CS60050 - MACHINE LEARNING ASSIGNMENT 2
DECISION TREES - PART 1

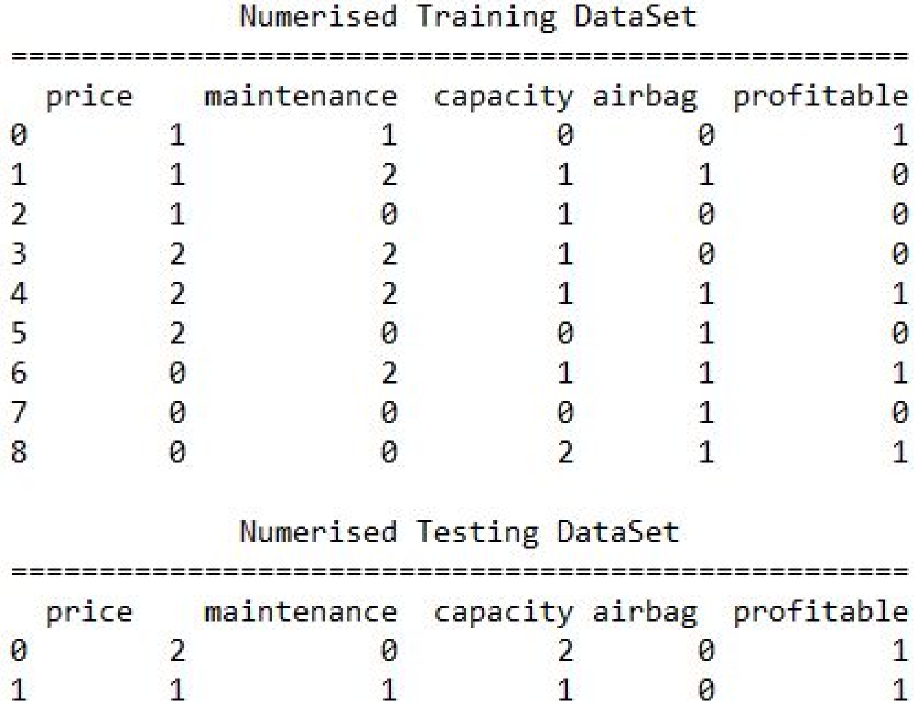
NAME: **Himanshu Mundhra**
ROLL No. : **16CS10057**

In this folder, we build a Decision Tree on a Small DataSet for Car Sales.

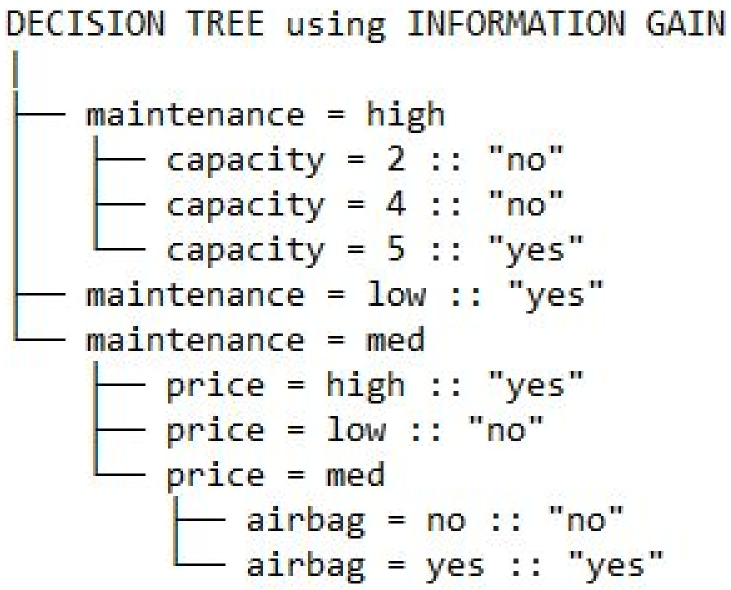
We have 4 features of **'price', 'maintenance', 'capacity', 'airbag'** which take values in the form of strings/integers. Our Target Class **'profitable'** takes a binary truth value, whether the car sale is profitable or not, given the features.

The presence of strings renders a need to normalise the data into numbers. We **Numerise** the Data in such a way that we assign values starting from 0 till we cover all the unique possible outcomes of a feature.

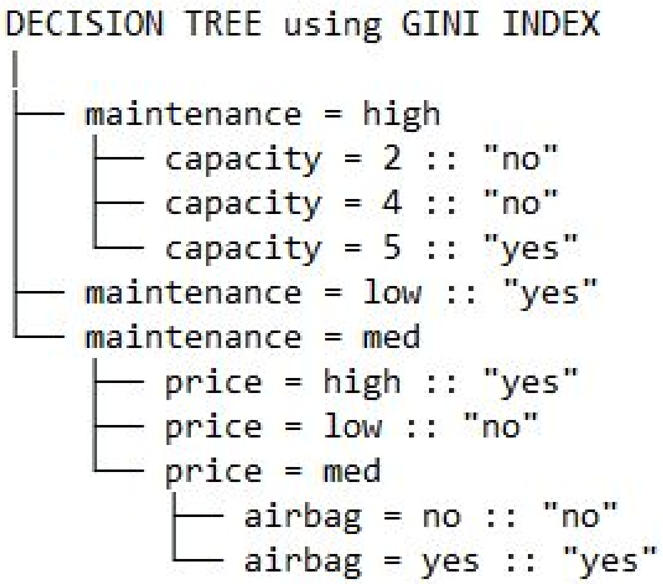
We use the criterion of ***entropy*** and ***gini index*** to grow a full tree and then use the testing data to report the accuracy of our trees compared to Scikit-Learn's models.



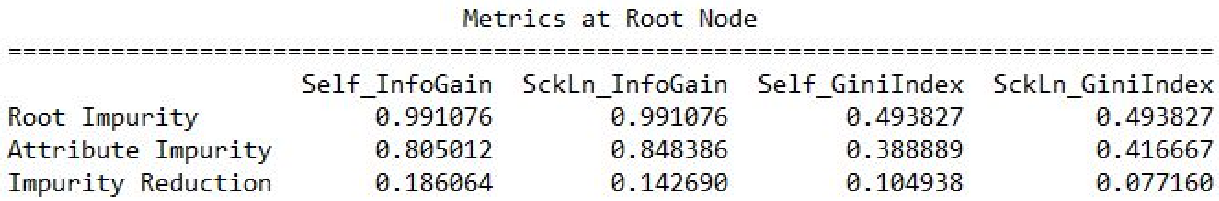
```
                 Numerised Training DataSet
=====================================================
  price    maintenance  capacity airbag  profitable
0      1             1         0      0           1
1      1             2         1      1           0
2      1             0         1      0           0
3      2             2         1      0           0
4      2             2         1      1           1
5      2             0         0      1           0
6      0             2         1      1           1
7      0             0         0      1           0
8      0             0         2      1           1

                 Numerised Testing DataSet
=====================================================
  price    maintenance  capacity airbag  profitable
0      2             0         2      0           1
1      1             1         1      0           1
```



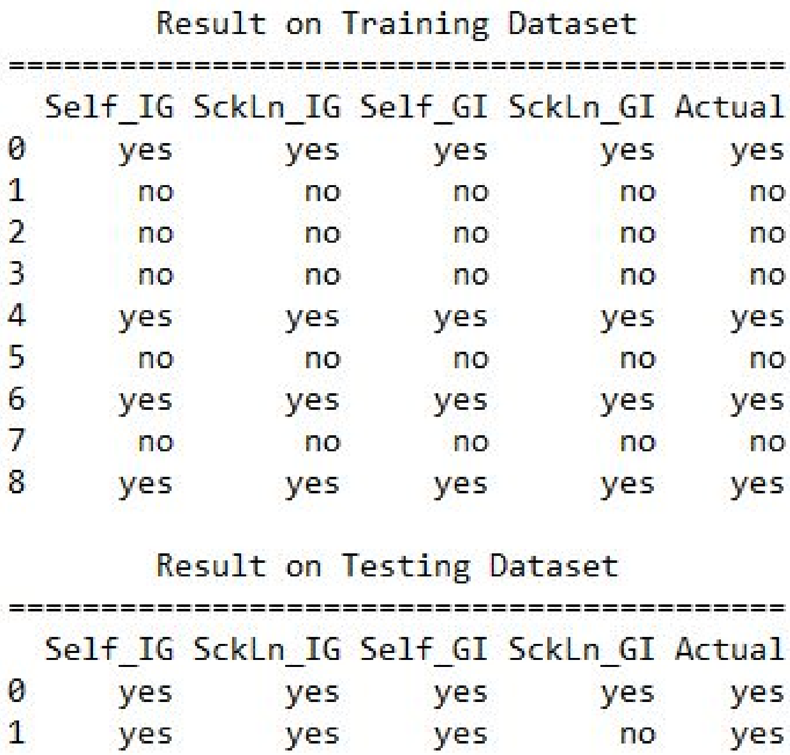
```
DECISION TREE using INFORMATION GAIN
|
├── maintenance = high
│   ├── capacity = 2 :: "no"
│   ├── capacity = 4 :: "no"
│   └── capacity = 5 :: "yes"
├── maintenance = low :: "yes"
└── maintenance = med
    ├── price = high :: "yes"
    ├── price = low :: "no"
    └── price = med
        ├── airbag = no :: "no"
        └── airbag = yes :: "yes"
```



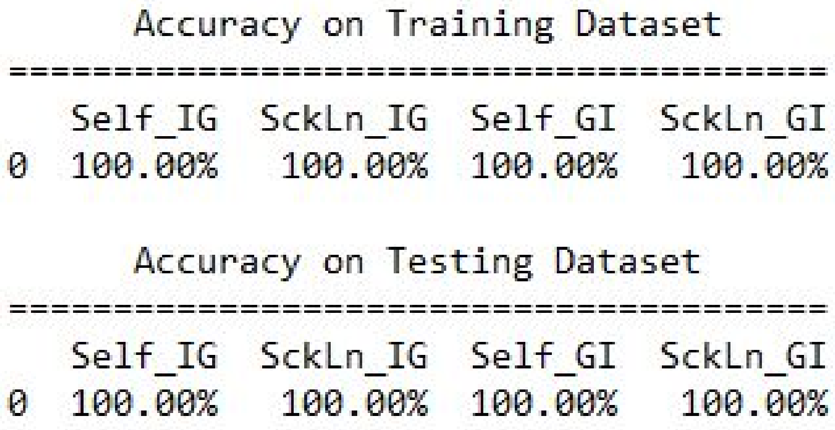
```
DECISION TREE using GINI INDEX
|
├── maintenance = high
│   ├── capacity = 2 :: "no"
│   ├── capacity = 4 :: "no"
│   └── capacity = 5 :: "yes"
├── maintenance = low :: "yes"
└── maintenance = med
    ├── price = high :: "yes"
    ├── price = low :: "no"
    └── price = med
        ├── airbag = no :: "no"
        └── airbag = yes :: "yes"
```



```
                                  Metrics at Root Node
===============================================================================
                    Self_InfoGain  SckLn_InfoGain  Self_GiniIndex  SckLn_GiniIndex
Root Impurity            0.991076        0.991076        0.493827         0.493827
Attribute Impurity       0.805012        0.848386        0.388889         0.416667
Impurity Reduction       0.186064        0.142690        0.104938         0.077160
```



```
          Result on Training Dataset
==========================================
  Self_IG SckLn_IG Self_GI SckLn_GI Actual
0     yes      yes     yes      yes    yes
1      no       no      no       no     no
2      no       no      no       no     no
3      no       no      no       no     no
4     yes      yes     yes      yes    yes
5      no       no      no       no     no
6     yes      yes     yes      yes    yes
7      no       no      no       no     no
8     yes      yes     yes      yes    yes

          Result on Testing Dataset
==========================================
  Self_IG SckLn_IG Self_GI SckLn_GI Actual
0     yes      yes     yes      yes    yes
1     yes      yes     yes       no    yes
```



```
     Accuracy on Training Dataset
=======================================
   Self_IG  SckLn_IG  Self_GI  SckLn_GI
0  100.00%   100.00%  100.00%   100.00%

     Accuracy on Testing Dataset
=======================================
   Self_IG  SckLn_IG  Self_GI  SckLn_GI
0  100.00%   100.00%  100.00%   100.00%
```