Assignment No: 2

Title: Decision tree classifier

Problem statement: A dataset collected in cosmetic shop showing details of customers or whether or not they responded to a special offer to buy a new lip stick is shown in table below. Use the dataset to build a decision tree, with buys as a target variable, to help in buying clip sticks in future.

- Find root node of decision tree
- According to the decision tree you have made previous training dataset, what is decision tree for test data.

[Age<25, Income = low, Gender = Female, Marital status = Married]

Learning objective: i] Learn how to apply decision tree classi to find root node of decision tree.

ii] Make decisions based on decision tree

Learning outcome: i) After completion of assignment, students are able to implement code for creating decision tree for given dataset.

Software requirement: Anaconda 3, python 3.7

Concept related theory :

i] Decision tree : A decision tree is a flow chart like structure in which each internal node represents a "test" on an attribute, each branch represents a class label i.e. outcome of test. The path from root to leaf represents classification rules -

- A decision tree consists of 3 types of nodes -

i) Decision nodes : Commonly represented as squares
ii] Chance nodes : Represented by circles.
iii] End nodes : Represented by triangles.

Algorithm used : ID3 (Iterative dichotomiser 3) is an ~~at~~ algorithm invented by Ross Quinlan used to generate decision tree from dataset.

Steps :

1. Calculate entropy of every attribute using the dataset.
2. Split set into subsets using attributes from which entropy is minimum.
3. Make decision tree node containing that attribute.
4. Recurse on subset using remaining attributes.

B] Input data set :

Age	Income	Gender	Marital status	Buys
≤21	High	M	Single	N
<21	High	M	married	N
21-35	High	M	Single	Y
>35	Medium	M	Single	Y
>35	low	F	Single	Y
>35	low	F	married	N
21-35	low	F	married	Y
<21	Medium	M	Single	N
<21	low	F	married	Y
>35	Medium	F	single	Y
<21	medium	F	married	Y
21-35	medium	M	married	Y
21-35	High	F	single	Y
>35	medium	M	married	N

Step 1: Calculate class entropy

$P = 'Y' = 9$

$M = 'N' = 5$

$$\text{Entropy (class)} = \frac{-P}{P+M} \log_2\left(\frac{P}{P+M}\right) - \frac{N}{P+M} \log_2\left(\frac{N}{P+M}\right)$$

$$= \frac{-9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right)$$

$= 0.409 + 0.530$
$= 0.940$

Step 2: Calulate gain for each attribute.

For age:

	P	N	$I(P_i, N_i)$
<21	2	3	0.970
21-35	4	0	0
>35	3	2	0.970

$$\text{Entropy} = 0.970 \times \frac{5}{14} + 0 + 0.970 \times \frac{5}{14}$$
$$= 0.692$$

Gain = Entropy (Buys) − Entropy

$= 0.940 - 0.692$
$= 0.248$

Similarly,

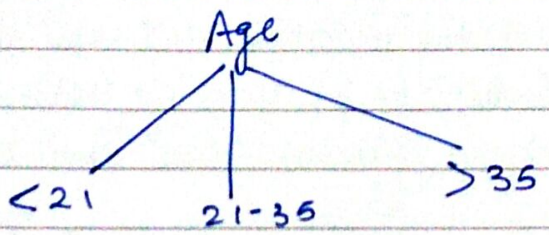
For 'income': Gain $= 0.940 - 0.910$
$= 0.030$

For 'gender': gain $= 0.940 - 0.7875$
$= 0.1525$

For 'maritial status': gain $= 0.940 - 0.923$
$= 0.017$

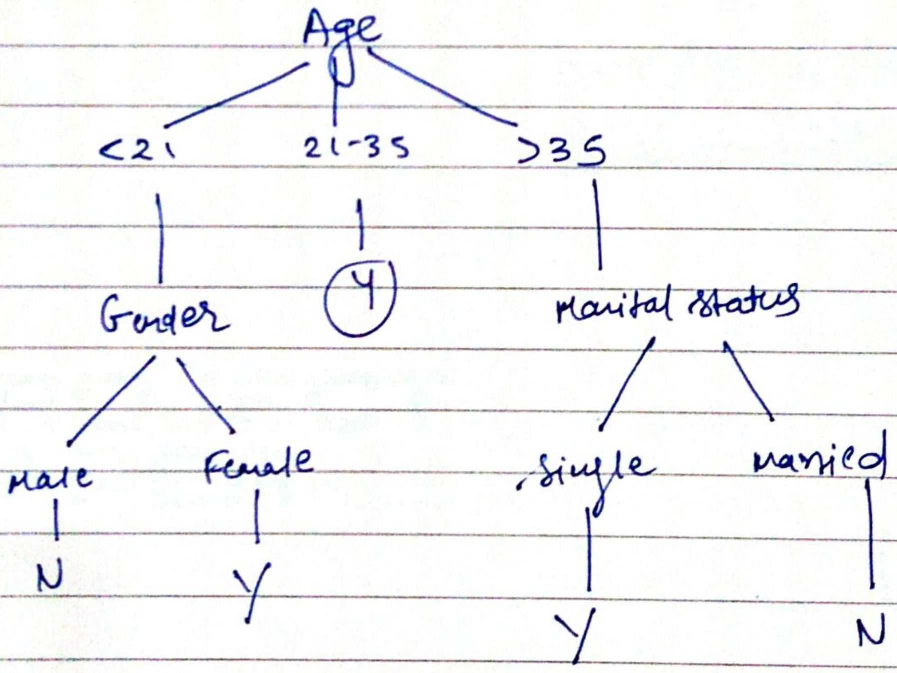
Highest information gain for attribute : Age.

The partial decision tree -



For every branch, take out subset and repeat step ① & step ②

After traversing each branch, final decision tree becomes -



Conclusion: Thus, we have build decision tree for given data.