**Assignment 3**

**Clustering**

**Submitted By: Neel Vijay Patil (11436004)**

**Submitted To: Dr. Junhua Ding**

**INFO 5505 Applied Machine Learning for Data Science**

**UNT**

**Dataset**

I have collected the ALS Training and Testing dataset from the Kaggle.

The objective of this assignment is to train a k-Means clustering on the given data.

The main goal of assignment is to design ML model using K-Means Clustering and also to evaluate the performance of models.

**Step 1:** Importing Required Libraries

Graphical user interface, text, application

Description automatically generated

**Step 2:** Importing ALS\_TrainingData and creating the dataframes

Graphical user interface, text, application

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Importing ALS\_Testingdata and creating dataframe

Graphical user interface, text, application

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

**Step 3:** Checking the statistics of the train and test data using the describe()

Graphical user interface, table

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

**Step 4:** Checking the Null values , data values, Range, Memory usage using the info() and isnull()

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

**Step 5:** Checking skewness for each column using hist()

Graphical user interface, text, application

Description automatically generated

Calendar

Description automatically generated with medium confidence

Here I have used density plot on training and testing data to visualize attributes vs there density.

Text, letter

Description automatically generated

Calendar

Description automatically generated with medium confidence

**Step 6:** Carrying out log transformation on the attributes in columns

Text

Description automatically generated

A picture containing electronics

Description automatically generated

**Step 7: To find the correlation between the different attributes of the train dataset I have utilized the Heatmap**

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

**Step 8:** Here we can see that the training dataset have large set of features. So, I have carried out Principal Component Analysis (PCA) which reduces the dimensions of the current dataset without losing data. Here, I have imported the PCA and then send the n-components to 0.95 and carried out fit transformation on the training dataset.

Graphical user interface, text

Description automatically generated

**Step 9:** Here, I have **used** the standard scaler function to convert the feature values into there respective Z-score values.

Graphical user interface, text, application

Description automatically generated

**Step 10:** Here, I have used the Elbow method and inertia which executes the clustering (k-means ) on training dataset over k values in some range and calculates AVG score for every k value.

Graphical user interface, text, application

Description automatically generated

Chart, line chart

Description automatically generated

The line graph indicates the plot of WCSS to number of clusters and from graph we can conclude that k-values from 2 to 5 were found to be optimal values.

**Step 11:** I have dropped the labels like ID, SubjectID from the data\_test and applied the standardscaler transformation on the test data. Designed the K-means clustering model for the best valus from k =2 to 5.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

**Step 12:** Scatter plot distributions between 2 different features

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

**Step 13: Model Evaluation**

Here, I have used the Silhouette score for the evaluation of the k-means clustering model. It is used to assess nature of clusters. We have evaluated for range of k-values from 2 to 5.

The Silhouette Score for:

K (2) =0.59

K (3) =0.51

K (4) =0.44

K (5) =0.39

Graphical user interface, text, application, email

Description automatically generated

**Graphical user interface, text

Description automatically generated**

**Conclusion: From the evaluation we can infer that Silhouette values is highest when values of k is 2. As the value of the k increases from 2 to 5 the silhouette score decreases from its previous k values. So, the designed model where the value of k is 2 is found to be best models with respect to other k values**