The Open University of Sri Lanka
Faculty of Engineering Technology
Department of Electrical and Computer Engineering

038

Study Programme	: Bachelor of Science Honours in Engineering
Name of the Examination	: Final Examination
Course Code and Title	**: EEX7244 Data Mining**
Academic Year	: 2023/2024
Date	: 13th February 2025
Time	: 0930-1230hrs
Duration	: **3 hours**

General Instructions

1. Read all instructions carefully before answering the questions.
2. This question paper consists of **four (4)** questions in **four (4)** pages.
3. Answer all four questions. All questions carry equal marks.
4. Answer for each question should commence from a new page.
5. No charts/ codes are provided.
6. This is a Closed Book Test (CBT).
7. Answers should be in clear handwriting.
8. Do not use a Red colour pen.

Question 01

a) Use a suitable diagram to explain the knowledge discovery process **[05 Marks]**

b) Consider the income range from LKR 30,000 to LKR 450,000. Use Min-Max normalization and map LKR 180,000 to the new normalized value between [0.0, 1.0]. **[04 Marks]**

c) Consider a scenario where correlation coefficient is calculated as –1. What can you say about the two variables involved? Use a suitable diagram to depict the correlation between the variables. **[04 Marks]**

d) What is the use of Principal Component Analysis in data mining? **[04 Marks]**

e) Portfolio investment involves buying financial assets with the expectation of earning returns. Common asset classes for portfolio investment include stocks and bonds. Kirihami is an investor. His portfolio primarily tracks the performance of the ABC company. Kirihami now wants to add the stock of XYZ company. Before adding stock to his portfolio, he wants to assess the directional relationship between the stock of XYZ and ABC. Use a suitable method to identify whether the price of the XYZ and the ABC tend to move in the same direction using the data below. **[08 Marks]**

Year	ABC	XYZ
2019	250	50
2020	300	100
2021	275	120
2022	320	150
2023	360	170

Table 1

Question 02

a) Briefly explain the steps in "Binning" with respect to handling of noisy data. **[04 Marks]**

b) Consider three employee databases of a reputable Hotel Chain that need to be cleaned before they are used for data mining process.

Answer the following questions considering the above statement. **[06 Marks]**

i. Briefly explain a situation where data inconsistency can be present with a suitable example.

ii. Considering the attribute **salary**, briefly describe a situation where Noisy data can be present.

iii. State two reasons why missing data can occur.

c) Suppose that a data warehouse consists of three dimensions: - time, doctor, and patient, and two measures count and charge. The count is the number of patients a doctor visits each day. Charge is the fee that a doctor charges a patient for a visit. Part of the lattice of cuboids (from apex to base cuboid) for the above data warehouse is given in Figure 1.

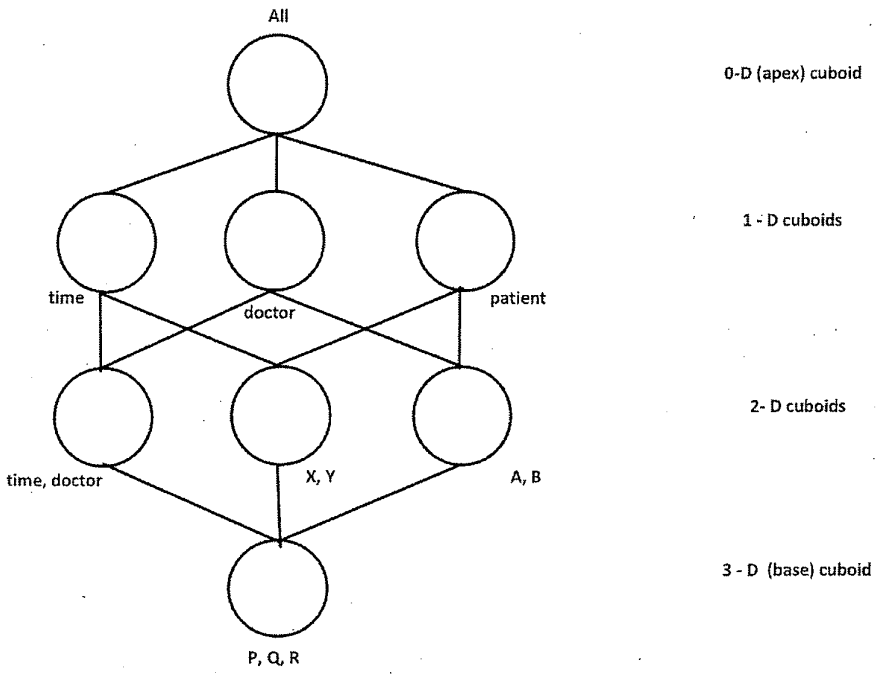

Figure 1

i. Complete the lattice by identifying the values of X, Y, A, B, P, Q, R of cuboids considering the information given. **[05 Marks]**

ii. Draw a star schema diagram for the above data warehouse. For each dimension, include the appropriate attributes (conceptual hierarchies). **[10 Marks]**

Question 03

a) With the help of a suitable diagram, explain the process of text mining. Describe each step with the techniques that can be used in each step. **[07 Marks]**

b) Consider Table 2 given below which consists of transaction details at a super store. Find frequent sets of items that are purchased by the customers and generate the association rules for them. Use the Apriori algorithm for this given question considering minimum-support as 3 and the minimum-confidence as 70%. **[18 Marks]**

Transaction ID	Items Purchased
T1000	1, 2, 3
T1001	1, 4
T1002	4, 5
T1003	1, 2, 3
T1004	6, 3
T1005	6
T1006	6, 4
T1007	1, 2, 7, 3
T1008	7, 5
T1009	1, 2

Table 2

Question 04

Sepsis is reported to be a leading cause of death in general Intensive Care Units [ICU]. Early predictions of cardiac arrest in Sepsis patients would help the medical practitioners take necessary actions to minimize the fatality rate. Data mining techniques can be used to develop classification systems to detect whether a Sepsis patient will face a cardiac arrest or not.

a) Name two techniques that can be used for the above classification **[04 Marks]**

b) Select one such classification technique and briefly explain its theory and working principle. You may assume that the dataset of the above scenario contains 35 attributes. **[05 Marks]**

c) Use a suitable diagram to support your answer in b). **[04 Marks]**

d) Imagine that you used the above selected technique for classification and decided to use the 10-fold cross validation technique. Use a suitable diagram to illustrate how, in each fold, the data is split for training and testing. Assume that 4500 data instances are available in the Sepsis patient dataset. **[04 marks]**

e) Imagine you received the confusion matrix below during the testing process. Considering the confusion matrix, identify the values for i – ii and calculate the percentage values for iii – iv. **[08 Marks]**

	Predicted: No	Predicted: Yes
Actual: No	2600	293
Actual: Yes	400	1207

Figure 2

i. True Positive value
ii. False Negative value
iii. Accuracy
iv. Sensitivity

END OF PAPER