Project Report – Project 1

Authors

Jørgen Finsveen - s231469
Even J. P. Haslerud - s231473

Contribution

Report

Introduction	Jørgen Finsveen	
Dataset Description	Jørgen Finsveen	Even J. P. Haslerud
Data Attributes	Jørgen Finsveen	Even J. P. Haslerud
Data Visualization	Jørgen Finsveen	Even J. P. Haslerud
Discussion		Even J. P. Haslerud

Solutions

Question 1		Even J. P. Haslerud
Question 2	Jørgen Finsveen	
Question 3	Jørgen Finsveen	
Question 4		Even J. P. Haslerud
Question 5		Even J. P. Haslerud
Question 6	Jørgen Finsveen	

Code

Plotting and Analyzing	Jørgen Finsveen	Even J. P. Haslerud

Table of Contents

INTRODUCTION 3
DATASET DESCRIPTION 3
DATA ATTRIBUTES 3
DATA VISUALIZATION 4
Outlier Analysis 5
Distribution of Attributes 5
Correlation Analysis 6
PCA 6
DISCUSSION 7
QUESTIONS 7
Question 1 7
Question 2 7
Question 3 9
Question 4 9
Question 5 9
Question 6 9
REFERENCES 10

Introduction

This report is written by two employees of an imaginary corporation. The employees has received a dataset, and this report is intended to give a useful description of the dataset with some basic plots from the data, which the co-workers of the corporation may find useful.

Dataset Description

The dataset in question in this project is named *Heart Disease*[1] and was donated to the University of California Irvine Machine Learning Repository in 1988. UCI describes the data as a set of four datasets from Cleveland, Hungary, Switzerland, and VA Long Beach. It consists of 14 attributes which may have some correlation to the risk of developing heart diseases. UCI has categorized the dataset to be suitable for classification and regression. The dataset was created by Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano[2].

The dataset has been studied before by several persons. One of these studies was conducted by Rob Harrand and published on Kaggle[3]. Harrand uses the dataset and attempts to create a model which is able to classify a person as having a heart disease or not by using a random forest model. He concluded his report by stating that the dominating factors came as a surprise for him, as he expected attributes such as cholesterol and age to be dominating but observed that i.e. ECG was a dominating attribute.

Upon performing our own analysis on the dataset, it would be appropriate to apply regression and classification. This report will focus on how this should be done rather than an actual implementation. The main motivation of applying such techniques is to learn how the different attributes in the dataset correlate to each other. The heart disease data is of special interest, and this attribute will be compared to the others in order to discover how it changes based on changes in other attributes such as age, maximum heartrate, sex, resting blood pressure, and ECG. When it comes to classification, it would be an interesting task to try classifying heart disease based on the above attributes. For the sake of simplicity, this project will focus on the Cleveland dataset.

Data Attributes

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them, which are:

- Age: Represent the age of the person and is a integer type. This variable is a ratio and discrete.
- Sex: Represent the sex of the person with a categorical type, where 1 is male and 0 is female. This variable is nominal and discrete.
- Cp: Represent the chest pain type, which consist of 4 values:
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic

  This variable is nominal and discrete.
- Trestbps: Represent the resting blood pressure (in mm Hg on admission to the hospital). This variable is ratio and continuous.
- Chol: Represent serum cholesterol measured in mg/dl. This variable is ratio and continuous.
- Fbs: Represent the person's fasting blood sugar (> 120 mg/dl, 1 = true, 0 = false). This variable is nominal and discrete.
- Restecg: Representing electrocardiographic measurement, where we have 3 different values:
  - 0 = normal
  - 1 = having ST-T wave abnormality
  - 2 = showing probable or definite left ventricular hypertrophy by Este's criteria.

  This variable is ordinal and discrete.

[1] University of California Irvine Machine Learning Repository (2018)

[2] University of California Irvine Machine Learning Repository (2018)

[3] Harrand (2018)

- Thalach: Represents the person's maximum heart rate achieved. This variable is ratio and discrete.
- Exang: Represents exercise induced angina, where we have 1 equal to yes and 0 equal to no. This variable is nominal and discrete.
- Old peak: Represent the ST depression induced by exercise to rest, and this variable is ratio and continuous.
- Slope: Representing the scope of the peak exercise ST segment with three values:
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: down sloping

  This variable is nominal and discrete.
- Ca: Represent the number of major vessels (0-3), where the variable is ordinal and discrete.
- Thal: Represent a blood disorder called thalassemia, with three values:
    - Value 3: normal
    - Value 6: fixed defect
    - Value 7: reversable defect.

  This variable is ordinal and discrete.
- Num: Represent a heart disease where 0 equals to no and 1 equal to yes. The variable is nominal and discrete.

The dataset was taken from the UCI Machine Learning Repository, where it says that it has 6 missing values of 303 instances. However, the dataset contains 14 attributes that describe what they stand for and if they are discrete/continuous, nominal/ordinal/interval/ratio. Under it shows and plot of the data table with the given attributes:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0
1	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0
2	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0
3	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0
4	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0

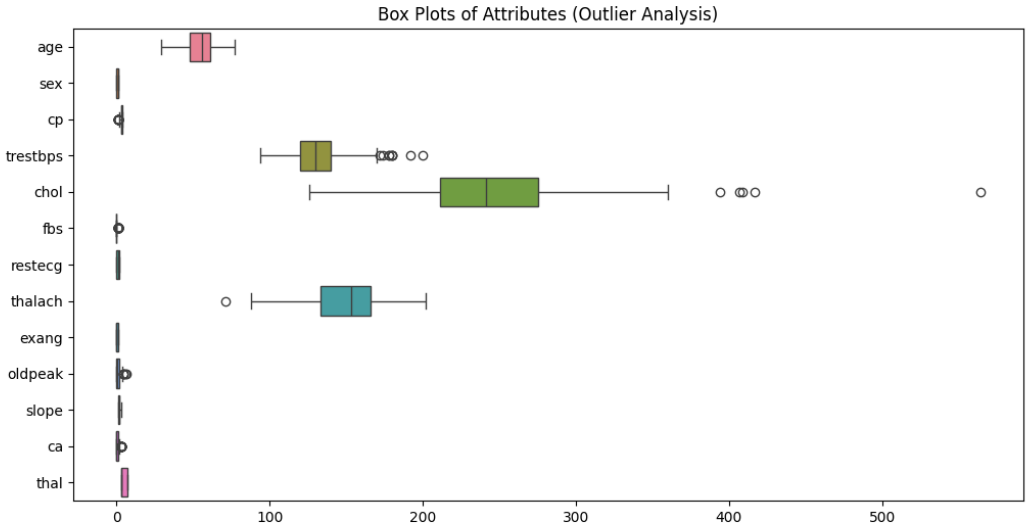
(Figure 1)

Data Visualization

In this section, we have done visualization and analysis on the heart disease dataset with the given attributes mentioned above. The analysis has been conducted through effective visualization techniques and Principal Component Analysis (PCA). Throughout the whole process of analyzing the dataset, we have kept in mind the ACCENT principles and Tufte's guidelines in order to ensure that our visualization is accurate, clear, and effective.

Outlier Analysis

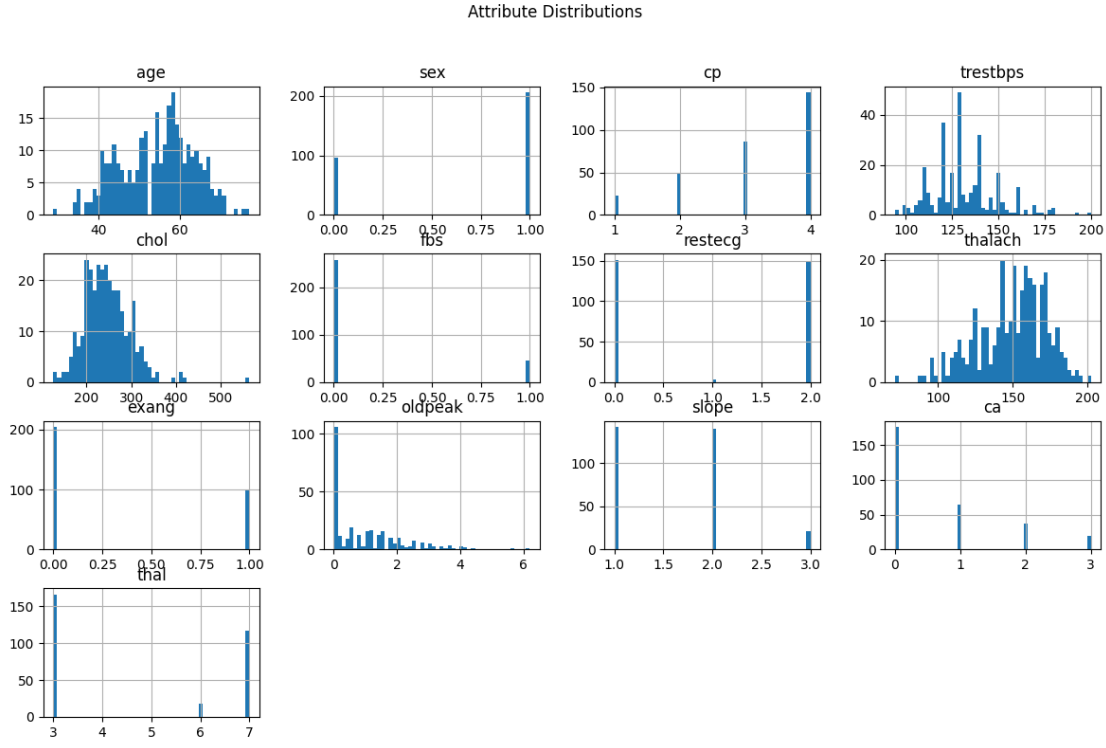
Since outliers can significantly impact analysis, we have used box plots to identify potential outliers. Outliers are data points that fall outside the “whiskers” of the box plots, and as we can see below at figure 2, chol, trestbps, has the biggest gaps when it comes to outliers. Addressing and observing these outliers is essential for getting an accurate PCA result and for other machine learning tasks as well. One important observation we can notice here is that one of the values at chol is over 500, which is very high.



(Figure 2)

Distribution of Attributes

The distribution of the dataset is crucial in data analysis. We will therefore implement attribute histograms to visualize the distribution of each attribute in the dataset. The histograms displays the frequency of values within specific intervals, allowing us to observe whether attributes follow a normal distribution or other patterns. This helps us understand the characteristics of the data. As shown in figure 3 below:

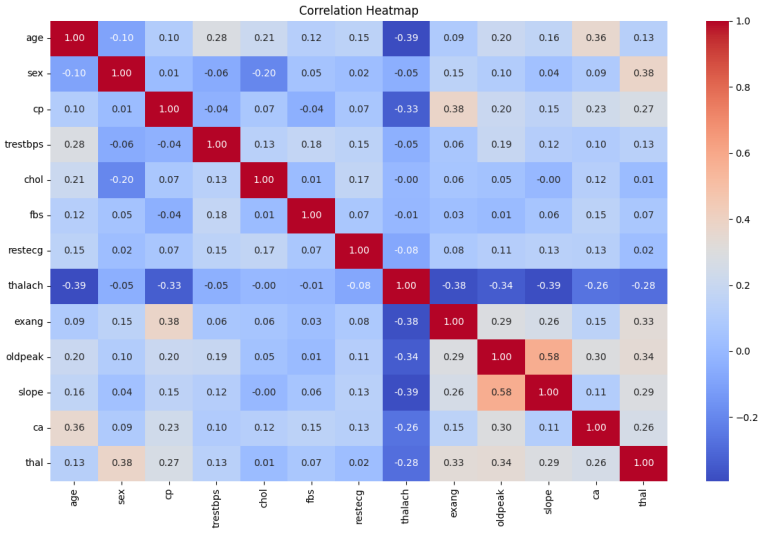


(Figure 3)

Correlation Analysis

In order to unveil correlations between different attributes in the dataset, we utilized a correlation heatmap. The heatmap in figure 4 provides a clear representation of pairwise correlations, strong positive correlations (red), negative correlations (blue), or weak correlation (light colors).

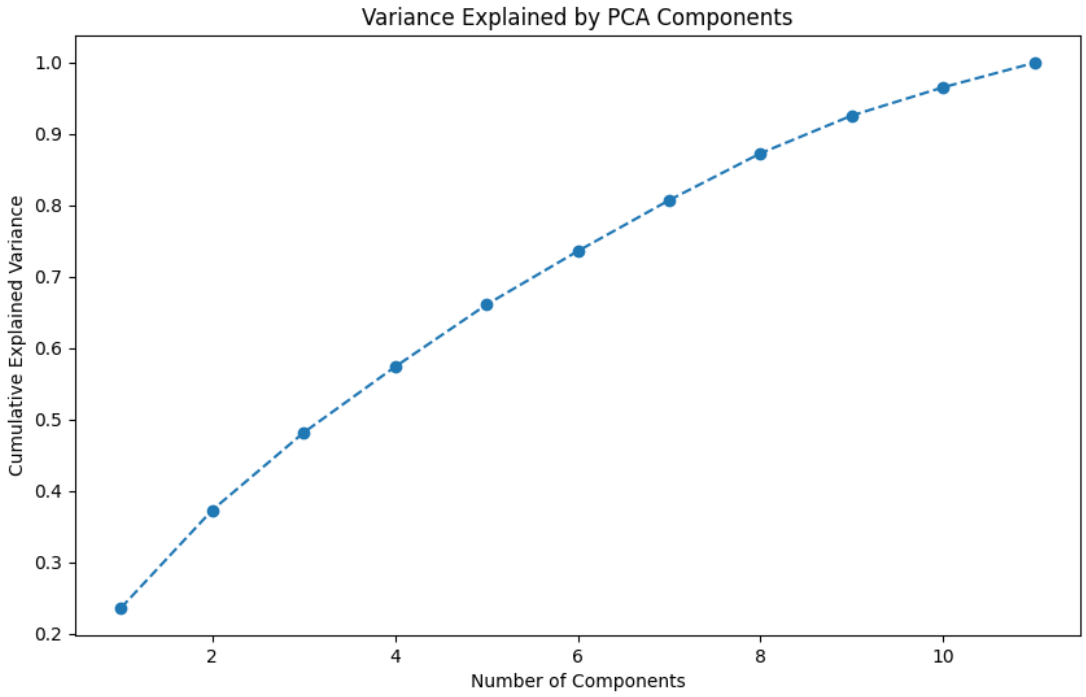
The heatmap allows us to observe that different attributes correlates with each other. Some attributes has a strong positive degree of correlation such as slope and oldpeak, and others has a strong negative correlation such as thalac and age. There are some attributes for which the median value of the variable remains independent from other attributes, i.e. restecg.



(Figure 4)

PCA

From Figure 5 below, it's clear to say that the first seven principal components covers almost 80% of the data, and more than 80% of the variation is covered by the first seven principal components. This observation tells us that the data can be projected on to lower components and still keep the majority of the variance.



(Figure 5)

Discussion

We have learned that visualizing data through different plots and diagrams, it is easier to recognize patterns and correlations in the dataset. By looking at the generated diagrams, we can see some correlations and relationships between different attributes, which indicates that the data could be used to create a model. The histograms shows that there is variations in some attributes in the dataset, while some does not possess a wide distribution, By looking at the correlation heatmap, we can see how different attributes has a high degree of correlation such as slope and fixed oldpeak. As shown in figure 5, the variance in the data is about 90% intact with around eight components.

Through different visualization methods, a dataset which seems meaningless to the naked eye became clear. This shows the effectiveness of such data analysis and how it can contribute to building models which generate correct, clear, and precise outputs in an effective manner. It has also shown that the attributes of the data has relations with each other, and that the data has a degree of variety. This makes it suitable for training of a machine learning model, and we are confident that we can proceed with our intended strategies for this project.

Questions

For this project, there were six questions provided as previous exam projects. Below are the attempted solutions for each of these questions by the authors.

Question 1

The problem is solved by simply thinking about what the attributes represent and comparing to the definition in the different types. Recall that:

- Nominal is a type that only allow comparison (equal or different)
- Ordinal allows ordering (but not differences)
- Interval allows differences but no (physically well-defined) zero.
- Ratio is a type with a zero with a well-defined meaning.

With these definitions, we see that:

- x_1 (Time of the day) is an interval.
- x_2 (Broken Truck) is a ratio.
- x_3 (Accident victim) is a ratio.
- x_4 (Immobilized bus) is a ratio.
- x_5 (Defects) is a ratio.
- x_6 (Traffic lights) is a ratio.
- x_7 (Running over) is a ratio.
- y (Congestion level) is ordinal.

Therefore, option D is correct.

Question 2

We define the sets *U* and *V* as the 14th and 18th observations of the Urban Traffic dataset:

$$U = x_{14} = \begin{bmatrix} 26 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad V = x_{18} = \begin{bmatrix} 19 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

According to the lecture notes[4] for Introduction to Machine Learning and Data Mining *p*-norm distance can be defined as:

$$d_p(x,y) = \|x-y\|_p = \begin{cases} (\sum_{i=1}^{M}|x_i - y_i|^p)^{\frac{1}{p}} & if\ 1 \leq p < \infty \\ max\{|x_1 - y_1|, |x_2 - y_2|, \dots, |x_M - y_M|\} & if\ p = \infty \end{cases}$$

By utilizing this definition, we can evaluate the following statements about the *p*-norm distance $d_p(.,.)$:

A. Statement: $d_{p=\infty}(U,V) = 7.0$

The norm $p = \infty$ the distances measures the largest difference between two corresponding elements in each observation. In this case, it will be $|U_1 - V_1| = |26 - 19| = 7.0$. As this is the biggest distance between corresponding attributes in the observations, this statement is considered to be **correct**.

B. Statement: $d_{p=3}(U,V) = 3.688$

This statement can be verified by performing the calculation by hand, or by writing a script which does the calculation for us. Both approaches will be demonstrated.

Manual calculation:

$$d_{p=3}(U,V) = (|26-19|^3 + |2-0|^3)^{\frac{1}{3}} = \sqrt[3]{7^3 + 2^3} = \sqrt[3]{351} \approx 7.054$$

By performing the calculation, we end up with an answer which indicates that the statement we are verifying is **incorrect**.

Computation by scripting:

We write the following script in Python to solve the task for us:

```
U = [26, 0, 2, 0, 0, 0, 0]
V = [19, 0, 0, 0, 0, 0, 0]

def p_distance(p, x, y):
    s = 0
    for i in range(len(x)):
        s += (x[i] - y[i]) ** p
    return s ** (1/p)

print(p_distance(3, U, V))
```

By running this script, we will receive the answer 7.0540 which is the same answer as received when doing the manual calculation. The statement is therefore **incorrect**.

C. Statement: $d_{p=1}(U,V) = 1.286$

The *p*-distance is calculated by using the Python-script made for validating statement B, but with a different parameter in the function-call:

```
print(p_distance(1, U, V))
```

The resulting number from running this script is 9.0 which indicates that statement C is **incorrect**.

[4] Herlau, Schmidt, and Mørup. 2022, p. 58.

D. Statement: $d_{p=4}(U,V) = 4.311$

The *p*-distance is once again calculated by using the Python-script. The function-call parameters are adjusted to match the statement:

```
print(p_distance(4, U, V))
```

The resulting number from running this script is 7.0116 which indicates that statement D must be **incorrect**.

Question 3

The correct answer is A. To see this, recall the variance explained by a given component k of the PCA is given by:

$$\frac{\sigma_k^2}{\sum_{j=1}^{M}\sigma_j^2}$$

Where M is the number of attributes in the dataset being analyzed. The values of σ_k can be read off as entry $\sigma_k = S_{kk}$ where S is the diagonal matrix of the SVD computed. We therefore find the variance explained by components x_1, x_2, x_3, x_4, x_5 is:

$$Var.Expl. = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2} = 0.8667.$$

Question 4

The correct answer is C. Focusing on the correct answer, note the projection onto principal component v_1 is:

$$b_1 = x^T * v_1 = [\![x_1, x_2, x_3, x_4, x_5]\!] \begin{bmatrix} 0.49 \\ 0.58 \\ 0.56 \\ 0.31 \\ -0.06 \end{bmatrix}$$

It is now a simple matter of observing that for this number to be (relatively large) and negative, this occurs if x_1, x_2, x_3, x_4 has large magnitudes and the sign convention given in option C.

Question 5

The correct answer is A. To compute the Jaccard Similarity between $s_1\ and\ s_2$, we need to do some observations of the set:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|2|}{|13|} = 0.153846.$$

So, the Jaccard similarity between the documents $s_1 and\ s_2$ is option A.

Question 6

The problem is solved by a simple application of Bayes' theorem:

$$= \frac{p(y=2|x_2=0)}{p(x_2=0|y=1)p(y=2)} \Big/ \sum_{k=2}^{4} p(x_2=0|y=k)p(y=k)$$

The values of $p(y)$ are given in the problem text and the values of $p(x_2|y)$ in the Table 2. Inserting the values, we see option B is correct.

References

Harrand, Rob (2018). *What Causes Heart Disease? Explaining the Model*. https://www.kaggle.com/code/tentotheminus9/what-causes-heart-disease-explaining-the-model. (Downloaded: 02.10.2023)

Janosi, Andras, Steinbrunn, William, Pfisterer, Matthias, and Detrano, Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X. (Downloaded: 02.10.2023)

Herlau, Tue, Schmidt, Mikkel N. and Mørup, Morten. Introduction to Machine Learning and Data Mining. Lecture notes, Spring 2022, version 1.0. 4.3 Measures of distance, page 58.