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

Study Programmes	: Bachelor of Software Engineering Honours
	: Bachelor of Technology Honours in Engineering
Name of the Examination	: Final Exam
Course Code and Title	: **EEX6377-Principles and Applications of Data Mining**
	: **EEX7244- Data Mining**
Academic Year	: 2021/22
Date	: 11th February 2023
Time	: 0930-1230 hrs
Duration	: **3 Hours**

General Instructions

1. Read all instructions carefully before answering the questions.
2. This question paper consists of **five (5) questions** in **four (4)** pages.
3. Answer all questions, each question will carry equal marks.
5. Answers for each question should commence from a new page.
6. This is a Closed Book Test **(CBT).**
7. Answers should be in clear handwriting.
8. Do not use a Red colour pen.
9. Use of Scientific Calculators is allowed.

Q1

A well established and reputed online store collects information on their daily sales of books, CDs, DVDs and stationery etc. for each customer. The store is having about 1 million customers and about 50,000 items for sale. The online store was exploring methods for storing these data and one suggestion was to store them in a one million row and a fifty thousand column matrix with non-zero entries. In the matrix i,j th entry will contain the information what the ith customer has bought.

The online store wants to expand the business by adopting strategies like recommending suitable items to buy at appropriate times. Further they want to suggest any items a new customer has not bought by comparing to a previous similar customer who had bought the particular item. Considering this scenario answer the following questions based on concepts and techniques of Data Mining.

(a) Describe the data preparation steps required in this scenario.

(b) Describe the sequence of data mining steps with respect to the above given scenario.

(c) Justify your answer whether storing data in a matrix is a good initiative if not propose and justify any other approaches.

(d) Describe the major challenges that the online store may face in mining such large datasets. (5×4=20 marks)

Q2

The ABC University is planning to store some data in a data warehouse and the four dimensions identified are student, course, semester, and instructor, and two measures are count and average grade (avg grade). When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the avg grade for the given combination. Based on this scenario answer the following questions.

(a) Draw a snowflake schema diagram for the data warehouse. (10 marks)

(b) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each student in the university. (6 marks)

(c) If each dimension has five levels (including all), such as "student < major < status < university< all", how many cuboids will this cube contain (including the base and apex cuboids)? Explain how you derive the answer. (4 marks)

Q3

(a) Association rules with 99% confidence are more interesting than rules with 100% confidence. Justify this statement with an example. (2 marks)

(b) Describe what is the Apriori property with an example? (2 marks)

(c) A stationary shop is selling different types of stationary and those are denoted by English alphabet letters for easy reference. The transaction data are stored in a database as follows. Below shown is a portion of the database with 4 transactions.

TID	Date	items_bought
T100	10/15/04	{K, A, D, B}
T200	10/15/04	{D, A, C, E, B}
T300	10/19/04	{C, A, B, E}
T400	10/22/04	{B, A, D}

Assuming a minimum level of support min_sup = 60% and a minimum level of confidence min_conf = 80%.

(i) Find all frequent itemsets that satisfy the minimum support upto 3 levels. That is C3/L3. You need to show all the intermediate working. (4 marks)

(ii) Identify all possible 3 member rules that can be derived from the frequent item set identified from C3/L3 in part (a). (4 marks)

(iii) Calculate the confidence for each rule derived in part (ii). (4 marks)

(iv) List all of the strong association rules, along with their support and confidence values, which match the following meta rule, where X is a variable representing customers and item *i* denotes variables representing items (e.g., "A", "B", etc.).
∀x ∈ transaction, buys(X, item1) ∧ buys(X, item2) ⇒ buys(X, item3)

Hint

One example rule derived is, buys(X, A) ∧ buys(X, B) → buys(X, D).
Based on the support and confidence strong association rules can be derived. (4 marks)

Q4

(a) Briefly describe the major steps in decision tree classification. (2 marks)

(b) The following data set is used to train a decision tree to decide whether a given fruit is edible or not based on the attributes shape, colour and the odor.

Shape	Colour	Odor	Edible
C	B	1	YES
D	B	1	YES
D	W	1	YES
D	W	2	YES
C	B	2	YES
D	B	2	NO
D	G	2	NO
C	U	2	NO

C	B	3	NO
D	W	3	NO

(i) Calculate the expected *information need* to classify a tuple for the class category Edible. (2 marks)

(ii) Calculate the expected information gain for each attribute shape, colour and odor. (2*3 =6 marks)

(iii) Calculate the information gain for each attribute. (3 marks)

(iv) Justify which attribute will be considered as the root node for developing the decision tree. (3 marks)

(v) Draw the full decision tree based on the answer given in part (iv) above. (4 marks)

Q5 Consider the below given two data mining scenarios.

Scenario 01

Diabetes is one of the most common long-lasting diseases that indicates how the food that you eat is converted to energy. It requires a lot of care and proper medication to keep the disease in control. Data mining techniques are used to develop classification systems to detect whether a patient has diabetes or not based on different attributes of patients.

Scenario 02

With the exponential credit card usage there is a high tendency for credit card frauds as well. Banks are trying to handle this issue using data mining techniques to classify each transaction as legitimate or a fraudulent transaction.

Identify a single scenario of your choice and answer the following questions. You have to clearly indicate the scenario you have chosen to answer the questions.

(a) Identify a suitable ML based classification technique to apply for the scenario chosen. Justify your selection. (3 marks)

(b) Explain the process of knowledge discovery and data mining (KDD) with respect to the scenario you have chosen. You have to clearly explain any assumptions that you make in each of the steps with respect to the scenario chosen. (6 marks)

(c) Describe how you can measure the accuracy of the model derived in Q5 (b). (4 marks)

(d) Explain the techniques that you can employ to improve the accuracy and performance of the classification model. (4 marks)

(e) Describe what model overfitting is and how you can avoid it. (3 marks)

END OF THE PAPER