Classification of Heart Disease Indicator

Tyler Echols

7/10/2022

Classification: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

We are looking at data of calculating the chance of people with heart disease.

```
# Reading in the CSV file and displaying the data
HDI <- read.csv("heart_2020.csv")
head(HDI)
```

```
##   HeartDisease   BMI Smoking AlcoholDrinking Stroke PhysicalHealth
MentalHealth
## 1           No 16.60     Yes              No     No              3
30
## 2           No 20.34      No              No    Yes              0
0
## 3           No 26.58     Yes              No     No             20
30
## 4           No 24.21      No              No     No              0
0
## 5           No 23.71      No              No     No             28
0
## 6          Yes 28.87     Yes              No     No              6
0
##   DiffWalking    Sex AgeCategory  Race Diabetic PhysicalActivity GenHealth
## 1          No Female       55-59 White      Yes              Yes Very good
## 2          No Female 80 or older White       No              Yes Very good
## 3          No   Male       65-69 White      Yes              Yes      Fair
## 4          No Female       75-79 White       No               No      Good
## 5         Yes Female       40-44 White       No              Yes Very good
## 6         Yes Female       75-79 Black       No               No      Fair
##   SleepTime Asthma KidneyDisease SkinCancer
## 1         5    Yes            No        Yes
## 2         7     No            No         No
## 3         8    Yes            No         No
## 4         6     No            No        Yes
## 5         8     No            No         No
## 6        12     No            No         No
```

```
nrow(HDI)
```

```
## [1] 319795
```

```r
# More data read-in varaibles
str(HDI)
```

```
## 'data.frame':    319795 obs. of  18 variables:
##  $ HeartDisease    : chr  "No" "No" "No" "No" ...
##  $ BMI             : num  16.6 20.3 26.6 24.2 23.7 ...
##  $ Smoking         : chr  "Yes" "No" "Yes" "No" ...
##  $ AlcoholDrinking : chr  "No" "No" "No" "No" ...
##  $ Stroke          : chr  "No" "Yes" "No" "No" ...
##  $ PhysicalHealth  : num  3 0 20 0 28 6 15 5 0 0 ...
##  $ MentalHealth    : num  30 0 30 0 0 0 0 0 0 0 ...
##  $ DiffWalking     : chr  "No" "No" "No" "No" ...
##  $ Sex             : chr  "Female" "Female" "Male" "Female" ...
##  $ AgeCategory     : chr  "55-59" "80 or older" "65-69" "75-79" ...
##  $ Race            : chr  "White" "White" "White" "White" ...
##  $ Diabetic        : chr  "Yes" "No" "Yes" "No" ...
##  $ PhysicalActivity: chr  "Yes" "Yes" "Yes" "No" ...
##  $ GenHealth       : chr  "Very good" "Very good" "Fair" "Good" ...
##  $ SleepTime       : num  5 7 8 6 8 12 4 9 5 10 ...
##  $ Asthma          : chr  "Yes" "No" "Yes" "No" ...
##  $ KidneyDisease   : chr  "No" "No" "No" "No" ...
##  $ SkinCancer      : chr  "Yes" "No" "No" "Yes" ...
```

```r
table(HDI$HeartDisease)
```

```
##
##     No    Yes
## 292422  27373
```

```r
yes <- which(HDI$HeartDisease == "Yes")
no <- which(HDI$HeartDisease == "No")

length(yes)
```

```
## [1] 27373
```

```r
length(no)
```

```
## [1] 292422
```

```r
no_downsample <- sample(no, length(yes))
HDI <- HDI[c(no_downsample, yes),]

str(HDI)
```

```
## 'data.frame':    54746 obs. of  18 variables:
##  $ HeartDisease    : chr  "No" "No" "No" "No" ...
##  $ BMI             : num  30 22.7 21.9 29.8 30.2 ...
##  $ Smoking         : chr  "No" "Yes" "No" "Yes" ...
##  $ AlcoholDrinking : chr  "No" "No" "No" "No" ...
##  $ Stroke          : chr  "No" "No" "No" "No" ...
##  $ PhysicalHealth  : num  5 0 0 0 14 0 0 0 0 0 ...
```

```
##  $ MentalHealth    : num  3 2 0 0 0 14 2 5 0 0 ...
##  $ DiffWalking     : chr  "No" "No" "No" "No" ...
##  $ Sex             : chr  "Female" "Female" "Male" "Female" ...
##  $ AgeCategory     : chr  "70-74" "75-79" "40-44" "65-69" ...
##  $ Race            : chr  "White" "White" "White" "White" ...
##  $ Diabetic        : chr  "Yes" "No" "No" "No, borderline diabetes" ...
##  $ PhysicalActivity: chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ GenHealth       : chr  "Fair" "Fair" "Very good" "Excellent" ...
##  $ SleepTime       : num  6 7 7 7 8 6 7 8 7 7 ...
##  $ Asthma          : chr  "No" "No" "No" "No" ...
##  $ KidneyDisease   : chr  "Yes" "No" "No" "No" ...
##  $ SkinCancer      : chr  "No" "Yes" "No" "No" ...
```

```
yes <- which(HDI$HeartDisease == "Yes")
no <- which(HDI$HeartDisease == "No")
length(yes)
```

```
## [1] 27373
```

```
length(no)
```

```
## [1] 27373
```

```
# Converting variables into factors, getting rid of unbalance Variables
HDI$AgeCategory[HDI$AgeCategory == "18-24"] <- 0
HDI$AgeCategory[HDI$AgeCategory == "25-29"] <- 1
HDI$AgeCategory[HDI$AgeCategory == "30-34"] <- 2
HDI$AgeCategory[HDI$AgeCategory == "35-39"] <- 3
HDI$AgeCategory[HDI$AgeCategory == "40-44"] <- 4
HDI$AgeCategory[HDI$AgeCategory == "45-49"] <- 5
HDI$AgeCategory[HDI$AgeCategory == "50-54"] <- 6
HDI$AgeCategory[HDI$AgeCategory == "55-59"] <- 7
HDI$AgeCategory[HDI$AgeCategory == "60=64"] <- 8
HDI$AgeCategory[HDI$AgeCategory == "65-69"] <- 9
HDI$AgeCategory[HDI$AgeCategory == "70-74"] <- 10
HDI$AgeCategory[HDI$AgeCategory == "75-79"] <- 11
HDI$AgeCategory[HDI$AgeCategory == "80 or older"] <- 12
HDI$AgeCategory <- as.factor(HDI$AgeCategory)

HDI$Diabetic[HDI$Diabetic == "Yes"] <- TRUE
HDI$Diabetic[HDI$Diabetic == "No"] <- FALSE
HDI$Diabetic[HDI$Diabetic == "Yes (during pregnancy)"] <- FALSE
HDI$Diabetic[HDI$Diabetic == "No, borderline diabetes"] <- TRUE
HDI$Diabetic <- as.factor(HDI$Diabetic)

HDI$DiffWalking[HDI$DiffWalking == "Yes"] <- TRUE
HDI$DiffWalking[HDI$DiffWalking == "No"] <- FALSE
HDI$DiffWalking <- as.factor(HDI$DiffWalking)



HDI$GenHealth[HDI$GenHealth == "Poor"] <- 0
```

```r
HDI$GenHealth[HDI$GenHealth == "Fair"] <- 1
HDI$GenHealth[HDI$GenHealth == "Good"] <- 2
HDI$GenHealth[HDI$GenHealth == "Very good"] <- 3
HDI$GenHealth[HDI$GenHealth == "Excellent"] <- 4
HDI$GenHealth <- as.factor(HDI$GenHealth)
HDI$GenHealth <- as.factor(HDI$GenHealth)


HDI$PhysicalActivity[HDI$PhysicalActivity == "Yes"] <- TRUE
HDI$PhysicalActivity[HDI$PhysicalActivity == "No"] <- FALSE
HDI$PhysicalActivity <- as.factor(HDI$PhysicalActivity)

HDI$Sex[HDI$Sex == "Male"] <- 0
HDI$Sex[HDI$Sex == "Female"] <- 1
HDI$Sex <- as.factor(HDI$Sex) # seems good

HDI$Smoking[HDI$Smoking == "Yes"] <- TRUE
HDI$Smoking[HDI$Smoking == "No"] <- FALSE
HDI$Smoking <- as.factor(HDI$Smoking) # seems good

HDI$AlcoholDrinking <- NULL
HDI$Stroke <- NULL
HDI$Race <- NULL
HDI$Asthma <- NULL
HDI$KidneyDisease <- NULL
HDI$SkinCancer <- NULL
HDI$MentalHealth <- NULL

colnames(HDI)[which(names(HDI) == "PhysicalHealth")] <- "InjuryRate"

names(HDI)
```

```
##  [1] "HeartDisease"     "BMI"              "Smoking"          "InjuryRate"
##  [5] "DiffWalking"      "Sex"              "AgeCategory"      "Diabetic"
##  [9] "PhysicalActivity" "GenHealth"        "SleepTime"
```

```r
# Graphs
summary(HDI)
```

```
##  HeartDisease            BMI         Smoking        InjuryRate
##  Length:54746       Min.   :12.16   FALSE:27817   Min.   : 0.000
##  Class :character   1st Qu.:24.41   TRUE :26929   1st Qu.: 0.000
##  Mode  :character   Median :27.80                 Median : 0.000
##                     Mean   :28.83                 Mean   : 5.385
##                     3rd Qu.:32.08                 3rd Qu.: 5.000
##                     Max.   :85.91                 Max.   :30.000
##
##  DiffWalking   Sex        AgeCategory     Diabetic     PhysicalActivity
##  FALSE:41540   0:28942   10     : 7286   FALSE:41427   FALSE:15635
##  TRUE :13206   1:25804   12     : 7205   TRUE :13319   TRUE :39111
```

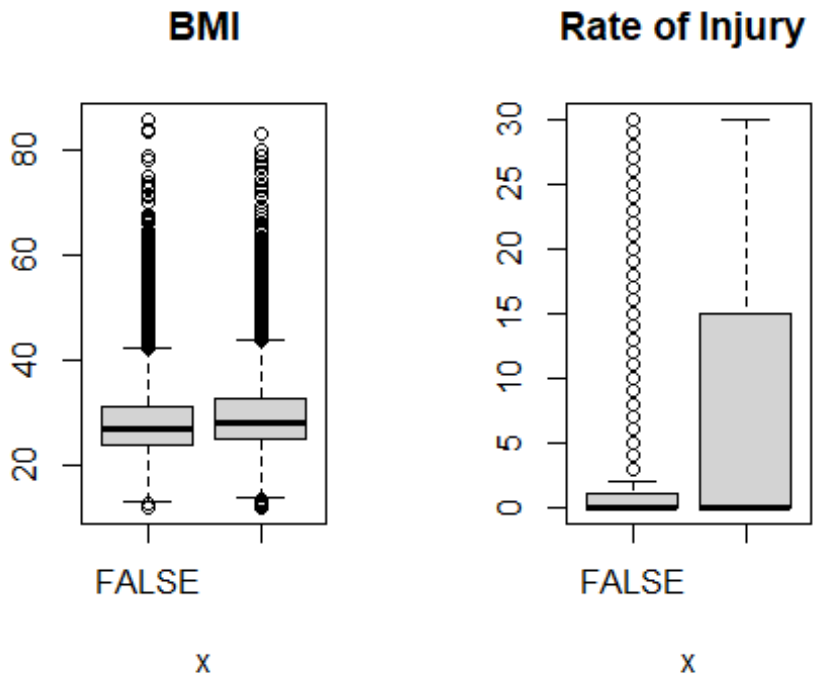
```
##                          9      : 6908
##                          60-64  : 6198
##                          11     : 5718
##                          7      : 4853
##                          (Other):16578
##  GenHealth   SleepTime
##  0: 4565   Min.   : 1.000
##  1: 9608   1st Qu.: 6.000
##  2:17454   Median : 7.000
##  3:15479   Mean   : 7.106
##  4: 7640   3rd Qu.: 8.000
##            Max.   :24.000
##
```

```
str(HDI)
```

```
## 'data.frame':    54746 obs. of  11 variables:
##  $ HeartDisease    : chr  "No" "No" "No" "No" ...
##  $ BMI             : num  30 22.7 21.9 29.8 30.2 ...
##  $ Smoking         : Factor w/ 2 levels "FALSE","TRUE": 1 2 1 2 2 1 1 1 2
2 ...
##  $ InjuryRate      : num  5 0 0 0 14 0 0 0 0 0 ...
##  $ DiffWalking     : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1
1 ...
##  $ Sex             : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 2 2 1 1 ...
##  $ AgeCategory     : Factor w/ 13 levels "0","1","10","11",..: 3 4 8 13 8
1 5 5 11 2 ...
##  $ Diabetic        : Factor w/ 2 levels "FALSE","TRUE": 2 1 1 2 1 1 2 1 1
1 ...
##  $ PhysicalActivity: Factor w/ 2 levels "FALSE","TRUE": 2 2 2 2 1 2 2 1 2
2 ...
##  $ GenHealth       : Factor w/ 5 levels "0","1","2","3",..: 2 2 4 5 5 4 4
3 3 5 ...
##  $ SleepTime       : num  6 7 7 7 8 6 7 8 7 7 ...
```

```
HDI$HeartDisease[HDI$HeartDisease == "Yes"] <- TRUE
HDI$HeartDisease[HDI$HeartDisease == "No"] <- FALSE
HDI$HeartDisease <- as.factor(HDI$HeartDisease)

par(mfrow=c(1,2))
plot(HDI$HeartDisease,HDI$BMI, main="BMI", ylab="", varwidth=TRUE)
plot(HDI$HeartDisease,HDI$InjuryRate, main="Rate of Injury ", ylab="",
varwidth=TRUE)
```




```
# Train and test Split
set.seed(1234)

x <- sample(1:nrow(HDI), nrow(HDI)*0.75, replace=FALSE)
train <- HDI[x,]
test <- HDI[-x,]
nrow(train)

## [1] 41059

nrow(test)

## [1] 13687

# Naive Bayes
library(e1071)
nb1 <- naiveBayes(HeartDisease~., data=train)
nb1

##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     FALSE      TRUE
```

```
## 0.4994033 0.5005967
##
## Conditional probabilities:
##        BMI
## Y           [,1]     [,2]
##   FALSE 28.23212 6.308781
##   TRUE  29.39794 6.587281
##
##        Smoking
## Y           FALSE      TRUE
##   FALSE 0.6023897 0.3976103
##   TRUE  0.4128637 0.5871363
##
##        InjuryRate
## Y           [,1]      [,2]
##   FALSE 2.952987  7.431252
##   TRUE  7.823051 11.504520
##
##        DiffWalking
## Y           FALSE      TRUE
##   FALSE 0.8831992 0.1168008
##   TRUE  0.6329668 0.3670332
##
##        Sex
## Y               0         1
##   FALSE 0.4677396 0.5322604
##   TRUE  0.5941909 0.4058091
##
##        AgeCategory
## Y                0           1           10           11           12
2
##   FALSE 0.069251402 0.056376494 0.089441600 0.060375518 0.063106559
0.064033163
##   TRUE  0.004670624 0.004719276 0.177143135 0.149362654 0.197090591
0.008076287
##        AgeCategory
## Y                3           4           5           6       60-64
7
##   FALSE 0.071348452 0.071884906 0.072031212 0.079248964 0.104218483
0.096854426
##   TRUE  0.010898122 0.017904058 0.027342610 0.050987642 0.122555220
0.079011385
##        AgeCategory
## Y                9
##   FALSE 0.101828822
##   TRUE  0.150238396
##
##        Diabetic
## Y           FALSE      TRUE
##   FALSE 0.8712997 0.1287003
```

```
##   TRUE  0.6420648 0.3579352
##
##        PhysicalActivity
## Y           FALSE      TRUE
##   FALSE 0.2088271 0.7911729
##   TRUE  0.3628004 0.6371996
##
##        GenHealth
## Y                0          1          2          3          4
##   FALSE 0.02643258 0.09280663 0.28792977 0.36888564 0.22394538
##   TRUE  0.14216211 0.25985210 0.34776686 0.19606889 0.05415004
##
##        SleepTime
## Y           [,1]     [,2]
##   FALSE 7.072031 1.392784
##   TRUE  7.132772 1.767472
```

```
nb.pred <- predict(nb1, newdata=test, type="class")
table(nb.pred, test$HeartDisease)
```

```
##
## nb.pred FALSE TRUE
##   FALSE  5390 2426
##   TRUE   1478 4393
```

```
nb.acc <- mean(nb.pred == test$HeartDisease)
print(paste("Accuracy: ", nb.acc))
```

```
## [1] "Accuracy:  0.71476583619493"
```

```
# kNN

library(class)

for (x in 1:ncol(HDI)){
    if(!is.numeric(HDI[1,x])) {
      HDI[,x] <- as.integer(HDI[,x])
    }
}

predictors <- c("BMI", "Smoking", "InjuryRate", "DiffWalking", "Sex",
"AgeCategory", "Diabetic", "PhysicalActivity", "GenHealth", "SleepTime")

normalize <- function(x) { (x - min(x))/(max(x) - min(x))}
HDI_normalized <- as.data.frame(lapply(HDI[,predictors], normalize))
summary(HDI_normalized)
```

```
##       BMI            Smoking         InjuryRate      DiffWalking
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1661   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.2121   Median :0.0000   Median :0.0000   Median :0.0000
```

```
##  Mean   :0.2260   Mean   :0.4919   Mean   :0.1795   Mean   :0.2412
##  3rd Qu.:0.2701   3rd Qu.:1.0000   3rd Qu.:0.1667   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       Sex          AgeCategory        Diabetic      PhysicalActivity
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.2500   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.5000   Median :0.0000   Median :1.0000
##  Mean   :0.4713   Mean   :0.5395   Mean   :0.2433   Mean   :0.7144
##  3rd Qu.:1.0000   3rd Qu.:0.8333   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##    GenHealth        SleepTime
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2500   1st Qu.:0.2174
##  Median :0.5000   Median :0.2609
##  Mean   :0.5549   Mean   :0.2655
##  3rd Qu.:0.7500   3rd Qu.:0.3043
##  Max.   :1.0000   Max.   :1.0000
```

```
set.seed(1234)
x <- sample(1:nrow(HDI_normalized), nrow(HDI_normalized)*0.75, replace=FALSE)
train <- HDI_normalized[x,]
test <- HDI_normalized[-x,]

train.labels <- HDI[x,"HeartDisease"]
test.labels <- HDI[-x,"HeartDisease"]

knn.pred <- knn(train, test, cl=train.labels, k=9)
results <- knn.pred == test.labels
knn.acc <- length(which(results == TRUE)) / length(results)
print(paste("Accuracy: ", knn.acc))
```

```
## [1] "Accuracy:  0.729378242127566"
```

```
table(results, knn.pred)
```

```
##        knn.pred
## results    1    2
##   FALSE 1568 2136
##   TRUE  4732 5251
```

```
#Train and test part 2
x <- sample(1:nrow(HDI), nrow(HDI)*0.75, replace=FALSE)
train <- HDI[x,]
test <- HDI[-x,]
nrow(train)
```

```
## [1] 41059
```

```
nrow(test)
```

```
## [1] 13687
```

```
# Logic Regression
# glm1 <- glm(HeartDisease~., data=train, family=binomial)
# summary(glm1)

# glm2 <-
glm(HeartDisease~Smoking+BMI+InjuryRate+Diabetic+GenHealth,data=train,
family="binomial")
# summary(glm2)

# glm3 <- glm(HeartDisease~.-AgeCategory-PhysicalActivity, data=train,
family="binomial")
# summary(glm3)

# glmprobs <- predict(glm1, newdata=test, type="response")

# glmpred <- rep(TRUE, nrow(test))
# glmpred[glmprobs<0.5] <- FALSE

# glmacc <- mean(glmpred == test$HeartDisease)
# print(glmacc)
# table(Predicted = glmpred, Actual = test$HeartDisease)
```

What Did I Learn:

That some variables and some algorithms need to have multiple instances of what is needed to be calculated for. Predictors have a hard time trying to predict values that need the result of other algorithms to work.