Updated Project Proposal

Jake Kaihewalu (4012753) Jasmine Kellogg (4266367) Andrew Nguyen (5495395)
Masaki Osato (4222154)

March 02, 2020

Research Question

Can the presence of cardiovascualr disease be accurately predicted based off of a combination of objective, subjective, and examination data?

Data

Our response variable will be the binary classification variable "cardio", which is 1 if the patient tests positive and 0 if they test negative. There are 11 predictor variables, 6 of which are categorical and 5 are numeric.

Analysis Plan

Firstly, we plan on using exploratory data analysis and graphics to find whether some predictors are highly correlated to one another or insignificant in its effect on the response variable. We then plan on using logistic regression, classification trees and random forests to classify observations into two binary groups, splitting based on largest reductions in impurity (using either the Gini index or entropy). For the logistic regression, we can fit a model based on a new variable (cardio) that indicates whether or not a patient tested positive for cardiovascular disease. We plan to use 70% of our dataset for training and 30% of our dataset for testing since we have a significant amount of observations.

References

Cardiovascular Disease Dataset. Ulianova, Svetlana. Kaggle, 20 Jan. 2019, [www.kaggle.com/sulianova/cardiovascular-disease-dataset.]

Data Overview

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
set.seed(9)

## read in data
cardio_full <- read_delim("cardio_train.csv", ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   age = col_double(),
##   gender = col_double(),
##   height = col_double(),
##   weight = col_double(),
##   ap_hi = col_double(),
##   ap_lo = col_double(),
##   cholesterol = col_double(),
##   gluc = col_double(),
##   smoke = col_double(),
##   alco = col_double(),
##   active = col_double(),
##   cardio = col_double()
## )
```

```
cardio_full$gender <- as.factor(cardio_full$gender)
cardio_full$cholesterol <- as.factor(cardio_full$cholesterol)
cardio_full$gluc <- as.factor(cardio_full$gluc)
cardio_full$smoke <- as.factor(cardio_full$smoke)
cardio_full$alco <- as.factor(cardio_full$alco)
cardio_full$active <- as.factor(cardio_full$active)
cardio_full$cardio <- as.factor(cardio_full$cardio)

dim(cardio_full)
```

```
## [1] 70000    13
```

```
head(cardio_full)
```

```
## # A tibble: 6 x 13
##      id   age gender height weight ap_hi ap_lo cholesterol gluc  smoke alco
##   <dbl> <dbl> <fct>   <dbl>  <dbl> <dbl> <dbl> <fct>       <fct> <fct> <fct>
## 1     0 18393 2         168     62   110    80 1           1     0     0
## 2     1 20228 1         156     85   140    90 3           1     0     0
```

```
## 3     2 18857 1         165     64   130    70 3           1     0     0
## 4     3 17623 2         169     82   150   100 1           1     0     0
## 5     4 17474 1         156     56   100    60 1           1     0     0
## 6     8 21914 1         151     67   120    80 2           2     0     0
## # ... with 2 more variables: active <fct>, cardio <fct>
```

```
## Going to be working with a random sample of n = 10,000 for computational purposes
cardio_sample <- sample_n(cardio_full, 10000)

## Split the data
id <- sample(1:nrow(cardio_sample), 0.75*nrow(cardio_sample))
cardio_sample.train <- cardio_sample[id,]
cardio_sample.test <- cardio_sample[-id,]
dim(cardio_sample.train)
```

```
## [1] 7500   13
```

```
dim(cardio_sample.test)
```

```
## [1] 2500   13
```

```
#sumary statistics
summary(cardio_sample.train)
```

```
##        id              age        gender       height          weight
##  Min.   :    0   Min.   :10798   1:4794   Min.   : 68.0   Min.   : 21.00
##  1st Qu.:25171   1st Qu.:17632   2:2706   1st Qu.:159.0   1st Qu.: 65.00
##  Median :50337   Median :19705            Median :165.0   Median : 72.00
##  Mean   :50033   Mean   :19471            Mean   :164.5   Mean   : 74.43
##  3rd Qu.:74615   3rd Qu.:21359            3rd Qu.:170.0   3rd Qu.: 82.00
##  Max.   :99999   Max.   :23673            Max.   :207.0   Max.   :200.00
##      ap_hi             ap_lo         cholesterol gluc     smoke    alco
##  Min.   : -115.0   Min.   :   0.00   1:5598      1:6358   0:6819   0:7105
##  1st Qu.:  120.0   1st Qu.:  80.00   2:1023      2: 588   1: 681   1: 395
##  Median :  120.0   Median :  80.00   3: 879      3: 554
##  Mean   :  131.3   Mean   :  97.63
##  3rd Qu.:  140.0   3rd Qu.:  90.00
##  Max.   :14020.0   Max.   :8099.00
##  active   cardio
##  0:1478   0:3746
##  1:6022   1:3754
##
##
##
##
```

```
#basic model
glm.fit <- glm(cardio ~ age + cholesterol + active, data = cardio_sample.train, family = binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = cardio ~ age + cholesterol + active, family = binomial,
##     data = cardio_sample.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.0289  -1.0890   0.5215   1.1155   1.6927
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.439e+00  2.038e-01 -16.873  < 2e-16 ***
## age           1.756e-04  1.012e-05  17.359  < 2e-16 ***
## cholesterol2  5.076e-01  7.034e-02   7.216 5.36e-13 ***
## cholesterol3  1.250e+00  8.579e-02  14.573  < 2e-16 ***
## active1      -2.286e-01  6.074e-02  -3.764 0.000167 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10397.2  on 7499  degrees of freedom
## Residual deviance:  9711.2  on 7495  degrees of freedom
## AIC: 9721.2
##
## Number of Fisher Scoring iterations: 4
```