Lecturer:	(Date)	Approved by:	(Date)
(Signature & Fullname)		(Signature, Position & Fullname)	

(The above part must be hidden when copying for exam)

UNIVERSITY OF TECHNOLOGY - VNUHCM FACULTY OF CSE	FINAL EXAM	Semester/Academic year	1	2022-2023
		Date	24/12/2022	
	Course title	Data mining		
	Course ID	CO3029		
	Duration	70 mins.	Question sheet code	

Notes:
- *Open book*
- *Do not use mobile phones, laptops or any electronic device*
- *7 pages: 30 multiple choice questions and 6 short writing question*
- ***Submit the question sheet together with the answer sheet***

Student Full name	
Student ID number	

ANSWER SHEET

SECTION 1 (7.0 points)

Code 100

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Answer	C	E	A	C	D	E	D	E	E	B	D	D	C	B	A
Question	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Answer	B	C	D	B	B	A	B	C	C	B	B	A	C	D	B

Code 200

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Answer	A	C	C	E	D	E	D	E	E	B	D	D	C	B	A
Question	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Answer	B	C	D	B	B	A	B	C	C	B	B	A	C	D	B

Part2 (3.0 points): short written questions

31. (L.O.3.2, 1.0 points) Given following three data points/objects **P1(3, 1, 2); P2(0, 2, 1); P3(3, 0, 5); P4(1, 1, 1); P5(4, 2, 2)**. If we use K-Means with k=2 and Euclidean distance to measure the similarity between two data objects. Let's initiate two centroids as C1(1, 0, 0) and C2(3, 0, 0).

a) (0.5 point) Write down data points in each cluster

C1={P2, P4}; C2={P1, P3, P5}

b) (0.5 point) What is the Centroid of each cluster?

C1(0.5, 1.5, 1), C2(3.33, 1, 3)

32. (L.O.3.3, 0.5 points) Given a classifier M built to classify images which are labelled in as "dog" or "cat". Let M works on a data set D of 10 "cat" images and 4 others, and it recognizes 9 cat's images. However, among those 9 images there are 6 images are correct while 3 incorrect ones come from other images.
Write the expressions and calculte following measures: TP (true positive), FP (false positive), FN (false negative), TN (true negative), P (precision), R (recall) and F_score.

TP=6, FP=3, FN=4, TN=1

=> P=6/(6+3) =6/9=0.667; R=6/(6+4)=6/10=0.6; F_score=(2*6/9*6/10)/(6/9+6/10)=0.63

33. (L.O.3.2, 1.0 point) Given a data set *D* as following table

RID	Tuoi	Thu_nhap	Sinh_vien	Tin_dung	Mua_may_tinh
1	tre	cao	no	kha	khong_mua
2	tre	cao	no	tot	khong_mua
3	trung	cao	no	kha	mua
4	cao	trung_binh	no	kha	mua
5	cao	thap	yes	kha	mua
6	cao	thap	yes	tot	khong_mua
7	trung	thap	yes	tot	mua
8	tre	trung_binh	no	kha	khong_mua
9	tre	thap	yes	kha	mua
10	cao	trung_binh	yes	kha	mua
11	tre	trung_binh	yes	tot	mua
12	trung	trung_binh	no	tot	mua
13	trung	cao	yes	kha	mua
14	cao	trung_binh	no	tot	khong_mua

a) (0.25 point) Write the expression to calculate the $Info_A(D)$ (Information gain of attribute A) in the decision tree method using Information gain as measure to select the spliting attribute.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$

b) (0.75 point) Write down the expression and calculate the value of $Gain_{Tuoi}$

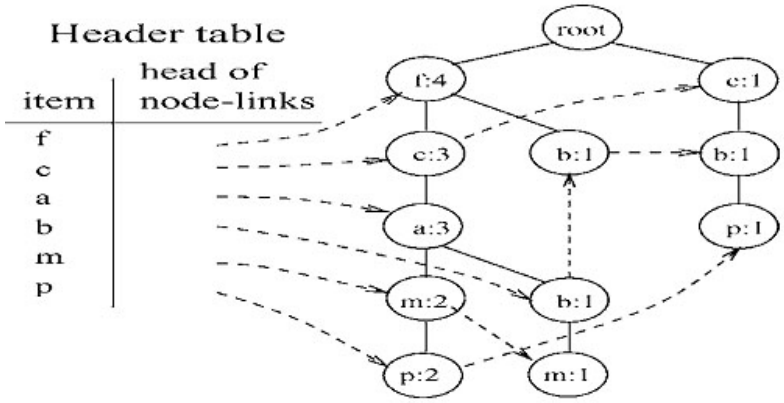
$$Info(D) = -\left(\frac{9}{14} log_2\left(\frac{9}{14}\right) + \frac{5}{14} log_2\left(\frac{5}{14}\right)\right) = 0.94 \qquad (4.5)$$

$$Info_{Tuoi}(D) = \frac{5}{14}\left(-\frac{2}{5} log_2 \frac{2}{5} - \frac{3}{5} log_2 \frac{3}{5}\right) + \frac{4}{14}\left(-\frac{4}{4} log_2 \frac{4}{4} - \frac{0}{4} log_2 \frac{0}{4}\right) + \frac{5}{14}\left(-\frac{3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5}\right) = 0.694 \qquad (4.6)$$

$$Gain_{Tuoi} = Info(D) - Info_{Tuoi}(D) = 0.94 - 0.694 = 0.246 \qquad (4.7)$$

34. (L.O.3.2, 0.5 points) Give a dataset bellow, draw the FP-tree from the above dataset with min_sup = 3?

TID	Items bought
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f , h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n



--- END ---