Assignment-1

Anjul Ravi Gupta
Roll No: 2018201021
M.Tech CSE

Decision Tree Report

q-1-1:

- Filtered the categorical data from the Input data

- Divided the Input file data into Training data (80%) and Validation data(20%)

- Used ID3 algorithm in which attribute with maximum info gain is selected as root node. Then for each unique value of the attribute, it is called recursively to form the decision tree

- Results -
  - Accuracy is : 75.75 %
  - precision : 1
  - recall : 1
  - F1 score : 1.0

q-1-2:

- Divided the Input file data into Training data (80%) and Validation data(20%)

- Used ID3 algorithm in which:

Divided the continuous data into different classes and then these classes are treated as categorical data.Then for each attribute,used ID3 algorithm in which attribute with maximum info gain is selected as root node. Then for each unique value of the attribute, it is called recursively to form the decision tree

- Results -
  - Accuracy is : 96.57 %
  - precision : 0.91
  - recall : 0.94
  - F1 score : 0.92

q-1-3:

- Using scikit learn -
  - Accuracy is : 97.45 %

```
                 precision    recall  f1-score   support

           0       0.98      0.98      0.98      1654
           1       0.94      0.95      0.95       507

   micro avg       0.97      0.97      0.97      2161
   macro avg       0.96      0.97      0.96      2161
weighted avg       0.97      0.97      0.97      2161
```

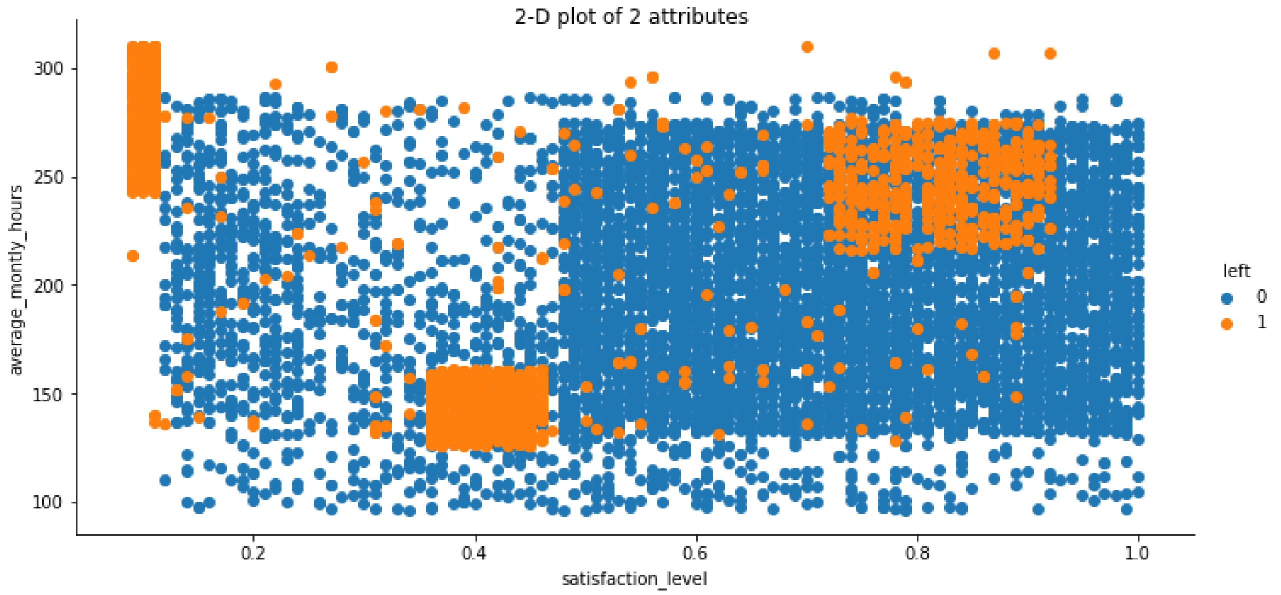
- Using Entropy -
  - Accuracy is : 96.57 %
  - precision : 0.91
  - recall : 0.94
  - F1 score : 0.92

- Using Gini Index -
  - Accuracy is : 96.21 %
  - precision : 0.92
  - recall : 0.91
  - F1 score : 0.92

- Using Misclassification -
  - Accuracy is : 96.17 %
  - precision : 0.90
  - recall : 0.93
  - F1 score : 0.91

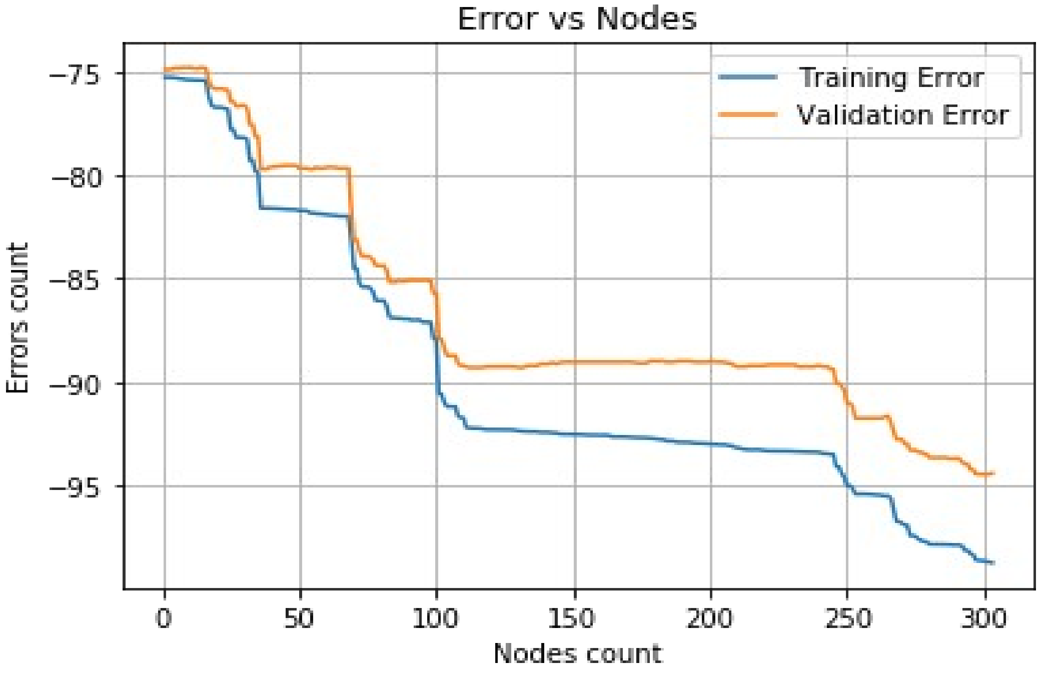
q-1-4:

Plot of graph between satisfactio level (x -axis) and average_monthly_hours (y -axis)



q-1-5:

Plot of graph for training and validation errors and number of nodes



q-1-6:

- Handling missing values in data in Decision Tree -

1.Ignore the missing values

- If the attribute with missing value is not significant for output then ignore missing values from it.

- If data set is large and attribute is important then ignore that row having missing value.

2.For categorical attribute, Fill missing values with the mode of that attribute(value with most occurence).

3.For continuous data attribute, Fill missing values with the mean for that attribute.

4.For continuous data attribute, Fill missing values with the median for that attribute.

5.Predecide a path for missing values and take always that path only (like move to left most child if missing value found).

6.Lazy Decision Tree (Reduced Feature Models/Known Value Strategy):- in this the prediction model is constructed at testing time based on the available test instance values. This is also known as 'Known values strategy'. During tree construction it uses only attributes whose values are known at testing. Hence it naturally handles the missing values at testing.