GHARDA INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Machine Learning Lab BE Computer (Semester-VII)

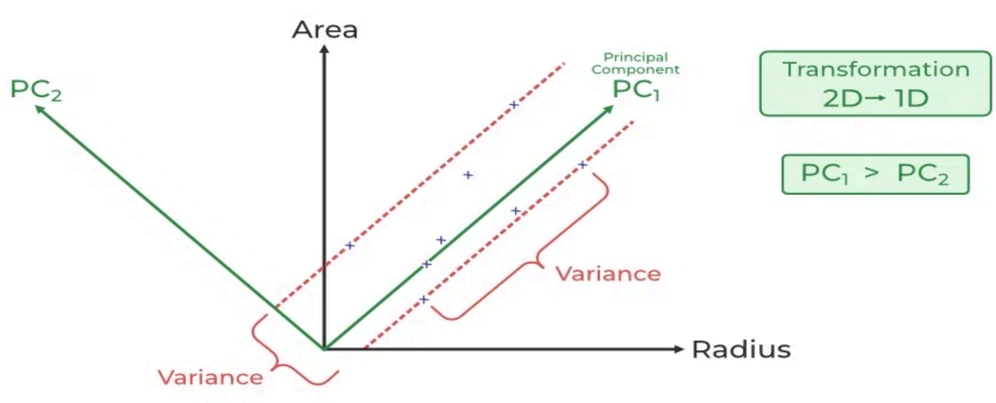
Experiment No.6: PCA (Principal Component Analysis)

Aim : To study, understand and implement a PCA (Principal Component Analysis), a dimensionality reduction technique used in machine learning.

Theory :

Principal Component Analysis (PCA) is a statistical method used to convert a dataset with correlated variables into a set of uncorrelated variables called **principal components**. These components are ordered such that the first principal component accounts for the largest possible variance in the data, the second accounts for the next largest variance, and so on.

1. **Standardization**: Transform the data to have zero mean and unit variance.
2. **Covariance Matrix**: Calculate the covariance matrix of the features in the dataset.
3. **Eigenvalues and Eigenvectors**: Compute the eigenvalues and corresponding eigenvectors of the covariance matrix. The eigenvectors represent the direction of the new feature space (principal components), while the eigenvalues determine their magnitude or importance.
4. **Principal Components**: Select the top kkk eigenvectors based on their corresponding eigenvalues to form a new feature subspace with reduced dimensions.
5. **Projection**: Transform the original dataset into this new subspace.



Discussion :

- **Dimensionality Reduction**: By reducing the number of features, PCA helps in speeding up model training and improving the performance of algorithms sensitive to the curse of dimensionality.
- **Data Visualization**: PCA allows for the visualization of high-dimensional data in two

or three dimensions, making it easier to detect patterns, clusters, or anomalies.

- **Noise Reduction**: By retaining only the components that capture the most variance, PCA helps to reduce noise in the data.
- **Interpretability**: While PCA simplifies data, interpreting the principal components can be challenging, as they are linear combinations of the original features.

Applications :

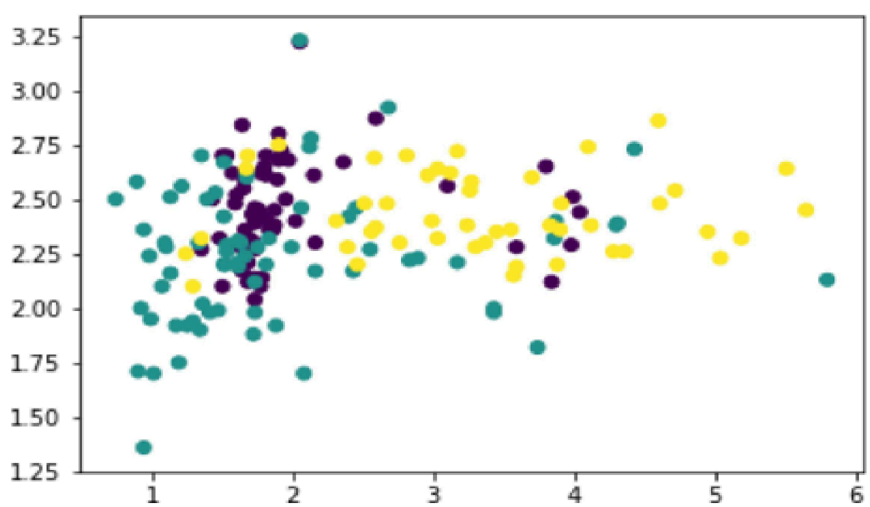
- **Image Compression**: PCA is used to reduce the dimensionality of image data, making storage and processing more efficient while maintaining image quality.
- **Face Recognition**: In computer vision, PCA is used for facial recognition by reducing the dimensionality of face images and retaining the most important features.
- **Finance**: PCA helps in reducing the number of variables in financial datasets for stock market predictions or portfolio management.
- **Bioinformatics**: PCA is applied in genomic and proteomic data to identify patterns or clusters in high-dimensional biological datasets.

Program Code :

In [3]:
```
from sklearn.datasets import load_wine
winedata = load_wine()
X, y = winedata['data'], winedata['target']
print(X.shape)
print(y.shape)
```

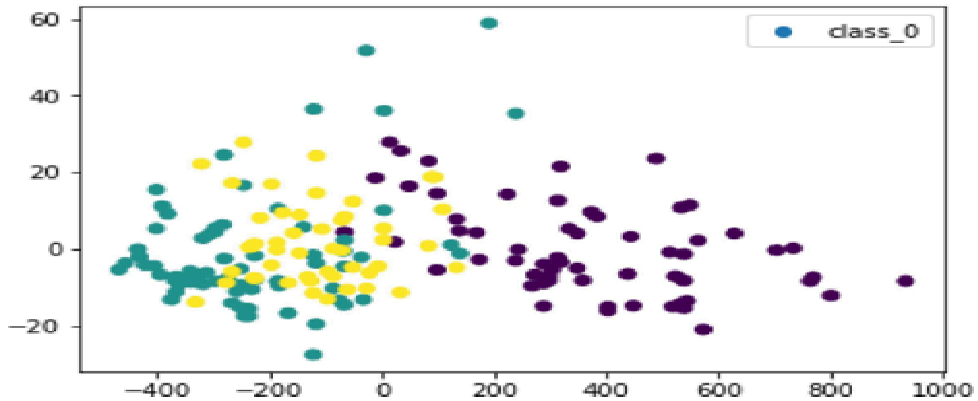
```
(178, 13)
(178,)
```

In [4]:
```
import matplotlib.pyplot as plt
plt.scatter(X[:,1], X[:,2], c=y)
plt.show()
```



In [7]:

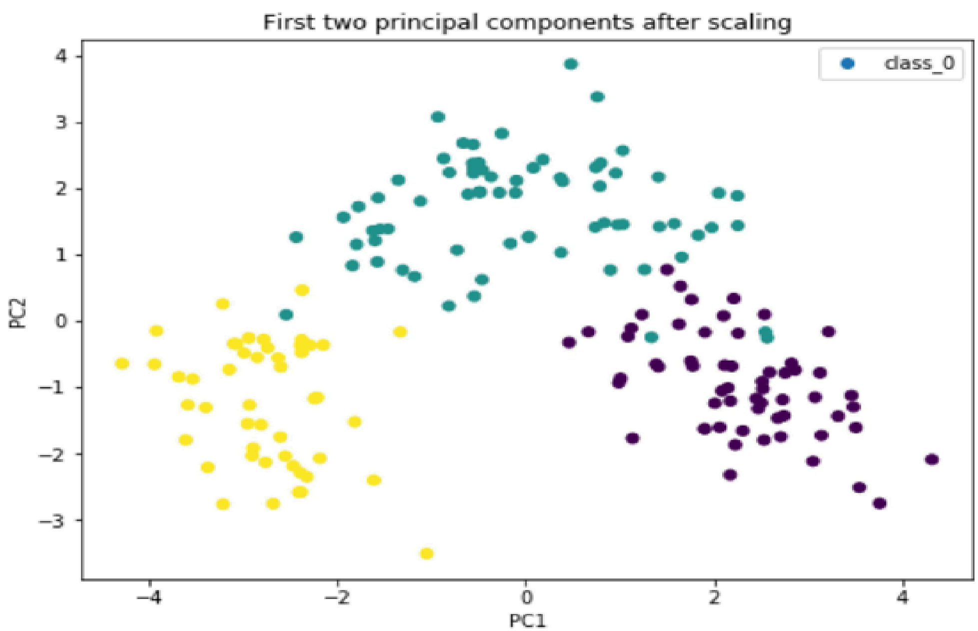
```
from sklearn.decomposition import PCA
pea = PCA()
Xt = pca.fit_transform(X)
plot = plt.scatter(Xt[:,0], Xt[:,1], c=y)
plt.legend(labels=list(winedata['target_names']))
plt.show()
```



In [11]:

```
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline

pca = PCA()
pipe = Pipeline([('scaler', StandardScaler()), ('pca', pca)])
plt.figure(figsize=(8,6))
Xt = pipe.fit_transform(X)
plot = plt.scatter(Xt[:,0], Xt[:,1], c=y)
plt.legend(labels=list(winedata['target_names']))
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("First two principal components after scaling")
plt.show()
```



Conclusion :

Principal Component Analysis (PCA) is a powerful technique for dimensionality reduction that simplifies high-dimensional datasets while preserving critical information. Despite some limitations in interpretability and potential loss of data, PCA remains a valuable tool in many machine learning applications, offering benefits such as improved computational efficiency and enhanced visualization of complex data.

References :

1. https://link.springer.com/book/10.1007/b98835
2. https://www.tandfonline.com/doi/abs/10.1080/14786440109462720
3. https://www.sciencedirect.com/science/article/abs/pii/0169743987800849
4. https://onlinelibrary.wiley.com/doi/full/10.1002/wics.101