Jagadeeswar Chimata

700731649

Assignment-5

Github link : https://github.com/JagadeeswarChimata/Assignment5.git

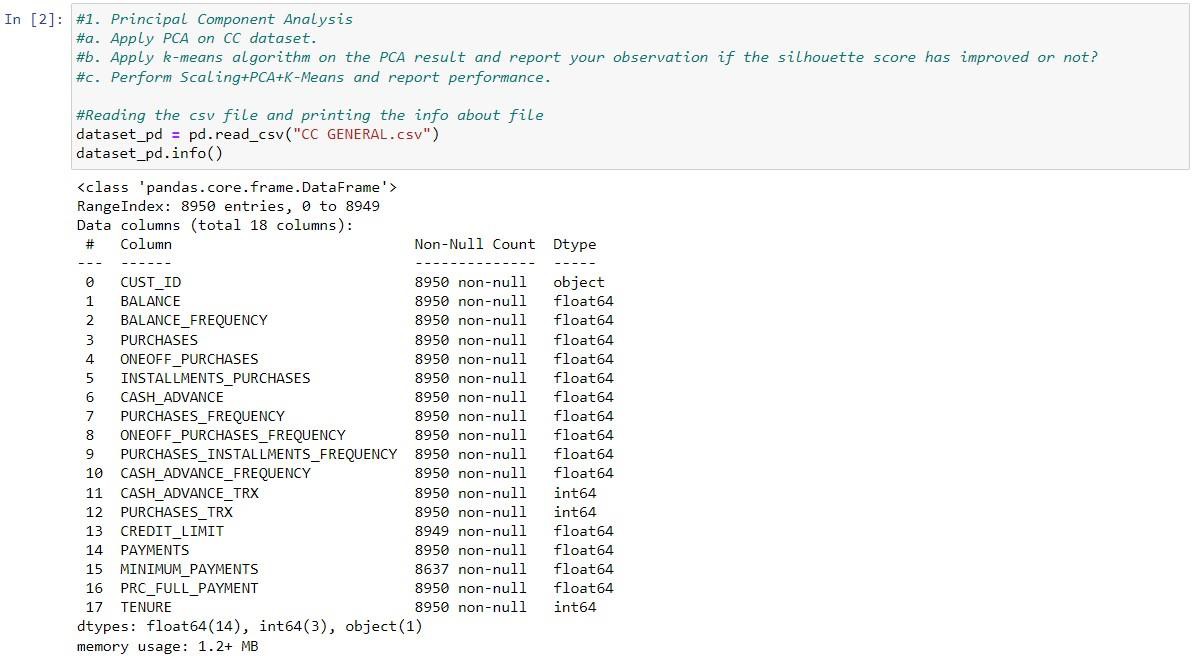
Video Link : https://drive.google.com/file/d/1c2utKo1DVNIJ3k4oewy2KM7BywLqUegA/view?usp=share\_link

Q1) PrincipalComponen**t** Analysis

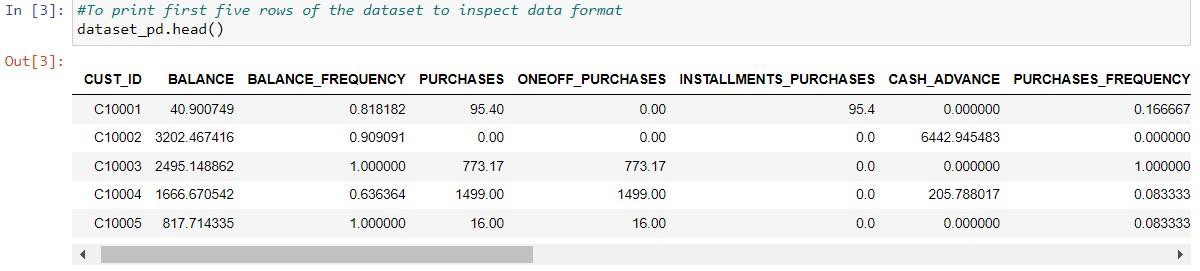
we import the libraries.



we read the csv file and print the info about file

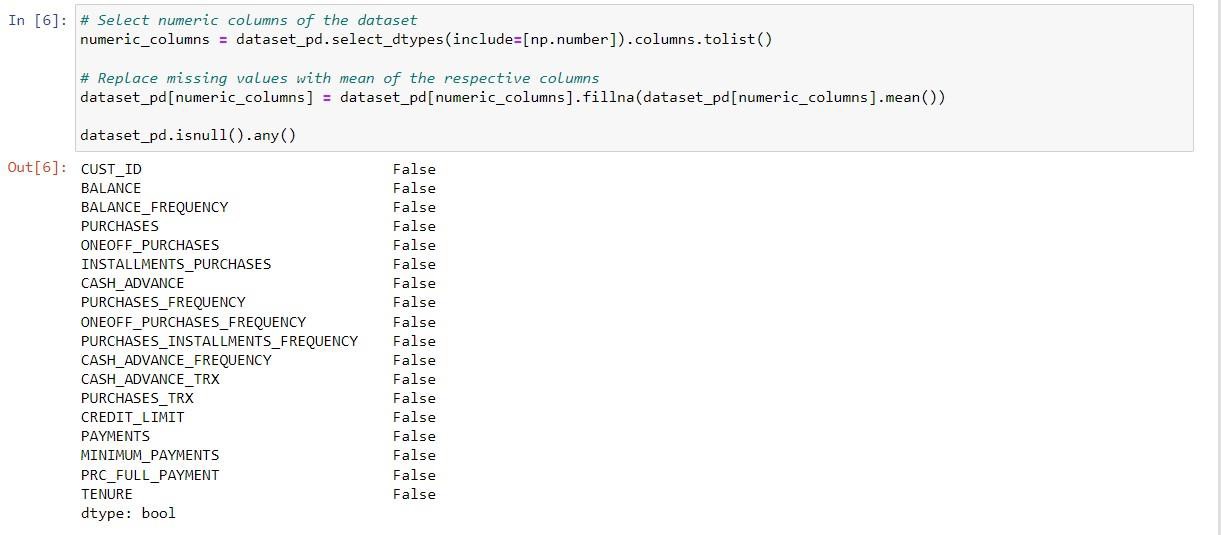


we print the first five rows of dataset to inspect data format

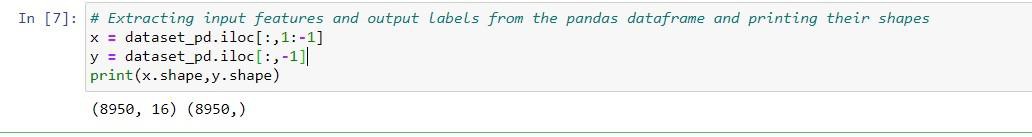


Then we check if there are any missing values in dataset



In the above output there are missing values ​​in the dataset, so we eliminate the missing values ​​by selecting numeric columns in the dataset and replacing the missing values ​​with the mean of the respective column

Then extract the input features and output labels from the pandas dataframe and output the shapes. The x input features and y output labels are:

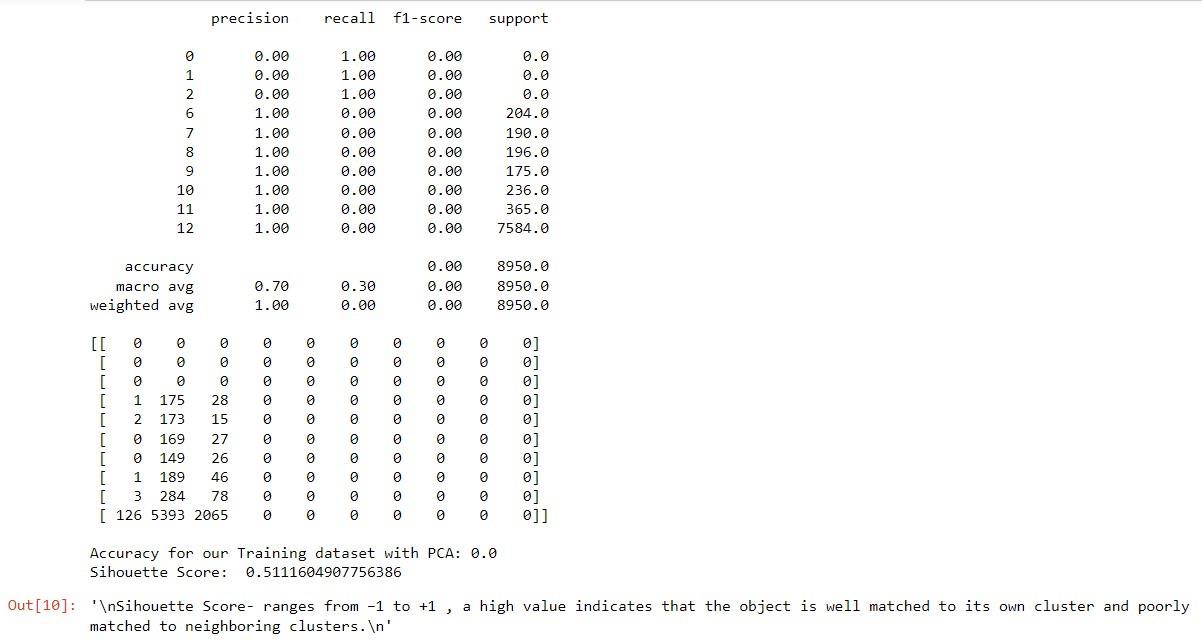


1a) Apply PCA on CC dataset



1b) Apply k-means algorithm on the PCA result and report your observation if the silhouette score has improved or not?

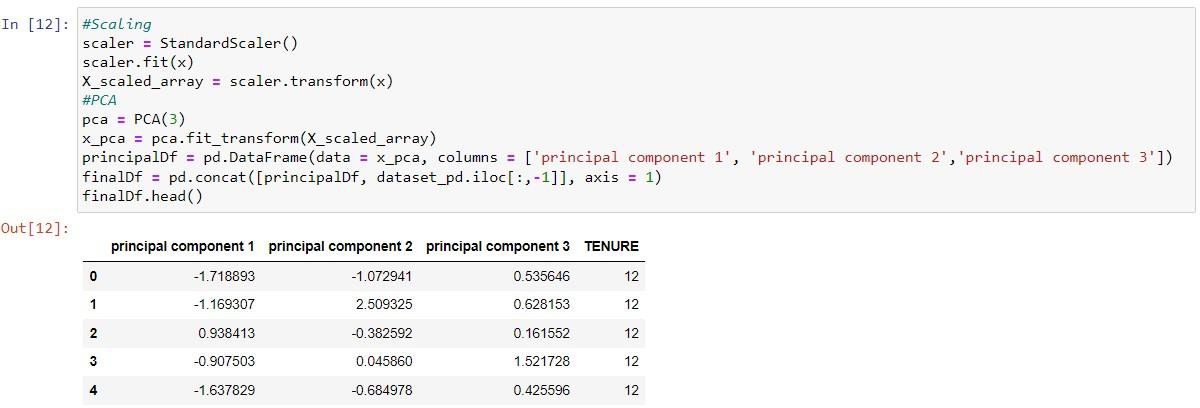


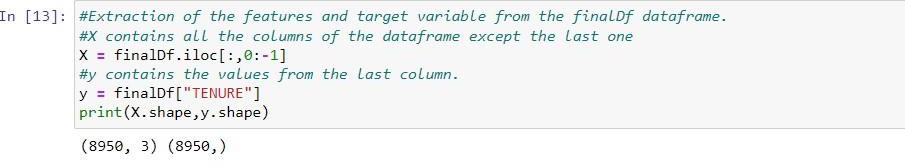


1c) Perform Scaling+PCA+K-Means and report performance.

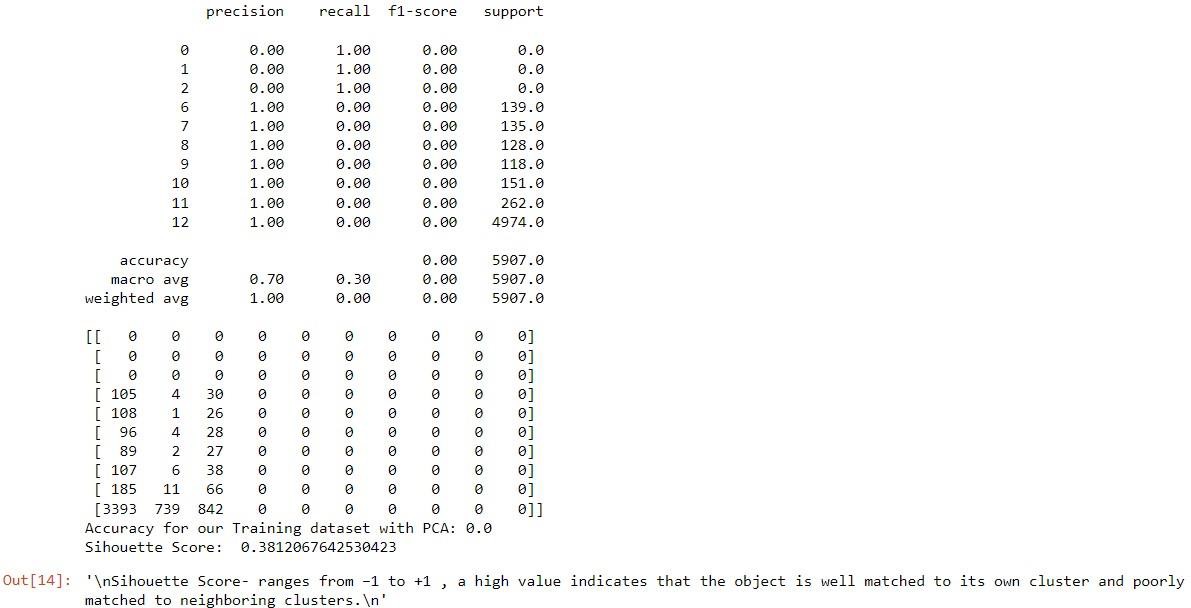
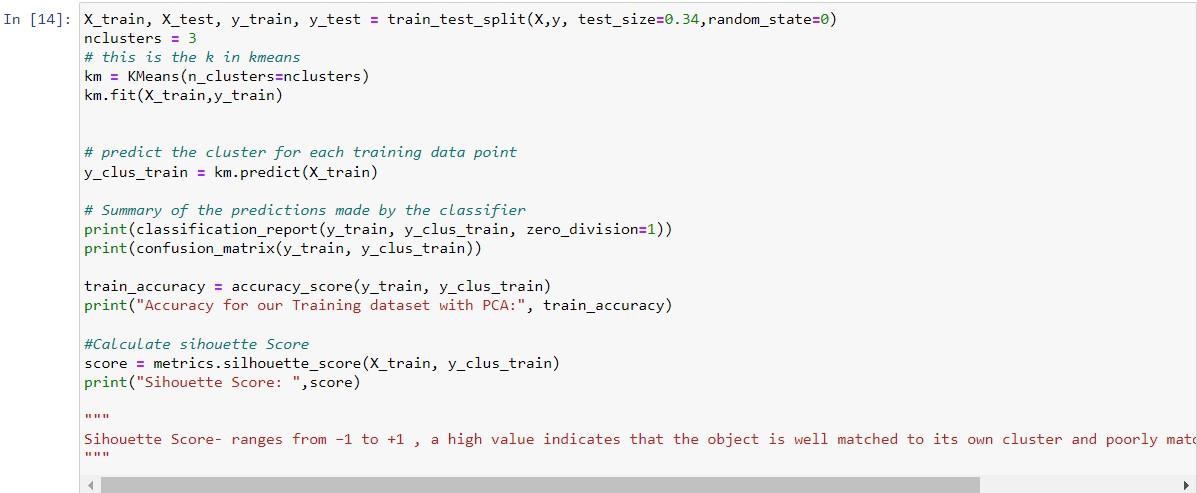


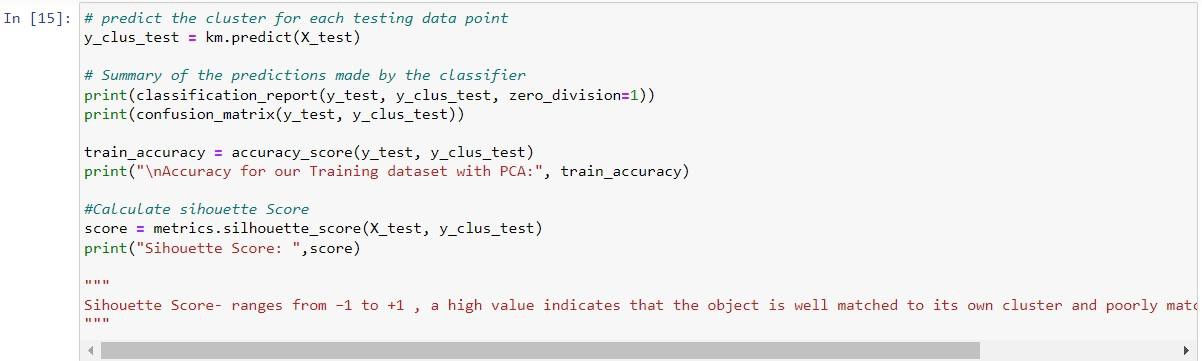
Here we perform Scaling and PCA

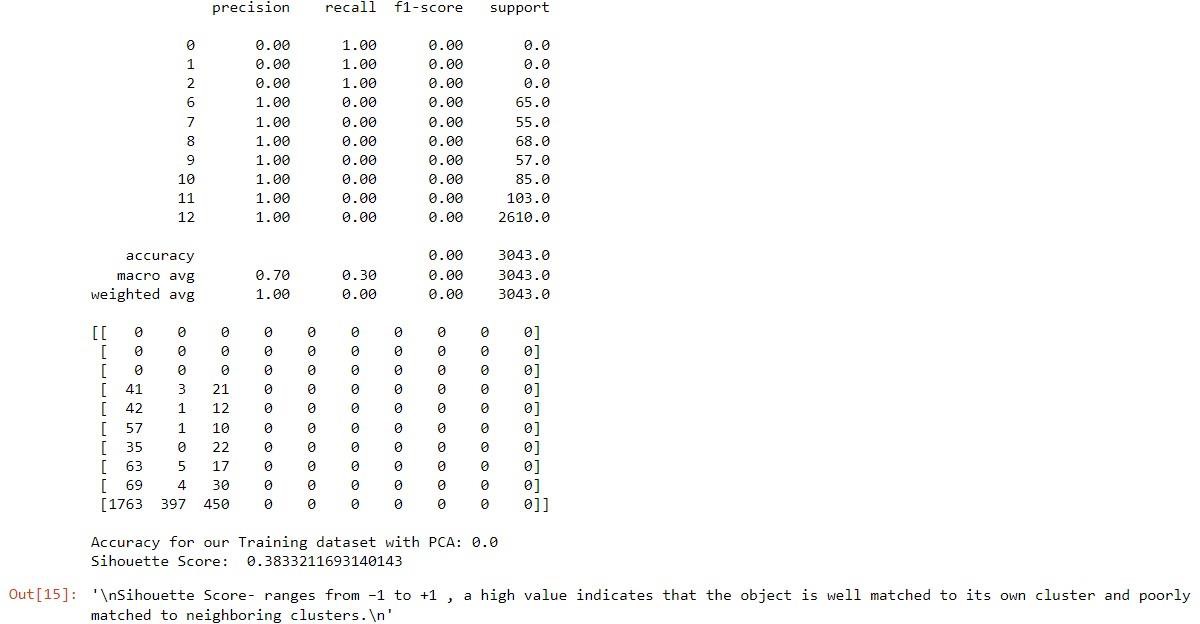


Here we extract the features and the target variable from the finalDf dataframe, where X contains all the columns of the dataframe except the last column and y contains the values ​​of the last column 

we perform k-means



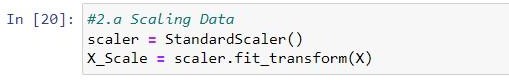




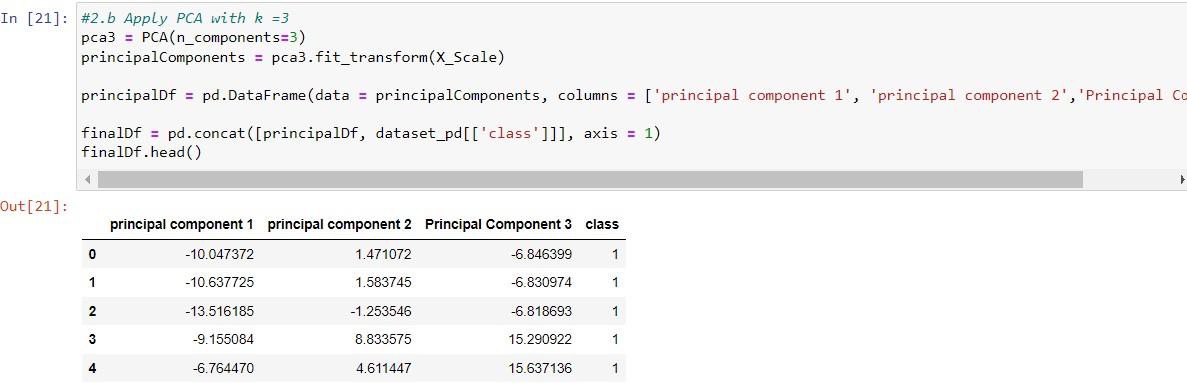
Here use pd\_speech\_features.csv



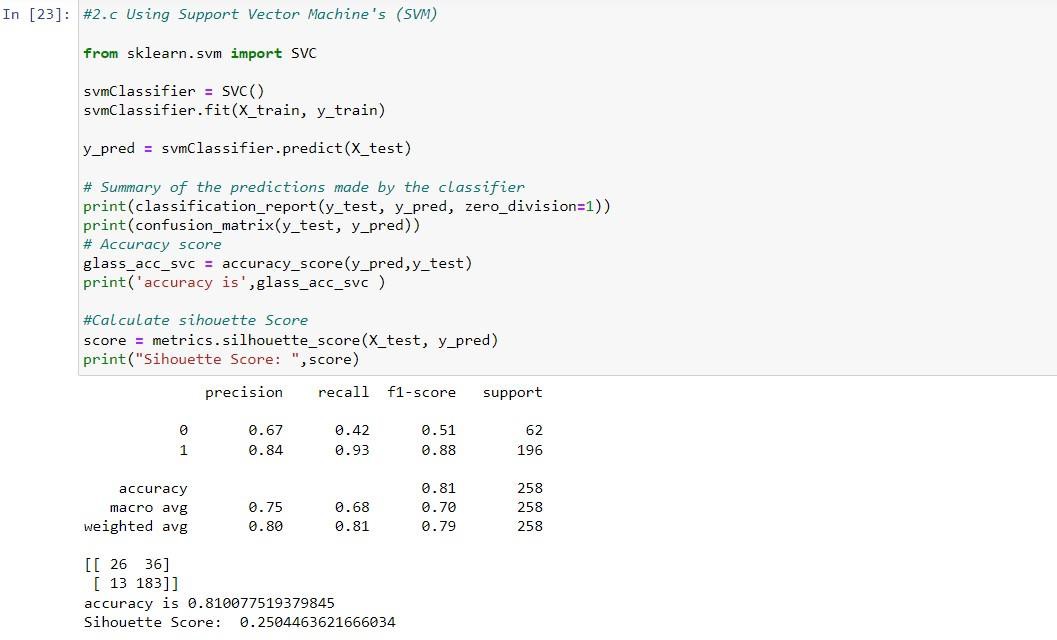
2a) Perform Scaling



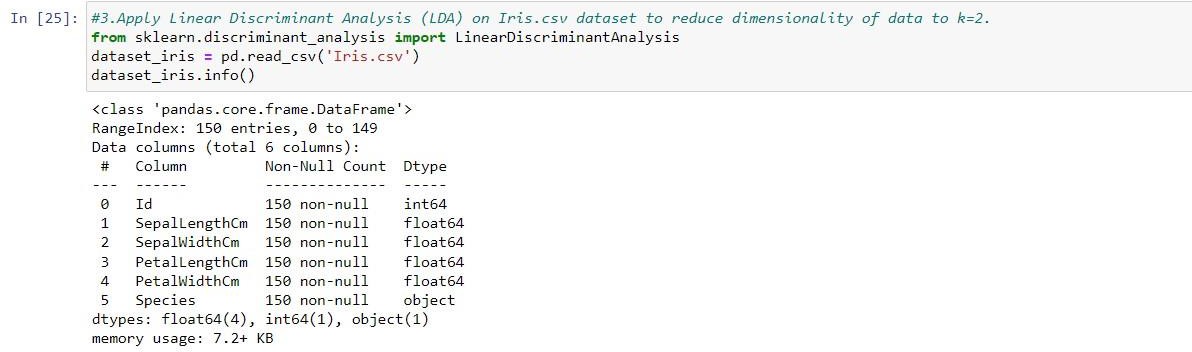
2b) Apply PCA (k=3)

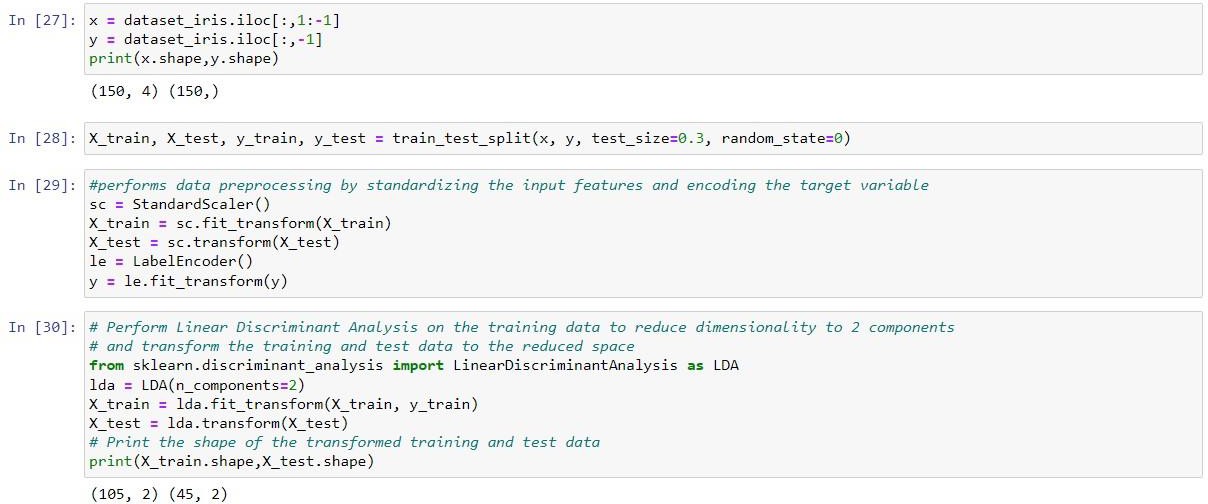


2c) Use SVM to report performance



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





4.Briefly identify the difference between PCA and LDA

Ans. By lowering the number of features or variables in a dataset while keeping the most crucial data, dimensionality reduction techniques like LDA and PCA are applied.

To reduce the original data's dimensionality, they both employ linear treatments.

Without respect to class labels, PCA is an unsupervised learning approach that locates the directions of largest variation in the data. It creates new features known as principle components that are orthogonal (i.e., uncorrelated) and capture the bulk of the data's variation. The most data variability is captured by the first principal component, followed by the second and third, and so on.

The LDA algorithm, on the other hand, is a supervised learning technique that seeks to increase the separability between various classes in the data. The variation within each category is minimized while the variance across distinct categories is maximized via certain linear discriminants. When determining the directions of maximal class separability, LDA takes into account the class labels included in the data.

Therefore, the primary distinction between PCA and LDA is that, while LDA focuses on identifying the directions that best separate various classes, PCA is concerned with collecting the largest variance in the data. While LDA explicitly uses the class labels to determine the best discriminative directions, PCA ignores any class distinctions.