Overview of Code and Implementation:

The project has been implemented using Python, utilizing the pandas library for working with datasets, matplotlib.pyplot for plotting graphs, and scikit-learn for implementing PCA and data standardization.

For each dataset, the following steps were performed. Firstly, the class of each dataset (y) was separated from the overall data (x), and then the other features were standardized. After standardizing the data, PCA with two components was applied to reduce each dataset to two dimensions and visualize it in a two-dimensional space. After applying PCA, the samples from each dataset were plotted on their respective graphs.

Data Standardization:

Since PCA varies with different scales, it is necessary to standardize the feature data of each dataset before applying PCA to ensure consistent scales. This standardization is only applied to columns that contain the values of the dataset's features.

In this project, the StandardScaler function from the scikit-learn library is used. StandardScaler is an important technique primarily used as a preprocessing step before many machine learning models to standardize the input dataset's performance range, working based on the z-scale. The resulting values from the z-scale method are often between -3 and 3. In this method, to calculate the new value, the current value is subtracted from the mean value and divided by the standard deviation. It generally works by centering and scaling the features, transforming them to have unit variance.

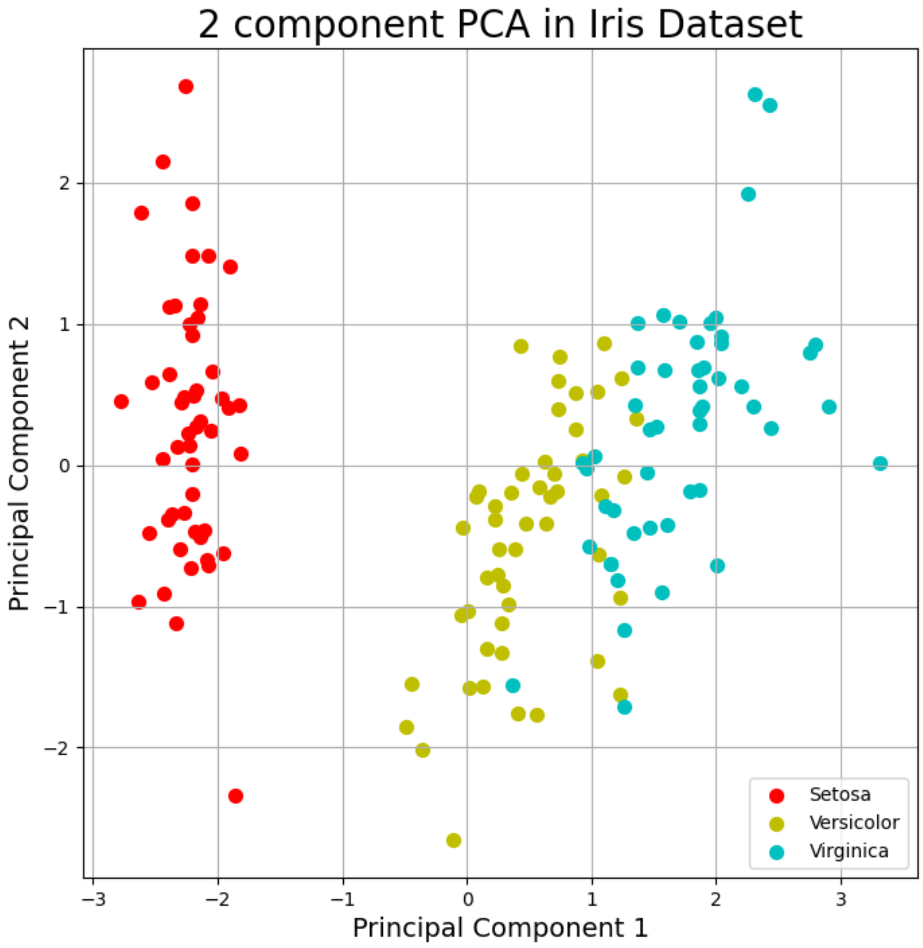
Conclusion:

By using explained variance, we can determine how much information (variance) can be attributed to each of the components selected by PCA. This value is important because when reducing the dimensionality of a dataset, we lose a significant amount of information. This value and the results obtained from applying PCA to each dataset are further examined.

Iris Dataset

The first dataset is the Iris Dataset, which contains 4 features. To better visualize and plot this dataset, PCA was used to reduce its 4 features to 2 and plot them in a two-dimensional space.

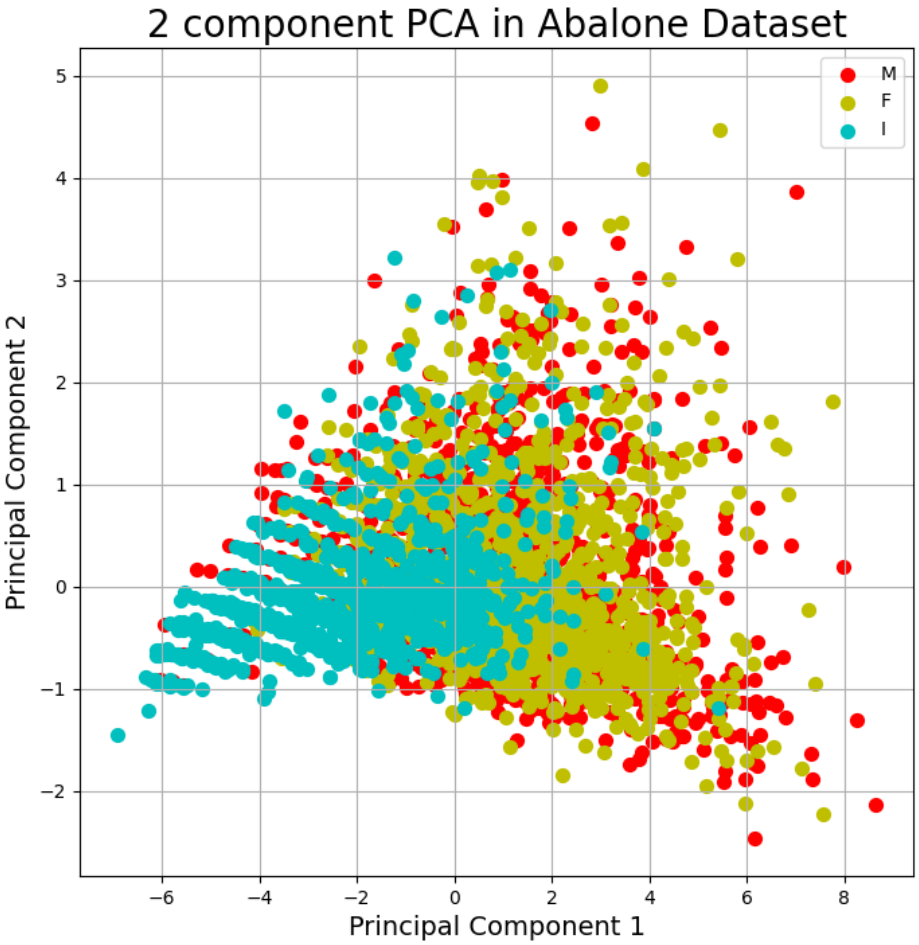
It can be observed that the first component captures approximately 72.96% of the information, and the second component captures approximately 22.85% of the information. Together, these two components represent 95.81% of the information, indicating that PCA has successfully preserved a significant amount of information. Based on the plot and the percentage of information, applying PCA to this dataset is suitable for classification purposes.



Abalone Dataset

The next dataset is the Abalone Dataset, which contains 8 features. To better visualize and plot this dataset, PCA was used to reduce its 8 features to 2 and plot them in a two-dimensional space.

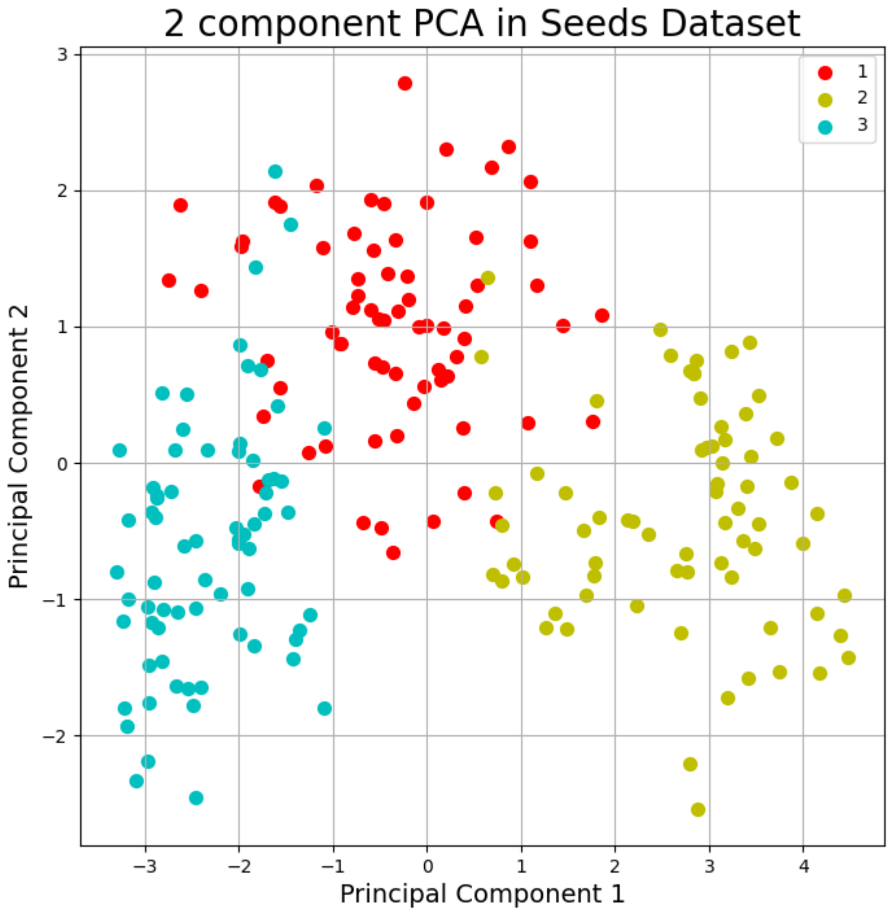
It can be observed that the first component captures approximately 83.90% of the information, and the second component captures approximately 8.69% of the information. Together, these two components represent 92.59% of the information. Despite having more features and data compared to the previous dataset, PCA has still preserved a considerable amount of information. Additionally, based on the plot, it appears that applying PCA to this dataset is suitable for classification purposes.



Seeds Dataset

The next dataset is the Seeds Dataset, which contains 7 features. To better visualize and plot this dataset, PCA was used to reduce its 7 features to 2 and plot them in a two-dimensional space.

It can be observed that the first component accounts for approximately 87.71% and the second component accounts for approximately 10.17% of the information, and these two components together represent 97.88% of the information, which is a higher amount of information loss compared to the previous two datasets. However, despite this, PCA has still managed to preserve a significant amount of information quite well. Additionally, based on the graph, it appears that applying PCA to this dataset would be suitable for classification purposes.


2 component PCA in Seeds Dataset
● 1
● 2
● 3
Principal Component 2
Principal Component 1