Page No.
Date

Assignment A2

Title:- To find the decision based on a given scenario from a dataset using decision tree classifier.

Date of completion:- 3rd Feb 2021

Problem statement :-

A dataset collected in a Cosmetics shop showing details of customers and whether or not they responded to special offer to buy a new lipstick is shown in table below.

Use this dataset to build a decision tree, with Buys as the target variable, to help in buying lipsticks in the future. Find the root node of decision tree. According to the decision tree you have made from previous training dataset, what is the decision for the test data: [Age <21, Income = Low, Gender = Female, Marital Status = Married] ?

Objective :

To understand how the decision tree classifier algorithm works on the given data set.

Outcome :-

Students will be able to :

Find the decision based on a given scenario of people with income, gender and marital status information from a dataset using DTC.

Page No.
Date

Software and Hardware Requirements :-
i3/i5/i7 64 bit processor, OS - Linux 64 bit os
Editor - gedit / Eclipse.
Software - Notebook / Python.

Theory :-

Decision tree

A decision tree is a simple representation for classifying examples. It is a supervised Machine learning where the data is continuously split according to a certain parameter.

The decision tree consists of :-

Nodes: Test for the value of a certain attribute.
Edges/Branch: Correspond to the outcome of a test and connect to the next node/leaf.

Leaf Node: Terminal nodes that predict the outcome (represent class labels or class distribution)

Attribute selection Measures

It is heuristic for selecting the splitting criterion that partitions the data into the best possible manner. It is also known as splitting rules because it helps us to determine breakpoints for tuples on a given node. ASM provides a rank for each feature by explaining

the given dataset. Best score attribute will be selected as a splitting attribute. In the case of a continuous-valued attribute, split points for branches also need to be defined.

(1) Information gain:-

Entropy refers to the impurity in a group of examples. Information gain is the decrease in entropy. It computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

p_i is the probability that an arbitrary tuple in D belongs to class C_i.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times info(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

(2) Gain Ratio

Gain Ratio handles the issue of bias by normalising the information gain using split Info.

$$\text{SplitInfo}(A) = \sum_{j=1}^{v} \frac{|A_j|}{|A|} \times \log_2\left(\frac{|A_j|}{|A|}\right)$$

where, $|A_j| \neq |A|$ as the weight of the j^{th} partition.

v is the number of discrete values in attribute A.

gain ratio can be defined as

$$\text{Gain ratio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

The attribute with the highest gain ratio is chosen as the splitting attribute.

(3) Gini Index

It is used to create split points.

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

where p_i is the probability that a tuple in A belongs to class C_i

The gini index considers a binary split for each attribute.

$$\text{Gini}(A) = \frac{|A_1|}{|A|}\text{Gini}(A_1) + \frac{|A_2|}{|A|}\text{Gini}(A_2)$$

Page No.
Date

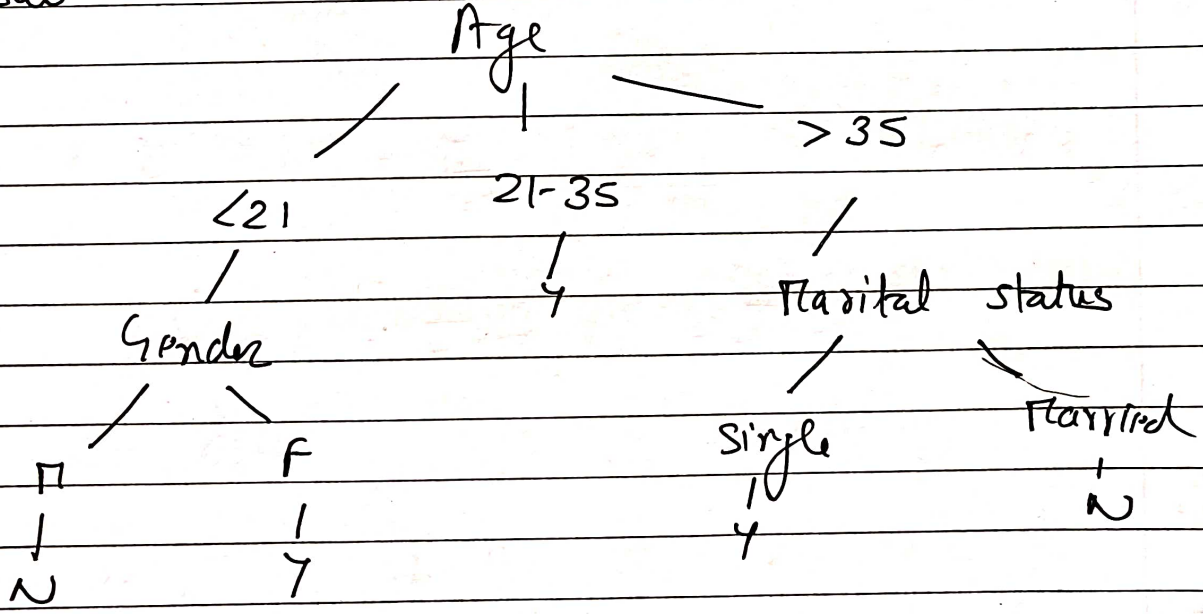
Δ Gini (A) = Gini (D) − Gini_A (D).

The attribute with minimum index is chosen as the splitting attribute.

Algorithm

(1) Calculate entropy of every attribute using the dataset.
(2) split the set into subsets using the attributes for which entropy is minimum (or in formation gain is maximum).
(3) Make a decision tree node containing that attribute.
(4) Recurse on subset using remaining attributes.

Result:



For the test data [Age < 21, Income = low, gender = Female, Marital status = Married] = "Yes".

Page No.
Date

Conclusion :

Successfully found the decision tree and built it using the decision tree classifier.