Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2019-2020
M.Tech (Data Science and Engineering)
Mid-Semester Test (EC-2 Make-up)

Course No.	: DSECLZC415
Course Title	: DATA MINING
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 90 Minutes
Date of Exam	: 04/01/2020 (AN)

No. of Pages = 2
No. of Questions = 4

Note:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Answer All the Questions (only in the pages mentioned against questions. if you need more pages, continue remaining answers from page 21 onwards)

Question 1: [4 + 3 = 7 marks] **[to be answered only in pages 2-6]**

a) Based on the information given in the table below, find most similar and most dissimilar persons among them. Apply min-max normalization on income to obtain [0,1] range. Consider profession and mother tongue as nominal. Consider native place as ordinal variable with ranking order of [Village, Small Town, Suburban, Metropolitan]. Give equal weight to each attribute.

Name	**Income**	**Profession**	**Mother tongue**	**Native Place**
Ram	70000	Doctor	Bengali	Village
Balram	50000	Data Scientist	Hindi	Small Town
Bharat	60000	Carpenter	Hindi	Suburban
Kishan	80000	Doctor	Bhojpuri	Metropolitan

b) How Classification, Association and Clustering can help bank? Briefly explain with help of examples.

Question 2: [3+3+2= 8 Marks] **[to be answered only in pages 7-11]**

a) Below is the dataset which is divided into 3 bins as follows:

Bin 1	5,10,11,13
Bin 2	15,35,50,55
Bin 3	7,29,22,04,215

Demonstrate how you will smooth the data using Bin Mean, Bin Median and Bin Boundary.

b) Below are the marks scored by a student in 2 subjects. In which subject the student has performed better? Justify your answer.

Subject	**Marks**	**Mean**	**Standard Deviation**
S1	70	60	15
S2	65	60	6

c) How does positively skewed data different from negatively skewed data in terms of central tendency?

Question 3: [3+5= 8 Marks] **[to be answered only in pages 12-16]**

a) In a manufacturing plant there are several conditions like abnormal temperature, pressure, humidity, electricity supply, machine failure, labor shortage etc which can impact the production. The management is interested to know what situation actually impact the production and hires data scientists to get a classifier prepared which predicts if the conditions will impact the production (Yes) or not (No). When the classifier was run on the test data, the following confusion matrix was obtained. Comment on the performance of the classifier using appropriate metric(s) to meet the management's objective.

Actual Class	**Predicted Class**	
	Yes	No
Yes	50	115
No	72	5000

b) The sales of a company (in million dollars) for each year are shown in the table below. **[4+1]**

x (year)	2005	2006	2007	2008	2009
Y (Sales)	12	19	29	37	45

1) Find the least square regression line y = a x + b.
2) Use the least squares regression line as a model to estimate the sales of the company in 2012.

Question 4: [3+2+2 = 7 Marks] **[to be answered only in pages 17-20]**

A shop on a railway platform captured the following five transactions. Answer the following questions, given that support threshold is 40% and confidence threshold is 80%:

T1:	Binoculars ,Umbrella, Juice
T2:	Binoculars, Umbrella, Juice, Snacks
T3:	Umbrella, Juice
T4:	Binoculars, Juice, Snacks
T5:	Snacks

1) Find all frequent itemsets using Apriori Algorithm.
2) Find all Closed Frequent and Maximal Frequent itemsets.
3) If two association rules mined from the given dataset are (i) {Juice, Snacks} → {Binoculars} (ii) {Binoculars, Juice} → {Umbrella}. Which one will you select and why?