Institute of Information Technology
Jahangirnagar University
Professional Masters in IT

Final Examination	Semester: Semester II, 2019
Duration: 3 hours	Full Marks: 60
Course Code: PMIT-6307	Course Title: Data Mining & Knowledge Discovery

Answer any 5 (five) of them. Figures in the right margin indicate marks.

1. a) Define Data Mining? What are the challenges in data mining? 4
   b) What are the difference between descriptive data mining and predictive data mining? 4
   c) Describe a situation where we can use a data mining technique. 4

2. a) What do you mean by aggregation? Why do we apply sampling on data? 4
   b) When do we follow "feature subset selection" to reduce data before data mining? Explain it in details. 4
   c) What is OLAP? Why do we need OLAP? Define the term "Slicing" and "Dicing". 4

3. a) What are the different methods of calculating similarity and dissimilarity? 4
   b) Suppose you want to calculate similarity between two customers who bought some goods from a supermarket. What are the formula you can use in this regards? 4
   c) To calculate dissimilarity between two data objects you can use Euclidian Distance and Mahalanobis Distance. Which one will you prefer and why? 4

4. a) What is classification in data mining? Draw the process f classification. 4
   b) Suppose, Asad, Kabir and Raihan bought groceries from shop. Calculate their similarities and evaluate who are very closest 4
   Asad 5 kg rice, 1 kg fish, 200 gm chilly and 1 kg milk
   Raihan 1 kg rice, 1 kg Meat, 200 gm chilly and 1 kg milk
   Kabir 2 kg rice, 1 kg fish, 200 gm salt and 1 kg milk
   c) What is the main principal of Gini index? – explain. When we have to use Gini index in splitting? 4

5. a) Write the procedures of tree based classification? How does one can check the validity of a tree? 4
   b) What do you mean by KNN classification? 4
   c) Suppose you have two variables (Gender and Marital status) to select one in constructing a tree. Among males there are Cheat=6 and No Cheat=9 and among females Cheat=10 and No Cheat=5 whereas under variable marital status it is 5, 7 among married and 10, 8 among unmarried respectively (assuming that there are no divorce). Which variable you should select at this point to grow up tree? Use Gini coefficient or entropy to provide your answer. 4

6. a) What do you mean by clustering? What are the different types of clusters? 4
   b) Shortly explain process of hierarchical clustering. 4
   c) Write the limitations K-means clustering. How can we minimize these limitations? 4

7. a) If you have a data set with class attribute and a new data without class attribute. Write the process of classifying this new data using SVM. 4
   b) If x1(1,2), x2(3,2), x3(2,4), x4(4,1), x5(3,5) x6(1,2) are six data points of only two dimensional. Draw dendrogram for hierarchical clustering. 4
   c) In density based clustering (DBSCAN), how do you estimate radius (eps) and minPoints (minPts) from a set of data? 4