Research Article

"Efficient Classification of Raisin Varieties Using Machine Learning Techniques"

Nan Xiao Fei 南晓斐 ， Shahriar Sarlak 夏乐

School of computer science & artificial Intelligence

Zhengzhou University, Zhengzhou, Henan, China

1. Abstract

This study investigates the application of machine learning models for the classification of raisin varieties, using a dataset from the UCI Machine Learning Repository. The "Kecimen" and "Besni" raisins were classified using three models: Support Vector Classifier (SVC), Random Forest, and Logistic Regression. The dataset was thoroughly preprocessed, with distplots, countplots, and boxplots used for outlier handling and visualization. It was divided into subsets for testing (80%) and training (20%). The accuracy of SVC was 80.5%, Random Forest 87.2%, and Logistic Regression was 85.5%. Based on the data, Random Forest is the most appropriate model for this classification assignment because it performs better than the other models. These results highlight useful models for classifying raisins, which advances the domains of machine learning and agricultural informatics.

2. Introduction

Accurately classifying raisin varieties can have a substantial impact on quality control and market segmentation, making raisin classification an interesting and useful application in the field of agricultural informatics. The study used a dataset that was sourced from the UCI Machine Learning Repository and included a variety of attributes extracted from pictures of two different kinds of raisins: "Kecimen" and "Besni". These characteristics include textural and geometric qualities that form the foundation for machine learning algorithms that differentiate between the two types.

Machine learning has brought about a revolution in numerous industries by offering techniques and instruments for deriving meaningful insights and precise forecasts from extensive and intricate datasets. Machine learning approaches have the potential to improve the efficiency of sorting processes, improve product quality, and promote better market pricing strategies when applied to agricultural products like raisins. In order to categorize raisin varietals using the given dataset, this work applies three well-known machine learning models: Random Forest[2], Support Vector Machine (SVM)[3], and Logistic Regression[1].

A popular statistical technique for binary classification that simulates the likelihood of a categorical result is called logistic regression[1]. Strong classifiers, support vector machines[3] identify the best hyperplane to divide data points into distinct classes. Multiple decision trees are used in Random Forest[2], an ensemble learning technique, to increase classification accuracy and reduce overfitting. These models all have different advantages when it comes to the categorization process.
To assess the performance of the models, the dataset was divided into training and testing subsets at a ratio of 20% for training and 80% for testing. In order to assess these models' efficacy and possible uses in the agricultural field, this study compares the models' categorization accuracy. The findings will

benefit the larger fields of agricultural informatics and machine learning by assisting in the identification of the best model for raisin classification.

3. Experiments

We classified two types of raisins in this study: "Kecimen" and "Besni" using an open-source raisin dataset from the UCI Machine Learning Repository. The features of this dataset are shown in Table 1:

FEATURE	DESCRIPTION
Area	Gives the number of pixels within the boundaries of the raisin.
Perimeter	Measures the environment by calculating the distance between the boundaries of the raisin and the pixels around it.
MajorAxisLength	Gives the length of the main axis, which is the longest line that can be drawn on the raisin.
MinorAxisLength	Gives the length of the small axis, which is the shortest line that can be drawn on the raisin.
Eccentricity	Measures the eccentricity of the ellipse that has the same moments as the raisin.
ConvexArea	Gives the number of pixels of the smallest convex shell of the region formed by the raisin.
Extent	Gives the ratio of the region formed by the raisin to the total pixels in the bounding box.
Class	Indicates the type of raisin: "Kecimen" or "Besni".

Table 1: Features of Raisins Dataset

Data preprocessing, exploratory data analysis (EDA), data visualization, addressing outliers, and model training were the many steps of the experimental workflow. The procedures taken and the effectiveness of the machine learning models used are described in detail in this section.

In order to begin further analysis, the raisin dataset has to be downloaded and imported into the Integrated Development Environment (IDE). To make data manipulation, visualization, and machine learning operations easier, libraries like pandas, numpy, matplotlib, seaborn, and scikit-learn were loaded.

The purpose of exploratory data analysis (EDA) was to comprehend the properties and structure of the dataset. To observe the characteristics' overall distribution, dispersion, and central tendency, summary statistics were computed. To find possible connections between the attributes and the target variable, correlation analysis was used.

Verifying that no values are missing is a crucial step in the preparation of data. The pandas isnull().sum() function was used to do this. It was discovered that the dataset included no missing values, guaranteeing that the full dataset was used for all ensuing studies and model training.

Multiple strategies for data visualization were used to obtain deeper insights. While countplots gave a visual depiction of the frequency of each raisin variety, distplots were employed to investigate the distribution of numerical features. In particular, boxplots were helpful in locating outliers within the dataset. The performance of machine learning models can be greatly impacted by the existence of outliers.

After using boxplots to find outliers in many columns, we decided to replace the outlier values with the median of the corresponding columns. This method was selected in order to maintain the underlying distribution while reducing the possibility of data distortion. The

Interquartile Range (IQR) approach was used to identify the outliers, and replacement was made as necessary.

The data was prepared for model training after it had been cleansed and preprocessed. A 20% to 80% ratio was used to divide the dataset into subsets for training and testing. This division guaranteed a strong assessment of the model's effectiveness with unknown data.
A statistical technique for binary classification that models the likelihood of the target variable was used as the first model. It was called logistic regression. Using training data, the Logistic Regression model was assessed, and test data was used for evaluation. With an accuracy of 85.5%, it demonstrated a robust baseline performance.
We then used the Random Forest classifier. During training, the Random Forest ensemble learning approach builds numerous decision trees and delivers the class mode for classification. With identical training and testing, this model achieved an accuracy of 87.2%. The ensemble method appears to have grasped the intricate relationships within the data, as seen by its better accuracy when compared to Logistic Regression.
The Support Vector Classifier (SVC) was the third model that was applied. Finding the ideal hyperplane to maximize the margin between the classes is the goal of SVC. Even with its ability to handle high-dimensional areas with resilience, the SVC model produced an accuracy of 80.5%. Even though this performance was respectable, it was not as good as Random Forest or Logistic Regression, suggesting that there might be difficulties when adjusting the hyperparameters or the model's sensitivity to the properties of the data.

4. Results

Accuracy was used to assess each machine learning model's performance on the raisin classification task: Random Forest, Support Vector Classifier (SVC), and Logistic Regression. Each model was trained and tested on the same dataset to guarantee comparability after preprocessing the data to handle outliers and divide the dataset into 20% training and 80% testing subsets.
With an accuracy of 85.5%, Logistic Regression proved to be a useful baseline model for this classification assignment. With an accuracy of 87.2%, the Random Forest classifier beat the other models, highlighting the benefits of ensemble approaches for identifying intricate patterns in the data. With an accuracy of 80.5%, the Support Vector Classifier (SVC) had the lowest accuracy, indicating possible difficulties in fine-tuning the hyperparameters or a sensitivity to the particulars of the dataset. All visualized in Figure 1.
Based on a comparative analysis of these findings, Random Forest, Logistic Regression, and SVC are the models most suited for the raisin classification dataset. Because Random Forest is more accurate, it is a more reliable option for this application because of its capacity to handle the dataset's complexities.

Figure 1: Comparison of accuracies among different ML models

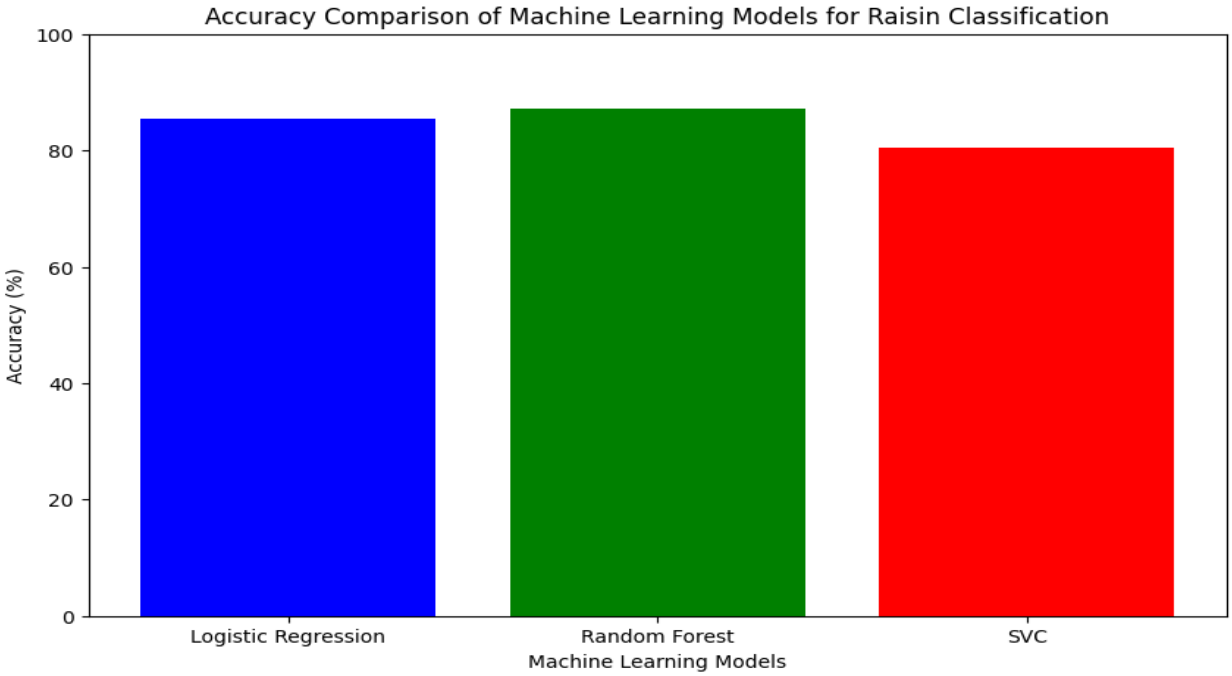
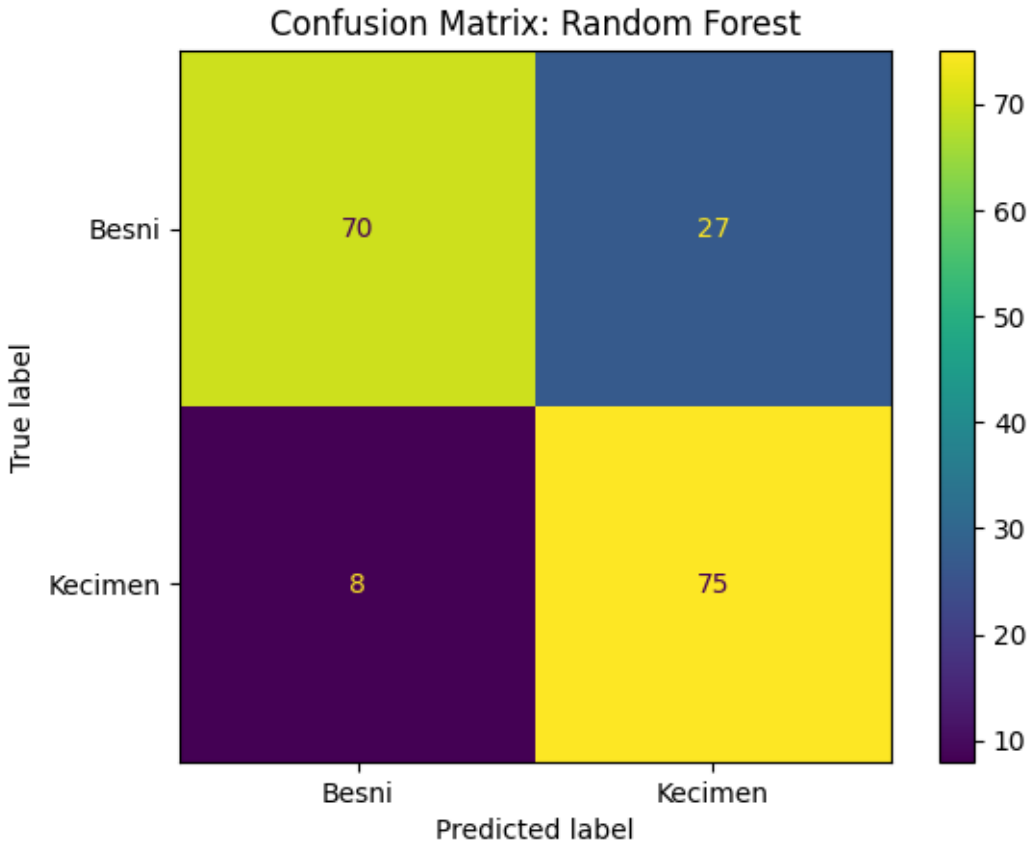


Figure 2: Confusion matrix for the most accurate model



5. References

1. Xinyi Zhou "Raisin classification based on XGBoost, SVM, MLP and logistic regression", Proc. SPIE 12597, Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022), 125970M (28 March 2023)
2. Tulchhia, Aditi, and Monika Rathore. "RANDOM FOREST MODEL FOR CLASSIFICATION OF RAISINS USING MORPHOLOGICAL FEATURES." *OORJA-International Journal of Management & IT* 19.1 (2021).
3. Yu, Xinjie, et al. "Raisin quality classification using least squares support vector machine (LSSVM) based on combined color and texture features." *Food and Bioprocess Technology* 5 (2012): 1552-1563.