DADS Experiment No: 2

Name: Abhishek S Waghchaure

PRN: **1032221714**

Dept: M Tech DSA

Aim:

Implementation of Feature Reduction using PCA in Python.

Objective:

Feature reduction using PCA on a Placement data of students.

Introduction:

Principal component analysis (PCA) is an unsupervised linear transformation technique which is primarily used for feature extraction and dimensionality reduction. It aims to find the directions of maximum variance in high-dimensional data and projects the data onto a new subspace with equal or fewer dimensions than the original one. The way PCA is different from other feature selection techniques such as random forest, regularization techniques, forward/backward selection techniques etc is that it does not require class labels to be present (thus called as unsupervised).

Dataset Used:

This data set consists of Placement data of students of some campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students.

Feature Description:

1. Gender- Male='M', Female='F'
2. ssc_p- Secondary Education percentage- 10th Grade
3. ssc_b - Board of Education- Central/ Others
4. hsc_p - Higher Secondary Education percentage- 12th Grade
5. hsc_b - Board of Education- Central/ Others
6. hsc_s - Specialization in Higher Secondary Education
7. degree_p - Degree Percentage
8. degree_t - Under Graduation (Degree type)- Field of degree education
9. workex - Work Experience
10. etest_p - Employability test percentage (conducted by college)
11. specialisation - Post Graduation (MBA)- Specialization
12. mba_p- MBA percentage
13. status- Status of placement- Placed/Not placed

14. salary - Salary offered by corporate to candidates

Sample data:

	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
0	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed	270000.0
1	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000.0
2	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed	250000.0
3	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed	NaN
4	5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed	425000.0

Code:

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.decomposition import PCA

data = pd.read_csv('Placement_Data.csv')
data.head()

data.drop(columns = ['sl_no'],inplace = True)
data.info()
data.isnull().sum()

# Filling Missing Values
data['salary'].fillna(0, inplace=True)
data.isnull().sum()
data.head()

# Change Categorical Variables into values
data["gender"] = data.gender.map({"M":0,"F":1})
data["hsc_s"] = data.hsc_s.map({"Commerce":0,"Science":1,"Arts":2})
data["degree_t"] = data.degree_t.map({"Comm&Mgmt":0,"Sci&Tech":1, "Others":2})
data["workex"] = data.workex.map({"No":0, "Yes":1})
data["status"] = data.status.map({"Not Placed":0, "Placed":1})
data["specialisation"] = data.specialisation.map({"Mkt&HR":0, "Mkt&Fin":1})
data['hsc_b'] = data.hsc_b.map({'Others':0, 'Central':1})
```

```
data['ssc_b'] = data.ssc_b.map({'Others':0, 'Central':1})
data.head()

x = data.loc[:,
['gender','ssc_p','ssc_b','hsc_p','hsc_b','hsc_s','degree_p','degree_t','workex','etest_p','specialisation','mba_p']]
y = data.loc[:,'status']

from sklearn.preprocessing import StandardScaler
x_std = StandardScaler().fit_transform(x)

from sklearn.preprocessing import StandardScaler
x_std = StandardScaler().fit_transform(x)

pca = PCA(n_components = 12)
x_pca = pca.fit_transform(x_std)
explained_variance = pca.explained_variance_ratio_
explained_variance

#PCA is known as Principal Component Analysis and is used to reduce the number of features and ultimately
the dimensionality.
pca = PCA(n_components=2)
X_new = pca.fit_transform(x)

#Let's plot the graphs before and after PCA
fig, axes = plt.subplots(1,2)

axes[0].scatter(x.iloc[:,0], x.iloc[:,1], c=y)
axes[0].set_xlabel('x1')
axes[0].set_ylabel('x2')
axes[0].set_title('Before PCA')

axes[1].scatter(X_new[:,0], X_new[:,1], c=y)
axes[1].set_xlabel('PC1')
axes[1].set_ylabel('PC2')
axes[1].set_title('After PCA')

plt.show()
```

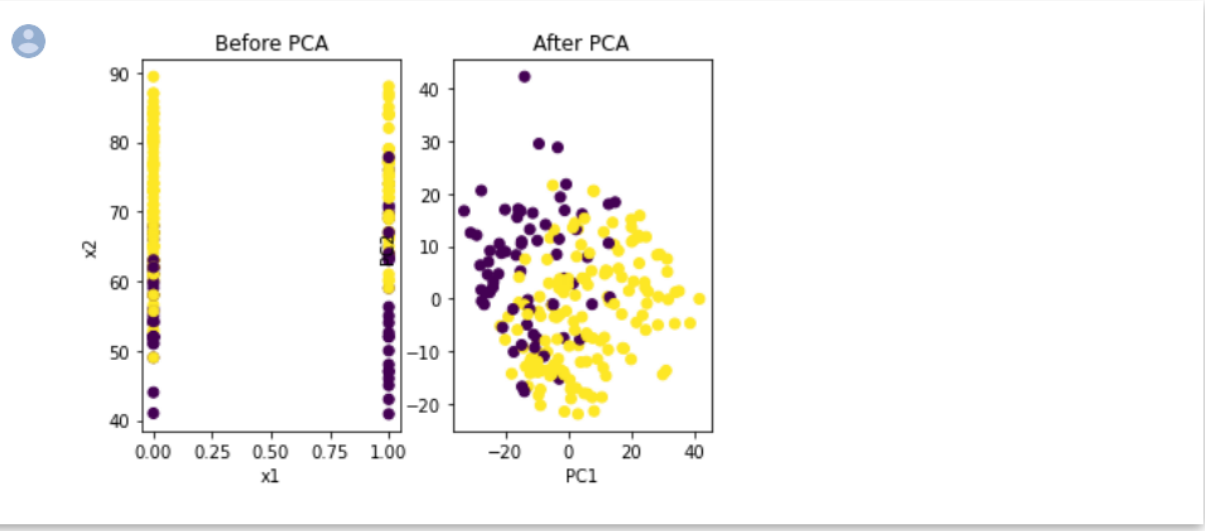
Output:

Below figure gives us the variance within the input variables to the target variable.

```
[ ]  explained_variance

     array([0.22485243, 0.14699017, 0.12962452, 0.10785902, 0.07799149,
            0.06879237, 0.06087671, 0.04598993, 0.04183893, 0.03825761,
            0.03206042, 0.02486641])
```

Below figure shows the output before applying PCA and after applying PCA.



Conclusion:

From given data we can say that Work Experience, Employability test percentage (conducted by college), Post-Graduation (MBA)- Specialization, MBA percentage matters most for placement of the students.

Reference:

1. https://www.projectpro.io/recipes/extract-features-using-pca-in-python
2. https://vitalflux.com/feature-extraction-pca-python-example/
3. Dataset:

   https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement?resource=download