ML Report Assignment-1
Mahak Sharma

Question 1 Data Visualisation

1.

covid 19 india.csv analysis:-

a. Number of columns: 9

b. Names of column: Sno, Date, Time, State/UnionTerritory, ConfirmedIndianNational, ConfirmedForeignNational, Cured, Deaths,Confirmed

c. Types of Column:-

Categorical Column	Continuous Column
State/Union Territory	Date Time ConfirmedIndianNational ConfirmedForeignNational Cured Deaths Confirmed

d**. Possible data values/range:-**

Column Name and Changes/Assumptions	**Possible Value/Range**
Date:- Column in the data set is of type object. So, first it is converted into date time format to get the max and min range of date available in dataset	In data: Minimum: 2020-01-30 Maximum: 2021-08-11 Possible range: any date after 2020-01-30 till today's date.
Time:- To get the range value_counts() is used and then by analysing minimum and maximum value is determined	In data: Minimum: 8:00 AM Maximum: 9:30 PM Possible range: Any time of the day
State/Union Territory:- All names of states/union territory Observations:- 1. Few names which are not the names of state/union territory 2. Few state/union territory names have '*' symbol after that	Kerala, Telengana, Delhi, Rajasthan, Uttar Pradesh, Haryana, Ladakh, Tamil Nadu, Karnataka, Maharashtra, Punjab, Jammu and Kashmir, Andhra Pradesh, Uttarakhand, Odisha, Puducherry, West Bengal, Chhattisgarh, Chandigarh, Gujarat, Himachal Pradesh, Madhya Pradesh, Bihar, Manipur, Mizoram, Andaman and Nicobar Islands, Goa, Unassigned, Assam, Jharkhand, Arunachal Pradesh, Tripura, Nagaland, Meghalaya, Dadra and Nagar Haveli and Daman and Diu,Sikkim, Daman & Diu, Lakshadweep, Telangana, Dadra and Nagar Haveli Himanchal Pradesh, Karanataka **Non state/union territory name:** Cases being reassigned to states, Unassigned **State names with '*':** Bihar****, Madhya Pradesh***, Maharashtra***

ConfirmedIndianNational:- This column has some values as '-' other than count of cases Range considered by excluding '-' entries in the column	Min-max present in the data: 0 - 177 Possible range: non-negative integers
ConfirmedForeignNational:- This column has some values as '-' other than count of cases Range considered by excluding '-' entries in the column	Min-max present in the data: 0 - 14 Possible range: non-negative integers
Cured	Min-max present in the data: 0-6159676 Possible range: non-negative integers till confirmed
Deaths	Min-max present in the data: 0-134201 Possible range: non-negative integers till confirmed
Confirmed	Min-max present in the data: 0-636442 Possible range: non-negative integers

covid vaccine statewise.csv analysis:

a. Number of columns: 24

b. Names of column: 'Updated On', 'State', 'Total Doses Administered', 'Sessions',' Sites ', 'First Dose Administered', 'Second Dose Administered','Male (Doses Administered)', 'Female (Doses Administered)', 'Transgender (Doses Administered)', ' Covaxin (Doses Administered)','CoviShield (Doses Administered)', 'Sputnik V (Doses Administered)','AEFI', '18-44 Years (Doses Administered)','45-60 Years (Doses Administered)', '60+ Years (Doses Administered)','18-44 Years(Individuals Vaccinated)','45-60 Years(Individuals Vaccinated)','60+ Years(Individuals Vaccinated)', 'Male(Individuals Vaccinated)','Female(Individuals Vaccinated)', 'Transgender(Individuals Vaccinated)','Total Individuals Vaccinated'

c. Types of Column:-

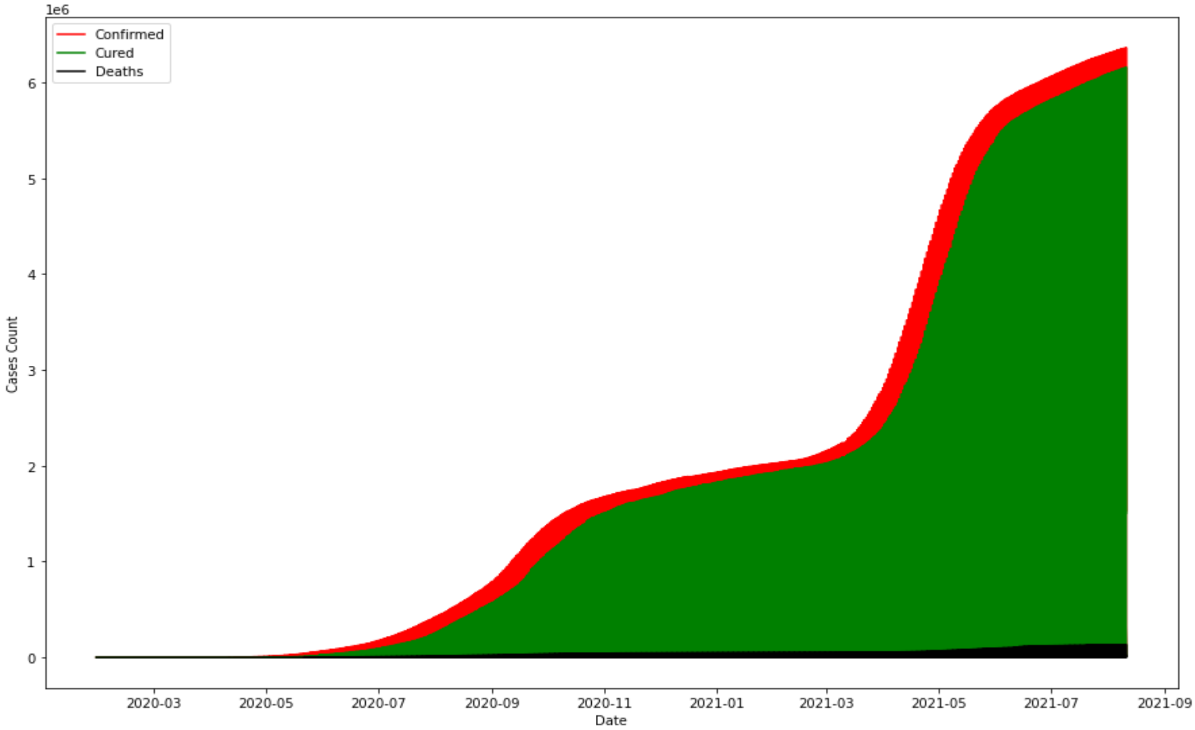
Categorical Column	Continuous Column
State	Updated On, Total Doses Administered, Sessions, Sites , First Dose Administered, Second Dose Administered, Male (Doses Administered), Female (Doses Administered), Transgender (Doses Administered), Covaxin (Doses Administered), CoviShield (Doses Administered), Sputnik V (Doses Administered), AEFI, 18-44 Years (Doses Administered), 45-60 Years (Doses Administered), 60+ Years (Doses Administered),18-44 Years(Individuals Vaccinated),45-60 Years(Individuals Vaccinated),60+ Years(Individuals Vaccinated), Male(Individuals Vaccinated),Female(Individuals Vaccinated), Transgender(Individuals Vaccinated),Total Individuals Vaccinated

d**. Possible data values/range:-**

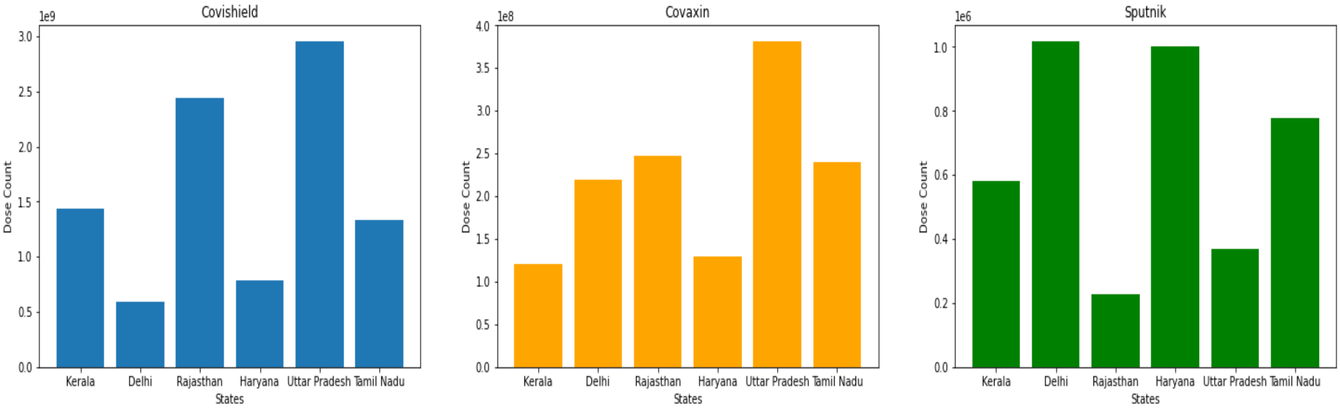
Column Name	Possible Value/Range(Min-Max)
Updated On: This is a date column. So, to get correct comparison b/w dates this is converted to date time format and then min and max range is considered	2021-01-02---2021-12-08 Possible range: Any time of the day
State: Names of state	Delhi, Ladakh, Tripura, Madhya Pradesh, Mizoram, Nagaland, Maharashtra, Uttar Pradesh, West Bengal, Dadra and Nagar

	Haveli and Daman and Diu, Karnataka, Goa, Andaman and Nicobar Islands, Punjab, Andhra Pradesh, Rajasthan, Puducherry, Himachal Pradesh, Chhattisgarh, India, Telangana, Lakshadweep, Manipur, Sikkim, Gujarat, Odisha, Assam, Arunachal Pradesh, Meghalaya, Kerala, Jharkhand, Tamil Nadu, Uttarakhand, Chandigarh, Jammu and Kashmir, Haryana, Bihar
Total Doses Administered	Min-max present in the data: 7 - 5.13228e+08 Possible range: non-negative integers
Sessions	Min-max present in the data: 0 - 3.50103e+07 Possible range: non-negative integers
Sites	Min-max present in the data: 0 - 73933 Possible range: non-negative integers
First Dose Administered	Min-max present in the data: 7 - 4.0015e+08 Possible range: non-negative integers
Second Dose Administered	Min-max present in the data: 0 - 1.13078e+08 Possible range: non-negative integers
Male (Doses Administered)	Min-max present in the data: 0 - 2.70164e+08 Possible range: non-negative integers
Female (Doses Administered)	Min-max present in the data: 2 - 2.39519e+08 Possible range: non-negative integers
Transgender (Doses Administered)	Min-max present in the data: 0 - 98275 Possible range: non-negative integers
Covaxin (Doses Administered)	Min-max present in the data: 0 - 6.23674e+07 Possible range: non-negative integers
CoviShield (Doses Administered)	Min-max present in the data: 0 - 4.46825e+08 Possible range: non-negative integers
Sputnik V (Doses Administered)	Min-max present in the data: 7 - 588039 Possible range: non-negative integers
AEFI	Min-max present in the data: 0 - 26542 Possible range: non-negative integers
18-44 Years (Doses Administered)	Min-max present in the data: 26624- 9.22431e+07 Possible range: non-negative integers
45-60 Years (Doses Administered)	Min-max present in the data: 16815 - 1.66757e+08 Possible range: non-negative integers
60+ Years(Individuals Vaccinated)	Min-max present in the data: 558 - 6.7311e+07 Possible range: non-negative integers
Male(Individuals Vaccinated)	Min-max present in the data: 23757 - 1.34942e+08 Possible range: non-negative integers
Female(Individuals Vaccinated)	Min-max present in the data: 24517 - 1.15668e+08 Possible range: non-negative integers
Transgender(Individuals Vaccinated)	Min-max present in the data: 2 - 46462 Possible range: non-negative integers
Total Individuals Vaccinated	Min-max present in the data: 7 - 2.50657e+08 Possible range: non-negative integers

2. Trend of confirmed, and cured cases along with the number of deaths in India



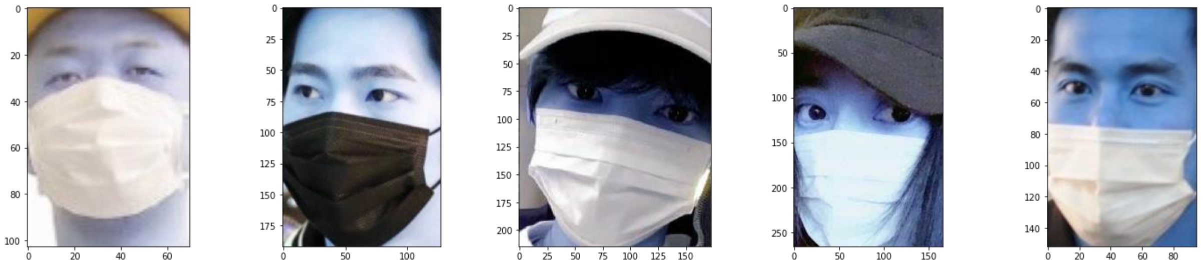
3. Plot the total number of doses administered of Covishield, Covaxin, and Sputnik



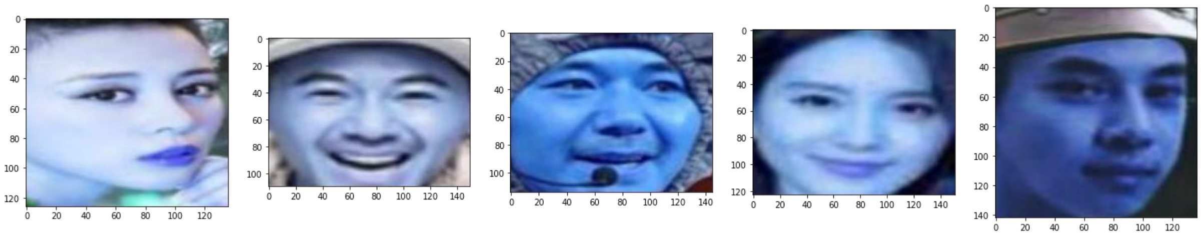
4.

a.

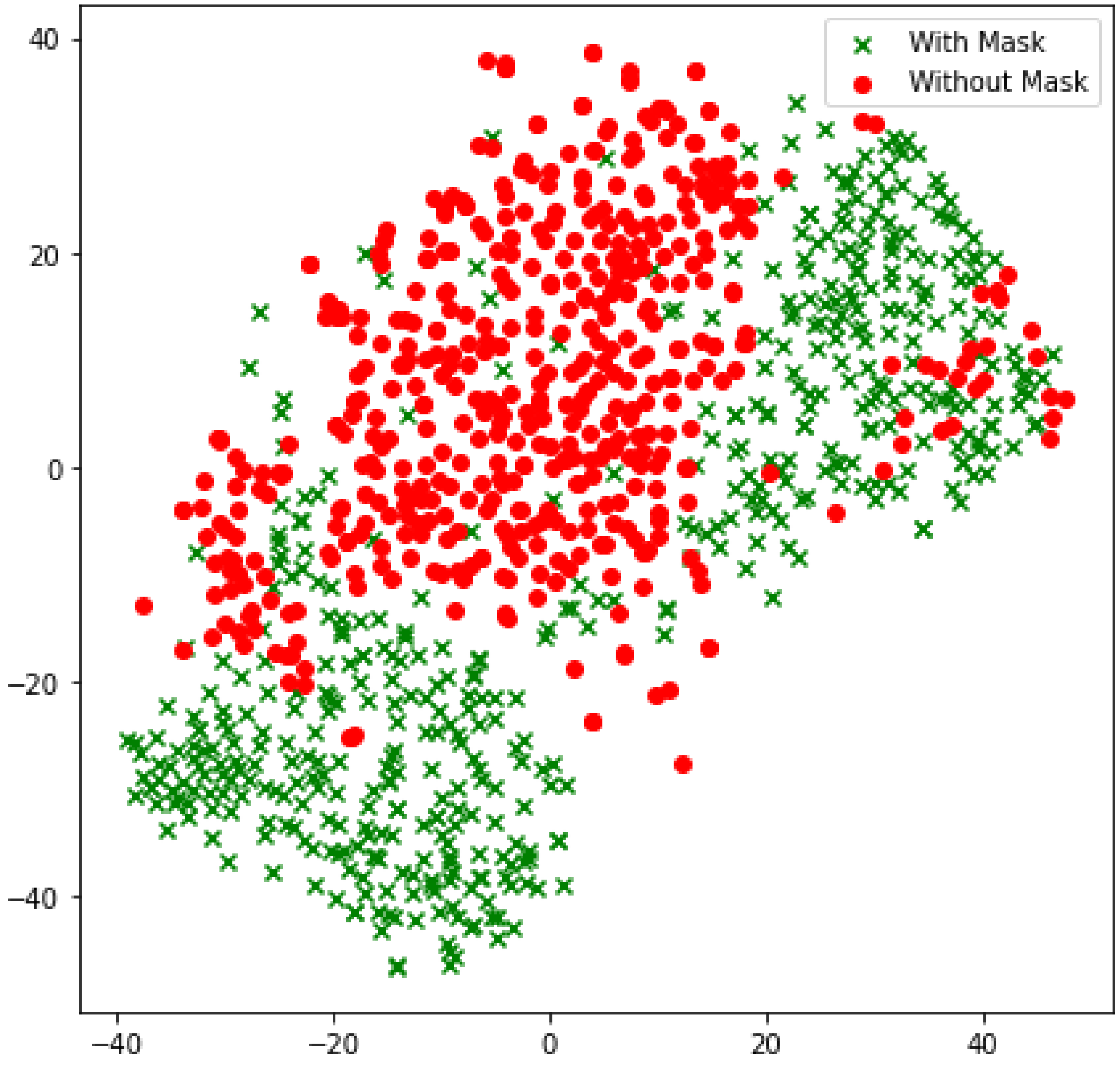
With Mask:-



Without Mask:-



b.
T-Sne:



Question 2.

Approach:-

1. **Converted categorical features into one hot encoding**
2. **This function splits 70% of data into training set and rest into testing set**

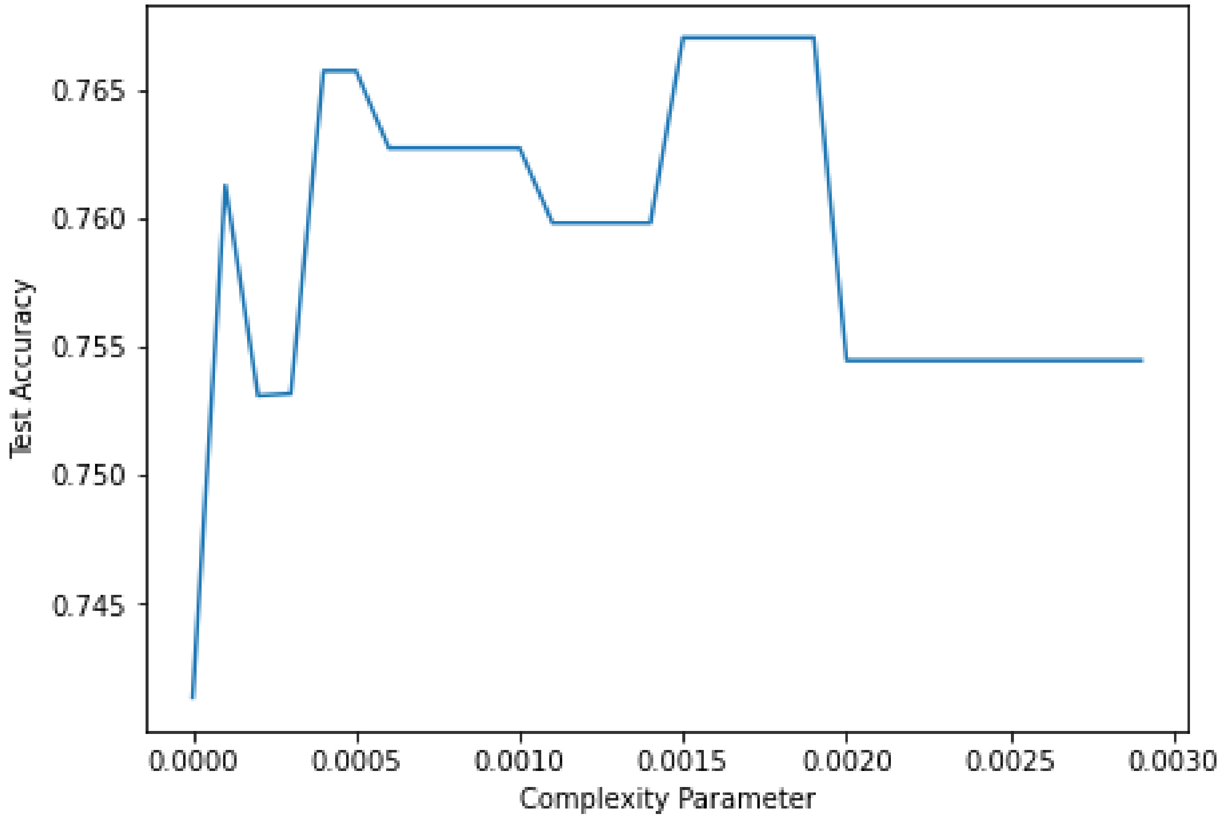
```
def splitting(features, target, p):
    splitting_index = int(p * features.shape[0])
    train_X = features[0:splitting_index]
    train_y = target[0:splitting_index]
    test_X = features[splitting_index:]
    test_y = target[splitting_index:]
    return train_X,train_y,test_X,test_y
```

3. **Function to calculate accuracy**

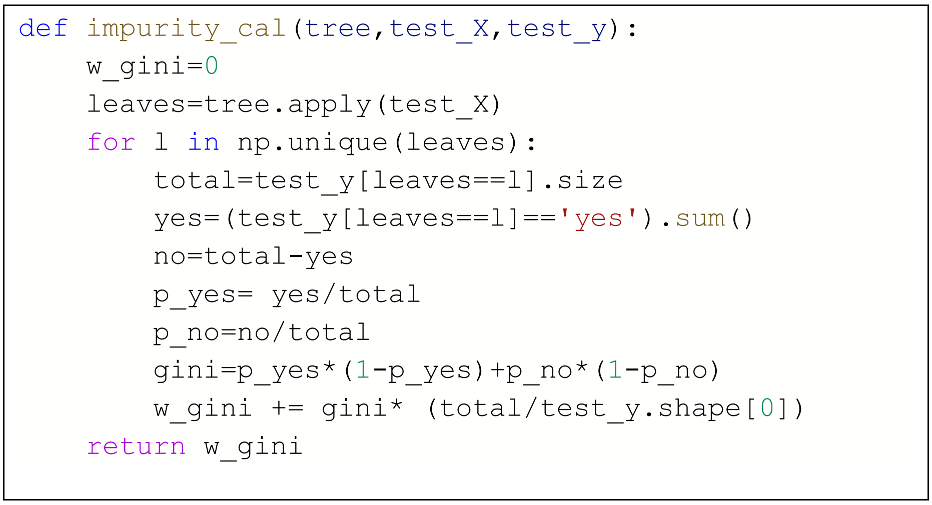
```
def accuracy(expected,predicted):
    correct_predictions = expected == predicted
    cp_sum=correct_predictions.sum()
    accuracy= cp_sum/expected.shape[0]
    return accuracy
```

1.

Curve between Complexity parameter and testing accuracy

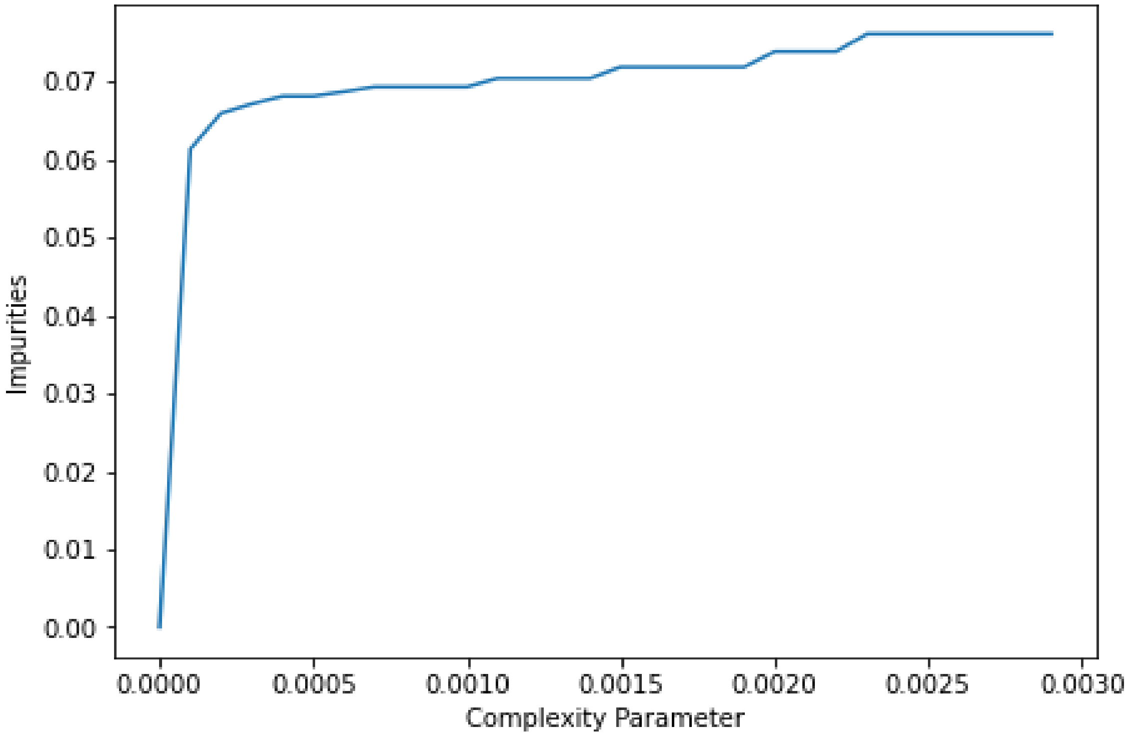


This function calculates sum of impurities of leaves using gini index and returns w_gini which is the sum



```
def impurity_cal(tree,test_X,test_y):
    w_gini=0
    leaves=tree.apply(test_X)
    for l in np.unique(leaves):
        total=test_y[leaves==l].size
        yes=(test_y[leaves==l]=='yes').sum()
        no=total-yes
        p_yes= yes/total
        p_no=no/total
        gini=p_yes*(1-p_yes)+p_no*(1-p_no)
        w_gini += gini* (total/test_y.shape[0])
    return w_gini
```

Curve between Complexity parameter and sum of impurity of leaf nodes.



2.
CCP_Alpha does the pruning in the tree model. Due to which model can sometime overfit and underfit

CCP Alpha Value	**Train Accuracy**	**Test Accuracy**	**Comment:Overfit/Underfit**
0.0	1.0	0.7413611718054544	**Overfit** As, model accuracy was 100% during training while testing accuracy in low.
0.0001	0.9549790156428843	0.7612689164036578	**Better fit** As, model accuracy was dropped though testing accuracy increased. So, it is a better fit

0.0002	0.9523429641705109	0.7530954115076475	**Underfit** There is decrease in training accuracy as well test accuracy on this value of ccp_alpha
0.0003	0.9521348548437446	0.753176337298697	**Better fit** There is slight decrease in training data accuracy but increase in test data accuracy
0.0004	0.9514758419756512	0.7657198349113863	**Better fit**
0.0005	0.9514758419756512	0.7657198349113863	**No change**
0.0006	0.950920883770941	0.7627255806425508	**Underfit** Both training and test accuracy decreased
0.0007	0.950920883770941	0.7627255806425508	**No change**
0.0008	0.950920883770941	0.7627255806425508	**No change**
0.0009	0.950920883770941	0.7627255806425508	**No change**

3.

CCP Alpha Value	**Training Accuracy**	**Testing Accuracy**
0.0	1.0	0.7413611718054544
2.601366584579099e-05	0.9986126044882245	0.7394189528202638
3.030955663498314e-05	0.9976067427421872	0.7429796876264465
3.1448164394220426e-05	0.9964621414449725	0.7433843165816946
3.220739580907457e-05	0.9950747459331969	0.7440317229100915
3.303359491908052e-05	0.9940341992993653	0.7491300477462167
6.705744973581678e-05	0.9689223405362284	0.7529335599255482
7.74703736123076e-05	0.9645173597863411	0.7529335599255482
8.737288401538688e-05	0.9576497520030522	0.7615116937768067
0.00016626125562309893	0.9528632374874267	0.7526098567613498

No, there is not much deviation between the results came from my implementation and inbuilt function but we have took random values of ccp_alpha so our test accuracy is in fluctuating order but in in-built function values of ccp_alpha are such that our pruning increasing our test accuracy also increased .

Question 3.

Pre-Processing: 1. NAN values of column A are replaced by its mean.
2.One hot encoding is applied for E column as it is categorical
3. Data is shuffled randomly to keep the training and testing data distribution Similar

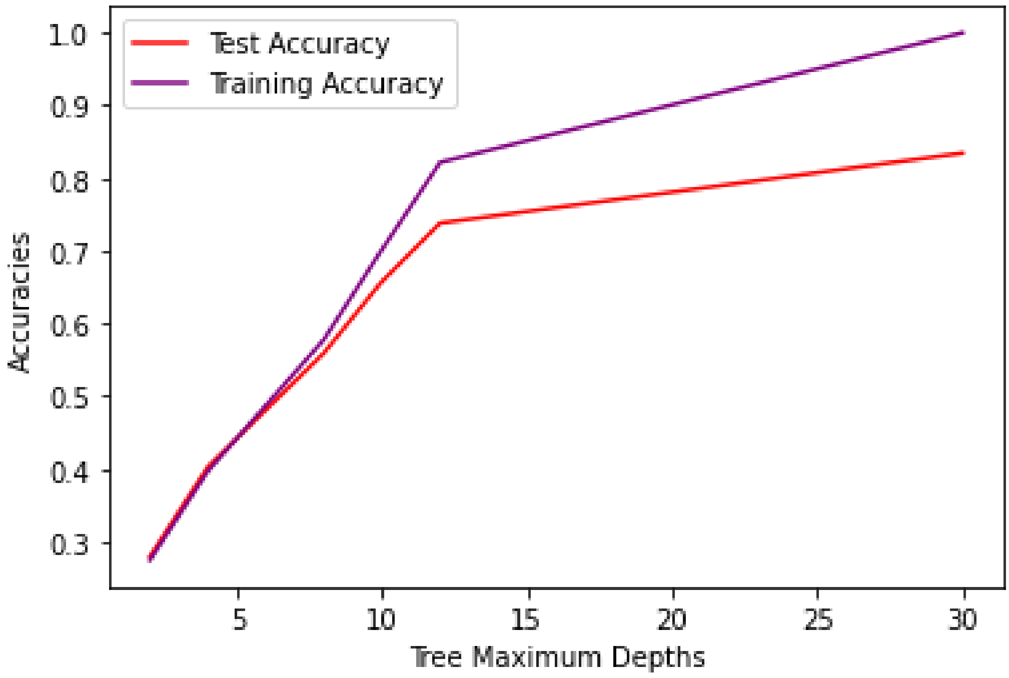
a. Gini Index Accuracy: 0.8243
Entropy Accuracy: 0.8344

Difference is not much in both the models but we have selected the model with entropy criteria as it has bit higher accuracy.

b. The best value of depth by using testing and training accuracy is

```
Depth:30
Training Accuracy:1.0
test Accuracy:0.83442862358525
```

Plot the curve between training and testing accuracy and depth



c. Accuracy in ensembled model is: 0.4155, which is significantly low compared to models trained in part (a) and (b). This is due to a relatively low max depth (=4) of decision stumps

d. On applying grid search on “number of trees” and “max depth”, we got the model with “number of trees” = 75 and “max depth” = 30 as the best model.
The training accuracy came out was: 0.9418
The testing accuracy came out was: 0.8353
The model is best so far based on testing accuracy and is performing slightly better than the models trained in part (a). This model in performing far better than model trained in part (c) due to more freedom in selecting max_depth.