LABWORK 1

PRINCIPAL COMPONENT ANALYSIS

Mai Hải Đăng - BI12-076

01.12.2023

Machine Learning and Data Mining II

Table of Contents

1. Introduction

- This report is dedicated to the understand the basic knowledge about Data Mining
- Discover structure inside unstructured data; extract meaning from noisy data; understand trends, patterns, correlations.
- To visualize data distribution in 2D
- Answering questions such as:
    - Determine which feature is discrete or continuous? Is it qualitative or quantitative?
    - What is the mean, variance, covariance, correlation of the selected datasets.
    - What is the most correlated couple of features of each dataset?
- Apply PCA on the selected datasets to see:
    - How much of the total variation in the data is explained by the first two principal components?
    - How well are the individual classes separated in the case we use two principal components?
    - Obtained results when increasing the number of principal components.
- The original dataset are Wine Quality and Dry Bean

2. Concepts

2.1. Discrete vs. Continuous Data

Discrete Data	Continuous Data
Take on distinct, separate values	Take on an infinite number of values within a specified range
Countable (can count the number of different values it can take)	Measurable
★ Number of students in a classroom ★ Number of cars in a parking lot. ★ Number of books on a shelf.	★ Height of individuals. ★ Weight of objects. ★ Temperature of a room.

2.2. Qualitative or quantitative Data

Qualitative	Quantitative
Descriptive in nature and deals with qualities or characteristics.	Measurable quantities and is expressed in numerical terms.

It is non-numerical and is often expressed in words. Relies on the interpretation of meanings, feelings, or attributes.	Can be continuous (measured on a scale) or discrete (countable)
★ **Categories:** Qualitative data is often categorical and falls into categories. ★ **Nominal or Ordinal:** It can be nominal (categories with no inherent order) or ordinal (categories with a specific order).	★ **Continuous or Discrete:** Quantitative data can be continuous (measured on a scale) or discrete (countable). ★ **Ratio or Interval:** It can be ratio (with a true zero point) or interval (without a true zero).

2.3. Statistical Measurements

2.3.1. Mean (Average)

- The mean is a measure of central tendency and represents the average value of a set of numerical values
- Formula:

$$\boxed{Mean = \frac{Sum\ of\ values}{Number\ of\ values}}$$

2.3.2. Variance

- Variance measures the spread or dispersion of a set of values from the mean. It gives an indication of how much individual data points differ from the average.
- Formula:

$$\boxed{Variance = \frac{Sum\ of\ squared\ differences\ from\ the\ mean}{Number\ of\ values}}$$

2.3.3. Covariance

- Covariance measures how two variables change together. A positive covariance indicates a positive relationship, while a negative covariance indicates a negative relationship.
- Formula:

$$\boxed{Cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{2}}$$

2.3.4. Correlation

- Correlation is a standardized measure of the strength and direction of a linear relationship between two variables. It is a unitless measure that ranges from -1 to 1.
- Formula:

$$Correlation = \frac{Cov(X, Y)}{STD(X) * STD(Y)}$$

2.4. PCA (Principal Component Analysis)

- The PCA objective is to project the data onto a lower dimensional linear space such that the variance of the projected data is maximized.
- **Standardize the Data:**
    - If the features in the dataset are measured on different scales, it's common to standardize the data (subtract mean and divide by standard deviation) to ensure that all features have the same influence on the analysis.
- **Compute the Covariance Matrix:**
    - Calculate the covariance matrix of the standardized data. The covariance matrix provides information about the relationships between different features.
- **Compute Eigenvectors and Eigenvalues:**
    - Compute the eigenvectors and eigenvalues of the covariance matrix. Eigenvectors represent the directions of maximum variance, and eigenvalues indicate the magnitude of variance in those directions.
- **Sort Eigenvectors by Eigenvalues:**
    - Sort the eigenvectors in descending order based on their corresponding eigenvalues. The eigenvector with the highest eigenvalue corresponds to the principal component with the most significant variance.
- **Select Principal Components:**
    - Choose the top k eigenvectors to form the basis of the new subspace, where kk is the desired dimensionality of the reduced space.
- **Project Data onto the New Subspace:**
    - Multiply the original data matrix by the selected eigenvectors to obtain the new lower-dimensional representation of the data.

3. Data Analysis

3.1. Wine Quality

Table 1. General information Wine Quality Dataset.

Data	Role	Data Type	Non-null	Description
Fixed acidity	Feature	Continuous, Quantitative	True	Physicochemical

Volatile acidity	Feature	Continuous, Quantitative	True	Physicochemical
Citric acid	Feature	Continuous, Quantitative	True	Physicochemical
Residual sugar	Feature	Continuous, Quantitative	True	Physicochemical
Chlorides	Feature	Continuous, Quantitative	True	Physicochemical
Free sulfur dioxide	Feature	Continuous, Quantitative	True	Physicochemical
Total sulfur dioxide	Feature	Continuous, Quantitative	True	Physicochemical
Density	Feature	Continuous, Quantitative	True	Physicochemical
pH	Feature	Continuous, Quantitative	True	Physicochemical
Sulfates	Feature	Continuous, Quantitative	True	Physicochemical
Alcohol	Feature	Continuous, Quantitative	True	Physicochemical
Quality	Target	Discrete (1-10), Quantitative	True	Sensory assessors

Table 2. First eight samples from the Wine Quality Dataset (total: 6497).

	Fixed acidity	Volatile acidity	Citric acid	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	pH	Sulfates	Alcohol	Quality
0	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
4	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
5	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
6	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7

Table 3. Statistical Measurements of Wine Quality Dataset.

Data	Mean	Standard Deviation	Variance
Fixed acidity	7.215307	1.296434	1.680740
Volatile acidity	0.339666	0.164636	0.027105
Citric acid	0.318633	0.145318	0.021117
Residual sugar	5.443235	4.757804	22.636696
Chlorides	0.056034	0.035034	0.001227
Free sulfur dioxide	30.525319	17.749400	315.041192
Total sulfur dioxide	115.744574	56.521855	3194.720039
Density	0.994697	0.002999	0.000009
pH	3.218501	0.160787	0.025853
Sulfates	0.531268	0.148806	0.022143
Alcohol	10.491801	1.192712	1.422561
Quality	5.818378	0.873255	0.762575

Table 4. Total variation in the wine quality dataset explained by four principal components.

PC	Variance Explained (%)	Sum
PC1	9.53758252e-01	9.94386006e-01
PC2	4.06277547e-02	
PC3	4.06277547e-02	
PC4	4.63879237e-04	

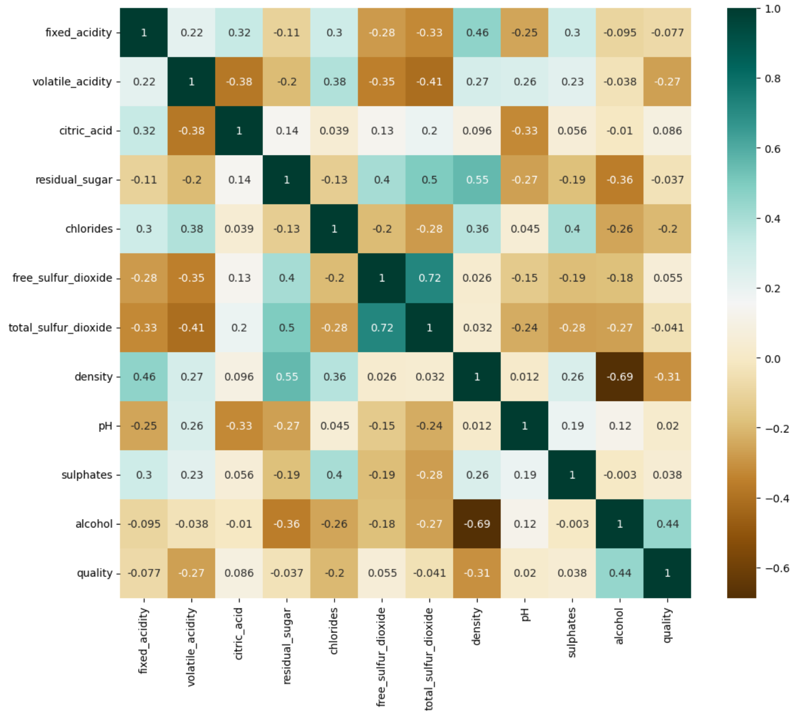

Figure 1. Wine quality database correlation matrix.

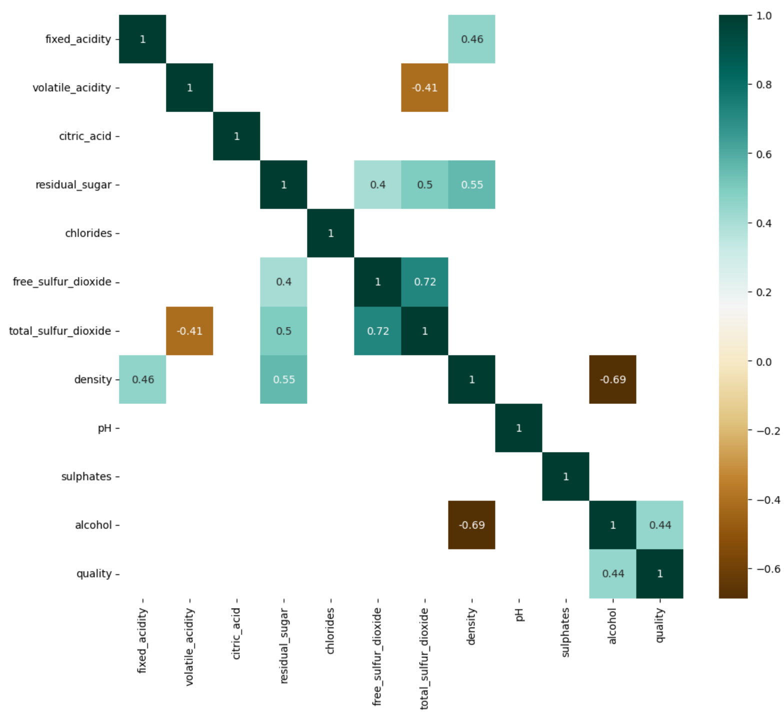

Figure 2. Wine quality database correlation matrix with corr > 0.4, corr < -0.4. The most correlated couple is: Free sulfur dioxide & total sulfur dioxide (0.72).

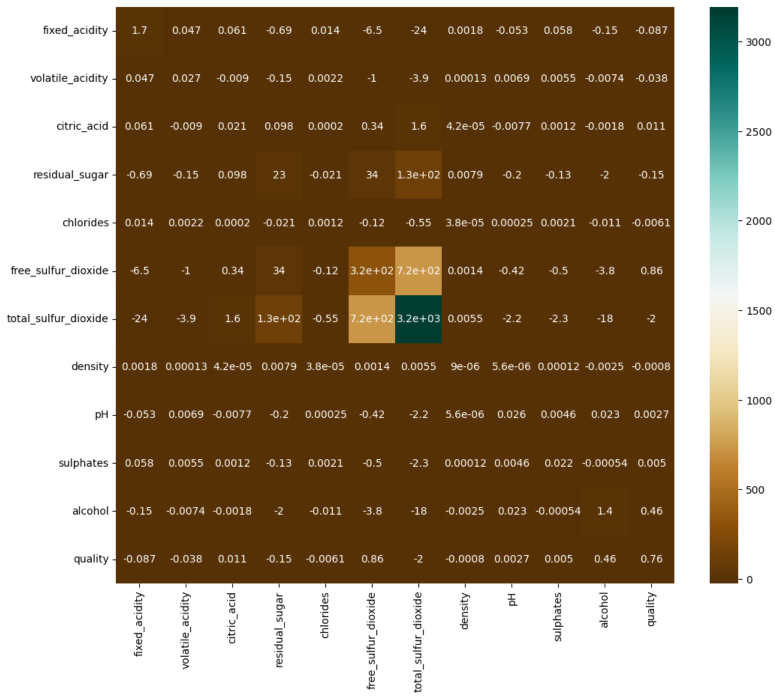

Figure 3. Wine quality database covariance matrix.

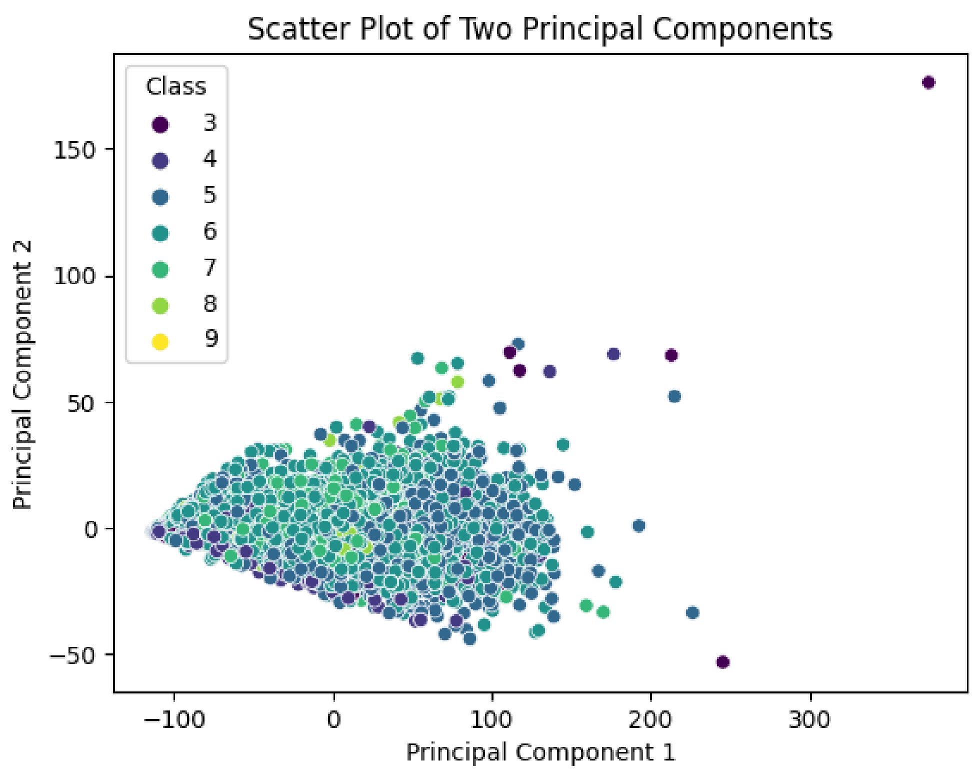

Figure 4. Wine quality database PCA. Individual classes separated in the case we use two principal components are not well distinguished from each other.

3.2. Dry Bean

- Please refer to the Dry Bean variables table for further data descriptions.
- Total number of instances: 13611.

Table 4. General information Dry Bean Dataset.

Data	Role	Data Type	Non-null
Area	Feature	Discrete, Quantitative	True
Perimeter	Feature	Continuous, Quantitative	True
MajorAxisLength	Feature	Continuous, Quantitative	True
MinorAxisLength	Feature	Continuous, Quantitative	True
AspectRatio	Feature	Continuous, Quantitative	True
Eccentricity	Feature	Continuous, Quantitative	True
ConvexArea	Feature	Discrete, Quantitative	True
EquivDiameter	Feature	Continuous, Quantitative	True
Extent	Feature	Continuous, Quantitative	True
Solidity	Feature	Continuous, Quantitative	True
Roundness	Feature	Continuous, Quantitative	True
Compactness	Feature	Continuous, Quantitative	True
ShapeFactor1	Feature	Continuous, Quantitative	True
ShapeFactor2	Feature	Continuous, Quantitative	True
ShapeFactor3	Feature	Continuous, Quantitative	True
ShapeFactor4	Feature	Continuous, Quantitative	True
Class	Target	Categorical, Qualitative	True

Table 5. Statistical Measurements Dry Bean Dataset.

Data	Mean	Standard Deviation	Variance
Area	53048.284549	29324.095717	8.599026e+08
Perimeter	855.283459	214.289696	4.592007e+04
MajorAxisLength	320.141867	85.694186	7.343494e+03
MinorAxisLength	202.270714	44.970091	2.022309e+03
AspectRatio	1.583242	0.246678	6.085026e-02
Eccentricity	0.750895	0.092002	8.464324e-03
ConvexArea	53768.200206	29774.915817	8.865456e+08
EquivDiameter	253.064220	59.177120	3.501932e+03
Extent	0.749733	0.049086	2.409471e-03
Solidity	0.987143	0.004660	2.171913e-05
Roundness	0.873282	0.059520	3.542617e-03
Compactness	0.799864	0.061713	3.808552e-03
ShapeFactor1	0.006564	0.001128	1.272380e-06
ShapeFactor2	0.001716	0.000596	3.550668e-07
ShapeFactor3	0.643590	0.098996	9.800238e-03
ShapeFactor4	0.995063	0.004366	1.906595e-05

Table 4. Total variation in the dry bean dataset explained by four principal components.

PC	Variance Explained (%)	Sum
PC1	9.99967207e-01	9.999978243e-01
PC2	3.06176794e-05	
PC3	1.92111562e-06	
PC4	2.29430254e-07	

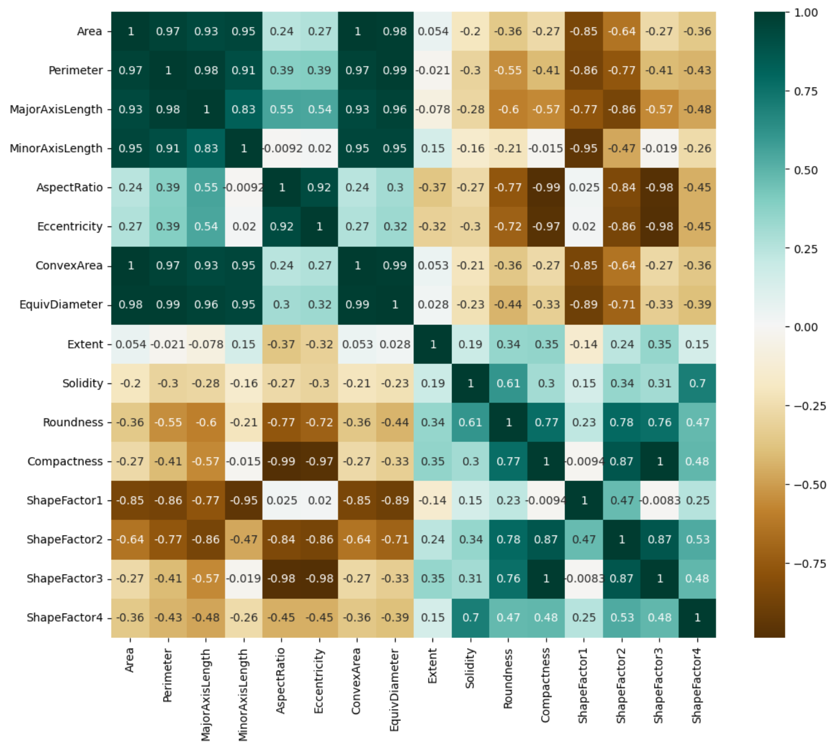

Figure 6. Dry Bean database correlation matrix

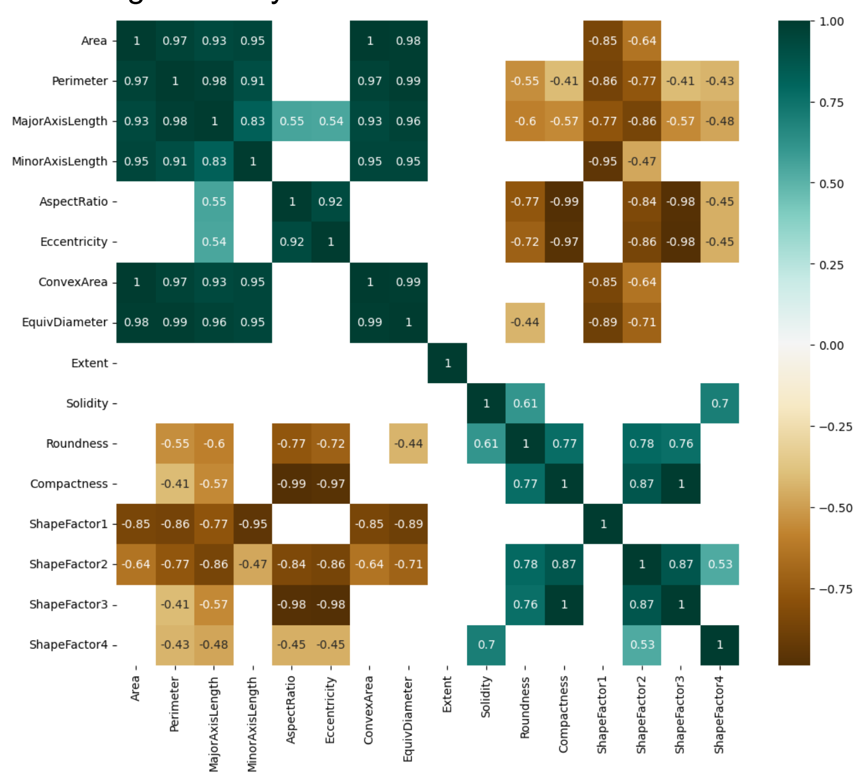

Figure 7. Dry Bean database correlation matrix with corr > 0.4, corr < -0.4.
The most correlated couples are: EquivDiameter & Area (0.98);
EquivDiameter & Perimeter (0.99); EquivDiameter & ConvexArea (0.99).

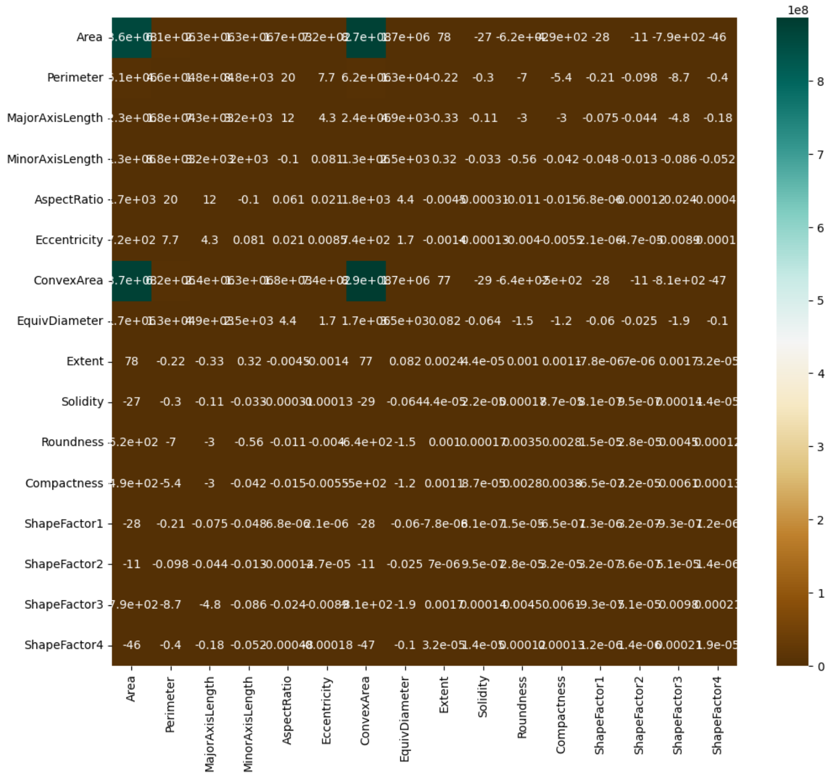

Figure 8. Dry Bean database covariance matrix.

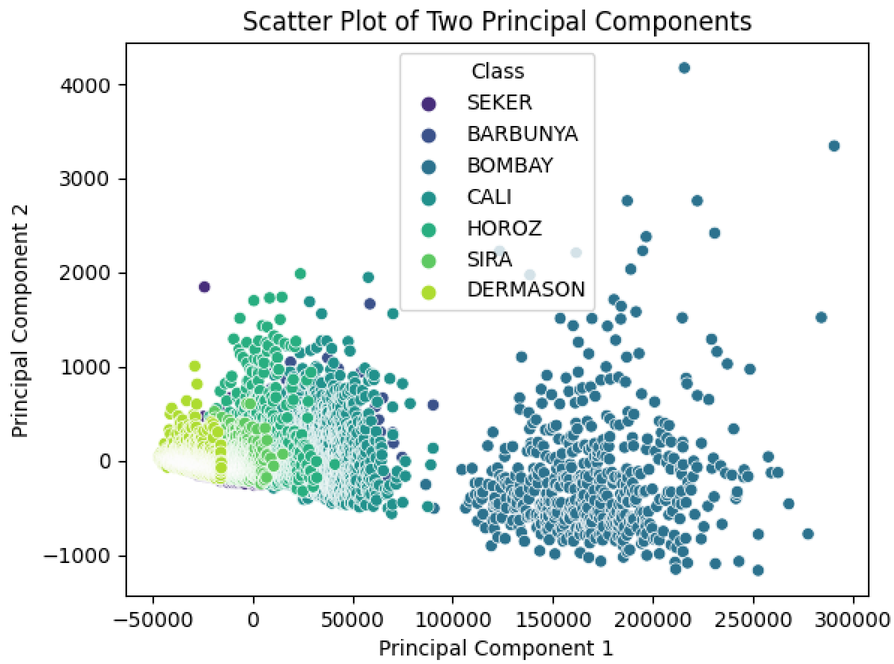

Figure 9. Dry Bean database PCA. Individual classes separated in the case we use two principal components. We can see a better distinction between BOMBAY especially, but the rest are a bit difficult to make out.

References

- Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.
- Dry Bean Dataset. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C50S4B.
- Cortez, Paulo & Teixeira, Juliana & Cerdeira, António & Almeida, Fernando & Matos, Telmo & Reis, José. (2009). Using Data Mining for Wine Quality Assessment. 66-79. 10.1007/978-3-642-04747-3_8.