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

Course No.	: DSECLZC415
Course Title	: Data Mining
Nature of Exam	: Open Book
Weightage	: 30%
Duration	: 90 minutes
Date of Exam	: 21/06/2020 (AN), 2:00 pm to 3:30 pm

No. of Pages = 3
No. of Questions = 4

Note to Students:

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.

Q.1. You have been given a task to perform the data preprocessing of the data retrieved from multiple sources, before you start applying the data mining task. Identify, (atleast 5) data quality issues with the sample data set retrieved from the master data set. Suggest, how do you resolve these quality issues (python code is not required)? **[5]**

TXN-ID	NAME	AGE	HEIGHT	WEIGHT	BLOOD GROUP	COVID-19 RESULT
T001	RAMA	45	145	62kg	O+ve	Positive
T002	SEETHA	43	168	45kg	B+ve	Negative
T003	Akbar	38	172	60kg	Iam+ve	Positive
T004	BIRBAL	45	168	52kg	AB+ve	Negative
T005	THenali	22	157	78kg	B-ve	1
T006	Venkat	36	157	54kg	O-ve	Negative
T007	Rajuu	350	132	48kg	O+ve	Positive
T008	HARI	32	180	120lbs	AB-ve	Negative
T009	Inba	25		85kg	O+ve	0
T010	SysUsr789	20	165	68kg	O-ve	Negative

Q.2. Answer the following:

a) Find Minkowski distance of order = 3, between two objects represented by the coordinates (12, 36, 42, 20) and (17,35, 43, 26). **[2]**

b) Apply equi-width and equi-depth binning method on following dataset to create sets of 3 bins. [23, 8, 2, 20, 11, 1, 29, 30, 21] **[3]**

c) Suppose the stock closing price of two companies A and B are as follows **[4]**

Company	Mon	Tues	Wed	Thurs	Fri
A	100	110	105	100	95
B	150	148	155	155	100

What inference can you draw from this dataset on the dependency of stock prices in companies A and B? Explain how you arrive at this.

Q.3. Answer the following: **[5+3+3]**

a) Consider the following training data set (with three attributes, such as Past trend, Open Interest, Trading Volume and class/target variable is "return") for a binary class problem. The attributes are nominal with two possible values. We intend to create a decision tree model using Information Gain. Which attribute would the decision tree induction algorithm choose for the root node?

Past Trend	Open Interest	Trading Volume	Return
Positive	Low	High	Up
Negative	High	Low	Down
Positive	Low	High	Up
Positive	High	High	Up
Negative	Low	High	Down
Positive	Low	Low	Down
Negative	High	High	Down
Negative	Low	High	Down
Positive	Low	Low	Down
Positive	High	High	Up

b) Suppose a training set consists of 100 positive examples and 100 negative examples for each of the rules R1 and R2. Considering R1 covers 30 positive examples and 90 negative examples; R2 covers 60 positive examples and 60 negative examples, Identify which rule is better, using FOIL gain.

c) Consider the confusion matrices for two models M1 and M2 are given below. Evaluate the performance of models using F1-score, and Identify which one is better.

M1:

Predicted Class → Actual Class ↓	Positive	Negative	
Positive	1800	200	2000
Negative	400	1600	2000
	2200	1800	4000

M2:

Predicted Class → Actual Class ↓	Positive	Negative	
Positive	1600	400	2000
Negative	800	1200	2000
	2400	1600	4000

Q.4. Answer the following: **[1+4]**

a) Consider the following data describing three customers (A, B, C) and their preferences for four products P1, P2, P3, P4 where "1" indicates the customer prefers that product. Which similarity measure is appropriate in this case to measure the similarity between customers?

b) Identify the customer pairs that are more similar with respect to the rest of them by computing the similarity measure.

	P1	**P2**	**P3**	**P4**
A	0	1	1	0
B	1	1	0	0
C	1	1	0	1