Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2022/2023

COURSE NAME	:	DATA MINING
COURSE CODE	:	BIT 33603
PROGRAMME CODE	:	BIT / BIP
EXAMINATION DATE	:	FEBRUARI 2023
DURATION	:	3 HOURS

INSTRUCTION

1. ANSWER ALL QUESTIONS.
2. THIS FINAL EXAMINATION IS CONDUCTED VIA **CLOSED BOOK.**
3. STUDENTS ARE **PROHIBITED** TO CONSULT THEIR OWN MATERIAL OR ANY EXTERNAL RESOURCES DURING THE EXAMINATION CONDUCTED VIA CLOSED BOOK.

THIS QUESTION PAPER CONSISTS OF **FOUR (4)** PAGES

Q1 Differentiate between:

(a) discrete and continuous data.

(4 marks)

(b) histogram and boxplot.

(4 marks)

Q2 Discuss the life cycle of data mining projects.

(10 marks)

Q3 Given the following data (in increasing order) for the attribute `age`:

```
13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33,
35, 35, 35, 35, 36, 40, 45, 46, 52, 70
```

(a) Use the min-max normalization to transform the value `35` for `age` onto the range [0.0,1.0].

(5 marks)

(b) Use z-score normalization to transform the value `35` for `age`, where the standard deviation of `age` is 12.94 years.

(5 marks)

Q4 Based on Figures **Q4(i)**.and **Q4(ii)**, answer **Q4(a) – Q4(c)**. Figures **Q4(i)** and **Q4(ii)** represent two different data mining tasks.

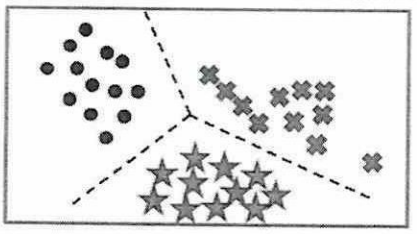

Figure Q4(i)

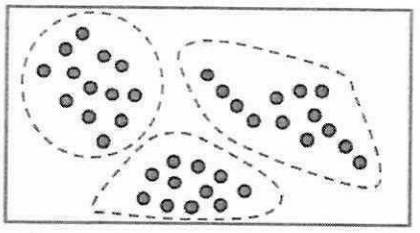

Figure Q4(ii)

(a) Discuss the goal of each data mining task.

(4 marks)

(b) Suggest the type of data needed for each task.

(4 marks)

(c) Suggest one suitable algorithm/technique to be used for each task.

(4 marks)

Q5 (a) Calculate the impurity of each node in **Figure Q5(a)** using:

(i) Gini index

(ii) Entropy

Node N1	Count
Class = 0	0
Class = 1	6

Node N2	Count
Class = 0	1
Class = 1	5

Node N3	Count
Class = 0	3
Class = 1	3

Figure Q5(a)

(12 marks)

(b) **Figure Q5(b)** shows a decision tree for the Financial Status of employees at ABC company. Construct **FIVE (5)** classification rules from the tree.

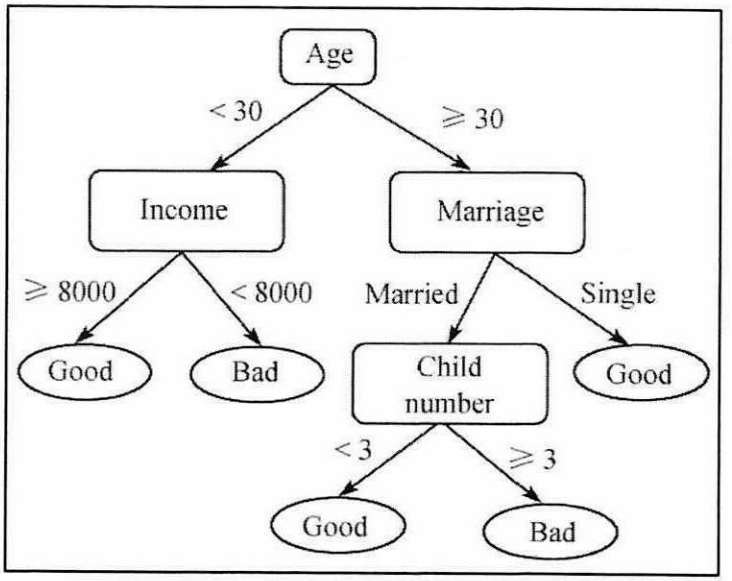

Figure Q5(b)

(10 marks)

Q6 **Figure Q6** shows the attributes of Concrete data taken from UCI Machine Learning Repository. Sketch and label the input/output mapping for the prediction of Concrete compressive strength using a Multilayer Perceptron.

```
Name / Data Type / Measurement Unit / Description
-----------------------------
Cement (component 1) -- kg in a m3 mixture
Blast Furnace Slag (component 2) -- kg in a m3 mixture
Fly Ash (component 3) -- kg in a m3 mixture
Water (component 4) -- kg in a m3 mixture
Superplasticizer (component 5) -- kg in a m3 mixture
Coarse Aggregate (component 6) -- kg in a m3 mixture
Fine Aggregate (component 7) -- kg in a m3 mixture
Age (component 8) -- Day (1-365)
Concrete compressive strength -- MPa
```

Figure Q6

(10 marks)

Q7 Based on **Table Q7**, answer **Q7(a) – Q7(c)**. Using the Training set in **Table Q7** and the Euclidean distance.

Table Q7

Attribute 1	Attribute 2	Class
0.8	6.3	-
2.1	7.4	-
2.6	14.3	+
6.8	12.6	-
8.8	9.8	+
9.2	11.6	-
10.8	9.6	+
11.8	9.9	+
12.8	1.1	-
14.2	18.5	-
15.6	17.4	-

(a) Calculate the distance of each instance in the Training set from the Unseen set, which consists of `Attribute 1: 9.1`, and `Attribute 2: 11.0`, respectively.

(12 marks)

(b) Calculate the optimum `k` values for the dataset. Explain your answer.

(6 marks)

(c) Using the `k` value from **Q7(b):**

(i) examine and identify the nearest neighbours for the Unseen set.

(6 marks)

(ii) what is the class label for the Unseen set.

(4 marks)

-END OF QUESTIONS-