MACHINE INTELLIGENCE AND EXPERT SYSTEMS

AUTUMN SEMESTER - 2021

ASSIGNMENT-2 (Decision Tree)

Special Instructions: ***All the answers should be brief and to the point. All the parts of a question should be in one place. If you answer the same question multiple times and do not pen through them, only one of them will be considered for the correction and marking. Unfair answer scripts and unfair practices will be penalized.***

Q1 . Consider the following data collected by a business to understand their customers' behavior and whether they become repeat customers or not.

Serial No	Gender	Age Group	Pays by EMI?	Has visited shop more than once?
1	M	20-25	Yes	Yes
2	F	35-40	Yes	No
3	M	40-45	No	Yes
4	F	30-35	Yes	Yes
5	F	25-30	No	Yes
6	F	30-35	No	No
7	F	45-50	No	No
8	F	30-35	Yes	No

Assume all customers are between the ages of 20 to 50.

If an attribute does not provide any information gain or a leaf node cannot deterministically predict the output, specify the probabilistic behavior for that leaf. For instance, a person with gender = X, age group = Y and pays by EMI = Z will be a repeat customer with probability p (where p can be anything between 0 and 1).

a. Using **Information Gain** as the splitting decision, construct a decision tree and predict whether the following customers will visit the shop again or not.

Serial No	Gender	Age Group	Pays by EMI?	Has visited shop more than once?
8	F	25-30	Yes	??
9	M	35-40	No	??
10	F	45-50	Yes	??

b. Do you think the tree obtained in Q3a is a good decision tree? What could be the reason behind it having the structure it has. What would be a way to fix this issue?

Q2 . Apart from information gain calculated by the decrease in entropy, another criterion that can be used to construct decision trees is 'GainRatio', defined as:

$$GainRatio(S, A) := \ Gain(S, A) \ / \ SplitInformation(S, A)$$

where

$$SplitInformation(S, A) := \ \Sigma \ \frac{|Si \ |}{|S|} \ log \ \frac{|Si \ |}{|S|}$$

where Si is the subset of S for which A has value vi.

a. Using the same data given in the previous question and **GainRatio** as the splitting decision, construct a decision tree and predict whether the following customers will visit the shop again or not. Compare it with the tree from the previous question and how it differs.

Serial No	Gender	Age Group	Pays by EMI?	Has visited shop more than once?

8	F	25-30	Yes	??
9	M	35-40	No	??
10	F	45-50	Yes	??

b. If you were to use this tree and the tree from Q3 on real data, which do you think will perform better and why? In general, how would you decide which criterion to use when building a decision tree for a real problem: information gain or gain ratio?