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	: 40%
Duration	: 3 Hours
Date of Exam	: 09/11/2019 (AN)

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

Q1. **[2+5 = 7 M]**

A. What is the purpose of aggregation when preprocessing raw data to obtain a dataset? Why is it, sometimes, desirable to aggregate sets of attributes/ objects into a single attribute/object?

B. 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 2500 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. **[7+5+4=16 M]**

A. Assume you have 100 unique values, and use equal width binning with 10 bins.

(i) What is the largest number of records that could appear in one bin?

(ii) What is the smallest number of records that could appear in one bin?

(iii) If you use equal frequency binning with 10 bins, what is largest number of records that can appear in one bin?

(iv) If you use equal frequency binning with 10 bins, what is smallest number of records that can appear in one bin?

(v) Now assume that the maximum value frequency is 20. What is the largest number of records that could appear in one bin with equal width binning (10 bins)?

(vi) What about with equal height binning (10 bins)?

B. An oceanographer wants to compare the whales found in the Indian and the Pacific Ocean. The following attributes have been captured for 20,000 whales: No of teeth, Length of fin, Width of Fin, Lifespan in years, Length in meters, Mass in Kilograms and Gestation period in months. Based on these measurements, do you think that any preprocessing techniques need to be applied on these features and what sort of similarity measures would you use to compare these whales? Justify your answer and explain any special circumstances.

C. Explain wrapper based approach using Stepwise backward feature selection. What is the major disadvantage of this approach? When do you suggest to use this approach despite the disadvantage?

Q3.

A. Explain how Principal Component Analysis (PCA) can be used for eliminating noise. **[3 M]**

B. Given the following data shown in **Figure-1**, answer questions **i-iv** **[2+2+1+2 = 7M]**

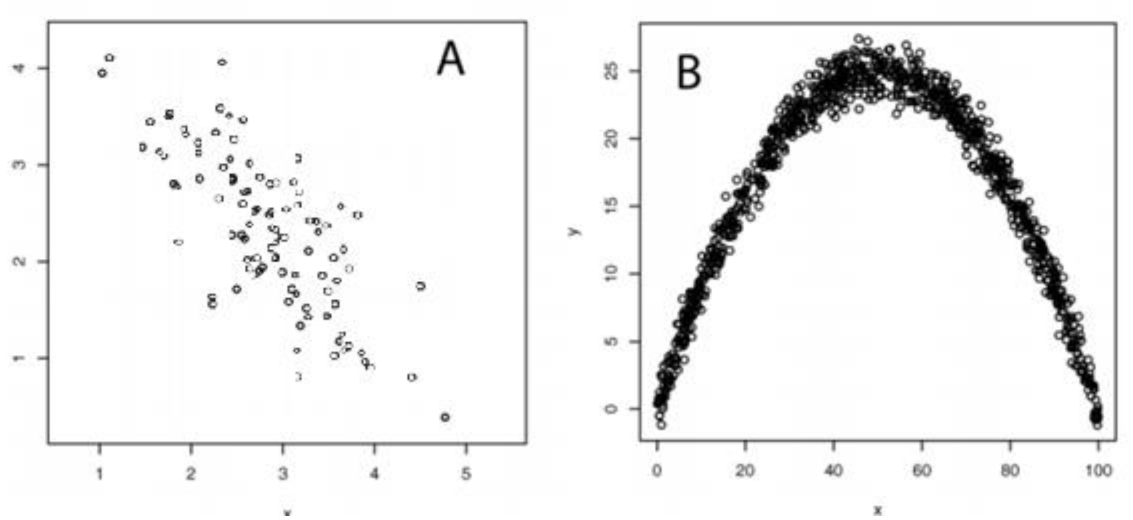

Figure-1

i) Pictorially represent a pair of eigenvectors that you would expect to obtain from a Principal Components Analysis (PCA) for the data shown in Panel A of Figure-1.

ii) Which eigenvector you would expect to have the larger corresponding eigenvalue and why?

iii) What do the numerical values of the eigenvalues tell you about the data?

iv) Do you think it is appropriate to use PCA to reduce the dimensionality of the dataset shown in panel B? Why or why not?

C. We have discussed two techniques for feature selection namely PCA and feature subset selection. Justify which technique is better in terms of interpreting the output as a user. **[2 M]**

Q4. **[3+2=5M]**

A. Describe the key benefits of any 3 of the following charts for univariate analysis:

- Column Chart
- Pie Chart
- Line Chart
- Tree chart
- Bar Chart

B. Describe parallel coordinates and scatter plot approach.

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