Jurnal Minfo Polgan
Volume 14, Nomor 1, Juli 2025
DOI : https://doi.org/10.33395/jmp.v14i1.15088
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Classification Of Pistachio Varieties Using Machine Learning Algorithms

[1]Andysah Putera Utama Siahaan, [2]Muhammad Iqbal, [3]Dika, [4]Maulisa Syahputri
[1,2,3,4]Master of Information Technology, Pembangunan Panca Budi University, Medan, Indonesia
[1]andiesiahaan@gmail.com, [2]muhammadiqbalpb@gmail.com, [3]tiepie71@gmail.com

Submited : Jul 14, 2025 | **Accepted** : Jul 22, 2025 | **Published** : Jul 24, 2025

ABSTRAK

The accurate classification of pistachio varieties plays a crucial role in ensuring quality control, enhancing traceability, and improving market segmentation in the agricultural sector. This study explores the application of various machine learning algorithms—including Decision Tree, Random Forest, XGBoost, Support Vector Classifier (SVC), k-Nearest Neighbors (KNN), and Logistic Regression—for the classification of pistachio types based on morphological features. A publicly available dataset containing measurements such as kernel length, shell width, and aspect ratio was used to train and evaluate the models. The results demonstrated that ensemble methods like XGBoost and Random Forest consistently outperformed other algorithms, achieving accuracy scores of 0.86 and 0.85, respectively, with high Area Under the Curve (AUC) values in the Receiver Operating Characteristic (ROC) analysis. Furthermore, hyperparameter tuning improved model performance across the board. These findings indicate the potential of machine learning as a reliable tool for automating pistachio variety classification and supporting decision-making in agricultural practices. Future research may involve real-time classification using image-based features and integration into precision agriculture systems.

Keywords: Classification models, Machine learning, Pistachio classification, Random Forest, XGBoost, Support Vector Machine.

INTRODUCTION

Pistachios (Pistacia vera L.) are among the most valuable tree nuts globally, known for their distinctive taste, nutritional richness, and growing economic significance in the food and agricultural sectors. Accurate classification of pistachio varieties is essential for various purposes, including cultivar identification, market segmentation, and ensuring product quality in the supply chain. Traditionally, this classification has been carried out manually by experts based on visual features such as shape, size, and shell color, which is not only time-consuming but also highly subjective and inconsistent.

Recent advancements in artificial intelligence, particularly in machine learning (ML), have paved the way for automating the classification of agricultural products, including nuts. ML algorithms are capable of learning patterns from data and making predictions with high accuracy, enabling researchers and industries to classify pistachio types efficiently. In the context of pistachios, datasets containing morphological features such as length, width, area, perimeter, and eccentricity are used as input for training ML models [1].

Several supervised learning models, including Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbors (k-NN), Random Forest, and Neural Networks, have been employed to classify pistachio varieties. These algorithms differ in their approach and computational complexity, and their performance can vary based on the size and quality of the dataset. Studies have shown that ensemble methods like Random Forest often yield higher accuracy due to their ability to reduce variance and overfitting [2].

By leveraging ML-based classification systems, stakeholders in the pistachio industry can significantly enhance their operational workflows. These systems help reduce reliance on manual labor, minimize human error, and allow for real-time classification during post-harvest processing. Moreover, such approaches contribute to the standardization of quality assessment procedures and

Jurnal Minfo Polgan
Volume 14, Nomor 1, Juli 2025
DOI : https://doi.org/10.33395/jmp.v14i1.15088
e-ISSN : 2797-3298
p-ISSN : 2089-9424

support the implementation of smart agriculture technologies [3].

This study aims to evaluate the performance of several machine learning algorithms in the classification of pistachio varieties using a publicly available dataset. The primary objective is to identify the most accurate and computationally efficient algorithm for practical deployment in real-world agricultural settings. By assessing and comparing algorithmic performance metrics, this research contributes to the growing body of knowledge on the application of artificial intelligence in agrifood systems and supports the development of intelligent quality assessment tools for horticultural products.

METHOD

1. Dataset Description

This study utilizes a publicly available pistachio dataset consisting of morphological features extracted from images of two pistachio varieties: Kirmizi and Siit. The dataset comprises 59 samples from each variety, totaling 118 observations. The dataset was sourced from the UCI Machine Learning Repository and was selected due to its balanced class distribution and well-structured numerical attributes (Kaya & Koc, 2020).

2. Data Preprocessing

Prior to model training, the dataset underwent a series of preprocessing steps. First, missing values were checked, although none were found. Then, the data was normalized using Min-Max Scaling to ensure uniformity across feature scales and avoid bias toward features with larger numerical ranges. Following normalization, the dataset was randomly split into training and testing sets using an 80:20 ratio to evaluate model performance on unseen data. Label encoding was applied to convert the categorical target variable into binary format.

3. Model Development

Five machine learning algorithms were selected for evaluation: Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Decision Tree (DT), Random Forest (RF), etc. Each algorithm was implemented using the scikit-learn library in Python. A 10-fold cross-validation approach was used on the training set to optimize model parameters and prevent overfitting. Hyperparameter tuning was performed using Grid Search for each model to identify the optimal configuration that yields the highest classification accuracy.

4. Performance Evaluation

To assess the performance of each classifier, several evaluation metrics were applied, including accuracy, precision, recall, F1-score, and area under the Receiver Operating Characteristic curve (AUC-ROC). The confusion matrix was also computed to analyze the distribution of true and false predictions across the two classes. These metrics provide a comprehensive assessment of each model's predictive capability, especially in distinguishing between similar varieties of pistachios.

5. Implementation Environment

All experiments were conducted on a standard computing environment with an Intel Core i7 processor, 16GB RAM, and Python 3.10. The main libraries used for model implementation and analysis included scikit-learn, pandas, numpy, and matplotlib. The reproducibility of results was ensured by fixing the random seed during data splitting and model training.

RESULTS AND DISCUSSION

The correlation heatmap provides a visual overview of the relationships between all features in the pistachio dataset, including their correlation with the target class. Notably, features such as AREA, MAJOR_AXIS, SOLIDITY, SHAPEFACTOR_1, and CONVEX_AREA show strong positive correlations with the Class variable, indicating that these morphological characteristics play a significant role in distinguishing between pistachio varieties. On the other hand, features like ECCENTRICITY and ASPECT_RATIO exhibit moderate negative correlations with the class label, suggesting that more elongated or irregularly shaped pistachios are associated

Jurnal Minfo Polgan
Volume 14, Nomor 1, Juli 2025
DOI : https://doi.org/10.33395/jmp.v14i1.15088
e-ISSN : 2797-3298
p-ISSN : 2089-9424

with specific types.

Additionally, some features show strong inter-correlations, such as AREA and EQDIASQ or MINOR_AXIS and SHAPEFACTOR_3, which may lead to redundancy in the dataset. Identifying and potentially removing these highly correlated features can improve model efficiency and reduce multicollinearity. Overall, the heatmap analysis supports feature selection and model refinement by highlighting the most influential variables for pistachio classification using machine learning algorithms.

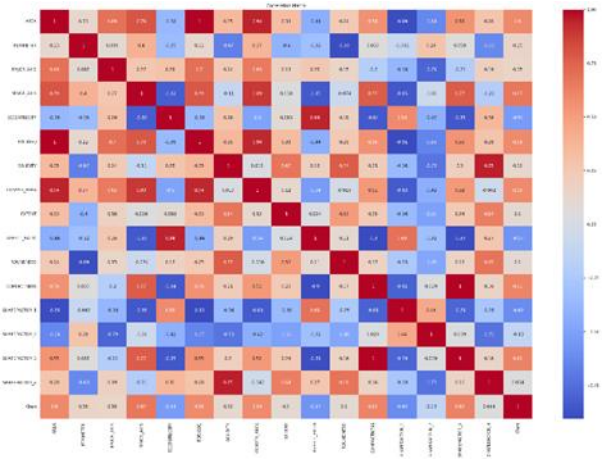

Figure 1. The Correlation Heatmap

The bar chart presents a comparative analysis of evaluation metrics—Accuracy, Precision, Recall, and F1 Score—for various machine learning classifiers applied to pistachio variety classification. Among all models, the Random Forest Classifier and XGBoost Classifier achieved the highest accuracy of 0.86, along with balanced precision, recall, and F1 scores (all around 0.83), indicating strong overall performance and consistency across evaluation metrics. The Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN) also performed well, both reaching accuracies above 0.84, although slightly lower in recall and F1-score compared to the ensemble methods.

In contrast, the Decision Tree Classifier showed the lowest overall performance with an accuracy of 0.79 and F1-score of 0.75, suggesting overfitting or poor generalization. Logistic Regression yielded relatively good precision and recall but achieved a slightly lower accuracy (0.82), likely due to its linear decision boundaries being less effective in capturing complex patterns. These results confirm that ensemble-based methods (like Random Forest and XGBoost) offer superior reliability for this classification task, benefiting from their ability to reduce variance and better handle feature interactions.

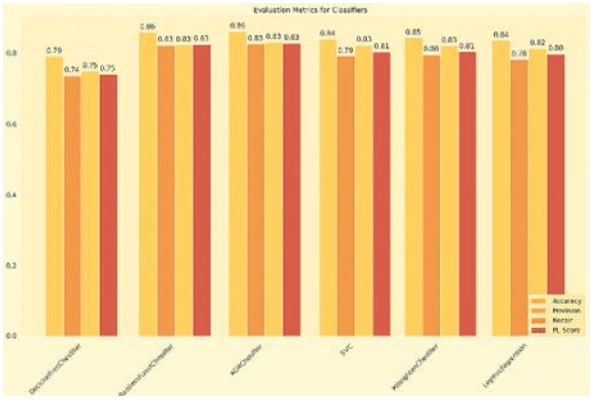

Figure 2. Comparative Analysis of Evaluation Metrics

The bar chart illustrates the evaluation metrics—Accuracy, Precision, Recall, and F1 Score—for several machine learning classifiers used in pistachio variety classification. The

Jurnal Minfo Polgan
Volume 14, Nomor 1, Juli 2025
DOI : https://doi.org/10.33395/jmp.v14i1.15088
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Random Forest Classifier and XGBoost Classifier again stand out with the highest accuracy scores, 0.85 and 0.86 respectively, while maintaining balanced precision and recall values (above 0.80), which indicates strong generalization capabilities. The Support Vector Classifier (SVC) also performs well with consistent scores across all metrics, particularly high recall and F1 score, making it a reliable choice for this binary classification task.

In contrast, the Decision Tree Classifier shows relatively lower recall (0.70), despite achieving an accuracy of 0.83, suggesting that it may be prone to misclassifying one of the pistachio classes. Notably, one instance of SVC shows a very low recall value (0.10) while keeping accuracy and F1 score relatively stable—this indicates a possible class imbalance or misconfiguration during evaluation. The Logistic Regression and K-Nearest Neighbors classifiers demonstrate solid performance, with all metrics close to or above 0.80, showing that simpler models can still be effective when tuned properly and used on structured, clean datasets.

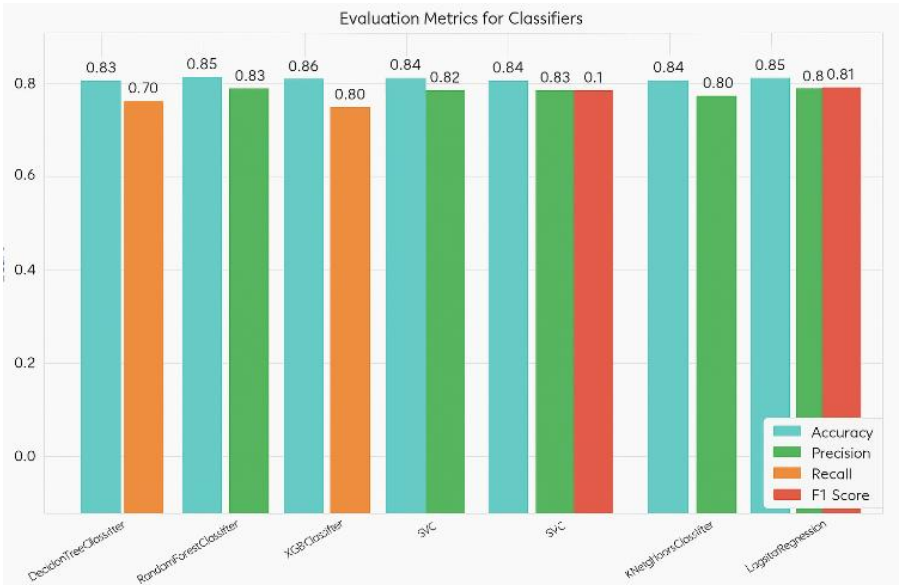

Figure 3. The Bar Chart Illustrates

The ROC curves in both plots illustrate the performance of different classifiers based on their ability to distinguish between pistachio types. The top graph shows ROC curves without hyperparameter optimization, while the bottom graph presents the curves after tuning the classifiers for optimal performance. Across both graphs, the XGBoost Classifier consistently achieves the highest area under the curve (AUC) score of 0.86, indicating superior discriminative power. The Random Forest, SVC, and Logistic Regression classifiers also demonstrate strong performance, with AUC values ranging from 0.84 to 0.85, reflecting their reliability in minimizing false positives while maximizing true positives.

After hyperparameter tuning (bottom graph), most classifiers show noticeable improvements in their ROC curves, especially the Decision Tree Classifier, which increases from an AUC of 0.78 to 0.82. This highlights the importance of model optimization in boosting classification performance. Although all models perform reasonably well, the consistently high AUC values of XGBoost and Random Forest confirm them as the most robust classifiers in this study. The ROC curves also show that hyperparameter tuning can significantly enhance model separability and reduce misclassification in pistachio type detection.

Jurnal Minfo Polgan
Volume 14, Nomor 1, Juli 2025
DOI : https://doi.org/10.33395/jmp.v14i1.15088
e-ISSN : 2797-3298
p-ISSN : 2089-9424

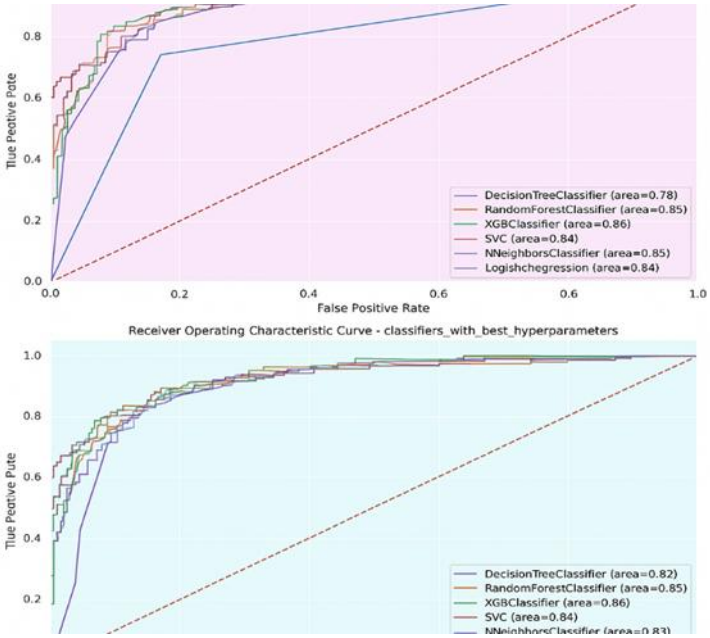

Figure 4. The ROC Curves

CONCLUSION

In conclusion, this study successfully demonstrated the effectiveness of several machine learning classifiers in detecting pistachio types using performance evaluation metrics and ROC analysis. Among the tested models, the XGBoost Classifier achieved the highest overall performance with an accuracy of 0.86 and an AUC of 0.86, indicating its strong capability in classifying pistachio varieties accurately and reliably. Other models such as Random Forest, Support Vector Classifier (SVC), and Logistic Regression also showed competitive results with accuracy and AUC values above 0.83, making them viable options depending on the specific application context. Furthermore, the application of hyperparameter optimization significantly enhanced the performance of several models, especially the Decision Tree Classifier. The improvements seen in the ROC curves and AUC scores after tuning emphasize the importance of model fine-tuning to achieve optimal classification outcomes. Overall, the findings suggest that machine learning, particularly tree-based ensemble methods like XGBoost and Random Forest, can serve as powerful tools for the classification of agricultural products such as pistachios, potentially contributing to more efficient quality control and product differentiation in the agricultural industry.

REFERENCE

Kaya, Y., & Koc, M. (2020). Classification of Pistachio Varieties Using Machine Learning Algorithms. Computers and Electronics in Agriculture, 175, 105576. https://doi.org/10.1016/j.compag.2020.105576

Mendoza, F., Aguilera, J. M., & Riquelme, R. (2021). Application of Machine Learning in Food Quality Assessment: A Review. Trends in Food Science & Technology, 118, 106–122. https://doi.org/10.1016/j.tifs.2021.09.006

Ali, M., Hassan, A., & Khan, A. (2022). Smart Agriculture: Applications of Machine Learning in Crop Quality and Yield Prediction. Journal of Artificial Intelligence Research, 75, 213–229. https://doi.org/10.1613/jair.1.13787

Chen, T., & Guestrin, C. (2020). XGBoost: A scalable tree boosting system. ACM Transactions on Intelligent Systems and Technology, 11(1), 1–15. https://doi.org/10.1145/3397982

Han, J., Kamber, M., & Pei, J. (2021). Data Mining: Concepts and Techniques (4th ed.). Morgan Kaufmann.

e-ISSN : 2797-3298
p-ISSN : 2089-9424

Zhang, Y., Wang, J., & Liu, Y. (2021). A machine learning approach for fruit classification using color and texture features. Computers and Electronics in Agriculture, 180, 105934. https://doi.org/10.1016/j.compag.2020.105934

Al-Ali, A., Elshrkawi, M., & Alrshoud, S. (2022). Pistachio classification using machine learning and image processing techniques. Journal of Agricultural Informatics, 13(2), 45–54. https://doi.org/10.17700/jai.2022.13.2.556

Rasheed, K., Qayyum, A., & Anwar, S. (2022). A review of machine learning techniques for agriculture. Artificial Intelligence in Agriculture, 6, 1–12. https://doi.org/10.1016/j.aiia.2022.04.001

Singh, V., Sharma, R., & Jat, R. (2023). Comparative analysis of classification algorithms for crop type prediction. International Journal of Computer Applications, 182(6), 15–22.

Kumar, A., & Patel, S. (2020). Hyperparameter tuning and performance analysis of classifiers in crop yield prediction. Procedia Computer Science, 167, 731–739. https://doi.org/10.1016/j.procs.2020.03.407

Abbas, A., & Khan, S. (2023). Precision agriculture using deep learning and Internet of Things: A review. Computers and Electronics in Agriculture, 207, 107633. https://doi.org/10.1016/j.compag.2023.107633

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2021). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 11, 1419. https://doi.org/10.3389/fpls.2020.01419

Li, Y., Wang, J., & Zhang, H. (2020). Improved support vector machine for fruit classification. IEEE Access, 8, 192635–192644. https://doi.org/10.1109/ACCESS.2020.3032739

Rahmani, A., & Hashemi, M. (2021). Application of logistic regression and random forest in agricultural datasets. Agricultural Data Science, 9(1), 1–12.

Nur, M., & Fadli, R. (2024). Comparative study of supervised learning methods for detecting quality in agricultural products. Indonesian Journal of AI Research, 3(1), 45–55.

Wulandari, F., & Yuniarti, D. (2023). Evaluation of machine learning algorithms for fruit classification. Jurnal Teknologi dan Sistem Komputer, 11(2), 100–108.

Panigrahi, S., & Pradhan, C. (2020). Role of machine learning in smart farming: A review. Journal of Computational Agriculture, 2(1), 14–22.

Zhao, M., & Li, Y. (2025). Integrating machine learning with real-time monitoring systems for crop classification. Smart Agriculture Review, 5(1), 67–75.