Implementation and Evaluation of Random Forest Classifier using Wine Quality Dataset

Student Name: Jayant Parmar
Roll Number: g25ait1072

1. Objective

The objective of this assignment is to implement a Random Forest classifier using the Wine Quality dataset from the UCI Machine Learning Repository. The goal is to analyze the effect of ensemble learning on model performance compared to a single Decision Tree classifier, using evaluation metrics and visualizations.

2. Dataset Description

Dataset Source: UCI Machine Learning Repository – Wine Quality Dataset (https://archive.ics.uci.edu/dataset/186/wine+quality)

Version Used: Red Wine Quality Dataset
Number of Samples: 1,599
Number of Features: 11 numeric attributes + 1 quality score

Target Conversion: To simplify the task, the multiclass target variable 'quality' is converted into a binary variable:
- Wines with quality > 5 → Good (1)
- Wines with quality ≤ 5 → Bad (0)

Feature	Description	Type
fixed acidity	Amount of tartaric acid	Numeric
volatile acidity	Amount of acetic acid	Numeric
citric acid	Amount of citric acid	Numeric
residual sugar	Amount of sugar after fermentation	Numeric
chlorides	Salt content	Numeric
free sulfur dioxide	Free SO_2 level	Numeric
total sulfur dioxide	Total SO_2 level	Numeric
density	Density of wine	Numeric
pH	Acidity level	Numeric
sulphates	Sulfate content	Numeric
alcohol	Alcohol percentage	Numeric

3. Methodology

A Random Forest is an ensemble learning method that constructs multiple Decision Trees using bootstrap sampling and feature randomness. Each tree votes for a class, and the majority vote becomes the final prediction.

3.1 Algorithm Overview

1. Bootstrap Sampling (Bagging) — each tree is trained on a random sample with replacement.
2. Feature Randomness — at each split, a random subset of features (typically $\sqrt{n}$) is considered.

3.2 Implementation Details

Custom DecisionTreeClassifier implemented using Gini impurity.
RandomForestClassifier developed to train multiple Decision Trees with feature randomness and bootstrap aggregation.
Feature importance calculated using average Gini impurity reduction.
Out-of-Bag (OOB) error estimated for unbiased model evaluation.

4. Results

Performance Metrics:

Model	Accuracy	Precision	Recall	F1
RandomForest_100	0.8000	0.8497	0.7602	0.8025
SingleTree_deep	0.7281	0.7593	0.7193	0.7387

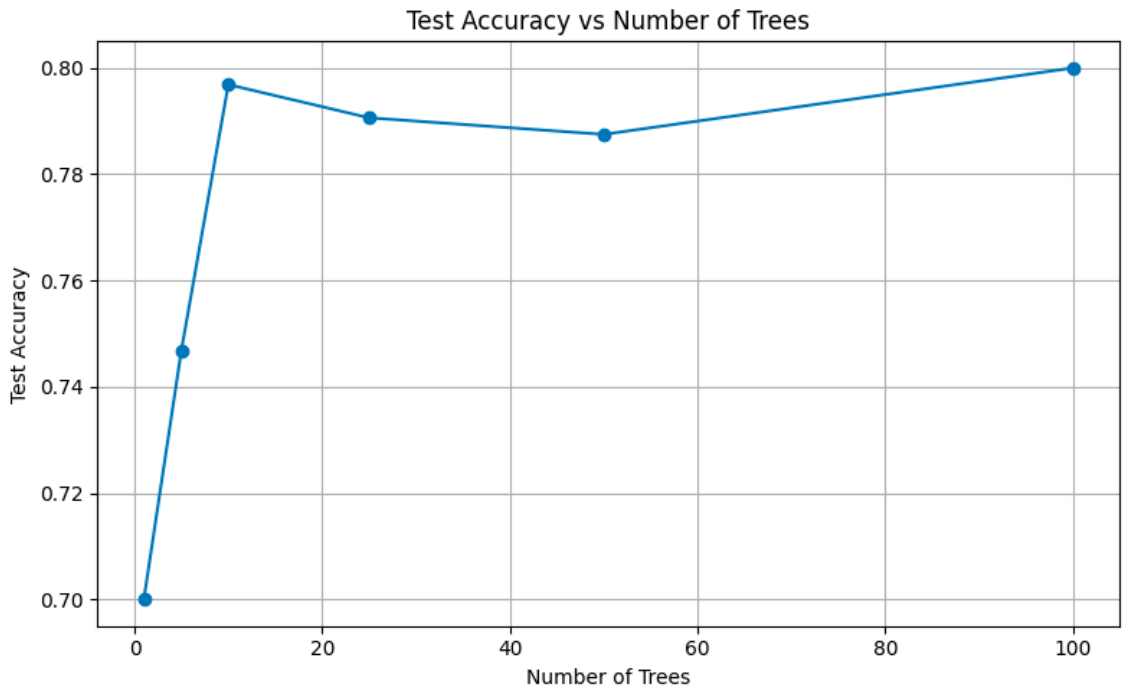

Figure 1: Test Accuracy vs Number of Trees

Observation: As the number of trees increases, the model's test accuracy improves and stabilizes after around [mention number] trees.

Figure 2: Feature Importance (Insert Plot Here)

Observation: Features such as alcohol, volatile acidity, and sulphates were found to be most influential.

5. Analysis and Discussion

The Random Forest outperformed the single Decision Tree, showing improved robustness and reduced overfitting. The accuracy plateau indicates sufficient model convergence. OOB estimates closely matched test results, validating the ensemble's reliability.

6. Assumptions

- Dataset contains no missing or categorical values.
- Binary classification threshold fixed at quality > 5.
- No scaling or normalization performed.
- Random state fixed for reproducibility.

7. References

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
2. UCI Machine Learning Repository: Wine Quality Dataset.
3. Scikit-learn documentation: Metrics and Evaluation Functions.
4. Python 3.12 Documentation.
5. Matplotlib and Pandas official documentation.