THE UNIVERSITY OF THE WEST INDIES
ST. AUGUSTINE

EXAMINATIONS OF DECEMBER 2021 (Alternative Assessment Exam)

Code and Name of Course: **COMP 3605 - Introduction to Data Analytics** Paper: 1

Date and Time: **December 13, 2021 from 9 a.m. to 12 Noon** Duration: **3 hours**

INSTRUCTIONS TO CANDIDATES: This paper has **4** pages and **4** questions

(This is hand-written online alternative assessment. It is to replace the 2-hour in-class final exam)

Answer ALL Questions

INSTRUCTIONS
1. Type or write your answers neatly.
2. Show all working of your answers.
3. Your solutions must be your own. You must not share your working or solutions with your peers.
4. You are not permitted to copy, summarize, or paraphrase the work of others in your solutions.
5. Submit your answers in a single zipped file named FinalExam_ID.zip via myElearning, where ID is replaced with your student ID. The file FinalExam_ID.zip contains
- a single PDF file containing all of your handwritten, typed, and screenshots answers.
- a signed and dated UWI Plagiarism Declaration indicating that the work submitted is your own.

PLEASE TURN TO THE NEXT PAGE

Total Mark: 50

Question 1 You are given a training dataset *D* shown in the table below for a classification problem. The class label attribute Diagnosis has three different values {Strep throat, Cold, Allergy}.

The class-labeled training dataset *D* for Disease Diagnosis

Patient ID	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

a. Compute the information gain for the attribute **Fever**. [4 marks]
b. Compute the gain ratio for the attribute **Congestion** using *Gain*(Congestion) = 0.446. [3 marks]
c. Compute the Gini index for the attribute **Swollen Glands**. [3 marks]

[Total mark: 10]

Question 2 You are given the following data set *D* containing $n = 6$ instances in Euclidean space

The data set *D*

Instance	*x*	*y*
1	1.2	1.5
2	1.1	4.3
3	2.5	1.5
4	2.3	3.5
5	2.5	3.5
6	4.5	5.5

Show the results of the first two iterations of the *K*-means algorithm to partition the given data set *D* into two clusters C_1 and C_2, where $K = 2$, the two randomly selected centroids are the instances 1 and 3 (i.e., $\boldsymbol{c}_1 = (1.2, 1.5)$, $\boldsymbol{c}_2 = (2.5, 1.5)$).

[Total mark: 12]

PLEASE TURN TO THE NEXT PAGE

Question 3 You are given the transactional database *D* shown in the table below. The database has five transactions. Let *min_sup* = 60%.

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}

Find all frequent itemsets in *D* using the horizontal Apriori algorithm.

[Total mark: 15]

Question 4

a. The following contingency table summarizes supermarket transaction data, where *hot dogs* refers to the transactions containing hot dogs, $\overline{hot\ dogs}$ refers to the transactions that do not contain hot dogs, *hamburgers* refers to the transactions containing hamburgers, and $\overline{hamburgers}$ refers to the transactions that do not contain hamburgers.

	hot dogs	$\overline{hot\ dogs}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
Σ_{col}	3000	2000	5000

i. Suppose that the association rule *hot dogs* → *hamburgers* is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

[3 marks]

ii. Based on the given data, is the purchase of *hot dogs* independent of the purchase of *hamburgers*? If not, what is the correlation between the two items hot dogs and hamburgers? [3 marks]

PLEASE TURN TO THE NEXT PAGE

b. The following multinomial logistic regression model predicts the TYPE of a retail customer (*single*, *family*, or *business*) based on the average amount that they spend per visit, SPEND, and the average frequency of their visits, FREQ:

$h_{\mathbf{w}_{single}}(\mathbf{q}) = g(0.7993 + (-15.9030) \times \text{SPEND} + 9.5974 \times \text{FREQ})$

$h_{\mathbf{w}_{family}}(\mathbf{q}) = g(3.6526 + (-0.5809) \times \text{SPEND} + (-17.5886) \times \text{FREQ})$

$h_{\mathbf{w}_{business}}(\mathbf{q}) = g(4.6419 + 14.9401 \times \text{SPEND} + (-6.9457) \times \text{FREQ})$

where g is the logistic function $g(x) = 1 / (1 + e^{-x})$.

Use the given logistic regression model to predict the following query instance

$$\mathbf{q} = (\text{SPEND}, \text{FREQ}) = (-0.43, -0.71).$$

[7 marks]
[Total mark: 13]

End of Question Paper