Project 1

Introduction to machine learning and data mining

Søren Meyer Nielsen	s173932
Peixuan Wang	s192176
Zixuan Li	s192169

Fall, 2019

02450

Department of Applied Mathematics and
Computer Science

1 Introduction to data set

The data we will be using for this report is a data set that contains medical data from males in western cape, South Africa. The data set is taken from *Standford University* and the data set can be found at https://web.stanford.edu/~hastie/ElemStatLearn/. The data is a sub data set taken from a larger data set, including 463 observations and 10 attributes. The attributes are focused on medical factors.

2 Data attributes

Below is a table that descibes the attributes aswell as giving a short description and the classification. From table you you will see that we will mostly work with continous ratios but a few discrete nominals aswell; familiy history of heart disease and if the person have had any heart disease. These are presented with 1/0 vaules, a binary variable.

Table 1: Introduction of different attributes

Attribute number	Name	Description	Classification
0	row.names	Observation ID	Discrete Nominal
1	sbp	Systolic blood pressure	Continues Ratio
2	tobacco	Cumulative tobacco consumption (kg)	Continues Ratio
3	ldl	Low density lipoprotein cholesterol	Continues Ratio
4	adiposity	Messurement of bodyfat %	Continues Ratio
5	famhist	Family history of heart disease (Present, Absent)	Discrete Nominal
6	typea	Type-A behavior	Discrete Interval
7	obesity	Body mass index	Continues Interval
8	alchohol	current alcohol cinsumption	Continues Ratio
9	age	age	Continues Ratio
10	chd	response, coronary heart disease	Discrete Nominal

While working with this dataset, we have made some changes to the data set. We have left out *row.names*, which is the index label of observations. It going from 1 up tp 426, and it has no effect on the persons health and therefore left out. The attribute *famhist* was in the data set presentet as a Present/Absent value. Because we wanted to work with *one-out-off-K coding* we changed this to a 1/0 value, where 1 mean present and 0 means absent. like stated earlier this data set is sub data set from a larger medical journal.

While the data set has no corrupt data or missing values, there are some complications with the data. Stanford University states:*"**Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments.**"*. This can make the data difficult to model with, details will be discussed further in the discussion.

Table 2: Statistics of attributes

Attribute	Mean	Median	Standard Deviation	number of outlier	max	min
sbp	138.32	134	20.47	15	218	101
tobacco	3.63	2	4.58	19	31.2	0
ldl	4.74	4.34	2.07	14	15.33	0.98
adiposity	25.40	26.11	7.77	0	42.5	6.74
typea	53.10	53	9.80	4	78	13
obesity	26.04	25.805	4.21	9	46.58	14.7
alcohol	17.04	7.51	24.45	33	147.19	0
age	42.81	45	14.59	0	6 4	15

we see that a lot of the variables have some outlier while others have none. Age has no outliers which is good because it means that we operate whithin a certian ratio. If some of the people were either much older or much younger would cause problems with the model, especially when age is so important when trying to determine whether a person will develope heart disease.
Something that is interresting to note is that tobacco and alcohol have alot of outliers, but typea, which is the last variable that behavior related has only 4.

The primary machine learning modelling aim to be used upon this dataset is classification tasks, as we are interested in classifying whether a male has heart disease or not.But we can also do some other tasks just like:
Regression:All the continuous attributes can do the regression analyze. For instance, we can predict ldl of a male based on his value of adiposity and obesity; Or predict sbp of a male based on ldl,age,adiposity and obesity.
Classification: Family history and chd can be predicted in classification task because they are discrete. Furthermore, these two attributes are binary so there will be only two classes at the output of the model.
Clustering: Without labels, so we need to analyze the similarity of all input data and decided how many groups we want. It will be useful to do some visualization work to figure out which number of groups is the best in this clustering task.

3 Data visualization

In this section,we are not only interested of the distribution of a single attribute,but also the relationships between the attributes.We only analyze the numerical attributes,so the famhist and chd are excluded.And at last,we will do a simple analysis of the correlation between famhist and chd.

We are aware that a PCA is sensitive to the outliers in attributes,so we give a look of the outliers in our dataset.

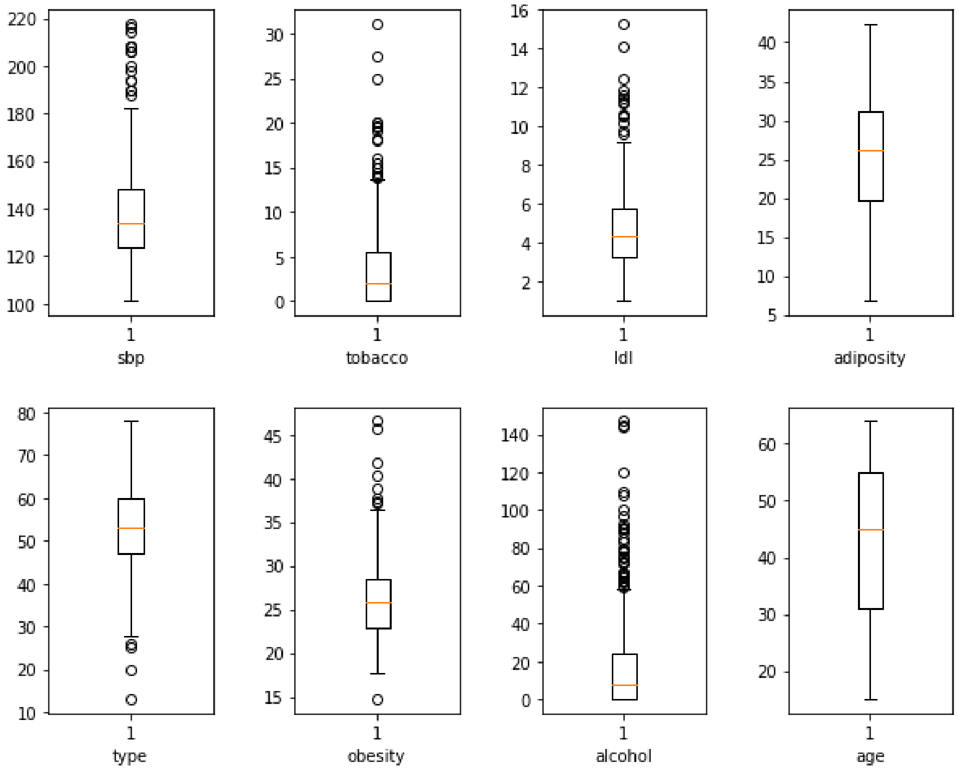

Figure 1: Box plots of numerical attributes

In Figure 1,bixplots of eight numerical attributes have been created. Six attributes have outliers as you can see from the box plots and the number of them can see in Table 2.These outliers affact the variances of each attributes so it may become a factor of driving the PCA compoments to wrong directions.

We are also interested in seeing the histogram distribution of individual variables.We have plotted in Figure 2.

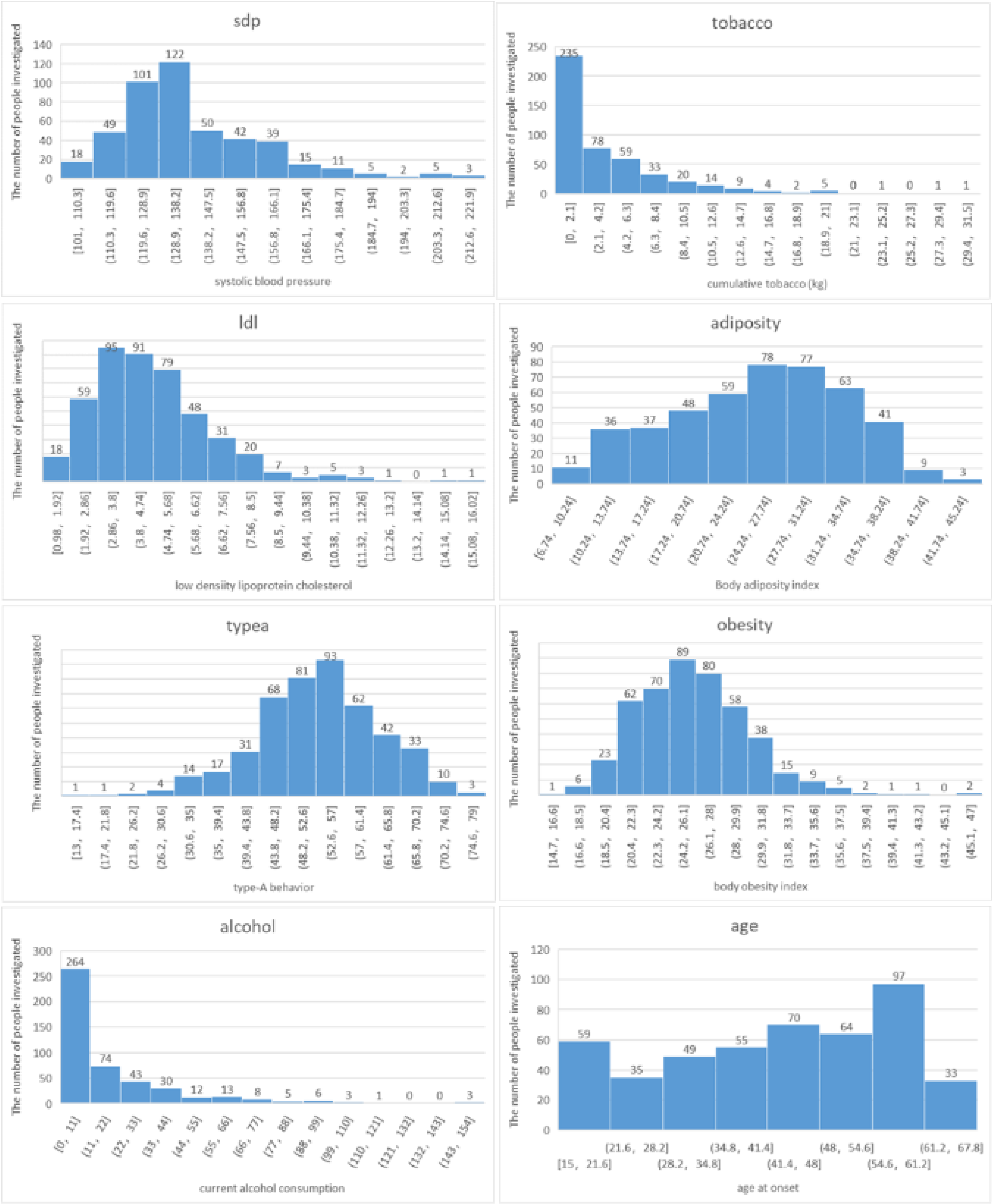

Figure 2: Distribution of attributes

From the histgrams,we know that the sbp, ldl, adiposity, typea and obesity are approximate normal distribution.Tobacco and alcohol are approximate exponential distribution.

Moreover,we are going to pair-wise correlation between the attributes.Represented in Figure 3, We conclude that only one pair: adiposity and obesity have high correlation,while other pairs of attributes have little correlation.It means that PCA will not as useful(since high correlation between some attributes meaning that such two attributes can be well-represented by one principal component.If there are little relationships between attributes,we can not reduce the complexity(dimension) of dataset by PCA). One thing we also can see is that most of the people with heart disease are older men, and that almost none of the younger boys had heart disease.

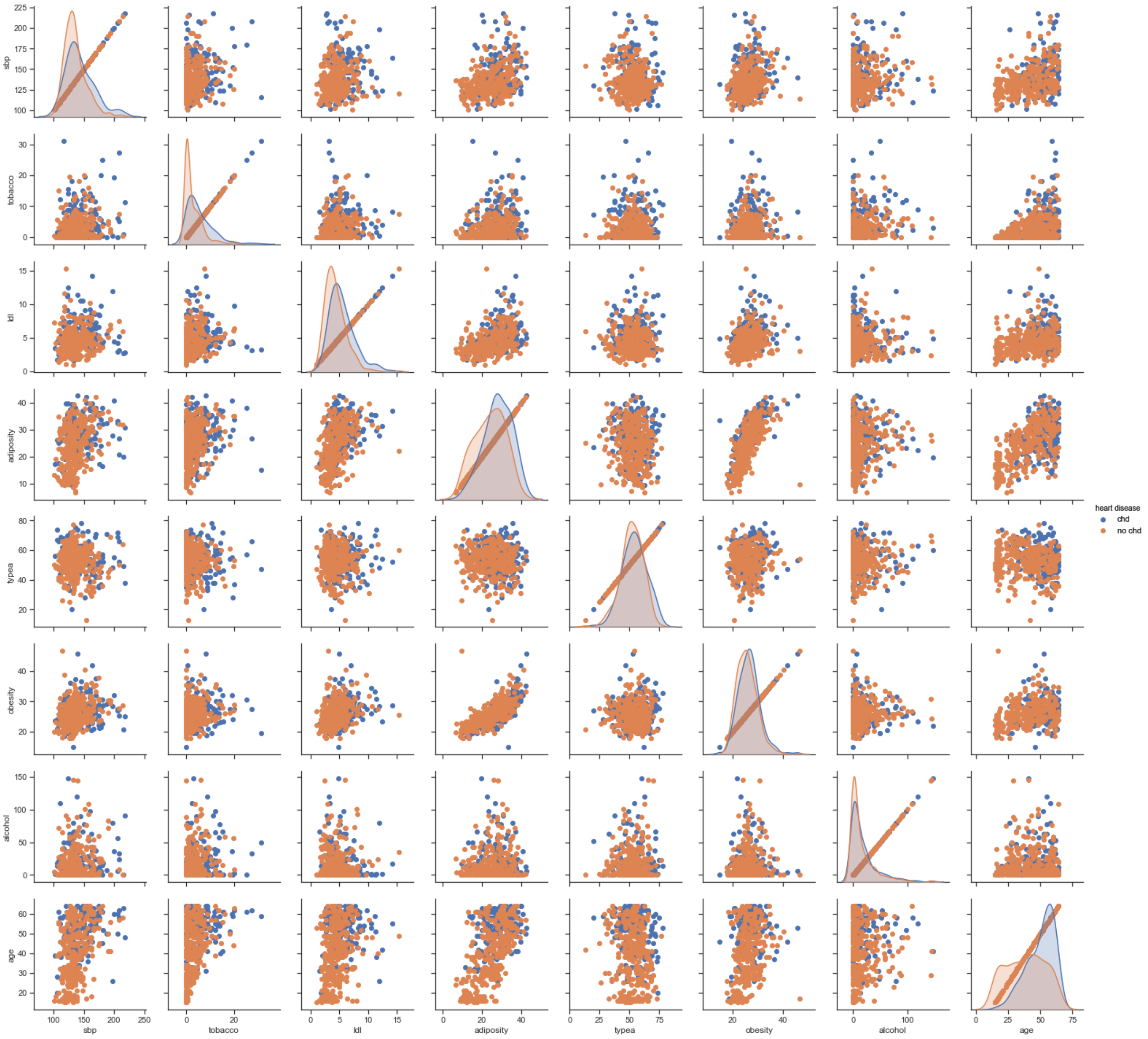

Figure 3: pair-wise correlation

Before we going to the PCA analysis,we focus on find out whether we need scaling of attributes.Figure 4 shows the standard deviations of attributes.

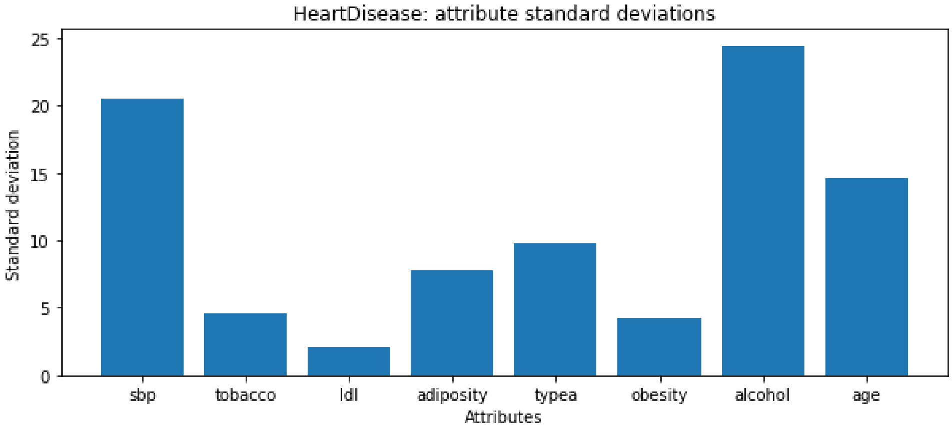

Figure 4: Attributes standard deviations

It is obvious that the standard deviations of variables differ largly,if we directly use the original data,the PCA components will be dominated by the attributes with large magnitudes so the result must be biased.So standardizing the dataset before performing PCA is quite important.

At last,we have done a similiarity analyze between famhist and heart disease(chd),showed in Figure 5.According to the plot,we can say that people who have family history of heart disease are more likely to get this disease.

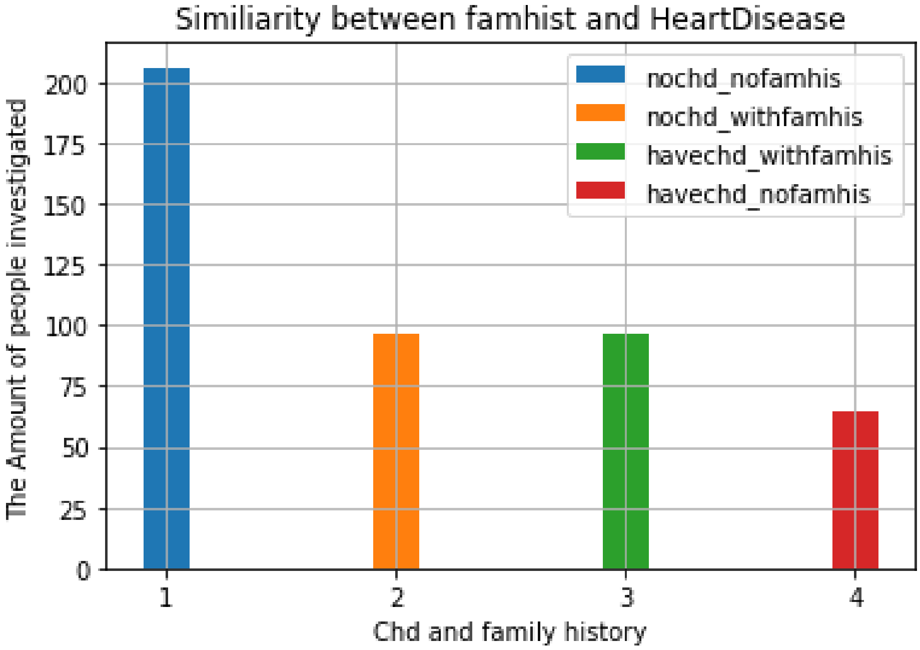

Figure 5: The correlation between famhist and chd

4 PCA Analysis

Generally speaking, variance can show how much information included about a dataset. Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.[1]

The principle components we are looking for are a couple of vectors which are the base victors of projection planes. These projection planes can be ordered by variances and covariances of the data. The first principle component means it points at the biggest variance direction which will cover the most information when we do the projection. Then we will go to the next axis and repeat this process n times until n equals to the number of attributes. While we project our dataset onto these planes, which called data reduction, we will get more visualized and more informatic data distribution maps.

We will use the PCA method to calculate the principle components of our dataset. The threshold we set is 90%.

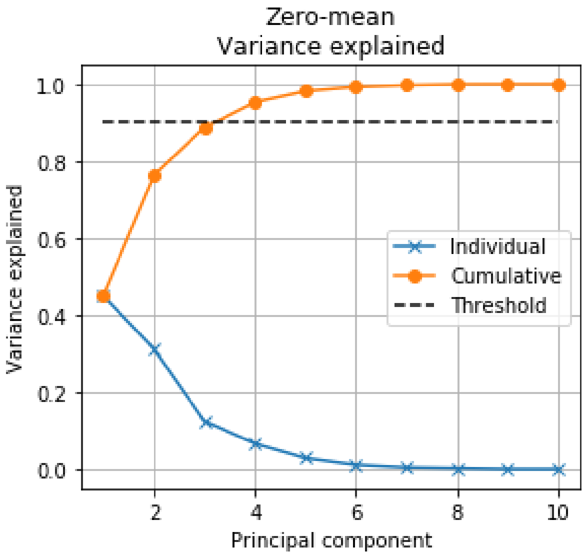

Figure 6: Principle Components of raw data

Totally, we have 10 variations, and according to figure 6, we can see that 4 principle components are able to explain 90% of the original data. But after a quick glance at our dataset, we found that our data includes different scales, it means we should make our data standardized.

[1] https://en.wikipedia.org/wiki/Principal_component_analysis

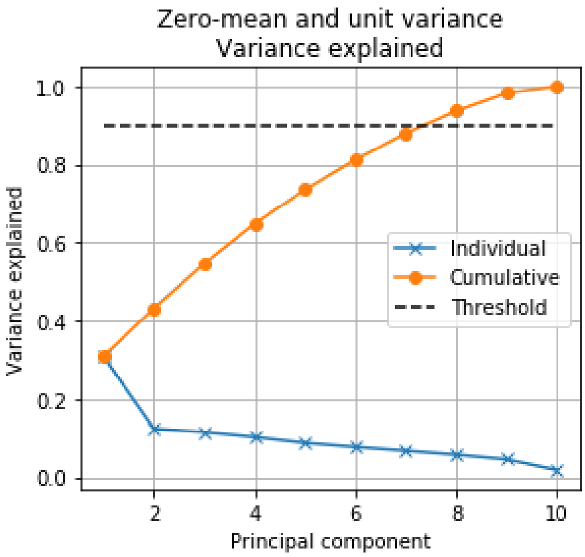

Figure 7: PCA after data standardization

From Figure 7, we can conclude we have to use over 7 principle components to explain 10 variances to avoid information-loss. The first principle component does not make more contribution than others significantly. In fact, they look like a linear function. And the data map projected in to the PC1-PC2 plane is shown on the figure below.

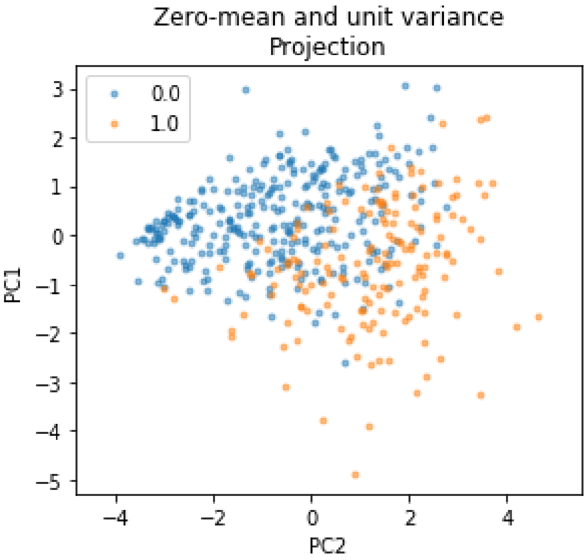

Figure 8: Data projected into PC1-PC2 subspace

The points represented in Figure 8 are not classified respectively, with an apparent overlapped area. After data standardization, each one in the dataset has been set into the same scale. Then, the principle directions of the considered PCA components are plotted inside a unit circle, shown in Figure 9.

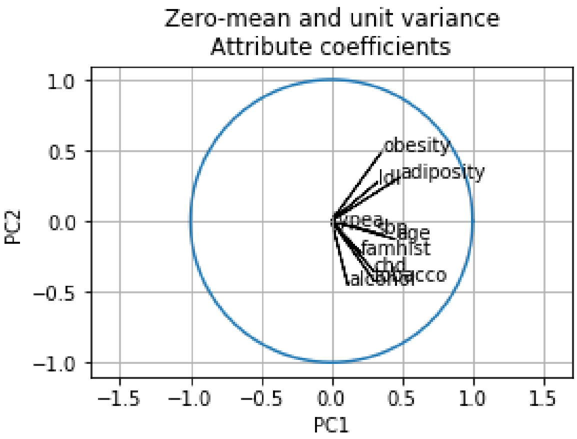

Figure 9: Principle directions of PCA

We can also make a brief view about the coefficient between our attributes and principle components by showing up each eigenvalue linked with different attributes (Figure 10). As family history and chd are binary attributes, there are no values here. From the figure, we find that it seems like every attribute can be projected to a certain principle component, other components only make a little influence on it. In other word, these attributes have low correlation between each other, which lead to this situation. PCA will find the biggest directions automatically in order to store the most amount of information, then it will come to an isolated attribute and products a component.

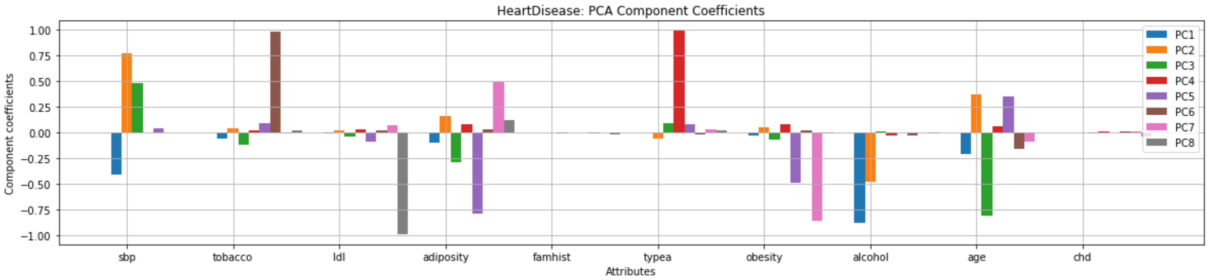

Figure 10: PCA Component coefficients

5 Discussion

The primary purpose we except is to classify if a person has heart disease by other attributions.Except doing the classification task, we can also try to use our dataset into regression models to quantificate some medical intuition.
As stated earlier, some of the men had undergone treatment while other haven't and some of the measurements were made after treatment. From plotting the attributes against each other and from looking at the results of PCA, we can see that there are low correlations of the variables between each other. Only adiposity and obesity seemes to have visible correlation.

At least 8 PCs are needed if the threshold is 90%. We could only explain less than 50% of the data by projecting the data into the first principle plane. It is inefficient to reduce dimensions of our dataset using PC analysis. This complicates the data and can be why it is so hard to find the correlation between the data.

Contribution specification

All group members have contributed to all sections, but the primary responsible is listed in a table below.

Section	Primary responsibility
1. Introduction to data set	Søren Meyer Nielsen
2. Data attributes	Søren Meyer Nielsen
3. Data visualization	Peixuan Wang
4. PCA Analysis	Zixuan Li
5. Discussion	All