Lab Report 1

Course Name: Machine Learning
Course Code: CSE 475
Section - 3

Assignment Name: Perform Mango leaf disease classification using Random Forest and Decision Tree.

Submitted By
Name: Tithi Paul
ID: 2021-2-60-057
Dept. of Computer Science & Engineering

Submitted To
Dr. Raihan Ul Islam
Associate Professor
Department of Computer Science and Engineering
East West University

Mango leaf disease classification using Random Forest and Decision Tree

Introduction

This project deals with designing a machine learning-based classification system that identifies diseases in mango leaves through images. Various conditions of the mango leaf are available in this dataset used for the analysis, which are then organized in respective classes. The system will be based on various techniques that deal with image processing and machine learning for effective classification of the images. The most important objectives are to apply EDA to the data and experiment with the two classification algorithms: Decision Tree and Random Forest.

Implementation

1. Libraries Used:

- cv2: For image processing and resizing.
- numpy and pandas: For data handling and manipulation.
- sklearn: For model selection, preprocessing, and classification algorithms.
- matplotlib and seaborn: For data visualization.

2. Data Loading and Preprocessing:

1. Data Loading: Images were loaded from different folders, each representing a unique class. Images were resized to a uniform size of 128x128 pixels.
2. Data Conversion: The loaded data and labels were converted into numpy arrays for efficient handling.
3. Data Visualization: The class distribution was visualized to check for imbalances, and sample images from each class were displayed.
4. Color Analysis: Mean RGB intensity for the dataset was calculated and visualized to understand the color profile of the images.
5. Data Normalization and Flattening: Images were flattened and normalized for input into machine learning models.
6. Label Encoding: Class labels were encoded into numeric values suitable for classification.

3. Algorithm:

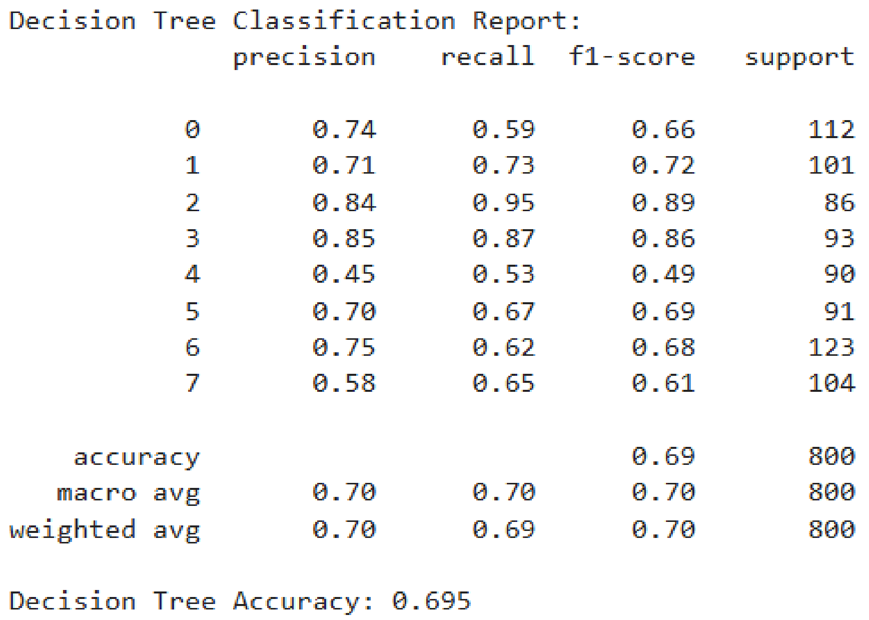
The study implements two primary algorithms for classification: Decision Tree and Random Forest.

Decision Tree: A simple, interpretable algorithm that splits the data into subsets based on feature values. It works by creating branches in a tree structure to reach a final decision on classification. Although efficient, it may be prone to overfitting.

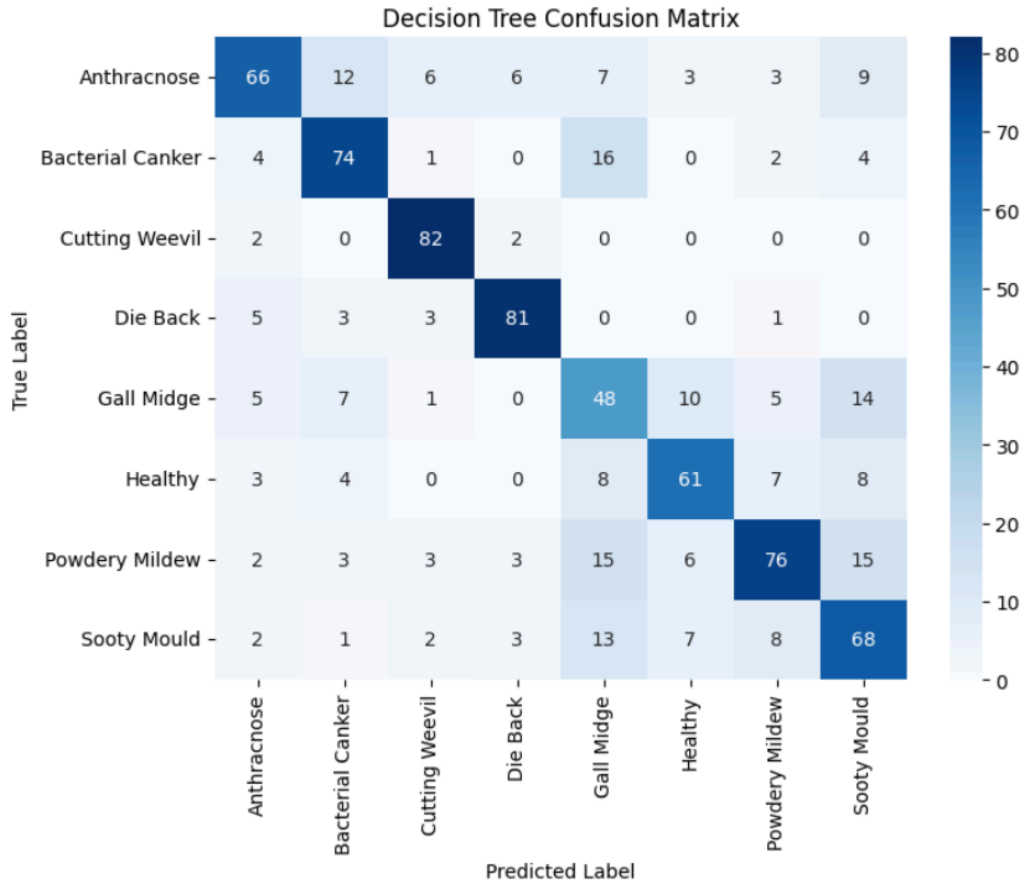
Random Forest: An ensemble of Decision Trees that combines multiple trees to improve generalization. By averaging the predictions of several trees, Random Forest tends to reduce overfitting and improve accuracy. This model is particularly robust in classification tasks with a complex feature space, like image data.

4. Simulation:

- **Decision Tree:**

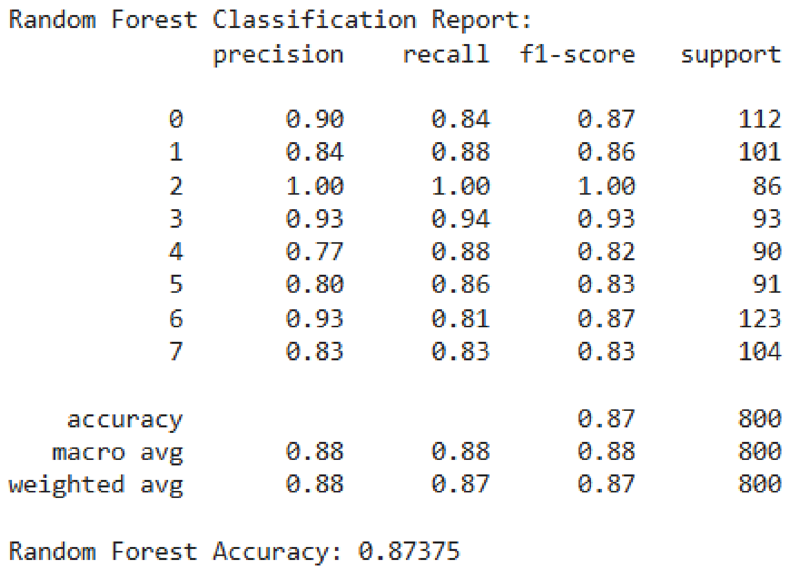


```
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.59      0.66       112
           1       0.71      0.73      0.72       101
           2       0.84      0.95      0.89        86
           3       0.85      0.87      0.86        93
           4       0.45      0.53      0.49        90
           5       0.70      0.67      0.69        91
           6       0.75      0.62      0.68       123
           7       0.58      0.65      0.61       104

    accuracy                           0.69       800
   macro avg       0.70      0.70      0.70       800
weighted avg       0.70      0.69      0.70       800

Decision Tree Accuracy: 0.695
```

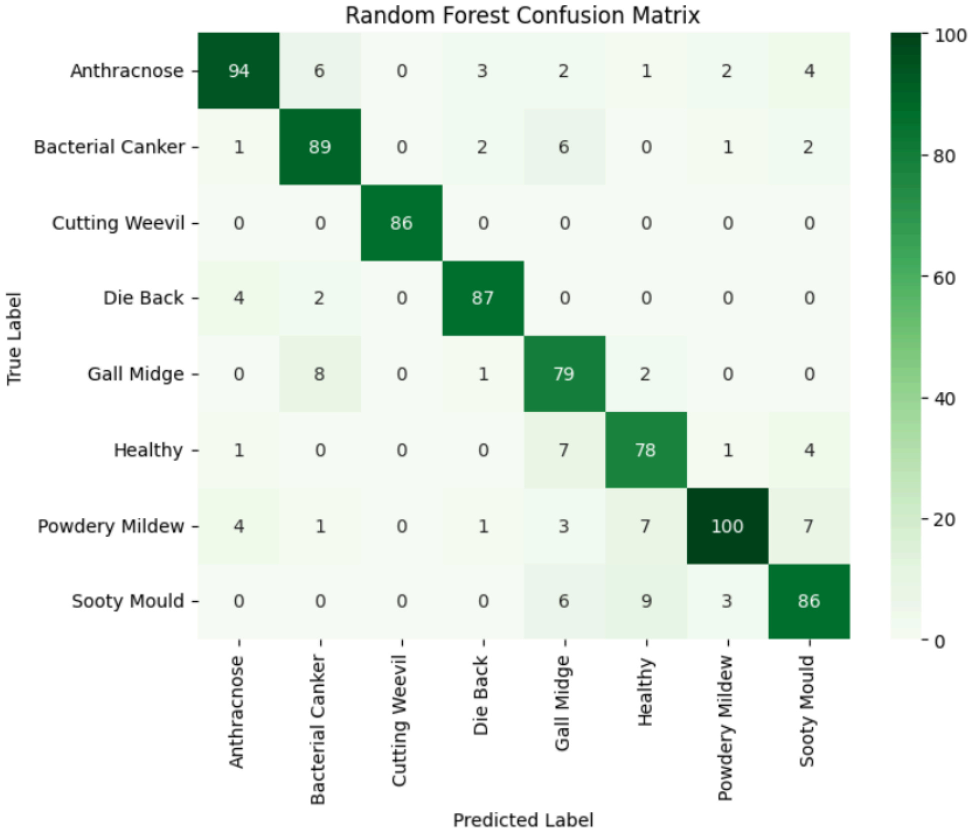


	Anthracnose	Bacterial Canker	Cutting Weevil	Die Back	Gall Midge	Healthy	Powdery Mildew	Sooty Mould
Anthracnose	66	12	6	6	7	3	3	9
Bacterial Canker	4	74	1	0	16	0	2	4
Cutting Weevil	2	0	82	2	0	0	0	0
Die Back	5	3	3	81	0	0	1	0
Gall Midge	5	7	1	0	48	10	5	14
Healthy	3	4	0	0	8	61	7	8
Powdery Mildew	2	3	3	3	15	6	76	15
Sooty Mould	2	1	2	3	13	7	8	68

- **Random Forest:**



```
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.84      0.87       112
           1       0.84      0.88      0.86       101
           2       1.00      1.00      1.00        86
           3       0.93      0.94      0.93        93
           4       0.77      0.88      0.82        90
           5       0.80      0.86      0.83        91
           6       0.93      0.81      0.87       123
           7       0.83      0.83      0.83       104

    accuracy                           0.87       800
   macro avg       0.88      0.88      0.88       800
weighted avg       0.88      0.87      0.87       800

Random Forest Accuracy: 0.87375
```



Random Forest Confusion Matrix

True Label \ Predicted Label	Anthracnose	Bacterial Canker	Cutting Weevil	Die Back	Gall Midge	Healthy	Powdery Mildew	Sooty Mould
Anthracnose	94	6	0	3	2	1	2	4
Bacterial Canker	1	89	0	2	6	0	1	2
Cutting Weevil	0	0	86	0	0	0	0	0
Die Back	4	2	0	87	0	0	0	0
Gall Midge	0	8	0	1	79	2	0	0
Healthy	1	0	0	0	7	78	1	4
Powdery Mildew	4	1	0	1	3	7	100	7
Sooty Mould	0	0	0	0	6	9	3	86

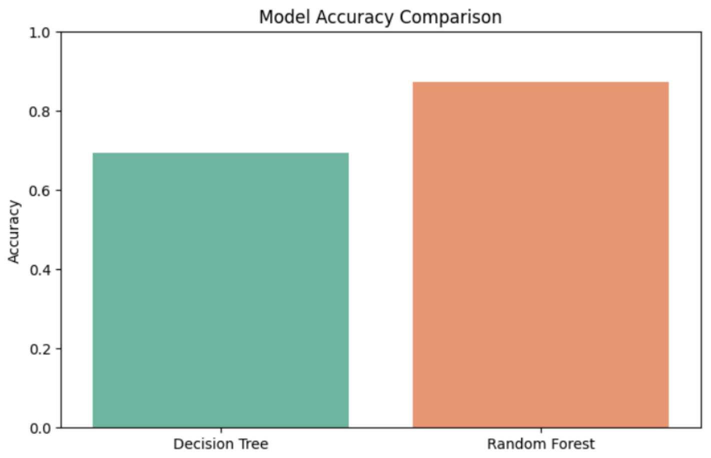
Features

Image Features:

- **Size Uniformity**: Each image was resized to 128x128 pixels, ensuring a consistent input size for models.
- **Flattening and Normalization**: The image data was flattened into a one-dimensional vector per image and normalized to a 0-1 range.

Analysis

1. **Exploratory Data Analysis (EDA)**:
   - **Class Distribution**: Using countplot, we visualized the distribution of classes in the dataset. This helped identify if the dataset was balanced across classes.
   - **Sample Images**: Displaying random samples from each class provided insights into the diversity of images within each category.
   - **Color Profile**: We calculated and displayed the mean RGB intensity across all images to analyze the color balance.
2. **Model Training and Evaluation**:
   - **Decision Tree**:
     - Trained a Decision Tree Classifier using the flattened image data.
     - Achieved an accuracy score, presented alongside a classification report.
     - Generated a confusion matrix to visualize performance across classes.
   - **Random Forest**:
     - Trained a Random Forest Classifier with 100 trees for comparison.
     - Achieved an accuracy score and generated a classification report similar to the Decision Tree.
     - Visualized the confusion matrix to observe improvements over the Decision Tree.
3. **Comparison of Model Performance**:
   - The accuracy scores of both models were plotted for a clear comparison.
   - Observations indicated that the Random Forest generally performed better due to its ensemble nature, which reduces variance and enhances generalization.



Conclusion

This project successfully demonstrated a machine-learning pipeline for classifying mango leaf diseases using image data. By employing image processing, feature extraction, and model comparison, we observed that:

- **Random Forest** outperformed **Decision Tree** in accuracy, likely due to its ensemble approach that mitigates overfitting.
- The EDA provided valuable insights into the dataset's structure and characteristics, guiding preprocessing and model selection decisions.

Future improvements could involve using convolutional neural networks (CNNs) for potentially higher accuracy due to their superior ability to handle image data.