Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2019-2020
Comprehensive Examination (Regular)

Course No.	: PCAM* ZC111
Course Title	: FEATURE ENGINEERING
Nature of Exam	: Closed Book
Weightage	: 40%
Duration	: 3 Hours
Date of Exam	: 02/11/2019 (AN)

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

Q1. **[2+2+2= 6 M]**

A. Suppose all students in the first cohort are friends on Facebook and the instructor wants to reach out to all students about a notice. To complete this task, the instructor sends the message to the most influential person (connected to many students of first cohort) so that he/she can propagate the message to all students. In this application what type of data would you use and how will you create the dataset?

B. Indigo is screening applications for its pilot positions and the dataset consist of height (in Centimeters) and Weight(in Kilograms) of applicants. The applicant is suitable if his/her BMI is between 21-24.
i. What concepts learnt from the data preprocessing you will use?
ii. How will you convert the given continuous values into a binary feature meaningful to the application?

C. You are given a dataset for classifying songs into classical and non-classical. The size of the data is given in Table-1. Out of 5000 samples there are 100 samples of classical songs (class label 0) and remaining are nonclassical (class label 1). You are asked to sample training data for building a classification model. What sort of sampling technique you will use and justify how your approach will help in reducing the training error?

Songid	**F1**	**F2**	**…….**	**F10**	**Class Label**
1					0
2					1
. .					
5000					1

Table-1

Q2. **[6+8=14M]**

A. A data set is described using 1000 features. The labels have been generated using the first 50 features. Another 50 features are exact copies of these features. The 900 remaining features are uninformative. Assume we have 100000 training instances/examples.
i) How many features will a filter based subset selection approach select and why?
ii) How many features will a wrapper based subset selection approach select and why?

B. A sample of 1000 students have been surveyed on Women's reservation. The respondents have been classified using gender (male or female) and by opinion (reservation for women, No reservation, or No opinion).Table - 2 shows the result of the survey:

	Opinion on Women's Reservation			Row Total
	Yes	No	Can't Say	
Male	200	150	50	400
Female	250	300	50	600
Column Total	450	450	100	1000

Table - 2

You are now asked to find whether the men's opinion on women's reservation differ significantly from that of women using chi2 test with 0.05 level of significance. From the standard chi2 test table the P-value for the chi square value you compute is 0.0003. Answer questions i-iv.

- **i.** State your hypothesis.
- **ii.** Compute the expected frequency (under the null hypothesis)
- **iii.** Select the test statistic
- **iv.** Make your conclusion based on the value.

Q3. **[3+3+2+2=10 M]**

A. Explain the concept of curse of dimensionality.

B. Principal component analysis (PCA) is a technique that is widely used for applications where dimensionality reduction is required. Answer questions i-iii related to PCA

i. Write the objective function of deriving PCA using Maximum Variance formulation.

ii. What are the constraints on the Principal Components?

iii. Suppose you are given the Eigen values and its corresponding Eigen Vectors. Using these Eigen values and their corresponding vectors can we get back our original data? If yes how, If No justify.

Q4. **[5+5=10 M]**

A. Describe 5 differences between Histogram and Box Plot

B. Explain 5 differences between t-SNE and PCA plots.

********** *That's all folks* **********