Assignment 1
Submitted By : Pranjal Patidar
E roll. : 2018201094
SMAI

Decision Tree :

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Algorithm :

The ID3 algorithm begins with the original set as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set and calculates the entropy (or information gain) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set is then split or partitioned by the selected attribute to produce subsets of the data.

Helper Functions :

train_validate_split :

This function separates the data into two catagories 1) Training data 2) Validation data.This takes arguments dataframe and validation size and returns two data frames training data frame and validation data frame.

is_pure :

This function checks purity of data means if data contains pure value ie. all 0’s or all 1’s. It takes argument as data and returns list or false. List contains only one value 0 or 1.

entropy_cal :

This function calculates Impurity in data by using various impurity measures like Entropy or Gini Index or Misclassification. This function is classified into two catagories one for catagorical data and one for numerical data. It takes aurguments as data and column for which impurity is to be calculated. It returns

Weighted impurity and dictionary or impurity of each value of that column. Two fumctions are as follows :

1- entropy_cal_for_catag
2- entropy_cal_for_num

overall_entropy :

This is similar function to entropy calculation it calculates impurity of overall data at perticular node. It takes aurgument only training data and returns the impurity.

information_gain :

This function compares impurities of all the columns and calculates the information gain and using that information gain selects the column with maximum information gain. This function acccepts arguments as training_data and column_list (list of remaining columns). It returns column which wins.

get_subtable :

It is useful in dividing the table into subtable. When at any node columns get splitted into various branches then to each branch we send data according to value on which they get split. The aurguments accepted by this function are training _data , column ,value and returns new_data(subtable).

decision_tree :

This function is helpful in building decision tree using above helper functions. Initially we calculate the winning column which will be the node at corresponding level. Then we calculate the unique values of that column. Then for each unique value we first get new subtable and then check purity of data of that subtable. If pure then save that pure value as result of that attribute esle call recursively decision tree. This is done till pure data is predicted. This function takes aurguments as training data , a column list and returns a tree.

predict :

This function iterates over the tree and predicts the result. In each iteration it checks value of true positive, true negative, false positive, false negative and increment accordingly. Then finally calculates recall, accuracy, f1_score, precision.

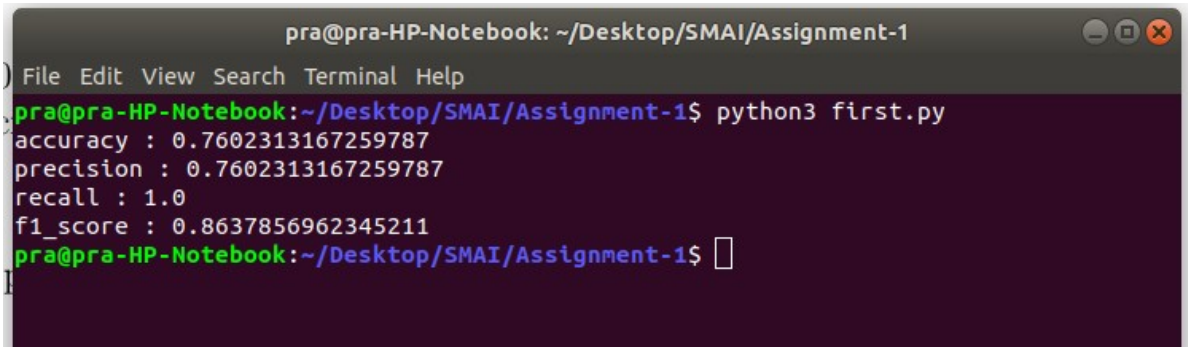
Libraries Used :

Matplotlib :
Pandas :
Numpy :
Math :
Pprint :

Results:

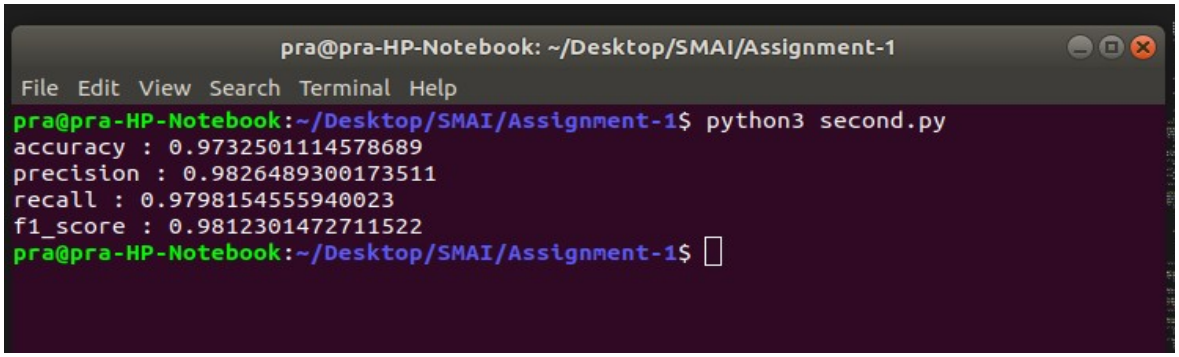
Part 1 :

Accuracy : 76.02 %
Precision : 76.02 %
Recall : 100%
F1-Score : 86.37%



Part 2 :

Accuracy : 97.32 %
Precision : 98.26 %
Recall : 97.98 %
F1-Score : 98.12 %



Part 3 :

Entropy :

Accuracy : 97.68 %
Precision : 95.04 %
Recall : 97.98 %

Gini :

Accuracy : 97.42 %
Precision : 94.47 %
Recall : 93.43 %

Misclassification :

Accuracy : 96.63 %

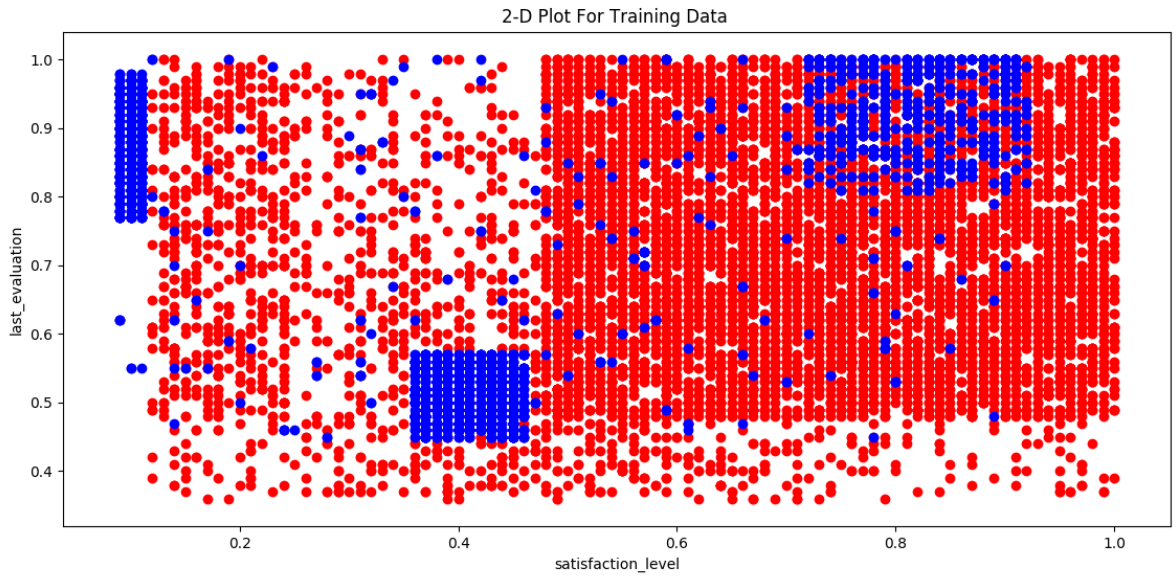
Precision : 92.75 %

Recall : 93.82 %

```
pra@pra-HP-Notebook: ~/Desktop/SMAI/Assignment-1
File  Edit  View  Search  Terminal  Help
pra@pra-HP-Notebook:~/Desktop/SMAI/Assignment-1$ python3 third.py
Entropy
accuracy : 0.9767753461366682
precision : 0.950381679389313
recall : 0.9343339587242027
Gini
accuracy : 0.9741648106904232
precision : 0.9446564885496184
recall : 0.9392789373814042
Misclassification rate
accuracy : 0.966282165039929
precision : 0.9274809160305344
recall : 0.9382239382239382
pra@pra-HP-Notebook:~/Desktop/SMAI/Assignment-1$
```

Part 4:

Graph Plot btw last_evaluation and satisfaction_level



Part 5 :

Graph btw depth vs training and validation error