Dept. of CSE, IIT Jammu
Machine Learning (Shaifu Gupta)
Lab - 1

Abhilash (Ph.D Scholar)
(ID3, CART)

April 6, 2022

1 Questions

1. Given a Vehicles data set with the attributes External Combustion Engines (EC), Reciprocating and Rotary along with the target column **Steam Engine**. The 'N' represents NO and 'Y' represents YES.

Table 1: Vehicles with three attributes with target column Steam Engine.

EC	Reciprocating	Rotary	**Steam Engine**
Y	Y	Y	Y
Y	N	Y	Y
Y	Y	Y	Y
N	Y	N	N
Y	N	Y	Y
N	Y	Y	Y
N	N	N	N
Y	Y	N	Y
N	Y	N	N
N	Y	Y	Y
Y	Y	N	N
N	Y	Y	N
Y	Y	N	N
Y	N	Y	Y

1 Construct the entire decision tree and figure out if the vehicle is Steam Engine or not if (EC = YES, Reciprocating = NO, Rotary = NO).

2 Construct the entire decision tree using **Gini impurity**. Check if both the trees are same.

2. The second column of the data set in Table 2 is converted into categorical values based on the threshold set by C4.5 algorithm. Your task is to find out the threshold value in the second column where the continuous values can be splitted into **less than or equal to** decision boundary.

Table 2: Marks of the Students

ID	Marks Out of 100	**Remarks**
1	75	PASS
2	70	FAIL
3	83	FAIL
4	60	PASS
5	58	PASS
6	55	FAIL
7	54	PASS
8	62	FAIL
9	59	PASS
10	65	PASS
11	62	PASS
12	71	PASS
13	61	FAIL

2 Pseudo-Code for ID3

1. Read the Data set (14×4) and assign it to **table**.
2. Print the Data set.
3. Call the recursive ID3 on the **table**(14×4).
4. **If** the **table** has ≤ 2 columns then set the first column as splitting node.
5. **Else** Find the parent entropy.
6. Find the splitting node by comparing each column with the target column.
7. Form the subset of the original table for each feature=$\{$N,Y$\}$.
8. Call the recursive ID3 for the current subset table.

NOTE:-

1. Build your own Python code.
2. Don't use inbuilt packages.