**MIDHUN KUMAR CHINTAPALLI 700739502**

CRN-23922

**Assignment 5**

**GitHub Link:** <https://github.com/midhun-ch/ML_Assignment5>

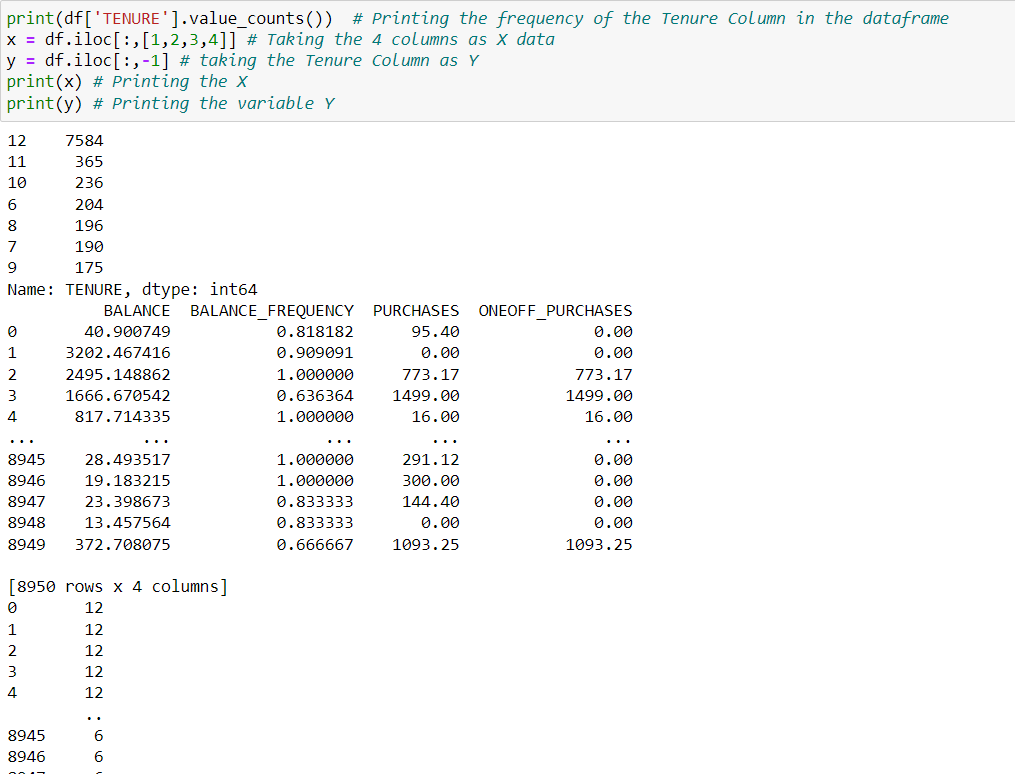
**Question 1:**

**Apply PCA on CC dataset.**

* We have imported necessary libraries and import the CC dataset csv file to code.

Text

Description automatically generated

* We have taken the ‘BALANCE’, ‘BALANCE\_FREQUENCY’, ‘PURCHASES’, ‘ONEOFF\_PURCHASES’ as X variables and tenure as Y variable.
* We applied principal component analysis with 2 components for the given data and stored the data in the dataframe.

Graphical user interface, application, Word

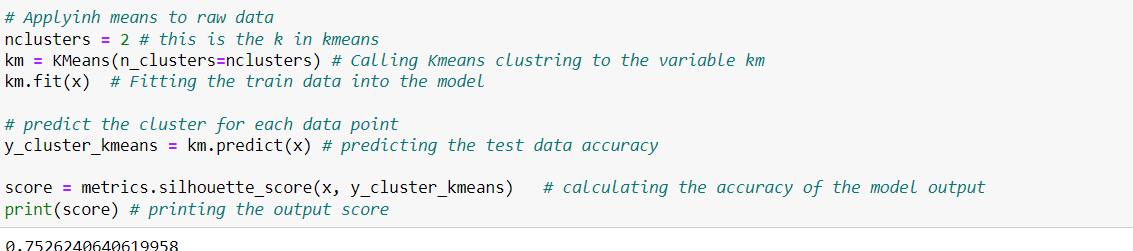
Description automatically generated

* We have applied Kmeans Clustering with 2 clusters to the above data and we have calculated the accuracy of the model.

Graphical user interface, text

Description automatically generated

* We are applying the Kmeans clustering to the raw data without PCA and calculating the Accuracy of the model.



* Now we are applying scaling to the raw data and finding the PCA for the scaled data and then storing the data into a dataframe.

Graphical user interface, text, application

Description automatically generated

* Now we are calculating the Kmeans Clustering model accuracy for the scaled pca applied data.

Text

Description automatically generated

* Now we are applying the Kmeans clustering to the scaled raw data for comparing the accuracy.

Text

Description automatically generated

Both Raw data and the Pca applied data gave almost the same accuracy which is almost 76 % whereas the scaled data’s accuracy is less when compared the above result which is around 70 %.

**Question 2**

**Use pd\_speech\_features.csv**

* Importing all the necessary libraries and and taking the file as input.

Text

Description automatically generated with low confidence

* Scaling the data after taking the entire dataset except class column as X and taking the class column as y and then Fitting the x data into the scaler method.
* And then Applying PCA with 3 components and storing the data into a dataframe.

A picture containing application

Description automatically generated

* Applying SVM model over the scaled PCA applied dataset to find the accuracy of the model by taking the necessary X and Y variables.

Text

Description automatically generated

* Applying the SVM model to the raw data to find the data accuracy.

Text

Description automatically generated

It seems that accuracy without pca is more than the accuracy when PCA is appled.

**Question 3:**

**Apply Linear Discriminant Analysis (LDA) on Iris.csv dataset to reduce dimensionality of data to k =2.**

* importing necessary libraires for executing the code.
* importing the iris csv file into df variable and printing the top 5rows of the data frame.

A picture containing text

Description automatically generated

* Taking the dataframe, applying scaling to it and then fitting the data into LDA model to get the dimension reduced data.

Text, letter

Description automatically generated

* Plotting the data which is discriminated.

Chart, scatter chart

Description automatically generated

**4. Briefly identify the difference between PCA and LDA**

Dimensionality reduction in machine learning refers to the process of collecting a collection of major variables to reduce the number of random variables being considered. Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two main algorithms in dimensionality reduction.

PCA is an unsupervised while LDA is a supervised dimensionality reduction technique.

PCA gets the results without depending on the output labels. PCA results a data set with maximum variance between the features by ignoring the duplicates of other features. Since the variance between the features is independent of the outcome, PCA does not consider the output labels.

LDA depends on the output labels. Based on the output labels information LDA reduces the feature set dimensions and finds a decision boundary. The data points are then projected to new dimensions so that the clusters are as distinct from one another as possible, and the individual components of a cluster are as near the cluster centroid as possible.