UTHM
Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
(ONLINE)
SEMESTER I
SESSION 2020/2021

COURSE NAME	:	DATA MINING
COURSE CODE	:	BIT 33603
PROGRAMME CODE	:	BIT
EXAMINATION DATE	:	JANUARY / FEBRUARY 2021
DURATION	:	3 HOURS
INSTRUCTION	:	1. ANSWER **ALL** QUESTIONS 2. STUDENTS SHOULD UPLOAD THE ANSWER BOOKLET (PDF/WORD FORMAT) WITHIN 30 MINUTES AFTER EXAMINATION PERIOD

THIS QUESTION PAPER CONSISTS OF **FIVE (5)** PAGES

TERBUKA

Q1 State whether or not each of the following activities is a data mining task. Explain your answer.

a) Monitoring the heart rate of a patient for abnormalities.

(? marks)

b) Computing the total profit of health insurance industry.

(2 marks)

c) Monitoring and predicting failures in a hydropower plant

(2 marks)

d) Dividing the customers of a company according to their salary.

(2 marks)

e) Extracting the frequencies of a sound wave

(? marks)

f) Identifying the characteristics of successful used car salesperson.

(2 marks)

Q2 Discuss the Knowledge Discovery process involved when solving a data mining problem for scenario given in **Figure Q2.**

A telco company wants to promote a new network service. Due to mailing costs, it wants to send the promotional material to a limited number of potential customers. The company decided to use data mining techniques to find out what attributes, if any, and group the customers into two or more clusters. It can then check historical records to determine which group is more likely to have higher acceptance of the offer. Based on that, the company may decide to mail the promotional material to a particular group of customers only.

FIGURE Q2

(18 marks)

Q3 a) Explain the goal of dimensionality reduction techniques.

(3 marks)

b) Are the two clusters shown in **Figure Q3(b)** are well separated? Justify your answer.

TERBUKA

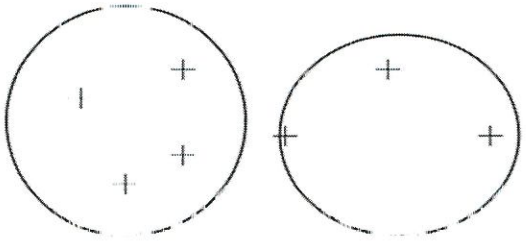

FIGURE Q3(b)

(3 marks)

Q4 Consider the following data in **Table Q4.** Use Euclidean Distance directly on the dataset to determine how far are B, C, and D from A. Who is the closest and who is the furthest? Please show all your calculations.

TABLE Q4: Customer Dataset

Customer	Age	Salary	Number of Children
A	26	51000	0
B	51	57000	3
C	33	60500	1
D	42	52500	2

(8 marks)

Q5 Based on the following scenario in **Figure Q5 ,** draw and label a diagram of a neural network architecture by considering the optimal number of weights for the network model is 48

Malaysian rubber industry is having a major problem in detecting symptoms of disease cause by fungus that destroys thousands hectares of Malaysia rubber plantings every year. Only 10,000 raw data had been collected manually in order to diagnose the symptoms whether infected, non-infected or neutral. Some major attributes that express the disease symptoms had also been identified as follows:

(1) The aggressiveness of fungus
(2) Size of fungus
(3) Type of fertilizer
(4) Humidity
(5) location

You have been chosen by Malaysia palm oil industry as data mining expert to solve their problem by using a neural network

FIGURE Q5

(8 marks)

TERBUKA

Q6 Consider the following dataset with 4 transactions given in **Table Q6**. Each line corresponds to one transaction. Assume that Support count (σ) = 2 and minConf = 80%. Using an Apriori method, construct and find all the association rules that satisfy the thresholds.

TABLE Q6: Transaction Dataset

Transactions	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

(24 marks)

Q7 Given a LaptopSales.csv dataset (refer to the uploaded file in Author UTHM) which contains the details of the sales for some laptops at a number of stores, during the first 10 days of January 2008. Provide the R codes for all the following steps. Determine the output generated from each of the step.

a) Set the working directory, and import the LaptopSales.csv dataset, give it a proper name.

(4 marks)

b) What is the average price of a laptop?

(4 marks)

c) What is the average price of a laptop with RAM.GB = 2?

(4 marks)

d) What is the configuration type with the highest price?

(4 marks)

e) How many laptops have a retail price greater than 600?

(4 marks)

f) What is the total value in $, of all the laptops sold in this dataset? (In other words, what is the sum of all the prices)

(4 marks)

TERBUKA

- END OF QUESTIONS –