FINAL

ALTERNATIVE ASSESSMENT

Examination Paper

(COVER PAGE)

Session	:	**April 2022**
Programme	:	Diploma In Computer Science (DCS)
Course	:	**DCS2106 : Data Mining**
Date of Examination	:	August 1, 2022 (Monday)
Time	:	12.00pm – 2.30pm Reading Time : Nil
Duration	:	2 Hours : 30 Minutes

Note: 30 minutes is added into the duration of the examination to factor in any connectivity matters and for you to scan and upload your scripts.

Special Instructions :

This paper consists of **FOUR (4)** questions. Answer **ALL FOUR (4)** questions

Materials permitted : Non-Programmable Scientific Calculator

Materials provided : Nil

Examiner(s) : **Ms Yogeswari Suppiah** and Dr Fakhitah Ridzuan

Chief Moderator : Ms Siti Hajar Khairuddin

*This paper consists of **6** printed pages, including the cover page*

DIPLOMA IN COMPUTER SCIENCE PROGRAMME (DCS)
DCS2106 : DATA MINING
FINAL ALTERNATIVE ASSESSMENT : APRIL 2022 SESSION

Instruction: This paper consists of **FOUR (4)** questions. Answer **ALL** the questions.

Question 1 (25 marks)

(a) Suppose you are employed as data mining consultant for a bakery. Describe how data mining can help the bakery to get more sales by applying data mining technique.

(4 marks)

(b) Data mining can be divided into supervised and unsupervised learning. Provide **THREE(3)** differences between these two.

(6 marks)

(c) Draw a contingency table for each of the following rules using the transactions in table below.

(5 marks)

Transaction ID	Items Bought
1	*{a, b, d, e}*
2	*{b, c, d}*
3	*{a, b, d, e}*
4	*{a, c, d, e}*
5	*{b, c, d, e}*
6	*{b, d, e}*
7	*{c, d}*
8	*{a, b, c}*
9	*{a, d, e}*
10	*{b, d}*

Rules: *{b}* → *{c}*, *{a}*-→ *{d}*, *{b}* → *{d}*, *{e}* → *{c}*, *{c}* → *{a}*.

(b) A database has 5 transactions. Let min sup = 60% and min conf = 80%

TID	*items bought*
T100	M, O, N, K, E, Y
T200	D, O, N, K, E, Y
T300	M, A, K, E
T400	M, U, C, K, Y
T500	C, O, O, K, I ,E

i. Find all frequent itemsets using Apriori.

(7 marks)

ii. In your opinion, between Apriori and FP-Growth, which one is more efficient for this type of data? Explain your answer. (3 marks)

Question 2 (25 marks)

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

(a) Consider the following dataset for a binary classification problem.

(i) Show the contingency tables after splitting on attributes.

(4 marks)

(ii) Calculate the information gained when splitting on *A* and *B*.

(6 marks)

(iii) Which attribute would the decision tree induction algorithm choose?

(2 marks)

(b) Why is *tree pruning* useful in decision tree induction?

(3 marks)

(c) What is a drawback of using a separate set of tuples to evaluate pruning? (4 marks)

(d) Explain **TWO (2)** possible reasons for model overfitting.

(6 marks)

Question 3 (25 marks)

(a) Data can be divided into two categories, which are qualitative and quantitative. What are the differences between these two data types?

(4 marks)

(b) What is the attribute type of the following example?

(3 marks)

i . Temperature in a room

ii . Type of living accommodation: House, Apartment etc

iii . Socioeconomic status: poor, middle, class, rich

(c) What is outlier? How is it different from noise?

(3 marks)

(d) Define clustering?

(3 marks)

(e) Briefly describe and give examples of **TWO (2)** type of clustering approaches.

(6 marks)

(f) You are given a data set with 100 records and are asked to cluster the data. You use K-means to cluster the data, but for all values of K, $1 \leq K \leq 100$, the K-means algorithm returns only one non-empty cluster. You then apply an incremental version of K-means, but obtain exactly the same result. How is this possible?

(2 marks)

(g) Based on your answer in f), explain how would single link and DBSCAN handle such data?

(4 marks)

Question 4 (25 marks)

(a) Consider a binary classification problem with the following set of attributes and attribute values:

- *Air Conditioner = {Working, Broken}*
- *Engine = {Good, Bad}*
- *Mileage = {High, Medium, Low}*
- *Rust = {Yes, No}*

Suppose a rule-based classifier produces the following rule set:

Mileage = High -⟶ Value = Low
Mileage = Low -⟶ Value = High
Air Conditioner = Working, Engine = Good -⟶ Value = High
Air Conditioner = Working, Engine = Bad -⟶ Value = Low
Air Conditioner = Broken -⟶ Value = Low

i. Are the rules mutually exclusive?

(1 mark)

ii. Is the rule set exhaustive?

(1 mark)

iii. Is the ordering needed for this set of rules? Justify your answer.

(2 marks)

iv. Do you need a default class for the set? Why is it so?

(2 marks)

(b) For each of the Boolean functions given below, state whether the problem is linearly separable (4 marks)

i . A AND B AND C

ii . NOT A AND B

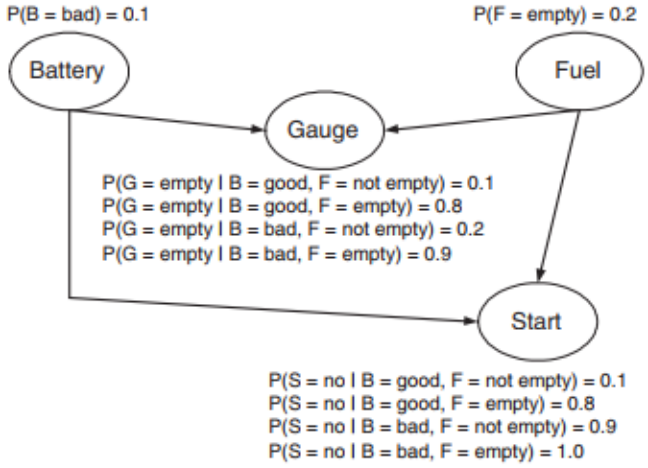
iii . (A OR B) AND (A OR C)

iv . (A XOR B) AND (A OR B)

(c) Suppose the fraction of undergraduate students who smoke is 5% and the fraction of graduate students who smoke is 13%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?

(5 marks)

(d) Figure below illustrates the Bayesian network, compute the following probabilities:



i . P(B = good, F = empty, G = empty, S = yes) (3 marks)

ii . P(B = bad, F = empty, G = not empty, S = no) (3 marks)

iii . Given that the battery is bad, compute the probability that the car will start. (4 marks)

-THE END-