

## 1. Principal Component Analysis

a. Apply PCA on CC dataset.

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

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

Silhouette Score- ranges from  $-1$  to  $+1$  , a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters

BB 13758115 x JupyterLite x Courses x +

← → ↻ jupyter.org/try-jupyter/lab/ 🔍 📄 ⚙️ 🌐

File Edit View Run Kernel Tabs Settings Help

Filter files by name 🔍

Name Last Modified

- Assignments\_... 42 minutes ago
- CC.csv an hour ago
- Clustering11... an hour ago
- Data Descript... 7 days ago
- Data Descript... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- lrs1.pdf an hour ago
- ML\_Assignme... 4 minutes ago
- ML\_ASSIGNME... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_7007571... 7 days ago
- PCA\_Feature R... an hour ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

```
[1]: # importing required libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from sklearn import preprocessing, metrics
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
# Importing PCA from scikit learn
from sklearn.decomposition import PCA
# Importing K-means from scikit learn
from sklearn.cluster import KMeans
sns.set(style='white', color_codes=True)
import warnings
warnings.filterwarnings("ignore")

[3]: #importing cc dataset and reading all the data from dataset
dataset_CC = pd.read_csv('datasets//CC.csv')
dataset_CC.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8950 entries, 0 to 8949
Data columns (total 18 columns):
 #   Column                Non-Null Count  Dtype
---  -
 0   CUST_ID               8950 non-null   object
 1   BALANCE               8950 non-null   float64
 2   BALANCE_FREQUENCY    8950 non-null   float64
 3   PURCHASES             8950 non-null   float64
 4   ONEOFF_PURCHASES     8950 non-null   float64
```

Simple 0 6 Python (Pyodide) | Idle Mode: Edit Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BB 13758115 x JupyterLite x Courses x +

← → ↻ jupyter.org/try-jupyter/lab/ 🔍 📄 ⚙️ 🌐

File Edit View Run Kernel Tabs Settings Help

Filter files by name 🔍

Name Last Modified

- Assignments\_... 43 minutes ago
- CC.csv an hour ago
- Clustering11... an hour ago
- Data Descript... 7 days ago
- Data Descript... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- lrs1.pdf an hour ago
- ML\_Assignme... 4 minutes ago
- ML\_ASSIGNME... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_7007571... 7 days ago
- PCA\_Feature R... an hour ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

```
4 ONEOFF_PURCHASES 8950 non-null float64
5 INSTALLMENTS_PURCHASES 8950 non-null float64
6 CASH_ADVANCE 8950 non-null float64
7 PURCHASES_FREQUENCY 8950 non-null float64
8 ONEOFF_PURCHASES_FREQUENCY 8950 non-null float64
9 PURCHASES_INSTALLMENTS_FREQUENCY 8950 non-null float64
10 CASH_ADVANCE_FREQUENCY 8950 non-null float64
11 CASH_ADVANCE_TRX 8950 non-null int64
12 PURCHASES_TRX 8950 non-null int64
13 CREDIT_LIMIT 8949 non-null float64
14 PAYMENTS 8950 non-null float64
15 MINIMUM_PAYMENTS 8637 non-null float64
16 PRC_FULL_PAYMENT 8950 non-null float64
17 TENURE 8950 non-null int64

dtypes: float64(14), int64(3), object(1)
memory usage: 1.2+ MB

[4]: dataset_CC.head()

[6]:
```

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUEI
0	C10001	40.900749	0.818182	95.40	0.00	95.4	0.000000	0.166667	0.000
1	C10002	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	0.000000	0.000

Simple 0 6 Python (Pyodide) | Idle Mode: Edit Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BD 13758115 x JupyterLite x Courses x +

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments\_... 44 minutes ago
- CC.csv an hour ago
- Clustering(1... an hour ago
- Data Descript... 7 days ago
- Data Descript... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- Iris1.pdf an hour ago
- ML\_Assignme... 5 minutes ago
- ML\_ASSIGNNM... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_7007571... 7 days ago
- PCA\_Feature R... an hour ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

2 C10003 2495.148862 1.000000 773.17 773.17 0.0 0.000000 1.000000 1.000

3 C10004 1666.070542 0.039364 1499.00 1499.00 0.0 205.788017 0.083333 0.083

4 C10005 817.714335 1.000000 16.00 16.00 0.0 0.000000 0.083333 0.083

```
[5]: #checking if dataset values has null values or not
dataset_cc.isnull().any()

[5]: CUST_ID False
BALANCE False
BALANCE_FREQUENCY False
PURCHASES False
ONEOFF_PURCHASES False
INSTALLMENTS_PURCHASES False
CASH_ADVANCE False
PURCHASES_FREQUENCY False
ONEOFF_PURCHASES_FREQUENCY False
PURCHASES_INSTALLMENTS_FREQUENCY False
CASH_ADVANCE_FREQUENCY False
CASH_ADVANCE_TRX False
PURCHASES_TRX False
CREDIT_LIMIT True
PAYMENTS False
MINIMUM_PAYMENTS True
PRC_FULL_PAYMENT False
TENURE False
dtype: bool

[6]: dataset_cc.fillna(dataset_cc.mean(), inplace=True)
dataset_cc.isnull().any()

[6]: CUST_ID False
BALANCE False
BALANCE_FREQUENCY False
PURCHASES False
```

Mode: Edit Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BD 13758115 x JupyterLite x Courses x +

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments\_... 44 minutes ago
- CC.csv an hour ago
- Clustering(1... an hour ago
- Data Descript... 7 days ago
- Data Descript... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- Iris1.pdf an hour ago
- ML\_Assignme... 5 minutes ago
- ML\_ASSIGNNM... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_7007571... 7 days ago
- PCA\_Feature R... an hour ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

```
[6]: CUST_ID False
BALANCE False
BALANCE_FREQUENCY False
PURCHASES False
ONEOFF_PURCHASES False
INSTALLMENTS_PURCHASES False
CASH_ADVANCE False
PURCHASES_FREQUENCY False
ONEOFF_PURCHASES_FREQUENCY False
PURCHASES_INSTALLMENTS_FREQUENCY False
CASH_ADVANCE_FREQUENCY False
CASH_ADVANCE_TRX False
PURCHASES_TRX False
CREDIT_LIMIT False
PAYMENTS False
MINIMUM_PAYMENTS False
PRC_FULL_PAYMENT False
TENURE False
dtype: bool

[7]: s = dataset_cc.iloc[:,1:-1]
t = dataset_cc.iloc[:, -1]
print(s.shape, t.shape)

(8958, 16) (8958,)

[8]: #2.a Applying PCA on CC Dataset
pca = PCA(3)
s_pca = pca.fit_transform(s)
principalDF = pd.DataFrame(data = s_pca, columns = ['principal component 1', 'principal component 2', 'principal component 3'])
finalDF = pd.concat([principalDF, dataset_cc.iloc[:, -1]], axis = 1)
finalDF.head()

[9]:
principal component 1 principal component 2 principal component 3 TENURE
0 -4326.383979 921.566582 183.708383 12
1 4118.916665 -2432.846346 2369.969289 12
2 1497.907641 -1997.578694 -2125.631328 12
```

Mode: Edit Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BD 13758115 x JupyterLite x Courses x +

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments\_... 44 minutes ago
- CC.csv an hour ago
- Clustering(1... an hour ago
- Data Descript... 7 days ago
- Data Descript... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- lrs1.pdf an hour ago
- ML\_Assignme... 6 minutes ago
- ML\_ASSIGNM... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_700757... 7 days ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

3 1394.548536 -1488.743453 -2431.799549 12

4 -3743.351896 757.342657 512.476492 12

```

[10]: #1.b Applying K Means on PCA Result
      S = finalDf.iloc[:,0:-1]
      t = finalDf.iloc[:, -1]

[11]: nclusters = 3 # this is the k in kmeans i.e., 3
      km = KMeans(n_clusters=nclusters)
      km.fit(S)

      # predict the cluster for each data point
      t_cluster_kmeans = km.predict(S)

      # Summary of the predictions made by the classifier
      print(classification_report(t, t_cluster_kmeans, zero_division=1))
      print(confusion_matrix(t, t_cluster_kmeans))

      train_accuracy = accuracy_score(t, t_cluster_kmeans)
      print("\nAccuracy for our Training dataset with PCA:", train_accuracy)

      #Calculating silhouette Score
      score = metrics.silhouette_score(S, t_cluster_kmeans)
      print("Silhouette Score: ", score)

```

	precision	recall	f1-score	support
0	0.00	1.00	0.00	0.0
1	0.00	1.00	0.00	0.0

Mode: Edit Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BD 13758115 x JupyterLite x Courses x +

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments\_... an hour ago
- CC.csv an hour ago
- Clustering(1... an hour ago
- Data Descript... 7 days ago
- Data Descript... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- lrs1.pdf an hour ago
- ML\_Assignme... 6 minutes ago
- ML\_ASSIGNM... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_700757... 7 days ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

2 0.00 1.00 0.00 0.0

6 1.00 0.00 0.00 284.0

7 1.00 0.00 0.00 190.0

8 1.00 0.00 0.00 196.0

9 1.00 0.00 0.00 175.0

10 1.00 0.00 0.00 236.0

11 1.00 0.00 0.00 365.0

12 1.00 0.00 0.00 7584.0

accuracy 0.00 8950.0

macro avg 0.70 0.30 0.00 8950.0

weighted avg 1.00 0.00 0.00 8950.0

```

[[ 0 0 0 0 0 0 0 0 0 0]
 [ 0 0 0 0 0 0 0 0 0 0]
 [ 0 0 0 0 0 0 0 0 0 0]
 [ 175 1 28 0 0 0 0 0 0 0]
 [ 173 2 15 0 0 0 0 0 0 0]
 [ 169 0 27 0 0 0 0 0 0 0]
 [ 149 0 26 0 0 0 0 0 0 0]

```

Mode: Edit Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BB 13758115 JupyterLite Courses

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments\_... an hour ago
- CC.csv an hour ago
- Clustering1... an hour ago
- Data Descripti... 7 days ago
- Data Descripti... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- Ist1.pdf an hour ago
- ML\_Assignme... 7 minutes ago
- ML\_ASSIGNM... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_700757... 7 days ago
- PCA\_Feature R... an hour ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

```

[14]: S = finalDf.iloc[:,0:-1]
      t = finalDf['TENURE']
      print(S.shape,t.shape)

(8958, 3) (8958,)

[15]: S_train, S_test, t_train, t_test = train_test_split(S,t, test_size=0.34,random_state=0)
      nclusters = 3
      # this is the k in kmeans
      km = KMeans(n_clusters=nclusters)
      km.fit(S_train,t_train)

      # predict the cluster for each training data point
      t_clus_train = km.predict(S_train)

      # Summary of the predictions made by the classifier
      print(classification_report(t_train, t_clus_train, zero_division=1))
      print(confusion_matrix(t_train, t_clus_train))

train_accuracy = accuracy_score(t_train, t_clus_train)
print("Accuracy for our Training dataset with PCA:", train_accuracy)

#Calculate silhouette Score
score = metrics.silhouette_score(S_train, t_clus_train)
print("Silhouette Score: ",score)

precision    recall  f1-score   support

0           0.00      1.00      0.00         0
1           0.00      1.00      0.00         0
2           0.00      1.00      0.00         0
6           1.00      0.00      0.00      139.0

```

Mode: Edit Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BB 13758115 JupyterLite Courses

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments\_... an hour ago
- CC.csv an hour ago
- Clustering1... an hour ago
- Data Descripti... 7 days ago
- Data Descripti... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- Ist1.pdf an hour ago
- ML\_Assignme... 7 minutes ago
- ML\_ASSIGNM... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_700757... 7 days ago
- PCA\_Feature R... an hour ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

```

Accuracy for our Training dataset with PCA: 0.0

Silhouette Score: 0.5189387274319468
[11]: "\nSilhouette Score- ranges from -1 to +1 , a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.\n"

[12]: #!c
      s = dataset_CC.iloc[:,1:-1]
      t = dataset_CC.iloc[:,1]
      print(x.shape,y.shape)

(8958, 16) (8958,)

[13]: # performing Scaling
      scaler = StandardScaler()
      scaler.fit(s)
      S_scaled_array = scaler.transform(s)
      # performing PCA
      pca = PCA(3)
      s_pca = pca.fit_transform(S_scaled_array)
      principalDf = pd.DataFrame(data = s_pca, columns = ['principal component 1', 'principal component 2','principal component 3'])
      finalDf = pd.concat([principalDf, dataset_CC.iloc[:,1]], axis = 1)
      finalDf.head()

[13]: .....
principal component 1 principal component 2 principal component 3 TENURE
0 -1.718893 -1.072939 0.535670 12
1 -1.169306 2.509320 0.628027 12
2 0.938414 -0.382600 0.161198 12
3 -0.907503 0.045859 1.521689 12
4 -1.637830 -0.684975 0.425658 12

```

Mode: Edit Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BD 13758115 JupyterLite Courses

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments\_... an hour ago
- CC.csv an hour ago
- Clustering(1... an hour ago
- Data Descripti... 7 days ago
- Data Descripti... an hour ago
- Iris.csv an hour ago
- K-Mean\_Datas... 7 days ago
- Iris1.pdf an hour ago
- ML\_Assignme... 8 minutes ago
- ML\_ASSIGNMEN... 14 days ago
- ML4\_7007403... 7 days ago
- ML4\_7007571... 7 days ago
- PCA\_Feature R... an hour ago
- PCA\_Feature R... an hour ago
- PCA.py an hour ago
- PCA+Classifier... an hour ago
- pd\_speech\_fea... an hour ago
- README.md 14 days ago
- Salary\_Data.csv 7 days ago
- Scaling+PCA+... an hour ago
- Untitled1.ipynb 21 days ago
- Untitled2.ipynb 21 days ago
- Untitled3.ipynb 21 days ago
- Untitled5.ipynb 7 days ago
- Untitled6.ipynb an hour ago

7 1.00 0.00 0.00 135.0

8 1.00 0.00 0.00 128.0

9 1.00 0.00 0.00 118.0

10 1.00 0.00 0.00 151.0

11 1.00 0.00 0.00 262.0

12 1.00 0.00 0.00 4974.0

accuracy 0.00 5907.0

macro avg 0.70 0.30 0.00 5907.0

weighted avg 1.00 0.00 0.00 5907.0

[[ 0 0 0 0 0 0 0 0 0 0 0]

[ 0 0 0 0 0 0 0 0 0 0 0]

[ 0 0 0 0 0 0 0 0 0 0 0]

[ 105 30 4 0 0 0 0 0 0 0 0]

[ 108 26 1 0 0 0 0 0 0 0 0]

[ 96 28 4 0 0 0 0 0 0 0 0]

[ 89 27 2 0 0 0 0 0 0 0 0]

[ 107 38 6 0 0 0 0 0 0 0 0]

[ 105 66 11 0 0 0 0 0 0 0 0]

Simple 0 6 Python (Pyodide) | idle Mode: Edit Ln 16, Col 34 ML\_Assignment\_S\_700740357.ipynb

Accuracy for our Training dataset with PCA: 0.0

Silhouette Score: 0.3812876198524835

[15]: '\nSilhouette Score- ranges from -1 to +1 , a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.\n'

```
[16]: # predict the cluster for each testing data point
t_clus_test = km.predict(S_test)

# Summary of the predictions made by the classifier
print(classification_report(t_test, t_clus_test, zero_division=1))
print(confusion_matrix(t_test, t_clus_test))

train_accuracy = accuracy_score(t_test, t_clus_test)
print("\nAccuracy for our Training dataset with PCA:", train_accuracy)

#Calculating silhouette Score
score = metrics.silhouette_score(S_test, t_clus_test)
print("Silhouette Score: ", score)
```

	precision	recall	f1-score	support
0	0.00	1.00	0.00	0.0
1	0.00	1.00	0.00	0.0
2	0.00	1.00	0.00	0.0
6	1.00	0.00	0.00	65.0
7	1.00	0.00	0.00	55.0
8	1.00	0.00	0.00	68.0
9	1.00	0.00	0.00	57.0
10	1.00	0.00	0.00	85.0

accuracy 0.00 3843.0

macro avg 0.70 0.30 0.00 3843.0

weighted avg 1.00 0.00 0.00 3843.0

```
[[ 0 0 0 0 0 0 0 0 0 0]
 [ 0 0 0 0 0 0 0 0 0 0]
 [ 0 0 0 0 0 0 0 0 0 0]
 [ 41 21 3 0 0 0 0 0 0 0]
 [ 42 12 1 0 0 0 0 0 0 0]
 [ 57 10 1 0 0 0 0 0 0 0]
 [ 35 22 0 0 0 0 0 0 0 0]
 [ 63 17 5 0 0 0 0 0 0 0]
 [ 69 30 4 0 0 0 0 0 0 0]
 [1763 450 397 0 0 0 0 0 0 0]]
```

Accuracy for our Training dataset with PCA: 0.0

Silhouette Score: 0.383322340968064

[16]: '\nSilhouette Score- ranges from -1 to +1 , a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.\n'

```
[18]: dataset_pd = pd.read_csv('datasets/pd_speech_features.csv')
dataset_pd.info()
```

2. Use pd\_speech\_features.csv
  - a. Perform Scaling
  - b. Apply PCA (k=3)
  - c. Use SVM to report performance

BB 13758115 JupyterLite Courses

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments...
- CC.csv
- Clustering(1)...
- Data Descripti...
- Data Descripti...
- Iris.csv
- K-Mean\_Datas...
- Iris1.pdf
- ML\_Assignme...
- ML\_ASSIGNMEN...
- ML4\_7007403...
- ML4\_700757...
- PCA\_Feature R...
- PCA\_Feature R...
- PCA.py
- PCA+Classifier...
- pd\_speech\_fea...
- README.md
- Salary\_Data.csv
- Scaling+PCA+
- Untitled1.ipynb
- Untitled2.ipynb
- Untitled3.ipynb
- Untitled5.ipynb
- Untitled6.ipynb

accuracy 0.00 3043.0

macro avg 0.70 0.30 0.00 3043.0

weighted avg 1.00 0.00 0.00 3043.0

```
[[ 0 0 0 0 0 0 0 0 0 0 0]
[ 0 0 0 0 0 0 0 0 0 0 0]
[ 0 0 0 0 0 0 0 0 0 0 0]
[ 41 21 3 0 0 0 0 0 0 0 0]
[ 42 12 1 0 0 0 0 0 0 0 0]
[ 57 10 1 0 0 0 0 0 0 0 0]
[ 35 22 0 0 0 0 0 0 0 0 0]
[ 63 17 5 0 0 0 0 0 0 0 0]
[ 69 30 4 0 0 0 0 0 0 0 0]
[1763 450 397 0 0 0 0 0 0 0 0]]
```

Accuracy for our Training dataset with PCA: 0.0

Silhouette Score: 0.383322340968964

```
[36]: 'The Silhouette Score- ranges from -1 to +1 , a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.\n'
```

```
[18]: dataset_pd = pd.read_csv('datasets/pd_speech_features.csv')
dataset_pd.info()
```

Mode: Command Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb

BB 13758115 JupyterLite Courses

jupyter.org/try-jupyter/lab/

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name Last Modified

- Assignments...
- CC.csv
- Clustering(1)...
- Data Descripti...
- Data Descripti...
- Iris.csv
- K-Mean\_Datas...
- Iris1.pdf
- ML\_Assignme...
- ML\_ASSIGNMEN...
- ML4\_7007403...
- ML4\_700757...
- PCA\_Feature R...
- PCA\_Feature R...
- PCA.py
- PCA+Classifier...
- pd\_speech\_fea...
- README.md
- Salary\_Data.csv
- Scaling+PCA+
- Untitled1.ipynb
- Untitled2.ipynb
- Untitled3.ipynb
- Untitled5.ipynb
- Untitled6.ipynb

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 756 entries, 0 to 755
Columns: 755 entries, id to class
dtypes: float64(749), int64(6)
memory usage: 4.4 MB
```

```
[19]: dataset_pd.head()
```

```
[19]:
```

	id	gender	PPE	DFA	RPDE	numPulses	numPeriodsPulses	meanPeriodPulses	stdDevPeriodPulses	locPctJitter	...	tqwt_kurtosisValue_dec_28	tqwt_kurtosisValue_dec_29	tqwt_kurtosisValue_dec_30
0	0	1	0.85247	0.71026	0.57227	240	239	0.008054	0.000087	0.00218	...	1.5620	2.6445	...
1	0	1	0.76686	0.69481	0.53966	234	233	0.008258	0.000073	0.00195	...	1.5589	3.6107	...
2	0	1	0.85083	0.67604	0.58982	232	231	0.008340	0.000060	0.00176	...	1.5643	2.3308	...
3	1	0	0.41121	0.79672	0.59257	178	177	0.010858	0.000183	0.00419	...	3.7805	3.5664	...
4	1	0	0.32790	0.79782	0.53028	236	235	0.008162	0.002669	0.00535	...	6.1727	5.8416	...

5 rows x 755 columns

```
[20]: dataset_pd.isnull().any()
```

```
[20]:
```

	id	gender	PPE	DFA	RPDE	numPulses	numPeriodsPulses	meanPeriodPulses	stdDevPeriodPulses	locPctJitter	...	tqwt_kurtosisValue_dec_28	tqwt_kurtosisValue_dec_29	tqwt_kurtosisValue_dec_30
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False

Mode: Command Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb



3115 x JupyterLite x Courses x +

jupyter.org/try-jupyter/lab/

File View Run Kernel Tabs Settings Help

files by name

Last Modified

Assignments... an hour ago

csv an hour ago

iterusing(1... an hour ago

a Descripti... 7 days ago

a Descripti... an hour ago

csv an hour ago

lean\_Datas... 7 days ago

.pdf an hour ago

Assignme... 9 minutes ago

.ASSIGNME... 14 days ago

1\_7007403... 7 days ago

1\_7007571... 7 days ago

Feature R... an hour ago

Feature R... an hour ago

upy an hour ago

+Classifier... an hour ago

speech\_fea... an hour ago

DME.md 14 days ago

ry\_Data.csv 7 days ago

ling+PCA+... an hour ago

itled1.ipynb 21 days ago

itled2.ipynb 21 days ago

itled3.ipynb 21 days ago

itled5.ipynb 7 days ago

itled6.ipynb an hour ago

Untitled6.ipynb x PCA\_Feature Reduction-X Assignments\_70074035 X ML\_Assignment\_5\_7007 X ML\_ASSIGNMENT\_7007 X CC.csv x Clustering(1).py x +

Python (Pyodide)

```
...
, tqwt_kurtosisValue_dec_33 False
, tqwt_kurtosisValue_dec_34 False
, tqwt_kurtosisValue_dec_35 False
, tqwt_kurtosisValue_dec_36 False
, class False
, length: 755, dtype: bool

+ [21]: S = dataset_pd.drop('class', axis=1).values
t = dataset_pd['class'].values

+ [22]: #Scaling Dataset
scaler = StandardScaler()
S_Scale = scaler.fit_transform(S)

+ [23]: # Applying PCA with k = 3
pca3 = PCA(n_components=3)
principalComponents = pca3.fit_transform(S_Scale)

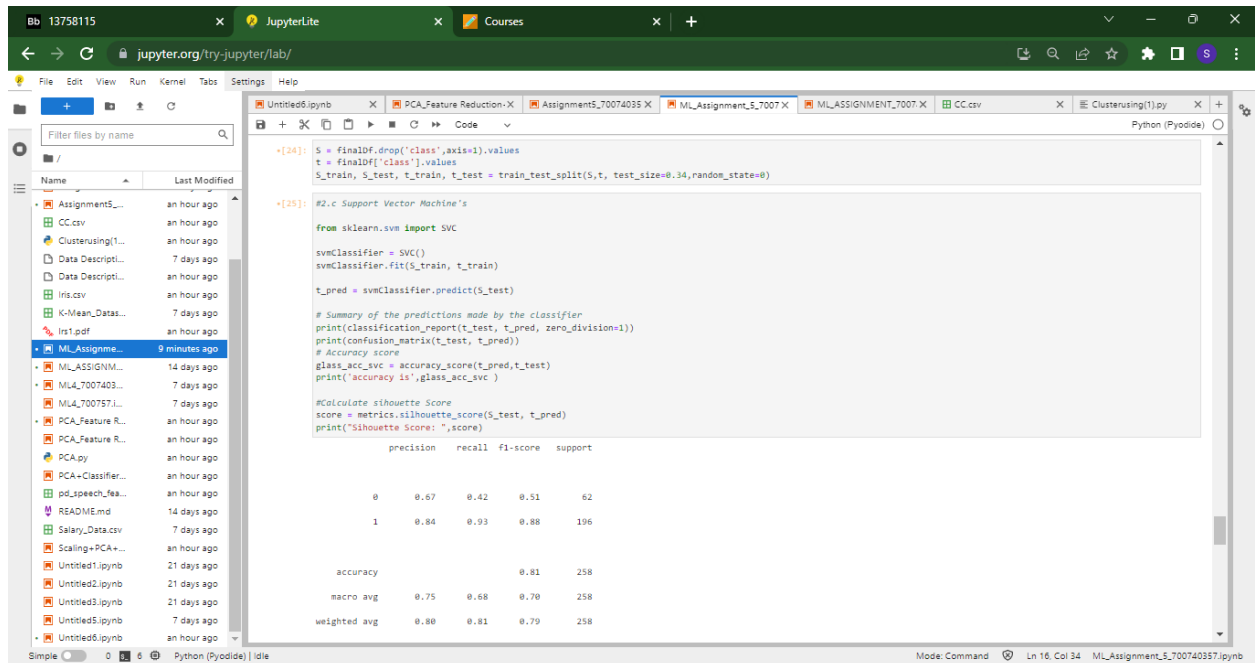
principalDf = pd.DataFrame(data = principalComponents, columns = ['principal component 1', 'principal component 2', 'Principal Component 3'])

finalDf = pd.concat([principalDf, dataset_pd[['class']]], axis = 1)
finalDf.head()

[23]: .....
principal component 1 principal component 2 Principal Component 3 class
0 -10.047372 1.471076 -6.846402 1
1 -10.637725 1.583749 -6.830976 1
2 -13.516185 -1.253542 -6.818696 1
3 -9.155084 8.833601 15.290906 1
4 -6.764470 4.611468 15.637121 1

+ [24]: S = finalDf.drop('class', axis=1).values
t = finalDf['class'].values
```

0 6 Python (Pyodide) | Idle Mode: Command Ln 16, Col 34 ML\_Assignment\_5\_700740357.ipynb



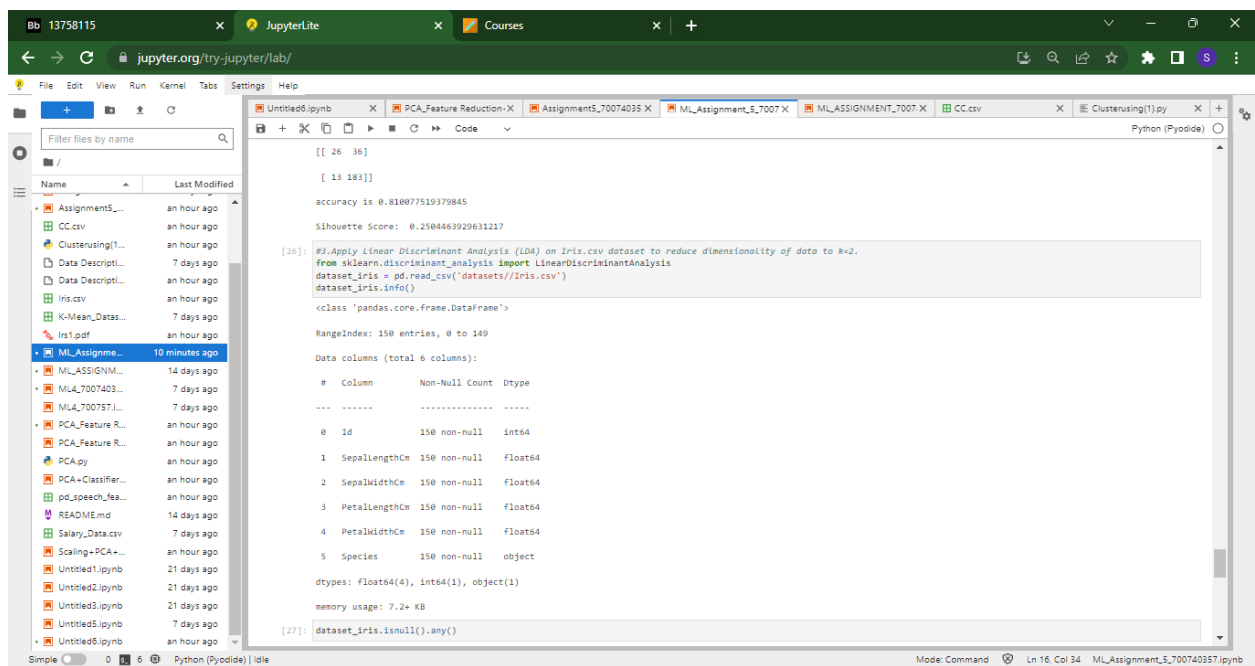
```
24: S = finalDF.drop("class",axis=1).values
    t = finalDF["class"].values
    S_train, S_test, t_train, t_test = train_test_split(S, t, test_size=0.34, random_state=0)

25: #2.c Support Vector Machine's
    from sklearn.svm import SVC
    svmClassifier = SVC()
    svmClassifier.fit(S_train, t_train)
    t_pred = svmClassifier.predict(S_test)

    # Summary of the predictions made by the classifier
    print(classification_report(t_test, t_pred, zero_division=1))
    print(confusion_matrix(t_test, t_pred))
    # Accuracy score
    glass_acc_svc = accuracy_score(t_pred, t_test)
    print("accuracy is", glass_acc_svc)

    #Calculate silhouette Score
    score = metrics.silhouette_score(S_test, t_pred)
    print("Silhouette Score: ", score)
```

	precision	recall	f1-score	support
0	0.67	0.42	0.51	62
1	0.84	0.93	0.88	196
accuracy			0.81	258
macro avg	0.75	0.68	0.70	258
weighted avg	0.80	0.81	0.79	258



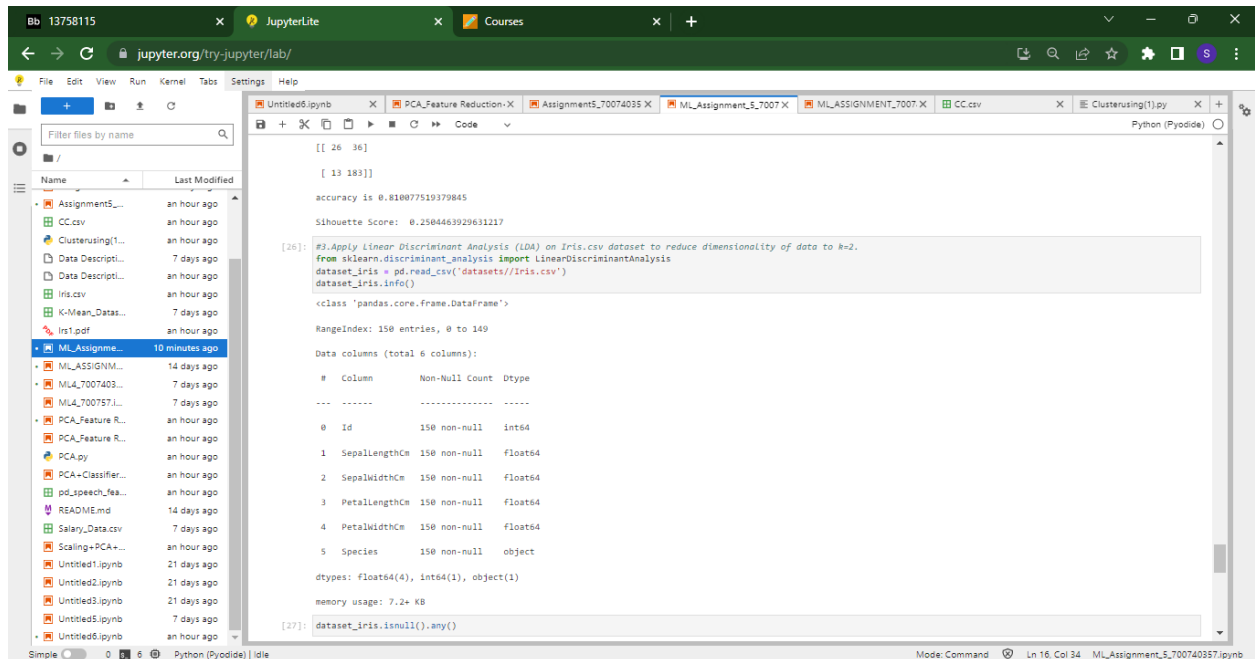
```
26: #3. Apply Linear Discriminant Analysis (LDA) on Iris.csv dataset to reduce dimensionality of data to k=2.
    from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
    dataset_iris = pd.read_csv('datasets/Iris.csv')
    dataset_iris.info()

    <class 'pandas.core.frame.DataFrame'>

    RangeIndex: 150 entries, 0 to 149
    Data columns (total 6 columns):
     #   Column      Non-Null Count  Dtype
    ---  -
     0   Id           150 non-null     int64
     1   SepalLengthCm 150 non-null     float64
     2   SepalWidthCm  150 non-null     float64
     3   PetalLengthCm 150 non-null     float64
     4   PetalWidthCm  150 non-null     float64
     5   Species       150 non-null     object
    dtypes: float64(4), int64(1), object(1)
    memory usage: 7.2+ KB

27: dataset_iris.isnull().any()
```

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



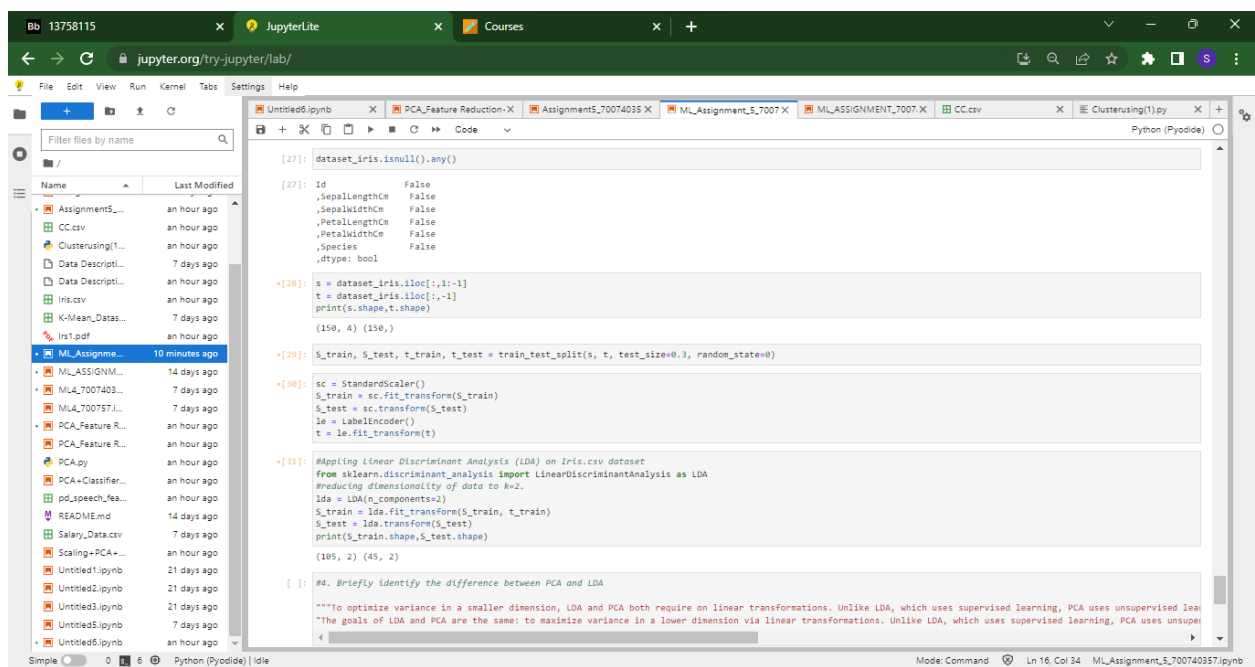
```
[[ 26 36]
 [ 13 183]]
accuracy is 0.810077519379845
Silhouette Score: 0.2584463929631217

[26]: #3. Apply Linear Discriminant Analysis (LDA) on Iris.csv dataset to reduce dimensionality of data to k=2.
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
dataset_iris = pd.read_csv('datasets/Iris.csv')
dataset_iris.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  --
 0   Id           150 non-null    int64  
 1   SepallengthCm 150 non-null    float64
 2   SepalwidthCm  150 non-null    float64
 3   PetallengthCm 150 non-null    float64
 4   PetalwidthCm  150 non-null    float64
 5   Species       150 non-null    object  
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB

[27]: dataset_iris.isnull().any()
```



```
[27]: dataset_iris.isnull().any()

[27]: Id           False
      SepallengthCm False
      SepalwidthCm  False
      PetallengthCm  False
      PetalwidthCm   False
      Species        False
      dtype: bool

+ [28]: s = dataset_iris.iloc[:,1:-1]
      t = dataset_iris.iloc[:,1]
      print(s.shape,t.shape)
      (150, 4) (150,)

+ [29]: S_train, S_test, t_train, t_test = train_test_split(s, t, test_size=0.3, random_state=0)

+ [30]: sc = StandardScaler()
      S_train = sc.fit_transform(S_train)
      S_test = sc.transform(S_test)
      le = LabelEncoder()
      t = le.fit_transform(t)

+ [31]: #Applying Linear Discriminant Analysis (LDA) on Iris.csv dataset
      from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
      #reducing dimensionality of data to k=2.
      lda = LDA(n_components=2)
      S_train = lda.fit_transform(S_train, t_train)
      S_test = lda.transform(S_test)
      print(S_train.shape,S_test.shape)
      (105, 2) (45, 2)

[ ]: #4. Briefly identify the difference between PCA and LDA

""""To optimize variance in a smaller dimension, LDA and PCA both require on linear
transformations. Unlike LDA, which uses supervised learning, PCA uses unsupervised learning.
Accordingly, LDA discovers directions of maximal class separability while PCA discovers
directions of maximum variance irrespective of class labels."
"The goals of LDA and PCA are the same: to maximize variance in a lower dimension via linear
transformations. Unlike LDA, which uses supervised learning, PCA uses unsupervised learning."
```

#### 4. Briefly identify the difference between PCA and LDA?

""""To optimize variance in a smaller dimension, LDA and PCA both require on linear transformations. Unlike LDA, which uses supervised learning, PCA uses unsupervised learning. Accordingly, LDA discovers directions of maximal class separability while PCA discovers directions of maximum variance irrespective of class labels."

"The goals of LDA and PCA are the same: to maximize variance in a lower dimension via linear transformations. Unlike LDA, which uses supervised learning, PCA uses unsupervised learning."

This indicates that although LDA identifies pathways of maximum class separability, PCA finds directions of maximum variance regardless of class labels.

The screenshot displays a JupyterLab environment. On the left, a file browser shows a directory structure with various files and notebooks. The main area on the right is a code editor for a file named 'Untitled5.ipynb'. The code within the editor is as follows:

```
[27]: dataset_iris.isnull().any()

[27]: Id          False
      ,SepalLengthCm  False
      ,SepalWidthCm   False
      ,PetalLengthCm  False
      ,PetalWidthCm   False
      ,Species        False
      ,dtype: bool

[28]: s = dataset_iris.iloc[:,1:-1]
      t = dataset_iris.iloc[:,1]
      print(s.shape, t.shape)
      (150, 4) (150,)

[29]: S_train, S_test, t_train, t_test = train_test_split(s, t, test_size=0.3, random_state=0)

[30]: sc = StandardScaler()
      S_train = sc.fit_transform(S_train)
      S_test = sc.transform(S_test)
      le = LabelEncoder()
      t = le.fit_transform(t)

[31]: #Applying Linear Discriminant Analysis (LDA) on Iris.csv dataset
      from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
      #reducing dimensionality of data to k=2.
      lda = LDA(n_components=2)
      S_train = lda.fit_transform(S_train, t_train)
      S_test = lda.transform(S_test)
      print(S_train.shape, S_test.shape)
      (105, 2) (45, 2)

[ ]: #1. Briefly identify the difference between PCA and LDA

      """To optimize variance in a smaller dimension, LDA and PCA both require on linear transformations. Unlike LDA, which uses supervised learning, PCA uses unsupervised learning.
      The goals of LDA and PCA are the same: to maximize variance in a lower dimension via linear transformations. Unlike LDA, which uses supervised learning, PCA uses unsupervised learning."""
```