Assignment-1 Report

Principal Component Analysis (PCA) on Iris Dataset

PADAMATI PRANAV SAI
Roll No: 2201AI49
Course: CS502

September 19, 2025

Contents

1 Introduction

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms correlated features into a smaller number of uncorrelated variables called principal components. In this assignment, PCA is applied to the Iris dataset to reduce its dimensionality and visualize the data in two dimensions.

2 Dataset Description

The Iris dataset contains 150 samples of iris flowers from three species:

- Setosa
- Versicolor
- Virginica

Each sample has 4 features: sepal length, sepal width, petal length, and petal width. The target variable indicates the species type.

3 Implementation in Python

The following Python code demonstrates PCA step by step.

3.1 Importing Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
```

3.2 Loading Dataset

```
iris = load_iris()
X = iris.data
y = iris.target
feature_names = iris.feature_names
target_names = iris.target_names

df = pd.DataFrame(X, columns=feature_names)
df['target'] = y
print(df.head())
```

3.3 Standardization

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

3.4 Applying PCA

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

print("Explained Variance Ratio:", pca.explained_variance_ratio_)
```

3.5 Visualization with PC1 & PC2

```
pca_df = pd.DataFrame(data=X_pca, columns=['PC1', 'PC2'])
pca_df['target'] = y

plt.figure(figsize=(8,6))
for i, target_name in enumerate(target_names):
    plt.scatter(pca_df.loc[pca_df['target']==i, 'PC1'],
                pca_df.loc[pca_df['target']==i, 'PC2'],
                label=target_name)

# Draw PC1 and PC2 axes
plt.axhline(0, color='gray', linestyle='--')
plt.axvline(0, color='gray', linestyle='--')

plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('PCA on Iris Dataset with PC1 and PC2')
plt.legend()
plt.show()
```

3.6 Explained Variance Plot

```
plt.figure(figsize=(6,4))
plt.plot(np.cumsum(pca.explained_variance_ratio_), marker='o')
plt.xlabel("Number of Components")
plt.ylabel("Cumulative Explained Variance")
plt.title("Explained Variance by PCA Components")
plt.grid()
plt.show()
```

4 Results and Discussion

- The explained variance ratio for the first two components is approximately:
  - PC1: 72%
  - PC2: 23%

  Together, these two components preserve about 95% of the dataset variance.
- The scatter plot shows clear separation of Setosa, while Versicolor and Virginica overlap slightly.

- By drawing PC1 and PC2 axes, we see how they capture the directions of maximum variance in the dataset.
- PCA effectively reduces 4D data into 2D for visualization without significant loss of information.

5 Conclusion

PCA is an effective dimensionality reduction method. For the Iris dataset, reducing from 4 dimensions to 2 dimensions retained 95% of the variance, making the dataset easier to visualize and interpret.

6 Screenshots

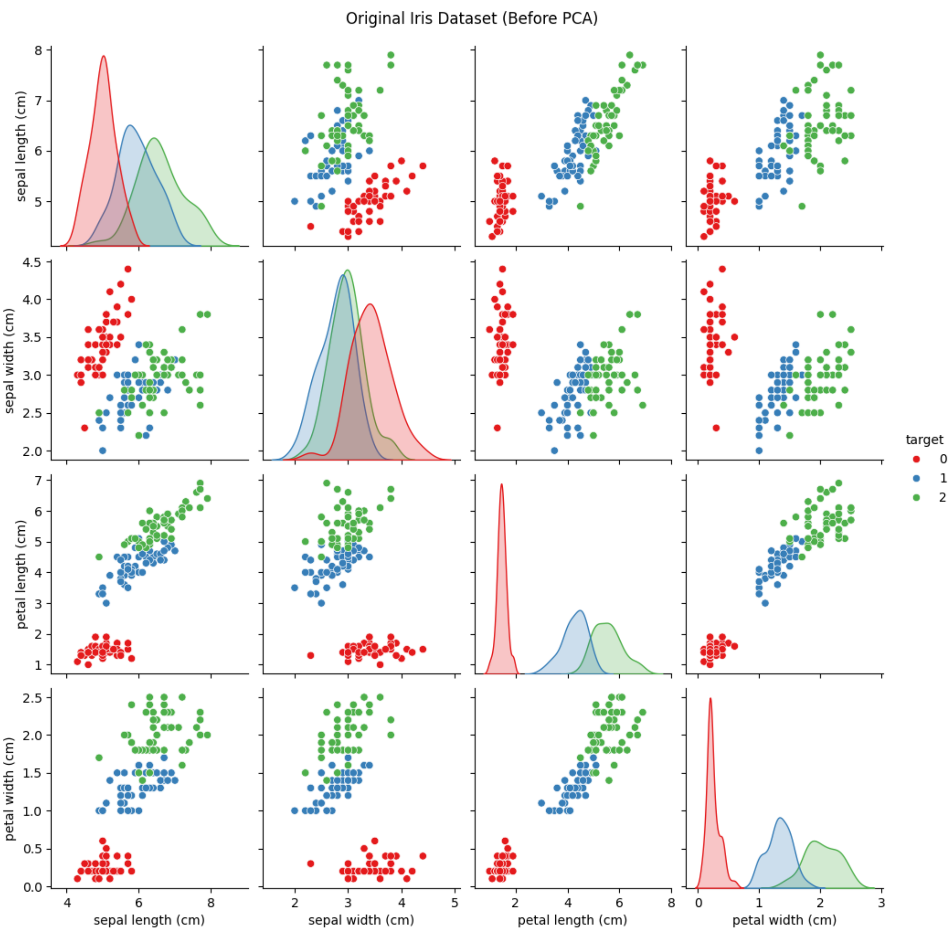

Figure 1: Pairplot of Iris dataset

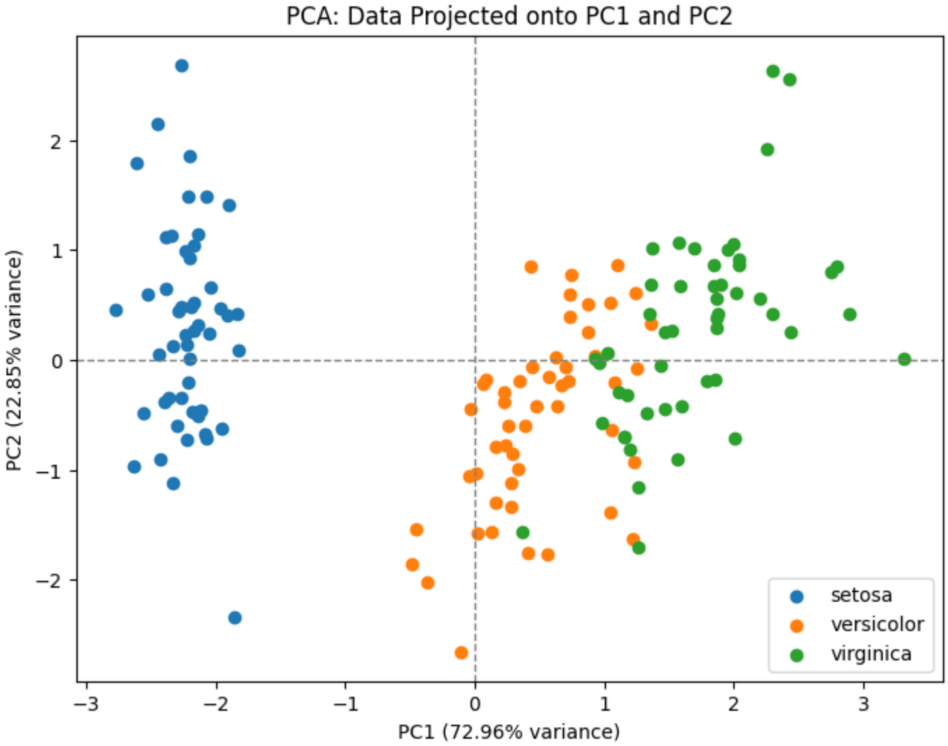

Figure 2: PCA scatter plot with PC1 and PC2 axes

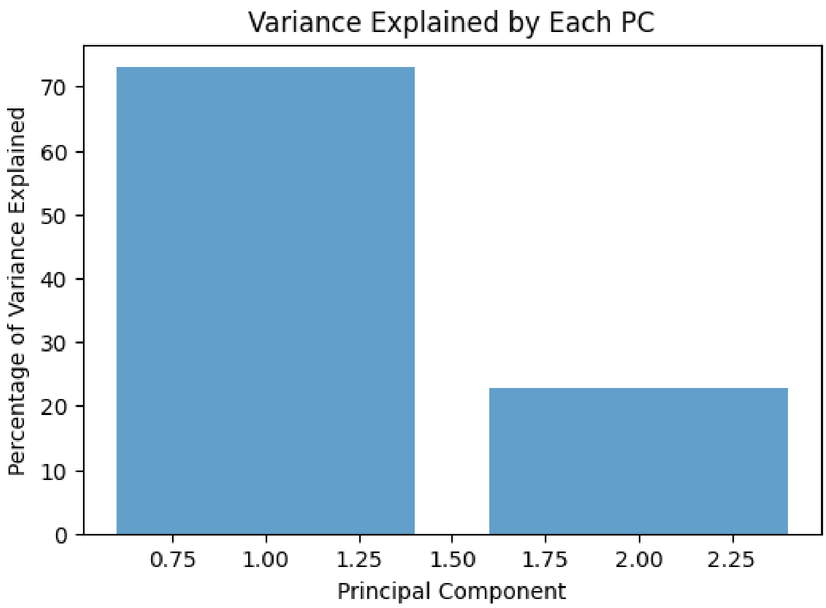

Figure 3: Explained variance by PCA components