Data Analytics
Assignment -2

Name : R Siva Girish SRN : PES1201700159

Dataset : Pima Indians Diabetes Database

Dataset Link:https://www.kaggle.com/uciml/pima-indians-diabetes-database

- ❖About the Dataset
  - ➢Dataset contain 768 rows and 9 columns

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

- ❖Principal Component Analysis
  - ➢Each Principal component is a linear combination of the original predictor variables which captures the maximum variance in the dataset.
  - ➢Normalization of data is very important as pca calculates a new projection of the dataset.A variable with higher variance will have a higher weight for calculation of axis.Therefore if you normalize data all variables will have the same Standard deviation and weights.Thereby PCA will calculate the relevant axis.

❖Analysis upon normalization of data :

Eigenvalues and EigenVectors :

```
[16]:
from sklearn.preprocessing import StandardScaler
X_std = StandardScaler().fit_transform(X)
```

```
[17]: import numpy as np
covariance_matrix=np.cov(X_std.T)
eigen_values, eigen_vectors = np.linalg.eig(covariance_matrix)
```

```
[18]: print("Eigen Vectors : %s\n"  %eigen_vectors)
print("Eigen Values : %s\n"  %eigen_values)
```

```
Eigen Vectors : [[ 0.38158207  0.49414673  0.65222658 -0.14857713 -0.15379938 -0.36232433
   0.0824671   0.03170909]
 [ 0.37434867 -0.11175705 -0.01078906 -0.29872009  0.60295038  0.20216056
   0.56666246 -0.18027463]
 [ 0.38712137  0.1410168   0.1934107   0.31014481  0.07783483  0.66757149
  -0.44838419 -0.20951244]
 [ 0.36598526 -0.31312345 -0.13906149  0.41546805 -0.15475509 -0.43895253
   0.04388519 -0.59791498]
 [ 0.20494708 -0.46075221  0.11600963 -0.46057938  0.25527554 -0.25524647
  -0.6169501   0.09575335]
 [ 0.39121669 -0.37684671  0.07198996  0.38125392 -0.10406382  0.02267586
   0.19803826  0.70981429]
 [ 0.14514665 -0.29641765  0.0398598  -0.47436462 -0.70365704  0.34050277
   0.17594858 -0.14975033]
 [ 0.46295645  0.42805125 -0.70534297 -0.19333373 -0.10798649 -0.08896106
  -0.13532944  0.17224287]]

Eigen Values : [2.15201252 1.54892928 0.34804074 1.03109859 0.92287482 0.85587968
 0.62525396 0.52672122]
```

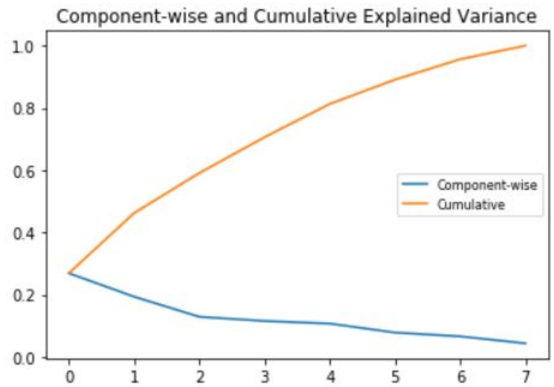
Code For PCA:

```
In [150]: from sklearn.decomposition import PCA

pca = PCA(n_components = 8)

#X_train = pca.fit_transform(X_train)
#X_test = pca.transform(X_test)
X_pca=pca.fit_transform(X)

explained_variance = pca.explained_variance_ratio_
```

```
In [151]: plt.plot(range(8), pca.explained_variance_ratio_,label="Component-wise")
plt.plot(range(8), np.cumsum(pca.explained_variance_ratio_),label="Cumulative")
plt.title("Component-wise and Cumulative Explained Variance")
plt.legend(loc='center right', fancybox=True, fontsize=8)
print("Component - wise variance explained",pca.explained_variance_ratio_)
print("Cumulative Variance explained : ",np.cumsum(pca.explained_variance_ratio_)[7])
```

We calculate the Component Wise variance as well as the cumulative variance and plot them on a graph and we can notice the component wise variance explained ratio.

```
Component - wise variance explained [0.26863854 0.19335487 0.12871339 0.11520367 0.10684058 0.07805127
 0.0657513  0.04344638]
Cumulative Variance explained :  1.0
```



It can be noticed that the first principal component accounts for around 27% of the variance in the data. The first 6 principal components account for around 90% of the variance.It can also be noted that the last principal component accounts for only 4% of the variance.

Therefore we can conclude that if we remove the last Principal component (Which accounts for only 4% of the variance) we can still manage to retain most of the data comfortably by not compromising the analysis.Thereby we can reduce the dimension of the dataset by 1.