Week 8 – Decision Tree Lab (Manual Calculations)

Dataset Description
The dataset contains 10 records. The target variable is Accident (Yes/No). The input attributes are Weather Condition, Road Condition, Traffic Condition, and Engine Problem. There are 5 instances with Accident = Yes and 5 instances with Accident = No.

1. Entropy of the Dataset
Entropy is used to measure the uncertainty in the dataset.

Total instances = 10
Yes = 5, No = 5

p(Yes) = 5/10 = 0.5
p(No) = 5/10 = 0.5

$\text{Entropy}(D) = -[0.5 \log_2(0.5) + 0.5 \log_2(0.5)] = 1.0$

This means the dataset is perfectly balanced.

2. Information Gain Calculation
Information Gain measures how much an attribute reduces uncertainty.
Gain(A) = Entropy(D) − Weighted Entropy after split

Weather Condition
Clear: 4 instances (2 Yes, 2 No) → Entropy = 1.000
Rain: 3 instances (1 Yes, 2 No) → Entropy ≈ 0.918
Snow: 3 instances (2 Yes, 1 No) → Entropy ≈ 0.918

Weighted Entropy = 0.951
Information Gain = 1.000 − 0.951 = 0.049

Road Condition
Bad: 4 instances (3 Yes, 1 No) → Entropy ≈ 0.811
Good: 4 instances (1 Yes, 3 No) → Entropy ≈ 0.811

Average: 2 instances (1 Yes, 1 No) → Entropy = 1.000

Weighted Entropy = 0.849
Information Gain = 1.000 − 0.849 = 0.151

Traffic Condition
High: 4 instances (3 Yes, 1 No) → Entropy ≈ 0.811
Light: 3 instances (1 Yes, 2 No) → Entropy ≈ 0.918
Normal: 3 instances (1 Yes, 2 No) → Entropy ≈ 0.918

Weighted Entropy = 0.875
Information Gain = 1.000 − 0.875 = 0.125

Engine Problem
Yes: 4 instances (3 Yes, 1 No) → Entropy ≈ 0.811
No: 6 instances (2 Yes, 4 No) → Entropy ≈ 0.918

Weighted Entropy = 0.875
Information Gain = 1.000 − 0.875 = 0.125

Based on Information Gain, Road Condition has the highest value and is selected as the root node.

3. Gini Impurity
Gini Impurity measures the probability of incorrect classification.

$Gini(D) = 1 - (0.5^2 + 0.5^2) = 0.5$

Weighted Gini values after splitting:
Weather = 0.467
Road = 0.400
Traffic = 0.417
Engine = 0.417

The lowest Gini value is for Road Condition, confirming it as the best split.

4. Final Decision Tree

The final decision tree uses Road Condition as the root node. Traffic Condition is used for the next level of splitting, resulting in clear decision outcomes for predicting whether an accident will occur.