DATA AND TEXT MINING
DOCUMENTATION RESEARCH UNIT, ISI-BC
FINAL EXAMINATION-2023

Total Marks: 63 **Time: 3 hrs.**

ANSWER ANY NINE

1. [3+4]
   a. Explain cross validation, and its significance?
   b. Discuss different methods of cross validation.

2.
   a. What is interquartile range and find it for the following vector.
      X=[1, 7, 36, 14, 4, 15, 53, 25, 11] [1+4]
   b. Check through interquartile range for the presence of outlier present in the above vector. [2]

3. Find the Mahalanobis distance of a point X=[1,2,1], from the data set [7]

$$Y = \begin{bmatrix} 1,3,4 \\ 2,0,1 \\ 2,1,1 \\ 1,2,4 \\ 2,1,2 \end{bmatrix}$$

4. Write the K- medoid algorithm with its advantage and disadvantages. In what way it is different from K-means algorithm. [5+2]

5.
   (a) What is the interpretation of the confusion matrix (CM) of a model for the data set with TWO classes C1 and C2? [4]
   (b) How CM helps to find out Precision, Recall of the same model. [3]

6. Describe the processing steps of finding the principal component analysis (PCA) of a data set. What are the significances of PCA in a decision making process? [4+3]

7. What is positive definiteness property of a matrix? How the positive definite aspect of a covariance matrix affects the probability density function of a Gaussian distribution? [3+4]

8. Describe the operational steps of KNN data mining algorithm. Describe its merits and demerits. [5+2]

9.
   a. What are generalization and over-fitting aspects in data mining? [4]
   b. List out the challenges in the design of a clustering model. [3]

10. What is a linear classifier? Describe the developmental steps (including the finding of decision boundary) of a Linear classifier for a TWO-Class problem. [2+4]
11. Describe four important information pre-processing steps with examples in a PR system. [7]
12. Describe with equations; the measures of Location, spread, Shape and dependency. [7]