Danmarks Tekniske Universitet

DTU

02450 Introduction to Machine Learning and Data Mining

Group 31 - Project 1

Adikrishna Murali Mohan - s212940
Spyridon Safikos - s223409
Mohamad Ashmar - s176492

Student ID	Section 1	Section 2	Section 3	Exam Questions
s212940	30%	40%	30%	33.33%
s223409	40%	30%	30%	33.33%
s176492	30%	30%	40%	33.33%

Date: 7th March 2023

Contents

1 Description of the data

The data used for analysis is the South African Heart Disease Data, taken from a larger dataset[1], described in Rousseauw et al, 1983, South African Medical Journal[2]. It contains medical data from patients who received medical examinations for heart disease at a South African medical center.

After summarizing and understanding previous analysis of Exploring Machine Learning Techniques for Coronary Heart Disease Prediction[1], it is important to note that the researchers had to perform data preparation, exploration, and cleaning in order to achieve their goal of predicting the occurrence of Coronary Heart Disease (CHD) by using machine learning algorithms. They performed exploratory data analysis (EDA) to look at the distribution of each of the attributes, then selected the most strongly associated features with CHD by using ANOVA method, correlation feature selection (CFS) and wrapper feature selection (WFS) algorithms. They focused on visualizing the features with respect to the CHD event.

Our plan when working with this dataset is to handle any missing values, scale the data, and perform dimensionality reduction techniques. This will help us to better understand the relationships between attributes and how they may contribute to the occurrence of CHD.

Thus the overall problem of interest in the South African Heart Disease dataset is to identify risk factors associated with the occurrence of heart disease in South African individuals.By analyzing these risk factors and their association with the occurrence of heart disease, we can better understand the underlying causes of heart disease in the South African population and potentially identify preventive measures to reduce the risk of heart disease in this population.

We can use classification and regression techniques to accomplish the following:

- Classification techniques: we can predict whether a patient is likely to develop CHD or not. By building a classification model, we can identify the most important risk factors for CHD and how they relate to each other. The classification model can be used to identify high-risk patients, and appropriate interventions can be implemented to reduce their risk of developing CHD. Additionally, we can use classification to identify subgroups of patients that are at higher risk of developing CHD.
- Regression techniques: can be used to predict the level of risk of developing CHD for an individual. By building a regression model, we can estimate the probability of developing CHD based on different risk factors. It can be used to identify the most significant risk factors for CHD and provide insight into the relationships between different risk factors.

By analyzing the data with these techniques, we can gain insights into the underlying causes of CHD and develop effective strategies for managing and preventing this condition in the South African population.

[1] `https://philarchive.org/archive/KHDEML`

In the context of CHD, we can use classification techniques to predict whether a patient is affected or not. For the regression part from the scatter plot and heatmap it seems that obesity and adiposity are highly correlated so choosing one as the predictor and excluding the other like predict obesity from adiposity. However, we can make the task more challenging and interesting by attempting to predict systolic blood pressure (SBP) using a combination of predictors known to be associated with high blood pressure, such as diabetes, age, obesity, tobacco use, and excessive alcohol consumption, as identified in a source such as the article 'Know Your Risk for High Blood Pressure'[2].

2 Attributes of the data

The dataset consists of a data frame with 462 rows and 10 variables: systolic blood pressure (sbp), cumulative tobacco in kg (tobacco), low-density lipoprotein cholesterol (ldl), adiposity, family history of heart disease (famhist), type-A behavior (typea), obesity, current alcohol consumption (alcohol), age at onset (age), and coronary heart disease (chd).

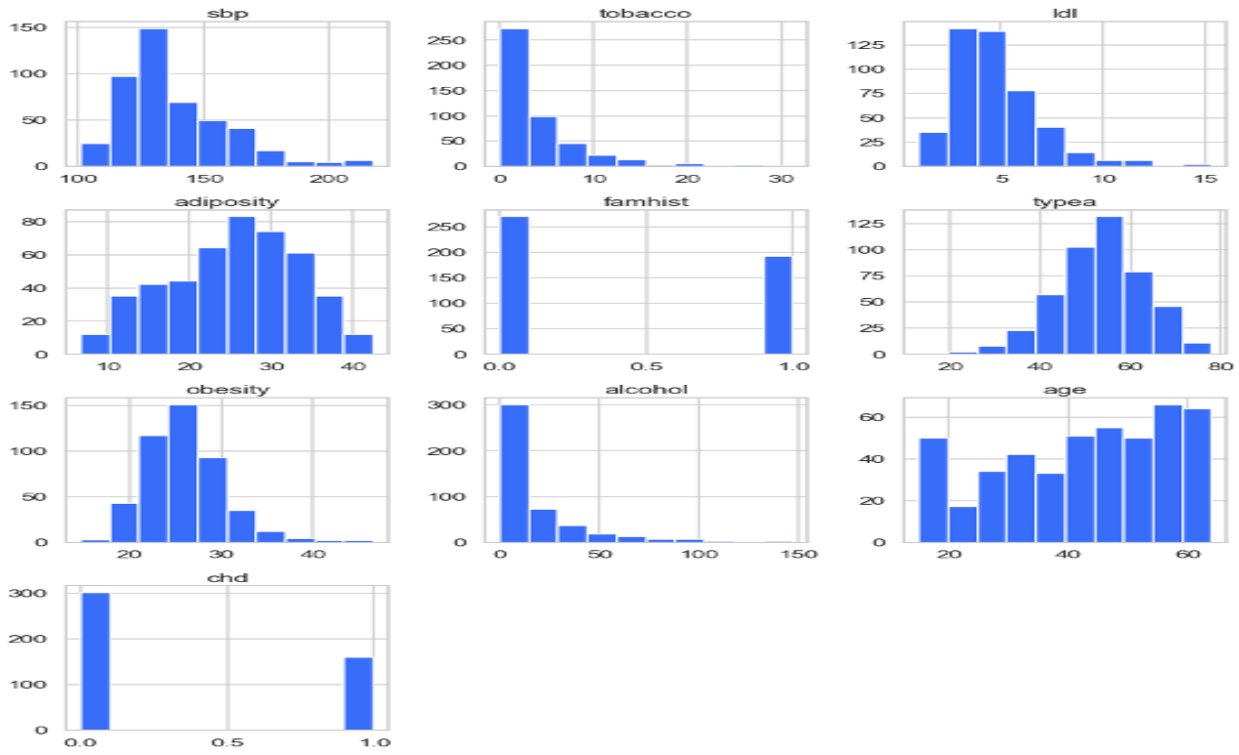

Figure 1: Data distribution for each variable

Among these variables we consider 'chd' as the target and the rest of the 9 variables are considered the features.

Missing values : We can use isnull() function to check for any missing values in the dataset. This returns a DataFrame of the same shape as the original dataset, with each element being

[2] https://www.cdc.gov/bloodpressure/risk_factors.htm

either True or False depending on whether it is a missing value or not. Finally, we can use the sum() function to count the total number of missing values in each column of the dataset. Upon inspection it can be confirmed that our data contains no missing values.

2.1 Classification of the attributes

The attributes in the South african heart disease dataset can be classified into the following categories:

Discrete : sbp, typea, age. They includes whole, finite set of numbers with specific and fixed data values determined by counting.
Continuous : ldl . It takes any value in a given range.
Nominal: famhist. It is binary and take values "Present" or "Absent". Nominal Attributes only provide enough attributes to differentiate between one object and another
Ordinal: typea. It has a natural order
Interval: discrete value and there is order and the difference between two values is meaningful.
Ratio: sbp, tobacco, ldl , adiposity, alcohol, obesity all have a true zero point. Ratio variables are characterized by having a meaningful zero point, which represents the complete absence of the attribute being measured, and a continuous scale with equal intervals between values.

2.2 Summary statistics

The summary statistics of the attributes give us the information about the data in the sample. It include the total number of values(count), mean, standard deviation, minimum value, maximum value and percentages corresponding to the data collection.We can use this information to understand the distribution and range of each attribute, and confirm for any unexpected values or missing values. The table below shows the statistical overview of our dataset:

Index	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
count	462	462	462	462	462	462	462	462	462	462
mean	138.33	3.64	4.74	25.41	0.42	53.1	26.04	17.04	42.82	0.35
std	20.5	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
25%	124	0.05	3.28	19.77	0	47	22.98	0.51	31	0
50%	134	2	4.34	26.12	0	53	25.8	7.51	45	0
75%	148	5.5	5.79	31.23	1	60	28.5	23.89	55	1
max	218	31.2	15.33	42.49	1	78	46.58	147.19	64	1

Figure 2: Summary statistics

The number of observations for each attribute is 462. It shows that 35 percent of the total 462 in our sample are diagnosed with the heart disease. 'chd' is a binary variable, where 1 indicates the presence of CHD and 0 for the absence of CHD. Therefore, the resulting table shows the mean value of each attribute for individuals with and without CHD.

chd	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age
0	135.46	2.63474	4.34424	23.9691	0.317881	52.3675	25.7375	15.9314	38.8543
1	143.738	5.52487	5.48794	28.1202	0.6	54.4937	26.6229	19.1453	50.2938

Figure 3: Summary statistics based on 'chd'

Upon analyzing the summary statistics of the whole data, it shows that the those who has chd(1) also has high value for adiposity and obesity. However the rates for alcohol and tobacco consumption is low.

3 Data Visualization

To check for the presence of outliers in our data, we can use boxplots which display the distribution of each attribute and the presence of any outliers. Outliers can be identified as individual points that are far from the rest of the distribution.

If there are outliers present in the data, we need to handle them appropriately before carrying out further analysis otherwise they can impact the model accuracy or the conclusion about relationships between attributes, where outliers can skew the distribution and cause an effect over the mean and standard deviation.

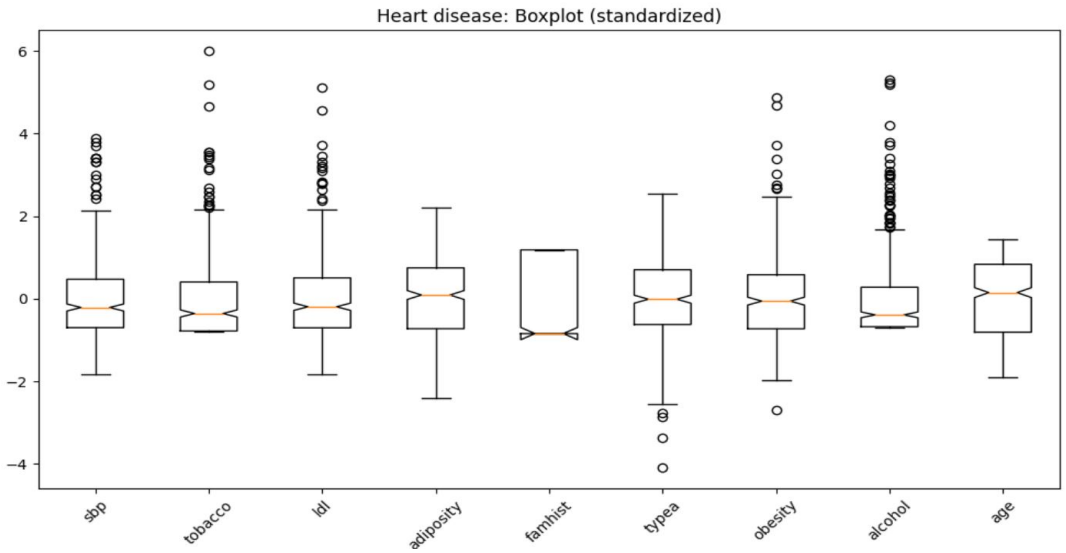

Figure 4: The figure shows the boxplot for each independent attribute

As you can see from the above boxplot, there are some points that appear above the whiskers, and we may claim that those are outliers. However, when we check the attribute description domain, the official range for every attribute, those points(outliers) are still located inside the range. BUT we are still not certain that we do not have outliers in our dataset, because those points appear above the whiskers and they might be far away from the majority of the distribution. Therefore, the interquartile range (IQR) method, which calculates the difference between the upper quartile (Q3) and lower quartile (Q1), can be used to identify outliers and any data points that fall below Q1-1.5IQR or above Q3+1.5IQR are almost considered outliers and should be removed[3].

As shown in the histograms below, tobacco and alcohol have a right-skewed distribution, while sbp, ldl, adiposity, typea, and obesity have a normal distribution. However, age seems to have a left-skewed distribution. Those distributions become more normalized after removing the outliers, as can be seen by comparing Figure 1 and Figure 5 which show the distributions before and after outlier removal.

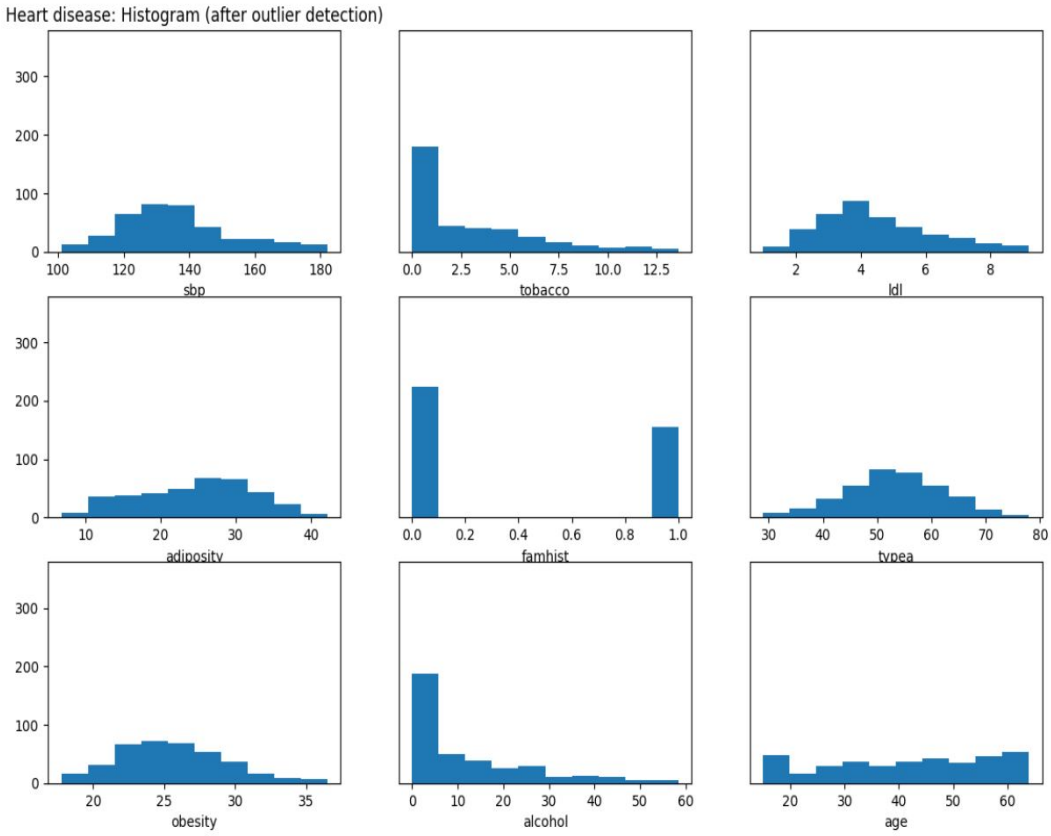

Figure 5: Histogram shows the distribution of the values after removing the outlier

The heatmap below shows the correlation between different attributes, with a high correlation indicating a strong linear relationship. Adiposity and obesity have a high correlation of 0.72, which suggests they are highly correlated with each other. However, including both

attributes in a predictive model could lead to multicollinearity issues since they have a similar effect on CHD, causing both variables to capture almost the same information. This can overwhelm a logistic regression model.

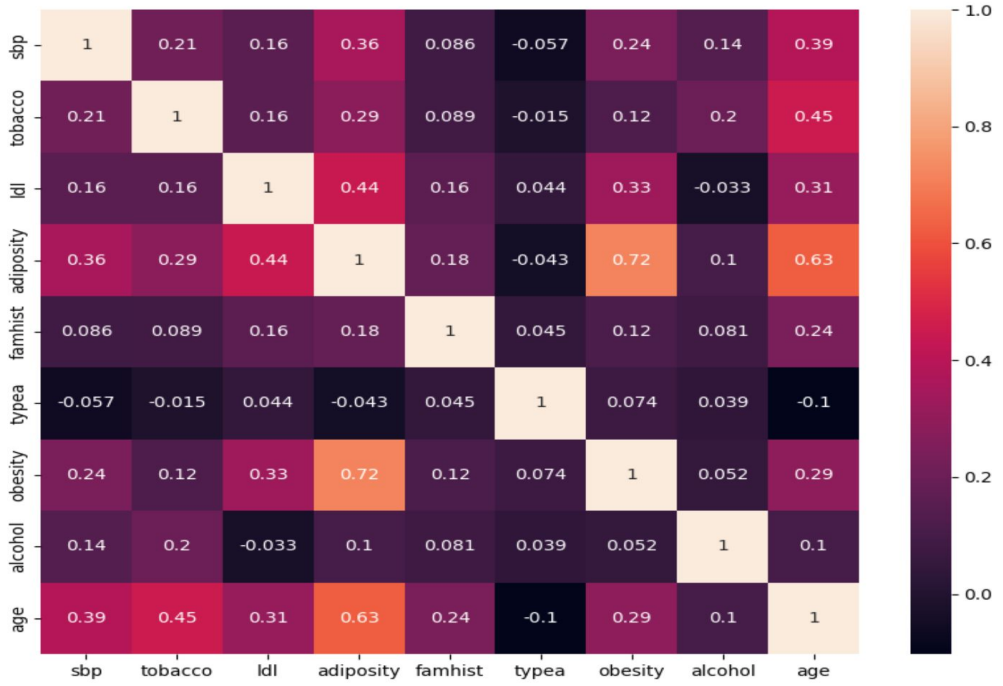

Figure 6: Heatmap shows the correlation between attributes

After the outlier removal we can see from the paired scatter plot for attribute, the correlation between the attributes with regard to chd. for example in case of the plot between tobacco and sbp, we can differentiate the presence or absence of chd .

Regarding the regression prediction part, we can see the high correlation between two variables from the heatmap above and the paired scatter plot below, such as adiposity and obesity, so it may be possible to predict one from the other.

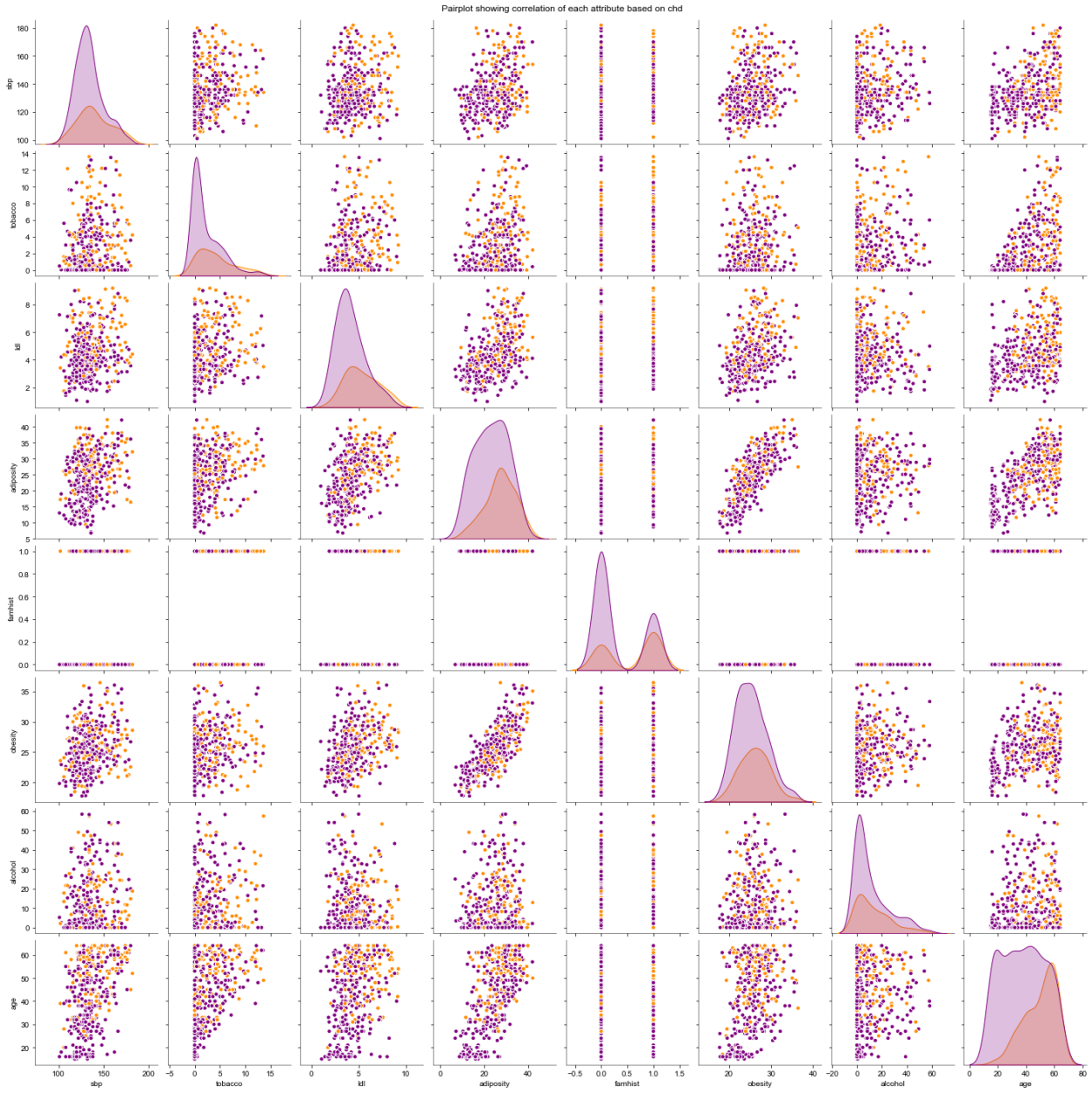

Figure 7: Scatter pairplot to show the correlation between the attributes

4 PCA

Two ways to conclude if the data need to be standardized or not. The first way by calculating the range of each attribute and indicating how much they are likely different. Like "sbp" attribute has range of 117, but "famhist" has range of 1 or use the second way which uses bar plot to display the attributes' standard deviations.

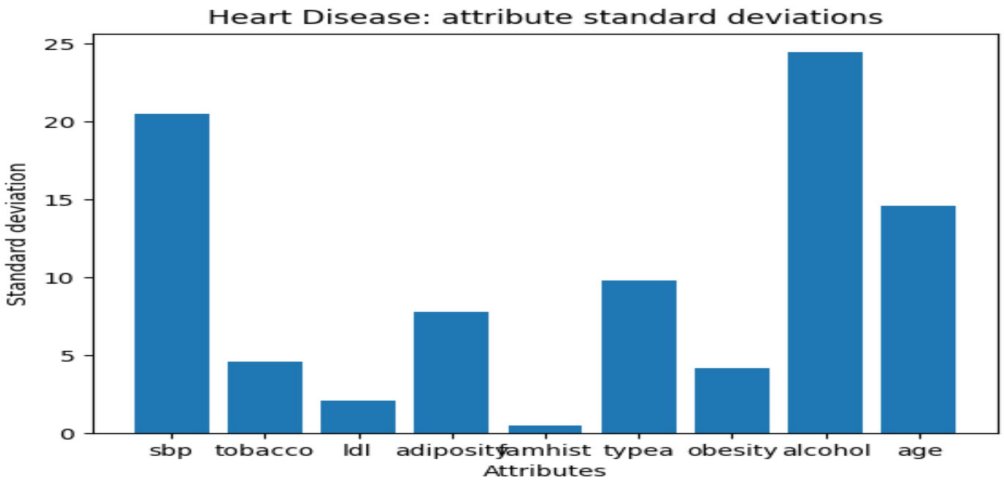

Figure 8: Attributes' standard deviations

The barplot above shows that the attributes have different scale and need to be standardized before doing PCA calculation in order to avoid dominating our covariance matrix by the attribute who has higher scale than other. To standardize data which transform the mean to zero and standard deviation of one, we subtract the mean from each sample and then divide it by the standard deviation.

To reduce the number of attributes by removing redundant variables or handling computational challenges, we can use a technique called Principal Component Analysis (PCA). PCA transforms high-dimensional data into a lower-dimensional space, simplifying complex data while still retaining the most important information from our attributes.

As you can see from the principal component plot below, after performing standardization on the data, we can see that the variance of the data is explained by the first 7 principal components, which means explaining almost 90% of the variance and this is what the dashline refers to as the 0.9 threshold.

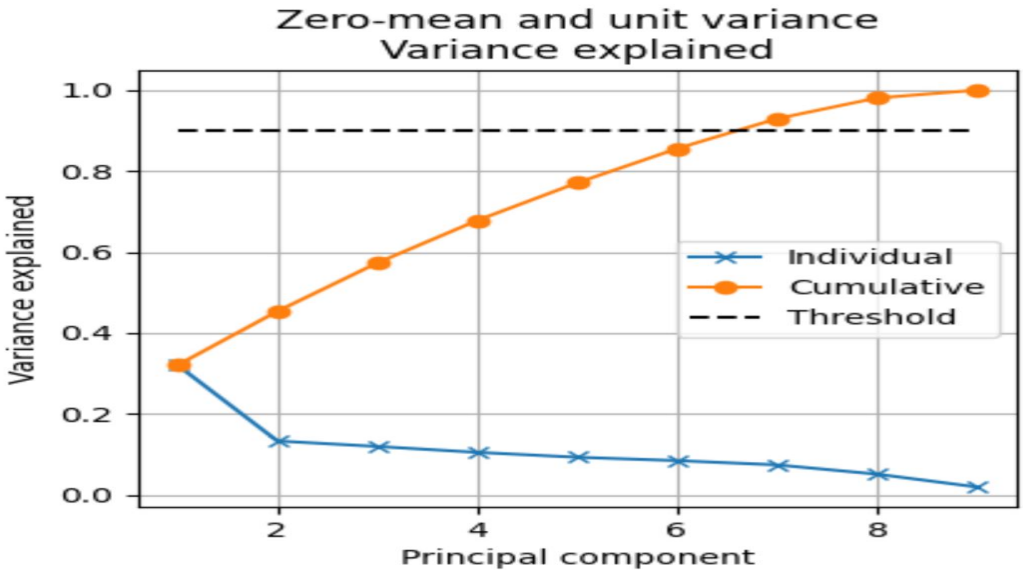

Figure 9: Variance explained by each principal component after standardization

After plotting all the combinations of the first 7 components, we can observe that the scatter plot below of PCA3 vs. PCA4 shows the best separation between the two classes, despite some overlapping, and hence, could be a good predictor for CHD.

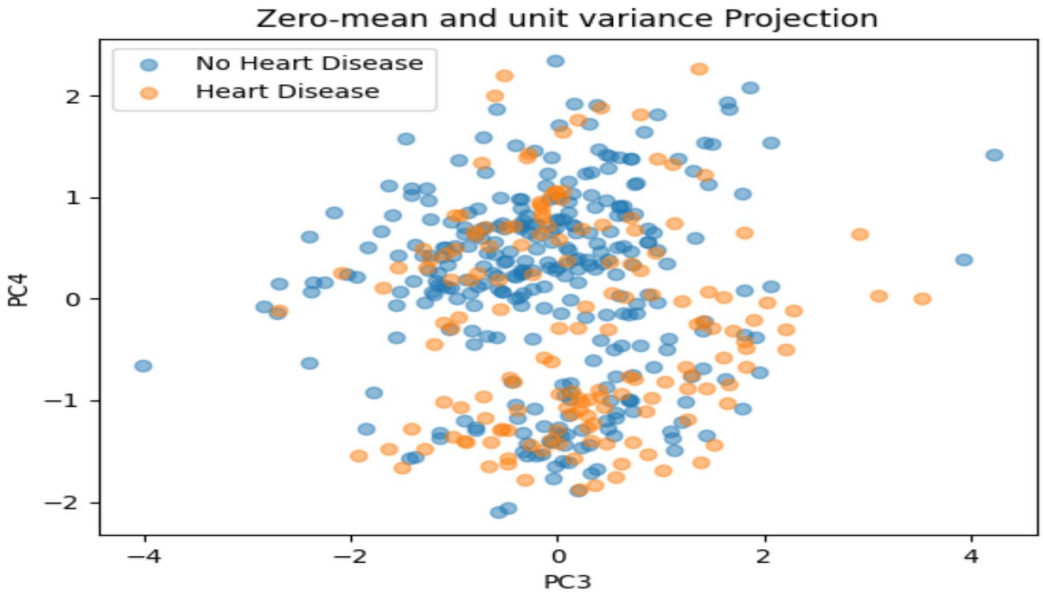

Figure 10: Scatter plot over variance direction

5 Discussion

From the analysis of the data, we could see that the data does not have any missing values or corrupted data, however the 9 attributes are of varying scales. Thus we performed standardization of the values to get an accurate result.

Based on the analysis and visualization of the data, it appears that our primary machine learning aim is to predict coronary heart disease (CHD). We have gained a good understanding of the relationships between the different attributes and how they relate to the target variable. However, we need to take into consideration the right attributes to avoid multicollinearity issues and consider the impact of outliers on the model's accuracy.

PCA was used to simplify the data and reduce the number of attributes while retaining the most important information. Before performing standardization on the data, almost 90% of the variance of the data was being explained by the first two principal components. However, after performing standardization to avoid having attributes with higher scales dominating the covariance matrix, we observed that the first 7 principal components explained almost 90% of the variance in the data.

After visualizing all the combinations of the first 7 principal components, we observed that PCA3 vs. PCA4 showed the best separation between the two classes, so these two are almost the best candidate for predicting CHD.

6 Exam questions

Problem 1: Answer: option C

Types of the attributes in the Urban Traffic dataset:
x1 (Time of day) is ordinal : values are ordered from 1 to 27
x6 (Traffic lights) is ratio : is a numerical value
x7 (Running over) is ratio : is a numerical value
y (Congestion level) is ordinal : it has categories such as low and high

Problem 2: Answer: option A
When $p = \infty$, the formula to compute the p-norm distance simplifies to:

$$d_\infty(x_{14}, x_{18}) = \max_{i=1}^{7} |x_{14,i} - x_{18,i}|$$

For $x_{14} = [26\ 0\ 2\ 0\ 0\ 0\ 0]$ and $x_{18} = [19\ 0\ 0\ 0\ 0\ 0\ 0]$, we have:

$$d_\infty(x_{14}, x_{18}) = \max\left(|26-19|, |0-0|, |2-0|, |0-0|, |0-0|, |0-0|, |0-0|\right) = 7$$

Therefore, $d_\infty(x_{14}, x_{18}) = 7$ and the correct answer is statement (A).

Problem 3: Answer: option A

Where the sum of the first four singular values over the sum of the squares of all five singular values is shown below:

$$\text{where } \frac{\sum_{i=1}^{4} \sigma_i^2}{\sum_{i=1}^{5} \sigma_i^2} = \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.85$$

Problem 4: Answer: option C

An observation with a low value of Time of day (0.49), a high value of Broken Truck (0.71), a low value of Accident victim (0.25), and a low value of Defects (0.41), and if the value of the projection onto principal component number 4 is close to zero, it means that the value considered as low and vice versa.

Problem 5: Answer: option A

To calculate Jaccard similarity between documents s1 and s2:
Convert the two text documents s1 and s2 containing the text into two set of unique words respectively.
s1 = {the, bag, of, words, representation, becomes, less, parsimonious}
s2 = {if, we, do, not, stem, the, words}

Find the intersection and union of unique words
The intersection of s1 and s2 = {the, words} which has a size of 2.
The union of s1 and s2 = {the, bag, of, words, representation, becomes, less, parsimonious, if, we, do, not, stem} which has a size of 13.
Jaccard similarity between s1 and s2 = intersection/ union = 2/13 = 0.154.

Problem 6: Answer: option B

Using Bayes' theorem we can calculate the posterior probability of x^2 = 0 given light congestion (y = 2) as:
p(x^2 = 0 | y = 2) = p(y = 2 | x^2 = 0) * p(x^2 = 0) / p(y = 2)
probability of light congestion given x^2 = 0, from the table
p(x^2=0) = P(x^2 = 0, x^7 = 0) + P(x^2 = 0, x^7 = 1) = 0.84
prior probability of light congestion, which is given as p(y=2) = 0.23
therefore:
p(x^2 = 0|y=2) = 0.84 * 0.23 / 0.23 = 0.84

References

[1] "Replication Data for: South African Heart Disease," 2016.

[2] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning."

[3] G. Holland, "How to find outliers in python," Feb 2023.