*Assignment-5*

Name: Vijay Kumar Kamsani

ID:700737910

Github Link:

https://github.com/Vijaykamsani/ML\_Assignment\_-70037910

Video:

https://drive.google.com/file/d/1QkEhMzGbarb1aIdUSnbgB3YX0GkvRXce/view?usp=share\_link

Programming:

Question1:



Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Comment:

We read the csv file and use the imputer function to fill the null values with the mean value. Following that, we apply PCA to the dataset using the k mean approach, and we obtain the resultant PCA of the silhouette function. Using the Standard scaler function, we may arrange the values between 0 and 1. The output of the scale PCA with the k-mean algorithm is displayed. When comparing the two results, the score is reduced after conducting the PCA, hence this data does not need to be processed using PCA.

Question:2

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated with medium confidence

Table

Description automatically generated

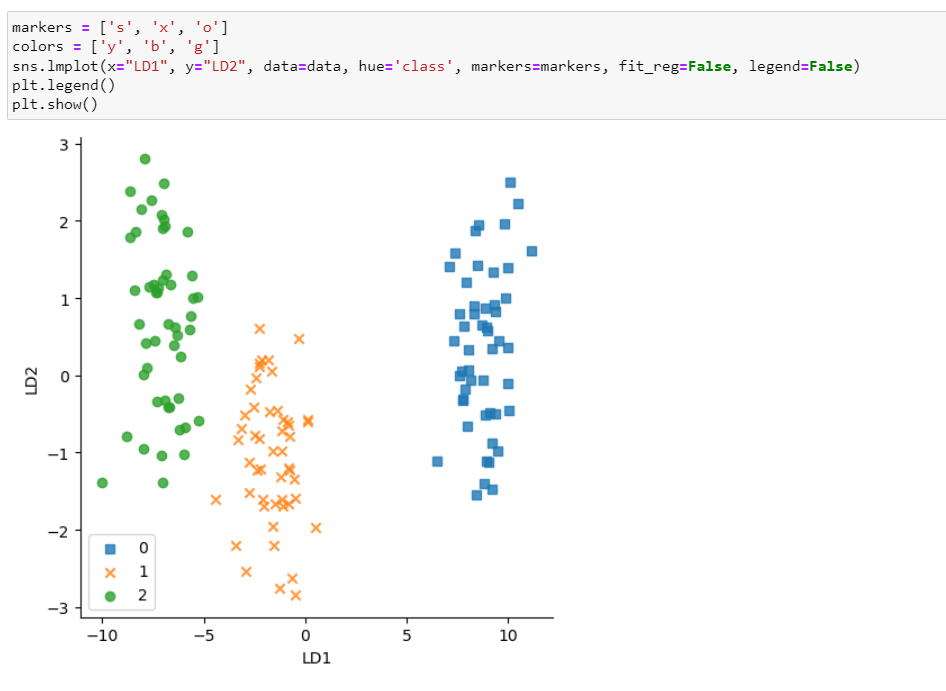
Comment:

First, the program selects all columns of the speech\_df DataFrame except for the first column are stored in the x variable. Next, the StandardScaler function module is used to standardize the features in x by removing the mean and scaling to unit variance and transformed. The resulting scaled features are stored in the speech\_x\_scaler variable.Then, the PCA function is used to perform PCA with 3 components, then the resulting transformed data is stored in the speech\_x\_pca variable.Next, the transformed data is stored in a new dataframe speech\_df2, and the original labels from speech\_df are concatenated with this new dataframe along the columns axis using the concat method. The resulting dataframe speech\_finaldf contains the transformed features as well as the original labels. Finally, the resulting dataframe speech\_finaldf is printed to the console to display the transformed data. The SVM module is then used to randomly split the data into training and testing sets, with 80% utilized for training and 20% for testing. The split data is stored in the variables X\_train, X\_test, y\_train, and y\_test. After getting the resultant values we are compared with predicated and tested value.

Question:3

Graphical user interface, text, application, email

Description automatically generated



Comment:

First, we modified the features using the transform method using the StandardScaler function, which we used to standardize the features in the independent variable columns of the iris\_df dataframe. The scaled features that result are saved in the X\_train\_std variable. Following that, we use the LabelEncoder function to encode the dependent variable column 'Species' of the iris\_df dataframe. The LabelEncoder object's fit\_transform method is used to assign a numerical value to each unique category in the 'Species' column, and the encoded labels are saved in the y variable. The Linear Discriminant Analysis function is then applied to the scaled features X\_train\_std and encoded labels y to perform LDA. The LinearDiscriminantAnalysis object's fit\_transform method is computed into the new coordinate system specified by these components. The modified data is saved in the variable X\_train. Next, the transformed data is stored in a new dataframe data, and the original encoded labels from y are concatenated with this new dataframe along the columns axis using the concat method. The resulting dataframe data contains the transformed features as well as the original encoded labels, with the column names set to "LD1", "LD2", and "class". The resultant graph is shown using the lmplot function.

Question4:

PCA performs better in case where number of samples per class is less. Whereas LDA works better with large dataset having multiple classes; class separability is an important factor while reducing dimensionality. PCA finds directions of maximum variance regardless of class labels while LDA finds directions of maximum class separability.