Number of pages	5
Number of questions	5

UNIVERSITY OF SRI JAYEWARDENEPURA
Faculty of Technology

Bachelor of Information and Communication Technology Honours Degree

Third Year Second Semester End Examination

March 2022

ITS3173 Data Mining and Warehousing

me allowed: Three (03) Hours — Answer ALL questio

Read and follow the instructions given below:

- **"Index number" and course code "ITS3173"** should be written on top of each page of the answer script, and pages must be **numbered appropriately**.

Question 1 (Total 20 Marks)

i. 'Data mining is a combination of multiple disciplines.' Justify this statement with four (4) reasons. (6 Marks)

ii. Illustrate the central tendency measures of following *a,b,c* scenarios using three (3) diagrams.

a. Median, Mean and Mode of symmetric distribution. (1 Marks)

b. Median, Mean and Mode of positively skewed distribution. (1 Marks)

c. Median, Mean and Mode of negatively skewed distribution. (1 Marks)

iii. 'Data retrieving is not data mining.' Explain this statement. (6 Marks)

iv. Describe the importance of data visualization providing five (5) reasons. (5 Marks)

Question 2

(Total 20 Marks)

i. ' Discrepancy detection is the first step in a data cleaning process.'

   a. List down four (4) factors that can cause discrepancies in the data cleaning process. (4 Marks)

   b. Describe two (2) techniques that can aid in the discrepancy detection step. (4 Marks)

ii. 'High Dimensionality of a dataset is a major challenge of any data mining project.' Briefly explain how data mining has addressed this challenge. (4 Marks)

iii. List down five (5) data discretization methods (5 Marks)

iv. Consider the following dataset for an attribute price in a data analysis activity. Smooth this dataset using *smoothing by bin boundaries* technique. Assume a bin-depth of three.

4, 8, 15, 21, 21, 24, 25, 28, 34 (3 Marks)

Question 3

(Total 20 Marks)

- Following dataset in Table 1 has five (5) transactions of a supermarket. Answer *parts i, ii, iii*, and iv based on this dataset.

Table 1

Transaction ID	Items bought
1	Tea, Cake, Soft Drink
2	Eggs, Tea, Soft Drink
3	Tea, Cake, Soft Drink, Coffee, Milk
4	Tea, Cake, Milk, Eggs, Soft Drink
5	Cake, Coffee, Soft Drink, Milk, Eggs

i. Calculate the **absolute support** for itemset {Tea}, itemset {Cake}, and itemset {Soft Drink} respectively.

ii. Calculate the **relative support** for itemset {Soft Drink}, itemset {Coffee}, itemset {Eggs}, itemset {Milk} respectively. (3 Marks)

iii. Complete the following **association rules** by calculating support and confidence. (4 Marks)

a. { Tea, Cake } → { Soft Drink }: (s = ..., c =.....) (2 Marks)

b. { Cake, Milk } →{Eggs}: (s = ..., c =.....) (2 Marks)

c. {Tea}→ {Cake}: (s = ..., c =.....) (2 Marks)

iv. List all the frequent itemsets in dataset of Table 2, using the Apriori Algorithm. Assume that the minimum support is 2. Show the steps clearly.

(7 Marks)

Table 2

Transaction_ID	Items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E

Question 4

(Total 20 Marks)

- Following Table 3 dataset D indicates buying behaviour of a computer based on several attributes.
  - Class P: buy_computer = "yes"
  - Class N: buy_computer = "no"

Using the following Table 3 dataset and formulas attached in annexure, answer the *parts i, ii, iii, iv, v.*

Table 3

age	income	student	credit_rating	buy_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no

31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

i. Calculate the Entropy - ***Info(D)*** for the above dataset D showing the steps clearly. (4 Marks)

ii. Calculate entropy for each value of attribute **income** E(income=high), E(income=medium), E(income=low) using above dataset D, clearly showing the steps. (4 Marks)

iii. Compute the **average information entropy** - $Info_A(D)$ for attribute **income**. (3 Marks)

iv. If the **information gain** ***Gain(A)*** for **student** attribute is given as Gain (student)=0.151 and for **credit_rating** is given as Gain (credit_rating)=0.048, determine the best splitting attribute to be used in creating a decision tree for dataset D, using Information Gain. Justify your selection (3 Marks)

v. Using the above data set D, calculate the **Gain Ratio (income)** (6 Marks)

Question 5 (Total 20 Marks)

i. Define Cluster Analysis and state the importance using a practical example. (4 Marks)

ii. Briefly explain how clustering can be used for outlier detection. (4 Marks)

iii. 'Even though clustering has a long history with large numbers of clustering algorithms in Data Mining and other fields, significant challenges still remain.' Briefly explain four (4) such challenges. . (4 Marks)

iv. Explain four (4) major Clustering approaches with their characteristics. (8 Marks)

Annexure

Entropy calculation formula

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Simplified above formula for entropy calculation

$$Entropy = \frac{-p}{p+n} log_2(\frac{p}{p+n}) - \frac{n}{p+n} log_2(\frac{n}{p+n})$$

Average information entropy

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

Simplified above formula for average information entropy

$$I(income) = \frac{P(high) + n(high)}{p+n} Entropy\ (income = high)$$
$$+ \frac{P(medium) + n(medium)}{p+n} Entropy\ (income = medium)$$
$$+ \frac{P(low) + n(low)}{p+n} Entropy\ (income = low)$$

Information Gain formula

$$Gain(A) = Info(D) - Info_A(D)$$

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$$

Gain Ratio

- GainRatio(A) = Gain(A)/SplitInfo(A)

****** End of the Paper******