Principal Component Analysis on a dataset containing chemical analysis of wines

Saurav S Sankhe (ET21MTECH11003), Rahul Ghuge (ET21MTECH11002)

Energy Science and Technology, Indian Institute of Technology, Hyderabad

Abstract

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue/eigenvector problem, and the new variables are defined by the dataset at hand, not a priori, hence making PCA an adaptive data analysis technique. This article applies the PCA technique to a dataset containing chemical analysis of wines grown in a particular region in Italy but derived from three different cultivars (or classes). PCA has been implemented using two different libraries to check whether the output reproduces the same result. The dataset contained 13 dimensions which were reduced to two principal dimensions that represented the entire dataset with minimum losses.

Introduction

Principal Component Analysis (PCA) is basically a statistical procedure to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. Each of the principal components is chosen in such a way so that it would describe most of them still available variance and all these principal components are orthogonal to each other. In all principal components, the first principal component has a maximum variance.

PCA is a standard tool in modern data analysis - in diverse fields from neuroscience to computer graphics - because it is a simple, non-parametric method for extracting relevant information from confusing data sets. With minimal effort, PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it. [1]

Uses of PCA is used in various applications, some of them are as follows:

- It is used to find inter-relation between variables in the data. It is used to interpret and visualize data.
- The number of variables is decreasing which makes further analysis simpler.
- It's often used to visualize genetic distance and relatedness between populations.

These are basically performed on a square symmetric matrix. It can be a pure sum of squares and cross-products matrix or Covariance matrix or Correlation matrix. A correlation matrix is used if the individual variance differs much.

Objectives of PCA:

1. It is basically a non-dependent procedure in which it reduces attribute space from a large number of variables to a smaller number of factors.
2. PCA is basically a dimension reduction process but there is no guarantee that the dimension is interpretable.
3. The main task in this PCA is to select a subset of variables from a larger set, based on which original variables have the highest correlation with the principal amount.

Basics of PCA

Understanding the details of PCA requires knowledge of linear algebra. This section will explain only the basics with a simple graphical representation of the data. [4]

In Plot 1A below, the data are represented in the X-Y coordinate system. The dimension reduction is achieved by identifying the principal directions, called principal components, in which the data varies.

PCA assumes that the directions with the largest variances are the most "important" (i.e., the most principal).

In the figure below, the PC1 axis is the first principal direction along which the samples show the largest variation. The PC2 axis is the second most important direction, and it is orthogonal to the PC1 axis.

The dimensionality of our two-dimensional data can be reduced to a single dimension by projecting each sample onto the first principal component (Plot 1B)

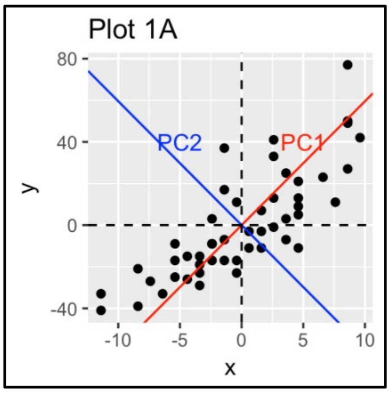

Figure 1: Plot 1A

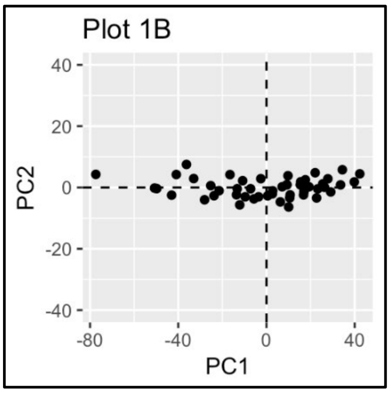

Figure 2: Plot 1B

Note that, the PCA method is particularly useful when the variables within the data set are highly correlated. Correlation indicates that there is redundancy in the data. Due to this redundancy, PCA can be used to reduce the original variables into a smaller number of new variables (principal components) explaining most of the variance in the original variables.

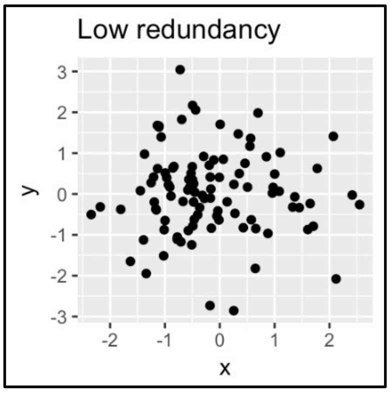

Figure 3: Low Redundancy data

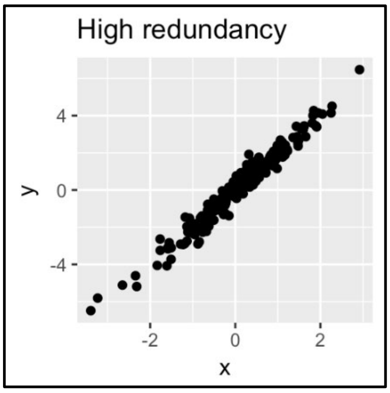

Figure 4: High Redundancy data

Taken together, the main purpose of the principal component analysis is to:

- identify a hidden pattern in a data set,
- reduce the dimensionality of the data by removing the noise and redundancy in the data,
- identify correlated variables.

Data standardization

In principal component analysis, variables are often scaled (i.e., standardized). This is particularly recommended when variables are measured in different scales (e.g.: kilograms, kilometers, centimeters); otherwise, the PCA outputs obtained will be severely affected.

The goal is to make the variables comparable. Generally, variables are scaled to have a standard deviation of one and a mean equal to zero.

The standardization of data is an approach widely used in the context of gene expression data analysis before PCA and clustering analysis. We might also want to scale the data when the mean and/or the standard deviation of variables are largely different.

When scaling variables, the data can be transformed as follow:

$$z = \frac{x - \bar{x}}{\sigma}$$

where:

$\bar{x} = mean$

$\sigma = standard\ deviation$

Methodology

Dataset was downloaded from the UCI Machine learning repository This data contains the results of a chemical analysis of wines grown in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

A] *Wine Recognition-Dataset Description*:

This data was used in many papers for comparing various classifiers (http://archive.ics.uci.edu/ml/datasets/Wine). In a classification context, it is a well-posed problem with somewhat "well-behaved" class structures. The dataset is the result of chemical analysis (see paper Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification, and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.) of wines grown in a particular region in Italy but derived from three different cultivars (or classes). The analysis determined the quantities of 13 constituents found (i.e., 13-D feature space) in each of the three types of wines. The attributes are:

1) Alcohol content
2) Malic acid
3) Ash
4) Alcalinity of ash (Ash alcalinity)
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline

The number of Instances for each class is Class 1: 59, Class 2: 71, and Class 3: 48, i.e., total of 178 samples. Note that the first number in each string of 14 numbers in the data file “Wine” represents the class label.

The dataset was imported in python using read_csv() command and a summary of the dataset was calculated using dataset.info() and dataset.describe() command available in pandas library.

B] *Visualization and Interpretation:*

Using the Seaborn library, the dataset was initially visualized to understand the variability within the datasets. sns.pairwiseplot() gave a better visualization for all the 13 dimensions as shown in Figure 5

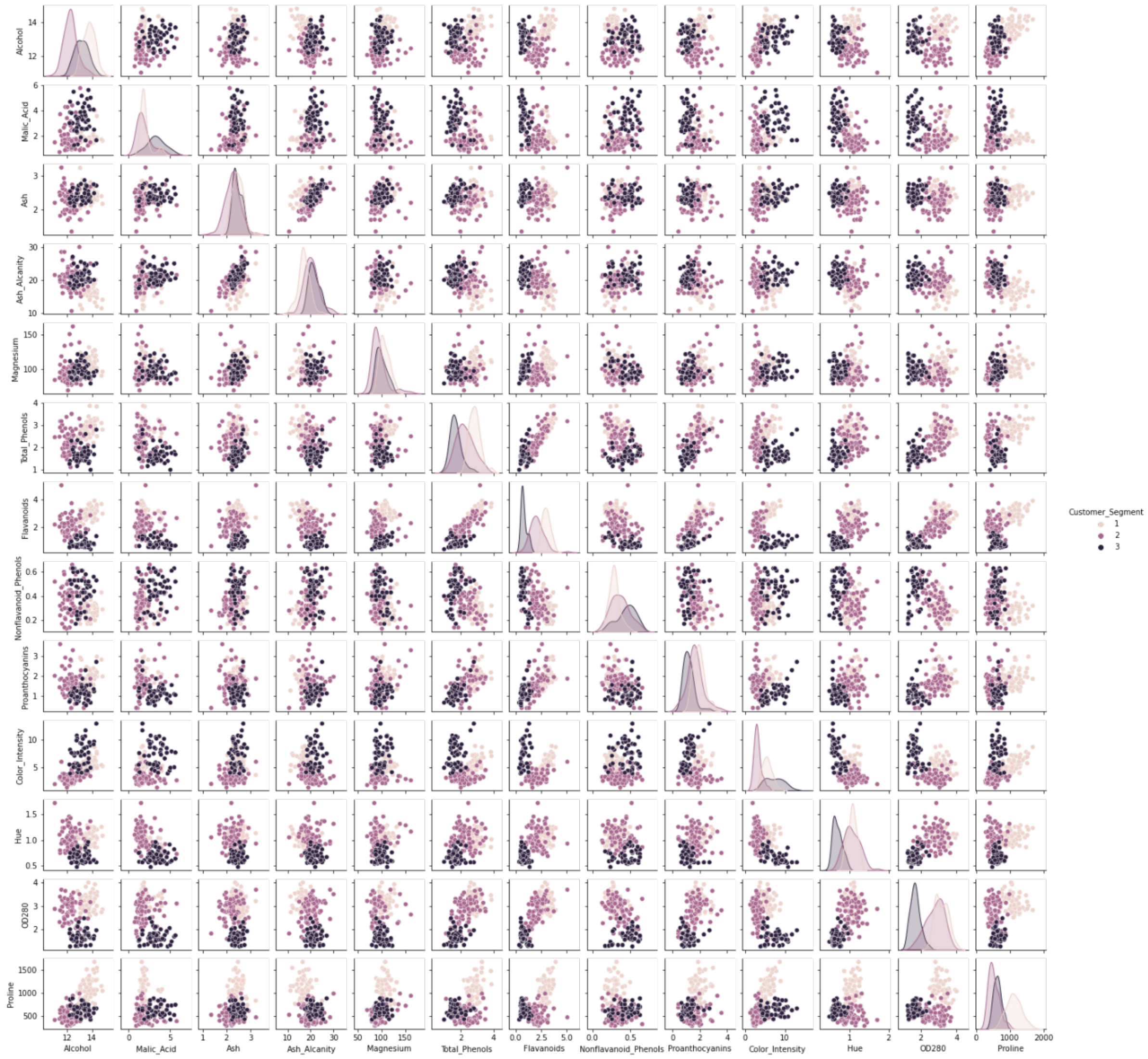

Figure 5: Pairwise plot for the dataset

Before calculating the covariance matrix, the dataset was standardized. After standardizing the dataset, the covariance matrix was calculated and plotted using sns.heatmap function available in the seaborn library. (See figure 6)

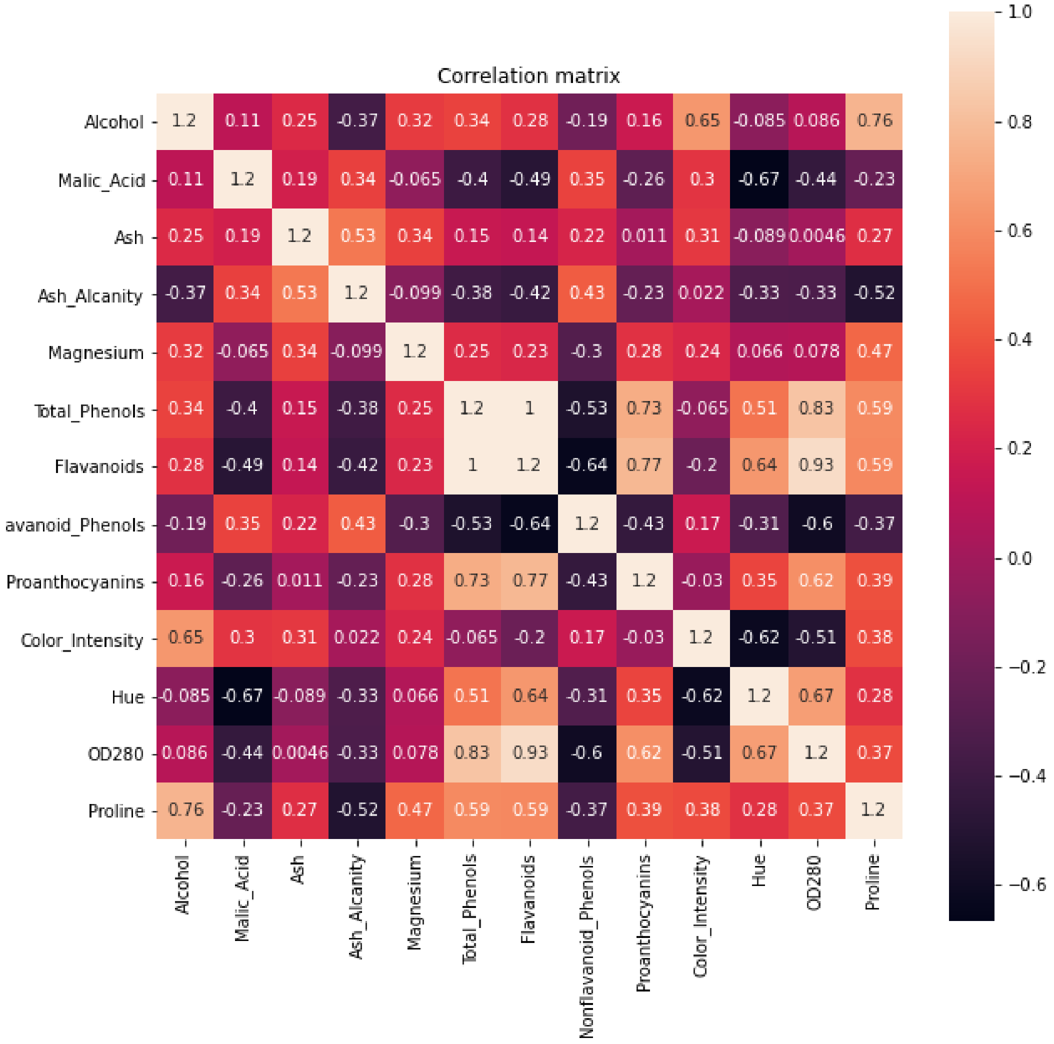

Figure 6: Correlation matrix

C] *Eigenvalues / Variances*

As described in previous sections, the eigenvalues measure the amount of variation retained by each principal component. Eigenvalues are large for the first PCs and small for the subsequent PCs. That is, the first PCs correspond to the directions with the maximum amount of variation in the data set.

We examine the eigenvalues to determine the number of principal components to be considered. The eigenvalues and the proportion of variances (i.e., information) retained by the principal components (PCs) can be extracted using the function (np.linalg.svd and np.linalg.eig).

Eigenvalues can be used to determine the number of principal components to retain after PCA

- An eigenvalue > 1 indicates that PCs account for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cut-off point for which PCs are retained. This holds true only when the data are standardized.
- We can also limit the number of the component to that number that accounts for a certain fraction of the total variance. For example, if we are satisfied with 70% of the total variance explained then use the number of components to achieve that.

Unfortunately, there is no well-accepted objective way to decide how many principal components are enough. This will depend on the specific field of application and the specific data set. In practice, we tend to look at the first few principal components to find interesting patterns in the data.

The scree plot can be produced (Figure 7) showed that the first two dimensions would be enough to represent the data.

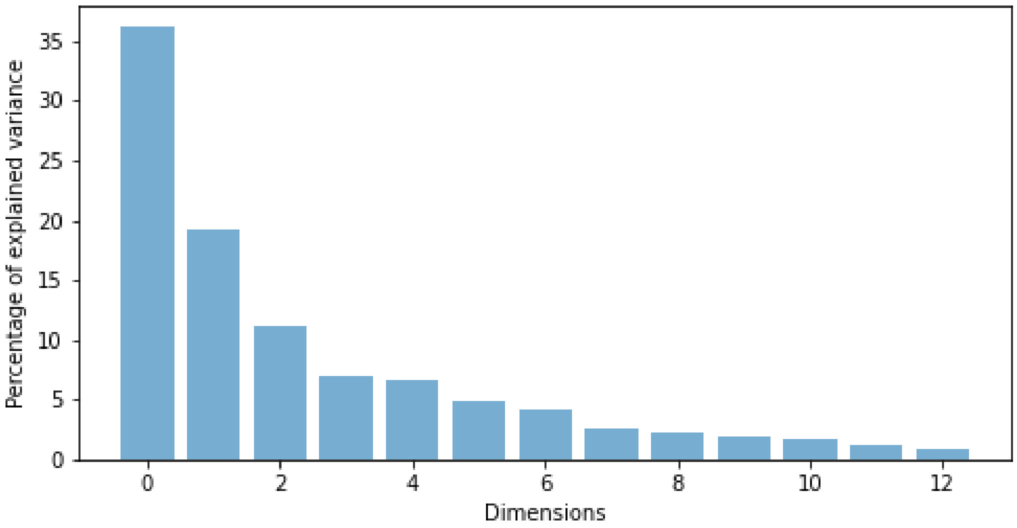

Figure 7: Bar plot between the percentage of explained variance with the number of principal components.

The first two principal components were then calculated using dot product between the eigenvectors and the standardized dataset X and a scatterplot was plotted with the numbering of all the data points. (See figure 8).

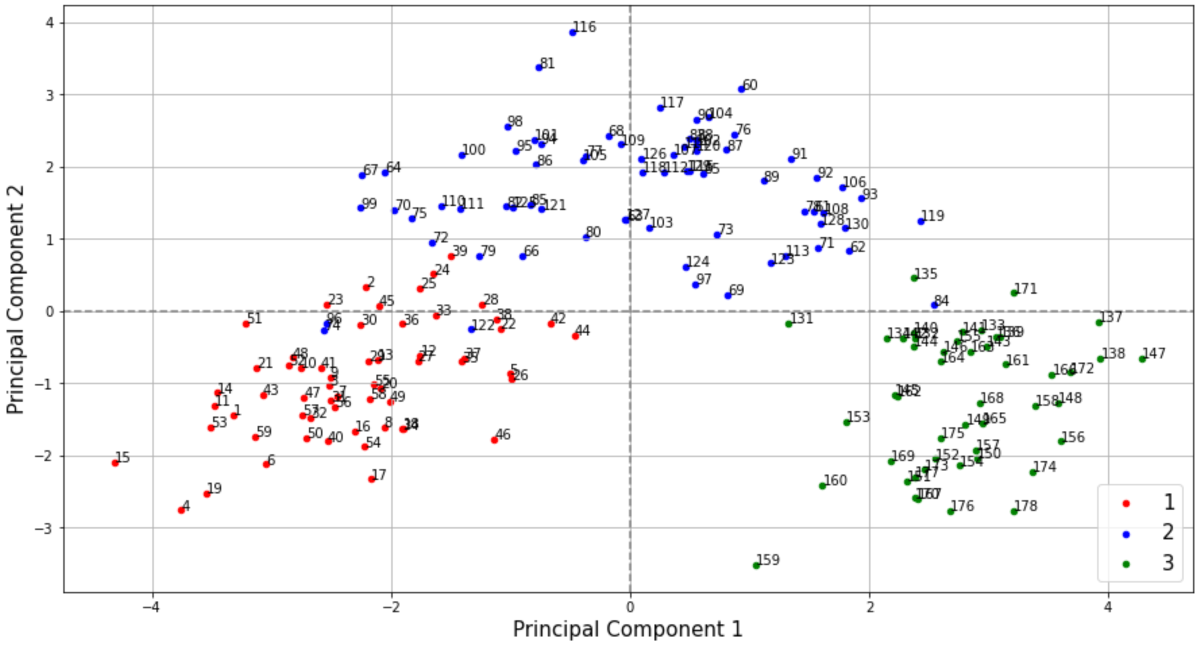

Figure 8: Plot of two principal components (without sklearn library)

D] *Correlation circle*

The correlation between a variable and a principal component (PC) is used as the coordinates of the variable on the PC. The representation of variables differs from the plot of the observations: The observations are represented by their projections, but the variables are represented by their correlations.

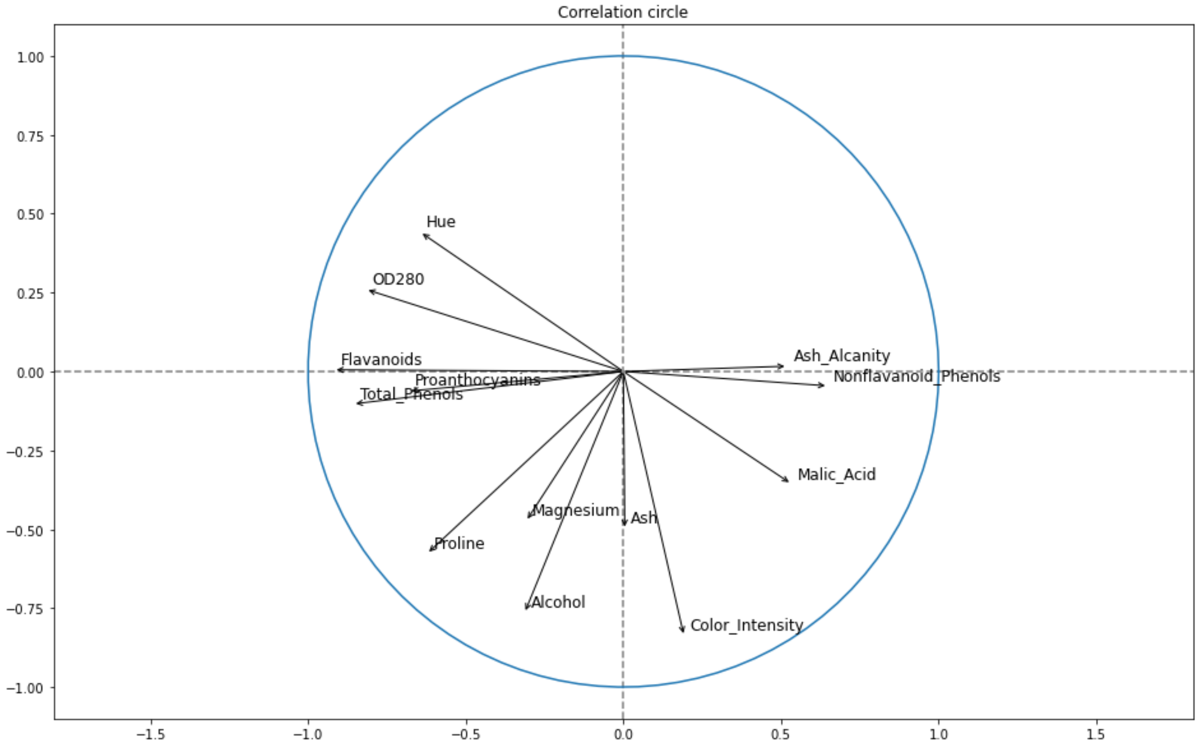

Figure 9: Correlation circle

The plot above is also known as the variable correlation plot. It shows the relationships between all variables. It can be interpreted as follow:

- Positively correlated variables are grouped together.
- Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants).
- The distance between variables and the origin measures the quality of the variables on the factor map. Variables that are away from the origin are well represented on the factor map.

In the dataset, Flavonoids & Ash Alcanity were the main things that have maximum variance in the entire dataset and were important principal factors/dimensions in the dataset.

E] *Verification using SK Learn Library*

The same dataset was used and PCA was implemented using SK learn library and the plots were compared and visualized.

The plot resembled the exact mirror copy of the previous version. The reason behind the mirror image of the plot is the method of calculation of the principal components.

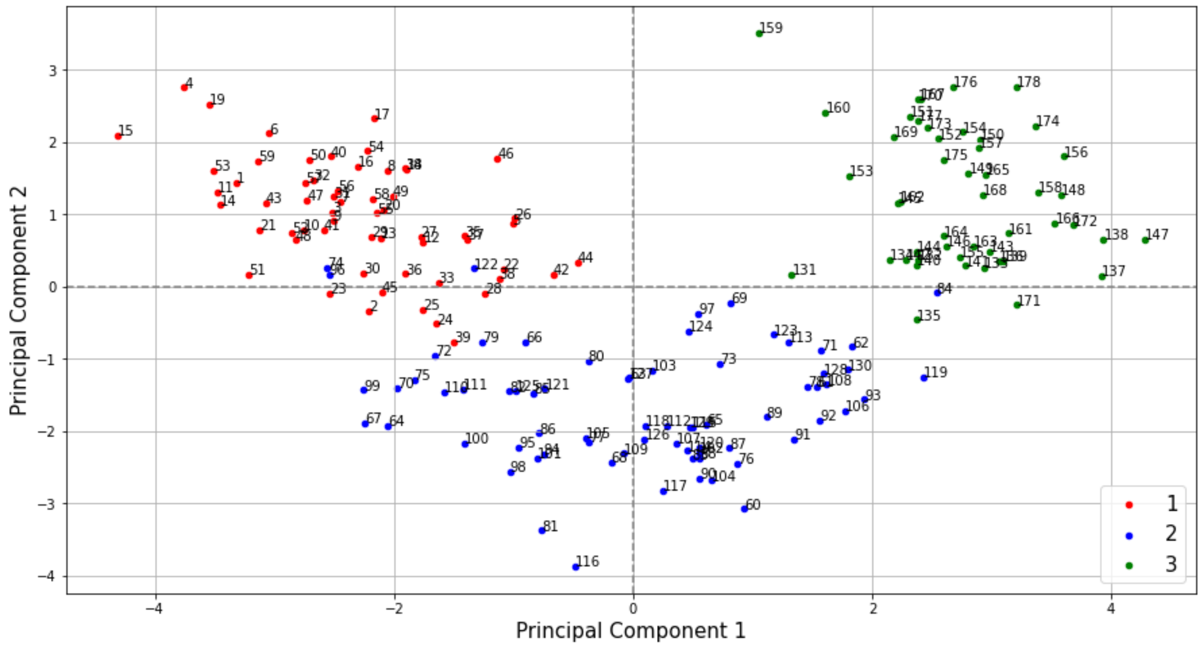

Figure 10: Plot of two principal components (with sklearn library)

Conclusion

Principal component analysis was successfully implemented on the dataset. The first two principal components (Flavonoids & Ash Alcanity) were able to represent the data perfectly with the good grouping of the classes. This dimensionality reduction technique can be used as a preliminary technique to further problems in classification and clustering models.

References

1] A Tutorial on Principal Component Analysis
2] Wine recognition data
3] https://www.kaggle.com/karimsaieh/pca-principal-component-analysis-without-sklearn
4] Principal Component Methods in R: Practical Guide

Appendix: Python Code

GitHub Link: https://github.com/SSSANKHE/PH6130-Project