Q1: Information Gain Calculation

Calculate the information gain. Given the training dataset with 8 records (4 **Low** risk and 4 **High** risk), the entropy of the parent node is:

$$E(\text{parent}) = -\sum_i p_i \log_2 p_i = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

After splitting on `CreditScore` at 650, the dataset is divided into two groups:

- **Group A (CreditScore ≥ 650):** 5 records (4 Low, 1 High)
- **Group B (CreditScore < 650):** 3 records (0 Low, 3 High)

Entropy for Group A:

$$p(\text{Low}) = \frac{4}{5}, \quad p(\text{High}) = \frac{1}{5}$$

$$E(A) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right)$$

Numerically, this gives:

$$E(A) \approx -\left(0.8 \times (-0.3219) + 0.2 \times (-2.3219)\right) \approx 0.722$$

Entropy for Group B: Since all records are High risk:

$$E(B) = 0$$

Weighted Entropy After the Split:

$$E_{\text{split}} = \frac{5}{8}E(A) + \frac{3}{8}E(B) = \frac{5}{8}(0.722) + \frac{3}{8}(0) \approx 0.451$$

Information Gain:

$$\text{Gain} = E(\text{parent}) - E_{\text{split}} = 1 - 0.451 \approx 0.549$$