Assignment 2 Report Part -1

Name: Sai Saketh Aluru
Roll No. 16CS30030

a.) Using the given toy dataset and training a decision tree classifier using
I. Information Gain
II. Gini index

The given dataset is:

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

Where the first 0-8 columns are training data and the last two are test data.

The obtained decision tree using information gain as the criteria, printed in the specified format is:

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

The class labels for the test data obtained using this decision tree to predict are:

1. data: prediction: yes
   price med
   maintenance high
   capacity 5
   airbag no

2. data: prediction: yes
   price low
   maintenance low
   capacity 4
   airbag no

The accuracy of this prediction, using the given ground truth values is 1.0, or 100%.

ii. The obtained decision tree using Gini index as the splitting criteria is:

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

1. data: prediction: yes
   price med
   maintenance high
   capacity 5
   airbag no

2. data: prediction: yes
   price low
   maintenance low
   capacity 4
   airbag no

The accuracy obtained using this classifier on the test data is 1.0, or 100%.

The values of the Information gain and GINI index of the root node using my model is:
i. Information gain - 0.18606356007860825 for the feature maintenance
ii. Gini Index - 0.38888888888888895 for the feature maintenance

The decision trees are then built using scikit learn library using Information gain and Gini index as the criteria. The obtained accuracies on the trees are:

i. Information gain - 1.0 or 100%.
ii. Gini index - 1.0 or 100%.

The predictions on the test data by the scikit learn model are:
Using both information gain and gini index criteria.

1. data: prediction: yes
   price med
   maintenance high
   capacity 5
   airbag no

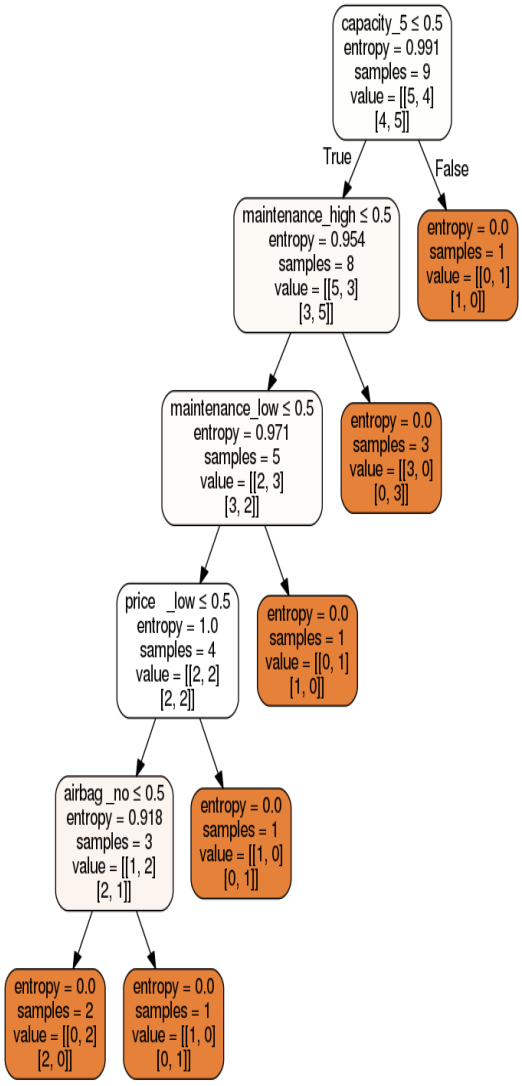
2. data: prediction: yes
   price low
   maintenance low
   capacity 4
   airbag no

The value of the information gain and gini index of the root node using scikit learn's model is:

i. Information gain:
   Entropy of parent node - 0.991 Sample size - 9
   Entropy of child 1 - 0.954 Sample size - 8
   Entropy of child 2 - 0.0 Sample size - 1
   Information gain - 0.991 - 0.954*8/9 = 0.143
ii. Gini index: 0.49382716049382713

The trees obtained are:

Information gain:



Gini index: