Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2019-2020

Comprehensive Examination (Makeup)

Course No.	: PCAM* ZC111
Course Title	: FEATURE ENGINEERING
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 3 Hours
Date of Exam	: ~~Friday, 20/09/2019 (AN)~~

No. of Pages =**4**
No. of Questions = **5**

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.

Q.1. Consider the following dataset for answering the following feature subset selection related questions. **[2 + 3 = 5]**

Feature1 (F1)	Feature2(F2)	Feature3(F3)
A	C	X
A		Y
	C	Z
B	C	Q

a) What the two stopping criteria's that are most commonly used in SFS algorithm.
b) Draw a lattice structure that will represent the different models that will be generated when the above mentioned dataset is used.

Q.2. The given dataset has information on some loans disbursed by a Peer to Peer lending platform. Refer below description of the columns. **[2 + 4 = 6]**

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)

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.

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

Q3. Apply equi-width binning method on following dataset with three number of bins. **[4]**
[24, 0, 6, 60, 63, 30, 87, 90, 87]
Also, List the disadvantages of equi-width binning.

Q.4. 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. Using Information Gain (with Gini Index method), find out the first attribute that will be used for the splitting. **[10]**

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

Q.5 Consider the following dataset consisting of three rows and three dimensions. Using the steps mentioned in the PCA algorithm, obtain the characteristics equation which can be used for the getting the eigen values and vectors. **[5]**

1	2	3
2	1	2
5	3	3

****