CIS 31041 - Practical for Data Mining

Continuous Assessment - 03

INSTRUCTIONS TO CANDIDATES:

- Create a folder on the desktop with your index number. (Eg. **ICTxxx**)
- Create sub-folders names **Q01,** and **Q02**.
- 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 files frequently.
- To answer Question 01 and Question 02 use the **Google Colab tool**.
- Marks given in brackets are indicative of the weight given to each part of the question.
- All the **required dataset** files are given in the folder named as **CIS31041.**

Question 01:

a) Import the **numpy**, **pandas** and **matplotlib** libraries.

(15 Marks)

b) Copy the dataset **Company_ABC_HumanResource.csv** file into the google colab. Load the dataset into colab notebook.

(10 Marks)

c) Display the first **five records** of the dataset and take a screenshot. Rename as **Q01_c.jpg**

(05 Marks)

d) Plot the "**Position**" (X axis) against **"Salary"** (Y axis) attribute. Capture the screenshot and rename as **Q01_d.jpg**

(10 Marks)

e) Develop a bar chart to find out the **number of employees** in different "**Department"**. Take the screenshot and rename as **Q01_e.jpg**

(10 Marks)

f) Implement the scatter plot for **"EmployerSatisfaction"** (X axis) against **"Salary"** (Y axis) attribute. Capture the screenshot and rename as **Q01_f.jpg**

(10 Marks)

g) Find out the attributes which have missing values. Capture the screenshot and rename as **Q01_g.jpg**

(10 Marks)

h) Fill the missing values of **"MaritalStatus"** attribute using the **mode value**.

(10 Marks)

i) Fill the missing values of **"Salary"** attribute using the **mean value**.

(10 Marks)

j) Save the dataset as **PreprocessedCompany_ABC_HumanResource.csv** and include in the answer folder. Also include the colab notebook in the answer folder.

(10 Marks)

[100 Marks]

Continuous Assessment - 04

Question 02:

a) Import the **numpy**, **pandas** and **matplotlib.pyplot** libraries.

(05 Marks)

b) Copy the dataset **Components_of_Fertilizer.csv** file into the google colab. Load the dataset into colab notebook.

(05 Marks)

c) Remove the **"Proline"** attribute from the dataset and display the first **five records** of the dataset. Then take a screenshot. Rename as **Q02_c.jpg**

(10 Marks)

d) Make an array of dataset to use in the upcoming clustering steps.

(10 Marks)

e) Implement the **Elbow method diagram** and find out the suitable number of clusters. Capture the screenshot of Elbow diagram and rename as **Q02_e.jpg**

(20 Marks)

f) Apply the **k-means++** algorithm for the dataset.

(25 Marks)

g) Plot the cluster results in a scatter plot and capture the screenshot. Rename the image as **Q02_g.jpg.** Also include the colab notebook in the answer folder.

(25 Marks)

[100 Marks]