

```

In [1]: import numpy as np
import matplotlib.pyplot as plt
import sklearn.preprocessing

# Importing the data set
dataset = pd.read_csv('data.csv', index_col = 0)
X = dataset[['fea1','fea2','fea3','fea4','fea5','fea6','fea7','fea8','fea9','fea10','...','fea776','fea777','fea778','fea779','fea780','fea781','fea782','fea783','fea784','fea785','fea786','fea787','fea788','fea789','fea790','fea791','fea792','fea793','fea794','fea795','fea796','fea797','fea798','fea799','fea800','fea801','fea802','fea803','fea804','fea805','fea806','fea807','fea808','fea809','fea810','fea811','fea812','fea813','fea814','fea815','fea816','fea817','fea818','fea819','fea820','fea821','fea822','fea823','fea824','fea825','fea826','fea827','fea828','fea829','fea830','fea831','fea832','fea833','fea834','fea835','fea836','fea837','fea838','fea839','fea840','fea841','fea842','fea843','fea844','fea845','fea846','fea847','fea848','fea849','fea850','fea851','fea852','fea853','fea854','fea855','fea856','fea857','fea858','fea859','fea860','fea861','fea862','fea863','fea864','fea865','fea866','fea867','fea868','fea869','fea870','fea871','fea872','fea873','fea874','fea875','fea876','fea877','fea878','fea879','fea880','fea881','fea882','fea883','fea884','fea885','fea886','fea887','fea888','fea889','fea890','fea891','fea892','fea893','fea894','fea895','fea896','fea897','fea898','fea899','fea900','fea901','fea902','fea903','fea904','fea905','fea906','fea907','fea908','fea909','fea910','fea911','fea912','fea913','fea914','fea915','fea916','fea917','fea918','fea919','fea920','fea921','fea922','fea923','fea924','fea925','fea926','fea927','fea928','fea929','fea930','fea931','fea932','fea933','fea934','fea935','fea936','fea937','fea938','fea939','fea940','fea941','fea942','fea943','fea944','fea945','fea946','fea947','fea948','fea949','fea950','fea951','fea952','fea953','fea954','fea955','fea956','fea957','fea958','fea959','fea960','fea961','fea962','fea963','fea964','fea965','fea966','fea967','fea968','fea969','fea970','fea971','fea972','fea973','fea974','fea975','fea976','fea977','fea978','fea979','fea980','fea981','fea982','fea983','fea984','fea985','fea986','fea987','fea988','fea989','fea990','fea991','fea992','fea993','fea994','fea995','fea996','fea997','fea998','fea999']]
y = dataset['l1'].values

In [3]: dataset.head()

Out[3]:
   fea1 fea2 fea3 fea4 fea5 fea6 fea7 fea8 fea9 fea10 ... fea776 fea777 fea778 fea779 fea780 fea781 fea782 fea783 fea784 fea785 fea786 fea787 fea788 fea789 fea790 fea791 fea792 fea793 fea794 fea795 fea796 fea797 fea798 fea799 fea800 fea801 fea802 fea803 fea804 fea805 fea806 fea807 fea808 fea809 fea810 fea811 fea812 fea813 fea814 fea815 fea816 fea817 fea818 fea819 fea820 fea821 fea822 fea823 fea824 fea825 fea826 fea827 fea828 fea829 fea830 fea831 fea832 fea833 fea834 fea835 fea836 fea837 fea838 fea839 fea840 fea841 fea842 fea843 fea844 fea845 fea846 fea847 fea848 fea849 fea850 fea851 fea852 fea853 fea854 fea855 fea856 fea857 fea858 fea859 fea860 fea861 fea862 fea863 fea864 fea865 fea866 fea867 fea868 fea869 fea870 fea871 fea872 fea873 fea874 fea875 fea876 fea877 fea878 fea879 fea880 fea881 fea882 fea883 fea884 fea885 fea886 fea887 fea888 fea889 fea890 fea891 fea892 fea893 fea894 fea895 fea896 fea897 fea898 fea899 fea900 fea901 fea902 fea903 fea904 fea905 fea906 fea907 fea908 fea909 fea910 fea911 fea912 fea913 fea914 fea915 fea916 fea917 fea918 fea919 fea920 fea921 fea922 fea923 fea924 fea925 fea926 fea927 fea928 fea929 fea930 fea931 fea932 fea933 fea934 fea935 fea936 fea937 fea938 fea939 fea940 fea941 fea942 fea943 fea944 fea945 fea946 fea947 fea948 fea949 fea950 fea951 fea952 fea953 fea954 fea955 fea956 fea957 fea958 fea959 fea960 fea961 fea962 fea963 fea964 fea965 fea966 fea967 fea968 fea969 fea970 fea971 fea972 fea973 fea974 fea975 fea976 fea977 fea978 fea979 fea980 fea981 fea982 fea983 fea984 fea985 fea986 fea987 fea988 fea989 fea990 fea991 fea992 fea993 fea994 fea995 fea996 fea997 fea998 fea999
0  1.0107711  0.9667814  0.3595936  ...  -1.03428476  1.04733254

In [4]: # Normalizing scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X = sc.fit_transform(X)

In [5]: print(X)

[[ 1.0107711  0.9667814  0.3595936  ...  -1.03428476  1.04733254
  1.4943731  1.2195292  0.91426816  ...  0.373791401  -1.70201871
  1.4817476  1.2195292  0.91426816  ...  0.373791401  -1.70201871
  0.98435881  ...
  [-1.01222007  0.30121327 -1.64108987  ...  -1.03428476  -0.2620708
  0.9435588]
...
  [-0.34412101  0.30121327 -0.30730086  ...  -1.03428476  -0.91677246
  0.3224815]
  1.4817476  0.9667814  0.3595936  ...  -1.03428476  1.04733254
  0.98435881  ...
  [ 0.373791401  0.30121327 -0.97419536  ...  -1.03428476  0.32629087
  -1.03428476]]

In [6]: # Applying the PCA transformation
from sklearn.decomposition import PCA
pca = PCA(n_components=X.shape[1]*0.95)
X2 = pca.fit_transform(X)

In [7]: X2 = X[:,1:]
X2.shape

Out[7]: (2066, 2)

In [8]: dataset['gnd'].value_counts()

Out[8]:
gnd
0    1146080  -0.489307  0
1    360911  -0.489307  0
3    431248  -0.402886  0
5    1806120  -1.862436  0
-    -
2062    1.975760  -0.939748  4
2063    0.525752  -2.646714  4
2064    -0.349422  -0.903681  4
2065    -3.115283  -2.004704  4
2066    -5.604704  0.246167  4
-    -
2066 rows x 1 columns

In [9]: prin2 = PCAFrame(X2, columns=['pc1','pc2'], index = range(1,2067))
prin2_gnd = pd.concat([prin2,dataset[['gnd']]], axis = 1)
prin2_gnd

Out[9]:
   pc1  pc2 gnd
0  9.970892 -6.18722  0
1 11.416600 -0.489307  0
3 3.609111 -0.489307  0
4 3.712408 -0.402886  0
5 18.061200 -1.862436  0
-    -
2062 1.975760 -0.939748  4
2063 0.525752 -2.646714  4
2064 -0.349422 -0.903681  4
2065 -3.115283 -2.004704  4
2066 -5.604704 0.246167  4
-    -
2066 rows x 3 columns

In [10]: data2 = prin2_gnd[prin2_gnd['gnd']==0]
data1 = prin2_gnd[prin2_gnd['gnd']==1]
data3 = prin2_gnd[prin2_gnd['gnd']==2]
data4 = prin2_gnd[prin2_gnd['gnd']==3]
data5 = prin2_gnd[prin2_gnd['gnd']==4]

In [11]: plt.figure(figsize=(8,6))

plt.scatter(data0['pc1'],data0['pc2'],label = 0)
plt.scatter(data1['pc1'],data1['pc2'],label = 1)
plt.scatter(data2['pc1'],data2['pc2'],label = 2)
plt.scatter(data3['pc1'],data3['pc2'],label = 3)
plt.scatter(data4['pc1'],data4['pc2'],label = 4)
plt.legend()
plt.xlabel('1st PCA Component')
plt.ylabel('2nd PCA Component')

Out[11]: Text(0, 0.5, '2nd PCA Component')



In [12]: X3 = XX[:,1:]
X3.shape

Out[12]: (2066, 2)

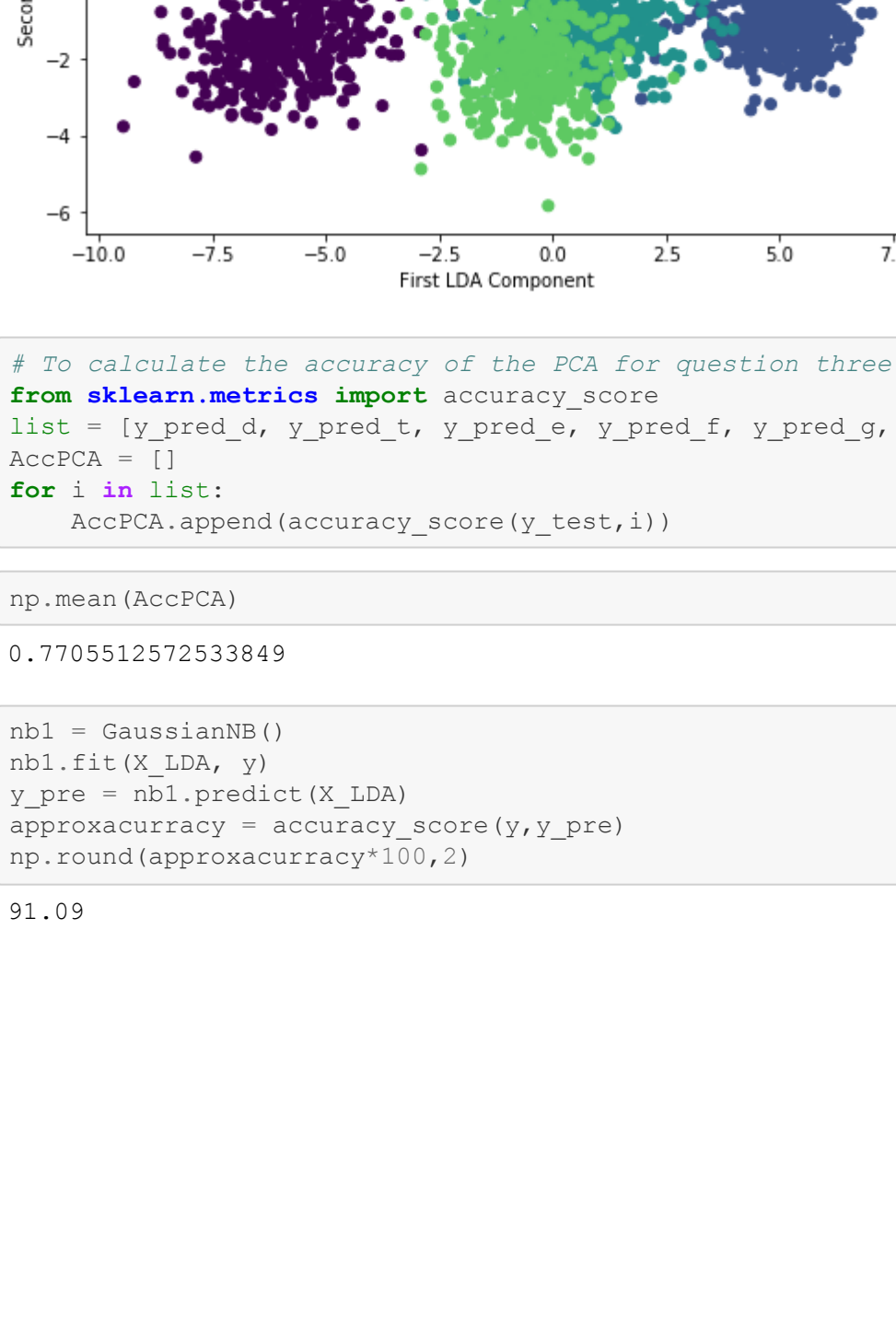
In [13]: prin3 = pd.DataFrame(X3, columns=['pc3','pc4'], index = range(1,2067
```



```
In [38]: X_LDA
Out[38]: array([[ -5.27723328, -2.0529124 ],
               [ -5.91372692, -1.95348169],
               [-4.15454288, -0.86888763],
               ...,
               [-1.31128997,  5.39808628],
               [  0.21324713,  5.1609646 ],
               [  0.40452267,  4.31739633]])

In [39]: plt.figure(figsize = (8,6))
plt.scatter(X_LDA[:,0],X_LDA[:,1],c = dataset['gnd'])
plt.xlabel('First LDA Component')
plt.ylabel('Second LDA Component')

Out[39]: Text(0, 0.5, 'Second LDA Component')
```



```
In [40]: # To calculate the accuracy of the PCA for question three
from sklearn.metrics import accuracy_score
list = [y_pred_d, y_pred_t, y_pred_e, y_pred_f, y_pred_g, y_pred_h, y_pred_3, y_pred_k]
AccPCA = []
for i in list:
    AccPCA.append(accuracy_score(y_test,i))

In [45]: np.mean(AccPCA)

Out[45]: 0.7705512572533849

In [42]: nb1 = GaussianNB()
nb1.fit(X_LDA, y)
y_pre = nb1.predict(X_LDA)
approxaccuracy = accuracy_score(y,y_pre)
np.round(approxaccuracy*100,2)

Out[42]: 91.09
```