Homework 1

CS461

Age	Gender	Car_Ownership	Will_Buy_Car
50	M	Yes	Yes
50	M	Yes	No
50	M	Yes	Yes
30	M	No	No
30	M	Yes	Yes
10	M	No	No
10	M	No	No
10	M	Yes	No

Question 1 (30 Points)

1. At root node, what is the Information Gain if we split by ‘Age‘, ‘Gender‘, and 'Car_Ownership' respectively? (10 points)

$$\text{Entropy}(S) \equiv -p_{\oplus} log_2 p_{\oplus} - p_{\ominus} log_2 p_{\ominus}$$

$\text{Entropy}(\text{Age}) \equiv -\frac{3}{8}\log_2\frac{3}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{3}{8}\log_2\frac{3}{8} \approx 1.56127812446$
$\text{Entropy}(\text{Gender}) \equiv -\frac{8}{8}\log_2\frac{8}{8} = 0$
$\text{Entropy}(\text{Car Ownership}) \equiv -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} \approx 0.954434002925$

2. Which column is used to make the first split? (10 points)

Age is used to make the first split.

3. Once trained, What will be the output on a new sample with Age = 50, Car Ownership = No, Gender = M ? Can you train the same tree without the Gender column, why or why not? (10 points)

The output on the new sample will be No. The same tree can be trained without the Gender column because the Gender column contains only a single value.