Project 1

Introduction to Machine Learning and Data Mining

Haotian Liu 212645
Yi Zhan 222504
Diego Rodríguez Gordo 222883

October 3, 2022

Contents

0.1 Contributions

Section	Responsability
What the data is about?	Diego Rodríguez Gordo 222883
Previous work	Yi Zhan 222504
Explanation of prediction and classification	Haotian Liu 212645
Description variables	Yi Zhan 222504
Basic summary statistics	Yi Zhan 222504
Plots variables	Diego Rodríguez Gordo 222883
Outliers	Haotian Liu 212645
Normality tests	Diego Rodríguez Gordo 222883
Correlation	Diego Rodríguez Gordo 222883
Feasibility of ML models	Diego Rodríguez Gordo 222883
PCA	Haotian Liu 212645
Conclusion	Yi Zhan 222504
Exercise 1, 2	Haotian Liu 212645
Exercise 3, 4	Diego Rodríguez Gordo 222883
Exercise 5, 6	Yi Zhan 222504

Chapter 1

Description of the data set

1.1 What is the data about?

The data was obtained from the UCI Machine Learning Repository. There is a wide range of diversity of dry beans and one of the most important aspects to provide the principles of sustainable agricultural systems is their quality. Therefore, is essential to obtain the standard features of the different kinds of dry beans.

The data set is obtained with this purpose using different computer vision techniques. Once the numerical features are extracted, there has to be different analysis in order to be able to classify each dry bean into its class.

1.2 Previous work

A group of Turkish researchers in [1] have used computer vision and machine learning techniques to take photos of dry beans and classify these dry beans into seven types by their features of the form, shape, type and structure, in order to divide these beans automatically and efficiently, which can help farmers cultivate them and increase their market value.

Actually, there have been a few researches on applying computer vision systems to classify the dry bean seeds in the last decades. The group of Turkish researchers went through their classification methods and summarized their accuracy. On the basis of previous studies, researchers decided to create Multilayer Perceptron, Support Vector Machine, k-Nearest Neighbors, Decision Tree classification models to determine the types of beans and provide parameters of them, comparing the performance of these models at the same time.

First of all, they built a computer vision system which contains a camera, an illumination box and a lens mount to capture images of 13,611 dry bean samples and then preprocessing and feature extraction with Matlab was needed to obtain the 12 dimensional and 4 shape features of the beans, including the area, perimeter, aspect ratio, etc., whose values are all in pixel count. In this way, the dataset of dry

beans was set up.

Moreover, 10-fold cross validation was used to divide the dataset into test set and training set. MP, SVM, kNN and DT algorithm, which are most frequently used, were respectively applied to creating classification models and confusion matrix was introduced to determine whether the model is good enough. The accuracy, error rate, precision, specificity, recall, F1-score classification performance metrics were calculated by using the confusion matrix of each model.

Analyzing the result, they found that all models except DT have classification success with the rate of over 90 percent and the SVM classifier works best. As a result, the classifier can work effectively on dry beans from different regions. However, it's a bit more difficult to distinguish Sira beans from Dermason beans, since their flatness and roundness features are similar, and the rate of success would rise if the dimension of the suture axis of the beans or the coefficient of variance in the shape and size variables of each cultivar were included. And they were also looking forward to turning the classifier into a mobile application and building an image bank of dry beans.

1.3 Predictions

1.3.1 Classification

Since the purpose of this study is to distinguish seven different varieties of dry beans with similar features, we decided to apply classification on seven classes of dry beans based on other 16 attributes. However, we filter attributes based on their correlations. For attributes with strong positive correlation, we will choose one of the attributes to use.

1.3.2 Regression

For regression prediction task, we decided to choose Aspect Ration to predict. For some attributes are calculated using Aspect Ration, it is important to avoid using related attributes to prevent pointless predictions. For example, Aspect Ration is calculated using Max Axis Length and Minor Axis Length. If Max Axis Length and Minor Axis Length are used to predict Aspect Ration, then the value of Aspect Ration can be calculated directly by dividing Max Axis Length by Minor Axis Length, and the prediction will be pointless. Due to this reason, attributes of Minor Axis Length, Minor Axis Length, Compactness and Shape Factor 1-4 should be ignored. Aspect Ration can be predicted in the regression based on Area, Perimeter, Eccentricity, Convex Area, Equiv Diameter, Extent, Solidity and Roundness.

Chapter 2

Attributes of the data

2.1 Descriptions of the attributes

1. **Area** (A) (discrete, ratio): The area of a bean zone and the number of pixels within its boundaries.
2. **Perimeter** (P) (continuous, ratio): Bean circumference is defined as the length of its border.
3. **Major axis length** (L) (continuous, interval): The distance between the ends of the longest line that can be drawn from a bean.
4. **Minor axis length** (l) (continuous, interval): The longest line that can be drawn from the bean while standing perpendicular to the main axis.
5. **Aspect ratio** (K) (continuous, ratio): Defines the relationship between L and l.

$$K = \frac{L}{l}$$

6. **Eccentricity** (Ec) (continuous, ratio): Eccentricity of the ellipse having the same moments as the region.
7. **Convex area** (C) (discrete, ratio): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
8. **Equivalent diameter** (Ed) (continuous, ratio): The diameter of a circle having the same area as a bean seed area.

$$E_d = \sqrt{\frac{4A}{\pi}}$$

9. **Extent** (Ex) (continuous, ratio): The ratio of the pixels in the bounding box to the bean area.

$$Ex = \frac{A}{A_B}$$

Where A_B = Area of bounding rectangle.

10. **Solidity** (S) (continuous, ratio): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.

$$S = \frac{A}{C}$$

11. **Roundness** (R) (continuous, ratio): Calculated with the following formula:

$$R = \frac{4\pi A}{P^2}$$

.

12. **Compactness** (CO) (continuous, ratio): Measures the roundness of an object:

$$CO = \frac{E_d}{L}$$

.

13. **ShapeFactor1** (SF1) (continuous, ratio): Calculated with the following formula:

$$SF_1 = \frac{L}{A}$$

.

14. **ShapeFactor2** (SF2) (continuous, ratio): Calculated with the following formula:

$$SF_2 = \frac{l}{A}$$

.

15. **ShapeFactor3** (SF3) (continuous, ratio): Calculated with the following formula:

$$SF_3 = \frac{A}{\frac{L}{2} \cdot \frac{L}{2} \cdot \pi}$$

.

16. **ShapeFactor4** (SF4) (continuous, ratio): Calculated with the following formula:

$$SF_4 = \frac{A}{\frac{L}{2} \cdot \frac{l}{2} \cdot \pi}$$

.

17. **Class** (discrete, nominal): Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira.

2.2 Basic summary statistics of the attributes

Data issues including missing values and corrupted data, which means there are data values which are not defined or stored for an attribute in a data object, can be caused by information that is not collected or measured and attribute that is not applicable. Missing values in the dataset can make errors in the data analysis results. We can use isnull() function to detect whether there are missing values and handle them. As a result, we found that there are no missing values in the dry beans dataset.

Basic summary statistics like the count, mean, standard deviation, median, minimum, maximum and range data of the features of all dry bean samples are also given in Table 2.1. The table gives the features of the attributes in general. We can see that the standard deviation of the variable Area and ConvexArea is much larger than those of any other attributes.

Table 2.1: Basic summary statistics of dry bean varieties

	count	mean	std	median	min	max	range	missing
Area	13611.0	53048.284549	29324.095717	44652.000000	20420.000000	254616.000000	234196.000000	0
Perimeter	13611.0	855.283459	214.289696	794.941000	524.736000	1985.370000	1460.634000	0
MajorAxisLength	13611.0	320.141867	85.694186	296.883367	183.601165	738.860154	555.258989	0
MinorAxisLength	13611.0	202.270714	44.970091	192.431733	122.512653	460.198497	337.685844	0
AspectRation	13611.0	1.583242	0.246678	1.551124	1.024868	2.430306	1.405438	0
Eccentricity	13611.0	0.750895	0.092002	0.764441	0.218951	0.911423	0.692472	0
ConvexArea	13611.0	53768.200206	29774.915817	45178.000000	20684.000000	263261.000000	242577.000000	0
EquivDiameter	13611.0	253.064220	59.177120	238.438026	161.243764	569.374358	408.130594	0
Extent	13611.0	0.749733	0.049086	0.759859	0.555315	0.866195	0.310880	0
Solidity	13611.0	0.987143	0.004660	0.988283	0.919246	0.994677	0.075431	0
roundness	13611.0	0.873282	0.059520	0.883157	0.489618	0.990685	0.501067	0
Compactness	13611.0	0.799864	0.061713	0.801277	0.640577	0.987303	0.346726	0
ShapeFactor1	13611.0	0.006564	0.001128	0.006645	0.002778	0.010451	0.007673	0
ShapeFactor2	13611.0	0.001716	0.000596	0.001694	0.000564	0.003665	0.003101	0
ShapeFactor3	13611.0	0.643590	0.098996	0.642044	0.410339	0.974767	0.564428	0
ShapeFactor4	13611.0	0.995063	0.004366	0.996386	0.947687	0.999733	0.052046	0

Chapter 3

Data visualizations

3.1 Outliers

Outliers is a data point which differs significantly from other observations. Outliers have serious negative influences on classification and regression predictions. Therefore, it is important to distinguish and discard outliers. We decided to use z-score/standard-score to distinguish outliers.

First we assume that the distribution of each attribute of each class of dry beans is a normal distribution. Then mean, standard deviation and variance of each attribute can be calculated. According to three sigma theory, $P(x \leq 3\sigma) = 99.7\%$, which means that points that are more than three standard deviation from the mean are considered low-confidence points which can be seen as outliers. Hence, we decided to discard low-confidence points. Taking Solidity as an example, the comparison between data with outliers and data without outliers is shown in figure A.6. It is obvious that the point of SEKER at the bottom in figure A.6(b) has been discard in figure A.6(b).

3.2 Normality of the attributes

There are multiple ways to check if a variable follows a normal distribution. The first option would be a formal test such as Shapiro Wilk. However, this type of procedures are quite sensitive to any small difference between the tested distribution and the null hypothesis distribution (Normal distribution in this case). They are useless for large samples because every deviation from the theoretical distribution is meaningful and leads to the rejection of the null hypothesis. A better option would be to check the normality with some graphic procedures. The first and most basic one is the histogram plot A.3. Visually, any of the variables follows a normal distribution. The majority seems to follow a bimodal distribution (i.e. with two peak instead of one). This type of distributions are the result of the mix of two unimodals.On the other hand, some of them (Area,EquivDiameter, ConvexArea...)

have long tails to the right (right-skewed variables) while others are left-skewed (Solidity and ShapeFactor4). The latter most common transformations in order to obtain a normal variable include square root, cube root and log In this case two transformations were calculated:

$$\begin{aligned} \text{Solidity\_sqrt} &= \sqrt[3]{K_1 - \text{Solidity}}, \\ \text{ShapeFactor4\_sqrt} &= \sqrt[3]{K_2 - \text{ShapeFactor4}}. \end{aligned} \tag{3.1}$$

with $K_1 = 0.9946$ and $K_2 = 0.9997$, and they are also plotted in A.3.

Other interesting graphic method is the *QQ-plot*. In this procedure, the theoretical quantiles of the tested distribution are plotted against the ones from a Normal distribution. A.4 shows the mentioned tails of the different variables. In some cases the absence of normality is completely clear while in others could be more arguably. For example the new transformed variables.

Nevertheless, due to the big amount of data we can assume the normality of the variables. Actually what the Central Limit Theorem states is that the distribution of different sub samples means approximates a normal distribution as the size of the sample gets larger, regardless its original distribution. This assumption is one of the requisites for most of statistics procedures (*t-test* or linear regression).

3.3 Correlation of variables

In this data set most of the variables are correlated one with each other as it can be seen in A.1. Apart from the ones which are calculated directly from others, there is also a strong correlation between other variables except from Extent and Roundness. This results can be ratified with a Pearson test, which measures the linear correlation between every pair of variables A.5.

3.4 Feasibility of the Machine Learning models

In A.2 the box-plots show that the beans from some classes take separated values from the rest in some variables. For example, the variable MinorAxisLength would be an almost a perfect variable for classifying into a bean of the class Bombay and the rest. Other important variable for the beans of the class Horoz is AspectRation. This type of beans take larger values than the rest. Also there are other cases in which the differences are lower.

On the other hand, in 3.1 we can appreciate the predominance from the beans type Sira and Dermanson and the lower percentage of Bombay beans. This problem (imbalanced dataset) could be a problem for the classification task and it probably will require a thorough analysis in order to evaluate the performance of the classifier (both for training and final results).

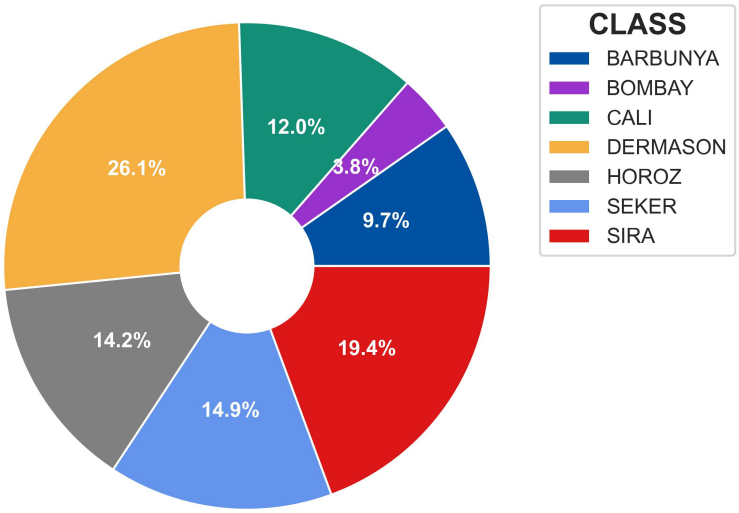


Figure 3.1. Amounts of beans from each class.

3.5 Principal Component Analysis

It is straightforward for humans to intuitively understand or visualize data in two or three dimensions. But for datasets with higher dimensions, visualization is almost impossible. However, making a dataset comprehensible by removing some attributes or features loses too much information. In order to make a dataset comprehensible and easy to visualize while preserving the maximum amount of information, PCA was applied.

Principal component analysis (PCA) is a popular technique for analyzing large datasets containing a high number of attributes per observation. PCA reduces the dimensionality of a dataset by linearly transforming the original data points into a 2- or 3-dimensional subspace, where the variance of all transformed data points is maximized.

Due to the different scales of each attribute, it is necessary to standardize all attributes with zero mean and unit variance before conducting PCA. Figure A.7 shows the difference between zero-mean standardization A.7(a) and zero-mean and unit variance standardization A.7(b). Obviously, due to the different scales of attributes, the area coefficient and convex area coefficient are much larger than other attributes, which will lead to overlooking of other attribute information. Comparing to figure A.7(a), standardizing with zero-mean and unit variance makes all attributes more balance so that most of information can be preserved.

PCA analysis can be conducted after zero-mean and unit variance standardization. First of all, all 16 attributes are explained as a function of 16 PCA components. From PCA it can be found out how much of the variation in the data each PCA

component accounts for, which is given by

$$\rho_m = \frac{s_{mm}^2}{\sum_{m'=1}^{M} s_{m'm'}^2}$$

According to A.8, with a threshold equal to 0.9, it is clear that the first three principal components can explain most of the information for all 16 attributes.

According to figure A.8, three principal directions can be obtained, which are v_1, v_2, v_3.

$$v_1 = \begin{bmatrix} 0.28245796 \\ 0.24588202 \\ -0.06144668 \\ 0.03154619 \\ -0.09132562 \\ 0.36639003 \\ -0.12504486 \\ 0.07174792 \\ -0.03506657 \\ 0.39041952 \\ -0.17768648 \\ 0.05448423 \\ -0.04629489 \\ 0.65572795 \\ -0.13319028 \\ 0.23143593 \end{bmatrix} v_2 = \begin{bmatrix} 0.31089112 \\ 0.17930292 \\ -0.0188526 \\ 0.0424679 \\ 0.08181987 \\ 0.01025082 \\ -0.0815297 \\ 0.03172951 \\ 0.15750117 \\ -0.34438307 \\ 0.19945362 \\ -0.75054998 \\ -0.31792027 \\ 0.08139011 \\ -0.01265847 \\ 0.01461438 \end{bmatrix} v_3 = \begin{bmatrix} 0.32582398 \\ 0.10075652 \\ -0.08469191 \\ 0.00679308 \\ -0.04421631 \\ 0.01490919 \\ -0.11816255 \\ -0.20094701 \\ 0.35236645 \\ -0.10199648 \\ 0.17363968 \\ 0.027355 \\ 0.68530197 \\ -0.18625119 \\ -0.17443158 \\ 0.34601942 \end{bmatrix} \tag{3.2}$$

Figure A.9 shows the linear transformation of raw data into a 3-dimension PCA subspace.

Chapter 4

Conclusions

To sum up, we have learned a lot about the data. In our dataset, there are seven different kinds of dry beans with 16 attributes. The dataset is of high quality after we discarded the outliers, since it has no missing values or corrupted data itself. We have obtained the derivation and basic summary statistics of each attribute, which help us understand the dataset more deeply and clearly. For all the attributes, we can view them as following the normal distribution while most of them are bimodal distributed.

In terms of correlations between the attributes, apart from variables that can be calculated directly from others, there is also a strong correlation between other variables except from Extent and Roundness. Based on this, we will apply one of the attributes with strong positive correlation to classification and Aspect Ration attribute to regression in the future work.

To lower the dimension of the original data, we used PCA methods to analyze the large dataset. First, we standardized the attributes with zero mean and unit variance in order to balance the weight of attributes in PCA. Then we found the three principal directions of the considered PCA components.

Combining with the previous work conducted by other researchers, our classification and regression tasks on the dry beans dataset are feasible based on our visualization, and more work should be done in the future.

Chapter 5

Exercises

1. **Question 1**. Solution: **A**.

   Time of day: nominal

   Broken Truck: ratio

   Running over: ratio

   Congestion level: ordinal

2. **Question 2**. Solution: **A**.

$$
\begin{aligned}
d_{p=\infty}(x_14, x_18) &= \sqrt[\infty]{(26-19)^\infty + 2^\infty} = \max\{|26-19|, |2-0|\} = 7.0 \\
d_{p=3}(x_14, x_18) &= \sqrt[3]{(26-19)^3 + 2^3} = 7.054 \\
d_{p=1}(x_14, x_18) &= |26-19| + |2-0| = 9 \\
d_{p=4}(x_14, x_18) &= \sqrt[4]{(26-19)^4 + 2^4} = 7.012
\end{aligned}
$$

3. **Question 3**. Solution: **A**.

$$
\text{Explained Variance} = \frac{||\mathbf{X}'||_F^2}{||\tilde{\mathbf{X}}||_F^2} = \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2} \overset{\substack{n=5 \\ M=4}}{=} \frac{\sum_{i=1}^{5} \sigma_i^2}{\sum_{i=1}^{4} \sigma_i^2} =
$$

$$
= \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2} =
$$

$$
= \boxed{0.8668 \geq 0.8}
$$

4. **Question 4**. Solution: **D**.

$$
v_2 = \begin{bmatrix} -0.5 \\ 0.23 \\ 0.23 \\ 0.09 \\ 0.8 \end{bmatrix} \begin{matrix} \text{Low Time of day} \\ \text{High Broken Truck} \\ \text{High Accident victim} \\ \\ \text{High value Defects} \end{matrix}
$$

5. **Question 5**. Solution: **A**.

   $S_1 \cup S_2 = \{$ the, bag, of, words, representation, becomes, less, parsimonious, if, we, do, not, stem $\}$

   The number of elements in $S_1 \cup S_2$ is 13.

   $S_1 \cap S_2 = \{$ the, words $\}$

   The number of elements in $S_1 \cap S_2$ is 2.

   So Jaccard similarity of S_1 and S_2 is $\frac{2}{13} = 0.153846$

6. **Question 6**. Solution: **B**.

   $p\left(\hat{x}_2 = 0, \hat{x}_7 = 0 \mid y = 2\right) = 0.81$

   $p\left(\hat{x}_2 = 0, \hat{x}_7 = 1 \mid y = 2\right) = 0.03$

   $p(y = 2) = 0.23$

   $p\left(\hat{x}_2 = 0 \mid y = 2\right) = \frac{p(\hat{x}_2=0, y=2)}{p(y=2)} = \frac{0.81 \times 0.23 + 0.03 \times 0.23}{0.23} = 0.84$

Appendix A

Graphics

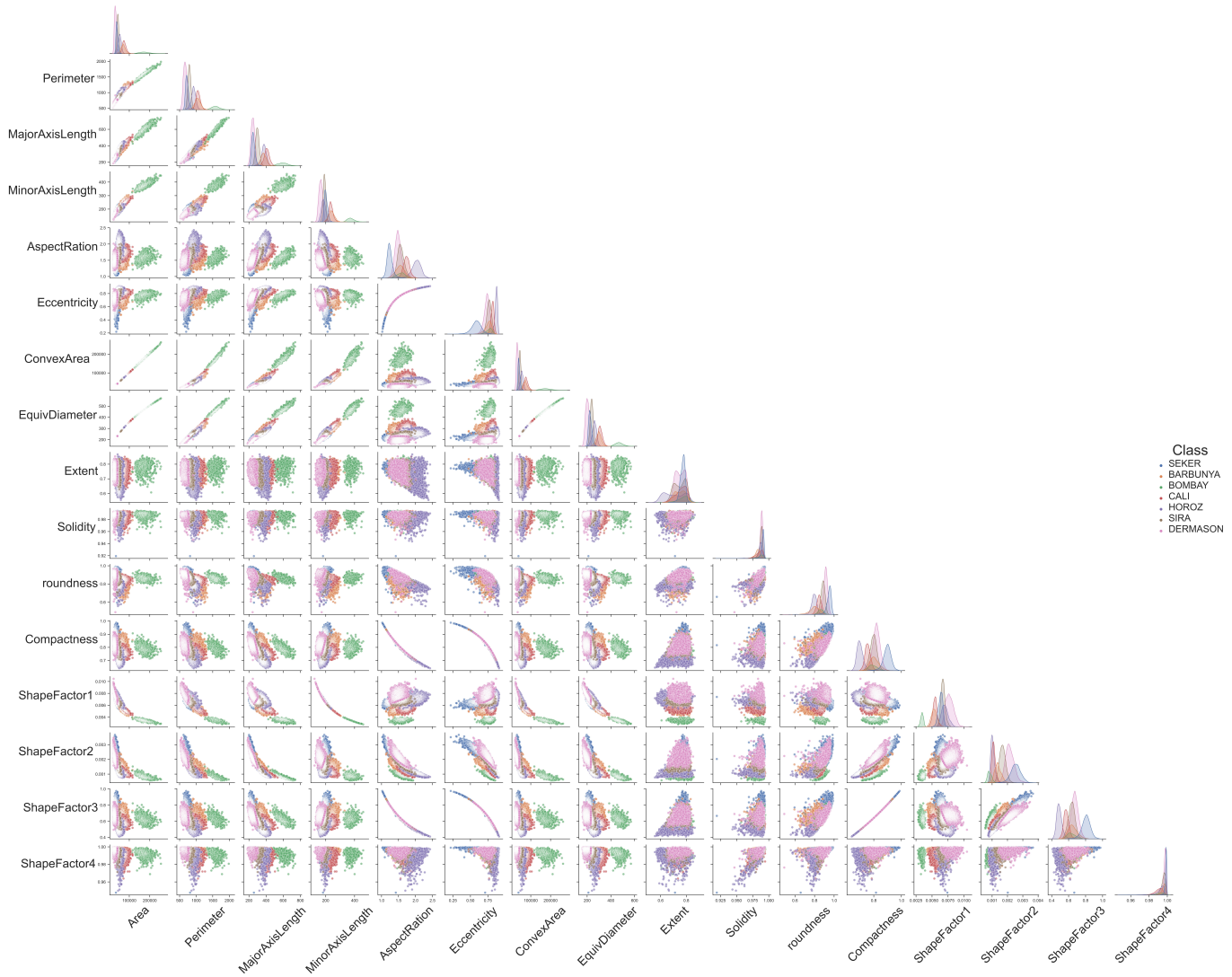


Figure A.1. Pairplot.

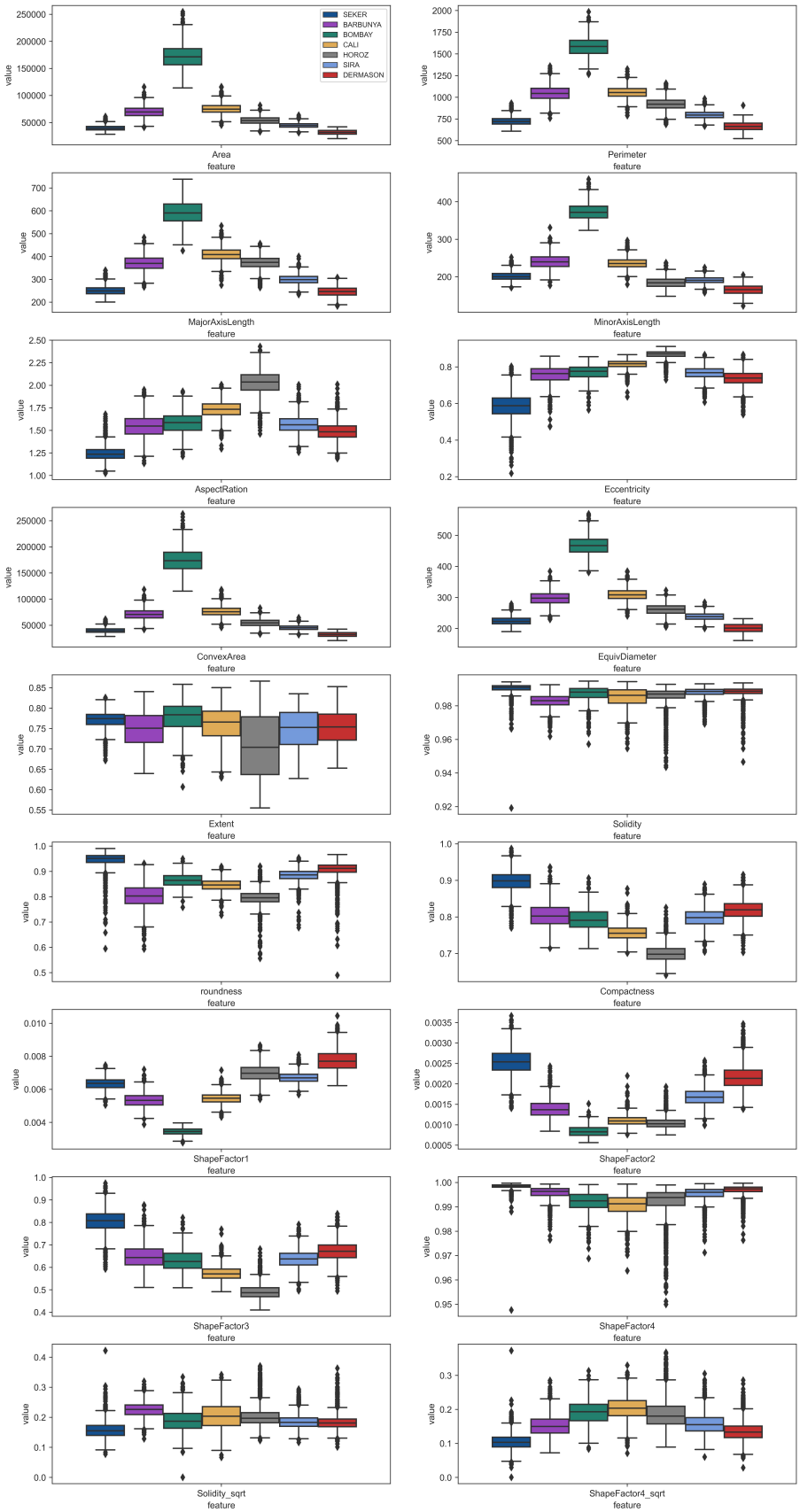


Figure A.2. Boxplots of every variable splitted by class type.

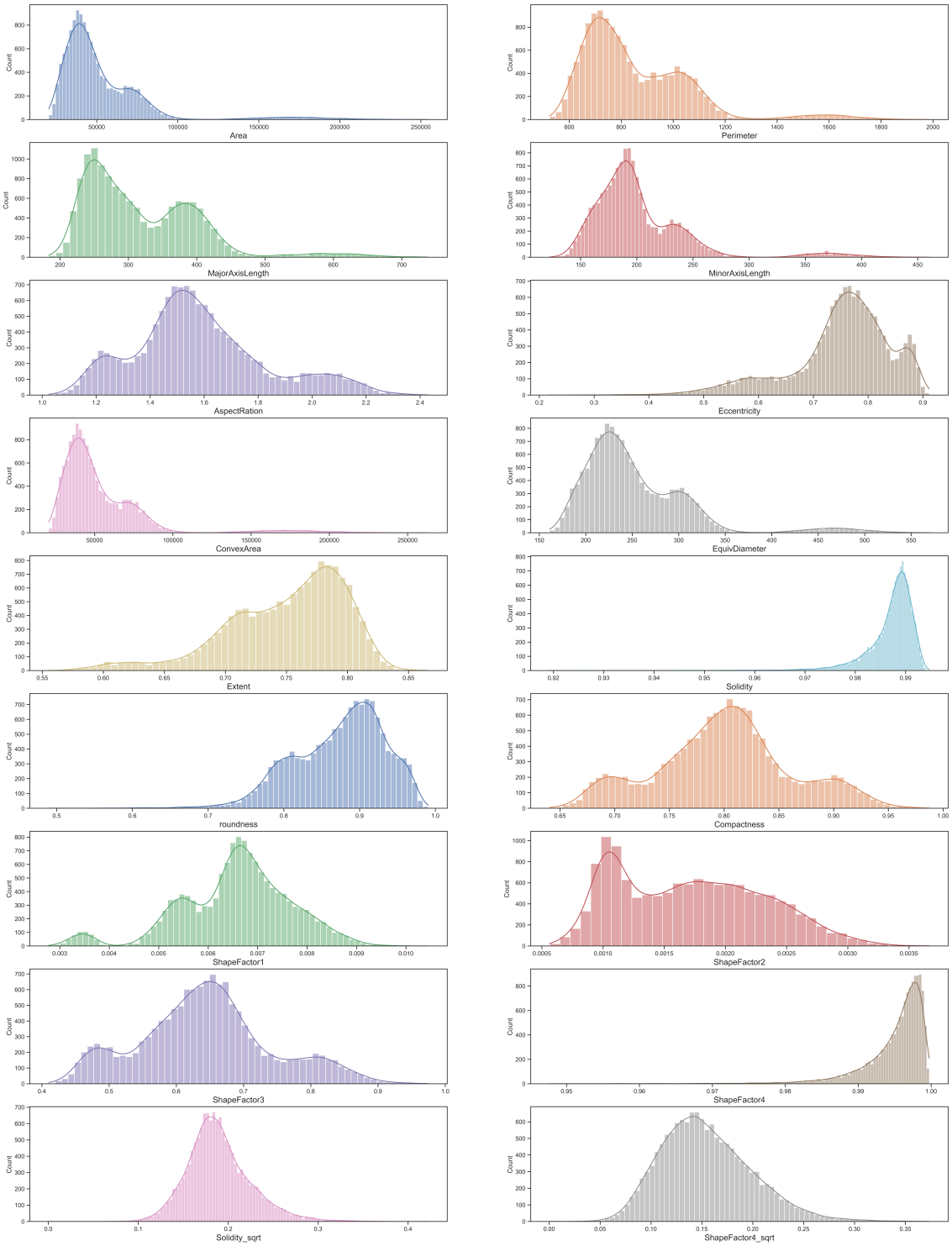


Figure A.3. Histograms for all the 16 numeric variables and the two transformations.

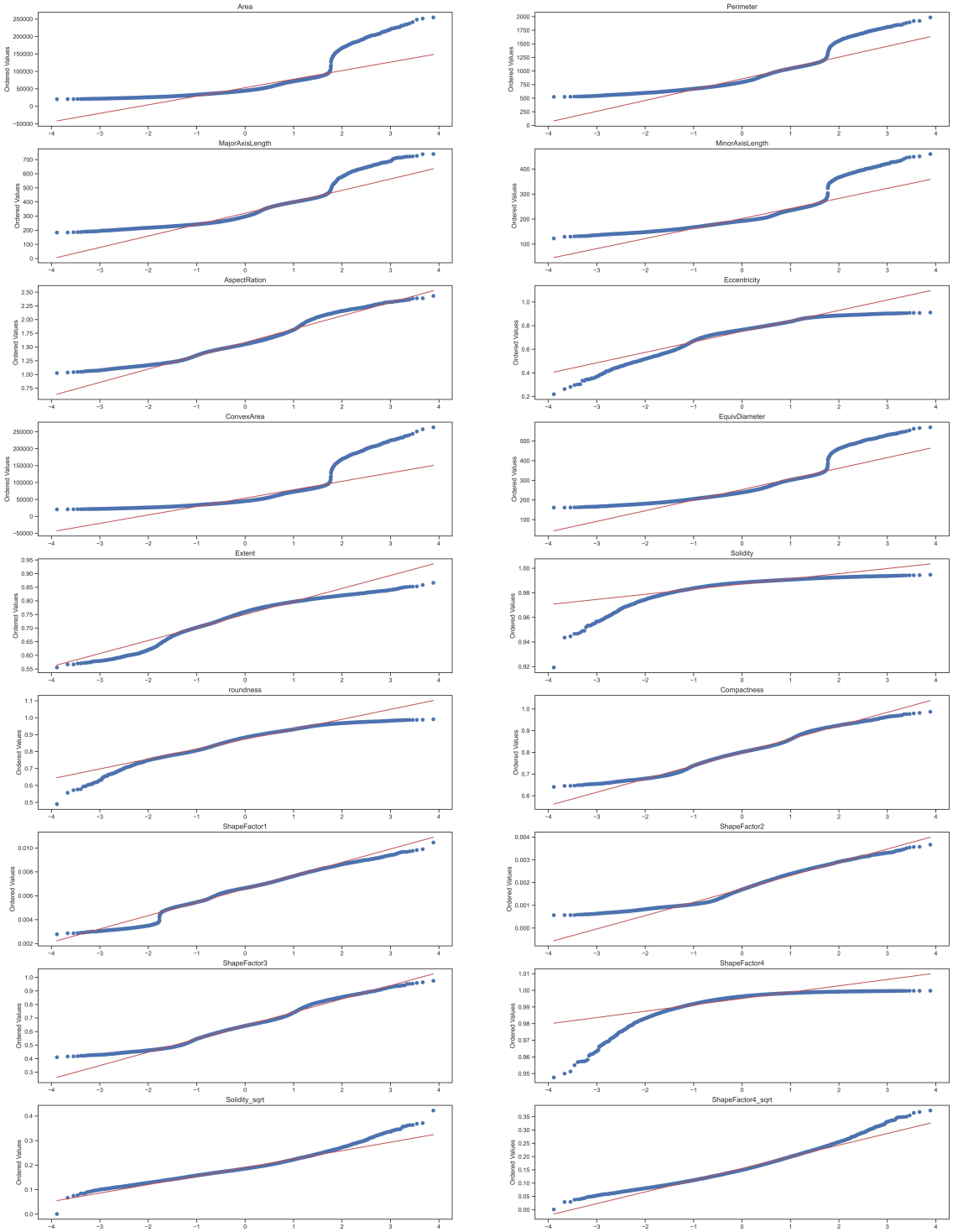


Figure A.4. QQ-plot from the variables. The quantiles from the normality-tested variable (y axis) against the quantiles of the normal distribution (x axis). Actually, the x axis shows the position of them instead of the values.

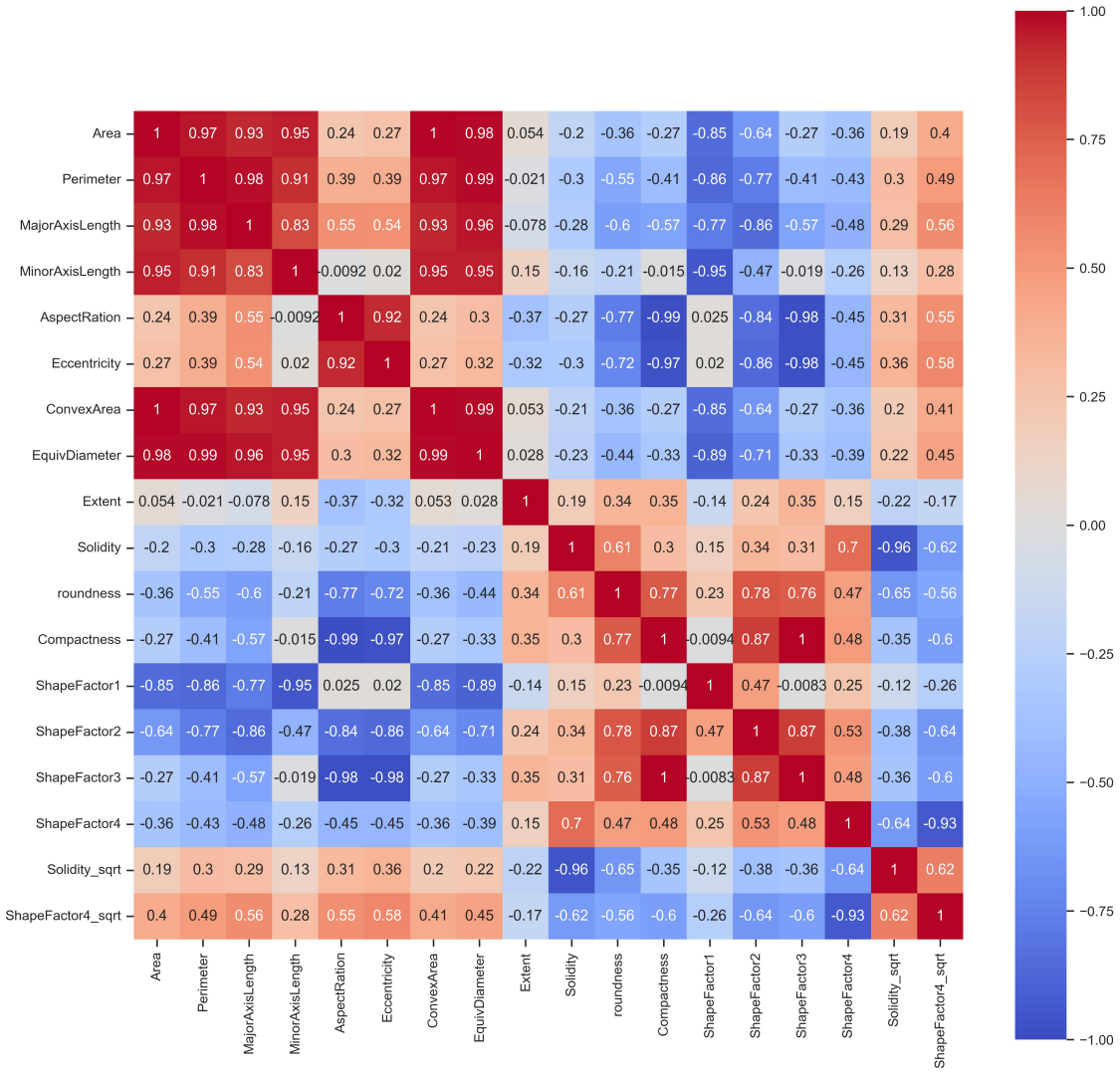
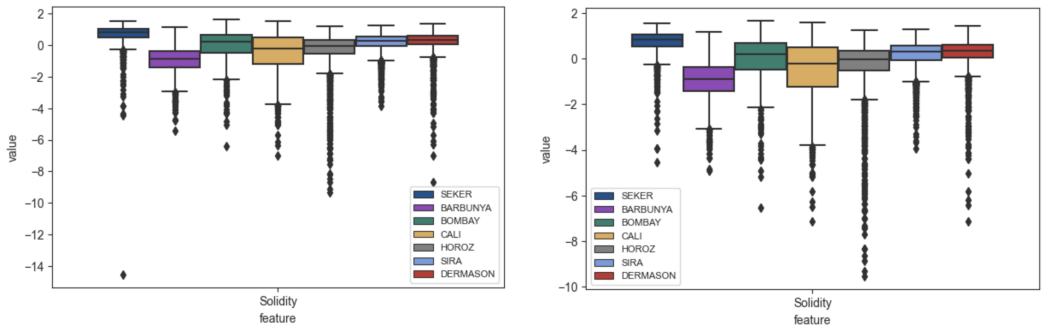
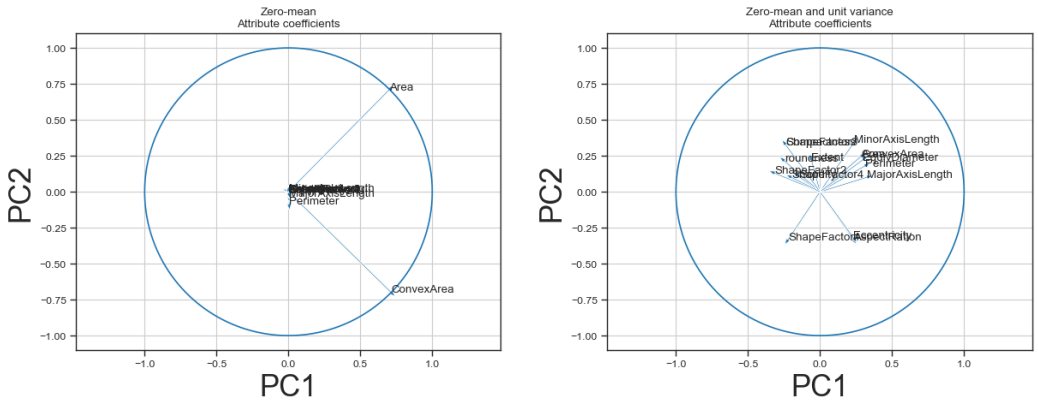


Figure A.5. Correlation plot. Each position (i, j) of the matrix indicates the linear correlation between the variables i and j with the Pearson Correlation test.



(a) With Outliers (b) Without Outliers

Figure A.6. Outliers Identification



(a) Zero-mean (b) Zero-mean and unit variance

Figure A.7. Comparison of Different Standardization

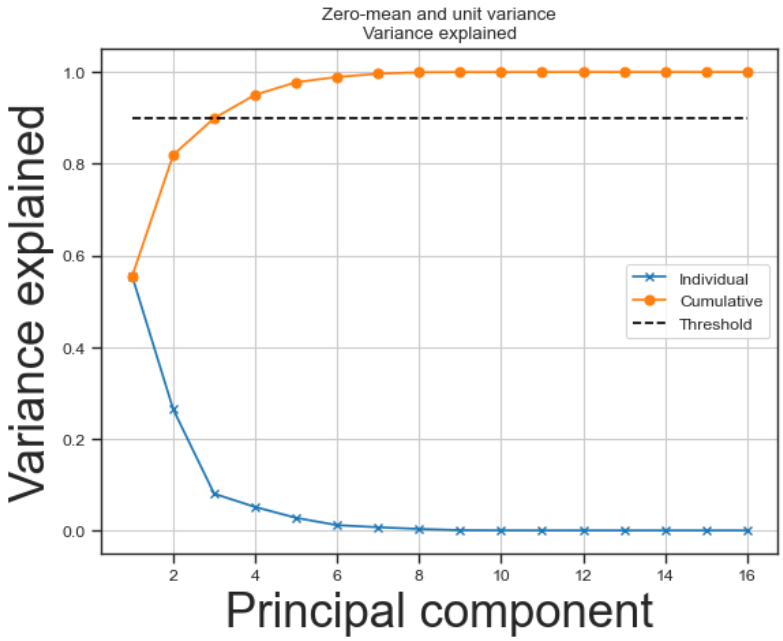


Figure A.8. PCA Analysis

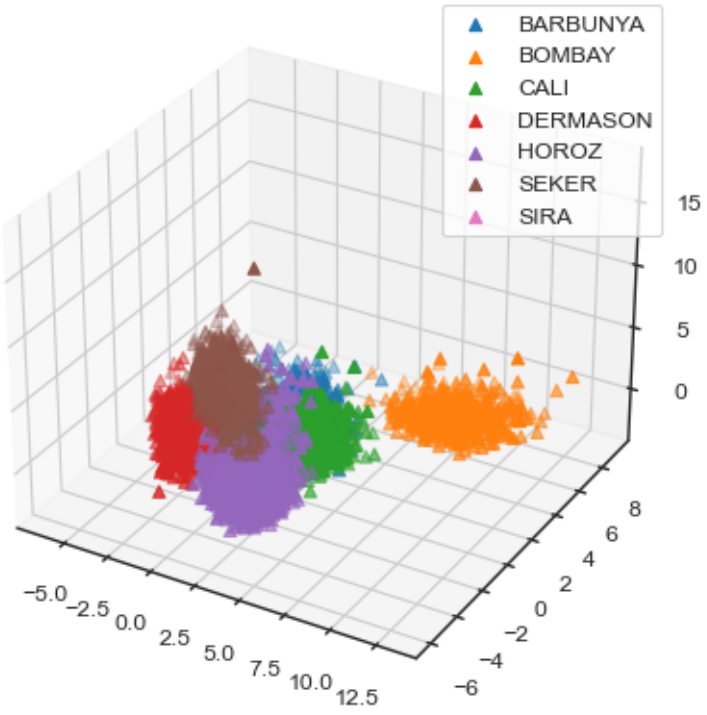


Figure A.9. 3-dimension PCA subspace

Bibliography

[1] M. Koklu and I. A. Ozkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," *Computers and Electronics in Agriculture*, vol. 174, p. 105507, 7 2020.