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 Regular)

Course No.	: DSECF ZC415
Course Title	: DATA MINING
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 90 Minutes
Date of Exam	: 23/06/2019 (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 20 onwards)

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

Q.1 (a) Identify which of the following activities is a data mining activity. Justify your answer.

1) Computing the total sales of a company.
2) Predicting the outcomes of tossing a (fair) pair of dice.
3) Predicting the future stock price of a company using historical records.
4) Dividing the customers of a company according to their demographics.

Q.1 (b) Consider that following employee data set made available to you for carrying out some data mining activity. What are the four potential issues with this dataset?

Name	Age	DateOfJoining	Designation	DateOfBirth
A	34	15-Jan-2015	Sr Engineer	Feb 24, 1981
B	33	27-Jan-2015		Mar 27, 1982
A	34	15-Jan-2015	Sr Engineer	Feb 24, 1981
C	32	30-Jan-2015	Staff Engineer	Nov 25,1982

Q.1 (c) Statistical inference may indirectly facilitate the data preparation phase while pre-processing. In light to this how the following inference can be used?
"A variable X is positively & uniformly correlated with another variable Y"

Question 2: [2+1+1+4 = 8 Marks] **[to be answered only in pages 5-9]**

Q.2. Consider the following ordered list of observations of a variable. Answer the following:
25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 41, 42, 42, 99

(a) What is the five-number summary for the given data?
(b) Draw boxplot.
(c) Identify the outliers if any.
(d) Explain, how do the outliers affect the measures of central tendency (Mean, and Median) of data? Comment using the given data set.

Question 3: [5+3 = 8 Marks] **[to be answered only in pages 10-13]**

Q.3 (a) The following dataset describes the data received on 14 patients and their status on diabetes. Considering decision tree classifier to this binary classification problem, identify the attribute selection among Exercise, Blood pressure level at the root node using Entropy and information gain computation.

Sl#	Exercise	Blood pressure level	Follow good diet ?	Class
1	Yes	High	No	- ve
2	Yes	High	Yes	+ve
3	No	High	No	+
4	Moderate	High	No	+
5	Moderate	Normal	No	+
6	Moderate	Normal	Yes	-
7	No	Normal	Yes	+
8	Yes	High	No	-
9	Yes	Normal	No	+
10	Moderate	Normal	No	+
11	Yes	Normal	Yes	+
12	No	High	Yes	+
13	No	Normal	No	+
14	Moderate	High	Yes	-

Q.3 (b) Following table shows results of classification for a 2-class problem. Consider ‘Y’ and ‘N’ as two classes. Calculate the F-score of the classifier for class ‘Y’.

Predicted Class ➡ Actual Class ⬇	Y	N	
Y	900	100	1000
N	200	800	1000
	1100	900	2000

Question 4: [3+3+2 = 8 Marks] **[to be answered only in pages 14-17]**

Q.4. Pantaloons wants to identify possible cases of bundle pricing. Consider the following dataset with five transactions and the associated list of items. Answer the following questions using Apriori algorithm:

Transaction ID	Itemset
1	SHIRT, SHOES, BELT, SOCKS, WATCH, SHOEPOLISH
2	TROUSERS, SHOES, BELT, SOCKS, WATCH, SHOEPOLISH
3	SHIRT, JEANS, SHOEPOLISH, SOCKS
4	SHIRT, TIE, SANDAL, SHOEPOLISH, WATCH
5	SANDAL, SHOES, SHOEPOLISH, SOCKS

(a) Given the minimum support count of 3, Generating all the frequent itemsets.

(b) Considering the minimum confidence threshold as 75%, Identify the Association Rules from the frequent itemsets generated in (a) above.

(c) Identify the products which are suitable for bundle pricing based on the final association rules, using the min support and min confidence values as provided above.