Part 1

By Gini Index:

Prediction on Train Data

	price	maintenance	capacity	airbag	profitable	Predicted
0	low	low	2	no	yes	yes
1	low	med	4	yes	no	no
2	low	high	4	no	no	no
3	med	med	4	no	no	no
4	med	med	4	yes	yes	yes
5	med	high	2	yes	no	no
6	high	med	4	yes	yes	yes
7	high	high	2	yes	no	no
8	high	high	5	yes	yes	yes

Accuracy = 100%

Prediction on Test Data

	price	maintenance	capacity	airbag	profitable	Predicted
0	med	high	5	no	yes	yes
1	low	low	4	no	yes	yes

Accuracy = 100%

By Information Gain:

Prediction on Train Data

	price	maintenance	capacity	airbag	profitable	Predicted
0	low	low	2	no	yes	yes
1	low	med	4	yes	no	no
2	low	high	4	no	no	no
3	med	med	4	no	no	no
4	med	med	4	yes	yes	yes
5	med	high	2	yes	no	no
6	high	med	4	yes	yes	yes
7	high	high	2	yes	no	no
8	high	high	5	yes	yes	yes

Accuracy = 100%

Prediction on Test Data

	price	maintenance	capacity	airbag	profitable	Predicted
0	med	high	5	no	yes	yes
1	low	low	4	no	yes	yes

Accuracy = 100%

DECISION TREE (GINI INDEX)

```
| maintenance = high
        | capacity = 2 : no
        | capacity = 5 : yes
        | capacity = 4 : no
| maintenance = med
        | price = med
                | airbag = yes : yes
                | airbag = no : no
        | price = low : no
        | price = high : yes
| maintenance = low : yes
```

DECISION TREE (INFORMATION GAIN)

```
| maintenance = high
        | capacity = 2 : no
        | capacity = 5 : yes
        | capacity = 4 : no
| maintenance = med
        | price = med
                | airbag = yes : yes
                | airbag = no : no
        | price = low : no
        | price = high : yes
| maintenance = low : yes
```

Without dummies:

1. Entropy of Root Node by my model: 0.9910760598382222 ≈ 0.991
2. Information Gain of Root by My model: 0.18606356007860758 ≈ 0.186
3. Gini Index of Root by My model: 0.49382716049382713 ≈ 0.494

   Gini of Split at Root by My model: 0.38888888888888884 ≈ 0.388

With dummies:

1. *Entropy of Root Node by Scikit-Learn:* *0.991* ***≈ 0.991***
2. *Entropy of Root Node by my model:* *0.9910760598382222* ***≈ 0.991***

3. *Information Gain of Root by My model:* *0.14269027946047563* ***≈ 0.143***
4. *Information Gain of Root by Scikit-Learn:* *0.143* ***≈ 0.143***

5. *Gini Index of Root by Scikit-Learn:* *0.49382716049382713* ***≈ 0.494***
6. *Gini Index of Root by My model:* *0.49382716049382713* ***≈ 0.494***

   Gini of Split at Root by My model: 0.4166666666666667 ≈ 0.417

My model

DECISION TREE (GINI INDEX)

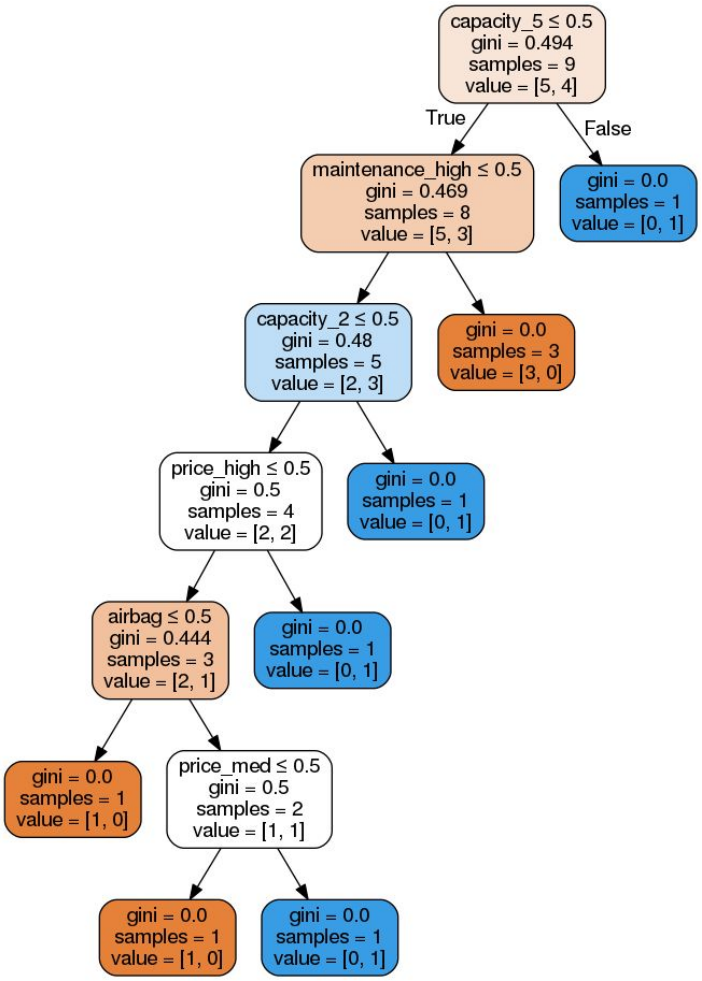
```
| capacity_5 = 0
        | maintenance_high = 0
                | capacity_4 = 1
                        | price_low = 0
                                | airbag = 1 : yes
                                | airbag = 0 : no
                        | price_low = 1 : no
                | capacity_4 = 0 : yes
        | maintenance_high = 1 : no
| capacity_5 = 1 : yes
```

DECISION TREE (INFORMATION GAIN)

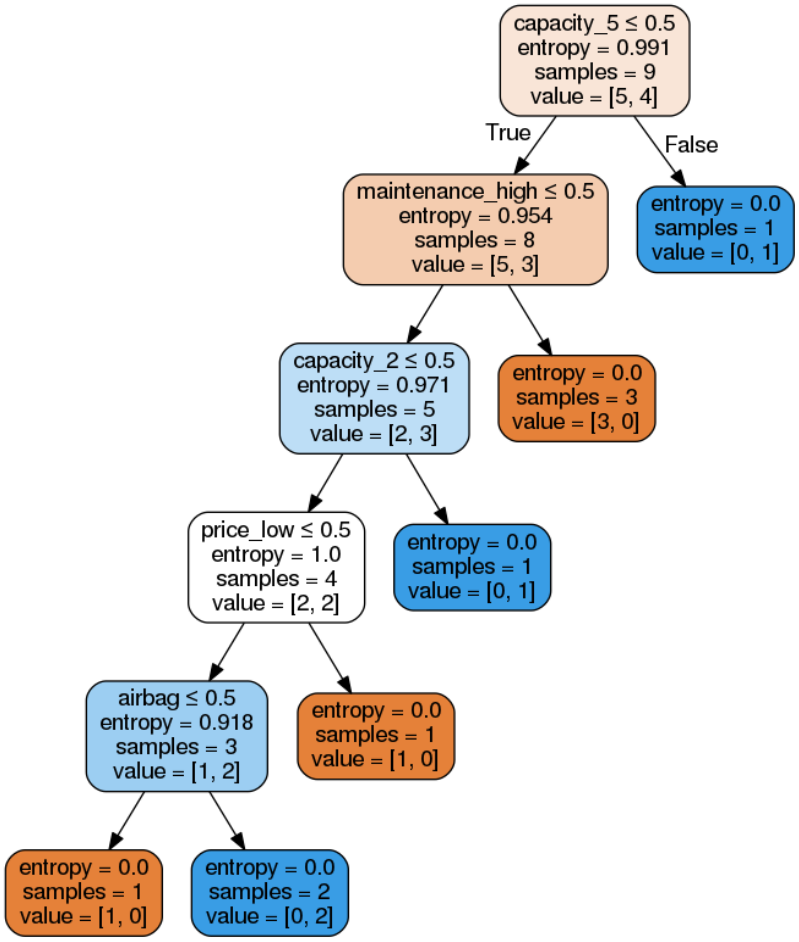
```
| capacity_5 = 0
        | maintenance_high = 0
                | capacity_4 = 1
                        | price_low = 0
                                | airbag = 1 : yes
                                | airbag = 0 : no
                        | price_low = 1 : no
                | capacity_4 = 0 : yes
        | maintenance_high = 1 : no
| capacity_5 = 1 : yes
```

Scikit Learn Model

DECISION TREE (GINI INDEX)



DECISION TREE (INFORMATION GAIN)



PREDICTIONS

My Model (Gini Index) (Test data)

	airbag	price_high	price_low	price_med	maintenance_high	maintenance_low	maintenance_med	capacity_2	capacity_4	capacity_5	profitable	Predicted
0	0	0	0	1	1	0	0	0	0	1	1	yes
1	0	0	1	0	0	1	0	0	1	0	1	no

Accuracy: 50%

My Model (Information Gain) (Test data)

	airbag	price_high	price_low	price_med	maintenance_high	maintenance_low	maintenance_med	capacity_2	capacity_4	capacity_5	profitable	Predicted
0	0	0	0	1	1	0	0	0	0	1	1	yes
1	0	0	1	0	0	1	0	0	1	0	1	no

Accuracy: 50%

Scikit Learn (Test data)

```
**********Prediction on TEST DATA**********

Model Prediction on Test Data (Scikit-Learn) (Gini Index) (dummy)

                [1 0]

Model Prediction on Test Data (Scikit-Learn) (Information Gain) (dummy)

                [1 0]
```

Therefore Accuracy is 50% in both cases.

My Model (Gini Index) (Train data)

	airbag	price_high	price_low	price_med	maintenance_high	maintenance_low	maintenance_med	capacity_2	capacity_4	capacity_5	profitable	Predicted
0	0	0	1	0	0	1	0	1	0	0	1	yes
1	1	0	1	0	0	0	1	0	1	0	0	no
2	0	0	1	0	1	0	0	0	1	0	0	no
3	0	0	0	1	0	0	1	0	1	0	0	no
4	1	0	0	1	0	0	1	0	1	0	1	yes
5	1	0	0	1	1	0	0	1	0	0	0	no
6	1	1	0	0	0	0	1	0	1	0	1	yes
7	1	1	0	0	1	0	0	1	0	0	0	no
8	1	1	0	0	1	0	0	0	0	1	1	yes

Accuracy: 100%

My Model (Information Gain) (Train data)

	airbag	price_high	price_low	price_med	maintenance_high	maintenance_low	maintenance_med	capacity_2	capacity_4	capacity_5	profitable	Predicted
0	0	0	1	0	0	1	0	1	0	0	1	yes
1	1	0	1	0	0	0	1	0	1	0	0	no
2	0	0	1	0	1	0	0	0	1	0	0	no
3	0	0	0	1	0	0	1	0	1	0	0	no
4	1	0	0	1	0	0	1	0	1	0	1	yes
5	1	0	0	1	1	0	0	1	0	0	0	no
6	1	1	0	0	0	0	1	0	1	0	1	yes
7	1	1	0	0	1	0	0	1	0	0	0	no
8	1	1	0	0	1	0	0	0	0	1	1	yes

Accuracy: 100%

Scikit Learn (Train data)

```
**********Prediction on TRAIN DATA**********

Model Prediction on Train Data (Scikit-Learn) (Gini Index) (dummy)

                [1 0 0 0 1 0 1 0 1]

Model Prediction on Train Data (Scikit-Learn) (Information Gain) (dummy)

                [1 0 0 0 1 0 1 0 1]
```

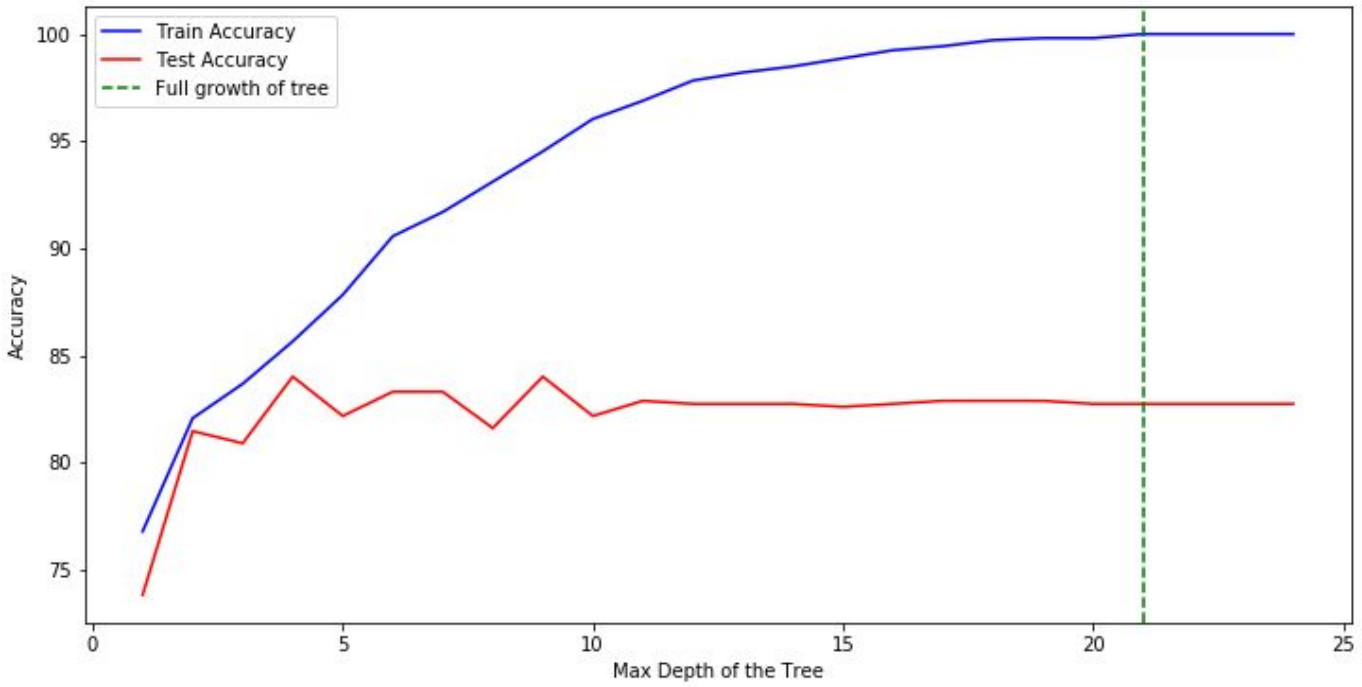
Therefore Accuracy is 100%

Part 2

All are Percentage Accuracy (in %)

Max_Depth	Train_Accuracy	Test_Accuracy
1	76.792453	73.833098
2	82.075472	81.471004
3	83.679245	80.905233
4	85.660377	84.016973
5	87.830189	82.178218
6	90.566038	83.309760
7	91.698113	83.309760
8	93.113208	81.612447
9	94.528302	84.016973
10	96.037736	82.178218
11	96.886792	82.885431
12	97.830189	82.743989
13	98.207547	82.743989
14	98.490566	82.743989
15	98.867925	82.602546
16	99.245283	82.743989
17	99.433962	82.885431
18	99.716981	82.885431
19	99.811321	82.885431
20	99.811321	82.743989
21	100.000000	82.743989
22	100.000000	82.743989
23	100.000000	82.743989
24	100.000000	82.743989

Tree of **maximum depth=4** and **maximum depth=9** achieved maximum Testing accuracy of **84.016973%**



Tree of **maximum depth=4** and **maximum depth=9** achieved maximum Testing accuracy of **84.016973%**

Yes, overfitting does occur. Since Accuracy on Training Data reached to 100% whereas Accuracy achieved on Testing Data is less than 85%. Also, Maximum testing accuracy was achieved at max_depth=4 and max_depth=9. At these depths, we can see that training error was around 85.66% and 94.52%. Now as the max_depth increases accuracy on only training data increases. No improvement has been recorded on Testing data.

Overfitting starts occurring from depth>=10.

Various Node asks question on one of the following words given below

If a document is related to **comp.graphics** newsgroup, then it is very likely that these documents will contain words like given below. These words resemble that if a document contains such words then it must belong to **comp.graphics** newsgroup.

- **graphics**
- **graphic**
- **image**
- **online**
- **disk**
- **time**
- **mac**
- **password**
- **program**
- **slow**
- **algorithm**
- **format**
- **acm**
- **comp**
- **windows**
- **port**
- **circle**

Also, if a document is related to **alt.atheism** newsgroup, then it is very likely that these documents will contain words like

- **god**
- **says**
- **don**
- **religious**
- **face**
- **wrote**
- **dwyer**
- **claiming**
- **bible**
- **bill**
- **thanks**
- **addition**
- **book**
- **cheers**
- **people**

Infact there are some nodes in this decision tree which can belong to **alt.atheism** and in general, are less seen in **comp.graphics** newsgroup. These nodes are:

- **your**
- **you**
- **what**
- **how**
- **he**
- **who**
- **am**
- **have**

Our Decision Tree contains few nodes which upon looking, even humans can't distinguish and predict the category of any document. For these nodes, we humans ideally can't assign any probability to any document for any newsgroup. These are some words which we normally do not expect our tree to use and take any decision. These nodes might have occurs in the decision tree due to repeated use of these words in various documents. These are some word which can be present in both the newsgroup. These nodes are:

- **if**
- **in**
- **by**
- **july**
- **get**
- **that**
- **just**
- **either**
- **with**
- **will**
- **use**
- **coming**
- **can**
- **an**
- **there**
- **for**
- **or**
- **to**
- **came**

Also there are few nodes which don't make sense to use for making/predicting decisions. These nodes are:

- **au**
- **mp**
- **ve**

An overview of the tree created by my model when Max_depth=9

```
| writes = 0
    | god = 0
        | that = 0
            | bible = 0
                | bill = 0
                    | if = 0
                        | book = 0
                            | cheers = 0
                                | snm = 0 : yes
                                | snm = 1 : no
                            | cheers = 1 : no
                        | book = 1
                            | to = 1 : yes
                            | to = 0 : no
                    | if = 1
                        | thanks = 0
                            | you = 0
                                | tao = 0 : yes
                                | tao = 1 : no
                            | you = 1
                                | there = 0 : yes
                                | there = 1 : yes
                        | thanks = 1 : yes
                | bill = 1 : no
            | bible = 1 : no
        | that = 1
            | wrote = 0
                | people = 0
                    | religious = 0
                        | an = 0
                            | he = 0
                                | face = 0 : yes
                                | face = 1 : no
                            | he = 1
                                | graphic = 0 : no
                                | graphic = 1 : yes
                        | an = 1
                            | can = 1
                                | dwyer = 0 : yes
                                | dwyer = 1 : no
                            | can = 0
                                | how = 0 : yes
                                | how = 1 : no
                    | religious = 1 : no
                | people = 1
                    | windows = 0 : no
                    | windows = 1 : yes
            | wrote = 1
                | came = 0
```

```
                    | ve = 0
                        | online = 0 : no
                        | online = 1 : yes
                    | ve = 1 : yes
                | came = 1 : yes
    | god = 1
        | use = 0 : no
        | use = 1
            | will = 1 : no
            | will = 0 : yes
| writes = 1
    | graphics = 0
        | image = 0
            | that = 1
                | program = 0
                    | comp = 0
                        | csd = 0
                            | slow = 0
                                | uchicago = 0 : no
                                | uchicago = 1 : yes
                            | slow = 1 : yes
                        | csd = 1 : yes
                    | comp = 1 : yes
                | program = 1 : yes
            | that = 0
                | keith = 0
                    | god = 0
                        | who = 0
                            | have = 0
                                | am = 0 : no
                                | am = 1 : yes
                            | have = 1
                                | rosenau = 0 : yes
                                | rosenau = 1 : no
                        | who = 1
                            | disk = 0 : no
                            | disk = 1 : yes
                    | god = 1 : no
                | keith = 1 : no
        | image = 1 : yes
    | graphics = 1 : yes
```

Please look at **“Max_Accuracy_Tree.txt”** file for getting a clear view.

Tree Build by Scikit Learn

For details please look at "**output_entropy_3.png**" file for details.

Scikit Learn Model:

Number of misclassification on Training Data: 0 out of 1060 (100% Accuracy)
Number of misclassification on Testing Data: 162 out of 707 (77.086% Accuracy)

Fully Grown Tree Build by My Model

```
| writes = 0
    | god = 0
        | that = 0
            | bible = 0
                | bill = 0
                    | if = 0
                        | book = 0
                            | cheers = 0
                                | snm = 0
                                    | sole = 0 : yes
                                    | sole = 1 : no
                                | snm = 1 : no
                            | cheers = 1 : no
                        | book = 1
                            | to = 1 : yes
                            | to = 0 : no
                    | if = 1
                        | thanks = 0
                            | you = 0
                                | tao = 0
                                    | addition = 0 : yes
                                    | addition = 1 : no
                                | tao = 1 : no
                            | you = 1
                                | there = 0
                                    | time = 0
                                        | your = 0
                                            | for = 1 : yes
                                            | for = 0
                                                | circle = 0
                                                    | mp = 0 : no
                                                    | mp = 1 : yes
                                                | circle = 1 : yes
                                        | your = 1 : no
                                    | time = 1 : no
                                | there = 1 : yes
                        | thanks = 1 : yes
                | bill = 1 : no
            | bible = 1 : no
        | that = 1
            | wrote = 0
                | people = 0
                    | religious = 0
                        | an = 0
                            | he = 0
                                | face = 0
                                    | tammy = 0
                                        | claiming = 0
                                            | doubt = 0 : yes
```

```
                                        | doubt = 1 : no
                                    | claiming = 1 : no
                                | tammy = 1 : no
                            | face = 1 : no
                        | he = 1
                            | graphic = 0 : no
                            | graphic = 1 : yes
                    | an = 1
                        | can = 1
                            | dwyer = 0
                                | coming = 0 : yes
                                | coming = 1 : no
                            | dwyer = 1 : no
                        | can = 0
                            | how = 0
                                | know = 0
                                    | will = 0
                                        | port = 0 : no
                                        | port = 1 : yes
                                    | will = 1 : yes
                                | know = 1 : yes
                            | how = 1 : no
                | religious = 1 : no
            | people = 1
                | windows = 0 : no
                | windows = 1 : yes
        | wrote = 1
            | came = 0
                | ve = 0
                    | online = 0 : no
                    | online = 1 : yes
                | ve = 1 : yes
            | came = 1 : yes
| god = 1
    | use = 0 : no
    | use = 1
        | will = 1 : no
        | will = 0 : yes
| writes = 1
| graphics = 0
    | image = 0
        | that = 1
            | program = 0
                | comp = 0
                    | csd = 0
                        | slow = 0
                            | uchicago = 0
                                | acm = 0
                                    | tiff = 0
```

```
                                        | algorithm = 0
                                            | format = 0
                                                | in = 1
                                                    | don = 0 : no
                                                    | don = 1
                                                        | with = 1 : no
                                                        | with = 0
                                                            | what = 0
                                                                | either = 0
                                                                    | rahul = 0
                                                                        | here = 0 : no
                                                                        | here = 1
                                                                            | atheists = 1 : no
                                                                            | atheists = 0 : yes
                                                                    | rahul = 1 : yes
                                                                | either = 1 : yes
                                                            | what = 1 : no
                                                | in = 0
                                                    | just = 0
                                                        | tracing = 0 : no
                                                        | tracing = 1 : yes
                                                    | just = 1
                                                        | says = 0 : yes
                                                        | says = 1 : no
                                            | format = 1 : yes
                                        | algorithm = 1 : yes
                                    | tiff = 1 : yes
                                | acm = 1 : yes
                            | uchicago = 1 : yes
                        | slow = 1 : yes
                    | csd = 1 : yes
                | comp = 1 : yes
            | program = 1 : yes
        | that = 0
            | keith = 0
                | god = 0
                    | who = 0
                        | have = 0
                            | am = 0
                                | time = 0
                                    | by = 0
                                        | with = 0
                                            | you = 0
                                                | july = 0
                                                    | mac = 0
                                                        | au = 0
                                                            | get = 0
                                                                | password = 0 : no
                                                                | password = 1 : yes
```

```
                                                            | get = 1 : yes
                                                        | au = 1 : yes
                                                    | mac = 1 : yes
                                                | july = 1 : yes
                                            | you = 1 : no
                                        | with = 1 : yes
                                    | by = 1 : no
                                | time = 1 : yes
                            | am = 1 : yes
                        | have = 1
                            | rosenau = 0
                                | your = 0 : yes
                                | your = 1
                                    | or = 0 : yes
                                    | or = 1 : no
                            | rosenau = 1 : no
                    | who = 1
                        | disk = 0 : no
                        | disk = 1 : yes
                | god = 1 : no
            | keith = 1 : no
        | image = 1 : yes
    | graphics = 1 : yes
```

Please look at **“Full_Grown_Tree.txt”** file for getting a clear view.

********** **THE END** **********

By Vaibhav Poddar
16CS10051