UNIVERSITY OF THE WESTERN CAPE
DEPARTMENT OF COMPUTER SCIENCE

CSC311 2022 - UNSUPERVISED LEARNING PROJECT

1. Initialisation

a. [Download and] import the required dataset file into Jupyter notebook - 2 marks
   See download link below
b. Display the first 10 rows on the dataset - 1 mark
c. Set Y as the "class" column and X as all other columns - 5 marks
d. In Y, replace all 1 with 0, 2 with 1 and 3 with 2 - please follow this order when replacing to avoid errors later - 2 marks
e. Use a scatter plot to show "Area" (x-axis) vs "Perimeter" (y-axis) - 2 marks
f. Scale the dataset using StandardScaler - 3 marks

Q1 TOTAL MARK = 15%

2. KMEANS

a. Use K-Means Elbow method to determine the number of clusters in dataset -
   i. Perform at least 10 iterations
   ii. Show your Elbow graph
   5 marks
b. Perform KMeans clustering with 3 clusters.
   i. Plot a scatter plot for "Area" and "Perimeter" clusters and their corresponding Centroids - **5 marks**
   ii. Plot a scatter plot for "Length" and "Width" clusters and their corresponding Centroids - **5 marks**
   iii. Repeat KMeans but let k = the number of clusters which you selected from your Elbow graph in 2a above
   iv. Plot a scatter plot for "Area" and "Perimeter" clusters and their corresponding Centroids - **3 marks**
   v. Plot a scatter plot for "Length" and "Width" clusters and their corresponding Centroids - **3 marks**
c. Calculate the classification accuracy of KMeans by comparing the output of KMeans in 2Bi above with the original labels of the dataset Y. **4 marks**
   Hint, if you have not done so earlier, you might need to replace 1 with 0, 2 with 1 and 3 with 2 in Y

Q2 TOTAL MARK = 25%

3. AHC

a. Use AHC dendrogram to determine the number of clusters in the dataset (set method = 'Ward') - **7.5 marks**
b. Perform AHC clustering with 3 clusters. - **7.5 marks**
c. Plot a scatter plot for "Length" and "Width" clusters - **5 marks**
d. Calculate the classification accuracy by comparing the output of AHC in B above with the original labels of the dataset Y. - **5 marks**

Hint, if you have not done so earlier, you might need to replace 1 with 0, 2 with 1 and 3 with 2 in Y

Q3 TOTAL MARK = 25%

4. KNN

a. Perform KNN clustering with Neighbour size of 3. - **7.5 marks**
b. Plot a scatter plot for "Length" and "Width" clusters - **5 marks**
c. Calculate the classification accuracy of KNN by comparing the output of KNN in A above with the original labels of the dataset Y.
   Hint, if you have not done so earlier, you might need to replace 1 with 0, 2 with 1 and 3 with 2 in Y - **5 marks**
d. Using the output of KNN in A above and the original Y, draw a confusion matrix for KNN. - **7.5 marks**

Q4 TOTAL MARK = 25%

5. Exit

a. Draw a bar chart comparing the accuracies of KMeans, AHC and KNN when cluster size is set to 3.

Q5 TOTAL MARK = 10%

GRAND TOTAL = 15 + 25 + 25 + 25 + 10 = 100%

Other information

1. The dataset can be downloaded from:

https://drive.google.com/file/d/1PephnI7T5pB-rZFjX4pJFKcvpDmr56rs/view?usp=sharing

2. Kindly submit ONLY a Jupyter Notebook with your answer for each question embedded.
3. The due date is 3rd June 2022

Dr. Olasupo Ajayi

ooajayi at uwc dot ac dot za

CSC311 – ML (Unsupervised learning) Project, 2022

Department of Computer Science,

University of the Western Cape, Cape Town.