Machine Learning project 1

Danmarks Tekniske Universitet

October 2, 2023

Markus Kaad Heuer (S224933)
Simen Fjeld (S224920)
Bertram Hage (S224918)

Contents

Work distribution

We have all participated in making the Visuals, writing the sections and solving the problems.

Section	Markus	Simen	Bertram
Description of data set	40%	30%	30%
Attributes	30%	40%	30%
Data visualization	40%	30%	30%
PCA	30%	30%	40%
Discussion	40%	30%	30%
Exam questions	17%	67%	17%

1 Introduction

Dry beans are amoung the most cultivated crop globally and thus play a very important role in the global food supply. Correctly managing seed and correctly identifying varieties play an important role in the large scale cultivation of beans and with the use of technology these tasks can be performed faster, cheaper and more accurate.

In this report classification of 7 bean varieties are investigated using machine learning methods and relevant data visualization techniques.

2 Description of data set

The dataset contains data relating 7 different varieties of dry breans, namely Cali, Horoz, Dermason, Seker, Bombay, Barbunya and Sira. In the study pictures were taken with a high resolution camera of 13,611 samples of dry beans of one of the forementioned varieties. From the images 16 features were derived, 12 of them relating to dimensions and 4 were shape related.

The introductory paper primary objective was to develop a model able to distinguish between the different varieties of beans. This would be important to the agricultural industry. They were able to accomplish this with a 93.31% accuracy with a Support vector machine model.

We have obtained the data through the UC Irvine Machine learning Repisatory. With this data we would like to develop a model with capabilities of distuinguishing and classifying the type of bean based on the data gotten from a picture. Analysis of many different attributes and their correlations to the specific type of bean will be carried out before a distinction can be made. There are no attribute that by itself can make this distinction.

Below is a visual to give an introduction to the data. The relation between Perimiter and Area is shown, and differences between the variety of bean can already be seen.

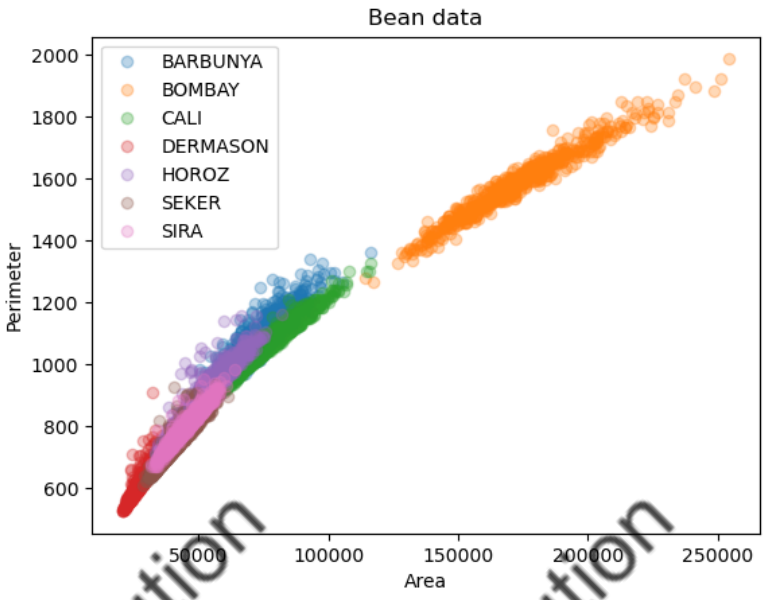


Figure 1

In the original collection of data they used pictures these to measure the many different attributes.

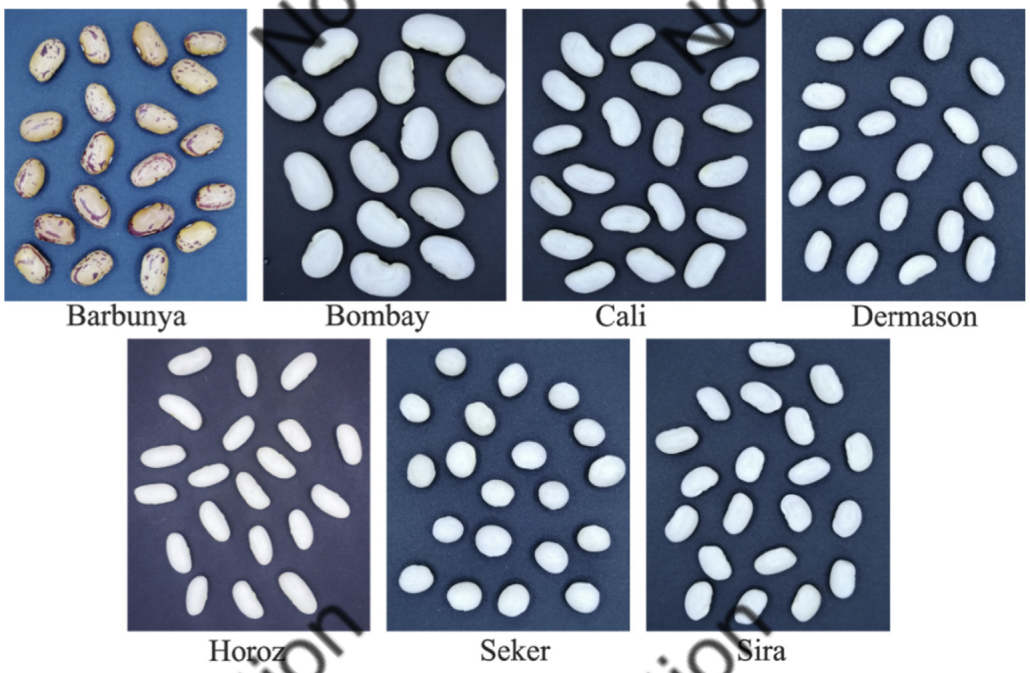


Figure 2: Sample of bean images. Source: M. Koklu, I.A. Ozkan, Multiclass classification of dry beans using computer vision and machine learning techniques, Comput. Electron. Agric. 174 (2020) 105507.

2.0.1 Attributes

There are 16 different attributes measured on the pictures taken of seven different types of beans. Altogether there are 13,611 data points, which of none are null values.

The acquisition of the data was achived by the processing of images of the dry beans. The images where processed by removal of bean shadows and converted to a binary images using a threshold determined by Otsu's global thresholding method. From here various features where extracted from each bean and put into the attributes.

Of the attributes almost all are continuous variables except for Area and ConvexArea which are discrete. These two attributes are counts of the pixels, and are represented as integers. The other 14 attributes are all continous attributes, and has real numbers as attributes. All 16 attributes are numerical values, with zero representing the absence of the measured quantity. Consequently, all the attributes are classified as ratio.

9 of the attributes are dimensional attributes calculated from the picture and computer vision system out of pixel count. 7 of the attributes are shape features and are calculated using the dimensional attributes.

- **Area**: The area of a bean is calculated from the number of pixels within the perimeter of the bean.
- **Perimeter**: The perimeter of a bean is calculated by the length of the border of the bean.
- **MajorAxisLength**: The length of the longest possible axis of the bean.

- **MinorAxisLength**: The length of the perpendicular standing axis to the major axis.
- **AspectRatio**: The ratio between MajorAxisLength and MinorAxisLength.
- **Eccentricity**: The eccentricity of an ellipse with the same moments as the bean.
- **ConvexArea**: The number of pixels within the smallest possible convex polygon containing the bean.
- **EquivDiameter**: The diameter of a circle with the same area as the bean.
- **Extend**: The ratio of the number of pixels in the rectangle bounding the bean and the brean area.
- **Solidity**: Ratio between Area and ConvexArea, calculated by $\frac{\text{Area}}{\text{ConvexArea}}$.
- **Roundness**: Calculated by $\frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$.
- **Compactness**: Measures the roundness by $\frac{\text{EquivDiameter}}{\text{MajorAxisLength}}$.
- **ShapeFactor1**: Calculated by $\frac{\text{MajorAxisLength}}{\text{Area}}$.
- **ShapeFactor2**: Calculated by $\frac{\text{MinorAxisLength}}{\text{Area}}$.
- **ShapeFactor3**: Calculated by $\frac{\frac{\text{Area}}{\text{MajorAxisLength} \cdot \text{MajorAxisLength} \cdot \pi}}{4}$.
- **ShapeFactor4**: Calculated by $\frac{\frac{\text{Area}}{\text{MajorAxisLength} \cdot \text{MinorAxisLength} \cdot \pi}}{4}$.

We have chosen to calculate the MEAN, VARIANCE and STANDARD DEVIATION for all the attributes to give a further understanding.

Index	Variable name	Type1	Type2	Mean	VAR	STDEV
1	Area	Descrete	Ratio	53048.28	859839412	29323.02
2	Perimeter	Continuous	Ratio	855.28	45916.70	214.28
3	MajorAxisLength	Continuous	Ratio	320.14	7342.95	85.69
4	MinorAxisLength	Continuous	Ratio	202.27	2022.16	44.97
5	AspectRatio	Continuous	Ratio	1.58	0.061	0.25
6	Eccentricity	Continuous	Ratio	0.75	0.0085	0.092
7	ConvexArea	Descrete	Ratio	53768.20	886480477	29773.82
8	EquivDiameter	Continuous	Ratio	253.06	3501.67	59.17
9	Extent	Continuous	Ratio	0.75	0.0024	0.049
10	Solidity	Continuous	Ratio	0.99	∼0	0.0047
11	Roundness	Continuous	Ratio	0.87	0.0035	0.060
12	Compactness	Continuous	Ratio	0.80	0.0038	0.062
13	ShapeFactor1	Continuous	Ratio	0.0066	∼0	0.0011
14	ShapeFactor2	Continuous	Ratio	0.0017	∼0	0.00060
15	ShapeFactor3	Continuous	Ratio	0.64	0.0098	0.099
16	ShapeFactor4	Continuous	Ratio	1	∼0	0.0044

The table shows all of the attributes, their respective types and the basic summary statistics.

3 Data visualization

Looking into outliers in the data we have initially chosen to remove values that are more than 3 standard deviations away from the center. In mathematical statistics this is a standard within normal distribution. A lot of natural phenomena is normally distributed and these values as well appear to be normally distributed.

In order to calculate mean and deviations of our data it has to be evaluated in smaller parts, and evaluated bean after bean. We also have to look at one attribute at a time. Calculations have been made on the values of all attributes and all types of beans and the results are shows in the table below.

Total values is a sum of all the individual outliers across all attributes for a certain bean.

Since some rows with measurements on one bean have outliers in multiple categories we dont have to remove as many rows as total outliers. Rows removed are the number of rows with at least 1 outlier.

Type	Total Values	Rows Removed	Total Observations	% removed
BARBUNYA	121	74	1322	**0.056**
BOMBAY	56	27	522	**0.052**
CALI	172	65	1630	**0.040**
DERMASON	288	164	3546	**0.046**
HOROZ	309	129	1928	**0.067**
SEKER	307	126	2027	**0.062**
SIRA	217	117	2636	**0.044**
SUM	**1470**	**702**	**13611**	**0.052**

If we delete all the rows with at least one outlier, we will altogether be removing 702 rows of data which is 5.2% of the total amount. It has to be noted that a row with an outlier consists of 15 other values that may not be outliers in their category. Therefore it can be

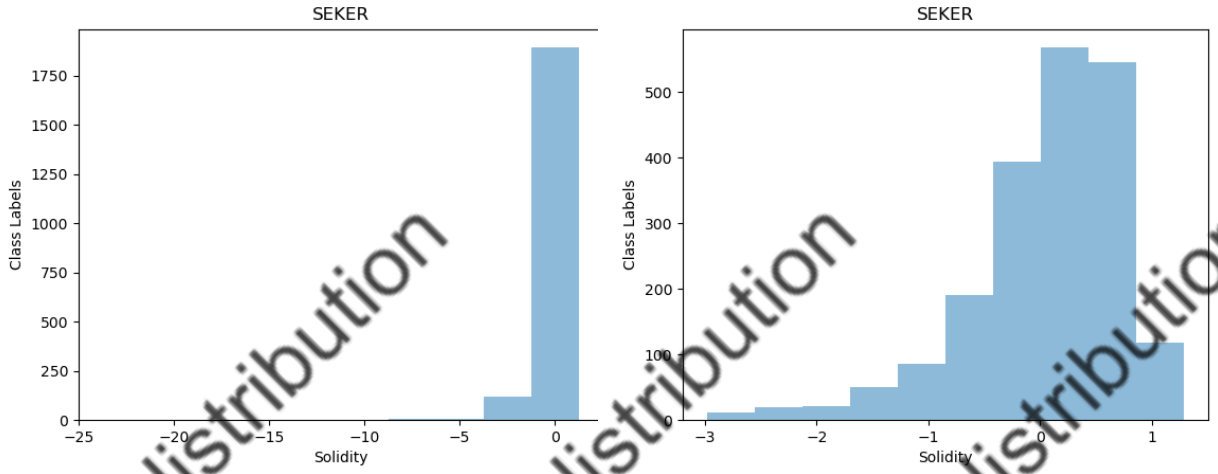
discussed weather or not to remove the rows or assigning new values to the outliers. We could assign the average value to the outlying values.

Even though it seems like a very high percentage of outliers it is not. The whole dataset consists of $13611 \cdot 16$ values. If we calculate the percentage of outliers.

$$\frac{1470}{13611 \cdot 16} = 0.006750055102 \approx 0.68\%$$

The whole dataset has 0.68% outliers further away than 3 standard deviations, which strongly suggests a normal distribution.

We can plot the distribution of the Bean SEKER with regard to attribute Solidity. Th



(a) Before removing Outliers (b) After removing outliers

Figure 3: Plot of: Bean SEKER, attribute Solidity

When the outliers are corrected for we will be able to do proper analysis and further model creation in the next report. Removing all the rows with any outliers seem excessive and will be removing too much data from the dataset.

Many of the Attributes are correlated which is shown in figure 3

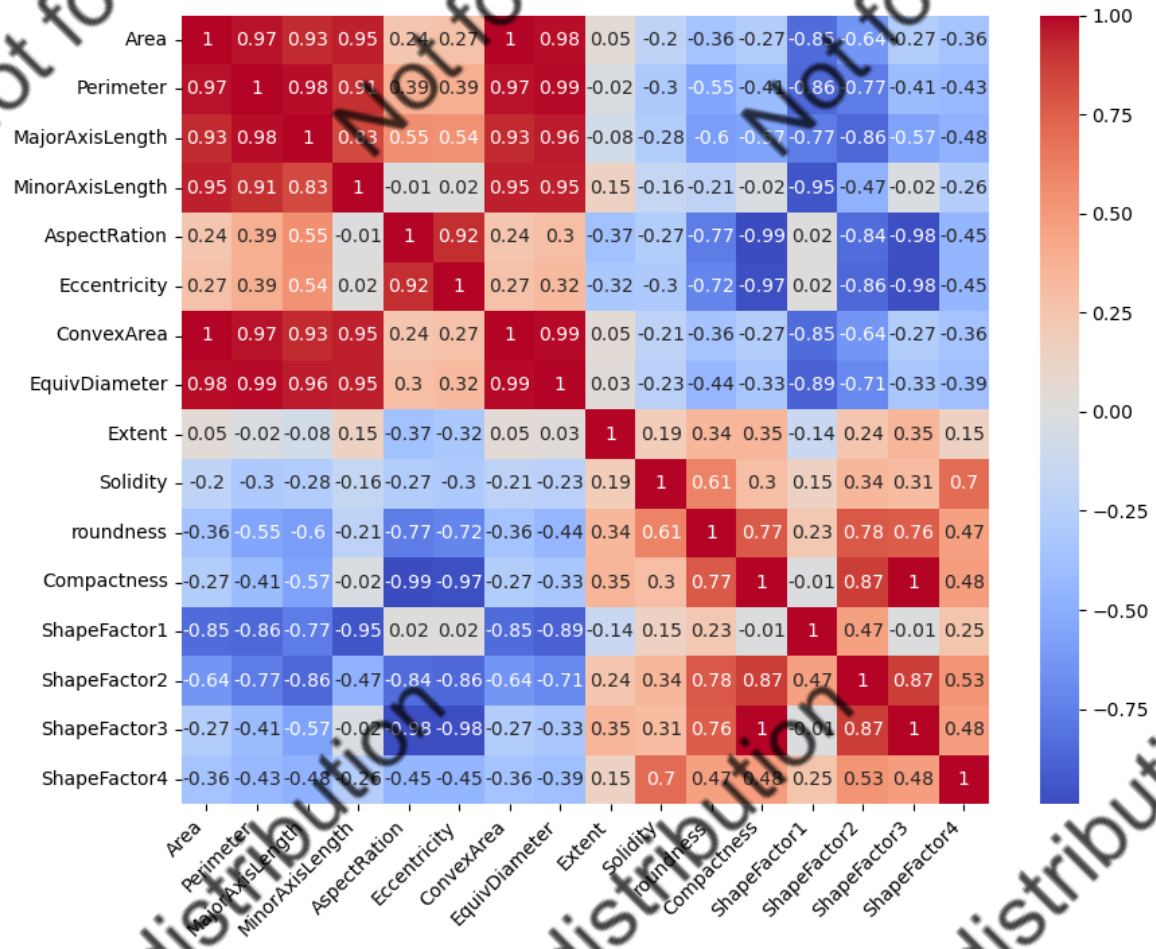


Figure 4: Heatmap of correlation

Values close to 1 means that they are highly correlated, fx. the Area and the Perimiter. Negative 1 means that they are opposites correlated, meaning that when one value is high there is a very high correlation to the other value being low. Close to Zero means little or no correlation.

4 Principal Component Analysis (PCA)

In order to reduce the dimensionality of our data we carry out a PCA analysis.

We start by standardizing our data by subtracting the mean and dividing by the standard deviation. This makes it possible for us to better compare across the different attributes. We then carry out our Singular Value Decomposition (SVD). From the SVD we can calculate the explained variance by each of the principal components:

$$\text{Explained variance} = \frac{\sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2},$$

where σ_i is the singular value for the i'th principal component.

In Figure 5 the amount of explained variance is plotted as a function of the number of principal components included. The cumulative explained variance of the first three PCA

components is approximately 0.899 and thus barely cutting a threshold of 0.9. Including the first four PCA components, however, brings the cumulative explained variance to approximatly 0.95.

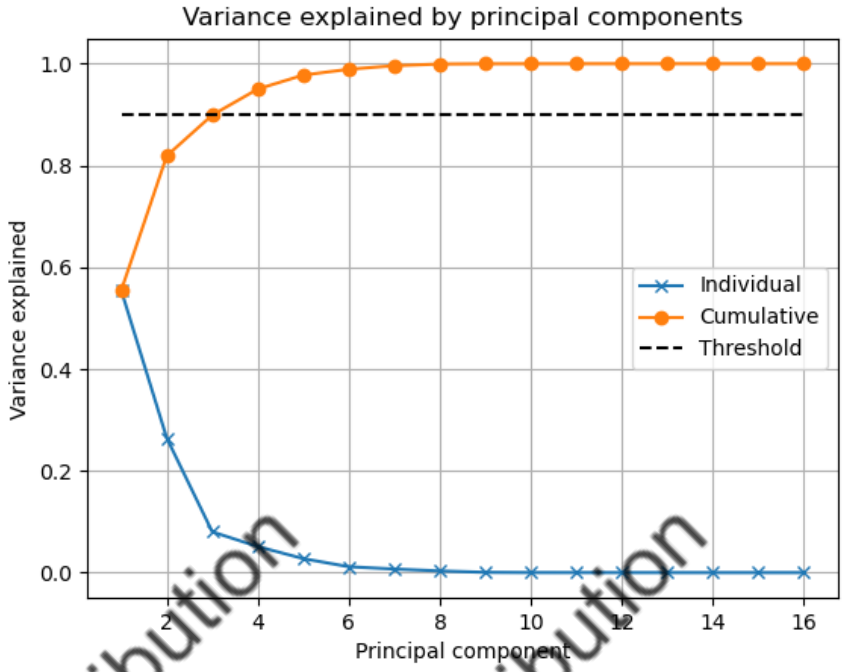


Figure 5: Individual and cumulative explained variance by principal components.

In Figure 6 the component coefficients of the first three PCA components are displayed from the V-matrix from the SVD.

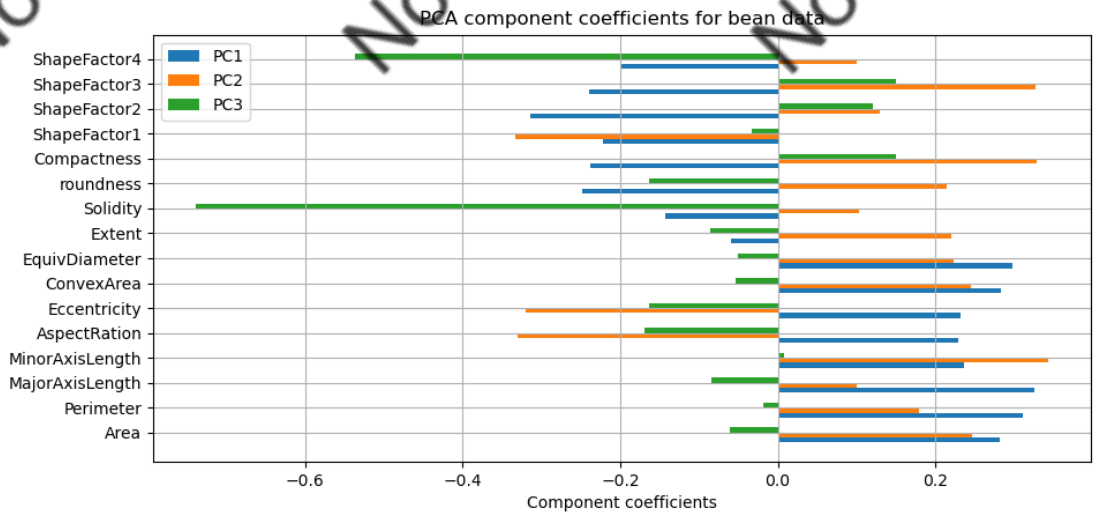


Figure 6: Coefficients for principal components.

It can be seen that PC1 has positive coeficcients for attributes relating to primarily size and negative coefficients for attributes like roundness, compactness and the shape factors. This sugggests that PC1 discrimitates between large beans with irregular shapes,

like Bombay, and small beans with more round and compact shapes, like Seeker (see Figure 2).

PC2 has positive coefficients for shape related coefficients suggesting a more round and compact shape while having negative coefficients for attributes suggesting a more elongated shape. More particular PC2 has especially positive coefficients for ShapeFactor3, compactness and MinorAxisLength while having especially negative coefficients for aspect ratio, eccentricity and ShapeFactor1. Thus elongated beans, like Horoz score low on PC2 while more round and compact beans, like Seker and Bombay score high on PC2.

PC3 has a particular negative coefficient for solidity and ShapeFactor4. Thus PC3 seems to be sensitive to how much the bean deviates from a convex shape.

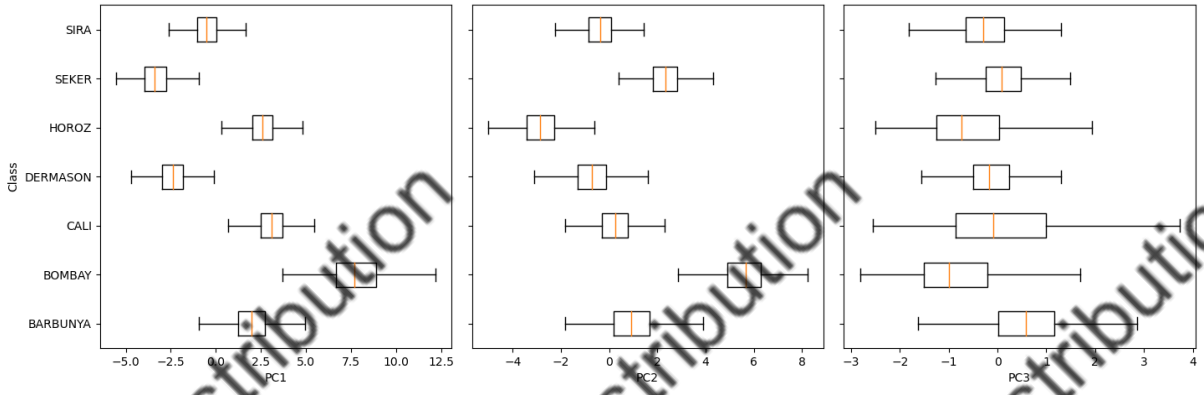


Figure 7: Boxplots for principal components.

In Figure 7 boxplots for PC1, PC2 and PC3 are shown for the 7 individual classes. Note that outliers are not represented on the figure as these have already been adressed and leaving them out gives a clearer view. Here it is confirmed that a large bean like Bombay score high on PC1 while a small and round bean like Seeker score low on PC1 but higher on PC2. A bean like Horoz, despite its relatively smalle size, still score high due to it's elongated shape which is also a contribute to PC1. It is also clear from Figure 7 that PC3 is less able to explain variance in the beans than the other two principal components which goes in line with the fact that the explained variance for PC3 is lower than that of PC1 and PC2.

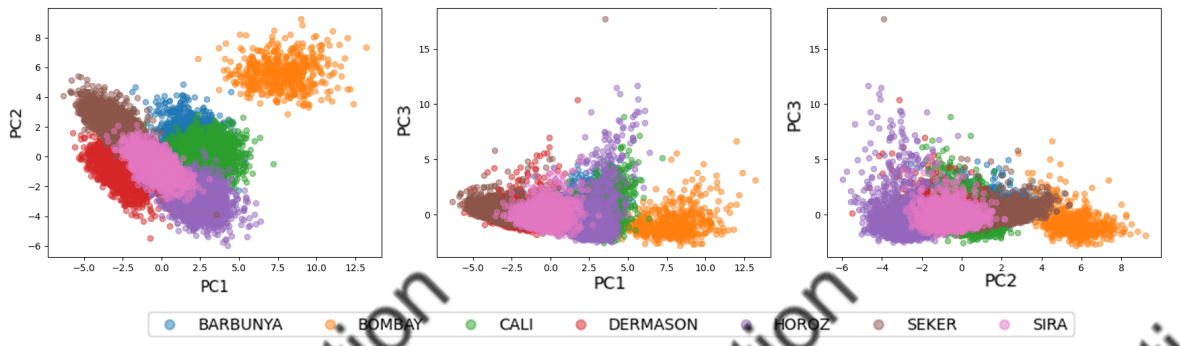


Figure 8: Scatter plot of principal components in 2 dimensions.

In Figure 8 the data is plotted onto two principal components at a time and thus in 2 dimensions. It can be seen on the left most figure that PC1 and PC2 combined is able to distinguish the beans relatively well which goes in line with the fact that their cumulative explained variance is approximatly 0.82.

5 Discussion

Through preliminary analysis a lot of variation in the bean types can be seen across the attributes, which gives confidence towards building a model in future reports. The visualizations also made it obvious that there are some values in the dataset that are many standard deviations away from the others and therefore can be removed before further analysis. Since there are so many different attributes a decision has to be made on how to handle a single outlier, since much of the other data in a row does not have to be far from the mean. The PCA analysis gave insight into the ability to destinguish between the different types of beans. Especially PC1 and PC2 give a lot of explanation for the different types. It can thought also be seen that some beans are very much alike like the Cali and Barbunya bean, and it will be difficult to seperate those. Overall it seams to be possible to distinguish new beans in the future.

6 Problems

6.1 Question 1

The dataset consists of 7 attributes, x_1-x_7 and y.

x_1 describes the time of day, where each object corresponds to a 30-minute interval between 7:00 and 20:30. This attribute consists of the integers from 1 and up to 27. The type of this attribute is ordinal, because the objects can be ranked as the time of the day, but has no distance between them.

x_2-x_7 are all numbers of different observations. All of the objectives in these attributes are actual numbers, where zero means abcence of what is measured. Therefore these are attributes to be categorized as ratios.

The last attribute is y, which describes the level of congestion. y = 1 (corresponding to no congestion), y= 2 (corresponding to a light congestion), y = 3(corresponding to an intermediate congestion), and y = 4 (corresponding to a heavy congestion). All of the objects for this attribute belongs to a category that can be ranked, and therefor this is also an ordinal attribute.

Now that we know the type of all of the attributes, we can state that **C** is the corrcet statement.

6.2 Question 2

For this exersice we apply the formula for "General Minkowsky distance:

When p is a constant:

$$d_p(x,y) = (\sum_{j=1}^{M} |x_j - y_j|^p)^{\frac{1}{p}}$$

When p is equal to ∞:

$$d_\infty(x,y) \;=\; max\left(|x_1 - y_1|\;,\; |x_2 - y_2|\;,\; \ldots\;,\; |x_M - y_M|\right)$$

We now insert the p-value and the values from the vectors, and calcuate the p-norm distance:

$$A \Rightarrow d_{p=\infty}(x_{14}, x_{18}) = \; max\left(|26 - 19]\;,\; |2 - 0|\right) = 7$$

$$B \Rightarrow d_{p=3}(x_{14}, x_{18}) = \sum_{j=1}^{M} |x_j - y_j|^3)^{\frac{1}{3}} \Rightarrow ((|26 - 19|)^3 + (|2 - 0|)^3)^{\frac{1}{3}} = 7.054004063$$

$$C \Rightarrow d_{p=1}(x_{14}, x_{18}) = \sum_{j=1}^{M} |x_j - y_j|) \Rightarrow ((|26 - 19|) + (|2 - 0|)) = 9$$

$$D \Rightarrow d_{p=4}(x_{14}, x_{18}) = \sum_{j=1}^{M} |x_j - y_j|^4)^{\frac{1}{4}} \Rightarrow ((|26 - 19|)^4 + (|2 - 0|)^4)^{\frac{1}{4}} = 7.011632778$$

We can after these calculations see that statement **A** is the correct one.

6.3 Question 3

We are familliar with matrix V and matrix S. We are also informed that:

$$X = USV^T$$

We will now calculate the explained variance. The explained variance is given by this formula:

$$Explained\ var. = \frac{\sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$$

We also know that $S = \Sigma$, and that Σ is a diagonal matrix with the standard deviations for each of the PCA components. σ_i is the singular value for the i'th principal component

The numerator are the sum of all the relevant σ_i^2-values, and the denominator are all of the σ_i^2-values added together.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & ... & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & ... & \sigma_N \end{bmatrix},\ \sigma_1 \geq\ \sigma_2 \geq ... \geq \sigma_N \geq 0$$

Now that we are familliar with all the standard deviations, ve can calculate the explained variance using the formula:

$$\sigma_1 = 13.9\ ,\ \sigma_2 = 12.47\ ,\ \sigma_3 = 11.48\ \sigma_4 = 10.03\ ,\ \sigma_5 = 9.45\ ,$$

$$A \Rightarrow \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.8667931475$$

$$B \Rightarrow \frac{11.48^2 + 10.03^2 + 9.45^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.4798501562$$

$$C \Rightarrow \frac{13.9^2 + 12.47^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.5201498438$$

$$D \Rightarrow \frac{13.9^2 + 12.47^2 + 11.48^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.7167331912$$

We now see that statement **A** is the correct one.

6.4 Question 4

For this Exercise we asses the Matrix V and asses wheather different attributes have a positive or negative effect on each of the five Principal components.

The way to do solve this question is to look at attribute and see if the effect on the certain principal component is positive or negative. Then we are given weather the value for the attribute is high or low which can help us estimate wheather which effect the attributes has on the particular component.

In answer **D** we want the attributes to typically give a pisitive value of the projection of Principal component number 2

- **Low Value** Time of day x_1 is **Negative**

- **High Value** Broken truck x_2 is **Positive**
- **High Value** Accident victim x_3 is **Positive**
- **High Value** Defects x_5 is **Positive**

With the values given in the description and in matrix V, the conclusion can be made that **D** is right in their assumption

6.5 Question 5

We have the two text documents s_1 and s_2

The Jaccard similarity is calculated as follows:

$$J(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|}$$

$$\frac{|(the),(words)|}{|(the),(bag),(of),(words),(representation),(becomes),(less),(parsimoneous),(if),(we),(do),(not),(stem)|}$$

$$= \frac{2}{13} = 0.15384615384$$

The Jaccard similarity between the two text documents are 0.153846, and we can therefore see that **A** is the correct answer.

6.6 Question 6

We can see from Table 2 that $p(\hat{x_2} = 0, \hat{x_7} = 0|y = 2) = 0.81$ and $p(\hat{x_2} = 0, \hat{x_7} = 1|y = 2) = 0.03$. Since $p(\hat{x_7} = 0) + p(\hat{x_7} = 1) = 1$ it must mean that

$$p(\hat{x_2} = 0|y = 2) = p(\hat{x_2} = 0, \hat{x_7} = 0|y = 2) + p(\hat{x_2} = 0, \hat{x_7} = 1|y = 2) = 0.81 + 0.03 = 0.84.$$

Thus B is the correct answer

$$p(\hat{x_2} = 0|y = 2) = 0.84.$$