Task 5: Decision Trees and Random Forests Report

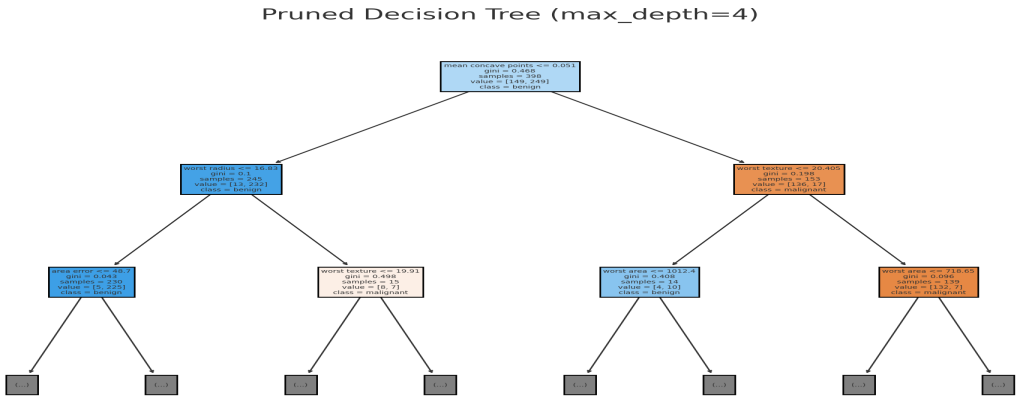
Dataset: Breast Cancer (sklearn.datasets)

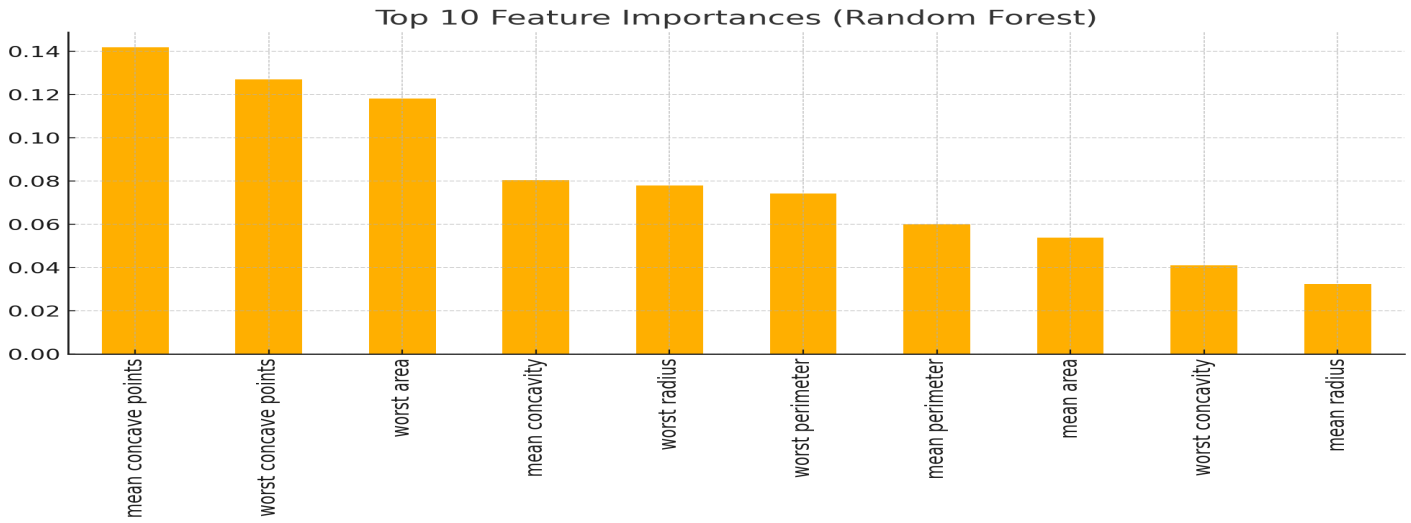
Unpruned Decision Tree Accuracy: Train = 1.00, Test = 0.94

Pruned Decision Tree (max_depth=4) Test Accuracy: 0.95

Random Forest Test Accuracy: 0.97

Random Forest 5-Fold CV Accuracy: 0.96 ± 0.02





1. How does a decision tree work?
A decision tree splits data by selecting features and thresholds that maximize class separation or regression accuracy. S

2. What is entropy and information gain?
Entropy measures impurity: -Σ p(i) log2 p(i). Information gain is the reduction in entropy after a split: IG = H(parent) -

3. How is random forest better than a single tree?
Random forests build many decision trees on bootstrap samples and random feature subsets. Averaging or voting reduc

4. What is overfitting and how do you prevent it?
Overfitting occurs when a model captures noise. Prevention in trees: prune the tree, limit max depth, set min samples p

5. What is bagging?
Bagging (bootstrap aggregating) trains models on different bootstrap samples and averages predictions to reduce varian

6. How do you visualize a decision tree?
Use tree plotting libraries (e.g., sklearn.tree.plot_tree, Graphviz). Display nodes, feature splits, thresholds, and class/val

7. How do you interpret feature importance?
Feature importance is the total reduction in impurity brought by that feature across all trees, normalized so importances

8. What are the pros/cons of random forests?
Pros: Robust to overfitting, handles high-dimensional data, provides feature importance. Cons: Less interpretable than s