Q5. Build a decision tree model based on the following dataset.

ID	X1	X2	Y
1	0.22	0.38	NO
2	0.58	0.32	Yes
3	0.57	0.28	Yes
4	0.41	0.43	Yes
5	0.6	0.29	NO
6	0.12	0.32	Yes
7	0.25	0.32	Yes
8	0.32	0.38	NO

0.12 0.22 | 0.25 0.32 | 0.41 0.57 | 0.58 0.6

3rd quantile

$\frac{0.57 + 0.58}{2} = 0.575 = 0.58$

1st split:

0.28 0.29 | 0.32 0.32 | 0.32 0.38 | 0.38 0.43

3rd quantile 0.38

2nd split on the left node:

0.28 0.29 | 0.32 0.32 0.32 | 0.38 0.38

3rd quantile

$\frac{0.32 + 0.38}{2} = 0.35$

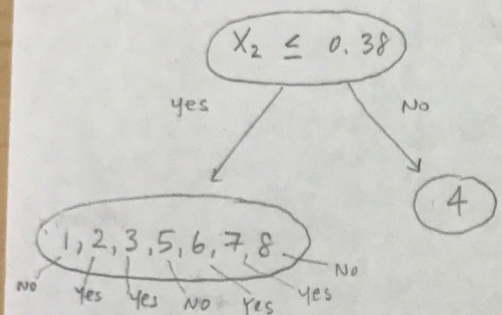
Formula for Information Gain (IG)

$$IG_1 = e_s - \sum_{i=1..n} W_i e_i$$

Formula for Entropy

$$e = \sum_{i=1..k} - P_i \log_2 P_i$$

Arbitrary Rule: splitting on the 4th quantiles.



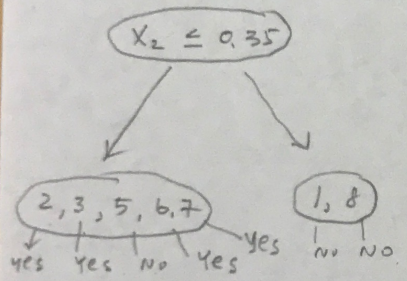
$$e_{root} = -\frac{5}{8}\log_2\left(\frac{5}{8}\right) - \frac{3}{8}\log_2\left(\frac{3}{8}\right) = 0.9544$$

$$e_{x_2 \leq 0.38} = -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right) = 0.9852$$

$$e_{x_2 > 0.38} = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0$$

$$IG_1 = e_{root} - \frac{7}{8}(e_{x_2 \leq 0.38}) - \frac{1}{8}(e_{x_2 > 0.3}) =$$

$$0.9544 - \frac{7}{8}(0.9852) - 0 = 0.0924$$



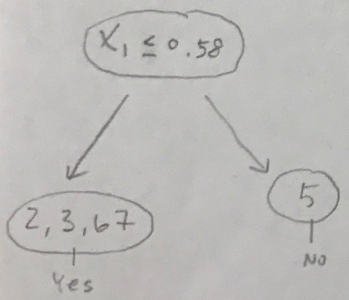
$$e_{root} = -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right) = 0.9852$$

$$e_{X_2 \leq 0.35} = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) = 0.7219$$

$$e_{X_2 > 0.35} = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$$

$$IG_1 = e_{root} - \frac{5}{7}(e \leq 0.35) - \frac{2}{7}(e_{x} > 0.35) =$$

$$0.9852 - 0.52 - 0 = 0.4696$$



$$e_{root} = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) = 0.7219$$

$$e_{x_1 \leq 0.58} = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) = 0$$

$$e_{x_1 > 0.58} = -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0$$

$$IG_1 = e_{root} - \frac{4}{5}(e_{x_1 \leq 0.58}) - \frac{1}{5}(e_{x_1 > 0.58}) = 0.7219$$