Program: B. Sc. Engg. in Computer Science and Engineering **Date: 20 June, 2021**
Semester: Winter 2020-2021 **Time: 2:30 PM – 4:00 PM**

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

Mid Semester Examination **Winter Semester: 2020-2021**
Course Number: CSE 4739 **Full Marks: 75**
Course Title: Data Mining **Duration: 1 Hour 30 Minutes**

There are **3 (THREE)** questions. Answer **ALL** of them. The symbols have their usual meanings. The examination is **Online** and **Closed Book**. Programmable calculators are not allowed. Marks of each question and corresponding CO and PO are written in the brackets.

---

1. a) What do you understand by spatiotemporal data? Give one example of spatiotemporal data streams explaining the knowledge that can be mined from such data. (8) [CO1, PO1]

   b) Suppose you have a COVID-19 dataset describing daily information on the number of affected cases, deaths, and recoveries. The detail description of the dataset is given below and a snapshot of data samples is given in Table 1. (12) [CO1, PO1]
   - SN - Serial number
   - Observation Date - Date of the observation in MM/DD/YYYY
   - Province/State - Province or state of the observation (Could be empty when missing)
   - Country/Region - Country of observation
   - Last Update - Time in UTC at which the row is updated for the given province or country.
   - Confirmed - Cumulative number of confirmed cases till that date
   - Deaths - Cumulative number of deaths till that date
   - Recovered - Cumulative number of recovered cases till that date

Table 1: Snapshot of COVID-19 dataset

# SNo	Observatio...	Province/S...	Country/R...	Last Update	# Confirmed	# Deaths	# Recovered
305856	05/29/2021	Andaman and Nicobar Islands	India	2021-05-30 04:20:55	6964.0	113.0	6660.0
305857	05/29/2021	Andhra Pradesh	India	2021-05-30 04:20:55	1671742.0	10738.0	1487382.0
305858	05/29/2021	Anguilla	UK	2021-05-30 04:20:55	109.0	0.0	109.0
305859	05/29/2021	Anhui	Mainland China	2021-05-30 04:20:55	1004.0	6.0	994.0
305860	05/29/2021	Antioquia	Colombia	2021-05-30 04:20:55	536911.0	11914.0	503302.0
305861	05/29/2021	Antofagasta	Chile	2021-05-30 04:20:55	53081.0	1039.0	50817.0

From the dataset in Table 1, describe the class/concept relationships. Explain with an example how you can derive these descriptions using data characterization and data discrimination.

   c) Name the types of data attributes present in the dataset of Table 1. List the most common statistical data dispersion measures for this dataset. (5) [CO1, PO1]

2. a) An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 questions with four possible answers each. How would you convert this data into a form suitable for association analysis? (7) [CO2, PO2]

   b) It is important to measure the proximity of attributes in data analysis. Such measures of similarity are used in different data mining applications such as clustering, outlier

analysis, recommendation, and so on. Equations of one of the most common similarity measures named as 'Cosine Similarity', is given in the equations (1) and (2). Suppose we have the following two-dimensional data set.

Table 2: A sample 2D dataset

	A_1	A_2
x_1	1.7	1.9
x_2	2.1	1.8
x_3	1.5	1.7
x_4	1.1	1.3
x_5	1.7	1.0

$$Cosine\ Similarity\ (x, y) = \frac{x^t \cdot y}{\|x\|\ \|y\|} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

$$Cosine\ Distance = 1 - Cosine\ Similarity \quad \ldots\ldots\ldots\ldots\ldots (2)$$

i. Given a new data point, $x_t = (1.3, 1.5)$ as a query, rank the data points based on similarity with the query according to minimum cosine distance. Use equations (1) and (2). (10) [CO2, PO2]

ii. Normalize the data set to make the norm of each data point equal to 1 and present the normalized data points in a table like Table 2. Rank the data points using Euclidean distance based on the normalized query. (8) [CO2, PO2]

3. a) Suppose you want to analyze COVID-19 cases from the dataset in Table 1. To analyze different dimensions you have to construct a data warehouse with respect to observation time, different cases (e.g. confirmed, death, recovered), city, province, country. These dimensions allow you to keep track of COVID-19 scenarios on a daily, monthly, or yearly basis. Assume that there is a table associated with each dimension. For this scenario answer the followings:

i. Draw a 4-D data cube representation of the data given in Table 1. Assume appropriate dimensions and measures. (6) [CO2, PO2]

ii. Draw an example Snowflake schema of COVID-19 cases' data warehouse. (6) [CO2, PO2]

b) Consider the transaction table given in Table 3. Let the $Min\_sup = 50\%$ and $Min\_conf = 60\%$.

Table 3: A sample transactional dataset

Transactions	List of items
T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	I_1, I_2, I_3, I_4

Answer the followings:

i. Find all the frequent itemsets using Apriori algorithm. (7) [CO2, CO3, PO2]

ii. Generate all the strong association rules. (6) [CO2, CO3, PO2]