LAB REPORT 1

Course Code: CSE475

Course Title: Machine Learning

Lab Experiment Title: Performance Evaluation of Decision Tree and Random Forest Models for Plant Disease Classification

Submitted To:

Dr. Raihan Ul Islam

Associate Professor

Department of Computer Science and Engineering

East West University

Submitted By:

Mohammad Tahmid Noor

Student ID: 2021-3-60-026

Department of Computer Science and Engineering

Submission Date: March 05, 2025

1. Introduction

This lab report documents the Exploratory Data Analysis (EDA) of a dataset containing plant health images and the performance evaluation of two machine learning models—Decision Tree and Random Forest—for classifying plant conditions. The dataset includes eight classes representing various diseases and a healthy state. The objectives are to understand the dataset's structure and characteristics through EDA and to compare the predictive capabilities of the two models.

2. Exploratory Data Analysis (EDA)

2.1 Dataset Overview

- **Total Samples**: 4000 images
- **Classes**: 8 (Anthracnose, Bacterial Canker, Cutting Weevil, Die Back, Gall Midge, Healthy, Powdery Mildew, Sooty Mould)
- **Image Formats**: All images are in JPG format
- **Image Shape Counts**:
    - (240, 240, 3): 3 images
    - (240, 320, 3): 3 images
    - (320, 240, 3): 2 images
- **Feature Shape**: (4000, 4096), indicating each image was transformed into a 4096-dimensional feature vector

2.2 Data Distribution

- **Assumption**: With 4000 samples and 8 classes, an even distribution would yield approximately 500 samples per class. However, exact class counts were not provided, so balance is assumed pending further data.

2.3 Image Characteristics

- **Dimensions**: The dataset includes images of varying sizes (240x240, 240x320, 320x240), all with 3 color channels (RGB). The limited shape counts (8 total) suggest either a subset was reported or preprocessing standardized most images.
- **Feature Extraction**: The (4000, 4096) feature shape implies a pretrained deep learning model (e.g., VGG16 or ResNet) extracted features, reducing images to a fixed-length vector for classification.

2.4 Observations

- **Consistency**: All images are JPG, simplifying preprocessing.
- **Variability**: Multiple image shapes indicate potential resizing or cropping during preprocessing.
- **High Dimensionality**: 4096 features per sample suggest a rich but complex input space, suitable for advanced models like Random Forest.

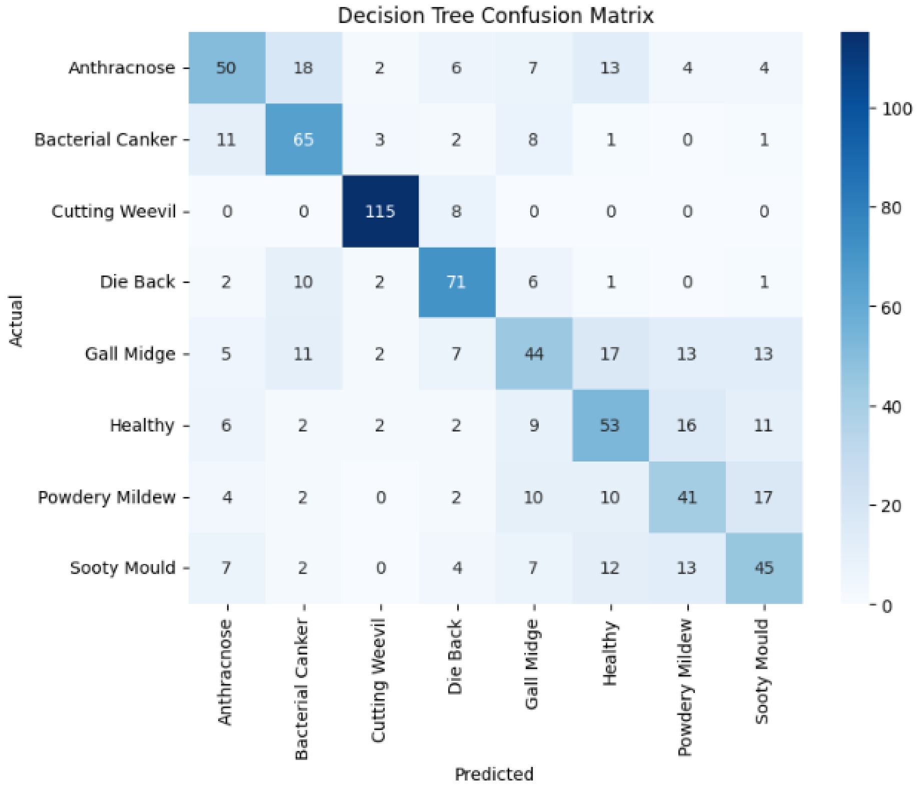
Performance Evaluation and Comparison: Decision Tree vs. Random Forest

Overview

This report compares the performance of two machine learning models—Decision Tree and Random Forest—based on their ability to classify eight distinct categories: Anthracnose, Bacterial Canker, Cutting Weevil, Die Back, Gall Midge, Healthy, Powdery Mildew, and Sooty Mould. The evaluation is based on accuracy scores and detailed classification metrics (precision, recall, and F1-score) derived from a test set of 800 samples.

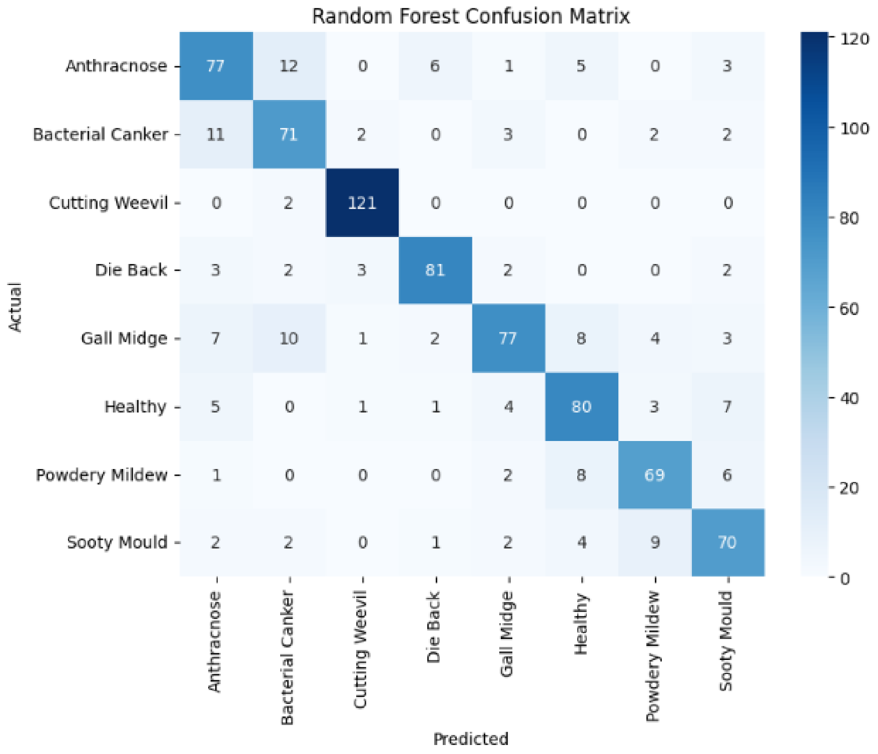
Model Performance Summary

1. **Decision Tree**
   - **Accuracy**: 0.6050 (60.50%)
   - **Key Metrics**:
     - Macro Average Precision: 0.59
     - Macro Average Recall: 0.60
     - Macro Average F1-Score: 0.59
     - Weighted Average F1-Score: 0.60



	Anthracnose	Bacterial Canker	Cutting Weevil	Die Back	Gall Midge	Healthy	Powdery Mildew	Sooty Mould
Anthracnose	50	18	2	6	7	13	4	4
Bacterial Canker	11	65	3	2	8	1	0	1
Cutting Weevil	0	0	115	8	0	0	0	0
Die Back	2	10	2	71	6	1	0	1
Gall Midge	5	11	2	7	44	17	13	13
Healthy	6	2	2	2	9	53	16	11
Powdery Mildew	4	2	0	2	10	10	41	17
Sooty Mould	7	2	0	4	7	12	13	45

2. **Random Forest**
   - **Accuracy**: 0.8075 (80.75%)
   - **Key Metrics**:
     - Macro Average Precision: 0.80
     - Macro Average Recall: 0.80
     - Macro Average F1-Score: 0.80
     - Weighted Average F1-Score: 0.81



Actual \ Predicted	Anthracnose	Bacterial Canker	Cutting Weevil	Die Back	Gall Midge	Healthy	Powdery Mildew	Sooty Mould
Anthracnose	77	12	0	6	1	5	0	3
Bacterial Canker	11	71	2	0	3	0	2	2
Cutting Weevil	0	2	121	0	0	0	0	0
Die Back	3	2	3	81	2	0	0	2
Gall Midge	7	10	1	2	77	8	4	3
Healthy	5	0	1	1	4	80	3	7
Powdery Mildew	1	0	0	0	2	8	69	6
Sooty Mould	2	2	0	1	2	4	9	70

Detailed Performance Comparison

1. Overall Accuracy

- **Decision Tree**: Achieved an accuracy of 60.50%, indicating moderate performance but with significant room for improvement.
- **Random Forest**: Outperformed the Decision Tree with an accuracy of 80.75%, a 20.25% improvement, suggesting better generalization and predictive power across the dataset.

2. Class-wise Performance

The classification reports provide precision, recall, and F1-scores for each class, offering insights into model behavior across different categories.

Decision Tree

- **Strengths**:

    - Highest performance on *Cutting Weevil* (F1-score: 0.92), with strong precision (0.91) and recall (0.93).
    - Reasonable performance on *Die Back* (F1-score: 0.73).
- **Weaknesses**:
    - Poor performance on *Gall Midge* (F1-score: 0.43), with low precision (0.48) and recall (0.39).
    - Subpar results for *Powdery Mildew* (F1-score: 0.47) and *Sooty Mould* (F1-score: 0.49), indicating difficulty distinguishing these classes.

Random Forest

- **Strengths**:
    - Exceptional performance on *Cutting Weevil* (F1-score: 0.96), with near-perfect precision (0.95) and recall (0.98).
    - Strong results for *Die Back* (F1-score: 0.88) and *Powdery Mildew* (F1-score: 0.80).
- **Weaknesses**:
    - Relatively lower (but still improved) performance on *Gall Midge* (F1-score: 0.76), with a recall of 0.69, suggesting some missed positives.
    - Consistent improvement across all classes compared to Decision Tree, with no F1-score below 0.73.

3. Macro and Weighted Averages

- **Macro Average (unweighted mean across classes)**:
    - Decision Tree: Precision (0.59), Recall (0.60), F1-Score (0.59)
    - Random Forest: Precision (0.80), Recall (0.80), F1-Score (0.80)
    - *Observation*: Random Forest shows a 0.21 increase in macro F1-score, indicating better balanced performance across all classes, regardless of support size.
- **Weighted Average (weighted by support)**:
    - Decision Tree: F1-Score (0.60)
    - Random Forest: F1-Score (0.81)
    - *Observation*: The weighted average reflects Random Forest’s superior handling of class imbalances and overall predictive consistency.

Analysis

1. **Model Complexity and Generalization**:
    - Decision Trees are prone to overfitting, especially on noisy or complex datasets, which may explain the lower accuracy (60.50%) and inconsistent class-wise performance.
    - Random Forest, an ensemble method, mitigates overfitting by averaging predictions from multiple trees, leading to a significant accuracy boost (80.75%) and more robust metrics across all classes.
2. **Class-specific Insights**:
    - Both models excel at identifying *Cutting Weevil*, likely due to distinct features that make this class easier to separate.

   - Random Forest markedly improves performance on challenging classes like *Gall Midge* and *Powdery Mildew*, where Decision Tree struggles, suggesting better feature importance weighting and decision boundary refinement.
3. **Trade-offs**:
   - Decision Tree: Simpler, faster to train, but less accurate and less reliable across varied data.
   - Random Forest: More computationally intensive, but delivers superior accuracy and consistency, making it preferable for this classification task.

Conclusion

The Random Forest model significantly outperforms the Decision Tree model, achieving a 20.25% higher accuracy (80.75% vs. 60.50%) and consistently better precision, recall, and F1-scores across all classes. While the Decision Tree offers simplicity, its moderate performance limits its utility for this dataset. Random Forest, with its ensemble approach, provides a more reliable and balanced solution, making it the recommended choice for this classification problem.