Experiment No. 10
Case Study: Applications of above algorithms as a case study (E.g. Hand Writing Recognition using MNIST data set, classification using IRIS data set, etc)
Date of Performance:
Date of Submission:
Marks:
Sign:

Aim: Dimensionality reduction on iris dataset.

Theory:

Dimensionality reduction is a technique employed in machine learning and data analysis to mitigate the curse of dimensionality by reducing the number of features in a dataset while preserving its essential information. This process aids in simplifying the dataset's complexity, improving computational efficiency, and mitigating the risk of overfitting. By transforming high-dimensional data into a lower-dimensional space, dimensionality reduction techniques like Principal Component Analysis (PCA) help uncover underlying patterns, relationships, and structures within the data, facilitating more efficient analysis, visualization, and model training.

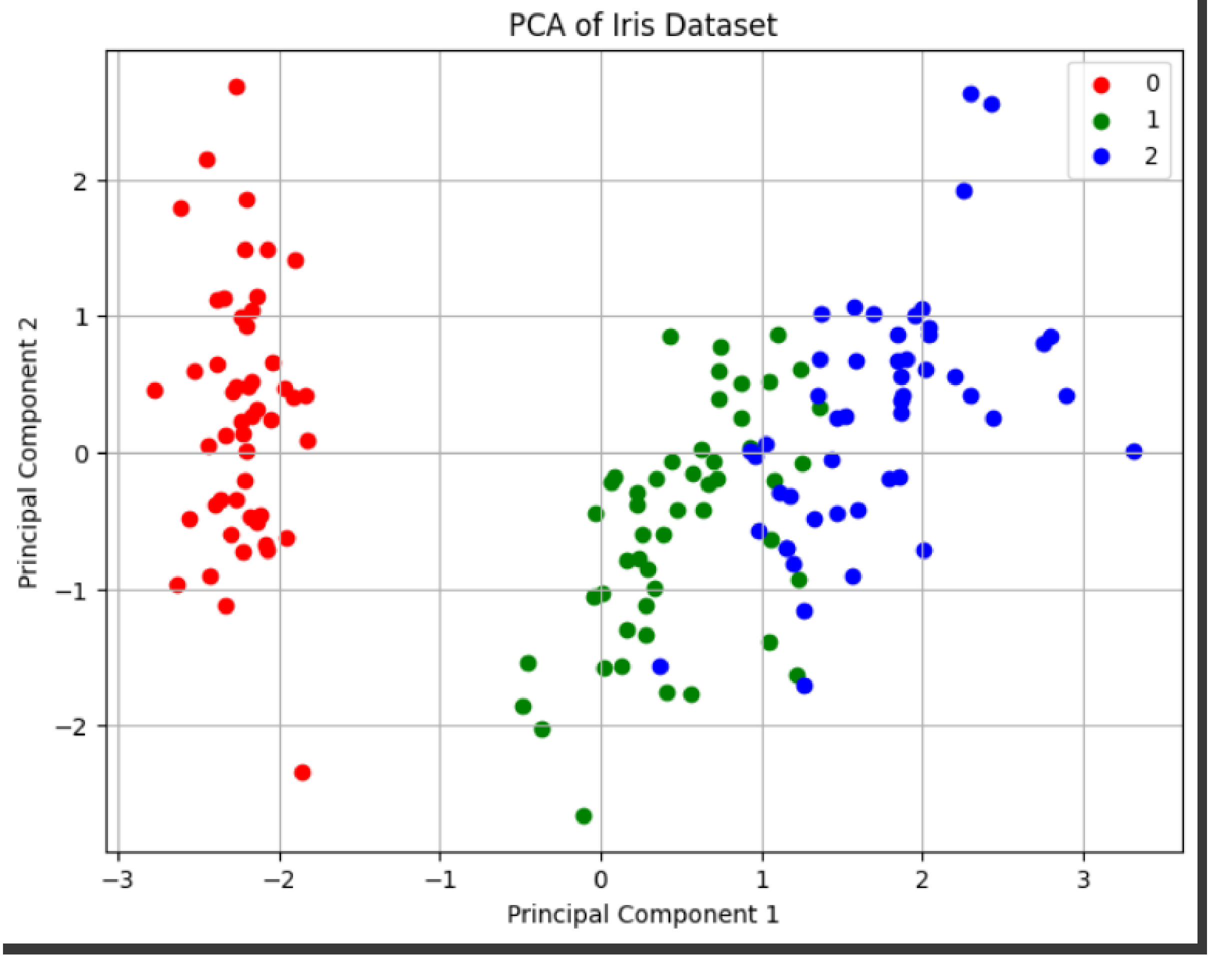
The Iris dataset is a classic example in the realm of machine learning and statistics, comprising 150 samples of iris flowers belonging to three different species: Setosa, Versicolor, and Virginica. Each sample is characterized by four features: sepal length, sepal width, petal length, and petal width. Despite its small size, the Iris dataset serves as a valuable benchmark for evaluating classification algorithms and dimensionality reduction techniques due to its simplicity, well-defined classes, and distinct feature characteristics.

In a recent study on the dimensionality reduction of the Iris dataset, Principal Component Analysis (PCA) was employed to reduce the dataset's four-dimensional feature space to a lower-dimensional representation while retaining most of the original variance. By extracting the principal components, which are orthogonal vectors representing the directions of maximum variance in the data, PCA transformed the dataset into a new coordinate system. This reduced feature space facilitated easier visualization, analysis, and modeling of the Iris dataset while maintaining the essential information necessary for accurate classification and pattern recognition tasks. Through dimensionality reduction, the study demonstrated the efficacy of PCA in simplifying complex datasets like Iris, enabling more efficient data analysis and model development processes.

Implementation:

```
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

# Load the Iris dataset

iris = load_iris()

X = iris.data

y = iris.target

# print(X)

# print(y)

# Standardize the features

scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)

# print(X_scaled)

# Perform PCA

pca = PCA(n_components=2)

X_pca = pca.fit_transform(X_scaled)
```

```
# print(X_pca)

# Create a DataFrame for visualization

df = pd.DataFrame(data=X_pca, columns=['PC1', 'PC2'])

df['Target'] = y

# print(df)

# Visualize the reduced dataset

plt.figure(figsize=(8, 6))

targets = np.unique(y)

colors = ['r', 'g', 'b']

for target, color in zip(targets, colors):

    indices_to_keep = df['Target'] == target

    plt.scatter(df.loc[indices_to_keep, 'PC1'], df.loc[indices_to_keep, 'PC2'], c=color,
label=target)

plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.title('PCA of Iris Dataset')

plt.legend(targets)

plt.grid(True)

plt.show()
```



Conclusion:

In conclusion, the implementation of dimensionality reduction using Principal Component Analysis (PCA) on the Iris dataset successfully transformed the original four-dimensional feature space into a lower-dimensional representation while preserving most of the dataset's variance. By standardizing the features and applying PCA, the dataset was visualized in a two-dimensional space, where the distinct clusters representing different iris species were clearly delineated. This reduced feature space facilitated easier visualization and analysis of the Iris dataset while maintaining essential information for accurate classification tasks. Through PCA, the study demonstrated the effectiveness of dimensionality reduction techniques in simplifying complex datasets like Iris, enabling more efficient data analysis, visualization, and model development processes.