Decision Trees

Instructor: Dr. Anam Qureshi

Introduction

- A **decision tree** is a machine learning algorithm used for both classification and regression tasks. It represents a series of decisions, making it easy to interpret the logic of the model. The tree consists of:
  - **Nodes**: Where a feature (attribute) is tested.
  - **Branches**: Outcomes of the test (Yes/No or True/False).
  - **Leaves**: Final outcomes or class labels.
- The main goal is to **split the data** into subsets that are more **homogeneous** (i.e., similar outcomes within the subset).

Homogenous (Pure/impure subsets)

- Examples of classification
- Examples of regression

Entropy and Information gain

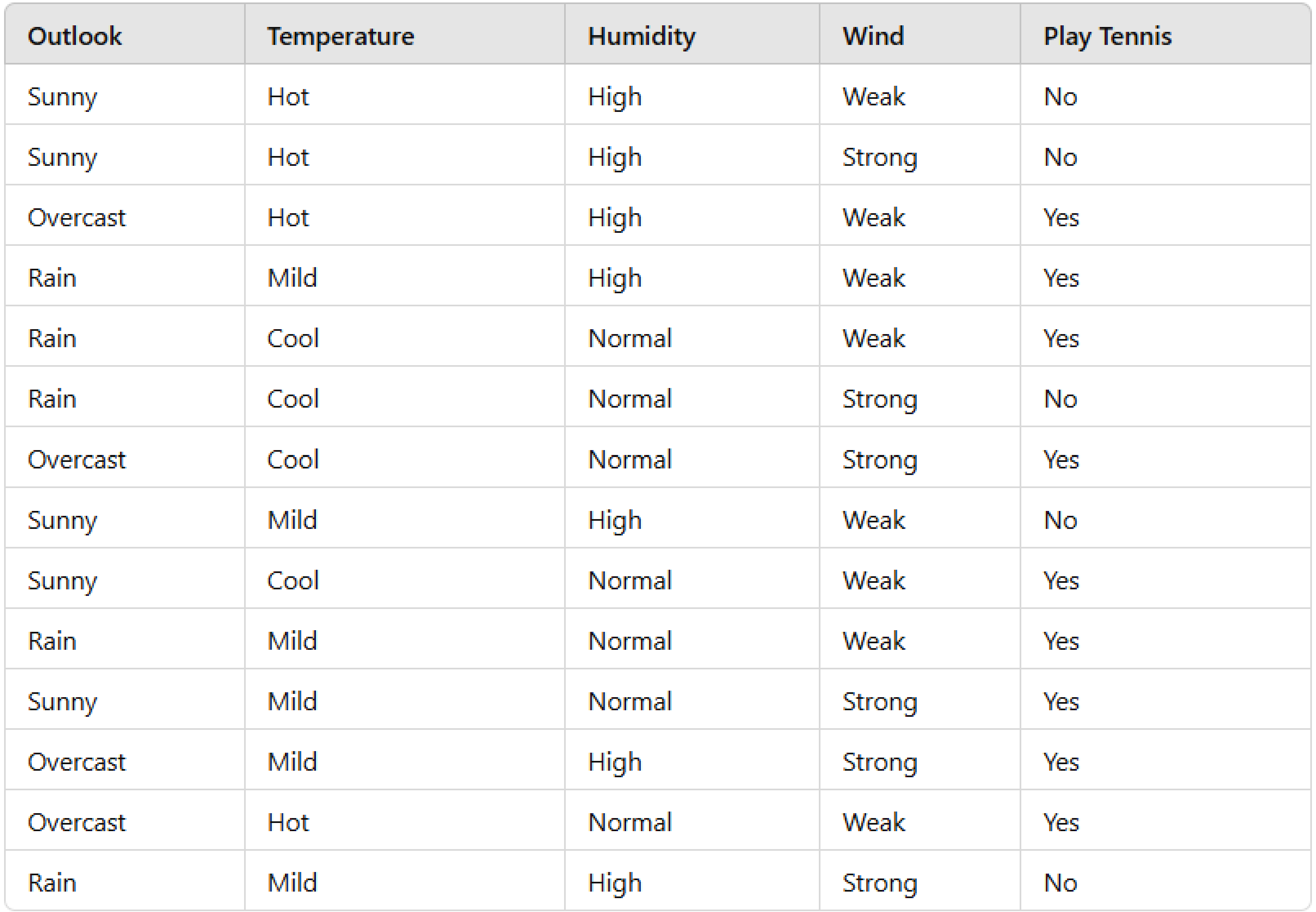
- **Entropy** measures the uncertainty or impurity in the data. The formula for entropy is:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

- **Information Gain** measures the reduction in entropy after splitting on an attribute:

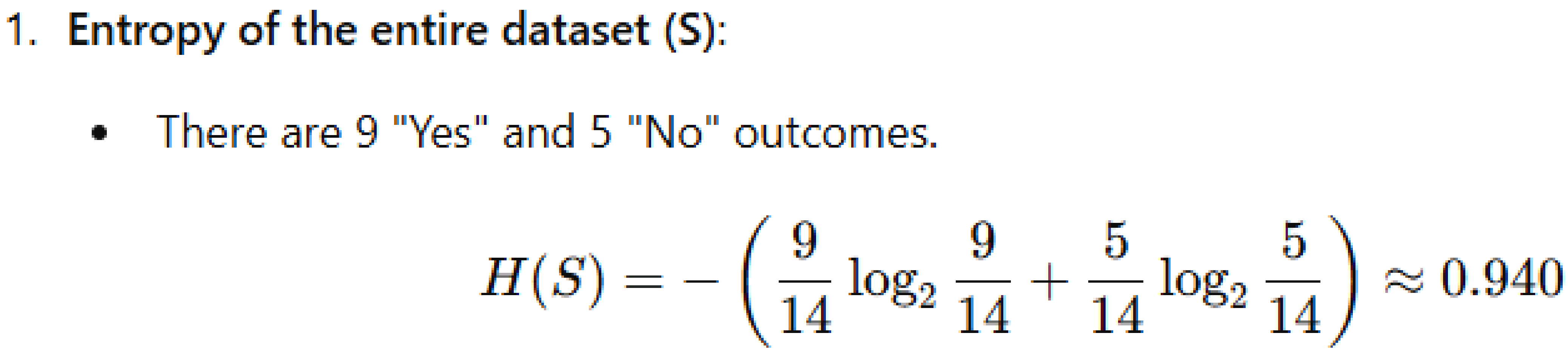
$$IG(S, A) = H(S) - \sum_{v \in A} \frac{|S_v|}{|S|} H(S_v)$$

Example (Classification Problem)



Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Step by Step Calculation of Information gain

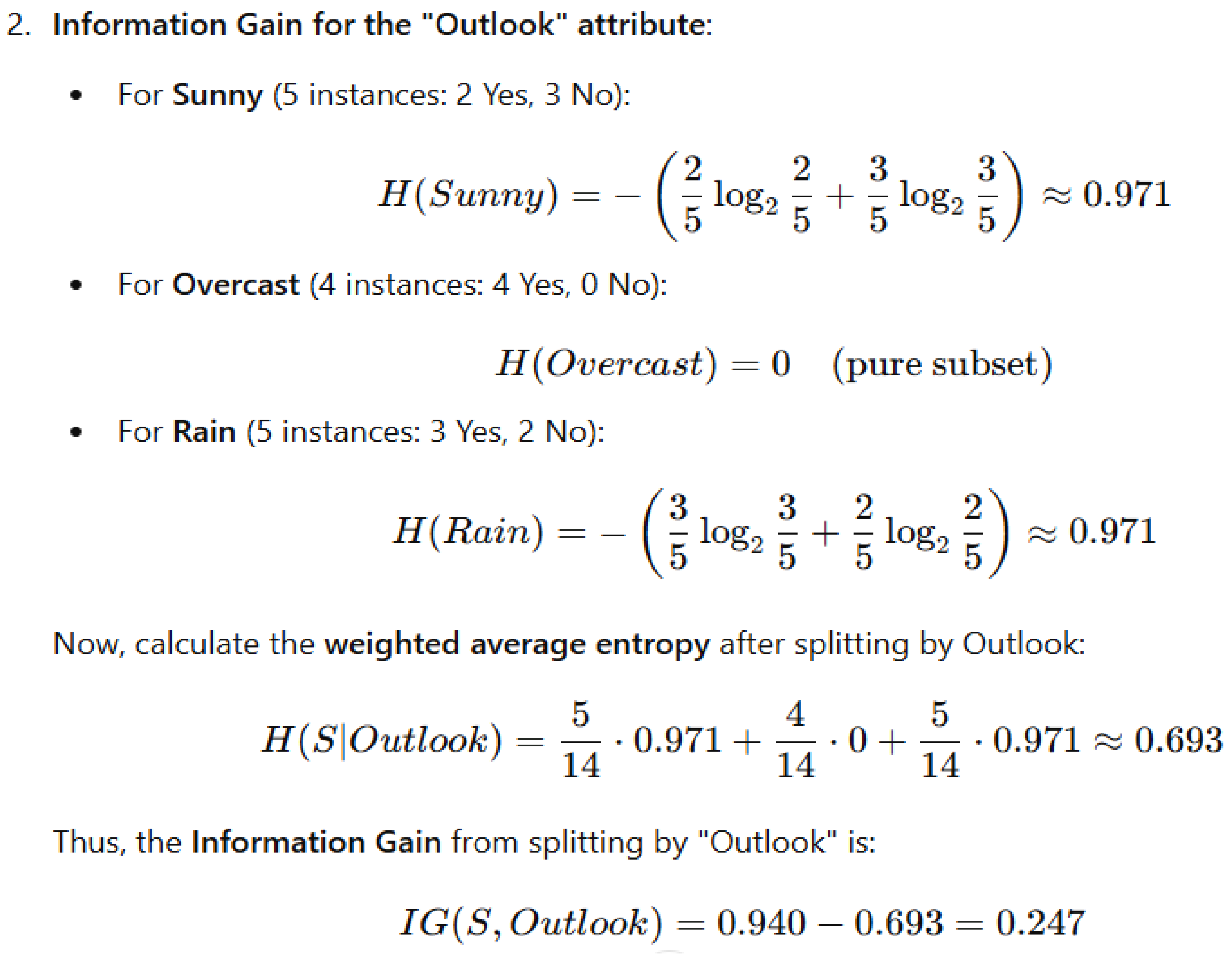


1. **Entropy of the entire dataset (S):**

- There are 9 "Yes" and 5 "No" outcomes.

$$H(S) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right) \approx 0.940$$

Information gain for “Outlook” attribute



2. **Information Gain for the "Outlook" attribute**:
   - For **Sunny** (5 instances: 2 Yes, 3 No):

$$H(Sunny) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.971$$

   - For **Overcast** (4 instances: 4 Yes, 0 No):

$$H(Overcast) = 0 \quad \text{(pure subset)}$$

   - For **Rain** (5 instances: 3 Yes, 2 No):

$$H(Rain) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \approx 0.971$$

Now, calculate the **weighted average entropy** after splitting by Outlook:

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

Thus, the **Information Gain** from splitting by "Outlook" is:

$$IG(S, Outlook) = 0.940 - 0.693 = 0.247$$

Information gain for “Temperature”, “Humidity”, “Wind”

- Temperature ??
- Humidity ??
- Wind ??

Root Node (highest information gain)

- Outlook

Further Calculations

- Outlook->Overcast-> Yes
- Outlook -> Sunny->Humidity
- Outlook-> Rain -> Wind

Activity

- When to use decision trees?
- What are the pros and cons of using decision trees?

Example (Regressor)

- To be continued