Danmarks Tekniske Universitet

Course 02450

Project 1

Group 90

Aldís Helga Björgvinsdóttir

s212956@dtu.dk

Helga Þórey Björnsdóttir

s213615@dtu.dk

Sara Húnfjörð Jósepsdóttir

s212952@dtu.dk

Supervised by

Jes Frellsen

October 5, 2021

Contents

List of Figures

1 The Dataset

Our data contains a sample of males from Western Cape, South Africa, which is a high-risk region for heart diseases. The problem of interest is the coronary heart disease (CHD) and which attributes contribute to the disease. The dataset was obtained from the website: `https://web.stanford.edu/~hastie/ElemStatLearn/` which was found from the list of recommended websites of datasets provided by the professor. The data from this website is taken from another, larger dataset in: Rousseauw et al, 1983, South African Medical Journal [1]. The dataset contains 462 observations and 9 attributes.

The study that the dataset is obtained from is a three-community Coronary risk-factor study, carried out in rural South Africa. As the dataset is taken from a larger dataset which was used in the aforementioned journal, there was not a lot of information specifically regarding the interpretation of the data being analyzed in this report. In the study, the writers defined certain cut-off points for the risk of CHD and for obesity. An individual was considered a "smoker" if the level of cigarette consumption was 10 or more cigarettes a day. For the assessment of type A behaviour, the Bortner Short Rating Scale was used and was modified to yield 12 seven-point bipolar scales for characteristics such as competitiveness, impatience and striving [2]. The possible total score was 12 - 84, and those who fell in the upper two-fifths, that is 55 points or more, were arbitrarily classified as exhibiting type A behaviour [1]. The study revealed that many risk factors were exceedingly common, such as smoking, coronary-prone type A behaviour, obesity and family history. They concluded that the study communities were at extremely high risk of CHD where few individuals were completely risk-free with the majority of the study population having one or more risk factors. They revealed that excess risk of CHD in the community could be adequatly explained by the interaction of high levels of lifestyle-induced risk factors with constitutional predisposition. Interestingly, they also found a much lower rate of CHD in women in spite of similar risk factor levels, with the exception of smoking and the use of contraceptive pill [1]. Nonetheless, the data analyzed in this report only takes into account individuals of the male gender.

In the next report, both classification and regression will be performed on the dataset. For the classification, attributes will be chosen that make the desired group form clusters. The desired group being subjects suffering from coronary heart disease. For the regression it would be interesting to see the correlation between various attributes. We would like to predict age based on risk factors of CHD, such as tobacco, obesity, adiposity, alcohol consumption and sbp. To carry out these tasks and before the data is analyzed it needs to be standardized, this is done by subtracting the mean and divide by the standard deviation.

About the data and reference	Previous analysis summary	Classification and regression
Helga (s213615)	Sara (s212952)	Aldís (s212956)

Table 1: Student contribution for section 1.

2 The Data Attributes

Table 2: A description of the CHD data attributes and their corresponding type.

Attribute	**Description**	**Continuous/Discrete/Binary**	**Type**
Sbp	Systolic blood pressure	Continuous	Ratio
Tobacco	Cumulative tobacco consumtion (in kg)	Continuous	Ratio
ldl	Low-density lipoprotein cholesterol	Continuous	Ratio
adiposity	Adipose tissue concentration	Continuous	Ratio
famhist	Family history of heart disease	Binary	Nominal
typea	Score on test to measure type-A-behaviour	Discrete	Interval
obesity	Obesity	Continuous	Ratio
alcohol	Current alcohol consumption	Continuous	Ratio
age	Age of subject	Continuous	Ratio
chd	Coronary heart disease at baseline	Binary	Nominal

Family history of heart disease and CHD, attributes famhist and chd, were given the value present or absent and are therefore considered binary. As obesity, sbp, ldl, age and adiposity could theoretically be 0 they are considered as ratio. After observing the dataset no data issues were detected, there were no missing values or missing information.

Table 3: Summary Statistics.

	sbp	**tobacco**	**ldl**	**alcohol**	**age**
Mean	138.327	3.636	4.74	17.044	42.816
Std	20.474	4.588	2.0687	24.454	14.593
Median	134	2	4.34	7.51	45

From table 3 it can be seen that the values of the summary statistics vary in size and are quite different depending on the attributes. The attributes have different scales and units so they were standardized to make them easier to compare and interpret.

To get the correlation matrix the covariance matrix was standardized. The correlation matrix can be seen in table 4. The table shows that age has some correlation with most of the other attributes and could therefore be used in regression analysis as mentioned in section 1. Also, obesity and adipostity have a very strong correlation but these attributes describe a similar feature.

Table 4: Correlation matrix.

	sbp	**tobacco**	**ldl**	**adiposity**	**typea**	**obesity**	**alcohol**	**age**
sbp	1	0.21224652	0.15829633	0.35650008	-0.05745431	0.23806661	0.14009559	0.3887706
tobacco	0.21224652	1	0.15890546	0.28664037	-0.01460788	0.12452941	0.20081339	0.45033016
ldl	0.15829633	0.15890546	1	0.44043175	0.04404758	0.33050586	-0.0334034	0.31179923
adiposity	0.35650008	0.28664037	0.44043175	1	-0.04314364	0.71655625	0.10033013	0.62595442
typea	-0.05745431	-0.01460788	0.04404758	-0.04314364	1	0.0740061	0.03949794	-0.10260632
obesity	0.23806661	0.12452941	0.33050586	0.71655625	0.0740061	1	0.05161957	0.29177713
alcohol	0.14009559	0.20081339	-0.0334034	0.10033013	0.03949794	0.05161957	1	0.10112465
age	0.3887706	0.45033016	0.31179923	0.62595442	-0.10260632	0.29177713	0.10112465	1

Table 5: Student contribution for section 2.

Explanation of attributes	Data issues	Summary statistics
Sara (s212952)	Helga (s213615)	Aldís (s212956)

3 Data Visualization and Principal Component Analysis

3.1 Data Visualization

From figure 1 it can be seen that there are no apparent outliers in the data, this can be confirmed by looking at histograms of each attribute, see figure 2. All of the attributes seem to follow a normal distribution except for age which seems almost uniform. However, the histograms of tobacco and alcohol are extremely skewed, but seem to be following a normal or even poisson distribution.

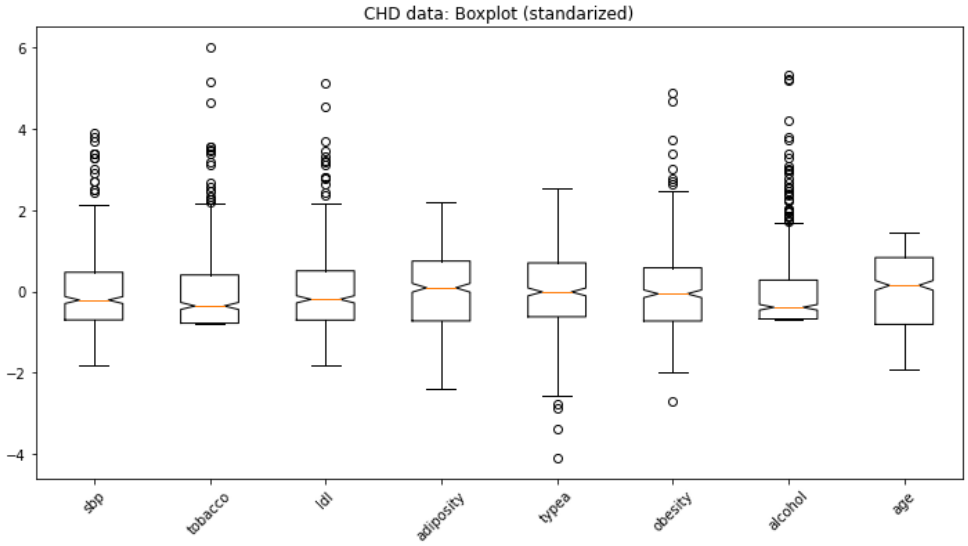

Figure 1: A boxplot of each attribute from the standardized data.

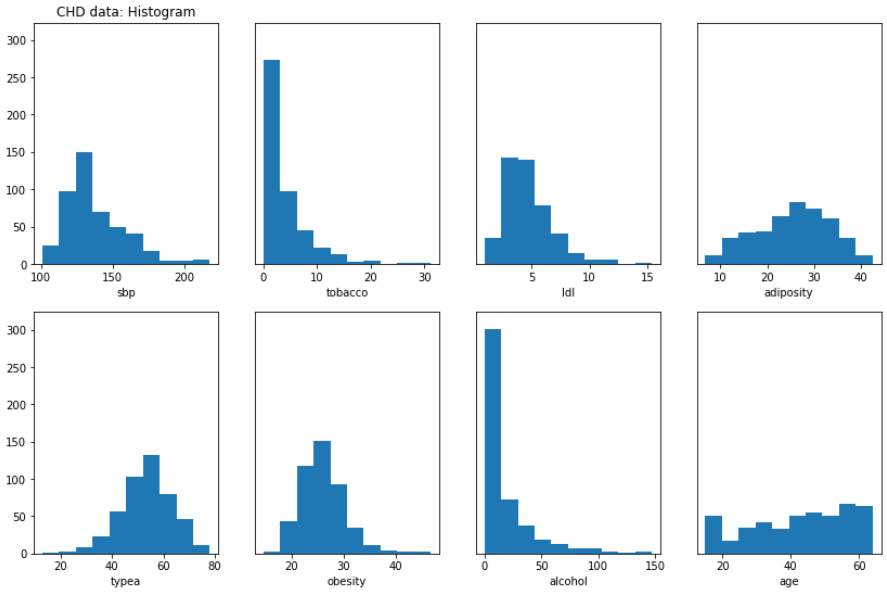

Figure 2: A histogram of all attributes in the CHD data.

Figure 3 shows a matrix plot of the correlation between each attribute, classified by presence of CHD (160 observations) and absence of CHD (302 observations). It can be seen that the adiposity and obesity attributes have the strongest correlation. Additionally, age seems to distinguish the most between the two classes, and could therefore prove useful in the classification task.

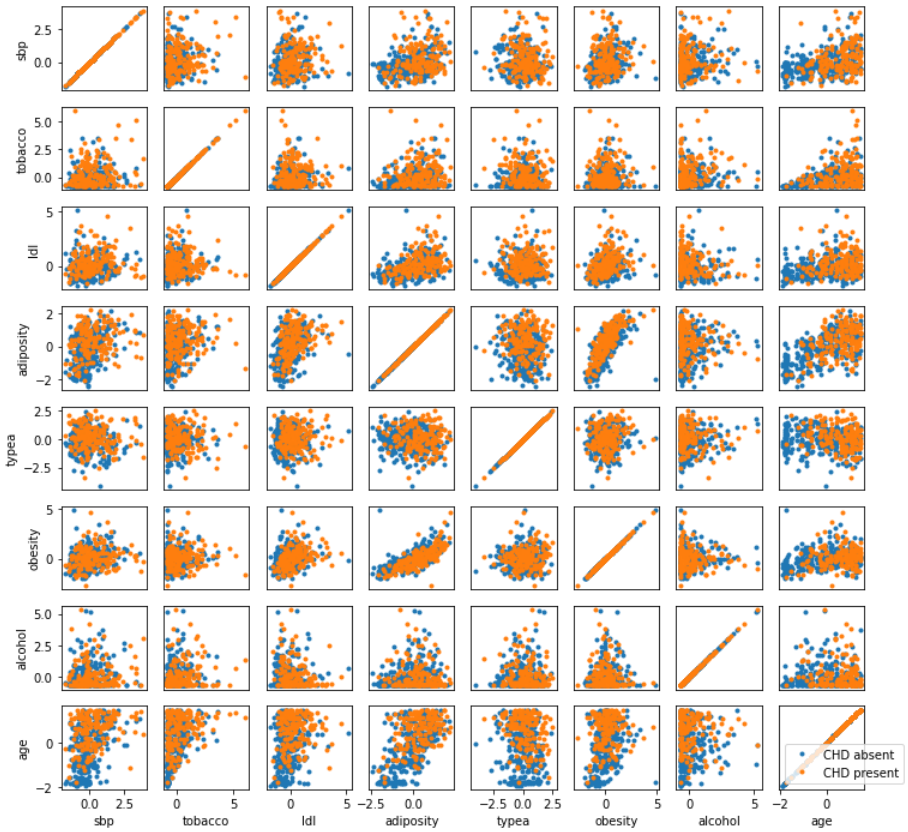

Figure 3: The correlation bewteen attributes for the classes CHD absent vs. present.

Figure 4 shows a clustered bar plot with the mean values of the standardized data for each attribute, classified by CHD absent vs. present. It can be seen that based on the mean values, the attributes sbp, tobacco and age distinguish the most between the absence and presence of CHD and could therefore prove useful in the classification task.

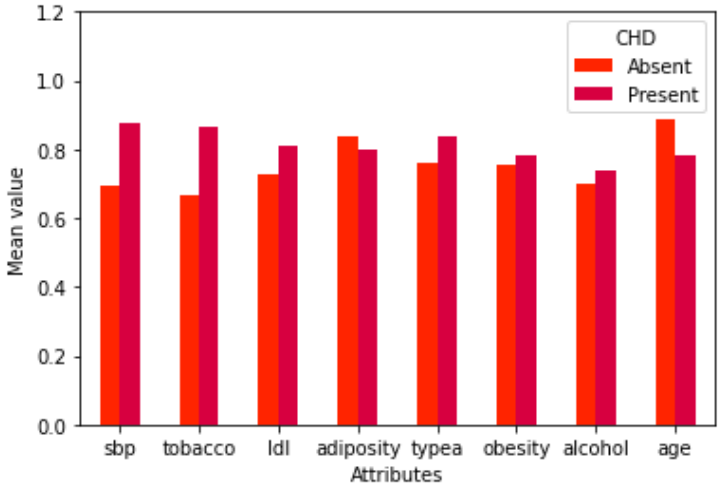

Figure 4: A bar plot of the mean value from the standardized data classified by CHD present and CHD absent.

3.2 Principal Component Analysis

Principal component analysis was done for better understanding and visualization of the data. Since the attributes of the data have different scales, the data was standardized before doing principal component analysis.

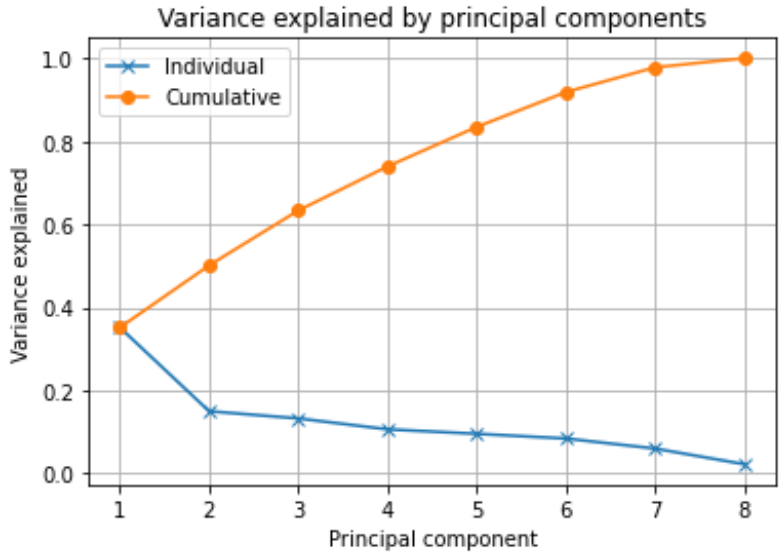

Figure 5: The amount of variance explained as a function of the number of PCA components included.

From figure 5 it can be seen that the slope describing the cumulative variance of the PCA components is almost linear. The slope for the individual variance flattens out after two principal components which would indicate that using the first two principal components would be sufficient but looking at the cumulative variance they only explain about 50% of the variance which is too small. Therefore it can be assumed that using five principal components would be efficient since they explain around 80% of the variance. For simplicity only the first two principal components were considered in the analysis made below.

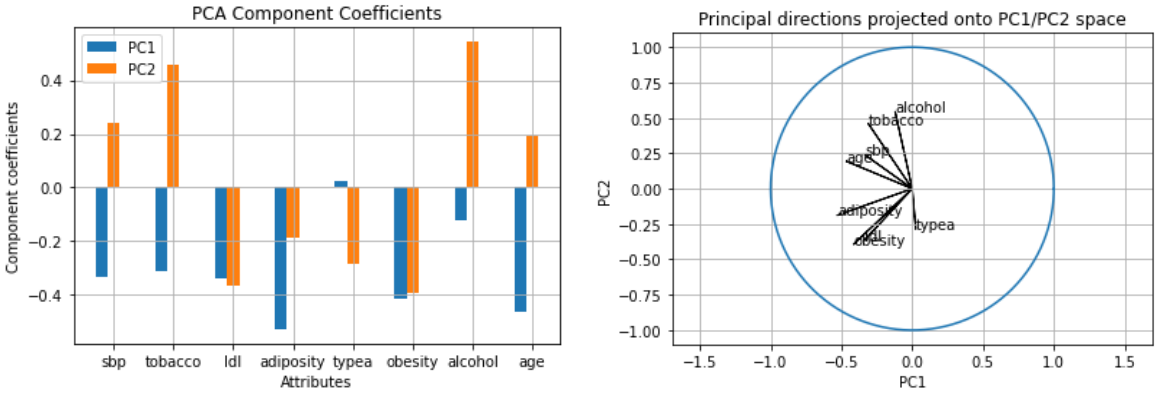

Figure 6: The principal directions of PC1 and PC2.

Table 6: The principal directions of the first two principal components.

	PC1	**PC2**
sbp	-0.3336685	0.2385342
tobacco	-0.3095857	0.45858007
ldl	-0.3371732	-0.3639169
adiposity	-0.5277529	-0.1874058
typea	0.02424838	-0.2826109
obesity	-0.4128036	-0.3917102
alcohol	-0.1203359	0.54281726
age	-0.4638294	0.19311399

The principal directions of PC1 and PC2 can be seen in the two plots in figure 6 and from table 3.2. In the left plot in figure 6 it can be seen that PC1 has almost only negative coefficients while PC2 has both negative and positive. The right plot is the direction coefficients plotted onto the PC1/PC2 plane in the unit circle. It can be observed that the alcohol and tobacco attributes are mostly described by PC2 in the positive direction while adiposity and age are mostly described by PC1 in the negative direction. This means that PC1 describes the least amount of the risk factors for CHD but PC2 has most weight for addiction such as alcohol and tobacco consumption.

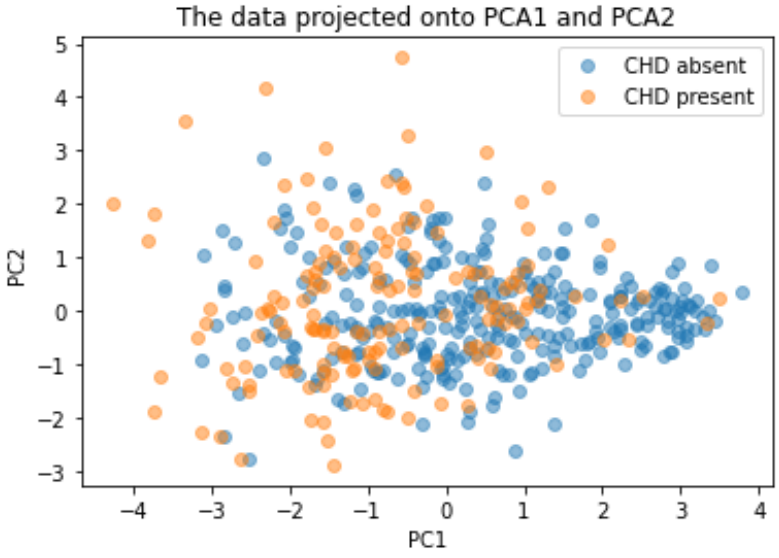

Figure 7: The data projected onto PC1 and PC2.

In figure 7 the data is projected onto PC1/PC2 space. Since the classes are not clustered it is hard to analyse the plot. However, it can be observed that subjects with CHD absent are mostly spread in the positive PC1 axis, indicating that subjects with absent CHD have low values of risk factors, which was expected.

For better description of the risk factors, such as obesity and ldl, more principal components are required.

Table 7: Student contribution for section 3.

Data Visualization	PCA
Sara (s212952)	Helga (s213615)

4 Discussion

Overall, the attributes are not very correlated, with adiposity and obesity having the strongest correlation. After analyzing the data it was decided that the attributes age, sbp, tobacco and type A were the best to use in the classification task of classifying subjects with CHD and without CHD. As age had the strongest correlations between attributes it seems achievable to do regression analysis to predict the age of a subject. The prediction can be done using the attributes that had the highest correlation coefficient for age: sbp, tobacco, ldl, and adiposity. From this it can be concluded that the machine learning aim appears to be feasable based on the visualization of the data and PCA.

Table 8: Student contribution for section 4.

Discussion
Aldís (s212956) & Helga (s213615) & Sara (s212952)

5 Problems

5.1 Question 1: Spring 2019 question 1

The correct answer is D. The reason is because Time of day, x_1, is an interval because it has physical meaning. As is given in the description in table 1 (see descriptsion for project 1), $x_1 = 1$ corresponds to 7:00-7:30 and $x_1 = 27$ corresponds to 20:00-20:30, which both have the same 30 minute differences. It is not ratio as the time of day cannot be 0. Running over, x_6, is ratio as it represents the number of run over accidents, so 0 means that there were no accidents. Lastly, Congestion level, y, is ordinal as the variable is ordered. It is based on the level of congestion which can be either high or low.

5.2 Question 2: Spring 2019 question 2

The correct answer is A. The answer was found using the following equation for the p-distance:

$$d_p(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_\infty = max\{|x_1 - y_1|, |x_2 - y_2|, ..., |x_n - y_n|\}$$

By subtracting $\mathbf{x}_{18}$ with $\mathbf{x}_{14}$ we get: $|\mathbf{x}_{14} - \mathbf{x}_{18}| = [7, 0, 2, 0, 0, 0, 0]^T$. From which the maximum component is 7.

5.3 Question 3: Spring 2019 question 3

The correct answer is A and is solved using eq. (3.18). Variance Explained $= \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$.
By inserting the eigenvalues from matrix $\mathbf{S}$ we obtain that the variance explained by the first four principal components is greater than 0.8.

5.4 Question 4: Spring 2019 question 4

The correct answer is D as the attributes Broken Truck, Accident Victim, and Defects have positive values on principal component 2, $\mathbf{v_2} = [-0.5, 0.23, 0.23, 0.09, 0.8]^T$. The Defects attribute has the most weight out of all the attributes, 0.8. Additionally, the Time of day attribute which has the second most weight and causes the projection to be negative is very low. Therefore the other three attributes all having positive values on principal components 2 and having high values have more impact.

5.5 Question 5: Spring 2019 question 14

The correct answer is A. It is found using the formula below for the Jaccard Similarity:

$$J(\mathbf{x}, \mathbf{y}) = \frac{f_{11}}{f_{11} + f_{10} + f_{01}}$$

Where:

f_{11}: Number of entries i where $x_i = 1$ and $y_i = 1$

f_{10}: Number of entries i where $x_i = 1$ and $y_i = 0$

f_{01}: Number of entries i where $x_i = 0$ and $y_i = 1$. For the two datasets s_1 and s_2 there are 2 common words, *the* and *words*, so $f_{11} = 2$. s_1 has 8 words total so there are 6 words in s_1 that are not in s_2, so $f_{10} = 6$. s_2 has 7 words total so there are 5 words in s_2 that are not in s_1, so $f_{01} = 5$. Using the equation above we get:

$$J(s_1, s_2) = \frac{2}{2 + 6 + 5} = 0.1538$$

5.6 Question 6: Spring 2019 question 27

The correct answer is B. It is found by summing the probabilities when $\hat{x_2} = 0$ and y=2. So we compute: $P(\hat{x_2} = 0|y = 2) = 0.81 + 0.03 = 0.84$

Questions	1 & 4	2 & 5	3 & 6
Student	Sara (s212952)	Helga (s213615)	Aldís (s212956)

Table 9: Student contribution for section 5.

References

[1] J. E. Rossouw et al., "Coronary risk factor screening in three rural communities. The CORIS baseline study," South Afr. Med. J. Suid-Afr. Tydskr. Vir Geneeskd., vol. 64, no. 12, pp. 430–436, Sep. 1983.

[2] "Edwardsetal1990b.pdf." [Online]. Available: `http://public.kenan-flagler.unc.edu/faculty/edwardsj/Edwardsetal1990b.pdf`. [Accessed: 30-Sep-2021]