Solution to Decision Tree Questions with Calculations

(a). Possible Functions

1. Total number of possible functions to map the four features to a boolean decision:
$$2^{3\times 3\times 2\times 2} = 2^{36}$$
There are 2^{36} possible functions.
2. $$2^{36-10} = 2^{26}$$

(b). Entropy of the Labels

The formula for entropy is:

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

From the dataset:

- Positive examples (+): 6
- Negative examples $(-)$ :: 4

$$p(+) = \frac{6}{10}, \quad p(-) = \frac{4}{10}$$

$$\begin{aligned} H(S) &= -(0.6 \cdot \log_2(0.6) + 0.4 \cdot \log_2(0.4)) \\ &= -(0.6 \cdot (-0.737) + 0.4 \cdot (-1.322)) \\ &\approx 0.971 \end{aligned}$$

The entropy of the labels is approximately 0.971.

(c). Information Gain

Information gain for each feature is computed as:

$$IG(T, A) = H(T) - \sum_{v \in Values(A)} \frac{|T_v|}{|T|} H(T_v)$$

The calculations for each feature are as follows:

Weather:

- Values: {Sunny, Cloudy, Rainy}

- Calculations:

$$H(Sunny) = -(\frac{1}{3}\log_2(\frac{1}{3}) + \frac{2}{3}\log_2(\frac{2}{3})) \approx 0.918$$
$$H(Cloudy) = -(\frac{3}{5}\log_2(\frac{3}{5}) + \frac{2}{5}\log_2(\frac{2}{5})) \approx 0.971$$
$$H(Rainy) = -(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) = 1.0$$

Weighted entropy:

$$H(Weather) = \frac{3}{10}(0.918) + \frac{5}{10}(0.971) + \frac{2}{10}(1.0) \approx 0.961$$

$$IG(Weather) = 0.971 - 0.961 = 0.010$$

Temp:

- Values: {Hot, Warm, Cold}
- Calculations:

$$H(Hot) = -(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) \approx 0.918$$
$$H(Warm) = -(\frac{3}{4}\log_2(\frac{3}{4}) + \frac{1}{4}\log_2(\frac{1}{4})) = 0.811$$
$$H(Cold) = -(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) \approx 0.918$$

Weighted entropy:

$$H(Temp) = \frac{3}{10}(0.918) + \frac{4}{10}(0.811) + \frac{3}{10}(0.918) \approx 0.875$$

$$IG(Temp) = 0.971 - 0.875 = 0.096$$

Crowd:

- Values: {Busy, Empty}
- Calculations:

$$H(Busy) = -(\frac{1}{5}\log_2(\frac{1}{5}) + \frac{4}{5}\log_2(\frac{4}{5})) \approx 0.722$$
$$H(Empty) = -(\frac{5}{5}\log_2(\frac{5}{5}) + 0) = 0$$

Weighted entropy:

$$H(Crowd) = \frac{5}{10}(0.722) + \frac{5}{10}(0) = 0.361$$

$$IG(Crowd) = 0.971 - 0.361 = 0.610$$

Time:

- Values: {Morning, Afternoon}
- Calculations:

$$H(Morning) = -(\frac{5}{7}\log_2(\frac{5}{7}) + \frac{2}{7}\log_2(\frac{2}{7})) \approx 0.863$$
$$H(Afternoon) = -(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) \approx 0.918$$

Weighted entropy:

$$H(Time) = \frac{7}{10}(0.863) + \frac{3}{10}(0.918) = 0.8795$$
$$IG(Time) = 0.971 - 0.8795 = 0.092$$

Feature	Information Gain
Crowd	0.610
Temp	0.096
Time	0.092
Weather	0.010

Table 1: Information Gain for Each Feature

(d). Root of the Decision Tree

Based on the information gain, the feature **Crowd** has the highest value (0.610). Therefore, **Crowd** should be selected as the root of the decision tree.

(e). Constructing the Decision Tree

Using **Crowd** as the root, the decision tree can be constructed as follows:

- If **Crowd** is *Empty*:
  - If **Weather** is *Sunny* or *Rainy*, the condition is +.
  - If **Weather** is *Cloudy*, the condition is +.
- If **Crowd** is *Busy*:
  - Based on **Time** and **Temp**, refine further decisions (details can be split accordingly).

(f). Predictions and Accuracy

Using the constructed decision tree, the predictions for the test dataset are:

Accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} = \frac{3}{3} = 100\%.$$

Weather	Temp	Crowd	Time	Actual	Predicted
Cloudy	Hot	Busy	Morning	$-$	$-$
Sunny	Cold	Busy	Afternoon	$-$	$-$
Rainy	Warm	Empty	Afternoon	$+$	$+$

Table 2: Predictions for the Test Dataset

2. Gini Impurity

(a). Entropy of the Labels

The Gini impurity is defined as:

$$GiniImpurity = 1 - \sum_i p_i^2$$

Calculations for each feature:

Weather:

- Values: {Sunny, Cloudy, Rainy}

$$G(Sunny) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) \approx 0.444$$

$$G(Cloudy) = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right) \approx 0.48$$

$$G(Rainy) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 0.5$$

Weighted Gini Impurity:

$$G(Weather) = \frac{3}{10}(0.444) + \frac{5}{10}(0.48) + \frac{1}{10}(0.5) \approx 0.473$$

$$IG(Weather) = 0.48 - 0.473 = 0.007$$

Crowd:

- Values: {Busy, Empty}

$$G(Busy) = 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right) \approx 0.32$$

$$G(Empty) = 1 - \left(\left(\frac{5}{5}\right)^2\right) = 0$$

Weighted Gini Impurity:

$$G(Crowd) = \frac{5}{10}(0.32) + \frac{5}{10}(0) = 0.16$$

$$IG(Crowd) = 0.48 - 0.16 = 0.32$$

Temp:

- Values: {Hot, Warm, Cold}

$$G(Hot) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) \approx 0.444$$
$$G(Warm) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) \approx 0.375$$
$$G(Cold) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) \approx 0.444$$

Weighted Gini Impurity:

$$G(Temp) = \frac{3}{10}(0.444) + \frac{4}{10}(0.375) + \frac{3}{10}(0.444) = 0.416$$
$$IG(Temp) = 0.48 - 0.416 = 0.064$$

Time:

- Values: {morning, afternoon}

$$G(morning) = 1 - \left(\left(\frac{5}{7}\right)^2 + \left(\frac{2}{7}\right)^2\right) \approx 0.408$$
$$G(afternoon) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) \approx 0.444$$

Weighted Gini Impurity:

$$G(Time) = \frac{3}{10}(0.444) + \frac{7}{10}(0.408) = 0.419$$
$$IG(Time) = 0.48 - 0.419 = 0.061$$

(b). Information Gain by Gini impurity

(c). Constructing the Decision Tree

Based on the information gain, **Crowd** should be selected as the root of the decision tree. Using **Crowd** as the root, the decision tree can be constructed as follows:

- If **Crowd** is *Empty*:

Feature	Information Gain (using Gini impurity)
Crowd	0.32
Temp	0.064
Time	0.061
Weather	0.007

Table 3: Information Gain using Gini Impurity

  - If **Weather** is *Sunny* or *Rainy*, the condition is +.
  - If **Weather** is *Cloudy*, the condition is +.

- If **Crowd** is *Busy*:
  - Based on **Time** and **Temp**, refine further decisions (details can be split accordingly).

Conclusion : The tree built by Gini impurity is the same as by Entropy