Name: Vaibhav Agrawal
Roll Number: 15CE10057
Assignment 2: Decision Tree (Part 1)
Machine Learning (CS60050)

Part 1

(Description) Consider that you are willing to buy a car and you have collected information having four attributes 'price', 'maintenance', 'capacity' and 'airbag', and are trying to predict whether a given car is 'profitable' or not. Assume all the four attributes are categorical, with discrete values.

(Dataset) Download the training and test data here . The sheet labelled "training data" contains the data to be trained on. The sheet named as "test data" contains the data on which you have to test your model.

(Tasks)
(A) Train your decision tree classifier on the train-data (where you will use "profitable"), using the impurity measure:

- Information Gain
- Gini Index
-

Test your model on test-data (where the "profitable" label is unseen). After prediction, report the individual accuracies on the test data obtained using (a) and (b). Note that the "profitable" field should not be used in the classification process.
For both cases, write your program such that it prints out the decision tree, in a particular format. For example, assume that your decision tree looks like the following - the attribute "price" is the root node and "maintenance" is the 2nd level node and "capacity" is the third level node (under maintenance = low). "yes" and "no" specifies the final value of "profitable". Then the program should print out the decision tree as follows:

price = low
| maintenance = low
| capacity = 4 : yes | maintenance = high : no

Where subsequent levels are at increasing indentations from the left.

(B) Repeat the experiment using the decision tree algorithm implemented in scikit learn , using both Information Gain and Gini index. Report the accuracies on test data.
(Deliverables) Your report should contain :

1. The decision tree in the format provided in (A)

2. The value of Information Gain and Gini Index of the root node using :
   - Your model
   - scikit learn

3. The labels generated on the test data and accuracy on the test data using :
   - Your model
   - scikit learn

1. Decision Tree in the prescribed format:

- **Information Gain Tree:**

```
maintenance = low : yesmaintenance != low
|price = low : no
|price != low
    |capacity = 2 : no
    |capacity != 2
        |airbag = no : no
        |airbag != no : yes
```

- **Gini Index Tree:**

```
maintenance = low : yesmaintenance != low
|capacity = 5 : yes
|capacity != 5
    |maintenance = high : no
    |maintenance != high
        |price = high : yes
        |price != high
            |price = low : no
            |price != low
                |airbag = no : no
                |airbag != no : yes
```

2. The value of Information Gain and Gini Index of the root node

a. Your Model:

Information Gain	0.8484
Gini Index	0.4167

a. Scikit Learn:

Information Gain	0.991
Gini Index	0.494

3. The labels generated on the test data and accuracy on the test data using

Orignal test data frame:

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

a. Your Model:

Classification using Information gain:

	price	maintenance	capacity	airbag	label	classification
0	med	high	5	no	yes	no
1	low	low	4	no	yes	yes

Classification using Gini Index:

	price	maintenance	capacity	airbag	label	classification
0	med	high	5	no	yes	yes
1	low	low	4	no	yes	yes

Impurity	Accuracy on test data
Information Gain	50%
Gini Index	100%

a. Scikit Learn:

Scikit Learn model has predicted the same result as my model.

Impurity	Accuracy on test data	Classification for 1st point	Classification for 2nd point
Information Gain	50%	no	Yes
Gini Index	100%	yes	Yes