Technical University of Denmark

02450
Introduction to Machine Learning and Data Mining

Report 1

March 25, 2021

Contents

1 Introduction

In this report, the Framingham Heart Study is selected as data set to analyze and apply machine learning techniques. This is a long term, ongoing cardiovascular study on residents of the town Framingham in Massachusetts that began in 1948 with 5209 adult subjects and now is on its fourth generation of participants [1]. Cardiovascular diseases (CVD) are a group of heart and blood vessels disorders that include the coronary heart disease, a disease of blood vessels supplying the heart muscle [2]. The objective of the study is to identify the main characteristics that have high correlation with cardiovascular diseases. This is achieved by monitoring a large group of people over a long period of time, who had not yet shown any symptoms [3].

The data set is obtained from the Kaggle website where is publicly available [4]. It provides patients' information in several categories such as demographic, behavioral and medical risk factors by including over 4,000 records and 15 attributes.

There are multiple papers over the years that analyzed the data in order to estimate which attributes contribute the most to coronary heart disease. In general, the results have shown that mainly cigarette smoking, cholesterol level, and blood pressure increase the risk of heart disease. In addition, it seems that men are more susceptible to heart disease than women as well as the age is an important factor. On the other hand, glucose causes a negligible change in heart disease risk.

The report aims to get a solid understanding and learn how to handle the selected data by using two widely spread machine learning techniques:

- Classification
- Regression

The main goal of classification is to create a model that is able to diagnose a ten year risk of coronary heart disease (CHD), based on the attributes given. In addition, the heart rate is also going to be predicted with a regression model, in order to compare the correlation of these two attributes with the rest of the data set.

Along the process, a Principal Component Analysis (PCA) is carried out to filter and visualize the data. Additionally, in many cases, it is required to standardize the data, since PCA is sensitive to features that reside on a different scale. One way to do so is to subtract the mean from the different values of each attribute, while another way to achieve standardization is to subtract the mean and also divide by the standard deviation.

2 Attributes

In this section, a detailed explanation of the data attributes is introduced. The selected data set consists of sixteen attributes, where all of them can be considered as a potential risk factor, except of the education level of each resident.

2.1 Type of attributes

At first, the attributes are categorized in different types, where they can be either discrete or continuous and then they are divided in more specific types such as nominal, ordinal, interval or ratio. The attributes of the data set as well as their type are included in Table 1.

Attribute	Symbol	Type
Sex	male	Discrete - Nominal
Age	age	Continuous - Ratio
Education	education	Discrete - Ordinal
Smoker	currentSmoker	Discrete - Nominal
Cigarettes per day	cigsPerDay	Continuous - Ratio
Blood pressure medication	BPMeds	Discrete - Nominal
Prevalent Stroke	prevalentStroke	Discrete - Nominal
Prevalent Hypertensive	prevalentHyp	Discrete - Nominal
Diabetes	diabetes	Discrete - Nominal
Total Cholesterol	totChol	Continuous - Ratio
Systolic blood pressure	sysBP	Continuous - Ratio
Diastolic blood pressure	diaBP	Continuous - Ratio
Body mass index	BMI	Continuous - Ratio
Heart rate	heartRate	Continuous - Ratio
Glucose level	glucose	Continuous - Ratio
10 year risk of CHD	TenYearCHD	Discrete - Nominal

Table 1: Description of data attributes.

2.2 Data Issues

The data set contains 4240 observations where 582 of them have at least one missing value. These observations appear to have lack of information in one or multiple attributes and the total number of missing values per attribute is summarized in Table 2.

Attribute	No of missing values
Education	105
Cigarettes per day	29
Blood pressure medication	53
Total Cholesterol	50
Body mass index	19
Heart rate	1
Glucose level	388
Total	645

Table 2: Missing values per attribute.

To handle this issue, an imputation method is implemented to 380 observations by replacing the missing values with the mean of all the observed instances of the corresponding attribute. On the contrary, the rest of the observations are deleted, due to the fact that either more than one attribute is missing or the missing attributes have binary values, so it does not make sense to replace them with the mean. In addition, several possible outliers are also detected, but at this initial stage, it is decided not to remove them, since we can not be totally sure that they are actually outliers.

2.3 Basic Summary Statistics

At last, several basic summary statistics are calculated for the continuous attributes of the data set. More specifically, the mean, the standard deviation, the variance as well as the minimum and maximum values are estimated and the results are shown in Table 3.

Attributes	Mean	Std	Variance	Min Value	Max Value
Age	49.48	8.53	72.92	32.00	70.00
Cigarettes per day	9.06	11.88	141.27	0.00	70.00
Total Cholesterol	236.60	44.02	1938.22	107.00	600.00
Systolic blood pressure	132.23	21.98	483.16	83.50	295.00
Diastolic blood pressure	82.85	11.89	141.51	48.00	142.50
Body mass index	25.77	4.07	16.56	15.54	56.80
Heart rate	75.89	12.08	146.06	44.00	143.00
Glucose level	81.92	23.05	531.50	40.00	394.00

Table 3: Summary statistics of the continuous attributes.

3 Data Visualization

Data visualization is one of the most important steps in the process of exploring the data. By different visualization methods, new relationships between the data can be discovered. Four different visualization plots are used for this scope.

3.1 Scatter Plots

Scatter plots can be used for the continuous attributes in order to diagnose a linear relationship between them. As it is easily observed in Fig. 1, there is no clear linear relationship between the attributes, with the only exception being between systolic and diastolic blood pressure.

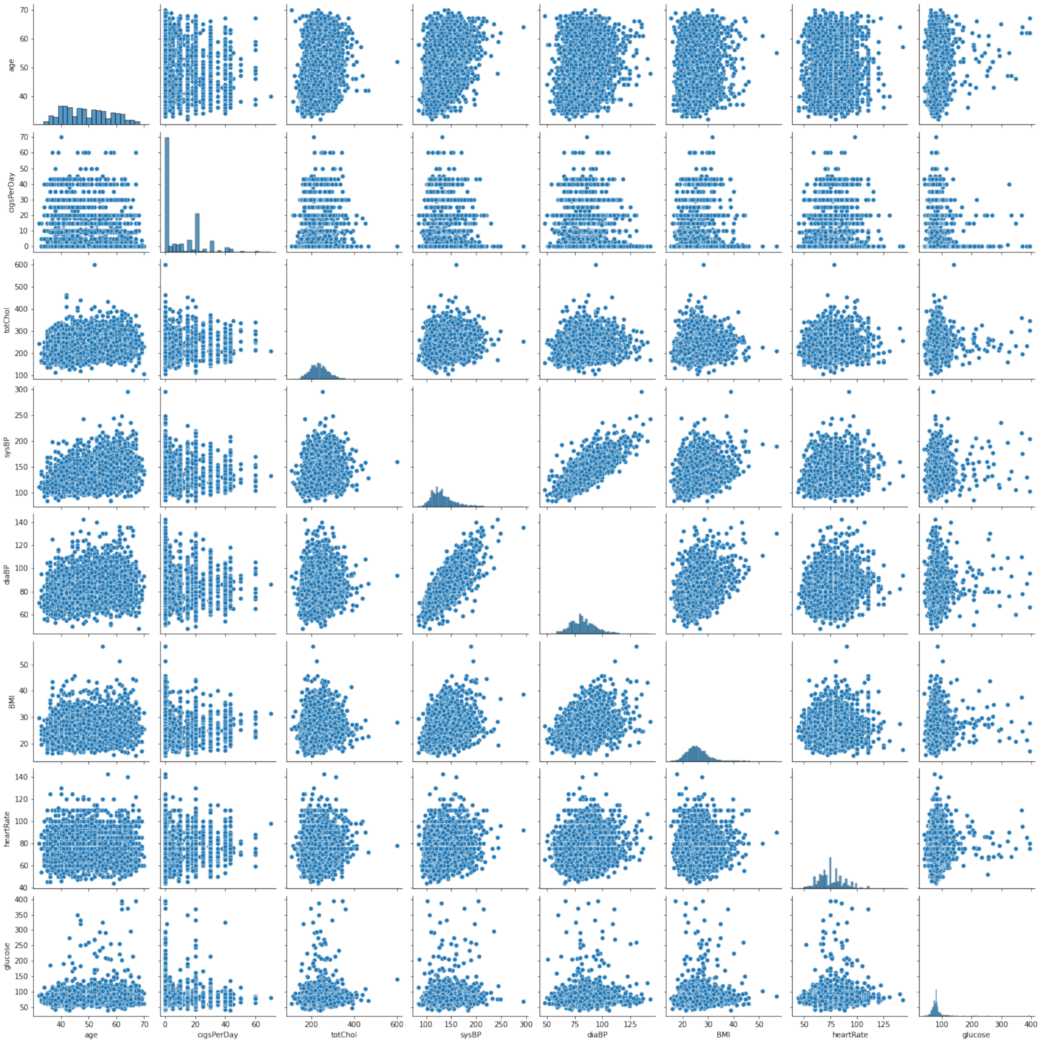


Figure 1: Scatter plot for the continuous attributes.

3.2 Heat maps

Heat maps are useful visualization tools to estimate the correlation coefficients between attributes. The heat map in Fig. 2 shows the correlation between the desired attribute ten year risk for coronary heart disease (TenYearCHD) with the other attributes. Four attributes seem to affect more the TenYearCHD which are the age, the prevalent hypertensive, the systolic and the diastolic blood pressure with correlation coefficients 0.23, 0.18, 0.22 and 0.15, respectively.

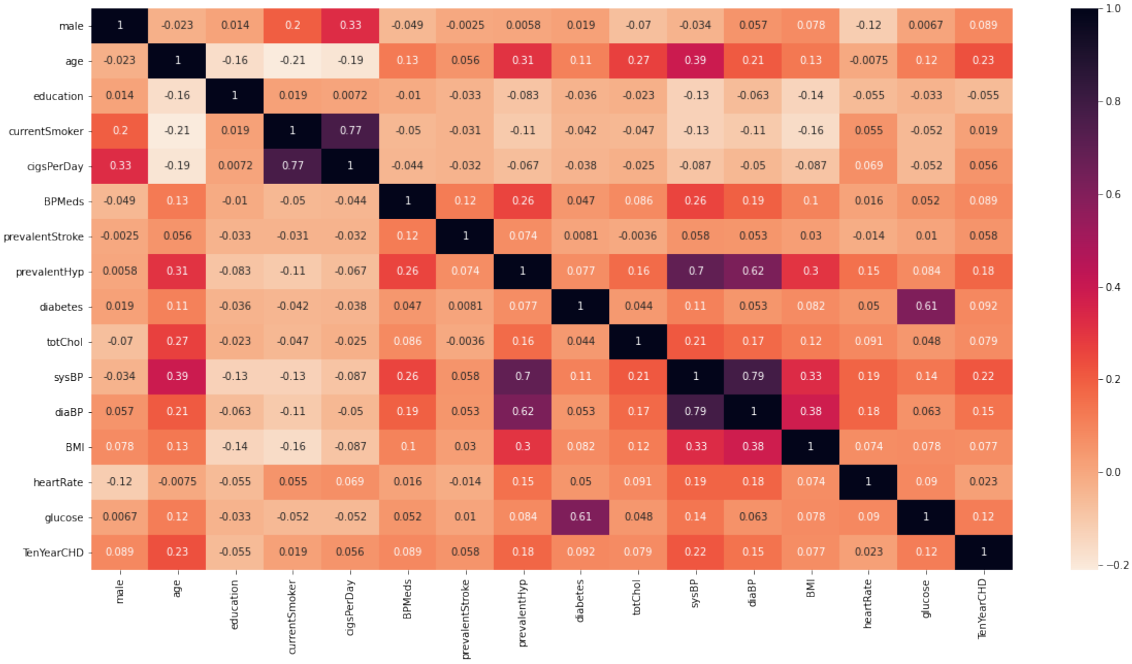


Figure 2: Correlation coefficients between attributes in the form of a heat map.

3.3 Box Plots

Among the useful information that can be extracted from the box plots, this visualization tool can help identify the outliers included in each attribute. As seen in Fig. 3, almost every attribute shows several outliers, except of the age (top left plot). However, due to our lack of knowledge in medicine issues, dropping these values would be a bad decision as they might be realistic.

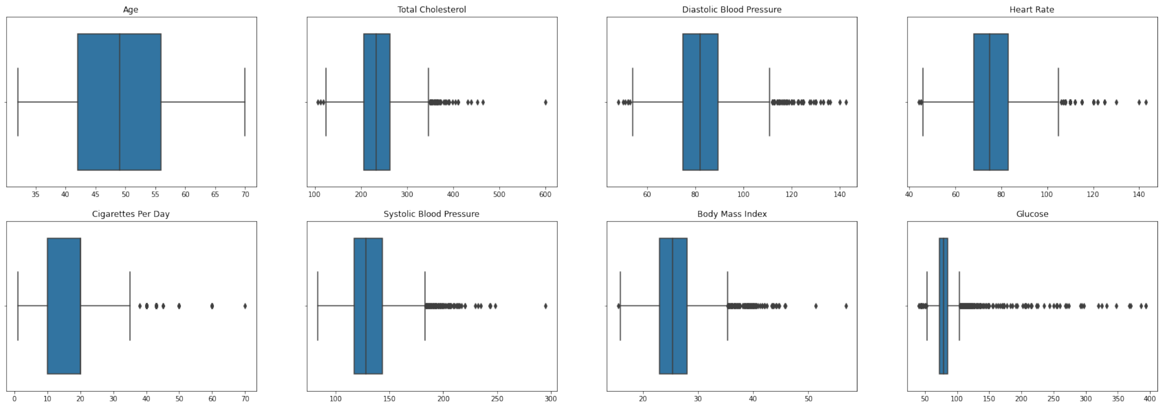

Age
Total Cholesterol
Diastolic Blood Pressure
Heart Rate
Cigarettes Per Day
Systolic Blood Pressure
Body Mass Index
Glucose

Figure 3: Box plots of the continuous attributes.

3.4 Histograms

The histograms are plotted in order to estimate the distribution of each continuous attribute. As seen in Fig. 4, it is clear that most of the observations are concentrated around the mean and, in that way, the attributes follow a Gaussian distribution.

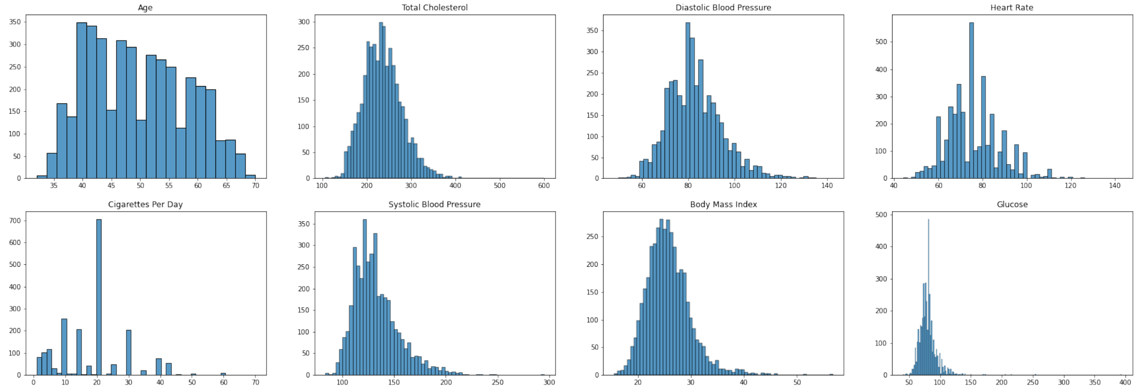

Age
Total Cholesterol
Diastolic Blood Pressure
Heart Rate
Cigarettes Per Day
Systolic Blood Pressure
Body Mass Index
Glucose

Figure 4: Histograms the continuous attributes.

4 Principal Component Analysis

The principal component analysis (PCA) is a mathematical algorithm that can be used to find a lower-dimensional representation of the data set, while retaining most of the variation of the data, as it is necessary for the classification of the residents with a risk of a coronary heart disease.

In order to carry out a PCA, the outliers of the data need to be removed, but most importantly every attribute of the data must be standardized. To standardize the data, as mentioned above, the mean of every attribute is subtracted from every observation that maps to the respective attribute and then these values are divided with the standard deviation of the attribute. This process is necessary since the attributes have substantially different scale as shown in Fig. 5.

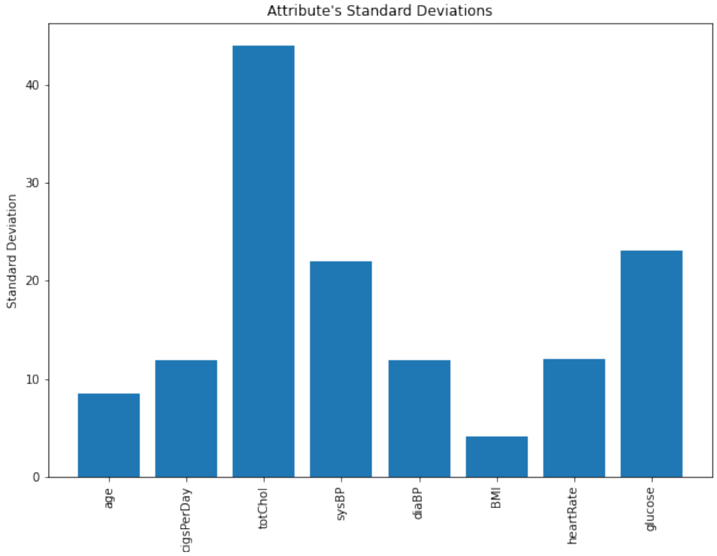


Figure 5: Standard deviation of the continuous attributes.

Every component of the PCA represents a different percentage of the overall variance. In Fig. 6, it is observed that nine principal components are needed in order to reach the threshold of 0.9. This means that only the first nine out of eighteen principal components are needed to cover the 90% of the variance. Additionally, in Fig. 6 can be seen that PC1 captures the 24.6% of variance, while PC2 and PC3 capture 13.7% and 10.04%, respectively. Therefore, three principal components generate almost the half of the whole variance.

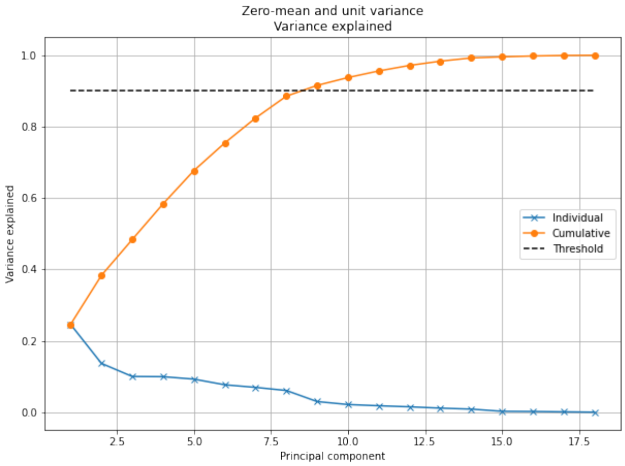
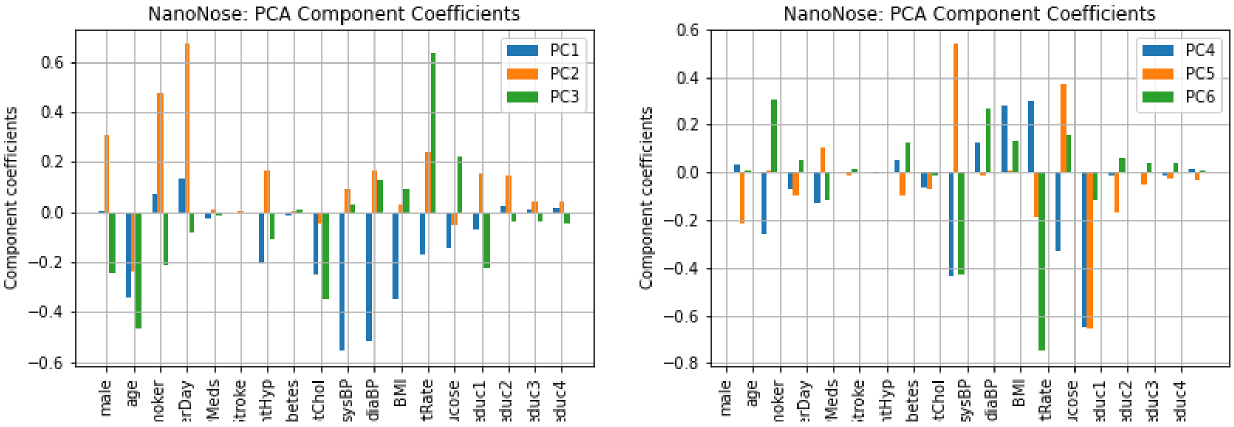


Figure 6: Cumulative variance for principal component analysis.

Further information can be obtained from Fig. 7, where it can be easily observed that every principal component gathers variance from almost every attribute except blood pressure medication, prevalent stroke, diabetes and education. This analysis may hint that these attributes could be disregarded. However, they will not be removed until further machine learning techniques are implemented.

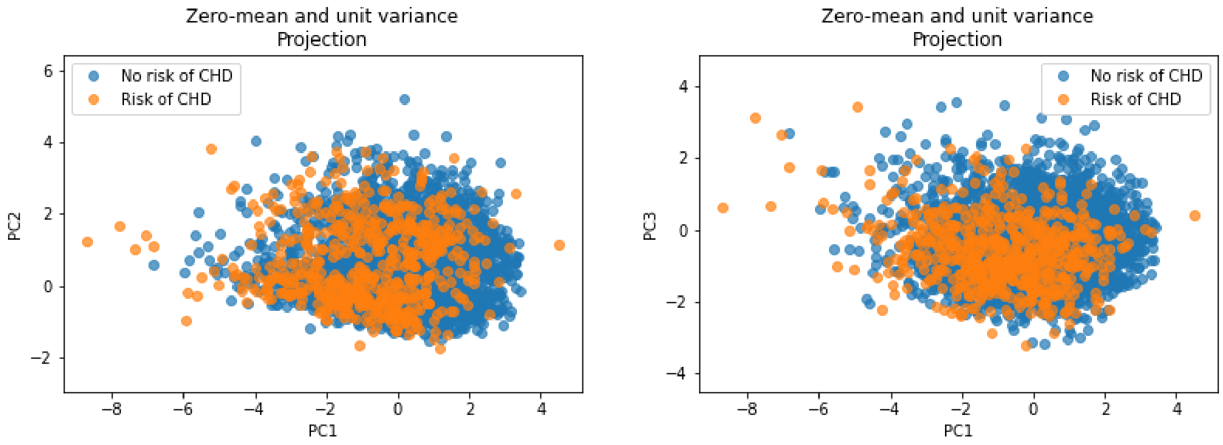


(a) PC1, PC2 and PC3

(b) PC4, PC5 and PC6

Figure 7: PCA component coefficients: (a) for principal components 1, 2 and 3; (b) for principal components 4, 5 and 6.

Moreover, by plotting the projections of the data from two principal components, a colored scatter plot is created depending on the predicted attribute, as shown in Fig. 8. In these graphs as well as in additional combinations that were investigated (i.e. PC1-PC4, PC2-PC3 etc), it is evident that there is not a clear separation of the observations. In that way, it can be concluded that two principal components are not enough for a proper classification of the data.

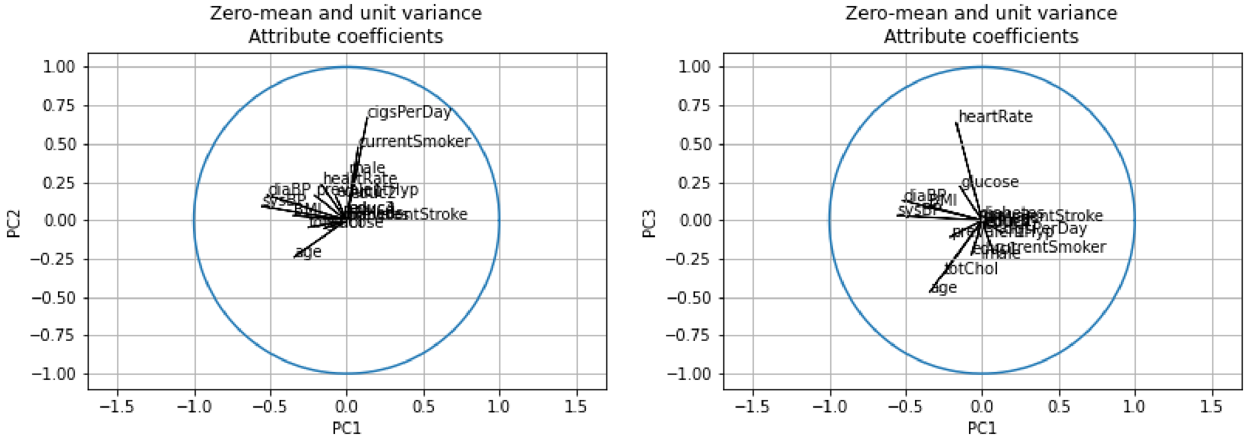


(a) PC1 - PC2

(b) PC1 - PC3

Figure 8: Projections of the data: (a) from principal components 1, 2; (b) from principal components 1, 3.

Finally, the attribute coefficients are shown in Fig. 9 along with their direction. In these graphs, the larger the length of the vector the higher the variation it contributes to the principal component. Moreover, the closer the attributes to each other the larger the variation they share. Based on that, along with the information obtained from Fig. 7, some attributes could be discarded in further analysis.



(a) PC1 - PC2

(b) PC1 - PC3

Figure 9: Attribute coefficients: (a) from principal components 1, 2; (b) from principal components 1, 3.

5 Discussion

In the present report, the Framingham heart study was investigated with a primary goal to get a solid understanding and prepare the data in order to apply machine learning techniques in the upcoming report. At first, an initial analysis of the data set was implemented as the attributes were divided into different types and an investigation was made for data issues (i.e. missing values and corrupted data). Subsequently, several visualization tools were used in order to analyze the data and discover a correlation between the different attributes. At last, the principal component analysis (PCA) was performed to investigate the possibility of finding a lower-dimensional representation of the selected data set. PCA showed that a dimensionality reduction could be achieved by discarding several attributes as proved in Chapter 4. Overall, the first report was a useful introduction to the machine learning philosophy, since we learned how to handle, analyze, modify and visualize our data in order to get numerous useful information about it.

References

[1] Wikipedia, Framingham Heart Study, viewed 10 February 2021, https://en.wikipedia.org/wiki/Framingham-Heart-Study.

[2] World Health Organization, Cardiovascular diseases (CVDs), viewed 10 February 2021, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[3] Framingham Heart Study, About the Framingham Heart Study, viewed 10 February 2021, https://framinghamheartstudy.org/fhs-about/.

[4] Kaggle, Framingham Heart study dataset, viewed 10 February 2021, https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset.