Assignment-2_Aparna Bharathi Suresh

Question 3:

The following table lists a sample of data from a census.

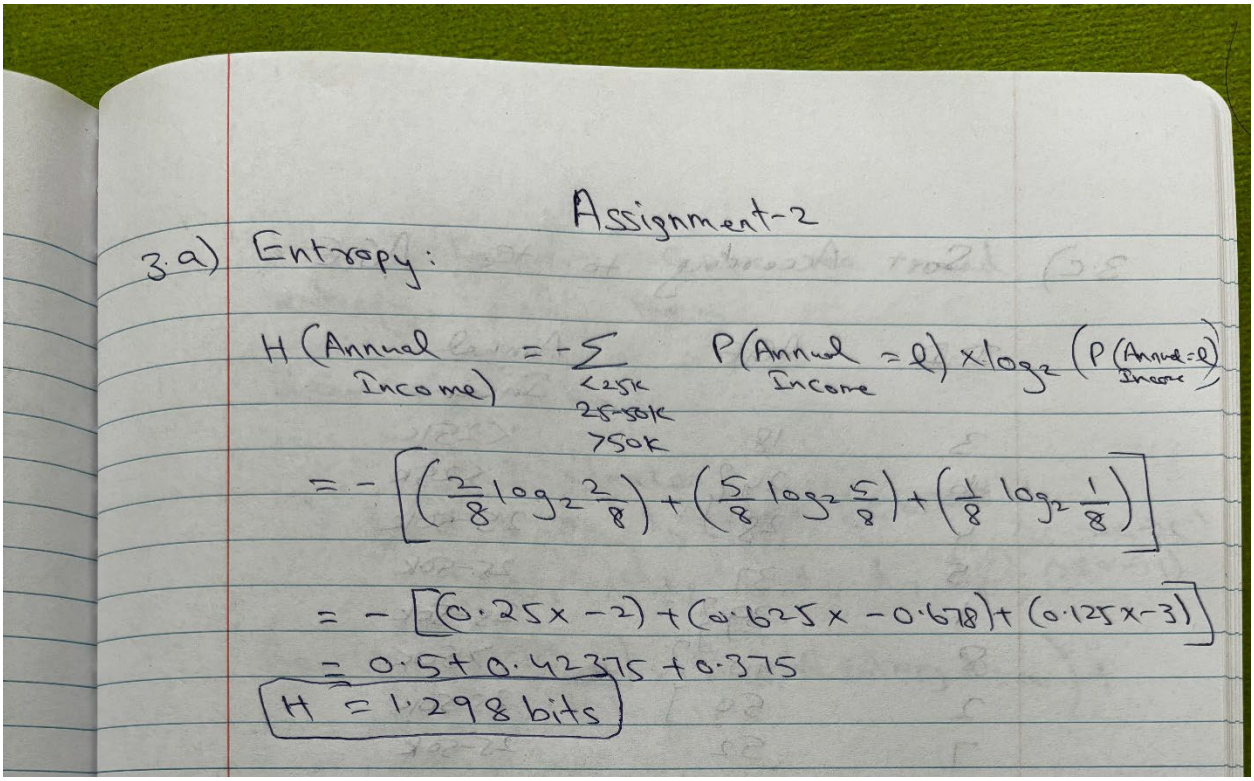
There are four descriptive features and one target feature in this dataset, as follows:

- AGE, a continuous feature listing the age of the individual;
- EDUCATION, a categorical feature listing the highest education award achieved by the individual (high school, bachelors, doctorate);
- MARITAL STATUS (never married, married, divorced);
- OCCUPATION (transport = works in the transportation industry; professional = doctor, lawyer, or similar; agriculture = works in the agricultural industry; armed forces = is a member of the armed forces);
- ANNUAL INCOME, the target feature with 3 levels (25K, 25K–50K, ¡50K).

Part (a)

Calculate the entropy for this dataset.

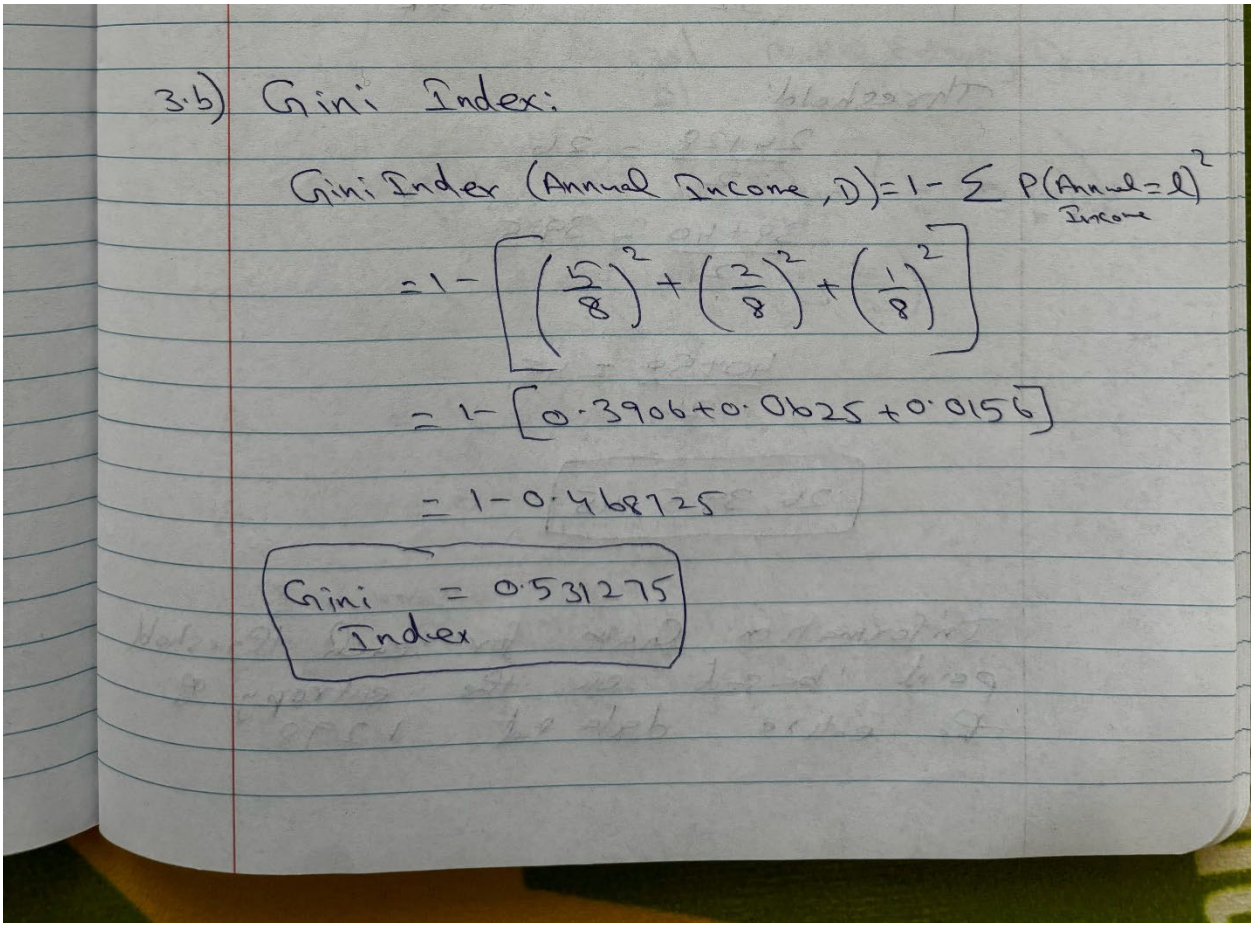
Answer 3.a:



Part (b)

Calculate the Gini index for this dataset.

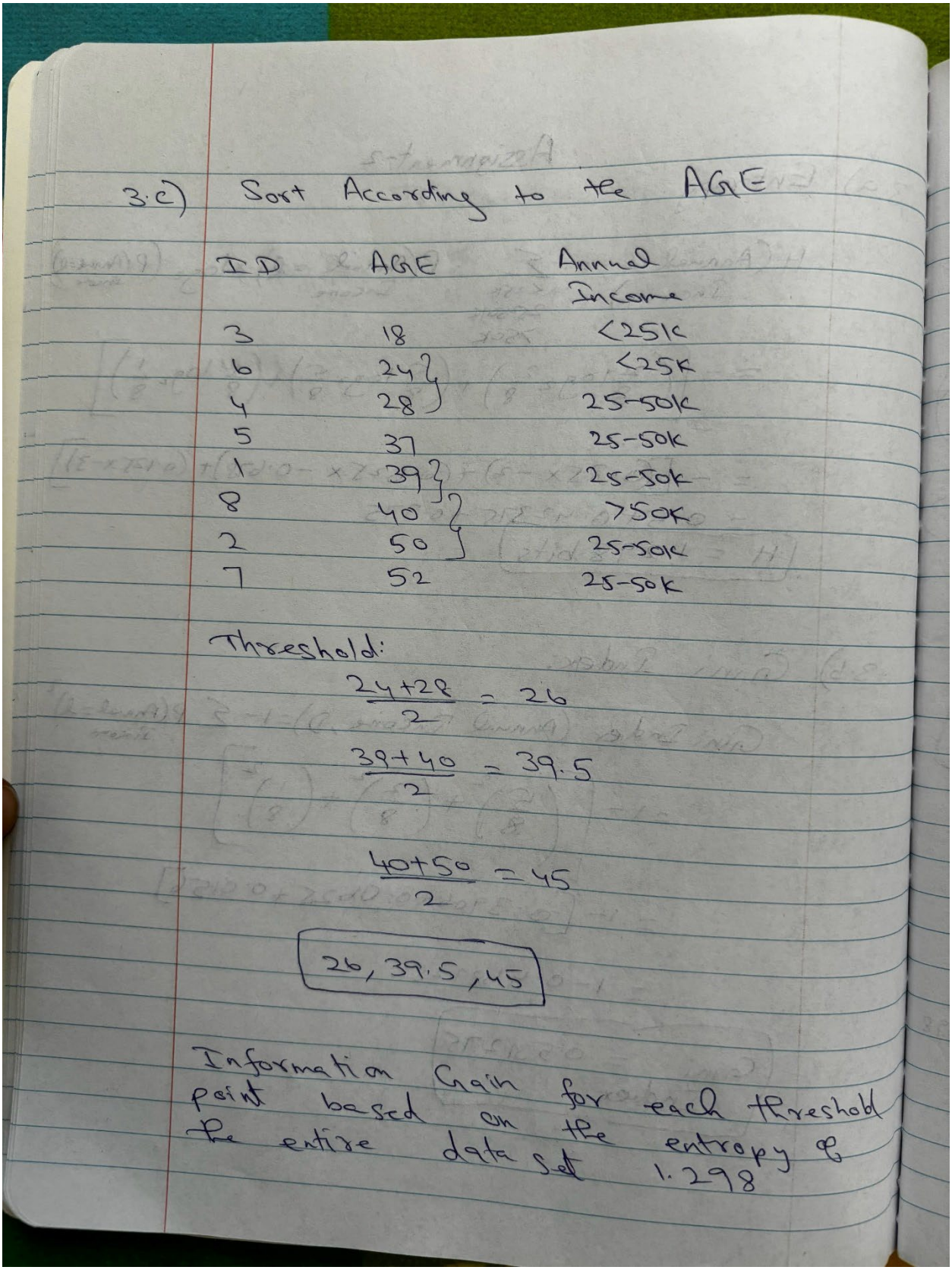
Answer 3.b:



Part (c)

In building a decision tree, the easiest way to handle a continuous feature is to define a threshold around which splits will be made. What would be the optimal threshold to split the continuous AGE feature (use information gain based on entropy as the feature selection measure)?

Answer 3.c:



3.c) Sort According to the AGE

ID	AGE	Annual Income
3	18	<25k
6	24	<25K
4	28	25-50K
5	37	25-50K
1	39	25-50k
8	40	>50K
2	50	25-50K
7	52	25-50K

Threshold:

$$\frac{24+28}{2} = 26$$

$$\frac{39+40}{2} = 39.5$$

$$\frac{40+50}{2} = 45$$

26, 39.5, 45

Information Gain for each threshold point based on the entropy of the entire data set 1.298

For Information Gain we need entropy and Rem.

Rem of >26:

Instances

>26 → (<=26) d_3, d_6 (<25K → 2)

(>26) $d_1, d_2, d_4, d_5, d_7, d_8$ (25-50K → 5), (>50 → 1)

$$Rem(>26) = \left(\frac{|D_{>26}|}{|D|} \times \text{Partial Entropy } D_{>26}\right) + \left(\frac{|D_{<26}|}{|D|} \times \text{Partial Entropy } D_{<26}\right)$$

$$= \frac{6}{8} \times \left(-\left(\frac{5}{6}\log\frac{5}{6}\right) + \left(\frac{1}{6}\log\frac{1}{6}\right)\right) + \frac{2}{8} \times \left(-\frac{2}{2}\log\frac{2}{2}\right)$$

$$= \frac{6}{8} \times 0.6500 + 0$$

$$= 0.4875$$

Information Gain(>26) $= 1.298 - 0.4875$

$= \boxed{0.8095}$

<39.5 → $d_1, d_3, d_4, d_5, d_6 \Rightarrow$ 25-50k → 3, <25 → 2

>39.5 → $d_2, d_7, d_8 \Rightarrow$ 25-50k → 2, >50 → 1

39.5

$$Rem(>39.5) = \left(\frac{|D_{>39.5}|}{|D|} \times \text{Partial Entropy}_{>39.5}\right) + \left(\frac{|D_{<39.5}|}{|D|} \times \text{Partial Entropy}_{<39.5}\right)$$

$$= \left(\frac{3}{8} \times \left(-\left(\frac{2}{3}\log\frac{2}{3}\right) + \left(\frac{1}{3}\log\frac{1}{3}\right)\right)\right) + \frac{5}{8} \times \left[-\left(\frac{3}{5}\log\frac{3}{5}\right) + \left(\frac{2}{5}\log\frac{2}{5}\right)\right]$$

$$= \frac{3}{8} \times 0.91829 + \frac{5}{8} \times 0.970944$$

$$= 0.60684 + 0.34435$$

$$= 0.951198$$

Information Gain $(>39.5) = 1.298 - 0.951198$

$$= \boxed{0.3458}$$

>45 → $d_1, d_3, d_4, d_5, d_6, d_8$ ⇒ 25-50 → 3, <25 → 2, >50 → 1

>45 → d_2, d_7 ⇒ 25-50 → 2

$$Rem(>45) = \left(\frac{|D_{>45}|}{|D|} \times \text{Partial Entropy}_{>45}\right) + \left(\frac{|D_{<45}|}{|D|} \times \text{Partial Entropy}_{<45}\right)$$

$$= \frac{2}{8} \times \left[-\left(\frac{2}{2}\log\frac{2}{2}\right)\right] + \frac{6}{8}\left[-\left(\left(\frac{3}{6}\log\frac{3}{6}\right) + \left(\frac{2}{6}\log\frac{2}{6}\right) + \left(\frac{1}{6}\log\frac{1}{6}\right)\right)\right]$$

$$= 0 + \frac{6}{8}(0.5 + 0.528332 + 0.430826)$$

$$= \frac{6}{8}(1.45914)$$

$$= 1.094355$$

Information Gain (>45) = 1.298 − 1.094355

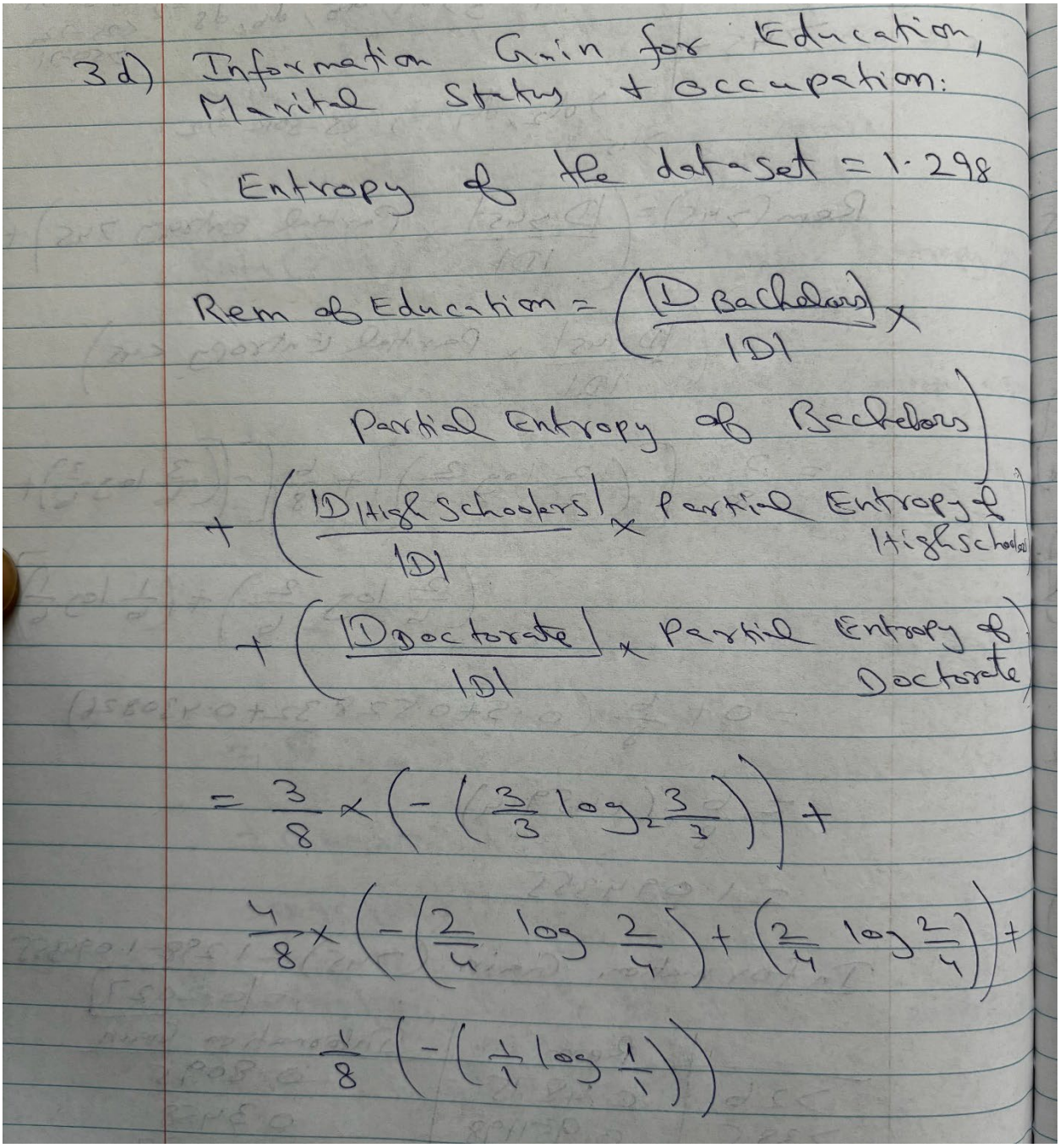
= **0.2027**

	Rem	Information Gain
>26	0.4875	0.8095
>38.5	0.951198	0.3458
>45	1.09435	0.2027

Part (d)

Calculate information gain (based on entropy) for the EDUCATION, MARITAL STATUS, and OCCUPATION features.

Answer 3.d:



$= 0 + \frac{1}{2}(1) + 0$

$= 0.5$

Information Gain of Education = 1.298 - 0.5

$= 0.797$

Rem of Marital Status =

$$\left(\frac{|D_{\text{Never Married}}|}{|D|} \times \text{Partial Entropy of Never Married}\right)$$

$$+\left(\frac{|D_{\text{Married}}|}{|D|} \times \text{Partial Entropy of Married}\right)$$

$$+\left(\frac{|D_{\text{Divorced}}|}{|D|} \times \text{Partial Entropy of Divorced}\right)$$

$$= \frac{3}{8} \times \left[-\left(\left(\frac{2}{3}\log\frac{2}{3}\right)+\left(\frac{1}{3}\log\frac{1}{3}\right)\right)\right]$$

$$+\frac{4}{8} \times \left[-\left(\left(\frac{3}{4}\log\frac{3}{4}\right)+\left(\frac{1}{4}\log\frac{1}{4}\right)\right)\right]$$

$$+\frac{1}{8} \times \left[-\left(\frac{1}{1}\log\frac{1}{1}\right)\right]$$

$= 0.35561 + 0.4056$

$= 0.761245$

Information Gain of Marital Status $= 1.298 - 0.761245$

$= 0.5375$

Rem of Occupation $= \left(\frac{|D_{transport}|}{|D|} \times \text{Partial Entropy of transport}\right)$

$+ \left(\frac{|D_{professional}|}{|D|} \times \text{Partial entropy of professional}\right)$

$+ \left(\frac{|D_{Agriculture}|}{|D|} \times \text{Partial entropy of Agriculture}\right)$

$+ \left(\frac{|D_{Armedforces}|}{|D|} \times \text{Partial Entropy of Armed forces}\right)$

$$= \frac{2}{8} \times \left[-\left(\frac{2}{2} \log \frac{2}{2} \right) \right] + \frac{3}{8} \left[-\left(\frac{2}{3} \log \frac{2}{3} \right) + \left(\frac{1}{3} \log \frac{1}{3} \right) \right] + \frac{2}{8} \left[-\left(\frac{1}{2} \log \frac{1}{2} \right) + \left(\frac{1}{2} \log \frac{1}{2} \right) \right] + \frac{1}{8} \left[-\frac{1}{1} \log \frac{1}{1} \right]$$

$$= 0.34435 + \frac{1}{4}$$

$$= 0.59435$$

Information Gain of Occupation = 1.298 - 0.5943

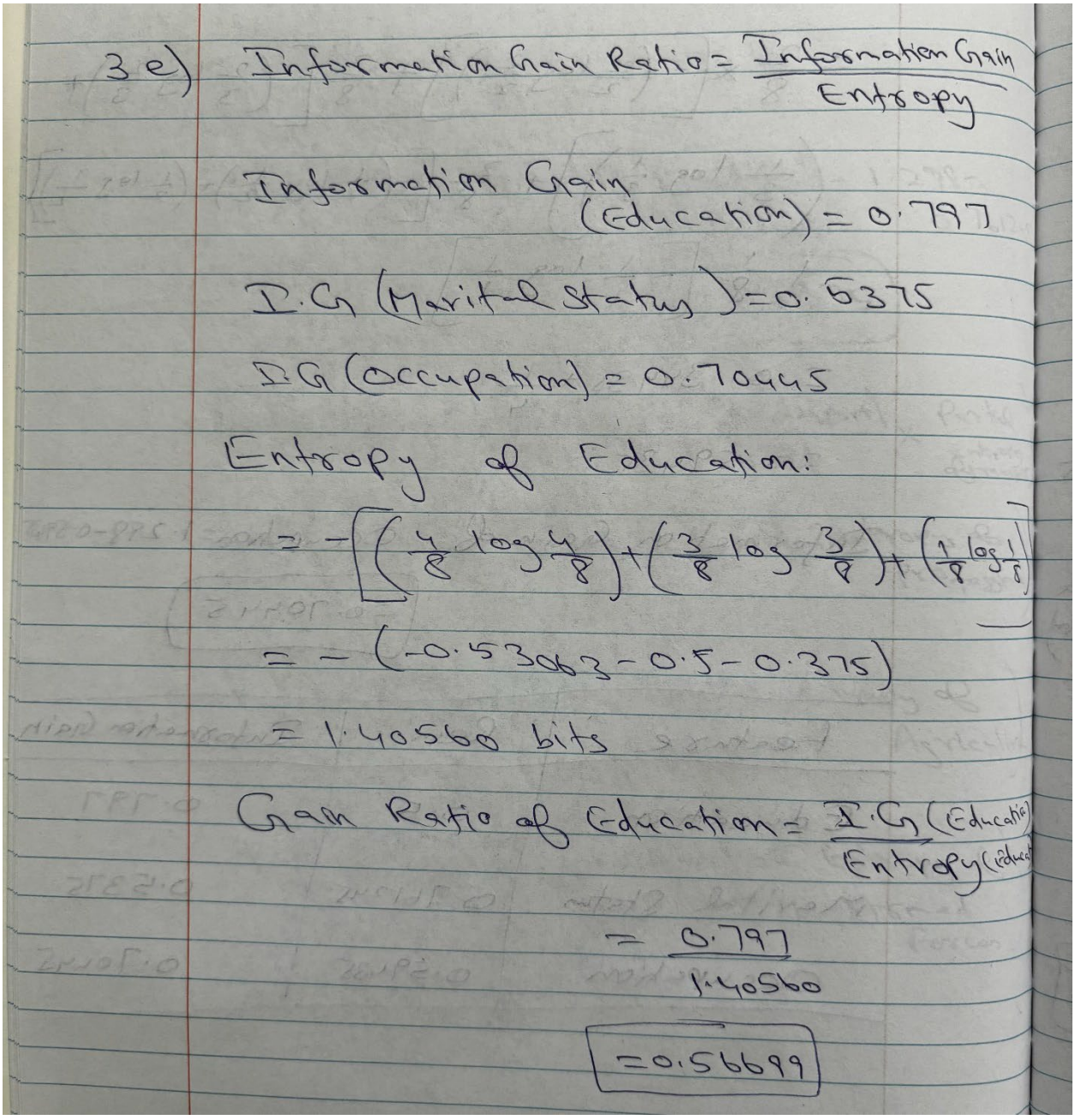
$$= 0.70445$$

Feature	Rem	Information Gain
Education	0.5	0.797
Marital Status	0.761245	0.5375
Occupation	0.59435	0.70445

Part (e)

Calculate the information gain ratio (based on entropy) for EDUCATION, MARITAL STATUS, and OCCUPATION features.

Answer 3.e:



Entropy of Marital Status =

$$-\left[\left(\frac{3}{8}\log\frac{3}{8}\right)+\left(\frac{4}{8}\log\frac{4}{8}\right)+\left(\frac{1}{8}\log\frac{1}{8}\right)\right]$$

$= 1.40560$ bits

Gain Ratio of Marital Status = $\frac{\text{I.G of Marital}}{\text{Entropy of Marit.}}$

$$= \frac{0.5375}{1.40560}$$

$$= 0.38238$$

Entropy of occupation:

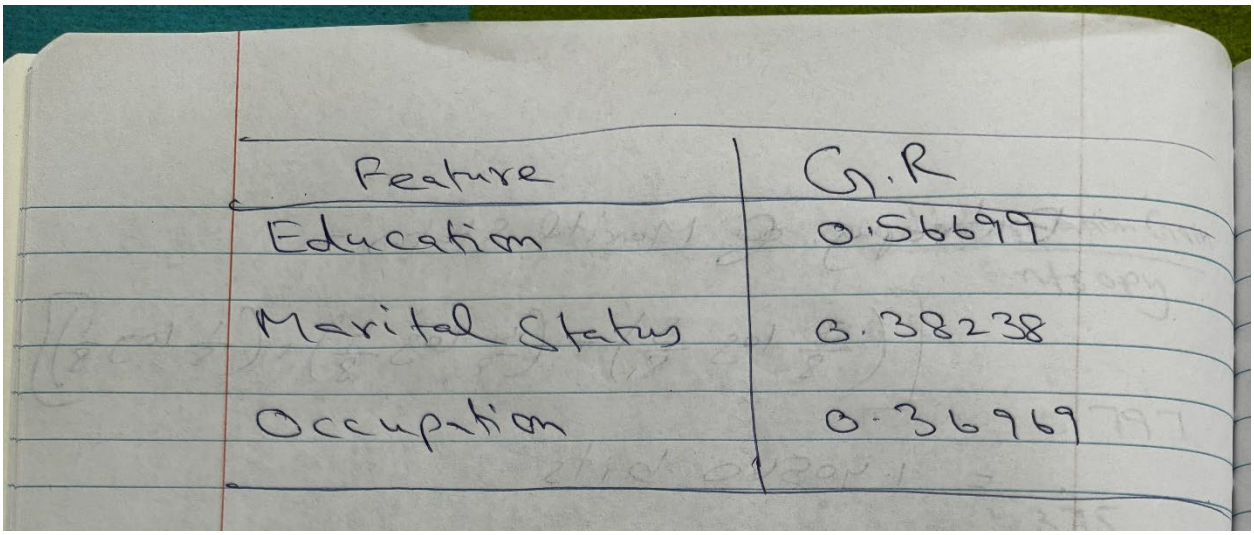
$$= -\left[\left(\frac{2}{8}\log\frac{2}{8}\right)+\left(\frac{3}{8}\log\frac{3}{8}\right)+\left(\frac{2}{8}\log\frac{2}{8}\right)+\left(\frac{1}{8}\log\frac{1}{8}\right)\right]$$

$$= -(-0.5-0.53063-0.5-0.375)$$

$= 1.90563$

G.R of Occupation = $\frac{0.7045}{1.90563}$

$= 0.36969$

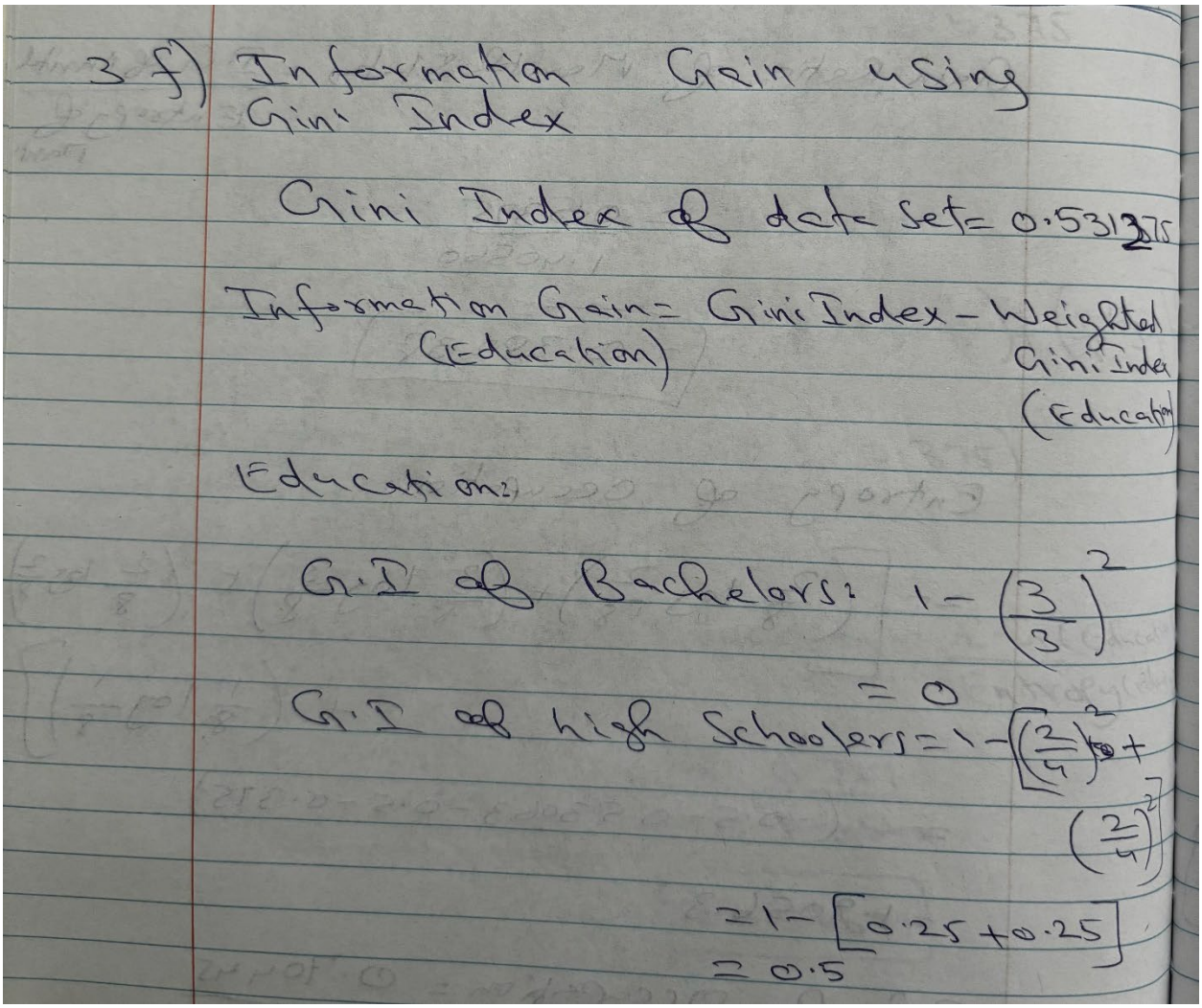


Feature	G.R
Education	0.56699
Marital Status	0.38238
Occupation	0.36969

Part (f)

Calculate information gain using the Gini index for the EDUCATION, MARITAL STATUS, and OCCUPATION features.

Answer 3.f:



3 f) Information Gain using Gini Index

Gini Index of data set = 0.531375

Information Gain = Gini Index − Weighted Gini Index (Education)

Education:

G.I of Bachelors: $1 - \left(\frac{3}{3}\right)^2$

$= 0$

G.I of high Schoolers $= 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$

$= 1 - [0.25 + 0.25]$

$= 0.5$

G.I of Doctorate $= 1 - \left(\frac{1}{1}\right)^2$

$= 0$

Weighted G.I $= \left(\frac{3}{8} \times 0\right) + \left(\frac{4}{8} \times 0.5\right) + \left(\frac{1}{8} \times 0\right)$

$= 0.25$

Information Gain of Education $= 0.53125 - 0.25$

$= 0.28125$

Marital Status:

G.I of Never Married $= 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]$

$= 0.444$

G.I of Married $= 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$

$= 0.375$

G.I of Divorced $= 1 - \left(\frac{1}{1}\right)^2$

$= 0$

Weighted G.I $= \left(\frac{3}{8} \times 0.444\right) + \left(\frac{4}{8} \times 0.375\right) + \left(\frac{1}{8} \times 0\right)$

$= 0.354$

Information Gain of Marital Status $= 0.531275 - 0.354$

$= 0.177275$

Occupation:

G.I of transport $= 1 - \left(\frac{2}{2}\right)^2$

$= 0$

G.I of Agriculture $= 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$

$= 1 - 0.5$

$= 0.5$

G.I of Armed force $= 1 - \left(\frac{1}{1}\right)^2$

$= 0$

G.I of Professional $= 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right]$

$= 0.444$

Weighted G.I $= \left(\frac{2}{8} \times 0\right) + \left(\frac{3}{8} \times 0.444\right) + \left(\frac{2}{8} \times 0.5\right) + \left(\frac{1}{8} \times 0\right)$

$= 0.1665 + 0.125$

$= 0.2915$

Information Gain of Occupation $= 0.531275 - 0.2915$

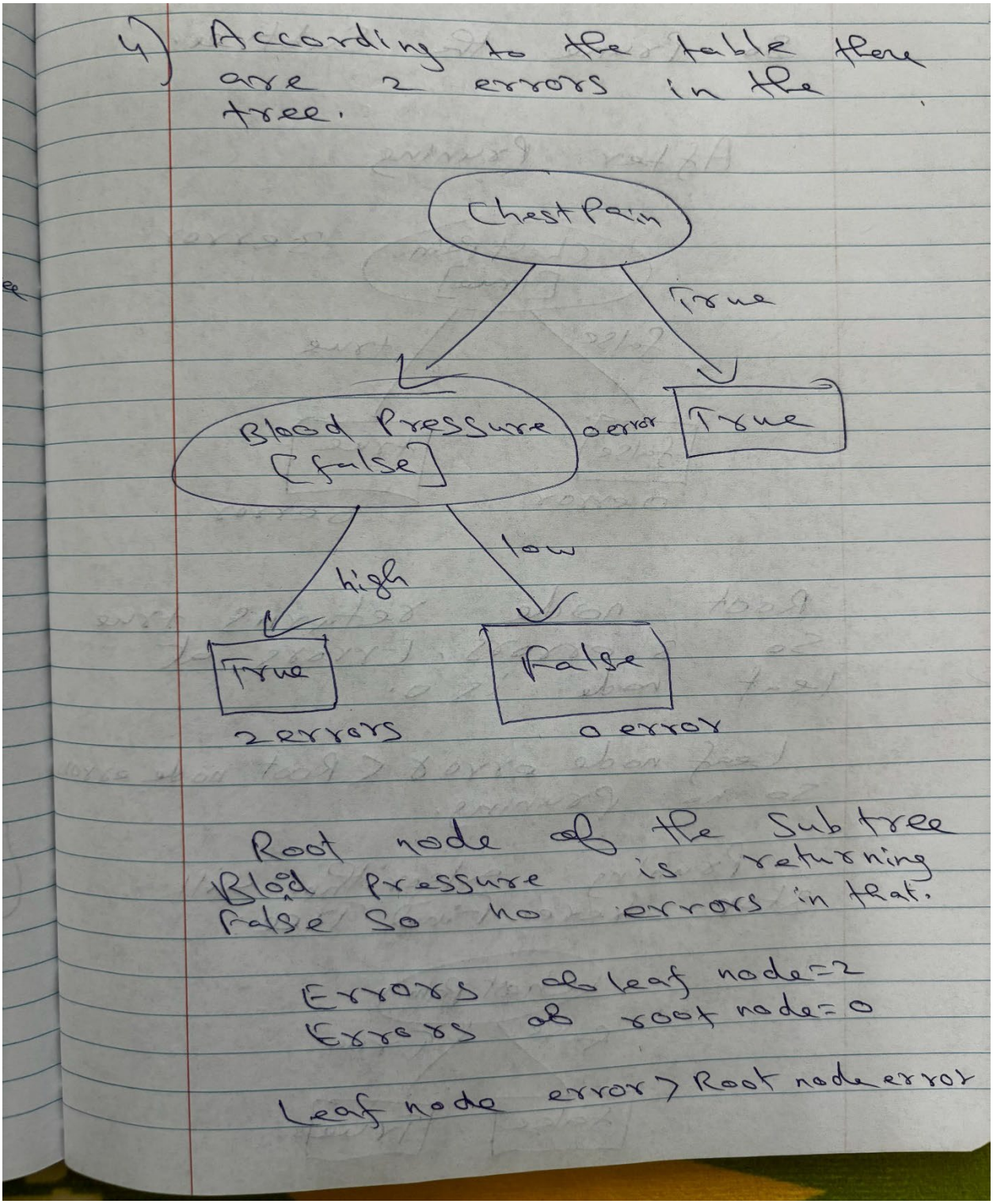
$= 0.239795$

Feature	Rem	Info Gain
Education	0.25	0.28125
Marital Status	0.354	0.177275
Occupation	0.2915	0.239795

Question 4:

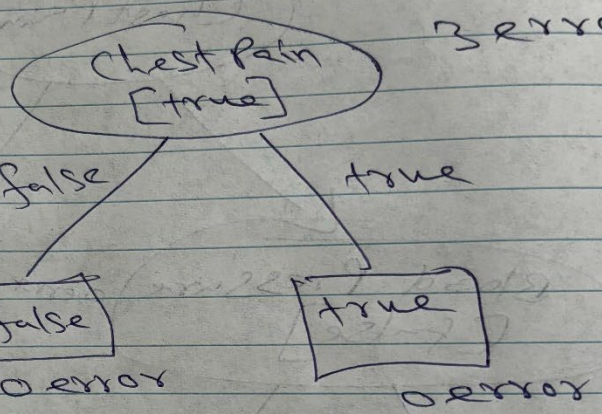
The following diagram shows a decision tree for the task of predicting heart disease.3 The descriptive features in this domain describe whether the patient suffers from chest pain (CHEST PAIN) and the blood pressure of the patient (BLOOD PRESSURE). The binary target feature is HEART DISEASE. The table beside the diagram lists a pruning set from this domain.

Answer 4:


4) According to the table there are 2 errors in the tree.
Chest Pain
True
Blood Pressure
[false]
0 error
True
high
low
True
2 errors
False
0 error
Root node of the Sub tree Blod Pressure is returning false So no errors in that.
Errors of leaf node=2
Errors of root node=0
Leaf node error > Root node error

So Prune the Subtree.

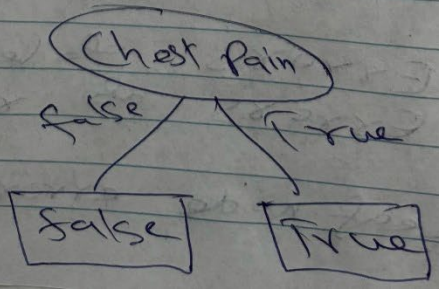
After Pruning



Root node returns true So 3 errors. Errors at leaf node is 0.

Leaf node error < Root node error So no Pruning.

Pruning algorithm will stop here. Final Tree:



Question 5:

The following table4 lists a dataset containing the details of five participants in a heart disease study, and a target feature RISK, which describes their risk of heart disease.

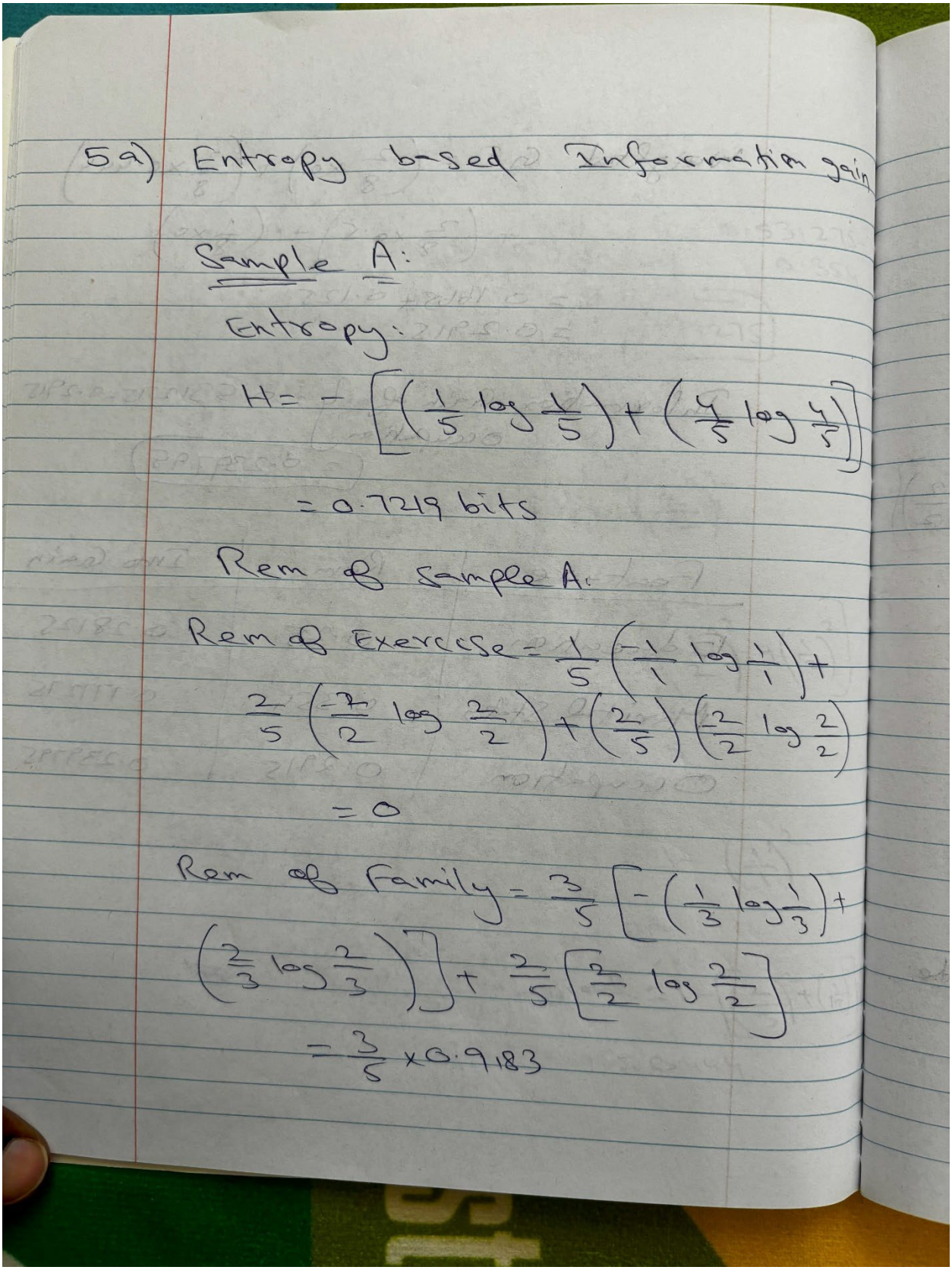
Each patient is described in terms of four binary descriptive features

- EXERCISE, how regularly do they exercise
- SMOKER, do they smoke
- OBESE, are they overweight
- FAMILY, did any of their parents or siblings suffer from heart disease

Part a:

As part of the study, researchers have decided to create a predictive model to screen participants based on their risk of heart disease. You have been asked to implement this screening model using a random forest. The three tables below list three bootstrap samples that have been generated from the above dataset. Using these bootstrap samples, create the decision trees that will be in the random forest model (use entropy-based information gain as the feature selection criterion).

Answer 5.a:



5a) Entropy based Information gain

Sample A:

Entropy:

$$H = -\left[\left(\frac{1}{5}\log\frac{1}{5}\right) + \left(\frac{4}{5}\log\frac{4}{5}\right)\right]$$

$$= 0.7219 \text{ bits}$$

Rem of sample A:

$$\text{Rem of Exercise} = \frac{1}{5}\left(\frac{-1}{1}\log\frac{1}{1}\right) + \frac{2}{5}\left(\frac{-2}{2}\log\frac{2}{2}\right) + \left(\frac{2}{5}\right)\left(\frac{-2}{2}\log\frac{2}{2}\right)$$

$$= 0$$

$$\text{Rem of Family} = \frac{3}{5}\left[-\left(\frac{1}{3}\log\frac{1}{3}\right) + \left(\frac{2}{3}\log\frac{2}{3}\right)\right] + \frac{2}{5}\left[\frac{2}{2}\log\frac{2}{2}\right]$$

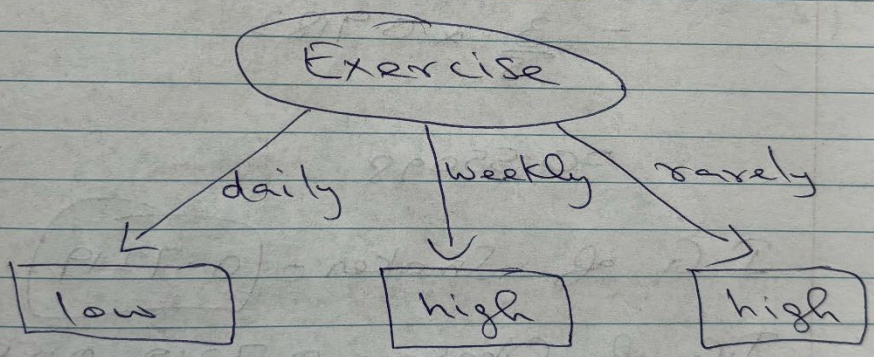
$$= \frac{3}{5} \times 0.9183$$

$= 0.5510$

IG of Exercise $= 0.7219 - 0$
$= \boxed{0.7219}$

I.G of Family $= 0.7219 - 0.5510$
$= \boxed{0.17092}$

Information gain of Exercise is high. So Exercise should be root node.



Sample B:

Entropy:

$$H = -\left[\left(\frac{1}{5}\log\frac{1}{5}\right) + \left(\frac{4}{5}\log\left(\frac{4}{5}\right)\right)\right]$$

$= 0.7219$ bits

Rem of Smoker $= \frac{1}{5}\left(-\frac{1}{1}\log\frac{1}{1}\right) +$

$\frac{4}{5}\left(-\frac{4}{4}\log\frac{4}{4}\right)$

$= 0$

Rem of Obese $= \frac{3}{5}\left[-\left(\frac{1}{3}\log\frac{1}{3}\right) + \left(\frac{2}{3}\log\frac{2}{3}\right)\right] + \frac{2}{5}\left(-\frac{2}{2}\log\frac{2}{2}\right)$

$= \frac{3}{5} \times 0.9183$

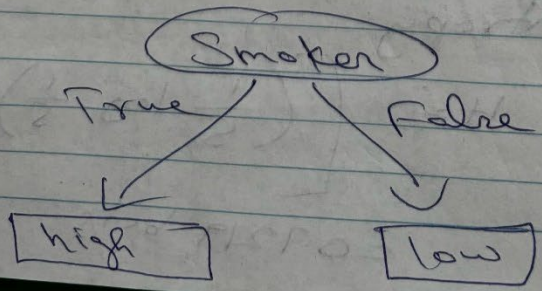
$= 0.55098$

I.G of Smoker = 0.7219

I.G of Obese = $0.7219 - 0.55098$

$= 0.17092$

I.G of Smoker is high.
So Smoker is the root node.



Sample c:

Entropy:

$$H = -\left(\left(\frac{2}{5}\log\frac{2}{5}\right)+\left(\frac{3}{5}\log\frac{3}{5}\right)\right)$$

$$= 0.528768 + 0.442176$$

$$= 0.970944 \text{ bits}$$

$$\text{Rem Obese} = \frac{3}{5} \times \left[-\left(\frac{2}{3}\log\frac{2}{3}\right)+\left(\frac{1}{3}\log\frac{1}{3}\right)\right]$$

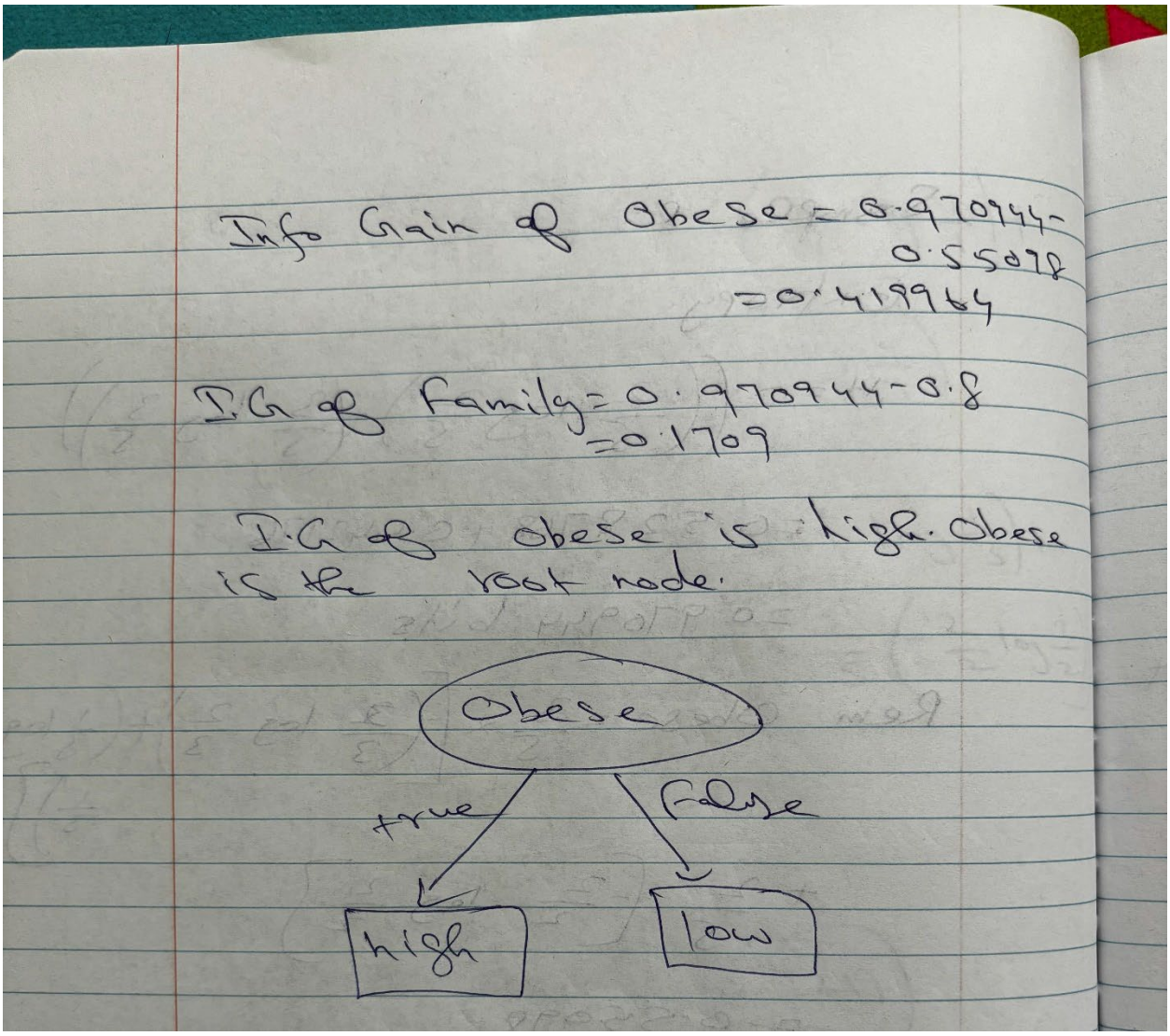
$$+ \frac{2}{5}\left[-\frac{2}{2}\log\frac{2}{2}\right]$$

$$= 0.55098$$

$$\text{Rem Family} = \frac{4}{5}\left[-\left(\frac{2}{4}\log\frac{2}{4}\right)+\left(\frac{2}{4}\log\frac{2}{4}\right)\right]$$

$$+ \frac{1}{5}\left[-\frac{1}{1}\log\frac{1}{1}\right]$$

$$= 0.8$$

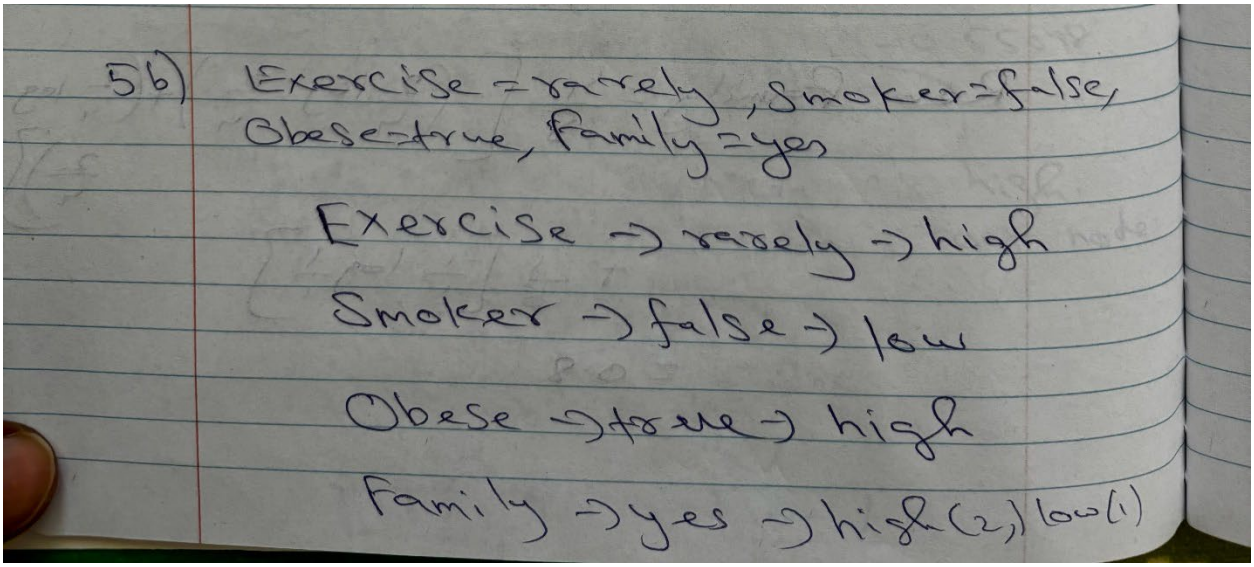


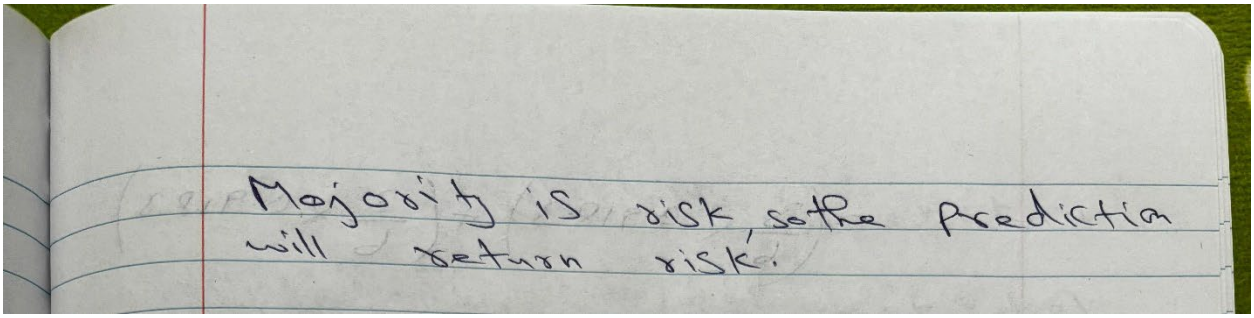
Part b:

Assuming the random forest model you have created uses majority voting, what prediction will it return for the following query:

EXERCISE=rarely, SMOKER=false, OBESE=true, FAMILY=yes

Answer 5.b:





Tree 1→Exercise=Rarely→Risk: High

Tree 2→ SMOKER=false→ RISK=low

Tree 3→ OBESE=true→RISK=high

So, the majority vote is for RISK=high, so the prediction will return high risk.

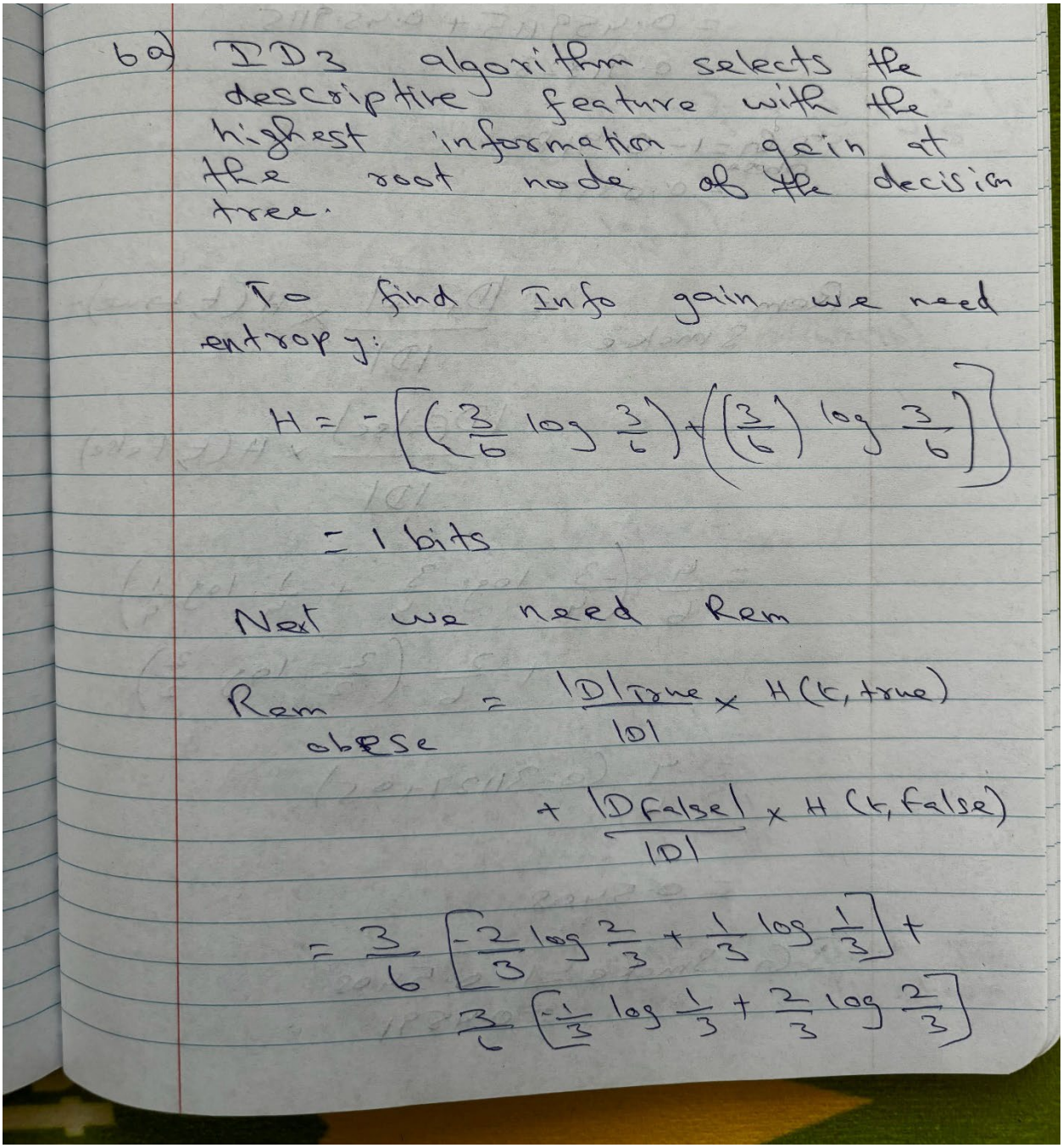
Question 6:

The following table lists a dataset containing the details of six patients. Each patient is described in terms of three binary descriptive features (OBESE, SMOKER, and DRINKS ALCOHOL) and a target feature (CANCER RISK).

Part a)

Which of the descriptive features will the ID3 decision tree induction algorithm choose as the feature for the root node of the decision tree?

Answer 6.a:



6a) ID3 algorithm selects the descriptive feature with the highest information gain at the root node of the decision tree.

To find Info gain we need entropy:

$$H = -\left[\left(\frac{3}{6}\log\frac{3}{6}\right)+\left(\left(\frac{3}{6}\right)\log\frac{3}{6}\right)\right]$$

$$= 1 \text{ bits}$$

Next we need Rem

$$\text{Rem}_{\text{obese}} = \frac{|D|_{\text{True}}}{|D|} \times H(t, \text{true}) + \frac{|D_{\text{False}}|}{|D|} \times H(t, \text{false})$$

$$= \frac{3}{6}\left[-\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right] + \frac{3}{6}\left[-\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right]$$

$$= \left(\frac{3}{6} \times 0.9183\right) + \left(\frac{3}{6} \times 0.9183\right)$$

$$= 0.459115 + 0.459115$$
$$= 0.91823$$

$$I.G_{obs} = 1 - 0.91823$$
$$= 0.0817$$

$$Rem_{Smoke} = \frac{|D_{true}|}{|D|} \times H(t, true) + \frac{|D_{false}|}{|D|} \times H(t, false)$$

$$= \frac{4}{6}\left(\frac{-3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right) + \frac{2}{6}\left(\frac{2}{2}\log\frac{2}{2}\right)$$

$$= \frac{4}{6}(0.31127 + 0.5)$$

$$= 0.5408$$

$$I.G\ Smoke = 1 - 0.5408$$
$$= 0.4591$$

Rem Alcohol = $\frac{|D_{True}|}{|D|} \times H(t, true) + \frac{|D_{false}|}{|D|} \times H(t, False)$

$= \frac{5}{6} \times \left[-\frac{2}{5} \log \frac{2}{5} + \frac{3}{5} \log \frac{3}{5} \right] +$

$\frac{1}{6} \times \left(\frac{-1}{1} \log \frac{1}{1} \right)$

$= \frac{5}{6} (0.528768 + 0.442176)$

$= 0.80912$

I.G Alcohol $= 1 - 0.80912$
$= 0.19088$

Feature	I.G
Obese	0.0817
Smoke	0.4591
Alcohol	0.19088

Smoker has the highest IG so ID3 algorithm will choose

Smoker as the root node.

Part b)

In designing a dataset, it is generally a bad idea if all the descriptive features are indicators of the target feature taking a particular value. For example, a potential criticism of the design of the dataset in this question is that all the descriptive features are indicators of the CANCER RISK target feature taking the same level, high. Can you think of any descriptive features that could be added to this dataset that are indicators of the low target level?

Answer 6.b:

1. EXERCISES REGULARLY (true/false)

People who exercise regularly generally have a lower risk of cancer. This feature could help identify patients with **low cancer risk**.

2. HEALTHY DIET (true/false)

A person with a healthy diet (e.g., rich in fruits and vegetables, low in processed foods) may have a lower cancer risk.

3. NON-EXPOSURE TO POLLUTANTS (true/false)

A person who lives or works in an environment with less exposure to pollutants (e.g., non-industrial areas, good air quality) may have a lower cancer risk.

4. NON-FAMILY HISTORY OF CANCER (true/false)

A person without a family history of cancer could be at a lower genetic risk for developing cancer.

Question 7:

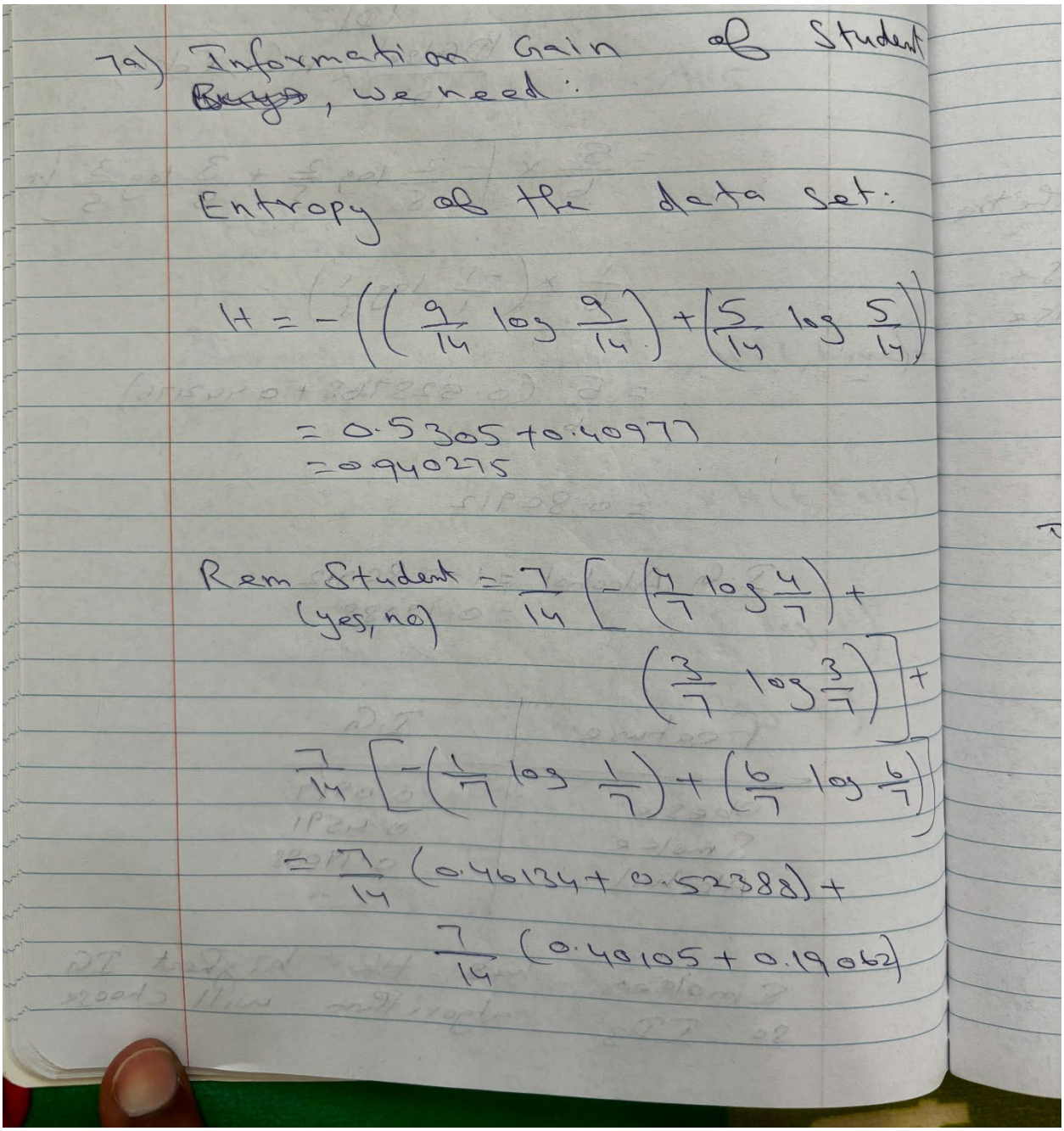
The following table lists a dataset collected in an electronics shop showing details of customers and whether they responded to a special offer to buy a new laptop.

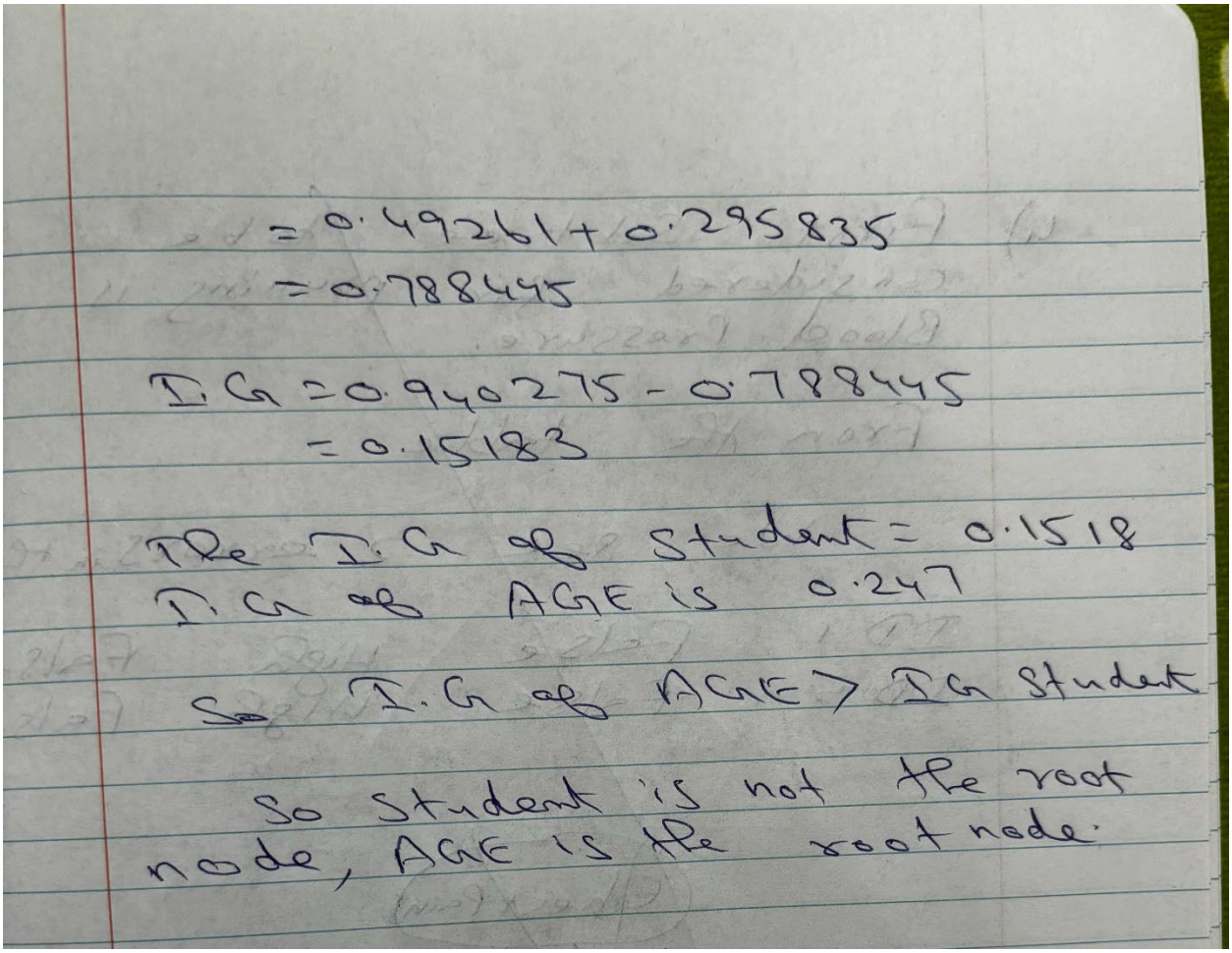
This dataset has been used to build a decision tree to predict which customers will respond to future special offers. The decision tree, created using the ID3 algorithm, is the following:

Part a:

The information gain (calculated using entropy) of the feature AGE at the root node of the tree is 0:247. A colleague has suggested that the STUDENT feature would be better at the root node of the tree. Show that this is not the case.

Answer 7.a:





Part b:

Yet another colleague has suggested that the ID feature would be a very effective at the root node of the tree. Would you agree with this suggestion?

Answer 7.b:

Each instance has a unique value for the ID feature. So, ID would not be a good feature for the root node of the tree because it ID doesn't provide any information, and the resulting decision tree would be overfitted to the training data.