Assignment 1: Decision Tree Classifier on Iris Dataset

CS564

1 Objective

The goal of this assignment is to implement a **Decision Tree Classifier from scratch** and evaluate it on the **Iris dataset**. You are expected to:

- Understand the working of decision trees for classification.
- Implement impurity-based splitting criteria: *Gini Index, Entropy, and Misclassification Error*.
- Analyze the impact of tree depth on underfitting and overfitting.
- Evaluate performance using appropriate classification metrics.

2 Dataset

The assignment will use the **Iris dataset** available at the UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data`

Details

- **Attributes (features):**
  1. Sepal length (cm)
  2. Sepal width (cm)
  3. Petal length (cm)
  4. Petal width (cm)
- **Target (class):**
  1. Iris-setosa
  2. Iris-versicolor
  3. Iris-virginica

This is a 3-class classification problem.

3 Tasks

Part A: Implementation of the Decision Tree

- Implement a class `DecisionTreeClassifierScratch` with methods:
  - `fit(X, y)`: Build the tree recursively.
  - `train_test_split(X, y,test_ratio)`: Randomly split the dataset, assigning len(X) $*$ test_ratio into test set and remaining as train set
  - `predict(X)`: Predict labels for given samples.
  - `print_tree()`: Display the structure of the learned tree.
- Implement the following splitting criteria:
  1. Gini Index
  2. Entropy (Information Gain)
  3. Misclassification Error
- Handle continuous attributes by identifying the best threshold for splits.
- Implement stopping conditions:
  - Stop when all records at a node belong to the same class.
  - Stop when the maximum tree depth is reached.
  - Stop when the number of samples at a node is below a minimum threshold.

Part B: Model Evaluation

- Divide the dataset into training (70%) and testing (30%).
- Train and test the decision tree using each impurity measure.
- Evaluate performance using:
  - Confusion Matrix
  - Accuracy
  - Precision, Recall, and F1-score

Part C: Effect of Tree Depth

- Train the classifier with different maximum depths: {1, 2, 3, 5, unlimited}.
- Plot training and testing accuracy against tree depth.

Part D: Analysis and Discussion

- Which impurity measure provided the best classification performance?
- How does increasing depth affect generalization?
- What is the simplest tree depth that achieves good performance on the test set?

4 Experimental Settings

- Programming Language: **Python**
- Allowed Libraries: **NumPy, Pandas, Matplotlib**
- Restriction: **Do not use** `sklearn.tree.DecisionTreeClassifier` or any equivalent built-in tree implementation.
- The function `train_test_split` from `sklearn.model_selection` is permitted.
- Use random seed

5 Deliverables

A Jupyter Notebook (`.ipynb`) containing

- Implementation.
- Results table comparing impurity measures.
- Plots of accuracy vs. depth.
- Short discussion of results and conclusions.
- Note: All the outputs, tables, analysis must be clearly visible in the notebook. DO NOT clear the cell outputs before submission