

Question 1.

The dataset contain 4 Low and 4 High Risk Level.

$$\Rightarrow I_E([4;4]) = -\frac{4}{8} \log_2 \frac{4}{8} - \frac{4}{8} \log_2 \frac{4}{8} = 1.$$

Split at 650, we Credit Score = 650, we have: 4

Left Subset: (Credit Score \leq 650) = ID: [2, 4, 6, 8]: 4 high

Information gain $I_E([0;4]) = 0$

Right Subset: (Credit Score $>$ 650) = ID: [1, 3, 5, 7]: 4 low

$$I_E([4;0]) = 0$$

$$I_E([0;4], [4;0]) = 0$$

\Rightarrow Information gain of the split:

$$I_E = I_E([4;4]) - I_E([0;4], [4;0]) = 1 - 0 = 1.$$

If hi. Then split at Credit Score = 650 is good because

We can clearly split ~~two~~ classify two type of Risk Level