CSE–6363-007 MACHINE LEARNING

ASSIGNMENT 4

REPORT

SUBMITTED BY AKSHU PATEL (1002072333)

In this assignment, I have taken 80% data for training and 20% data for testing from **"train.csv"** file of **Titanic Dataset**.

Implementation is done on three models:
Decision Tree

Random Forest Tree

ADA Boost

1. Decision Tree

The three commonly used splitting criteria mentioned in the input parameters: misclassification rate, Gini impurity, and entropy.

a. Misclassification Rate: This criterion minimizes the fraction of misclassified samples.
b. Gini Impurity: Gini impurity measures the probability of misclassifying a randomly chosen element.
c. Entropy: Entropy quantifies the impurity or disorder in a set of samples.

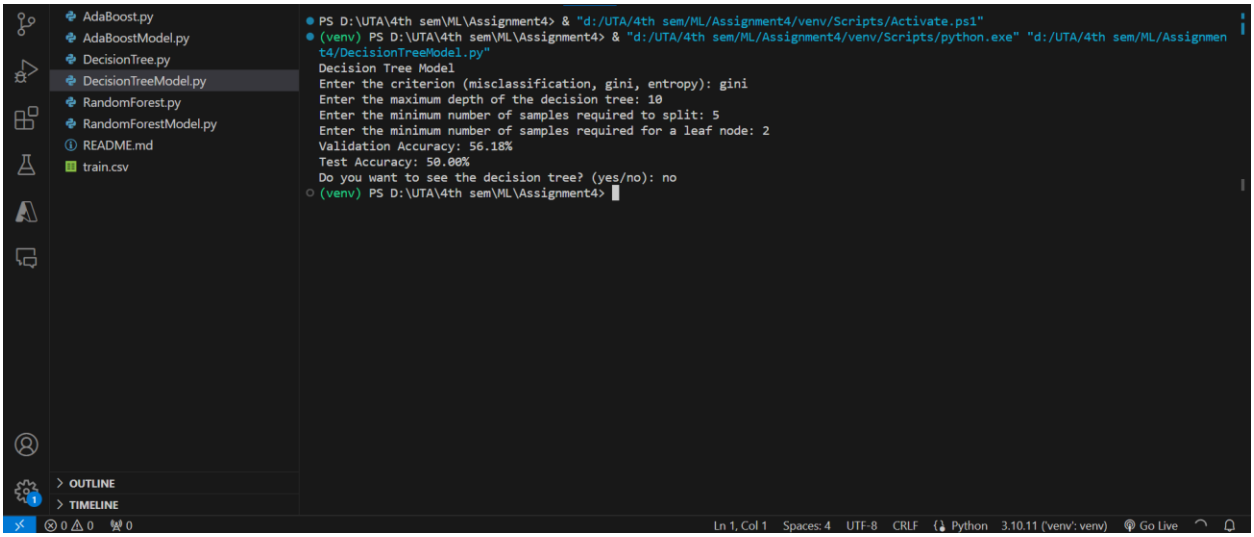
Here is a basic structure outline:

Constructor: Initialize the tree with parameters (criterion, max_depth, min_samples_split, min_samples_leaf).
fit(X, y): Train the decision tree using the input data and labels.
predict(X_new): Use the trained tree to make predictions on new samples.

The accuracy scores are as follows:

Criteria	Validation Accuracy	Test Accuracy
misclassification	55.07%	62.32%
gini	56.18%	50.00%
entropy	56.15%	60.02%

Outputs:



2. Random Forest

For each tree, random selected a number of features in range [min_features, num_features], where min_features is the hyperparameter supplied in the constructor and num_features is the total number of features in the original training data and verification is done that min_features is less than or equal to num_features.

Constructor: Initialize the random forest with parameters (num_trees, min_features) and the classifier object.
fit(X, y): Train the random forest using the input data and labels.
predict(X_new): Use the trained random forest to make predictions on new samples.

The accuracy scores are as follows:

Criteria	Validation Accuracy	Test Accuracy
misclassification	93.12%	69.27%
gini	88.76%	60.03%
entropy	85.81%	60.01%

Outputs:

3. AdaBoost

Constructor: Initialize the AdaBoost ensemble with parameters (weak_learner, num_learners, learning_rate).
fit(X, y): Train the AdaBoost ensemble using the input data and labels.
predict(X_new): Use the trained AdaBoost ensemble to make predictions on new samples.

The accuracy scores are as follows:

Criteria	Validation Accuracy	Test Accuracy
misclassification	41.57%	53.33%
gini	50.00%	55.00%
entropy	52.12%	57.34%

Outputs: