Teacher Initial: NA

Daffodil International University

Fall 2021

Department of Computer Science and Engineering

Midterm Open Book Examination Answer Script

Full Marks: 25 Allowed, Time: 2:30 hrs.

Date: Thursday 18, November 2021

General Information (must be filled by the student)

COURSE CODE: CSE 450 **SECTION:** PC-B **PROGRAM:** DAY EVEN

STUDENT ID: 181-15-1905 **TIME STARTED:** **TIME ENDED:**

[Student must either TYPE or HAND WRITE the answers in this template; In case needed just write your detail on the paper using hand]

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905

Date .....

Answers to the question no. 1

a) Corona create a huge change in our daily life, economy, our personal life. In this new normal we need to reconsider a lot of activites we did previous.

By data analysis we can take a lot of decision. Analysing our future data we can predict future step and upcoming effect of predicted activites.

for detecting post corona challenges, data analysis method are huge impactful. By collecting recent data our data analyist will take care of that data and can give us impactful

Student ID:181-15-1905

Date .....

information from that data. Now there some data analysis method that help to gain ~~predict~~ our information.

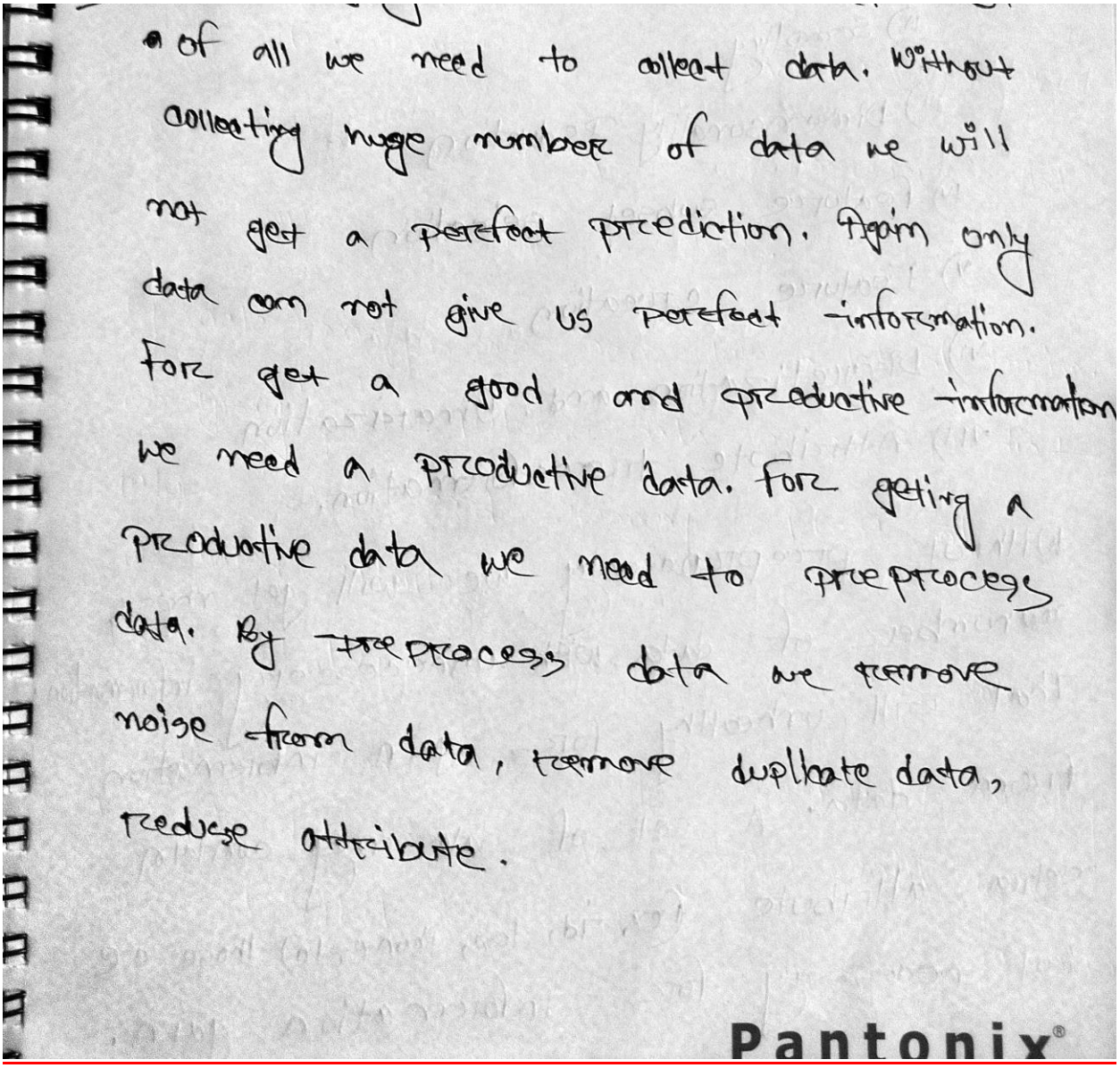
Cluster analysis that helps to identify a group. and from that group we can gain information.

Regression analysis is get a continues value that is dependent on some other independent value.

Neural network are learn like human. It segment data into different part and learn from that give us useful information. from using those methodology we can get useful information that helps to fight

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905



of all we need to collect data. Without collecting huge number of data we will not get a perfect prediction. Again only data can not give us perfect information. For get a good and productive information we need a productive data. For geting a productive data we need to preprocess data. By preprocess data we remove noise from data, remove duplicate data, reduce attribute.

Pantonix®

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

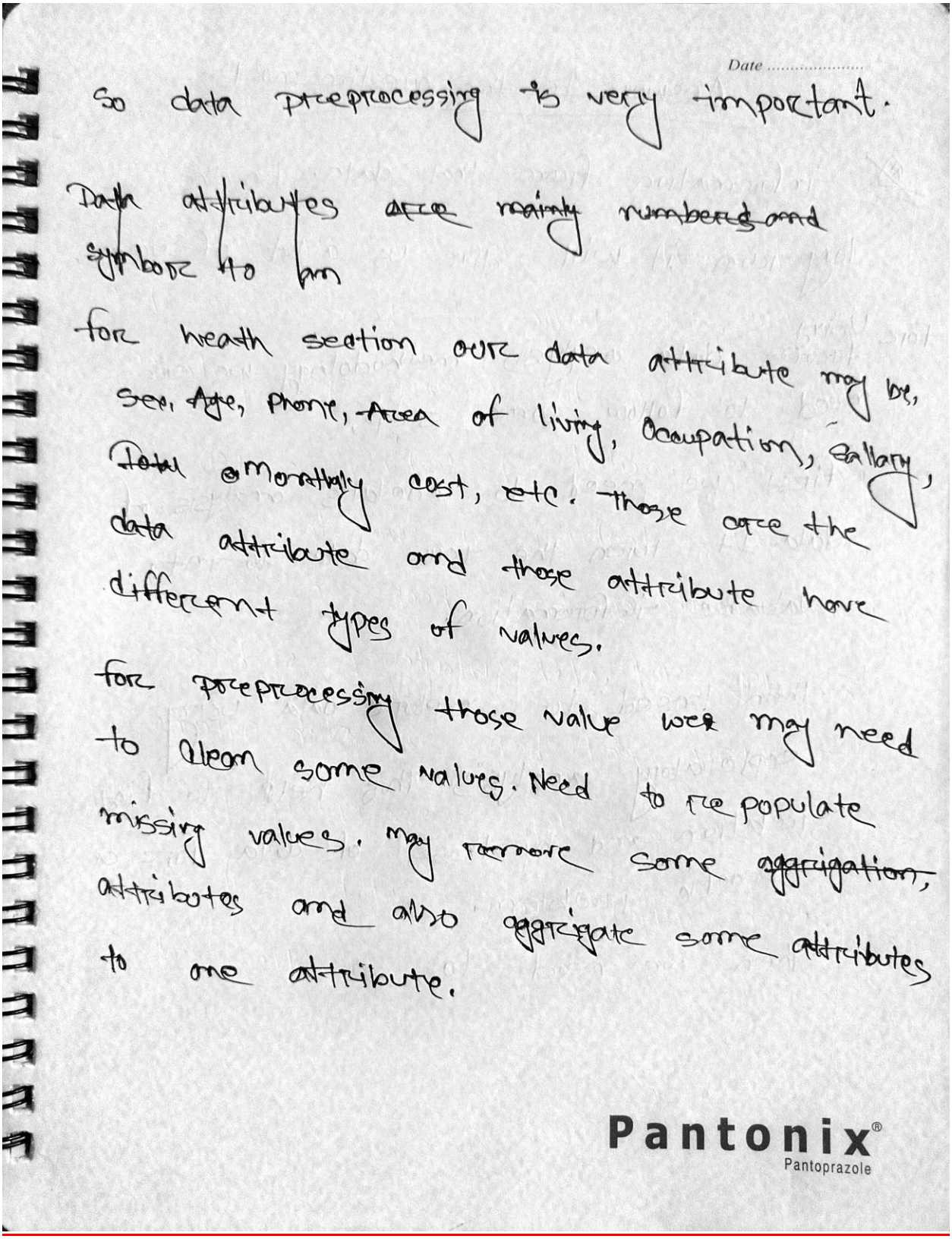
Student ID:181-15-1905

There are some data preprocessing ~~technique~~ technique,

i) ~~Aggregation~~
ii) Sampling
iii) Dimensionality Reduction
iv) ~~feature~~ feature subset selection
v) Feature creation
vi) Discretization and Binarization
vii) ~~Attribute~~ Attribute transformation.

Without preprocessing we will get ~~mess~~ number of data, with unnecessary information. That will unhealthy for gain information from data. A set of data may contain extra attribute (ex. id, tax, phone etc) those are not necessary for information gain.

Student ID:181-15-1905



Date .....

So data preprocessing is very important.

Data attributes are mainly numbers and symbols to us

for health section our data attribute may be, Sex, Age, Phone, Area of living, Occupation, Salary, Total monthly cost, etc. those are the data attribute and those attribute have different types of values.

for preprocessing those value we may need to clean some values. Need to re populate missing values. may remove some aggrigation, attributes and also aggrigate some attributes to one attribute.

Pantonix®
Pantoprazole

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905

Answer to the question no 2

Attribute class is Transportation Mode,

Hence,

Bus = 4

Train = 3

Car = 3

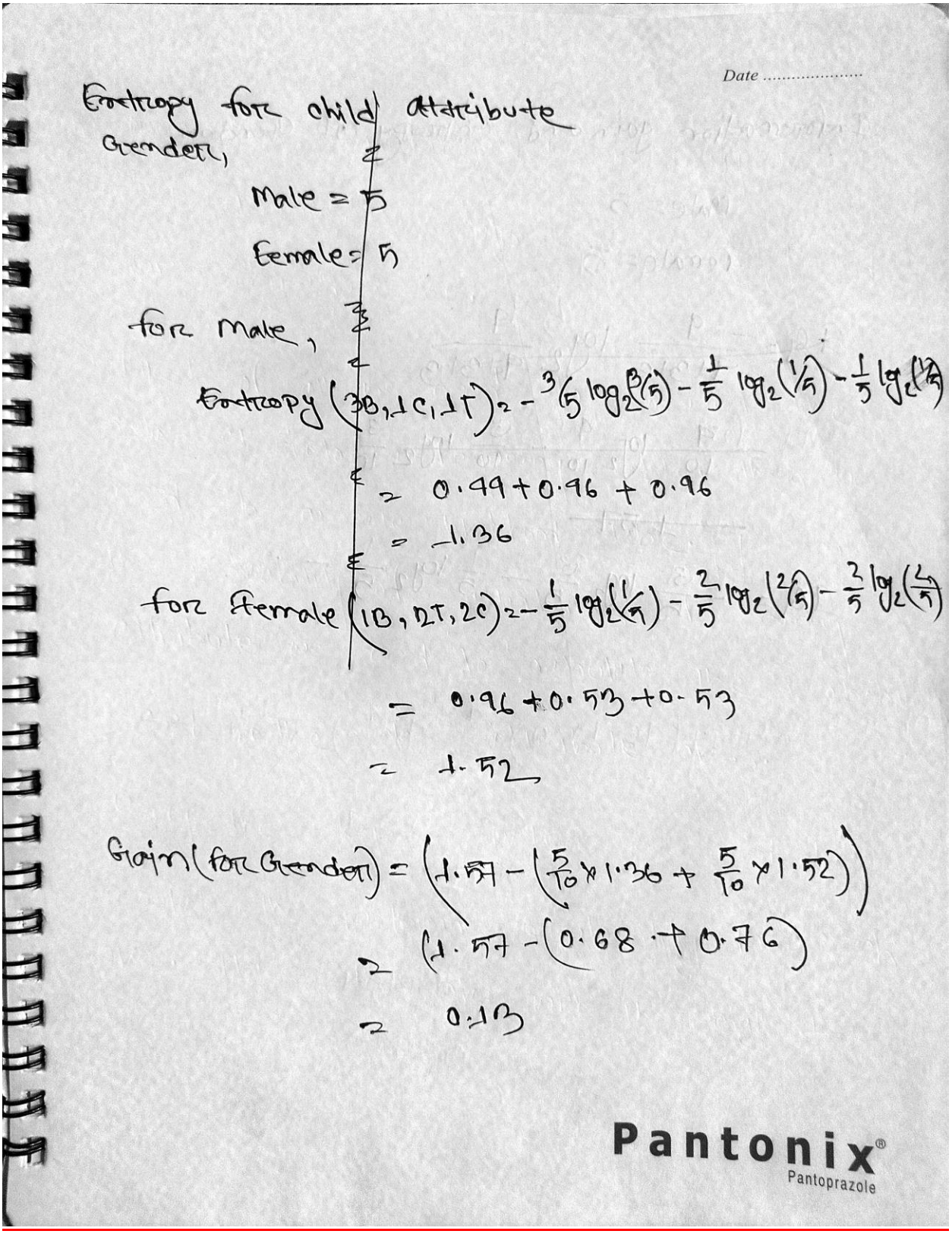
$$\text{Entropy}(4B, 3T, 3C) = \left(-\frac{4}{10}\log_2\frac{4}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{3}{10}\log_2\frac{3}{10}\right)$$

$$= 0.53 + 0.52 + 0.52$$

$$= 1.57$$

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905



Date .....

Entropy for child attribute Gender,

Male = 5

Female = 5

for Male,

Entropy $(3B, 1C, 1T) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)$

$= 0.44 + 0.46 + 0.46$

$= 1.36$

for female $(1B, 2T, 2C) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$= 0.46 + 0.53 + 0.53$

$= 1.52$

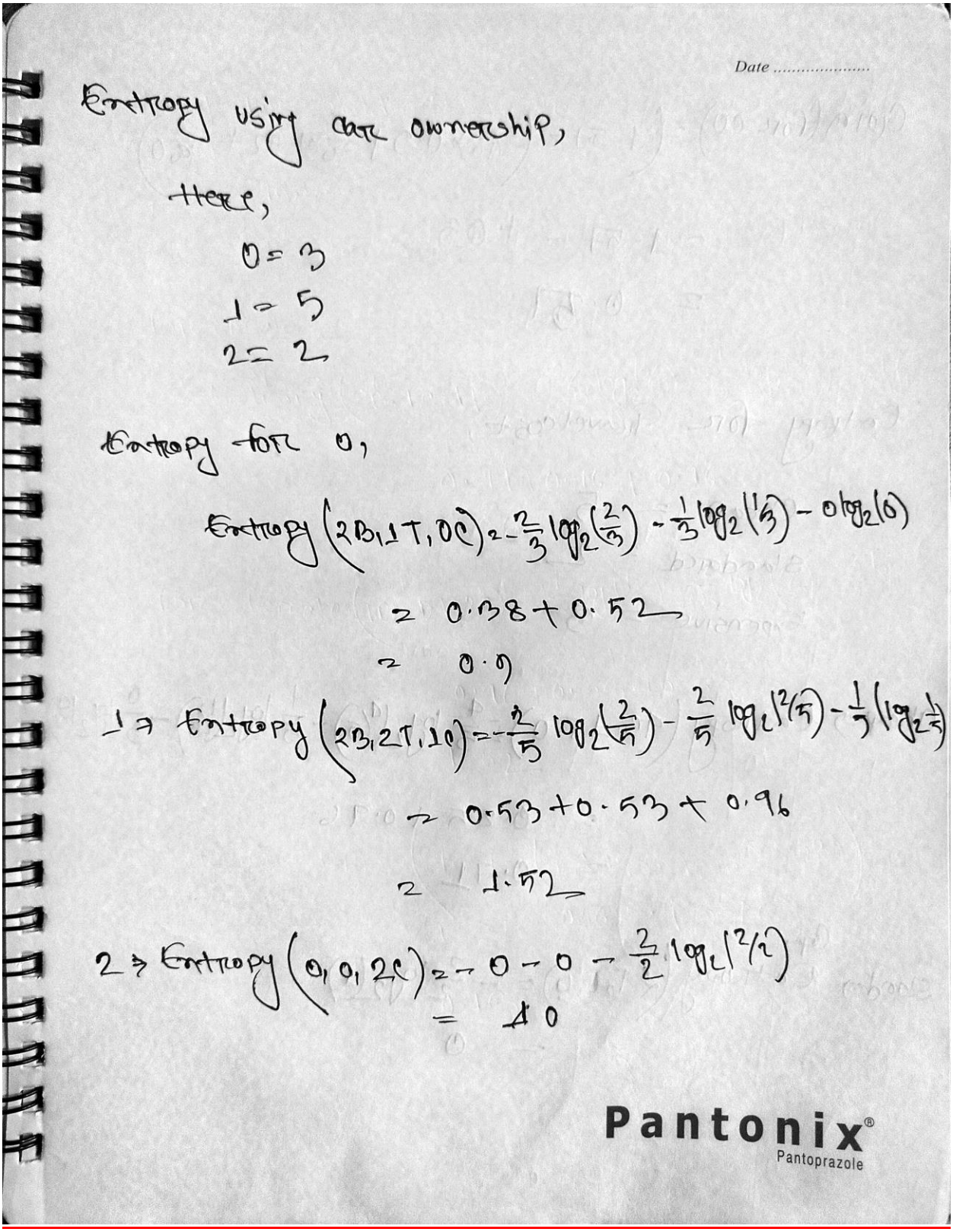
Gain(for Gender) $= \left(1.57 - \left(\frac{5}{10} \times 1.36 + \frac{5}{10} \times 1.52\right)\right)$

$= (1.57 - (0.68 + 0.76))$

$= 0.13$

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905



Date .....

Entropy using car ownership,

Here,

0 = 3

1 = 5

2 = 2

Entropy for 0,

$$\text{Entropy}(2B, 1T, 0C) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) - 0\log_2(0)$$

$$= 0.38 + 0.52$$

$$= 0.9$$

$$1 \Rightarrow \text{Entropy}(2B, 2T, 1C) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)$$

$$= 0.53 + 0.53 + 0.46$$

$$= 1.52$$

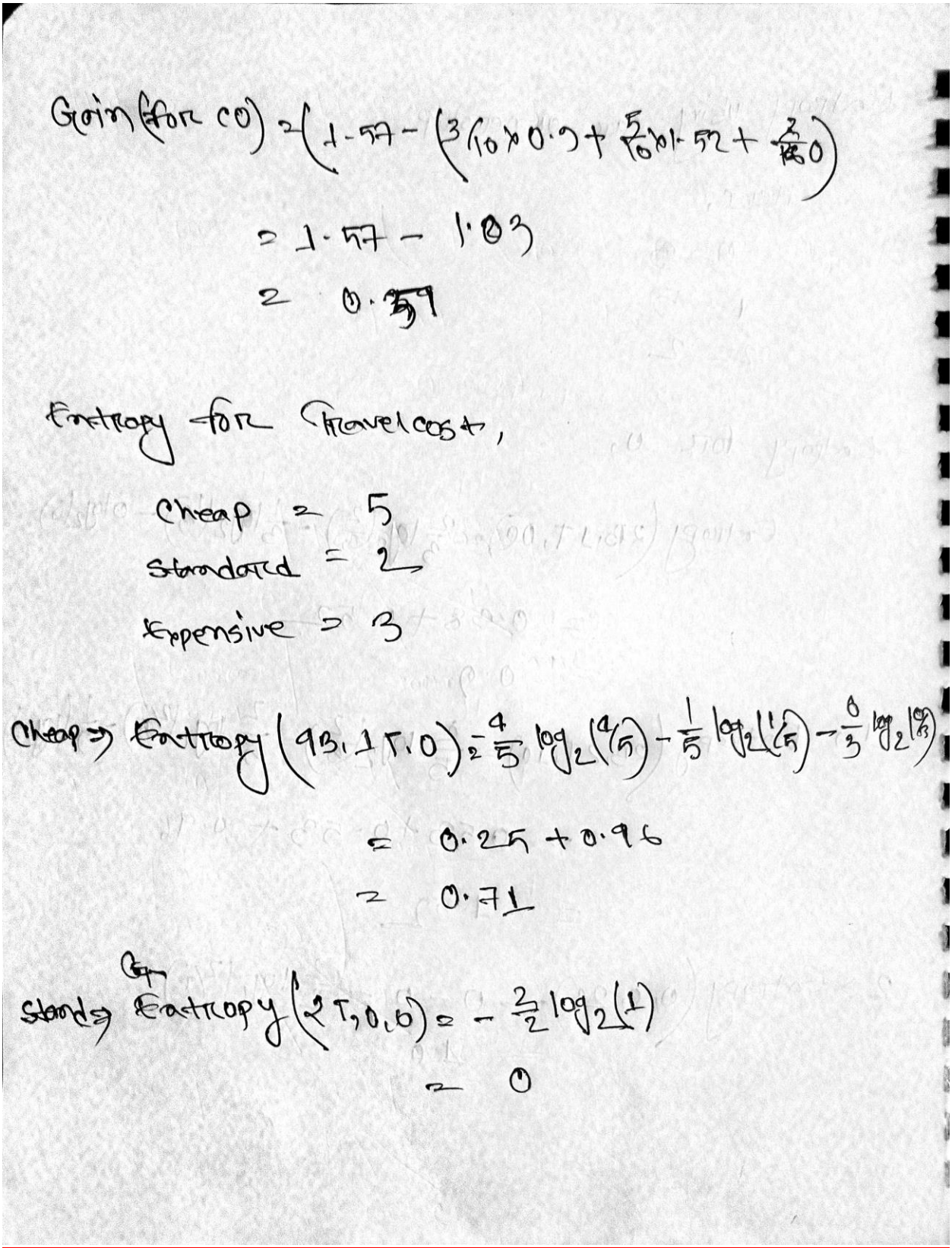
$$2 \Rightarrow \text{Entropy}(0, 0, 2C) = -0 - 0 - \frac{2}{2}\log_2\left(\frac{2}{2}\right)$$

$$= 0$$

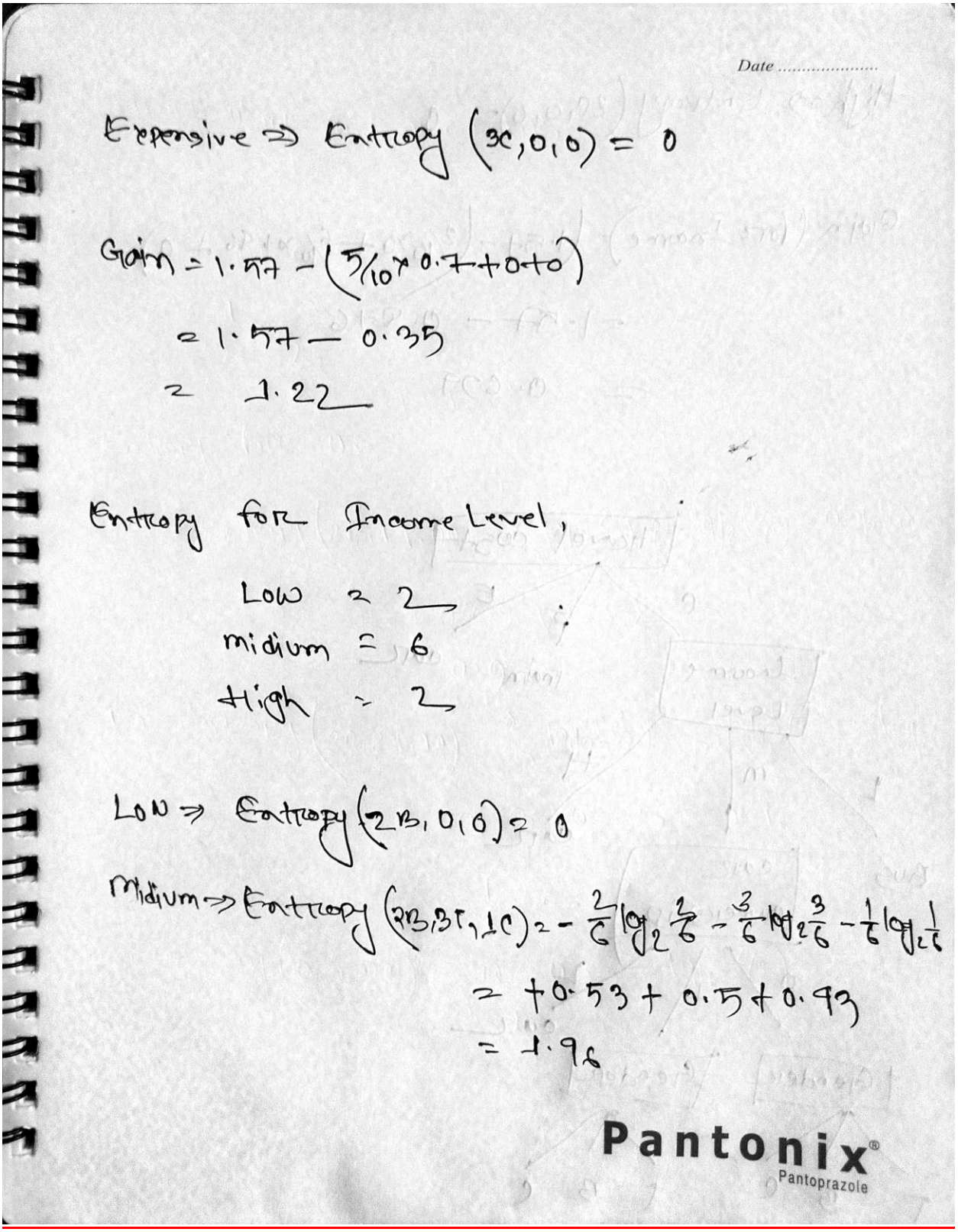
Pantonix®
Pantoprazole

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905



Student ID:181-15-1905



Date .....

Expensive ⇒ Entropy (3C, 0, 0) = 0

Gain = 1.57 − (5/10 × 0.7 + 0 + 0)

= 1.57 − 0.35

= 1.22

Entropy for Income Level,

Low = 2

midium = 6

High = 2

Low ⇒ Entropy (2B, 0, 0) = 0

Midium ⇒ Entropy (2B, 3T, 1C) = $-\frac{2}{6}\log_2\frac{2}{6} - \frac{3}{6}\log_2\frac{3}{6} - \frac{1}{6}\log_2\frac{1}{6}$

= +0.53 + 0.5 + 0.43

= 1.46

Pantonix®
Pantoprazole

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

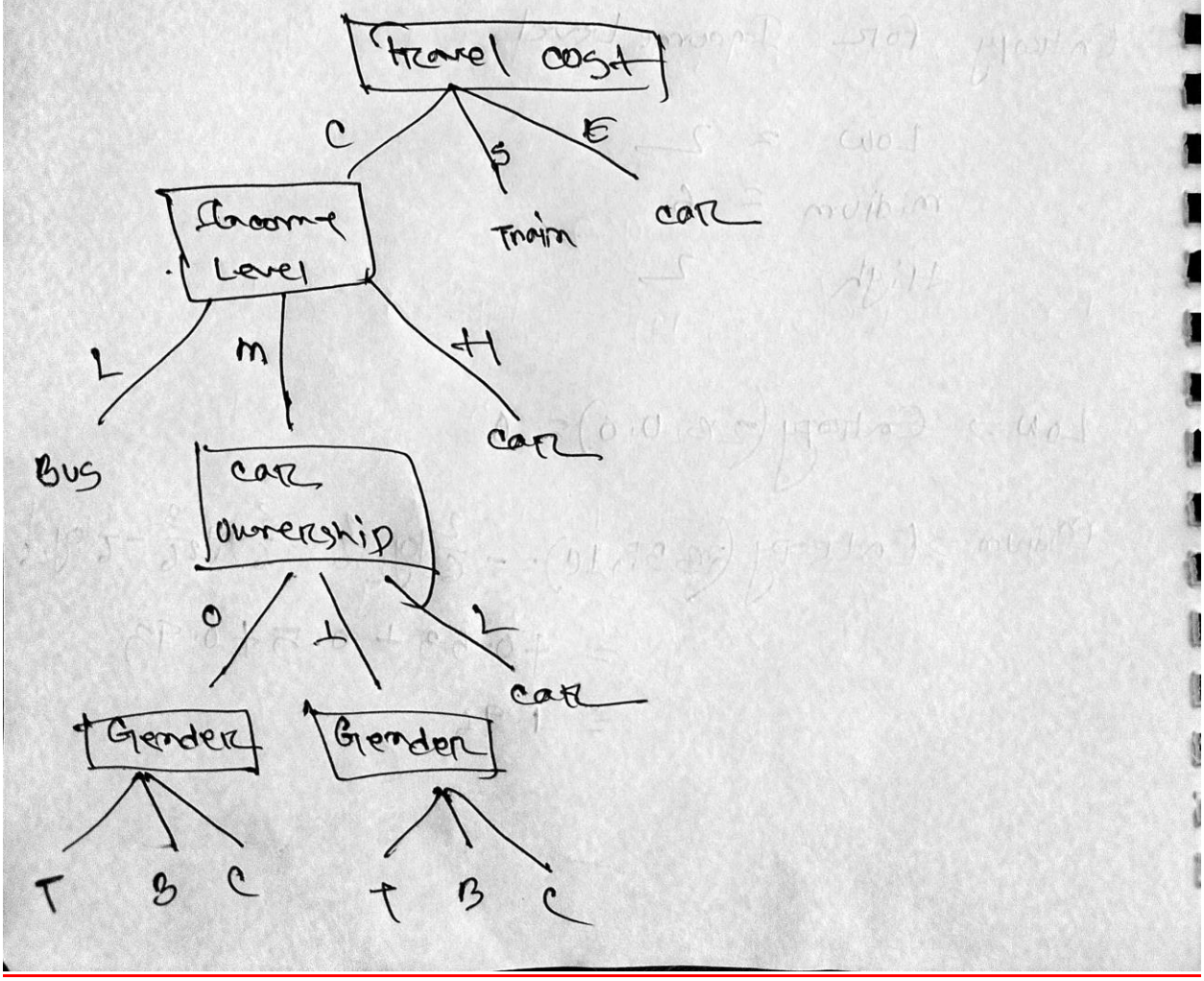
Student ID:181-15-1905

High ⇒ Entropy (20,0,0) = 0

Gain (for Income) = (1.57 − (2/10 × 0 + 6/10 × 1.46 + 0))

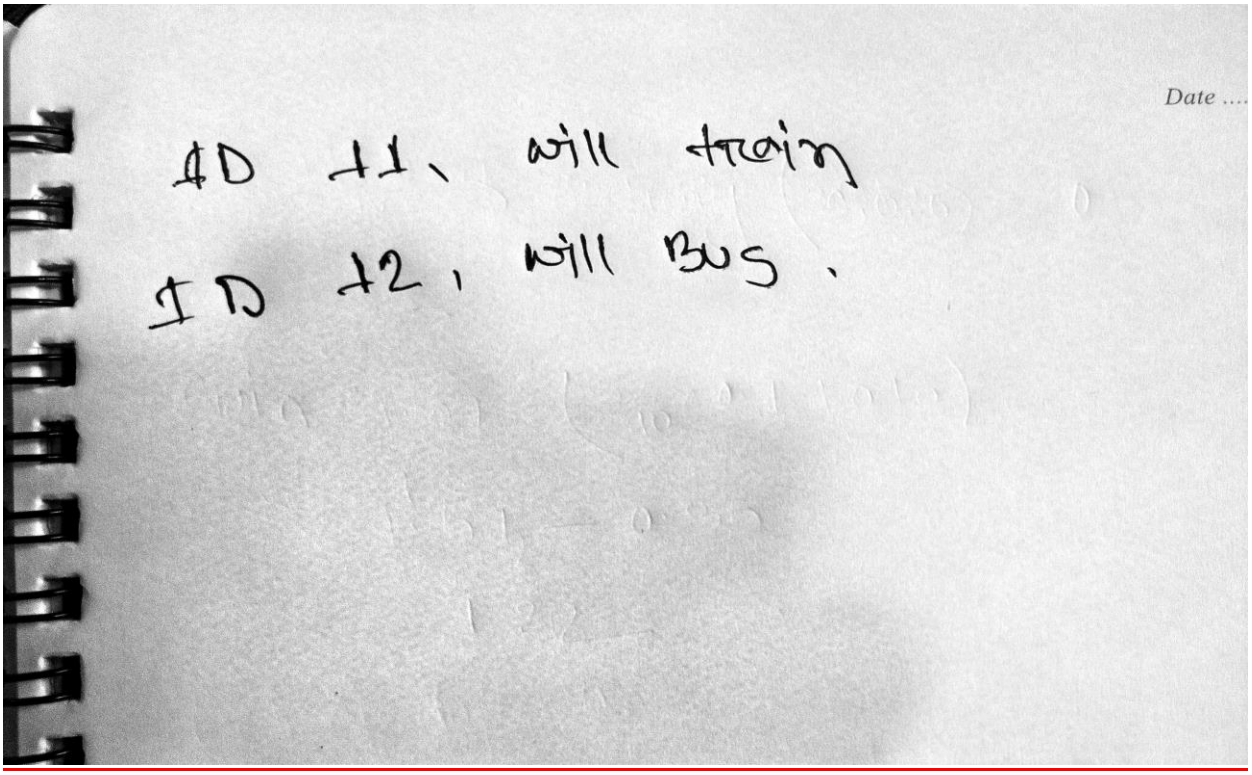
= 1.57 − 0.876

= 0.694



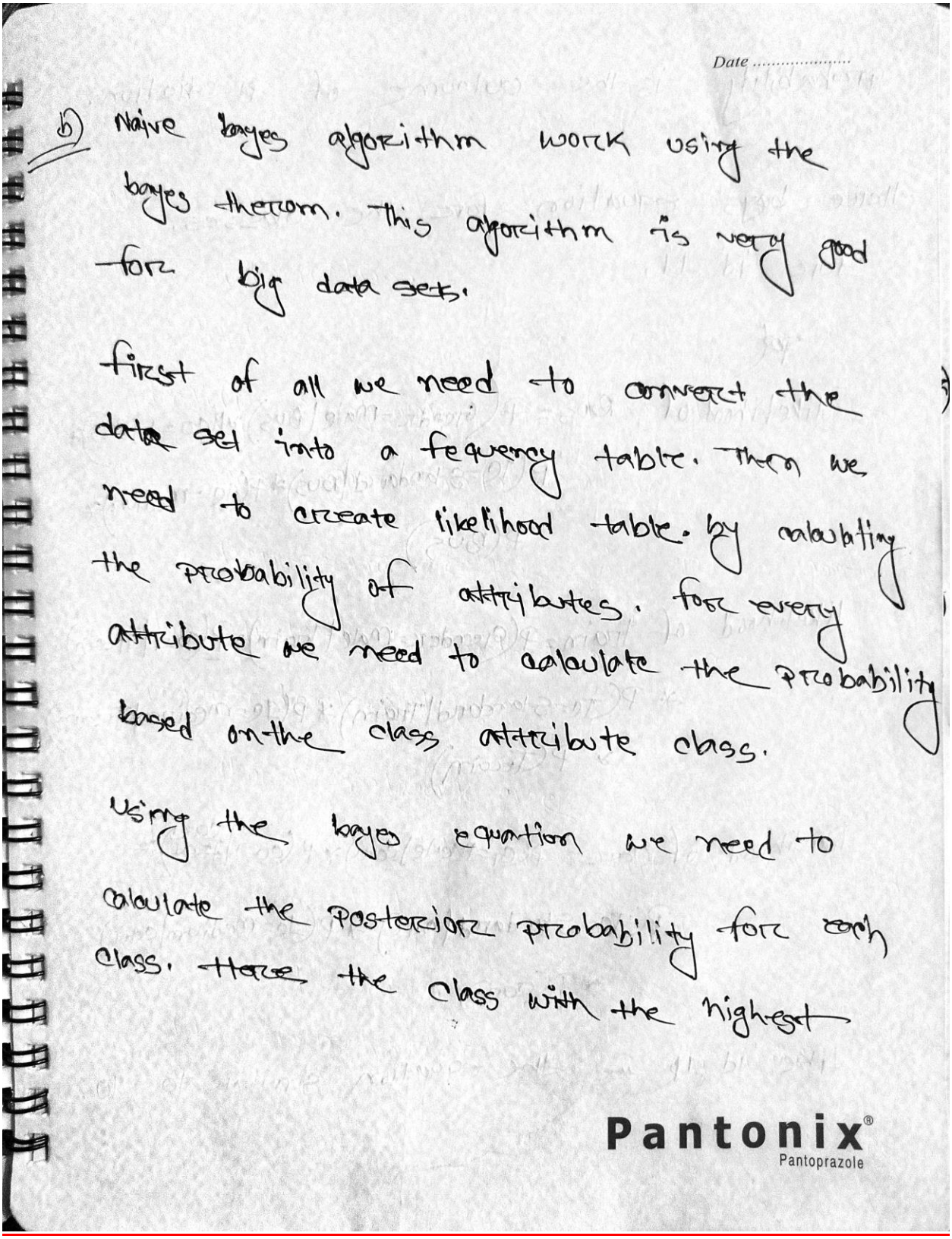
**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905



**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905



Date .....

b) Naive bayes algorithm work using the bayes theorm. this algorithm is very good for big data sets.

first of all we need to convert the data set into a fequency table. then we need to create likelihood table. by calculating the probability of attributes. for every attribute we need to calculate the probability based on the class attribute class.

using the bayes equation we need to calculate the posterior probability for each class. there the class with the highest

Pantonix®
Pantoprazole

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**

Student ID:181-15-1905

probability is the outcome of prediction.

Here bays equation for the classes,

for id 11,

likelihood of Bus = P(Gender=Male|Bus) * P(CO=1|Bus) * P(TC=Standard|Bus) * P(TC=medium) * P(Bus)

likelihood of Train = P(Gender=Male|Train) * P(CO=1|Train) * P(TC=Standard|Train) * P(TC=medium|Train) * P(Train)

likelihood of Cars = P(G=Male|Cars) * P(CO=1|Cars) * P(TC=Standard|Cars) * P(TC=Medium|Cars) * P(Cars)

like id 10 the equation similar to 10.

**** Plagiarism will be checked while you submit your response. You are advised to be honest during the open book exam.**