Danmarks
Tekniske
Universitet

Introduction to Machine Learning and Data Mining - Project 1

Authors

Ion Chetratru - s204710
Irene Abigail Sotomayor Munoz - s205720
Zihao Fu - s202533

Distribution of work	Ion	Abigail	Zihao
1. Description of data set	40%	60%	-
2. Description of attributes in the data set	60%	40%	-
3.1 Data visualization	45%	30%	25%
3.2 PCA analysis	45%	30%	25%
4. Discussion	20%	50%	30%
5. Problem 1 and 6	-	-	100%
6. Problem 2 and 5	-	100%	-
7. Problem 3 and 4	100%	-	-

4th October 2022

1 Description of Data set

The data chosen for this project is the South African Heart Disease. The data consists of a total of 10 attributes and 462 observation. The purpose of our project is to analyse and find out if there exists a relationship between one's lifestyle i.e tobacco and alcohol consumption, and a heart disease problem.

The data set was obtained via the webpage "The Elements of Statistical Learning".[3] However, the original data was extracted from a larger data set of a medical journal.

The original data set was used in a medical journal in 1893 [4] for the sole purpose of investigating whether there were any factors that induce coronary heart disease. The study was held in the south-western Cape rural community. In this study the data was bigger as it compared the risk factor for men and women. The studies shoes that the population aged 44 years or older present the majority of of the high risk factors such as hypercholesterolaemia, hypertension and smoking as well as some low risk factor such as obesity and coronary-prone behaviour.

In the next report for the classification part, our goal and interest would be in finding out whether a person will have or not a coronary heart disease based on the values of systolic blood pressure, and the amount of cumulative tobacco the person has had in their life. For the regression part we are interested in predicting the value of low density lipoprotein cholesterol of a person given the measurements of the other attributes e.g tobacco, obesity.

In order to achieve this goal we would like to transform some of our data. For example, the value given for *famhist* is either "Present" or "Absent" and for analysis purposes we want to give them binary values where 0 is *Absent* and 1 is *Present*.

2 Attributes of the Data set

In this section we are going to further understand and describe the data given in the data set. For this reason we start by presenting a table where abbreviations, types of attributes and units of measurement are shown for each attribute.

Table 1: Description of the attributes in the data set

Attribute	Abbreviation	Description and type of attribute	Measure
Systolic blood pressure	sbp	continuous and ratio	mmHg
Cumulative tobacco	tobacco	continuous and ratio	kg
Low density lipoprotein cholesterol	ldl	continuous and ratio	$\frac{mmol}{L}$
Adiposity	adiposity	continuous and interval	Percentage of body fat
Family history of heart disease	famhist	discrete and nominal	'Absent' or 'Present'
Type-A behaviour	typea	continuous and interval	Bortner rating scale
Obesity	obesity	continuous and interval	BMI
Current alcohol consumption	alcohol	continuous and ratio	Alcohol concentration
Age on set	age	discrete and ratio	years
Coronary heart disease	chd	discrete and nominal	'1' = yes or '0' = no

Table 1 show the different attributes that are going to be used for the following report. For better understanding of the data it is worth mentioning that those people with a BMI > 30 are considered to be obese. Those who have and sbp $\geqslant 160$ are considered to have hypertension. Those who have a value typea > 55 are considered to have type-A behaviour. Finally, people with ldl $> 6{,}5$ are considered to have hypercholesterolaemia.

As for corrupted data or missing values. We do not encounter any missing values in the data set. All the values seem to be reasonable and for this reason we would argue that there is not corrupted data. However, it is mentioned that some measurements were taken after the person has undergone treatment. For example some people have had a heart disease and undergone diet treatment, after the treatment measurement for this data set were taken. For that reason if any correlation is found in the data there would be some values which do not follow the correlation.

We are now going to take a look into the summary statistics for the data set. Therefore we have calculated for each attribute the mean value (μ) the standard deviation (σ), the minimun and maximun value and the median. This is shown in Table 2

Table 2: Summary statistics of data set

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
μ	138.33	3.64	4.74	25.41	0.42	53.10	26.04	17.04	42.82	0.35
σ	20.50	4.59	2.07	7.78	0.49	9.82	4.21	24.48	14.61	0.48
min	101	0	0.98	6.74	0	13	14.7	0	15	0
50%	134	2	4.34	26.12	0	53	25.81	7.51	45	0
max	218	31.2	15.33	42.49	1	78	46.58	147.19	64	1

3 Data Visualization

The first thing we would like to investigate in our data is whether there are outliers or not. In order to do this we will plot a box plot of each attribute and by definition we will find outliers by looking at the points that lie outside the whiskers of the box plots.

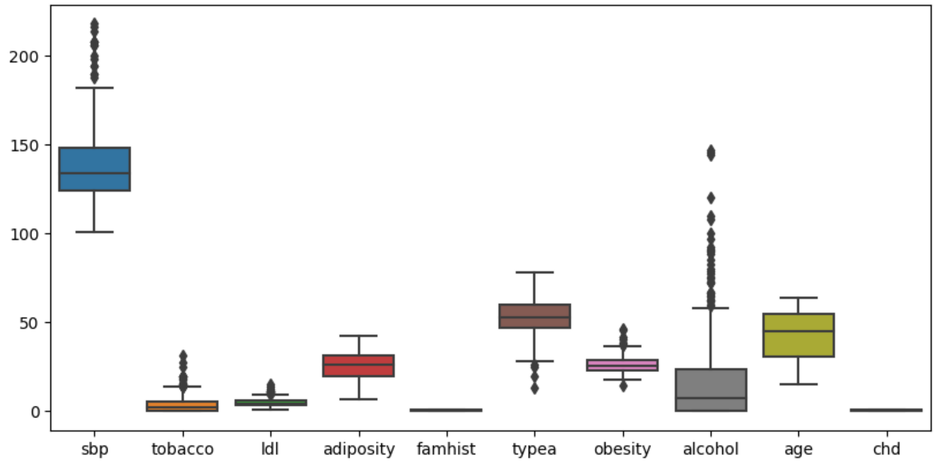


Figure 1: Boxplots of all attributes

Figure 1, shows that some attributes such as (*sbp*, *tobacco*, *typea*, *obesity*, *alcohol*), have some outliers in their data. However, if we take a closer look at this attributes and the dataset description, we can show that these attributes do not actually have to be treated as outliers since those point belong to the data set and do not represent any illogical data. In order to show this we can plot histograms and double check for outliers.

By plotting each of the above mentioned attributes as histograms Figure 2 we can say that the data does not have outliers. Finally it is important to notice that we can discard *famhist* and *chd* since they represent binary data.

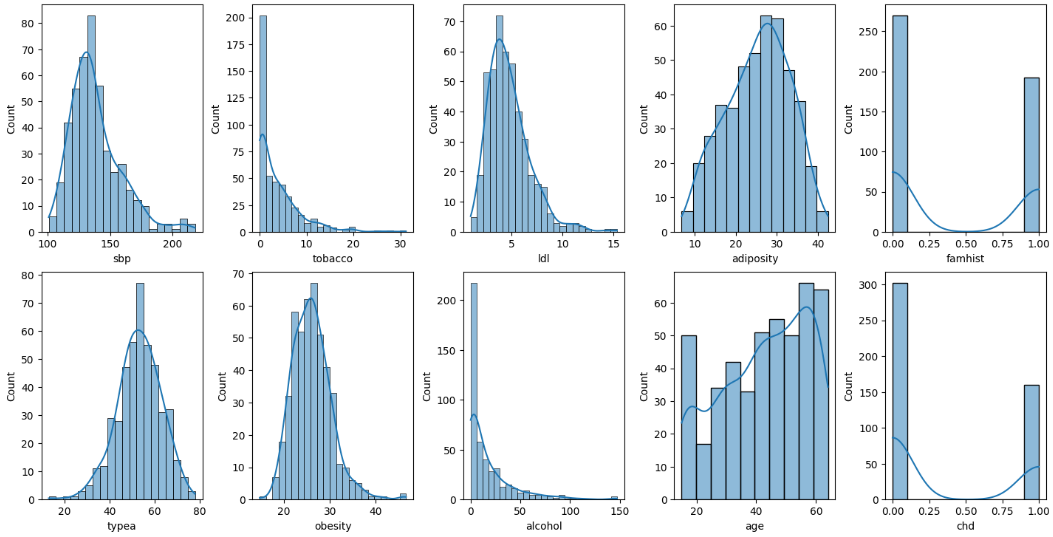
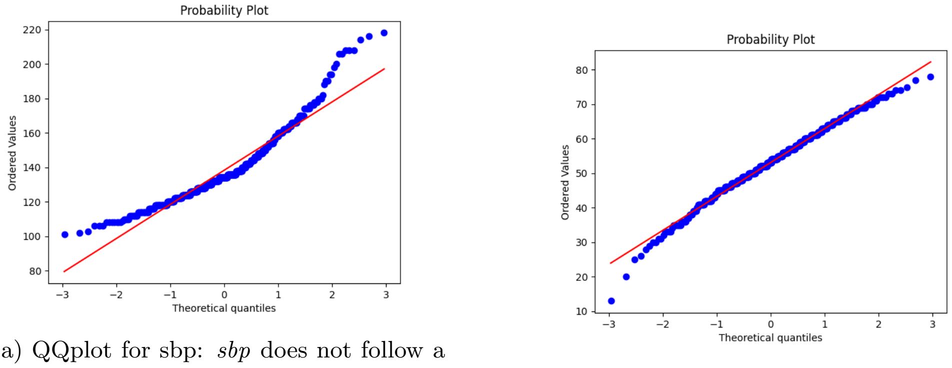


Figure 2: Histograms for all attributes

From Figure 1, we see that *sbp*, *ldl*, *adiposity*, *typea*, *age*, appear to be normal distributed, however, if we take a look at Figure 2 we clearly see that *sbp*, *ldl* and *age* do not follow a normal distribution, but *adiposity* and *typea* do. This can be confirmed by performing a normality test - *qqplot* in Figure 3.



(a) QQplot for sbp: *sbp* does not follow a normal distribution, the distribution seems to be *right-skewed*

(b) QQplot for typea: *typea* does follow a normal distribution

Figure 3: Normality test

As we can see in Figure 3 the graph to the left (3a) the points do not lie on the straight line which indicates that *sbp* does not follow a normal distribution. However, the graph on the right (3b) do contain points laying on the line and for that reason we can conclude that *typea* follows a normal distribution.

We have chosen to write down the distribution of the other attributes in the following table.

Attribute	Distribution
sbp	Right-skewed
tobacco	Exponential
ldl	Log-normal
typea	Normal
adiposity	Normal
obesity	Right-skewed
alcohol	Exponential
age	Uniform

We now proceed in finding the correlation between pairs of attributes. To do this we will plot a correlation matrix in which the coefficient indicates the correlation between the two attributes. From theory we know that values closer to 1 would show a strong correlation between attributes. The Matrix can be found on the figure below.

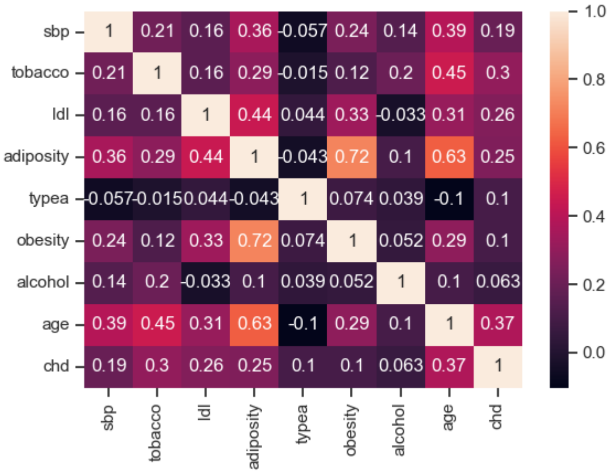


Figure 4: Correlation Matrix

From Figure 4 it can be seen that there is a **strong** correlation between the pairs of the attributes: (*adiposity, age*); (*adiposity, obesity*); and a **moderate** correlation between (*ldl, adiposity*) and (*age, tobacco*). Another way to show this correlation is by plotting a scatter plot in where a completely straight line will show strong correlation.

Figure 5 confirms the correlation as we can see that the scatter plot of the pairs mentioned above almost show a straight line.

After having described our data we will start performing PCA analysis. The data will be standardized by subtracting the mean value from the original value. In our case we will not divide by the standard deviation since different attributes follow different distributions which was shown previously. The standardization was done by the following formula:

$$\mathbf{z} = \mathbf{x} - \mu$$

Where $\mathbf{x}$ is the matrix containing our data, 1 column corresponds to an attribute, and μ is the mean of the data corresponding to each attribute.
After standardizing we performed singular value decomposition where we are going to obtain a matrix $\mathbf{V}$ such that each column corresponds to an eigen vector. We will also obtain a Σ matrix which is a diagonal matrix where each value is an eigenvalue corresponding to the eigen vector in the $\mathbf{V}$ matrix.

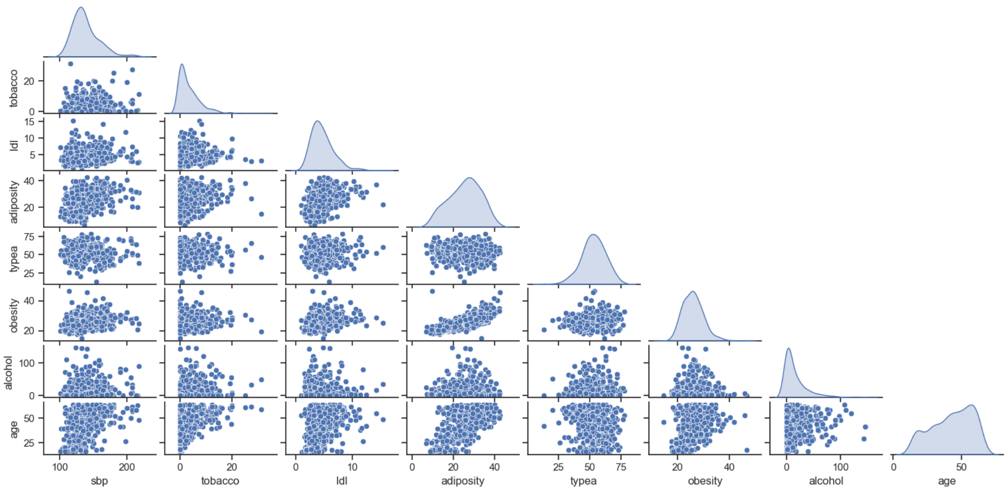


Figure 5: Scatter plots between attributes with significant correlation

Having performed PCA we can now find out the variance explained by the principal components. We chose to plot a graph in where the cumulative variance explained by the k'th principal components is found by using the formula:

$$\text{Variance explained} = \frac{\Sigma_{i=1}^{k}\sigma_i^2}{\Sigma_{i=1}^{M}\sigma_i^2}$$

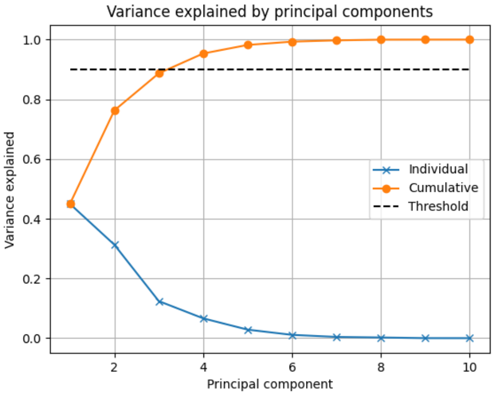


Figure 6: Amount of variance explained as a function of the number of PCs included

Figure 6 shows both the cumulative variance explained as well as the individual variance explained. It also shows that by taking the first three Principal Components *88.70%* of the variance from the overall data will be explained.

For this reason we choose to analyse the first three principal components. Its coefficients can be found in the table below where we have also chosen to write the attribute corresponding to each coefficient. In order to talk about each component we have chosen values of correlation to be significant if they are above $|0.4|$.

Table 3: Coefficients of the first three PCs

Attributes	PC1	PC2	PC3
sbp	**-0.4146**	**0.7738**	**0.4774**
tobacco	-0.0568	0.0418	-0.1153
ldl	-0.0064	0.0245	-0.0393
adiposity	-0.0966	0.1638	-0.2893
famhist	-0.0026	0.0022	-0.0068
typea	-0.0005	-0.0549	0.0892
obesity	-0.0309	0.0555	-0.0727
alcohol	**-0.8785**	**-0.4760**	0.0082
age	-0.2069	0.3734	**-0.8125**
chd	-0.0032	0.0050	-0.0086

In Table 3 we can see that the first principal component has a high negative correlation with the attributes *alcohol* and *sbp*, which means that the higher the values for these attributes, the smaller the projection on PC1 will be. So, if we choose to project the data, then those points to the left of the graph - assuming PC1 is on the x-axis - will have as a property a high *alcohol* and *sbp* value. Therefore this component can be seen as a measure of alcohol consumption and sbp value. Likewise PC2 show a high positive coefficient for *sbp* and a high negative coefficient for *alcohol* which means that a high projection onto PC2 will mean a high value of *sbp* and a low value of *alcohol* and vise versa. Therefore, this component can be seen as a measure of how high sbp value is in a person who does not consume much alcohol. Finnaly, PC3 shows a high positive coefficient value of *sbp* and a high negative value of *age* which means that a high projection onto PC3 will correspond to a high value of *sbp* and a low value of *age* and vise versa. So, we can say that the PC3 component captures those people who have a high value of *sbp* and a young age.

This high and low coefficients values can be easily be shown in Figure 7.

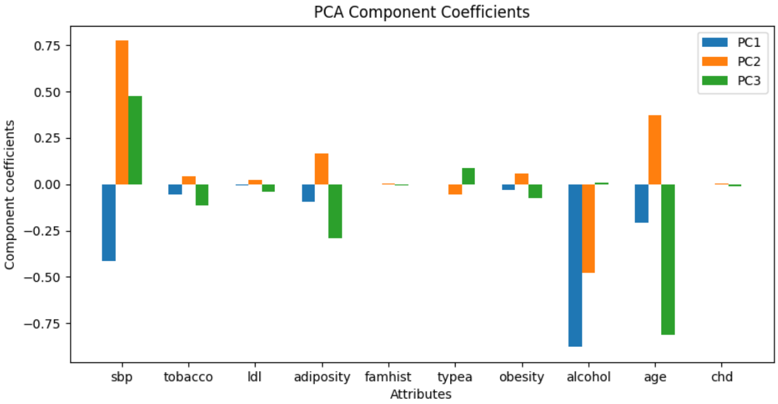
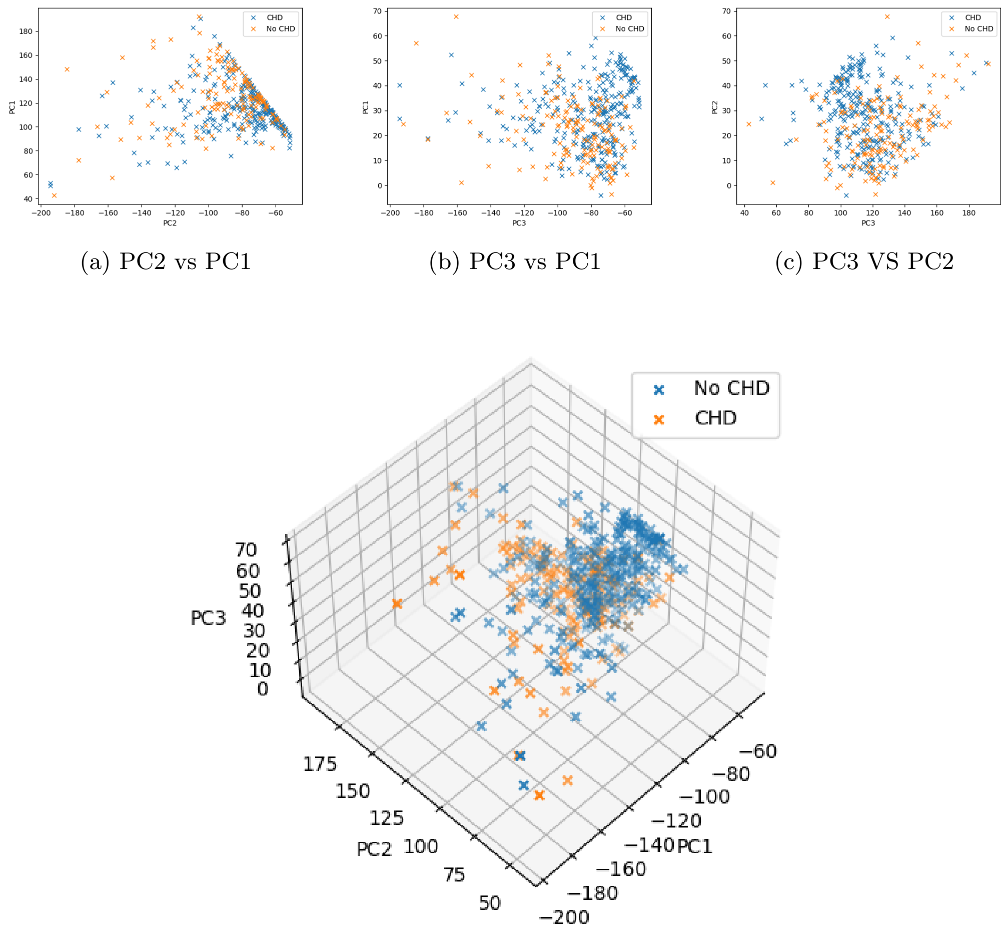


Figure 7: Coefficients of principal components

Now that we have analysed the different principal components we would like to plot them against each other and see how the data will look with respect to the different components.



(a) PC2 vs PC1

(b) PC3 vs PC1

(c) PC3 VS PC2

(d) 3D graph PC1. PC2 and PC3

Figure 8: Principal components plots

Figure 8 shows the different principal components plotted against each other. Firstly, in figure 8a we can see that all the data gathers in the right most part of the graph and forms a straight line. However for Figure 8b and Figure 8c the data is a bit more spread out which indicates there is less similarities between the observations as compared with the first graph.

We found in this case the principal components analysis would not help us to classify whether people have coronary heart disease or not. We did a research and found that PCA perform poorly on Non-Gaussian distribution data-set.[5] Therefore, We decided to transfer our Non-Gaussian distribution data to Gaussian distribution before PCA and standardize the data by standard deviation to see if we could get better result. *Sbp* and *obesity* follows right-skewed distribution, so we used reciprocal method to transfer them into normal distribution.[2].Using box-cox[1] method to process *tobacco*, *alcohol* and *ldl* data to get a normal distribution. For age, it follows uniform distribution and is hard to transfer to normal distribution. We tried multi methods, and decided to use QuantileTransformer which provided by sklearn to transfer it into normal distribution.

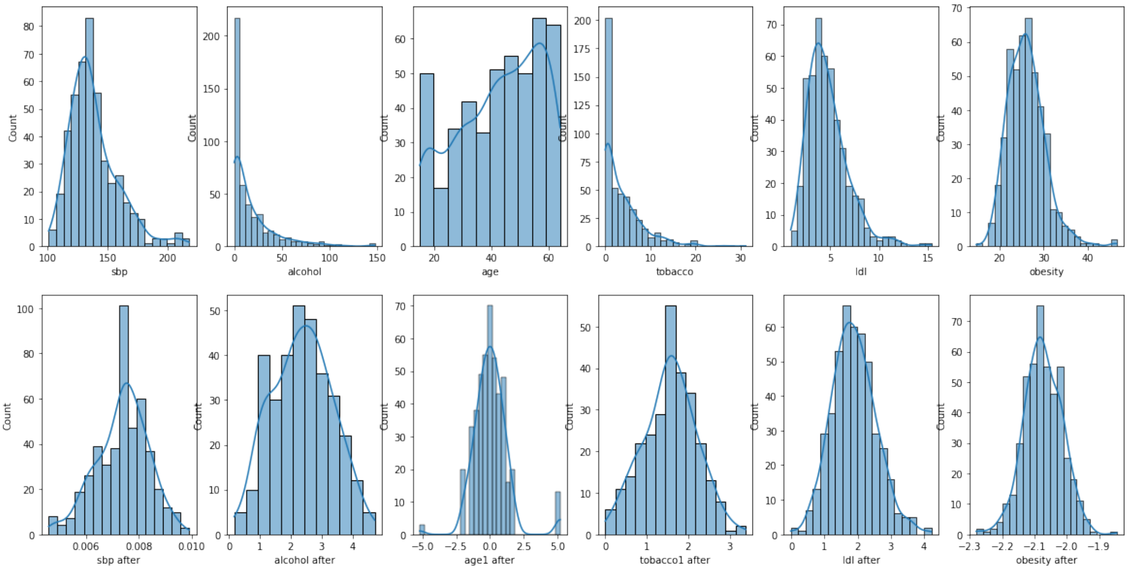


Figure 9: Histogram before and after transferring

Figure 9 shows the the histogram of Non-Gaussian distribution data in our data-set before and after transferring

After processing our data, we standardize the data by standard deviation. We found the first six Principal Components needed to be taken to explain 80.30% of the data. Compared to the previous PCA result, we need more Principal Component at this time. However, after visualizing and analyzing the Principal Components against each other, we found this new Principal Components would be helpful for classifying whether people had or did not have a coronary heart disease.

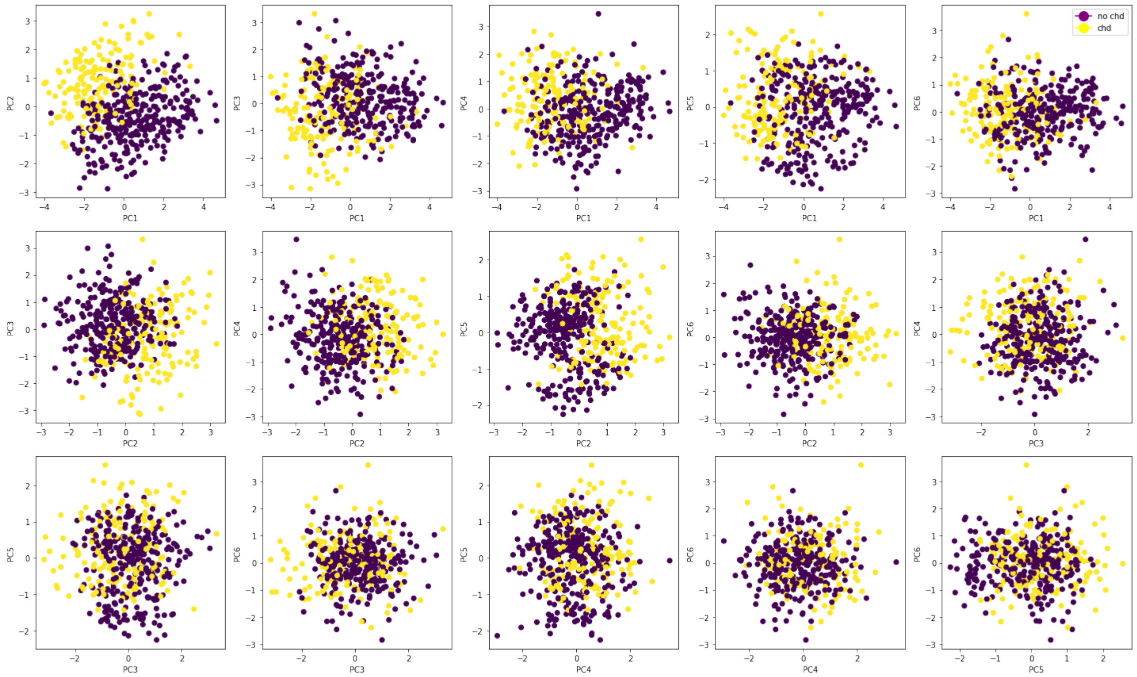


Figure 10: Principal components plots

Figure 10 shows the different principal components plotted against each other. We found chd and no chd can be distinguished clearly in many plots i.e. PC1 vs PC2, PC1 vs PC3, PC2 vs PC5 and so on. It indicates that the new Principal Components can be used for classification task.

4 Discussion

Finally, in the previous section we have chosen to make a distinction between the people that had and did not have a coronary heart disease - chd - and we can say that in this case the principal component analysis without dividing by standard deviation would not help us achieve the goal of classifying people based on data since the principal components do not have a correlation with the attribute *chd*.

However, if we chose to make the distinction between those who have a high pressure and a low pressure we would be able do find a clear dividing line in the data set as the principal components one and two capture this attribute. This can be visualized in the following figure.

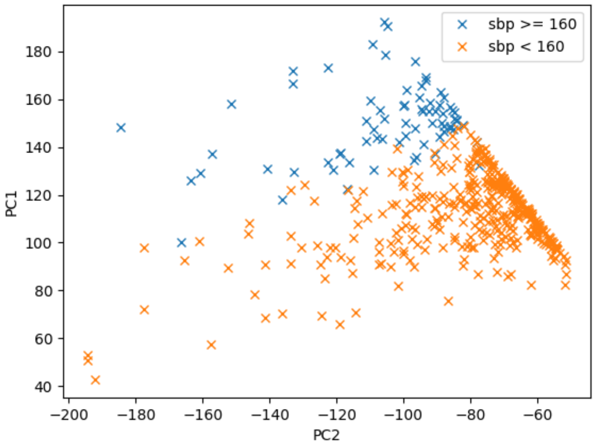


Figure 11: PC1 vs PC2

Figure 11 shows that we would be able to make a classification model on this data set if we wanted to classify people based on their value of blood pressure but this is not our goal.
After normalizing the data and dividing it by standard deviation, although we cannot reduce the dimension of the data-set as low as we do in the model without dividing by standard deviation, principal component analysis would help us to classify whether people had a coronary heart disease or not but another machine learning model would be more helpful for us since in our case our PCA did not perform well in terms of reducing dimensions.

Two important things we learned from the data. The first thing is principal component analysis really performs poorly on Non-Gaussian distribution data-set sometimes. The pre-processing of data for machine learning is important and can change a lot. The second thing is when using PCA to reduce data-set dimension, the different methods of data standardization would lead to different results. Choosing how to standardize or normalize the data is critical.

5 Exam problems

5.1 Question 1

The description of x_1(Time of day) is a coded number of 30-minute interval, so the x_1(Time of day) is interval. x_6(Traffic lights) shows the rate of broken traffic lights, so x_6(Traffic lights) is ratio. y(Congestion level) is a rank(level) of congestion so y(Congestion level) is ordinal.

Correct answer: D.

5.2 Question 2

For this question we want to find out what is the p-norm distance for the right p corresponding to the vectors

$$\mathrm{x}_{14} = \begin{bmatrix} 26 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{x_{17}} = \begin{bmatrix} 19 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

To do this we use the formula

$$d_p(x,y) = \begin{cases} (\Sigma_{i=1}^{M} |x_i - y_i|)^{\frac{1}{p}} & 1 \leq p < \infty \\ max\{|x_1 - y_1|, |x_2 - y_2|, ...|x_M - y_M|\} & p = \infty \end{cases} \tag{1}$$

A. for p = ∞, d_p(x,y) = $max\{|26 - 19|, |2 - 0|\} = 7.00$

B. for p = 3, d_p(x,y) = $\sqrt[3]{(26-19)^3 + (2-0)^3} = 7.052$

C. for p = 1, d_p(x,y) = $7 + 2 = 9$

D. for p = 4, d_p(x,y) = $\sqrt[4]{(26-19)^4 + (2-0)^4} = 7.01$

Correct answer: A.

5.3 Question 3

The variance explained by the first k PCs can be found using the formula:

$$\text{Variance explained} = \frac{\Sigma_{i=1}^{k} S_{i,i}^2}{\Sigma_{i=1}^{M} S_{i,i}^2}$$

A. Variance $PC_{1-3} = \dfrac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.87 > 0.8$

B. Variance $PC_{3-5} = \dfrac{11.48^2 + 10.03^2 + 9.45^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.48 < 0.51$

C. Variance $PC_{1-2} = \dfrac{13.9^2 + 12.47^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.52 > 0.5$

D. Variance $PC_{1-3} = \dfrac{13.9^2 + 12.47^2 + 11.48^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.72 > 0.7$

Hence, the correct answer is A.

5.4 Question 4

High values of positive coordinates and low values of negative coordinates will result in a more positive projection on the Principal Component. *(and vice versa for negative projections)*

$$PC{=}\left|\begin{array}{c}\text{Time of day}\\ \text{Broken Truck}\\ \text{Accident victim}\\ \text{Immobilized bus}\\ \text{Deffects}\end{array}\right|$$

A. sign($\mathbf{PC_5}$)= $\{+,+,-,-,+\}$: low values of *Time of day*, *Broken Truck* and *Deffects*, and high values of *Accident Victim* and *Immobilized bus* will result in a negative projection on *PC5*. **Wrong**.

B. Similar to A, **wrong**.

C. The projection on PC_4 will be positive.**Wrong.**

D. sign($\mathbf{PC_2}$)= $\{-,+,+,+,+\}$: low value of *Time of day*, and high values of *Broken Truck*, *Accident victim* and *Deffects* will result in a positive projection on **PC2**. **Correct**.

Answer: **D**

5.5 Question 5

In this question we want to find the right Jaccard similarity. For this we first do a term-document representation

	the	bag	of	words	representation	becomes	less	parsimoneous	if	we	do	not	stem
s$_1$	1	1	1	1	1	1	1	1	0	0	0	0	0
s$_2$	1	0	0	1	0	0	0	0	1	1	1	1	1

To find Jaccard similarity we use the formula

$J(\mathbf{x},\mathbf{y}) = \dfrac{f_{11}}{\mathbf{k} - f_{00}}$, where k is the number of terms in the term-document representation.

Therefore, Jaccard similarity is $\dfrac{2}{13} = 0.154$.

Correct answer: A.

5.6 Question 6

$p\left(\hat{x}_2 = 0 \mid y = 2\right) = p\left(\hat{x}_2 = 0, \hat{x}_7 = 0 \mid y = 2\right) + p\left(\hat{x}_2 = 0, \hat{x}_7 = 1 \mid y = 2\right)$

From table 2 we can get $p\left(\hat{x}_2 = 0, \hat{x}_7 = 0 \mid y = 2\right) = 0.81$, and $p\left(\hat{x}_2 = 0, \hat{x}_7 = 1 \mid y = 2\right) = 0.03$,

So $p\left(\hat{x}_2 = 0 \mid y = 2\right) = 0.81 + 0.03 = 0.84$

Correct answer: B.

References

[1] George EP Box and David R Cox. 'An analysis of transformations'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2 (1964), pp. 211–243.

[2] Pratap Dangeti. *Statistics for machine learning*. Packt Publishing Ltd, 2017.

[3] *Elements of Statistical Learning*. URL: https://hastie.su.domains/ElemStatLearn/.

[4] A.J.S.BENAD P.C.J.JORDAAN J.P.KOTZ P.L.JOOSTE J. J.FERREIRA J.E.ROSSOUW J.P.DUPLESSIS. 'Coronary risk factor screening in three rural communities'. In: *SA Medical Journal* 94 (1983), 430–436.

[5] Jonathon Shlens. 'A tutorial on principal component analysis'. In: *arXiv preprint arXiv:1404.1100* (2014).