Name – Sameer Vivek Pande
Entry Number – 2017CS10371

Assignment 4 – Report

For this assignment I had used two types of impurity measures:

a) Entropy
b) Gini Index

The way of splitting the data for continuous variables was set to **Median Splitting**

For discrete variables, each node was split into **number of distinct values the discrete variable** could take.

Functions used to decide which feature to split upon were

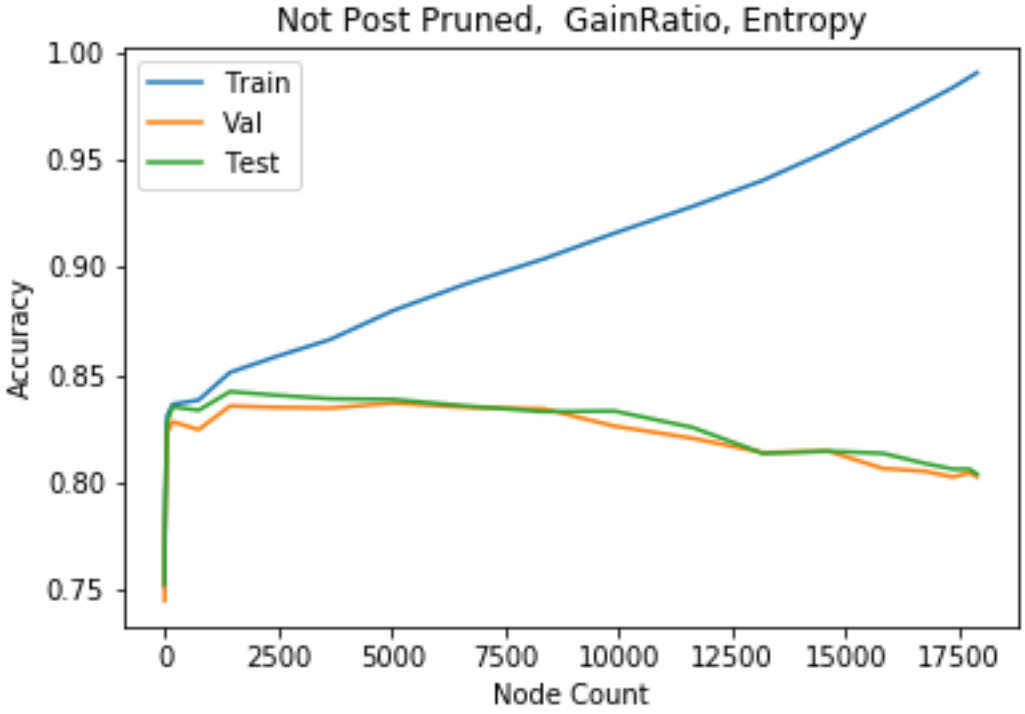
a) Information Gain
b) Gain Ratio

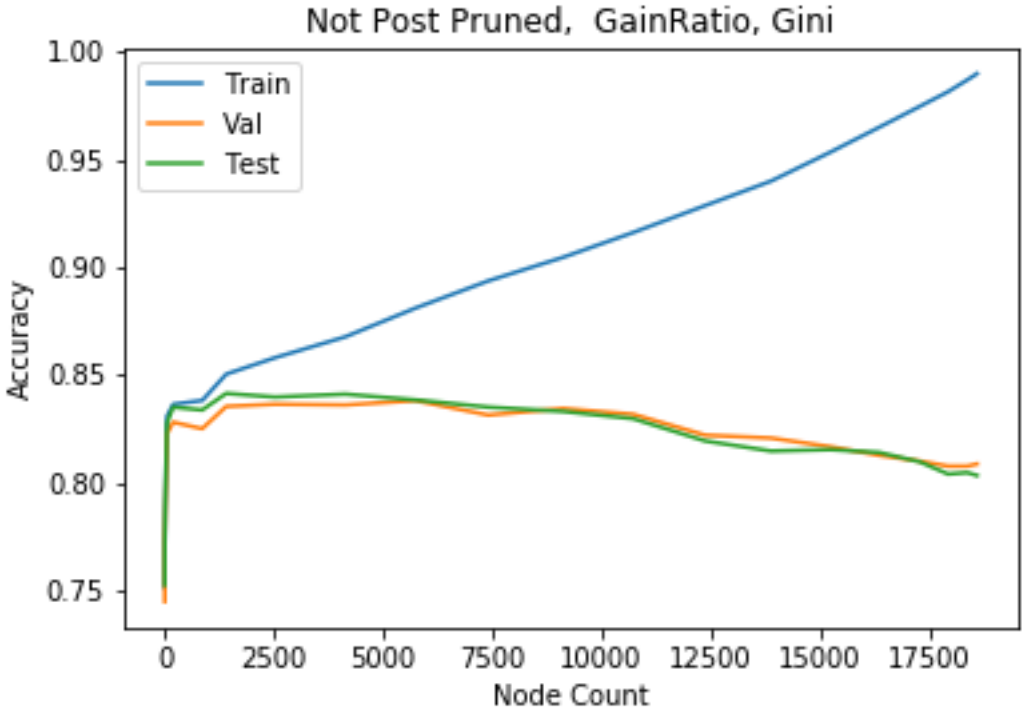
Part 1 : - Before Pruning

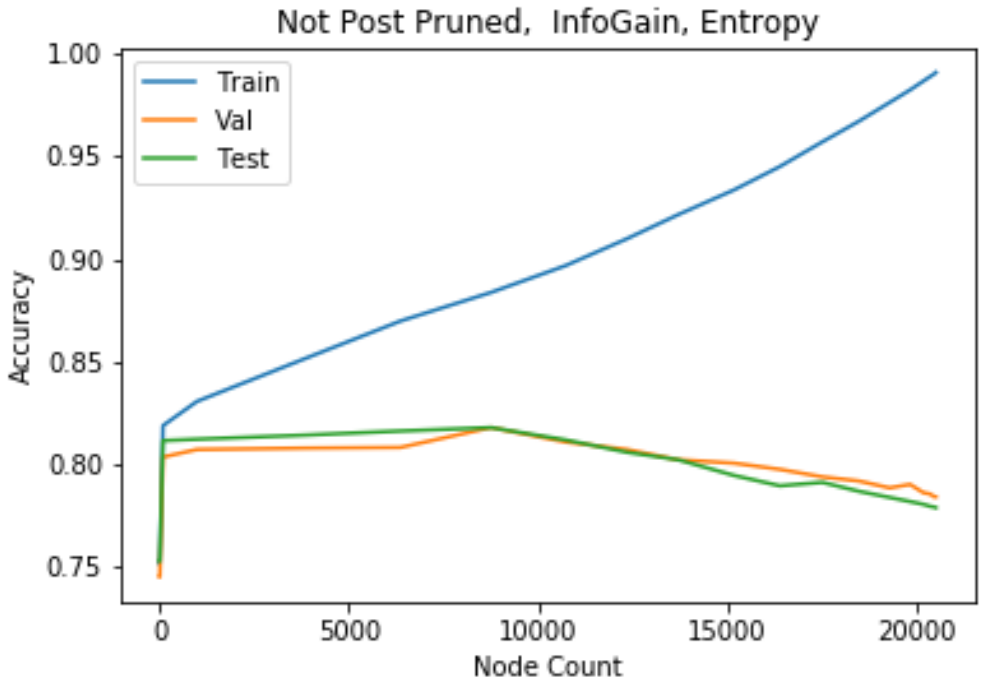
In this part, trees are created using a “maxDepth” parameter that is being passed while creating tree using “createTree()” function. It means that decision trees of different depths were created to obtain certain observations. Since the depth of the tree was varied the “Node Count” was automatically varied.

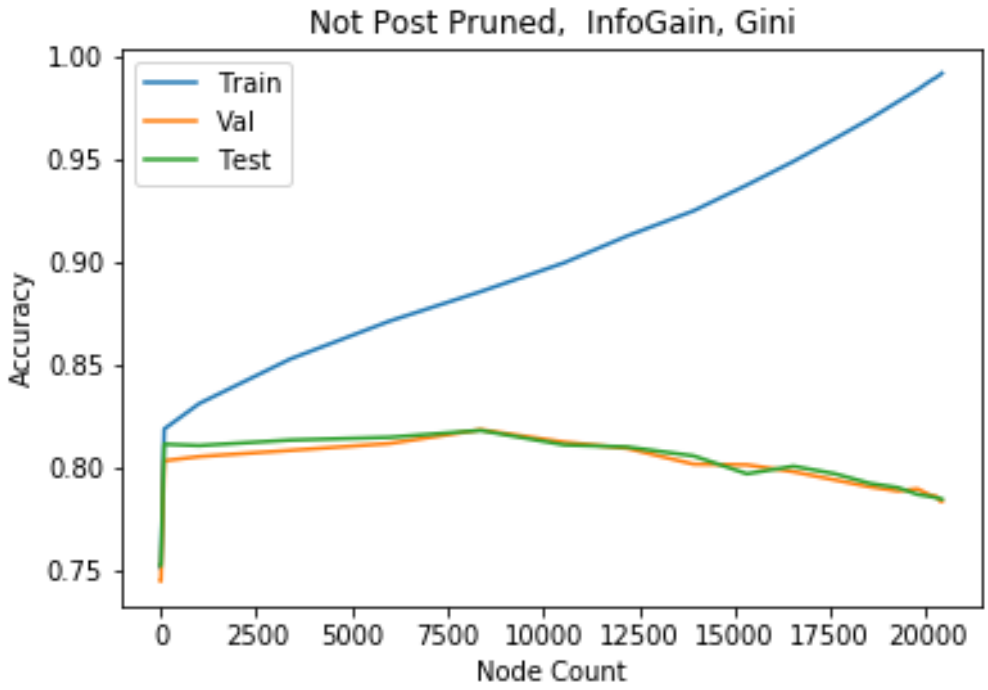
Here are the results of some experiments:

(Data was generated for this part by creating trees till depth = 21, without Pruning)


Not Post Pruned, GainRatio, Entropy
Train
Val
Test
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0
2500
5000
7500
10000
12500
15000
17500
Node Count


Not Post Pruned, GainRatio, Gini
Train
Val
Test
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0
2500
5000
7500
10000
12500
15000
17500
Node Count



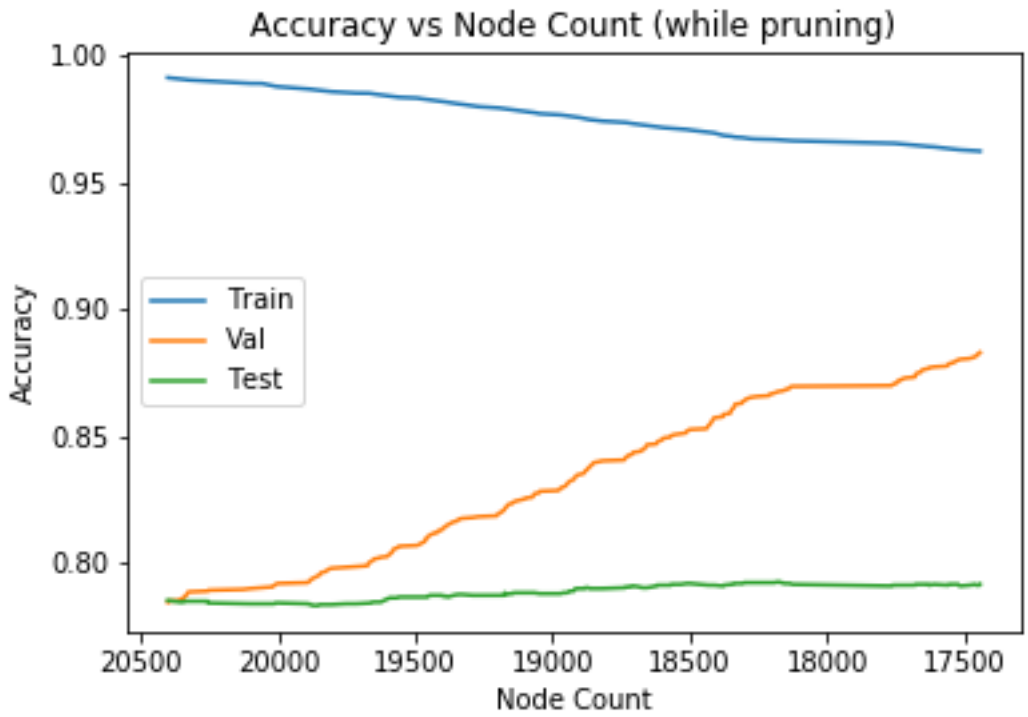


Observations:

1.) Gain Ratio gives better performance than Information Gain since we can observe that the number of labels were distributed in ratio approximately 3:1 in the train data. Hence Gain Ratio which accounts for such cases improves the accuracy

2.) Entropy and Gini show almost similar results for this dataset

Part 2 : - Post – Pruning



Whole tree was created using “createFullTree()” function in the code.

Data was recorded for the entire train, test and validation data sets each time a node was merged.

The whole data was passed through the entire modified decision tree every time a node was merged.

(The whole data has been recorded in single pruning of the Decision Tree)

As we move right along x-axis, Node Count decreases which signifies pruning of nodes.

For the given tree the height of the fully grown tree (before post- pruning) was **20**.