shapes.geometric, arrows

Problem Statement

Objective: Develop a Decision Tree model to predict whether a candidate will be hired based on categorical attributes.
Attributes:

- Degree (Bachelor's, Master's, PhD)
- Work Experience (None, 1-2 Years, 2-4 Years, 3-5 Years, 5+ Years)
- Technical Skills (Basic, Intermediate, Expert)
- Leadership (Yes, No)
- Target Variable: Hired (Yes, No)

Dataset

Degree	Work Exp	Technical Skills	Leadership	Hired
Bachelor's	None	Basic	No	No
Master's	2-4 Years	Intermediate	Yes	Yes
PhD	5+ Years	Expert	Yes	Yes
Bachelor's	1-2 Years	Intermediate	No	No
Master's	3-5 Years	Expert	Yes	Yes
Bachelor's	3-5 Years	Intermediate	Yes	Yes
PhD	None	Expert	No	No
Bachelor's	5+ Years	Intermediate	Yes	Yes
Master's	None	Basic	No	No
PhD	2-4 Years	Expert	Yes	Yes

Entropy Calculation

The entropy for a binary classification problem is given by:

$$H(S) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \tag{1}$$

where:

- p_1 is the probability of class "Yes" (Hired)
- p_2 is the probability of class "No" (Not Hired)

Based on the dataset:

$$H(S) = -\frac{5}{10} \log_2 \frac{5}{10} - \frac{5}{10} \log_2 \frac{5}{10} = 1.0 \tag{2}$$

This represents the maximum uncertainty in classification.

Information Gain Calculation for Degree

The information gain for splitting on **Degree** is calculated as:

$$IG(S, \text{Degree}) = H(S) - \sum_{v \in \text{Values}} \frac{|S_v|}{|S|} H(S_v) \tag{3}$$

where S_v represents subsets of data partitioned by Degree.
Based on the dataset:

$$H(Bachelor's) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$$

$$H(Master's) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$$

$$H(PhD) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$$

Weighted entropy:

$$H(S|\text{Degree}) = \frac{4}{10} \times 1.0 + \frac{3}{10} \times 0.918 + \frac{3}{10} \times 0.918 = 0.95 \tag{4}$$

Information Gain:

$$IG(S, \text{Degree}) = 1.0 - 0.95 = 0.05 \tag{5}$$

Information Gain Calculation for Work Experience

The information gain for splitting on **Work Experience** is calculated as:

$$IG(S, \text{Work Experience}) = H(S) - \sum_{v \in \text{Values}} \frac{|S_v|}{|S|} H(S_v) \quad (6)$$

where S_v represents subsets of data partitioned by Work Experience.

Based on the dataset:

$$H(None) = 0.0$$
$$H(1 - 2\,Years) = 0.0$$
$$H(2 - 4\,Years) = 0.0$$
$$H(3 - 5\,Years) = 0.0$$
$$H(5 + Years) = 1.0$$

Weighted entropy:

$$H(S|\text{Work Experience}) = \frac{3}{10} \times 0 + \frac{1}{10} \times 0 + \frac{2}{10} \times 0 + \frac{2}{10} \times 0 + \frac{2}{10} \times 1.0 = 0.2 \quad (7)$$

Information Gain Calculation for Technical Skills

The information gain for splitting on **Technical Skills** is calculated as:

$$IG(S, \text{Technical Skills}) = H(S) - \sum_{v \in \text{Values}} \frac{|S_v|}{|S|} H(S_v) \tag{9}$$

where S_v represents subsets of data partitioned by Technical Skills. Based on the dataset:

$$H(Basic) = 0.0$$
$$H(Intermediate) = 0.971$$
$$H(Expert) = 0.918$$

Weighted entropy:

$$H(S|\text{Technical Skills}) = \frac{2}{10} \times 0.0 + \frac{5}{10} \times 0.971 + \frac{3}{10} \times 0.918 = 0.76 \tag{10}$$

Information Gain:

$$IG(S, \text{Technical Skills}) = 1.0 - 0.76 = 0.24 \tag{11}$$

Information Gain Calculation for Leadership

The information gain for splitting on **Leadership** is calculated as:

$$IG(S, \text{Leadership}) = H(S) - \sum_{v \in \text{Values}} \frac{|S_v|}{|S|} H(S_v) \tag{12}$$

where S_v represents subsets of data partitioned by Leadership.
Based on the dataset:

$$H(Yes) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.722$$
$$H(No) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$$

Weighted entropy:

$$H(S|\text{Leadership}) = \frac{5}{10} \times 0.722 + \frac{5}{10} \times 0.722 = 0.722 \tag{13}$$

Information Gain:

$$IG(S, \text{Leadership}) = 1.0 - 0.722 = 0.278 \tag{14}$$

Information Gain for All Attributes

Attribute	Information Gain
Degree	0.05
Work Experience	0.80
Technical Skills	0.24
Leadership	0.278

Decision Tree Construction

Based on the highest information gain, the root node is chosen as:

- **Root Node:** Work Experience (IG = 0.80)

Decision Tree Splitting:

- If Work Experience = None or 1-2 Years, then Hired = No
- If Work Experience = 2-4 Years or 3-5 Years, then Hired = Yes
- If Work Experience = 5+ Years, further split on Leadership:
  - If Leadership = Yes, then Hired = Yes
  - If Leadership = No, then Hired = No

Decision Tree Visualization