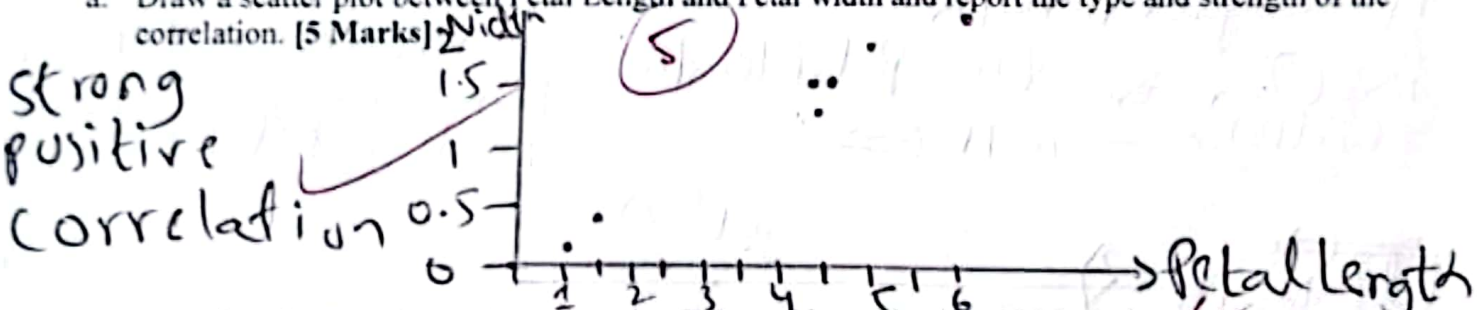
Course:	Introduction to Data Science	CourseCode:	DS 2001
Program:	BS(DS)	Semester:	Fall 2023
Duration:	1 Hour	Total Marks:	60
Paper Date:	10-11-2023	Page(s):	6
Section:	BS (DS) A, B, C	Section:	DS 3A
Exam:	Mid II	Roll No:	22L-7482

Instructions: Answer in the space provided. **You can ask for rough sheets, but they will not be graded or marked.** In case of confusion or ambiguity make a reasonable assumption. Questions during exam are not allowed.

Problem 1: Answer the following question related to the data sample given above.

Sr. No	Sepal Length	Sepal Width	Petal Length	Petal Width
1	6.3	2.3	4.4	1.3
2	6.8	3.2	5.9	2.3
3	5.6	2.7	4.2	1.3
4	5.1	3.8	1.5	0.3
5	5.9	3	4.2	1.5
6	6.3	2.7	4.9	1.8
7	6.2	2.2	4.5	1.5
8	4.3	3	1.1	0.1
Σ	46.5	22.9	30.7	10.1
μ	5.81	2.86	3.84	8.42

a. Draw a scatter plot between Petal Length and Petal width and report the type and strength of the correlation. **[5 Marks]**



b. Normalize the Sepal Length feature using Z Score Normalization. **[5 Marks]**

std = 0.79

Sr No			
1	0.620	6	0.620
2	1.253	7	0.496
3	−0.266	8	−1.911
4	~~0.114~~ −0.891		
5	0.114		

0006

c. Compute slope and intercept of a regression line between Sepal Length and Petal Length features. Take Sepal Length as independent feature and Petal Length as dependent feature. **[10 Marks]**

Sum of squares (SS_X) = 4.4488
Sum of products (SP_{XY}) = 8.8463

X /10 Y

$$\text{Slope} = b = \frac{SS_{xy}}{SS_{xx}} = \frac{8.8463}{4.4488}$$

$$b = 1.9884$$

$$\bar{Y} = b\bar{x} + a$$

$$a = \bar{Y} - b\bar{x}$$

$$a = 3.84 - (1.9884 \times 5.81)$$

$$a = -7.71 \rightarrow \text{intercept}$$

$$Y = 1.9884 \times 8 - 7.7$$

d. Interpret the values of slope and intercept. **[5 Marks]** /5

intercept

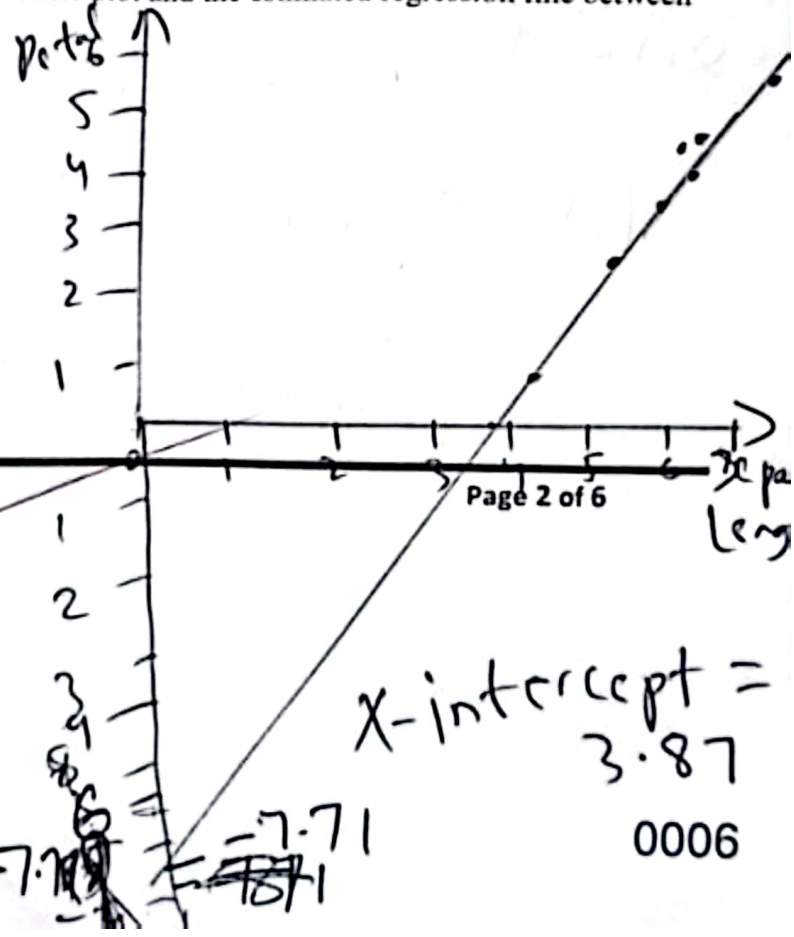
if sepal length is 0, the petal length will be −7.71

slope

for every change in x, there is 1.9884 times change in y. that is petal length

e. Using computed slope and intercept draw the scatter plot and the estimated regression line between the two features. **[5 Marks]** /5

X	Y
4.8	5
5.8	6
3.4	7
2.4	8
	4.08
	4.8
	4.6
	0.84



Problem 2: Answer the following question related to the data sample given in Question 1.

a. Using Sepal'Length and Sepal Width features perform K-Means clustering and divide group the data into 2 cluster. Take Data point 2 & 8 as initial seed values. Perform one Iteration and fill the table below. **[10 Marks]** /10

	Iteration 1	
Data Point	Distance (Seed1)	Distance (Seed2)
1	$\sqrt{1.06}$	$\sqrt{4.49}$
2	0	$\sqrt{6.29}$
3	$\sqrt{1.69}$	$\sqrt{1.78}$
4	$\sqrt{3.25}$	$\sqrt{1.28}$
5	$\sqrt{0.85}$	$\sqrt{2.56}$
6	$\sqrt{0.5}$	$\sqrt{4.09}$
7	$\sqrt{1.36}$	$\sqrt{4.25}$
8	$\sqrt{6.29}$	0
	Cluster1 Data Points= 1, 2, 3, 5, 6, 7 Cluster2 Data Points = 4, 8 Cluster1 Center = 6.18, 2.68 Cluster 2 Center = 4.7, 3.4	

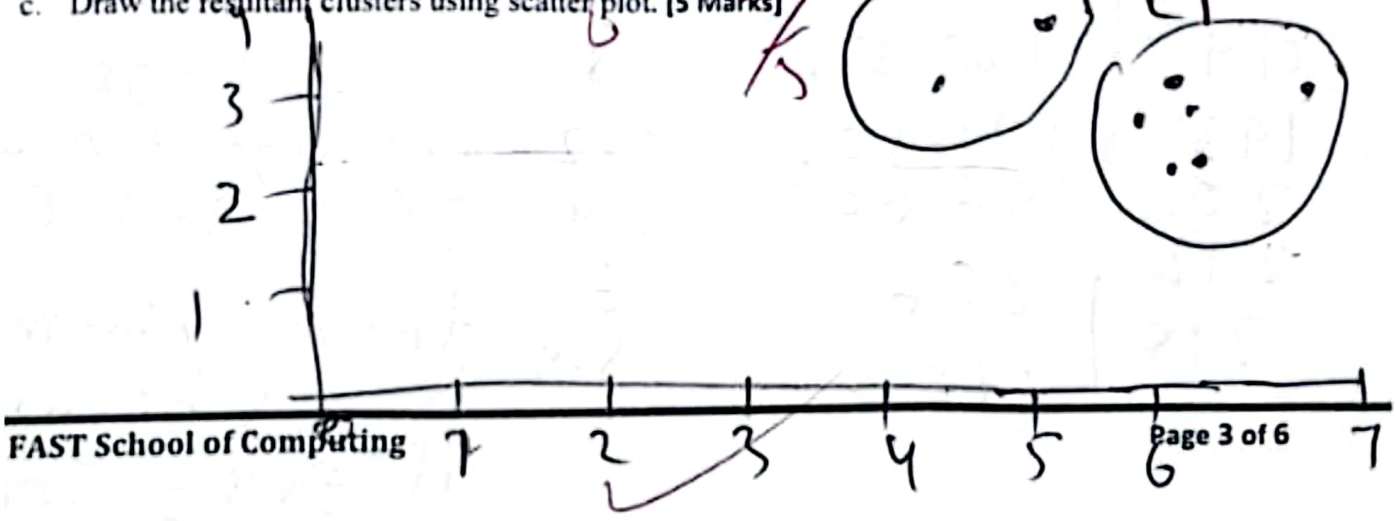
b. How would you classify a new data point (5,2) into one of the clusters created in part (a). Show your working. **[5 Marks]** /3

(5,2)

d S1	d S2
1.36	1.43

classify(5,2) in cluster 1

c. Draw the resultant clusters using scatter plot. **[5 Marks]** /5



0006

Problem 3: Given the following dataset, your job is to train a multilinear regression model that determines the impact of salary and age on the house size. After training the model, the following relationship is identified: /10

Predicted House Size = -6.867 + 3.148Salary – 1.656Age

Salary (x1)	Age (x2)	House size
60	22	140
62	25	155
67	24	159
70	20	179
71	15	192
72	14	200
75	14	212
78	11	215

a. Using the information given, determine the quality of the machine learning model trained on the given dataset. **[10 Marks]**

House size = Y

Y	$\hat{Y}$	$\lvert Y-\hat{Y}\rvert$
140	145.581	5.581
155	146.909	8.1
159	164.305	5.3
179	180.373	1.373
192	191.801	0.2
200	196.605	3.4
212	206	6
215	220.5	5.5

$$MSE = \frac{\sum (Y-\hat{Y})^2}{8} = 25.6$$

$RMSE = 5.06$

Considering the range of house size in 100s, the machine model is good because it gives ± 5.06 error

0006