Homework 3
Theory Part

Question 1: Given the dataset (where each sample has 3 feature values: a, b, and c) below, compute the Gini impurity for the condition S1: a > 10. Please show the estimations of Pr (class = +1|a > 10) and Pr (class = −1|a > 10). If you do not show these two but a result, you will get no point.

Sample Number	feature a	feature b	feature c	class
1	12	3	5	+1
2	4	7	6	+1
3	5	4	8	+1
4	6	6	7	+1
5	7	5	1	-1
6	8	2	2	-1
7	9	6	3	-1
8	11	8	1	-1

Answer: Gini impurity GS is
$g(S) = \sum_{c=\pm 1} \Pr(class = c \mid S) * (1 - \Pr(class = c \mid S)$

$$pr(class = +1 \mid a > 10) = 1/2$$
$$pr(class = -1 \mid a > 10) = 1/2$$
$$g(S) = P(a > 10)$$
$$= 0.5 * (1 - 0.5) + 0.5 * (1 - 0.5)$$
$$= 0.25 + 0.25$$
$$= 0.5$$

Question 2: Do the same for a condition S2: a ≤ 5. Again, intermediate steps need to be shown.

$$pr(class = +1 \mid a \le 5) = 2/2 = 1$$
$$pr(class = -1 \mid a \le 5) = 0$$
$$g(S) = P(a \le 5)$$
$$= 1 * (1 - 1) + 0 * (1 - 0)$$
$$= 0 + 0$$
$$= 0$$

Question 3: Based on the results from the two problems above, compute the expectation for Gini Impurity for the feature and threshold 5. Please show the estimation of the probabilities of both conditions, i.e., P (a > 5) 1 and P(a ≤ 5). If you just show a final result, no point.

$$pr(class = +1 \mid a > 5) = 2/6$$
$$pr(class = -1 \mid a > 5) = 4/6$$
$$g(S) = P(a > 5)$$
$$= \frac{2}{6} * (1 - \frac{2}{6}) + \frac{4}{6} * (1 - \frac{4}{6})$$
$$= \frac{2}{6} * \frac{4}{6} * 2$$
$$= \frac{4}{9}$$

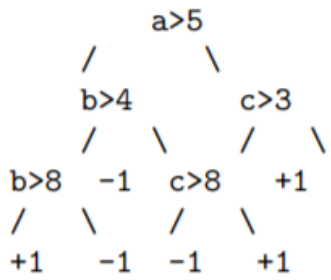
And:

$$P(a \le 5)$$
$$= 1 * (1 - 1) + 0 * (1 - 0)$$
$$= 0 + 0$$
$$= 0$$

The expectation for Gini Impurity for the feature and the threshold 5.
Expectation of impurity: E(F, T) = P(F > T)g(F > T) + P(F ≤ T)g(F ≤ T)

$$= \frac{6}{8} * \frac{4}{9} + \frac{2}{8} * 0$$
$$= \frac{1}{3}$$

Question 4: Using the decision tree below, decide the classification outcomes for all samples in Problem 1. Left branch is True and right branch is False. Present your result as a two-column table.



Sample Number	Prediction
1	-1
2	+1
3	+1
4	-1
5	-1
6	-1
7	-1
8	-1