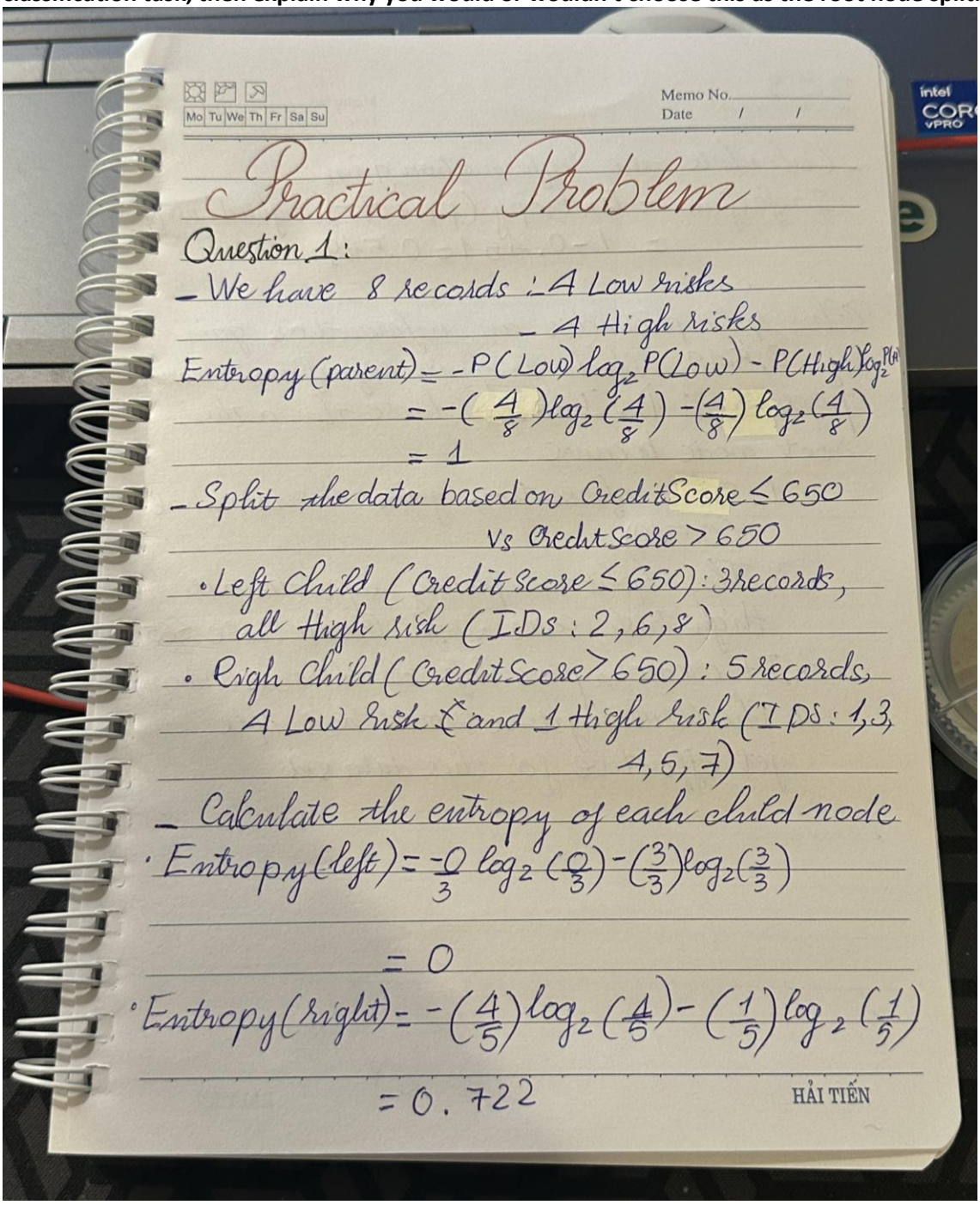
Question 1: Calculate the information gain for splitting CreditScore at 650 in a decision tree classification task, then explain why you would or wouldn't choose this as the root node split.



Practical Problem

Question 1:

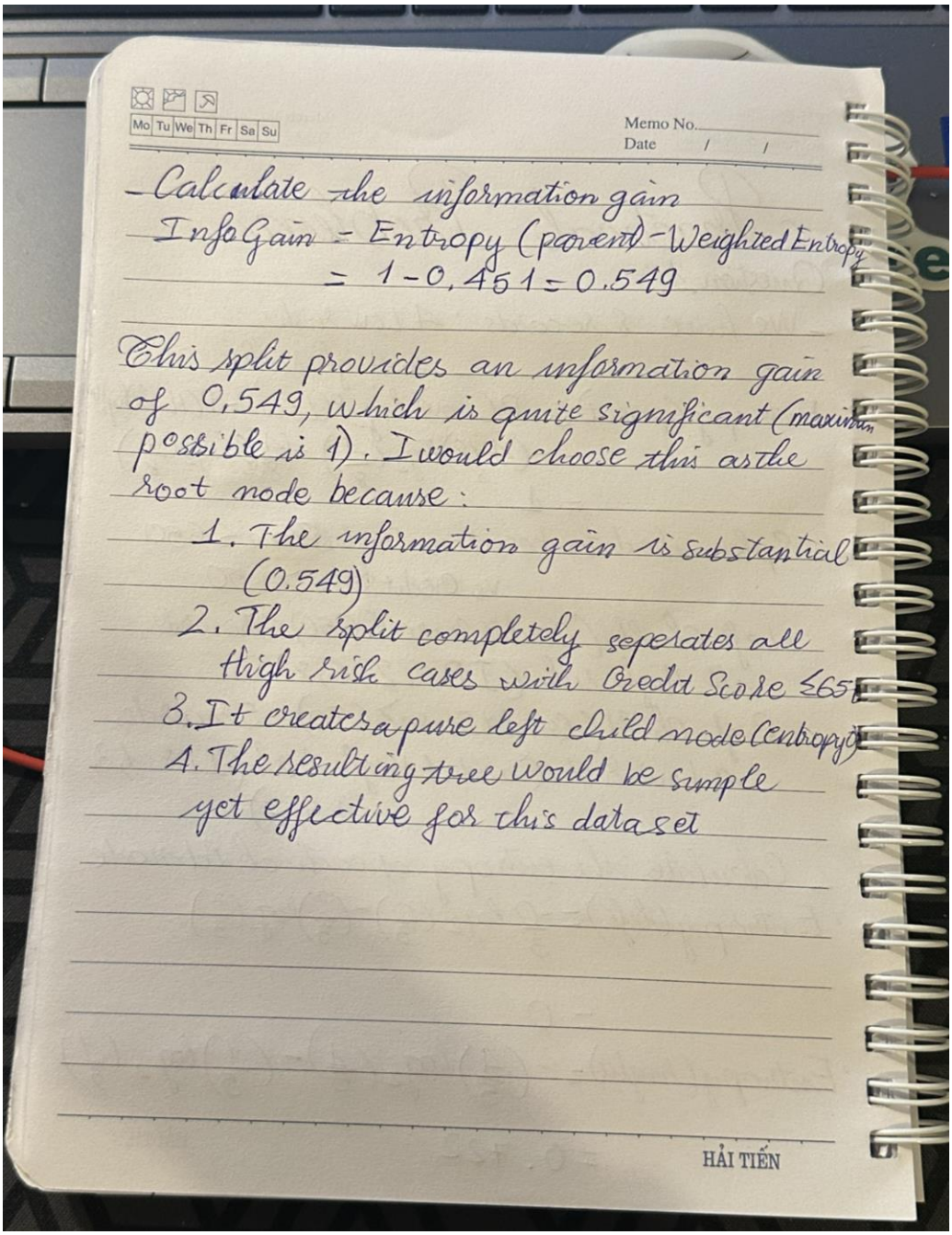
- We have 8 records : 4 Low risks
  - 4 High risks

$$\text{Entropy(parent)} = -P(\text{Low})\log_2 P(\text{Low}) - P(\text{High})\log_2 P(\text{High})$$
$$= -\left(\frac{4}{8}\right)\log_2\left(\frac{4}{8}\right) - \left(\frac{4}{8}\right)\log_2\left(\frac{4}{8}\right)$$
$$= 1$$

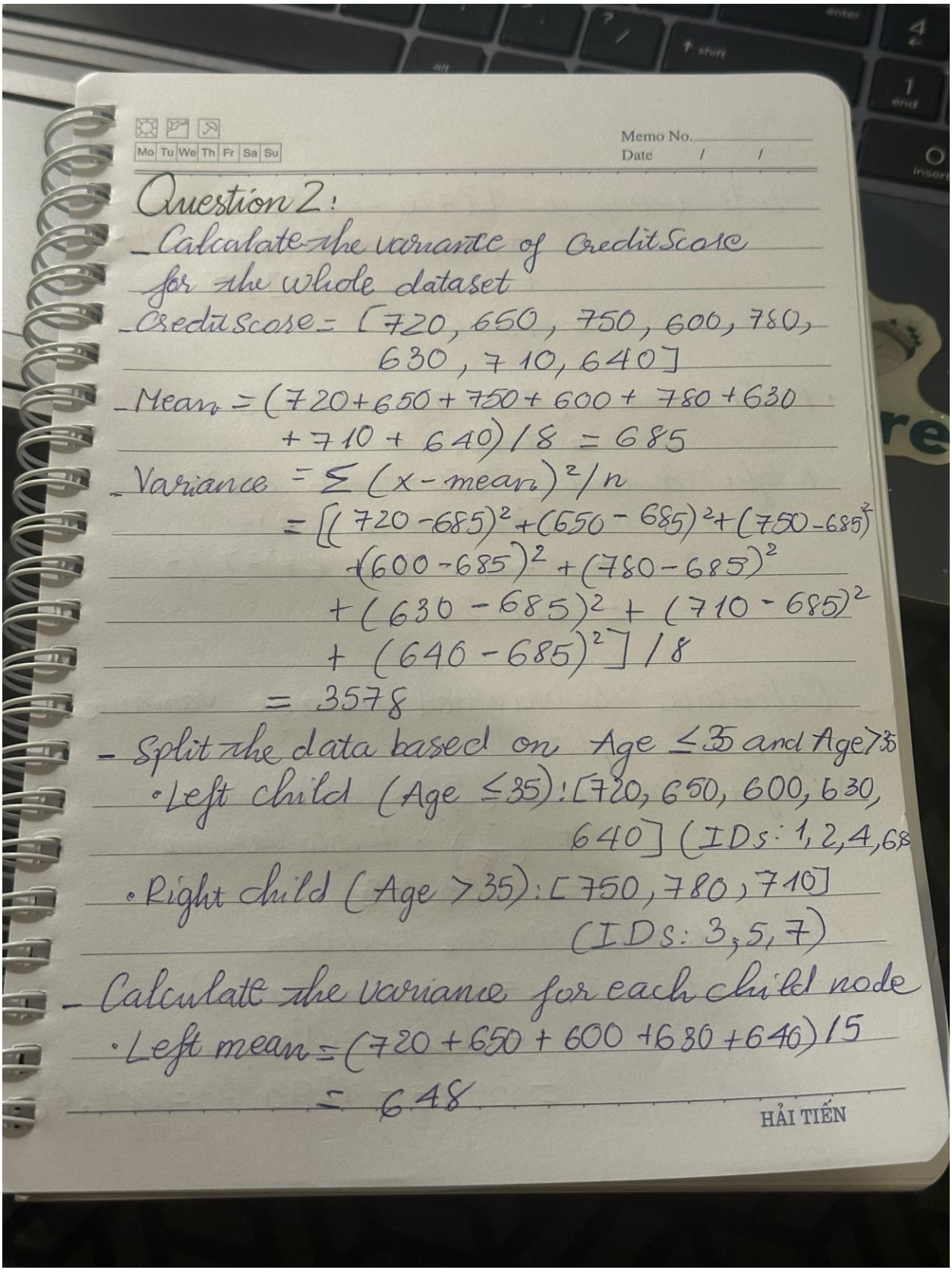
- Split the data based on CreditScore ≤ 650 Vs CreditScore > 650
  - Left Child (CreditScore ≤ 650): 3 records, all High risk (IDs : 2, 6, 8)
  - Righ Child (CreditScore > 650) : 5 records, 4 Low risk and 1 High risk (IDs : 1, 3, 4, 5, 7)

- Calculate the entropy of each child node
  - $\text{Entropy(left)} = -\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right)$
  - $= 0$
  - $\text{Entropy(right)} = -\left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right) - \left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right)$
  - $= 0.722$

HẢI TIẾN


Mo Tu We Th Fr Sa Su
Memo No.
Date / /
- Calculate the information gain
InfoGain = Entropy (parent) - Weighted Entropy
= 1 - 0.451 = 0.549
This split provides an information gain of 0.549, which is quite significant (maximum possible is 1). I would choose this as the root node because:
1. The information gain is substantial (0.549)
2. The split completely seperates all high risk cases with Credit Score ≤65
3. It creates a pure left child node (entropy 0)
4. The resulting tree would be simple yet effective for this dataset
HẢI TIẾN

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.



Question 2:

- Calculate the variance of CreditScore for the whole dataset

CreditScore = [720, 650, 750, 600, 780, 630, 710, 640]

Mean = (720+650+750+600+780+630+710+640)/8 = 685

Variance = $\Sigma (x - \text{mean})^2 / n$

$= [(720-685)^2 + (650-685)^2 + (750-685)^2 + (600-685)^2 + (780-685)^2 + (630-685)^2 + (710-685)^2 + (640-685)^2] / 8$

$= 3578$

- Split the data based on Age ≤ 35 and Age > 35
  - Left child (Age ≤ 35): [720, 650, 600, 630, 640] (IDs: 1, 2, 4, 6, 8)
  - Right child (Age > 35): [750, 780, 710] (IDs: 3, 5, 7)
- Calculate the variance for each child node
  - Left mean = (720+650+600+630+640)/5 = 648

HẢI TIẾN

Mo Tu We Th Fr Sa Su

Memo No.
Date / /

- Left variance = $[(720-648)^2 + (650-648)^2 + (600-648)^2 + (630-648)^2 + (640-648)^2] / 5$
  $= 1576$

- Right mean = $(750 + 780 + 710)/3$
  $= 746.67$
- Right variance = $[(750-746.67)^2 + (780-746.67)^2 + (710-746.67)^2] / 3$
  $= 822.22$

- Calculate the weighted average variance after the split:

Weight Variance = $\left(\frac{5}{8}\right) \times$ Left Variance $+ \left(\frac{3}{8}\right) \times$ Right Variance
$= 1293.33$

- Calculate the variance reduction

Variance Reduction = Variance (parent) - Weighted Variance
$= 3575 - 1293.33 = 2281.67$

HẢI TIẾN

Mo Tu We Th Fr Sa Su

Memo No.

Date / /

- The difference between variance reduction and information gain:

• Variance reduction is used for regression problems where the target is a continuos value (Credit score)

• Information gain is used for classification problems where the target is a categorical value (Risk value)

• Variance reduction measures the homogeneity of numberic value after spliting

• Information gain massure the purity of class ~~lab~~ labels after spliting

The variance reduction of 228 1.67 is significant, showing ~~the~~ Age = 35 split reduces the variance by about 64% of the orginal variance, making ~~this~~ split effective for a regression tree

HẢI TIẾN