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

030

Study Programme	: Bachelor of Software Engineering Honours
Name of the Examination	: Final Examination
Course Code and Title	**: EEX4373 Data Science**
Academic Year	: 2021/22
Date	:22nd February 2023
Time	: 0930-1230hrs
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. All questions carry equal marks.
5. Answer for each question should commence from a new page.
6. This is a Closed Book Test **(CBT).**
7. Answers should be in clear hand writing.
8. Do not use Red colour pen.

Question 1

Assume that you are working as a Data Scientist for a large online retailer. Your team has been tasked with analyzing customer purchasing behavior in order to identify trends and make recommendations for improving sales. Based on this scenario answer the below questions.

(a) Identify six (06) attributes (features) in the above scenario that can be used as variables for building the prediction model. You need to justify the selection of the attributes. (4 marks)

(b) Describe the process of Exploratory Data Analysis (EDA) that you can carry out for the attributes identified in part(a). (4 marks)

(c) Describe three (03) common techniques used in EDA. (3 marks)

(d) Describe three (03) challenges that can arise in carrying out EDA techniques described in part (c). (3 marks)

(e) Clearly describe two techniques that can be used to detect outliers considering the real-world implementation of the above given scenario. (6 marks)

Question 2

You are given a dataset containing information about customers of a local supermarket, and your task is to build a classification model to predict whether a customer will purchase a certain product or not. The dataset has the following features:

- Age (numeric)
- Gender (categorical: male/female)
- Income (numeric)
- Education (categorical: primary, secondary, bachelor's, master's, PhD)
- Marital Status (categorical: married, single, divorced)
- Purchase (binary: 0 or 1)

(a) Based on the above given features of the data set, identify 3 potential limitations/errors that can occur in the dataset and give a solution to overcome each of the identified limitation/error. (6 marks)

(b) List down 04 possible classification algorithms that can be applied to the above data set. Out of the identified classification algorithms select the most suitable algorithm with justification. (8 marks)

(c) Describe three (03) possible evaluation metrics that can be used to measure the performance of the classification model built in part (b). (3 marks)

(d) Describe three (03) methods that can be used to improve the accuracy of the classification model you build in part(b). (3 marks)

Question 3

Assume that you are given a dataset containing information about customers of a supermarket, and your task is to perform customer segmentation using clustering techniques. The dataset has the following features:

- Customer ID (numeric)
- Age (numeric)
- Gender (categorical: male/female)
- Income (numeric)
- Education Level (categorical: primary, secondary, graduate)
- Marital Status (categorical: single, married, divorced)
- Number of Children (numeric)
- Total Spending (numeric)

(a) Out of the above features identify the best feature for clustering with justification. (3 marks)

(b) How would you choose the number of clusters to use in your clustering model? (3 marks)

(c) Describe 3 clustering methods and identify the most suitable clustering method for the above scenario with justification. (6 marks)

(d) With respect to the most suitable clustering method identified in part (c), clearly describe the process of clustering for this given scenario. (4 marks)

(e) Describe why high dimensional data clustering is difficult and explain a method to address the problem. (4 marks)

Question 4

(a) Given a dataset {0, 2, 4, 6, 24, 26}, initialize the k-means clustering algorithm with 2 cluster centers c1 = 3 and c2 = 4.

(i) What are the values of c1 and c2 after one iteration of k-means? You may consider Euclidian distance as the distance measure. (4 marks)

(ii) What are the values of c1 and c2 after the second iteration of k-means? (3 marks)

(iii) Describe with an example to what kind of situations a k-means clustering suits better. (3 marks)

(b) Assume that you are working as a data analyst for an online retailer that sells a variety of products. You have been asked to create a decision tree to help the company decide which products to promote on its homepage, based on historical sales data. After analyzing the data, you need to identify several variables that are likely to be good predictors of future sales. Based on this scenario answer the following questions.

(i) Based on the question "Should we promote a product on the homepage?" develop a decision tree for the above scenario. (5 marks)

(ii) Critically analyze the use of decision tree for the above scenario. (5 marks)

Question 5

This question is based on your knowledge in R scripting language.

(a) What is a data frame in R, and how can you create a new data frame? Give an example of a data frame with at least two columns. (4 marks)

(b) Below table represents marks of 5 students for 3 subjects.

Student Name	Data Structures	Data Science	Python
Amal	75	67	68
Dulani	66	64	59
Anura	63	60	62
Saman	72	52	50
Ruvinie	53	70	67

Based on the above table write R Scripts to

(i) Store these data in a Matrix
(ii) Find the average mark of the subject Data Science
(iii) Find the mean of the Python subject (3×3=9 marks)

(c) Assume that you are required to store the below data set in a 3×3 matrix.

2, 4, 6, 8, 10, 12, 14, 16, 18

Write R scripts to

(i) Create a matrix that contains the response variable and the predictor variables. The response variable should be in the first column of the matrix. (3 marks)

(ii) Fit a linear regression model for the matrix created in Part(i). (2 marks)

(iii) Write any assumptions that you make in generating the linear regression model. (2 marks)

End of the Paper