Final Project DSE6111

Adeline Casali

2023-12-05

# Section 1: What subset of predictors can be used for preventative health screening for heart disease?

**Loading data and packages**

heart\_df <- read.csv("Data/heart\_disease\_health\_indicators\_BRFSS2015.csv")  
library(corrplot)

## corrplot 0.92 loaded

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(MASS)

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(e1071)  
library(class)  
library(tree)  
library(randomForest)

## randomForest 4.7-1.1  
## Type rfNews() to see new features/changes/bug fixes.  
##   
## Attaching package: 'randomForest'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine  
##   
## The following object is masked from 'package:ggplot2':  
##   
## margin

library(gbm)

## Loaded gbm 2.1.8.1

library(kableExtra)

##   
## Attaching package: 'kableExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## group\_rows

**Data exploration and tidying**  
The only predictors that have a correlation stronger than 0.5 are physical health and general health with a correlation of 0.52.

# View the dataset and summary statistics  
head(heart\_df)

## HeartDiseaseorAttack HighBP HighChol CholCheck BMI Smoker Stroke Diabetes  
## 1 0 1 1 1 40 1 0 0  
## 2 0 0 0 0 25 1 0 0  
## 3 0 1 1 1 28 0 0 0  
## 4 0 1 0 1 27 0 0 0  
## 5 0 1 1 1 24 0 0 0  
## 6 0 1 1 1 25 1 0 0  
## PhysActivity Fruits Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost  
## 1 0 0 1 0 1 0  
## 2 1 0 0 0 0 1  
## 3 0 1 0 0 1 1  
## 4 1 1 1 0 1 0  
## 5 1 1 1 0 1 0  
## 6 1 1 1 0 1 0  
## GenHlth MentHlth PhysHlth DiffWalk Sex Age Education Income  
## 1 5 18 15 1 0 9 4 3  
## 2 3 0 0 0 0 7 6 1  
## 3 5 30 30 1 0 9 4 8  
## 4 2 0 0 0 0 11 3 6  
## 5 2 3 0 0 0 11 5 4  
## 6 2 0 2 0 1 10 6 8

summary(heart\_df)

## HeartDiseaseorAttack HighBP HighChol CholCheck   
## Min. :0.00000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:1.0000   
## Median :0.00000 Median :0.000 Median :0.0000 Median :1.0000   
## Mean :0.09419 Mean :0.429 Mean :0.4241 Mean :0.9627   
## 3rd Qu.:0.00000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.00000 Max. :1.000 Max. :1.0000 Max. :1.0000   
## BMI Smoker Stroke Diabetes   
## Min. :12.00 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:24.00 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000   
## Median :27.00 Median :0.0000 Median :0.00000 Median :0.0000   
## Mean :28.38 Mean :0.4432 Mean :0.04057 Mean :0.2969   
## 3rd Qu.:31.00 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:0.0000   
## Max. :98.00 Max. :1.0000 Max. :1.00000 Max. :2.0000   
## PhysActivity Fruits Veggies HvyAlcoholConsump  
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :1.0000 Median :1.0000 Median :1.0000 Median :0.0000   
## Mean :0.7565 Mean :0.6343 Mean :0.8114 Mean :0.0562   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## AnyHealthcare NoDocbcCost GenHlth MentHlth   
## Min. :0.0000 Min. :0.00000 Min. :1.000 Min. : 0.000   
## 1st Qu.:1.0000 1st Qu.:0.00000 1st Qu.:2.000 1st Qu.: 0.000   
## Median :1.0000 Median :0.00000 Median :2.000 Median : 0.000   
## Mean :0.9511 Mean :0.08418 Mean :2.511 Mean : 3.185   
## 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:3.000 3rd Qu.: 2.000   
## Max. :1.0000 Max. :1.00000 Max. :5.000 Max. :30.000   
## PhysHlth DiffWalk Sex Age   
## Min. : 0.000 Min. :0.0000 Min. :0.0000 Min. : 1.000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 6.000   
## Median : 0.000 Median :0.0000 Median :0.0000 Median : 8.000   
## Mean : 4.242 Mean :0.1682 Mean :0.4403 Mean : 8.032   
## 3rd Qu.: 3.000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:10.000   
## Max. :30.000 Max. :1.0000 Max. :1.0000 Max. :13.000   
## Education Income   
## Min. :1.00 Min. :1.000   
## 1st Qu.:4.00 1st Qu.:5.000   
## Median :5.00 Median :7.000   
## Mean :5.05 Mean :6.054   
## 3rd Qu.:6.00 3rd Qu.:8.000   
## Max. :6.00 Max. :8.000

str(heart\_df)

## 'data.frame': 253680 obs. of 22 variables:  
## $ HeartDiseaseorAttack: num 0 0 0 0 0 0 0 0 1 0 ...  
## $ HighBP : num 1 0 1 1 1 1 1 1 1 0 ...  
## $ HighChol : num 1 0 1 0 1 1 0 1 1 0 ...  
## $ CholCheck : num 1 0 1 1 1 1 1 1 1 1 ...  
## $ BMI : num 40 25 28 27 24 25 30 25 30 24 ...  
## $ Smoker : num 1 1 0 0 0 1 1 1 1 0 ...  
## $ Stroke : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Diabetes : num 0 0 0 0 0 0 0 0 2 0 ...  
## $ PhysActivity : num 0 1 0 1 1 1 0 1 0 0 ...  
## $ Fruits : num 0 0 1 1 1 1 0 0 1 0 ...  
## $ Veggies : num 1 0 0 1 1 1 0 1 1 1 ...  
## $ HvyAlcoholConsump : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AnyHealthcare : num 1 0 1 1 1 1 1 1 1 1 ...  
## $ NoDocbcCost : num 0 1 1 0 0 0 0 0 0 0 ...  
## $ GenHlth : num 5 3 5 2 2 2 3 3 5 2 ...  
## $ MentHlth : num 18 0 30 0 3 0 0 0 30 0 ...  
## $ PhysHlth : num 15 0 30 0 0 2 14 0 30 0 ...  
## $ DiffWalk : num 1 0 1 0 0 0 0 1 1 0 ...  
## $ Sex : num 0 0 0 0 0 1 0 0 0 1 ...  
## $ Age : num 9 7 9 11 11 10 9 11 9 8 ...  
## $ Education : num 4 6 4 3 5 6 6 4 5 4 ...  
## $ Income : num 3 1 8 6 4 8 7 4 1 3 ...

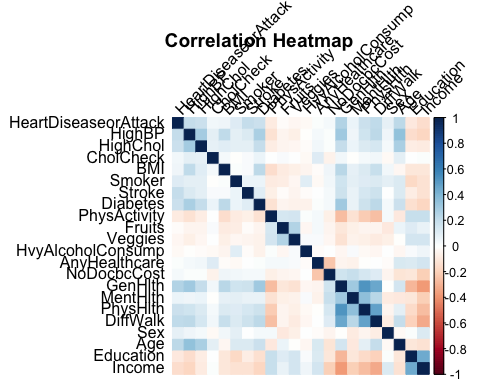
dim(heart\_df)

## [1] 253680 22

names(heart\_df)

## [1] "HeartDiseaseorAttack" "HighBP" "HighChol"   
## [4] "CholCheck" "BMI" "Smoker"   
## [7] "Stroke" "Diabetes" "PhysActivity"   
## [10] "Fruits" "Veggies" "HvyAlcoholConsump"   
## [13] "AnyHealthcare" "NoDocbcCost" "GenHlth"   
## [16] "MentHlth" "PhysHlth" "DiffWalk"   
## [19] "Sex" "Age" "Education"   
## [22] "Income"

# Check for correlated variables  
cor\_matrix <- cor(heart\_df)  
corrplot(cor\_matrix, method = "color", tl.col = "black", tl.srt = 45)  
title("Correlation Heatmap")



cor\_matrix <- cor(heart\_df)  
cor\_matrix\_filtered <- cor\_matrix  
cor\_matrix\_filtered[abs(cor\_matrix) <= 0.5] <- NA  
print(cor\_matrix\_filtered)

## HeartDiseaseorAttack HighBP HighChol CholCheck BMI Smoker  
## HeartDiseaseorAttack 1 NA NA NA NA NA  
## HighBP NA 1 NA NA NA NA  
## HighChol NA NA 1 NA NA NA  
## CholCheck NA NA NA 1 NA NA  
## BMI NA NA NA NA 1 NA  
## Smoker NA NA NA NA NA 1  
## Stroke NA NA NA NA NA NA  
## Diabetes NA NA NA NA NA NA  
## PhysActivity NA NA NA NA NA NA  
## Fruits NA NA NA NA NA NA  
## Veggies NA NA NA NA NA NA  
## HvyAlcoholConsump NA NA NA NA NA NA  
## AnyHealthcare NA NA NA NA NA NA  
## NoDocbcCost NA NA NA NA NA NA  
## GenHlth NA NA NA NA NA NA  
## MentHlth NA NA NA NA NA NA  
## PhysHlth NA NA NA NA NA NA  
## DiffWalk NA NA NA NA NA NA  
## Sex NA NA NA NA NA NA  
## Age NA NA NA NA NA NA  
## Education NA NA NA NA NA NA  
## Income NA NA NA NA NA NA  
## Stroke Diabetes PhysActivity Fruits Veggies  
## HeartDiseaseorAttack NA NA NA NA NA  
## HighBP NA NA NA NA NA  
## HighChol NA NA NA NA NA  
## CholCheck NA NA NA NA NA  
## BMI NA NA NA NA NA  
## Smoker NA NA NA NA NA  
## Stroke 1 NA NA NA NA  
## Diabetes NA 1 NA NA NA  
## PhysActivity NA NA 1 NA NA  
## Fruits NA NA NA 1 NA  
## Veggies NA NA NA NA 1  
## HvyAlcoholConsump NA NA NA NA NA  
## AnyHealthcare NA NA NA NA NA  
## NoDocbcCost NA NA NA NA NA  
## GenHlth NA NA NA NA NA  
## MentHlth NA NA NA NA NA  
## PhysHlth NA NA NA NA NA  
## DiffWalk NA NA NA NA NA  
## Sex NA NA NA NA NA  
## Age NA NA NA NA NA  
## Education NA NA NA NA NA  
## Income NA NA NA NA NA  
## HvyAlcoholConsump AnyHealthcare NoDocbcCost GenHlth  
## HeartDiseaseorAttack NA NA NA NA  
## HighBP NA NA NA NA  
## HighChol NA NA NA NA  
## CholCheck NA NA NA NA  
## BMI NA NA NA NA  
## Smoker NA NA NA NA  
## Stroke NA NA NA NA  
## Diabetes NA NA NA NA  
## PhysActivity NA NA NA NA  
## Fruits NA NA NA NA  
## Veggies NA NA NA NA  
## HvyAlcoholConsump 1 NA NA NA  
## AnyHealthcare NA 1 NA NA  
## NoDocbcCost NA NA 1 NA  
## GenHlth NA NA NA 1.0000000  
## MentHlth NA NA NA NA  
## PhysHlth NA NA NA 0.5243636  
## DiffWalk NA NA NA NA  
## Sex NA NA NA NA  
## Age NA NA NA NA  
## Education NA NA NA NA  
## Income NA NA NA NA  
## MentHlth PhysHlth DiffWalk Sex Age Education Income  
## HeartDiseaseorAttack NA NA NA NA NA NA NA  
## HighBP NA NA NA NA NA NA NA  
## HighChol NA NA NA NA NA NA NA  
## CholCheck NA NA NA NA NA NA NA  
## BMI NA NA NA NA NA NA NA  
## Smoker NA NA NA NA NA NA NA  
## Stroke NA NA NA NA NA NA NA  
## Diabetes NA NA NA NA NA NA NA  
## PhysActivity NA NA NA NA NA NA NA  
## Fruits NA NA NA NA NA NA NA  
## Veggies NA NA NA NA NA NA NA  
## HvyAlcoholConsump NA NA NA NA NA NA NA  
## AnyHealthcare NA NA NA NA NA NA NA  
## NoDocbcCost NA NA NA NA NA NA NA  
## GenHlth NA 0.5243636 NA NA NA NA NA  
## MentHlth 1 NA NA NA NA NA NA  
## PhysHlth NA 1.0000000 NA NA NA NA NA  
## DiffWalk NA NA 1 NA NA NA NA  
## Sex NA NA NA 1 NA NA NA  
## Age NA NA NA NA 1 NA NA  
## Education NA NA NA NA NA 1 NA  
## Income NA NA NA NA NA NA 1

# Convert HeartDiseaseorAttack to factor  
heart\_df$HeartDiseaseorAttack <- as.factor(heart\_df$HeartDiseaseorAttack)

**Logistic Regression**  
The linear model with all predictors shows that the variables HighBP, HighChol, CholCheck, Smoker, Stroke, Diabetes, HvyAlcoholConsump, NoDocbcCost, GenHlth, DiffWalk, Sex, Age, and Income were the most significant predictors, followed by PhysActivity, Veggies, and MentHlth. The model created with all of the significant predictors has an accuracy of 89.78%, and the model created with only the most significant predictors has the same accuracy of 89.78%.

# Null model  
heart\_lm\_null <- glm(HeartDiseaseorAttack ~ HeartDiseaseorAttack, data = heart\_df, family = binomial)

## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on  
## the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 1 in  
## model.matrix: no columns are assigned

summary(heart\_lm\_null)

##   
## Call:  
## glm(formula = HeartDiseaseorAttack ~ HeartDiseaseorAttack, family = binomial,   
## data = heart\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.263567 0.006797 -333 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 158355 on 253679 degrees of freedom  
## Residual deviance: 158355 on 253679 degrees of freedom  
## AIC: 158357  
##   
## Number of Fisher Scoring iterations: 5

# Model with all predictors  
heart\_lm <- glm(HeartDiseaseorAttack ~ ., data = heart\_df, family = binomial)  
summary(heart\_lm)

##   
## Call:  
## glm(formula = HeartDiseaseorAttack ~ ., family = binomial, data = heart\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.9124089 0.1028054 -76.965 < 2e-16 \*\*\*  
## HighBP 0.5245081 0.0177520 29.546 < 2e-16 \*\*\*  
## HighChol 0.6112677 0.0164496 37.160 < 2e-16 \*\*\*  
## CholCheck 0.5248111 0.0662510 7.922 2.35e-15 \*\*\*  
## BMI 0.0009744 0.0012122 0.804 0.4215   
## Smoker 0.3629524 0.0157323 23.071 < 2e-16 \*\*\*  
## Stroke 0.9783443 0.0244338 40.041 < 2e-16 \*\*\*  
## Diabetes 0.1465123 0.0089720 16.330 < 2e-16 \*\*\*  
## PhysActivity 0.0398975 0.0171991 2.320 0.0204 \*   
## Fruits 0.0060203 0.0163274 0.369 0.7123   
## Veggies 0.0426327 0.0189359 2.251 0.0244 \*   
## HvyAlcoholConsump -0.2940714 0.0392877 -7.485 7.15e-14 \*\*\*  
## AnyHealthcare -0.0070009 0.0412820 -0.170 0.8653   
## NoDocbcCost 0.2528463 0.0268896 9.403 < 2e-16 \*\*\*  
## GenHlth 0.4907058 0.0095105 51.596 < 2e-16 \*\*\*  
## MentHlth 0.0024628 0.0009778 2.519 0.0118 \*   
## PhysHlth 0.0010542 0.0008766 1.202 0.2292   
## DiffWalk 0.2947780 0.0193855 15.206 < 2e-16 \*\*\*  
## Sex 0.7611811 0.0160326 47.477 < 2e-16 \*\*\*  
## Age 0.2556493 0.0036439 70.158 < 2e-16 \*\*\*  
## Education 0.0110314 0.0081696 1.350 0.1769   
## Income -0.0431551 0.0042476 -10.160 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 158355 on 253679 degrees of freedom  
## Residual deviance: 120933 on 253658 degrees of freedom  
## AIC: 120977  
##   
## Number of Fisher Scoring iterations: 6

# Create a training and testing set  
set.seed(123)  
heart\_train <- heart\_df %>%   
 sample\_frac(0.7)  
heart\_test <- anti\_join(heart\_df, heart\_train)

## Joining with `by = join\_by(HeartDiseaseorAttack, HighBP, HighChol, CholCheck,  
## BMI, Smoker, Stroke, Diabetes, PhysActivity, Fruits, Veggies,  
## HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth,  
## DiffWalk, Sex, Age, Education, Income)`

# Model with significant predictors  
heart\_lm\_sig <- glm(HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck + Smoker + Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost + GenHlth + DiffWalk + Sex + Age + Income + PhysActivity + Veggies + MentHlth, data = heart\_train, family = binomial)  
summary(heart\_lm\_sig)

##   
## Call:  
## glm(formula = HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck +   
## Smoker + Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost +   
## GenHlth + DiffWalk + Sex + Age + Income + PhysActivity +   
## Veggies + MentHlth, family = binomial, data = heart\_train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.835845 0.101326 -77.333 < 2e-16 \*\*\*  
## HighBP 0.526391 0.020942 25.136 < 2e-16 \*\*\*  
## HighChol 0.623145 0.019618 31.764 < 2e-16 \*\*\*  
## CholCheck 0.543923 0.079393 6.851 7.33e-12 \*\*\*  
## Smoker 0.354107 0.018672 18.965 < 2e-16 \*\*\*  
## Stroke 0.946965 0.029093 32.549 < 2e-16 \*\*\*  
## Diabetes 0.142707 0.010519 13.567 < 2e-16 \*\*\*  
## HvyAlcoholConsump -0.284262 0.046573 -6.104 1.04e-09 \*\*\*  
## NoDocbcCost 0.234995 0.031695 7.414 1.22e-13 \*\*\*  
## GenHlth 0.495146 0.010365 47.773 < 2e-16 \*\*\*  
## DiffWalk 0.317315 0.022180 14.306 < 2e-16 \*\*\*  
## Sex 0.762415 0.019058 40.006 < 2e-16 \*\*\*  
## Age 0.252476 0.004166 60.610 < 2e-16 \*\*\*  
## Income -0.041475 0.004690 -8.843 < 2e-16 \*\*\*  
## PhysActivity 0.038610 0.020277 1.904 0.0569 .   
## Veggies 0.038037 0.021958 1.732 0.0832 .   
## MentHlth 0.003395 0.001132 3.000 0.0027 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 111242 on 177575 degrees of freedom  
## Residual deviance: 84942 on 177559 degrees of freedom  
## AIC: 84976  
##   
## Number of Fisher Scoring iterations: 6

# Test the model and calculate accuracy  
predict\_lm <- predict(heart\_lm\_sig, newdata = heart\_test)  
binary\_predict\_lm <- ifelse(predict\_lm > 0.5, 1, 0)  
results <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,   
 Predicted = binary\_predict\_lm  
)  
results$Correct <- results$Actual == results$Predicted  
confusion\_matrix\_lm <- table(Predicted = results$Predicted, Actual = results$Actual)  
print(confusion\_matrix\_lm)

## Actual  
## Predicted 0 1  
## 0 59853 6623  
## 1 234 391

accuracy\_lm <- (59853 + 391) / (59853 + 6623 + 234 + 391)  
error\_lm <- 1 - accuracy\_lm  
cat("Accuracy:", accuracy\_lm, "\n")

## Accuracy: 0.8978108

cat("Error Rate:", error\_lm, "\n")

## Error Rate: 0.1021892

# Model with only the most significant predictors  
heart\_lm\_sig2 <- glm(HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck + Smoker + Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost + GenHlth + DiffWalk + Sex + Age + Income, data = heart\_train, family = binomial)  
summary(heart\_lm\_sig2)

##   
## Call:  
## glm(formula = HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck +   
## Smoker + Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost +   
## GenHlth + DiffWalk + Sex + Age + Income, family = binomial,   
## data = heart\_train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.758065 0.098684 -78.615 < 2e-16 \*\*\*  
## HighBP 0.524694 0.020937 25.060 < 2e-16 \*\*\*  
## HighChol 0.625225 0.019603 31.894 < 2e-16 \*\*\*  
## CholCheck 0.546769 0.079366 6.889 5.61e-12 \*\*\*  
## Smoker 0.355084 0.018655 19.034 < 2e-16 \*\*\*  
## Stroke 0.947827 0.029083 32.590 < 2e-16 \*\*\*  
## Diabetes 0.141732 0.010513 13.482 < 2e-16 \*\*\*  
## HvyAlcoholConsump -0.280701 0.046548 -6.030 1.64e-09 \*\*\*  
## NoDocbcCost 0.244870 0.031503 7.773 7.67e-15 \*\*\*  
## GenHlth 0.498179 0.010072 49.462 < 2e-16 \*\*\*  
## DiffWalk 0.317142 0.021823 14.532 < 2e-16 \*\*\*  
## Sex 0.758174 0.018949 40.012 < 2e-16 \*\*\*  
## Age 0.250268 0.004088 61.223 < 2e-16 \*\*\*  
## Income -0.041128 0.004639 -8.867 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 111242 on 177575 degrees of freedom  
## Residual deviance: 84958 on 177562 degrees of freedom  
## AIC: 84986  
##   
## Number of Fisher Scoring iterations: 6

# Test the second model and calculate accuracy  
predict\_lm2 <- predict(heart\_lm\_sig2, newdata = heart\_test)  
binary\_predict\_lm2 <- ifelse(predict\_lm2 > 0.5, 1, 0)  
results2 <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,   
 Predicted = binary\_predict\_lm2  
)  
results2$Correct <- results2$Actual == results2$Predicted  
confusion\_matrix\_lm2 <- table(Predicted = results2$Predicted, Actual = results2$Actual)  
print(confusion\_matrix\_lm2)

## Actual  
## Predicted 0 1  
## 0 59847 6618  
## 1 240 396

accuracy\_lm2 <- (59847 + 396) / (59847 + 6618 + 240 + 396)  
error\_lm2 <- 1 - accuracy\_lm2  
cat("Accuracy:", accuracy\_lm2, "\n")

## Accuracy: 0.8977959

cat("Error Rate:", error\_lm2, "\n")

## Error Rate: 0.1022041

**LDA**  
The LDA model containing all significant predictors has an accuracy of 89.14%.

# Model with significant predictors  
heart\_lda <- lda(HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck + Smoker + Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost + GenHlth + DiffWalk + Sex + Age + Income + PhysActivity + Veggies + MentHlth, data = heart\_train)  
heart\_lda

## Call:  
## lda(HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck + Smoker +   
## Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost + GenHlth +   
## DiffWalk + Sex + Age + Income + PhysActivity + Veggies +   
## MentHlth, data = heart\_train)  
##   
## Prior probabilities of groups:  
## 0 1   
## 0.90532504 0.09467496   
##   
## Group means:  
## HighBP HighChol CholCheck Smoker Stroke Diabetes  
## 0 0.3949205 0.3955799 0.9595059 0.4254684 0.02793536 0.2554614  
## 1 0.7501784 0.7031882 0.9888175 0.6190221 0.16404949 0.6846895  
## HvyAlcoholConsump NoDocbcCost GenHlth DiffWalk Sex Age Income  
## 0 0.05785499 0.08118733 2.420691 0.1427496 0.4269177 7.813808 6.147091  
## 1 0.03598620 0.10908875 3.372175 0.4182132 0.5739353 10.118011 5.143053  
## PhysActivity Veggies MentHlth  
## 0 0.7689532 0.8162400 3.038952  
## 1 0.6411492 0.7635023 4.732512  
##   
## Coefficients of linear discriminants:  
## LD1  
## HighBP 0.2934865419  
## HighChol 0.3740542652  
## CholCheck 0.1465449354  
## Smoker 0.2194045131  
## Stroke 1.7900323795  
## Diabetes 0.2102071299  
## HvyAlcoholConsump -0.1675486525  
## NoDocbcCost 0.0514054892  
## GenHlth 0.3356617796  
## DiffWalk 0.5344940710  
## Sex 0.5037227569  
## Age 0.1109019681  
## Income -0.0317635706  
## PhysActivity 0.0456651232  
## Veggies 0.0465340124  
## MentHlth 0.0009576617

# Test the model and calculate accuracy  
lda\_predictions <- predict(heart\_lda, newdata = heart\_test)$class  
lda\_results <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,  
 Predicted = lda\_predictions  
)  
lda\_results$Correct <- lda\_results$Actual == lda\_results$Predicted  
lda\_confusion\_matrix <- table(Predicted = lda\_results$Predicted, Actual = lda\_results$Actual)  
print(lda\_confusion\_matrix)

## Actual  
## Predicted 0 1  
## 0 58497 5694  
## 1 1590 1320

accuracy\_lda <- (58497 + 1320) / (58497 + 5694 + 1590 + 1320)  
error\_lda <- 1 - accuracy\_lda  
cat("Accuracy:", accuracy\_lda, "\n")

## Accuracy: 0.8914472

cat("Error Rate:", error\_lda, "\n")

## Error Rate: 0.1085528

**QDA**  
The QDA model containing all significant predictors has an accuracy of 83.25%.

# Model with significant predictors  
heart\_qda <- qda(HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck + Smoker + Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost + GenHlth + DiffWalk + Sex + Age + Income + PhysActivity + Veggies + MentHlth, data = heart\_train)  
heart\_qda

## Call:  
## qda(HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck + Smoker +   
## Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost + GenHlth +   
## DiffWalk + Sex + Age + Income + PhysActivity + Veggies +   
## MentHlth, data = heart\_train)  
##   
## Prior probabilities of groups:  
## 0 1   
## 0.90532504 0.09467496   
##   
## Group means:  
## HighBP HighChol CholCheck Smoker Stroke Diabetes  
## 0 0.3949205 0.3955799 0.9595059 0.4254684 0.02793536 0.2554614  
## 1 0.7501784 0.7031882 0.9888175 0.6190221 0.16404949 0.6846895  
## HvyAlcoholConsump NoDocbcCost GenHlth DiffWalk Sex Age Income  
## 0 0.05785499 0.08118733 2.420691 0.1427496 0.4269177 7.813808 6.147091  
## 1 0.03598620 0.10908875 3.372175 0.4182132 0.5739353 10.118011 5.143053  
## PhysActivity Veggies MentHlth  
## 0 0.7689532 0.8162400 3.038952  
## 1 0.6411492 0.7635023 4.732512

# Test the model and calculate accuracy  
qda\_predictions <- predict(heart\_qda, newdata = heart\_test)$class  
qda\_results <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,  
 Predicted = qda\_predictions  
)  
qda\_results$Correct <- qda\_results$Actual == qda\_results$Predicted  
qda\_confusion\_matrix <- table(Predicted = qda\_results$Predicted, Actual = qda\_results$Actual)  
print(qda\_confusion\_matrix)

## Actual  
## Predicted 0 1  
## 0 52651 3802  
## 1 7436 3212

accuracy\_qda <- (52651 + 3212) / (52651 + 3802 + 7436 + 3212)  
error\_qda <- 1 - accuracy\_qda  
cat("Accuracy:", accuracy\_qda, "\n")

## Accuracy: 0.8325211

cat("Error Rate:", error\_qda, "\n")

## Error Rate: 0.1674789

**Naive Bayes**  
The Naive Bayes model containing all significant predictors has an accuracy of 80.39%.

# Model with significant predictors  
heart\_nb <- naiveBayes(HeartDiseaseorAttack ~ HighBP + HighChol + CholCheck + Smoker + Stroke + Diabetes + HvyAlcoholConsump + NoDocbcCost + GenHlth + DiffWalk + Sex + Age + Income + PhysActivity + Veggies + MentHlth, data = heart\_train)  
heart\_nb

##   
## Naive Bayes Classifier for Discrete Predictors  
##   
## Call:  
## naiveBayes.default(x = X, y = Y, laplace = laplace)  
##   
## A-priori probabilities:  
## Y  
## 0 1   
## 0.90532504 0.09467496   
##   
## Conditional probabilities:  
## HighBP  
## Y [,1] [,2]  
## 0 0.3949205 0.4888351  
## 1 0.7501784 0.4329225  
##   
## HighChol  
## Y [,1] [,2]  
## 0 0.3955799 0.4889764  
## 1 0.7031882 0.4568665  
##   
## CholCheck  
## Y [,1] [,2]  
## 0 0.9595059 0.1971157  
## 1 0.9888175 0.1051575  
##   
## Smoker  
## Y [,1] [,2]  
## 0 0.4254684 0.4944154  
## 1 0.6190221 0.4856416  
##   
## Stroke  
## Y [,1] [,2]  
## 0 0.02793536 0.1647882  
## 1 0.16404949 0.3703315  
##   
## Diabetes  
## Y [,1] [,2]  
## 0 0.2554614 0.6548011  
## 1 0.6846895 0.9338205  
##   
## HvyAlcoholConsump  
## Y [,1] [,2]  
## 0 0.05785499 0.2334698  
## 1 0.03598620 0.1862613  
##   
## NoDocbcCost  
## Y [,1] [,2]  
## 0 0.08118733 0.2731234  
## 1 0.10908875 0.3117598  
##   
## GenHlth  
## Y [,1] [,2]  
## 0 2.420691 1.025722  
## 1 3.372175 1.085493  
##   
## DiffWalk  
## Y [,1] [,2]  
## 0 0.1427496 0.3498184  
## 1 0.4182132 0.4932802  
##   
## Sex  
## Y [,1] [,2]  
## 0 0.4269177 0.4946317  
## 1 0.5739353 0.4945181  
##   
## Age  
## Y [,1] [,2]  
## 0 7.813808 3.046982  
## 1 10.118011 2.237037  
##   
## Income  
## Y [,1] [,2]  
## 0 6.147091 2.034289  
## 1 5.143053 2.207500  
##   
## PhysActivity  
## Y [,1] [,2]  
## 0 0.7689532 0.4215036  
## 1 0.6411492 0.4796776  
##   
## Veggies  
## Y [,1] [,2]  
## 0 0.8162400 0.3872896  
## 1 0.7635023 0.4249439  
##   
## MentHlth  
## Y [,1] [,2]  
## 0 3.038952 7.196951  
## 1 4.732512 9.251938

# Test the model and calculate accuracy  
nb\_predictions <- predict(heart\_nb, newdata = heart\_test)  
nb\_results <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,  
 Predicted = nb\_predictions  
)  
nb\_results$Correct <- nb\_results$Actual == nb\_results$Predicted  
nb\_confusion\_matrix <- table(Predicted = nb\_results$Predicted, Actual = nb\_results$Actual)  
print(nb\_confusion\_matrix)

## Actual  
## Predicted 0 1  
## 0 50161 3234  
## 1 9926 3780

accuracy\_nb <- (50161 + 3780) / (50161 + 3234 + 9926 + 3780)  
error\_nb <- 1 - accuracy\_nb  
cat("Accuracy:", accuracy\_nb, "\n")

## Accuracy: 0.8038777

cat("Error Rate:", error\_nb, "\n")

## Error Rate: 0.1961223

**KNN (K-Nearest Neighbors)**  
The KNN model containing all significant predictors and K=3 has an accuracy of 98.13%.

# Model with all significant predictors and K = 3  
predictors <- c(  
 "HighBP", "HighChol", "CholCheck", "Smoker", "Stroke", "Diabetes",  
 "HvyAlcoholConsump", "NoDocbcCost", "GenHlth", "DiffWalk",  
 "Sex", "Age", "Income", "PhysActivity", "Veggies", "MentHlth"  
)  
train\_data <- heart\_train[, c(predictors, "HeartDiseaseorAttack")]  
test\_data <- heart\_test[, c(predictors, "HeartDiseaseorAttack")]  
knn\_predictions <- knn(  
 train = train\_data[, -length(predictors)],  
 test = test\_data[, -length(predictors)],  
 cl = train\_data$HeartDiseaseorAttack,  
 k = 3  
)  
knn\_results <- data.frame(  
 Actual = test\_data$HeartDiseaseorAttack,  
 Predicted = knn\_predictions  
)  
knn\_results$Correct <- knn\_results$Actual == knn\_results$Predicted  
knn\_confusion\_matrix <- table(Predicted = knn\_results$Predicted, Actual = knn\_results$Actual)  
print(knn\_confusion\_matrix)

## Actual  
## Predicted 0 1  
## 0 60081 1245  
## 1 6 5769

accuracy\_knn <- (60082 + 5763) / (60082 + 1251 + 5 + 5763)  
error\_knn <- 1 - accuracy\_knn  
cat("Accuracy:", accuracy\_knn, "\n")

## Accuracy: 0.9812819

cat("Error Rate:", error\_knn, "\n")

## Error Rate: 0.01871805

# Model with all significant predictors and K = 5  
knn\_predictions2 <- knn(  
 train = train\_data[, -length(predictors)],  
 test = test\_data[, -length(predictors)],  
 cl = train\_data$HeartDiseaseorAttack,  
 k = 5  
)  
knn\_results2 <- data.frame(  
 Actual = test\_data$HeartDiseaseorAttack,  
 Predicted = knn\_predictions2  
)  
knn\_results2$Correct <- knn\_results2$Actual == knn\_results2$Predicted  
knn\_confusion\_matrix2 <- table(Predicted = knn\_results2$Predicted, Actual = knn\_results2$Actual)  
print(knn\_confusion\_matrix2)

## Actual  
## Predicted 0 1  
## 0 60086 1419  
## 1 1 5595

accuracy\_knn2 <- (60085 + 5606) / (60085 + 1408 + 2 + 5606)  
error\_knn2 <- 1 - accuracy\_knn2  
cat("Accuracy:", accuracy\_knn2, "\n")

## Accuracy: 0.9789869

cat("Error Rate:", error\_knn2, "\n")

## Error Rate: 0.0210131

# Model with all significant predictors and K = 10  
knn\_predictions3 <- knn(  
 train = train\_data[, -length(predictors)],  
 test = test\_data[, -length(predictors)],  
 cl = train\_data$HeartDiseaseorAttack,  
 k = 10  
)  
knn\_results3 <- data.frame(  
 Actual = test\_data$HeartDiseaseorAttack,  
 Predicted = knn\_predictions3  
)  
knn\_results3$Correct <- knn\_results3$Actual == knn\_results3$Predicted  
knn\_confusion\_matrix3 <- table(Predicted = knn\_results3$Predicted, Actual = knn\_results3$Actual)  
print(knn\_confusion\_matrix3)

## Actual  
## Predicted 0 1  
## 0 60087 1679  
## 1 0 5335

accuracy\_knn3 <- (60087 + 5340) / (60087 + 1674 + 0 + 5340)  
error\_knn3 <- 1 - accuracy\_knn3  
cat("Accuracy:", accuracy\_knn3, "\n")

## Accuracy: 0.9750525

cat("Error Rate:", error\_knn3, "\n")

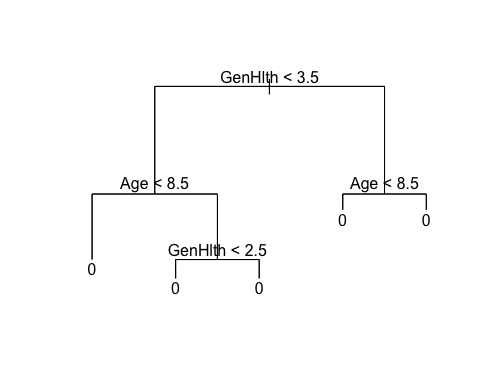
## Error Rate: 0.02494747

**Classification Trees**  
The classification tree containing all predictors and pruned with cross-validation has an accuracy of 89.55%.

# Model with all predictors  
heart\_tree <- tree(HeartDiseaseorAttack ~ ., heart\_train)  
summary(heart\_tree)

##   
## Classification tree:  
## tree(formula = HeartDiseaseorAttack ~ ., data = heart\_train)  
## Variables actually used in tree construction:  
## [1] "GenHlth" "Age"   
## Number of terminal nodes: 5   
## Residual mean deviance: 0.5353 = 95050 / 177600   
## Misclassification error rate: 0.09467 = 16812 / 177576

plot(heart\_tree)  
text(heart\_tree, pretty = 0)



# Test the model and calculate accuracy  
tree\_predict <- predict(heart\_tree, heart\_test, type = "class")  
tree\_results <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,  
 Predicted = tree\_predict  
)  
tree\_results$Correct <- tree\_results$Actual == tree\_results$Predicted  
tree\_confusion\_matrix <- table(Predicted = tree\_results$Predicted, Actual = tree\_results$Actual)  
print(tree\_confusion\_matrix)

## Actual  
## Predicted 0 1  
## 0 60087 7014  
## 1 0 0

accuracy\_tree <- (60087 + 0) / (60087 + 7014 + 0 + 0)  
error\_tree <- 1 - accuracy\_tree  
cat("Accuracy:", accuracy\_tree, "\n")

## Accuracy: 0.895471

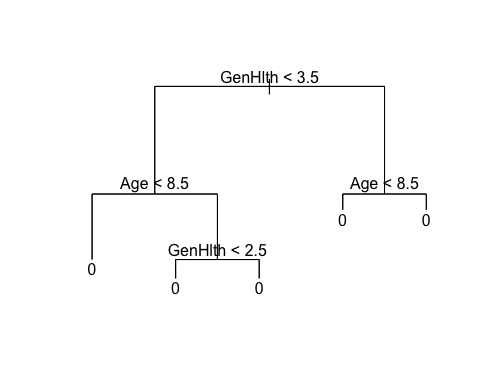
cat("Error Rate:", error\_tree, "\n")

## Error Rate: 0.104529

# Prune the tree with cross-validation  
set.seed(123)  
cv\_heart <- cv.tree(heart\_tree, FUN = prune.misclass)  
cv\_heart

## $size  
## [1] 5 1  
##   
## $dev  
## [1] 16812 16812  
##   
## $k  
## [1] -Inf 0  
##   
## $method  
## [1] "misclass"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"

prune\_heart <- prune.misclass(heart\_tree, best = 5)  
plot(prune\_heart)  
text(prune\_heart, pretty = 0)



# Test the pruned tree and calculate accuracy  
tree\_predict2 <- predict(prune\_heart, heart\_test, type = "class")  
tree\_results2 <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,  
 Predicted = tree\_predict2  
)  
tree\_results2$Correct <- tree\_results2$Actual == tree\_results2$Predicted  
tree\_confusion\_matrix2 <- table(Predicted = tree\_results2$Predicted, Actual = tree\_results2$Actual)  
print(tree\_confusion\_matrix2)

## Actual  
## Predicted 0 1  
## 0 60087 7014  
## 1 0 0

accuracy\_tree2 <- (60087 + 0) / (60087 + 7014 + 0 + 0)  
error\_tree2 <- 1 - accuracy\_tree2  
cat("Accuracy:", accuracy\_tree2, "\n")

## Accuracy: 0.895471

cat("Error Rate:", error\_tree2, "\n")

## Error Rate: 0.104529

# The pruned and unpruned trees are the same

**Bagging**  
The bagged classification tree containing all predictors has an accuracy of 87.75%. The most important predictors in the tree are (1) Age, (2) BMI, and (3) Income.

# Model with all predictors  
set.seed(123)  
heart\_bag <- randomForest(HeartDiseaseorAttack ~ ., data = heart\_train, mtry = 16, importance = TRUE, ntree = 25)  
heart\_bag

##   
## Call:  
## randomForest(formula = HeartDiseaseorAttack ~ ., data = heart\_train, mtry = 16, importance = TRUE, ntree = 25)   
## Type of random forest: classification  
## Number of trees: 25  
## No. of variables tried at each split: 16  
##   
## OOB estimate of error rate: 10.92%  
## Confusion matrix:  
## 0 1 class.error  
## 0 155100 5661 0.03521376  
## 1 13724 3088 0.81632167

importance(heart\_bag)

## 0 1 MeanDecreaseAccuracy MeanDecreaseGini  
## HighBP -9.453797 19.8925326 -0.3090534 631.2097  
## HighChol -12.549170 26.4123276 -3.2835389 564.9037  
## CholCheck 5.092407 -2.6031519 4.1494782 126.9510  
## BMI 7.533636 -3.8052444 4.9277909 5530.3945  
## Smoker -3.230946 7.6994319 0.4178417 869.6693  
## Stroke 10.338214 32.1654268 22.6743155 715.7023  
## Diabetes 3.808568 8.6099910 7.6360371 957.7022  
## PhysActivity 6.080891 -4.6828963 3.4617302 956.7861  
## Fruits 5.676729 -1.1941288 4.0209661 1082.7540  
## Veggies 2.003716 -0.1069718 1.8029961 846.1092  
## HvyAlcoholConsump 2.437638 -0.6126569 1.6702023 324.2037  
## AnyHealthcare 4.009127 -2.8926022 2.7306862 232.6367  
## NoDocbcCost 8.240178 -0.2863122 6.9116860 450.2840  
## GenHlth 16.273986 17.8019342 26.9336732 2279.5800  
## MentHlth 10.310916 -3.6418811 7.3903003 1950.9674  
## PhysHlth 14.775446 -2.4436951 14.2738733 2535.3968  
## DiffWalk 1.541262 12.2072850 9.2774139 860.0739  
## Sex 7.924247 11.9507307 13.6468996 645.5515  
## Age 11.125472 34.9689543 24.5523960 2777.3193  
## Education 19.012976 -6.2168920 14.6522430 2207.2180  
## Income 14.589597 -2.0897785 13.9938751 3157.4731

# Test the model and calculate accuracy  
tree\_predict\_bag <- predict(heart\_bag, heart\_test, type = "class")  
tree\_results\_bag <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,  
 Predicted = tree\_predict\_bag  
)  
tree\_results\_bag$Correct <- tree\_results\_bag$Actual == tree\_results\_bag$Predicted  
tree\_confusion\_matrix\_bag <- table(Predicted = tree\_results\_bag$Predicted, Actual = tree\_results\_bag$Actual)  
print(tree\_confusion\_matrix\_bag)

## Actual  
## Predicted 0 1  
## 0 58207 5894  
## 1 1880 1120

accuracy\_tree\_bag <- (57539 + 1341) / (57539 + 5673 + 2548 + 1341)  
error\_tree\_bag <- 1 - accuracy\_tree\_bag  
cat("Accuracy:", accuracy\_tree\_bag, "\n")

## Accuracy: 0.8774832

cat("Error Rate:", error\_tree\_bag, "\n")

## Error Rate: 0.1225168

**Random Forests**  
The random forest classification tree containing all predictors has an accuracy of 89.45%. The most important predictors in the tree are (1) Age, (2) BMI, and (3) GenHlth.

# Model with all predictors  
set.seed(123)  
heart\_rf <- randomForest(HeartDiseaseorAttack ~ ., data = heart\_train, mtry = 4, importance = TRUE, ntree = 25)  
heart\_rf

##   
## Call:  
## randomForest(formula = HeartDiseaseorAttack ~ ., data = heart\_train, mtry = 4, importance = TRUE, ntree = 25)   
## Type of random forest: classification  
## Number of trees: 25  
## No. of variables tried at each split: 4  
##   
## OOB estimate of error rate: 9.84%  
## Confusion matrix:  
## 0 1 class.error  
## 0 158101 2661 0.01655242  
## 1 14810 2001 0.88097079

importance(heart\_rf)

## 0 1 MeanDecreaseAccuracy MeanDecreaseGini  
## HighBP -9.6305526 12.5565119 0.6843860 809.81440  
## HighChol -10.3694707 14.8229002 0.9644538 679.93561  
## CholCheck 2.6103315 -0.4093466 2.5155214 97.11157  
## BMI 12.9636999 -2.9705478 9.8569128 3263.23924  
## Smoker -9.5435946 9.6846098 -0.3882864 502.72400  
## Stroke 5.6253149 39.7938864 18.8813333 763.19208  
## Diabetes -0.6267632 9.1181822 3.6660343 673.13682  
## PhysActivity 4.7218669 -2.5749532 2.2937435 620.59727  
## Fruits 3.8365262 -1.7856203 2.1357503 668.07655  
## Veggies 2.0719579 -2.0302961 1.0935410 591.07056  
## HvyAlcoholConsump 2.6230475 0.3989772 2.9316309 209.55416  
## AnyHealthcare 4.4841208 -0.5483266 4.0833164 182.03580  
## NoDocbcCost 5.4151921 -2.0114670 3.8731648 354.61536  
## GenHlth 7.0065827 26.9871170 16.3452153 1762.41789  
## MentHlth 12.8490718 -7.6722303 10.5400008 1425.18542  
## PhysHlth 10.0921953 -1.8453169 9.8522780 1872.28068  
## DiffWalk 1.4581647 12.2251124 9.6233747 682.55238  
## Sex 9.2676631 10.2017891 11.2873441 595.04416  
## Age 6.4873333 19.8988746 16.6811849 2487.90392  
## Education 11.2803123 -3.7764568 8.2815047 1442.37075  
## Income 6.3686110 2.1722698 7.6344434 1940.88147

# Test the model and calculate accuracy  
tree\_predict\_rf <- predict(heart\_rf, heart\_test, type = "class")  
tree\_results\_rf <- data.frame(  
 Actual = heart\_test$HeartDiseaseorAttack,  
 Predicted = tree\_predict\_rf  
)  
tree\_results\_rf$Correct <- tree\_results\_rf$Actual == tree\_results\_rf$Predicted  
tree\_confusion\_matrix\_rf <- table(Predicted = tree\_results\_rf$Predicted, Actual = tree\_results\_rf$Actual)  
print(tree\_confusion\_matrix\_rf)

## Actual  
## Predicted 0 1  
## 0 59362 6335  
## 1 725 679

accuracy\_tree\_rf <- (59362 + 679) / (59362 + 6335 + 725 + 679)  
error\_tree\_rf <- 1 - accuracy\_tree\_rf  
cat("Accuracy:", accuracy\_tree\_rf, "\n")

## Accuracy: 0.8947855

cat("Error Rate:", error\_tree\_rf, "\n")

## Error Rate: 0.1052145

**Overview of models**  
Overall, the model with the highest accuracy is kNN with k=3. This model has an accuracy of 98.13%.

# Table with models and relative accuracies  
classification\_overview <- data.frame(  
 Method = c("Logistic Regression", "LDA", "QDA", "Naive Bayes", "kNN (k=3)", "CV Tree", "Bagging", "Random Forest"),  
 Accuracy = c("89.78%", "89.14%", "83.25%", "80.39%", "98.13%", "89.55%", "87.75%", "89.45%")  
)  
classification\_table <- kable(classification\_overview, "markdown") %>%  
 kable\_styling(full\_width = FALSE) %>%  
 column\_spec(1, bold = TRUE)

## Warning in kable\_styling(., full\_width = FALSE): Please specify format in  
## kable. kableExtra can customize either HTML or LaTeX outputs. See  
## https://haozhu233.github.io/kableExtra/ for details.

## Warning in column\_spec(., 1, bold = TRUE): Please specify format in kable.  
## kableExtra can customize either HTML or LaTeX outputs. See  
## https://haozhu233.github.io/kableExtra/ for details.

classification\_table

| Method | Accuracy |
| --- | --- |
| Logistic Regression | 89.78% |
| LDA | 89.14% |
| QDA | 83.25% |
| Naive Bayes | 80.39% |
| kNN (k=3) | 98.13% |
| CV Tree | 89.55% |
| Bagging | 87.75% |
| Random Forest | 89.45% |

# Section 2: What indicators can be used to monitor for HIV response in children and adolescents?