Example : <https://towardsdatascience.com/random-forest-in-r-f66adf80ec9>

# Section 1: Getting Started

## Introduction

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This analysis intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression and random forest algorithms.

## Research question

We would like to do Exploratory Data Analysis on the data to get an understanding of the different features and how they are related to one another. We would also look at the variables and see if they are related to one another.We are trying to build a classification problem to answer our main question about the chances of having a heart disease.

1. What are the chances of having a heart disease based on the Vital data collected from a patient?
2. What is the most significant predictor for a Heart disease?
3. Is their a correlation between Age and Sex with the chances of having a heart disease?
4. Is there a possibility that based on the analysis of the vital data if we could prevent a possible heart failure from happening?

## Approach

I am planning to use Logistic Regression algorithm using glm and random forest in order to do a comparison to the output. The random forest algorithm works by aggregating the predictions made by multiple decision trees. I would be using bootstrapped dataset created from the original dataset.

## How my approach addresses (fully or partially) the problem

When the random forest is used for classification and is presented with a new sample, the final prediction is made by taking the majority of the predictions made by each individual decision tree in the forest. In the event, it is used for regression and it is presented with a new sample, the final prediction is made by taking the average of the predictions made by each individual decision tree in the forest.

## Data

I have looked into various sources for this dataset that could be helpful in predicting the possible heart failure for an individual based on patient data that are collected for each of the patients on a regular checkup. Finally the below dataset was found to be most suitable. It was available in kaggle also.

Link to the dataset: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

This dataset integrates all the databases present in Heart Disease Dataset available at UCI Machine Learning Repository. Original one contains 4 databases: Cleveland, Hungarian, Long Beach, and Switzerland. Most of the work has been done using Cleveland dataset only. The authors of the databases have requested:

...that any publications resulting from the use of the data include the   
 names of the principal investigator responsible for the data collection  
 at each institution. They would be:  
  
 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.  
 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.  
 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.  
 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:  
 Robert Detrano, M.D., Ph.D.

## Required Packages

I am planning to use the following packages to start with and some more might be needed as I start the work: package : tidyr package : dplyr package : ggplot2 package : broom package : randomForest package : caTools

## Plots and Table

Tables: Showing the first few rwos of the data set

Plot: scatter plot showing the distribution of the important features 3-d plot if needed for the data histogram

## Questions for future steps

As a starting point I would like to answer the following questions: 1) What is the distribution of the data for the important features? 2) How does the features have there values normalized or do they have biases? 3) Is there any collinearity between the variables/features? 4) What are most imporatnt features from the list of all the features available in the dataset?

# Section 2 (Week 10)

Loading the required libraries.

library(tidyr)  
library(dplyr)  
library(ggplot2)  
library(broom)  
library(randomForest)

## How to import and clean my data

### About the source data

The actual database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. But we would be using all the other datasets also. The “goal” field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

The names and social security numbers of the patients were removed from the database, replaced with dummy values. We are using all four processed files which also exist in the dataset directory.

### Summary of the various datasets are given below.

Database: # of instances:  
 Cleveland: 303  
 Hungarian: 294  
 Switzerland: 123  
 Long Beach VA: 200

### Attribute Information: The attributes that are defined in the below datasets are defined here. It also shows the position of the attributes in the actual files.

– Only 14 used – V1. #3 (age) : Age in years  
– V2. #4 (sex) : sex (1 = male; 0 = female) – V3. #9 (cp) : chest pain type (1:typical angina, 2:atypical angina, 3:non-anginal pain, 4: asymptomatic) – V4. #10 (trestbps) : resting blood pressure (in mm Hg on admission to the hospital) – V5. #12 (chol) : serum cholestoral in mg/dl – V6. #16 (fbs) : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) – V7. #19 (restecg) : resting electrocardiographic results 0: normal, 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: showing probable or definite left ventricular hypertrophy by Estes’ criteria – V8. #32 (thalach) : maximum heart rate achieved – V9. #38 (exang) : exercise induced angina (1 = yes; 0 = no) – V10. #40 (oldpeak) : ST depression induced by exercise relative to rest  
– V11. #41 (slope) : the slope of the peak exercise ST segment 1: upsloping 2: flat 3: downsloping – V12. #44 (ca) : number of major vessels (0-3) colored by flourosopy – V13. #51 (thal) : 3 = normal; 6 = fixed defect; 7 = reversable defect  
– V14. #58 (num) : the predicted attribute diagnosis of heart disease (angiographic disease status) 0: < 50% diameter narrowing (No heart disease) 1: > 50% diameter narrowing ( Yes Heart disease)

I would start by importing each of the 4 files which have already being processed, into 4 different data frames and looking at those separately in order to get a better understanding of the values of the attributes and how they are distributed.

#populating the cleveland file into dataframe  
wd <- getwd()  
cleveland\_f <- "processed.cleveland.data"  
path\_to\_file <- paste(wd,'/dataset/',cleveland\_f, sep = "")  
path\_to\_file  
  
cleveland\_df <- read.csv(path\_to\_file, header = FALSE)  
## dim(cleveland\_df)  
## summary(cleveland\_df)  
## head(cleveland\_df)  
  
#populating the Hungarian file into dataframe  
hungarian\_f <- "processed.hungarian.data"  
path\_to\_file <- paste(wd,'/dataset/',hungarian\_f, sep = "")  
path\_to\_file  
  
hungarian\_df <- read.csv(path\_to\_file, header = FALSE)  
## dim(hungarian\_df)  
## summary(hungarian\_df)  
## head(hungarian\_df)  
  
#populating the Switzerland file into dataframe  
switzerland\_f <- "processed.switzerland.data"  
path\_to\_file <- paste(wd,'/dataset/',switzerland\_f, sep = "")  
path\_to\_file  
  
switzerland\_df <- read.csv(path\_to\_file, header = FALSE)  
## dim(switzerland\_df)  
## summary(switzerland\_df)  
## head(switzerland\_df)  
  
#populating the Long Beach, CA data file into dataframe  
long\_beach\_f <- "processed.va.data"  
path\_to\_file <- paste(wd,'/dataset/',long\_beach\_f, sep = "")  
path\_to\_file  
  
long\_beach\_df <- read.csv(path\_to\_file, header = FALSE)  
## dim(long\_beach\_df)  
## summary(long\_beach\_df)  
## head(long\_beach\_df)  
  
dim(cleveland\_df)  
dim(hungarian\_df)  
dim(switzerland\_df)  
dim(long\_beach\_df)

After having a initial look at the datasets, now adding the 4 individual dataframes into separate data datasets and also adding the column names for each dataset in order to do some more analysis on each of those, as we know that the data do not have the column names.

# Adding the new column for each of the data source name. Also at the same time adding an extra column to keep the Source name along with the data set. Now looking at the data briefly.

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal out  
## 1 63 1 1 145 233 1 2 150 0 2.3 3 0.0 6.0 0  
## 2 67 1 4 160 286 0 2 108 1 1.5 2 3.0 3.0 2  
## 3 67 1 4 120 229 0 2 129 1 2.6 2 2.0 7.0 1  
## 4 37 1 3 130 250 0 0 187 0 3.5 3 0.0 3.0 0  
## 5 41 0 2 130 204 0 2 172 0 1.4 1 0.0 3.0 0  
## 6 56 1 2 120 236 0 0 178 0 0.8 1 0.0 3.0 0  
## datasrc  
## 1 Cleveland  
## 2 Cleveland  
## 3 Cleveland  
## 4 Cleveland  
## 5 Cleveland  
## 6 Cleveland

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal out  
## 1 28 1 2 130 132 0 2 185 0 0 ? ? ? 0  
## 2 29 1 2 120 243 0 0 160 0 0 ? ? ? 0  
## 3 29 1 2 140 ? 0 0 170 0 0 ? ? ? 0  
## 4 30 0 1 170 237 0 1 170 0 0 ? ? 6 0  
## 5 31 0 2 100 219 0 1 150 0 0 ? ? ? 0  
## 6 32 0 2 105 198 0 0 165 0 0 ? ? ? 0  
## datasrc  
## 1 Hungarian  
## 2 Hungarian  
## 3 Hungarian  
## 4 Hungarian  
## 5 Hungarian  
## 6 Hungarian

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal out  
## 1 32 1 1 95 0 ? 0 127 0 .7 1 ? ? 1  
## 2 34 1 4 115 0 ? ? 154 0 .2 1 ? ? 1  
## 3 35 1 4 ? 0 ? 0 130 1 ? ? ? 7 3  
## 4 36 1 4 110 0 ? 0 125 1 1 2 ? 6 1  
## 5 38 0 4 105 0 ? 0 166 0 2.8 1 ? ? 2  
## 6 38 0 4 110 0 0 0 156 0 0 2 ? 3 1  
## datasrc  
## 1 Switzerland  
## 2 Switzerland  
## 3 Switzerland  
## 4 Switzerland  
## 5 Switzerland  
## 6 Switzerland

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal out  
## 1 63 1 4 140 260 0 1 112 1 3 2 ? ? 2  
## 2 44 1 4 130 209 0 1 127 0 0 ? ? ? 0  
## 3 60 1 4 132 218 0 1 140 1 1.5 3 ? ? 2  
## 4 55 1 4 142 228 0 1 149 1 2.5 1 ? ? 1  
## 5 66 1 3 110 213 1 2 99 1 1.3 2 ? ? 0  
## 6 66 1 3 120 0 0 1 120 0 -0.5 1 ? ? 0  
## datasrc  
## 1 Long Beach VA  
## 2 Long Beach VA  
## 3 Long Beach VA  
## 4 Long Beach VA  
## 5 Long Beach VA  
## 6 Long Beach VA

Here for our problem, we are only going to attempt to distinguish the presence of heart disease (values 1,2,3,4) from absence of heart disease (value 0). Therefore, we replace all labels greater than 1 by 1. Then taking summary of each of the data sets. Showing below our findings on each of the various datasets.

We see that the data set is showing mean for categorical variables also. Hence we need to re-specify the column types. Also we find that there missing values for

## age sex cp trestbps   
## Min. :29.00 Min. :0.0000 Min. :1.000 Min. : 94.0   
## 1st Qu.:48.00 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:120.0   
## Median :56.00 Median :1.0000 Median :3.000 Median :130.0   
## Mean :54.44 Mean :0.6799 Mean :3.158 Mean :131.7   
## 3rd Qu.:61.00 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:140.0   
## Max. :77.00 Max. :1.0000 Max. :4.000 Max. :200.0   
## chol fbs restecg thalach   
## Min. :126.0 Min. :0.0000 Min. :0.0000 Min. : 71.0   
## 1st Qu.:211.0 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:133.5   
## Median :241.0 Median :0.0000 Median :1.0000 Median :153.0   
## Mean :246.7 Mean :0.1485 Mean :0.9901 Mean :149.6   
## 3rd Qu.:275.0 3rd Qu.:0.0000 3rd Qu.:2.0000 3rd Qu.:166.0   
## Max. :564.0 Max. :1.0000 Max. :2.0000 Max. :202.0   
## exang oldpeak slope ca thal   
## Min. :0.0000 Min. :0.00 Min. :1.000 ? : 4 ? : 2   
## 1st Qu.:0.0000 1st Qu.:0.00 1st Qu.:1.000 0.0:176 3.0:166   
## Median :0.0000 Median :0.80 Median :2.000 1.0: 65 6.0: 18   
## Mean :0.3267 Mean :1.04 Mean :1.601 2.0: 38 7.0:117   
## 3rd Qu.:1.0000 3rd Qu.:1.60 3rd Qu.:2.000 3.0: 20   
## Max. :1.0000 Max. :6.20 Max. :3.000   
## out datasrc   
## Min. :0.0000 Length:303   
## 1st Qu.:0.0000 Class :character   
## Median :0.0000 Mode :character   
## Mean :0.4587   
## 3rd Qu.:1.0000   
## Max. :1.0000

## age sex cp trestbps chol   
## Min. :28.00 Min. :0.0000 Min. :1.000 120 :65 ? : 23   
## 1st Qu.:42.00 1st Qu.:0.0000 1st Qu.:2.000 130 :54 230 : 5   
## Median :49.00 Median :1.0000 Median :3.000 140 :50 246 : 5   
## Mean :47.83 Mean :0.7245 Mean :2.983 150 :23 275 : 5   
## 3rd Qu.:54.00 3rd Qu.:1.0000 3rd Qu.:4.000 110 :21 196 : 4   
## Max. :66.00 Max. :1.0000 Max. :4.000 160 :20 211 : 4   
## (Other):61 (Other):248   
## fbs restecg thalach exang oldpeak slope ca thal   
## ?: 8 ?: 1 150 : 29 ?: 1 Min. :0.0000 ?:190 ?:291 ?:266   
## 0:266 0:235 140 : 21 0:204 1st Qu.:0.0000 1: 12 0: 3 3: 7   
## 1: 20 1: 52 130 : 17 1: 89 Median :0.0000 2: 91 6: 10   
## 2: 6 170 : 14 Mean :0.5861 3: 1 7: 11   
## 160 : 13 3rd Qu.:1.0000   
## 120 : 11 Max. :5.0000   
## (Other):189   
## out datasrc   
## Min. :0.0000 Length:294   
## 1st Qu.:0.0000 Class :character   
## Median :0.0000 Mode :character   
## Mean :0.3605   
## 3rd Qu.:1.0000   
## Max. :1.0000   
##

## age sex cp trestbps chol   
## Min. :32.00 Min. :0.0000 Min. :1.000 115 :14 Min. :0   
## 1st Qu.:51.00 1st Qu.:1.0000 1st Qu.:4.000 120 :13 1st Qu.:0   
## Median :56.00 Median :1.0000 Median :4.000 160 :11 Median :0   
## Mean :55.32 Mean :0.9187 Mean :3.699 110 :10 Mean :0   
## 3rd Qu.:61.50 3rd Qu.:1.0000 3rd Qu.:4.000 130 :10 3rd Qu.:0   
## Max. :74.00 Max. :1.0000 Max. :4.000 140 :10 Max. :0   
## (Other):55   
## fbs restecg thalach exang oldpeak slope ca thal   
## ?:75 ?: 1 120 : 9 ?: 1 0 :42 ?:17 ?:118 ?:52   
## 0:43 0:85 128 : 5 0:68 2 :11 1:33 1: 2 3:19   
## 1: 5 1:30 110 : 4 1:54 1 :10 2:61 2: 3 6:10   
## 2: 7 115 : 4 ? : 6 3:12 7:42   
## 122 : 4 1.5 : 6   
## 100 : 3 .5 : 5   
## (Other):94 (Other):43   
## out datasrc   
## Min. :0.000 Length:123   
## 1st Qu.:1.000 Class :character   
## Median :1.000 Mode :character   
## Mean :0.935   
## 3rd Qu.:1.000   
## Max. :1.000   
##

## age sex cp trestbps chol   
## Min. :35.00 Min. :0.00 Min. :1.000 ? :56 0 : 49   
## 1st Qu.:55.00 1st Qu.:1.00 1st Qu.:3.000 120 :16 ? : 7   
## Median :60.00 Median :1.00 Median :4.000 130 :15 203 : 4   
## Mean :59.35 Mean :0.97 Mean :3.505 140 :10 220 : 4   
## 3rd Qu.:64.00 3rd Qu.:1.00 3rd Qu.:4.000 110 : 9 258 : 4   
## Max. :77.00 Max. :1.00 Max. :4.000 150 : 9 186 : 3   
## (Other):85 (Other):129   
## fbs restecg thalach exang oldpeak slope ca   
## ?: 7 Min. :0.000 ? : 53 ?:53 ? :56 ?:102 ?:198   
## 0:125 1st Qu.:0.000 120 : 12 0:52 0 :40 1: 16 0: 2   
## 1: 68 Median :1.000 140 : 12 1:95 2 :25 2: 53   
## Mean :0.735 110 : 8 1.5 :21 3: 29   
## 3rd Qu.:1.000 112 : 6 1 :18   
## Max. :2.000 130 : 6 3 :13   
## (Other):103 (Other):27   
## thal out datasrc   
## ?:166 Min. :0.000 Length:200   
## 3: 4 1st Qu.:0.000 Class :character   
## 6: 8 Median :1.000 Mode :character   
## 7: 22 Mean :0.745   
## 3rd Qu.:1.000   
## Max. :1.000   
##

data1 %>% group\_by(out) %>% summarise(n())

# stacking the dataframes together now

combined.data <- rbind(data1,data2,data3,data4)

combined.data %>% group\_by(datasrc,out) %>% summarise(n())

head(combined.data) ``` ## What does the final data set look like With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible. What do you not know how to do right now that you need to learn to import and cleanup your dataset? Discuss how you plan to uncover new information in the data that is not self-ent to answer? Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain. How could you summarize your data to answer key questions?vident. What are different ways you could look at this data to answer the questions you wa What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

## Questions for future step

What do you not know how to do right now that you need to learn to answer your questions?  
Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.