UNIVERSITY OF INFORMATION TECHNOLOGY
Faculty of Information Systems

FINAL EXAMINATION
Semester I
Academic year: 2022-2023
Subject: Data mining
During: 90 minutes

***(Students are allowed to use 1 A4 paper of material*)**

Question 1: (2.0 scores) Select ONE of the following questions:

1. Present 2 difficult problems when collecting real data for data mining issue. Give examples to clarify your ideas.
2. Give one example of data mining application in the field of ***logistics***. Based on your example, what kind of data, and data mining method you can use?

Question 2: (6.0 scores)

Suppose that a "**Ready to Test ChatGPT"** dataset as in the following table (Let *Test ChatGPT* be the decision attribute).

*Note: Students can use abbreviations (for example: A for **Age**) to present the examination.*

	Academic degree (AD)	**Occupation (O)**	**Gender (G)**	**Test ChatGPT (TC)**
1	Doctor	Lecturer	Male	Yes
2	Master	Researcher	Female	No
3	Bachelor	Programmer	Male	Yes
4	Master	Lecturer	Female	Yes
5	Master	Lecturer	Male	Yes
6	Doctor	Researcher	Female	Yes
7	Doctor	Researcher	Male	Yes
8	Bachelor	Lecturer	Male	No
9	Bachelor	Programmer	Female	No
10	Bachelor	Programmer	Male	No

1. Let *min_sup=40%* and *min_conf =50%.* Using Apriori, find all frequent itemsets. Then, list all association rules of ONE maximal frequent itemset. *(1.5 scores)*
2. Suppose B = {*Occupation, Gender*}, X={1, 3, 4, 5, 6, 7} (**Test ChatGPT** = "Yes"). Use rough set to compute: upper approximation, lower approximation, and quality coefficient *(1.5 score)*
3. Determine the root of Decision Tree using Gini Index. *(1.5 scores)*
4. Given a sample X = *(Academic Degree = "Doctor", Occupation= "Lecturer", Gender**=** "Male")*, what would a Naïve Bayesian classification using Laplacican correction of the ***Test ChatGPT*** for sample X be? *(1.5 scores)*

Question 3: (2.0 scores)

Suppose that 8 points as: P1=(1, 4), P2=(5, 1), P3=(2, 6), P4=(8, 5), P5=(7, 5), P6=(4, 2), P7=(10, 4), P8=(3, 1). And the matrix M_0 is:

M_0	P1	P2	P3	P4	P5	P6	P7	P8
C1	1	0	0	0	0	0	0	0
C2	0	1	0	0	0	0	0	0
C3	0	0	1	1	1	1	1	1

Cluster the data to 3 clusters using K_means algorithm and Euclidean distance.

Note: Show 3 steps:

- *Step 1: calculate center of each cluster.*
- *Step 2: calculate distances.*
- *Step 3: show matrix* M_1.

END

Ho Chi Minh city, January 30th, 2023.

Faculty of Information Systems

Lecturer