TUNKU ABDUL RAHMAN UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

ACADEMIC YEAR 2023/2024

JANUARY EXAMINATION

AACS1573 INTRODUCTION TO DATA SCIENCE

TUESDAY, 16 JANUARY 2024 TIME: 9.00 AM – 11.00 AM (2 HOURS)

DIPLOMA IN COMPUTER SCIENCE

DIPLOMA IN INFORMATION TECHNOLOGY

DIPLOMA IN INFORMATION SYSTEMS

Instructions to Candidates:

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

Question 1

a) Define what is Data in terms of Data Science. (2 marks)

b) Give **TWO (2)** types of data in Data Science. Then, describe each with an example. (3 + 3 marks)

c) Give **FOUR (4)** types of Data Analytics that can be applied in Data Science. Then, describe each. (4 + 4 marks)

d) Data Science has a wide range of applications across various industries, and its impact continues to grow as organisations leverage data to gain insights and make informed decisions.

Give **THREE (3)** examples of Data Science applications. Then, provide a concise explanation for each. (3 + 6 marks)

[Total: 25 marks]

Question 2

a) Based on case study in Table 1 and Figure 1, answer the following question:

Table 1: Customer Details

id	customer_name	purchase_date	customer_emailaddress
01	Choo Yun Huoy	15032023	chooyh@gmail.co,
01		20230411	Jacky_22#gmail.com
02	Ong Jia Le	13042023	ong@gmail.com
03	Aryssara Anand	16052023	sarasara@gmail.com
04	Darween	17-June-2023	
05	Alexander	23062023	Alex_00@gmail.com
06	Simon Lian	03072023	Lian88@gmail.com
07	Aarav Viran	04072#23	aaravvian@gmail.com
08	Gardenia	14072023	
08	Farah	20082023	Farah#gmail.com

Case Study:
Customer Data Analysis

Background:
You work for a retail company that has collected customer data over the past year. The dataset includes information such as id, customer_name, purchase_date, and customer_emailaddress. Assume the data is stored in a CSV file as shows in Table 1.

Figure 1: Case Study for Customer Data Analysis

Question 2 a) (Continued)

(i) Upon initial inspection of the dataset in Table 1, there are several issues requiring attention that become apparent in the Data Preparation stage. Give **THREE (3)** issues. (6 marks)

(ii) Give **THREE (3)** ways to handle the issues presented in Question 2 a)(i). (6 marks)

(iii) Give **THREE (3)** fields that require analysis using the customer data. (3 marks)

b) Feature selection in the context of Machine Learning refers to the process of choosing a subset of relevant and important features from a larger set of features available in a dataset. Based on the case study provided in Figure 2, answer the following question:

Case Study:
Healthcare Predictive Modelling

Background:
You are working with a healthcare provider that is interested in implementing predictive modelling to improve patient outcomes. The goal is to predict the likelihood of readmission within 30 days for patients with chronic conditions.

Data:
The dataset includes patient demographics, medical history, prescribed medications, number of hospitalisations in the past year, and details about the previous hospitalisations (if any). The target variable is binary, indicating whether the patient was readmitted within 30 days (1 for readmission, 0 for no readmission).

Objectives:
Develop a predictive model to identify patients at risk of readmission within 30 days. Assess the model's performance and discuss its potential impact on patient care. Investigate the key features contributing to the model's predictions.

Figure 2: Case Study of Healthcare prediction Modelling

(i) Give **FIVE (5)** features that you think would be most relevant for predicting readmission in this healthcare context. (5 marks)

c) Define Data Stream. Then, give **TWO (2)** ways that queries get asked about streams. (3 + 2 marks)

[Total: 25 marks]

Question 3

a) Define Machine Learning. (3 marks)

Question 3 (Continued)

b) K-means and Mean Shift are two examples of clustering methods. Therefore, answer the following question:

(i) Discuss the **FOUR (4)** important steps used in each of the clustering methods. (4 + 4 marks)

(ii) Discuss **TWO (2)** limitations for each of the clustering methods. (2 + 2 marks)

c) In general, the algorithm of Machine Learning can be categorised into two main types. Define the Supervised Learning and Unsupervised Learning. (2 + 2 marks)

d) Descriptive analytics is commonly likened to unsupervised learning since it lacks a target variable to guide the learning process. Explain the purpose of the following algorithm used in Descriptive Analytics with **ONE (1)** example each.

(i) Association Rules. (3 marks)

(ii) Sequence Rules. (3 marks)

[Total: 25 marks]

Question 4

a) Demonstrate the step-by-step calculation of the surprise number using the Alon-Matias-Szegedy Algorithm for the given question:

(i) 1, 5, 1, 5, 5, 1, 3 (2 marks)

(ii) 6, 6, 6, 6, 6, 6 (2 marks)

(iii) a, b, c, c, c, a, b, d (2 marks)

b) Discuss the **FOUR (4)** rules for forming bucket according to the Datar-Gionis-Indyk-Motwani algorithm. (8 marks)

c) Define PageRank. (2 marks)

Question 4 (Continued)

d) Calculate the PageRanks (first iteration) for the transition matrix given follows:

$$\begin{vmatrix} 0 & 0 & ½ \\ ½ & 0 & ½ \\ ½ & 1 & ½ \end{vmatrix}$$

(9 marks)

[Total: 25 marks]