48

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

MID SEMESTER EXAMINATION **WINTER SEMESTER, 2022-2023**
DURATION: 1 HOUR 30 MINUTES **FULL MARKS: 75**

CSE 4775: Data Mining

Programmable calculators are not allowed. Do not write anything on the question paper.
Answer **all 3 (three)** questions. Figures in the right margin indicate full marks of questions whereas corresponding CO and PO are written within parentheses.

---

1. a) Home values among the 8,000 homeowners of Town X are normally distributed, with a standard deviation of $11,000 and a mean of $90,000. Find the number of homeowners in Town X whose home value is greater than $112,000. — 10 (CO2) (PO1)

   b) What is the curse of dimensionality? Describe a dimensionality reduction technique. — 4 + 6 (CO1) (PO1)

   c) Describe the property that the Apriori algorithm exploits to work efficiently. — 5 (CO1) (PO1)

2. a) Why do we need a separate Data Warehouse? Distinguish between OLTP and OLAP. — 4 + 6 (CO1) (PO1)

   b) What is meta-data repository? Describe the four properties of a data warehouse. — 3 + 5 (CO1) (PO1)

   c) The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. — 7 (CO1) (PO1)

   Give an overview of variable datatypes (i.e. nominal, binary, ordinal, numeric etc) from the data represented in Table 1.

Table 1: Variable definitions (Data for Question 2.c)

Variable	Definition
survival	Survival (0 = No, 1 = Yes)
pclass	Ticket class/Economic status (1 = 1st, 2 = 2nd, 3 = 3rd)
Gender	gender
Age	Age in years
sibsp	# spouses aboard the Titanic (Siblings or Spouse)
parch	# children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation (C=Cherbourg, Q=Queenstown, S=Southampton)

3. a) If a set of data consists of only the first ten positive multiples of 5, Find the interquartile range of the set? 7 (CO2) (PO1)

b) Answer the following questions: 4 + 4 (CO1) (PO1)

i. What is association analysis?

ii. Why do we need to clean data?

c) A researcher is interested in examining the effects of a new drug on pain reports. A total of 90 participants were randomly assigned to one of three conditions (control/drug/placebo) and asked at the end of one week of treatment whether they were experiencing arthritis pain (yes/no). The results are presented in Table 2. 10 (CO2) (PO1)

Table 2: Findings (Data for Question 3.c)

Condition	No pain	Pain
control	6	24
drug	20	10
placebo	8	22

Conduct a chi-square test of independence on the data and report whether the condition is significantly related to the pain report. Take the level of significance at 5%. The partial chi-square distribution is shown in Table 3.

Table 3: Chi-square distribution table (Data for Question 3.c)

df	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	6.63
2	0.02	0.10	0.21	0.58	1.39	2.77	4.61	5.99	9.21
3	0.11	0.35	0.58	1.21	2.37	4.11	6.25	7.82	11.34
4	0.30	0.71	1.06	1.92	3.36	5.39	7.78	9.49	13.28