Spring 2023 5710 Machine Learning: Assignment 5

Murali Yaswanth Motamarri

(700741647)

Video link: <https://drive.google.com/file/d/1u3ZsBs6DRCWJ_xZ1TY8gtqU8qjAM6RFu/view?usp=sharing>

Github link: <https://github.com/Yaswanthmm/machine-learning>

**1. Principal Component Analysis**

Text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, table

Description automatically generated

Graphical user interface, table

Description automatically generated

**Description**: In the first question, I imported the libraries and loaded the dataset “CC General.csv” using read\_csv function. Checked if there are any null values in the data using isnull() function and filled null values with the mean using fillna() function and loaded the dependent and target variables to x and y.

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

Graphical user interface, text

Description automatically generated

Description: Used PCA, fit\_transform() method and passed the principal components to a dataframe.

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

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

**Description**: From the obtained output we can see that the sihouette score is positive. And close to +1. The sihouette score ranges from -1 to +1. We can say that the object is closely matched to its own cluster.

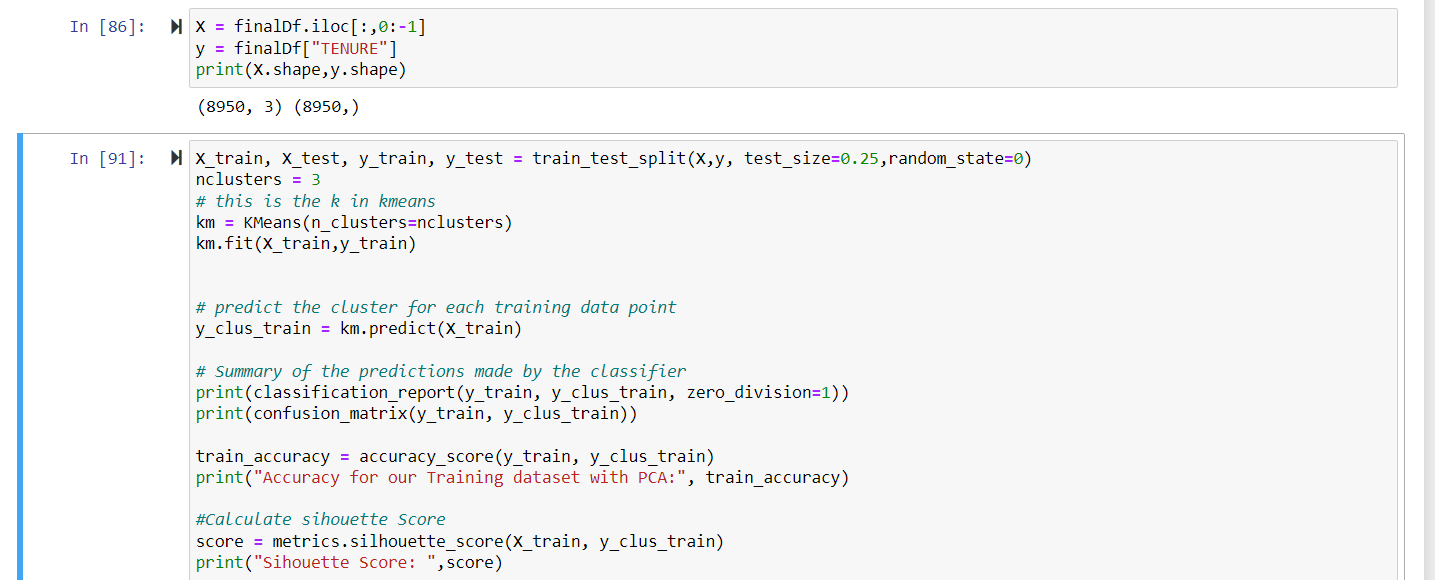
In this the n\_clusters value is 3. Loaded the k\_means model with n\_cluster value as 3 and fitted the model and predicted the target variable by passing the dependent variable as parameter. Printed the classification report and printed the confusion matrix, Calculated the silhouette score using metrics. Silhouette\_score()

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

**Graphical user interface, text

Description automatically generated**

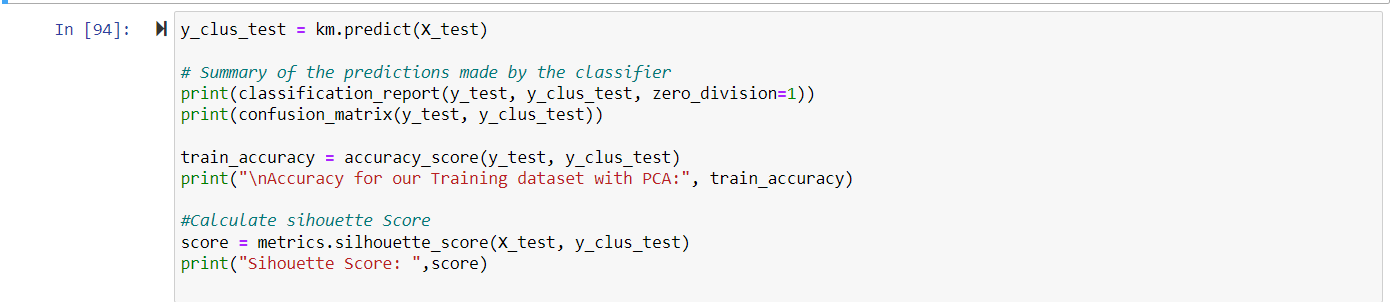
**Description**: Again, loaded the dependent variables and independent variable. Using the StandardScalar() and fitted the model with the independent variable and used the scaler.transform to get the x-scaled array. Loaded PCA with n\_clusters value as 3 and fit transformed the model by passing the s-scaled array as parameter and loaded the data into a Pandas DataFrame



Table

Description automatically generated

**Description**: Loaded the first three columns into a new dataframe and Tenure to a new variable. Splitted the data into training and testing dataset, using the kmeans, fitted the model with training data and predicted the output. The classification report and confusion matrix has been calculated and printed the accuracy for our training dataset with PCA and printed the Sihouette score.

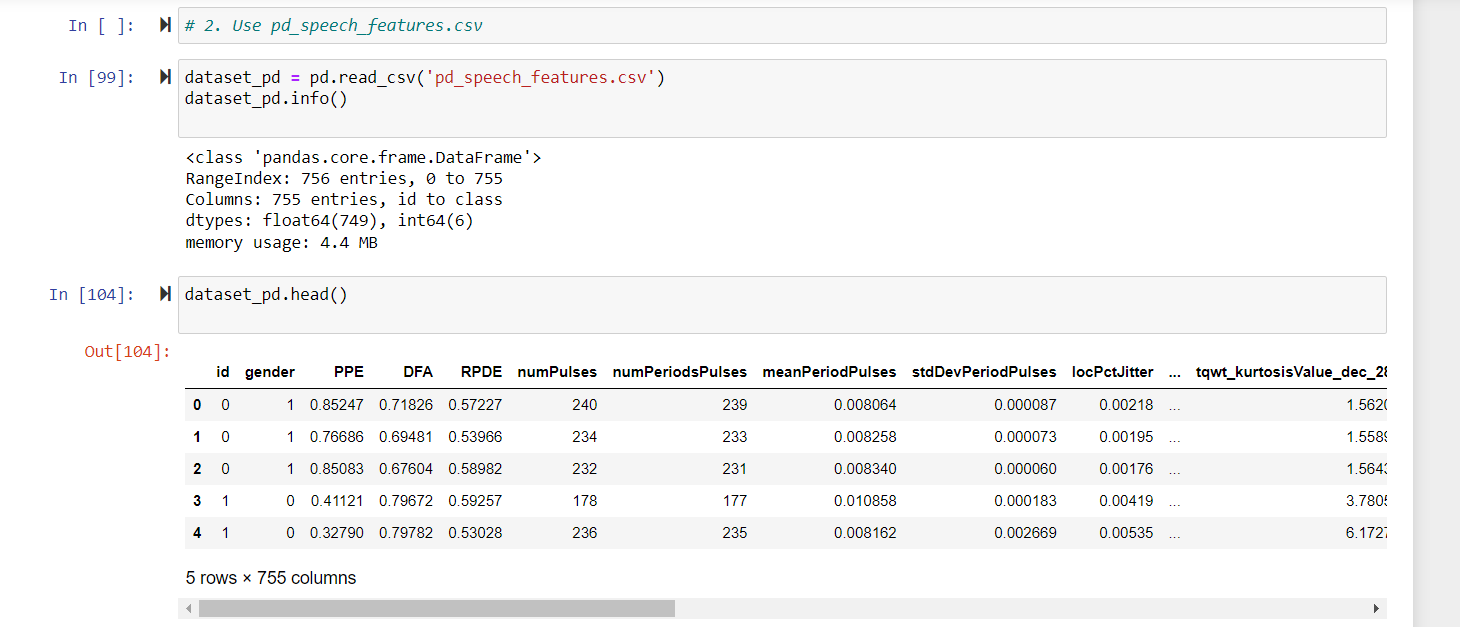


Table

Description automatically generated

**Description**: We have summarized the predictions made by the classifier and printed the classification report, confusion matrix along with the accuracy for our Training dataset with PCA and Sihouette score.

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



Graphical user interface, text, application

Description automatically generated

Description: We have loaded the dataset using the read\_csv() method. Found out if there are any null values and there aren’t any. We have loaded all the columns into x excluding the class row and loaded the class row into a new variable.

**2.a) Perform Scaling** 

Description: I have used the Standard scalar and fit transform method is used to perform Scaling

**2.b) Apply PCA with k =3**

Graphical user interface

Description automatically generatedDescription: We have loaded PCA with n\_components as 3 and fit transformed with parameter as x\_scale and loaded into principal DF and used this to get a pandas dataframe and finally displayed the first 5 rows using thehead()

**2.c) Use SVM to report performance.**

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface

Description automatically generated with low confidence

Description: After loading the the output column into y and the rest of the columns into x , we have splitted the data into training and testing data. We loaded the SVC and fitted the model with training and testing data and calculated the classification report and confusion matrix and printed the accuracy and the Sihouette score as well.

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

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Description: We have loaded the dataset using the Panda’s read\_csv() method. Found out if there are any null values in the dataset by using isnull(). any() and here we don’t have any null values. The last column which is the species column is loaded into a new variable which is the target variable and the rest into a new dataframe. We have split the dataset into testing and training dataset. We have loaded the StandardScaler and fit transformed the x-train data and transformed the x-test dataset. We have fit transformed the y using label encoder. We have imported the LinearDiscriminantAnalysis and loaded with n\_components as 2 and fit transformed the x-train and y-train dataset and transformed the x-test.

4) Briefly identify the difference between PCA and LDA

Description: Here both the PCA and LDA depend on linear transformation and tries to increase the variance in a lower dimensionality. PCA is an unsupervised learning algorithm , where as LDA is a supervised learning algorithm. PCA tries to identify the closeness to the class label whee are LDA tries to separate the commonness between the datapoints and PCA tries to minimize the separation with in its own class.