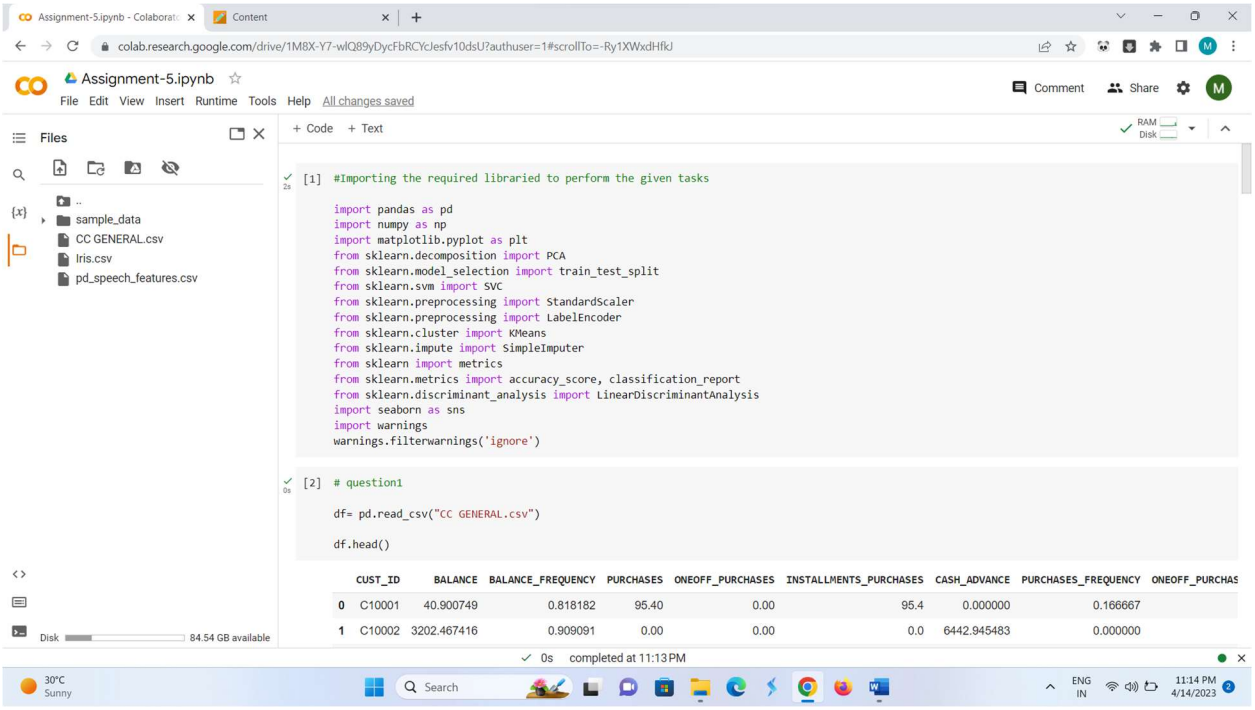
Name: Mythresh Maddina

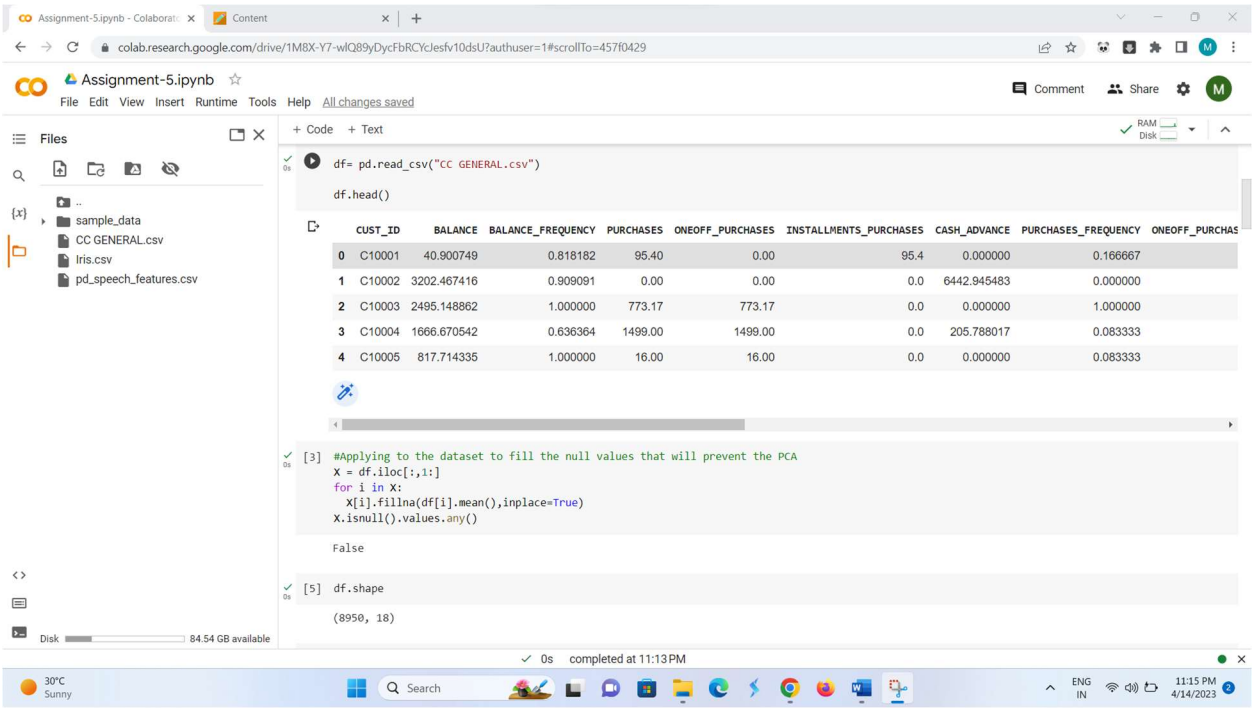
700: 700741162

Github Link: https://github.com/MythreshM/CS5710_Assignment5

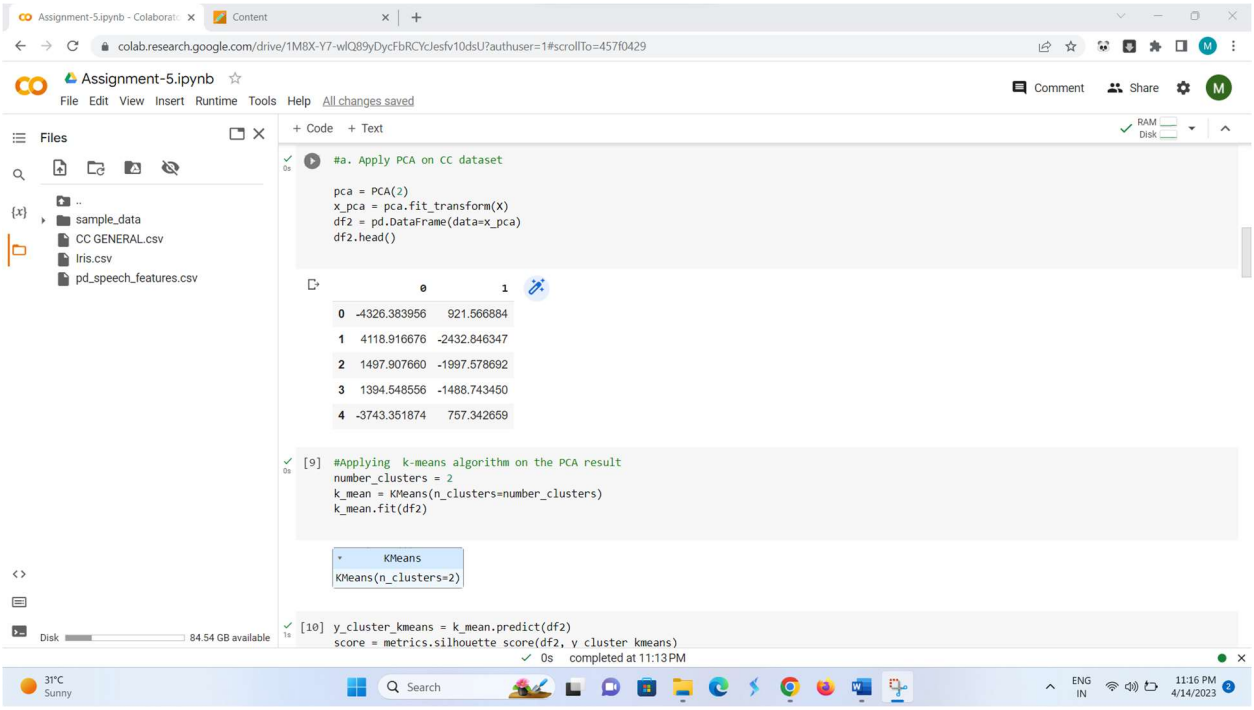
Video Link: https://drive.google.com/file/d/1THm1NRhQdvQE-fCxqRLLzxce66oqon5_/view?usp=sharing

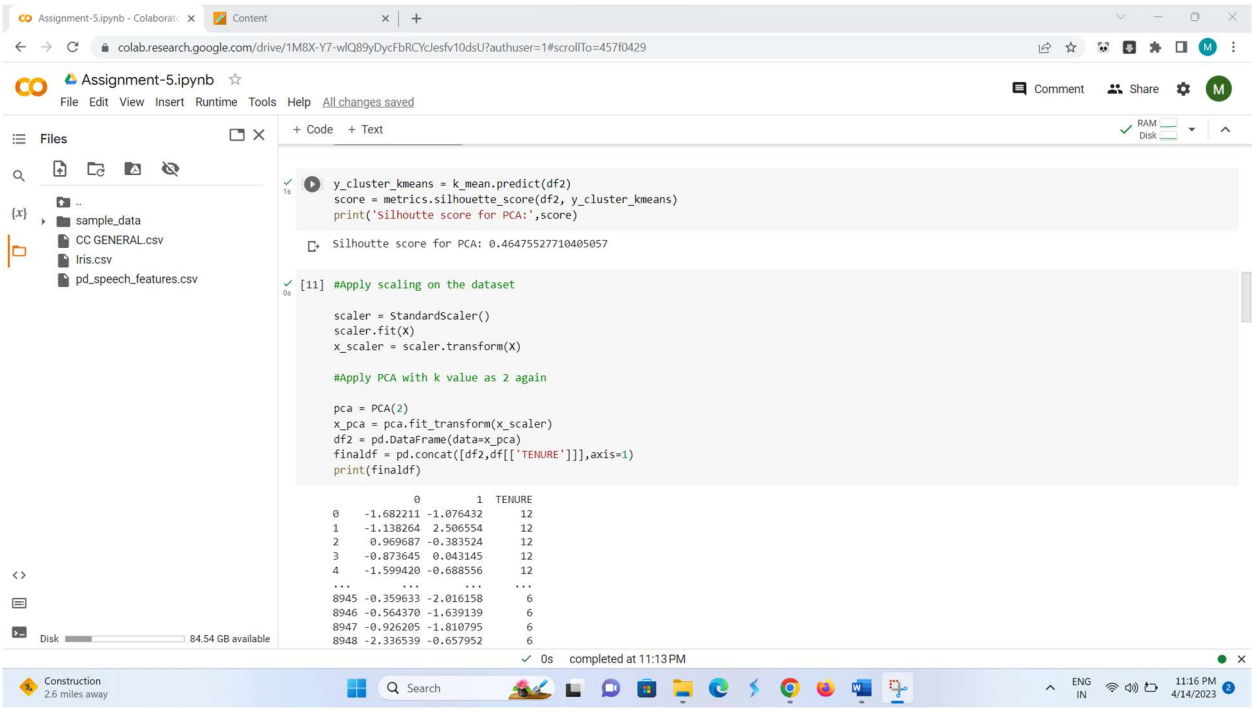
1.

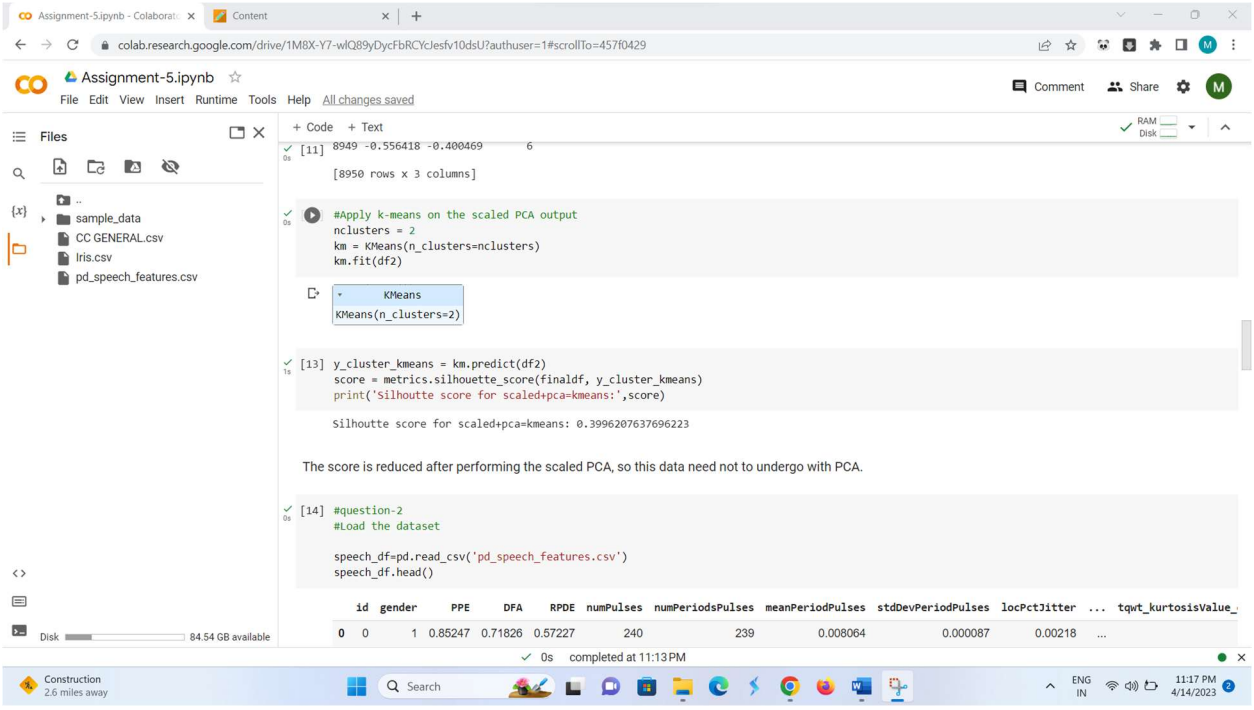




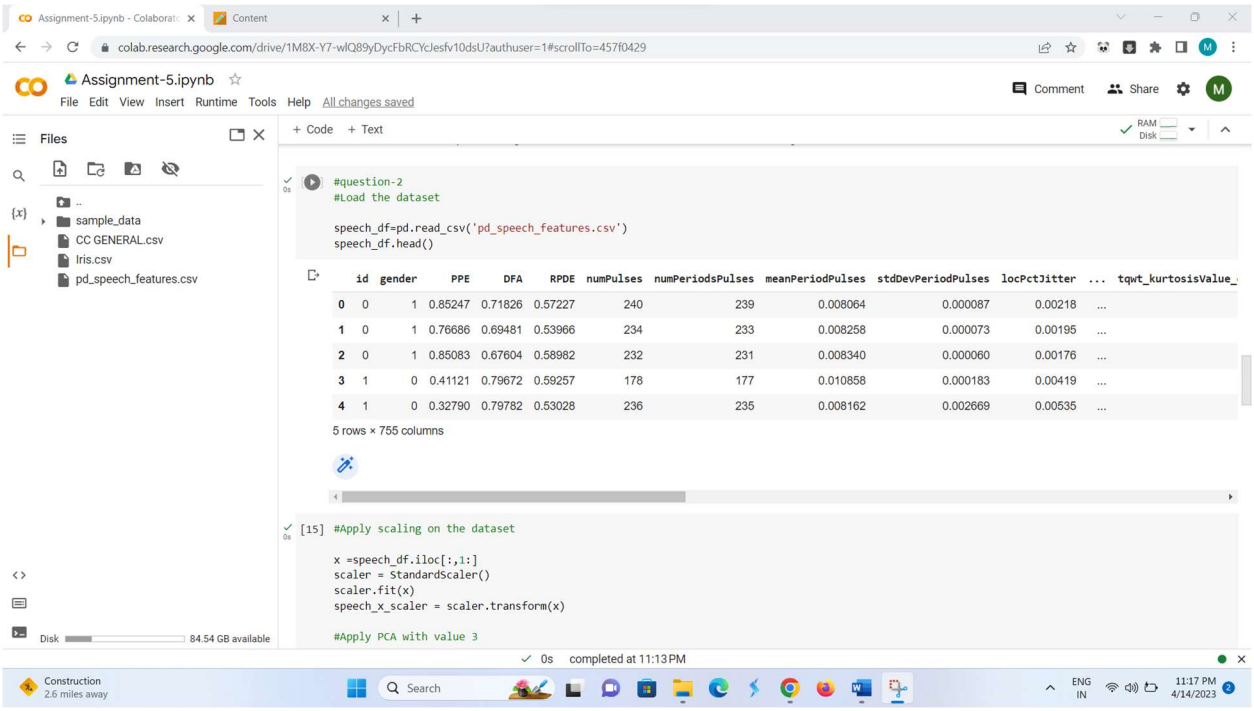
a) Applying PCA on CC

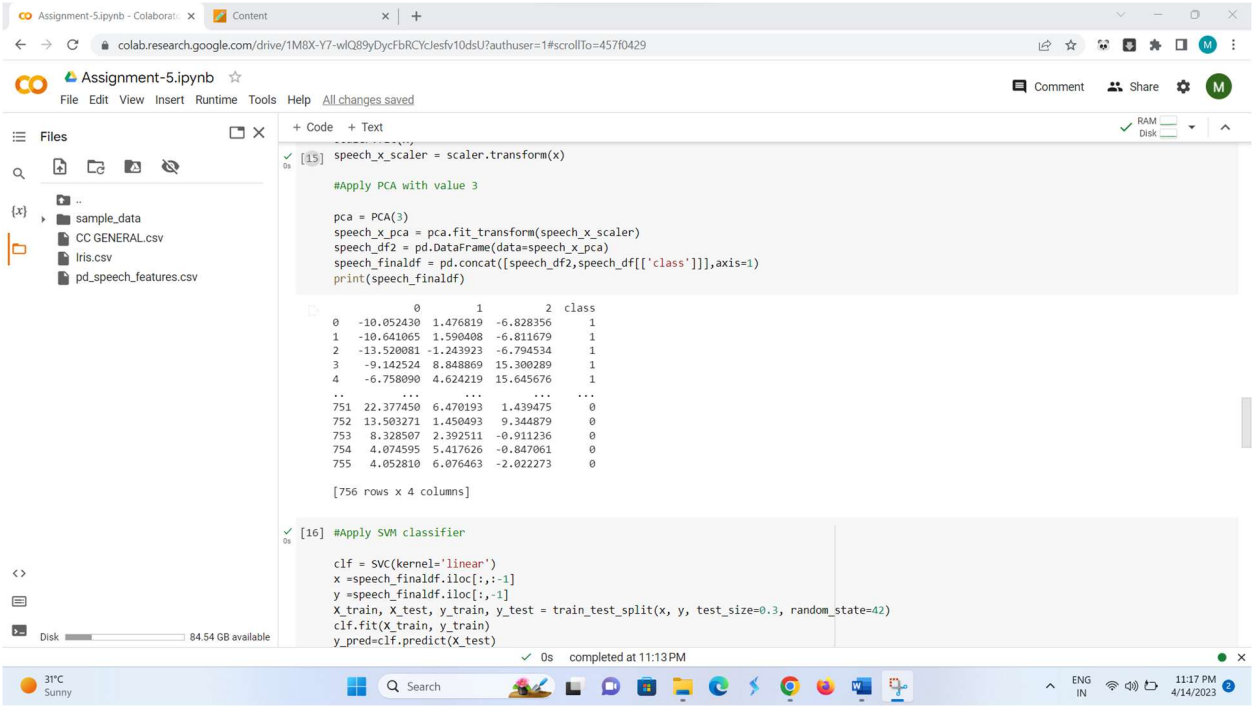


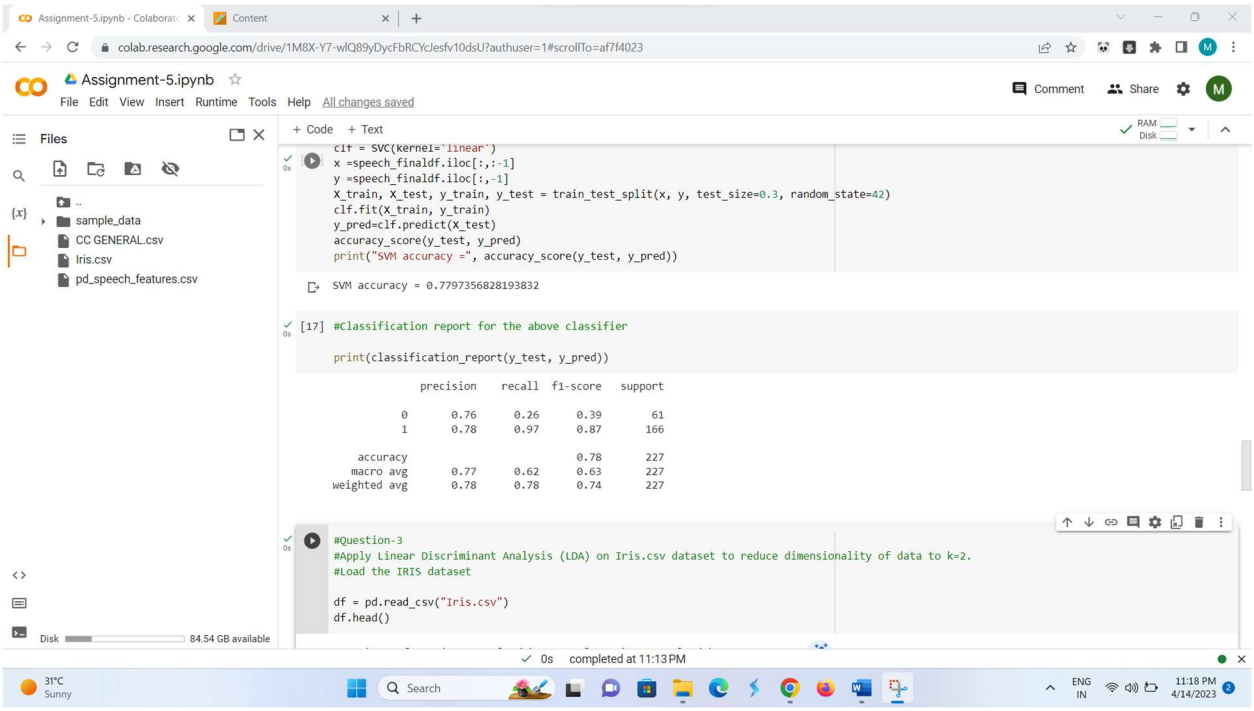

Assignment-5.ipynb - Colaborat
Content
colab.research.google.com/drive/1M8X-Y7-wlQ89yDycFbRCYcJesfv10dsU?authuser=1#scrollTo=457f0429
Assignment-5.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
CC GENERAL.csv
Iris.csv
pd_speech_features.csv
Disk 84.54 GB available
+ Code + Text
RAM
Disk
y_cluster_kmeans = k_mean.predict(df2)
score = metrics.silhouette_score(df2, y_cluster_kmeans)
print('Silhoutte score for PCA:',score)
Silhoutte score for PCA: 0.46475527710405057
[11] #Apply scaling on the dataset
scaler = StandardScaler()
scaler.fit(X)
x_scaler = scaler.transform(X)
#Apply PCA with k value as 2 again
pca = PCA(2)
x_pca = pca.fit_transform(x_scaler)
df2 = pd.DataFrame(data=x_pca)
finaldf = pd.concat([df2,df[['TENURE']]],axis=1)
print(finaldf)
0 1 TENURE
0 -1.682211 -1.076432 12
1 -1.138264 2.506554 12
2 0.969687 -0.383524 12
3 -0.873645 0.043145 12
4 -1.599420 -0.688556 12
... ...
8945 -0.359633 -2.016158 6
8946 -0.564370 -1.639139 6
8947 -0.926205 -1.810795 6
8948 -2.336539 -0.657952 6
0s completed at 11:13 PM
Construction
2.6 miles away
Search
ENG
IN
11:16 PM
4/14/2023


Assignment-5.ipynb - Colaborat
Content
colab.research.google.com/drive/1M8X-Y7-wlQ89yDycFbRCYcJesfv10dsU?authuser=1#scrollTo=457f0429
Assignment-5.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
CC GENERAL.csv
Iris.csv
pd_speech_features.csv
Disk 84.54 GB available
+ Code + Text
RAM
Disk
[11] 8949 -0.556418 -0.400469 6
[8950 rows x 3 columns]
#Apply k-means on the scaled PCA output
nclusters = 2
km = KMeans(n_clusters=nclusters)
km.fit(df2)
KMeans
KMeans(n_clusters=2)
[13] y_cluster_kmeans = km.predict(df2)
score = metrics.silhouette_score(finaldf, y_cluster_kmeans)
print('Silhoutte score for scaled+pca=kmeans:',score)
Silhoutte score for scaled+pca=kmeans: 0.3996207637696223
The score is reduced after performing the scaled PCA, so this data need not to undergo with PCA.
[14] #question-2
#Load the dataset
speech_df=pd.read_csv('pd_speech_features.csv')
speech_df.head()
id gender PPE DFA RPDE numPulses numPeriodsPulses meanPeriodPulses stdDevPeriodPulses locPctJitter ... tqwt_kurtosisValue_
0 0 1 0.85247 0.71826 0.57227 240 239 0.008064 0.000087 0.00218 ...
0s completed at 11:13 PM
Construction
2.6 miles away
Search
ENG
IN
11:17 PM
4/14/2023

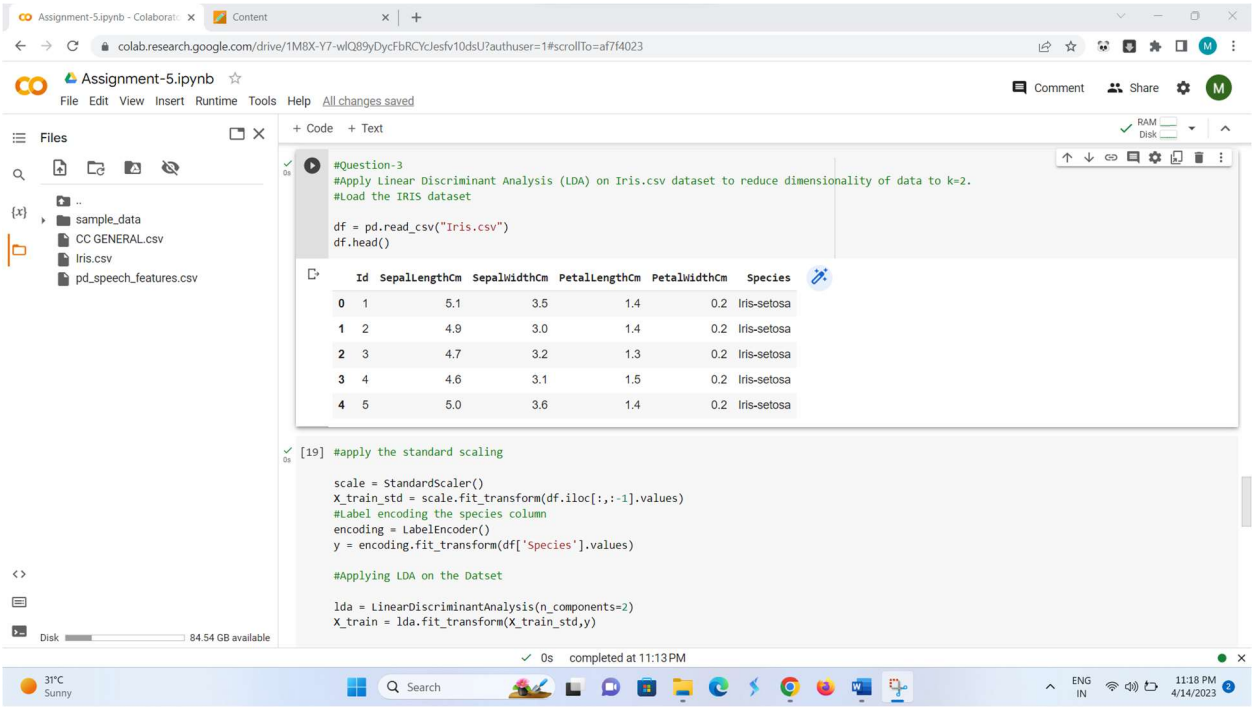
2)


Assignment-5.ipynb - Colaborat
Content
colab.research.google.com/drive/1M8X-Y7-wlQ89yDycFbRCYcJesfv10dsU?authuser=1#scrollTo=457f0429
Assignment-5.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
CC GENERAL.csv
Iris.csv
pd_speech_features.csv
Disk 84.54 GB available
+ Code + Text
RAM
Disk
#question-2
#Load the dataset
speech_df=pd.read_csv('pd_speech_features.csv')
speech_df.head()
id gender PPE DFA RPDE numPulses numPeriodsPulses meanPeriodPulses stdDevPeriodPulses locPctJitter ... tqwt_kurtosisValue_
0 0 1 0.85247 0.71826 0.57227 240 239 0.008064 0.000087 0.00218 ...
1 0 1 0.76686 0.69481 0.53966 234 233 0.008258 0.000073 0.00195 ...
2 0 1 0.85083 0.67604 0.58982 232 231 0.008340 0.000060 0.00176 ...
3 1 0 0.41121 0.79672 0.59257 178 177 0.010858 0.000183 0.00419 ...
4 1 0 0.32790 0.79782 0.53028 236 235 0.008162 0.002669 0.00535 ...
5 rows × 755 columns
[15] #Apply scaling on the dataset
x =speech_df.iloc[:,1:]
scaler = StandardScaler()
scaler.fit(x)
speech_x_scaler = scaler.transform(x)
#Apply PCA with value 3
0s completed at 11:13 PM
Construction
2.6 miles away
Search
ENG
IN
11:17 PM
4/14/2023


Assignment-5.ipynb - Colaborat
Content
colab.research.google.com/drive/1M8X-Y7-wlQ89yDycFbRCYcJesfv10dsU?authuser=1#scrollTo=457f0429
Assignment-5.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
CC GENERAL.csv
Iris.csv
pd_speech_features.csv
Disk 84.54 GB available
+ Code + Text
RAM
Disk
[15] speech_x_scaler = scaler.transform(x)
#Apply PCA with value 3
pca = PCA(3)
speech_x_pca = pca.fit_transform(speech_x_scaler)
speech_df2 = pd.DataFrame(data=speech_x_pca)
speech_finaldf = pd.concat([speech_df2,speech_df[['class']]],axis=1)
print(speech_finaldf)
0 1 2 class
0 -10.052430 1.476819 -6.828356 1
1 -10.641065 1.590408 -6.811679 1
2 -13.520081 -1.243923 -6.794534 1
3 -9.142524 8.848869 15.300289 1
4 -6.758090 4.624219 15.645676 1
... ...
751 22.377450 6.470193 1.439475 0
752 13.503271 1.450493 9.344879 0
753 8.328507 2.392511 -0.911236 0
754 4.074595 5.417626 -0.847061 0
755 4.052810 6.076463 -2.022273 0
[756 rows x 4 columns]
[16] #Apply SVM classifier
clf = SVC(kernel='linear')
x =speech_finaldf.iloc[:,:-1]
y =speech_finaldf.iloc[:,-1]
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)
clf.fit(X_train, y_train)
y_pred=clf.predict(X_test)
0s completed at 11:13 PM
31°C
Sunny
Search
ENG
IN
11:17 PM
4/14/2023


Assignment-5.ipynb - Colaborat
Content
colab.research.google.com/drive/1M8X-Y7-wlQ89yDycFbRCYcJesfv10dsU?authuser=1#scrollTo=af7f4023
Assignment-5.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
CC GENERAL.csv
Iris.csv
pd_speech_features.csv
+ Code + Text
clf = SVC(kernel='linear')
x =speech_finaldf.iloc[:,:-1]
y =speech_finaldf.iloc[:,-1]
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)
clf.fit(X_train, y_train)
y_pred=clf.predict(X_test)
accuracy_score(y_test, y_pred)
print("SVM accuracy =", accuracy_score(y_test, y_pred))
SVM accuracy = 0.7797356828193832
[17] #Classification report for the above classifier
print(classification_report(y_test, y_pred))
precision recall f1-score support
0 0.76 0.26 0.39 61
1 0.78 0.97 0.87 166
accuracy 0.78 227
macro avg 0.77 0.62 0.63 227
weighted avg 0.78 0.78 0.74 227
#Question-3
#Apply Linear Discriminant Analysis (LDA) on Iris.csv dataset to reduce dimensionality of data to k=2.
#Load the IRIS dataset
df = pd.read_csv("Iris.csv")
df.head()
Disk 84.54 GB available
0s completed at 11:13 PM
31°C Sunny
Search
ENG IN
11:18 PM 4/14/2023

3)


Assignment-5.ipynb - Colaborat
Content
colab.research.google.com/drive/1M8X-Y7-wlQ89yDycFbRCYcJesfv10dsU?authuser=1#scrollTo=af7f4023
Assignment-5.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
CC GENERAL.csv
Iris.csv
pd_speech_features.csv
+ Code + Text
#Question-3
#Apply Linear Discriminant Analysis (LDA) on Iris.csv dataset to reduce dimensionality of data to k=2.
#Load the IRIS dataset
df = pd.read_csv("Iris.csv")
df.head()
Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species
0 1 5.1 3.5 1.4 0.2 Iris-setosa
1 2 4.9 3.0 1.4 0.2 Iris-setosa
2 3 4.7 3.2 1.3 0.2 Iris-setosa
3 4 4.6 3.1 1.5 0.2 Iris-setosa
4 5 5.0 3.6 1.4 0.2 Iris-setosa
[19] #apply the standard scaling
scale = StandardScaler()
X_train_std = scale.fit_transform(df.iloc[:,:-1].values)
#Label encoding the species column
encoding = LabelEncoder()
y = encoding.fit_transform(df['Species'].values)
#Applying LDA on the Datset
lda = LinearDiscriminantAnalysis(n_components=2)
X_train = lda.fit_transform(X_train_std,y)
Disk 84.54 GB available
0s completed at 11:13 PM
31°C Sunny
Search
ENG IN
11:18 PM 4/14/2023

Assignment-5.ipynb - Colaborat | Content

colab.research.google.com/drive/1M8X-Y7-wlQ89yDycFbRCYcJesfv10dsU?authuser=1#scrollTo=ea7afb15

Assignment-5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

Files

.. sample_data CC GENERAL.csv Iris.csv pd_speech_features.csv

Disk 84.54 GB available

+ Code + Text

RAM Disk

```
#Applying LDA on the Datset

lda = LinearDiscriminantAnalysis(n_components=2)
X_train = lda.fit_transform(X_train_std,y)

data=pd.DataFrame(X_train)
data['class']=y
data.columns=["LD1","LD2","class"]
data.head()
```

	LD1	LD2	class
0	-10.036763	-0.451330	0
1	-9.172930	-1.477234	0
2	-9.480989	-0.979693	0
3	-8.818119	-1.408602	0
4	-9.960200	-0.112546	0

```
[20] markers = ['s', 'x', 'o']
     colors = ['y', 'b', 'g']
     sns.lmplot(x="LD1", y="LD2", data=data, hue='class', markers=markers, fit_reg=False, legend=False)
     plt.legend()
     plt.show()
```

0s completed at 11:13 PM

31°C Sunny

Search

ENG IN 11:18 PM 4/14/2023

Assignment-5.ipynb - Colaborat | Content

colab.research.google.com/drive/1M8X-Y7-wlQ89yDycFbRCYcJesfv10dsU?authuser=1#scrollTo=42034e94

Assignment-5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

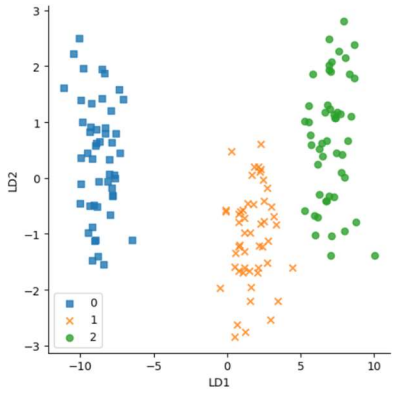
Files

.. sample_data CC GENERAL.csv Iris.csv pd_speech_features.csv

Disk 84.54 GB available

+ Code + Text

RAM Disk



1.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.

2.PCA finds directions of maximum variance regardless of class labels while LDA finds directions of maximum class separability.

0s completed at 11:13 PM

31°C Sunny

Search

ENG IN 11:19 PM 4/14/2023