Predicting a Person Diagnose a Heart Disease based on Cleveland Database

MATH 2319 Machine Learning Applied Project Phase I

Praneetha Meegahalanda Durage (s3685754)

08 April 2018

Introduction

The objective of this project is to build a model to predict whether a person diagnose a heart disease based on pattern extracted from analysing 14 descriptive features found in Cleveland data set from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). The project consists of two phases. Phase I focuses on data preprocessing and exploration, as covered in this report. The model building, validation and prediction are presented in Phase II. The rest of this report is organised as follow. Section 2 describes the data sets and their attributes. Section 3 covers data pre-processing. In Section 4, we explore each attribute and their inter-relationships. The last section ends with a summary.

Data Set

The UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disease) provides four databases linked to the above source namely, Cleveland, Hungary, Switzerland and the VA Long Beach with 76 attributes. Out of these 4 sets Cleveland data set was the benchmark for many researchers as it is more robust and easy to use compared to other 3. Out of 76 attributes 14 were selected as many of them repeat similar information and some attributes are not related with target attribute. Details of 303 patients information were included in this data set. The data was initially donated by David W. Aha (aha '@' ics.uci.edu).

Target Feature

The response feature is, the presence of heart disease (num). It takes 5 levels based on angiographic disease status. 0-Healthy, 1-diagnosed with stage 1, 2-diagnosed with stage 2, 3-diagnosed with stage 3, 4-diagnosed with stage 4. For this project we just consider whether a person diagnose with a heart disease or not. Therefore, levels of the target feature is reduced to 2 by combining stage 1-4 to a single level and give a new column name as "target" with two levels of "Yes" and "No" (0=No, 1-4=Yes).

Descriptive Features

Only 13 descriptive features named below are used.

- age: Age of the patient (Continuous)
- sex: Sex of the patient (Categorical with 2 levels- Male, Female)
- cp: Chest pain type (Categorical with 4 levels-Type 1, Type 2, Type 3 and Type 4) Type 1:typical angina Type 2:atypical angina Type 3:non-anginal pain Type 4:asymptomatic
- trestbps: Resting blood pressure-in mm Hg on admission to the hospital(Continuous)
- chol: Serum cholesterol in mg/dl (Continuous)
- fbs: Fasting blood sugar > 120 mg/dl (Categorical with 2 levels-True,False)
- restecg: Resting electrocardiographic results (Categorical with 3 levels-N(Normal), L1(Level 1), L2(Level 2))
- thalach: Maximum heart rate achieved (Continuous)
- exang: Exercise induced angina (Categorical with 2 levels-Yes, No)
- oldpeak: ST depression induced by exercise relative to rest (Continuous)
- slope: The slope of the peak exercise ST segment (categorical with 3 levels-Up, Flat, Down)
- ca: Number of major vessels (0-3) colored by flourosopy (Categorical with 4 levels-0, 1, 2, 3)
- thal: The heart status as retrieved from Thallium test (Categorical with 3 levels-N(normal),FD(fixed defect), RD(reversible defect)

Data Pre-processing

Preliminaries (Optional)

In this project, we used the following `R` packages.

```
library(knitr)
library(readr)
library(dplyr)
library(ggplot2)
library(mlr)
library(cowplot)
```

We read the data to R and checked the internal structure of the data and got a column wise summary in order to make sure that data is free of anomalies.

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : int  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp      : int  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: int  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 ...
##  $ restecg : int  2 2 2 0 2 0 2 0 2 2 ...
##  $ thalach : int  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : int  0 1 1 0 0 0 0 1 0 1 ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope   : int  3 2 2 3 1 1 3 1 2 3 ...
##  $ ca      : Factor w/ 5 levels "?","0","1","2",..: 2 5 4 2 2 2 4 2 3 2 ...
##  $ thal    : Factor w/ 4 levels "?","3","6","7": 3 2 4 2 2 2 2 2 4 4 ...
##  $ num     : int  0 2 1 0 0 0 3 0 2 1 ...
```

Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
age	integer	0	54.4389439	9.0386624	56.0	8.89560	29	77.0	0
sex	integer	0	0.6798680	0.4672988	1.0	0.00000	0	1.0	0
cp	integer	0	3.1584158	0.9601256	3.0	1.48260	1	4.0	0
trestbps	integer	0	131.6897690	17.5997477	130.0	14.82600	94	200.0	0
chol	integer	0	246.6930693	51.7769175	241.0	47.44320	126	564.0	0
fbs	integer	0	0.1485149	0.3561979	0.0	0.00000	0	1.0	0
restecg	integer	0	0.9900990	0.9949713	1.0	1.48260	0	2.0	0
thalach	integer	0	149.6072607	22.8750033	153.0	22.23900	71	202.0	0
exang	integer	0	0.3267327	0.4697945	0.0	0.00000	0	1.0	0
oldpeak	numeric	0	1.0396040	1.1610750	0.8	1.18608	0	6.2	0