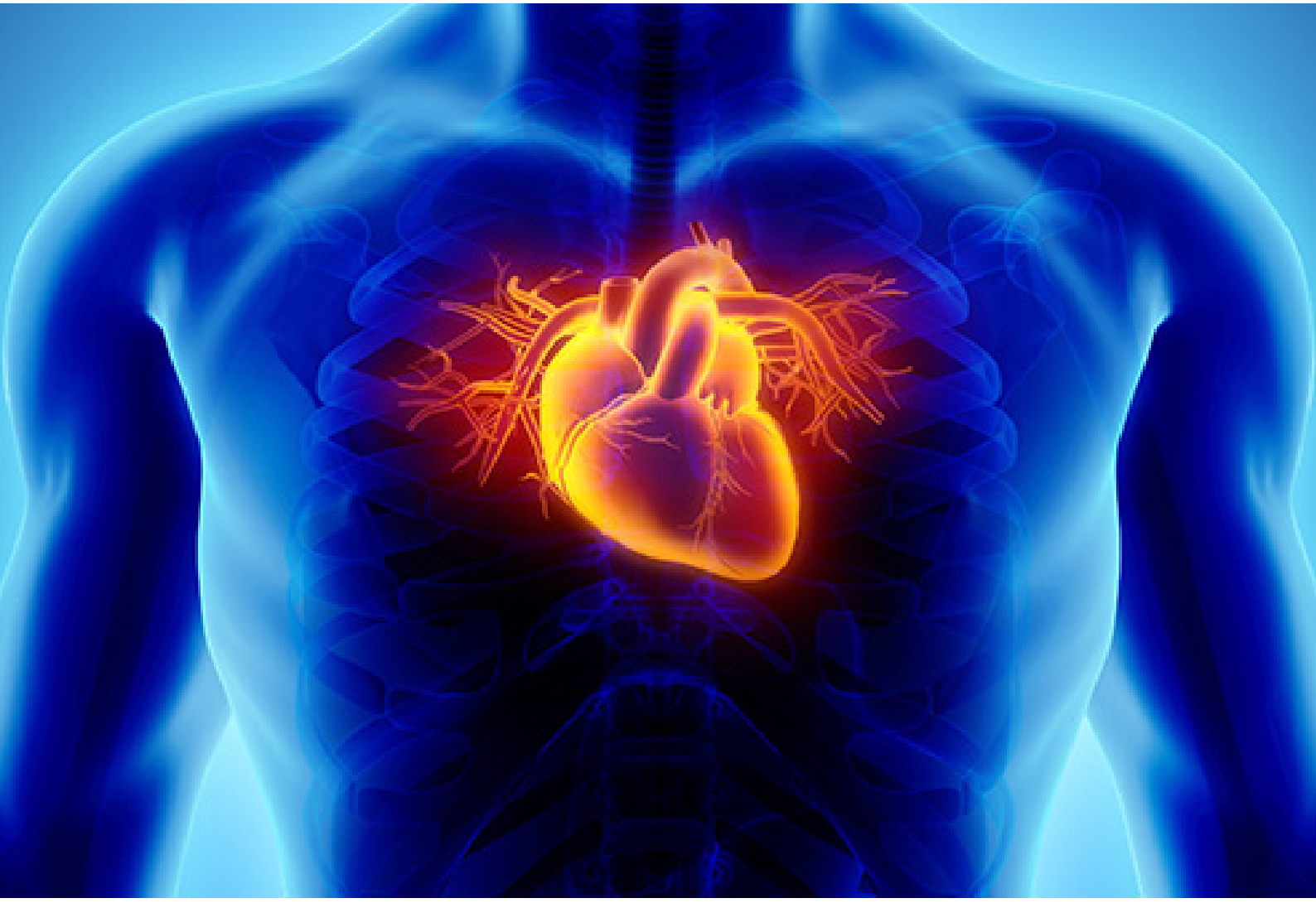
DTU Compute
Department of Applied Mathematics and Computer Science



**Group 70**
**Name & Student number:**
Krishna Harish Mohinani, s192182
Simon Winther Schor, s202817
Mikkel Stilling, s040038

**Course 02450**
**Introduction to Machine Learning & Data Mining**
Technical University of Denmark
9. March 2021

# Preface

This report is the first of two reports in a study, conducted during the spring of 2021 as part of the course "02450 - Introduction to Machine Learning & Data Mining" offered at the Technical University of Denmark (DTU).
The objective of this report is to apply methods learned during the course on a data set, giving a basic understanding of the data prior to further analysis. This further analysis is the objective of the second report. All students have contributed to the project equally.

Table 1: Student contributions to the report

| **Section** | **Authors** |
|---|---|
| 1. Description of data set | Simon/Krishna/Mikkel (30%/30%/40%) |
| 2. Explanation of data attributes | Simon/Krishna/Mikkel (40%/30%/30%) |
| 3. Data visualization | Simon/Krishna/Mikkel (30%/40%/30%) |
| 4. Discussion | Simon/Krishna/Mikkel (40%/30%/30%) |

# Contents

# 1. Description of data set

## 1.1 What the data is about

The data used for this study is a subset of the CORIS (Corinary risk factor screening) baseline study, consisting of results of a survey from three rural areas of the Western Cape, South Africa carried out in 1979. The CORIS baseline study set out to quantify coronary heart disease (chd) risk factors in this Afrikaans-speaking community [1].
The data set consists of a $463 \times 11$ matrix, containing measurements of 10 physiological and psychological features for 462 males in the region. These features are referred to as attributes and the males as observations throughout this report.

The problem of interest of this study is to predict chd risk based on knowledge of the other 9 features.

## 1.2 Data reference

The data was obtained from Stanford University. The reader may access the data following this link:
`https://web.stanford.edu/~hastie/ElemStatLearn/datasets/SAheart.data`

## 1.3 Previous analysis of the data

The original article on the CORIS baseline study, published in 1983, presents a number of findings [1]. Briefly summarized:

- *Cholesterol levels* rose with age, however after reaching a mean peak ($\pm$ SD) of $6.39 \pm 1.27$ mmol/l in males aged 45-54 years levels declined slightly by the ages of 55-64 years.
- *Mean systolic blood pressure* rose with age.
- *Smoking* was highest in the age groups 35-44 years (50.2% of males) and declined thereafter.
- *Mean BMI* rised from approximately 23 at 15-24 years to 28-29 at 55-64 years. In the higher age groups the mean BMI exceeded the 'overweight' cut-off points.
- *Type A behaviour* was present in about 40% of all subjects, and the peak prevalence was at age 25 - 34.
- *Type A behaviour* was less prevalent in younger subjects (15 - 24 years) and in those over 54 years.
- Overall 59,4% of the males had one or more of the major risk factors.

## 1.4 Context of problem of interest

As mentioned in section 1.1 the aim is to predict chd based on the remaining attributes of the data set thereby accomplishing binary classification. By using every available attribute the aim is to determine a set of Principal Components (PCs), which have a sufficiently covering level of variance explained.

To handle the data it was needed to remove the *row.names* attribute and transform the attribute values of *famhist* for family history heart disease cases from *present* and *absent* to *1* and *0* respectively.

Additionally, it could be relevant to assess correlation among the attributes, using regression, to be able to predict a response to a continuous parameter.

# 2. Explanation of data attributes

## 2.1 Description of attributes

The data set attributes are presented in table 2.

Table 2: Attributes of the data set

| **Attribute** | **Short Description** | **Attribute type** |
|---|---|---|
| row.names | Sample ID number. Unique for each observation. | Discrete, nominal |
| sbp | Systolic blood pressure (mmHg)* | Discrete, interval |
| tobacco | Cumulative tobacco use (kg) | Continuous, ratio |
| ldl | Low density lipoprotein cholesterol (mmol/L) | Continuous, interval |
| adiposity | Adiposity index | Continuous, interval |
| famhis | Family history of heart disease (0: Absent, 1: Present) | Binary, nominal |
| typea | Type A behavior (from 0-100%) | Discrete, interval |
| obesity | Body mass index (BMI) ($kg/m^2$) | Continuous, interval |
| alcohol | Current alcohol consumption (L/year) | Continuous, ratio |
| age | Age at onset (year)* | Discrete, ratio |
| chd | Response, coronary heart disease (0: No, 1: Yes) | Binary, nominal |

*Continuous feature, reduced to discrete variable in the source data set.

As most of the units of the attributes provided are not explicitly given, neither in the data set itself, nor in the source description of the data set, best guesses from the authors of this report have been made. Since the purpose of this report is learning to work with machine learning methods, and considering the fact that data handling performed is independent of attribute units, it was not found necessary to further check up on the validity of these best guesses of the units.

## 2.2 Data issues

It was not found necessary to remove any observations of the data set. While some attribute values of some observations were suspicious, they were still in the range of being physically possible and thus should not be removed or altered. For instance, one of the observations had a BMI of 46.58 and an adiposity of 9.74. This apparent inconsistency could be explained by the male being a muscular bodybuilder, and was thus included unaltered in the analysis.

As mentioned, the data set consists of 462 observations and 11 attributes, including the target which is aimed to be predicted, coronary heart disease (chd). The attributes are presented in table 2. While most of the attributes are self explanatory, some deserve a brief description:

*Adiposity* is a measure of a person's obesity *not* dependant on body weight, rather typically dependant on height and hip circumference [2]. In the study conducted for this data set, however, adiposity was determined by measurement of skin folds at biceps, triceps, with subscapular and supra-iliac with Harpenden skinfold calipers [1].

*Type A behaviour* people are ambitious, organised, impatient, punctual, tend to be irritable, are hard working and career oriented [3]. Type A indices of the observed people were determined by each respondent taking a self-administered personality test, following the Bortner Short Rating Scale for coronary-prone (type A) behaviour [1].

## 2.3 Summary statistics

To get an overview of the data set, brief summary statistics of the attributes were made. The results are shown in table 3.

Table 3: Summary statistics of attributes

| | **sbp** | **tobacco** | **ldl** | **adiposity** | **typea** | **obesity** | **alcohol** | **age** | **famhist** | **chd** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Mean** | 138.33 | 3.64 | 4.74 | 25.41 | 53.10 | 26.04 | 17.04 | 42.82 | 0.42 | 0.35 |
| **SD** | 20.50 | 4.59 | 2.07 | 7.78 | 9.82 | 4.21 | 24.48 | 14.61 | 0.49 | 0.48 |
| **Min** | 101.00 | 0.00 | 0.98 | 6.74 | 13.00 | 14.70 | 0.00 | 15.00 | 0 | 0 |
| **Q1** | 124.00 | 0.05 | 3.28 | 19.78 | 47.00 | 22.99 | 0.51 | 31.00 | 0 | 0 |
| **Median** | 134.00 | 2.00 | 4.34 | 26.12 | 53.00 | 25.81 | 7.51 | 45.00 | 0 | 0 |
| **Q3** | 148.00 | 5.50 | 5.79 | 31.23 | 60.00 | 28.50 | 23.89 | 55.00 | 1 | 1 |
| **Max** | 218.00 | 31.20 | 15.33 | 42.49 | 78.00 | 46.58 | 147.19 | 64.00 | 1 | 1 |

# 3. Data visualizations

## 3.1 Outliers

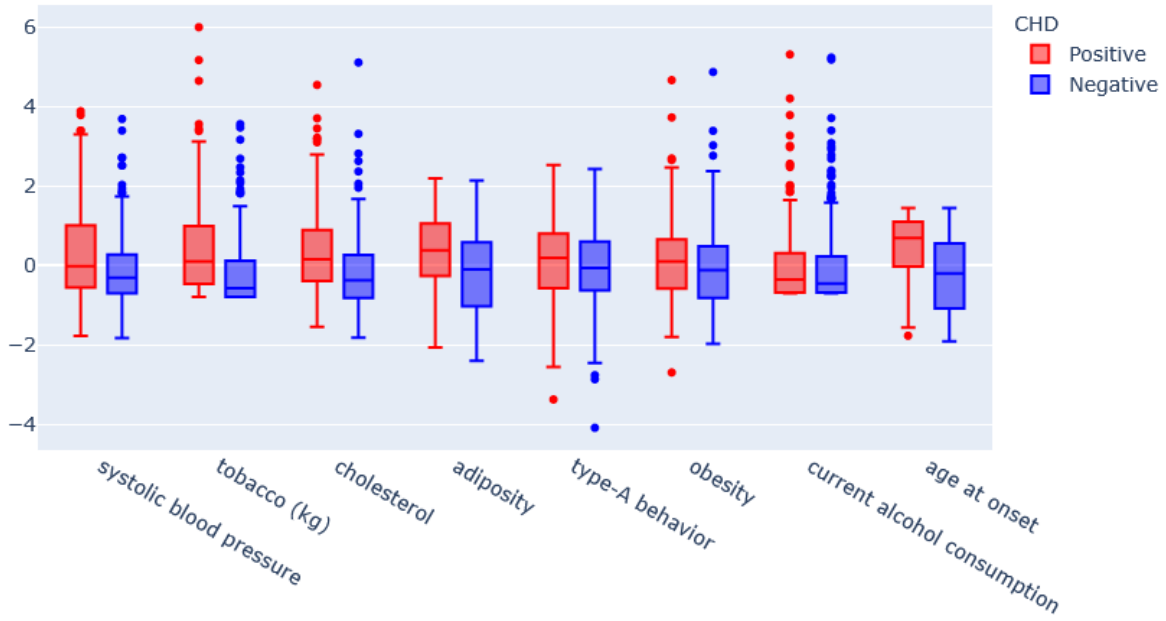



Figure 1: Box plot specifically highlighting the outliers in the data set

There is large variation in scales between attributes that are continuous and discrete. This is clearly shown in Table 3. For example comparing *sbp* & *tobacco* or *ldl* & *typea* highlights the difference in scales. This can be a setback because so much variation in data (scales) can be hard to visualize and may baffle the machine learning model. To overcome this setback the data has been standardized by subtracting the mean and dividing by the standard deviation. This is visualized in figure 1 where the box plots are grouped by the target variable *chd* for comprehensive understanding. It is crucial to mention that the Box plot shows outliers for some attributes. Before removing anything from the data set, it is very important to acknowledge that discarding outliers without clear reasons is a cause for concern since this may affect conclusions drawn from the data. In this specific data set the outliers seem to be logical and thus, again, nothing is discarded.

## 3.2 Distribution

Formal analysis of skewness was not performed for this study. The following comments on attribute distributions were done solely by visual inspection of histograms in figure 2 in supplement to summary statistics in table 3.

For *sbp*, the lower and upper quartiles are 124 and 148. The sample data is positively skewed.

For *tobacco*, the observations are clearly positively skewed, beginning from 0, intuitively having no negative values. There are a few outliers with large smoking habits. 12 out of 462 observations were above 15, and of these 3 were above 20.

The *ldl* data is positively skewed as well.

For *adiposity*, the observations are quite evenly distributed around the median of 26.12, with a mean of 25.41. The distribution of this attribute is considered normal.

For *typea*, the observations are slightly negatively skewed, showing a majority of type A personalities among the sample, not more so than considered a normal distribution.

For *obesity*, the upper and lower quartiles are approximately 23 and 28.5. The data is positively skewed, with larger extreme values of high BMI, than of low values. Even so, this is also considered a normal distribution, as there are just a few positive outliers. There are 15 observations of BMI of 30 or larger, of these 4 observations were 40 or larger.

For *alcohol*, the observation distribution is fairly similar to the *tobacco*. The observations are clearly positively skewed. The data set has some outliers with 31 observations of 60 or above, 15 of 80 or above and 7 of 100 or above.

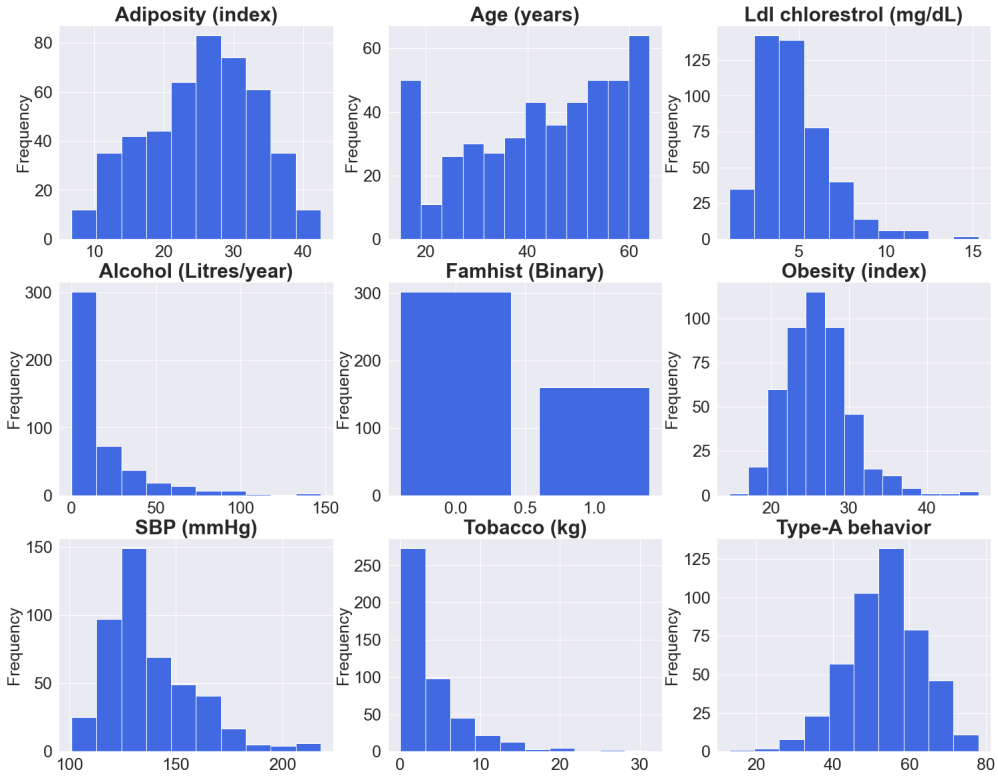



Figure 2: Histograms

Regarding *age*, the average age of the observed males in the data set was 42.8 years, with 50% lying within 31 and 55 years of age. The distribution neither appears to be normal nor equally distributed among all ages. There is a relatively low sample frequency among young adults. The sample frequency is somewhat steadily rising until the highest values around 60 years of age. However, an over-representation of young males below 20 years of age has a sample frequency which coincides with the age intervals around the upper 50 year-olds.

Lastly, concerning *chd*, about 41.5% of the observations have cases of heart disease in the family history and about 34.6% tested positive.

## 3.3 Correlation

In order to obtain an effective machine learning model, ideally the non-target attributes provided must both be good descriptors for the target which is aimed to be investigated, and be completely non-correlated to each other.

To assess correlation between the different attributes, a correlation matrix was constructed, each value in this correlation matrix calculated according to the definition of the empirical correlation of variables $x$ and $y$ [4]:

$$\hat{cor}[x, y] = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\hat{\sigma_x}\hat{\sigma_y}}$$

where $i$ is observation number, $N$ is the total number of observations, $\hat{\mu}_x$ and $\hat{\mu}_x$ are the empirical means of $x$ and $y$ respectively, and $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the empirical standard deviations of $x$ and $y$.

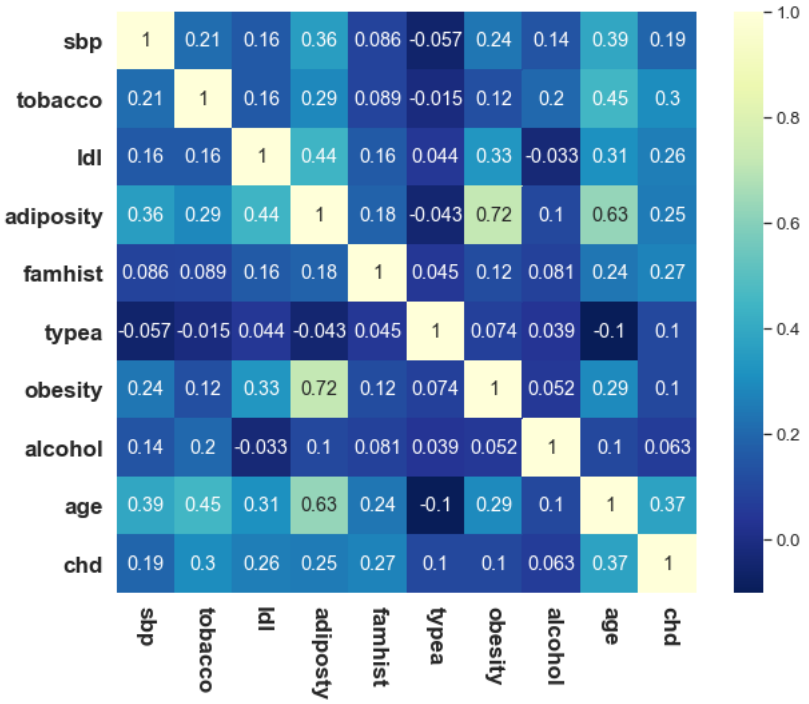



Figure 3: Correlation matrix

As indicated in figure 3, the most correlated attributes are *adiposity* & *age* and *adiposity* & *obesity*, with scores of 0.63 and 0.72 respectively.

Additionally, it is crucial to highlight *age*, *tobacco*, *ldl*, *adiposity* and *sbp* as the attributes which individually explain *chd* the most, with correlations of 0.37, 0.3, 0.26, 0.25, and 0.19 respectively.

## 3.4 Feasibility of modeling aim

Assessing whether the primary machine learning modelling aim of this study appears to be feasible, based on vizualisations of the data, several aspects were considered:

- Assessing the box plot in figure 1 (in supplement to summary statistics and manual inspection of the data), no observations of the data set were assessed to be in need of being removed or altered in any way.
- Assessing the histograms in figure 2 there's no apparent cause for concern that any attributes have distributions of a sort that are not easy to handle.
- Assessing the correlation matrix in figure 3, in general correlations between attribute-attribute pairs are non-zero and relatively low. Specifically, correlations between target (*chd*) and all other attributes are non-zero and with no single attribute being dominantly correlated to the target, which is nearly ideal for the modelling aim.

Given these considerations, the modelling aim appears feasible.

## 3.5 Principal component analysis

The goal is to prepare the data set, to be able to make sound classification of chd cases. For this, it is beneficial to make a linear representation of less dimensions, but still with a high level of variance explained, from the original high dimensional data set given. To do this Principal Component Analysis (PCA) is applied, creating reasonable input parameters for the machine learning model to be used. With PCA the data set is projected onto a set of principal components (PCs) which are orthonormal, hence not correlated to each other.

The algorithm to be followed is first to perform linear transformation of the data by subtracting the mean of each attribute from the observations, followed by division by the standard deviation. If attribute values and distributions are similar, this initial step can be passed. Then, to be able to compute the values of the set of eigenvectors, that is the PCs, Singular Value Decomposition is applied.

In the data set chosen for this report, it is clear that the scales of the values are markedly different. As illustrated in section 3.2, *sbp* values range approximately from 100 to 220, and the standard deviation (SD) listed in table 3 has a value of 20.5, while *typea* range is 0-100 and has SD 9.82. Further, *obesity* approximately ranges from 15 to 50 with an SD of 4.21. The standard deviations of the attributes used for the model differ from 2.07 to 24.48.

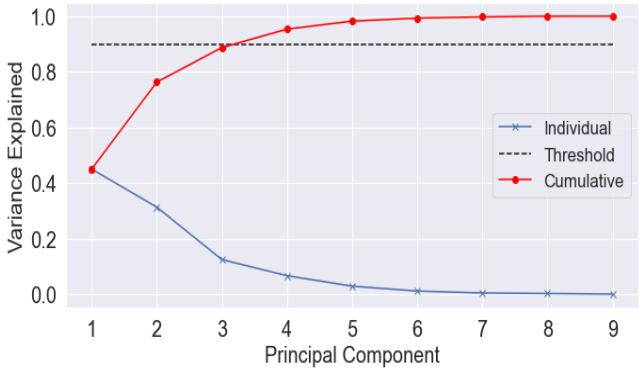
To investigate the consequences of leaving out the optional step in the PCA-algorithm, the PCA with a chosen threshold for variance explained of 90% was performed both excluding and including this step. The results are shown in the six panes of figure 4.

Inspecting pane 4a which represents the variance explained for the ordered principal directions from PCA1-PCA9, it is clear that the cumulative variance will exceed the chosen threshold by applying a model with just four PCs, as they account for 95.4% of the variance explained. However, recalling that this pane represents an assumption that the attributes in the data set are previously standardized by nature.

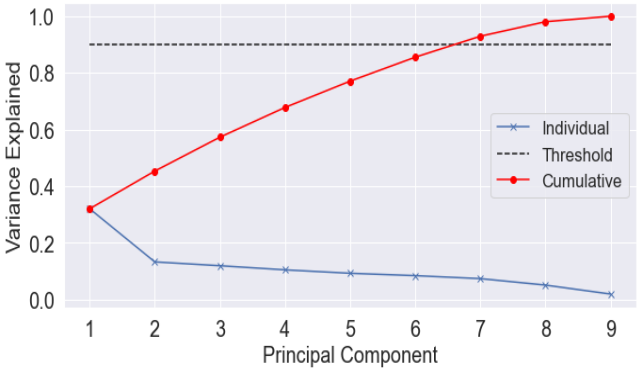
For this reason, the data samples projected by the attributes onto the PCA1 and PCA2-directions (which is shown in pane 4c) are inspected. From this, the distribution of projections is visibly distorted. Looking further, onto pane 4e, it is clear that the attributes *sbp*, *alcohol*, *age*, *typea* and *adiposity* drive the selected PCs. The standard deviation of these attributes are quite larger than those of the other non-target attributes. For this reason standardization of the data set is needed in order to create sufficiently predicting PCs.

Investigating pane 4d, where the optional step of dividing by the standard deviation has been taken, the distribution of projections onto PCA1 and PCA2 are less distorted. Additionally, while not being a text book example of separation of positive and negative target values, they seem to have slighty different centers of mass, with the positive chd-results (red dots) having slightly lower values on the x-axis (PCA1). Additionally, consulting pane 4f, contrary to pane 4e, all attributes included in the model seem to have an effect on the PCs.

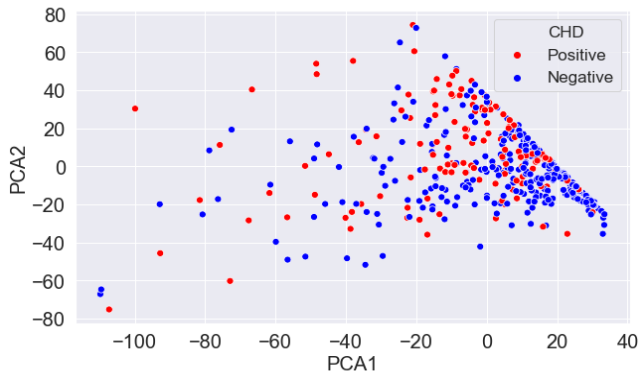
Assessing pane 4b, it is clear that the individual principal components are not as dominating as first concluded from pane 4a. Ultimately, PCA1-PCA7 are needed to exceed the threshold, as they account for 92.9% variance explained.



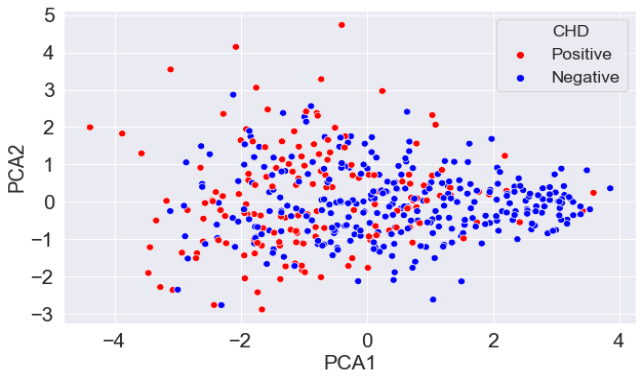

(a) Variance explained by Principal directions $Data - Mean$



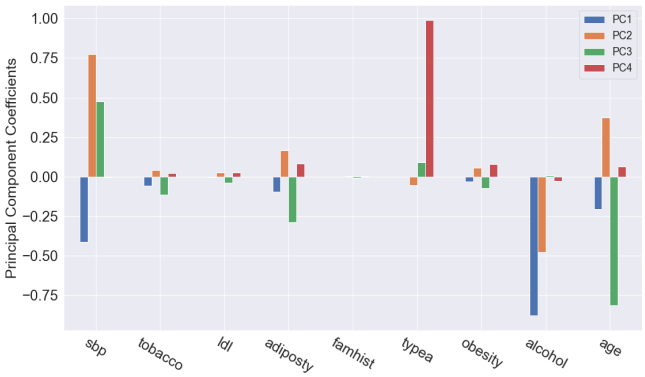

(b) Variance explained by Principal directions $(Data - Mean)/SD$



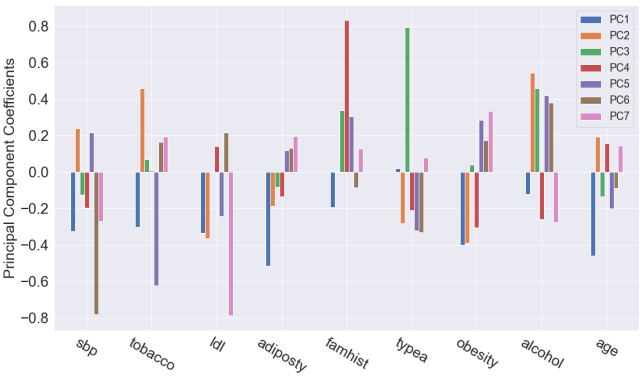

(c) Data projection & Principal directions $Data - Mean$




(d) Data projection & Principal directions $(Data - Mean)/SD$




(e) Principal Component coefficients $Data - Mean$




(f) Principal Component coefficients $(Data - Mean)/SD$

Figure 4: Plots from Principal Component Analyses. PCA without standardization (left panes) & PCA including standardization (right panes).

# 4. Discussion - Learnings about the data

The key objective of this study is to predict whether a patient is chd positive or negative. Thus, in terms of the data set, chd is the target attribute. The whole data set will be used, nothing will be discarded except the *row.names* attribute.

From the assessment of the correlation of the data, it could be seen that there is a strong correlation between *adiposity* and *obesity*. Discarding one of these in the further study, to reduce dimensions, is thus justifiable and can be considered an option.

Nevertheless, the data set consists of only two binary attributes *famhist* and *chd*, this is a binary classification problem since there are only two choices. Although, *chd* will be the target attribute and techniques like logistic regression or decision tress can be applied to predict a binary response. Also, the majority of data set consists of continuous variables, so it is possible to predict all of these attributes through regression. Nevertheless, the team will focus only on predicting *adiposity* because *adiposity* shows highest correlation with other attributes.

Studying the attributes further has shown that the varying forms of distributions - from vastly positively skewed to next to perfect normal distribution - make it clear that data centering is necessary. The markedly different standard deviations among the attributes shown in table 3 make it a necessity to standardize when performing the PCA. This, in order to make PCs of a precision which is sufficient to make sound and feasible predictions of chd cases, including the best 7 PCs in the model.

# References

[1] J. E. Roussouw et al. "Coronary risk factor screening in three rural communities". In: *South African Medical Journal 64: 430–436* (1983).

[2] D. S. Freedman et al. *The body adiposity index (hip circumference ÷ height$^{1.5}$) is not a more accurate measure of adiposity than is BMI, waist circumference, or hip circumference.* `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3477292/`
Visited on 02-03-2021. 2012.

[3] Rayman W. Bortner. *A short rating scale as a potential measure of pattern A behavior.* `https:/https://www.sciencedirect.com/sdfe/pdf/download/eid/1-s2.0-0021968169900617/first-page-pdf`
Visited on 08-03-2021. 1969.

[4] Tue Herlau et al. *Introduction to Machine Learning and Data Mining.* 2021.