TECHNICAL UNIVERSITY OF DENMARK

Project 1

02450 Introduction to Machine Learning and Data Mining

Authors

Jan Bures - s237197
Mario Calio - s237090
Oscar Wilkins - s236761

February 29, 2024

Contents

1 Report contribution weights **1**

2 Description of the dataset **1**
2.1 Attributes description 2

3 Attributes of the data **2**
3.1 Attribute Classification 2
3.2 Data Issues 3
3.3 Summary Statistics 3

4 Data visualization **5**
4.1 Initial data visualizations 5
4.2 Principal component analysis 6
4.2.1 Variance explained 7
4.2.2 Significant principle components 8

5 Discussion of the results **11**

6 Answers to Exam Questions **11**

1 Report contribution weights

Student ID	Section 1	Section 2	Section 3	Exam quest.
237197	25%	25%	50%	33.33%
237090	50%	25%	25%	33.33%
236761	25%	50%	25%	33.33%

Table 1: Distribution of section contribution weights and exam questions.

2 Description of the dataset

The dataset used in this report is a subset of a wider study named CORIS (Coronary risk factor screening in three rural communities), conducted in 1982 on White-only rural Afrikaans-speaking South Africans in the Western Cape region. More information on the data collection and methods used in the study can be found in the original paper by J.E. Rossouw et al [1].

In this preliminary paper, the data collection methods are explained and an initial analysis of the data is performed. The study was carried out to investigate on an unusual high incidence of ischaemic heart disease (IHD)[2] in the region. Coronary hearth disease CHD and IHD are often used as interchangable terms. The research was conducted within a community predominantly comprising individuals of white ethnicity, representing a minority of less than ten percent within the broader South African population. This demographic specificity can be attributed to the historical context of the apartheid regime that was in place during the study period. Consequently, the findings may not be extrapolated to provide a comprehensive representation of the entire population of South Africa. The original paper then analyses the data collected using simple threshold methods and highlights an high prevalance of risk factor (both genetic and behavioural) in the said population.

The dataset used in this paper, instead, relates the presence of CHD with the values of specific risk factors taken into account. The dataset has been retrieved from the The Elements of Statistical Learning: Data Mining, Inference, and Prediction book by Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome [3]. The dataset was publicly available through the course's website [4].

During the course of the next reports, classification and regression methods will be applied to the dataset. Classification methods will allow the construction of a model capable of predicting the presence of the disease in a subject, given the other attributes. Regression methods will provide a prediction of a continuous parameter (e.g. systolic blood pressure) based on the other attributes.

The attribute named *"famhist"*, present a *Present/Absent* data which has been transformed into a *0/1* value using a simple *python* script.

2.1 Attributes description

The following attributes are taken into account:

- **Systolic blood pressure (sbp)**: Blood pressure in mmHg.
- **Cumulative tobacco (tobacco)**: Consumption of tobacco measured in kilograms. It is not stated if it is yearly or lifetime consumption.
- **Low density lipoprotein cholesterol (ldl)**: LDL is one of the five major groups of lipoprotein that transport all fat molecules around the body in extracellular water. It is sometimes called the "bad" cholesterol because a high LDL level leads to a buildup of cholesterol in the arteries. The unit of measurement is not known [5].
- **Adiposity (adiposity)**: Adiposity is the fact or condition of having much or too much fatty tissue in the body. Its unit of measure in not known.
- **Family history of heart disease (famhist)**: Presence or absence of heart diseases in family history.
- **Type-A behaviour (typea)**: Type-A personality defines individuals who are competitive, hostile, and excessively driven. The relation between type-A personality and heart diseases have been discussed in the past century. [6]
- **Obesity (obesity)**: Obesity measured using a BMI scale.
- **Current alcohol consumption (alcohol)**: Alcohol consumption of the individual. The unit of measurement and time interval of measurement in not known.
- **Age at onset (age)**: Age of the subject when the survey started.

Reponse, coronary heart disease (chd): This is our response parameter, indicates the presence of CHD in the patient.

3 Attributes of the data

3.1 Attribute Classification

The attribute types for this dataset are classified in the following table:

Attribute Name	Disc./Cont.	Attribute Type
Systolic Blood Pressure	Discrete	Ratio
Cumulative Tobacco	Continuous	Ratio
Low Density Lipoprotein Cholesterol	Continuous	Ratio
Adiposity	Continuous	Ratio
Family History of Heart Disease	Discrete (Binary)	Nominal
Type A Behaviour	Discrete	Ordinal
Obesity	Continous	Ratio
Alcohol	Continous	Ratio
Age at Onset	Discrete	Ratio

Figure 1: Attribute Classification Table

The classification of discrete and continuous attributes in this data set can be contested. The formal definition of continuous is that the attribute can take on any value to infinite precision. However, this dataset only features (at most) 2 decimal places of precision for attributes which otherwise would be thought of as continuous. Therefore, an attribute will be classified as continuous in this dataset if it has two decimal places of precision, and discrete otherwise. Another point of interest is that Family History of Heart Disease is discrete and binary. The extra binary classification simply demonstrates that this is a present/absent indicator type of attribute, in other words, a nominal attribute.

Type A Behaviour is an ordinal attribute as the attribute indicates the extent to which someone has a type A personality. As there is no definitive measure of this, Type A Behaviour is more of a qualitative ranking and thus suits the ordinal classification.

The remaining attributes are ratio attributes. This is because a 0 value would indicate an absence of what is measured and it would not make sense for these values to dip into the negative values.

3.2 Data Issues

The chosen data set is of a high quality, however, there are minor data issues. Firstly, an initial inspection of the data revealed that a row of data is missing. The original data rows are labelled from 1 to 463 and it can be seen that row 262 is missing. This should not significantly affect the results of the investigation, but it is worth noting.

Further, there are data issues in the lack of clarity of the attribute descriptions. For the ldl, alcohol and adiposity attributes, units of measurements aren't known. Additionally, for the alcohol and tobacco attributes, an interval of consumption is not defined. Clarity of attribute definitions would help provide real life context for the results of the data analysis, however, it should not affect the results of the analysis.

3.3 Summary Statistics

The following table details the attribute summary statistics:

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

Figure 2: Attribute Summary Statistics Table

The summary statistics provide a good initial overview of the attribute data. Taking in all of the summary data, it can be observed that there are obvious differences in the magnitudes of different variables. The extreme example of this is Systolic Blood Pressure (sbp) and Low Density Lipoprotein Cholesterol (ldl). Sbp contains values between 101 and 218 whereas ldl contains values between 0.98 and 15.33. These values are across significantly different ranges, thus highlighting the importance of standardisation. Without standardisation, it is difficult to make comparisons between attributes. However, one interesting characteristic that can easily be observed is the high standard deviations of the tobacco and alcohol attributes. The maximum value of tobacco is 6 standard deviations from the mean and the maximum value of alcohol is 5.3 standard deviations from the mean. It is likely these attributes have high standard deviations due to the nature of the attribute - people are usually smokers/drinkers or not.

4 Data visualization

4.1 Initial data visualizations

First, we examine the distributions of the attributes. Based on the Figure 3 it can be concluded that none of the attributes perfectly follow a normal distribution. However, certain attributes, such as sbp, adiposity, typea and obesity, are relatively closer to a normal distribution but still show signs of skewness. Conversely, the distributions of the other attributes - tobacco and alcohol in particular - are heavily right-skewed and appear not to follow a normal distribution. Further analysis of Figure 3 reveals that tobacco, ldl, and alcohol exhibit potential outliers at higher values. However, these outliers were considered minor and were thus not removed during the data preprocessing.

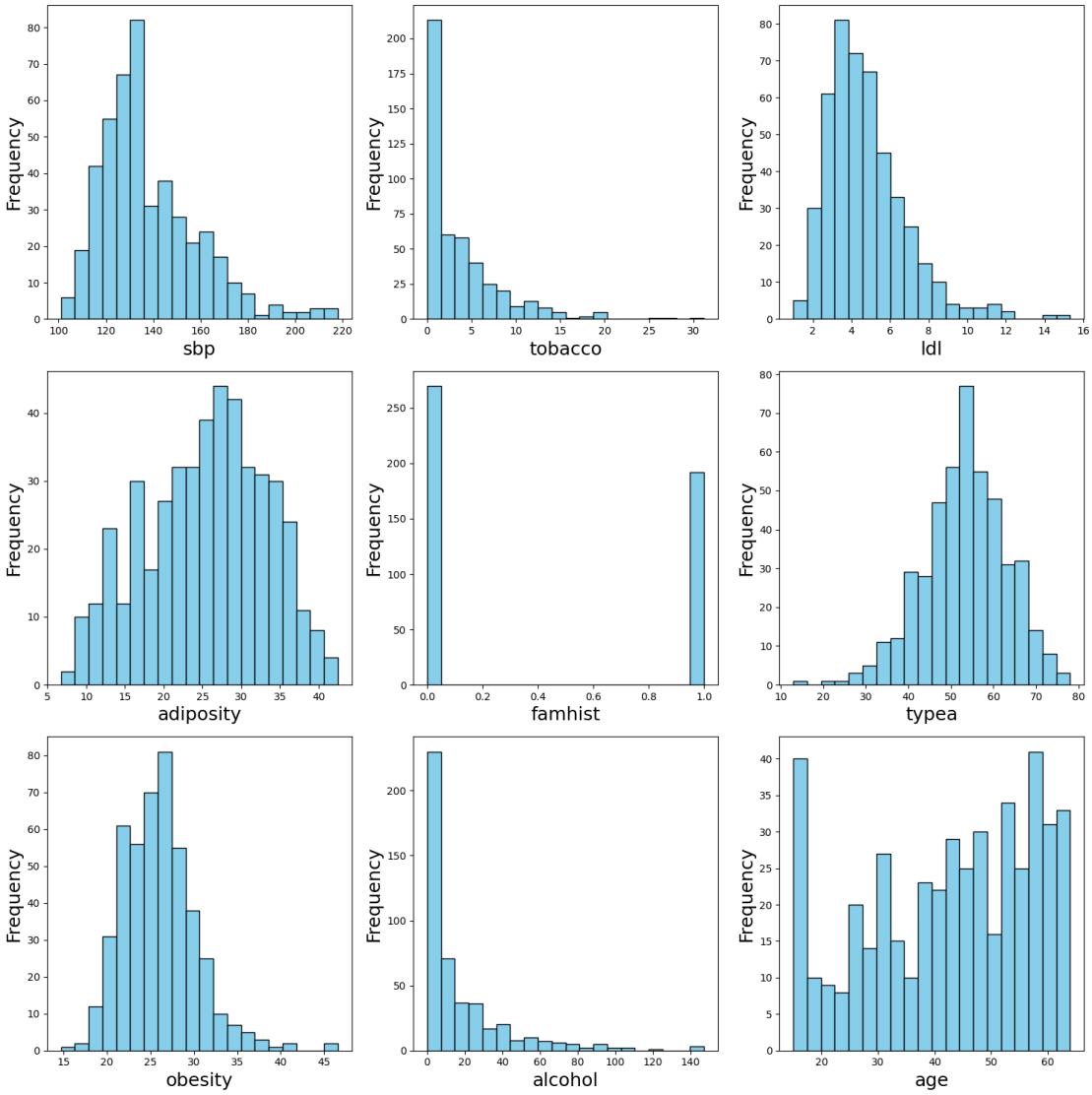

Figure 3: Distributions of the attributes.

Next, the correlation matrix is presented in the Figure 4. Due to the nature of the attributes, adiposity and obesity appear to be correlated. Furthemore, it can be observed that the only other notable non-trivial correlations visible are between adiposity and age, as well as tobacco and age. Beyond these relations, no significant correlation can be seen.

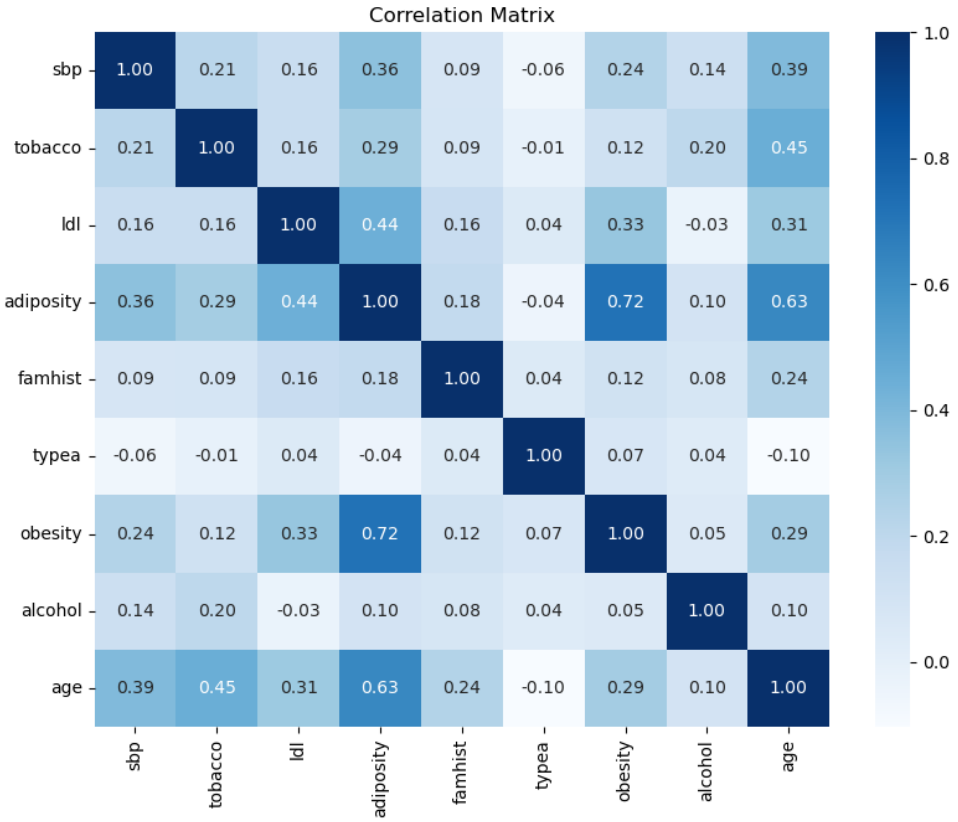

Figure 4: Correlation matrix.

Considering all the information gained from the dataset analysis it can be concluded that the dataset is of high quality and is suitable for the machine learning modeling aim that was set.

4.2 Principal component analysis

In this section, we perform the principal component analysis (PCA) and subsequently analyze the results. We denote the data matrix as $\mathbf{X} \in \mathbf{R}^{N \times M}$.

Before carrying out the PCA, we standardize the data. First, we calculate the mean values and subtract them from $\mathbf{X}$. Second, we divide the resulting matrix by the corre-

sponding standard deviations resulting in the standardized data matrix, which is denoted by $\tilde{\mathbf{X}}$. Formally, we express this as

$$\tilde{\mathbf{X}} = \begin{bmatrix} \cdots & \frac{(X_{1j}-\hat{\mu}_j)}{\hat{\sigma}_j} & \cdots \\ \cdots & \frac{(X_{2j}-\hat{\mu}_j)}{\hat{\sigma}_j} & \cdots \\ & \vdots & \\ \cdots & \frac{(X_{Nj}-\hat{\mu}_j)}{\hat{\sigma}_j} & \cdots \end{bmatrix},$$

where

$$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij}, \quad \hat{\sigma}_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_{ij}-\hat{\mu}_j)^2}, \quad j = \{1, 2, \ldots, M\}.$$

4.2.1 Variance explained

Using the standardized data matrix $\tilde{\mathbf{X}}$ we perform the singular value decomposition (SVD) expressed as

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbf{R}^{N\times N}$, $\mathbf{\Sigma} \in \mathbf{R}^{N\times M}$, and $\mathbf{V}^\top \in \mathbf{R}^{M\times M}$. Additionally, $\mathbf{\Sigma}$ is a rectangular diagonal matrix with non-negative singular values on the diagonal. We denote the singular values as $s_1, s_2, \ldots, s_M$. Note, that it follows from the properties of SVD that the singular values satisfy $s_1 \geq s_2 \geq \cdots \geq s_M$. We calculate the cumulative explained variance of the first K principle components as

$$\frac{\sum_{i=1}^{K} s_i^2}{\sum_{i=1}^{M} s_i^2}, \tag{1}$$

the individual variance explained by the j'th principal component is then calculated as

$$\frac{s_j^2}{\sum_{i=1}^{M} s_i^2}. \tag{2}$$

Equations 1 and 2 are used to calculate the explained variance, which is illustrated in Figure 5. Values of explained variance for individual components is presented in Table 2.

It can be seen, that in order to explain more than 90 % of the total variance, we must include at least seven principal components (PCs). Moreover, the usage of only the two most significant PCs accounts for only 45.35 % of the total variance, which means that limiting our further analysis to these two components results in a significant loss of variance.

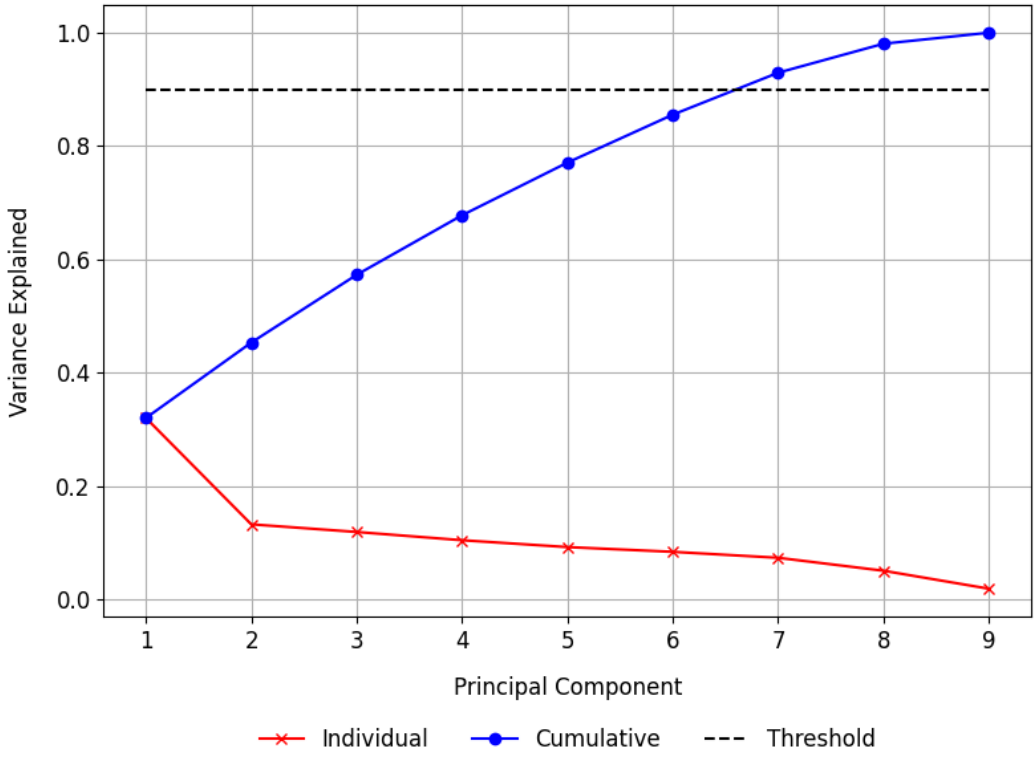

Figure 5: Variance explained by principal components. Threshold equal to 90 % of explained variance was chosen as a reference value.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
32.05%	13.30%	11.95%	10.50%	9.29%	8.45%	7.40%	5.12%	1.94%

Table 2: Explained variance of each principal component.

4.2.2 Significant principle components

In this section, the analysis is restricted to the first two principal components denoted as PC1 and PC2. To interpret the meaning of PC1 and PC2, we examine the coefficients that form the eigenvectors. The magnitude of these coefficients reflects the different levels of importance of the attributes. The coefficient values are presented in Figure 6, the direction of the attributes coefficients can be seen in Figure 7.

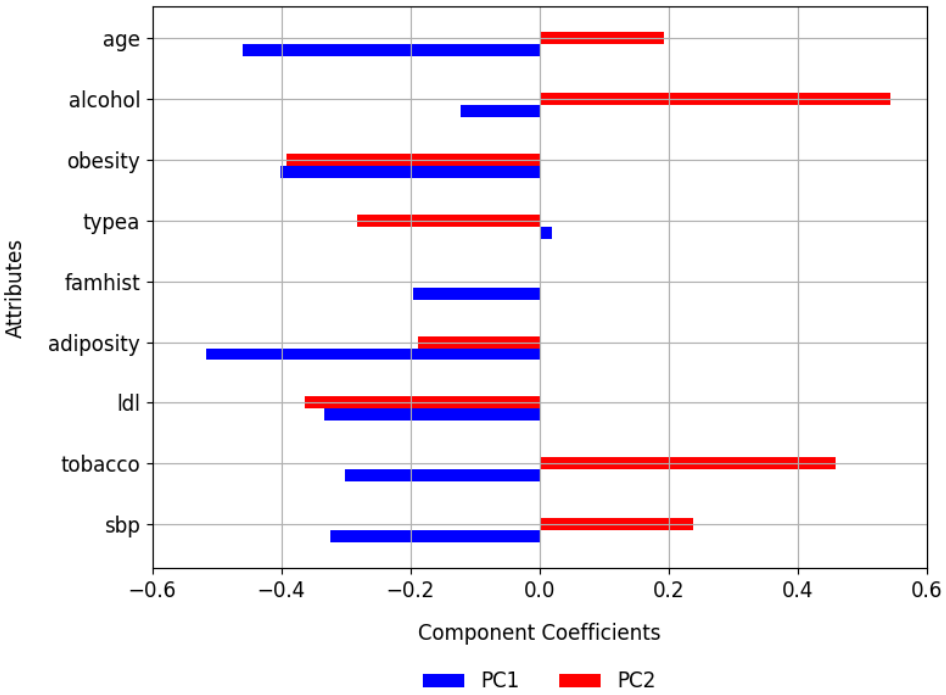

Figure 6: PC1 and PC2 coefficients.

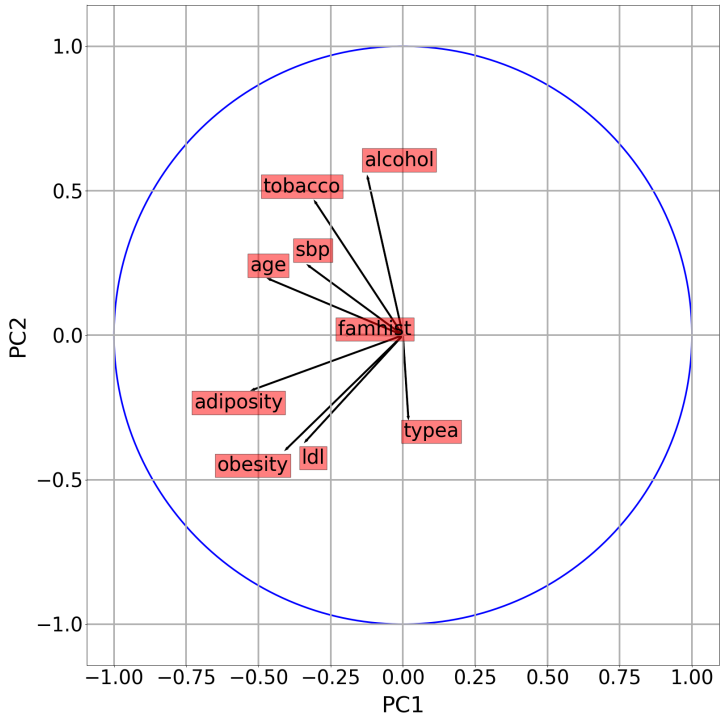

Figure 7: Direction of the attributes coefficients in the subspace generated by PC1 and PC2.

Figure 6 shows that PC1 mainly reflects characteristics associated with older age and higher levels of body fat - both obesity and adeposity (or it can also represent the absence of these traits). Furthermore, PC2 appears to capture the variance associated with the consumption of unhealthy substances such as alcohol and tobacco. Similarly, the importance of the attributes can be examined by inspecting Figure 7, where the main factors are the length and direction (alignment with specific PC) of each attribute.

Finally, the data can be projected onto the subspace generated by PC1 and PC2. This projection is shown in Figure 8. It can be seen that there is no apparent separation between the patients with CHD and without CHD when projected onto the first two PCs. This may be linked to the fact that PC1 and PC2 alone do not explain majority of variance, as previously discussed. It can, however, be observed that patients with CHD present have the tendency to have higher values of PC1 and values of PC2 closer to zero. This is consistent with the previously discussed interpretation of PC1 being the component mainly reflecting the age and body fat level of individuals - naturally, it is reasonable to expect that younger people with lower level of body fat are less likely to be diagnosed with CHD. On the contrary, people with diagnosed CHD are spread across the entire plot following no apparent rule - this suggests a higher variance within the group of diagnosed individuals. Thus, it can be concluded that the presence of CHD may not be fully explained using just the first two principal components.

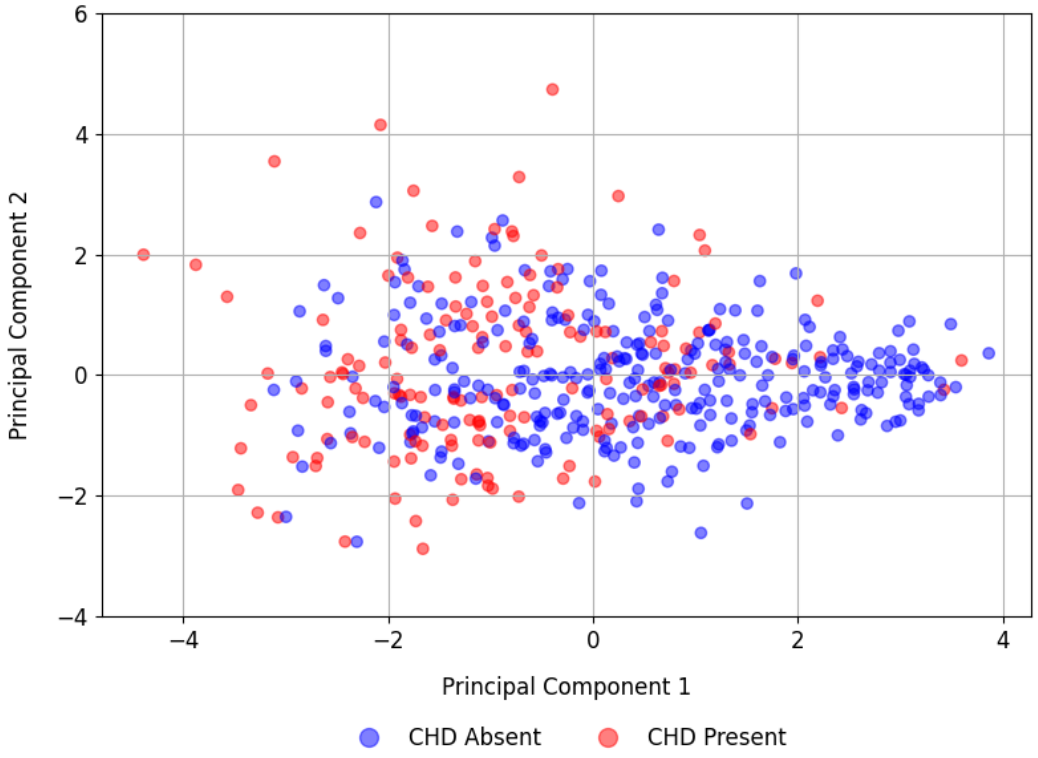

Figure 8: Data projected onto the subspace generated by PC1 and PC2. The points are distinguished by different color based on the presence of coronary heart disease (CHD).

5 Discussion of the results

This preliminary report presented an analysis of a dataset derived from the CORIS study, focusing on the correlation between various risk factors and the presence of Coronary Heart Disease (CHD). The dataset, originally collected in 1982, involved a predominantly white Afrikaans-speaking population in the Western Cape region, thus limiting the generalizability of findings to the broader South African population due to demographic specificity.

The dataset proved to be of great quality from an analytical point of view and required almost no cleaning. The dataset, though, was lacking some clarity surrounding the attribute definitions, which might make real-life interpretation of the findings slightly more difficult.

A careful description of the attributes was carried out along with basic statistical analysis such as computation of mean, standard deviation and correlation. The summary statistics were in line with what was expected. Namely, there was high variation in alcohol and tobacco consumption.

Data visualisations further supported the suitability of the data. The histograms indicated that a majority of attributes followed a loose normal distribution. In addition, the correlation matrix showed that whilst there were strong correlations between attributes (e.g. adiposity and obesity), the attributes were different enough so as to provide the dataset with additional information and dimensions for machine learning models to explore.

Principal Component Analysis (PCA) was then deployed, revealing that seven principal components were needed to explain more than 90% of the total variance. The analysis of the first two principal components (PC1 and PC2), which together account for less than 50% of the total variance, struggled to highlight the presence of CHD based solely on these components, emphasizing the need for more advanced approaches. Such approaches will be subjects of further report and might include, for example, classification and regression tasks.

Given the high quality of the dataset and the successful, informative visualisations given by the data analysis and PCA visualisations, it can be concluded that this dataset is appropriate for our machine learning aim - predicting coronary heart disease.

6 Answers to Exam Questions

Question 1

Question 1 asks us to determine the statement which correctly types the attributes from the Urban Traffic dataset. The following method of deduction was used:

- A is incorrect. This is because Time of Day is not a nominal attribute. A nominal attribute is one which contains categories. The half an hour intervals of time of day are not categories. Time of day is rather an interval attribute.
- B is incorrect. Immobilised Bus is not a nominal attribute. A nominal attribute is one which contains categories. The number of buses which are immobilised is not a category-type attribute. Immobolised Bus is more of a ratio attribute
- C is incorrect. As discussed before, Time of Day is interval. Time of day is not ordinal.

- D is correct. Time of day is interval (as discussed), Traffic lights is ratio as the number of broken traffic lights is a count where 0 represents an absence of what is measured, Running over is a ratio for the same reason and congestion level is ordinal as congestion level is ranked somewhat qualitatively from low to high.

Question 2 Question 2 asks us to determine which p-norm distance statement is correct. The following method of deduction was used:

- A is correct. $d_{p=\infty}$ tells us to find the max norm distance between the two vectors. The max norm distance occurs in the first element of both vectors (26-19 = 7). Therefore, $d_{p=\infty}(x_{14}, x_{18}) = 7$.
- B is incorrect. Using General Minkowsky Distance ($d_p = (\sum_{i=1}^{M} |x_j - y_j|^p)^{1/p}$), we have $d_{p=3} = (9^3 + 2^3)^{1/3} = 9.08$. The solution suggests the answer is 3.688 which is incorrect.
- C is incorrect. Using General Minkowsky Distance ($d_p = (\sum_{i=1}^{M} |x_j - y_j|^p)^{1/p}$), we have $d_{p=1} = (9^1 + 2^1)^{1/1} = 1.286$. The solution suggests the answer is 3.688 which is incorrect.
- D is incorrect. Using General Minkowsky Distance ($d_p = (\sum_{i=1}^{M} |x_j - y_j|^p)^{1/p}$), we have $d_{p=4} = (9^4 + 2^4)^{1/4} = 9.005$. The solution suggests the answer is 4.311 which is incorrect.

Question 3 Question 3 asks us to identify which statement about explained variance is correct. Recall that the S matrix contains the standard deviations of each of the principal components in order. Also recall that the formula for explained variance is $\frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$. Knowing this, the following method of deduction was used:

- A is correct. Using the above explained variance formula, $\frac{13.9^2+12.47^2+11.48^2+10.03^2}{13.9^2+12.47^2+11.48^2+10.03^2+9.45^2} = 0.867 > 0.8$.
- B is incorrect. Using the above explained variance formula, $\frac{11.48^2+10.03^2+9.45^2}{13.9^2+12.47^2+11.48^2+10.03^2+9.45^2} = 0.48$. 0.48 is not greater than 0.51
- C is incorrect. Using the above explained variance formula, $\frac{13.9^2+12.47^2}{13.9^2+12.47^2+11.48^2+10.03^2+9.45^2} = 0.52$. 0.52 is not less than 0.5.
- D is incorrect. Using the above explained variance formula, $\frac{13.9^2+12.47^2+11.48^2}{13.9^2+12.47^2+11.48^2+10.03^2+9.45^2} = 0.716$. 0.716 is not less than 0.7

Question 5 Question 5 asks us to determine the Jaccard similarity between two vectors with bag-of-words encoding. Recall that the Jaccard similarity formula is $J(x, y) = \frac{f_{11}}{k - f_{00}}$. f_{11} denotes words which are shared between both vectors. In this case, the words 'the' and 'words' are shared so f_{11} is 2. k denotes the total number of possible words. Assuming a total vocabulary of 20,000, k = 20,000. f_{00} is the number of words which are in the vocabulary (k) but are contained in neither vector. f_{00} is 20,000 - 11 - 2 = 19,987. Therefore, using the formula, $J(s_1, s_2) = \frac{2}{20{,}000 - 19{,}987} = 0.1538$. This corresponds to answer A. Therefore, A is correct.

References

[1] J. E. Rossouw, J. P. Du Plessis, A. J. S. Betadje, P. L. Jooste, P. C. J. Jordaan, and J. P. Kotze, "Coronary risk factor screening in three rural communities. the coris baseline study.," *South African Medical Journal*, vol. 64, no. 12, pp. 430–436, 1983.

[2] "Coronary heart disease." `https://www.nhs.uk/conditions/coronary-heart-disease/`.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, NY, 2 ed., 2009.

[4] "The elements of statistical learning: Data mining, inference, and prediction." `https://hastie.su.domains/ElemStatLearn/`.

[5] "LDL: The "Bad" Cholesterol." `https://medlineplus.gov/ldlthebadcholesterol.html`.

[6] S. Sahoo, S. K. Padhy, B. Padhee, N. Singla, and S. Sarkar, "Role of personality in cardiovascular diseases: An issue that needs to be focused too!," *Indian Heart J.*, vol. 70, no. Suppl 3(Suppl 3), pp. S471–S477, 2018.