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.	: DSECLZC415
Course Title	: DATA MINING
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 90 Minutes
Date of Exam	: 22/12/2019 (AN)

No. of Pages = 3
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: [2 + 4 = 6 marks] **[to be answered only in pages 2-6]**

a) You are the chief selector of the Indian cricket team and you are tasked with selecting the best all-rounder for the Indian world cup squad. Below is the list of all-rounders who are available for selection and their respective batting, bowling and fielding stats

Player	Batting Average	Bowling Average	Catches/runouts per match
Hardik Pandya	29	40	0.3
Kedar Jadhav	43	37	0.25
Ravindra Jadeja	31	36	1.2
Stuart Binny	29	22	0.1
Vijay Shankar	32	53	0.15

1) Identify the player who is an outlier on a given criterion. Explain How did you decide the outlier.
2) Below are Kapil Dev's stats (fielding stats not available). Using Manhattan distance find which all-rounder is most "Kapil-like". Does the result change if you use cosine similarity instead of Manhattan distance?

Player	Batting Average	Bowling Average
Kapil Dev	24	27

Question 2: [2+4 +2= 8 Marks] **[to be answered only in pages 7-11]**

a) The given dataset has information on some loans disbursed by a Peer to Peer lending platform. Refer below description of the columns. **[2+4]**
Tot Loan Repayments – Total number of loans previously paid by the borrower on the platform
Credit Score – Credit score of the borrower of the loan
State – The state in which the loan was issued

Monthly EMI – Monthly EMI paid by the borrower
Monthly Income – Monthly income of the borrower
Liquidity Ratio – Monthly EMI/ Monthly Income
Defaulted – If the borrower defaulted on the loan (This is the dependent variable)

Tot Loan Repayments	Credit Score	State	Monthly EMI	Monthly Income	Liquidity Ratio	Defaulted
2	99	AP	13000	27000	High	N
1	580	GUJ	29000	32000	Very High	N
0	710	AP	3000	22000	Low	N
0	660	MAH	68000	20000	Low	Y
0	580	AP	32000	40000	Very High	Y
3	720	MAH	23000	25000	Very High	N
0	680	AP	15000	32000	High	Y
1	600	GUJ	16000	20000	Very High	N
0	700	GUJ	9000	45000	0.2	N
2	720	MAH	4000	34000	Low	N
0	650		23000	33000	High	Y

1) Categorize attributes as Nominal, Ordinal and Numeric.
2) In the context of data pre-processing, Identify and explain four opportunities of problematic data which will require cleaning.

b) Apply equi-width binning method on following dataset with three number of bins. **[2]**
[24, 0, 6, 60, 63, 30, 87, 90, 87]
Also, List the disadvantages of equi-width binning.

Question 3: [6+2 = 8 Marks] **[to be answered only in pages 12-16]**

a) After the parliament passed a bill on stringent traffic regulation, the following data was captured on a busy and representative traffic signal for a specific period. Consider Crash Severity as the class of interest and use multiway split for the discrete-valued attributes.

Weather Condition	Driver Condition	Rule Violation	Seat Belt?	Crash Severity
Good	Alcohol	Speed	No	Major
Bad	Sober	None	Yes	Minor
Good	Sober	Red Signal	Yes	Minor
Good	Sober	Speed	Yes	Major
Bad	Sober	Other Rules	No	Major
Good	Alcohol	Red Signal	Yes	Minor
Bad	Alcohol	None	Yes	Major
Good	Sober	Other Rules	Yes	Major
Good	Alcohol	None	No	Major
Bad	Sober	Other Rules	No	Major
Good	Alcohol	Speed	Yes	Major
Bad	Sober	Red Signal	Yes	Minor

Using ID3 algorithm, find out: **[2+4]**

1) The expected information needed to classify a tuple in the data
2) The first attribute that will be used for the splitting.

b) Why is tree pruning useful in decision tree induction? **[2]**

Question 4: [2+3+3 = 8 Marks] **[to be answered only in pages 17-20]**

BigBasket saw the following transactions from its customers and based on it they wish to identify possible cases of bundle pricing.

Transaction ID	Itemset
1	Apple, Banana, Basil, Kiwi, Watermelon, Orange
2	Grapes, Bananas, Basil, Kiwi, Watermelon, Orange
3	Apple, Jackfruit, Orange, Kiwi
4	Apple, Tiramisu, Pears, Orange, Watermelon
5	Pears, Bananas, Orange, Kiwi

Using Apriori algorithm

1) Explain, how apriori property helps in optimizing the computation efforts while mining for frequent itemsets?
2) Given the minimum support of 3, apply Apriori algorithm for generating all the frequent itemsets.
3) Considering the minimum confidence threshold as 80%, Identify the Association Rules from the frequent itemsets generated in 2) above.