SOUTH EASTERN UNIVERSITY OF SRI LANKA

THIRD EXAMINATION IN BACHELOR OF INFORMATION AND COMMUNICATION TECHNOLOGY - 2021/2022

SEMESTER - I, SEPTEMBER/OCTOBER - 2024

CIS 31041 - Practical for Data Mining

Answer All (04) Questions **Time Allowed: Three (03) hours**

INSTRUCTIONS TO CANDIDATES:

- Create a folder on the desktop with your index number. (Eg. **ICTxxx**)
- Create sub-folders named with **Q01, Q02, Q03,** and **Q04**.
- All answer files should be saved within the folder you created.
- All answer files should be named as per the instructions given in each question.
- Save your works frequently.
- To answer the Question 01 and Question 02 use the **Weka Explorer**.
- To answer the Question 03 and Question 04 use the **Anaconda Navigator** tool. Same as **Google Colab** platform.
- Search **Anaconda Navigator** in the Start Menu and click Anaconda Navigator tool. In the home screen, click **Launch** button of **Jupyter Notebook** icon. It opens in a browser screen. Select Desktop > Select folder with your index number > Select Q03 folder.
- Then, click **New** > select **Python 3.** It opens a new browser screen. It is the interface to write your codes.
- Create separate notebooks for Q03 and Q04. Rename the notebooks as **Q03** and **Q04**.
- To rename: Go to **File** > select **Rename...**
- Marks given in brackets are indicative of the weight given to each part of the question.
- All the **required dataset** files are available in the **Desktop** with folder name **CIS31041.** The Password to open the dataset file: **2019_2020_ICT**
- Compress the folder and submit it to the examiner.
- **Internet access is prohibited** during the exam.

Question 01:

You are given the task of analyzing a dataset containing information about employees and their performance ratings. Complete the following tasks in Weka:

a) Using the information in **Table 1**, manually create an **Employee.arff** file using a text editor. Ensure that the file follows the correct ARFF format. Save the file as **Employee.arff**.

(30 Marks)

b) Open the **Employee.arff** file in Weka Explorer. Identify and fill the missing values using a suitable approach. After processing, save the file as **Q01_b.arff** in your answer folder.

(20 Marks)

c) Focus on the **Performance_Rating** attribute and answer the following questions. Type your answers in a text file and save it as **Q01_c.txt**.

1. What is the type of this attribute?

(05 Marks)

2. How many ***unique values*** does this attribute contain? List them.

(10 Marks)

3. What is the most ***frequently occurring value***? How many times does it appear?

(10 Marks)

d) Visualize the **Department** against **Gender** in a bar chart. Save the image as **Q01_d.jpg**.

(05 Marks)

Table - 1 Employee Information

ID	Age	Gender	Department	Salary	Performance_Rating
1	29	Male	IT	60000	Excellent
2	34	Female	Marketing	75000	Good
3	28	Male	IT	55000	
4	40	Female	HR	80000	Excellent
5	32	Male	IT	70000	Good
6		Female	Marketing	68000	Good
7	29	Male	IT	61000	Poor
8	45	Female	HR		Excellent
9	31	Male	Marketing	65000	Poor
10	33	Female	IT	72000	Excellent

e) Open **Q01_b.arff** file and apply **discretization** to the **Salary** attribute. The requirement is **three(03)** distinct classes with **equal ranges**. Take the screenshot of nominally converted attribute in the Weka Explorer and save it as **Q01_e.jpg**. Finally, save the updated file as **Q01_e.arff** in the answer folder.

(20 Marks)

[100 Marks]

Question 02:

In this question, you are required to implement association rule mining and identify the best rules using the ***Apriori*** algorithm. You are provided with a retail dataset called **Retail_Transactions.arff**, which contains transactional data of customers. The dataset has **eight(08)** attributes describing different customer behaviors in a retail store.

a) Open the **Retail_Transactions.arff** file in Weka Explorer and perform all necessary **preprocessing** steps. Create a Word document named **Q02_a.docx** that thoroughly explains each preprocessing step taken, including relevant screenshots to illustrate your actions. Finally, save the processed file as **Q02_a.arff** in the answer folder.

(50 Marks)

b) Open the **Q02_a.arff** file in Weka Explorer and apply association rule mining using the ***Apriori*** algorithm. Use the following parameters and find the top **ten (10)** rules. Save the results as **Q02_b.txt** in the answer folder.

- **Minimum Support = 20%**
- **Confidence = 75%**

(20 Marks)

c) Select one rule from the best ten rules implemented in the previous step. Explain the rule in your own words. Save the answer as **Q02_c.txt** in the answer folder.

(10 Marks)

d) Modify the ***Apriori*** algorithm parameters as follows:

- **Minimum Support: 20%**
- **Confidence: 85%**

Re-apply the ***Apriori*** algorithm on the **Q02_a.arff** dataset and find the top **ten (10)** rules. Save the results as **Q02_d1.txt** in the answer folder. Compare the new set of rules with the existing rules and describe any changes in the rules. Write your observations in a text file named **Q02_d2.txt**.

(20 Marks)

[100 Marks]

Question 03:

This question evaluates your understanding of **k-means** clustering with the help of dataset related to famous tourist places. The dataset contains **seven(07)** attributes such as ***Tourist_Place_Id, Total_Reviews, Average_Stay_Duration_Hours, Entry_Fee_USD, Nearby_Hotels_Count, Accessibility_Rating,*** and ***Distance_From_City_Center_km.***

a) Import the necessary libraries.

(15 Marks)

b) Copy the dataset named **Famous_Tourist_Places.csv** file into the same folder **(Q03)**. And load the dataset into Jupyter notebook.

(05 Marks)

c) Select the suitable attributes from the dataset to cluster **popular tourist places** based on relevant characteristics. Remove all other attributes and display the first **ten(10)** records of the filtered dataset. Take a screenshot and name it as **Q03_c.jpg**.

(10 Marks)

d) Convert the filtered dataset into an **array** to prepare for clustering analysis.

(10 Marks)

e) Implement the **Elbow method** and determine the optimal number of clusters. Capture the screenshot of Elbow diagram and rename it as **Q03_e.jpg**

(20 Marks)

f) Apply the ***k-means++*** algorithm to cluster the dataset based on the selected attributes.

(20 Marks)

g) Plot the cluster results in a scatter plot between ***Total_Reviews*** and ***Accessibility_Rating***. Capture the screenshot and named it as **Q03_g.jpg**.

(20 Marks)

Note: Ensure that you attach **Q03.ipynb** file in your answer folder.

[100 Marks]

Question 04:

This question is designed to assess your ability to **preprocess** and **visualize** data. It focuses on handling missing values, and visualizing data to better understanding.

a) Import the necessary libraries and load the dataset **WeatherWaterLevel.csv** into your Jupyter Notebook.

(10 Marks)

b) Select appropriate **independent variables** from the dataset to classify the target variable **WeatherType**. Display the first **ten(10)** records of independent variable. Capture the screenshot and rename it as **Q04_b.jpg**.

(10 Marks)

c) Plot the "**MaxTemperature**" (X-axis) against "**WaterLevel**" (Y-axis) attribute. Capture the screenshot and rename it as **Q04_c.jpg**.

(10 Marks)

d) Develop a bar chart to display the **number of records** per "**WeatherType**". Take the screenshot and rename it as **Q04_d.jpg**.

(10 Marks)

e) Implement a scatter plot for "**Inflow**" (X-axis) against "**WaterLevel**" (Y-axis) attribute. Capture the screenshot and rename it as **Q04_e.jpg**.

(10 Marks)

f) Identify the attributes that contain missing values. Capture the screenshot and rename it as **Q04_f.jpg**

(10 Marks)

g) Fill the missing values with suitable approach.

(30 Marks)

h) Save the final dataset as **FinalDataset.csv** and include it in the answer folder.

(10 Marks)

Note: Ensure that you attach **Q04.ipynb** file in your answer folder.

[100 Marks]

[Total Marks: 400]