ASSIGNMENT NO.3

Title:

Decision Tree Classifier

Problem Definition:

A 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

Prerequisite:

Basic Python programming, Concept of Decision Tree Classifier

Software Requirements:

Jupyter notebook python 2.7/3.5, ubuntu OS

Hardware Requirement:

2GB RAM, 500 GB HDD, Laptop or pc

Objectives:

To Apply Decision Tree Classifier to find the root node and generate complete decision tree.

Outcomes:

Identification of root and generate decision tree

Theory:

A **decision tree** is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation.

Common terms used with Decision trees:

1. **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
4. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
6. **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

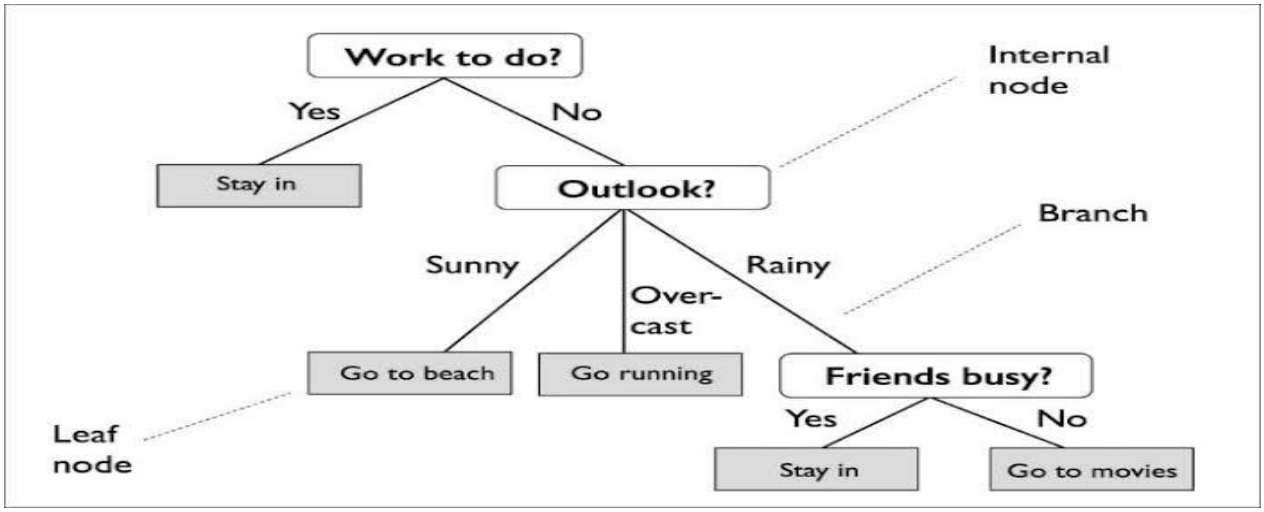

Fig 1.Example of Decision Tree

What are splitting measures?

With more than one attribute taking part in the decision-making process, it is necessary to decide the relevance and importance of each of the attributes, thus placing the most relevant at the root node and further traversing down by splitting the nodes. As we move further down the tree, the level of impurity or uncertainty decreases, thus leading to a better classification or best split at every node. To decide the same, splitting measures such as Information Gain, Gini Index, etc. are used.

1.Gini index or Gini impurity:

It measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. But what is actually meant by 'impurity'? If all the elements belong to a single class, then it can be called pure. The degree of Gini index varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes.

Formula for Gini Index

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

2.Entropy

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogeneous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is equally divided then it has entropy of one.

For example,

if we have items as number of dice face occurrence in a throw event as 1123,

the entropy is

p(1) = 0.5

p(2) = 0.25

p(3) = 0.25

entropy = - (0.5 * log(0.5)) - (0.25 * log(0.25)) -(0.25 * log(0.25)

= 0.45

3.Information Gain

Suppose we have multiple features to divide the current working set. What feature should we select for division? Perhaps one that gives us less impurity.

Suppose we divide the classes into multiple branches as follows, the information gain at any node is defined as

Information Gain (n) = Entropy(source/dataset) — ([weighted average] * entropy(children for feature))

Entropy(source/dataset) =-p(yes)*log p(yes)- p(no)*log p(no)
=-(9/14) * log(9/14)- (5/14) *log (5/14)
=0.94

Entropy of Age				
Age	**YES**	**NO**	**Entropy Calculation**	**Information Gain**
<21	2	3	-2/5 log(2/5)-3/5 log (3/5)=0.971	0.94-[(5/14)*0.971+(4/14)*0+(5/14)*0.97] =0.94-0.70
21-35	4	0	-4/4 log(4/4)-0/4 log(0/4)=0	
>35	3	2	-3/5 log(3/5)-2/5 log(2/5)=0.97	
Gain of Age=0.24				

Entropy of Income				
Income	**YES**	**NO**	**Entropy Calculation**	**Information Gain**
High	2	2	-2/4 log(2/4)-2/4 log(2/4)=1	0.94-[4/14*1+6/14 *0.932+4/14*0.811] =0.94-0.91
Medium	4	2	-4/6 log(4/6)-2/6 log(2/6)=0.932	
Low	3	1	-3/4 log(3/4)-1/4 log(1/4)=0.811	
Gain of Income=0.03				

Entropy of Gender				
Gender	**YES**	**NO**	**Entropy Calculation**	**Information Gain**
Male	3	4	-3/7 log(3/7)-4/7 log(4/7)=0.98488	0.94-[7/14 * 0.98488+7/14 * 0.59167] =0.94-0.788
Female	6	1	-6/7 log(6/7)-1/7 log(1/7)=0.59167	
Gain of Gender=0.152				

Entropy of Marital Status				
Marital Status	**YES**	**NO**	**Entropy Calculation**	**Information Gain**
Single	5	2	5/7 log(5/7)-2/7 log (2/7)=0.863	0.94-[7/14 *0.863 + 7/14*0.98488] =0.94-0.9237
Married	4	3	4/7 log(4/7)-3/7 log(3/7)=0.98488	
Gain of Marital status=0.0163				

Algorithm

- Import the Required Packages
- Read Given Dataset
- Separate independent and dependent features.
- Perform the label Encoding which will convert string value into Numerical values
- Import and Apply Decision Tree Classifier ()
- Predict value for the given Expression like [Age < 21, Income = Low, Gender = Female, Marital Status = Married]? In encoding Values [1,1,0,0]
- Find the root node by visualizing decision tree.

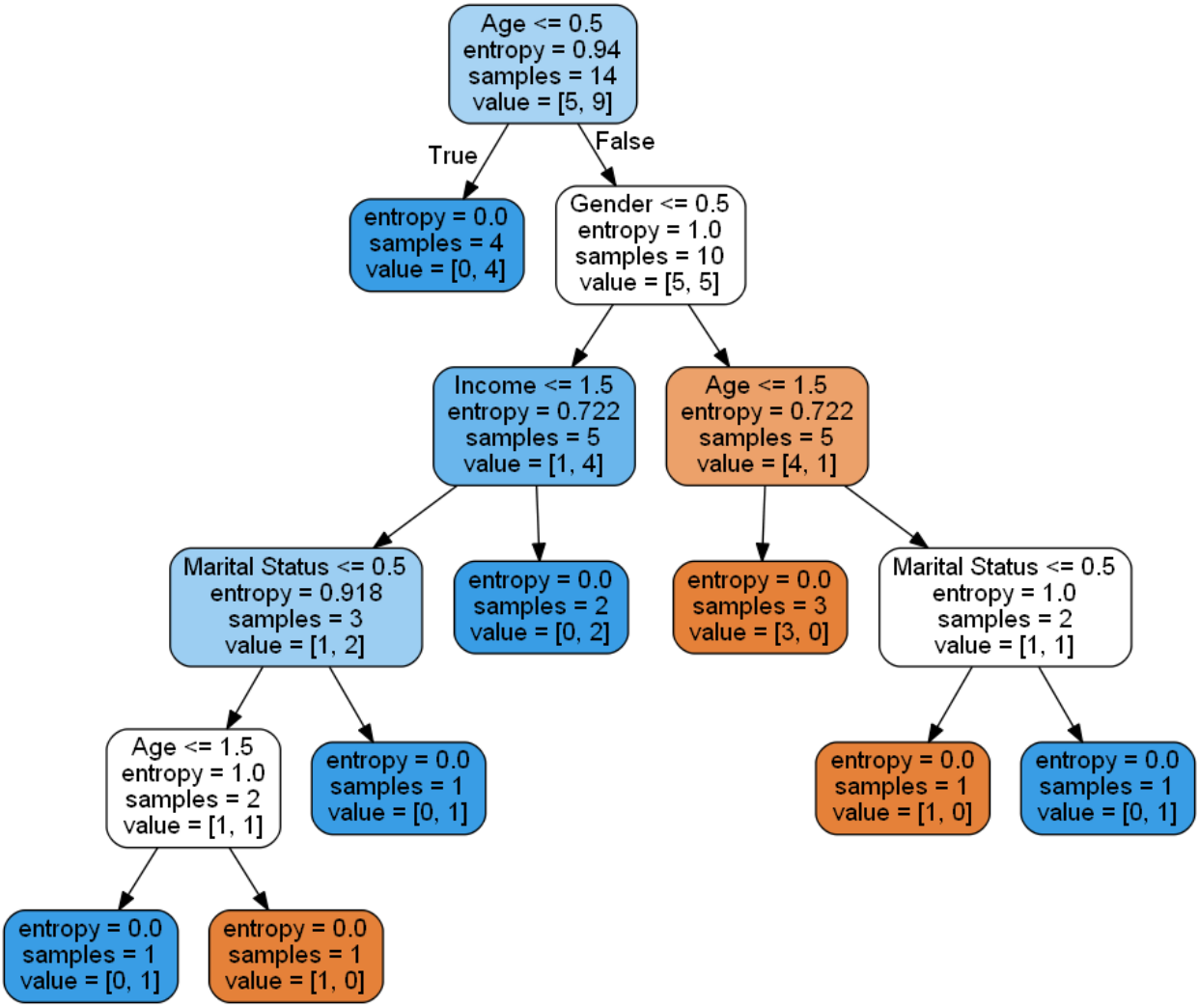

Fig 2.Decision tree visualization

Conclusion :

In this way, we studied what is decision tree and its important terminologies , how algorithm works and how to calculate entropy, information gain and on basis of this how to identify the root of the decision tree.