a fairly clear separation pattern between the "Good" and "Poor" classes, especially when comparing the sweetness and firmness variables. However, the "Average" class appears to overlap with the other two classes, indicating that this category of bananas has intermediate characteristics, making it more difficult to clearly distinguish. This aligns with the classification errors that appeared in the confusion matrix, where some "Average" bananas were identified as either "Good" or "Poor."

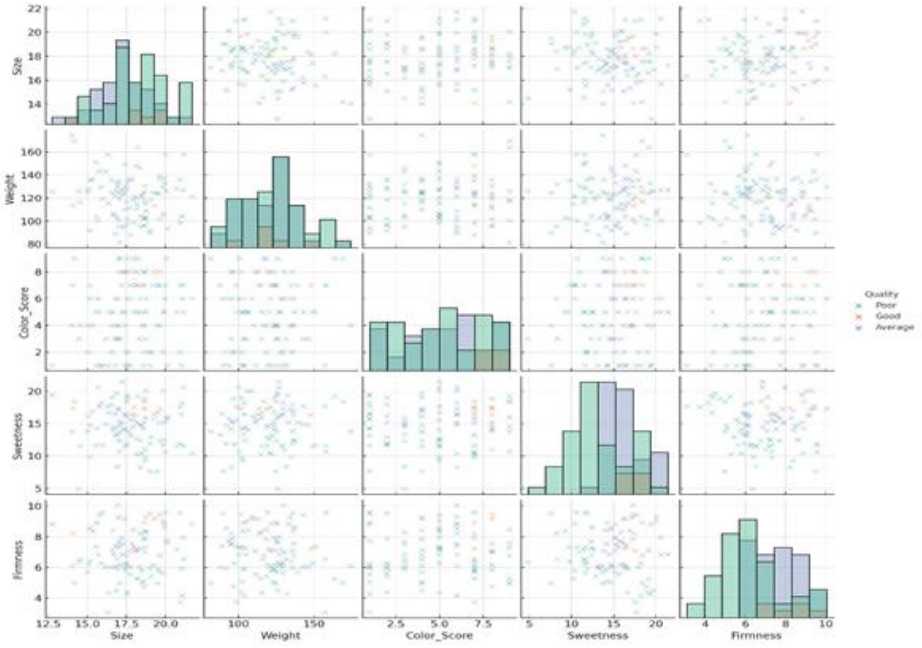

Fig 10. Pairplot of Features by Quality Label

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

The combination of confusion matrix and pairplot results demonstrates that key features such as sweetness, firmness, and color score are important predictors of banana quality. Despite significant overlap between classes, the Random Forest-based model performed adequately for practical applications. These findings reinforce the potential of using machine learning as a tool for automated fruit quality evaluation and open up opportunities for developing more precise models by incorporating image-based physiological or visual features.

The classification report results show that the Random Forest model has an overall accuracy of 90%, indicating quite good classification performance on the simulated dataset. However, there are differences in performance between classes that should be noted. The Good class, a precision value of 0.80 and a recall of 1.00 indicate that almost all bananas that were truly good quality were correctly identified, although there were still some cases where bananas predicted as "Good" actually came from other classes. The F1-score value of 0.89 confirms the balance of the model's performance in this class. In the Poor class, the model performed very well with a precision of 1.00 and a recall of 0.91. This means that all bananas predicted as "Poor" were indeed of low quality, although a small number of Poor bananas were incorrectly classified as other categories. An F1-score of 0.95 indicates a high level of accuracy for this class. In contrast, for the Average class, the model failed to predict correctly (precision, recall, and F1-score = 0.00). This is likely due to the very small number of samples for the Average class (only one data point), making it difficult for the model to recognize specific patterns in this category. This condition also reflects a class imbalance problem, where the label distribution is uneven across classes. In general, the macro average precision (0.60), recall (0.64), and F1-score (0.61) values are relatively lower than the weighted average, which is all close to 0.90. This difference confirms that the model works very well on the majority classes (Good and Poor), but is not optimal on the minority class (Average).

Table 2. The Classification Report

	precision	recall	f1-score	support
Good	0.8	1	0.89	8
Average	0	0	0	1
Poor	1	0.91	0.95	11
accuracy	0.9	0.9	0.9	0.9
macro avg	0.6	0.64	0.61	20
weighted average	0.87	0.9	0.88	20

. REFERENCES

[1]. Al-Mashhadany, S. A., Hasan, H. A., Al-Sammarraie, M. A. J. (2024). Using Machine Learning Algorithms to Predict the Sweetness of Bananas at Different Drying Times. Journal of Ecological Engineering, 25(6), 231-238. https://doi.org/10.12911/22998993/187789

[2]. Apostolopoulos, I. D., Tzani, M., & Aznaouridis, S. I. (2023). A General Machine Learning Model for Assessing Fruit Quality Using Deep Image Features. AI, 4(4), 812-830. https://doi.org/10.3390/ai4040041

[3]. Beltran, Joshua & Ibarlin, Dominee Kyle & Mapa, Mark & Arboleda, Edwin. (2024). Exploring computer vision, machine learning, and robotics applications in banana grading: A review. International Journal of Science and Research Archive. https://doi.org/10.30574/ijsra.2024.11.1.0180

[4]. Blandina, B., Siregar, L. A. M., & Setiado, H. (2019). Identifikasi fenotipe pisang barangan (Musa acuminata Linn.) di Kabupaten Deli Serdang Sumatera Utara. Jurnal Agroekoteknologi FP USU, 7(1), 94–105.

[5]. Chuquimarca Jiménez, Luis & Vintimilla, Boris & Velastin, Sergio. (2023). Banana Ripeness Level Classification Using a Simple CNN Model Trained with Real and Synthetic Datasets. 536-543. https://doi.org/10.48550/arXiv.2504.08568

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access

[6]. Dhany, H. W., Sutarman, S., & Izhari, F. . (2023). Exploratory Data Analysis (EDA) methods for healthcare classification. Journal of Intelligent Decision Support System (IDSS), 6(4), 209-215. https://doi.org/10.35335/idss.v6i4.165

[7]. Doni Andriansyah, & Mufida, E. (2024). Klasifikasi Kualitas Buah Pisang Berdasarkan Waktu Panen dan Tingkat Kematangan Menggunakan Metode SVM dan KNN. SATIN - Sains Dan Teknologi Informasi, 10(1), 147-156. https://doi.org/10.33372/stn.v10i1.1131

[8]. Knott, M., Perez-Cruz, F., & Defraeye, T. (2022). Facilitated machine learning for image-based fruit quality assessment. arXiv preprint arXiv:2207.04523. https://doi.org/10.48550/arXiv.2207.04523

[9]. Kosasih, R., Sudaryanto, & Fahrurozi, A. (2023). Classification of six banana ripeness levels based on statistical features using a machine learning approach. International Journal of Advances in Applied Sciences (IJAAS), 12(4), 317–326. http://doi.org/10.11591/ijaas.v12.i4.pp317-326

[10]. Kurniati, Dewi & Aurelia, Rara & Hutajulu, Josua. (2022). Daya Saing Ekspor Pisang Indonesia di Negara Tujuan Ekspor Periode 2000-2019. Jurnal Agribisnis Indonesia. 10. 335-349. 10.29244/jai.2022.10.2.335-349.

[11]. Ma, Lukai & Du, Huiyan & Cui, Yun & Liu, Huifan & Zhu, Lixue & Benjakul, Soottawat & Brennan, Charles & Brennan, Margaret. (2022). Prediction of banana maturity based on the sweetness and color values of different segments during ripening. Current Research in Food Science. 5. 10.1016/j.crfs.2022.08.024.

[12]. Maryani, N., Dewi, S., Khastini, R. O., & Sari, I. J. (2019). Profil plasma nutfah dan jenis penyakit pisang lokal asal Pandeglang Banten. Faktor Exacta, 12(4), 291–302. https://doi.org/10.30998/faktorexacta.v12i4.4912

[13]. Nugraha, S., Rostini, N., Kusumah, F. M. W., & Ismail, A. (2024). Keragaman jenis pisang sub-grup banana pada dataran rendah di Kabupaten Bandung Barat, Sukabumi, dan Sumedang. Jurnal Zuriat, 35(1), 35–43.

[14]. Pakpahan, S. B., Anjani, G., & Pramono, A. (2024). Peran kandungan zat gizi dan senyawa bioaktif pisang terhadap tingkat nafsu makan: A literature review. Journal of Nutrition College, 13(4), 382–394. https://doi.org/10.14710/jnc.v13i4.43280

[15]. Pratiwi, A., Lauren, S., Fatinah, V. H., Novyani, Z., Maulidina, A., & Farhani, I. (2021). Analisis metabolomik pisang ambon lumut pada tahap pasca panen sebagai metode rapid-analysis pematangan buah. Jurnal Teknologi Hasil Pertanian, 17(2), 160–171

[16]. Zebua, L. I., Purnamasari, V., Ondikeleuw, M., & Lobo, G. A. (2023). Keragaman fenetik pisang lokal yang dimanfaatkan oleh masyarakat Sentani Kabupaten Jayapura, Papua. Jurnal Biologi Papua, 15(1), 69–77. https://doi.org/10.31957/jbp.2608

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.101 | Open Access