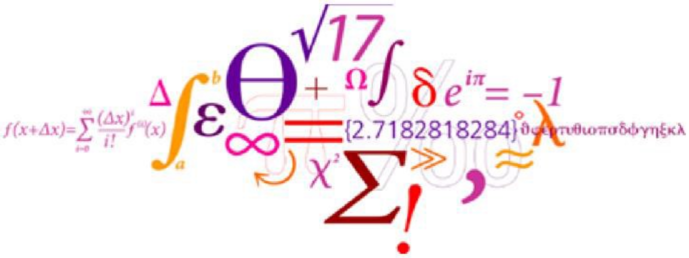
02450 Introduction to Machine Learning and Data Mining

Data: Feature extraction, and Visualization

Report 1

Longfei Lin, s185882	Contribution:33.3%
Sai Chaitanya Annadatha, s181739	Contribution:33.3%
Iraklis Chrysikopoulos, s182995	Contribution:33.3%



05/03/2019

1. Description of the data set

1.1 Problem of interest

This data set consists of factors that are responsible for heart disease. The goal to confirm whether the person is having heart disease or not, based on several test results. It is easy to identify the cause and predict future data to prevent the disease in its initial stage.

1.2 Source

This data is collected from UCI machine learning data set based on the Cleveland database which is used by the ML researchers. Initially, this data has 76 attributes but the only subset of 14 attributes are used in the previous experiments. It seems that the personal details of the patients were removed due to some security reasons and most of the attributes are removed due to some inaccuracies in data [1].

1.3 Previous work on data

A new discriminant function model was introduced based on the test results of 303 patients undergoing angiography at Cleveland clinic in Cleveland for estimating the probability of the heart disease. This data was also applied to various institutions in 3 different countries namely Hungary, California, Switzerland and compared with the CADENZA (application of Bayesian algorithm). Upon comparing it is said that both algorithms overpredict the data but on digging deep the new discriminant model less overpredicts than CADENZA. It is finally concluded that the Cleveland model is reliable and clinically useful when applied to patients with chest pain syndromes and intermediate disease prevalence [2].

1.4 Aim for the data

On observing our data is based on classification technique but these attributes can also be used for other techniques in machine learning. We defined attribute ‘target’ as our class label as we want to present the existence of the heart disease. We can use all attributes except two which are removed due to lack of quality in data but, for better diagnosis we assume to use attributes ‘Cp’, ‘Thalch’, ‘Exang’, ‘Oldpeak’,’Slope’ in future as they have strong correlation with attribute ‘Target’ which is our class label. Besides, based on our data it can be observed that all our attributes are clustered. Here it is possible to clearly visualize the data properly so, we did not use any transformation.

2. A detailed explanation of the attributes of the data

2.1 Attributes description

Our data set has 14 attributes, these attributes have different types and features which can be described by the following table.

Attribute	Age	Sex	CP	Trestbps	Chol	FBS	Restecg
	discrete	discrete	continuous	continuous	continuous	discrete	discrete
	ratio	nominal	nominal	ratio	ratio	nominal	nominal
Attribute	**Thalach**	**Exang**	**Oldpeak**	**Slope**	**CA**	**Thal**	**Target**
	continuous	discrete	continuous	discrete	discrete	discrete	discrete
	ratio	nominal	ratio	nominal	nominal	nominal	nominal

Table 1: Attributes Description

2.2 Data Issue

The accuracy of a machine learning algorithm depends on the quality of a data set. Data quality is absolute one of the most important conditions for solving a data-driven problem. However, it is inevitable for a data set which include much of different observations and attributes to have some data issues, such as Irrelevant or Spurious attributes, Outliers, and Missing Data.

In our data set, we also find some data issues. First, with respect to CA, it just has four different values according to the data description, but in our data set, it has five different values, and we don’t have any evidence for the extra data, another attribute Thal also meets the same issue. Moreover, we also meet some data dislocation problems, in our data description, the data ranges of CP and Slope are 1~4 and 1~3, but in our data set, the ranges are 0~3 and 0~2.

These are the data issues in our data set, in order to improve the quality of the data set, we discarded CA and Thal, and also made a relatively minor modification on the value of the CP and Slope to make them more consistent with the data requirement.

2.3 Summary statistics

Summary statistics are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible [3]. For most of the definitions, summary statistics consist of three main categories:
(1) Measures of location.
(2) Measures of spread.
(3) Graphs/charts.

In our data set, we focus on the first two parts. The results are shown by the following table.

Attribute	**Age**	**Sex**	**CP**	**Trestbps**	**Chol**	**FBS**
Count	303	303	303	303	303	303
Mean	54.36634	0.683168	0.966997	131.6238	246.264	0.148515
Std	9.082101	0.466011	1.032052	17.53814	51.83075	0.356198
Min	29	0	0	94	126	0
25%	47.5	0	0	120	211	0
50%	55	1	1	130	240	0
75%	61	1	2	140	274.5	0
Max	77	1	3	200	564	1
Attribute	**Restecg**	**Thalach**	**Exang**	**Oldpeak**	**Slope**	**Target**
Count	303	303	303	303	303	303
Mean	0.528053	149.6469	0.326733	1.039604	1.39934	0.544554
Std	0.52586	22.90516	0.469794	1.161075	0.616226	0.498835
Min	0	71	0	0	0	0
25%	0	133.5	0	0	1	0

50%	1	153	0	0.8	1	1
75%	1	166	1	1.6	2	1
Max	2	202	1	6.2	2	1

Table 2: Summary statistics of Dataset

3. Data visualizations

3.1 Data visualization based on original data

Data visualization is a very efficient way to convey information through graphics. We can learn kinds of features and relations from different attributes through various pictures such as histogram, pie, and heatmap.

At first, we use histograms to show the distribution of some continuous attributes.

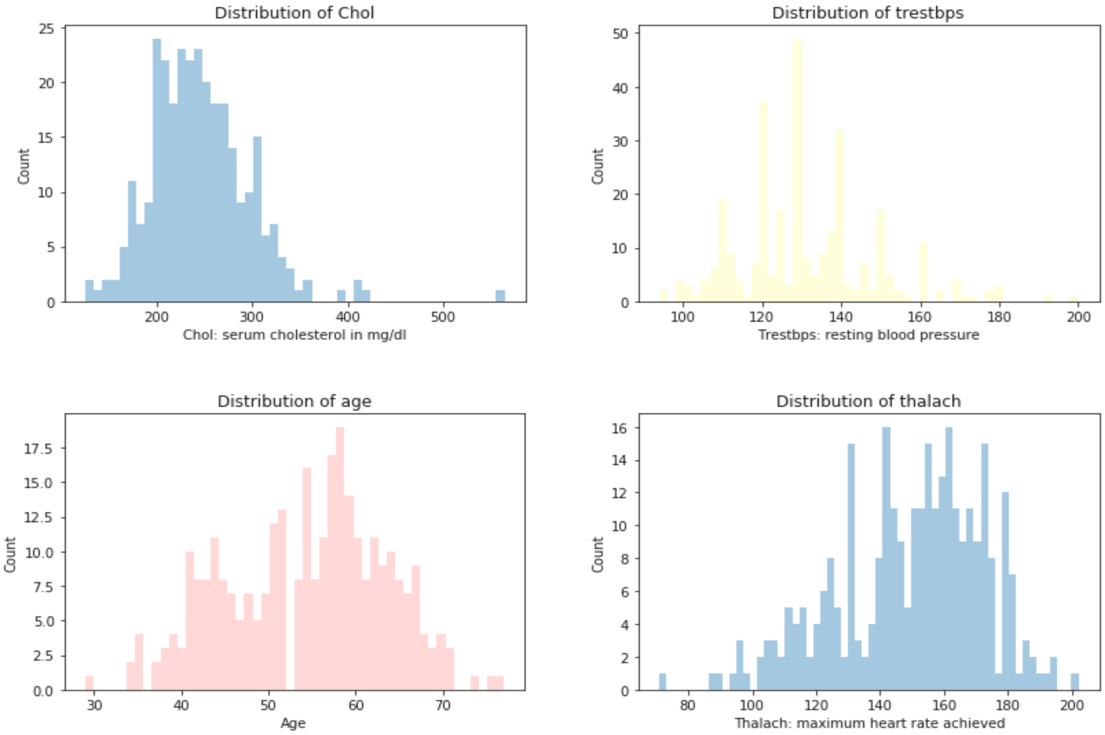


Fig:1 Histogram of continuous attributes

We can also use histograms to show the features of discrete attributes and the percentage of having the disease in each feature.

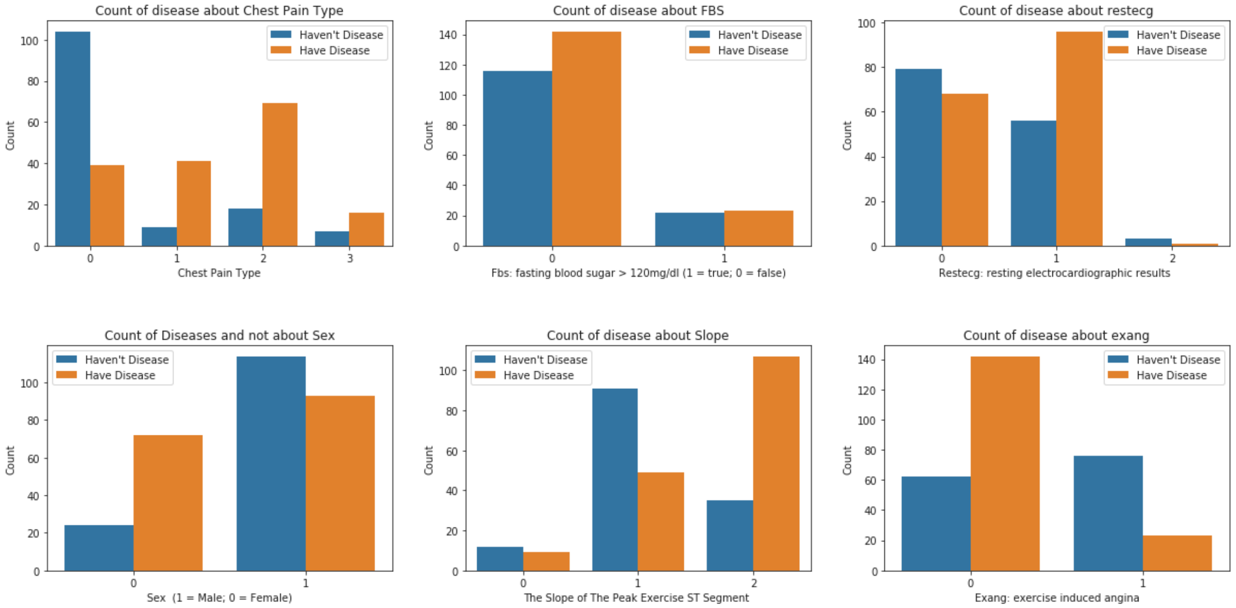
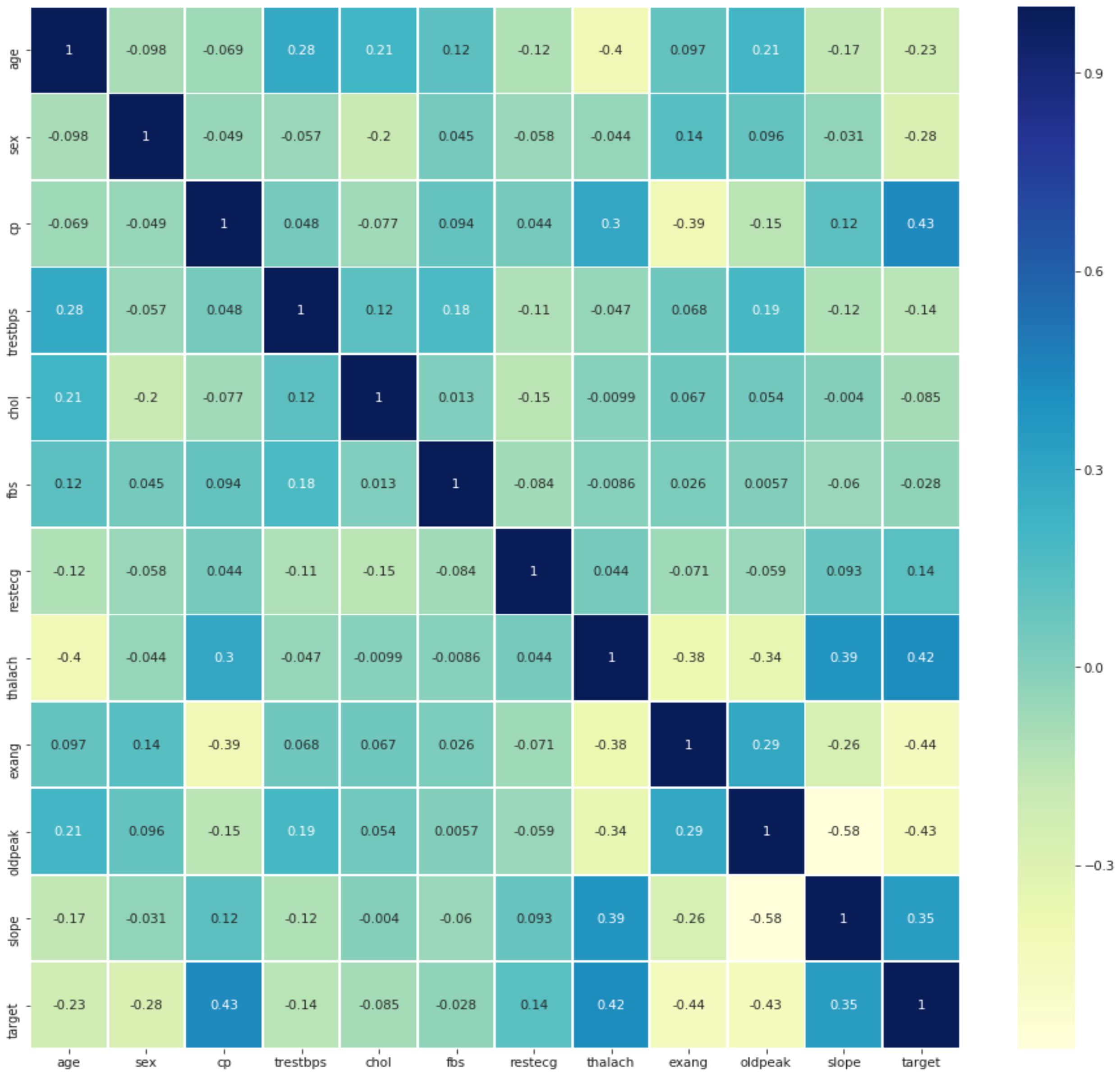


Fig:2 Histogram of Discreate attributes

The correlation of different attributes can be described by a heatmap. We use a different color to present different value of correlation coefficient, the value range is between -1 and 1.

Based on previous data visualizations, we can quickly get some conclusions from the data.
(1) There is an outlier in Chol (Serum Cholesterol in mg/dl).
(2) Most of the continuous attributes appear to be normally distributed.
(3) Combine the result of histogram and heatmap, we found that there are five attributes ‘Cp’, ‘Thalach’, ‘Slope’, ‘Exang’ and ‘Oldpeak’ which have a strong correlation with ‘Target’, and for independent variables, there is a strong correlation between ‘Slope’ and ‘Oldpeak’.



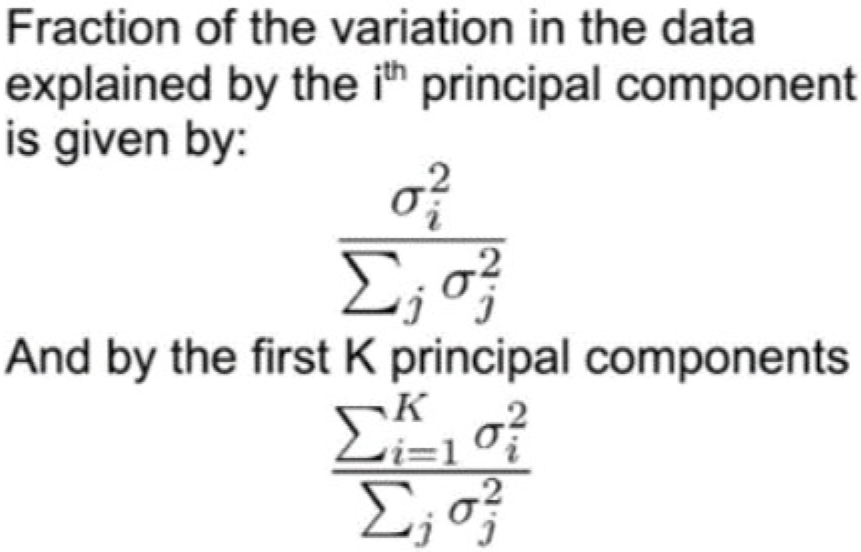
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	target
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	-0.23
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	-0.28
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	0.43
trestbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	-0.14
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	-0.085
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	-0.0086	0.026	0.0057	-0.06	-0.028
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	0.14
thalach	-0.4	-0.044	0.3	-0.047	-0.0099	-0.0086	0.044	1	-0.38	-0.34	0.39	0.42
exang	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	-0.44
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	-0.43
slope	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	0.35
target	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	1

Fig:3 Heatmap of the correlation coefficient

3.2 Data visualization base on PCA

3.2.1 The amount of variation

The fraction of variance explained by a principal component is the ratio between the variance of that principal component and the total variance.



Fraction of the variation in the data explained by the i^{th} principal component is given by:

$$\frac{\sigma_i^2}{\sum_j \sigma_j^2}$$

And by the first K principal components

$$\frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_j \sigma_j^2}$$

Fig 4: Expression of Variance[4]

In figure 5, it is observed that the variance explained for 90% of our data needs 9 principal components. For 95% we need 11 principal components. This number is quite high in comparison to the examples in the exercises, but the kind of our dataset needs a high value of principal components.

3.2.2 The principal directions of PCA components

In figure 6, the coefficients of each attribute of the first 3 PCA are shown. Each component coefficient starts from the origin and has a positive or negative magnitude. The direction and the magnitude define how the data from each attribute is projected onto the PC1/PC2 space. The first PCA contributes negatively to a high level at the attributes age, trestbps, exang, oldpeak and the second PCA in sex, restecg, exang. On the other side, the positive contribution in PCA1 is in cp, thalach and slope whereas PCA2 in age, cp, trestbps, chol and fbs. Eventually, our target has positive contributions in both PCA 1 and 2 with the first one to have a double impact than the second.

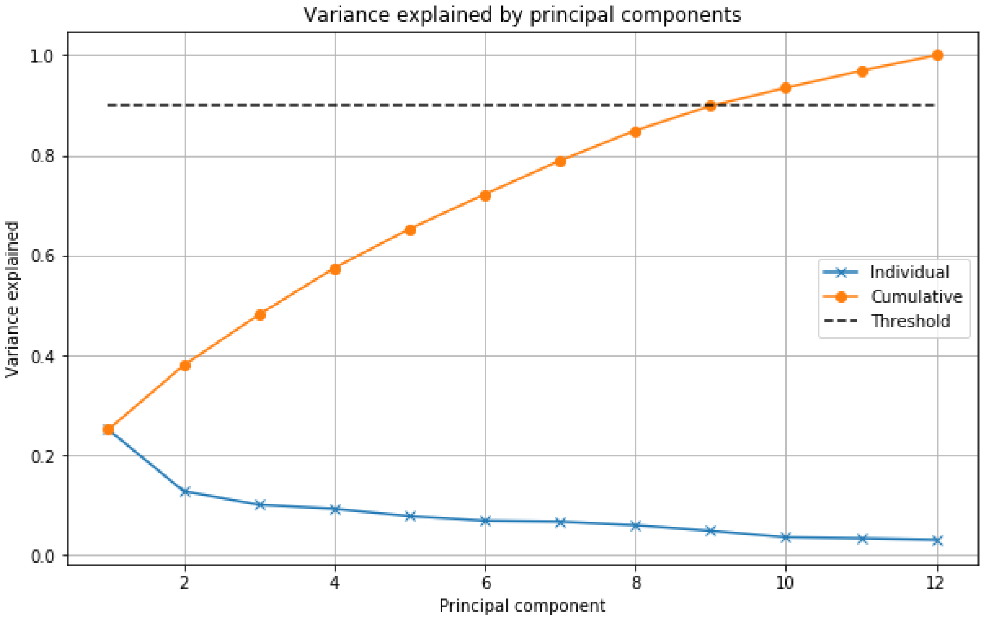


Fig:5 Variance explained by Principal Components

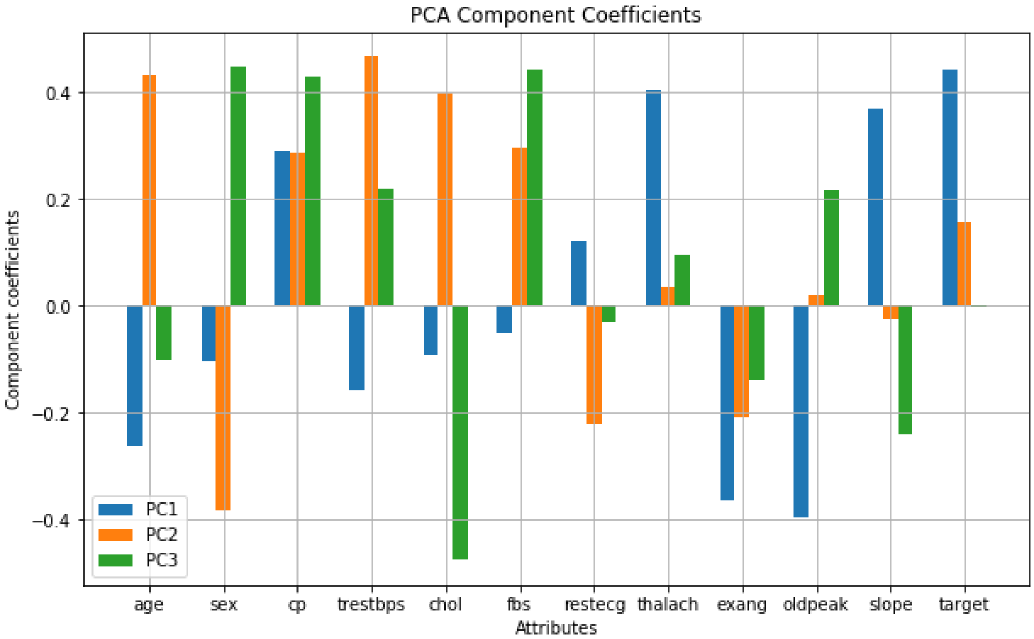


Fig:6 PCA component Coefficients

3.2.3 Data projection onto the principal components

PCA 1 has the most information. So the projections of PCA 1 and 2 are very high or low due to the change of their eigenvalues after the projections of them.

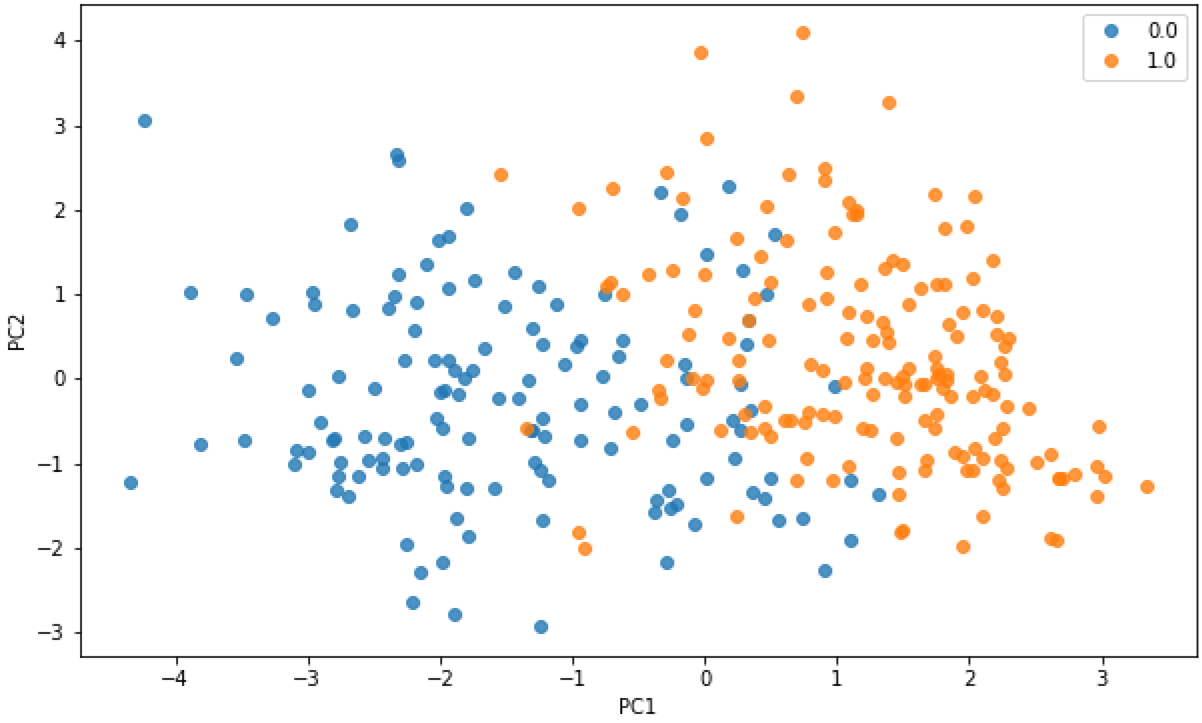


Fig:7 Projection on the first two principal components

In figure 7, the data projected onto the first 2 principal components is shown. It is observed that the data are somehow separated into two groups so we can use the classification method.

4. Conclusion

1. The first thing is that with PCA, the goal is to protect our data in a lower dimension so we can get more insights, better perspective, and less complexity. Also in that way, the visualization of our data is better because there are only 2 dimensions and not as much as the value of our attributes. To get 90% usage of our data, the variance explained was given from 9 principal components. Furthermore, an important result is the reduction in the size of our data.

2. The second thing we learned is that it is very important to check the scales of data. In our data set the magnitudes are different in different attributes, some of them are between 0 and 3 and others are between 100 and 200. If we don't consider any techniques before using PCA, we will get a result with a lot of errors. So, we used the script to subtract the means and normalize each attribute by further dividing each attribute by its standard deviation, only in this way can the

first principal component will not be highly driven by some attributes which have very large variance compared to others. This is shown in Fig:8.

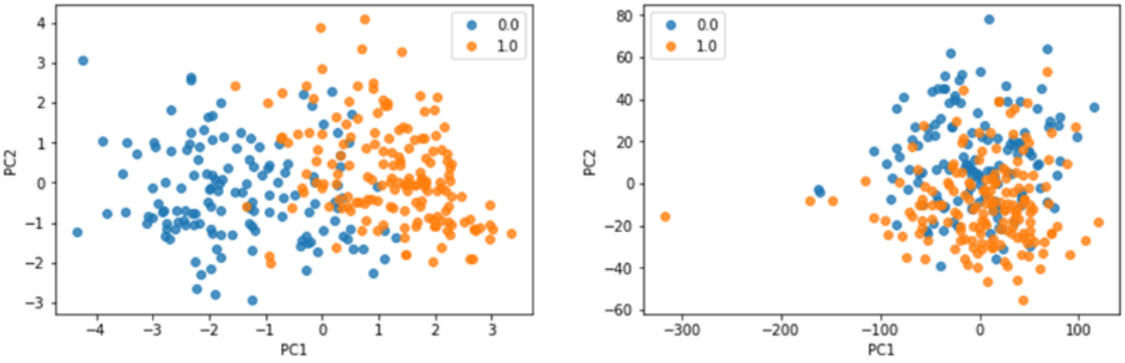


Fig: 8 Zero-Mean and unit variance projection vs Zero-Mean Projection

3. The third thing is on observing our data, we conclude that our data is based on the classification technique. Our PCA results are divided into two different clusters and it clearly classifies our class label ¨Target¨ i.e. the existence of the disease.

References

[1] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310

[3] https://en.wikipedia.org/wiki/Data_transformation_(statistics)

[4] Introduction to Machine Learning and Data Mining, Technical University of Denmark. Tue Herlau, Mikkel N. Schmidt and Morten Morup