Sanjivani Rural Education Society's

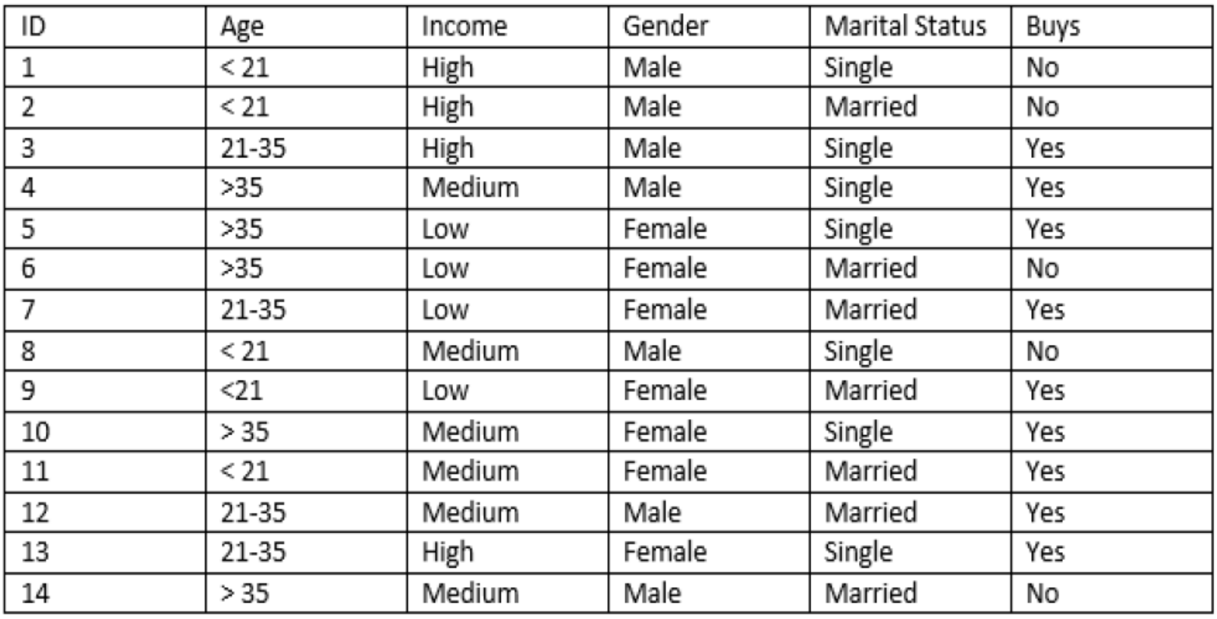
College of Engineering, Kopargaon-423603

DEPARTMENT OF COMPUTER ENGINEERING

Instruction No. 01 and 02 ML Lab / Sr. No.01 and 02 Rev 00 Date: 27/12/17	**Title: Assignment on Decision Tree**

3. Assignment on Decision Tree Classifier:

dataset collected in a cosmetics shop showing details of customers and whether or not they responded to a special offer to buy a new lip-stick is shown in table below. Use this dataset to build a decision tree, with Buys as the target variable, to help in buying lip-sticks in the future. Find the root node of decision tree. According to the decision tree you have made from previous training data set, what is the decision for the test data: [Age < 21, Income = Low, Gender = Female, Marital Status = Married]?



ID	Age	Income	Gender	Marital Status	Buys
1	< 21	High	Male	Single	No
2	< 21	High	Male	Married	No
3	21-35	High	Male	Single	Yes
4	>35	Medium	Male	Single	Yes
5	>35	Low	Female	Single	Yes
6	>35	Low	Female	Married	No
7	21-35	Low	Female	Married	Yes
8	< 21	Medium	Male	Single	No
9	<21	Low	Female	Married	Yes
10	> 35	Medium	Female	Single	Yes
11	< 21	Medium	Female	Married	Yes
12	21-35	Medium	Male	Married	Yes
13	21-35	High	Female	Single	Yes
14	> 35	Medium	Male	Married	No

Solution

Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility.

Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]
2. Result [End Nodes]

A decision tree can be visualized. A decision tree is one of the many Machine Learning

algorithms.
It’s used as classifier: given input data, it is class A or class B? In this lecture we will visualize a decision tree using the Python module **pydotplus and the module graphviz.**

If you want to do decision tree analysis, to understand the decision tree algorithm / model or if you just need a decision tree maker - you’ll need to visualize the decision tree.

Decision Tree

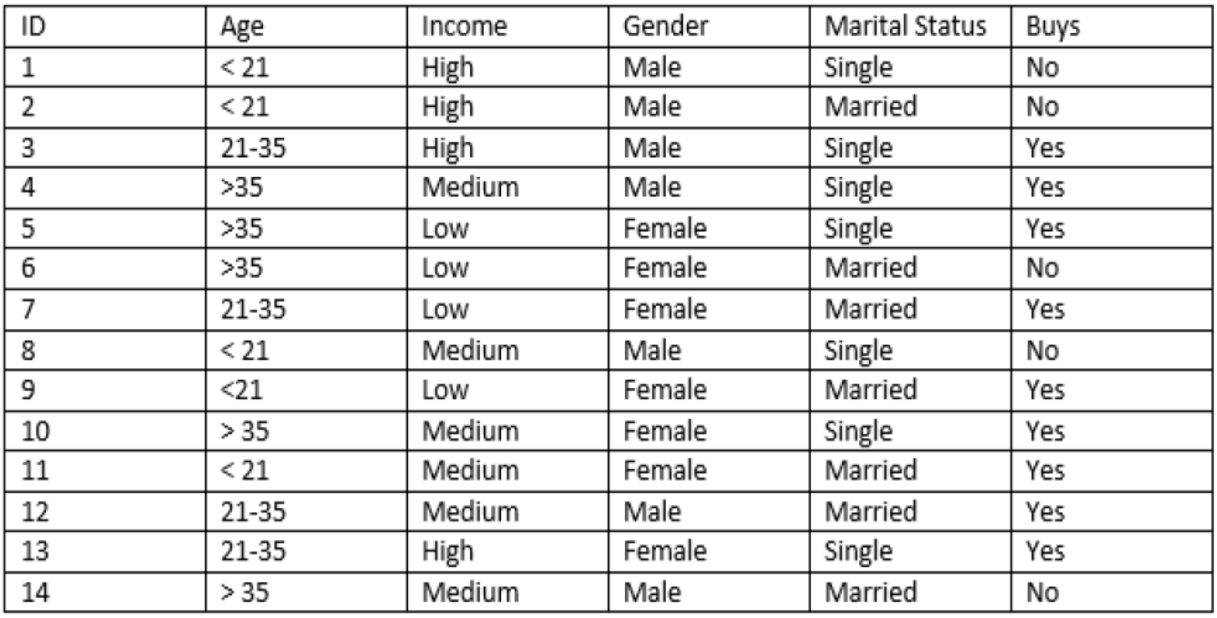
Install

You need to install pydotplus and graphviz. These can be installed with your package manager and pip.
Graphviz is a tool for drawing graphics using dot files. Pydotplus is a module to Graphviz’s Dot language.

Data Collection

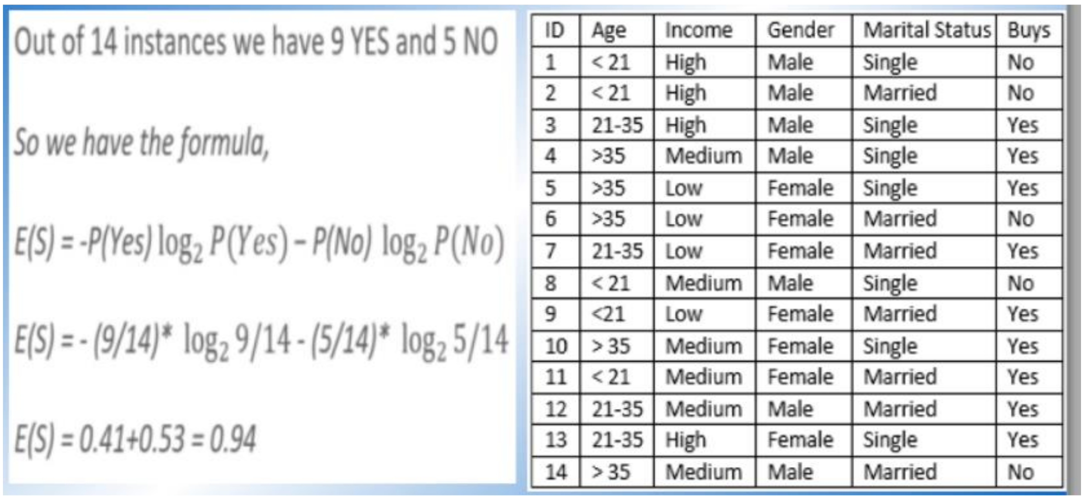
We start by defining the code and data collection. Let’s make the decision tree on Yes or No for Buys.

We start with the training data:



ID	Age	Income	Gender	Marital Status	Buys
1	< 21	High	Male	Single	No
2	< 21	High	Male	Married	No
3	21-35	High	Male	Single	Yes
4	>35	Medium	Male	Single	Yes
5	>35	Low	Female	Single	Yes
6	>35	Low	Female	Married	No
7	21-35	Low	Female	Married	Yes
8	< 21	Medium	Male	Single	No
9	<21	Low	Female	Married	Yes
10	> 35	Medium	Female	Single	Yes
11	< 21	Medium	Female	Married	Yes
12	21-35	Medium	Male	Married	Yes
13	21-35	High	Female	Single	Yes
14	> 35	Medium	Male	Married	No

Step1-Compute Entropy for Data Set



Out of 14 instances we have 9 YES and 5 NO

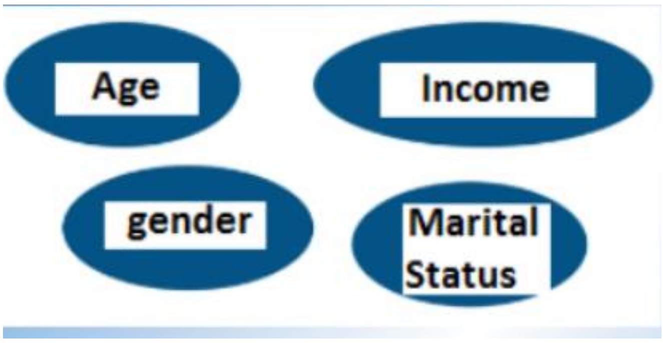
So we have the formula,

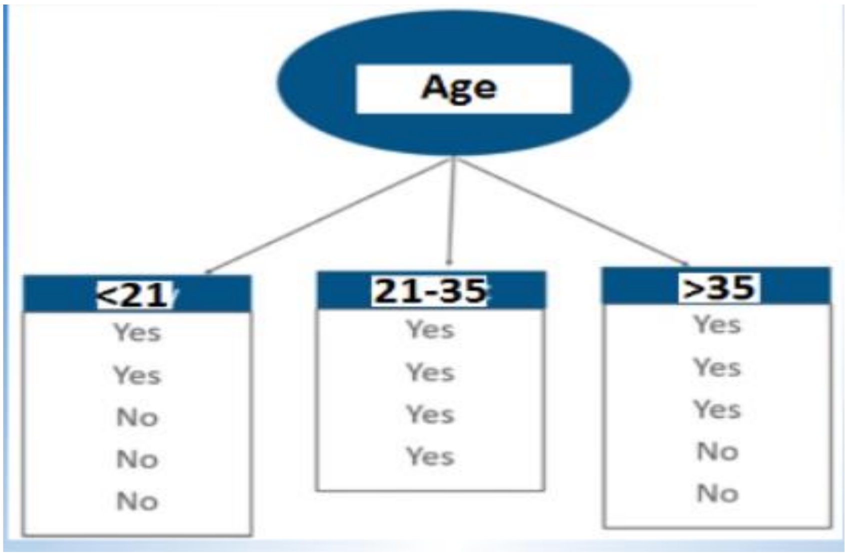
$E(S) = -P(Yes)\log_2 P(Yes) - P(No)\log_2 P(No)$

$E(S) = -(9/14)^{*}\log_2 9/14 - (5/14)^{*}\log_2 5/14$

$E(S) = 0.41+0.53 = 0.94$

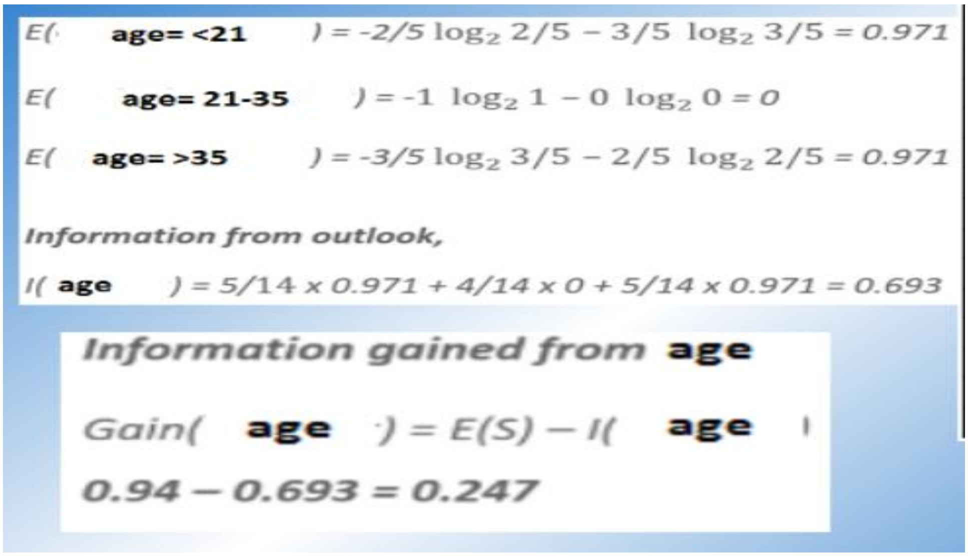
ID	Age	Income	Gender	Marital Status	Buys
1	< 21	High	Male	Single	No
2	< 21	High	Male	Married	No
3	21-35	High	Male	Single	Yes
4	>35	Medium	Male	Single	Yes
5	>35	Low	Female	Single	Yes
6	>35	Low	Female	Married	No
7	21-35	Low	Female	Married	Yes
8	< 21	Medium	Male	Single	No
9	<21	Low	Female	Married	Yes
10	> 35	Medium	Female	Single	Yes
11	< 21	Medium	Female	Married	Yes
12	21-35	Medium	Male	Married	Yes
13	21-35	High	Female	Single	Yes
14	> 35	Medium	Male	Married	No





Step2-Which Node to select as Root

Step3:Find Maximum Gain As Root



$E($ **age= <21** $) = -2/5 \log_2 2/5 - 3/5 \log_2 3/5 = 0.971$

$E($ **age= 21-35** $) = -1 \log_2 1 - 0 \log_2 0 = 0$

$E($ **age= >35** $) = -3/5 \log_2 3/5 - 2/5 \log_2 2/5 = 0.971$

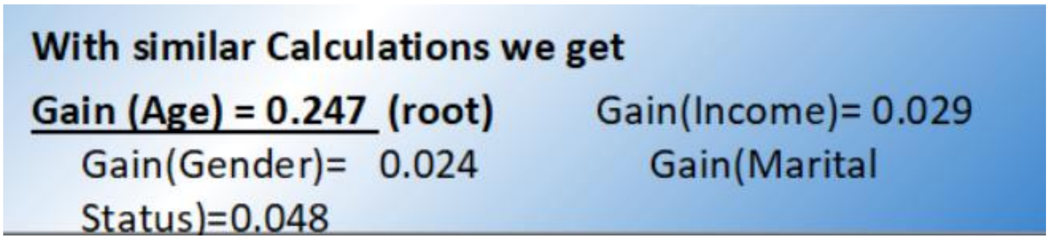
Information from outlook,

$I($ **age** $) = 5/14 \times 0.971 + 4/14 \times 0 + 5/14 \times 0.971 = 0.693$

Information gained from **age**

$Gain($ **age** $) = E(S) - I($ **age** $)$

$0.94 - 0.693 = 0.247$



With similar Calculations we get

<u>Gain (Age) = 0.247 </u> (root) Gain(Income)= 0.029

Gain(Gender)= 0.024 Gain(Marital Status)=0.048

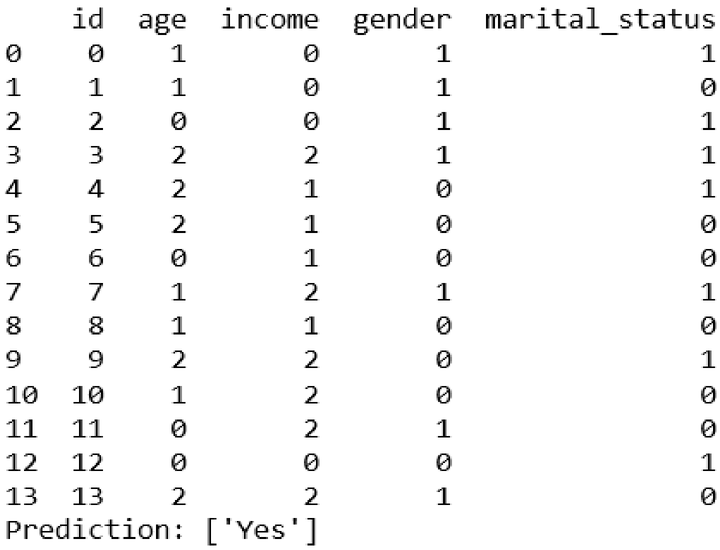
AGE IS Root Node Which has maximum Gain

Source Code :

```
#import packages
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

#reading Dataset
dataset=pd.read_csv("data.csv")
X=dataset.iloc[:,:-1]
y=dataset.iloc[:,5].values

#Perform Label encoding
from sklearn.preprocessing import LabelEncoder
labelencoder_X = LabelEncoder()

X = X.apply(LabelEncoder().fit_transform)
print (X)

from sklearn.tree import DecisionTreeClassifier
regressor=DecisionTreeClassifier()
```

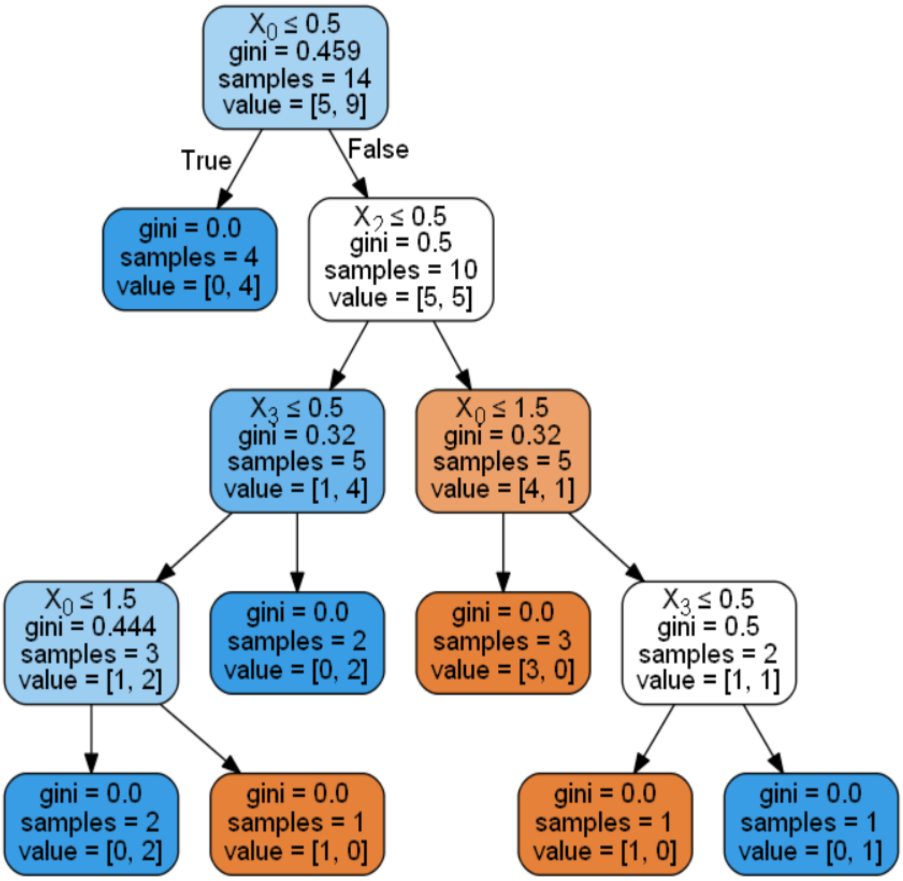
```
regressor.fit(X.iloc[:,1:5],y)

#Predict value for the given expression
X_in=np.array([1,1,0,0])

y_pred=regressor.predict([X_in])
print ("Prediction:", y_pred)

from sklearn.externals.six import StringIO
from IPython.display import Image
from sklearn.tree import export_graphviz
import pydotplus
# Create DOT data
dot_data = StringIO()

export_graphviz(regressor, out_file=dot_data, filled=True, rounded=True, special_characters=True)
# Draw graph
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
graph.write_png('Decision_Tree.png')
# Show graph
Image(graph.create_png())
```

Output:



```
    id  age  income  gender  marital_status
0    0    1       0       1               1
1    1    1       0       1               0
2    2    0       0       1               1
3    3    2       2       1               1
4    4    2       1       0               1
5    5    2       1       0               0
6    6    0       1       0               0
7    7    1       2       1               1
8    8    1       1       0               0
9    9    2       2       0               1
10  10    1       2       0               0
11  11    0       2       1               0
12  12    0       0       0               1
13  13    2       2       1               0
Prediction: ['Yes']
```

Out[2]:



DECISION TREE ADVANTAGES

1. Decision trees are powerful and popular tools for classification and prediction.
2. Simpler and ease of use.
3. They are able to handle both numerical and categorical attributes
4. Easy to understand.
5. State is recorded in memory.
6. Provide a clear indication of which fields are most important for prediction or classification.
7. Can be learned.

DECISION TREE DISADVANTAGES

1. Each tree is “unique” sequence of tests, so little common structure.
2. Perform poorly with many class and small data.
3. Need as many examples as possible.
4. Higher CPU cost - but not much higher.
5. Learned decision trees may contain errors.
6. Hugely impacted by data input.
7. Duplicate in sub trees

DECISION TREE APPLICATIONS

1. Medical diagnosis.
2. Credit risk analysis.
3. Library book use.

Conclusion: Thus, Students can learn how to create Decision Tree based on given decision, Find the Root Node of the tree using Decision tree Classifier successfully.

Prepared by:
Dr.T.Bhaskar
Subject Teacher

Approved by:
Dr. D.B. Kshirsagar
HoD-Computer