TUNKU ABDUL RAHMAN UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

ACADEMIC YEAR 2023/2024

JANUARY EXAMINATION

AACS2383 INTRODUCTION TO DATA MINING

MONDAY, 15 JANUARY 2024 TIME: 9.00 AM – 11.00 AM (2 HOURS)

DIPLOMA IN COMPUTER SCIENCE

Instructions to Candidates:

Answer **ALL** questions. All questions carry equal marks.

Question 1

a) List down **EIGHT (8)** types of data that can be mined. (8 marks)

b) One of the major issues in Data Mining is related to Mining Methodology. Discuss this issue by providing **ONE (1)** example in your discussion. (3 marks)

c) Discuss **FOUR (4)** issues that require data cleansing. Provide an example for each. (8 marks)

d) List **THREE (3)** data reduction strategies with an example for each. (6 marks)

[Total: 25 marks]

Question 2

a) Describe **FOUR (4)** key distinctions between Online Transaction Processing and Online Analytical Processing. (8 marks)

b) Binning is grouping a set of continuous or numerical data points into a smaller number of discrete "bins" or intervals to simplify the data and reduce its dimensionality. Table 1 shows the data for bestselling books in the TAR UMT Bookstore.

Table 1: Bestselling books in the TAR UMT Bookstore

Book	RM
The Girl with No Name	20
Wings of Fire	35
Fear	25
Trust in Me	40
Life in Woods	98
Call Me by Your Name	48
Living in Venice	57
Insanity	22
The Name Jar	57
Made to Stick	25
The Unfolding	18
Easy as ABC	11

(i) Assuming that the $number\ bins = 3$. Compute by partition into equal-frequency bins. (4 marks)

(ii) Assuming that the $number\ of\ bins = 3$. Compute by partition into equal-width bins. (4 marks)

Question 2 (Continued)

c) Analyse the attributes of a dimension table and a fact table within the context of data warehousing. Support your analysis with an example. Then, illustrate your insights by creating a star schema diagram representing a data warehouse with four dimensions: Product, Store, Sales, and Customer. The fact table in this schema will be Sales. (4 + 5 marks)

[Total: 25 marks]

Question 3

a) Imagine you are advising a local store that wants to improve its business using Data Mining. Apply your knowledge by proposing and explaining **THREE (3)** specific applications of Data Mining in the real world. For each application, provide a practical example of how it could benefit the store. Consider any potential challenges and suggest strategies to address them, demonstrating a practical understanding of how data mining can be actively applied to enhance business operations. (9 marks)

b) Discuss data quality in terms of Accuracy, Consistency, Completeness, and Timeliness. Explain each criterion and provide an example for each to support your answer. (12 marks)

c) Contrast supervised learning with unsupervised learning from **TWO (2)** aspects. (4 marks)

[Total: 25 marks]

Question 4

a) Examine **TWO (2)** characteristics that contribute to the generation of high-quality clusters in clustering algorithms. (4 marks)

b) Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering method used in Data Mining to separate high-density clusters from low-density.

(i) Provide an analysis of **TWO (2)** strengths and limitations of DBSCAN. (4 marks)

(ii) In the context of DBSCAN, evaluate the significance of a "core point". Then, describe the characteristics that define a point as a core point, and explain how core points contribute to the clustering process in DBSCAN. (1 + 2 + 2 marks)

Question 4 (Continued)

c) K-means is a popular clustering algorithm that partitions data points into clusters based on distance measures and iterative centroid updating.

(i) Briefly explain in **FOUR (4)** steps how K-means partitions data points into clusters. (4 marks)

(ii) List **TWO (2)** distance measures used by the K-means algorithm for different variables. (2 marks)

(iii) Assuming that A and B are two data points. Calculate the Euclidean Distance and Manhattan Distance between $A(1, 3)$ and $B(2, 3)$. (6 marks)

[Total: 25 marks]