Training Dataset (8 records)

ID	Age	CreditScore	Education	RiskLevel
1	35	720	16	Low
2	28	650	14	High
3	45	750	missing	Low
4	31	600	12	High
5	52	780	18	Low
6	29	630	14	High
7	42	710	16	Low
8	33	640	12	High

Test Dataset (2 records)

ID	Age	CreditScore	Education
T1	37	705	16
T2	30	645	missing

Question 1. Calculate the information gain for splitting CreditScore at 650 in a decision tree classification task, then explain why you would or wouldn't choose this as the root node split.

Consider a split on CreditScore = 650

- Calculate entropy for entire dataset: (before split)

- 4 low risk & 4 high risk

$$\Rightarrow H(\text{data}) = -\sum_{i \in \{\text{low, high}\}} p_i \log_2 p_i = -\left(\frac{4}{8}\log_2\frac{4}{8} + \frac{4}{8}\log_2\frac{4}{8}\right) = 1$$

- Entropy after split:

Group A (≤ 650): IDs = {2, 4, 6, 8} → 4 high risk, 0 low risk

$\Rightarrow Ent(A) = 0$

Group B (> 650): IDs = {1, 3, 5, 7} → 4 low risk, 0 high risk

$\Rightarrow Ent(B) = 0$

- Weighted average entropy after split:

$$H_{after} = p_A Ent(A) + p_B Ent(B) = \frac{4}{8}\cdot 0 + \frac{4}{8}\cdot 0 = 0$$

• Information gain = $H_{before} - H_{after} = 1 - 0 = 1$

$H_{before} = H(data)$

⇒ This is a maximum information gain ⇒ Exellent split

I would choose this split because it perfectly separate the 2 Risk level without scratch.

Question 2. For a regression decision tree predicting CreditScore, calculate the variance reduction when splitting on Age=35, and describe how this splitting criterion differs from information gain in classification trees.

• Variance before split (Age = 35)

$$\mu = \frac{\sum \text{CreditScore}}{|data|} = 685$$

$$Var(data) = \frac{1}{8}\sum (x_i - \mu) = 3468.75$$

• Split on Age = 35

Group A (≤ 35) : IDs = {1, 2, 4, 6, 8}

→ CreditScores = [720, 650, 600, 630, 640]

$\mu_A = 648$, $Var(A) = 1704$

Group B (> 35) ...

$\mu_B = 746.67$, $Var(B) = 822.22$

• Weighted Variance after split

$$Var_{after} = P_A\, Var(A) + P_B\, Var(B)$$

$$= 1370.8$$

. Variance Reduction = $Var_{before} - Var_{after}$

$= 2097.95$

* Variance reduction measures how much prediction error is reduced

~ Differ from information gain "which is based on entropy"

Question 3. Using both CreditScore and Age patterns in the training data, determine the probability of T2 being High Risk given its missing Education value, then propose a method to handle similar missing values in future cases.

- Using nearest neighbor to estimate

T2:
. Age : 30
. CreditScore : 645
. Education : missing

Training record
Age ~ 30
CreditScore ~ 645
} IDs = {2, 4, 6, 8}

All high risk

=> Estimate prob of T2 is High Risk = 100%.

Missing Data :

Opt 1: Based on mean/mode on similar users

Opt 2: Use k-NN to find most similar users & base decision on them

Question 4. Implement batch gradient descent to find the optimal weights for predicting CreditScore using Age as input. Starting with initial parameters $\theta_0 = 500$, $\theta_1 = 5$, compute the cost function and one iteration of gradient descent updates using learning rate $\alpha = 0.01$. Interpret the direction of the parameter updates.

$$\hat{y} = \theta_0 + \theta_1 \cdot x_{Age}$$

Calculate error (MSE)

$$J(\theta) = \frac{1}{2m} \sum (y_i - \hat{y}_i)^2 = 810.94$$

Calculate Gradient

$$\frac{dJ(\theta)}{d\theta_0} = \frac{dJ(\theta)}{d\hat{y}} \cdot \frac{d\hat{y}}{d\theta_0} = \frac{1}{m}\sum(\hat{y}_i - y_i) = 0.625$$

($\frac{d\hat{y}}{d\theta_0} = 1$)

$$\frac{dJ(\theta)}{d\theta_1} = \frac{dJ(\theta)}{d\hat{y}} \cdot \frac{d\hat{y}}{d\theta_1} = \frac{1}{m}\sum(\hat{y}_i - y_i)\, x_i = 15.625$$

($\frac{d\hat{y}}{d\theta_1} = x_{age}$; x_i = Age)

Update params:

$$\theta_0 = \theta_0 - \alpha \frac{dJ(\theta)}{d\theta_0} = 500 - 0.01 \times 0.625 = 499.99375$$

$$\theta_1 = \theta_1 - \alpha \frac{dJ(\theta)}{d\theta_1} = 5 - 0.01 \times 15.625 = 4.84375$$

- Direction:
  - Small update in θ_0, larger decrease in θ_1
  → Model will decrease slope, as prediction were slightly too steep.