Universitat Politècnica de Catalunya

Grau en Ciència i Enginyeria de Dades

Prediction of male body fat from body measurements

AD - Anàlisi multivariant (GCED)

Sonia Castro & Lucia De Pineda

June 2022

Table of contents

1 Introduction

1.1 Goals of the project

The main goal of this project is to analyze a dataset using different methods seen in class. The aim is to understand better how these methods work and how they help us when analyzing data. In particular, the methods we are going to present are: Principal Component Analysis (PCA), Multidimensional Scaling (MDS) and Cluster Analysis.

1.2 Dataset description

The dataset used in this project has been obtained from this Repository. It has 252 samples with information of 15 decimal variables: the response variable *percentage of body fat*, *age*, *weight*, *height*, *density* and ten body circumference measurements. It also includes a variable named *density* which description is not included in the dataset. We checked its correlation with the response variable and saw that it is really high (0.98888). Therefore, we believe that it is a calculation made from the rest of the variables, so we will not use it for future analysis.

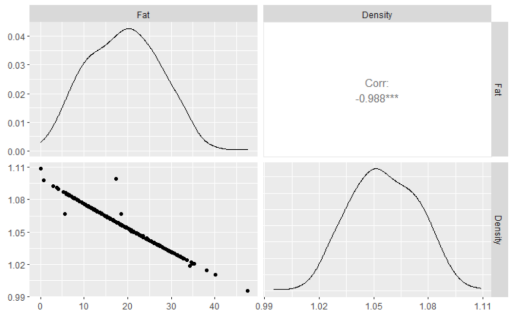

Figure 1: Correlation between percentage of body fat and density.

2 Methods and results

We have analyzed the previously described dataset with three different methods: Principal Component Analysis (PCA), Multidimensional scaling (MDS) and cluster analysis.

2.1 PCA

The first method is Principal Component Analysis, which is a statistical technique for finding low-dimensional summaries of high-dimensional datasets, so they can be more easily visualized and analysed. These summaries are called Principal Components (PCs), they are uncorrelated variables that are linear combinations of the original variables. PCs help us interpret the importance of each original variable, as well as how they relate to each other. The idea is that most of the variability in the data can be explained by the first few k principal components, this way we reduce dimensionality of our data. Therefore, what we are going to do is analyze the first PCs that account for most of the variability and interpret their meaning.

The first step is to check the correlations between variables. We compute all of them and we see that the most correlated with the response variable ($percentage\ of\ body\ fat$) are the *abdomen* and *chest*. The most correlated in general are the variable *weight* with different body measurements.

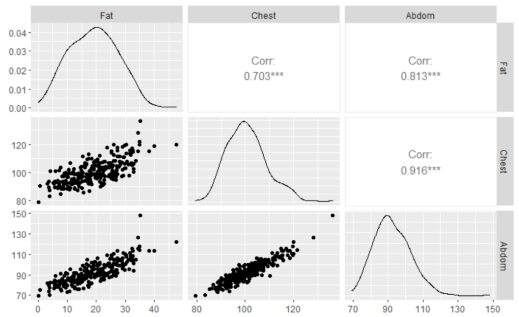

Figure 2: Scatterplot, distribution and correlation values between percentage of body fat, abdomen and chest.

As the correlations between variables are quite high, we believe the PC analysis will be useful for our dataset and we will be able to reduce dimensionality.

Then, we standardize the data and leave out the variable fat, we run PCA and calculate the proportion of variance explained by the first PCs. The proportion explained by the first PC is of 0.618 and by the second of 0.104. The third one explains 0.077 of the variance and the fourth 0.048. Combined the first four PCs explain 0.847 the variance, which is a considerable amount.

To represent the proportions of variance we just stated, we build a scree plot:

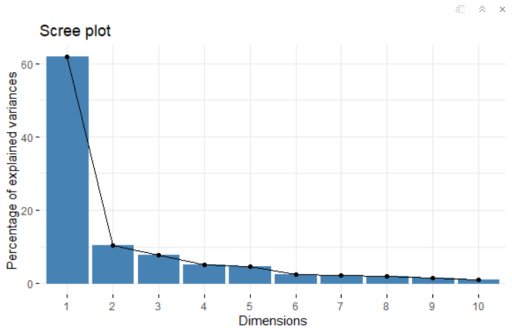

Figure 3: Percentages of variances explained by the PCs.

In this plot we can see what we commented before about the proportions of variance each PC from 1 to 10 explains. The first PC account for the biggest proportion, with a really important difference compared to the other ones.

Then, we plot the loadings of the first two PCs, as well as the biplot.

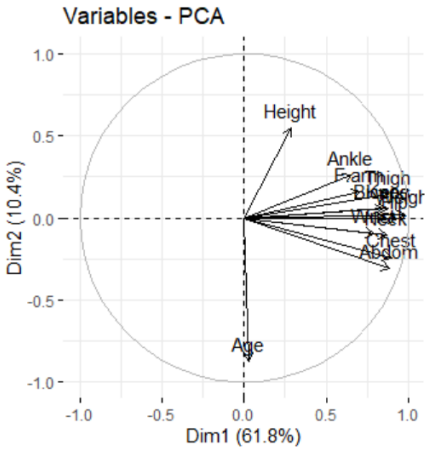

Figure 4: Loadings plot of PC1 and PC2.

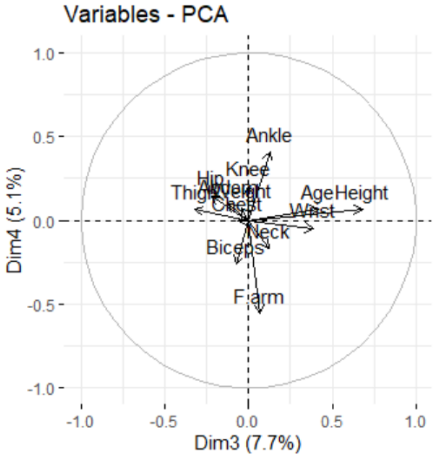

Figure 5: Loadings plot of PC3 and PC4.

In the loadings plot we can see that in PC1 all values are positive but Height and Age are given lower weight, as we expected since they are less correlated with the rest of the variables (including body fat). So in PC1 the variables more taken into account are the body measurements and weight.

In PC2 we have positive and negative values. We can see that *Height* and *Age* have the highest positive and negative weights. In the PC3 and PC4 the interpretation is more difficult as we cannot identify a clear relationship between the variables with positive/negative weights in each one.

The biplot combines the loadings and scores plot into one plot. We already interpreted the loadings of the first two PCs. For the score plot, each point represents the score of a subject for those PCs (PC1 and PC2). The score for a subject for a PC is the value calculated with the coefficients for each variable. As the variables are centered, the points close to the origin are the ones closer to the score average. The points that are far away from the origin, have scores that stand out from the rest.

If two subjects are close in the biplot, it means that they have similar scores. As the first two PCs explain most of the variability, it also means that their original values are similar too.

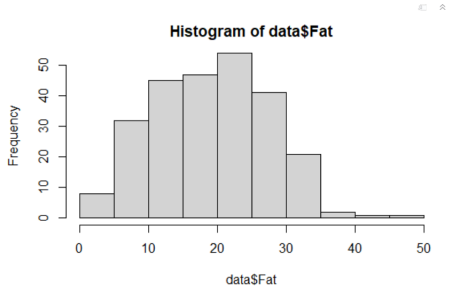

Figure 6: Histogram of the variable fat

To make the interpretation easier we have divided our response variable fat into 5 categories depending on their value: [0,10],[10,20],[20,30], [30,40] and [40,50]. And we have colored the plot to see if there is a relation between the position of the samples and their level of body fat.

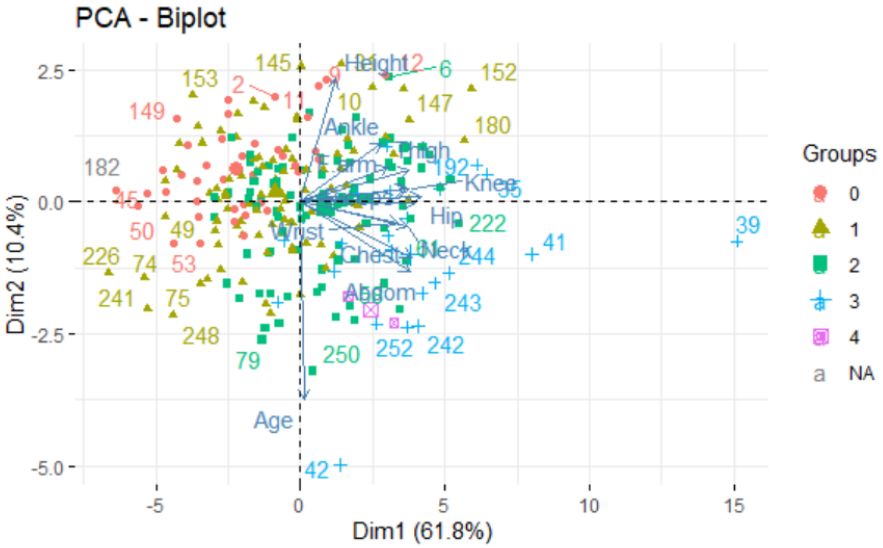

Figure 7: Biplot of the PCA.

As we can see most of the samples in class 0 (lowest body fat, red) are on the left part of the plot indicating that their values in body measurements and weight (the coefficients with more weight in PC1) are lower than the mean. The ones on the right of the plot are also high which means their measurements may have been higher than the mean but they are very tall and young.The same happens with class 1 (yellow), that we find on the left or top right sides of the plot.

If we check class 2 (green) we see how this one is more centered as it has observations on both sides of the mean. On the other hand, class 3 only has points on the right side of the PC1 so they have higher body measurements which makes sense. In class 4 (blue) we can see that most of them are the ones more on the right since they have the highest body measurements and weight. A few of them are on the left but down so maybe they have less weight and body measurements but they are older and shorter.

Lastly, we can see 3 persons in class 5 (highest body fast), they are not on an extreme of the plot, so it can be difficult to guess why they have those values. Anyways they are on the right bottom side, which makes sense.

To finish the PCA, we analyse the PC loadings and values of the variables of some subjects that stand out: 39,41,42.

We can conclude that subject 39's score is very far from the mean in PC1, that's because a lot of variables relating to body measurements have higher value than the mean, specially *weight* and *hip*. In the case of 41 is the same reason, but the difference is lower so it is no as far of the mean for PC1, it also has a lower height. Finally, subject 42 is far from the mean in PC2, as its *height* has a lower value than the mean by far, and height has one of the highest weights in PC2.

2.2 MDS

The idea of Multidimensional Scaling is to find a good lower-dimensional approximation of the original data in a Euclidean space. We suppose we have a matrix D containing the distances between observations. The objective is to find a k-dimensional Euclidean space that preserves the distances d_{ij} as well as possible. So, we will have k-dimensional points $(v_1,v_2,...,v_n)$ so that $d_E(v_i,v_j) \approx d_{ij}$.

An observation is that if the distances d_{ij} are Euclidean distances between rows of a dataset, then MDS is equivalent to PCA. However, the distances d_{ij} might not be Euclidean, so MDS generalizes PCA. Another advantage of MDS is the fact that it can be used even if the original dataset is not available, but we do have the distance matrix.

To perform MDS on our dataset, the first step is to calculate the distance matrix D. Once we have our D matrix, we perform MDS on it, using the *cmdscale* function. The last step will be to calculate the residual matrix between D and $\hat{D}$.

So, the first thing was to decide which notion of distance to use. We decided to try different methods available with the 'dist' function in R. We calculated Euclidean, Manhattan, Canberra and Minkowski distances and calculated the residual for each of them. The best one was the Euclidean one, as it calculates the distance between two real-valued vectors and all of our variables are numeric. Therefore, we will have a matrix containing the Euclidean distance between each pair of rows of the dataset.

As we said before, if the distance used is the Euclidean one, MDS is equivalent to PCA, so the proportions of the variance explained by the axes will be the same. We remind them: 0.6185 (1st), 0.1042 (2nd), 0.0771 (3rd) and 0.0513 (4th). The first four axes combined account for 85% of the variance. So, we use k=4 to perform MDS and see how well it approximates the distances. The graphical representation of the two first axes of MDS is:

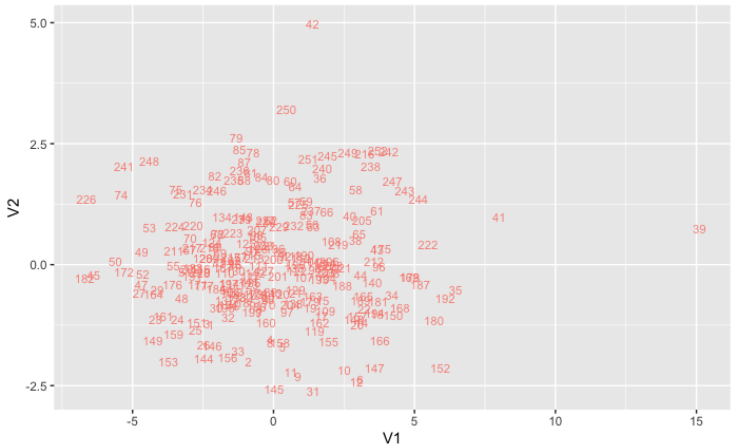

Figure 8: MDS plot first and second axes

Then, we compute the distances on the MDS, building the $\hat{D}$ matrix. To see how accurate the approximation is, we calculate the difference between $\hat{D}$ and D, building the residual matrix. Also, to analyze better the values the errors take, we print the histogram:

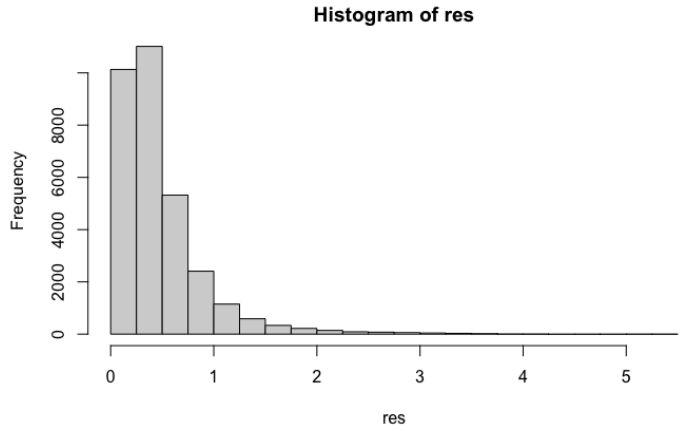

Figure 9: Histogram of the residuals

We can see how most of the residuals are concentrated between 0 and 0.5, which is acceptable. We also find some large values, going all the way to 2. To have a representative measure, we check the mean value of the error, which is 0.4786. We can consider it is a good approximation, and it reduces the dimensionality to 4.

2.3 Cluster

The last method we are going to use is cluster analysis. The goal of cluster analysis is to identify groups of observations (or variables) that are similar to each other, where "similarity" between each pair of subjects means some global measure over the whole set of characteristics. To do the cluster assignment there are different algorithms, in this project we will use K-means and agglomerative hierarchical clustering. The output of clustering methods depends on whether we standardize the data or not, in our case we will standardize it to avoid giving higher weights to variables with units with large values.

2.3.1 K-means

The first method is k-means. It works in the following way: it initializes k cluster centroids (average of observations within a cluster) and assigns observations to clusters with the closest centroid. Then it finds new centroids by averaging data assigned to those clusters. It iterates until the cluster assignment doesn't change.

Starting from our standardized data, we will classify it using this algorithm and we will represent the PCA scores plot with the clustering assignment. First of all, we have to choose the number of clusters k. To do it we will use the total within sum of squares (TWSS).

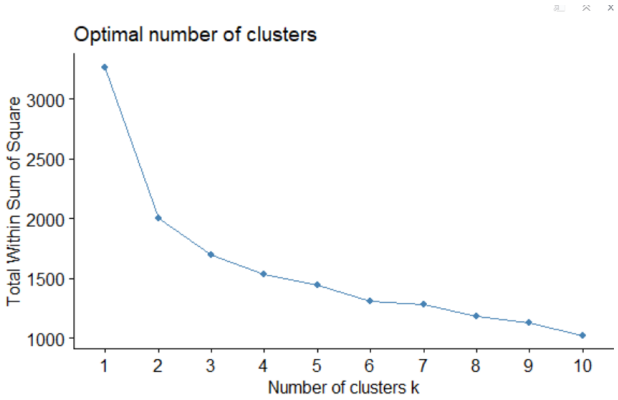

Figure 10: Plot of TWSS and the number of clusters k.

We track TWSS as a function of k and we choose the smallest value of k with a relatively small TWSS. Observing our function, we pick $k = 4$. Then we plot the cluster assignment:

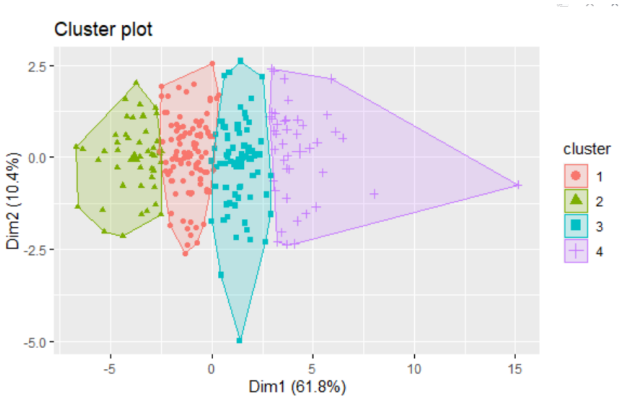

Figure 11: Cluster plot.

We can observe how with this value of k the clusters created are easily differentiable since there are no intersections.

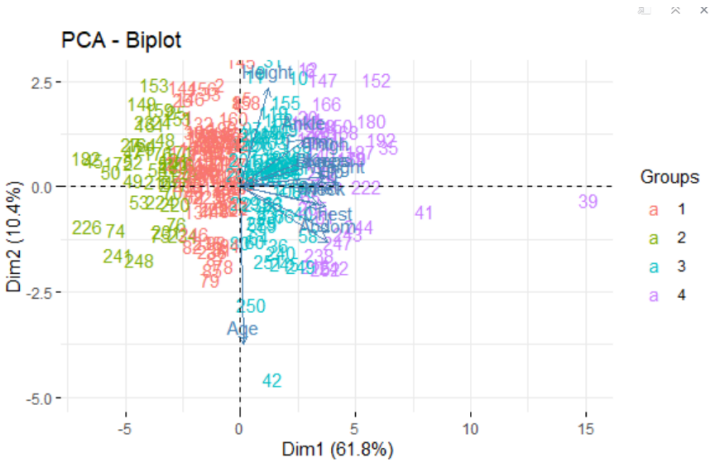

Figure 12: PCA biplot that includes the clustering assignment, PCA scores, and PCA loadings.

As we can see, we divided the data into 4 separate groups according to the values in PC1, two to the left of the sample and the other two to the right.

At first it makes some sense to group the individuals along the PC1 as it represents most of the variance. However, if we remember the plot with the true values in classification in PCA in figure 7, these clusters they are not alike at all. A lot of clusters would be a a mixture of individuals from different groups according to real values of body fat.

2.3.2 Agglomerative hierarchical clustering

The next cluster method we are going to use is Agglomerative hierarchical clustering. We start with each observation being its own cluster. Then, iteratively, we merge the two closest clusters. We'll define distance between clusters by looking at the pairs of distances between the observations in the clusters. There are four main ways of defining distances:

- Single linkage: distance between closest points

$$d(C1, C2) = \min 1 \leq i \leq n, 1 \leq j \leq m d(X_i, Y_j)$$

- Complete linkage: distance between points that are farthest apart

$$d(C1, C2) = \max 1 \leq i \leq n, 1 \leq j \leq m d(X_i, Y_j)$$

- Average linkage: average distance between points

$$d(C1, C2) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{d(X_i, Y_j)}{nm}$$

- Centroid linkage: distance between averages

$$d(C1, C2) = d(\hat{X}, \hat{Y})$$

As before, the first thing to do is choose the value of k and the do the cluster assignment:

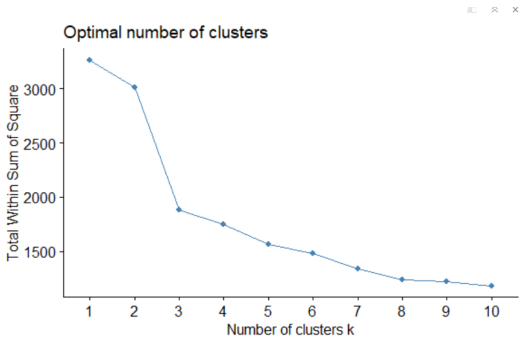

Figure 13: Plot of TWSS and the number of clusters k.

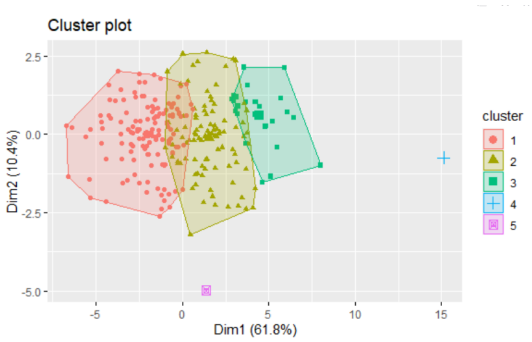

Figure 14: Cluster plot.

In this case we choose k=5 as we consider it is the smallest value of k with a relatively small TWSS.

We try single, complete, average and centroid linkage distances. After trying it for each method, we arrive to the conclusion that the best one is complete linkage as the size of the groups is more balanced and the groups are more differentiable.

Now we will represent the dendrogram, which is a tree representation of clustering. It allow us to see the distance between clusters (axis y) when merging.

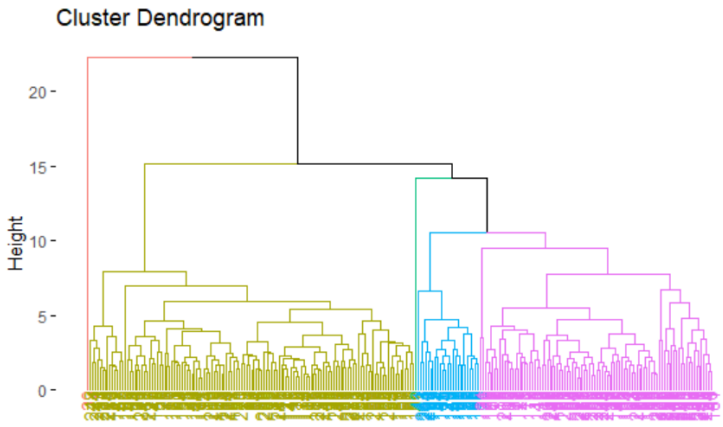

Figure 15: Dendogram.

As we can see the samples that are joined below are the most similar, so we can see how purple (yellow in the plot bellow) is similar to the blue (green bellow) which makes sense since they are near. The green one (purple below) and the red (blue below) are outliers, specially the red one since joins the rest on top of all.

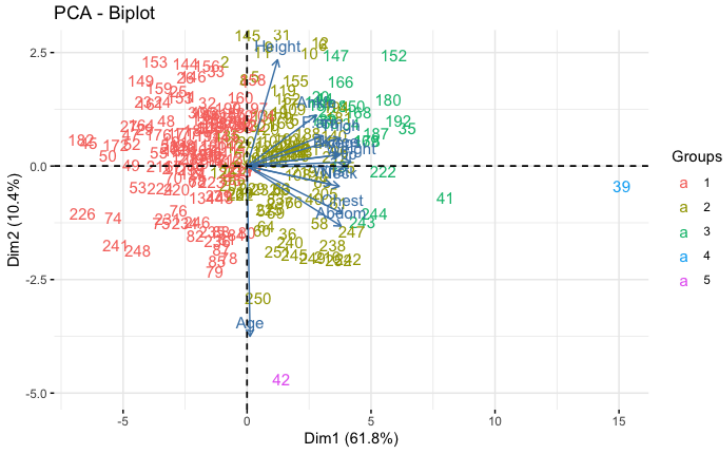

Figure 16: PCA biplot that includes the clustering assignment, PCA scores, and PCA loadings.

Again we can see how it divides most of the data in three groups along the PC1 and then it creates two clusters formed only by an outlier sample. Again it's not an accurate representation of the true classes as we have seen in figure 6.

3 Discussion and conclusions

Taking everything we have seen into account, we can say that our variables do not allow us to see properly how they affect the amount of body fat. With the PCA we found new variables that accounted for most of the variability. However, they did not help us much with the interpretation, as most of the variables have the same weight. Also, we found that MDS does a pretty good job when reducing dimensionality, and we managed to reduce our data to 4 dimensions with a mean error value of 0.479.

On the other hand, we discovered that if we try to separate our data into natural clusters, the groups created do not differentiate the level of body fat. It separates the samples into 5 groups, mostly divided by the PC1, but it does not correspond to the actual different classes we got. In every cluster there are observations with very different values of body fat, so these variables do not help us separate our target.

In addition, this project made us learn the importance of these statistical methods when analyzing a dataset, as it is essential to know how well do your variables explain the target. In this way, we will be aware of the limitations of our dataset, as well as possible improvements.

4 Bibliography

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/interpret-the-results/all-statistics-and-graphs/

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/dist

https://www.tibco.com/reference-center/what-is-cluster-analysis

Undoubtedly one of the main sources of information for accomplishing this project have been the presentations and notes given in class of this subject (AD).