Danmarks Tekniske Universitet

DTU

02450 Introduction to Machine Learning and Data Mining - Project 1

Authors

Maxime Swagel - s242033
Santiago Londoño - s250161
Boon Keng Leck - s247301

We confirm that this report is our own independent and original work.
The allocation of tasks is detailed in the table below.

Section	Contribution (%)		
	Maxime Swagel	**Santiago Londoño**	**Boon Keng Leck**
1	20	20	60
2	30	30	40
3	60	20	20
4	20	60	20

March 6, 2025

Contents

Introduction

The objective of this report is to present the results obtained after applying various data extraction and visualization techniques to the chosen dataset. For this analysis, the South African Heart Disease Study dataset was selected. This dataset is a retrospective sample of males from a high-risk heart disease region in the Western Cape, South Africa, and contains multiple features that make it well-suited for exploratory data analysis and predictive modeling.

This dataset is particularly valuable for applying machine learning techniques due to its mix of numerical and categorical variables. It includes key features such as systolic blood pressure (sbp), low-density lipoprotein cholesterol (ldl), adiposity, obesity, tobacco consumption, alcohol consumption, family history of heart disease (famhist), and Type-A behavior. These features allow for a comprehensive evaluation of different data preprocessing techniques, including handling missing values, feature scaling, and encoding categorical variables.

1 A description of the data set

1.1 Explain the overall problem of interest and the associated data.

The general problem focuses on the study of coronary heart disease (CHD) using data from male subjects in Western Cape, South Africa, examining the relationships between various health measurements (blood pressure, tobacco use, cholesterol, etc.) and the occurrence of CHD. The dataset includes both CHD cases and controls (2:1 ratio), though some measurements were taken after CHD-positive subjects had received treatments to reduce risk factors. There are 9 features in total and the breakdown (the meanings behind each attribute) is given in Table 1.

Variable	**Description**
bp	Systolic blood pressure (mmHg)
tobacco	Cumulative tobacco (kg)
ldl	Low-density lipoprotein cholesterol (mmol/l)
adiposity	Body fat percentage (%)
famhist	Family history of heart disease (Present/Absent)
obesity	BMI value (weight/height^2)
alcohol	Current alcohol consumption
age	Age at onset of CHD
typea	Type-A behavior (Bortner Short Rating Scale, where a value of more than 55 is considered Type A)
chd	Whether subject has coronary heart disease (0/1)

Table 1: Description of Variables in the South African Heart Disease Dataset

1.2 Provide a reference to where you obtained the data

The website from which we retrieved the data is https://hastie.su.domains/ElemStatLearn/. The data was taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal.

1.3 Summarize previous analysis of the data. (i.e. go through one or two of the original source papers and read what they did to the data and summarize their results).

The original paper (Coronary risk factor screening in three rural communities. The CORIS baseline study by Rousseauw et. al.) reveals a high prevalence of cardiovascular risk factors, particularly in middle-aged and older subjects. Risk factors are grouped into major - hypercholesterolemia, hypertension and smoking - and minor/non-reversible - obesity, type A behaviour, family history of CHD. As the data we are using is only a subset (contains only male subjects) of the original data, the relevant findings are that the high prevalence of major risk factors are compounded by similarly high prevalence of minor factors such as obesity and type A behaviour - i.e. they are most likely correlated to each other. In addition, due to a strong age trend for cholesterol, blood pressure and smoking, more men in the younger age groups are classified as absent for CHD.

1.4 You will be asked to apply (1) classification and (2) regression on your data in the next report. For now, we want you to consider how this should be done. Therefore:

1.4.1 Explain, in the context of your problem of interest, what you hope to accomplish/learn from the data using these techniques?

We seek to identify key risk factors associated with coronary heart disease (CHD) among males in the communities of West Cape, South Africa, and whether this corresponds to the findings from the original paper. In addition, alcohol consumption was not mentioned in the original paper which can give us different insights.

1.4.2 Explain which attribute you wish to predict in the regression based on which other attributes?

While the original paper treated type A as a binary variable (having type A behavior or not), it was originally measured as a continuous variable based on the Bortner Short Rating Scale for coronary prone behavior where a value of more than 55 is considered type A. Hence, we seek to use `'typea'` as the attribute to be predicted using regression based on other attributes.

1.4.3 Which class label will you predict based on which other attributes in the classification?

The original paper treats CHD as a binary variable of whether someone has CHD or not. Hence, we seek to use `'chd'` as the attribute to be predicted using classification based on other attributes.

1.4.4 Explain if you need to transform individual attributes in order to carry out these tasks (e.g. centering, standardization, discretization, log transform, etc.) and how you plan to do this.

We used boxplots to visualise whether we need to center and standardize our data, in the event they have different scales. Based on Figure 1, `'sbp'` has a much higher mean than the other variables and we intend to center all variables by subtracting the mean from each feature. In addition, the variables also have different ranges of data and we intend to normalize every variable using min-max scaling. As for discretization, it is foresee that they are not required as discretization introduces more complexity and we intend to keep the features simple. The variable *alcohol* will need to be log-transformed as it

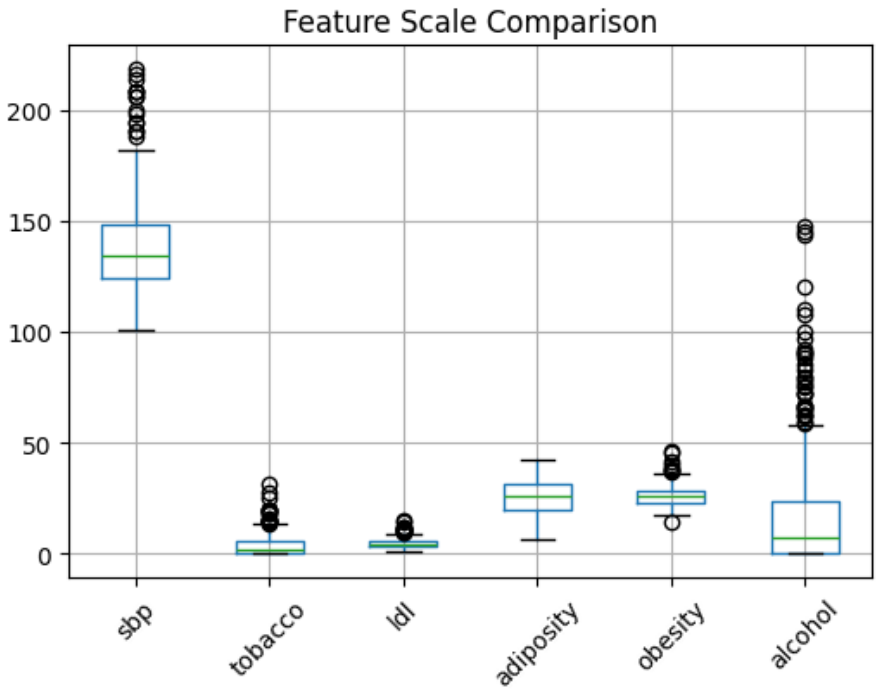

Figure 1: Boxplots of continuous variables in dataset

is heavily skewed with the outliers being about 2 orders of magnitude larger than the mean. Lastly, the variable *famhist* was encoded in text as *Present* or *Absent*. We intend to transform the data into binary values with 0 replacing *Absent* and 1 replacing *Present*.

2 A detailed explanation of the attributes of the data

2.1 Describe if the attributes are discrete/continuous and whether they are nominal/ordinal/interval/ratio.

Variable	Measurement Scale
bp	Continuous, Ratio
tobacco	Continuous, Ratio
ldl	Continuous, Ratio
adiposity	Continuous, Ratio
famhist	Discrete, Nominal (Binary)
obesity	Continuous, Ratio
alcohol	Continuous, Ratio
age	Discrete, Ratio
typea	Discrete, Ratio
chd	Discrete, Nominal (Binary)

Table 2: Variable Types and Measurement Scales

Table 2 shows the breakdown of the attributes on whether they are discrete/continuous and whether they are nominal/ordinal/interval/ratio

2.2 Give an account of whether there are data issues (i.e. missing values or corrupted data) and describe them if so and how you will handle them.

Since Panda library is being used to import the data in Python environment, a summary of the dataset is being presented as soon as the data is imported. The summary is shown in Table 3. Based on the information given by this table, there are no missing values in any of the rows (which totals up to 462 data points). The data types of each variable corresponds to their meaning and only irrelevant variables such as row.names will need to be removed prior to model training.
Based on the boxplot from the previous question, while alcohol consumption is highly skewed, it is not due to outliers or corrupted data as it is possible for a population to have people consuming high amounts of alcohol.

Column	Non-Null Count	Dtype
row.names	462	int64
sbp	462	int64
tobacco	462	float64
ldl	462	float64
adiposity	462	float64
famhist	462	object
typea	462	int64
obesity	462	float64
alcohol	462	float64
age	462	int64
chd	462	int64

Table 3: Data types of variables in South African Heart Disease Dataset

2.3 Include relevant summary statistics of the attributes. Reflect on the values

Statistic	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
count	462.00	462.00	462.00	462.00	462.00	462.00	462.00	462.00	462.00	462.00
mean	138.33	3.64	4.74	25.41	0.42	53.10	26.04	17.04	42.82	0.35
std	20.50	4.59	2.07	7.78	0.49	9.82	4.21	24.48	14.61	0.48
min	101.00	0.00	0.98	6.74	0.00	13.00	14.70	0.00	15.00	0.00
25%	124.00	0.05	3.28	19.77	0.00	47.00	22.98	0.51	31.00	0.00
50%	134.00	2.00	4.34	26.12	0.00	53.00	25.80	7.51	45.00	0.00
75%	148.00	5.50	5.79	31.23	1.00	60.00	28.50	23.89	55.00	1.00
max	218.00	31.20	15.33	42.49	1.00	78.00	46.58	147.19	64.00	-

Table 4: Summary Statistics for Variables in the South African Heart Disease Dataset

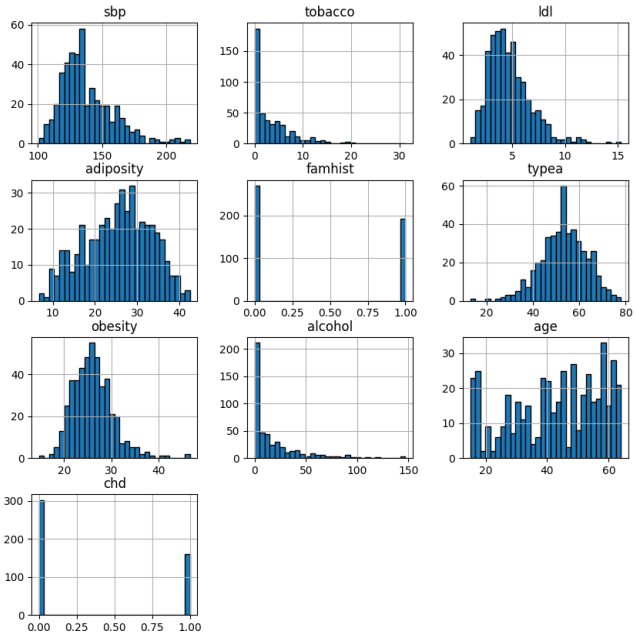

Figure 2: Histogram of all variables in dataset

Based on the summary statistics and their representations in Figure 2 and Table 4, it is shown how alcohol consumption and tobacco usage are right skewed with mean values closer to the lower quartile. This is understandable since most people in the population tend to have low to moderate levels of alcohol and tobacco use while a small number of individuals who consume excessive amounts. The variable *age* is quite spread out with peaks at the lower and upper quartile as well as at the median. Understandably, the histogram of *chd* and *famhist* (after it was transformed) will only show values of 0 and 1.

3 Data visualization(s) based on suitable visualization technique

3.1 Are there issues with extreme values or outliers in the data

From the statistics that we have calculated, we found no extreme values or outliers in the data. Only that the distribution of tobacco, sbp, ldl, obesity and alcohol are right tailed. Even if the alcohol distribution is severely right tailed, it shows no outliers.

3.2 How are the individual attributes distributed (e.g., normally distributed)?

To investigate this question, we plot the histograms of the density of our non-binary data after the transformations, this results are shown in Figure 3. For the alcohol data, we have used the transformation log(x+1) to avoid having logarithms of 0 which is prevalent in the variable.

We have added the normal distribution on top of the plot to first investigate if the data are normally distributed. Note that the KDE function (the purple line) is a nonparametric fit of the data. It can help us visually compare our histogram with the normal distribution (red line). We clearly see that all attributes follow a distribution that is far from normal. To complement these graphs, it was necessary to also plot the Q-Q plots for all the components of the dataset. These plots are shown in Figure 4. These results suggest that many variables in the South African high-risk heart disease

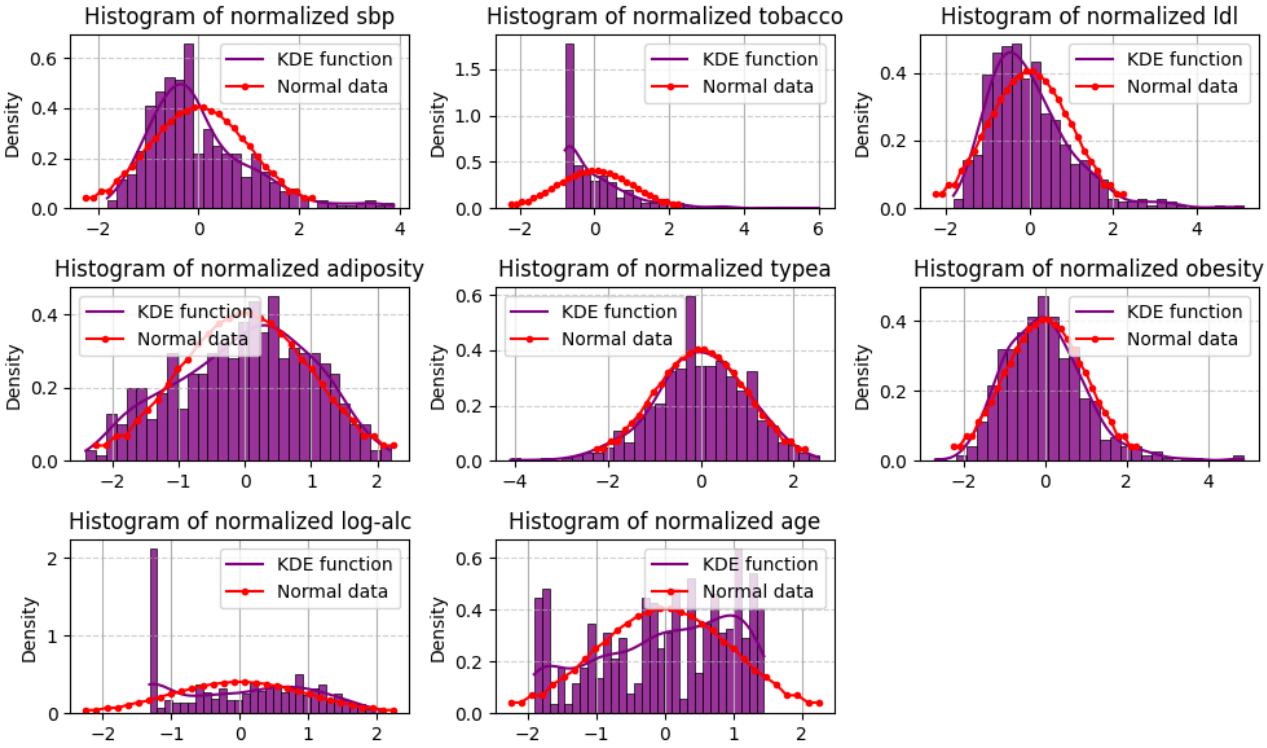

Figure 3: Normalized histogram of all variables in dataset

dataset deviate from a normal distribution, with several exhibiting skewness and heavy tails. While variables like typea appear to follow normality reasonably well, others such as tobacco, ldl, log-alc, and age show significant deviations.

This suggests that the dataset contains variables that may require transformation or robust statistical methods for analysis. The presence of skewed data, particularly in health-related metrics like cholesterol and alcohol consumption, aligns with expectations in epidemiological studies, where extreme values and non-normal distributions are common.

3.3 Are the attributes correlated?

To verify the correlation between the attributes, it was necessary to calculate the correlation matrix to verify the values and understand the correlation between variables. The correlation matrix provides information on the relationships between various health-related variables in the South African high-risk heart disease dataset. The results are shown in Table 5. In particular, adiposity and obesity have a strong positive correlation (0.717), suggesting that higher adiposity is closely associated with higher levels of obesity. Similarly, age shows a moderate to strong correlation with adiposity (0.626), tobacco use (0.450), and LDL cholesterol (0.312), indicating that older individuals in the dataset tend to have higher body fat, smoke more, and exhibit higher cholesterol levels. Systolic blood pressure (Sbp) is moderately correlated with adiposity (0.357) and age (0.389), which aligns with known cardiovascular risk factors. Interestingly, Type A personality traits (TypeA) exhibit weak or negligible correlations with most variables, implying limited direct associations with the measured health markers. The slight negative correlation between TypeA and age (-0.103) suggests that younger individuals in the dataset may have stronger Type A traits.

Overall, we can conclude that our data is significantly correlated. This can be seen as an indication

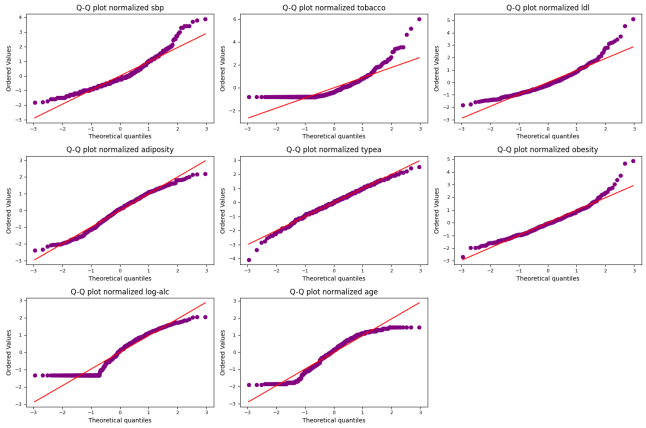

Figure 4: Q-Q Plots of different components

that our data is relevant for our experiment and that further analysis will likely give results. The matrix highlights key health risk factor relationships, with notable associations between age, adiposity, obesity, and cardiovascular markers.

	Sbp	Tobacco	LDL	Adiposity	TypeA	Obesity	Log-Alc	Age	Famhist
Sbp	1.000	0.212	0.158	0.357	-0.057	0.238	0.166	0.389	0.086
Tobacco	0.212	1.000	0.159	0.287	-0.015	0.125	0.198	0.450	0.089
LDL	0.158	0.159	1.000	0.440	0.044	0.331	-0.001	0.312	0.161
Adiposity	0.357	0.287	0.440	1.000	-0.043	0.717	0.126	0.626	0.182
TypeA	-0.057	-0.015	0.044	-0.043	1.000	0.074	0.020	-0.103	0.045
Obesity	0.238	0.125	0.331	0.717	0.074	1.000	0.089	0.292	0.116
Log-Alc	0.166	0.198	-0.001	0.126	0.020	0.089	1.000	0.140	0.076
Age	0.389	0.450	0.312	0.626	-0.103	0.292	0.140	1.000	0.239
Famhist	0.086	0.089	0.161	0.182	0.045	0.116	0.076	0.239	1.000

Table 5: Correlation Matrix

3.4 PCA Analysis

3.4.1 The principal directions of the considered PCA components. Plot and interpret the components in terms of the attribute

In this section, a biplot will be necessary in order to compare the proncipal directions of every attribute from the dataset. The PCA biplot, shown in Figure 5 provides insight into how the different variables contribute to the first two principal components (PC1 and PC2).

The length and direction of the arrows indicate the influence of each attribute on these principal components. Notably, adiposity, obesity, and age have relatively strong contributions, suggesting they play a significant role in the variance captured by these components. Sbp (systolic blood pressure) and tobacco use also contribute moderately, while type A personality (typea) appears to have a distinct direction, possibly capturing a different aspect of variation. The circular reference (unit circle) indicates that most attributes are well represented within the first two components. Overall, the PCA results suggest that variables related to body composition (adiposity, obesity) and lifestyle factors (tobacco,

alcohol) are key drivers of variance in this dataset.

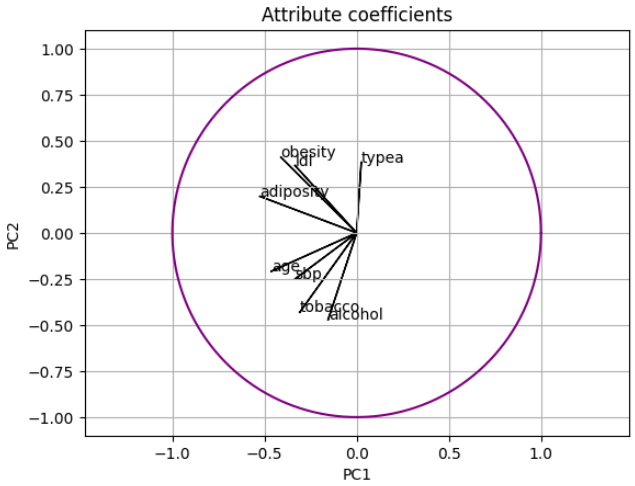

Figure 5: Biplot of attribute coefficients

3.4.2 The amount of variance explained as a function of the number of PCA components included.

Figure 6 shows the plot of the variance explained by the first 8 PCA components. It is concluded that the first two PCA components account for approximately 60 percent of the total variance in the dataset. While this indicates that a significant portion of the data's structure is captured, it also suggests that a two-dimensional representation may not be sufficient for effective visualization. PCA is often useful when the first two components provide a clear separation of data points, but in this case, the explained variance is not high enough to ensure meaningful clustering or pattern recognition. Since a larger number of components are required to capture over 90 percent of the variance, the use of PCA for dimensionality reduction in our machine learning objective may not be as beneficial as expected. This suggests that alternative feature selection or transformation methods might be needed to improve model performance.

3.4.3 The data projected onto the considered principal components, e.g. in 2D scatter plots

As expected, there is no distinct separation between the two classes in the 2D PCA space as shown in Figure 7. This suggests that a simple linear projection using the first two principal components does not provide an optimal separation of individuals with and without heart disease. However, subtle distribution differences can be observed: the *chd yes* points appear more concentrated toward the right side of the plot, while *chd no* points show a higher density on the left. This indicates that while PCA captures some variance relevant to classification, it may not be sufficient as a standalone feature reduction technique for a machine learning model. Non-linear dimensionality reduction methods or the inclusion of additional principal components, might be necessary to improve class separability.

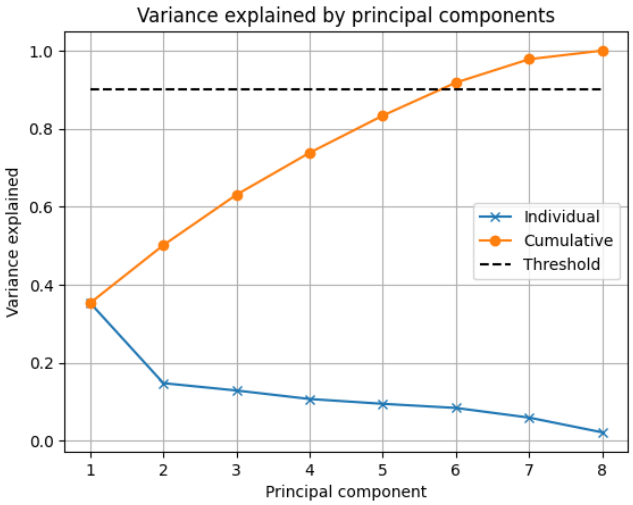

Figure 6: Variance by PCA Components

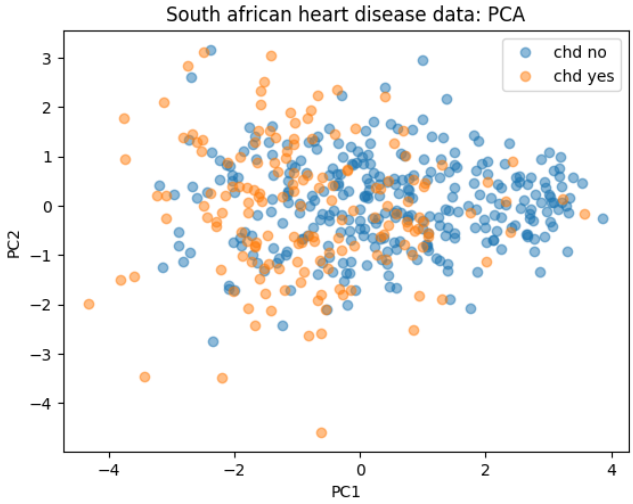

Figure 7: Variance by PCA Components

4 A discussion explaining what you have learned about the data.

4.1 Summarize the most important things you have learned about the data and give your thoughts on whether your primary machine learning aim appears to be feasible based on your visualization.

Based on our data analysis and visualizations, we can conclude that both regression and classification approaches are feasible given the structure and distribution of the dataset.

4.1.1 Regression: Predicting Type-A Behavior

We seek to predict typea as a continuous variable using regression. The feasibility of this approach is supported by the fact that typea was originally measured on the Bortner Short Rating Scale and treated as a continuous attribute before being binarized in the original study. Our analysis of attribute distributions and correlations indicates that several independent variables, such as age, tobacco, and adiposity, show potential relationships with typea. While some variables exhibit skewed distributions, standardization and transformations (e.g., log transformation for alcohol) will help normalize the data and improve model performance. The presence of moderate correlations suggests that a regression model should be able to capture patterns in the data effectively.

4.1.2 Classification: Predicting Coronary Heart Disease (CHD)

For classification, we aim to predict chd as a binary variable (presence or absence of coronary heart disease). Our data exploration highlights strong correlations among key risk factors, such as sbp, ldl, tobacco, and adiposity, all of which are known contributors to cardiovascular risk. The correlation analysis and PCA results indicate that some variables contribute significantly to variance in the dataset, meaning they are informative for classification. The right-skewed distribution of certain attributes suggests that transformations may improve predictive accuracy. Given the distinct clustering tendencies observed in PCA projections, albeit not fully separable, we expect that non-linear classification techniques (such as logistic regression, decision trees, or neural networks) will effectively distinguish individuals with and without CHD.

4.1.3 Overall feasibility of Machine Learning Aim

From the visualizations, correlation matrices, and statistical summaries, both regression and classification tasks appear viable. While PCA indicates that a simple linear projection does not fully separate CHD cases, machine learning models using multiple features should still be able to classify CHD effectively. Similarly, for typea, regression models should capture meaningful relationships despite the presence of skewness in some predictor variables. Proper data preprocessing, including normalization and transformation, will be essential for improving model performance.

Thus, our primary machine learning aims predicting typea through regression and chd through classification are feasible, provided that we carefully preprocess the data and consider non-linear relationships where necessary.