SMAI Assignment-1 Report

Part-1

Result: ACCURACY%: 76.20106761565836
RECALL: 0.1865671641791045
PRECISION: 99.99999999999997
F1-score: 0.3724394785847305

Observation: Accuracy came out to be 76.2% only as we used only categorical data for building the tree, this is because the real decision may depend upon numerical data too.

Part-2

Result: ACCURACY%: 97.68683274021352
RECALL: 95.80152671755725
PRECISION: 94.3609022556391
F1-score: 95.07575757575758

Observation: Accuracy came out to be 97.7% as we used the complete feature set for building the tree and the validation set gave a great result as compared to the previous part.

Part-3

Result:

Entropy: $i(V) = -\big(q \log q + (1-q)\log(1-q)\big)$

Gini index: $i(V) = 2q(1-q)$

Misclassification rate: $i(V) = \min(q, 1-q)$

Misclassification Rate

ACCURACY%: 97.06405693950178
RECALL: 89.69465648854961
PRECISION: 97.5103734439834
F1-score: 93.4393638170974

Gini Index

ACCURACY%: 97.41992882562278
RECALL: 96.31782945736434
PRECISION: 92.72388059701493
F1-score: 94.48669201520912

Entropy

ACCURACY%: 97.68683274021352
RECALL: 95.80152671755725
PRECISION: 94.3609022556391
F1-score: 95.07575757575758

Observation: Comparing the result from the three we observed the best result in case of the Gini Index and Entropy.

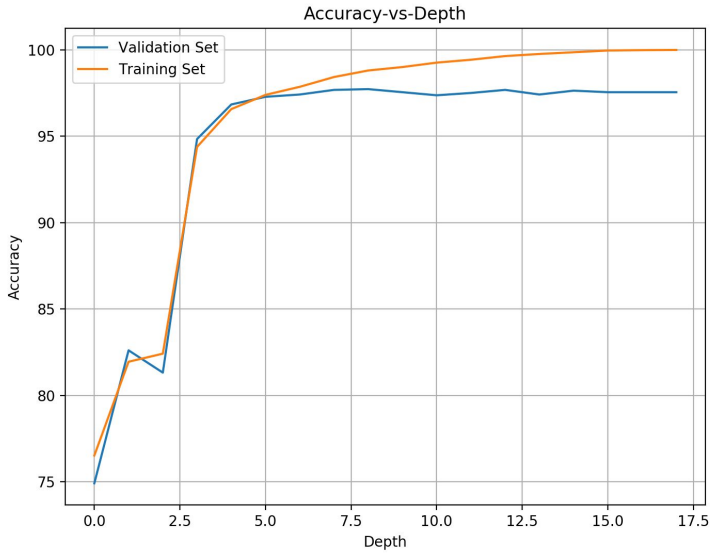
Part-4

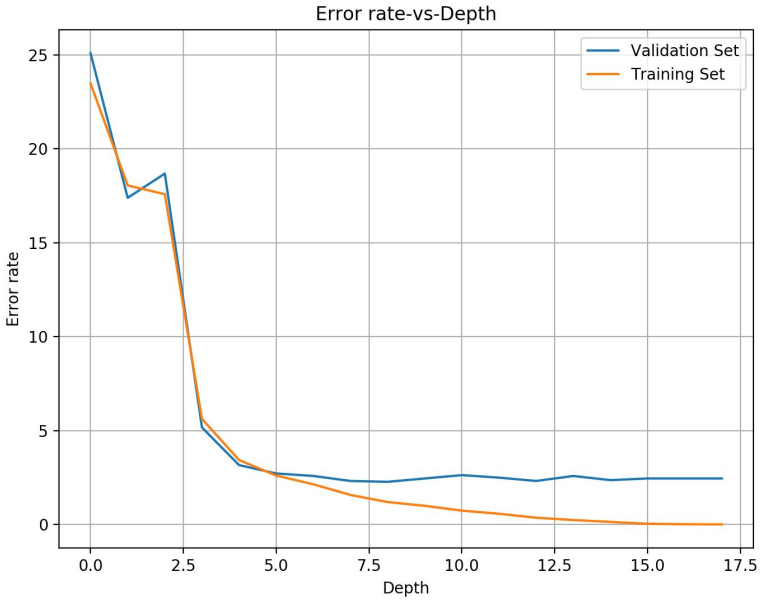
Result:

Observation: For visualizing the Training Data set I plotted all combinations of attributes on the x-axes and y-axes and found out the best data visualization I got from satisfaction_level vs last_evaluation plot.

Part-5.1

Result:

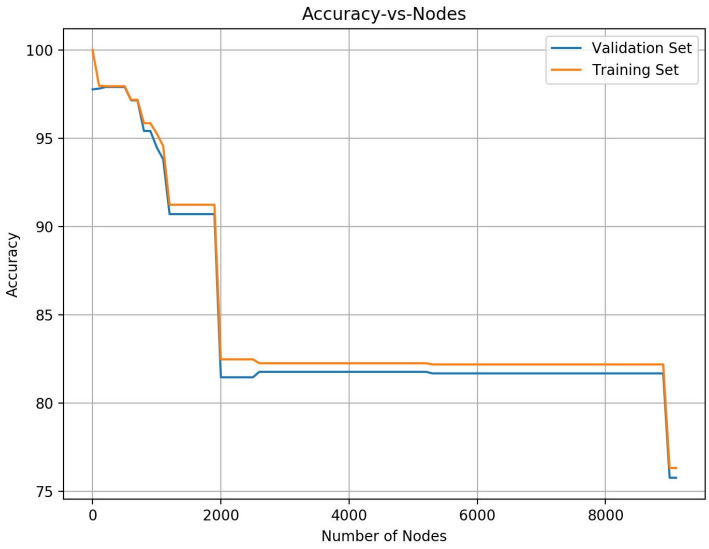


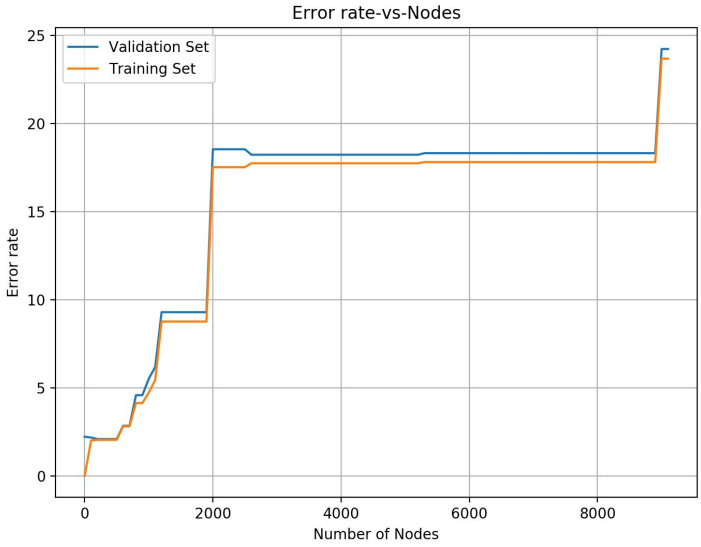


Observation: Comparing the Error rate vs Depth we found out that as obvious Training set gave better result in compared to the Validation set. If we restrict the depth to zero the decision will take on the root node itself that gave 75% accuracy. And if we allow the depth to take up the max depth of the tree it gave 97% accuracy with validation set and 100% accuracy with the training set.

Part-5.2

Result:





Observation: Comparing the Error rate vs Number of nodes in the data frame we found the above plot. If we allow the number of nodes to become zero the decision will be taken at the leaf node of the tree that gave 97% accuracy with validation set and 100% accuracy with the training set.. And if we restrict the number of nodes to make decision on the complete length of data frame we got 75% accuracy only.

Part-6

There are several methods for handling missing values in a Decision Tree as listed below-

1. We can predict the missing value of an attribute by selecting the value that occurred the most number of time in the training set in case of the categorical attribute. But doing this may not result in a good output.

2. Alternatively, we came to find another attribute that has the highest correlation with the attribute that has missing data and use it as a guide.

3. For numerical attributes, we can replace the missing values with the mean or median of the attribute column.

4. Alternatively, we can find another attribute that has the highest correlation with the attribute that has missing data for example- height and weight are correlated and one can be guessed using the other using linear regression.

5. we can recuse down the tree to all possible values of an attribute and vote the maximum probable result out of it. This adds overhead in terms of time complexity but may result in good accuracy.