Assignment 1: Decision Tree

- **Members (GR-G, 65)**
  - **Name:** Surajit Kundu, **Roll No:** 21MM91R09
  - **Name:** Ankur Kumar Jaiswal, **Roll No:** 21MM62R08

1. Build a decision-tree classifier by randomly splitting the dataset as 80/20 split. Use the impurity measures- 1) gini index and 2) information gain. Analyze the impact of using individual impurity measures on the prediction. Do not use a package for building the tree and implement this part on your own.

Response:

The decision tree is built on impurity measures, both gini index and information gain. Using entropy, we are getting an accuracy of 89.74 %. And using the gini index, we are getting an accuracy of 87.17%. So, we can conclude that impurity measure - entropy is better than impurity measure - gini index to build the decision tree of the given data set in our study.

2. Provide the accuracy by averaging over 10 random 80/20 splits. Consider that particular tree that provides the best test accuracy as the desired one.

Response:

The accuracy by averaging over 10 random 80/20 splits is 86.98% using entropy measure and 84.45% using Gini index. Again, we can conclude that impurity measure - entropy is better than impurity measure - gini index in the built decision tree of the given data set in our study.

The best test accuracy we get is 94.11%.
The decision tree is

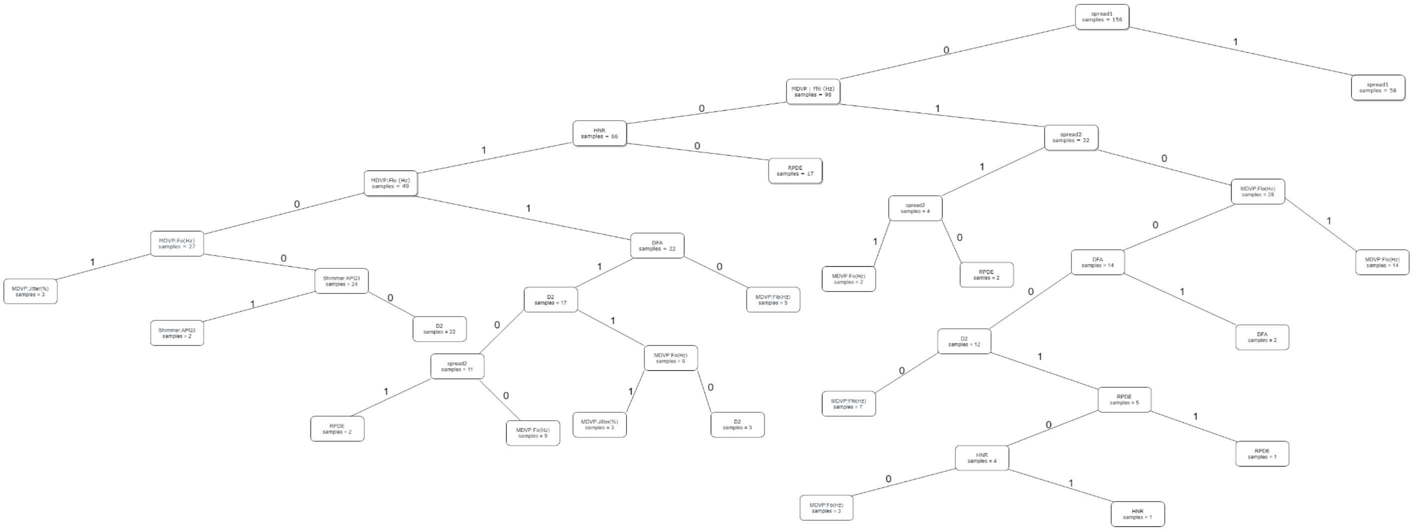


Figure 1: Decision tree with highest accuracy

3. What is the best possible depth limit to be used for your dataset? Provide a plot explaining the same. Also, provide a plot of the test accuracy vs. the total number of nodes in the trees.

Response:

The best possible depth limit to be used for your dataset is 8.

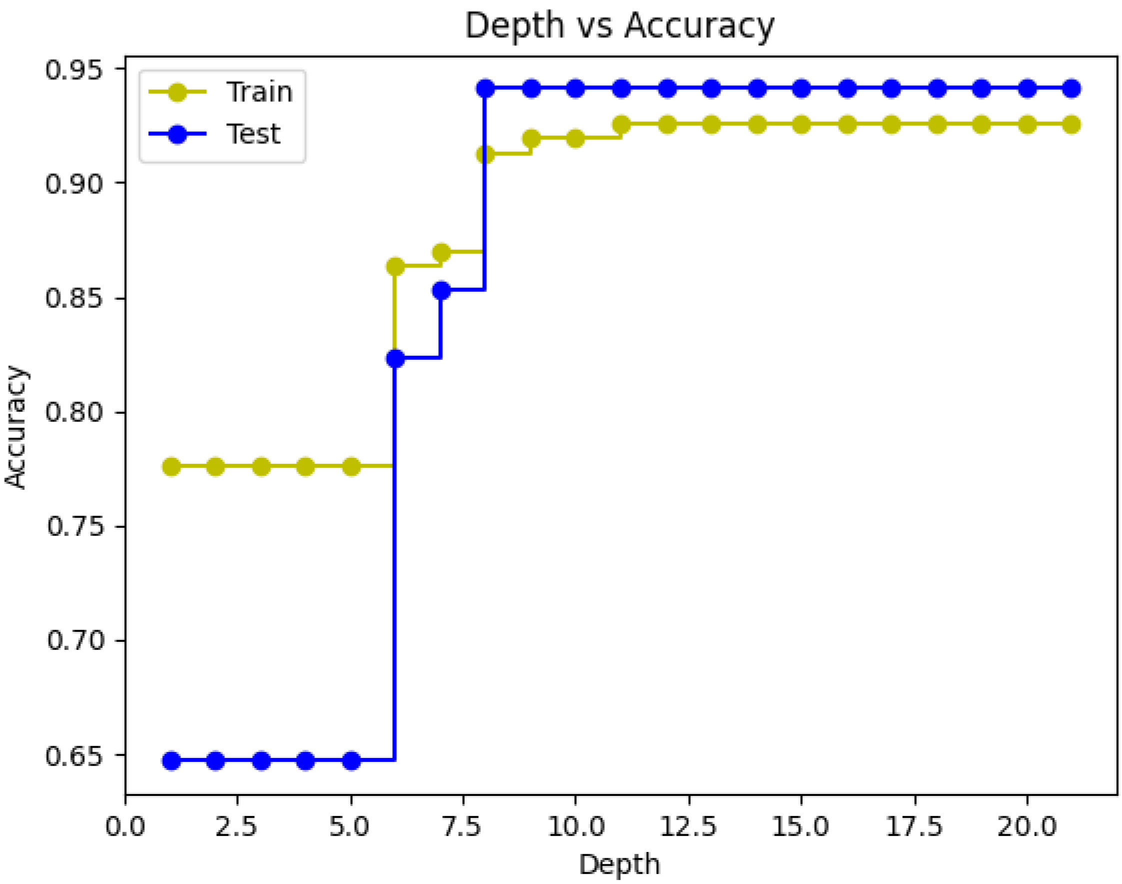
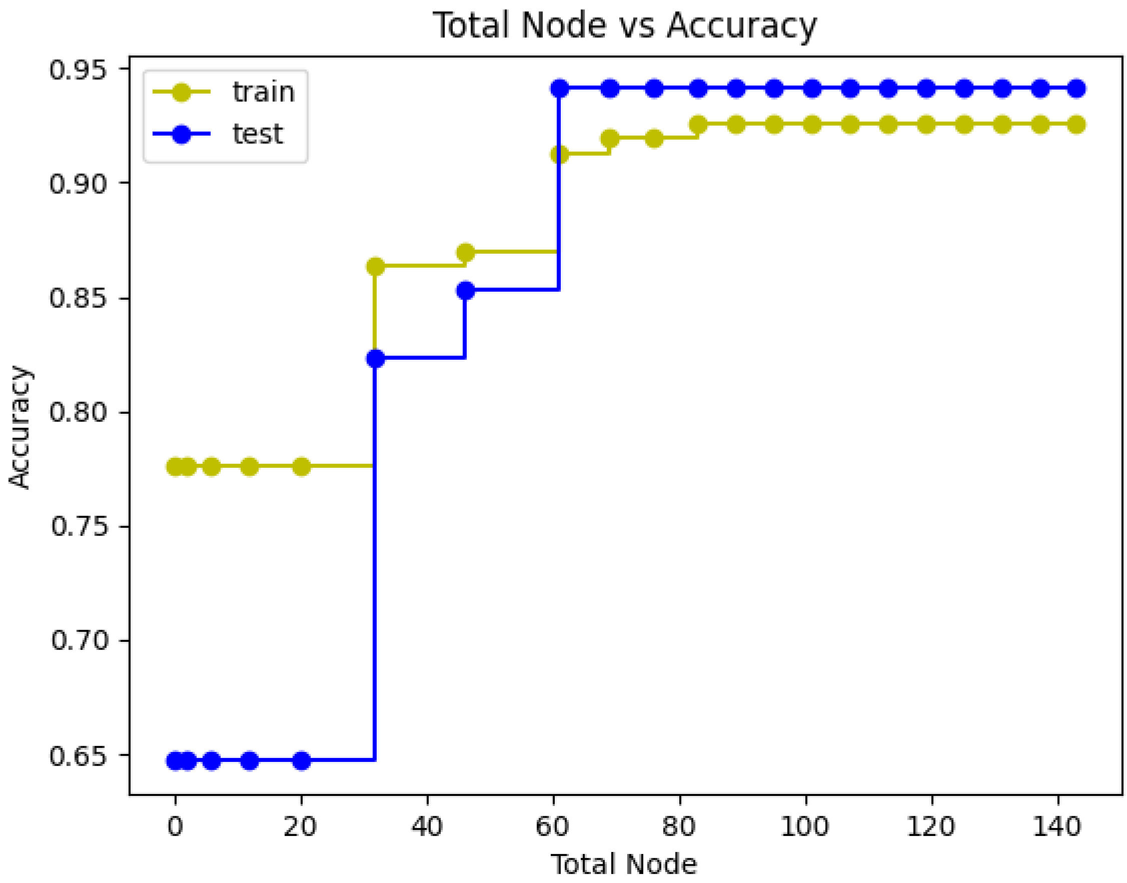


Figure 2: Tree depth vs accuracy



.

Figure 3: Total Node vs accuracy

As shown in Figure 2, We can see a graph between test accuracy and the depth of the decision tree. After analyzing the plot, we can conclude that as we increase the decision tree's depth, the accuracy of the data is also growing. But as we can see in the graph, after a certain depth(i.e., around 6) of the trees, the accuracy of the data becomes constant. So, **the best possible depth limit of the decision tree to be used for your dataset is 8.**

As shown in Figure 3, We can see a graph between test accuracy and the number of decision tree nodes. After analyzing the plot, we can conclude that as we increase the decision tree's nodes, the accuracy of the data is also growing. But as we can see in the graph, after a specific node of the trees, the accuracy of the data becomes constant.

4. Perform the pruning operation over the tree with the highest test accuracy in question 2 using a valid statistical test for comparison.

Response :

We have taken the same model where we achieve the highest accuracy over 10 random 80/20 splits. Then the pruning operation is performed on the node impurity. We set a list of threshold values for impurity measures. The threshold value decides when a new node is created. As shown in Figure 4, we take a minimum threshold value of impurity and increase its value continuously, then observe the test accuracy. And the test accuracy is maximum at a threshold value of 0.04.

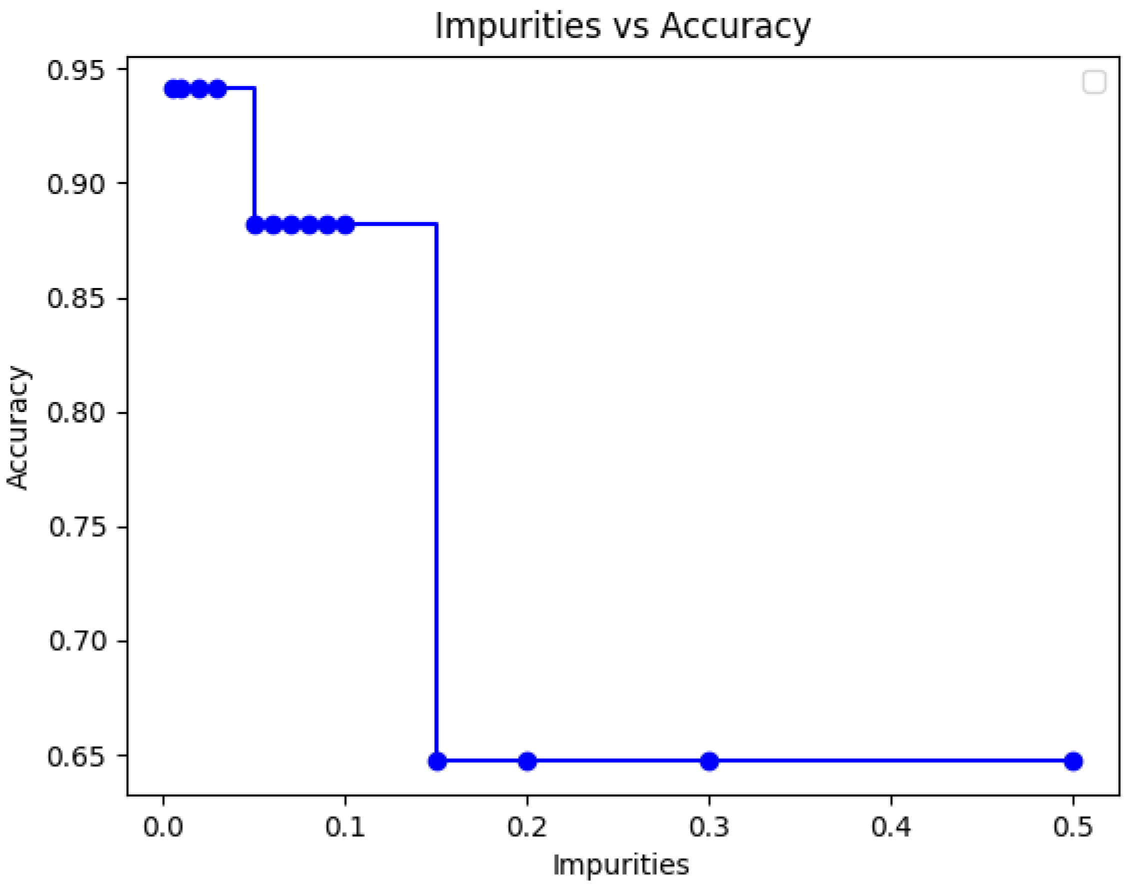


Figure 4: Accuracy vs Impurities

5. Print the final decision tree obtained from question 3 following the hierarchical levels of data attributes as nodes of the tree.

Response : Refer the below figure for the final decision tree with the maximum depth of the dataset.

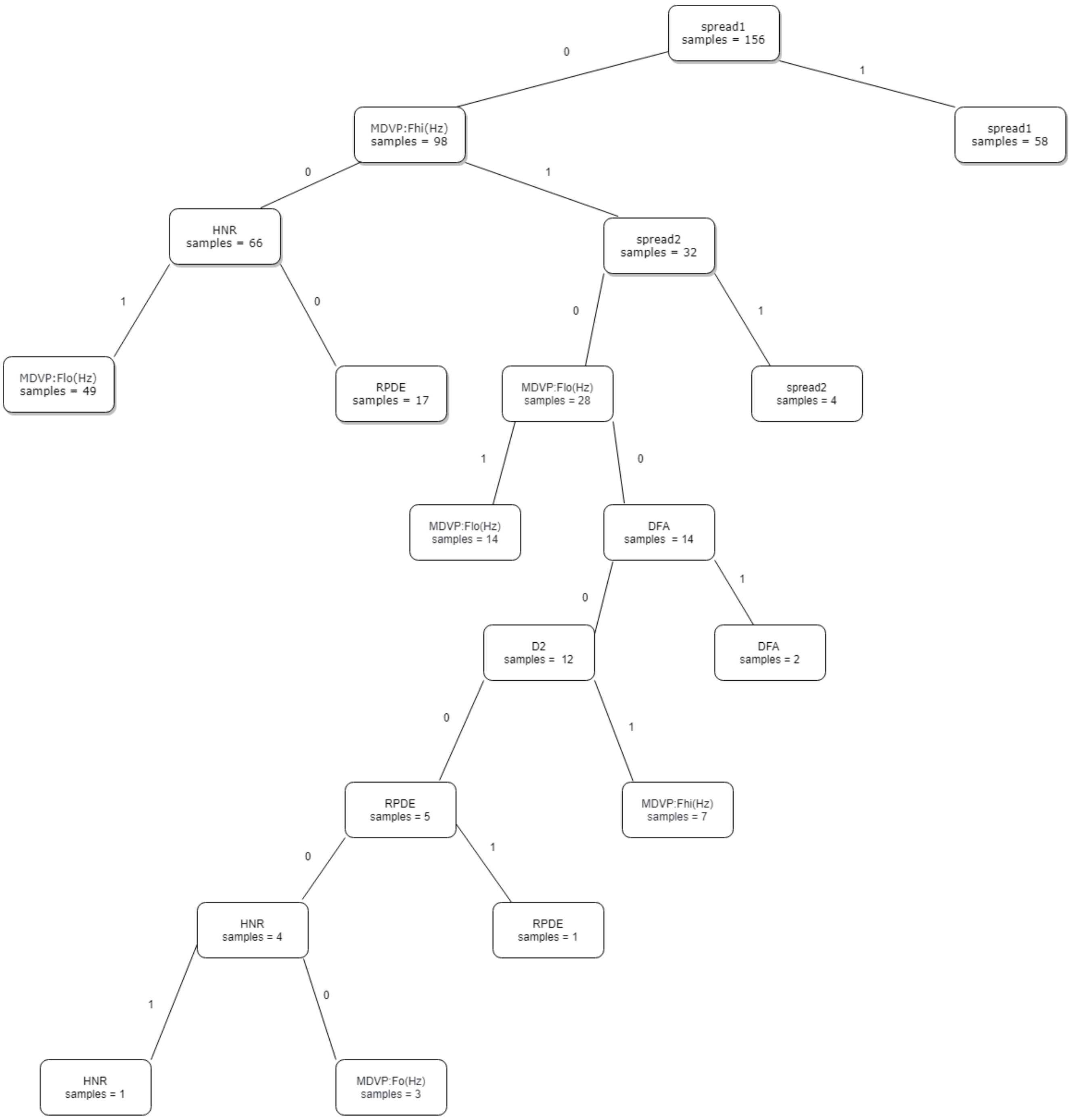


Figure 5 : Pruned Decision Tree

6. A brief report explaining the procedure and the results.

Response : Please refer to the document “Procedure and Results.pdf ” for procedure and results