**CS 5710**

**Machine Learning (Assignment # 5)**

E-mail: [gxm04390@ucmo.edu](mailto:gxm04390@ucmo.edu)

Name: Mogalapu Gideon Jaideep Sudarshan

Course: CS 5710

Assignment: Assignment 5

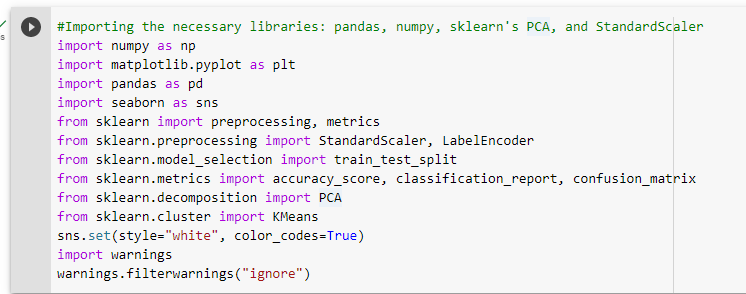
Git Hub: <https://github.com/JaideepMogalapu/Assignment5>

Video Link : <https://drive.google.com/drive/folders/1K8D8JAIg4bi7dxruvSXXnnMKY0nzu16o?usp=share_link>

**Question 1**

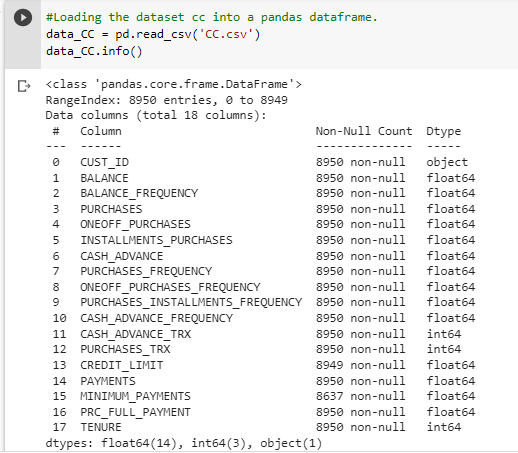
**Description :**

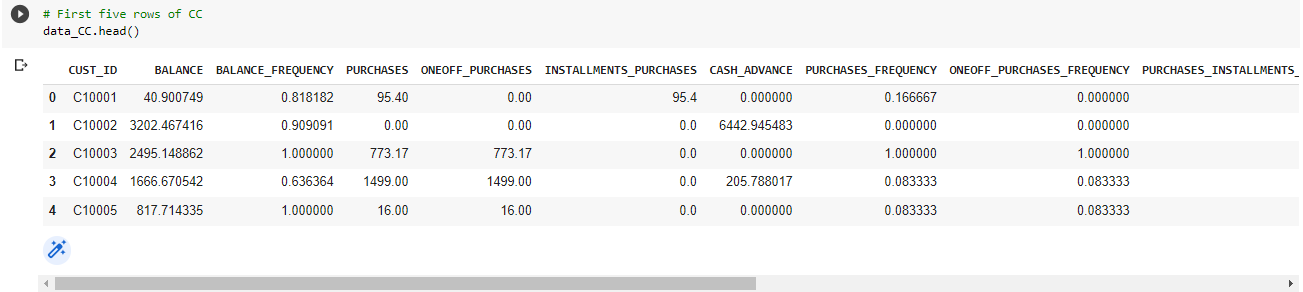
We have imported all the necessary libraries

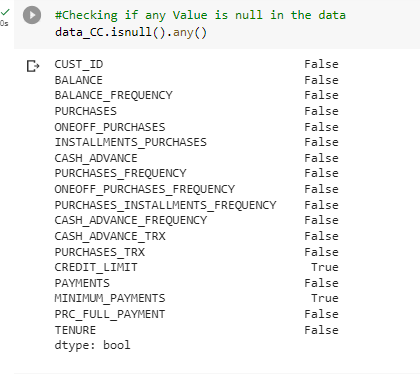


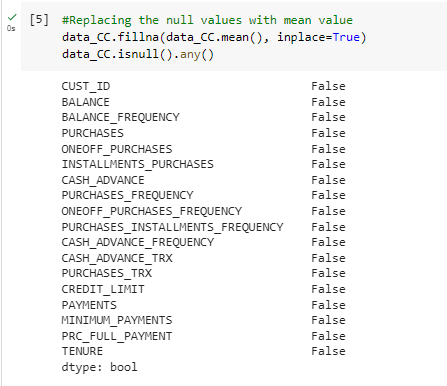
**1a) Apply PCA on CC dataset.**

**Description:** We have loaded the CC data set using panda and displayed the info of the data. We displayed the first 5 rows using head method. We checked for null values and replaced the null values with the mean value

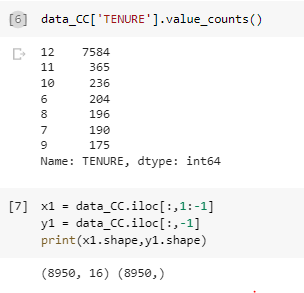




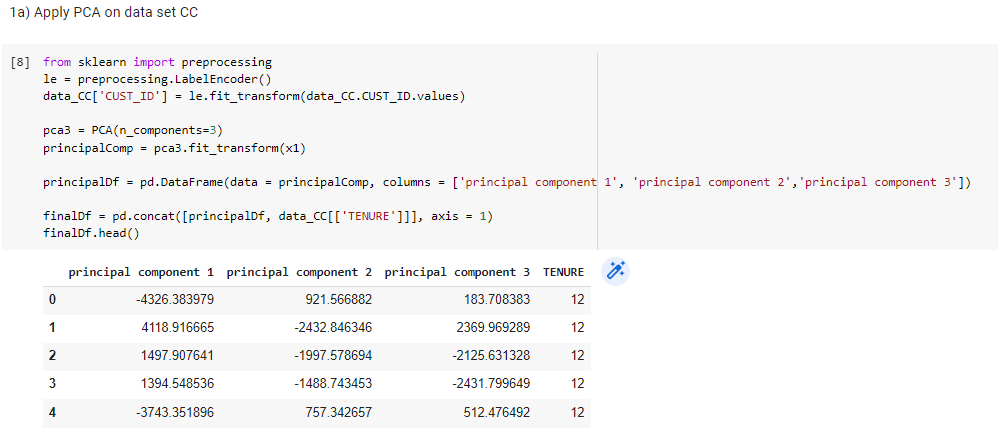


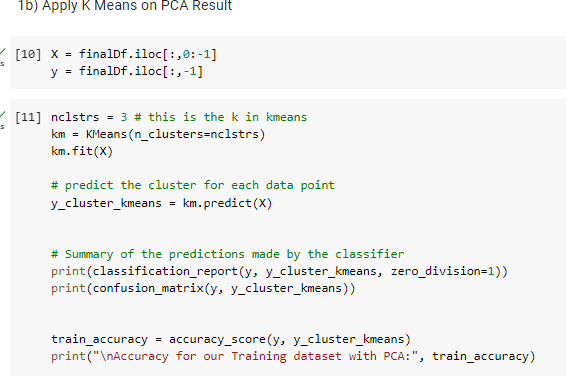


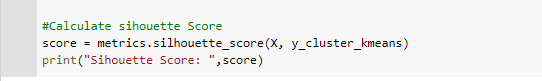
**Description:** We have counted the number of times each distinct values in the tenure column is present. And we have sliced the data set

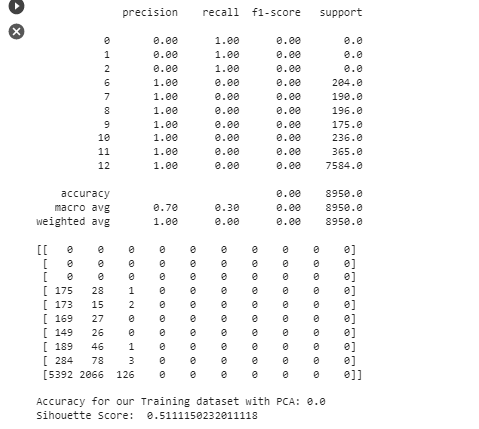


**Description:** Applied PCA on the scaled numerical variables



**Description:** After applying PCA on the CC dataset, we can use the k-means algorithm on the reduced-dimensionality space to cluster the data points. We printed the silhouett score

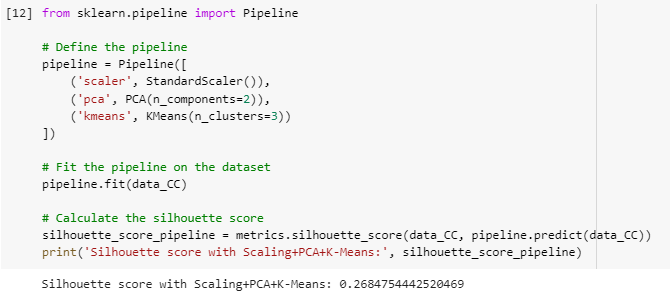




Sihouette Score- ranges from −1 to +1 , a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

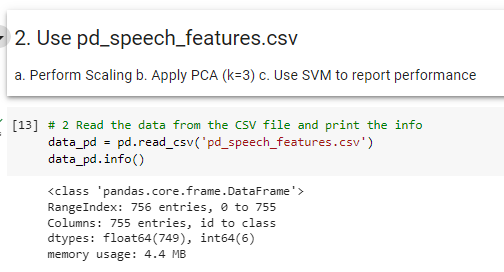
**1c) c. Perform Scaling+PCA+K-Means and report performance**

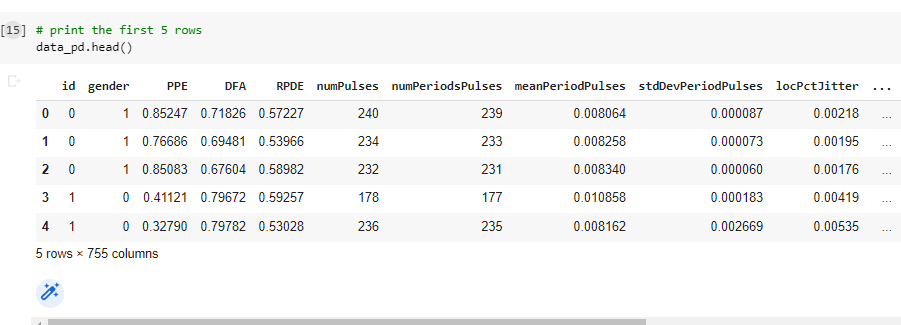
**Description:** Here we imported the pipeline. To further improve the performance of clustering, we performed scaling, PCA, and k-means in a pipeline

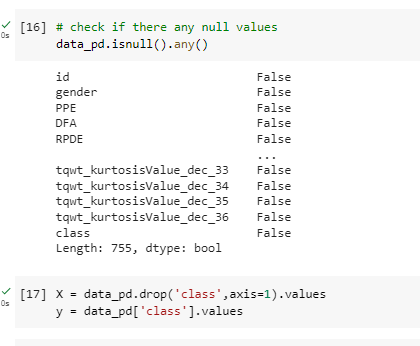


**Question 2:**

**Description:** Imported the data from the CSV file. Printed the info and first five rows of the data using info and head. We had checked for null values.

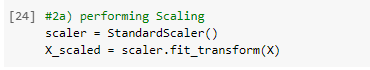






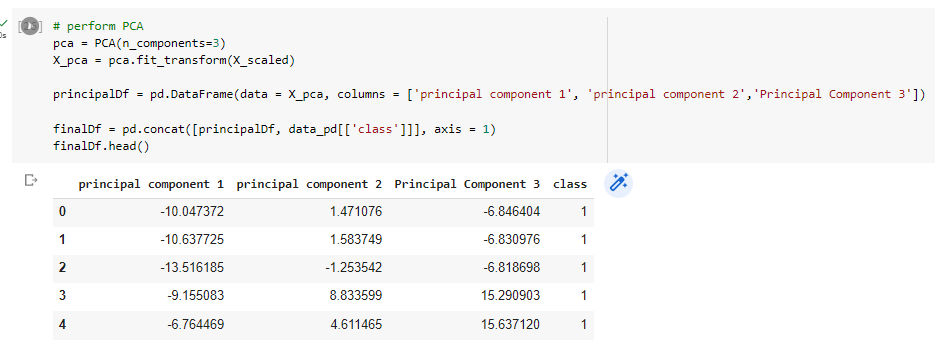
**2a) Perform Scaling**

**Description:** We dropped the class column and performed scaling on this data set

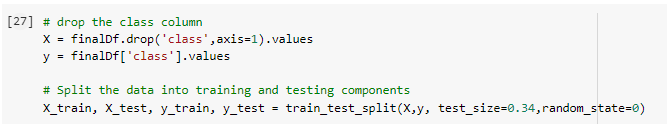


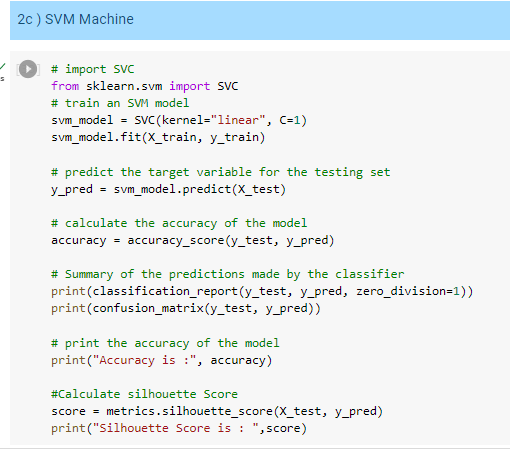
**2b) Apply PCA (k=3)**

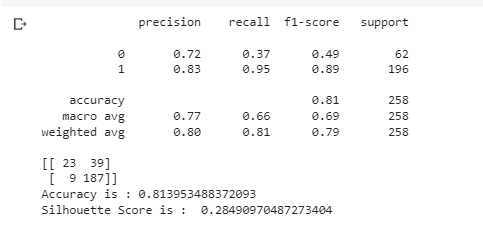
**Description:** We have applied PCA with 3 components for the above dataset and printed the final data set with 3 components and class column



**2c) Use SVM to report performanceDescription:** We dropped the class column from the above dataset and split the data into training and testing data sets. We then trained the data on SVM model and printed the accuracy, classification report, confusion matrix and silhouette score.

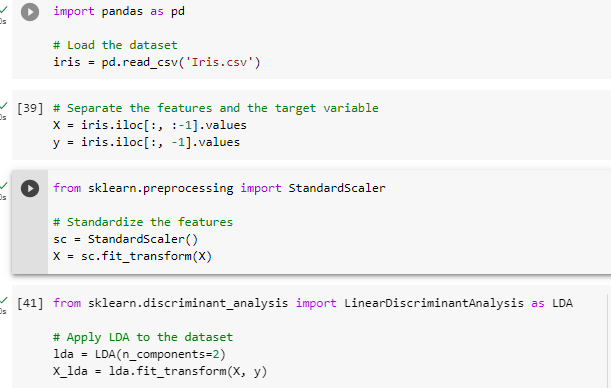


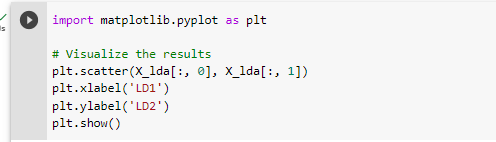


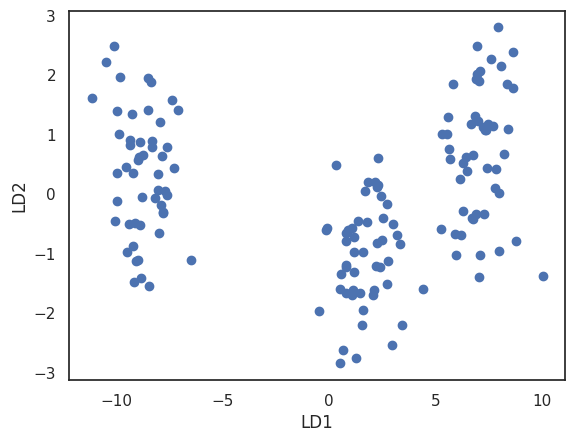


**Question 3**

**Description:** We have loaded the dataset and separated the features and target variables. We then standardized the features using standard scalar. We applied LDA to the data set and visualized the results using a scatter plot







**Question 4) Briefly identify the difference between PCA and LDA**

PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are both dimensionality reduction techniques used in machine learning, but they have different objectives and methods.

PCA is an unsupervised learning technique used to reduce the dimensionality of a dataset by identifying and extracting the most important features that capture the maximum amount of variance in the data. PCA aims to find a new set of uncorrelated variables called principal components, which are linear combinations of the original variables. The principal components are ranked by their explained variance, with the first principal component explaining the most variance in the data, and subsequent components explaining decreasing amounts.

LDA, on the other hand, is a supervised learning technique used to reduce the dimensionality of a dataset while preserving class-discriminatory information. LDA aims to find a new set of variables that maximize the separation between classes in the data. This is achieved by computing a set of discriminant functions that project the data onto a lower-dimensional space while maximizing the between-class scatter and minimizing the within-class scatter.

In summary, while both techniques aim to reduce the dimensionality of a dataset, PCA is unsupervised and focuses on maximizing variance, while LDA is supervised and focuses on maximizing class separability.