Assignment 1

- Submit the assignment by **25th September by 11:00 PM**. (12 days from today).
- There will be a penalty for late submissions.
- Answers should be solved individually (this is not a group assignment).
- Please be careful, do not copy the assignments. **Zero marks** will be awarded to both persons.
- Try to submit a soft copy (using a word file) since handwritten softcopy is sometimes hard to read.

Q1. What do you understand by proximity measures? Compute the following measures on given vectors X and Y below:

X = (0,1,0,0,1,1,0,0,1,1)
Y = (1,0,0,0,1,1,0,0,0,1)

a) Cosine Similarity
b) Jaccard Similarity
c) Pearson's correlation

Q2. Consider the training examples given below for a binary classification problem.

(a) Compute the Gini index for the overall collection of training examples.
(b) Compute the Gini index for the Customer ID attribute.
(c) Compute the Gini index for the Gender attribute.
(d) Compute the Gini index for the Car Type attribute using multiway split.
(e) Compute the Gini index for the Shirt Size attribute using multiway split.
(f) Which attribute is better, Gender, Car Type, or Shirt Size?
(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.
(h) Construct a decision tree by considering the algorithms ID3.
(i) Construct a decision tree by considering the algorithms C4.5.

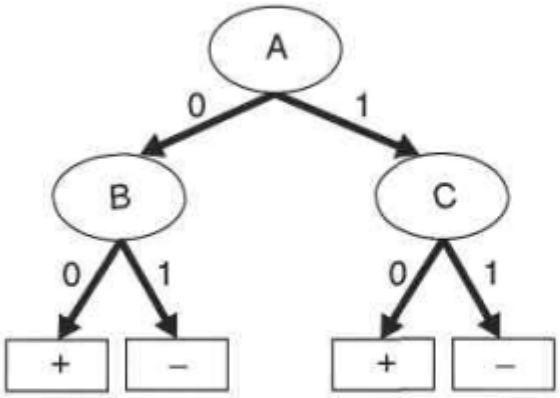
Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Q3. Consider the decision tree shown in the figure below:

(a) Compute the generalization error rate of the tree using the optimistic approach.

(b) Compute the generalization error rate of the tree using the pessimistic approach. (For simplicity, use the strategy of adding a factor of 0.5 to each leaf node.)

(c) Compute the generalization error rate of the tree using the validation set shown above.



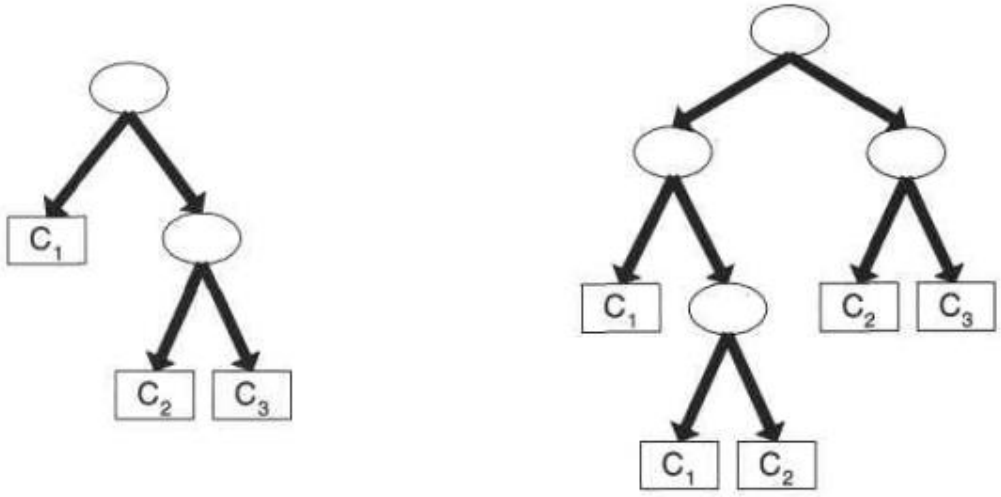
Training:

Instance	A	B	C	Class
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	–
5	1	0	0	+
6	1	0	0	+
7	1	1	0	–
8	1	0	1	+
9	1	1	0	–
10	1	1	0	–

Validation:

Instance	A	B	C	Class
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	–
15	1	0	0	+

Q4. Consider the decision trees shown in the figure below. Assume they are generated from a data set that contains 25 binary attributes and 3 classes, C1, C2, and C3.



Consider the left decision tree has 9 errors and the right decision tree has 6 errors. Compute the total description length of each decision tree according to the minimum description length principle. Which decision tree is better, according to the MDL principle?

Q5. Consider a binary classification problem with the following set of attributes and attribute values:

- Air Conditioner: {Working, Broken}
- Engine: {Good, Bad}
- Mileage: {High, Medium, Low}
- Rust: {yes, No}

Suppose a rule-based classifier produces the following rule set:

$$\text{Mileage} = \text{High} \longrightarrow \text{Value} = \text{Low}$$
$$\text{Mileage} = \text{Low} \longrightarrow \text{Value} = \text{High}$$
$$\text{Air Conditioner} = \text{Working, Engine} = \text{Good} \longrightarrow \text{Value} = \text{High}$$
$$\text{Air Conditioner} = \text{Working, Engine} = \text{Bad} \longrightarrow \text{Value} = \text{Low}$$
$$\text{Air Conditioner} = \text{Broken} \longrightarrow \text{Value} = \text{Low}$$

a) Are the rules mutually exclusive?
b) Is the rule set exhaustive?
c) Is ordering needed for this set of rules?
d) Do you need a default class for the rule set?