TECHNICAL UNIVERSITY OF DENMARK

Robert Mundziel (s222885)
Alberto Vendramini (s232103)
Emilie Grønborg Kristensen (s224923)

Project 1

02450 - Introduction to Machine Learning and Data Mining Fall 23

January 18, 2024

Name	Section 1	Section 2	Section 3	Exam Questions
Robert	30%	30%	40%	33%
Alberto	30%	40%	30%	33%
Emilie	40%	30%	30%	33%

1 Introduction and detailed description of the Data Set

The data collected allow us to examine the impact of various elements related to people's lives, such as smoking, alcohol consumption, and family medical history, on the chance of contracting heart disease. The data comes from a dataset from the South African Medical Journal by Rousseauw et al from 1983.[1] The project would try to answer the question "what is the probability of heart disease given the patient's parameters?". This would make it possible to determine the chances of such an incident. In addition, the topic of the study could be to examine the impact of specific environmental elements on the likelihood of heart disease.

We were not able obtain an analysis of the data set. However, it is implied on the website "Elements of Statistical Learning"[2] that high blood pressure, high cholesterol, obesity, smoking, excessive alcohol intake, old age[3], family history of heart disease [4] can lead to coronary heart disease. On the other hand there is no clear evidence in recent studies that Type A personality behaviour impacts heart disease. There is even some evidence that shows that the tobacco industry payed for a lot of the research involving Type A personality's effect on heart disease to undermine the role tobacco plays on the risk of heart disease. [5]

We wish to perform a supervised classification to see whether a person will get coronary heart disease (CHD) or not. We wish to explain CHD by the rest of the attributes in the data set except the 1st column "Row name".

The different attributes in the data set are described below:

Attribute name	**Explanation**	**Data Type**
SBP	Systolic Blood Pressure	Continuous, ratio
Tobacco	Intake pr person	Continuous, ratio
LDL	Low Density Lipoprotein Cholesterol	Continuous, ratio
Adiposity	Body Fat Percentage	Continuous, ratio
FamHist	Presence of heart disease in the family	Discrete, nominal
TypeA	Point on Bortner Short Rating Scale of Type-A personality	Discrete, interval
Obesity	BMI	Continuous, ratio
Alcohol	Current Alcohol Consumption	Continuous, ratio
Age		Discrete, ratio
CHD	Coronary Heart Disease	Discrete, nominal

Table 1: Explanation of all attributes

The data does not contain missing values and we cannot conclude that any of the data points are outliers. We did observe one observation for LDL that was 15 mmol/l which seemed to high, but we do not know enough to conclude that it is an outlier. The same was the case for the maximum value in alcohol consumption, which was 147.19.

[1] https://hastie.su.domains/ElemStatLearn/

[2] https://hastie.su.domains/ElemStatLearn/

[3] https://www.nia.nih.gov/health/heart-health-and-aging

[4] https://www.heartfoundation.org.au/bundles/your-heart/family-history-and-heart-disease

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3477961/

Attribute name	**Unit**	**Boundaries**
SBP	mmHG	]40;400[
Tobacco	Kg	Unknown
LDL	mmol/L	]0;20[
Adiposity	Percentage	]0;100[
FamHist	None	1: "Yes", 0: "No"
TypeA	None	[12;84]. More than 55 i considered "Type A behaviour"
Obesity	Kg/m^2	]0;300[
Alcohol	Probably volume/time	Unknown
Age	Years	[0;125]
CHD	None	1: "Yes", 2: "No"

Table 2: Basic statistics of all attributes

1.1 If you need to transform the data in order to carry out these tasks, explain roughly how you plan to do this

In order to carry out classification based on having Coronary Heart Disease we don't really need to change the data so much. The main thing to do is take out the last column, which is the main information and the attribute that we want to predict, and standardize the data (as written in Section 2.1). We also found a high correlation between "Adiposity" and "Obesity" so we can think to exclude one of them.

1.2 Basic Summary Statistics

Underneath you will see the summary statistics:

Attribute name	SBP	Tobacco	LDL	Adiposity	FamHist	TypeA	Obesity	Alcohol	Age	CHD
Range	117	31.2	14.35	35.75	1	65	31.88	147.19	49	1
Mean	138.3	3.64	4.74	25.4	0.416	53.1	26.0	17.0	42.8	0.346
Median	134	2.00	4.34	26.1	0	53.0	25.8	7.51	45.0	0
Standard deviation	20.5	4.59	2.07	7.77	0.493	9.81	4.21	24.5	14.6	0.476
Minimum value	101.0	0.0	0.98	6.74	0	13	14.7	0	15	0
Maximum value	218.0	31.2	15.33	42.49	1	78	46.58	147.19	64	1

Table 3: Basic statistics of all attributes

We will only tap into a few of the basic summary statistics. The first one to focus on, is that the median of CHD is 0 and the mean is 0.346, meaning that roughly $\frac{2}{3}$ of the people involved in the research project did not have coronary heart disease. On the other hand the mean of FamHist is 0.416, which implies that there were cases in the project where the person did not have a heart disease, but they had it in their family.
It may also be worth noting that the mean of Age is 42.8 with a standard deviation of 14.6 years and the mean of Obesity is 26 and median is 25.8 which is out of the normal area of BMI which ranges from 18.5-25. The population is therefore not that old, but their average BMI is high.

2 Data visualization based on PCA

2.1 Standardization

First of all, since our attributes have different scales, it is important to standardize them. To do it we subtract the mean of every attribute from that column's values and then divide by the Sample Standard Deviation.

$$\mu_j = \frac{\sum_{i=1}^{N} X_{i,j}}{N}$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{N} (X_{i,j} - \mu_j)^2}{N-1}}$$

$$S_{i,j} = \frac{X_{i,j} - \mu_j}{\sigma_j}$$

Where X is the raw data and S is the standardized data.

Since our goal is to do classification based on the last attribute, which is Coronal Heart Disease (Positive or Not), we will exclude that column from now on and consider it as our y vector.

2.2 Outliers

An outlier is an observation or data point that significantly deviates from the majority of the data in a data set. In other words, it's a data point that is either much smaller or much larger than most other data points in the data set. Before the Principal Component Analysis we checked if there were unexpected values. For doing so we utilized Box plots and Histograms to get a better view on suspect elements. It can be seen that there are no clear outliers, but surely there are elements which are far from the mean, for example in the tobacco and obesity attribute.

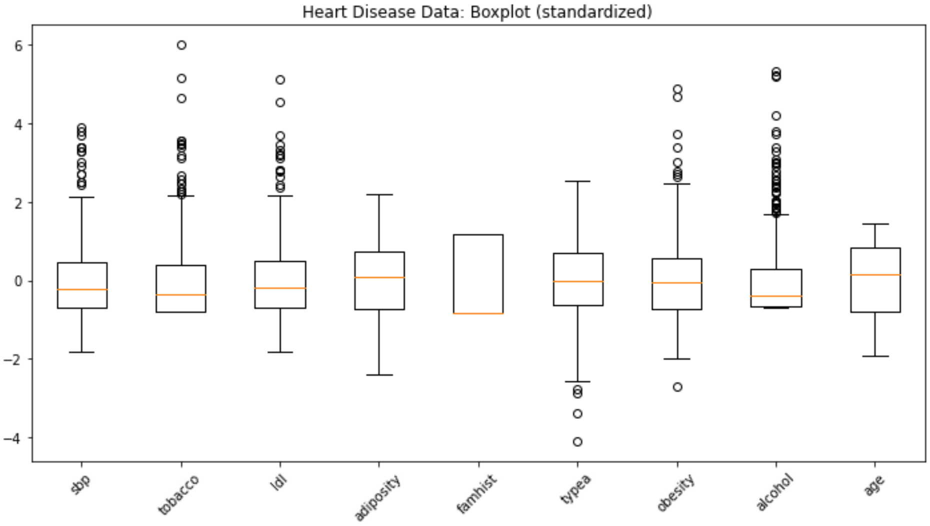

Figure 1: Box Plot of the attributes

Quality of data points that invalid data and outliers were already removed from the data set. We have not found any missing values or values that could be outliers. We decided to leave everything as it was since there is no evidence of mismeasuring or other type of errors.
We checked the odd elements and they all seem to be feasible measurements but with low likelihood, therefore we cannot exclude them.

2.3 Normal distribution of attributes

The normal distribution, also referred to as the Gaussian distribution or bell curve, is a fundamental statistical concept widely used in data analysis and modeling. In this report, we will delve into the characteristics and implications of a normal distribution within a given data set. Understanding the properties of a normal distribution is essential for making robust statistical inferences and predictions.

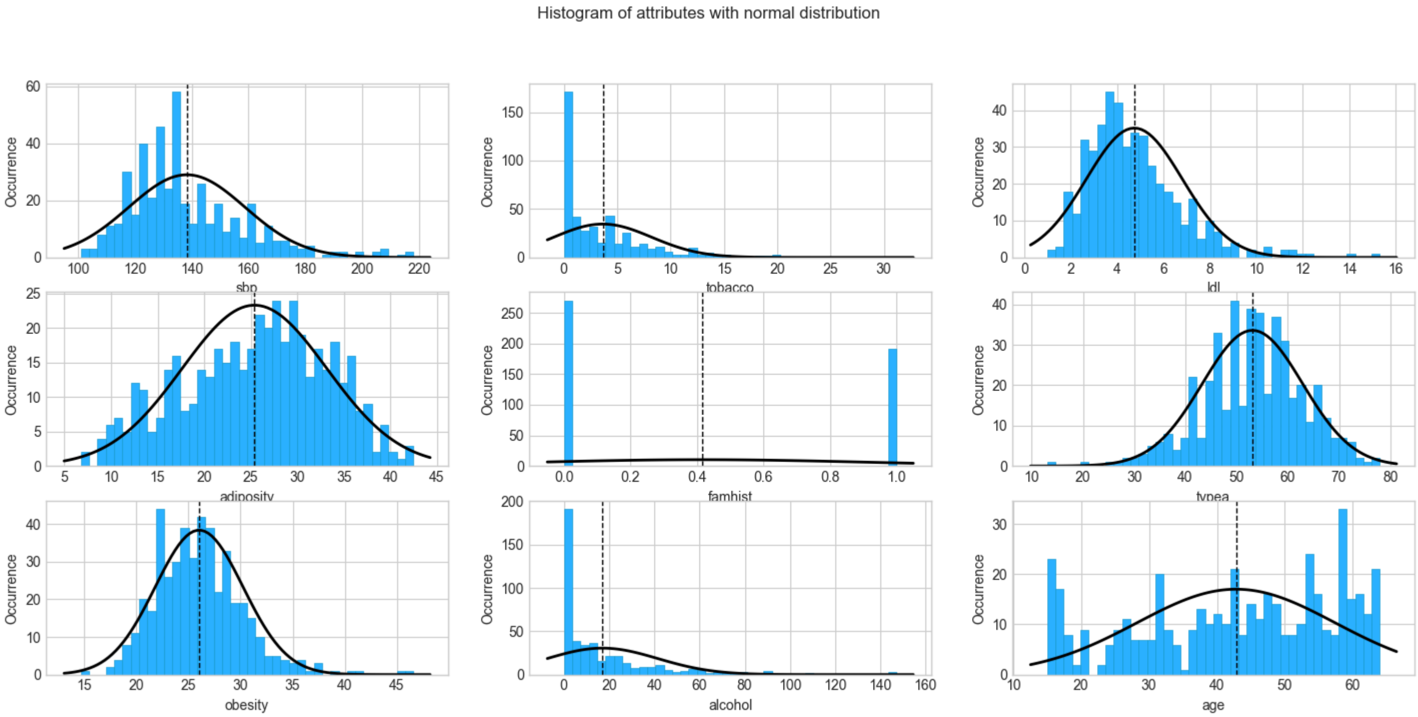

Figure 2: Histogram of attributes (normal distribution)

The chart above shows a histogram for each of the attributes included in the heart disease data set. You will notice that many of the distributions indicate that they are normally distributed. First up is the attribute "SBP," which describes the subject's blood pressure. It can be seen that they are arranged in an irregular normal distribution, for which many of the values from the groups with lower heart rates are apparent. The normal distribution, on the other hand, does not fit the consumption of tobacco. Many of the subjects are non-smokers, so the dominant value is 0. "LDL", or cholesterol results, indicate a right-skewed normal distribution with a mean value around 5.0. Adiposity also indicates an irregular normal distribution - mean values occur most often, while median values occur to a lesser degree. The history of heart disease contains only yes and no answers, so it cannot be concluded that this is a normal distribution. However, it can be indicated that most of the respondents' families did not face heart problems. Another attribute is personality type, which they indicate describes and generalizes people's behavior. It can be seen that the data is normally distributed. Next is BMI, which is weight divided by height squared. As can be seen, it too is normally distributed, with the mean around 25. The distribution of alcohol consumption is a similar case to that of tobacco consumption. Many of the people surveyed did not declare themselves to be alcohol drinkers. The last value studied is age, which is not normally distributed, and the average age of the subjects is greater than 40.

Data indicated as normally distributed retain the basic characteristics for a normal distribution. One of the defining features of a normal distribution is its symmetry. The distribution is symmetric around its mean, implying that the left and right tails of the distribution are mirror images of each other. Distributed data takes the shape of a smooth, symmetric bell curve. The peak of this curve corresponds to the mean, while the curve tapers off gradually as you move away from the mean in both directions. In the given dataset, you can point to attributes that distribute normally, which can enhance the accuracy and reliability of statistical inferences and predictions.

2.4 Correlation of variables

The dataset contains nine variables: sbp, tobacco, ldl, adiposity, famhist, typea, obesity, alcohol, and age. The correlation matrix provides insights into how these variables are related to each other, which is essential for understanding patterns and dependencies within the data.

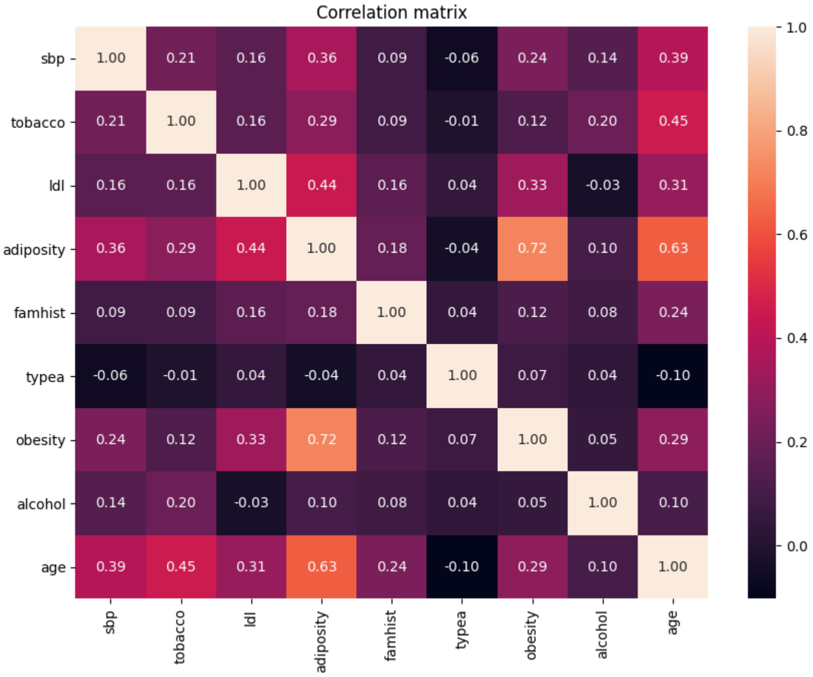

Figure 3: Correlation matrix

The correlation matrix, displayed above, represents the pairwise correlations between the variables. The values in the matrix represent the strength and direction of the linear relationships between pairs of variables. Correlation coefficients range from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. It's essential to assess whether the observed correlations are statistically significant. Further statistical tests may be needed to confirm the significance of these relationships. The strongest correlation in the matrix appears to be between the variables "obesity" and "adiposity" with a correlation coefficient of 0.72. This correlation is positive and moderate in strength. It suggests that there is a tendency for individuals who consume more food to have a higher adipose tissue, and vice versa. Also, age and tobacco use have a somewhat positive relationship, but it is not extremely strong. The weakest correlation in the matrix appears to be between the variables "typea" (Type A behavior) and "age" with a correlation coefficient of -0.10. This correlation is negative and weak. It suggests that there is a slight tendency for individuals with higher levels of Type A behavior to have less age, and vice versa. However, this relationship is not very strong, and the correlation is close to zero.

2.5 Data feasibility

Advantages:

- **Outliers:** The absence of outliers can be advantageous as it typically indicates that the data is relatively clean and free from extreme values that could adversely affect model training and performance
- **Normal Distribution:** Normally distributed attributes can simplify modeling because many machine learning algorithms assume that the data follows a normal distribution.
- **No Strong Correlations:** Lack of strong correlations (beside obesity and adiposity) between variables means that multicollinearity (high intercorrelations between predictors) is less likely to be a problem. This can simplify feature selection and interpretation.

Main goal is to implement supervised classification. It's important to consider whether the attributes are informative for predicting the target variable. Lack of correlation between variables does not necessarily imply that they are not useful for classification. While normally distributed attributes and lack of strong correlations can simplify modeling, they don't guarantee good model performance. The effectiveness of a classification model depends on the quality and relevance of the features

2.6 Variance explained by Principal Components

PCA is a widely used technique for reducing the dimensionality of data while retaining as much relevant information as possible. It achieves this by identifying and preserving the directions, known as principal components, along which the data varies the most.

PCA is fundamentally concerned with capturing the variance or spread of data. Variance measures how data points are dispersed from the mean or center. As you can see in Figure 4 the variance is very spread out on the principal components. The first one makes up for 32% of the variation and therefore explains the most of it. The following 8 principal components range from 13.3% to 1.9%.

We decided to put the threshold to 75%, so the variance is explained by 5 principal components.

This is what we will continue working with, since we would be including almost all principal components if choosing a larger threshold.

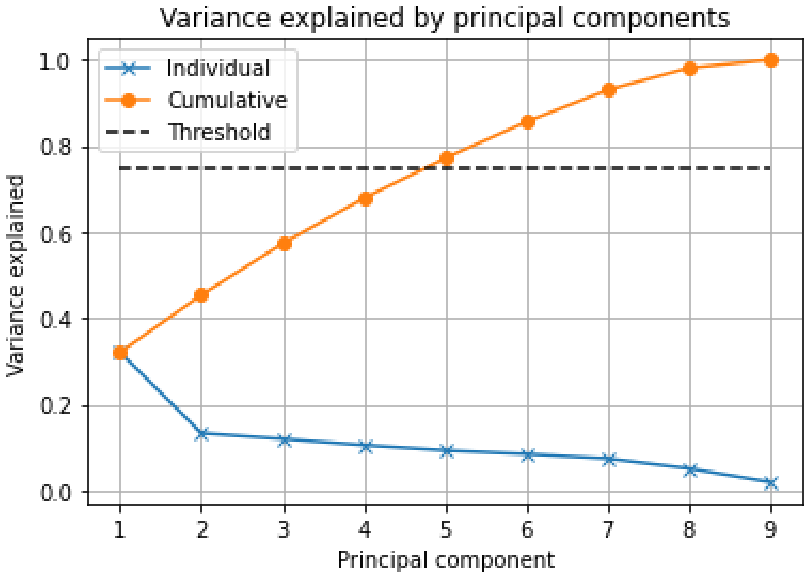

Figure 4: Variance of the Principal Components

Since we have to pick 5 out of 9 principal components to meet the threshold of 75% we can conclude that our data is not very structured. In the following sections we will try to produce more information about Principal Components using Coefficients and Data Projection.

2.7 Principal Components Coefficients

From the next plot you can clearly see the weight of every attribute in the first 5 principal components.

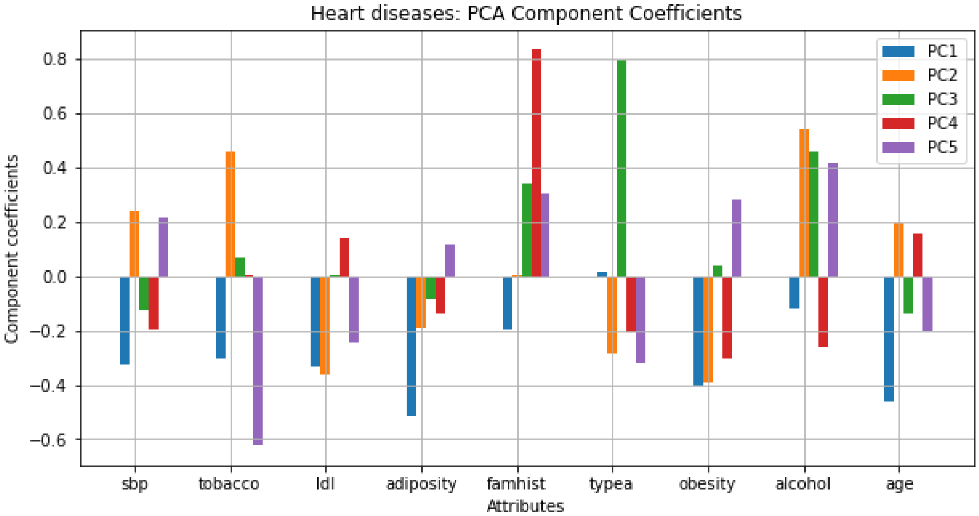

Figure 5: First 5 Principal Components attributes' weights

It can be seen that:

- PC1 has most of the weights negative, apart from "Type A" which is a low positive value.
- PC2 higher values are "Alcohol" and "Tobacco".
- PC3 is mostly influenced by "Type A" and secondly from "Alcohol" and "Family History"
- PC4 is mostly influenced by the "Family History" attribute and has negative weights in the others,exept for "Age" and "LDL"
- PC5 is mostly influenced by "Tobacco" which is negative and "Alcohol" which is positive

2.8 Data Projection

Main idae of Data Projection is to obtain valuable information about how the original data points are represented in a lower-dimensional space. From our analysis we need 5 Principal Components, it is not possible for us to plot it, the best that we can do is to just use the first 3, with the expense of a loss of accuracy. We tried rotating the 3D plot to see if there were sides were there was a good view.

As we expected we did not see a clear structure in the data, and there is zero separation from the Heart Disease patients and not. Unfortunately, PC couldn't find any linear combination of Principal Components.

We decided to check for any separation or structure based off two PCs. In orange are the patients whose CHD=1 and in blue whose CHD=0. As seen on the plots there are no two principal components describing the variance in a structured way. However, the first principal component does explain almost 32% of the variance, and it is possible to see that there is a tiny bit more of structure on plots where PC 1 appears as a variable. It supports our findings from the PCs variance analysis that led us to conclude that we need 5 principle components in order to explain the variance with a threshold of 75%. Thus, we cannot see any clear patterns by just plotting two principle components against each other.

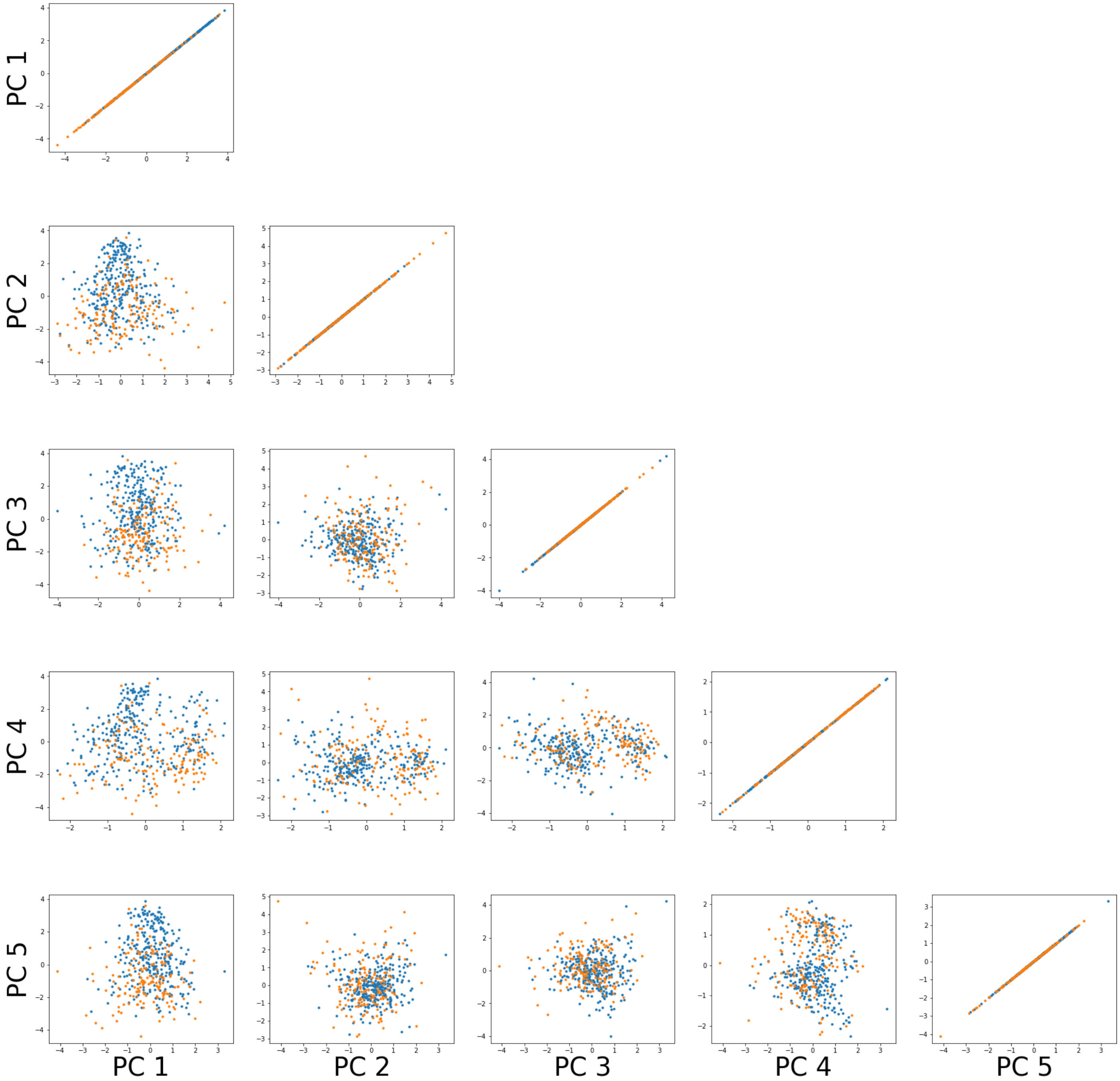

Figure 6: PC-PC plot

3 Learnings about the data

In this section, we will summarize the most important insights gained from the analysis of the dataset and evaluate the feasibility of our primary machine learning aim based on our visualizations and observations. The dataset contains information related to various aspects of people's lives, such as smoking, alcohol consumption, family medical history, and physiological attributes, whith the primary aim of exploring the probability of contracting coronary heart disease (CHD) based on these parameters.
Main goals:

- **Data Quality:** The dataset appears to be of relatively high quality, with no missing values or clear outliers. While there were some observations with extreme values, such as unusually high LDL levels and alcohol consumption, these values were not conclusive outliers.
- **Attribute Types:** We classified the attributes into various types, including continuous (ratio and interval), discrete (nominal and interval), and binary nominal. This categorization helped us understand the nature of the data.
- **Basic Summary Statistics:** We calculated and examined basic summary statistics for each attribute, shedding light on important characteristics of the dataset.
- **Data Standardization:** We discussed the importance of standardizing the data due to variations in scale among attributes, a crucial step for effective machine learning modeling.
- **Correlation Analysis:** We explored correlations between attributes, identifying both positive and negative relationships. This analysis helps us understand how attributes relate to each other, which can be valuable for feature selection and model building.
- **Variance Explained by Principal Components:** We performed Principal Component Analysis (PCA) and examined the variance explained by the principal components. We found that a substantial portion of the variance was distributed across several principal components, suggesting that the data lacks a clear, structured pattern.
- **Data Projection:** While we attempted to visualize the data using 3D plots based on PCA, we did not find a clear separation between individuals with and without CHD, indicating the complexity of the classification problem.

In conclusion, our analysis has provided valuable insights into the dataset, highlighting its strengths and challenges. While the data is of good quality and contains potentially relevant attributes, the lack of structured patterns may pose difficulties in achieving our primary machine learning aim. Further analysis and experimentation with different modeling approaches will be necessary to determine the feasibility of accurately predicting CHD based on this data. Our primary machine learning aim is to do classification on the Coronary Heart Disease attribute later, but it seems really hard to do with this data.

4 Solutions for exam problems

1. Answer D;
   The Time of Day attribute is the only one that varies. Since it is 30 minute intervals that progress during the day, each interval has a physical meaning (30 minutes) and it is ordered from low to high (1-27) which resets every day.

2. Answer A:
   We checked for the different options and found that if you use the max-norm distance, it corresponds to answer A: $d_p = inf(x_{14}, x_{18}) = max(|x_{14}[0] - x_{18}[0]|, ..., |x_{14}[6] - x_{18}[6]|) = max(|26 - 19|...) = 7.000$

3. Answer A:
   First of all we calculate the sum of all the variances.
   $\Sigma PC = 13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2 = 670.4047$
   Then we calculate the sum of the variances of the PC's we are working with
   $\Sigma PC_{1-4} = 13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 = 581.1022$
   Then we compute the percentage:
   $\sigma^2 = \frac{\Sigma PC_{1-4}}{\Sigma PC} = 86.6\%$

4. Answer D:
   The second Principal Component has negative weight only in "Time of Day" and positive in all the other ones, so having a low value in "Time of Day" and high values in the others will result in a positive value.

5. Answer A:
   the: 1, bag: 0, of: 0, words: 1, representation: 0, becomes: 0, less: 0, parsimonious: 0, if: 0, we: 0, do: 0, not: 0, stem: 0
   $f_{11} = 2, f_{00} = (20000 - 13), K = 20000$

   $\frac{f_{11}}{K - f_{00}} = \frac{2}{(20000-(20000-13))} = 0.153$

6. Answer B:
   $p(x_2 = 0|y = 2) = p(x_2 = 0, x_7 = 0|y = 2) + p(x_2 = 0, x_7 = 1|y = 2)$
   $= 0.81 + 0.03 = 0.84$