CSE 472 Assignment 2: Decision Trees, Random Forests, and Extra Trees

Department of Computer Science and Engineering

Introduction

Decision trees are intuitive and powerful models but often suffer from high variance and overfitting. Ensemble methods such as Random Forests and Extremely Randomized Trees (Extra Trees) address these issues by combining multiple trees using different sources of randomness.

In this assignment, you will implement tree-based learning algorithms from scratch and empirically analyze how ensemble methods improve generalization performance. You will also compare your implementations against optimized versions available in `scikit-learn`.

Objectives

By completing this assignment, you are expected to:

- Understand the working principles of decision trees
- Implement ensemble techniques such as Random Forests and Extra Trees
- Analyze bias–variance tradeoffs empirically
- Compare custom implementations with industrial-grade libraries

Problem Statement

You are required to implement the following algorithms **from scratch**:

1. Decision Tree
2. Random Forest
3. Extremely Randomized Trees (Extra Trees)

Your implementations must support:

- Classification

You must then compare:

- Decision Trees vs ensemble methods
- Your implementations vs `scikit-learn` implementations

Datasets

The following datasets may be used for development and experimentation.

- Iris
- Wine

All datasets are available through `sklearn.datasets`.

Implementation Rules

- You must **not** use any tree-based model from `scikit-learn` in your custom implementation.
- You may use `numpy` and optionally `pandas`.
- All randomness must be controlled using a fixed random seed.
- The following hyperparameters must be configurable:
  - Maximum tree depth
  - Minimum samples per split
  - Number of trees
  - Number of features considered per split

Evaluation Metrics

Classification:

- Accuracy
- F1-score
- AUROC

Required Comparisons

For each dataset, results must be reported for:

- Custom Decision Tree
- Custom Random Forest
- Custom Extra Trees
- `scikit-learn` Decision Tree
- `scikit-learn` Random Forest
- `scikit-learn` Extra Trees

Results should be presented in clearly labeled tables.

Marks Distribution

Component	**Marks**
Decision Tree implementation	15
Random Forest implementation	20
Extra Trees implementation	20
Evaluation on datasets	20
Comparison with sklearn models	15
Report and Analysis	10
Total	**100**

Submission Instructions

Submit a single compressed file containing:

- All source code
- A PDF report describing:
  - Algorithmic details
  - Experimental setup
  - Results and analysis

Late submissions will not be accepted.

Academic Integrity

This assignment must be completed individually.

- Discussion of ideas is allowed
- Sharing code is strictly prohibited
- Any external resources must be cited

Violation of these rules will result in disciplinary action.

Final Remarks

Write clean, modular, and well-documented code. Your implementation will be evaluated for correctness, robustness, and clarity. **You will not get any marks on any code that you fail to explain.**