



**UNIVERSITI TUN HUSSEIN ONN MALAYSIA**

**FINAL EXAMINATION  
SEMESTER I  
SESSION 2019/2020**

COURSE NAME : DATA MINING  
COURSE CODE : BIT 33603  
PROGRAMME CODE : BIT  
EXAMINATION DATE : DECEMBER 2019 / JANUARY 2020  
DURATION : 3 HOURS  
INSTRUCTION : A) ANSWER ALL QUESTIONS  
B) PLEASE WRITE YOUR  
ANSWERS IN THIS QUESTION  
BOOKLET

**TERBUKA**

THIS QUESTION PAPER CONSISTS OF SEVEN (7) PAGES

**Q1** State whether each of the following activities is a data mining task. Explain your answer.

- (a) Dividing the customers of a company according to their profitability. (2 marks)

**Answer:**

- (b) Computing the total sales of a company. (2 marks)

**Answer:**

- (c) Sorting a student database based on student identification numbers. (2 marks)

**Answer:**

- (d) Predicting the outcomes of tossing a (fair) pair of dice. (2 marks)

**Answer:**

- (e) Predicting the future stock price of a company using historical records. (2 marks)

**Answer:**

**TERBUKA**

Q2 Table 1 shows a dataset for making decision to buy computer.

Table 1: Buy Computer Dataset

ID	Age	Income	Student	Credit Rating	class: buy_computer
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

- (a) Build a decision tree using Information Gain as the attribute selection measure. The entropy for the root node is given in Table 2.

Table 2: Entropy for Root Node

Attribute	Average Entropy
Age	0.6935
Income	0.9110
Student	0.7885
Credit Rating	0.8922

(20 marks)

Answer:



- (b) Predict the class of the following new example using the decision tree answered in **Q2(a)**.

age $\leq$ 30, income=medium, student=yes, credit\_rating=fair.

(5 marks)

**Answer:**

- (c) Predict the class of the following new example using Naïve Bayes

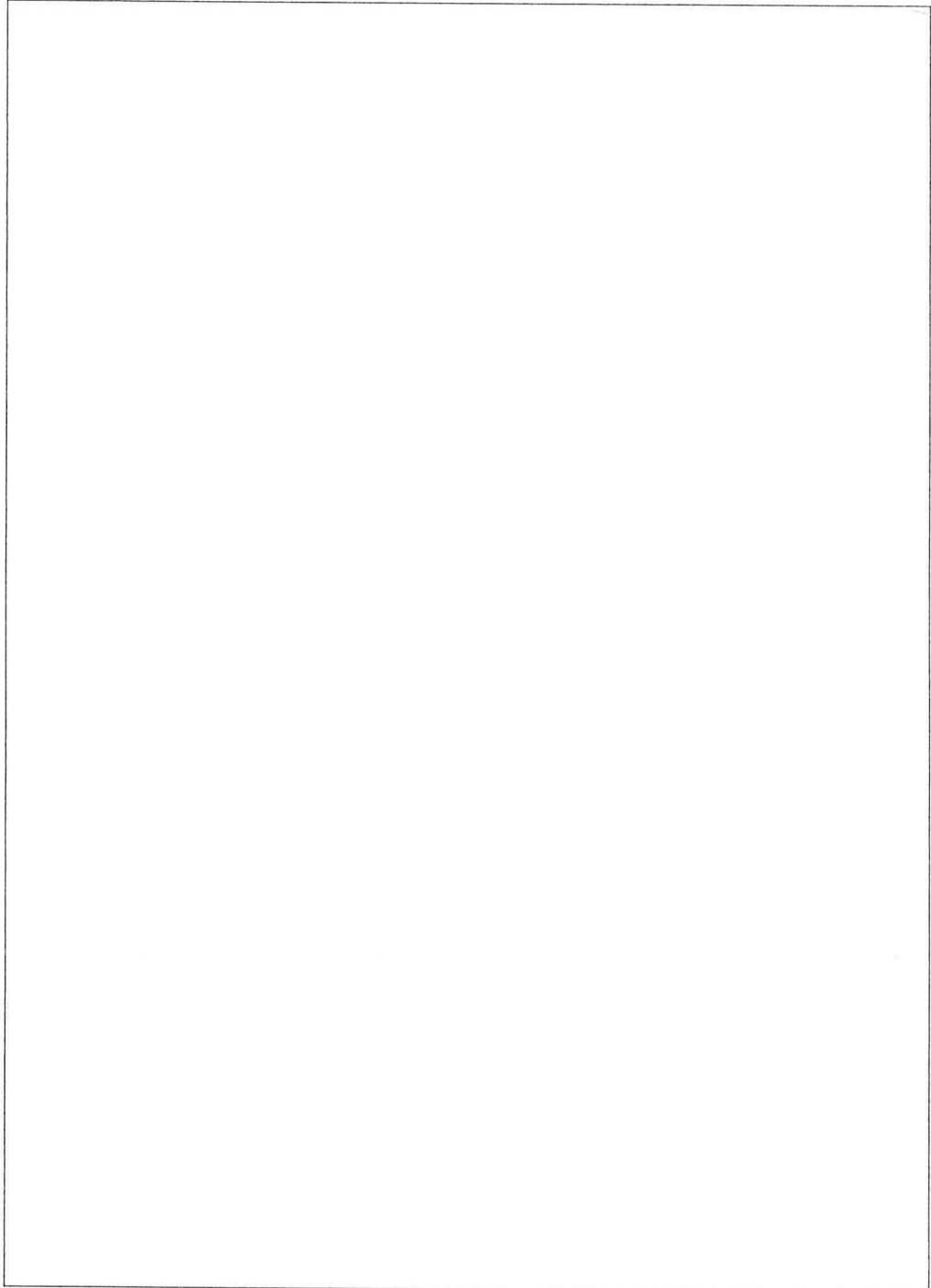
TERBUKA<sup>4</sup>

classification:

age<=30, income=medium, student=yes, credit\_rating=fair.

(20 marks: AN)

**Answer:**



**Q3** Figure **Q3** shows a distance matrix of a dataset. Suppose the initial seeds are  $A_1$ ,  $A_4$ , and  $A_8$ . Show the new clusters based on the  $k$ -means algorithm for 1 epoch

TERBUKA<sup>5</sup>

only.

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

FIGURE Q3

(20 marks)

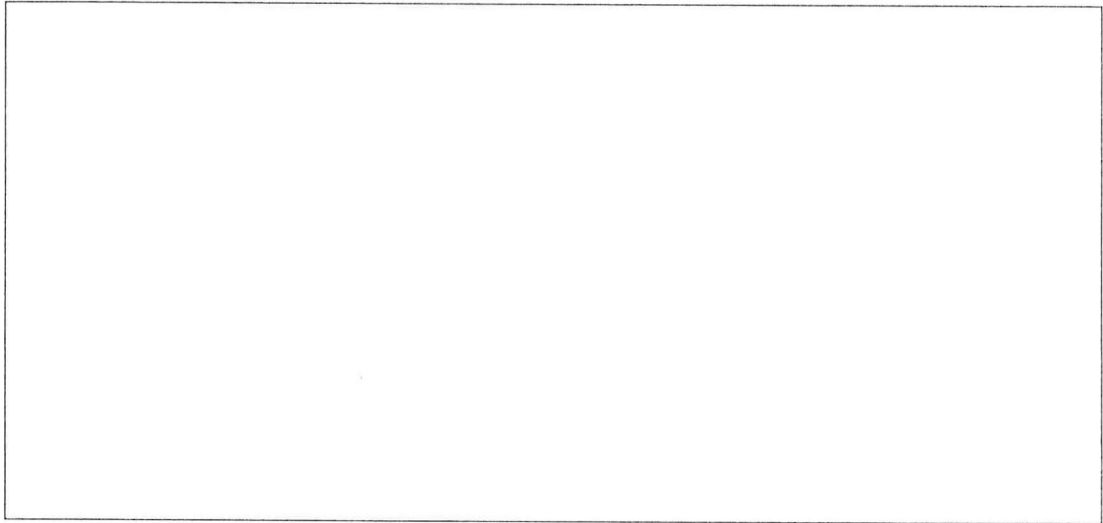
Answer:

**Q4** Outline the major research challenges of data mining in one specific application domain, such as stream/sensor data analysis, spatio-temporal data analysis, or bioinformatics. Choose **ONE (1)** domain only.

TERBUKA  
6

(5 marks)

**Answer:**



**- END OF QUESTIONS -**

**TERBUKA**

**CONFIDENTIAL**