02450 Introduction to Machine Learning and Data Mining

Report 1

Authors:
Andrew Ardueser (s182332)
Alexandros Dorodoulis (s182148)

March 5, 2019

Contents

1 A description of the data set
Contributing Author(s): Andrew Ardueser, Alexandros Dorodoulis

1.1 Problem of interest

Our main problem of interest is assessing which features in the dataset are most significant in the classification in regards to the criterion variable (the "target" variable in this dataset). In the case of our dataset, this goal will predict if a patient has heart disease (1=Present; 0=Absent for the target variable) based on several health metrics. Though both supervised and unsupervised learning can provide useful insights into the data, we will focus on supervised learning's ability to give us a target output. Regression and classification will both be useful to predict if a subject is likely to have heart disease based on the corresponding feature variable values. Machine learning methods such as logistic regression and linear discriminant analysis are useful tools in this situation, but this report will be restricted to summarizing the data and principal component analysis.

1.2 Data origin

The data was obtained through the platform Kaggle. Kaggle is an online platform that aims to be the home of data scientists. It has a large pool of well-maintained datasets, and it provides a lot of good examples of applied machine learning. The data set that was chosen is called "Heart Disease UCI" and can be obtained through the following link: here.

1.3 Previous data transformations

The dataset originally contained 76 features, but we have decided to use the commonly-used subset containing 14 of them. The most important fields removed from the original dataset contained patient personal information, including names and social security numbers. These were replaced with dummy variables before being released but were ultimately removed before we began our analysis.

1.4 Primary machine learning modeling aim

The primary machine learning modeling aim for this data set would likely be either regular classification or a regression via logistic regression. From a common sense viewpoint, several features in the data set would be useful in classifying or predicting the presence of heart disease in a patient. However, the most relevant features would likely be age, resting blood pressure, cholesterol, maximum heart rate achieved, the number of major blood vessels colored by flouroscopy, and the presence of heart defects. Each of these appear to be directly related to a heart's ability to healthily function. For example, an older person with high resting blood pressure and cholesterol levels would seem more likely to have/develop some form of heart disease. While there is likely a correlation between some of the features (age and cholesterol, for example), a successful model would be able to account for the collinearity while still attaining information not provided by other variables.

During the assignments based on analysis, we hope to discover which features are actually significant and use them to predict/classify if a person has a heart disease based on their feature values. After splitting the data set into training and test data, the target variable will be used to calculate the accuracy of our model.

In order to properly weigh each variable in the model (due to their varying scales and units of measurement), we will standardize each quantitative variable by subtracting their means and dividing by their

standard deviations, respectively. This will create a unified scale among our variables, preventing bias based on feature value ranges.

2 A detailed explanation of the attributes of the data

Contributing Author(s): Andrew Ardueser

2.1 Variable description

The dataset includes the following variables:

"age": age of subject, continuous and ratio variable

"sex": sex of subject (1 = male; 0 = female), discrete and nominal variable

"cp": chest pain type (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic), discrete and nominal variable

"trestbps":resting blood pressure (in mm Hg on admission to the hospital), continuous and ratio variable

"chol": serum cholesterol in mg/dl, continuous and ratio variable

"fbs": fasting blood sugar >120 mg/dl (1 = True, 0 = False), discrete and nominal variable

"restecg": resting electrocardiogram (ECG) results (Value 0: normal, Value 1: having ST-T wave abnormality, Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria), discrete and nominal variable

"thalach": maximum heartrate achieved, continuous and ratio variable

"exang": exercise-induced angina (1 = yes, 0 = no), discrete and nominal variable

"oldpeak": ST depression induced by exercise relative to rest, continuous and ratio

"slope": the slope of the peak exercise ST segment (Value 1: slope>0, Value 2: slope=0, Value 3: slope<0), discrete and ordinal variable

"ca": number of major vessels (0-3) colored by fluoroscopy, discrete and ordinal variable

"thal": 3 = normal; 6 = fixed defect; 7 = reversable defect, discrete and nominal variable

"target": presence of heart disease (1 = True, 0 = False), discrete and nominal variable

Finally, there appear to be no missing or corrupted data from the dataset. Each feature has a number of values equal to the number of observations in the dataset, and none of them have values which would not fit appropriate range (e.g. age of -100, sex of 'bear' or 17, etc.).

2.2 Basic statistics

To give a basic visual description of the feature value distributions, we created histograms for each.

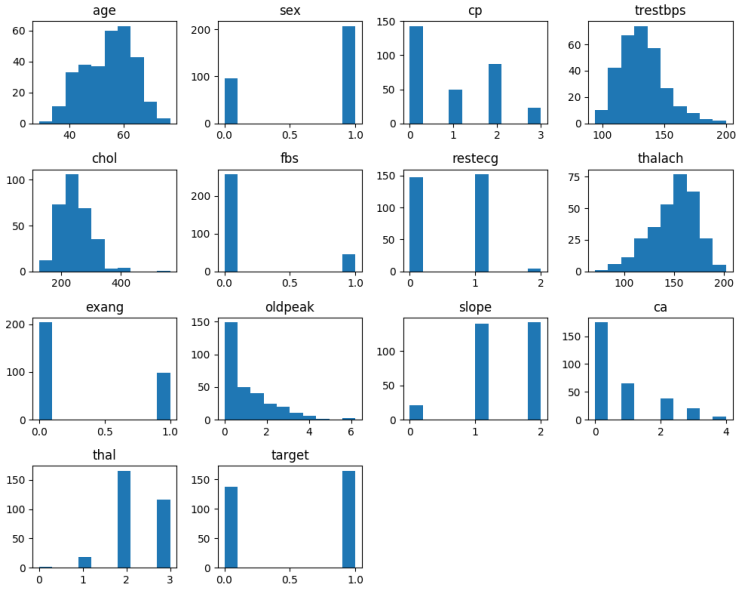

Figure 1: Feature Histogram

From the graphs, we see a large variety in most of the feature values. Other than the fasting blood sugar (fbs), there does not appear to be a single value dominating the others in representation.

To more easily analyze the features, we will split them into quantitative and qualitative groups. This will allow for a more balanced comparison. The qualitative feature group will include: sex, chest pain type (cp), resting ECG results, fasting blood sugar >120 mg/dl, exercise-induced angina (exang), ST segment slope, and heart defect presence (thal). The quantitative feature group will include: age, resting blood pressure, cholesterol, maximum heartrate achieved, ST depression peak, and number of colored major vessels. Below are summary statistics for the quantitative variables.

	Age	trestbps	chol	thalach	oldpeak	ca
Mean	54.366	131.623	246.264	149.646	1.039	0.729
Std Dev	9.082	17.538	51.830	22.905	1.161	1.022
Min	29	94	126	71	0	0
Median	55	130	240	143	0.8	0
Max	77	200	564	202	6.2	4

Table 1: Quantitative Feature Value Statistics

To further summarize the qualitative features, here are the distinct factor levels for each along with their corresponding frequencies.

sex		cp		fbs		exang		slope		thal		restecg	
0	96	0	143	0	258	0	204	0	21	0	2	0	147
1	207	1	50	1	45	1	99	1	140	1	18	1	152
		2	87					2	142	2	166	2	4
		3	23							3	117		

Table 2: Qualitative Feature Value Frequencies

Though correlation is an important part of dataset analysis, it will be covered in the PCA section.

3 Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA)
Contributing Author(s): Alexandros Dorodoulis

3.1 Outliers

Before we proceed further with the exploration of our dataset, we should identify any outliers. In order to achieve this, we leverage the following boxplot chart.

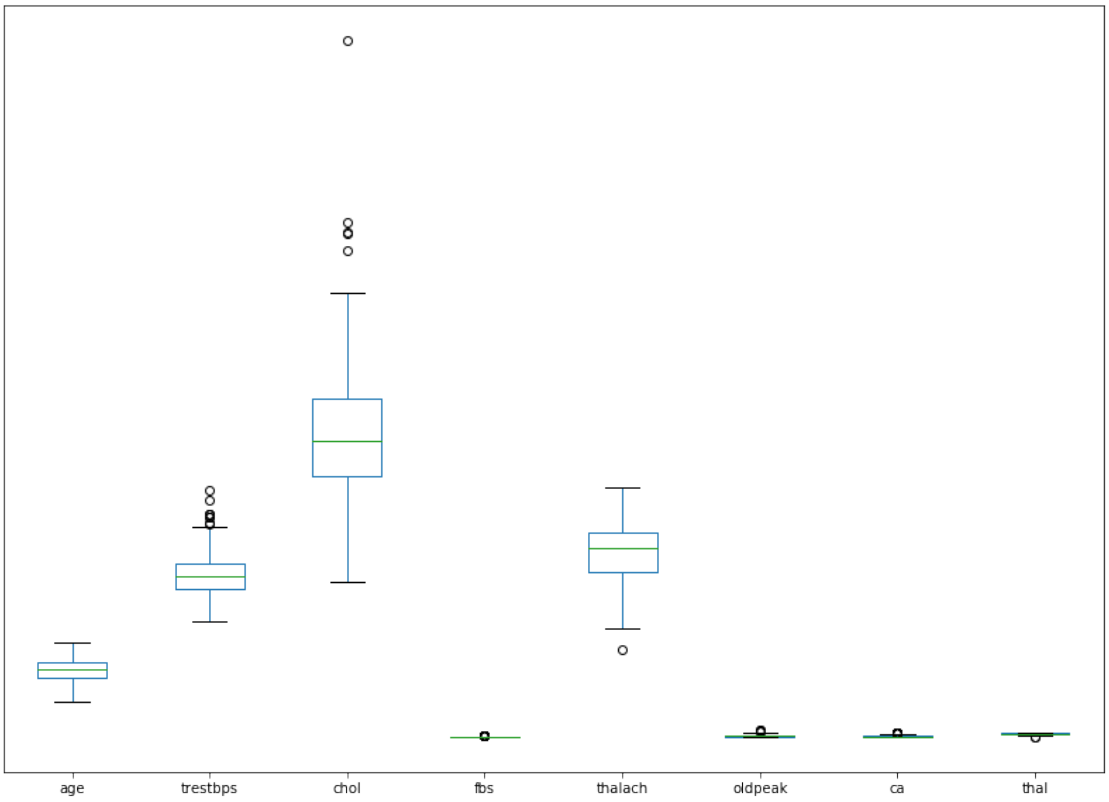

Figure 2: Boxplot of variables

There seem to be a few outliers for the trestbs, chol and thalac variables.

3.2 Normal distribution

For the most part, the variables do appear to be normally distributed. Outside the outliers in the chol and trestbps variables, there does not appear to be much skew. One exception is the thalach variable. The values appear to be skewed towards the higher end of the range. While the skew is not extreme, it is significant enough to be noted.

3.3 Correlation

As mentioned before, correlation between variables is an important aspect to consider when analyzing a data set. We created a heatmap matrix to display the correlation coefficients between each pair of features and the criterion variable ”target”. A higher correlation coefficient will have a more opaque color (whether

red for positive or blue for negative correlation). Obviously, we are able to ignore the diagonal of the matrix. These entries show that each feature is perfectly correlated to itself.

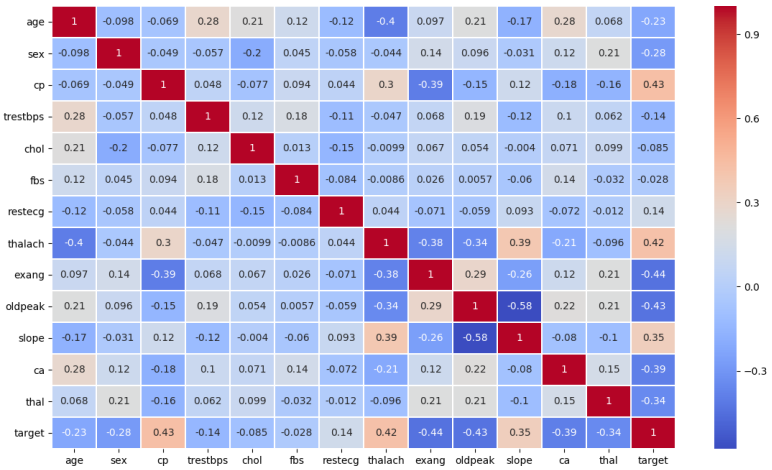

Figure 3: Correlation Matrix

We see there are a few entries that have relatively high levels of correlation. These pairs include slope and oldpeak, age and thalach, cp and exang, thalach and slope, and thalach and exang. Additionally, the target appears to have a slight correlation with cp, thalach, exang, oldpeak, and ca. Though the correlation coefficients are somewhat low, we would look for significant interactions as well when constructing our model.

Additionally, we created a bar graph to more easily show the relationship between each feature and the target variable. As seen below, there are several features with relatively high correlation while others appear to have nearly no correlation.

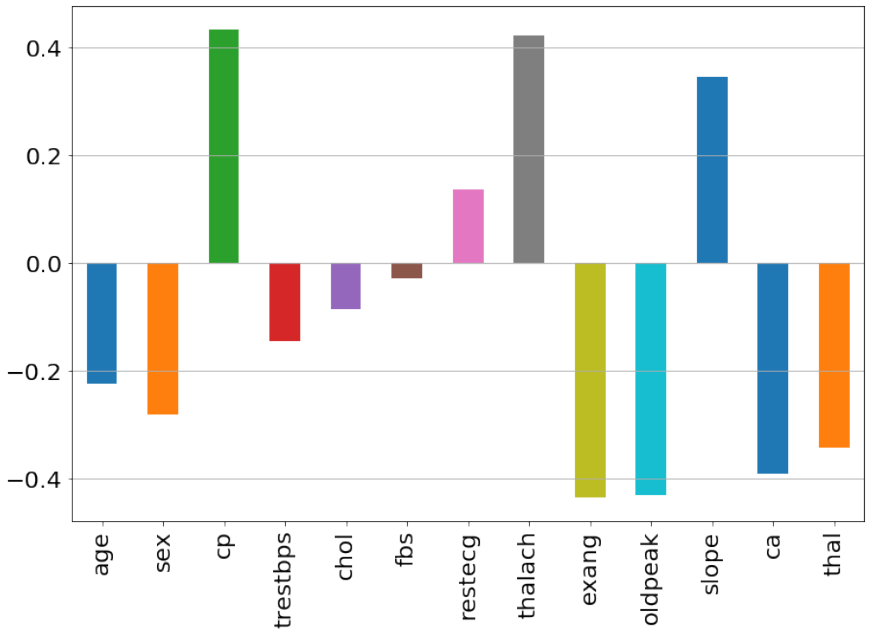

Figure 4: Correlation with Target

3.4 Primary machine learning modeling aim

Classification does appear to be viable based on our visualizations. Many of our features appear to have a significant level of correlation with the target, and the dataset's relatively normal distribution in its features will facilitate easier utilization of models which require the assumption of normally distributed data. Furthermore, there is a good variety of feature types. The dataset consists of multiple discrete and continuous, quantitative and qualitative features.

3.5 Principal component analysis (PCA)

Before we can apply a principal component analysis, we have to normalize our data and calculate which is the ideal number of variables for the analysis. In order to achieve that, we calculate the explained variance ratio of each additional variable, and we can visual it in the following plot:

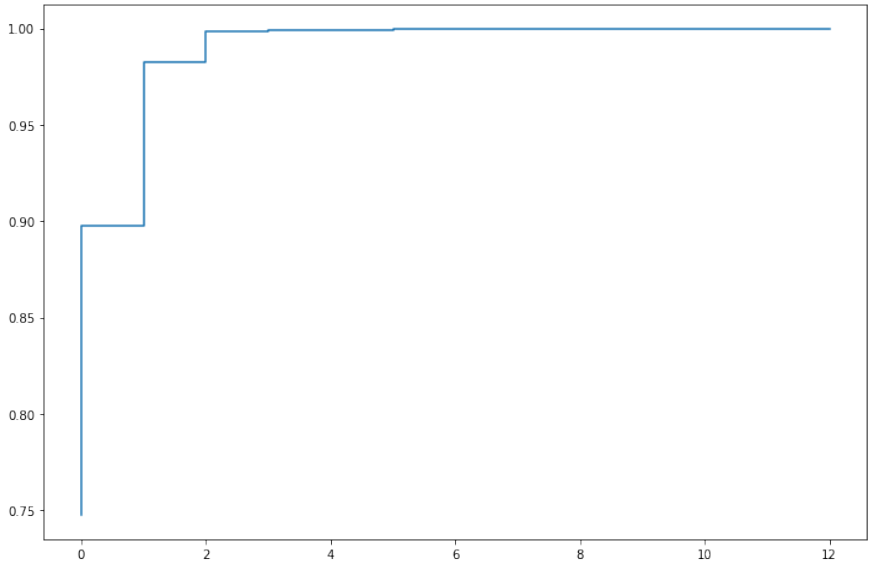

Figure 5: PCA explained

From the above plot we can derive that if we use 2 components we get almost 90% of the variance of the dataset. After applying the principal component analysis our data distribution is the following.

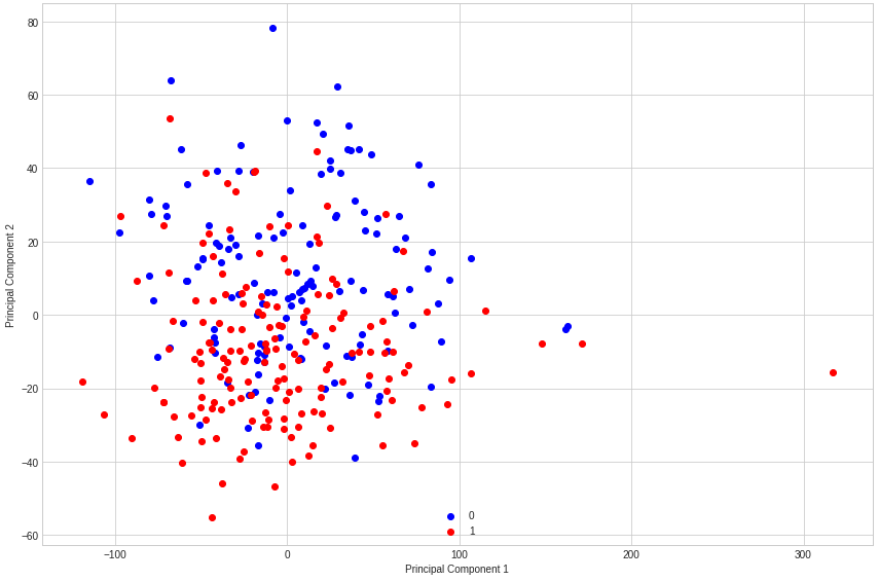

Figure 6: PCA scatter

Another important aspect to consider is the importance of each variable for each dimension of the principal component analysis. A break down of the variables can be found in the next plot.

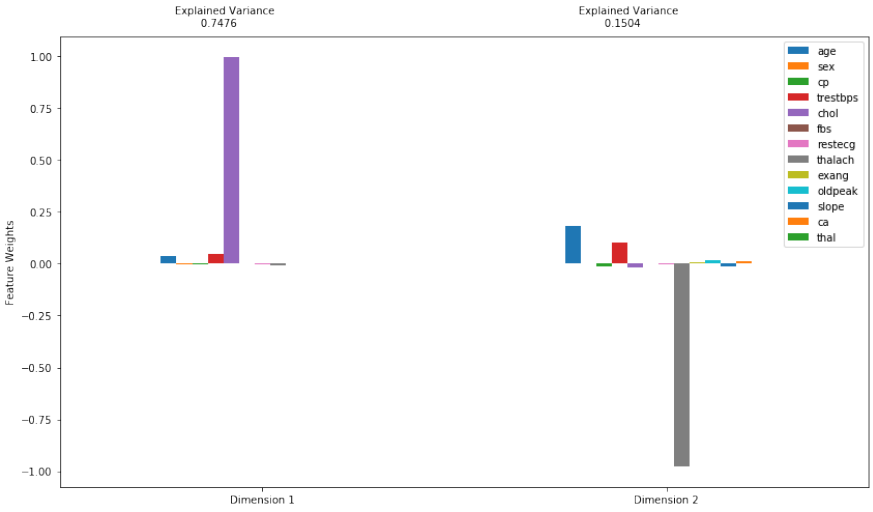

Figure 7: PCA weights analysis

From the above plots we can make the following observations:

* Regarding the first dimension, it is noticeable that ”chol” has a large and positive weight, while ”sex” and ”cp” are slightly positive. This means that people with a high value in sex and cp will have very little impact on heart disease, whereas people with higher cholesterol have greater chances of heart disease.

* For the second dimension, it is clear that the weight is large and negative for ”thalach” and slightly negative for ”cp”, ”chol” and ”slope”. Similar to the previously mentioned features, those with a lot of weight will have a much larger impact compared to features with smaller weights. However, the features in this case will have a cause the chance of heart disease to become smaller. Finally ”age” and ”trestbps” are moderately positive.

4 Learnings from the data set

Contributing Author(s): Andrew Ardueser, Alexandros Dorodoulis

From analyzing the features, we see that all of them appear to have a useable distribution of values. Outside of fbs, none of the features' value distributions are overly dominated by a single value. Furthermore, several features have a relatively high significant relationship with each other and the target variable. We see that most of the features are mostly normally distributed. Considering the relatively small number of observations, we would not have expected such smooth distributions. This aspect of our dataset will likely help in our later analysis. Some features such as chol and trestbps have some outliers, skewing their respective distributions. However, normalization of the data will hopefully assist with this issue.

During the data visualization phase we identified that our dataset is normally distributed, but it includes a few outliers. Additionally, there are also a few variables with a relatively high correlation between them.

During the principal component analysis, we identified that we could use 2 components to account for 90% of the variance in our data. We also identified that people with high cholesterol run a bigger risk of facing a heart disease, while those with high values for the thalach and maximum heart rate features have a lower risk of heart disease. The rest of the features had significantly lower impact on the risk, but we will do further analysis in the next report for further validation.