Danmarks
Tekniske
Universitet

DTU

Project 1

Authors

Filippo Bosi - s220015 (section 1 & 4)
Davide Venuto - s220331 (section 2)
Aleksander Nagaj - s220350 (section 3)

March 8, 2022

Contents

1 Description of the data set (s220015)

The *South African Heart Disease* data set collects samples regarding the risk factors linked to the Coronary Heart Disease (*chd*) in the region of the Western Cape, South Africa. The data-collection process was part of the Coronary Risk Factor Study (CORIS) which represents the first attempt to quantify *Ischemic Heart Disease* (IHD) risk factors in an Afrikaans-speaking community.

The data gathered here are a subset of a larger data set, described by Rousseauw et al, 1983 [1]. Furthermore, this data set was deeply studied in Hastie et al, 2009 [2], with the aim of generating a classification model for *chd*.

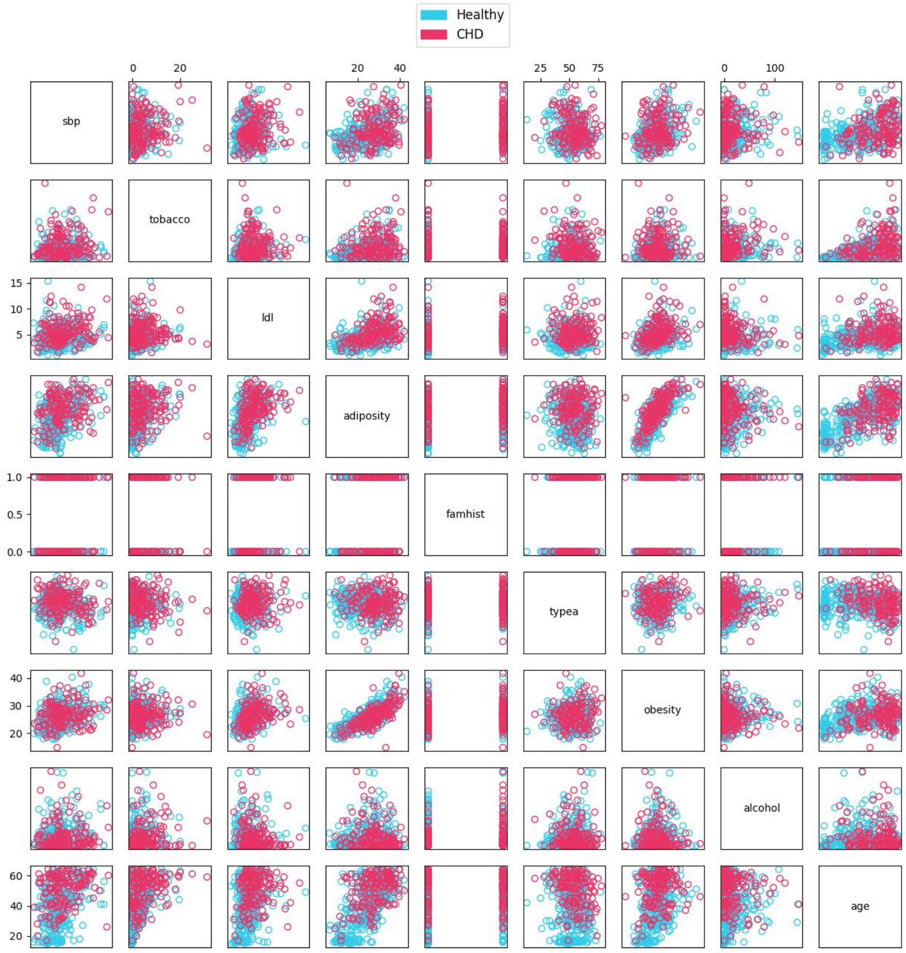


Figure 1: A scatterplot matrix of the attributes

In their paper, the authors fitted a linear logistic regression model by maximum likelihood to the data set. They calculated a Z score for each of the attributes, and by using the Wald test, they were able to discard some of the terms from the model. In fact, the information carried by these variables could be obtained from the other correlated attributes.

The Wald test is a method that can be used to test for exclusion of a term without requiring iterative fitting. This allows to perform efficient computations, making the model-building process quicker.

Their analysis proceeds with the selection of a model performed by dropping the least significant coefficients, and refitting the model until no further terms can be dropped from it. After this preliminary study, the authors implemented a nonlinear model by the means of natural splines. As a result, the previously excluded terms can now be included again in the final nonlinear model. These nonlinear effects are related to the fact that the measurements were collected some time after the patients suffered from *chd* and in several cases, they had already opted for a healthier lifestyle.

At this early stage we would like to perform classification of the *chd* attribute as well as attempt to predict the variable *tobacco* given all other terms. We plan to standardize the data by subtracting the mean and dividing it by the standard deviation. Furthermore, we are also considering carefully analyzing the outliers and possibly removing them based on an empirically derived threshold.

2 Attributes of the data (s220331)

The South African heart disease study use a subset of the "CORIS" data set which were collected through a coronary risk factor survey in the southwestern Cape Province in 1979; 3357 males and 3831 females, coming from the White population of the region between the age of 15 and 64 years, were recruited covering 82% of the target population[1].

The South African data set include a sample of only 462 people from the previous data set and includes 10 data Attributes. Which are: Systolic blood pressure (Sbp), Cumulative Tobacco (kg), Low-density lipoproteins (Ldl), Adiposity, Family history of heart disease (Famhist), Type A behaviour (more hard driving, ambitious and time conscious people), Obesity using body mass index (BMI), Alcohol (Current alcohol consumption)[2], Age, Coronary heart

[1] More data collection information: data were collected in cities with similar in cultural and socioeconomic structure:magistral districts of SwelIendam (White population 5860), Riversdale (5 540) and Robertson (5 320); mortality rates from IHD in economically active White males ,averaged 214/100000 (85/100 000 for females. Instead, the national figure for White SouthAfrican males was 240/100000 (72/100000 for females) in that year by means.

[2] The description of Alcohol, Adiposity, and Tobacco attributes in the reference papers are not complete. Adiposity values meaning is not clear described, they range between 6 and 43, after some researches medical papers let us suggest that it could be a ratio of body fat; Alcohol data measure unit is not specify; for Tobacco is not specify how data are collected and the relation between the cumulative data and time.

disease (Chd).

	Unit and Range	Variable Type	Additional Description
Systolic blood pressure (sbp)	mmHg	Discrete, Ratio	Blood pressures were measured after subjects had been seated for 5 minutes. A standard 12,5 x 23-cm cuff connected to a mercury manometer was used. The American Heart Association guidelines for measuring blood pressure.
Tobacco	Cumulative tobacco (kg)	Continuous, Ratio	Cumulative Time not specified.
Low-density lipoproteins (Ldl)	mg/dl	Continuous, Ratio	Non-fasting blood samples were taken with minimal stasisinto. Serum was separated within 2 hours and analysed manually on the same day for serum cholesterol by means of the Boehringer CHOD-PAP enzymatic method and dextran sulphate-magnesium chloride precipitation.
Adiposity	Bodyfat percentage	Continuous, Ratio	Interpretation of the value range.
Family history of heart diseas (Famhist)	Present:Absent	Binary	Presence of chd cases in family history.
TypeA behaviour	Index of pattern A behaviour. Higher value indicate as exhibiting type A characteristics	Discrete, Interval	Bonner RW. Ashort rating scale as a potential measure of pattern A behaviour. Chronic Dis 1969; 22: 87-91. They used a questionnaire with 14 questions to try to
Obesity Using Body mass index (BMI)	$weight/height^2$	Continuous, Interval	Bray GA. Definition, measurement, and classification of the syndromes of obesity. [ne] Obes 1978; 2: 99-112. BMI >= 30 scored as "obese" Rossouw (1983)
Alcohol	Current alcohol consumption	Continuous, Ratio	Unit of Measurement not clarified in the main paper.
Age	15–64 years	Discrete, Ratio	Sample of people between 15-64 years old.
Coronary heart disease (Chd)	1(heart disease):0	Binary	There are 160 cases in our data set, and a sample of 302 controls. At the time the overall prevalence of MI was 5.1% in this region). These data are described in more detail in Hastie and Tibshirani (1987).

Table 1: Attributes of South African Heart Disease Study.

	sbp	**tobacco**	**ldl**	**adiposity**	**famhist**
mean	138.33	3.64	4.74	25.41	0.42
std	20.47	4.59	2.07	7.77	0.49
min	101.0	0.0	0.98	6.74	0.0
25%	124.0	0.05	3.28	19.77	0.0
50%	134.0	2.0	4.34	26.12	0.0
75%	148.0	5.5	5.79	31.23	1.0
max	218.0	31.2	15.33	42.49	1.0

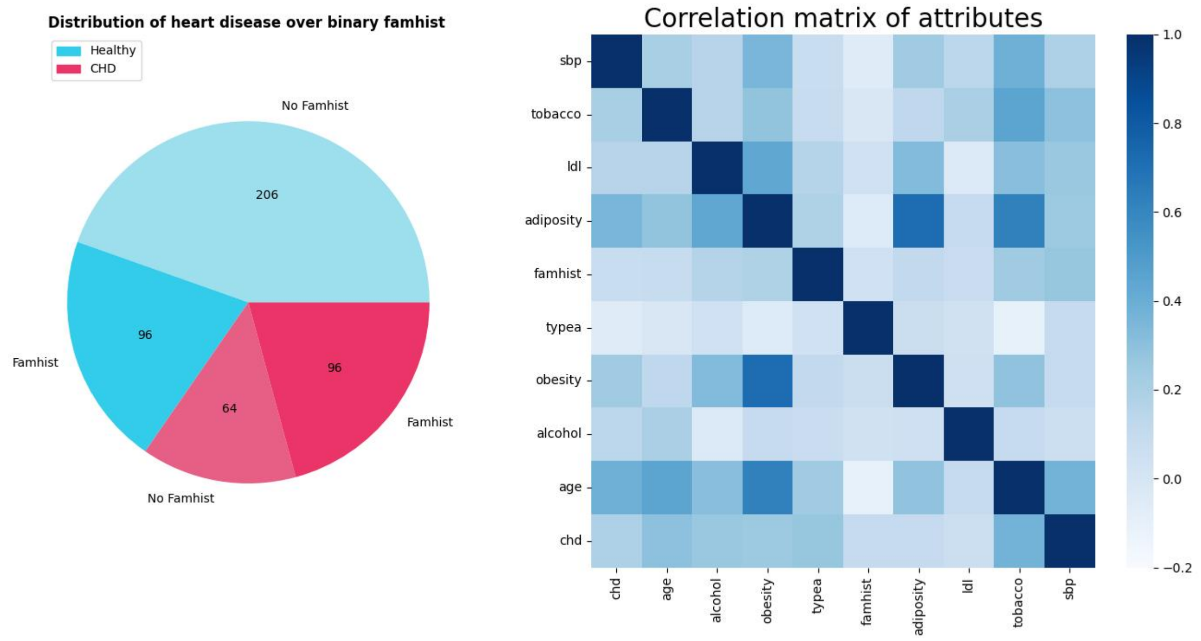
	typea	**obesity**	**alcohol**	**age**	**chd**
mean	53.1	26.04	17.04	42.82	0.35
std	9.81	4.21	24.45	14.59	0.48
min	13.0	14.7	0.0	15.0	0.0
25%	47.0	22.98	0.51	31.0	0.0
50%	53.0	25.8	7.51	45.0	0.0
75%	60.0	28.5	23.89	55.0	1.0
max	78.0	46.58	147.19	64.0	1.0

Table 2: Attributes statistics overview

In figure 2 basic statistics of the data attribute are provided.

Should be noted that 35% of the sample of 462 people have had heart disease problems in life and the average age of the sample is 43 years. In addition, it could be seen how overall *typeA* score is 53 from a 82 scale and 42 percent of the sample has a family history of heart disease; in addition in figure 2a (relation between chd and famhist) could be seen that half of the sample with a heart disease history have had *chd*; on the contrary only 24% of the sample without any *famhist* is in the *chd* positive group.

From the statistics summary could be also noticed how mean and standard deviation of each attribute changes a lot, for that reason before any further analysis (PCR) or machine learning method the data will be standardized (subtracting the mean and dividing for the standard deviation – this do not occur for binary variables).



(a) Chd-Famhist binary visualization

(b) Correlation matrix of attributes.

Figure 2

In figure 2b the correlation between attributes could be seen. It shows that obesity and adiposity are the most correlated factors. However, age and adiposity are the attribute that present the most overall correlation. In the other hand typeA is the one that seem to be less correlated with the other attributes. Furthermore, it could be noted that the correlation between factors in general doesn't seem to be to very marked.

Finally it appear also evident that chd (future dependent variable) does not present a strong mathematical correlation with all the risk factors[3].

[3] For the correlation between binary and interval-ratio variables the Point-Biserial Correlation Coefficient is used (famhist, chd).

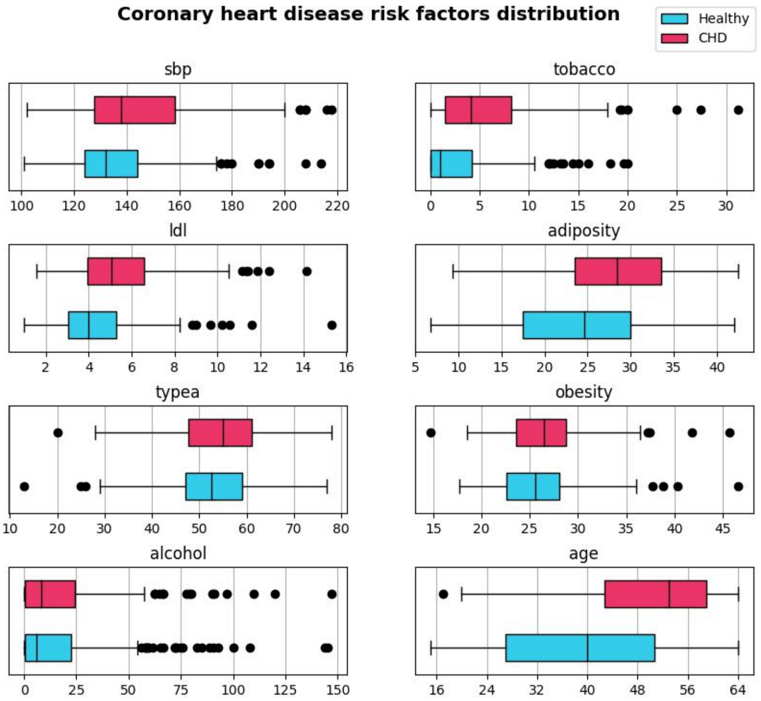


Figure 3: CHD risk factors distribution

To conclude the statistical data description and visualization, some box plots (figure 3) where chd crossed with attribute information are provided. Unfortunately, as it could be seen in the charts a clear visual classification could not be done; in fact the box plots of chd and healtz people have a wide intersection all over the risk factors. (the mean of chd is also often in the range 2575 quantile from the "healty people" plot). In conclusion it could be said that this visual analysis are not sufficient for have sufficient evidences of ,if and how, the risk factors affect chd variables. Further analysis methods like classification and regression would be useful to have a deeper insight on how chd is related with the risk factors taken in analysis.

3 Principal Component Analysis (s220350)

The goal of Principal Component Analysis, as mentioned in [3], is to reduce dimensionality of the data set where it is assumed that the lower-dimensional representations are linear. *South African Hearth Disease* data consists of $N = 462$ observations $x_1, x_2, ..., x_N \in \mathbb{R}^M$ each consisting of $M = 9$ attributes. To perform the analysis, there was selected number $n \leq M$ for which there was found n-dimensional representation of the data

$$b_1, b_2, ..., b_N \in \mathbb{R}^n$$

In order to transform vectors from a M to n dimensional space, they can be projected onto n-dimensional subspace V with orthonormal basis $v_1, v_2, ..., v_n$. Each b_i is defined as the projection of x_i onto this space. This is a moment where *Singular value decomposition* (SVD) comes in handy. It allows to decompose any $N \times M$ matrix $\mathbf{X}$ into three matrices

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

such that

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{X} \tag{1}$$

where $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_M$ are *singular values* of $\mathbf{X}$ and vectors $v_1, v_2, ..., v_M$ are orthonormal. If we then choose a subset of n eigenvectors corresponding to first n eigenvalues, which are simply squared singular values, we receive a desired subspace $\mathbf{V}$.

Last but not least, after the projection of M onto V has been performed, *variance explained* is computed. It is a measure of how much variance of the original data is preserved in n principal components.

$$VarianceExplained = \frac{\sum_{i=i}^{n} \sigma_i^2}{\sum_{i=i}^{M} \sigma_i^2} \tag{2}$$

All these steps were applied to *South African Heart Disease* data set in the following manner

1. Remove outliers for each attribute: $x_{ij} \geq \sigma_j$, where x_{ij} is the *i-th* observation for *j-th* attribute and $\sigma_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \tilde{x_{ij}}}$.

2. Subtract the mean: $\tilde{x_i} - m$, where $m = \frac{1}{N}\sum_{i=1}^{N} x_i$.
3. Divide by standard deviation: $\tilde{x_{ij}} = \frac{\tilde{x_{ij}}}{\sigma_j}$.
4. Compute the SVD (1) of $\tilde{X}$.
5. Calculate variance explained by principal components and find number of vectors v_i which hold 90% of variability in the data (figure 4)

3.1 PCA analysis

We performed PCA for both data with **zero mean** and **standardized** set. This way allowed a comparison of not only principal components but also effect of unit standard deviation.

There is a significant difference in a number of components accounting for 90% of variability, which we chose as a threshold for dimensionality reduction (figure 4). For the zero-mean data set only the first three components hold 88.25% of overall variance, whereas for standardized set the cumulative variance explained reaches the threshold for the seventh (out of nine) principal component. The leading three account for only 61.21% variance. On the one hand, data standardization enables unbiased comparison of attributes and finding their correlations. On the other hand, it significantly decreases the possibility to reduce the dimensionality.

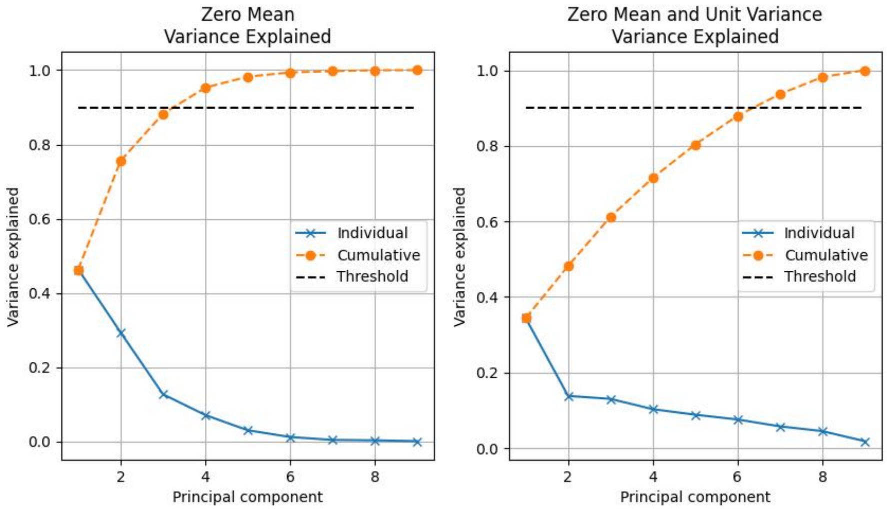


Figure 4: Variance Explained.

The **attribute coefficients vectors** for the first two principal components shows clearly why standardization is crucial. As shown at figure 5, the direction and magnitude of each attribute's eigenvector defines how much data from given attribute is projected onto PC1/PC2 space. For the zero-mean, most of value is stored in *alcohol*, *sbp* and *tobacco*, whereas for the standardized set, the data is more evenly distributed. The mentioned attributes have simply scale of higher magnitude compared to others, therefore the dominate a non-standardized data.

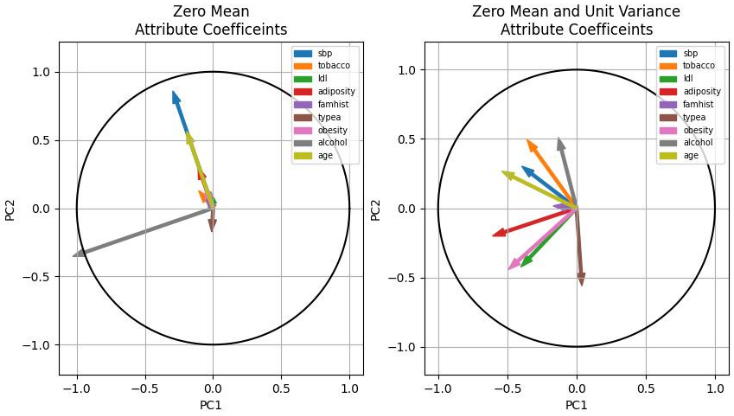


Figure 5: Attribute Coefficients Vectors at PC1/PC2

What is more, projection onto PC1/PC2 space neither for zero-mean nor standardized data allows to classify observations. It can be seen on figure 6 that standardized observations with diagnosed *chd* lean more towards negative PC1 values, but it is far from being a pattern which would allow to draw any distinction.

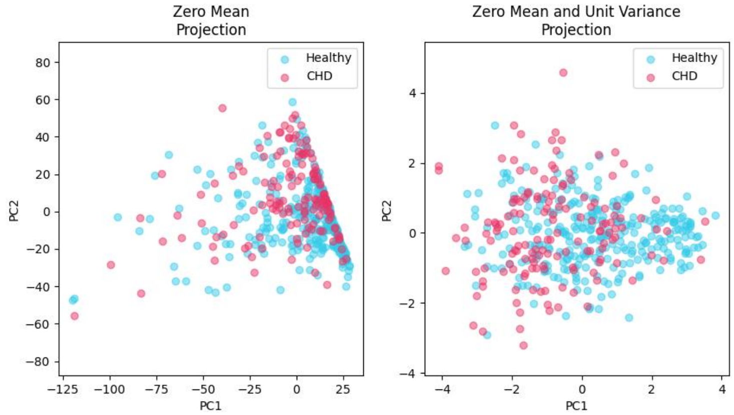


Figure 6: Projection onto PC1/PC2.

Taking a closer look into standardized data set attributes' coefficients for 6 PC's at figure 7, which account for nearly 90% of variance explained, it is difficult to find a dominant component. *Sbp* contributes most for *5-th* PC, *tobacco* for *4-th* and ldl for *6-th*. *Alcohol* has strong influence over each except the *1-st* PC, whereas *famhist* is barely projected on any PC. The last behaviour was expected, since *famhist* is a binary variable.

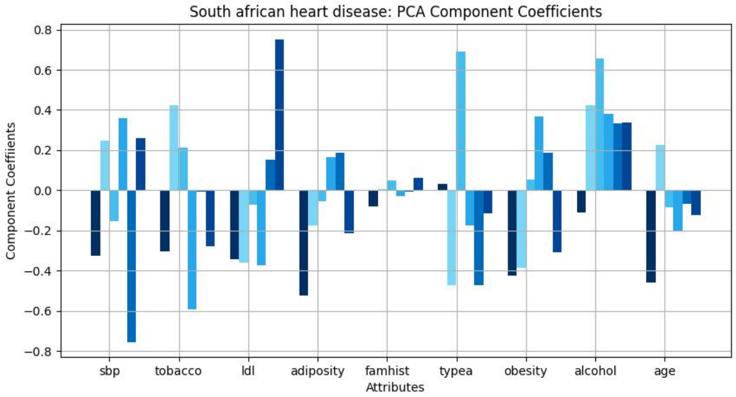


Figure 7: PCA Component Coefficients.

4 Final discussion (s220015)

With this preliminary analysis of the *South African Heart Disease* data set we were able to gain a deeper understanding of how the attributes are characterized and correlated. We have also found out that, due to the large difference in the standard deviation of the variables, the standardization of the data was necessary in order to perform a valuable PCA.

However, from the visual projection of PC1 and PC2 it was still difficult to draw a clear decision boundary between the presence and the absence of *chd*. Hence, it seems quite hard to carry out the classification task based only on the PCA itself.

Furthermore, by applying PCA we were forced to consider seven (out of nine) singular components to get at least the 90% of explained variance. As a result, the model does not reach the expected level of simplification.

After taking these considerations into account, we do not plan to carry out a PCA-based classification task. Instead, we plan to consider alternative methods such as Logistic Regression and other more advanced Machine Learning tools.

Finally, with this project, we have gained a better understanding of the importance of presenting results, which is often almost as important as the results themselves.

References

[1] J. Rousseauw, J. du Plessis, A. Benade, P. Jordaan, J. Kotze, P. Jooste, and J. Ferreira, “Coronary risk factor screening in three rural communities, south african medical journal 64,” pp. 430–436, 1983.

[2] T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning: Data mining, inference, and prediction. second edition.,” pp. 122–124, 146–148, February 2009.

[3] T. Herlau, M. N. Schmidt, and M. Morup, “Introduction to machine learning and data mining,” pp. 29–49, 2021.

A Exam Questions

A.1 Question 1

D is correct. *Time of fay* is interval since it is given in 30-minute intervals, there is no true zero; *Traffic lights* and *Running over* are both ratio, because there is a true zero; *Congestion level* is ordinal, because it is a discrete attribute which values can be ordered.

A.2 Question 2

A is correct. The equation for the $p-norm$ is given as

$$max\{|x_1 - y_1|, |x_2 - y_2|, ..., |x_M - y_M|\}$$

Substituting values of x_{14} and x_{18} we get

$$max\{|26 - 19|, |0 - 0|, |2 - 0|, |0 - 0|, ..., |0 - 0|\} = 7$$

A.3 Question 3

A is correct. In order to compute the variance explained by the first four components we use following formula.

$$VarianceExplained = \frac{\sum_{i=i}^{4} \sigma_i^2}{\sum_{i=i}^{5} \sigma_i^2}$$

Putting the values we obtain

$$VarianceExplained = \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.87$$

A.4 Question 4

D is correct. *Broken Truck*, *Accident victim*, *Defects* have a high value and a positive projection onto v_2, which will overwhelm the low negative projection of *Time of Day*. Hence PC2 will have a positive value.