Data Analytics
END SEMESTER EXAM - Spring 2021

Instructions:

1) This is a closed book, closed notes exam.
2) You should not discuss questions or answers with anyone (including outsiders)
3) You should have your **camera ON** at at all times and no headphones
4) Consists of Part-A and Part-B. Part-A consist of 4 descriptive questions (40 Marks). and Part-B consist of 10 MCQ questions (10 Marks)
5) For descriptive questions, write down your answers in A4 sheet. And be brief and to-the-point. Answers must be given in **ball point pen** only. Answers in pencils will not be checked.
6) You are allowed to use calculators. The required statistical tables are attached along with the question paper. So don't make any excuses in the middle of the examination.
7) You should submit the scanned copy of your answer sheet in google classroom.
8) The name of the scanned copy should be the Roll No_Set No.pdf. (e.g.,S20170010XYZ_SetB.pdf).
9) Write the name and the roll no. on each page of the answer sheets. If name or roll no. is missing, the paper won't be evaluated.
10) Follow all other instructions given by the faculty during the exam. Attempt all questions
11) Submit the answers in the given time. Penalties will be imposed for late submission.

Data Analytics
Descriptive Questions
END SEMESTER EXAM - Spring 2021

Duration: 1 hour 20 minutes
Total Marks: 40

SET – D

Question 1:

The average monthly electric power consumption (Y) at a certain manufacturing plant is considered to be linearly dependent on the average ambient temperature (x). Consider the 15 months data given in Table 1.

Table 1. Average monthly power consumption, Y (in thousands of kwh) and average ambient temperature, x, (in F)

x	82	73	95	66	84	89	51	82	75	90	60	81	34	49	87
Y	76	83	89	76	79	73	62	89	77	85	48	69	51	25	74

a) Obtain the simple linear regression analysis to predict the monthly electric power consumption (Y) from the average ambient temperature, x. **[6 Marks]**
b) It is suggested that if the regression is significant, then there is no need to measure electric power consumption in future. How you test the significance level of your regression analysis? **[4 Marks]**

Question 2:

Cognitive load is measured as low (L), Medium (M), High (H) and Very High (VH). A survey is conducted while playing a video game among a population of different age groups and cognitive load observed are recorded in Table 2.

Table 2. Cognitive Load (CL) and Age group (AG)

AG	90-100	80-90	10-20	20-30	30-40	40-50	50-60	60-70	70-80
CL	H	VH	VH	VH	M	L	L	M	H

a) Apply a suitable correlation analysis to check if there is any correlation exists between Cognitive load (CL) and Age group (AG). **[2 Marks]**
b) Calculate the coefficient of determination and interpret your result. **[8 Marks]**

Question 3:

a) Two documents (X and Y) are given with the frequency count of 10 words in each document. Calculate the similarity measure between X and Y. Also, mention the metric used. **[2 Marks]**

X	3	2	0	5	0	0	0	2	0	0
Y	1	0	0	0	0	0	0	1	0	2

b) Calculate the following for the classifier which is tested with a test set of size 100 and predicts 80 test tuples correctly. **[3 Marks]**

i) Observed frequency

ii) Standard error rate

iii) True accuracy. Assume T_α with confidence level α = 95% is 1.96.

c) Plot the ROC curve and clearly show the location of (i) ideal, (ii) worst (iii) ultra-liberal (iv)ultra-conservative and (v) random classifier in it. **[2 Marks]**

d) Consider the following confusion matrix. **[3 Marks]**

	Class A	**Class B**
Class A	80	25
Class B	15	70

Calculate the following clearly mentioning the formula of each metric.

i) Precision

ii) Recall

iii) Sensitivity

Question 4:

Consider the following data

Length	**Width**
5.76	3.31
5.55	3.33
5.29	3.34
5.32	3.37
5.65	3.56
5.38	3.31
6.19	3.56
5.99	3.48
6.15	3.93

a) Cluster the data with k = 3. Show your result with first three iterations. You should produce results in the tabular forms. Clearly mention the similarity measure you have followed in your working. **[6 Marks]**

b) Mention at least three situations when the k-means clustering fails to give good result. You should mention each situation clearly and explain why k-means algorithm fails. **[4 Marks]**

------------------------All the best--------------------------