Project 1

02450 Introduction to Machine Learning and Data Mining

Danmarks Tekniske Universitet

Group 4

	Section contribution percentage			
Name	Section 1	Section 2	Section 3	Section 4
Astrid Mantzius Jepsen (s214569)	60%	30%	10%	30%
Hoang Anh Do (s250220)	10%	60%	30%	30%
Bálint Kostyál (s250222)	30%	10%	60%	40%

March 5, 2025

Contents

1 Description of the data

In this project, data about males from the Western Cape, South Africa, will be analyzed. The observations of the males are obtained from [1], which are extracted from a larger dataset described by Rossouw et al. in 1983 in the South African Medical Journal [2].
The males are from a specific region of the Western Cape, where they are at high risk of getting sick from heart disease due to different reasons such as smoking, high BMI, or a history of coronary heart disease (CHD) in their families. The data have 10 attributes (besides the observation numbers):

Attribute Name	**Description**
Systolic Blood Pressure (sbp)	The pressure of the blood are measured in millimeters of mercury (mm Hg).
Tobacco	The amount of tobacco consumed in the lifetime of the male measured in kilograms.
Low-density lipoprotein cholesterol (ldl)	Also called cholesterol levels, is measured in mmol/l. The risk of coronary heart disease increases after 5.7 mmol/l, while high levels are diagnosed at 6.5 mmol/l [2].
Adiposity	A measurement related to weight, fat percent, and height, similar to BMI. The unit is unknown.
Family history	If someone in the family of the male has been diagnosed with coronary heart disease: 1 = yes, 0 = no.
Type-A	A score based on the test which tests the typical type-A behavior patterns such as impatience, competitiveness, and striving. Males with $\geq$ 55 points are classified as having Type-A behavior [2].
Obesity (BMI)	Described by BMI, which is a correlation between the weight and height of the males. Males with BMI $\geq 30\,\mathrm{kg/m^2}$ are classified as obese.
Alcohol	The unit is unknown and a unit such as liters pr. time unit or gallon.
Age	The age of the males in years.
Coronary heart disease (CHD)	If the males have been diagnosed with CHD before the data has been made: 1 = yes, 0 = no.

Table 1: List of attributes and their descriptions

The dataset includes males there has not been sick with coronary heart disease and males who have. Many of the males suffering from CHD have gone through treatments to reduce the risk of additional CHD attacks before the measurements made in the dataset [1]. The ratio of diagnosed to non-diagnosed males is approximately 2:1, meaning there are two controls for every male diagnosed with CHD [1].

In [2], Rossouw et al. compare minor and major risk factors for coronary heart diseases for both males and females in the age range 15 to 64 years. In addition to those listed above, they have other attributes such as other types of pressures and they categorized the tobacco use in smokers, cigarette smokers, and past cigarette smokers. In each gender, male and female, they separated the observations in groups according to their age such that observations aged 15-24 years were a group, and those aged 25-34 years were the second group and up to 64 years. In each age group, they found the means and standard deviations to

compare the behavior of the mean values when the people became older.
They did, for instance, find that the cholesterol level reached its maximum at 6.39±1.27 mmol/l for males in the age group 45-54 years. For females, the maximum was higher at 7.16±1.30 mml/l in the oldest age group (55-64 years).
In another table, they presented the percentage of the observed people in each age and gender group who excited the border value for each of the maker risk factors: BMI, blood pressure, cholesterol and smoking. They had set two limits for each of the factors, one string and one lower, but still significant. Exciting the lower limit often increases the risks of CHD, but not as much when exciting the highest limit. Rossouw et al. also counted the percentage of observers who exceeded multiple risk borders. The results from these observations were, for example, that 70% of the subjects over 44 years old had exceeded at least one of the risk factor limits.
Lastly, they summarized the ratio of males and females who had at least 1 or more risk factors in a histogram. Where the number of combined risk factors was at the x-axis and the ratio was up at the y-axis. At each number, two pillars describing the percentage of females and males, respectively, were drawn. The histogram gave a good overview of how much has met 1 or more of the risk factors, but you were not able to see the type combinations, and if some combinations of the risk factors are more common than others.

With this data set, we want to investigate the attributes and if there is some relation between them and the risk of being diagnosed with coronary heart disease. As a prediction based on regression, we want to predict the type-A points of the males based on their blood pressure, cholesterol number, obesity, smoking habits, and age of the males. By that, we can see how these attributes are related to their personality type.

A classification prediction we want to perform is finding a threshold above which there is a risk of getting diagnosed with coronary heart disease. This could be done by looking at the CHD-diagnosed male's tobacco consumption, blood pressure, cholesterol number, BMI, and age, and comparing those numbers to the non-diagnosed males. And by that predict at what age non-diagnosed males should be concerned about the risk of getting CHD, and further predict whether the non-diagnosed men are in the risk zone of coronary heart disease depending on their different attribute values, and age. We could also make predictions based on whether the family history and/or a type-A personality influences the risks of getting CHD.

To be able to compare and make predictions, we should standardize some of the attributes, as the mean of the sbp, tobacco consumption, ldl, obesity, and age are different in scale. Additionally, we could binarize the type-A values with the same threshold of 55 points as Rossouw et al. used in [2]. If the subjects have type-A points ≥ 55 points, then the males are classified as type-A persons with: type-A = 1, No type-A = 0. The same can be done for the family history: present = 1, Absent = 0. To see at what age the attributes give the highest risk of getting CHD, we might consider to classify the observations into age groups, as they did in [2].

2 Detailed explanation of the attributes of the data

In this dataset, there are different kinds of attributes as described in Table 1, which can be classified as follows: The blood pressure, tobacco consumption, cholesterol level, and alcohol consumption are all ratios, with all except for the blood pressure being continuous numbers, as they can take decimal numbers starting from zero. The blood pressure do only take discrete numbers. The adiposity, type-A personality score, and obesity are ordinal and discrete values, because they are unitless numbers of scores. The

attributes about the history of coronary disease in their families and whether they have been diagnosed with CHD are nominal and discrete, more specifically they are binary as they only can take the numbers 0 and 1. The attribute types are summarized in Table 2.

	Type of the attribute	Discrete / continuous
sbp	Ratio	Discrete
Tobacco	Ratio	Continuous
ldl	Ratio	Continuous
Adiposity	Ordinal	Continuous
Family history	Nominal	Discrete
Type-A	Ordinal	Discrete
Obesity	Ordinal	Continuous
Alcohol	Ratio	Continuous
Age	Interval	Discrete
CHD	Nominal	Discrete

Table 2: Type of the attributes and whether they are discrete or continuous.

The dataset consists of 463 observations [1] with no missing values in all of the attributes. Python was utilized for counting the values, the results are presented in Table 3. Furthermore, the data are looked

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
Count of missing values	0	0	0	0	0	0	0	0	0	0

Table 3: Counts of missing values per attribute on the dataset.

through for corrupted observations. Firstly, the minimum and maximum of the values in each category can be calculated, the results are shown in Table 4.

	sbp	**tobacco**	**ldl**	**adiposity**	**famhist**	**typea**	**obesity**	**alcohol**	**age**	**chd**
Min	101.00	0.00	0.98	6.74	0.00	13.00	14.70	0.00	15.00	0.00
Max	218.00	31.20	15.33	42.49	1.00	78.00	46.58	147.19	64.00	1.00

Table 4: Minimum and maximum values for each attribute

We cannot determine whether the dataset is corrupted due to our lack of medical expertise. However, the maximum blood pressure value is very high. According to [2], a blood pressure of 218 mmHg significantly increases the risk of coronary heart disease. Nevertheless, it still seems feasible to have this blood pressure. We do conclude the same about the value ranges of the other attributes in Table 4.

To gain insight into how the data points in each attribute are distributed, a histogram is created for each of them. In Figure 1, we can see that there are no extreme outliers in any of the attributes. The distributions of SBP, LDL, adiposity, type, and obesity are bell-shaped and likely follow a normal distribution. Tobacco and alcohol consumption are both right-skewed, with most observations showing very low or zero consumption. Famhist and CHD are binary attributes, and most observations indicate no family history of CHD; the same pattern is observed for the CHD attribute itself. The age distribution

of males is neither normal nor uniform but has a single mode around age 60.

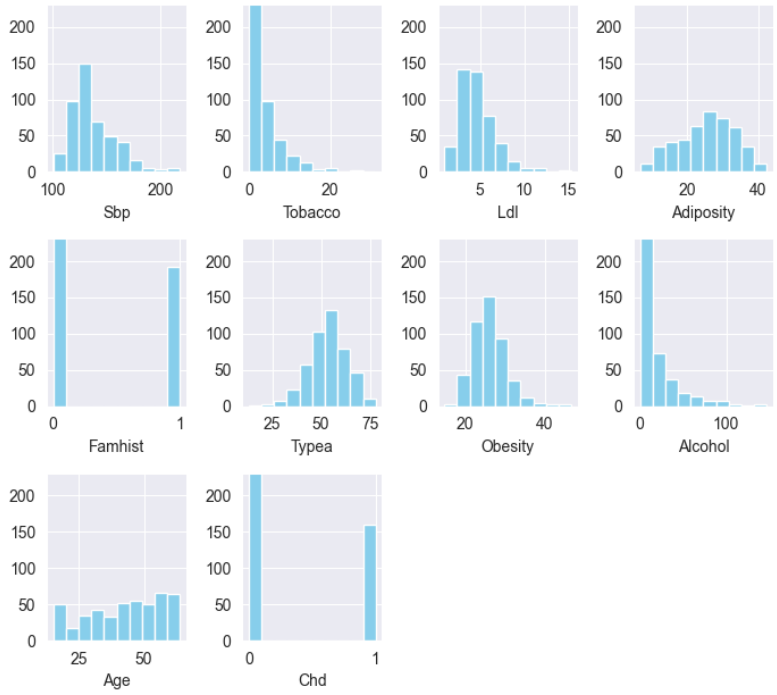
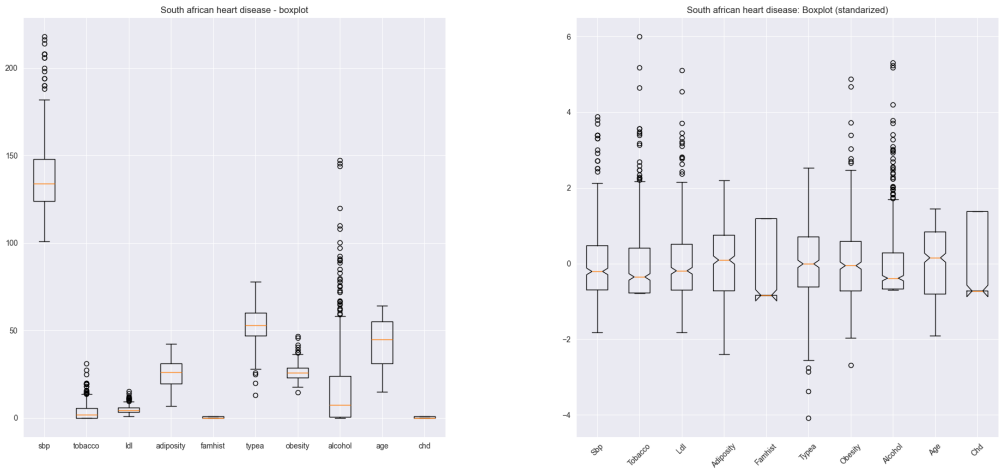


Figure 1: A histogram describing the distribution of the attributes.

To further see if the data contains any outliers, a boxplot (Figure 2) of the attributes is represented, both standardized and not.



(a) Given data of attributes.

(b) Standardized data of attributes.

Figure 2: Box plot of data, both given and standardized values.

In Figure 2a, some outliers in the attributes except for adiposity, famhist, age, and chd, are observed. After standardization, the distance from the mean is more clearly seen, where some of the outliers are more significant than others.

When we standardized the values, we observed that blood pressure (SBP), alcohol consumption, and age had relatively high variance (Figure 3). The high variances will influence how PCA prioritizes the attributes, giving more weight to those with higher variance over those with lower variance. Therefore, we need to divide all attributes by their variance to achieve unit variance across all attributes.

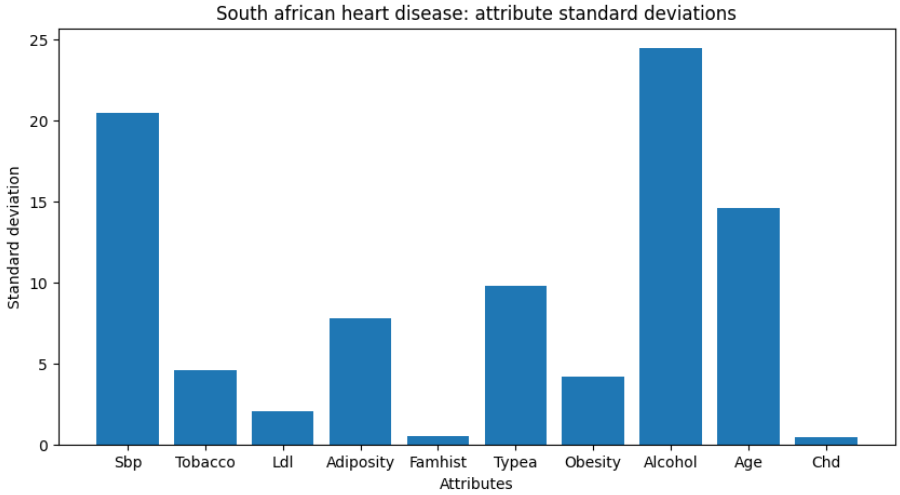


Figure 3: Standard deviation of attributes.

As a final way to analyze our attributes, we made a matrix of the standardized observations (Figure 4). None of the columns have different colors, and the different columns don't have different color patterns, so we can see that none of the attributes are more variable or have very different distributions.

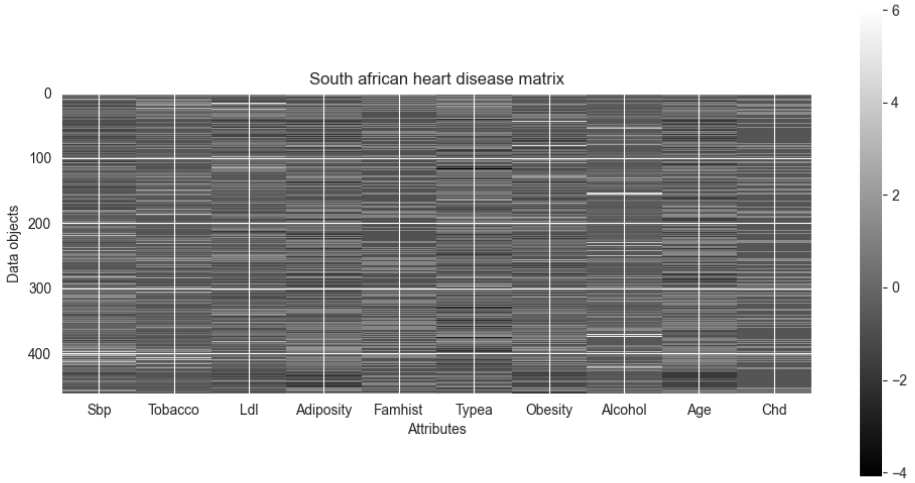


Figure 4: Standardized matrix of the attributes

3 Visualization of the data

3.1 Correlation on attribute basis

The previous section on summary statistics discusses the distribution of the attributes and the presence of any outliers. To find whether of the attributes are correlated, a pairwise scatter plot of M = 8 attributes[1] is made in Figure 5. The raw datapoints are classified into males with the CHD diagnosis (blue) and those without (orange), to see if these classes are affected by the attributes.

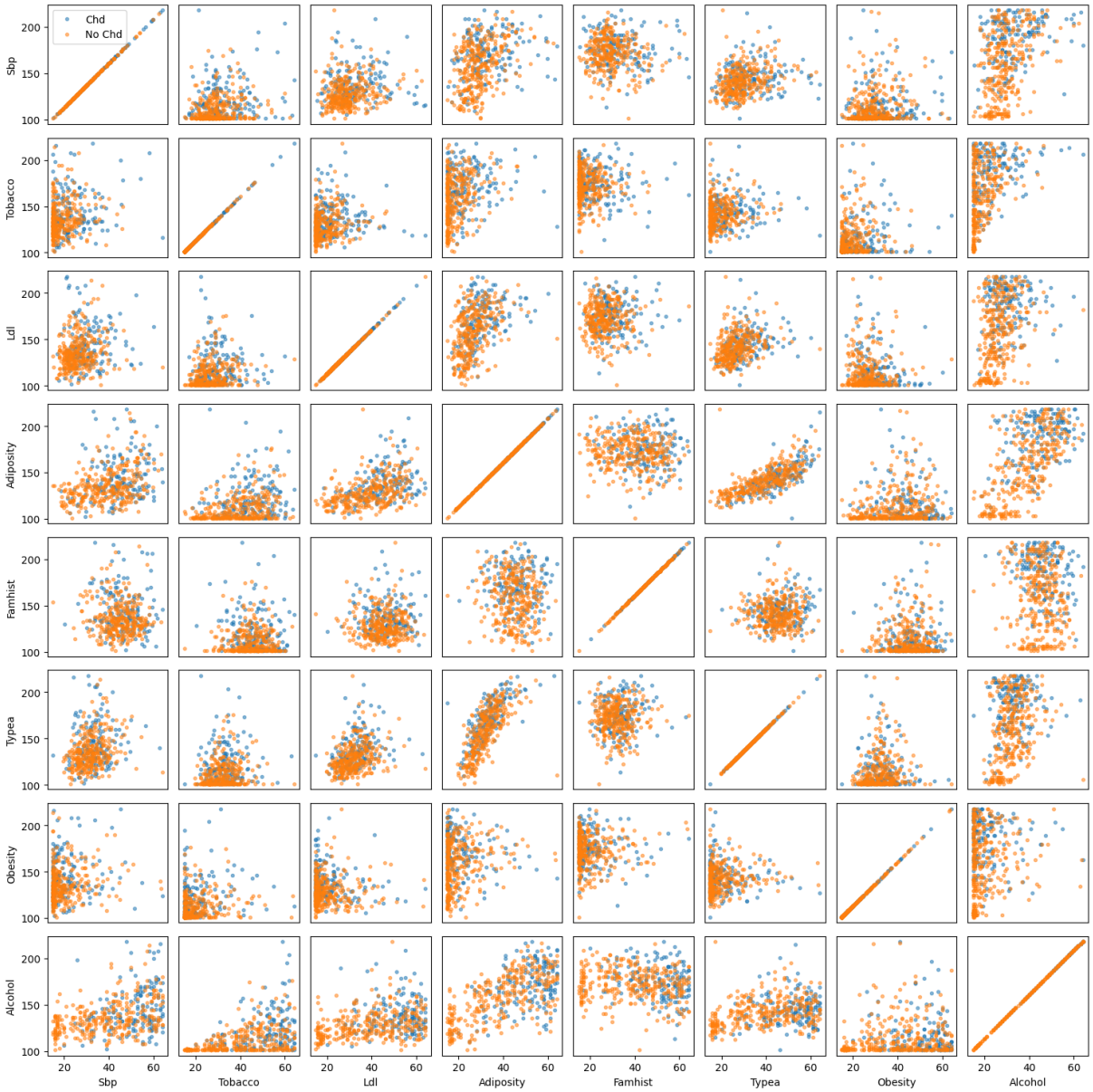


Figure 5: Correlation of attributes to each other and the classes: diagnosed with CHD = blue and not diagnosed with CHD = orange. (The attributes are not standardized nor normalized)

Overall, most of the attributes do not show a clear correlation with each other. However, a slight linear correlation can be observed between Type A and adiposity. The classification of the dots into CHD and non-CHD does not reveal any distinct pattern with the attributes.

[1] The binary attributes: famhist and chd are removed from this scatter matrix

3.2 Principal component analysis

For the principal component analysis, we need to consider the standardized data to get a better picture and PCA due to the different scales of the attributes. The eigenvalues that maximize the variance and give the lowest residual from the original plots are found by SVD. We have analyzed both the standardized and normalized data with SVD, where the attributes both have a mean value of 0 and a variance of one. In Figure 6, an overview of the first four PCA component coefficients is given (on the left), and attribute coefficients in the PC1, PC2 subspace (on the right).

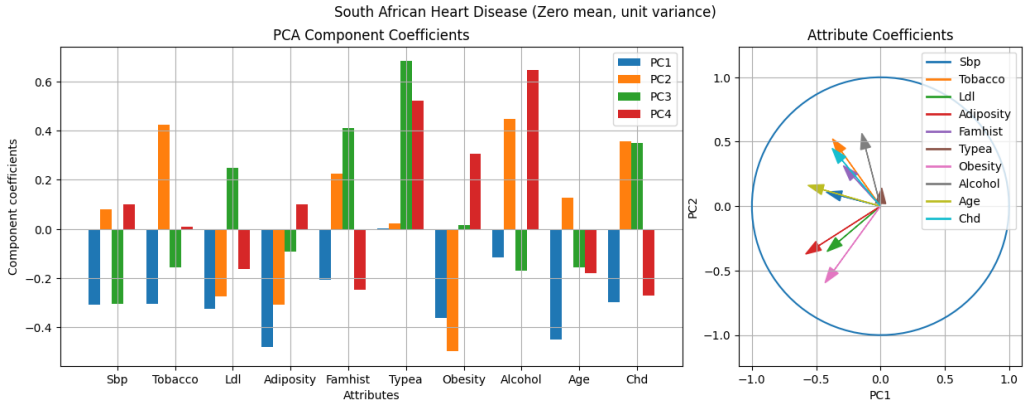


Figure 6: PCA component and attribute coefficients
(Zero mean, unit variance)

Both diagrams show that the first principal component (PC1) has only negative coefficients for all attributes, whereas the second principal component (PC2) includes both positive and negative values. The bar diagram on the left shows that both adiposity, obesity, and age contribute mostly to PC1, while the type-A score does not have any importance with a coefficient of nearly 0. For the second PC, the obesity, smoking, and drinking habits highly influence the component, with again no influence from the type-A score. The diagram to the right shows that the obesity and adiposity influence the PC1 vs. PC2 plots the most with having the highest magnitudes, while the family history and blood pressure have the lowest importance together with the type-A score.
In Figure 7, the cumulative variance explained is plotted as a function of the principal component with a threshold marked at 90%.

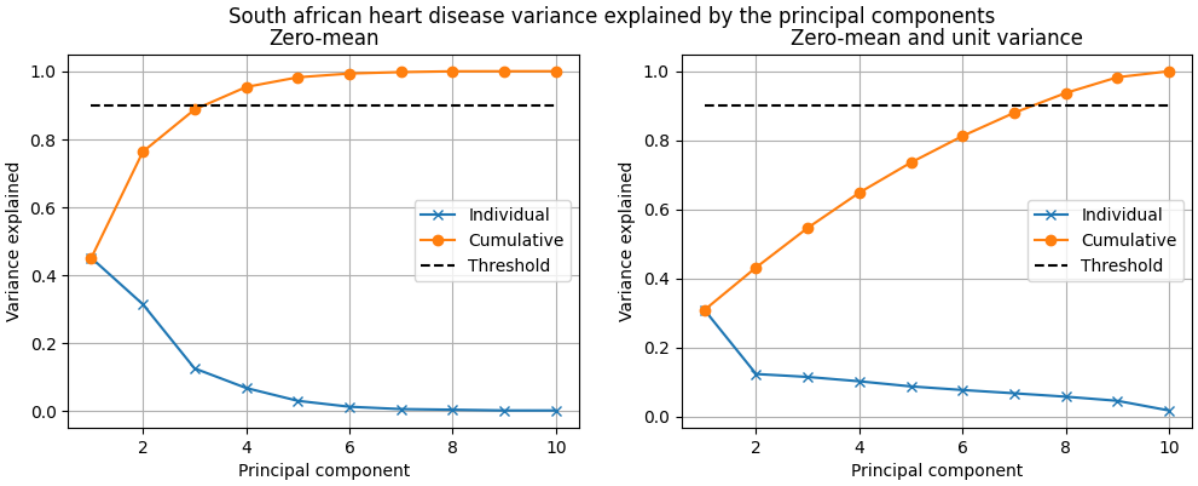


Figure 7: Variance explained by each principal component. The left shows it as a function of the PC's for the standardized data, and the right shows them for the normalized data with a unit variance.

With standardized data, we do only need 4 principal components to describe more than 90% of the data, whereas for the normalized data we need 8 PC's to do so. When we plot the first two principal components for both the standardized and normalized data, we obtain the plot in Figure 8.

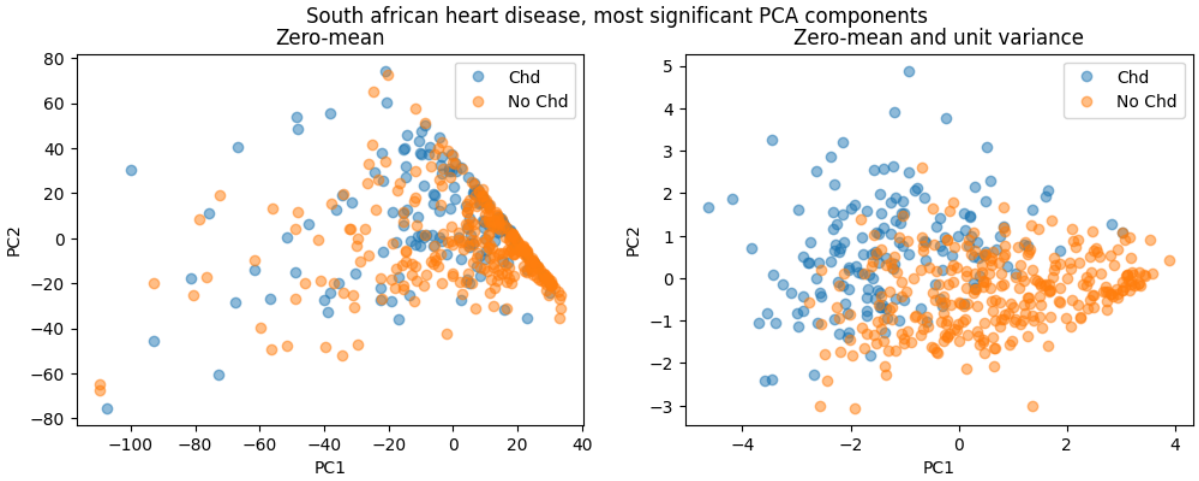


Figure 8: First two PC, raw dataset on the left, normalized by the mean on the right

When normalizing the data, we see some separation of the data between the two groups of males with and without CHD in the plot to the right. Whereas with the normalized and standardized data, the data still lie on top of each other independent of their class. We do also see some kind of line in the upper right corner of the scatter plot, although we are not sure why that is.

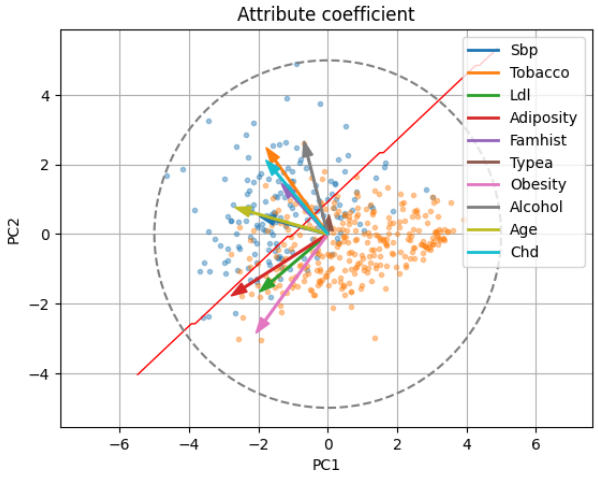


Figure 9: Attribute coefficients, PCA analysis and linear regression combined

We see a separation of the two classes (Chd 1 and 0) on Figure 9 indicated by the red line.[2] From the attribute coefficients, we can see that the attributes adiposity, Ldl, and obesity are generally parallel to the separation line. These attributes are not related to Chd, and the attributes that are perpendicular to the separation line (tobacco and famhist) are more relevant to the value of Chd. Also, alcohol, age and sbp are relevant to chd, but with a smaller coefficient.

We then tried to add the third PC on the normalized data to see the pattern of the data points in Figure 10. With the third PC, we do not get a further distribution/explanation of the data. [3]

[2]This separation line (shown in red) is the result of a linear classifier on the PC1-PC2 plane only.
[3]This also does not help when data is viewed in an interactive viewer

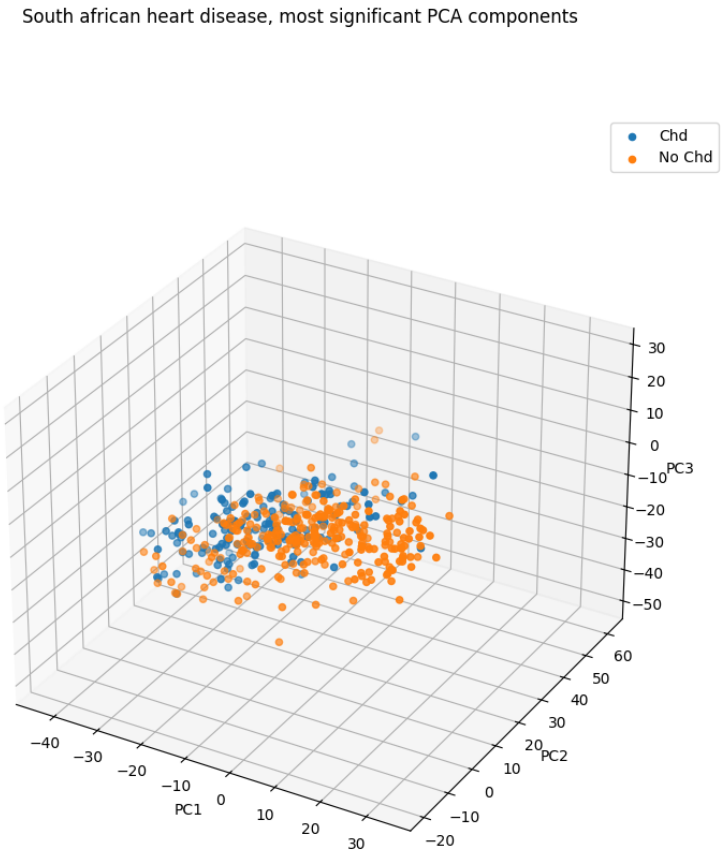


Figure 10: Observations viewed in the subspace defined by PC1, PC2, and PC3 normalized by the mean.

We also made Figure 11 like the above Figure 8 with the binarization of the attribute type-A values with the same threshold of 55 points as Rossouw et al. used in [2]. If the subjects have type-A points $\geq$ 55 points, then the males are classified as type-A persons with: type-A = 1, No type-A = 0. For that, we could not find any immediate separation (even with other PCs).

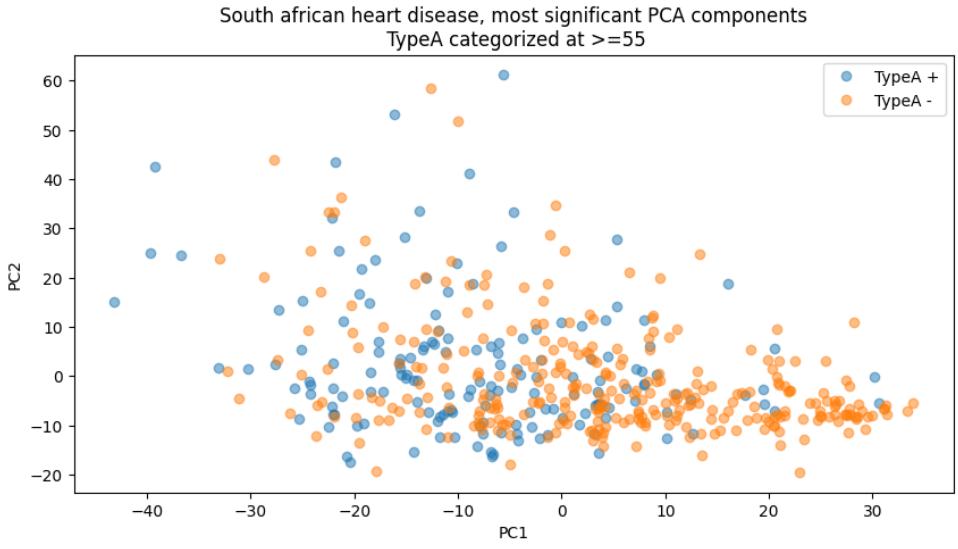


Figure 11: First two PC with categorized with Type-A

4 Discussion

The histograms and boxplots made in section 2 show that before analyzing the data, it needs to be normalized to prevent any mistakes when applying the machine learning methods due to the difference in ranges of the scales between the attributes. However, no clear outliers of the attributes are observed in the dataset. Hence, it is assumed that none of the observations are corrupted.

The pairwise scatter plot of the attributes (Figure 5) shows no direct correlation of the attributes, except for type-A and adiposity, and the classifications of the males in no-CHD and CHD groups are not separated by the attributes either. When we do the PCA on the normalized data, we see that the separation of the two groups happens when we plot PC1 against PC2. From looking at Figure 9, we can conclude that the presence of coronary heart disease (Chd = 1) is likely to be more related to the attributes of Tobacco and famhist. Whereas adiposity, Ldl, and obesity are not seen to be related to Chd. But the plot of the cumulative variance explained of the principal components in Figure 7, shows that we can get a better explanation of the standardized data with a lower set of PC's than for the normalized data.

A classification prediction we want to perform is finding a threshold above which there is a risk of getting diagnosed with coronary heart disease. We could already identify that the attributes of Tobacco and famhist have more influence, while type-A and adiposity seem to have less impact on getting coronary heart disease. We think that it is feasible to make a classification prediction as we have already found some separation just by performing PCA.

For the regression analysis, we will use Type-A values. Currently, we have visualized them using PCA, but we could not identify any immediate separation. Although on the scatter plot (Figure 5) we see that type-A and adiposity have some correlation. At first glance, it might be feasible to make regression with Type-A, even though we could not gather any insights based on the PCA. Although we think given the higher dimensionality of the data, regression can still be performed.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009. ISBN: 978-0-387-84858-7. URL: https://hastie.su.domains/ElemStatLearn/.

[2] J.J. Rossouw et al. "Coronary risk factor screening in three rural communities. The CORIS baseline study". In: *South African Medical Journal* 64 (1983), pp. 430–436. DOI: 10.10520/AJA20785135_9894.