**ICP6**

**BIG DATA ANALYTICS AND APPLICATIONS(CS5542)**

*Sravani Niharika Garapati (16328843)*

*Nikhila Chirumamilla (16323180)*

**QUESTION:**

Use a dataset and use the model provided in to perform Clustering. Must try 5 clusters based on elbow curve and for each cluster visualize the clustering results and report your findings in detail.

**What I have learned from this ICP:**

In this ICP, professor asked us to perform Clustering on a data set. In this ICP, we have created 5 clusters using K-Means clustering algorithm. The things which I have learned from this ICP are listed below

* I have learned many functions that were very useful in implementing the clustering using K-Means algorithm.
* I have known in detail about the taken data set and to perform analysis on that data set in a thorough manner.
* ***Iris Data Set:***

The dataset contains a set of 150 records under five attributes

1)sepal length

2)Sepal width

3)Petal length

4)Petal width

5)Species.

It contains three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification, and clustering examples and to test algorithms.

**ICP Description about the Task:**

In this ICP, we performed Clustering on Iris Dataset using K – Means Algorithm. This is done in the following steps.

* As a first step, I have imported all the necessary libraries that will be used in this programming.
* Then, I have uploaded my Iris dataset and then read the contents of the file.
* Extracted the numerical data columns such as Id, SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm into a new data frame by removing Species columns as it is string data type
* Then Performed scaling on the data so that the data will be in the closest range of the values
* Plotted a graph, known as an **Elbow Curve**, with the x-axis representing the number of clusters and the y-axis representing an assessment metric which is used to identify the optimal number of clusters that would be used in our problem.
* Created Models

***K- Means Clustering model for 2 clusters:***

* Defined the K-means function with initialization as k-means++ specifying the number of clusters as 2.
* Then fit the k means algorithm on the Scaled Data
* Then found the inertial of model having 2 clusters

***Inertia:*** 273.7275

***Cluster 1:*** Yellow

***Cluster 2:*** Blue

***K- Means Clustering model for 3 clusters:***

* Defined the K-means function with initialization as k-means++ specifying the number of clusters as 3.
* Then fit the k means algorithm on the Scaled Data
* Then found the inertial of model having 3 clusters

***Inertia:*** 177.0560

***Cluster 1:*** Yellow

***Cluster 2:*** Blue

***Cluster 3:*** Red

***K- Means Clustering model for 4 clusters:***

* Defined the K-means function with initialization as k-means++ specifying the number of clusters as 4.
* Then fit the k means algorithm on the Scaled Data
* Then found the inertial of model having 4 clusters

***Inertia:*** 150.5174

***Cluster 1:*** Yellow

***Cluster 2:*** Blue

***Cluster 3:*** Red

***Cluster 4:*** Black

***K- Means Clustering model for 5 clusters:***

* Defined the K-means function with initialization as k-means++ specifying the number of clusters as 5.
* Then fit the k means algorithm on the Scaled Data
* Then found the inertial of model having 5 clusters

***Inertia:*** 125.0576

***Cluster 1:*** Yellow

***Cluster 2:*** Blue

***Cluster 3:*** Red

***Cluster 4:*** Black

***Cluster 5:*** Green

**Challenges Faced in this ICP:**

The challenges which we faced while doing this ICP are listed below:

* It has become very difficult for us to find a dataset which would be good for clustering process.
* Also, it has become very difficult for us to find the labels on which the clustering should be performed.

**Screenshots of Execution of Code:**

*Importing required libraries*

**Graphical user interface, text, application

Description automatically generated**

*Importing the Iris Dataset*

**Graphical user interface, text, application

Description automatically generated**

*Reading the contents from the Iris dataset and description about the Iris Dataset*

**Graphical user interface, text, application, email

Description automatically generated**

*Depicting first 20 records of the dataset*

**Table

Description automatically generated with medium confidence**

*Statistics about the dataset*

**Table

Description automatically generated**

*Information about the dataset such as about column dataset*

*Text

Description automatically generated*

*Showing columns present in the taken dataset*

*Graphical user interface, text, application

Description automatically generated*

*Extracting numerical columns from the dataset by deleting columns with string data*

*Table

Description automatically generated*

*Showing the characteristics of the extracted data*

*Text

Description automatically generated*

*Scaling of the data*

*Table

Description automatically generated with medium confidence*

*Scaled Data*

**Graphical user interface, text, application

Description automatically generated**

**Text

Description automatically generated**

*Elbow Curve*

***Chart, line chart

Description automatically generated***

*K- Means initialization using 2 clusters*

***Graphical user interface, text, application

Description automatically generated***

*Visualizing the clusters of size 2*

*Text

Description automatically generated*

*Graphical depiction of clusters of size 2*

*Chart, scatter chart

Description automatically generated*

*K- Means initialization using 3 clusters*

*Graphical user interface, text, application

Description automatically generated*

*Visualizing the clusters of size 3*

*Text

Description automatically generated*

*Graphical depiction of clusters of size 3*

*Chart, scatter chart

Description automatically generated*

*K- Means initialization using 4 clusters*

*Graphical user interface, text, application

Description automatically generated*

*Visualizing the clusters of size 4*

*Text

Description automatically generated with medium confidence*

*Graphical depiction of clusters of size 4*

*Chart, scatter chart

Description automatically generated*

*K- Means initialization using 5 clusters*

*Graphical user interface, text, application

Description automatically generated*

*Visualizing the clusters of size 5*

*Table

Description automatically generated with medium confidence*

*Graphical depiction of clusters of size 5*

*Chart, scatter chart

Description automatically generated*

**Video Link**

https://youtu.be/iSycRxyb4rI