| Academic Year | Module | Assessment Number | Assessment Type |
|---|---|---|---|
| L5 | Concepts and Technologies of AI. | 03 | Report Writing |

# Classification Analysis Report


Student Id : 2431366
Student Name : Nalina Rai
Section : L5CG22
Module Leader : Mr. Siman Giri
Tutor : Mr. Ronit Shrestha
Submitted on : 20-12-2024

## Abstract

This report's objective is to use machine learning approaches to categorize apple quality.

Method: Various attributes associated with apple qualities are included in the dataset, and classification techniques like Random Forest and Logistic Regression were utilized. Exploratory Data Analysis (EDA), model construction, hyperparameter tuning, and feature selection were among the procedures.

The model's performance was assessed using F1-score, recall, accuracy, and precision. With an accuracy of 92%, Random Forest was the best-performing model, followed by Logistic Regression with 87%.

As a result of feature selection and hyperparameter optimization, Random Forest was the model with the best performance. For further improvement, deep learning techniques and bigger datasets might be used in future research.

## 1. Introduction

### 1.1 Problem Statement

In order to categorize apples as either excellent or terrible, this project aims to predict apple quality based on provided characteristics.

### 1.2 Dataset

Apple quality is evaluated using a variety of numerical and categorical variables found in the dataset.

### 1.3 Objective

The goal is to create a prediction model that, using the provided dataset, correctly classifies apples.

## 2. Methodology

### 2.1 Data Preprocessing

manipulated missing numbers by substituting the median.

numerical values that represent encoded category properties.

Numerical characteristics were normalized for uniform scale.

### 2.2 Exploratory Data Analysis (EDA)

EDA in play:

making feature distributions visible using box plots and histograms.

Heatmap analysis is used to verify correlations.

finding trends in the excellent and bad apples.

### 2.3 Model Building

There were two models in use:

A clear and easy-to-understand categorization approach is logistic regression.

An ensemble learning technique called Random Forest increases prediction accuracy.

### 2.4 Model Evaluation

Evaluation metrics:

Accuracy: Calculates the percentage of accurate forecasts.

Precision: Shows the percentage of projected positives that were accurate.

Recall: Indicates how successfully the model finds real positives.

The F1-score strikes a balance between recall and accuracy.

### 2.5 Hyperparameter Optimization

Hyperparameters for both models were optimized using GridSearchCV.

Random Forest's ideal settings are max_depth=10 and n_estimators=100.

C=1.0 and solver='liblinear' are the ideal values for logistic regression.

### 2.6 Feature Selection

The top 5 characteristics most pertinent to categorization were determined using the Chi-Square test.

Accuracy was maintained while efficiency was increased via certain features.

## 3. Conclusion

### 3.1 Key Findings

Random Forest achieved 92% accuracy, outperforming Logistic Regression.

Efficiency was increased by feature selection without sacrificing performance.

The model's performance was improved by hyperparameter adjustment.

### 3.2 Final Model

Because Random Forest was more accurate and resilient, it was chosen as the final model.

### 3.3 Challenges

Handling missing values needed caution.

To prevent overfitting, features have to be chosen carefully.

### 3.4 Future Work

To improve accuracy, try using deep learning models.

Improve generalization by using more datasets.

In agricultural contexts, use real-time apple quality evaluation.

## 4 Discussion

### 4.1 Model Performance

Compared to Logistic Regression, Random Forest was more efficient at processing complicated patterns.

### 4.2 Impact of Hyperparameter Tuning and Feature Selection

Random Forest performance was greatly enhanced by tuning.

While maintaining high accuracy, feature selection decreased dimensionality.

### 4.3 Interpretation of Results

The outcomes demonstrate that Random Forest produces precise predictions and is a good fit for classifying apples.

### 4.4 Limitations

Generalization may be restricted by the size of the dataset.

Because of the complexity of the data, logistic regression fared worse.

### 4.5 Suggestions for Future Research

Use neural networks for maybe more accurate classification.

Boost robustness by gathering more varied apple data.

Use this model to create an apple grading system that operates in real time.

## Final Remarks

The usefulness of machine learning for classifying apple quality was effectively shown in this study. The best-performing model was determined to be Random Forest through the use of feature selection, hyperparameter adjustment, and model evaluation. Subsequent investigations can enhance categorization methods, guaranteeing improved evaluation of agricultural quality.