DECISION TREE (ID3)

Dataset

Outlook	Temperature	Humidity	Windy	Match?
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rain	Mild	High	False	Yes
Rain	Cool	Normal	False	Yes
Rain	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rain	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rain	Mild	High	True	No

Assignment 1: Implementing a Decision Tree Using ID3 Algorithm

Objective: Implement the ID3 algorithm to classify a dataset based on given features.

1. **Instructions**:
    - Write a Python program to implement the ID3 algorithm.
    - Use a dataset with at least 5 features and 2 classes.
    - Calculate the entropy and information gain for each feature and use this to build the decision tree.
    - Visualize the resulting decision tree.
2. **Deliverables**:
    - Submit your Python code and a brief report explaining your implementation.
    - The report should include a visualization of the decision tree and a discussion of how the attributes were split.

Assignment 2: Evaluating Decision Trees with Different Feature Sets

Objective: Analyze how different feature sets impact the performance of a decision tree.

1. **Instructions**:
    - Using the same dataset from Assignment 1, build two decision trees using the ID3 algorithm.
    - In Tree 1, use all available features.
    - In Tree 2, remove two features that you believe are less significant.
    - Evaluate the performance of both trees using accuracy, precision, and recall.
    - Explain how removing features affects the tree structure and the classification results.
2. **Deliverables**:
    - Submit your Python code for both decision trees.
    - Provide a comparison of their performance and a discussion of the results.

Hints:

Step 1: Calculate Entropy for the Entire Dataset

Entropy is a measure of the uncertainty in the dataset. If all outcomes are the same (all "Yes" or all "No"), entropy is 0. If outcomes are perfectly mixed, entropy is 1.

The formula for entropy is:

$$H(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

Where:

- p_+ is the probability of "Yes"
- p_- is the probability of "No"

In our dataset:

- There are 9 "Yes" and 5 "No" outcomes.
- Total = 14 entries.

$$H(S) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right)$$

Let's compute it:

$$H(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$H(S) = -0.643 - 0.471 = 1.0$$

Step 2: Calculate Information Gain for Each Feature

Next, we calculate the information gain for each feature by calculating the entropy after the dataset is split based on that feature. We choose the feature that results in the highest information gain to be the root of the decision tree.

2.1: Information Gain for "Outlook"

"Outlook" has three possible values: Sunny, Overcast, and Rain. Let's split the dataset accordingly.

- **Sunny**: 5 examples → 2 Yes, 3 No.
  - $H(Sunny) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$
  - $H(Sunny) = 0.971$
- **Overcast**: 4 examples → 4 Yes, 0 No.
  - $H(Overcast) = 0$ (All outcomes are "Yes")
- **Rain**: 5 examples → 3 Yes, 2 No.
  - $H(Rain) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$
  - $H(Rain) = 0.971$

Now, we calculate the weighted average of the entropies:

$$H(Outlook) = \frac{5}{14}\cdot H(Sunny) + \frac{4}{14}\cdot H(Overcast) + \frac{5}{14}\cdot H(Rain)$$

$$H(Outlook) = \frac{5}{14}\cdot 0.971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.971 = 0.693$$

Information Gain for "Outlook":

$$IG(Outlook) = H(S) - H(Outlook) = 1.0 - 0.693 = 0.307$$

2.2: Information Gain for "Temperature"

"Temperature" has three values: Hot, Mild, and Cool.

- **Hot**: 4 examples → 2 Yes, 2 No.
  - $H(Hot) = 1$
- **Mild**: 6 examples → 4 Yes, 2 No.
  - $H(Mild) = 0.918$
- **Cool**: 4 examples → 3 Yes, 1 No.
  - $H(Cool) = 0.811$

$$H(Temperature) = \frac{4}{14} \cdot 1 + \frac{6}{14} \cdot 0.918 + \frac{4}{14} \cdot 0.811 = 0.911$$

Information Gain for "Temperature":

$$IG(Temperature) = 1.0 - 0.911 = 0.089$$

2.3: Information Gain for "Humidity"

"Humidity" has two values: High and Normal.

- **High**: 7 examples → 3 Yes, 4 No.
  - $H(High) = 0.985$
- **Normal**: 7 examples → 6 Yes, 1 No.
  - $H(Normal) = 0.592$

$$H(Humidity) = \frac{7}{14} \cdot 0.985 + \frac{7}{14} \cdot 0.592 = 0.789$$

Information Gain for "Humidity":

$$IG(Humidity) = 1.0 - 0.789 = 0.211$$

2.4: Information Gain for "Windy"

"Windy" has two values: True and False.

- **True**: 6 examples → 3 Yes, 3 No.
  - $H(True) = 1$
- **False**: 8 examples → 6 Yes, 2 No.
  - $H(False) = 0.811$

$$H(Windy) = \frac{6}{14} \cdot 1 + \frac{8}{14} \cdot 0.811 = 0.892$$

Information Gain for "Windy":

$$IG(Windy) = 1.0 - 0.892 = 0.108$$

Step 3: Selecting the Best Feature for the Root Node

We now compare the information gain values:

- $IG(Outlook) = 0.307$
- $IG(Humidity) = 0.211$
- $IG(Windy) = 0.108$
- $IG(Temperature) = 0.089$

Outlook has the highest information gain, so it will be the root of the decision tree.