Sanjivani Rural Education Society's

College of Engineering, Kopargaon-423603

DEPARTMENT OF COMPUTER ENGINEERING

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

Aim:

Implement Decision Tree Classification

Problem Statement:

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

Input:CSV Dataset

Theory-

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.

Entropy

Entropy is degree of randomness of elements or in other words it is measure of impurity. Mathematically, it can be calculated with the help of probability of the items as:

$$H = -\sum p(x) \log p(x)$$

p(x) is probability of item x.

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(x) — ([weighted average] * entropy (children for feature))

What is the decision for the test data: [Age < 21, Income = Low, Gender = Female, Marital Status = Married]?Answer is whether Yes or No?

Out of 14 instances we have 9 YES and 5 NO

So we have the formula,

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

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

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

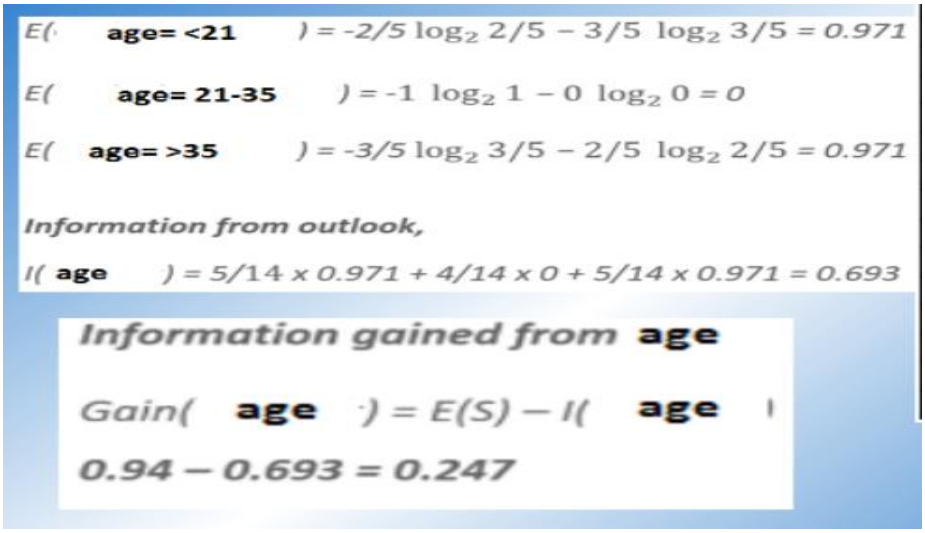
ID	Age	Income	Gender	Marital Status	Buys
1	< 21	High	Male	Single	No
2	< 21	High	Male	Married	No
3	21-35	High	Male	Single	Yes
4	>35	Medium	Male	Single	Yes
5	>35	Low	Female	Single	Yes
6	>35	Low	Female	Married	No
7	21-35	Low	Female	Married	Yes
8	< 21	Medium	Male	Single	No
9	<21	Low	Female	Married	Yes
10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

Step1-Compute Entropy for Data Set

Step2-Which Node to select as Root

A. Age
B. Income
C. Gender

D. Marital Status



Step3:Find Maximum Gain As Root

AGE IS Root Node Which has maximum Gain

Outcomes:

After completion of this assignment students are able to understand the decision tree classifier.

Algorithm

1. Import the Required Packages
2. Read Given Dataset
3. Perform the label Encoding Mean Convert String value into Numerical values
4. Import and Apply Decision Tree Classifier
5. Predict value for the given Expression like [Age < 21, Income = Low, Gender = Female, Marital Status = Married]? In encoding Values [1,1,0,0]
6. Import the packages for Create Decision Tree.
7. Check the Decision Tree Created based on Expression.

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']
```

i/p:1100 because [Age < 21(1), Income = Low(1) Gender = Female(0), Marital Status = Married(0)]

DECISION TREE ADVANTAGES

- Decision trees are powerful and popular tools for classification and prediction.
- Simpler and ease of use.
- They are able to handle both numerical and categorical attributes
- Easy to understand.

- State is recorded in memory.
- Provide a clear indication of which fields are most important for prediction or classification.
- Can be learned.

DECISION TREE DISADVANTAGES

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

DECISION TREE APPLICATIONS

- Medical diagnosis.
- Credit risk analysis.
- 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:	**Approved by:**
Dr.T.Bhaskar	Dr. D.B. Kshirsagar
Subject Teacher	Head, Computer Engineering