Technical University of Denmark

Machine Learning and Datamining

Cleveland database : Analysis and PCA

Authors:
Victor Niaussat

March 9, 2021

Contents

1 Introduction

The Cleveland database is a dataset of clinical results of 303 patients undergoing angiography at the Cleveland Clinic (Cleveland, Ohio) between May t1981 and September 1984. The original dataset can be obtained from the UCI Machine Learning Repository [1]. The average age of the patients was 54 years and 68 % were men. The full database contains 76 attributes; however we here focus on a subset of 13 attributes in line with previously published studies. The target feature is an integer value referring to the presence (1) or absence (0) of heart disease in the patient, where the "present" value is a collective term for the four different types of defects considered in the study. The remaining attributes include basic patient data (age and sex) as well as diagnostic data relating to cardiovascular health and will be discussed in detail in section 2.

The dataset was used in a study by Detrano et al. in 1989 [2]. In this study, a discriminant function model based on logistic regression for estimating probabilities of angiographic coronary disease was developed from the Cleveland database. The model was applied to 3 patient test groups, with varying disease prevalence, of 768 patients in total. The results were compared to results from a Bayesian algorithm based on published medical studies applied to the same patient groups. Both algorithms were found to overpredict the probability of disease, however, for the reported model the overprediction was less pronounced, in particular for the groups with low disease prevalence. The study concluded that the derived model was reliable for patients with chest pain syndromes and intermediate disease prevalence.

In another study by Gennari et al. from 1989 an unsupervised learning system for concept formation, CLASSIT, was developed and applied to the Cleveland database [3]. The CLASSIT system is based on a concept hierarchy approach to achieve classification of data. Given the 13 attribute values, CLASSIT created three classes, of which one was found to corrrespond well to the class of patients without heart disease, while the other two corresponded to patients with heart disease with 66.6 and 79.7 % accuracy.

The aim of this report is to do initial visualization of the Cleveland data prior to applying classification and regression models to the data in the following reports. In the classification task we expect to predict the presence or absence of heart disease based on the remaining 12 diagnostic attributes in the subset. We furthermore aim to use regression to predict resting blood pressure based on relevant descriptors, such as age and cholesterol.

2 Attributes

Table 1 shows the first five records of the data. All attributes in the dataset but five are discrete. The continuous attributes are Age, the resting blood pressure (in mm/Hg), Cholesterol (in mg/dL), maximum heart rate (in BPM) and ST depression (induced relative to rest). These values, however, all take integer values in the dataset.

Age	Sex	Chest_ Pain	Resting_ Blood_ Pressure	Colestrol	Fasting_ Blood_ Sugar	Rest_ ECG	MAX_ Heart_ Rate	Exercised_ Induced_ Angina	ST_ Depression	Slope	Major_ Vessels	Thalessemia	Target
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0

Table 1: First five records

The dataset has a total of six missing values in the *Thalessemia* column marked with a question mark. Even though this attribute is not going to be used for the analysis, the rows with the missing values have been removed from the dataset. With more than 300 records, removing ¡ 2% of the observations is likely not going to affect the end result significantly. The rest of the attributes can further be placed in one of four categories - nominal, ordinal, interval and ratio. This is done in table 2, together with a short explanation of attribute meanings.

Attribute	Attribute type
Age	Ratio - Measuring the age in years
Sex	Nominal, binary - with 1 = male and 0 = female
Chest_Pain	Nominal - Denoting which of four types of chest pains the patient is experiencing. (0 = Asymptomatic, 1 = atypical angina, 2 = non-anginal pain, 3 = typical angina)
Resting_Blood_ Pressure	Ratio - measuring the resting blood pressure in mmHg
Cholestrol	Ratio - Cholesterol level in in mg/dL
Fasting_ Bloodsugar	Ordinal, binary - Denoting whether fasting blood sugar >120 mg/dL. (1 = true, 0 = false).
Resting_ ECG	Nominal - Showing patient's resting electrocardiographic results. (0 = normal, 1 = having ST-T, 2 = hypertrophy)
Max_Heart_ Rate	Ratio - Maximum heart rate achieved in Beats Per Minute
Exercise_ Induced_ Angina	Nominal, binary - Exercise induced angina (1 = yes, 0 = no)
ST_depression	Ratio - ST depression induced by exercise relative to rest
Slope	Nominal - The slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)
# of major blood vessels	Ratio - Number of major vessels colored by fluorosopy test (0 - 3)
Thalessemia	***
Target	Nominal - Shows if the patient has a heart disease (1,2,3,4 = Different severity of heart disease, 0 = No heart disease)

Table 2: Attribute types

Since the analysis of this paper is focused on classifying the presence of heart disease, every positive value of the *Target* column, has been set equal to one. As mentioned, the *Thalessemia* attribute has been removed from the dataset for the analysis, as it is unclear what this feature measures, resulting in a total of 13 attributes (12 if the attribute we wish to classify is excluded). In table 3 basic summary statistics mean, 0.25-, 0.5- and 0.75-quantiles, minimum, maximum, variance and standard deviance for each attribute is shown.

Attribute	Mean	Quantile0.25	Median	Quantile0.75	Min	Max	Variance	Std. deviance
Age	54.5	48.0	56.0	61.0	29.0	77.0	81.6	9.0
Sex	0.7	0.0	1.0	1.0	0.0	1.0	0.2	0.5
Chest_pain	3.2	3.0	3.0	4.0	1.0	4.0	0.9	1.0
Resting_Blood_Pressure	131.7	120.0	130.0	140.0	94.0	200.0	314.5	17.7
Colestrol	247.4	211.0	243.0	276.0	126.0	564.0	2694.6	51.9
Fasting_Blood_Sugar	0.1	0.0	0.0	0.0	0.0	1.0	0.1	0.4
Rest_ECG	1.0	0.0	1.0	2.0	0.0	2.0	1.0	1.0
Max_Heart_Rate	149.6	133.0	153.0	166.0	71.0	202.0	524.5	22.9
Exercised_Induced_Angina	0.3	0.0	0.0	1.0	0.0	1.0	0.2	0.5
ST_Depression	1.1	0.0	0.8	1.6	0.0	6.2	1.4	1.2
Slope	1.6	1.0	2.0	2.0	1.0	3.0	0.4	0.6
Major_Vessels	0.7	0.0	0.0	1.0	0.0	3.0	0.9	0.9
Thalassemia	4.7	3.0	3.0	7.0	3.0	7.0	3.7	1.9
Target	0.9	0.0	0.0	2.0	0.0	4.0	1.5	1.2

Table 3: Summary statistics

From the summary statistics a large max value for the cholesterol attribute can be deduced. The box plot is a method for graphically depicting groups of numerical data through their quartiles. This allows for the detection of outliers. As shown on figure 1, there is an outlier in the cholesterol plot. A person has a cholesterol level of 584 mg/dl which is unreasonably high. Hence, this datapoint is removed from the dataset in the data analysis. The other extreme values are close enough to the minimum or maximum values of their respective columns to be included in the analysis.

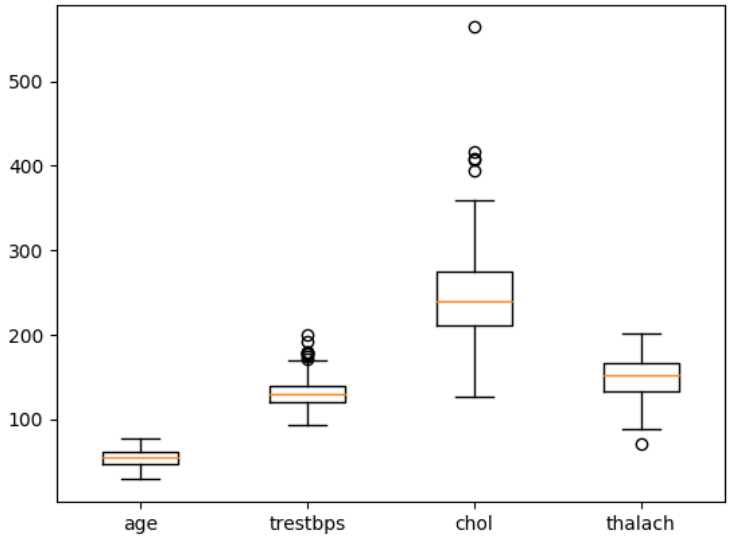

Figure 1: Box plot

3 Data visualization and PCA

3.1 Data visualization

Figure 2 shows histograms of each of the 12 attribute in the dataset. It is observed, that four of the six ratio type attributes, age, resting blood pressure, cholesterol and maximum heart rate to a good approximation follow a normal distribution, while the ST depression does not. As observed in the boxplot in Figure 1, the patient having a cholesterol concentration well above 500 mg/dl is clearly an outlier, and is removed from the dataset in the following.

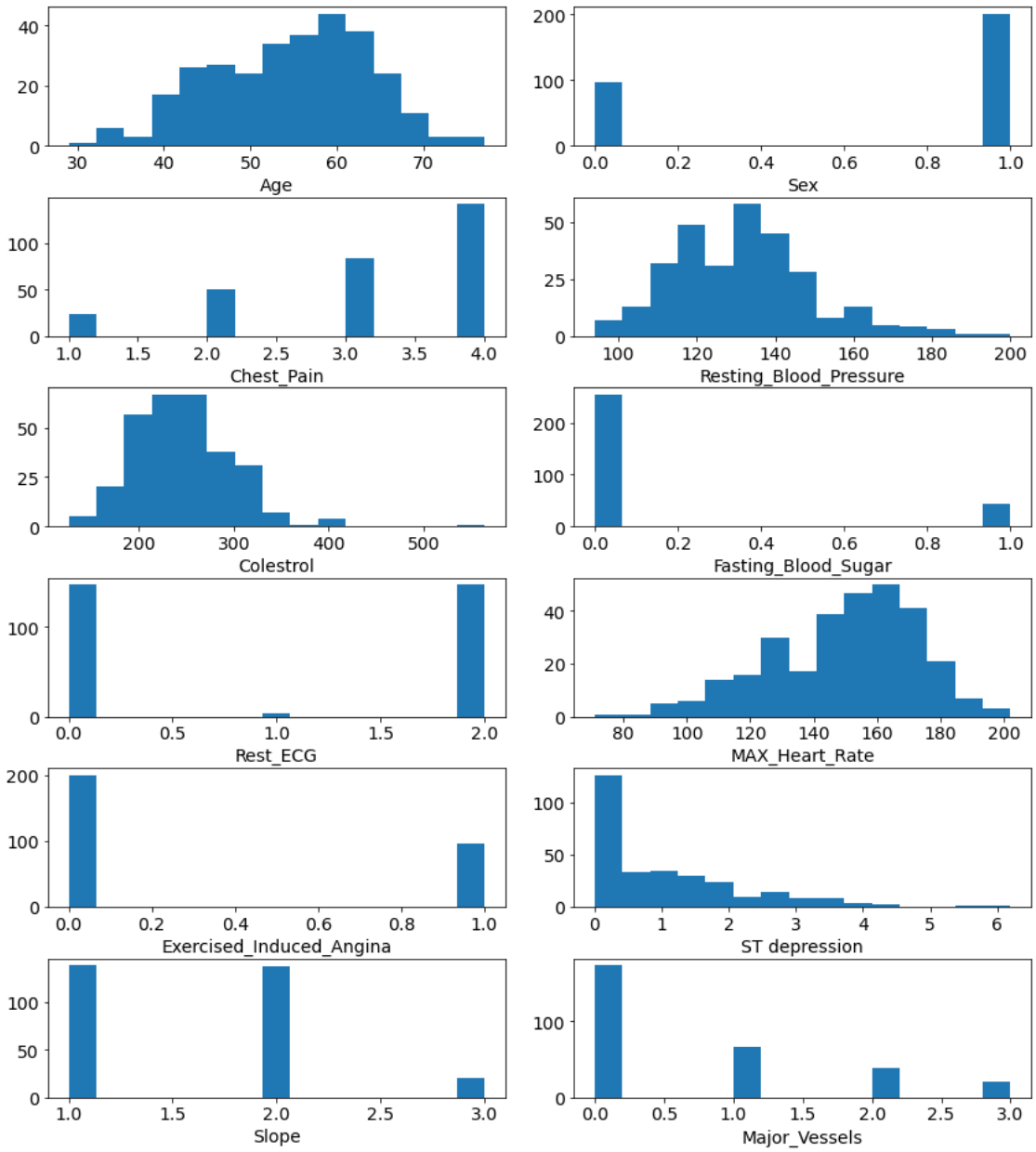

Figure 2: Histogram of each attribute in the dataset

The five attributes of the ratio type have been plotted in scatter plots to investigate possible correlations in Figure 3. The outlier in cholestorol has been removed from the dataset for clarity. It is observed that maximum heart rate, cholestorol and resting blood pressure seem to be correlated with age. Furthermore, weak correlations between cholestorol and maximum heart rate and resting blood pressure seem to be present, hence it is expected that regression of these atributes is feasible. Interstingly, several of the attributes seem to be reasonable descriptors to seperate patients with presence and absence of heart disease, including maximum heart rate in particular.

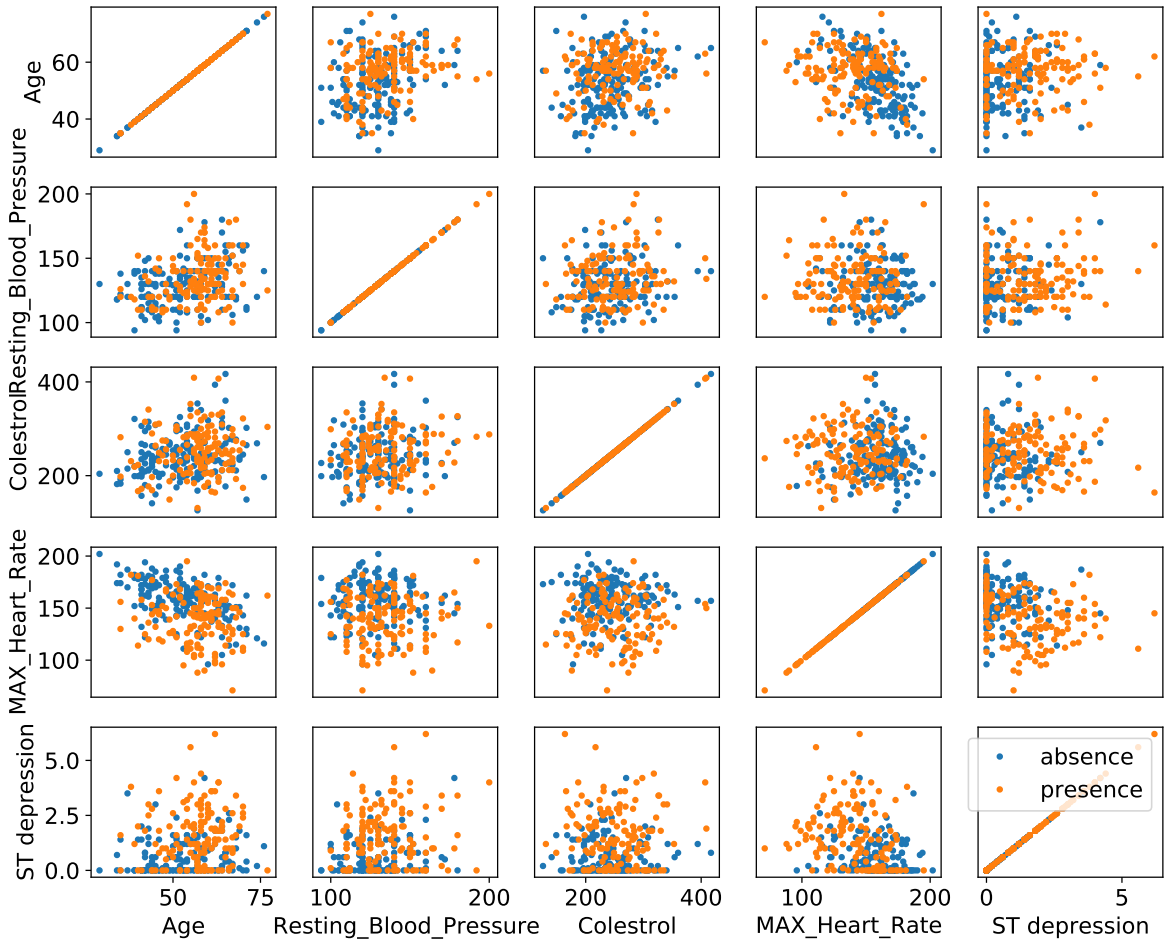

Figure 3: Scatter plots showing the correlations between all ratio type attributes

3.2 Principal component analysis

Principal component analysis (PCA) is a widely applicable technique where the goal is to find a lower-dimensional representation of a high-dimensional dataset.The purpose of this PCA is to reduce the dimension of this dataset (12) to a lower dimension (2 or 3) to detect and separate people with and without heart disease.

It should be noted that the present dataset contains a mix of continuous and discrete variables. This presents a problem, as the PCA method is designed for continuous data and might not handle discrete variables ideally. More appropriate methods for discrete data exists, such as categorial PCA (CATPCA), however it is beyond the scope of this report. An alternative approach would be to only include the continuous data in the analysis, but that would lead to the exclusion of attributes that are likely to be important risk indicators, such as chest pain and electrocardiography data. In the following we include continuous and discrete values, crudely ignoring the discreteness of discrete attributes.

The first step of the PCA algorithm is to subtract the mean of each attribute. this is followed by the computation of the singular value decomposition to project the dataset on principal direction. The attributes have different scales and different physical units, so we standardize by the standard deviation prior to the PCA analysis as shown in Figure 4.

The individual and cummulated variance explained by each principal component are shown in Figure 5. It

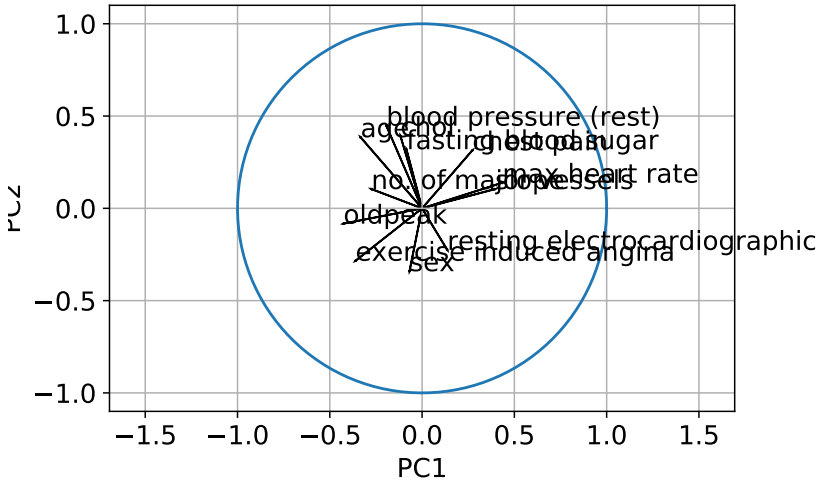

Figure 4: Attribute coefficients for standardized dataset

is a natural way to measure how much variance or information about data is retained in a reconstruction based on n principal components is the variance explained.

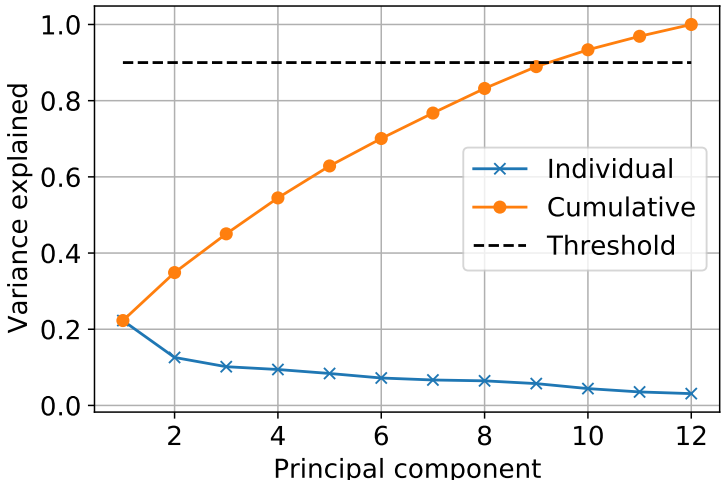

Figure 5: Variance explained for each principal component

The variance for the standardized dataset is explained by the different directions. Hence, we should project dataset on a higher-dimensional plane to represent more variations.

Nevertheless, we can make a clear distinction between people having a heart disease or not with a projection on PC1 and PC2 as shown in Figure 6 :

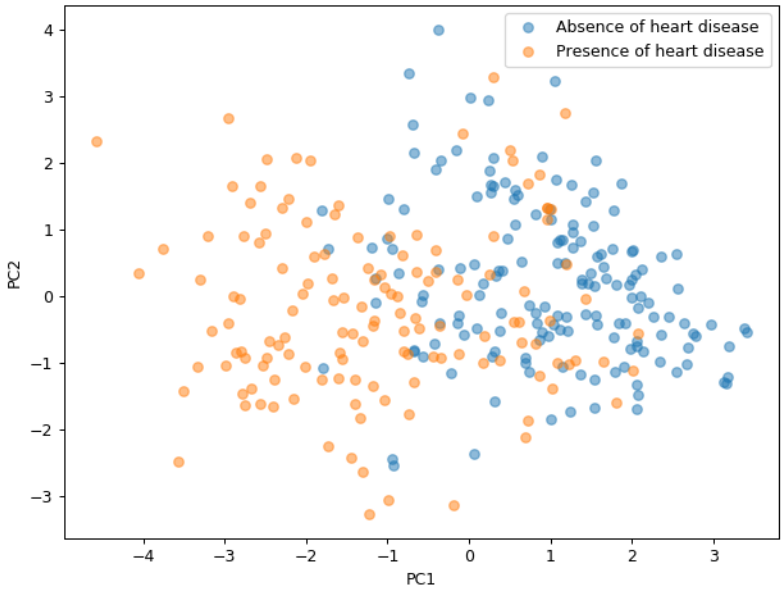

Figure 6: Data standardized projected onto PC1 and PC2.

In this figure, we can distinguish 3 areas:

- On the left side, nearly all people have a heart disease
- On the right side, nearly all do not heart disease
- In the middle, people are either sick or healthy. This is an area of uncertainty. A person can possibly have a heart disease

Moreover, it is interesting to project the dataset on the 3-dimensionally hyperplane having for principal directions v_1, v_2 and v_3 where it's a better representation approximation than a 2-dimensionally plane because there are more variations represented.

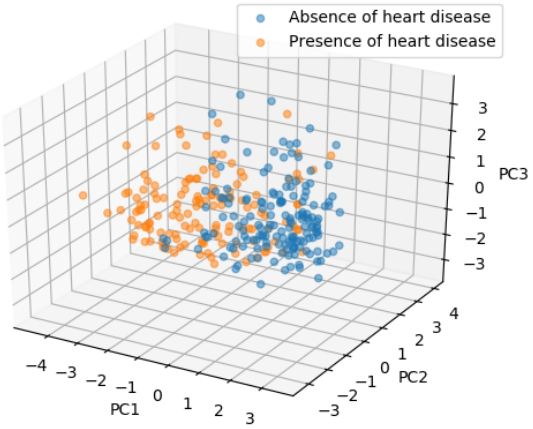

Figure 7: Data standardized projected onto PC1, PC2 and PC3.

4 Conclusion

The results show, that the PCA analysis does quite well in classifying the records into either having a heart disease or not, based on the attributes in the data. As expected, based on the previous studies using the dataset, this suggest that obtaining a machine learning model for classification based on the 12 selected attributes is feasible. While a machine learning model should be able to predict the mode outcome value for classification problems, the other part of this course is to predict mean outcome values for regression problems. Based on figure 3 there seem to be correlations between multiple variables, which may indicate that a regression model can be applied to predict one variable, such as the resting blood pressure from other relevant descriptors.

References

[1] UCI Machine Learning Repository. *Heart Disease Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[2] Robert Detrano et al. "International application of a new probability algorithm for the diagnosis of coronary artery disease". In: *The American Journal of Cardiology* 64.5 (1989), pp. 304–310. ISSN: 00029149. DOI: 10.1016/0002-9149(89)90524-9.

[3] John H. Gennari, Pat Langley, and Doug Fisher. "Models of incremental concept formation". In: *Artificial Intelligence* 40.1-3 (1989), pp. 11–61. ISSN: 00043702. DOI: 10.1016/0004-3702(89)90046-5.