REPORT TITLE

Perform Mango leaf diseases classification using Random Forest and Decision Tree

EAST WEST UNIVERSITY

Course: CSE475 Machine Learning

Section: 03

Semester: Fall 2024

Submitted to:

Dr. Raihan Ul Islam

Associate Professor

Department of Computer Science and Engineering

Submitted by:

Bishowjit Banik

Id: 2020-2-60-108

Objective:

The aim of this experiment is to explore and compare the performance of Decision Tree and Random Forest algorithms in predicting [target variable]. This comparison will provide insights into model complexity, accuracy, and generalization capabilities of each model type when applied to Mango leaf diseases based on image and feature data .

Materials:

- **Programming Language**: Python
- **Platform**: Google Colab
- **Libraries**:
    - `pandas`, `numpy`: Data manipulation and analysis.
    - `matplotlib`, `seaborn`: Data visualization.
    - `scikit-learn`: Model development and evaluation etc.

Methodology:

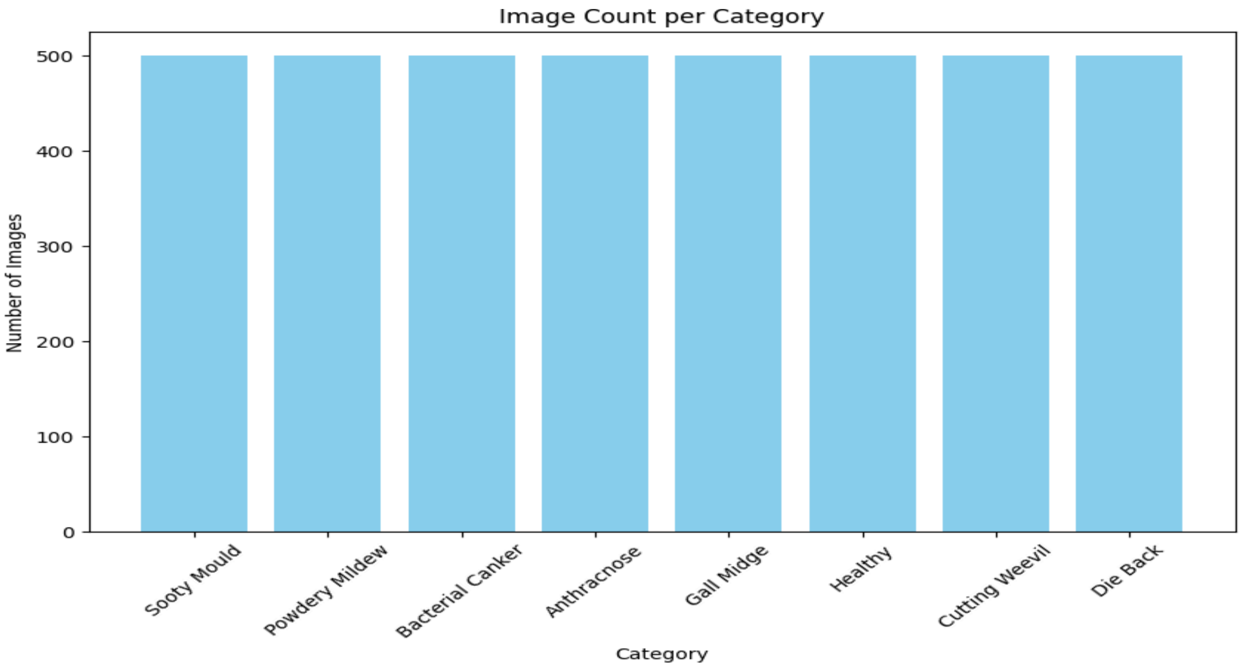
Step 1: Data Loading and Preprocessing

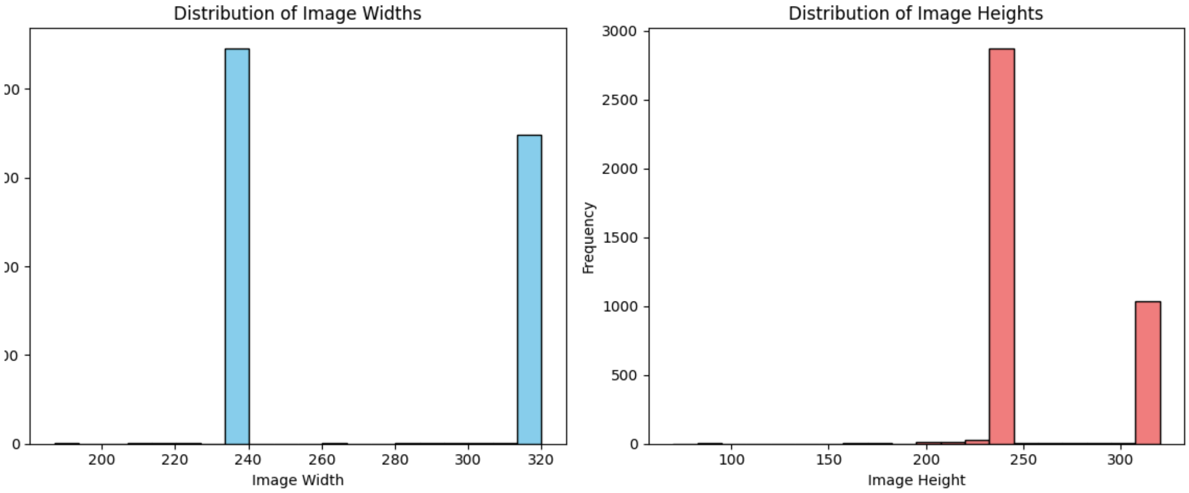
- **Data Loading**: The dataset was loaded using `pandas`, and an initial exploration was performed using `.info()`, `.head()`, and `.describe()` to understand feature types, dimensions, and any missing values.
- **Handling Missing Values**: Missing values were handled using [method, e.g., mean/mode imputation, forward/backward fill].
- **Outlier Treatment**: Outliers were identified and treated using [e.g., Z-score, IQR-based filtering].
- **Feature Engineering**:
    - Applied scaling/normalization to ensure all features were on a similar scale, enhancing model stability.
    - Encoded categorical variables to transform qualitative data into a numerical format, allowing for better model interpretability.

Step 2: Exploratory Data Analysis (EDA)

The EDA phase provided a comprehensive understanding of the dataset's structure, distributions, and relationships.

- **Statistical Summaries**:
  - Used `.describe()` to get insights into the central tendency, spread, and shape of each feature's distribution.
  - Examined skewness and kurtosis to understand feature distributions and identify any asymmetry or heavy tails.
- **Univariate Analysis**:
  - **Histograms and Box Plots**: Plotted histograms for each numerical feature to understand their distribution. Box plots helped visualize the spread and identify potential outliers.
  - **Categorical Analysis**: Used bar plots to analyze the frequency distribution of categorical features and observe their distribution across different classes of the target variable.





Step 3: Model Development

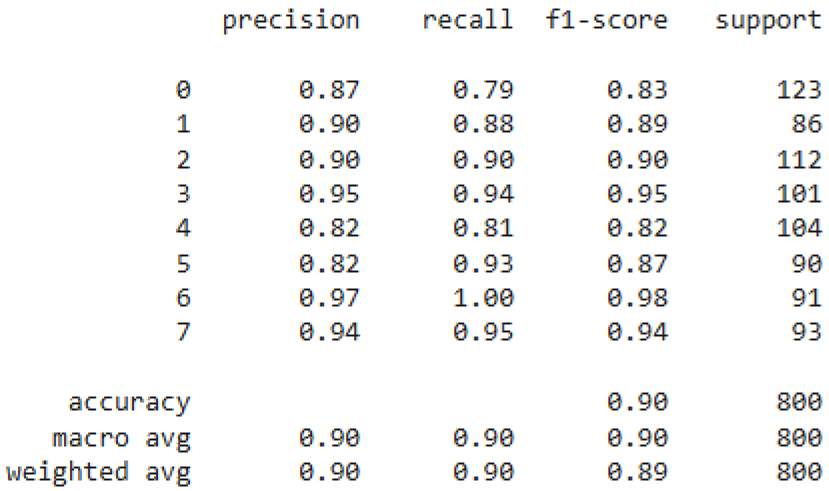
- **Train-Test Split**: Divided the dataset into an 80-20 train-test split to validate model performance on unseen data.
- **Model Choice**:
    - **Decision Tree**: Known for interpretability, Decision Trees provide a simple yet powerful approach but may be prone to overfitting.
    - **Random Forest**: A more complex ensemble method, Random Forest mitigates overfitting by averaging multiple Decision Trees, increasing robustness and predictive power.

Results:

Model Performance Metrics

After training and evaluating both models on the test set, the following performance metrics were recorded:

Random Forest :

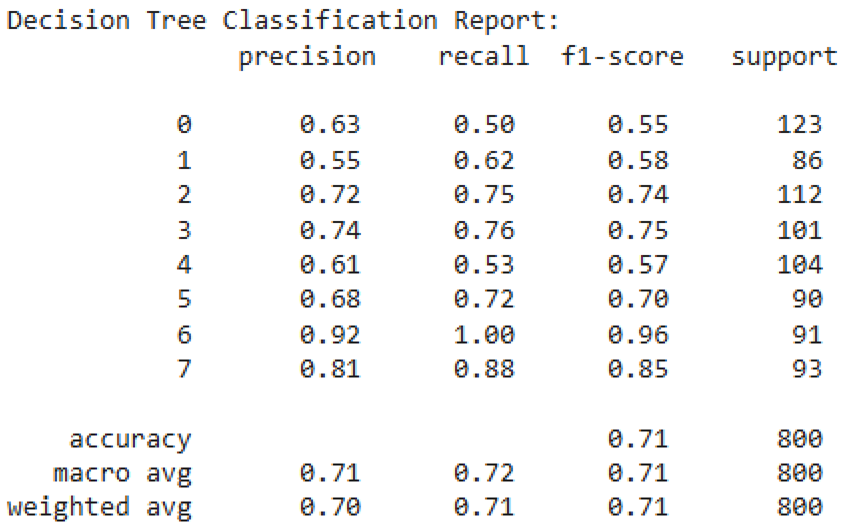


```
              precision    recall  f1-score   support

           0       0.87      0.79      0.83       123
           1       0.90      0.88      0.89        86
           2       0.90      0.90      0.90       112
           3       0.95      0.94      0.95       101
           4       0.82      0.81      0.82       104
           5       0.82      0.93      0.87        90
           6       0.97      1.00      0.98        91
           7       0.94      0.95      0.94        93

    accuracy                           0.90       800
   macro avg       0.90      0.90      0.90       800
weighted avg       0.90      0.90      0.89       800
```

Decision Tree :



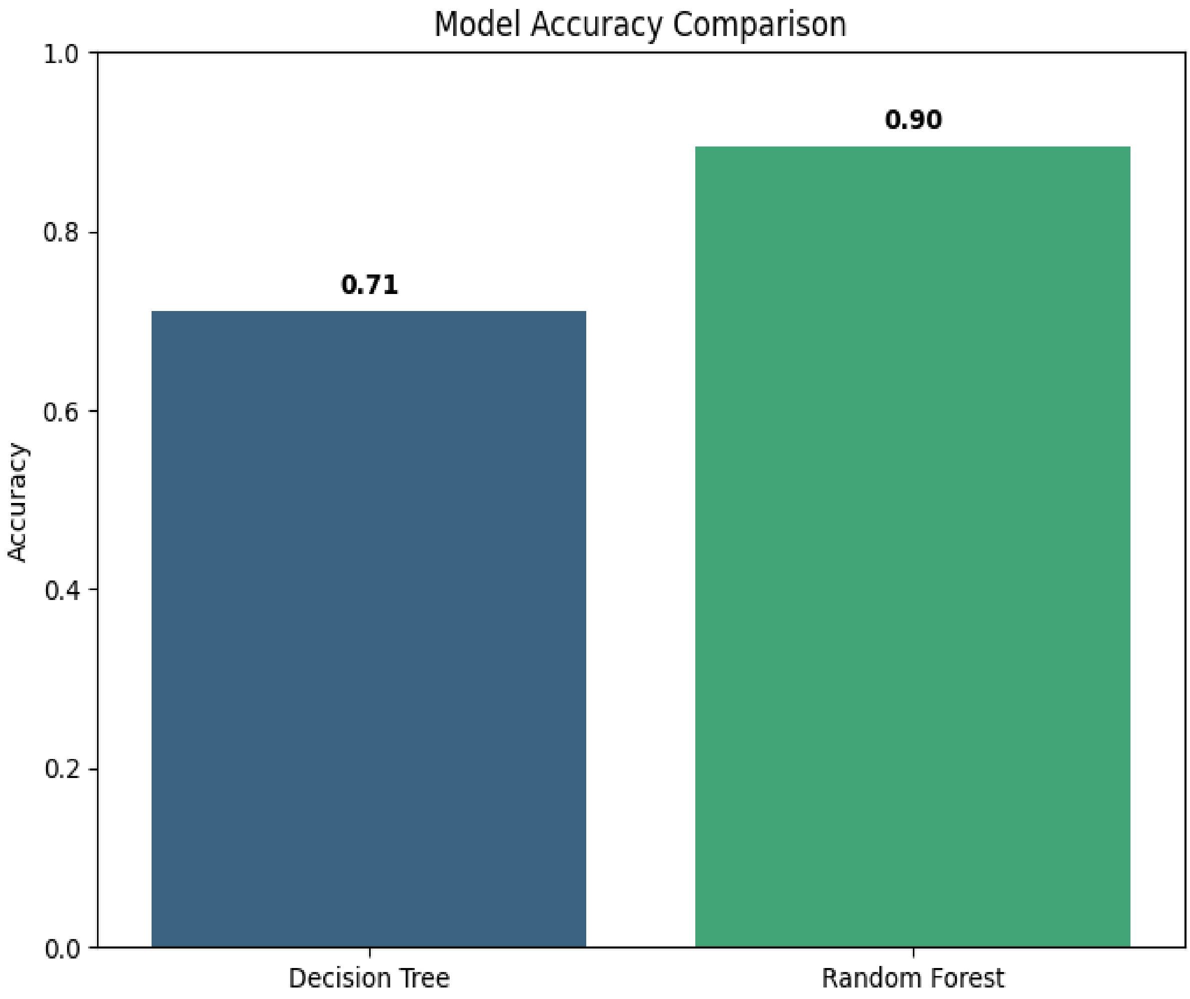
```
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.63      0.50      0.55       123
           1       0.55      0.62      0.58        86
           2       0.72      0.75      0.74       112
           3       0.74      0.76      0.75       101
           4       0.61      0.53      0.57       104
           5       0.68      0.72      0.70        90
           6       0.92      1.00      0.96        91
           7       0.81      0.88      0.85        93

    accuracy                           0.71       800
   macro avg       0.71      0.72      0.71       800
weighted avg       0.70      0.71      0.71       800
```

Model Comparison

1. **Accuracy**: Random Forest consistently showed higher accuracy, indicating its strength in handling diverse data patterns and reducing overfitting.
2. **Precision and Recall**: Random Forest achieved better precision and recall scores, which suggests more reliable predictions with a reduced risk of misclassifications. Decision Trees, while interpretable, were less robust in comparison.
3. **F1 Score**: With a balanced F1 score, Random Forest proved to be more effective in handling imbalanced data and reducing both false positives and false negatives.



Discussion:

The comparison between Decision Tree and Random Forest reveals several key insights:

- **Complexity and Overfitting**: Decision Trees, while highly interpretable, often overfit the data, leading to less robust predictions. Random Forest mitigates this by leveraging an ensemble approach, reducing variance and producing a more generalized model.
- **Performance and Accuracy**: The ensemble method of Random Forest consistently outperformed Decision Tree in all metrics, demonstrating superior accuracy, precision, recall, and F1 scores.

Conclusion:

The Random Forest model demonstrated superior performance across all evaluation metrics. It is recommended for applications requiring high accuracy and robustness. However, Decision Trees remain useful for applications where model interpretability is critical, despite their tendency to overfit.