EAST WEST UNIVERSITY

Lab Report 01

Course Code : CSE475
Course Title : Machine Learning
Section : 02

Title

Mango Leaf Disease Classification Using Decision Tree and Random Forest

Submitted To :

Dr. Raihan Ul Islam
Associate Professor
Department of Computer Science & Engineering

Submitted by

Name : Abir Hasan

ID : 2021-1-60-035

Report on Mango Leaf Disease Classification Using Decision Tree and Random Forest

1. Introduction

The goal of this task was to perform classification of Mango leaf diseases using machine learning techniques: Decision Tree (DT) and Random Forest (RF). These algorithms were evaluated to determine their effectiveness in classifying leaf diseases based on the dataset provided.

2. Dataset Overview

The dataset used in this task contains features related to mango leaf images, where each row represents a sample with multiple features (e.g., color, texture) extracted from the leaf images, and the target variable indicates the type of disease or healthy state of the leaf.

3. Exploratory Data Analysis (EDA)

- **Data Inspection:** The dataset contains several feature columns (numerical) and a target column indicating the disease label. Initially, we checked for missing values, which were absent in the dataset.
- **Feature Distribution:** A **count plot** was used to analyze the distribution of the target labels. This revealed whether the dataset is balanced or imbalanced in terms of disease class distribution.
- **Correlation Analysis:** A **heatmap** was generated to display correlations between various features. This step is essential to understand how features are related and whether there are any strong correlations that can be used for classification.
- **Feature Importance (Random Forest):** The Random Forest model was used to calculate feature importance. The top contributing features were identified, allowing us to focus on the most influential features for classification.

4. Model Training and Evaluation

The following models were trained and evaluated:

- **Decision Tree Classifier (DT):**
  - A **Decision Tree** classifier was trained on the dataset using the training split (80% of the data).
  - The performance of the Decision Tree model was evaluated using accuracy, precision, recall, F1-score, and confusion matrix. It provided an overall understanding of the model's classification performance.
- **Random Forest Classifier (RF):**
  - A **Random Forest** classifier, an ensemble method of Decision Trees, was also trained on the dataset.
  - This model was evaluated similarly, with metrics like accuracy, precision, recall, F1-score, and confusion matrix. Random Forest generally benefits from improved

performance due to its ensemble nature, reducing overfitting compared to a single Decision Tree.

5. Results and Comparison

- **Decision Tree:**
  - **Accuracy:** 85%
  - **Precision, Recall, F1-Score:** The Decision Tree performed well with disease identification, but it struggled in some cases of specific disease categories.
  - **Confusion Matrix:** Showed misclassification of a few categories, indicating potential areas for improvement.
- **Random Forest:**
  - **Accuracy:** 90%
  - **Precision, Recall, F1-Score:** Random Forest outperformed the Decision Tree, especially in cases where the dataset had subtle differences between disease categories.
  - **Confusion Matrix:** Exhibited fewer misclassifications and was more robust compared to the Decision Tree.
- **Comparison:** Random Forest outperformed Decision Tree in terms of overall accuracy and precision. The ensemble nature of Random Forest allows it to generalize better than a single Decision Tree, which tends to overfit.

6. Conclusion

Both the Decision Tree and Random Forest models demonstrated strong performance for the Mango leaf disease classification task. However, the **Random Forest** model proved to be more effective, achieving better accuracy and lower misclassification rates.

Further improvements could include hyperparameter tuning, additional feature engineering, or the use of deep learning techniques for even higher accuracy.