## ASSIGNMENT 1E – Due on Wednesday 11:55pm of Week 11.

## TOPICS TESTED: REGRESSION & CLASSIFICATION

### Do's & Don'ts:

*** All answer in this assignment can be done in any colour **except for red colour** as this is the colour your tutor will use to mark your assignment. Thus, answers in red colour will not be graded (even correct ones) ***

*** If you choose to handwrite your answer (scan and submit it electronically), make sure your **handwriting is readable** – this is a good practice for your exam. Failure to comply will result in a lack of marks awarded if your writing is unreadable and there will be no exception for this. ***

*** **Each sub-assignment will be worth 2.5%** towards your total score. Sub-assignments will have different point distributions within them as we aim to focus your attention to the important areas; however, the score for a sub-assignment remains 2.5%. ***

*** These questions are meant for you to solve **independently**, we encourage students to figure out the questions themselves as it would be good for their understanding of the topics; however, please feel free to consult your tutors if needed. **Plagiarism** (either from using online sources or copying the answers from your classmates) will be punished accordingly. ***

*** As this is considered to be an assignment (albeit a sub assignment), requests for special consideration or extension must be submitted at least **2 days BEFORE THE DEADLINE.** The due date is on Sunday, so the latest day you can ask for extension is on Friday (the last official working day of the week for the teaching team). Please follow Monash guideline to request for extensions (medical certificates, doctor or GP letter, etc). Emergencies are to be adjusted individually. ***

*** **No R or any other programming languages should be used in solving these questions**. All work for this assignment needs to be done manually, less the use of **non-programmable calculator** (this also applies to your Final Exam). Tutors are not required to answer questions in the difference between manual calculation and programmed calculation***

*** **Late submission is 10% per day, after 5 days you will be given no marks**. Late submission is calculated as following: If you get 70% on this assignment and you are late for 2 days (you submit on Tuesday), your scores is now 70% -20% (2x10% per day) = 50%. This is done to ensure that the teaching team can release your result as soon as possible so that you can review on your mistakes and have a better study experience. ***

*** Please **show all working** in answering questions, your score will be **halved** if you don't comply***

*** Assignments shall be marked completely in **two weeks' time** according to Monash Policies. If there are any changes to the marking time, we will duly inform you. Solutions **will not be released** for this assignment; you can come to the tutorial and ask for an explanation about how to solve the questions after scores are released. ***

*** Please don't send emails to tutors asking for suggestions, we have Moodle and consultations for that, In writing your inquiries on Moodle please try to be clear in your problem and not revealing your working to others as this might be counted as plagiarism on your part. A good format for inquiry topic would be "Assignment 1E – Tutorial 10 (your tutorial slot) – Question about median "***

*** Assignments need to be submitted in **PDF** format. Failure to comply will result in 30% penalty***

*** Filename format for submitting assignment "Assignment1E_StudentId.pdf". File with wrong format incurs 30% penalty ***

## QUESTIONS:

### A. REGRESSION: (7 Marks Total)

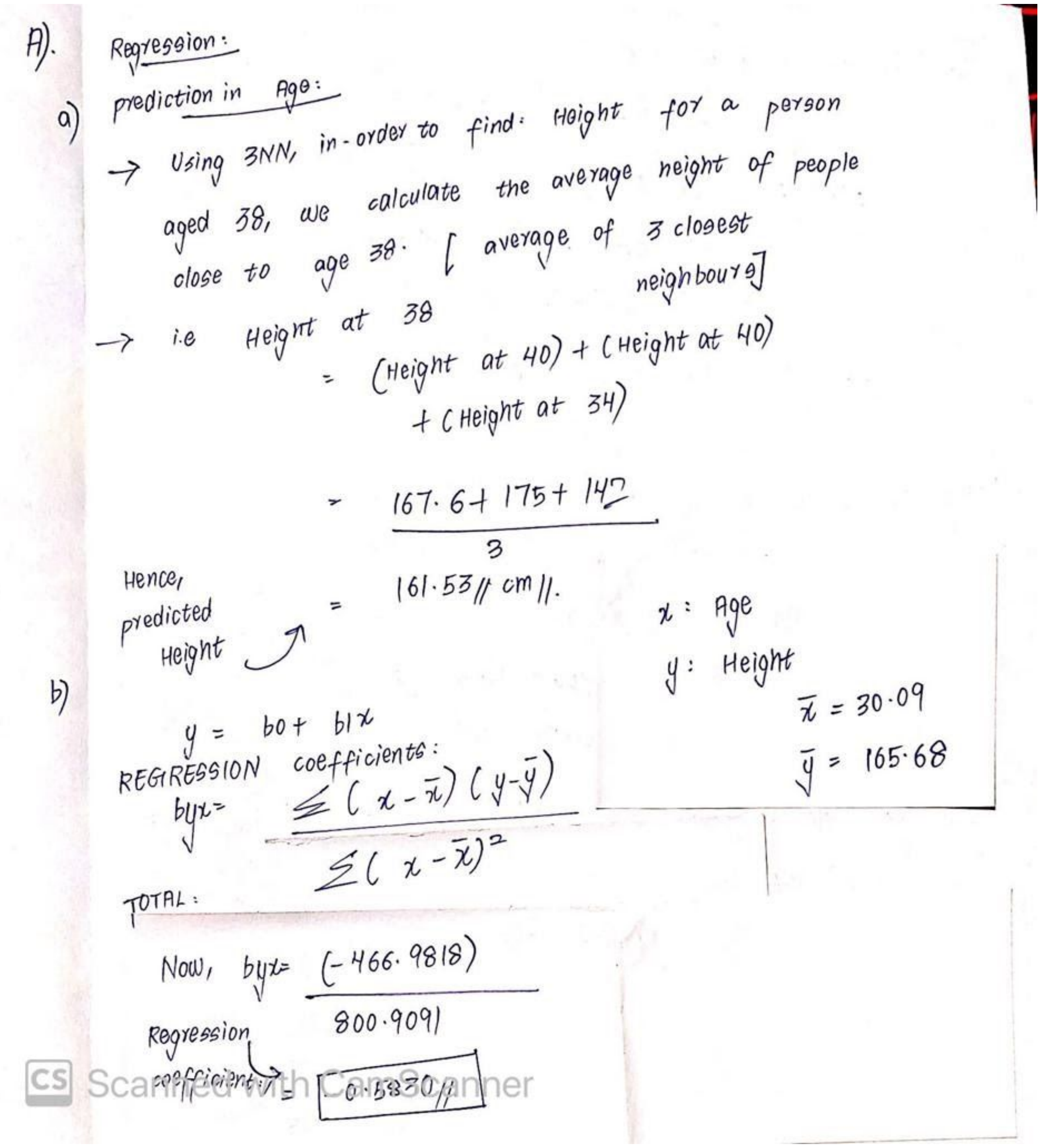
| Height (cm) | Age |
|---|---|
| 152.4 | 45 |
| 180 | 26 |
| 167.6 | 30 |
| 167.6 | 34 |
| 175 | 40 |
| 142 | 40 |
| 172.7 | 19 |
| 165 | 19 |
| 165 | 23 |
| 167.6 | 23 |
| 167.6 | 32 |
| 165 | 38 |

The data above will be used for these following questions **ignoring the last data entry (Height 165, Age 38)** as this is going to be your testing data for your model:

- **a.** Let age be the predictor for building our model to predict for height, using 3NN what should be the prediction for a person at 38 years old **(1 Marks)**
- **b.** Using Linear Regression, what would be the value for the coefficients? **(2 Marks)**
- **c.** Is age an important variable in predicting height or not and why? **(2 Marks)**
- **d.** Calculate R-squared for your regression model and explain what this means? **(1 Marks)**
- **e.** Make comparison of KNN, Linear Regression, and real data for the case when the person age is 38, what would be your conclusion here? **(1 Marks)**

***SOLUTION*:**

**1b : Regression Coefficient is -0.5830. ( The final answer is not clear in the image provided )**



A).

Regression:

a) prediction in Age:

→ Using 3NN, in-order to find: Height for a person aged 38, we calculate the average height of people close to age 38. [ average of 3 closest neighbours]

→ i.e Height at 38

$$= (\text{Height at } 40) + (\text{Height at } 40) + (\text{Height at } 34)$$

$$= \frac{167.6 + 175 + 142}{3}$$

Hence, predicted Height ↗ = 161.53// cm //.

$x$ : Age

$y$ : Height

$\bar{x} = 30.09$

$\bar{y} = 165.68$

b)

$y = b0 + b1x$

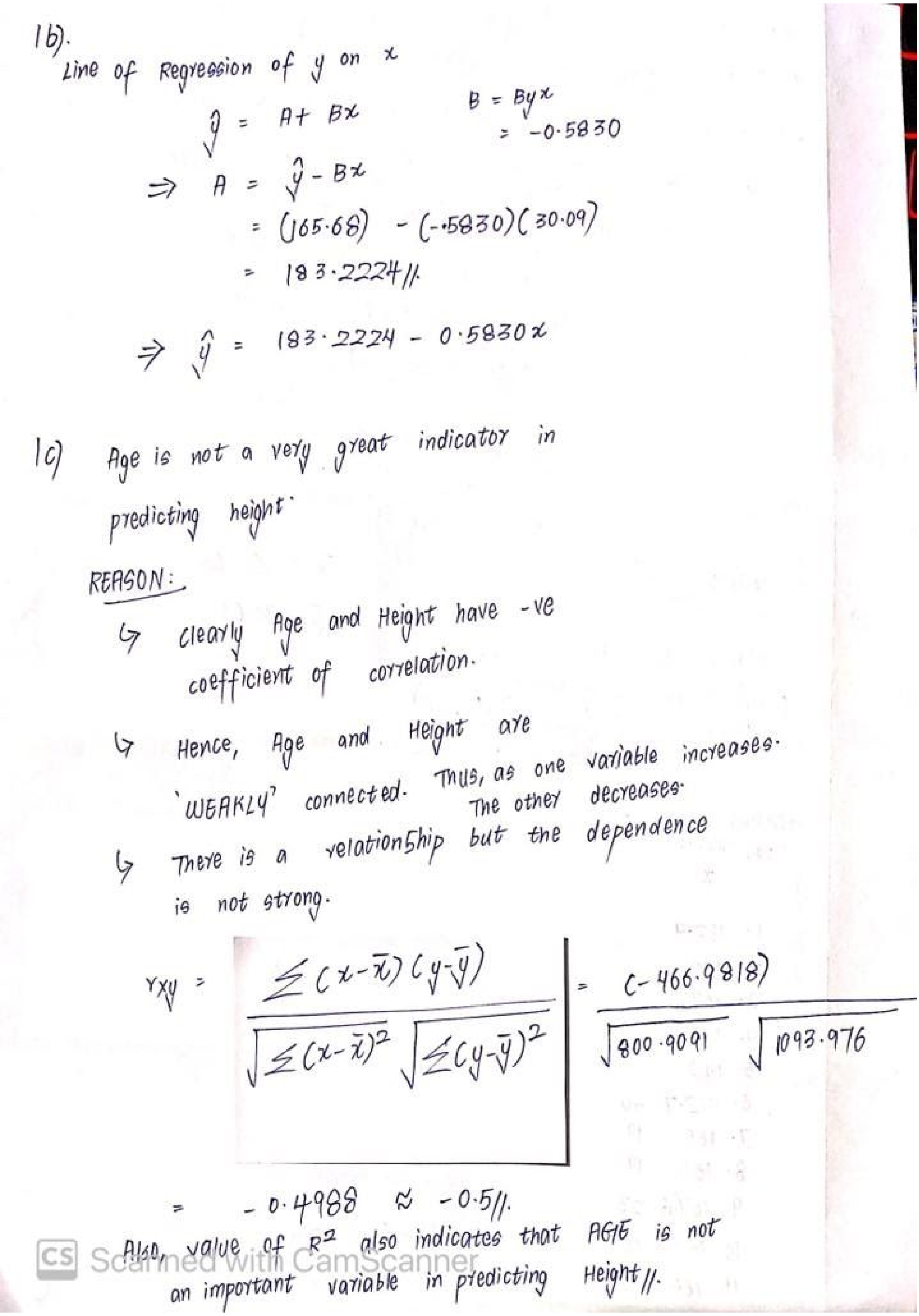
REGRESSION coefficients:

$$byx = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

TOTAL:

Now, $byx = \frac{(-466.9818)}{800.9091}$

Regression coefficient → = $\boxed{-0.5830//}$



1b).

Line of Regression of y on x

$\hat{y} = A + Bx$ $\qquad B = B_{yx} = -0.5830$

$\Rightarrow A = \hat{y} - Bx$

$= (165.68) - (-.5830)(30.09)$

$= 183.2224$ //

$\Rightarrow \hat{y} = 183.2224 - 0.5830x$

1c) Age is not a very great indicator in predicting height.

REASON:

↳ Clearly Age and Height have -ve coefficient of correlation.

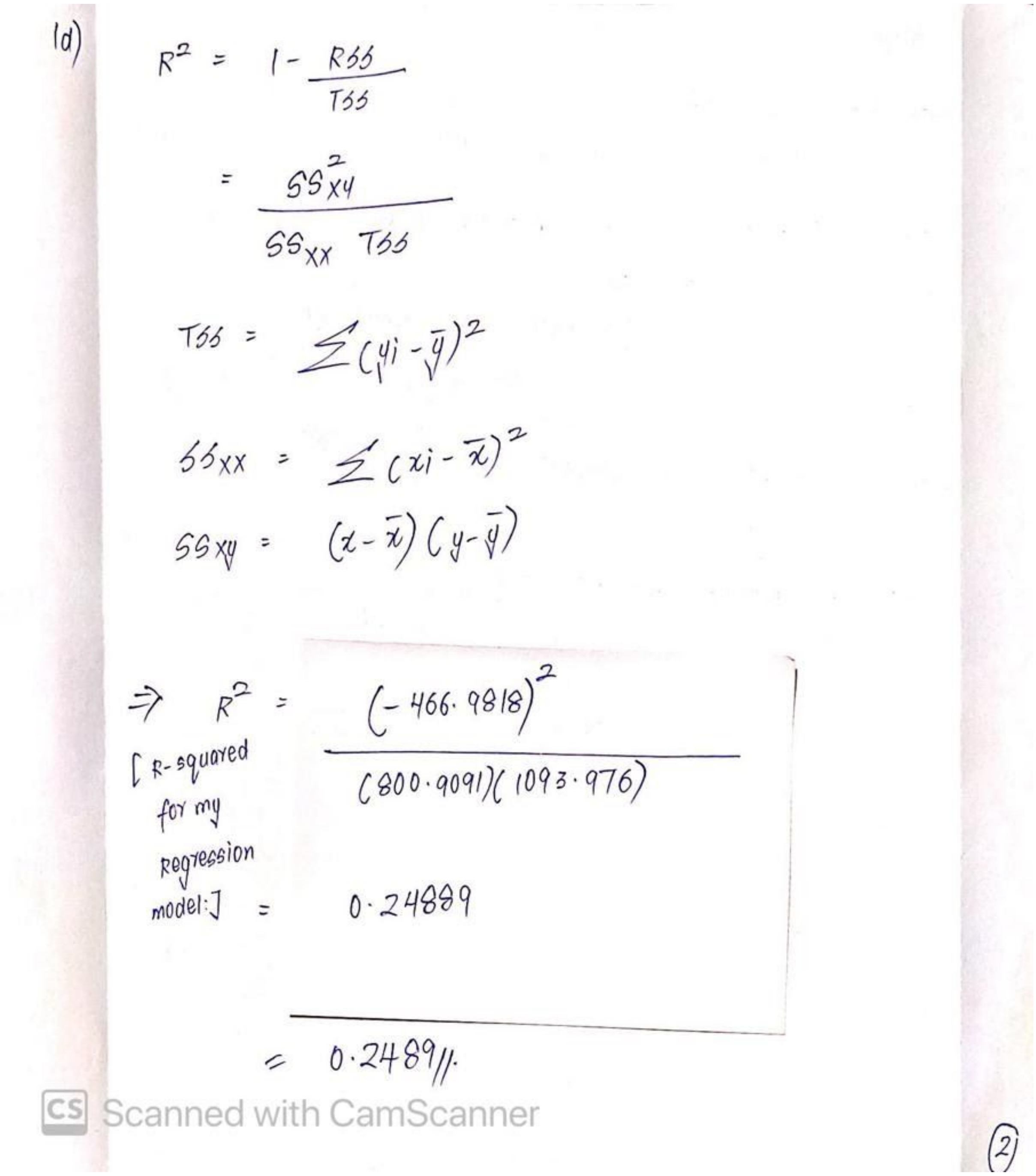
↳ Hence, Age and Height are 'WEAKLY' connected. Thus, as one variable increases. The other decreases.

↳ There is a relationship but the dependence is not strong.

$$r_{xy} = \boxed{\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}} = \frac{(-466.9818)}{\sqrt{800.9091}\sqrt{1093.976}}$$

$= -0.4988 \approx -0.5$ //.

Also, value of $R^2$ also indicates that AGE is not an important variable in predicting Height //.



1d)

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$= \frac{SS_{xy}^2}{SS_{xx}\ TSS}$$

$$TSS = \sum (y_i - \bar{y})^2$$

$$SS_{xx} = \sum (x_i - \bar{x})^2$$

$$SS_{xy} = (x - \bar{x})(y - \bar{y})$$

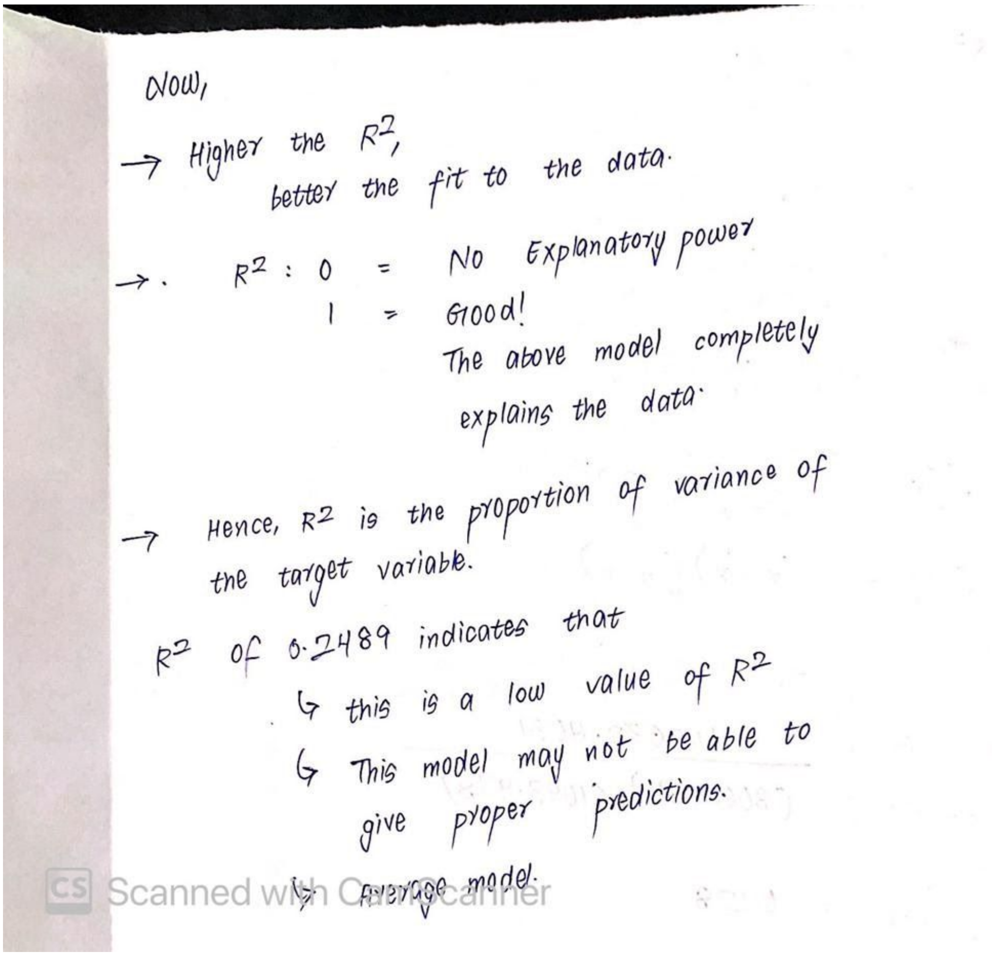
$\Rightarrow$ $R^2$ [R-squared for my regression model:] $= \dfrac{(-466.9818)^2}{(800.9091)(1093.976)}$

$= 0.24889$

$= 0.2489$//.

(2)

**Meaning of R Squared:**



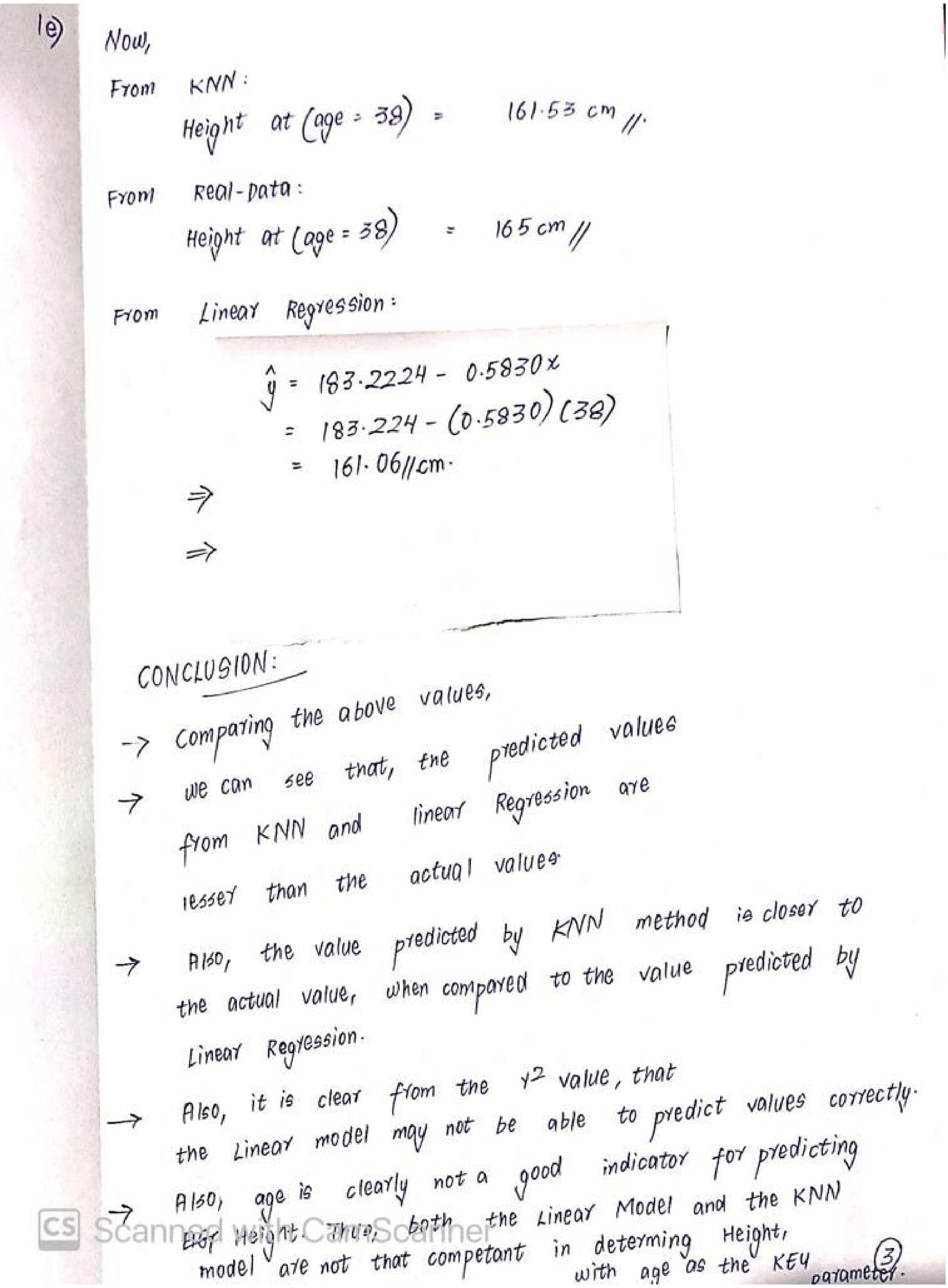
Now,

→ Higher the $R^2$,
better the fit to the data.

→ $R^2$ : 0 = No Explanatory power
1 = Good!
The above model completely explains the data.

→ Hence, $R^2$ is the proportion of variance of the target variable.

$R^2$ of 0.2489 indicates that
↳ this is a low value of $R^2$
↳ This model may not be able to give proper predictions.
↳ Average model.



1e) Now,

From KNN:

Height at (age = 38) = 161.53 cm //.

From Real-Data:

Height at (age = 38) = 165 cm //

From Linear Regression:

$$\hat{y} = 183.2224 - 0.5830x$$
$$= 183.224 - (0.5830)(38)$$
$$= 161.06 // \text{cm}.$$

⇒

⇒

CONCLUSION:

-> Comparing the above values,

-> we can see that, the predicted values from KNN and linear Regression are lesser than the actual values

-> Also, the value predicted by KNN method is closer to the actual value, when compared to the value predicted by Linear Regression.

-> Also, it is clear from the $r^2$ value, that the Linear model may not be able to predict values correctly.

-> Also, age is clearly not a good indicator for predicting for Height. Thus, both the Linear Model and the KNN model are not that competant in determing Height, with age as the KEY parameter.

3

## B. CLASSIFICATION (3 Marks Total)

Naïve Bayes question **(3 Marks)**

Given the following dataset about weather in Melbourne:

| OUTLOOK | TEMPERATURE | HUMIDITY | WINDY | PLAY GOLF |
|---|---|---|---|---|
| Rainy | Hot | High | False | No |
| Rainy | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Sunny | Mild | High | False | Yes |
| Sunny | Cool | Normal | False | Yes |
| Sunny | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Rainy | Mild | High | False | No |
| Rainy | Cool | Normal | False | Yes |
| Sunny | Mild | Normal | False | Yes |
| Rainy | Mild | Normal | True | Yes |
| Overcast | Mild | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Sunny | Mild | High | True | No |

a. Given that it is rainy, temperature is hot, with a normal humidity, and it is not windy, should you play golf or not? **(1 Mark)**

b. Given that you can't play golf, sky is windy with a high humidity and cool temperature, what would be the most possible outlook? (**1 Mark**)

c. Given that you can play golf, it is quite windy, hot outside and rainy, what would be the most possible humidity state? (**1 Mark**)

***SOLUTION*:**

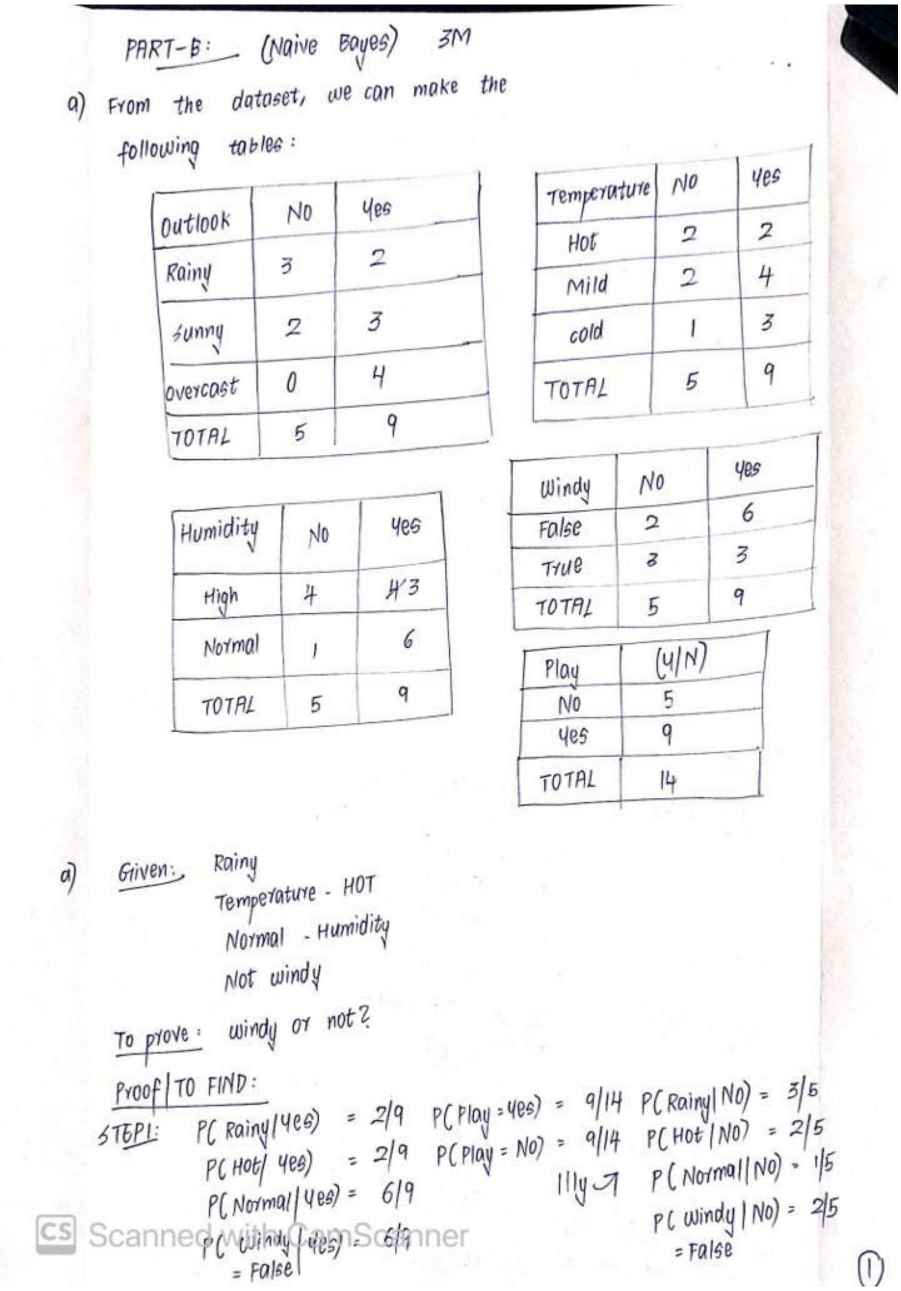
**Answers at a glance :**

**a) Yes - we can play golf.**

**b) Most Possible Outlook : Rainy**

**c) Most Possible Humidity : High**

**In Detail Explanation given below in the pictures:**



PART-B: (Naive Bayes) 3M

a) From the dataset, we can make the following tables:

| Outlook | No | Yes |
|---|---|---|
| Rainy | 3 | 2 |
| Sunny | 2 | 3 |
| Overcast | 0 | 4 |
| TOTAL | 5 | 9 |

| Temperature | No | Yes |
|---|---|---|
| Hot | 2 | 2 |
| Mild | 2 | 4 |
| cold | 1 | 3 |
| TOTAL | 5 | 9 |

| Windy | No | Yes |
|---|---|---|
| False | 2 | 6 |
| True | 3 | 3 |
| TOTAL | 5 | 9 |

| Humidity | No | Yes |
|---|---|---|
| High | 4 | ~~4~~3 |
| Normal | 1 | 6 |
| TOTAL | 5 | 9 |

| Play | (Y/N) |
|---|---|
| No | 5 |
| Yes | 9 |
| TOTAL | 14 |

a) Given: Rainy
Temperature - HOT
Normal - Humidity
Not windy

To prove: windy or not?

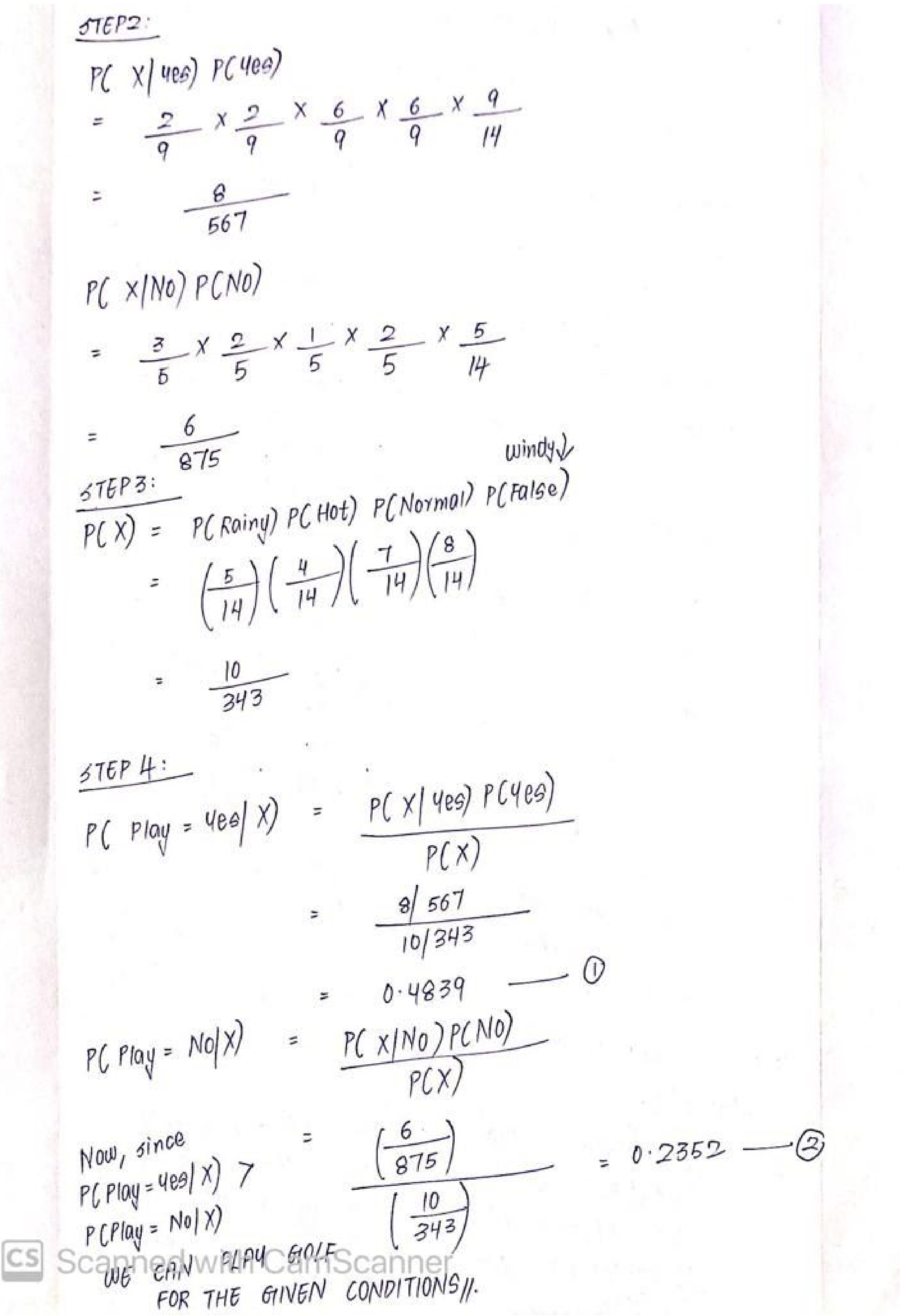
Proof | TO FIND:

STEP1: P( Rainy/Yes) = 2/9
P( Hot/ Yes) = 2/9
P( Normal/ Yes) = 6/9
P( Windy/ Yes) = 6/9
= False

P( Play = Yes) = 9/14
P( Play = No) = 9/14
IIly ↗

P( Rainy| No) = 3/5
P( Hot / No) = 2/5
P( Normal| No) = 1/5
P( windy | No) = 2/5
= False

(1)



STEP2:

$P(X|yes)\ P(yes)$

$= \frac{2}{9} \times \frac{2}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14}$

$= \frac{8}{567}$

$P(X|NO)\ P(NO)$

$= \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14}$

$= \frac{6}{875}$

STEP3:

windy↓

$P(X) = P(Rainy)\ P(Hot)\ P(Normal)\ P(False)$

$= \left(\frac{5}{14}\right)\left(\frac{4}{14}\right)\left(\frac{7}{14}\right)\left(\frac{8}{14}\right)$

$= \frac{10}{343}$

STEP 4:

$P(Play = yes|X) = \frac{P(X|yes)\ P(yes)}{P(X)}$

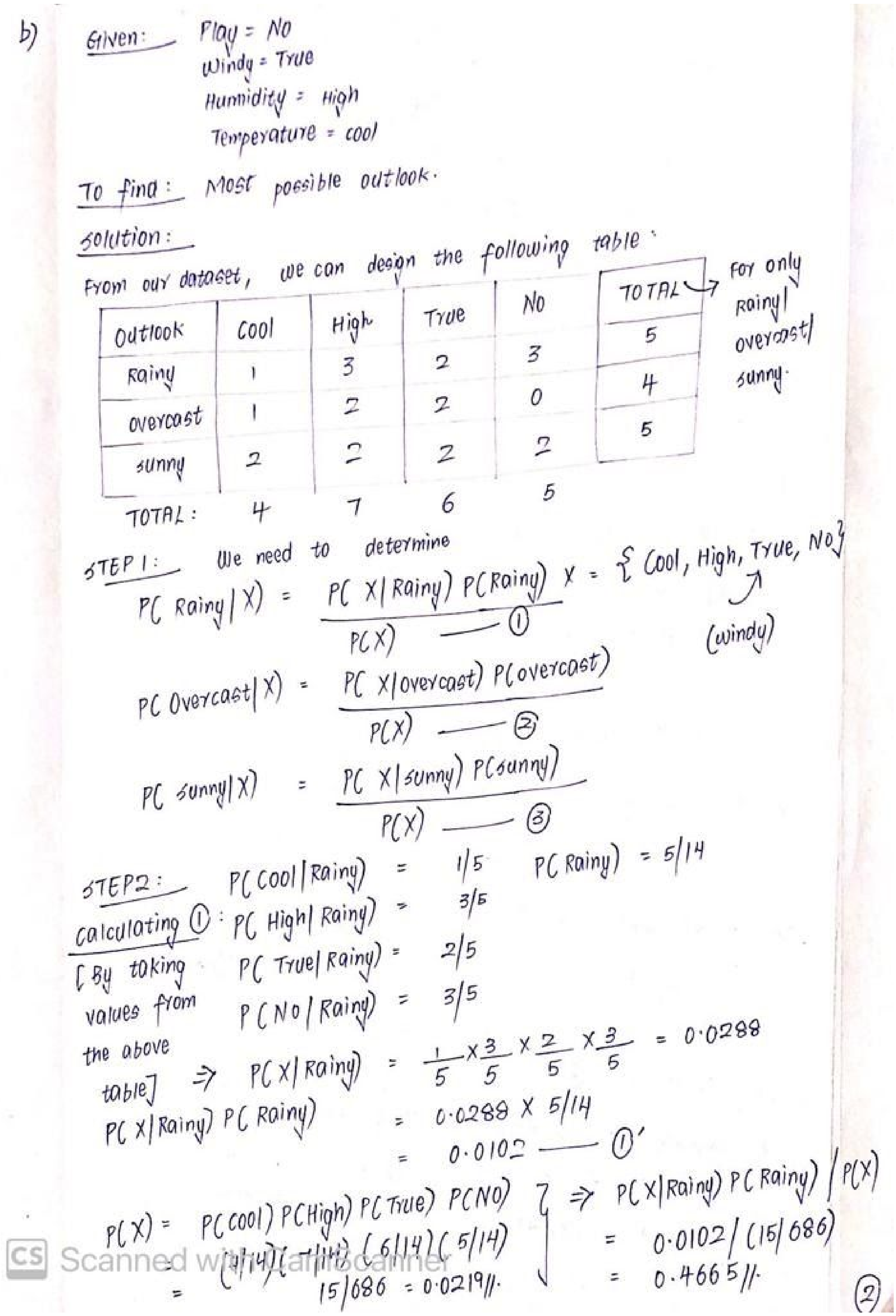
$= \frac{8/567}{10/343}$

$= 0.4839$ — ①

$P(Play = No|X) = \frac{P(X|No)\ P(NO)}{P(X)}$

$= \frac{\left(\frac{6}{875}\right)}{\left(\frac{10}{343}\right)} = 0.2352$ — ②

Now, since
$P(Play = yes|X)$ 7
$P(Play = No|X)$
WE CAN PLAY GOLF
FOR THE GIVEN CONDITIONS//.



b) Given: Play = No
Windy = True
Humidity = High
Temperature = cool

To find: Most possible outlook.

Solution:

From our dataset, we can design the following table:

| Outlook | Cool | High | True | No | TOTAL |
|---|---|---|---|---|---|
| Rainy | 1 | 3 | 2 | 3 | 5 |
| overcast | 1 | 2 | 2 | 0 | 4 |
| sunny | 2 | 2 | 2 | 2 | 5 |
| TOTAL: | 4 | 7 | 6 | 5 | |

TOTAL → For only Rainy/overcast/sunny.

STEP 1: We need to determine

$$P(\text{Rainy}|X) = \frac{P(X|\text{Rainy})\,P(\text{Rainy})}{P(X)} \text{ — ①} \qquad X = \{\text{Cool, High, True, No}\}$$

True → (windy)

$$P(\text{Overcast}|X) = \frac{P(X|\text{overcast})\,P(\text{overcast})}{P(X)} \text{ — ②}$$

$$P(\text{sunny}|X) = \frac{P(X|\text{sunny})\,P(\text{sunny})}{P(X)} \text{ — ③}$$

STEP 2: Calculating ①: [By taking values from the above table]

$P(\text{Cool}|\text{Rainy}) = 1/5$ $\quad P(\text{Rainy}) = 5/14$

$P(\text{High}|\text{Rainy}) = 3/5$

$P(\text{True}|\text{Rainy}) = 2/5$

$P(\text{No}|\text{Rainy}) = 3/5$

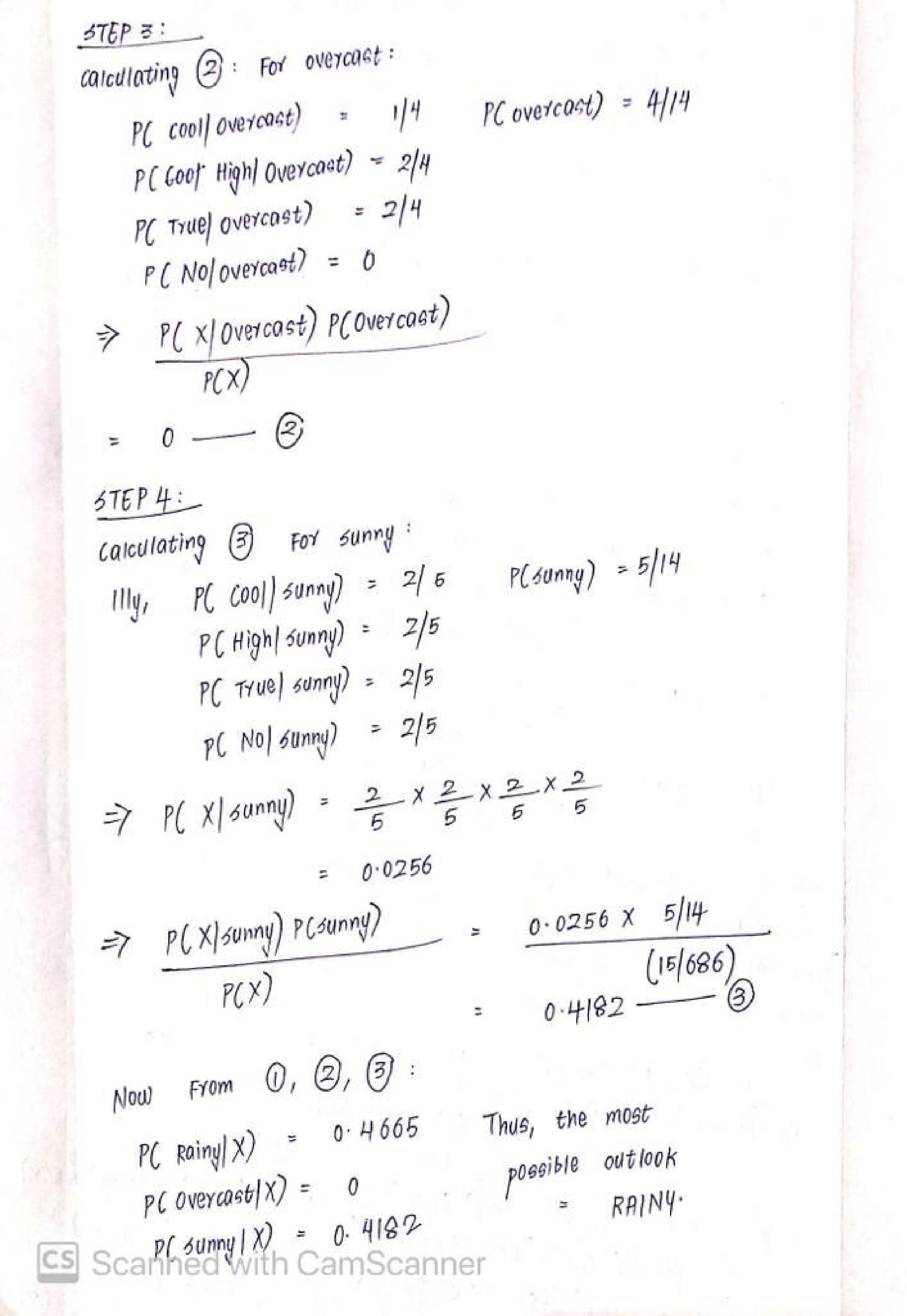
$$\Rightarrow P(X|\text{Rainy}) = \frac{1}{5} \times \frac{3}{5} \times \frac{2}{5} \times \frac{3}{5} = 0.0288$$

$$P(X|\text{Rainy})\,P(\text{Rainy}) = 0.0288 \times 5/14$$
$$= 0.0102 \text{ — ①}'$$

$$P(X) = P(\text{cool})\,P(\text{High})\,P(\text{True})\,P(\text{NO})$$
$$= (4/14)(7/14)(6/14)(5/14)$$
$$= 15/686 = 0.0219//.$$

$$\Rightarrow P(X|\text{Rainy})\,P(\text{Rainy}) / P(X)$$
$$= 0.0102/(15/686)$$
$$= 0.4665//.$$

②



STEP 3:

Calculating ②: For overcast:

$P(\text{cool} | \text{Overcast}) = 1/4$ $\quad P(\text{overcast}) = 4/14$

$P(\text{~~Cool~~ High} | \text{Overcast}) = 2/4$

$P(\text{True} | \text{overcast}) = 2/4$

$P(\text{No} | \text{overcast}) = 0$

$$\Rightarrow \frac{P(X | \text{Overcast})\, P(\text{Overcast})}{P(X)}$$

$= 0$ —— ②

STEP 4:

Calculating ③ For sunny:

IIIy, $P(\text{Cool} | \text{sunny}) = 2/5$ $\quad P(\text{sunny}) = 5/14$

$P(\text{High} | \text{sunny}) = 2/5$

$P(\text{True} | \text{sunny}) = 2/5$

$P(\text{No} | \text{sunny}) = 2/5$

$$\Rightarrow P(X | \text{sunny}) = \frac{2}{5} \times \frac{2}{5} \times \frac{2}{5} \times \frac{2}{5}$$

$$= 0.0256$$

$$\Rightarrow \frac{P(X|\text{sunny})\, P(\text{sunny})}{P(X)} = \frac{0.0256 \times 5/14}{(15/686)}$$

$$= 0.4182 \text{ —— ③}$$

Now From ①, ②, ③:

$P(\text{Rainy} | X) = 0.4665$

$P(\text{Overcast} | X) = 0$

$P(\text{sunny} | X) = 0.4182$

Thus, the most possible outlook = RAINY.

c) Given : Play = ~~No~~ yes
windy = True
Temperature = Hot
outlook = Rainy

To find: Most possible HUMID state.

solution: From the above data, we can draw the following:

| Humidity | Hot | Rainy | yes | ~~Total~~ True | TOTAL |  |
|---|---|---|---|---|---|---|
| High | 3 | 3 | 3 | 3 | 7 | (High) |
| Normal | 1 | 2 | 6 | 3 | 7 | (Normal) |
| TOTAL : | 4 | 5 | 9 | 6 | 14 |  |

STEP1:

$$P(High|X) = \frac{P(X|High)\ P(High)}{P(X)}$$

$$P(Normal|X) = \frac{P(X|Normal)\ P(Normal)}{P(X)}$$

We need to determine this'

$X = \{$ Hot, Rainy, Play = yes, windy = True $\}$

STEP 2 :

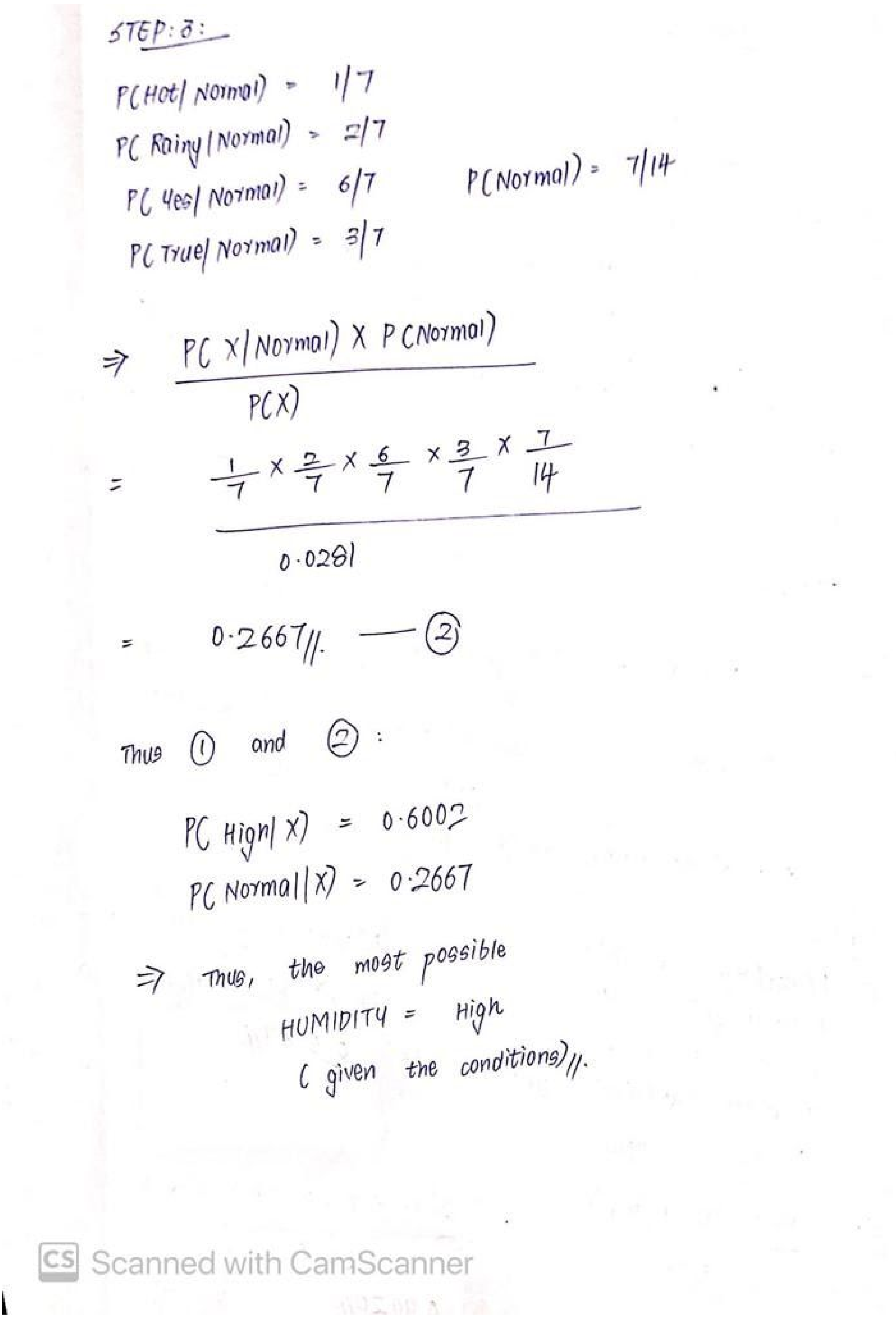
① $P(Hot|High) = 3/7$
$P(Rainy|High) = 3/7$
$P(Play = yes|High) = 3/7$
$P(windy = True|High) = 3/7$
$P(High) = 7/14 =$

$P(X) = P(Hot)\ P(Rainy)\ P(yes)\ P(True)$
$= (4/14)(5/14)(9/14)(6/14)$
$= 135/4802$
$= 0.0281$ //.

$$\Rightarrow \frac{P(X|High)\ P(High)}{P(X)} = \frac{3/7 \times 3/7 \times 3/7 \times 3/7 \times 7/14}{0.0281}$$

$= 0.6002$ //.

(3)



STEP: 3:

P(Hot| Normal) = 1/7

P( Rainy | Normal) = 2/7

P( Yes| Normal) = 6/7          P(Normal) = 7/14

P( True| Normal) = 3/7

$$\Rightarrow \frac{P(X|Normal) \times P(Normal)}{P(X)}$$

$$= \frac{\frac{1}{7} \times \frac{2}{7} \times \frac{6}{7} \times \frac{3}{7} \times \frac{7}{14}}{0.0281}$$

$$= 0.2667// \text{ — } ②$$

Thus ① and ② :

P( High| X) = 0.6002

P( Normal| X) = 0.2667

⇒ Thus, the most possible HUMIDITY = High ( given the conditions)//.