Institute of Information Technology
Jahangirnagar University
Professional Masters in IT

1st Trimester Final Examination, Fall 2019
Intake: Fall 2019 & Summer 2019
Duration: 3 Hours
Full Marks: 60
Course Code: PMIT-6307
Course Title: Data Mining & Knowledge Discovery

Do not write anything on the question paper.
There are **7 (Seven)** questions. Answer any **5 (Five)** of them.
Figures in the right margin indicate marks.

1. a) Define Data Mining? What are limitations of data mining? Write the differences between association rule discovery and sequential pattern discovery? 4
   b) What are the difference between classification and clustering in data mining? 4
   c) Why do we use test data in mining? 4

2. a) Define the terms: Why do we reduce dimensionality before applying mining techniques? 4
   b) What are the factors which affect data purity? Write the actions that you can take to impure your data? 4
   c) What do you mean by OLAP? Write the application of slicing and dicing? 4

3. a) Why and when do we use Gini coefficient or entropy? Write the Hunt's algorithm for decision tree. 4
   b) Suppose you have two variables (A and B) to select one in constructing a tree. Under variable A distribution of class variable are 5 Cheat and 15 No Cheat in one hand and in other hand it is 3, 10, whereas under variable B it is 4, 6 in one hand and 12 , 8 in other hand respectively. Which variable you should select? Use Gini coefficient or entropy to give your answer. 4
   c) If an attribute, like marital status (married, single and divorce), has three different values and you have to construct binary tree what will you have to do? 4

4. a) What do you mean by similarity and dissimilarity? Write on Euclidean distance and jaccard coefficient. 4
   b) What is ROC curve? What are the use of ROC curve? 4
   c) Consider a 2-class problem (cancer yes, cancer no) where actual number of yes was 6990 and number of no was 3010. Using the diagnostic machine it was found number of yes 5710 and number of no was 4290. Is this machine efficient and reliable? Calculate different parameters of validation and comment on the results. 4

5. a) What are the steps of KNN classification? 4
   b) Give an example of the application of clustering? 4
   c) What are the advantages of SVM classification? 4

6. a) What do you mean by clustering? What are the different types of clusters? 4
   b) Write the steps of K-means clustering in your own words sequentially. 4
   c) Write two limitations of k-means clustering. How can we minimize these limitations? 4

7. a) What do you mean by centroid in k-means clustering 4

b) How do we estimate epsilon (EPS) and minimum points (MinPoints) in Density Based Clustering? 4

c) In hierarchical clustering we have to calculate distance between two clusters, how do we calculate this distance? 4