

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



Study Programme	: Bachelor of Science Honours in Engineering/ Bachelor of Software Engineering Honours
Name of the Examination	: Final Examination
Course Code and Title	: EEX7244/ EEX6377 Principles and Applications of Data Mining
Academic Year	: 2024/25
Date	: 22 nd February 2026
Time	: 0930-1230h
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 questions.
4. This is a Closed Book Exam (CBE).
5. Answers should be in clear handwriting.
6. Do not use red colour pen.

Question 01

A financial analytics company has been hired to analyse large volumes of customer and stock market data to support decision-making. The company applies data mining techniques to discover hidden patterns, relationships, and trends that can support business intelligence and investment strategies.

Answer the following questions based on this scenario.

- a. Using a suitable diagram, explain the complete Knowledge Discovery in Databases (KDD) process. Clearly indicate each stage and its purpose. (5 marks)
- b. The company collects customer demographic information. Suppose the age of customers varies between 18 and 70 years. Using Min-Max normalization, transform the value 45 into the range [0,1]. Show all steps clearly. (4 marks)
- c. While analysing stock performance, the company computes a correlation coefficient of +0.85 between two variables. What does this indicate about the relationship between the variables? Use a suitable graph to illustrate your answer. (4 marks)

- d. The dataset contains over 50 attributes describing customers and stock indicators. Explain the importance of dimensionality reduction in large datasets and how it benefits the data mining process. (4 marks)
- e. An investor wants to determine whether Stock A and Stock B move together overtime before adding Stock B into his investment portfolio. Using covariance or correlation, determine the relationship between the two stocks using the data below given in Table 1. Show all steps clearly. (8 marks)

Table 1

Year	Stock A	Stock B
2020	150	75
2021	170	82
2022	165	90
2023	180	95
2024	200	110

Question 02

- a. Explain how clustering can be used to handle noise with suitable examples. (4 marks)
- b. Consider a hospital database that integrates patient records from three branches. (6 marks)
- Give one example of data redundancy
 - Give one example of inconsistent data
 - State two reasons for missing values
- c. A data warehouse consists of four dimensions: time, product, branch, and customer; and two measures: sales and profit.
- Draw and label a lattice of cuboids (from apex to base cuboid). (5 marks)
 - Draw a star schema for the above warehouse and identify fact and dimension tables with key attributes. (10 marks)

Question 03

- a. Briefly explain the difference between *Text mining* and *Web mining*. (3 marks)
- b. *Attribute Selection* and *Data Mining/Pattern Discovery* are 2 steps involved in the process of text mining. Name 2 other steps involved with text mining other than what is given. Discuss them with suitable examples. (4 marks)

- c. Consider Table 3 given below which consists of transaction details at a super store. Find frequent set of items that are purchased by the customers and generate the association rules for them. Use the Appropriate algorithm for this question considering minimum support value as 2 and the minimum confidence as 50%. (18 marks)

Table 3

Transaction ID	Items Purchased
T2000	A, B
T2001	B, D
T2002	B, C
T2003	A, B, D
T2004	A, C
T2005	B, C
T2006	A, C
T2007	A, B, C, E
T2008	A, B, C

Question 04

Kidney Disease symptoms tend to develop slowly over time. Chronic kidney disease (CKD) is the long-term, progressive loss of kidney function, often causing no symptoms until advanced stages. An early diagnostic methodology to determine kidney functionality is essential and extremely important in many cases. Data mining techniques can be used to develop systems to detect whether a patient will get CKD or not.

- Name two classification techniques that can be used for the above task. (4 marks)
- Select one such classification technique and explain its theory and working principle using not more than 10 lines. You may assume that the dataset of the above scenario contains 24 attributes. (5 marks)
- Use a suitable diagram to support your answer in b. (4 marks)
- 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, if 400 data instances are available in the CKD dataset. (4 marks)
- 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. (8 marks)

	Predicted: No	Predicted: Yes
Actual: No	170	28
Actual: Yes	35	167

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

All Rights Reserved

~End of the Paper~