MCKV Institute of Engineering

243, G.T. Road (N), Liluah, Howrah-711204

Machine Learning Lab

Assignment 21

Problem Statement:

Implement simple linear regression to predict student marks based on number of study hours using a small CSV dataset. Visualize regression line.

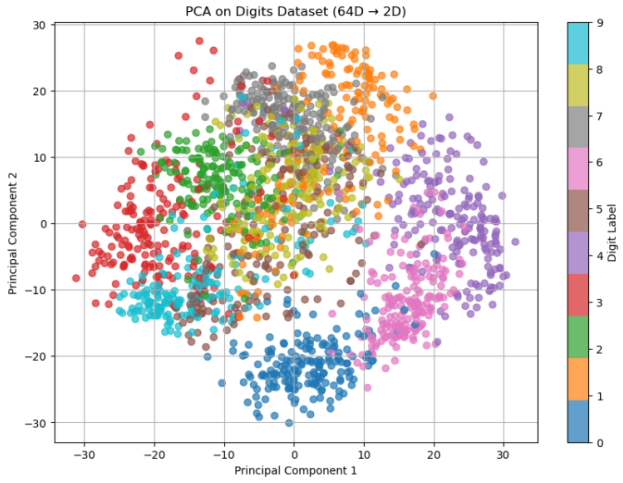
Program:

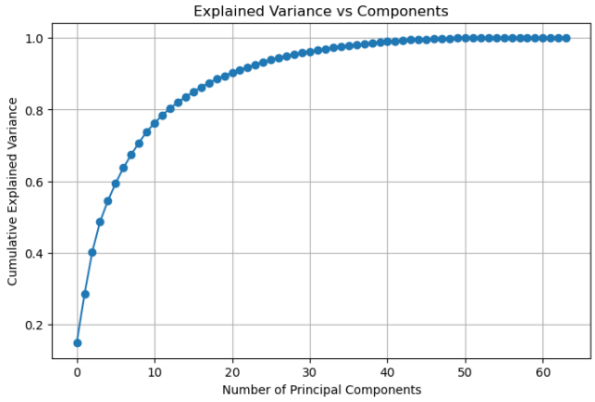
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.datasets import load_digits

# ------------------------------
# Step 1: Load Dataset
# ------------------------------
digits = load_digits()
X = digits.data      # (1797, 64) features (8x8 pixel images flattened)
y = digits.target    # digit labels (0 - 9)

print("Original shape:", X.shape)

# ------------------------------
# Step 2: Apply PCA
# ------------------------------
pca = PCA(n_components=2)  # reduce to 2D
X_pca = pca.fit_transform(X)

print("Reduced shape:", X_pca.shape)
print("Explained variance ratio:", pca.explained_variance_ratio_)

# ------------------------------
# Step 3: Visualization (2D scatter plot)
# ------------------------------
plt.figure(figsize=(10,7))
scatter = plt.scatter(X_pca[:,0], X_pca[:,1],
                      c=y, cmap="tab10", alpha=0.7, s=40)
plt.xlabel("Principal Component 1")
plt.ylabel("Principal Component 2")
```

```
plt.title("PCA on Digits Dataset (64D → 2D)")
plt.colorbar(scatter, label="Digit Label")
plt.grid(True)
plt.show()

# -------------------------------
# Step 4: Variance Explained Plot (Scree plot)
# -------------------------------
pca_full = PCA().fit(X)  # PCA with all components
plt.figure(figsize=(8,5))
plt.plot(np.cumsum(pca_full.explained_variance_ratio_), marker='o')
plt.xlabel("Number of Principal Components")
plt.ylabel("Cumulative Explained Variance")
plt.title("Explained Variance vs Components")
plt.grid(True)
plt.show()
```

Output:

```
Original shape: (1797, 64)
Reduced shape: (1797, 2)
Explained variance ratio: [0.14890594 0.13618771]
```





Viva Questions

Sl. No.	Question and Answer	Marks
1	**Q: What is Principal Component Analysis (PCA)?** **A:** PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving as much variance (information) as possible.	2
2	**Q: What is the main purpose of PCA?** **A:** The main purpose of PCA is to reduce the number of variables in a dataset while retaining most of the important information. It helps in visualization, noise reduction, and faster computation.	2
3	**Q: What are Principal Components?** **A:** Principal components are new variables formed as linear combinations of the original features. They are orthogonal to each other and ordered by the amount of variance they explain.	2
4	**Q: How does PCA determine the directions of maximum variance?** **A:** PCA computes the covariance matrix of the data and uses eigenvectors (directions) and eigenvalues (amount of variance) to identify the principal components.	2
5	**Q: What is the role of explained variance in PCA?** **A:** Explained variance tells us how much of the data's variability is captured by each principal component. It helps in deciding how many components should be kept.	2