Q2) Decision Tree

Y	X_1	X_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
−	T	T	0
−	T	F	1
−	F	T	3
−	F	F	5

① Sample entropy

$$H(Y) = -\sum p \log p$$
$$= -p_- \log p_- - p_+ \log p_+$$
$$= -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7}$$
$$= 0.98523$$

② Information gain

$$IG(X_1) = H(Y) - H(Y|X_1)$$
$$IG(X_2) = H(Y) - H(Y|X_2)$$

$$P(X_1 = T) = \frac{8}{21}$$
$$P(X_1 = F) = \frac{13}{21}$$

$$P(Y_+ | X_{1T}) = \frac{7}{8} \qquad P(Y_- | X_{1T}) = 1/8$$
$$P(Y_+ | X_{1F}) = 5/13 \qquad P(Y_- | X_{1F}) = 8/13$$

$$P(X_2 = T) = \frac{10}{21} \qquad P(X_2 = F) = \frac{11}{21}$$

$$P(Y_+ | X_{2T}) = \frac{1}{10} \qquad P(Y_- | X_{2T}) = \frac{3}{10}$$
$$P(Y_+ | X_{2T}) = \frac{5}{11} \qquad P(Y_- | X_{2F}) = \frac{6}{11}$$

$$H(Y|X_1) = -\sum_{j=1}^{v} P(X_1) \sum_{i=1}^{k} P(Y|X_1) \log P(Y|X_1)$$

$$= \frac{-8}{21}\left(\frac{7}{8}\log\frac{7}{8} + \frac{1}{8}\log\frac{1}{8}\right) - \frac{13}{21}\left(\frac{5}{13}\log\frac{5}{13} + \frac{8}{13}\log\frac{8}{13}\right)$$

$$= 0.802123$$

$$H(Y|X_2) = -\sum_{j=1}^{v} P(X_2) \sum_{i=1}^{k} P(Y|X_2) \log P(Y|X_2)$$

$$= \frac{-10}{21}\left(\frac{7}{10}\log\frac{7}{10} + \frac{3}{10}\log\frac{3}{10}\right) - \frac{11}{21}\left(\frac{5}{11}\log\frac{5}{11} + \frac{6}{11}\log\frac{6}{11}\right)$$

$$= 0.9403448$$

$\therefore H(Y|X_1) = 0.802123$

$H(Y|X_2) = 0.9403448$

$$IG(X_1) = H(Y) - H(Y|X_1)$$

$$= 0.98523 - 0.802123$$

$$= 0.183107$$

$$IG(X_2) = H(Y) - H(Y|X_2)$$

$$= 0.98523 - 0.9403448$$

$$= 0.044885$$

$\therefore IG(X_1) > IG(X_2)$

(3) Decision Tree