Part No.

Page 11

0094

Name: Mateeh Ullah Reg #: 19L-2313 Section: CS-8B

National University of Computer and Emerging Sciences, Lahore Campus

Course:	Data Science	Course Code:	CS4048
Program:	BS(Computer Science)	Semester:	Spring 2023
Duration:	180 Minutes	Total Marks:	90
Paper Date:	25-May-23	Weight	40
Section:	A & B	Page(s):	6
Exam:	FINAL		

Instruction/Notes: Attempt the examination on the answer sheets and write concise answers. Clearly write the question number and your answers in the answer Booklet(s) provided. You can use extra sheet for rough work. Do not attach extra sheets used for rough work with the answer sheets. Do not use pencil or red ink to answer the questions. In case of any confusion or ambiguity, make a reasonable assumption.

Question	1	2	3	4	5	6	Total
Marks	/20	/15	/19	/15	/5	/16	/90

Research problem

Question # 1: **[20 points]**

You have been hired to work on a research project. In this research project, you are tasked with investigating the association between movies and socioeconomic factors, particularly focusing on the UK film industry. Movies and films have a significant influence on society, reflecting our values and shaping our perceptions. They can also impact socioeconomic factors. The research aims to explore the effects of movie genres and box office gross on socioeconomic factors such as GDP, crime rate, hourly pay, health, education, and employment. Following are the research questions.

- What have been the Effects of having socio-Economic factors like GDP, Crime rate, Hourly Pay, Health, Education, and Employment on movie development in the UK from 2000 to 2022?
- Does Movie Genre affect any of the socioeconomic factors?
- Does Box Office Gross affect any of the socioeconomic factors?

1. Explain how would you support the topic originality with your literature review? **[5 points]**
2. Explain how you would identify and collect relevant data sources for both the movie industry and socioeconomic factors in the UK. Consider the data availability, reliability, and relevance to the research questions. **[5 points]**
3. Describe the steps you would take to clean and preprocess the collected data, addressing issues such as missing values, outliers, and inconsistencies. Also mention how you would integrate the different data sources, ensuring data compatibility. **[5 points]**
4. Describe the techniques and visualizations you would utilize to explore the data. Discuss your methodology how you would approach research question 2 & 3 to analyze the relationships between the variables. **[5 points]**

Name: ______________________ Reg #: ______________________ Section: ______

Regression Analysis

Question # 2: **[15 points]**

Research Question: Does the amount of time spent studying and the number of practice problems completed have a significant impact on a student's test score?

Sample Data:

Student	Time Spent Studying (hours) X_1	Practice Problems Completed X_2	Test Score (out of 100) y
1	4	10	75
2	5	12	83
3	3	8	68
4	6	15	90
5	2	5	60
6	7	18	95
7	4	9	73
8	5	13	85
	4.5	11.25	78.625

13.035
~~[illegible]~~
18.9868
111.9
128.4
344.18
265.865
31.546
40.49

In this analysis, you would use multiple linear regression to model the relationship between the independent variables, time spent studying and practice problems completed, and the dependent variable, test score. From the regression results, you will find the p-values associated with each coefficient. A low p-value (typically less than 0.05) indicates statistical significance.

$$b_1=\frac{\left(\sum x_2^2\right)\left(\sum x_1y\right)-\left(\sum x_1x_2\right)\left(\sum x_2y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right)-\left(\sum x_1x_2\right)}$$

$$b_2=\frac{\left(\sum x_1^2\right)\left(\sum x_2y\right)-\left(\sum x_1x_2\right)\left(\sum x_1y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right)-\left(\sum x_1x_2\right)}$$

$$a=b_0=\bar{Y}-b_1\bar{X}_1-b_2\bar{X}_2$$

a) Calculate the values of the coefficients and interpret their meaning. **[10 points]**

b) The p-values for Time Spent Studying and Practice Problems Completed are 0.0517 and 0.0243, respectively. State the significance of each attribute. **[5 points]**

$b_1 = 0.04446$

$b_2 = -6.153\times10^{-3}$

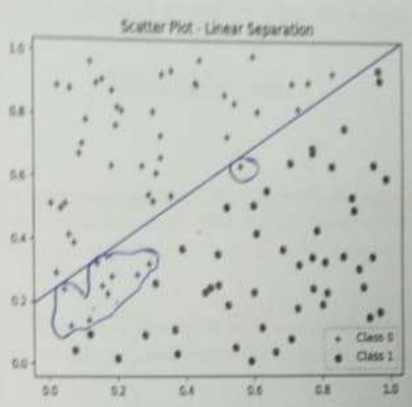
$b_0 = 78.49415$

0094

Name: ______________________ Reg #: ______________________ Section: ________

Question # 3: [19 points]

8



a) Using slope=1 and intercept=0.2 create a linear line to separate the two classes. **[5 points]**

b) How many points are going to be misclassified? Mark on the given graph. **[5 points]**

c) If true positives are 49, false positives are 16 and false negatives are 0, and total records are 100, calculate accuracy, precision and recall. **[9 points]**

Clustering

Question # 4: [15 points]

Using the given dataset (**given in question # 2**) of students' time spent studying and number of practice problems completed, apply the k-means clustering algorithm to identify distinct groups of students based on their study habits and performance.

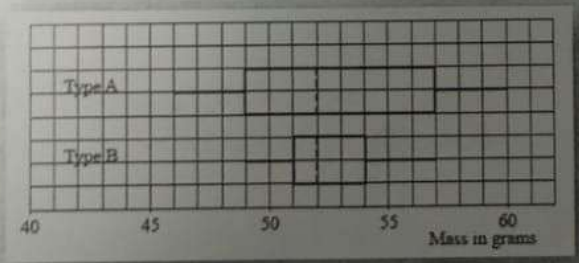
a) Your task is to apply the k-means algorithm to group the students into 2 clusters based on their study habits and performance. **[10 points]**

b) Create a scatter plot (on paper) and show the formed clusters. **[5 points]**

Visualization

Question # 5 [5 points]

A gardener collected data on two types of tomato. The box and whisker plot below shows data for the masses in grams of the tomatoes in the two samples. Compare and contrast the two types and advise the gardener which type of tomato he should grow in future.



Name: ____________________ Reg #: ____________________ Section: ________

Note: In this part you can attempt any one question of your own choice.

Deep Learning

[16 points]

Question # 6:

Convolutional Neural Networks (CNNs) have become a fundamental component of deep learning, particularly in the field of computer vision. They have revolutionized the field of image classification and object recognition by automating feature extraction. However, CNNs can be computationally intensive and often require GPUs for efficient training.

a) Explain the challenges with traditional multilayered neural networks.
b) Describe the purpose and functionality of convolutional layers, kernals, pooling layers, and fully connected layers in a CNN.
c) Given the following input matrix and a 3x3 filter perform a convolution operation with stride set to 1 and then 2. Show the resultant feature maps.

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

⊗

0	0	1
1	0	0
0	1	1

Name: ______________________ Reg #: __________________ Section: ______

Decision Tree

Question # 6: **[16 points]**

The following dataset will be used to learn a decision tree for predicting whether a mushroom is edible or not based on its shape, color and odor.

	Shape	Color	Odor	Edible
A1	C	B	1	Yes
A2	D	B	1	Yes
A3	D	W	1	Yes
A4	D	W	2	Yes
A5	C	B	2	Yes
A6	D	B	2	No
A7	D	G	2	No
A8	C	U	2	No
A9	C	B	3	No
A10	C	W	3	No
A11	D	W	3	No

a) What is entropy H(Edible|Odor = 1 or Odor = 3)? **[4 points]**

b) Which attribute would the ID3 algorithm choose to use for the root of the tree (no pruning)? **[4 points]**

c) Draw the full decision tree that would be learned for this data (no pruning). **[4 points]**

d) Suppose we have a validation set as follows. What will be the training set error and validation set error of the tree? Express your answer as the number of examples that would be misclassified. **[4 points]**

Shape	Color	Odor	Edible
C	B	2	No
D	B	2	No
C	W	2	Yes

Good Luck!

Q / Part No. Rough Work

11

Q#2)

(a)

$$b1 = \frac{\left[(10)^2+(12)^2+(8)^2+(15)^2+(5)^2+(18)^2+(9)^2+(13)^2\right]\left[(4)(75)+(5)(83)+(3)(68)+(6)(90)+(2)(60)+(7)(95)+(4)(73)+(5)(85)\right]-\left[(4)(10)+(5)(12)+(3)(8)+(6)(15)+(2)(5)+(7)(18)+(4)(9)+(5)(13)\right]\left[(10)(75)+(12)(83)+(8)(68)+(15)(90)+(5)(60)+(18)(95)+(9)(73)+(13)(85)\right]}{\left[4^2+5^2+3^2+6^2+2^2+7^2+4^2+5^2\right]\left[10^2+12^2+8^2+15^2+5^2+18^2+9^2+13^2\right]-\left[(4)(10)+(5)(12)+(3)(8)+(6)(15)+(2)(5)+(7)(18)+(4)(9)+(5)(13)\right]}$$

$$b1 = \frac{(1132)(2961)-(451)(7412)}{(180)(1132)-(451)}$$

$$b1 = \frac{9040}{203309} = 0.04446$$

$$b2 = \frac{(180)(7412)-(451)(2961)}{203309}$$

$$b2 = \frac{-1251}{203309} = -6.153\times10^{-3}$$

$$a = b_0 = 78.625-(0.04446)(4.5) - (-6.153\times10^{-3})(11.25)$$

$$a = bo = 78.625 - 0.20007 + 0.06922125$$

$$a = bo = 78.49415$$

$$J(b_0, b_1, b_2) = \frac{1}{2m}\sum_{i=0}^{m}(h_\theta(x)-y^{(i)})$$

$$h_\theta(x) = b_0 + b_1x_1 + b_2x_2$$

$$\frac{1}{2(8)}\left(\left(\left[78.49415+0.04446(4)+(-6.153\times10^{-3})(10)\right]-75\right)^2+\left(\left[78.49415+0.04446(5)+(-6.153\times10^{-3})(12)\right]-83\right)^2\right.$$

Page 2

Q / Part No. | Rough Work

$$
\begin{aligned}
&+ ([78.49415 + 0.04446(3) - (6.153\times10^{-3})(8)] - 68)^2 \\
&+ ([78.49415 + 0.04446(6) - (6.153\times10^{-3})(15)] - 90)^2 \\
&+ ([78.49415 + 0.04446(2) - (6.153\times10^{-3})(5)] - 60)^2 \\
&+ ([7.849415 + 0.04446(7) - (6.153\times10^{-3})(18)] - 95)^2 \\
&+ ([7.849415 + 0.04446(4) - (6.153\times10^{-3})(9)] - 73)^2 \\
&+ ([7.849415 + 0.04446(5) - (6.153\times10^{-3})(13)] - 85)^2
\end{aligned}
$$

$$= \frac{1}{16}[13.035 + 18.9868 + 111.9 + 128.4 + 344.18 + 265.865 + 31.546 + 40.49]$$

$J(b_0, b_1, b_2) = 59.65$

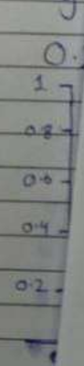
The values are $b_0 = 78.49415$, $b_1 = 0.04446$, $b_2 = -6.153\times10^{-3}$ and cost ~~will be~~ is 59.65. There is some error in b_0, b_1, b_2 which will be minimize using gradient descent. After that model will predict value with less or no error.

(b) The value 0.0517 and 0.0243 are minimized values. These values can be use to predict values because both have less or no error.

Q / Part No.

Q#

(a) y=



(b)

(c) Pred

Pr

Departm

Rough Work

$46(3)-(6.153\times10^{-3})(8)]-68)^2$

$)]-90)^2$

$-60)^2$

$-95)^2$

$-73)^2$

$85)^2$

0094

Section: CS-8B

ahore Campus

urse Code:	CS4048
mester:	Spring 2023
tal Marks:	90
eight	40
age(s):	6

ise answers. Clearly write the provided. You can use extra rough work with the answer s. In case of any confusion or

6	Total
/16	/90

[20 points]

ou are tasked with investigating g on the UK film industry. Movies ng our perceptions. They can also ie genres and box office gross on , and employment. Following are

P, Crime rate, Hourly Pay, Health, 000 to 2022?

re review? [5 points] for both the movie industry and bility, and relevance to the research

cted data, addressing issues such as u would integrate the different data

to explore the data. Discuss your analyze the relationships between the

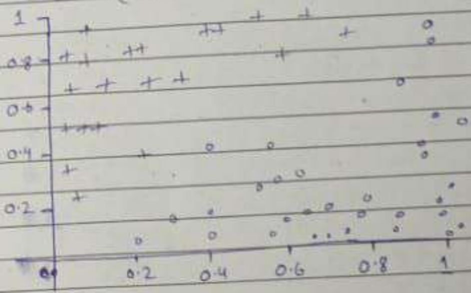
Q / Part No.

Rough Work

Q#3) 15

$y = mx + b$

(a) $0.2 = (1)x + b$



(b) On paper 10 points are going to be miscalssified which are mark on graph.

(c)

	Actual	
Prediction	1	0
1	true+ve	false +ve
0	false-ve	true-ve

Actual

Prediction ~~2~~	~~0~~1	0
1	49	16
0	0	35

8

$$Recall = \frac{true+ve}{true+ve+false-ve} = \frac{49}{49+0} = 1$$

Q / Part No. | Rough Work

$$\text{Precision} = \frac{\text{true +ve}}{\text{true+ve + false+ve}} = \frac{49}{49+16} = 0.753$$

2

$$\text{Accuracy} = \frac{2PR}{P+R} = \frac{2(0.753)(1)}{0.753+1} = 0.8598 \text{ or } 85.9\%$$

1

Q#4)
(a)

Time Spent (X_1)	No. of practice problem (X_2)
4	10
5	12
3	8
6	15
2	5
7	18
4	9
5	13

k = 2
Randomly initializing
$k_1 = (4, 10)$
$k_2 = (5, 12)$

$\sqrt{(3-4)^2 + (10-8)^2}$
$= 2.2$
$\sqrt{(3-5)^2 + (12-8)^2} = 4.4$
$\sqrt{(6-4)^2 + (15-10)^2} = 5.385$
$\sqrt{(6-5)^2 + (15-12)^2} = 3.16$
$\sqrt{(2-4)^2 + (5-10)^2} = 5.38$
$k_2 = 7.61$
$k_1 = 8.544, k_2 = 6.32$

Iteration 1

X_1	X_2	Cluster
4	10	k_1
5	12	k_2
3	8	k_1
6	15	k_2
2	5	k_1
7	18	k_2
4	9	k_1
5	13	k_2

Updating Cluster

$$k_1 = \left(\frac{4+3+2+4}{4}, \frac{10+8+5+9}{4}\right) = (3.25, 8$$

$$k_2 = \left(\frac{5+6+7+5}{4}, \frac{12+15+18+13}{4}\right) = (5.75, 14$$

Q / Part No.

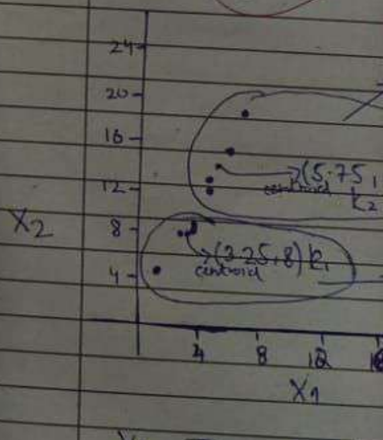
Iteration #2

X_1	X_2
4	10
5	12
3	8
6	15
2	5
7	18
4	9
5	13

Hence, No need
clusters

14

(b)



X_1 = Time spent
X_2 = No. of prac

Rough Work

$\frac{49}{49+16} = 0.753$

$1) = 0.8598$
or 85.9%

tice problem (X_2)

8

3

$\sqrt{(3-4)^2+(10-8)^2}$
$= 2.2$
$\sqrt{(3-5)^2+(12-8)^2} = 4.49$
$\sqrt{(6-4)^2+(15-10)^2} = 5.385$
$\sqrt{(6-5)^2+(15-12)^2} = 3.16$
$\sqrt{(2-4)^2+(5-10)^2} = 5.38$
$k_2 = 7.61$
$k_1 = 8.544, k_2 = 6.325$

er

$5+9\ ) = (3.25, 8)$

$+18+13) = (5.75, 14.5$
4

Q / Part No.

Rough Work

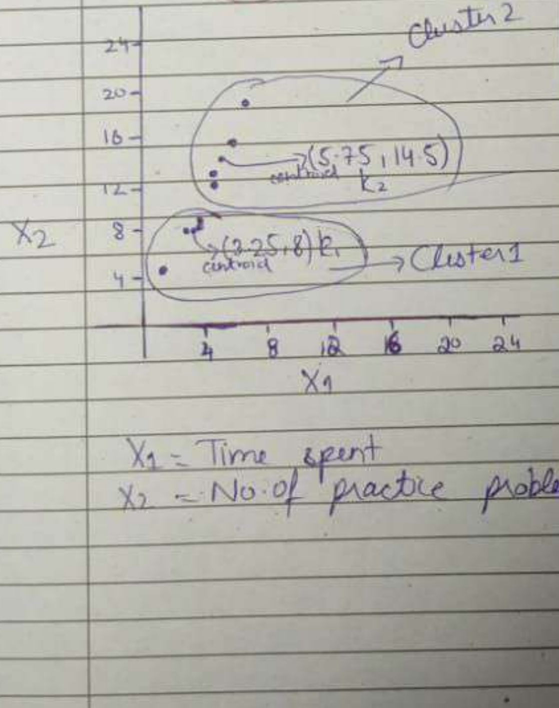
Iteration # 2 $k_1(3.25, 8)$ $k_2(5.75, 14.5)$

X_1	X_2	Cluster
4	10	k_1
5	12	k_2
3	8	k_1
6	15	k_2
2	5	k_1
7	18	k_2
4	9	k_1
5	13	k_2

Hence, No need to update clusters

14

(b)



X_1 = Time spent
X_2 = No. of practice problem

$\sqrt{(4-4)^2+(10-9)^2}$
$= 1$
$\sqrt{(4-5)^2+(12-9)^2}$
$(5, 13)$
$\sqrt{(5-4)^2+(13-10)^2}$
$\sqrt{1+9} = \sqrt{10}$
$= 3.162$
$\sqrt{}$

$\sqrt{(4-3.25)^2+(10-8)^2} = 2.136$

$\sqrt{(4-5.75)^2+(10-14.5)^2} = 4.8$
$\sqrt{(5-3.25)+(12-8}$
$k_2 = 2.61$
$(3, 8)$ $k_1 = 0.25$
$(6, 15)$ $k_2 = 0.$
$k_1 = 7$
$(2, 5)$ $k_2 = 1$
$k_1 = 3$
$(7, 18)$ $k_2 =$
k_1
$(4, 9)$ k_2
k_1
$(5, 13)$ k_2
k_1

Name: Mateeh Ullah

National University of Computer an

Course:	Data Science
Program:	BS(Computer S
Duration:	180 Minutes
Paper Date:	25-May-23
Section:	A & B
Exam:	FINAL

Instruction/Notes: Attempt the examination on the a
question number and your answe
sheet for rough work. Do not atta
sheets. Do not use pencil or red in
ambiguity, make a reasonable assu

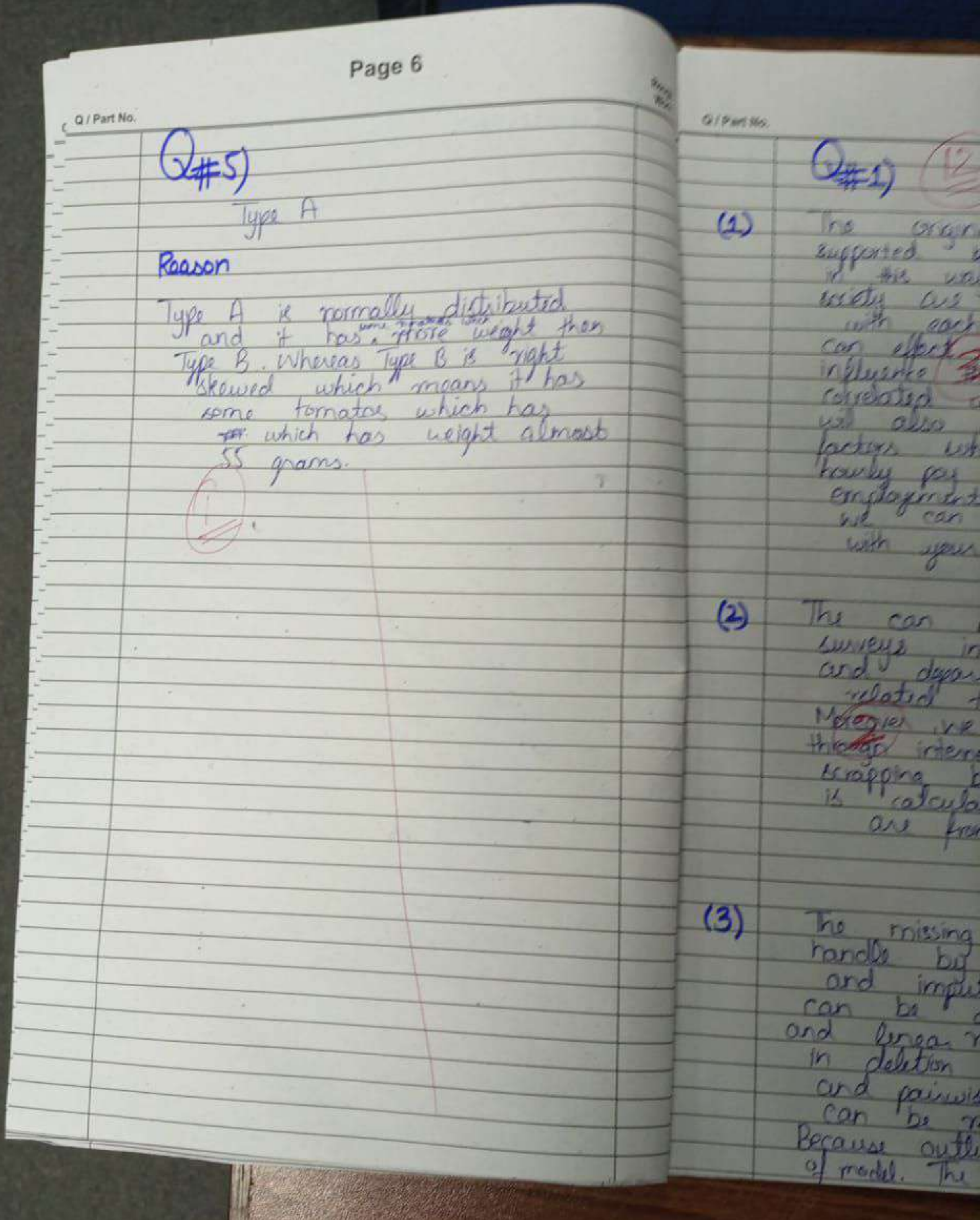
Question	1	2	3	
Marks	/20	/15	/19	

Question # 1: **Research pro**

You have been hired to work on a research project. In this r
the association between movies and socioeconomic factors, p
and films have a significant influence on society, reflecting our
impact socioeconomic factors. The research aims to explore th
socioeconomic factors such as GDP, crime rate, hourly pay, he
the research questions.

- What have been the Effects of having socio-Economic
Education, and Employment on movie development in
- Does Movie Genre affect any of the socioeconomic fact
- Does Box Office Gross affect any of the socioeconomic f

1. Explain how would you support the topic originality with y
2. Explain how you would identify and collect relevant da
socioeconomic factors in the UK. Consider the data availab
questions. **[5 points]**
3. Describe the steps you would take to clean and preprocess
missing values, outliers, and inconsistencies. Also mention
sources, ensuring data compatibility. **[5 points]**
4. Describe the techniques and visualizations you would u
methodology how you would approach research question 2 &
variables. **[5 points]**



Page 6

Q / Part No.

Q#5)

Type A

Reason

Type A is normally distributed and it has some tomatoes with ~~more~~ weight than Type B. Whereas Type B is right skewed which means it has some tomatoes which has ~~tor~~ which has weight almost 55 grams.

1

Q / Part No.

Q#1) 12

(1) The origin
supported
in the wa
society are
with each
can effect 3
influence
correlated
will also
factors wh
hourly pay
employment
we can
with your

(2) The can
surveys in
and depa
related
Moreover we 2
through intern
scrapping
is calcula
are from

(3) The missing
handle by
and impu
can be
and linea
in deletion
and pairwis
can be
Because outli
of model. The

Rough Work

distributed
e weight then
B is right
means it has
which has
weight almost

Q / Part No. | Rough Work

Q#1) 12

(1) The originality of topic is supported by literature review in this way that films and society are positively correlated with each other. Bad movie can effect ~~a~~ society. Because movies influence ~~the~~ societies As it correlated with sociteу so it will also correlated to socioeconomic factors which are GDP, crime rate, hourly pay, health, education, and employment. Thus, in this way we can support the topic origniality with your literature review

2

(2) The can be collected using surveys in society, film industry, and department which are related to socio economic factors. Moreover, we can collect data through internet resources by using scrapping but the data which is calculated through intrnet are from trusted sources

2

(3) The missing values can be handle by using partial deletion and imputation. The imputation can be done using Mean and linear regression method whereas in deletion we will use Listwise and pairwise deletion. The outliers can be remove or impute. Because outlier can effect accuracy of model. The inconsistencies can also be

Name: [redacted] Mc

National Univ

C
P
D
P
S
E

Instruction/Notes:

Question	1
Marks	/2

Question # 1:
You have been hired
the association betwe
and films have a signif
impact socioeconomi
socioeconomic factor
the research questior

- What have b
Education, ar
- Does Movie
- Does Box Off

1. Explain how w
2. Explain how
socioeconomi
questions. [5
3. Describe the
missing value
sources, ens
4. Describe th
methodolog
variables. [5

Department of

Q / Part No.

handle by removing we can use quartile or z-score for it. ~~the data from diff~~

3 The data from different sources can be integrating by doing web scrapping.

(3) ~~Firstly, we will analyze data by creating confusion matrix. Also we will find covariance and correlation b/w features~~

(4) Firstly, we will handle missing valubs. then we will analyze data by creating confusion matrix. scatter plot, Box plox and histogram. These methodologies will be use to transform data. We will also find covariance and correlation b/w feature to identify their relationship in research question 2 & 3.

Q / Part No.

10

Q#6

(a) H(Edible |

H(Odor) (

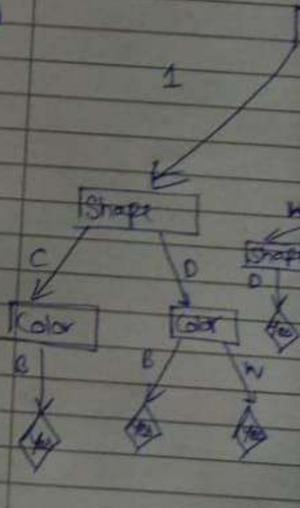
H(odor=1) [

H(odor=3) (

So entropy be 0.

(b) Odor wi node by

(c)



Q / Part No. Rough Work

10

Q#6)

(a) H(Edible | Odor = 1 or odor = 3)

$$H(\text{odor})(5+,6-) = -\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11}$$

$$= -0.4545(-1.1375) - 0.5454(-0.8744)$$

$$= 0.9938$$

H(odor = 1) [3+, 0-] = 0

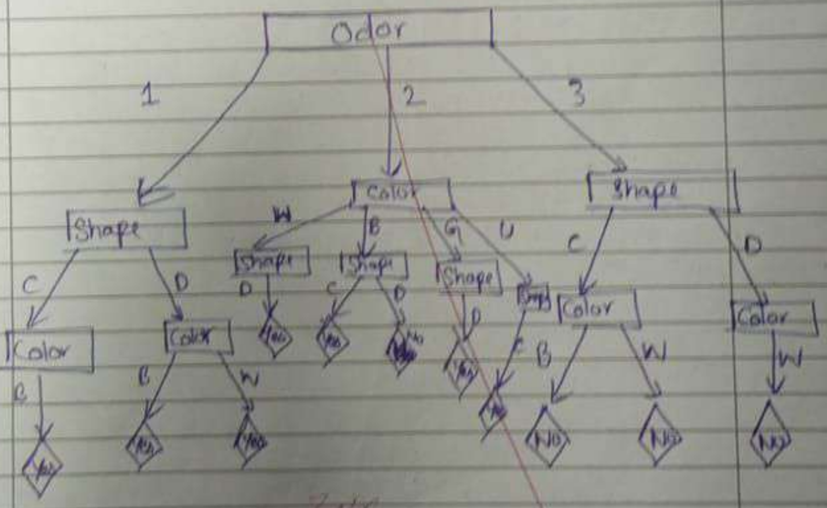
H(odor = 3) [0+, 3-] = 0

So entropy of Odor 1 or odor = 3 will be 0.

(b) Odor will be choose as next node by ID[3]. (Rough Work at end)

7

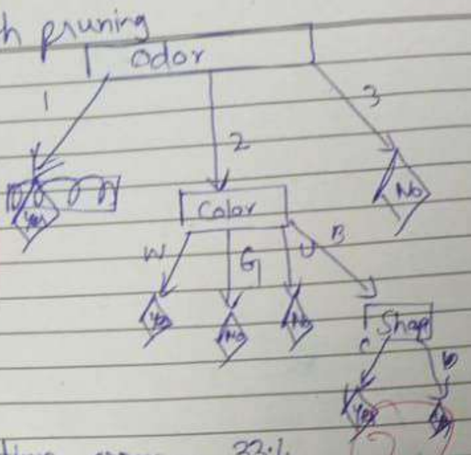
(c)



Page 10

Q / Part No. OR (with pruning

Rough Work



4

2

(d) Validation error = 33%

~~Whereas, miscalassified~~

training set error = ~~18.18%~~ 9.09%

Whereas misclassified are

Shape	Color	Odor	Edible
C	B	2	Yes