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:

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

Q2: Variance Reduction for Splitting on Age = 35

We consider the training dataset with CreditScore values:

$$\{720, 650, 750, 600, 780, 630, 710, 640\}$$

with corresponding Ages:

$$\{35, 28, 45, 31, 52, 29, 42, 33\}$$

Parent Node

The mean CreditScore is:

$$\mu = \frac{720 + 650 + 750 + 600 + 780 + 630 + 710 + 640}{8} = 685$$

The variance is computed as:

$$\sigma^2_{\text{parent}} = \frac{1}{8}\sum_{i=1}^{8}(x_i - 685)^2 = \frac{28600}{8} = 3575$$

Splitting on Age = 35

Group A (Age ≤ 35): Records: {1, 2, 4, 6, 8} with CreditScores:

$$\{720, 650, 600, 630, 640\}$$

The mean for Group A is:

$$\mu_A = \frac{720 + 650 + 600 + 630 + 640}{5} = 648$$

The variance for Group A is:

$$\sigma_A^2 = \frac{1}{5}\Big[(720 - 648)^2 + (650 - 648)^2 + (600 - 648)^2 + (630 - 648)^2 + (640 - 648)^2\Big] \approx 1576$$

Group B (Age > 35): Records: {3, 5, 7} with CreditScores:

$$\{750, 780, 710\}$$

The mean for Group B is:

$$\mu_B = \frac{750 + 780 + 710}{3} \approx 746.67$$

The variance for Group B is:

$$\sigma_B^2 = \frac{1}{3}\Big[(750 - 746.67)^2 + (780 - 746.67)^2 + (710 - 746.67)^2\Big] \approx 822.22$$

Weighted Variance After Split

The weighted variance after the split is:

$$\sigma_{\text{split}}^2 = \frac{5}{8}\sigma_A^2 + \frac{3}{8}\sigma_B^2 \approx \frac{5}{8}(1576) + \frac{3}{8}(822.22) \approx 1293.33$$

Variance Reduction

The variance reduction achieved by the split is:

$$\text{Reduction} = \sigma_{\text{parent}}^2 - \sigma_{\text{split}}^2 \approx 3575 - 1293.33 \approx 2281.67$$

Comparison with Information Gain in Classification

While **variance reduction** minimizes the mean squared error for continuous targets, **information gain** in classification trees measures the reduction in impurity (entropy) for categorical outcomes. Variance reduction focuses on reducing numerical dispersion, whereas information gain focuses on achieving purer class splits.

Q3: Predicting T2's Risk Level and Handling Missing Values

Part 1: Probability of T2 Being High Risk

Consider the training dataset with 8 records:

ID	Age	CreditScore	RiskLevel
1	35	720	Low
2	28	650	High
3	45	750	Low
4	31	600	High
5	52	780	Low
6	29	630	High
7	42	710	Low
8	33	640	High

The test record T2 has:

$$\text{Age} = 30, \quad \text{CreditScore} = 645, \quad \text{Education} = \text{missing}.$$

Since Education is missing, we rely on Age and CreditScore. We define similarity as:

$$\text{Absolute difference in Age} \leq 5 \quad \text{and} \quad \text{Absolute difference in CreditScore} \leq 25.$$

Based on this criterion, the similar training records are:

ID	Age	CreditScore	RiskLevel
2	28	650	High
6	29	630	High
8	33	640	High

Let:

$$n = \text{number of similar records} = 3, \quad n_H = \text{number of similar records with High Risk} = 3.$$

Then, the probability that T2 is High Risk is:

$$P(\text{High Risk} \mid \text{similar records}) = \frac{n_H}{n} = \frac{3}{3} = 1.$$

Part 2: Handling Missing Education Values

For future cases where the Education value is missing, several imputation methods can be used:

- **K-Nearest Neighbors (KNN) Imputation:** Estimate the missing Education value by averaging (or taking the mode, if categorical) the Education values of the most similar records, where similarity is determined using Age and CreditScore.
- **Regression Imputation:** Build a regression model that predicts Education based on other features (e.g., Age, CreditScore, and even RiskLevel) and use it to impute missing values.
- **Mean/Median Imputation:** Replace missing values with the mean or median Education value calculated from similar records or the entire dataset if appropriate.
- **Missingness Indicator:** Create an additional binary variable that indicates whether the Education value is missing. This allows the model to capture any information contained in the missingness itself.

In this example, because all similar records (based on Age and CreditScore) are classified as High Risk, T2 is predicted to be High Risk with probability 1, even though the Education value is missing.