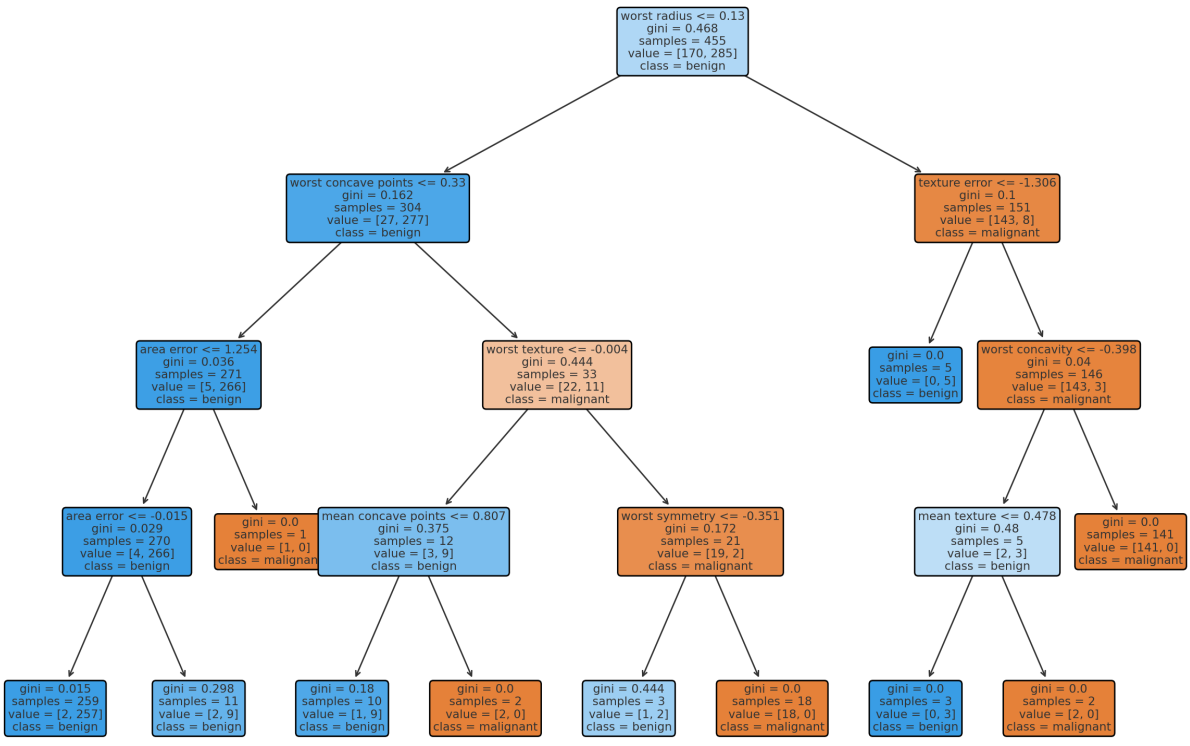
Decision Tree & Random Forests — Internship Task

Objective: Train Decision Tree and Random Forest classifiers; visualize the tree; analyze overfitting; evaluate with cross-validation.

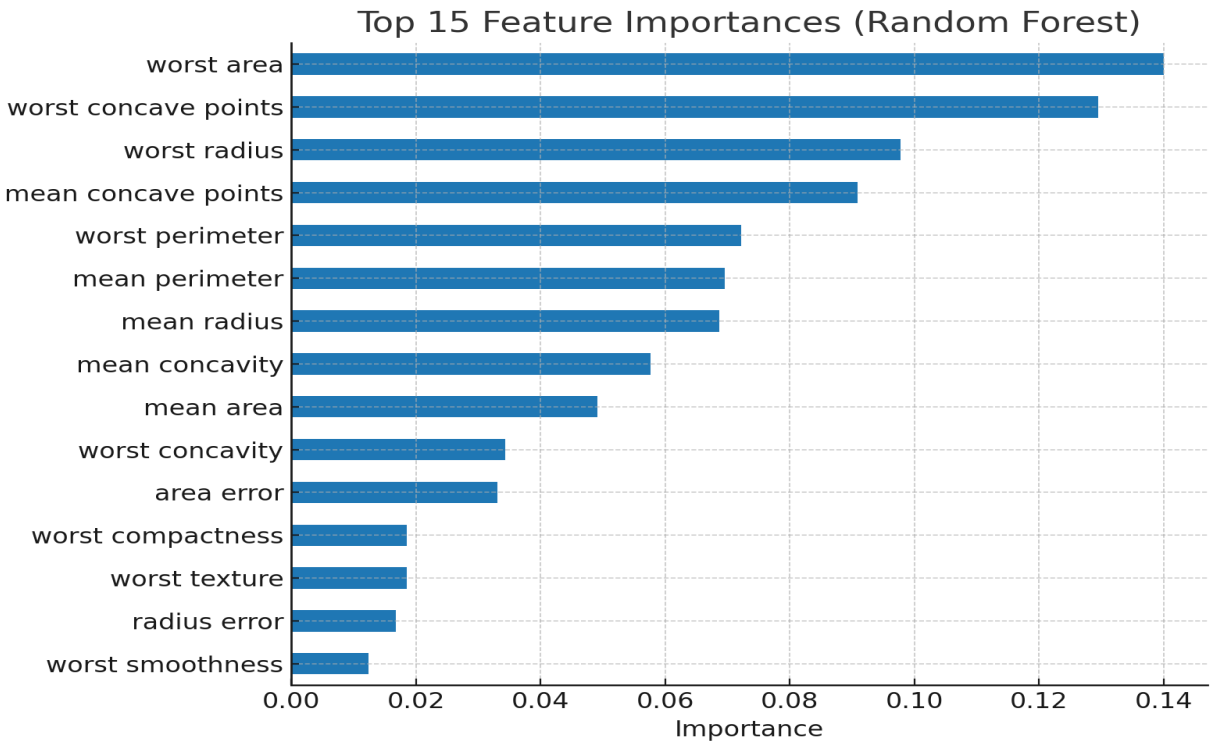
Steps performed: 1. Loaded the Breast Cancer dataset from scikit-learn. 2. Standardized features and split into train/test (80/20). 3. Trained Decision Tree (max_depth=4) and visualized the tree. 4. Trained Random Forest (100 trees) and computed feature importances. 5. Performed 5-fold stratified cross-validation for both models. 6. Saved plots and summarized results.

Results:
Decision Tree test accuracy: 0.9386
Random Forest test accuracy: 0.9561
Decision Tree CV mean accuracy: 0.9227 ± 0.0295
Random Forest CV mean accuracy: 0.9561 ± 0.0123
Decision Tree Confusion Matrix:
[[39, 3], [4, 68]]
Random Forest Confusion Matrix:
[[39, 3], [2, 70]]

Decision Tree Visualization:



Top Feature Importances (Random Forest):



Interview Questions & Answers (short):

1. **How does a decision tree work?** - Splits data using feature thresholds to create branches that increase purity.

2. **What is entropy and information gain?** - Entropy measures impurity; information gain is decrease in entropy after a split.

3. **How is random forest better?** - Aggregates many trees trained on bootstrap samples and random feature subsets; reduces overfitting and variance.

4. **What is overfitting?** - When a model learns noise; prevented by limiting depth, pruning, using ensembles, or cross-validation.

5. **What is bagging?** - Bootstrap aggregating: training multiple models on random samples and averaging predictions.

6. **How to visualize a tree?** - Use `plot_tree` or export to Graphviz format.

7. **Interpret feature importance?** - Higher importance means the feature contributed more to reducing impurity across trees.

8. **Pros/Cons of RF?** - Pros: robust, high accuracy. Cons: less interpretable, heavier compute.