B.Sc. Engg. CSE 7th Semester 27 May 2019

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
ORGANISATION OF ISLAMIC COOPERATION (OIC)
Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION **WINTER SEMESTER, 2018-2019**

DURATION: 3 Hours **FULL MARKS: 150**

CSE 4739: Data Mining

Programmable calculators are not allowed. Do not write anything on the question paper.

There are **8 (eight)** questions. Answer any **6 (six)** of them.

Figures in the right margin indicate marks.

---

1. a) Present an example where data mining is crucial to the success of a business. What *data mining* functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis? 10

   b) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order): 8
   13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   i. Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
   ii. Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.
   iii. Use normalization by decimal scaling to transform the value 35 for age.
   iv. Comment on which method you would prefer to use for the given data, giving reasons as to why.

   c) What are the common repositories used for Mining in Software Repositories (MSR)? 7

2. a) Why Data Integration is important? Discuss issues to consider during data integration. 10
   b) With the help of a diagram explain *Density-reachability* and *Density-connectivity*. 6
   c) What are the benefits of density-based clustering? How does DBSCAN find clusters? 3+6

3. Suppose that a data warehouse consists of the four dimensions *date*, *spectator*, *location*, and *game*, and the two measures *count* and *charge*, where *charge* is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.
   a) Draw a star schema diagram for the data warehouse. 10
   b) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should you perform in order to list the total charge paid by student spectators at GM Place in 2010? 8
   c) If each dimension has five levels (including all), such as "*date* < *week* < *month* < *year* < *all*", how many cuboids will this cube contain (including the base and apex cuboids)? 7

4. a) Today's consumers are faced with millions of goods and services when shopping online. Recommender systems help consumers by making product recommendations that are likely to be of interest to the user such as books, CDs, movies, restaurants, online news articles, and other services. Data mining plays a major role in these recommender systems. What are the approaches researchers use for these systems? What major challenges they face? 7+3

b) What is *Privacy-Preserving Data Mining* (PPDM)? Categorize the PPDM methods. 3+4

c) Briefly discuss the achievements of Mining Software Repositories (MSR). 8

5. A survey was performed in Facebook group "*Pavilion de IUT*", where group members voted their favorite combination of team IUT for Inter University Futsal Tournament. Result of that survey is listed in Table 1.
Let min_sup = 30% and min_conf = 60%.

Table 1: *Five-a-side* teams for Question 6.

Team ID	Team Players
T1	{Ishmam, Khalil, Tajdeed, Sabit, Shams}
T2	{Momin, Khalil, Tajdeed, Sabit, Sayem}
T3	{Fardin, Khalil, Tajdeed, Akik, Shams}
T4	{Momin, Khalil, Tajdeed, Sabit, Sayem}
T5	{Momin, Khalil, Tajdeed, Akik, Shams}
T6	{Momin, Khalil, Ishmam, Akik, Shams}
T7	{Fardin, Khalil, Ishmam, Sabit, Sayem}
T8	{Momin, Khalil, Tajdeed, Akik, Shams}
T9	{Momin, Fardin, Ishmam, Akik, Sayem}
T10	{Momin, Khalil, Tajdeed, Sabit, Shams}
T11	{Fardin, Khalil, Akik, Sifat, Sayem}
T12	{Momin, Khalil, Tajdeed, Sabit, Shams}
T13	{Momin, Khalil, Ishmam, Sabit, Sayem}
T14	{Fardin, Khalil, Akik, Sabit, Shams}
T15	{Momin, Khalil, Tajdeed, Sabit, Sayem}
T16	{Momin, Zakaria, Tajdeed, Sifat, Shams}
T17	{Momin, Khalil, Tajdeed, Akik, Shams}
T18	{Fardin, Khalil, Ishmam, Sabit, Sayem}
T19	{Momin, Akik, Ishmam, Sabit, Sayem}
T20	{Zakaria, Khalil, Ishmam, Sabit, Sayem}

a) Find all frequent sets of players using Apriori algorithm. 20

b) Is correlation measure really necessary for association rules? Give your opinion with proper justification. 5

6. a) Why is mobile crowdsensing preferred over traditional sensor networks? In which case it fails? 5+3

b) Discuss the major challenges in vehicular crowdsensing with appropriate example? 7

c) For the data of Table 1 in Question 5, List all the strong association rules (with support s and confidence c) matching the following meta rule, where X is a variable representing group member, and $Player_i$ denotes variables representing Players (e.g., "Momin," "Sabit"): 10

$\forall x \in Team, choose(X, Player_1) \wedge choose(X, Player_2) \Rightarrow choose(X, Player_3)\ [s,c]$

7. a) Give an application example where global outliers, contextual outliers, and collective outliers are all interesting. Based on your application answer the followings: 6+9
   i. What are the attributes?
   ii. What are the contextual and behavioral attributes?
   iii. How is the relationship among objects modeled in collective outlier detection?

   b) Define DB(r, π) – outliers. Why are they global? Explain a scenario where DB(r, π) fails. 3+3+4

8. Table 2 lists the points achieved by top six teams in 4 major football leagues.

Table 2: *Data for question 8*

Teams	EPL	La Liga	Serie A	Bundesliga
O1	98	86	89	75
O2	97	75	76	73
O3	72	68	66	66
O4	71	58	65	55
O5	70	58	62	55
O6	66	56	62	54

   a) Find the distance matrix using *Minkowski distance* between the objects in Table 2, using q = 4. 10

   b) Draw the dendograms of Hierarchical Agglomerative Clustering applied on the data in table 2 using the followings: 15
   i. Nearest-neighbor clustering algorithm
   ii. Farthest-neighbor clustering algorithm
   iii. Average linkage algorithm