Danmarks Tekniske Universitet

DTU

02450
Machine Learning and Data Mining
Project 1

Written by:

Ditte B. Gilsfeldt *(s210666)*

Mads B. Vejbæk *(s225783)*

William K. Stentzer *(s214645)*

Who did what				
	Section 1	**Section 2**	**Section 3**	**Exam Question**
Ditte	40%	20%	40%	33.3%
Mads	40%	40%	20%	33.3%
William	20%	40%	40%	33.3%

Date: 03.10.2023

Contents

Description

A description of the data set

The dataset is South Africa - Heart Disease[1], it consists of 462 observations with 10 attributes each. The dataset encompasses a cohort of males with ages ranging from 15 to 64. And covers three communities in the Western Cape of South Africa that have been heavily affected by coronary heart disease(CHD). The intent of the study was to find the reason for the high occurrence of CHD. By understanding the underlying reasons, being able to help the affected areas. The dataset is a subset of a larger dataset, taken from a medical study called "Coronary risk factor screening in three rural communities- the coris baseline study"[5].

The previous analysis of the data done in the original study found that the majority of the tested population had one or more major risk factors. They especially found a strong age trend for total cholesterol, blood pressure, and smoking. That would indicate the amount of risk factors rise with age. But they also mention that the cut off points for these three factors may be lower in a younger population. The cohort also had higher average serum cholesterol and systolic blood pressure values than those of their rural-age peers in the US. Smoking was also slightly more widespread than in the US comparatively. The elevated occurrence of significant risk factors is further intensified by comparably high occurrences of minor factors, including obesity, hyperuricemia, and coronary-prone behavior. There is no offsetting presence of potentially beneficial factors like physical activity to counteract the potential adverse effects of these risk factors. Which also contributes to the high prevalence of existing coronary disease. However, they did not know to what degree this cohort is representative of other African communities. They presume that the increased public awareness of Ischemic Heart Disease will reshape the habits that contribute to the risk factors. They then conclude with a recommendation for targeting the communities as a whole with measures of prevention. This is because the people at risk have been found in this study to be a majority of the population. Therefore targeting the entire community would be the most feasible solution [5].

A detailed explanation of the data attributes

The dataset, has no missing values or corrupt data and the objective for training a classification model using this dataset, would result in a possibility to predict if a person has CHD given the 9 other attributes[1]. Using a regression model it could predict the likelihood of a person having CHD, also given the 9 other attributes. However, in these scenarios, it may be considered illogical to have tested for the other 9 data points without having tested for coronary heart disease itself.

In the dataset, there are roughly two CHD controls per case. Some of these were done after the people received treatment to reduce blood pressure, as well as other treatments to reduce the risk factors. The treatments were done after their CHD incident. However, this still makes the dataset representative of the community since the reason for interest is the elevated level of CHD. It would be logical that those people in such communities received treatment.

The table below contains a description of dataset attributes.

Table 1 - The data set contains information about:[1]

Systolic Blood Pressure ***(sbp)***
This ratio variable is continuous, and it represents the systolic blood pressure of a person.

Cumulative Tobacco (kg) ***(tobacco)***
This ratio variable is continuous, and it represents the cumulative tobacco consumption in kg.

Low Density Lipoprotein Cholesterol ***(ldl)***
This ratio variable is continuous, and it represents the low-density lipoprotein cholesterol level of a person.

Adiposity ***(adiposity)***
This ratio variable is continuous, and it represents the amount of fat on a person as a percentage of overall bodyweight.

Family History Of Heart Disease (Present, Absent) ***(famhist)***
This nominal variable is binary and represents the presence or absence of heart disease in a family.

Type-A Behavior ***(typea)***
This interval variable is continuous, and it represents the behavior in Type A, which is characterized by impatience, a sense of urgency, competitiveness, and hostility.

Obesity ***(obesity)***
This ratio variable is continuous, and it represents the individual obesity level in a numerical way using BMI.

Current Alcohol Consumption ***(alcohol)*** This ratio variable is continuous, and it represents the yearly consumption of alcohol.

Age At Onset ***(age)***
This ratio variable is continuous, and it represents the age when the heart disease was diagnosed.

Respinse, Coronary Heart Disease ***(chd)***
This nominal variable is binary and represents the values, zero(0) for no disease and one(1) for disease. Furthermore, it represents when the person was diagnosed with coronary heart disease.

Certain variables need to be transformed, in response to the fact that some variables are not normally distributed.

Basic summary statistics are useful to check if an attribute average or median gives an indication for a possible risk factor. That is the reason why two different tables are created, one with every individual with heart disease, and one with every individual without heart disease.

The following tables illustrates the basic summary statistics of the attributes in the dataset.

Table 2 - Basic Summary Statistics - With Coronary heart disease

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
Count	160	160	160	160	160	160	160	160	160	160
Mean	143.74	5.52	5.49	28.12	0.60	54.49	26.62	19.15	50.29	1.0
Std	23.68	5.57	2.23	7.06	0.49	10.25	4.39	26.18	10.65	0.0
Min	102.00	0.00	1.55	9.39	0.00	20.00	14.70	0.00	17.00	1.0
25%	127.00	1.50	3.94	23.46	0.00	47.75	23.64	0.48	42.75	1.0
50%	138.00	4.13	5.07	28.41	1.00	55.00	26.48	8.33	53.00	1.0
75%	158.00	8.20	6.58	33.59	1.00	61.00	28.78	24.58	59.00	1.0
Max	218.00	31.20	14.16	42.49	1.00	78.00	45.72	147.19	64.00	1.0

Table 3 - Basic Summary Statistics - Without Coronary heart disease

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
Count	302	302	302	302	302	302	302	302	302	302
Mean	135.46	2.63	4.34	23.97	0.32	52.74	25.74	15.93	38.85	0.0
Std	17.99	3.61	1.87	7.77	0.47	9.52	4.09	23.50	14.88	0.0
Min	101.00	0.00	0.98	6.74	0.00	13.00	17.75	0.00	15.00	1.0
25%	124.00	0.00	3.06	17.51	0.00	47.00	22.60	0.51	27.00	0.0
50%	132.00	1.04	3.98	24.63	0.00	52.50	25.57	6.05	40.00	0.0
75%	144.00	4.20	5.28	29.96	1.00	59.00	28.07	22.42	50.75	0.0
Max	214.00	20.00	15.33	42.06	1.00	77.00	46.58	145.29	64.00	0.0

Tables 2 and 3 of basic summary statistics show the 10 attributes the dataset contains. While the difference in certain means is noticeable, it is not enough to conclude the possibility of predicting a higher probability of coronary heart disease. However, the 50% median shows certain larger differences, especially in tobacco use.

To measure if there are any similarities between attributes in the dataset, the cosine similarity can be used to find an answer. Before the similarities can be found, the data need to be standardized using the following formula[4]:

$$\tilde{x} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Every variable in the data goes through this standardization. When this is done the following cosine similarity formula is used[4]:

$$cos(\tilde{x}, y) = \frac{\tilde{x}^T y}{\| \tilde{x} \| \| y \|}$$

This results in a 8×8 matrix, which shows the similarity between the different attributes.

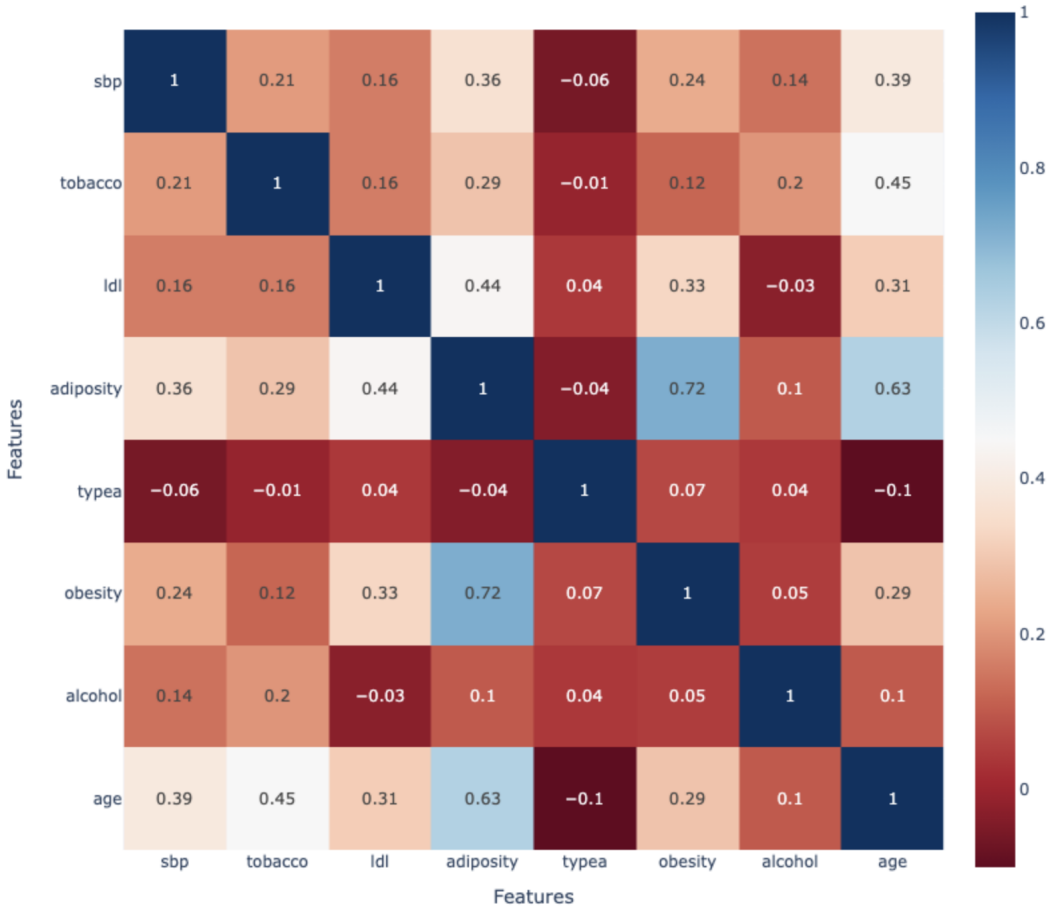

Figure 1: Cosine Similarity Matrix

However the *famhist* and the *chd*, have been removed, for the reason that these attributes are binary.

From the cosine similarity matrix 1, it can be observed that *obesity* and *adiposity* have the highest similarity. This is also very logical, since they both measure obesity. An argument can definitely be made to remove *obesity* because of the unreliability of BMI[3]. However, to avoid missing out on any of the representative data, the attribute will remain.

Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA)

In order to achieve a deeper insight into the data, it is important to look at the outliers. An outlier is a data object which is significantly different from most of the other data. The reason for outliers could be some measurement errors or a natural property of the data. A way to handle the outliers is to identify or exclude the outliers or model them.

A method to identify the outliers is to make a boxplot that shows the interquantile (IQR) which is measure the spread of the data. The outliers will be showed outside the IQR, which is between 75% percentile (Q3) and 25% percentile (Q1)[2].

Firstly, the data is standardized, then we plot it:

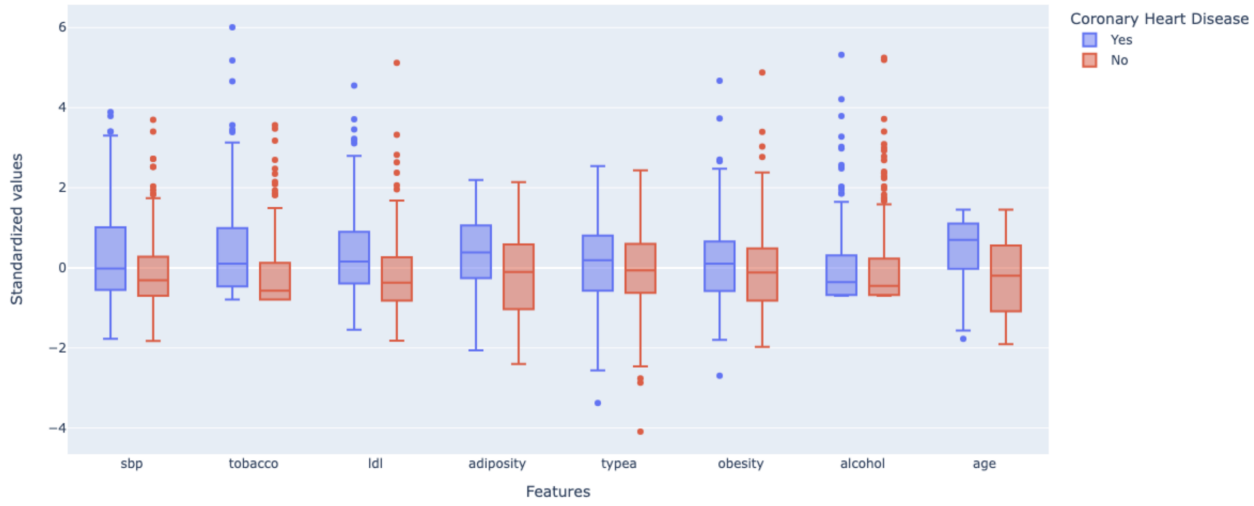

Figure 2: Boxplot of standardized data

The boxplot shows a significant amount of outliers, but that does not mean the outliers need to be removed. If the data is examined, the data points for the outliers are not dishonest, which means that the outliers should not be removed.

The South Africa - Heart Disease dataset contains a matrix 462×11, which is an 11 dimension dataset. The first column is just a numbering for each row, so we remove it. To work with such a high-dimension, the simplest way is to visualize the data in a 2D scatter plot 7.

The scatter matrix does not explicitly show us any significant outlier in terms of whether or not there is a connection between heart disease and a combined attribute. It tells nothing about the behavior of the risk factor. However, it can be useful to distinguishing between the heart disease or not, which can be used to terms of the classification problem [4].

To analyze and visualize the dataset, principal component analysis (PCA) is a technique to represent a high-dimensional dataset and transform it to a lower-dimension. To use this technique some circumstances or conditions for the lower-dimensional is required, in that some of the information can be lost. For handling high-dimensional data, an effective method is to visualize the data[4].

To transform the high-dimension to a lower-dimension, the high-dimension can be projected to a lower-dimension subspace by selecting a orthonormal basis vector of a n dimensinal subspace V, using the follow equation[4]:

$$\text{Project onto V: } b_i^T = \bar{x}_i^T V$$

The goal for PCA would be to select the vector, which maximizes the spread-outness for the projection of the data therefore the Singular Value Decomposition (SVD) need to be used, to help compute the eigenvectors that corresponds to the largest eigenvalues. By computing the n eigenvectors, which corresponds to the n largest eigenvalues, there will become three SVD matrices, showed below, for

the $N \times M$ matrix.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_M \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}, \quad U = [u_1, u_2, \dots, u_N], \quad V = [v_1, v_2, \dots, v_M]$$

This can be used to compute the PCA algorithm, the $N \times M$ matrix, $\mathbf{X}$, such that $U\Sigma V^T = \tilde{X}[4]$, which leads to the explained variation.
The explained variation of the PCA, is shown in the below scree plot.

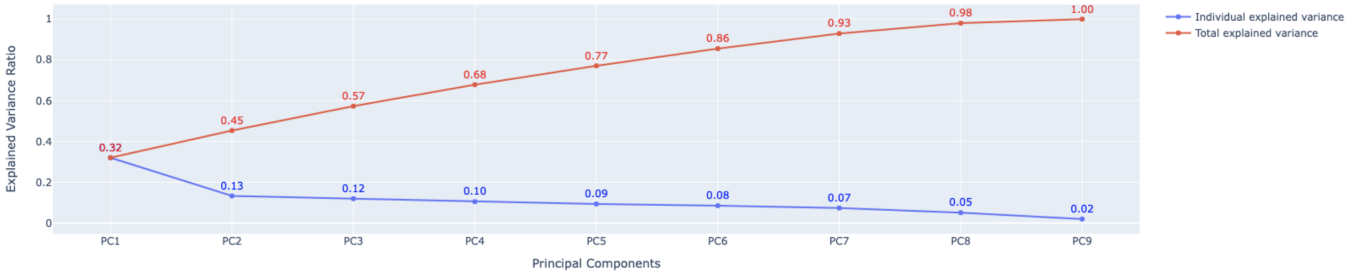

Figure 3: Scree Plot

The scree plot, shows the relation between each PCA and the explained variation ratio, for both the total- and the individual explained variation. The blue line shows how much information each individial component contains, while the red line shows the combined information of the principal components. It can be observed that PC6 surpasses 85% of the information, and should be enough.

To get an idea of the lost information in the dataset, the full dataset can be considered. The plots below shows the two first PC's and the 5th & 6th PC's:

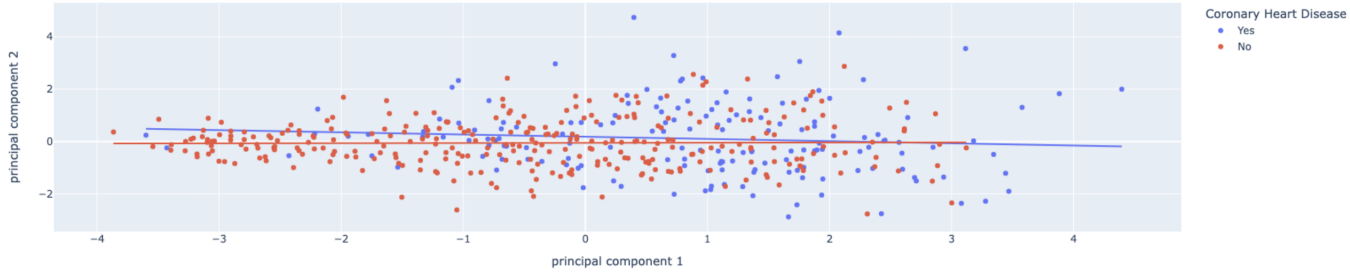

Figure 4: Plot of PC1 and PC2

The first graph does not show a significant visible pattern, even when it consists of 45% of the data.

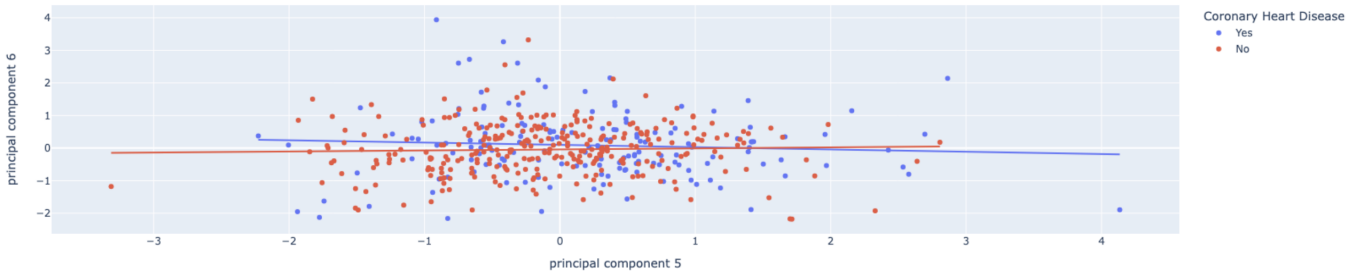

Figure 5: Plot of PC5 and PC6

However, even when including 86% of the data, no significant pattern emerges in the graph.

Discussion and conclusion

Through this project and working with the data set, the learning is to understand how the dataset is. Through the difference things in the project, such as missing data, similarity, PCA, etc. the dataset shows the difficulties with finding patterns about possible heart disease. The biggest factors are observed to be age, family history, and tobacco, however a certain visible pattern has not shown itself, not even after a principal component analysis.

The below plot shows a similar plot of the risk factors as in the original study[5].

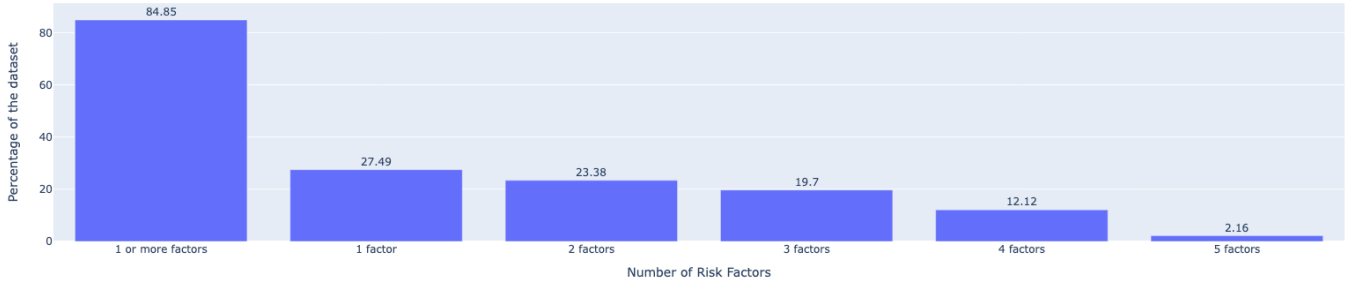

Figure 6: Risk Factors

This shows that in the communities in question a majority has at least one risk factor. However, these risk factors are limited, as the data could be skewed. The other factors could for example be diet, stress level, and a more comprehensive view of their level of physical activity.

A larger dataset involving more individuals and more possible risk factors could alleviate these limitations. Without a visible pattern, doing machine learning on this data would not be feasible and would not result in a reliable model.

Exam Question

Question 1. Spring 2019 question 1:

The correct answer is D, for the reason that some of the answers can be excluded.

A and C can be excluded, because the time of the day, x_1, is an interval, for the reason that each observation corresponds to a 30 minute interval ($x_1 = 1$, which is between 7:00 and 7:30 and up to $x_1 = 27$, which is between 20:00 and 20:30) and the measurement of the time for the ratio is recorded between 7:00 and 20:30.

The answer, B, can also be excluded, for the reason that the number of broken traffic lights, x_6, and Number of run-over accidents, x_7, are ratios, because they are both measurable variables.

The correct answer must therefore be D, which also can be underlined for the fact that the higher y is getting, the more heavy congetion there will be, which mean that the congestion level, y, is ordinal, for the fact that it can be ranked.

Question 2. Spring 2019 question 2:

The correct answer is A, for the reason that the calculation of the p-norm distance is 7.0.

The formula, that is used to calculate the p-norm distance is the following equation:

$$d_p(x,y) = ||x-y||_p = \begin{cases} \left(\sum_{i=1}^{M} |x_i - y_i|^p\right)^{\frac{1}{p}} & \text{if } i \leq p \leq \infty. \\ \max\{|x_1 - y_1|, |x_2 - y_2|, \dots |x_M - y_M|\}, & \text{if } p = \infty. \end{cases}$$

By using Scipy's linag.norm() function, in python, p-norm distance is, $d_{p=\infty}(x_{14}, x_{18}) = 7.0$, indicates that A is the correct answer.

Question 3. Spring 2019 question 3:

A is the correct answer, for the reason that the calculation shows that the PC4 is 0.87, which is greater than 0.8.

To find the result the formula of the variance explanied need to be used.

$$\text{Variance Explained} = \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$$

The S matrix, informed in the question is equal to the $\sum$-matrix, since both matrix contains the standard deviation diagonal.

Question 5. Spring 2019 question 14:

The correct answer is A, in behalf of the calculation of Jaccard similarity, which gives the result $J(s_1, s_2) = 0.153846$.

To calculate Jaccard similarity, the following formula is being used:

$$J(s_1, s_2) = \frac{S_1 \cap s_2}{S_1 \cup s_2}$$

The formula comparing the text document, that contains the bag-of-words, of their intersection and union, and find the Jaccard similarity of the to texts.

Question 6. Spring 2019 question 27:

Answer C, and it can be calculated using Bayes' theorem:

$$p(y_j|x_i, z_k) = \frac{p(x_i|y_j, z_k)p(y_j|z_k)}{\sum_j p(x_i|y_j, z_k)p(y_j|z_k)}$$

so the calculation is as follows:

$$p(\tilde{x}_2 = 0|y = 2) = \frac{p(y = 2|\tilde{x}_2 = 0)p(\tilde{x}_2 = 0)}{\sum p(y = 2|\tilde{x}_2 = 0)p(\tilde{x}_2 = 0)}$$

With the values from the question inserted:

$$p(\tilde{x}_2 = 0|y = 2) = \frac{0.23 \times 0.84}{0.274 \times 0.84 + 0.23 \times 0.84 + 0.244 \times 0.73 + 0.252 \times 0.77} \approx 0.243$$

References

[1] South african heart disease dataset. `https://hastie.su.domains/ElemStatLearn/datasets/SAheart.data`.

[2] Per B. Brockhoff, Jan K. Møller, Elisabeth W. Andersen Peder Bacher, and Lasse E. Christiansen. *Introduction to Statistics at DTU*, page 12 and 15. 2022.

[3] I. Gutin. In bmi we trust: Reframing the body mass index as a measure of health. *Soc Theory Health*, 16(3):256–271, Aug 2018. Epub 2017 Oct 25.

[4] Tue Herlau, Mikkel N. Schmidt, and Morten Mørup. *Introduction to Machine Learning and Data Mining*, pages 29, 34, 38–40, 59–60, and 121–122. Polyteknisk Kompendie, 2023.

[5] JE Rossouw, JP Du Plessis, AJS Benade, PCJ Jordaan, JP Kotze, PL Jooste, and JJ Ferreira. Coronary risk factor screening in three rural communities-the coris baseline study. *South African medical journal*, 64(12):430–436, 1983.

Appendix 1 - Plot from the decription section

Scatter matrix:

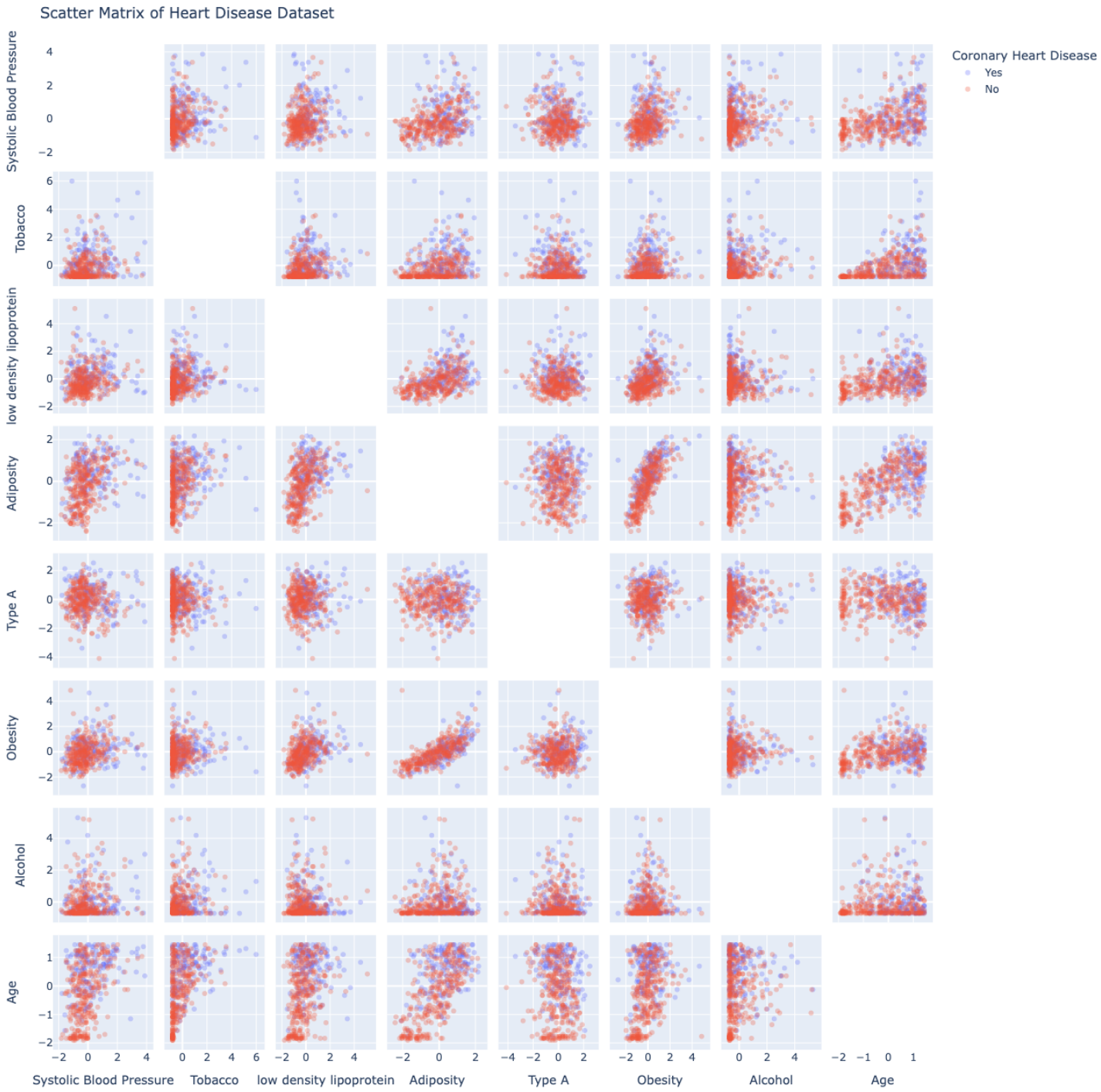

Figure 7: Scatter Matrix

Appendix 2 - Exam Question

Question 2:

```
import numpy as np

x1 = np.array([26, 0, 2, 0, 0, 0, 0])
x2 = np.array([19, 0, 0, 0, 0, 0, 0])

max_normal_distance = np.linalg.norm(x1 - x2, ord = np.inf)
print(max_normal_distance)
```

Question 3:

```
import numpy as np

sigma = np.array([13.9, 12.47, 11.48, 10.03, 9.45])
total_var_explained = np.power(sigma, 2).sum()
PCA_var = sigma.T * sigma / total_var_explained

print("first four:", np.cumsum(PCA_var[::1][:4]))
print("last three:", np.cumsum(PCA_var[:: -1][:3]))
print("first two:", np.cumsum(PCA_var[::1][:2]))
print("first three:", np.cumsum(PCA_var[::1][:3]))
```

Question 5:

```
def Jacard(set_1, set_2):
    words_set_1 = set(set_1.lower().split())
    words_set_2 = set(set_2.lower().split())

    Intersection = words_set_1.intersection(words_set_2)

    Union = words_set_1.union(words_set_2)

    return float(len(Intersection)) / len(Union)

s_1 = "the bag of words representation becomes less parsimoneous"
s_2 = "if we do not stem the words "

print(Jacard(s_1, s_2))
```