Danmarks Tekniske Universitet

DTU

Report 1 - Data: Feature Extraction and Visualization

Group 131

Elysia Gao - s222445
Marton Bertalan Limpek - s221835
Katja Naasunnguaq Jagd - s185395

October 4, 2022

Contents

1 Mandatory section

1.1 Contribution

Students	Section covered
Marton	Description of the dataset
Katja	Detailed explanation of the attributes of the data
Elysia	Data visualization
Marton, Katja, Elysia	Discussion and conclusion

1.2 MC questions

Q1 - D

As *Time of day* are intervals of equal size (30 minutes), the attribute is interval. 0 has a physical meaning for *Traffic lights* and *Running over* therefore they are both ratio. Finally Congestion level is ordinal as the congestion level can be ranked from low to high.

Q2 - A

$$\mathrm{d}_\infty(\boldsymbol{x},\boldsymbol{y}) = \max\left\{|x_1-y_1|, |x_2-y_2|, \ldots, |x_M-y_M|\right\} \; x_{14}-x_{18} = \begin{bmatrix} 7 & 0 & 2 & 0 & 0 & 0 & 0 \end{bmatrix}^T$$
$$\mathrm{d}_\infty\left(x_{14}, x_{18}\right) = 7$$

Q3 - A

$$\frac{\sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2+\sigma_5^2} = 0.867 > 0.8$$

Q4 - D

The second principal component is composed of 1 negative value followed by 4 positive. The magnitude of the attributes mentioned in D will result in a positive value projected onto this principal component.

$${v_2}^T = \begin{bmatrix} -0.5 & 0.23 & 0.23 & 0.09 & 0.8 \end{bmatrix}$$

Q5 - A

$$\begin{matrix} s_1 \\ s_2 \end{matrix}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$J\left(s_1, s_2\right) = \frac{f_{11}}{f_{11}+f_{10}+f_{01}} = \frac{2}{2+6+5} = 0.153846$$

Q6 - B

$$p\left(\hat{x}_2 = 0 \mid y = 2\right) = p\left(\hat{x}_2 = 0, \hat{x}_7 = 0 \mid y = 2\right) + p\left(\hat{x}_2 = 0, \hat{x}_7 = 1 \mid y = 2\right) = 0.84$$

2 Description of the Dataset

As data becomes readily available, the predominance of machine learning methods has arisen. In order to design and implement a good quality supervised machine learning methods, we first select a dataset.

2.1 Our dataset

We performed out analysis on the "Replication Data for: South African Heart Disease" [1] dataset. This dataset contains only a portion of a larger dataset and only contains data about males. The high incidence of ischaemic heart disease (IHD) among White South Africans has been extensively documented by Wyndham, using death certification data. The dataset contains the following attributes: CLASS, Sbp, Tobacco, Ldl, Adiposity, Famhist, Type A, Obesity, Alcohol, and Age. The detailed description of the data can be found in the next chapter. The dataset contains 462 samples. Our overall problem is to analyze 10 attributes listed above and determine if there's a direct correlation of those attributes and whether a patient has heart disease.

2.2 Previous analysis of the data

In the first report regarding the original dataset published in 1983 (in SA Medical Journal), health professionals believed that some groups of people in the white population are especially prone to the disease. This has been difficult to confirm because of the lack of language and cultural information on death certificates. They aimed to establish the prevalence and intensity of risk factors in an Afrikaner community and to evaluate the effectiveness of intervention towards lowering risk factors. The report is part of CORiS (Corona Risk Factor Study) which aimed to quantify IHD risk factors in an Afrikan speaking community. Therefore, data analysis performed on the data was in regards to summary statistics and correlations between attributes.
The article discusses the study population, the methods, the interpretation of data and the results in detail. In this chapter, we summarize these chapters briefly.

2.2.1 The study population

The data gathered was from three towns which are situated in the southwestern Cape Province and are similar in cultural and socioeconomic structure. In the original data, male and female date were collected. The study population was aged between 15-64 years.

2.2.2 Methods

Respondents had to fill out a questionnaire, which contained covered socioeconomic items, smoking habits, a family history of IHD, personal medical history and activity patterns, as well as items on risk factor knowledge, attitudes and actions. Furthermore, the participants

participated in interviews as well as medical examinations (e.g: height, blood pressure, blood samples).

2.2.3 Interpretation of data

In this chapter, some thresholds are compared against the values recommended by WHO, furthermore about data categorization.

2.2.4 Results

The results were summarized in tables and refers to groups/categories. Different conclusions and statements can be drawn from these results. For example, *"mean systolic and diastolic blood pressures rose with age, females eventually having higher pressures than males"* or *"cigarette smoking was more than twice as common among males than among females at all ages, and the mean daily cigarette consumption was also higher"*. We could also read in the article that for different categories which are the major risk factors.

2.3 Classification and Regression on the dataset

During the project we need to apply classification and regression on the data. For the classification exercise - this will be the main machine learning aim - the CLASS attribute will be predicted from all 9 other attributes. In terms of classification, we would like to predict from the other 9 attributes if a person likely has heart disease or not (discrete values).
After that, we would like to predict the systolic blood pressure (sbp) attribute in mmHg from the obesity, alcohol, and tobacco attributes (continuous values).
For information on feasibility of this task and pre-processing/data transformation, see section 4 "Data Visualization".

3 Detailed Explanation of Attribute Data

The observations in this study consist of males from the Western Cape in South Africa. The data is a subset of a larger dataset from 1983 [2] and report several medical factors as well as the presence or absence of heart disease for each test subject. It consists of 462 observations and 10 attributes. The attributes contain different lifestyle-induced risk factors and clinical predictors such as alcohol consumption and systolic blood pressure respectively. A summary of the attributes can be found in Table 1.

Table 1: Description of each attribute, and the class of focus (presence of heart disease). * = unit not stated in the source, therefore a possible unit has been suggested after assessment of the values in the dataset. See argumentation under subsection "Continuous attributes".

Attribute	Continuous /Discrete	Type of attribute	Description	Unit
CLASS	Discrete	Nominal	Diagnosis of heart disease	Present or absent
Sbp	Continuous	Interval	Systolic blood pressure	mm Hg
Tobacco	Continuous	Ratio	Lifetime cumulative tobacco	kg
Ldl	Continuous	Ratio	Low density lipoprotein cholesterol	mmol/L
Adiposity	Continuous	Interval	Reflects degree of obesity	BAI, %*
Famhist	Discrete	Nominal	Family history of heart disease	Present or absent
Typea	Discrete	Interval	Type-A behavior	Bortner scale
Obesity	Continuous	Interval	Body mass index	BMI
Alcohol	Continuous	Ratio	Current alcohol consumption	L/year*
Age	Discrete	Ratio	Age	years

Continuous attributes:
The following attributes were categorized as continuous; sbp, tobacco, ldl, adiposity, obesity and alcohol. Amongst these tobacco and ldl are ratio, as zero indicates no tobacco use or no presence of the low density lipoprotein in the test person's serum.
Obesity and sbp are both interval, as these are measurements of the body mass index and millimetre of mercury. BMI relates to a relationship between height and weight and sbp to excess pressure in relation to mercury. As both of these attributes do not have a "true zero" they are interval.

The evaluation of what type of attributes adiposity and alcohol correspond to has been done after assessing their possible units, as the units were not stated in the source. For adiposity the unit could be the body adiposity index, as the values in the data seem to be in the same range as the index. The index reflects the amount of body fat in a human. This would make the attribute interval, which is the same for BMI. The alcohol attribute's unit seems likely to be a measure of L/year. When comparing the average alcohol consumption of males in South Africa from 2022 to the average of this study group they seem to be in the same range. An average of 17 L/year (Table 2) in this dataset and an average of 15.74 L/year in the source from 2022 [3]. Therefore the alcohol attribute would be ratio, as zero would reflect no consumption similar to the tobacco attribute.

Discrete attributes:
The rest of the attributes, CLASS, typea, famhist and age are discrete. Here the diagnosis with heart disease and family history is given as binary attributes, which in this case makes them nominal. Type-A behavior is interval as the Bortner scale is an overall measurement of several factors which results in no "true zero". Lastly age is ratio as zero relates to not having been born yet.

The dataset contains no missing values, or peculiar values of the attributes for each test subject. Attributes containing values of 0, which might be suspected as missing values, are all ratio such as tobacco indicating no use.

Table 2: Summary statistics, of the 8 attributes (sbp, tobacco, ldl, adiposity, typea, obesity, alcohol, age).

Attribute	**Mean** μ	**Standard deviation** σ	$\mathbf{q}_{25}$	$\mathbf{q}_{50}$	$\mathbf{q}_{75}$	**Range**
Sbp	138.3	20.5	124	134	148	117
Tobacco	3.6	4.6	0.05	2	5.5	31.2
Ldl	4.7	2.1	3.3	4.3	5.8	14.4
Adiposity	25.4	7.8	19.8	26.1	31.2	35.8
Typea	53.1	9.8	47	53	60	65
Obesity	26.0	4.2	23	25.8	28.5	31.9
Alcohol	17.0	24.5	0.5	7.5	23.9	147.2
Age	42.8	14.6	31	45	55	49

Summary statistics of the attributes, sbp, tobacco, ldl, adiposity, typea, obesity, alcohol, and age, can be found in Table 2. Family history and presence of heart disease is omitted from the summary statistics as they are binary categories. Table 2 reveals that that the average BMI (obesity) of the test group classifies as "Overweight" [4], the average systolic blood pressure (sbp) classifies as "Hypertension Stage 1" [5] and the average level of low density lipoprotein cholesterol classifies as "High" [6]. These high values of possible risk factors indicate that the test group represent a subspace. A more diverse test group, would be optimal when determining the correlation between risk factors and the presence of heart disease.

4 Data Visualization

Pre-processing:
We normalized all attribute values which included subtracting the mean and dividing by the standard deviation for easier and more reliable graphical comparison. Since all values for the provided attributes were available and valid, no transformations were necessary to account for missing data. Class values for 'class'(no heart disease vs. heart disease) and 'famhist' (family history) were already encoded as binary values, thus, one-out-of-k encoding was unnecessary.

Data Visualization:
Figure 1 depicts box and whisker plots for each attribute which gives a summary of the following five statistics: minimum, first quartile, median, third quartile, and maximum. The points beyond on the whiskers typically depict outlier values. There are outlier values in the following attributes: sbp, tobacco, ldl, type a, obesity, and alcohol (as indicated by values that are more than 3 standard deviations away from the mean). There are, however, no issues with outliers as those in the indicated attributes are not a result of error but due to natural variation. For instance, tobacco, alcohol, and obesity attributes may greatly

vary to an individual's lifestyle and age. Therefore, outliers were included in data visualization and principal component analysis.

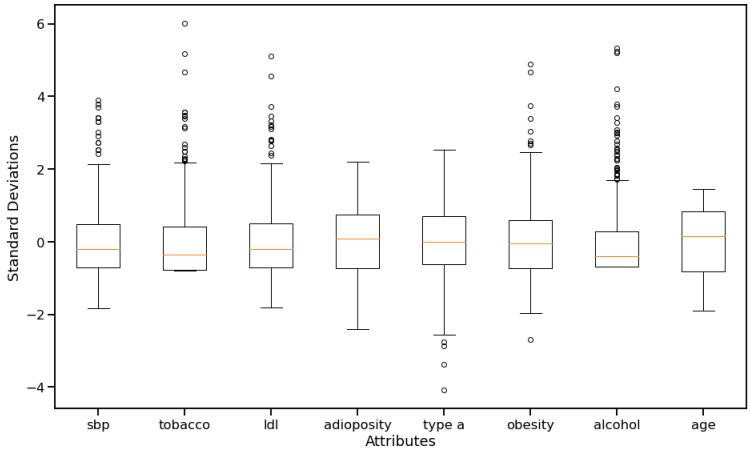

Figure 1: The figure depicts box and whisker plots for each given attribute. Plots were removed for class attributes 'class' and 'famhist' as box plots of those attributes did not provide useful information.

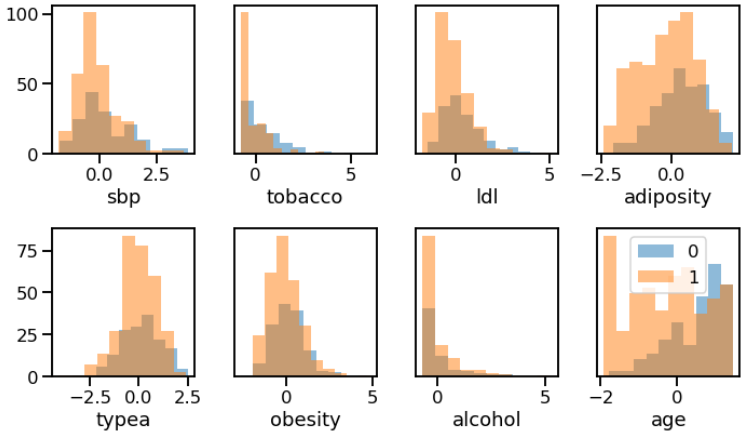

Figure 2: The figure depicts histograms for each given attribute. 0 (blue) and 1 (orange) correspond to patients without and with heart disease, respectively. Plots were removed for class attributes 'class' and 'famhist' as box plots of those attributes did not provide useful information.

Figure 2 depicts visualization of each single attribute through histogram plots. Histogram plots for attributes sbp, ldl, adiposity, type a, obesity, and age are roughly symmetric and bell-shaped for both no heart disease and heart disease patients, indicating these attributes are normally distributed. Histograms for tobacco and alcohol indicate data that is heavily skewed right but also normally distributed for both no heart disease and heart disease patients. Age for no heart disease patients is also skewed left and normally distributed.

These attributes values vary dramatically since alcohol and tobacco are measured by lifetime consumption which varies by age and an individual's lifestyle. From figure 2, the distributions between no heart disease and heart disease patients heavily overlap, indicating that a single attribute cannot determine a clear separation of data.

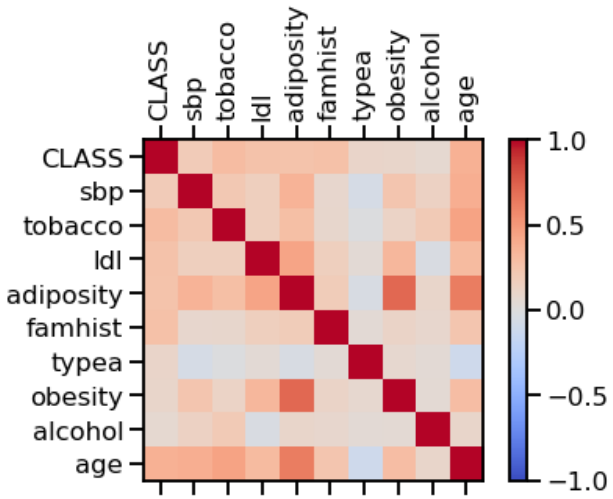

Figure 3: The figure depicts the correlation matrix of attributes.

Figure 3 shows correlation coefficients between any two attributes. Obesity and adioposity have a strong correlation with coefficient of 0.717, the highest correlation coefficient shown in the graph. Age and adioposity also show a moderate correlation with coefficient of 0.626. Correlation between any two other attributes are generally weaker, indicating attributes are fairly independent from one another.

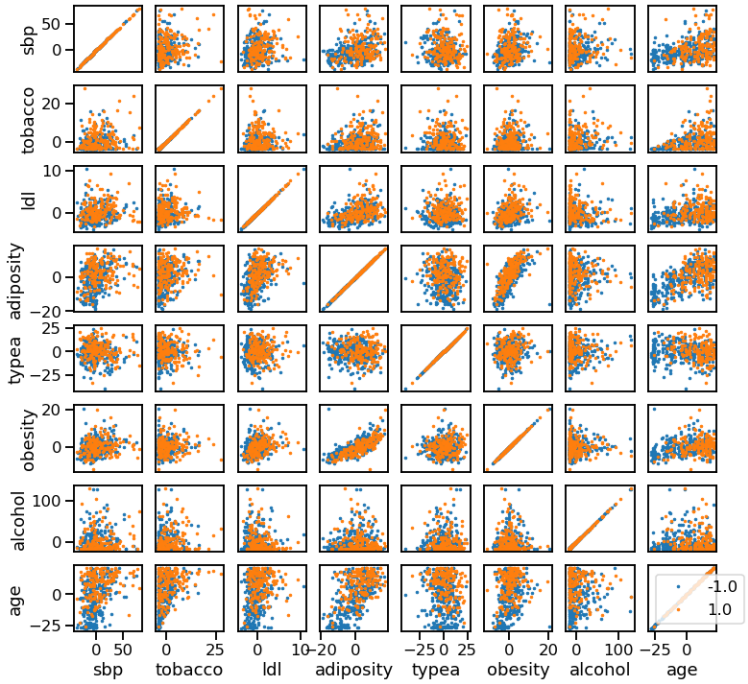

Figure 4: The figure depicts the matrix plot of selected attributes.

Figure 4 suggests difficulty in creating a non-linear separation between classes (non-heart disease and heart disease patients) as the data for both classes are fairly mixed in all plots.

Primary Machine Learning Aim:
The primary machine learning modeling aim seems not quite feasible for the classification method since the data depicted in figures 2 and 4 show overlapping data between classes, indicating it may be potentially difficult to classify a given set of attributes to the correct class. Regression regarding predicting systolic blood pressure seems more feasible as there is some at moderate linear correlation between systolic blood pressure and other attributes.

Principal Component Analysis:
The principal component analysis shows that the first three principal components account for more than 90% of the variance in the data. Thus, this dataset is a good candidate for PCA since we are able to discover a much lower-dimensional representation of the high-dimensional data set.

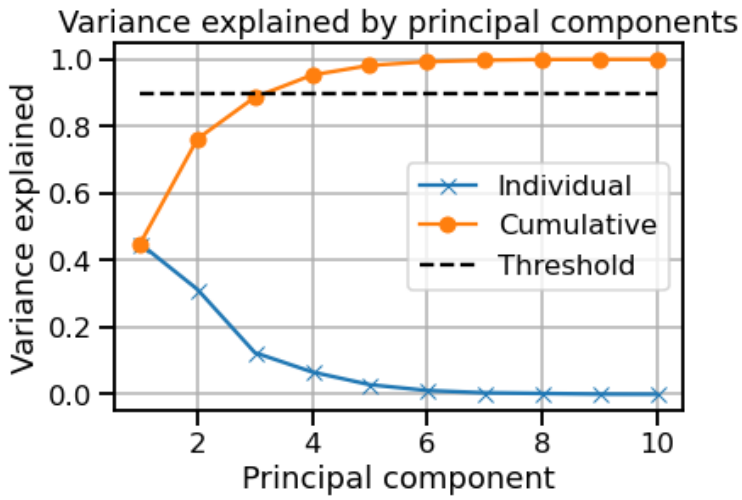

Figure 5: The figure depicts the proportion of variance for increasing number of components. The dotted line shows the cut off value of 90%.

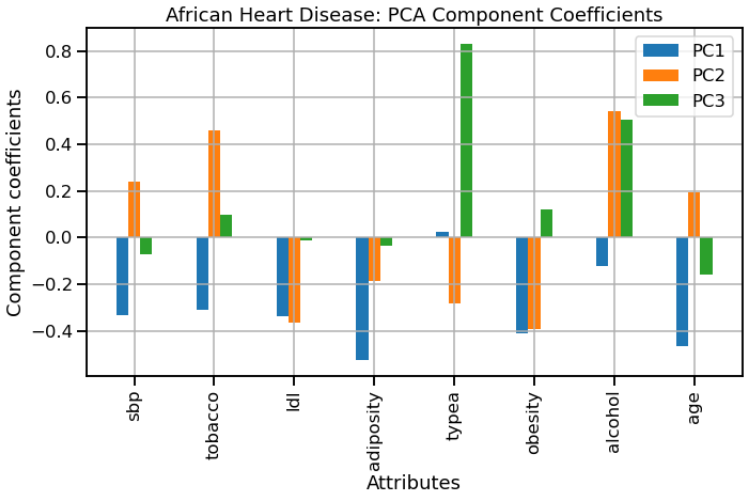

Figure 6: The figure depicts principal component analysis coefficient for the first three components. Three components are used as three components explains roughly 90% of the variance. Class attributes are omitted.

Figure 6 shows the component coefficients for the first three principal components for each attribute. For instance, the third principal component shows very high coefficients with attributes type a and alcohol and thus oriented towards those attributes.

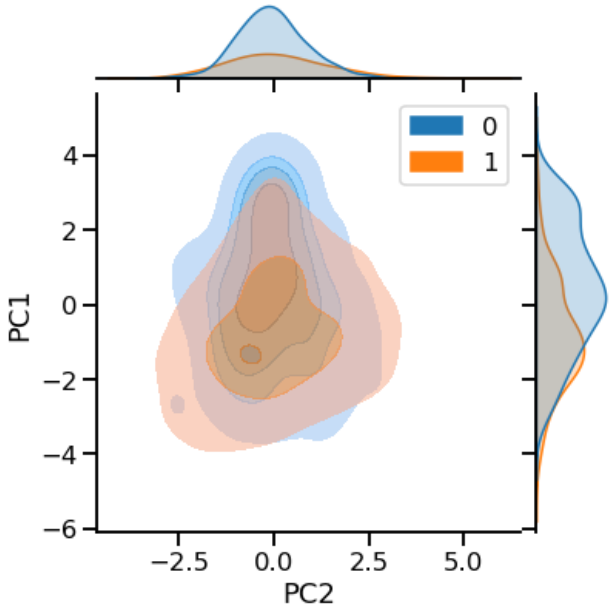

Figure 7: The figure depicts the density of each class (no heart disease vs. heart disease) projected onto the first two principal components.

Figure 7 depicts that there be potential complication with separation of classes which is consistent with previous figures. We can see that the class projections overlap with one another.

5 Discussion and conclusion

Assessing the summary statistics of the South African Heart Disease study suggested that the test group represents a subspace with respect to several attributes. This might interfere with the classification problem as the test group might not have an optimal variation. Further, we performed data visualization and principal component analysis on our dataset. Data visualization showed no issues with outliers/missing data, normally distributed attributes, as well as some correlated but mostly low correlation between attributes. Principal component analysis provided sufficient dimension reductionality as we can use 3 principal components to explain 90% of the variance in data. Projection of data onto PCA components showed overlap, indicating an unclear separation of data. As a result, we will pursue a classification method to determine whether there's a correlation between all the other attributes and the class attribute (heart disease vs. no heart disease). We will also evaluate the data using regression to predict the systolic blood pressure attribute from the obesity, alcohol, and tobacco usage attributes. From our evaluation of PCA and data visualization, our primary machine learning aim for classification may prove more difficult and less feasible, however, it is reasonable to perform regression on the dataset.

References

[1] "Replication Data for: South African Heart Disease." https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/76SIQD. Accessed: 2022-09-15.

[2] J. Rossouw, "Coronary risk factor screening in three rural communities," *South African medical journal*, 1983.

[3] "alcohol consumption by country 2022." =https://worldpopulationreview.com/country-rankings/alcohol-consumption-by-country. Accessed: 2022-10-02.

[4] F. Q. Nuttall, "Body mass index," *Nutrition Today*, vol. 50, no. 3, p. 117â128, 2015.

[5] T. F. LĂźscher, "What is a normal blood pressure?," *European Heart Journal*, vol. 39, no. 24, p. 2233â2240, 2018.

[6] "Ldl cholesterol." https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=ldl_cholesterol. Accessed: 2022-09-30.