UTHM
Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2018/2019

COURSE NAME	:	DATA MINING
COURSE CODE	:	BIT 33603
PROGRAMME CODE	:	BIT
EXAMINATION DATE	:	DECEMBER 2018 /JANUARY 2019
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ALL** QUESTIONS

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

Q1 Explain each issue in data quality as follows and give a solution to resolve the issue.

(a) Noise and outliers

(5 marks)

(b) Missing values

(5 marks)

Q2 The age values for the data tuples are (in increasing order) 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 min-max normalization to transform the value 35 for age onto the range [0:0;1:0].

(5 marks)

(b) Use normalization by decimal scaling to transform the value 35 for age.

(5 marks)

Q3 **Figure Q3** shows instances in a dataset.

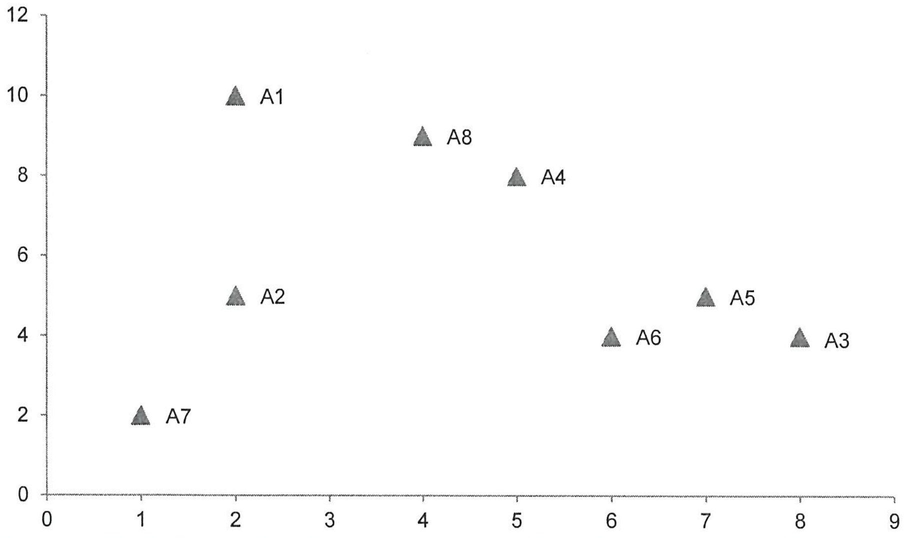

Figure Q3: Dataset

(a) Construct a distance matrix for the above instances using the following formula for Euclidean distance.

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$

(5 marks)

(b) Use the *k*-means algorithm and the Euclidean distance constructed in **Q3 (a)** to cluster the instances into 3 clusters with A1, A4 and A7 as the initial centroids for only the first epoch.

(10 marks)

Q4 Based on the following scenario:

A study on customer expenses is conducted and a dataset is given in Table 1. The study shows either customer will buy a house or not. Summary of the entropy calculation for root node is tabulated in Table 2.

Table 1: Customer dataset

ID	Age	Income	Government employee	Credit rating	Buy House
1	<=30	high	No	Fair	No
2	<=30	High	No	Good	No
3	31…40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Good	No
7	31…40	Low	Yes	Good	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<=30	Medium	Yes	Good	Yes
12	31…40	Medium	No	Good	Yes
13	31…40	High	Yes	Fair	Yes
14	>40	Medium	No	Good	no

Table 2: Entropy information for root node

Attribute	Average Entropy
Age	0.6935
Income	0.9110
Government Employee	0.7885
Expense history	0.8922

(a) Construct a decision tree for Table 1 based on entropy.

(15 marks)

(b) Convert the decision tree in **Q4(a)** to production rules.

(10 marks)

(c) What is the result of an old government employee with fair credit expenditure?

(5 marks)

Q5 How data mining techniques can be used in Web Mining? Give **TWO (2)** examples to support the answer.

(5 marks)

- END OF QUESTION –