Classification

Umar Ali-Salaam

9/25/2022

Source: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

This is a dataset based off of 70,000 records of patient data (Heart Related). Columns (13): ID, Age, Height(cm), Weight(kg), Gender, Systolic Blood Pressure (AP__HIGH), Diastolic Blood Pressure (AP LOW), Cholesterol, Glucose, Smoking, Alcohol Intake, Physical Activity, Presence or Absence of cardiovascular disease.

The .csv file needed to be edited a bit in Microsoft Excel before using it in R. I just performed a split column delimiter function around semicolons, to divide the singular column that existed into 13. Each row had 13 variables in 1 column separated by semicolons, the function I ran split it up into 13 columns, making a 70,000 x 13 table.

https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Visit the website above to better understand Systolic and Diastolic Blood Pressure

```
library(naivebayes)
```

```
## Warning: package 'naivebayes' was built under R version 4.1.3
```

```
## naivebayes 0.9.7 loaded
```

```
library(ggplot2)
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
```

```
# Read in .csv file
heart <- read.csv("cardio_train.csv")


# Clean out any rows that have an unrealistic blood pressure (AP_HIGH & AP_LOW)
```

```r
# They looked to be input errors by the person who made the data set
s <- subset(heart, AP_HIGH > 50)

s1 <- subset(s, AP_HIGH < 200)

s2 <- subset(s1, AP_LOW > 25)

s3 <- subset(s2, AP_LOW < 200)


# Removing little people(4'10") and Giants(7'3"), values are in cm
s4 <- subset(s3, HEIGHT > 147)

s5 <- subset(s4, HEIGHT < 220)


# Removing anyone below 90 lbs and above 375 lbs, the values are in kg
s6 <- subset(s5, WEIGHT > 40)

h1 <- subset(s6, WEIGHT < 180)

# AGE is in days so to get years i just divide by 365
h1$AGE <- (h1$AGE / 365)

# Removing people under 40
h <- subset(h1, AGE > 39)

# Checking for any NA values
# There is none
colSums(is.na(h))
```

```
##                ID               AGE            GENDER            HEIGHT
##                 0                 0                 0                 0
##            WEIGHT           AP_HIGH            AP_LOW       CHOLESTEROL
##                 0                 0                 0                 0
##           GLUCOSE             SMOKE           ALCOHOL PHYSICAL_ACTIVITY
##                 0                 0                 0                 0
##    CARDIO_DISEASE
##                 0
```

```r
# Everything that should be factored is factored
h$GENDER <- factor(h$GENDER)
h$CHOLESTEROL <- factor(h$CHOLESTEROL)
h$GLUCOSE <- factor(h$GLUCOSE)
h$SMOKE <- factor(h$SMOKE)
h$ALCOHOL <- factor(h$ALCOHOL)
h$PHYSICAL_ACTIVITY <- factor(h$PHYSICAL_ACTIVITY)
h$CARDIO_DISEASE <- factor(h$CARDIO_DISEASE)

# There is now 67,685 rows
```

Splitting the data into an 80/20 split

```r
#Split data in 80/20 train/test
splitt <- round(nrow(h) * 0.8)

train <- h[1:splitt,]

test <- h[(splitt + 1):nrow(h),]
```

Performing tests: Summary, str, Distribution of choleterol levels, Percent of people who have heart disease based on if the both smoke and drink vs not

```r
# Summary revealing distributions of the training data
summary(train)
```

```
##        ID             AGE        GENDER        HEIGHT          WEIGHT
##  Min.   :    0   Min.   :39.11   1:35215   Min.   :148.0   Min.   : 41.00
##  1st Qu.:19970   1st Qu.:48.34   2:18933   1st Qu.:159.0   1st Qu.: 65.00
##  Median :39994   Median :53.96             Median :165.0   Median : 72.00
##  Mean   :39968   Mean   :53.29             Mean   :164.7   Mean   : 74.25
##  3rd Qu.:59969   3rd Qu.:58.40             3rd Qu.:170.0   3rd Qu.: 82.00
##  Max.   :79861   Max.   :64.91             Max.   :207.0   Max.   :178.00
##     AP_HIGH          AP_LOW        CHOLESTEROL GLUCOSE   SMOKE     ALCOHOL
##  Min.   : 60.0   Min.   : 30.00   1:40700     1:46052   0:49358   0:51252
##  1st Qu.:120.0   1st Qu.: 80.00   2: 7332     2: 3977   1: 4790   1: 2896
##  Median :120.0   Median : 80.00   3: 6116     3: 4119
##  Mean   :126.4   Mean   : 81.33
##  3rd Qu.:140.0   3rd Qu.: 90.00
##  Max.   :197.0   Max.   :190.00
##  PHYSICAL_ACTIVITY CARDIO_DISEASE
##  0:10648           0:27414
##  1:43500           1:26734
##
##
##
##
```

```r
# Revealing which variable types are in the data set
str(train)
```

```
## 'data.frame':    54148 obs. of  13 variables:
##  $ ID               : int  0 1 2 3 4 8 9 12 13 14 ...
##  $ AGE              : num  50.4 55.4 51.7 48.3 47.9 ...
##  $ GENDER           : Factor w/ 2 levels "1","2": 2 1 1 2 1 1 1 2 1 1 ...
##  $ HEIGHT           : int  168 156 165 169 156 151 157 178 158 164 ...
##  $ WEIGHT           : num  62 85 64 82 56 67 93 95 71 68 ...
##  $ AP_HIGH          : int  110 140 130 150 100 120 130 130 110 110 ...
##  $ AP_LOW           : int  80 90 70 100 60 80 80 90 70 60 ...
##  $ CHOLESTEROL      : Factor w/ 3 levels "1","2","3": 1 3 3 1 1 2 3 3 1 1 ...
##  $ GLUCOSE          : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 1 3 1 1 ...
##  $ SMOKE            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ALCOHOL          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ PHYSICAL_ACTIVITY: Factor w/ 2 levels "0","1": 2 2 1 2 1 1 2 2 2 1 ...
##  $ CARDIO_DISEASE   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 1 1 ...
```

```r
# Percentage of people who smoke and drink, and have heart disease
smoke <- which(train$SMOKE == 1)

alcSmo <- which(train$ALCOHOL[smoke] == 1)

alcSmoCD <- which(train$CARDIO_DISEASE[alcSmo] == 1)

ascd <- round((length(alcSmoCD) / length(alcSmo)) * 100, 4)


# Percentage of people who don't smoke and drink, and have heart disease
noSmoke <- which(train$SMOKE == 0)

noAlcSmo <- which(train$ALCOHOL[noSmoke] == 0)

noAlcSmoCD <- which(train$CARDIO_DISEASE[noAlcSmo] == 0)

nascd <- round((length(noAlcSmoCD) / length(noAlcSmo)) * 100, 4)

cat("Interesting to see that not smoking and drinking doesn't have a difference\n
in heart disease versus smokers and drinkers.\n\n", "Percentage with Heart Disease\n",
"\n Smoke and Drink:", ascd, "%\n No Smoke or Drink:", nascd, "%")
```

```
## Interesting to see that not smoking and drinking doesn't have a difference
##
## in heart disease versus smokers and drinkers.
##
##  Percentage with Heart Disease
##
##  Smoke and Drink: 49.1216 %
##  No Smoke or Drink: 50.6547 %
```

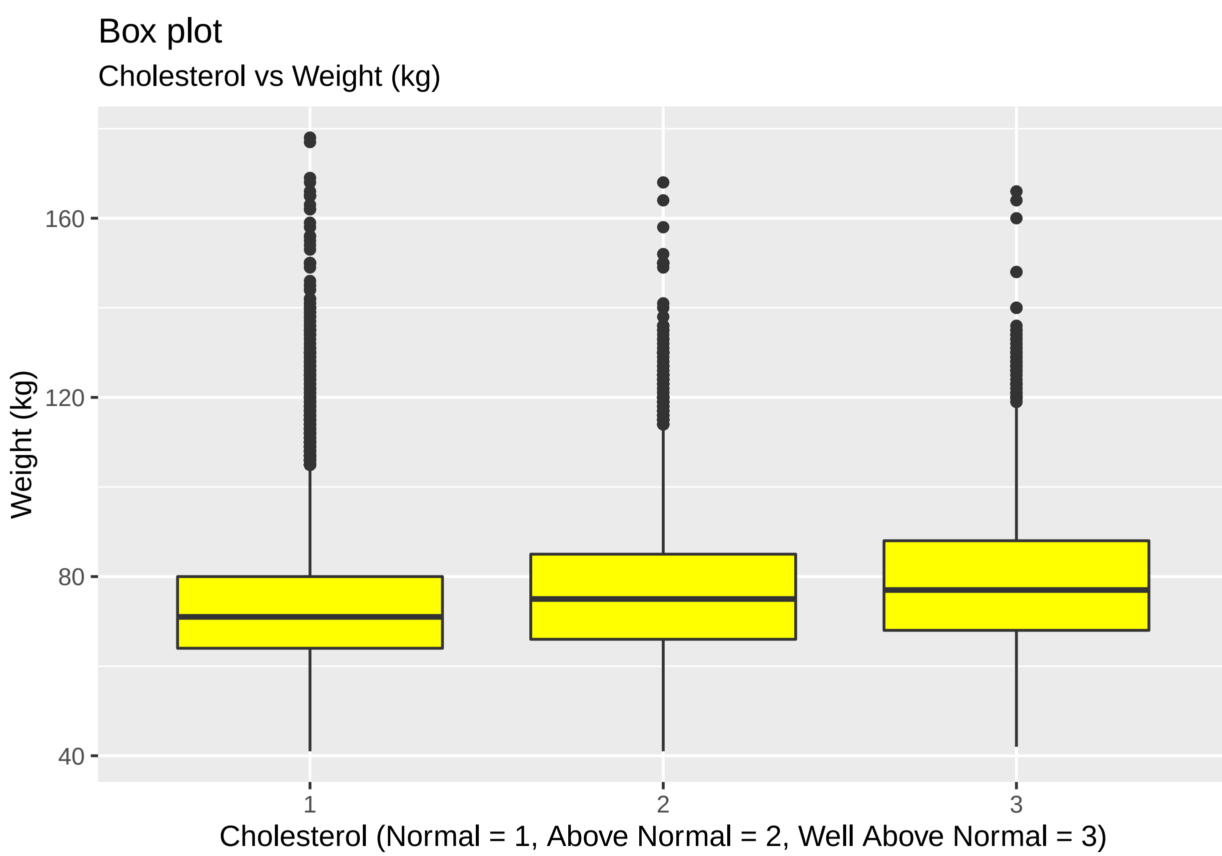
```r
# Distribution of cholesterol levels

cho1 <- which(train$CHOLESTEROL == 1)

cho2 <- which(train$CHOLESTEROL == 2)

cho3 <- which(train$CHOLESTEROL == 3)

chol1 <- round((length(cho1) / length(train$ID)) * 100, 4)

chol2 <- round((length(cho2) / length(train$ID)) * 100, 4)

chol3 <- round((length(cho3) / length(train$ID)) * 100, 4)

cat("Makes me happy to see that most people have a normal cholesterol, I worry\n
for the other 25%.\n\n",
"\nNormal:", chol1, "%\nAbove Normal:", chol2, "%\nWell Above Normal:", chol3, "%")
```

```
## Makes me happy to see that most people have a normal cholesterol, I worry
##
## for the other 25%.
```

```
##
##
## Normal: 75.1644 %
## Above Normal: 13.5407 %
## Well Above Normal: 11.295 %
```

Graphs for finding distributions and relationships

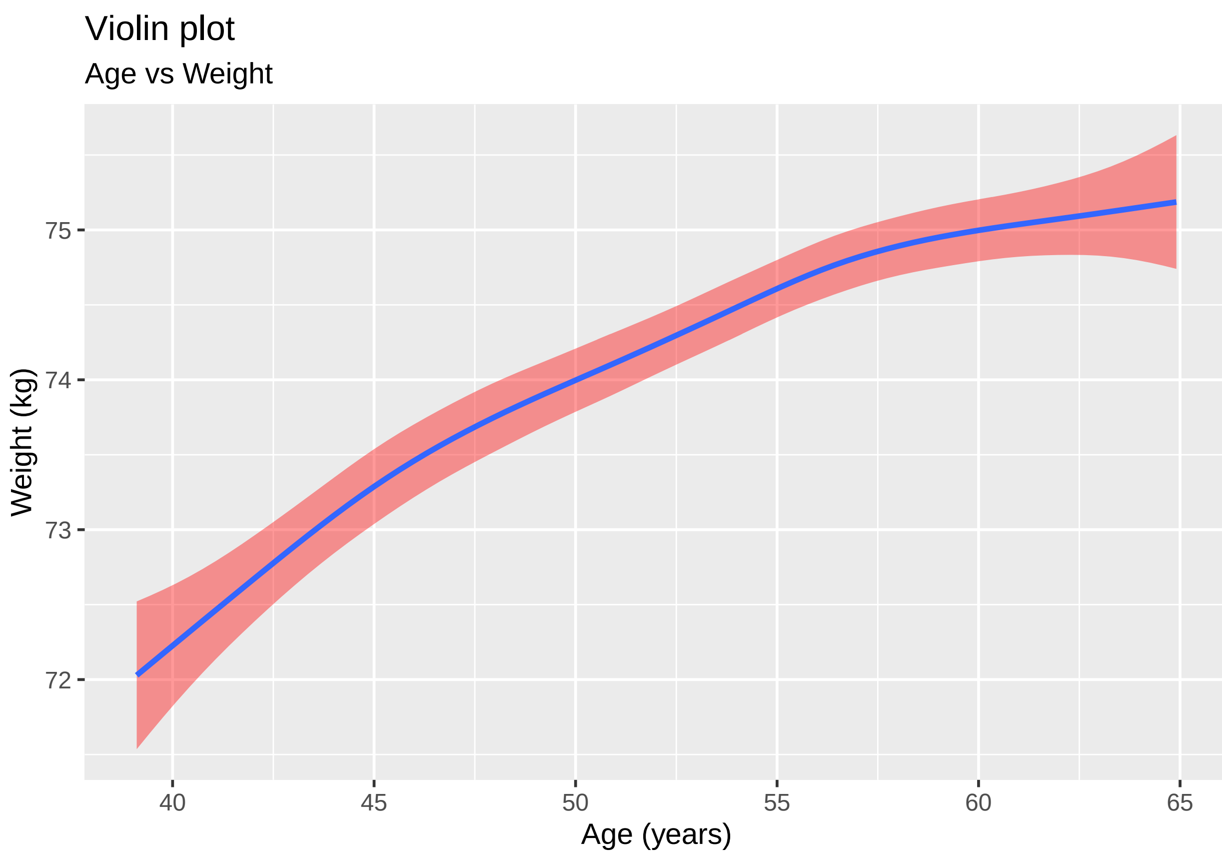
```r
# Correlation of Cholesterol and Weight
  ggplot(train, aes(CHOLESTEROL, WEIGHT)) +
  geom_boxplot(fill = "yellow") +
 labs(title = "Box plot",
      subtitle = "Cholesterol vs Weight (kg)",
      x = "Cholesterol (Normal = 1, Above Normal = 2, Well Above Normal = 3)",
      y = "Weight (kg)")
```



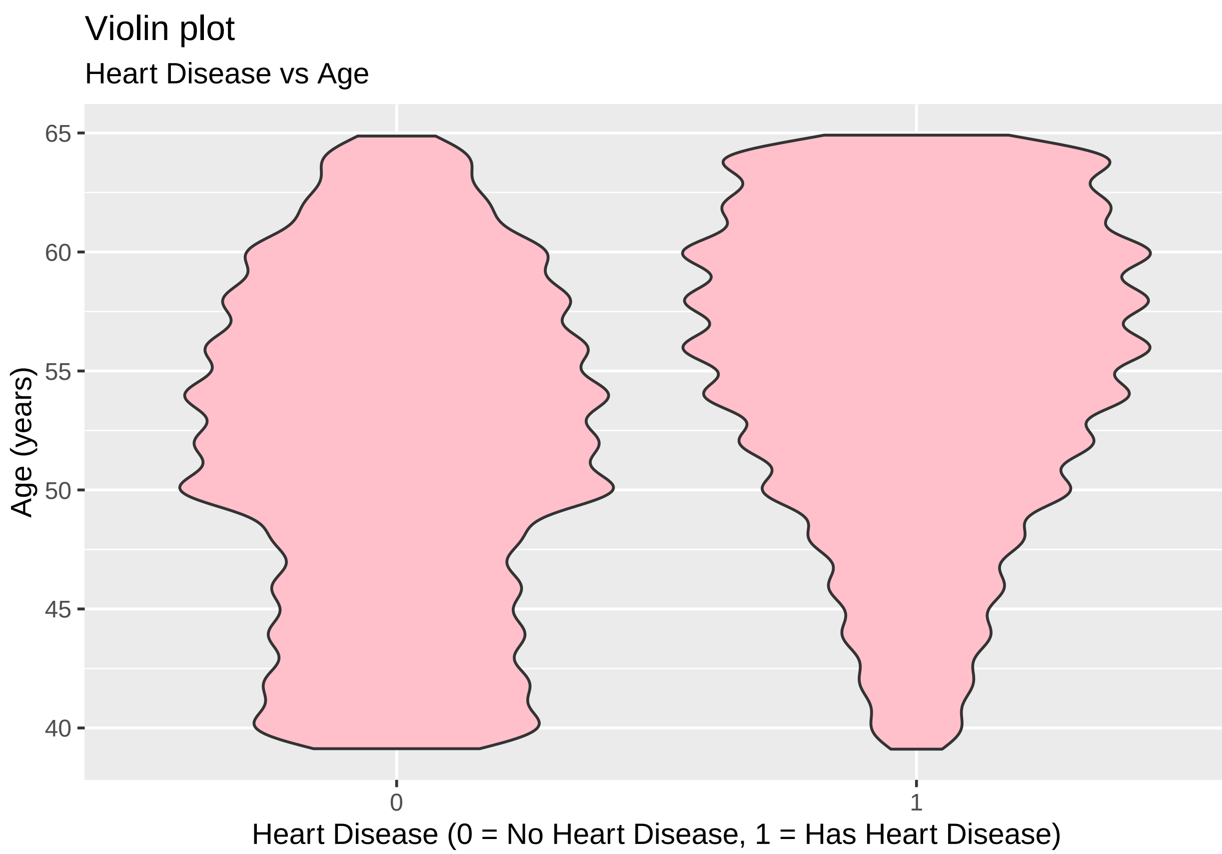

```r
# Correlation of Heart Disease and Weight
 ggplot(train, aes(CARDIO_DISEASE, WEIGHT)) +
  geom_violin(fill = "Pink") +
 labs(title = "Violin plot",
      subtitle = "Heart Disease vs Weight",
      x = "Heart Disease (0 = No Heart Disease, 1 = Has Heart Disease)",
      y = "Weight (kg)")
```


```
# Correlation of Age and Weight
  ggplot(train, aes(AGE, WEIGHT)) +
   geom_smooth(fill = "red") +
  labs(title = "Violin plot",
       subtitle = "Age vs Weight",
       x = "Age (years)",
       y = "Weight (kg)")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```r
# Distribution of Age per having and not having heart disease
  ggplot(train, aes(CARDIO_DISEASE, AGE)) +
   geom_violin(fill = "Pink") +
  labs(title = "Violin plot",
       subtitle = "Heart Disease vs Age",
       x = "Heart Disease (0 = No Heart Disease, 1 = Has Heart Disease)",
       y = "Age (years)")
```



Logistic Regression Model between Heart Disease and [Smoking, Alcohol Use, Execising]

```r
# Logistic Regression Model
lrm <- glm(formula = CARDIO_DISEASE ~ SMOKE + ALCOHOL + PHYSICAL_ACTIVITY, train, family = binomial)

summary(lrm)
```

```
##
## Call:
## glm(formula = CARDIO_DISEASE ~ SMOKE + ALCOHOL + PHYSICAL_ACTIVITY,
##     family = binomial, data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.237  -1.157  -1.088   1.198   1.269
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.13933    0.01957   7.119 1.09e-12 ***
## SMOKE1             -0.12200    0.03225  -3.783 0.000155 ***
## ALCOHOL1           -0.04274    0.04068  -1.051 0.293412
## PHYSICAL_ACTIVITY1 -0.18843    0.02168  -8.693  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

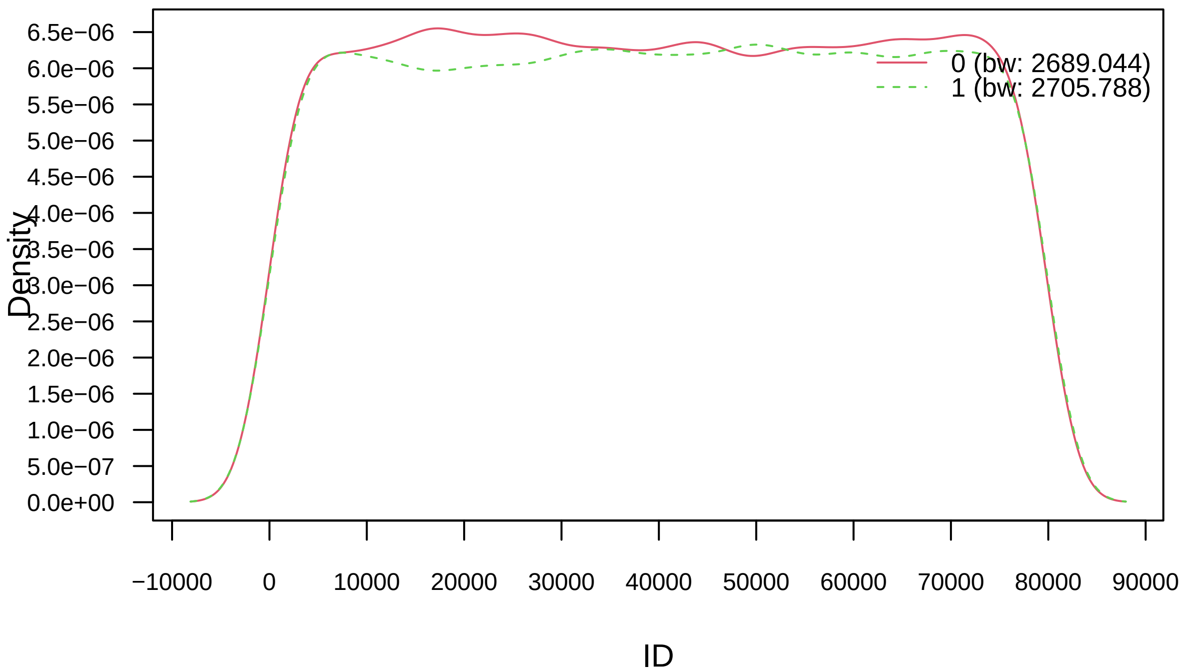
```
##
##     Null deviance: 75057  on 54147  degrees of freedom
## Residual deviance: 74958  on 54144  degrees of freedom
## AIC: 74966
##
## Number of Fisher Scoring iterations: 3
```

```r
# Interpretation
cat("We can see that Alcohol is not statistically signinficant, but Physical \n
Activity clearly is. That is determined by the p-value of the F-statisitic. \n
For Alcohol it's 0.293, 2e^(-16) for Physical Activity, and 0.000155 for Smoking.\n
The negative coefficient (-0.18843) for Physical Activity implies that if someone\n
has heart disease, they more likely are those who don't exercise. The Null deviance\n
gap is fairly large too, which is good, and implies stronger correlation. Looking\n
at the summary it looks pretty good for a correlation between Physical Activity and\n
Heart Disease.")
```

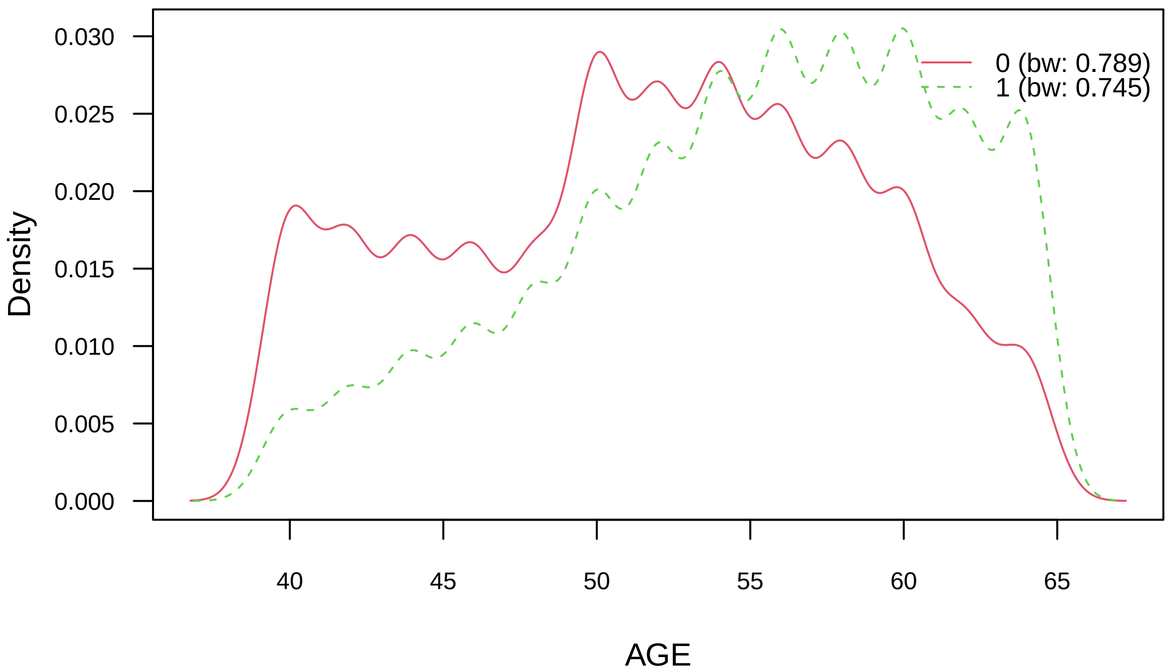
```
## We can see that Alcohol is not statistically signinficant, but Physical
##
## Activity clearly is. That is determined by the p-value of the F-statisitic.
##
## For Alcohol it's 0.293, 2e^(-16) for Physical Activity, and 0.000155 for Smoking.
##
## The negative coefficient (-0.18843) for Physical Activity implies that if someone
##
## has heart disease, they more likely are those who don't exercise. The Null deviance
##
## gap is fairly large too, which is good, and implies stronger correlation. Looking
##
## at the summary it looks pretty good for a correlation between Physical Activity and
##
## Heart Disease.
```

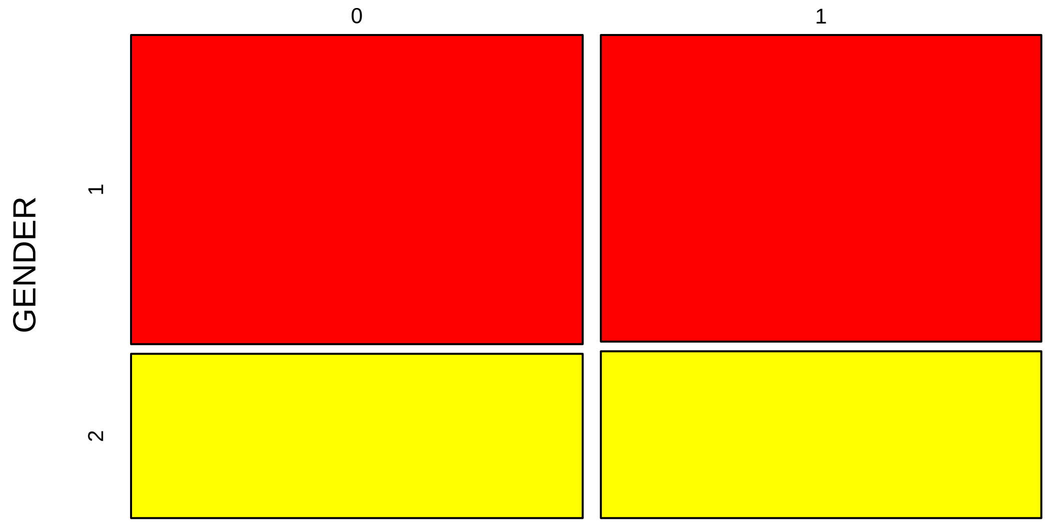
Naive Bayes Model

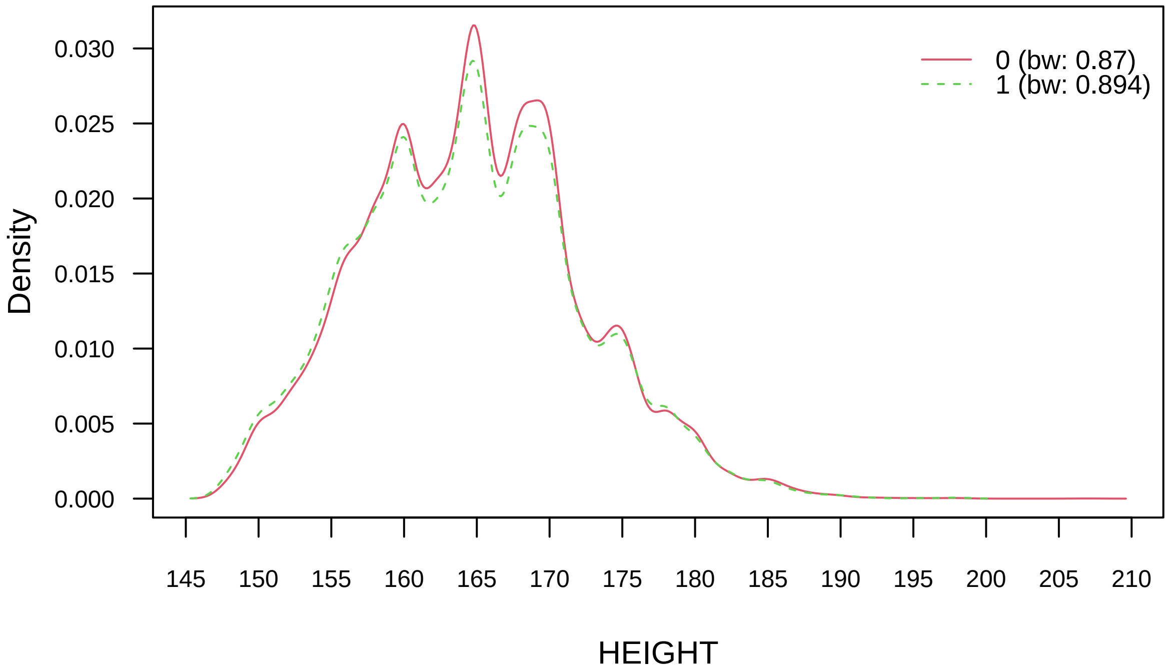
```r
# Naive Bayes Model
model <- naive_bayes(CARDIO_DISEASE ~ ., data = train, usekernel = T)

plot(model)
```



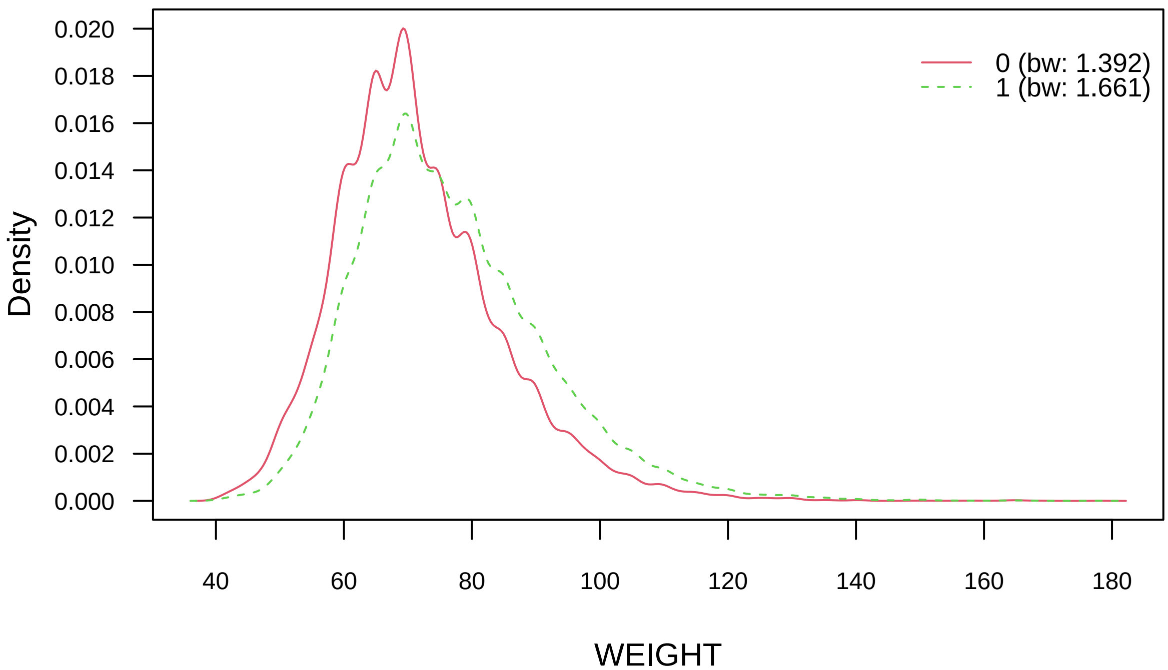
0 (bw: 2689.044)
1 (bw: 2705.788)
Density
ID

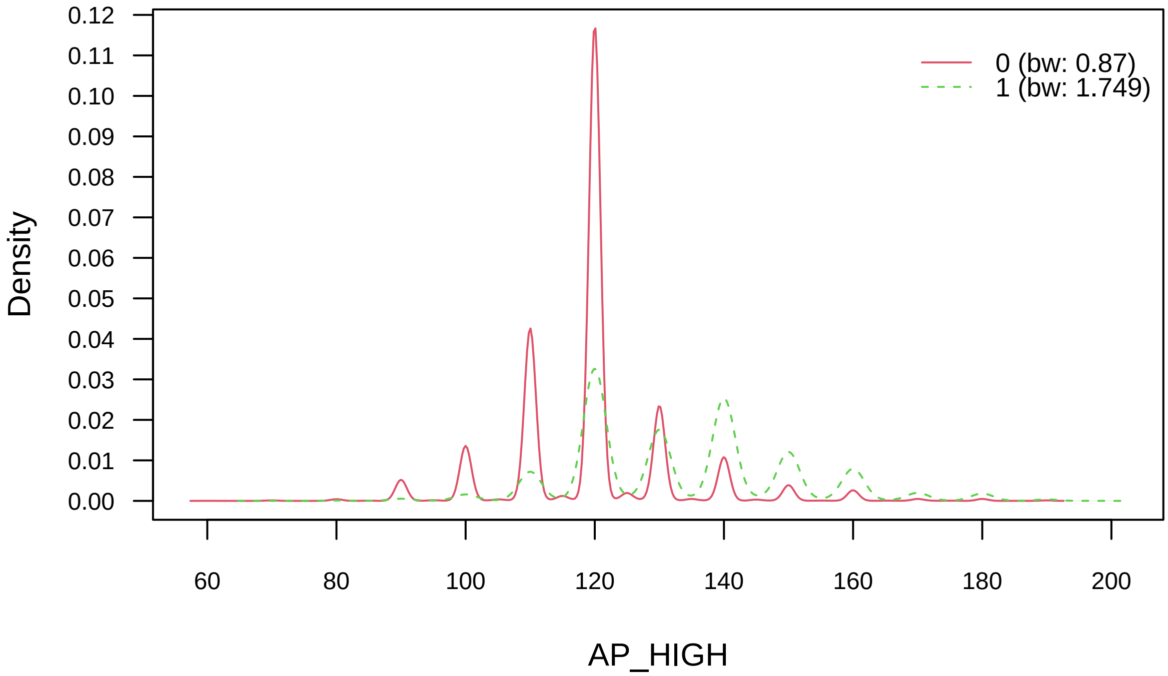

0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0 (bw: 0.789)
1 (bw: 0.745)
40
45
50
55
60
65
AGE


0
1
GENDER
1
2

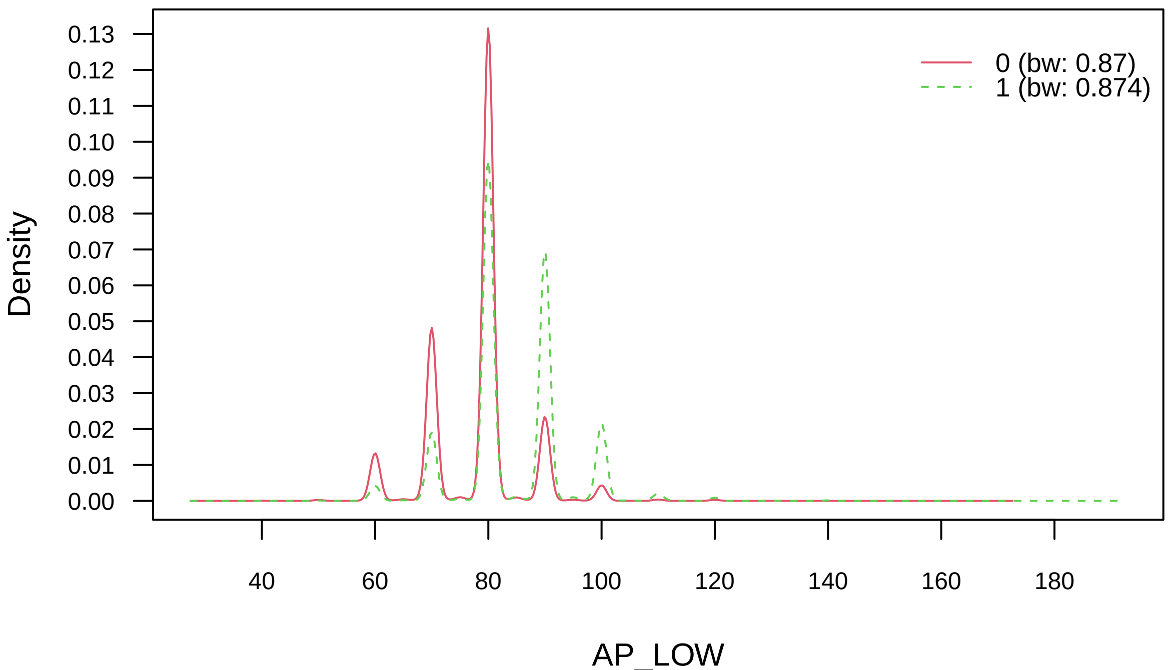


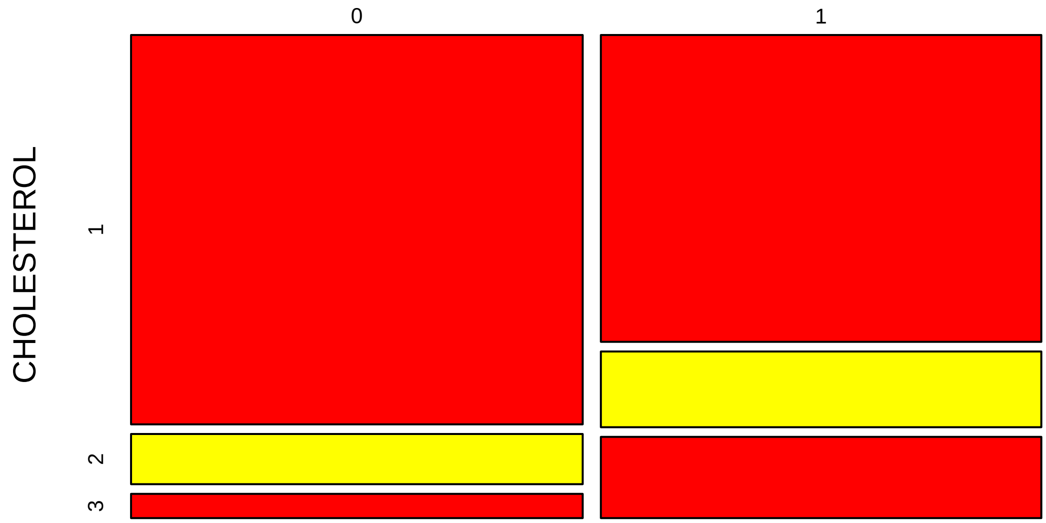
0 (bw: 0.87)
1 (bw: 0.894)
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
145
150
155
160
165
170
175
180
185
190
195
200
205
210
HEIGHT

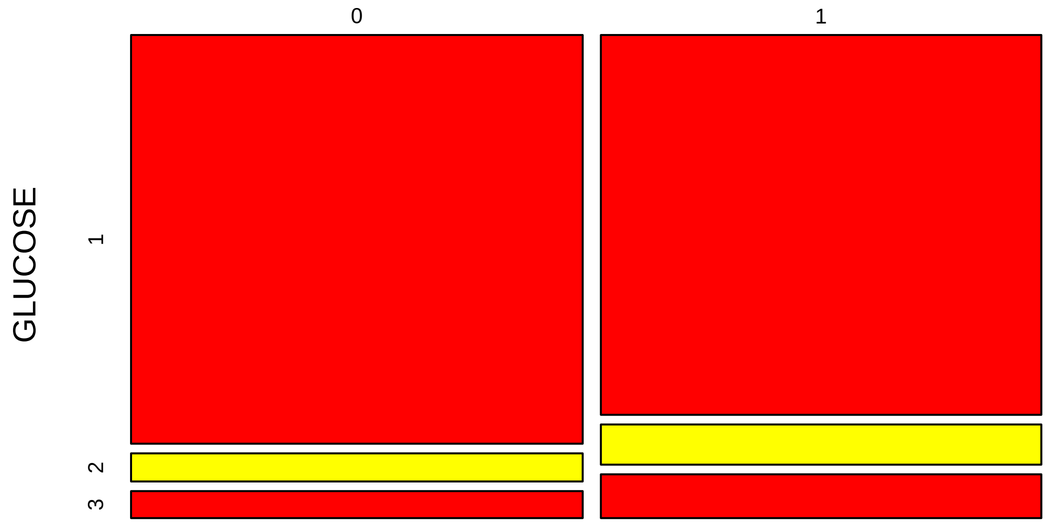

0 (bw: 1.392)
1 (bw: 1.661)
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0.018
0.020
40
60
80
100
120
140
160
180
WEIGHT

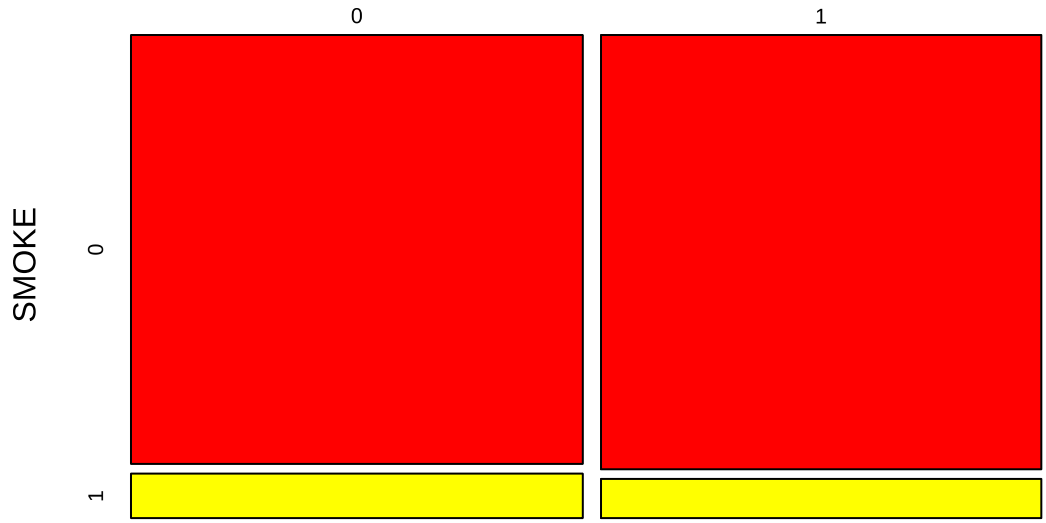


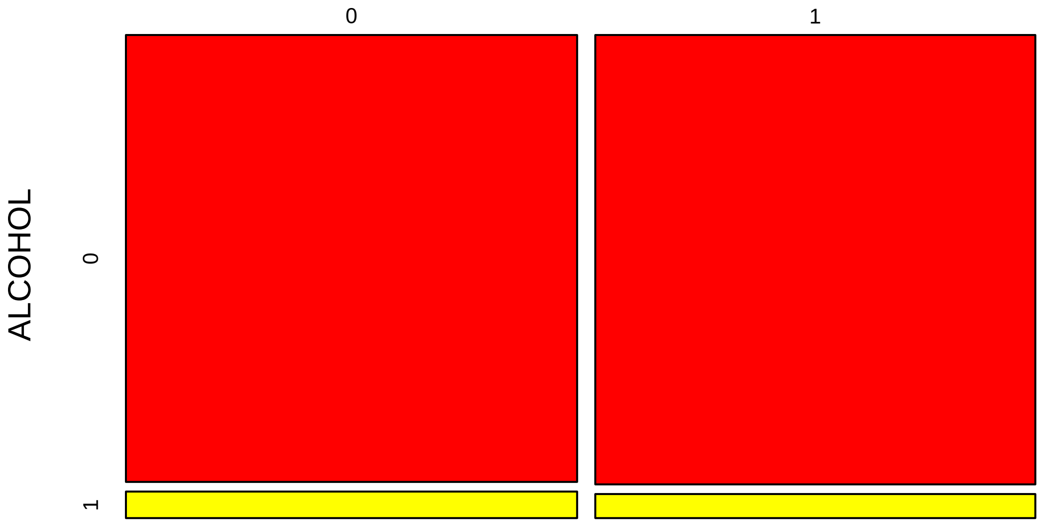
0 (bw: 0.87)
1 (bw: 1.749)
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11
0.12
60
80
100
120
140
160
180
200
AP_HIGH

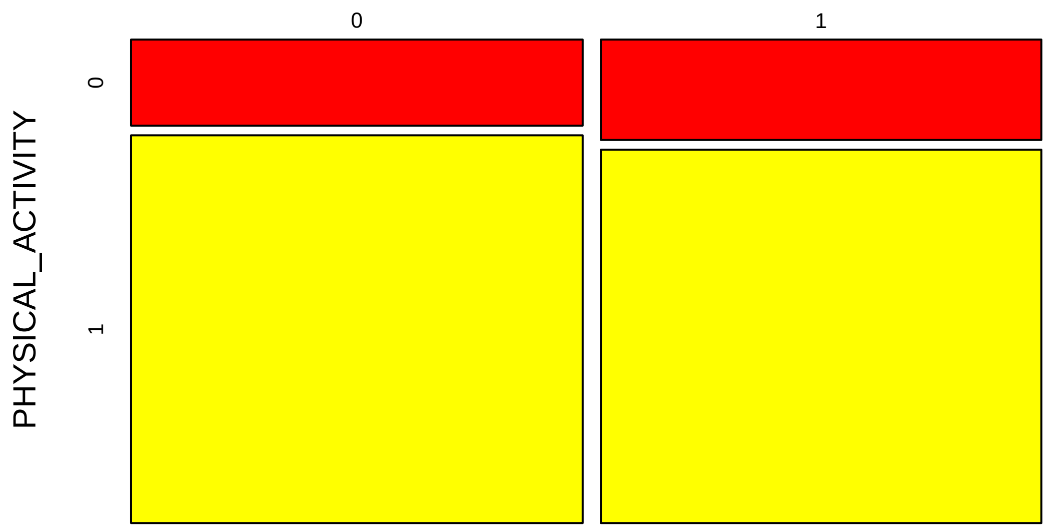

0 (bw: 0.87)
1 (bw: 0.874)
Density
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
40
60
80
100
120
140
160
180
AP_LOW


0
1
CHOLESTEROL
1
2
3


0
1
GLUCOSE
1
2
3


0
1
SMOKE
0
1


0
1
ALCOHOL
0
1




```r
summary(model)
```

```
##
## ================================== Naive Bayes ==================================
##
## - Call: naive_bayes.formula(formula = CARDIO_DISEASE ~ ., data = train,      usekernel = T)
## - Laplace: 0
## - Classes: 2
## - Samples: 54148
## - Features: 12
## - Conditional distributions:
##     - Bernoulli: 4
##     - Categorical: 2
##     - KDE: 6
## - Prior probabilities:
##     - 0: 0.5063
##     - 1: 0.4937
##
## ---------------------------------------------------------------------------------
```

```r
cat("Looking exclusively at the Mosaic graphs You can clearly see that high\n
cholesterol, high glucose, and not exercising, are the biggest contributors\n
for determining heart disease. Looking exclusively at the line graphs age, \n
weight, and maybe blood pressure has a correlation to heart disease.")
```

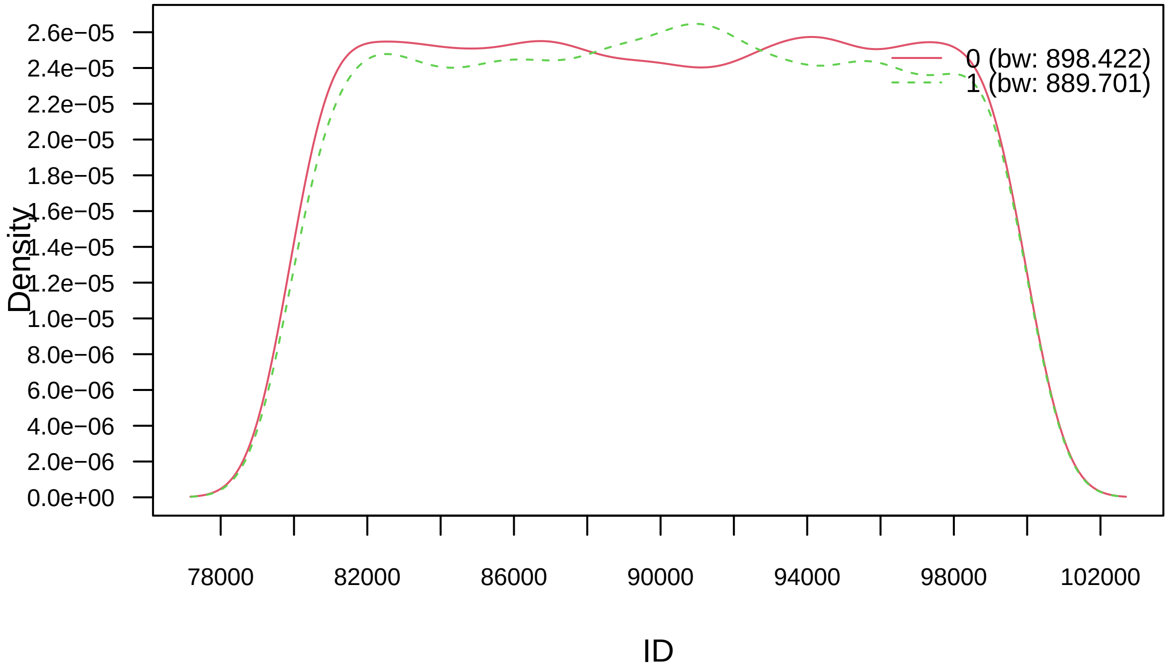
```
## Looking exclusively at the Mosaic graphs You can clearly see that high
##
## cholesterol, high glucose, and not exercising, are the biggest contributors
##
## for determining heart disease. Looking exclusively at the line graphs age,
##
## weight, and maybe blood pressure has a correlation to heart disease.
```

Logisitic Regression and Naive Bayes Model for test data

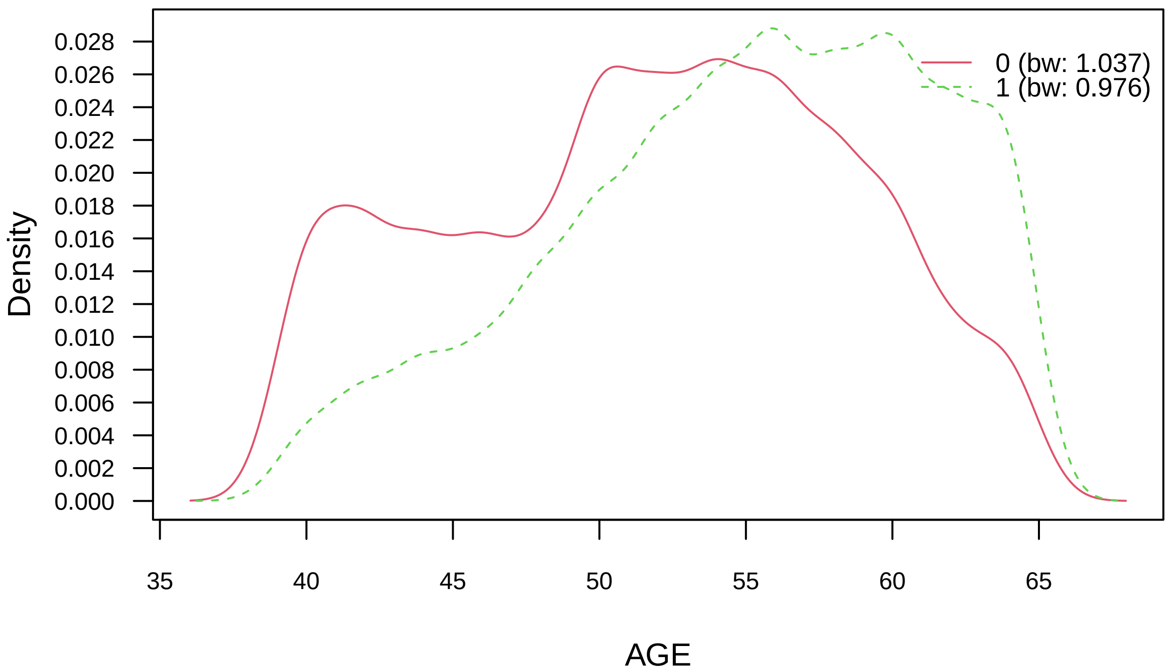
```r
# Logistic Regression based off of test data
tlrm <- glm(formula = CARDIO_DISEASE ~ SMOKE + ALCOHOL + PHYSICAL_ACTIVITY, test, family = binomial)

summary(tlrm)
```

```
##
## Call:
## glm(formula = CARDIO_DISEASE ~ SMOKE + ALCOHOL + PHYSICAL_ACTIVITY,
##     family = binomial, data = test)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.259  -1.152  -1.136   1.203   1.219
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.12912    0.03923   3.291 0.000998 ***
## SMOKE1             -0.03764    0.06428  -0.586 0.558146
## ALCOHOL1            0.06058    0.08052   0.752 0.451777
## PHYSICAL_ACTIVITY1 -0.18866    0.04338  -4.349 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18765  on 13536  degrees of freedom
## Residual deviance: 18745  on 13533  degrees of freedom
## AIC: 18753
##
## Number of Fisher Scoring iterations: 3
```

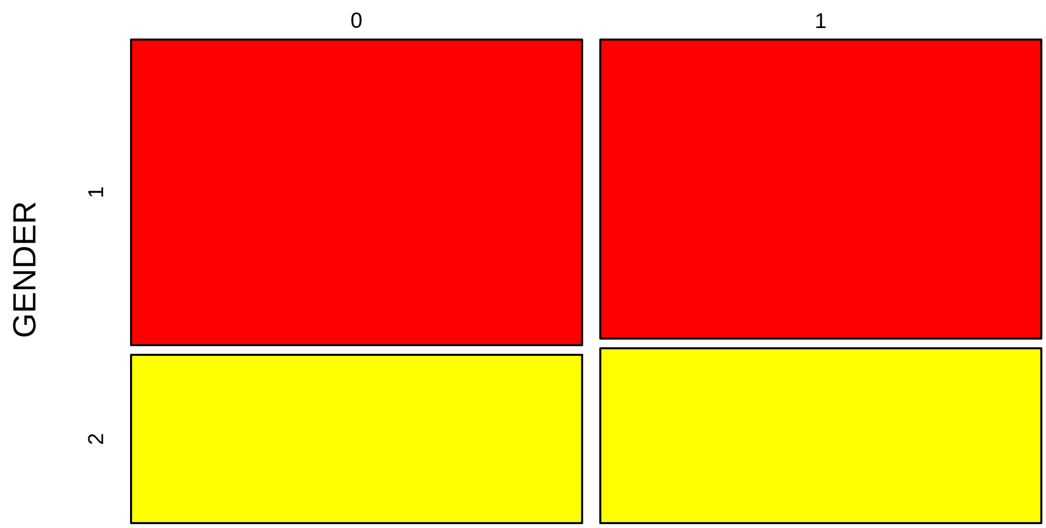
```r
# Naive Bayes Model based off of test data
tmodel <- naive_bayes(CARDIO_DISEASE ~ ., data = test, usekernel = T)

plot(tmodel)
```

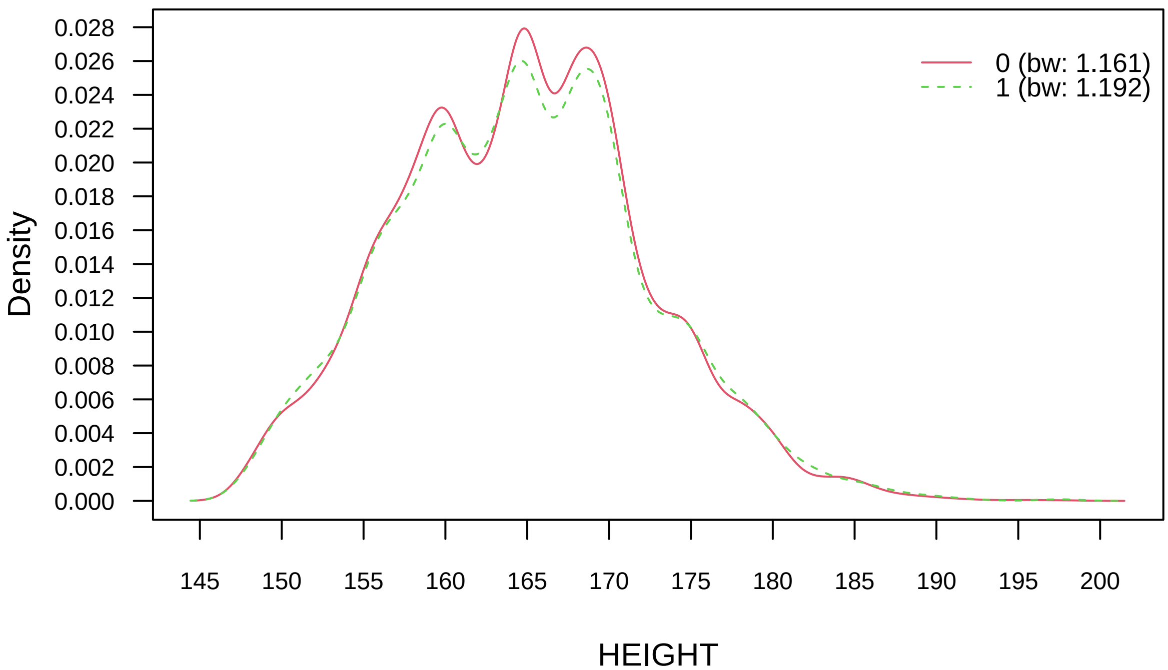


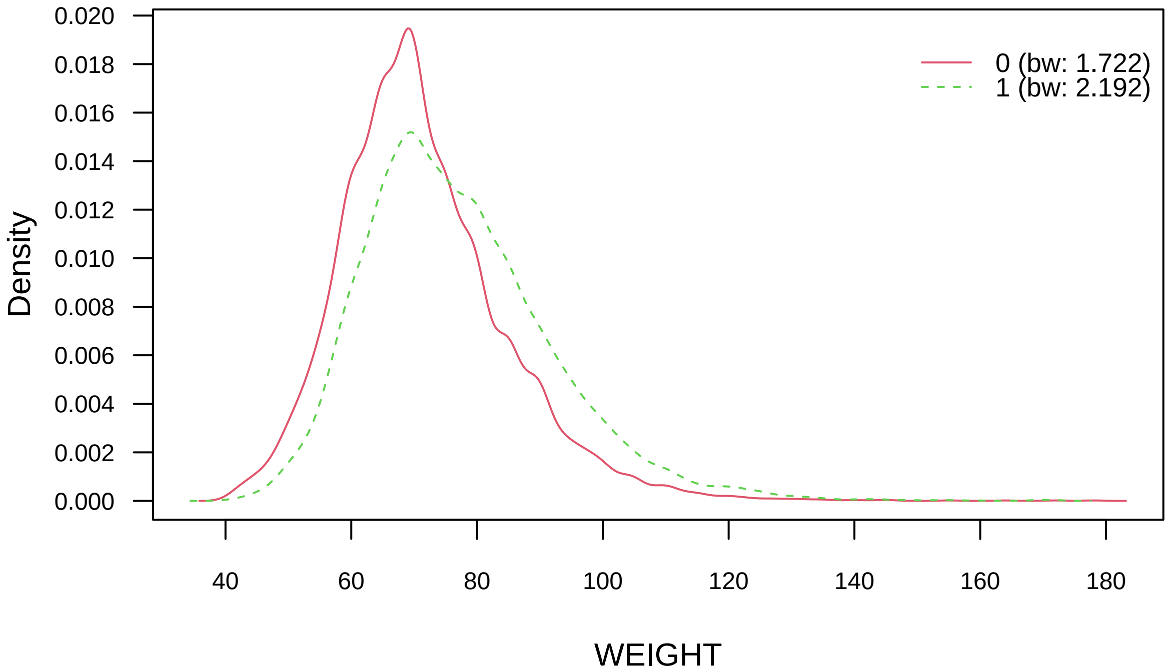
2.6e−05
2.4e−05
2.2e−05
2.0e−05
1.8e−05
1.6e−05
1.4e−05
1.2e−05
1.0e−05
8.0e−06
6.0e−06
4.0e−06
2.0e−06
0.0e+00
Density
78000
82000
86000
90000
94000
98000
102000
ID
0 (bw: 898.422)
1 (bw: 889.701)

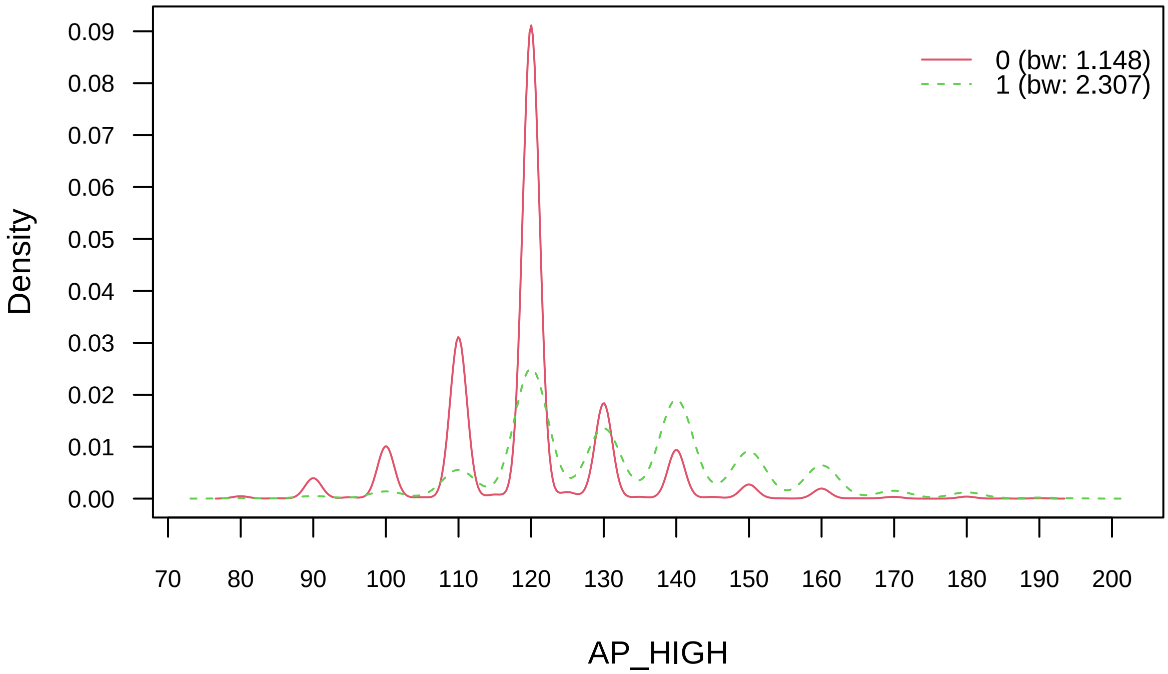


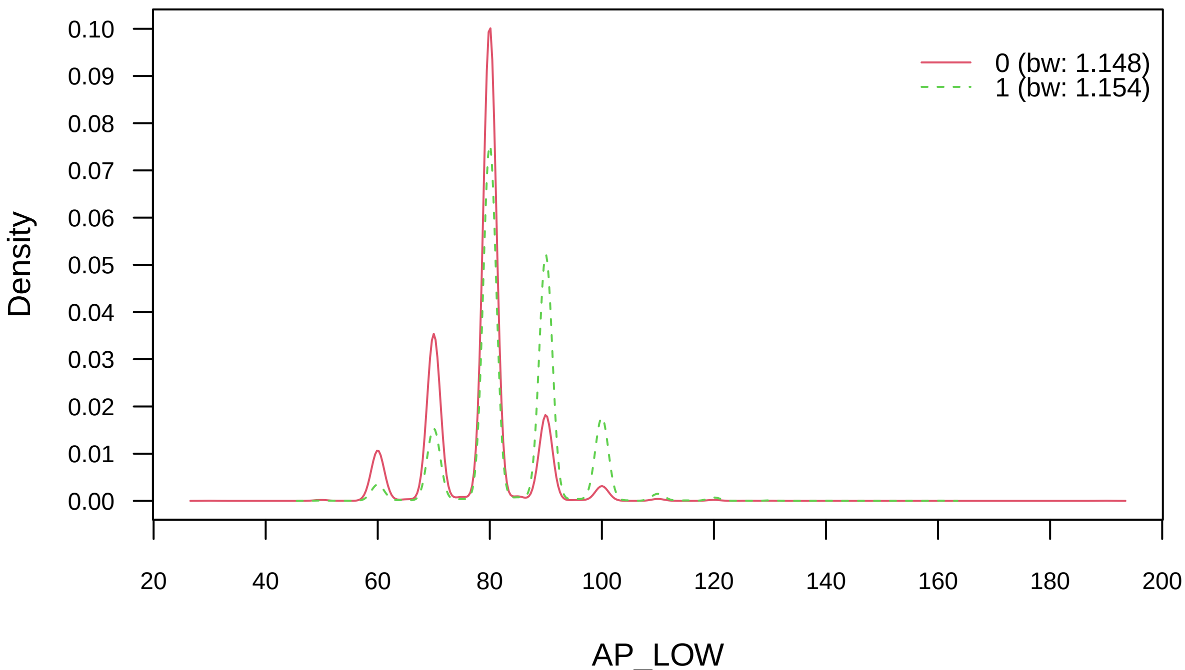
Density
0.028
0.026
0.024
0.022
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0 (bw: 1.037)
1 (bw: 0.976)
35
40
45
50
55
60
65
AGE

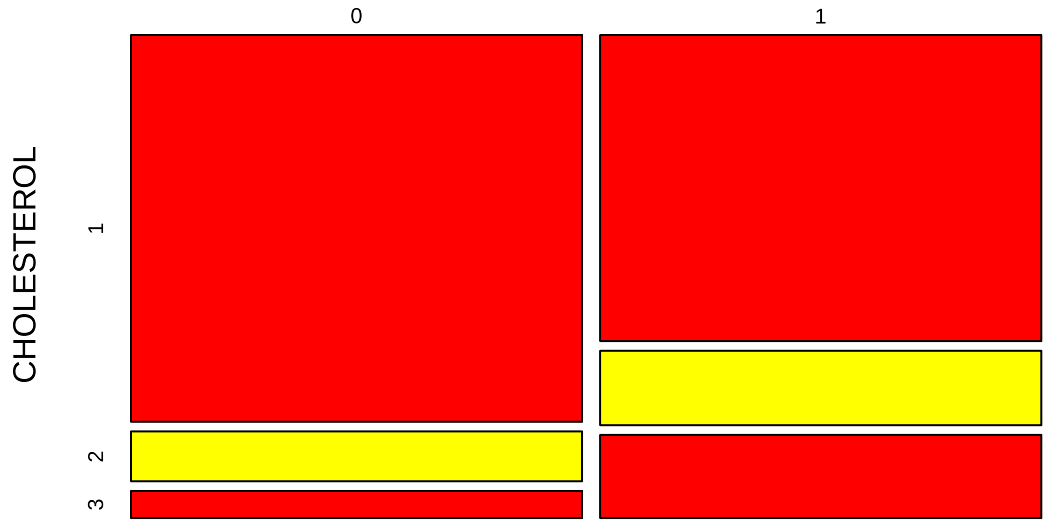

0
1
GENDER
1
2

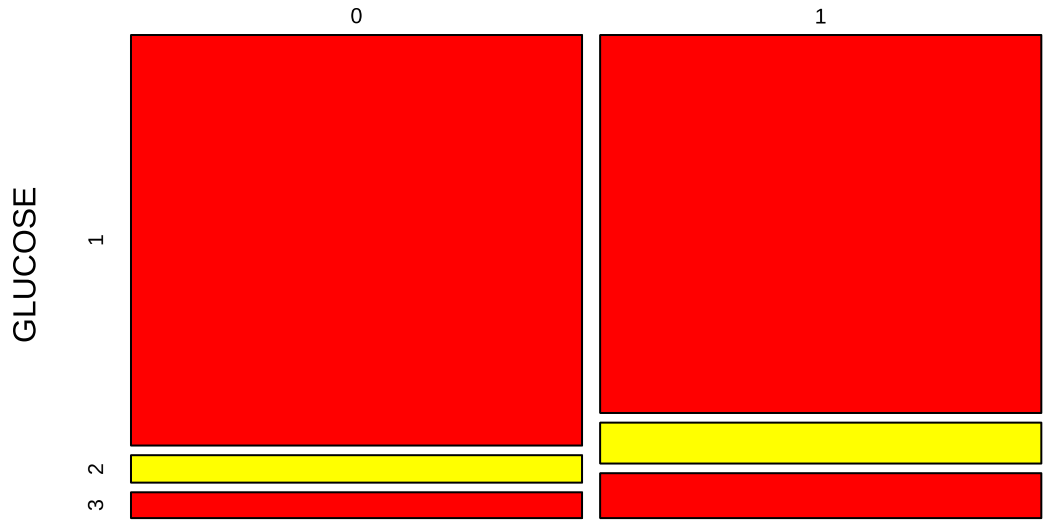

0.028
0.026
0.024
0.022
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0 (bw: 1.161)
1 (bw: 1.192)
145
150
155
160
165
170
175
180
185
190
195
200
HEIGHT

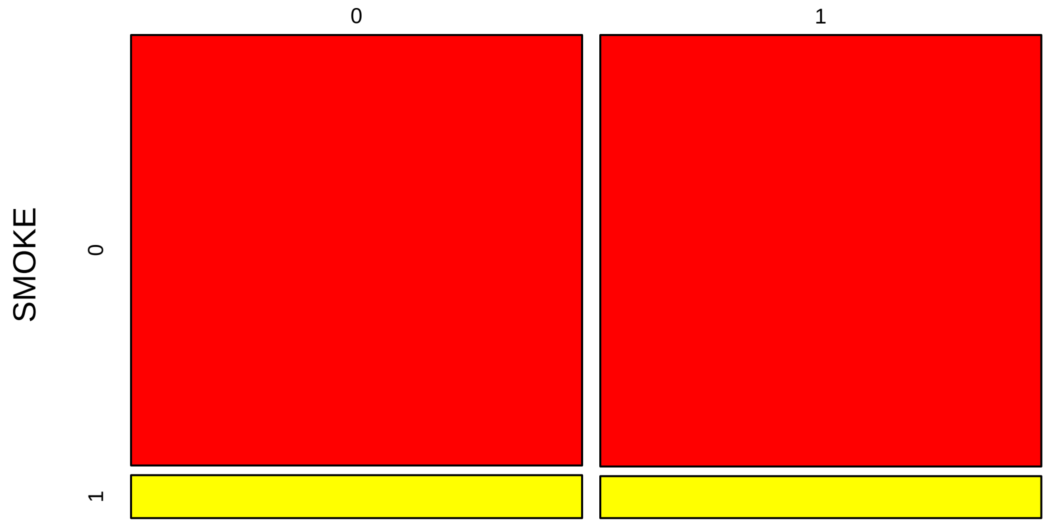

0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0 (bw: 1.722)
1 (bw: 2.192)
40
60
80
100
120
140
160
180
WEIGHT

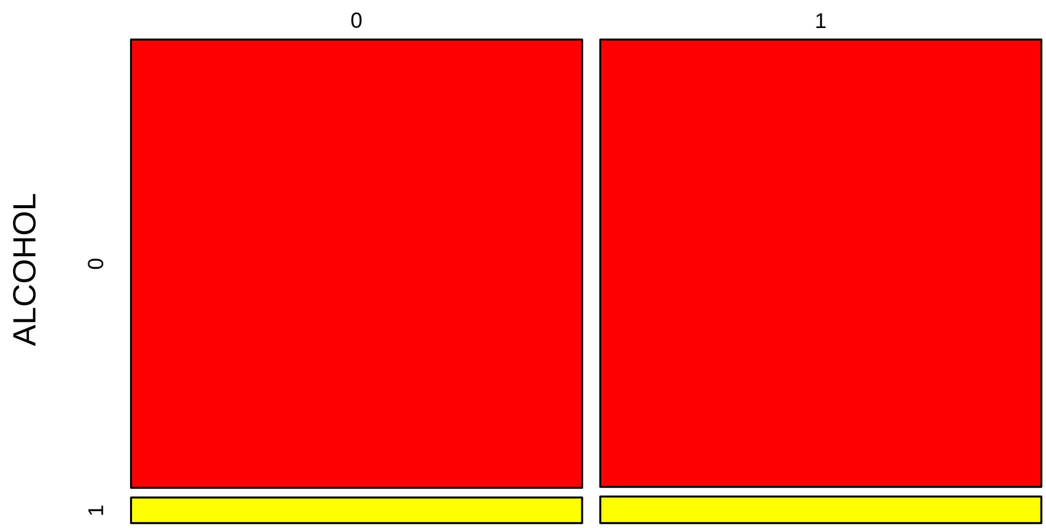

0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0 (bw: 1.148)
1 (bw: 2.307)
70
80
90
100
110
120
130
140
150
160
170
180
190
200
AP_HIGH

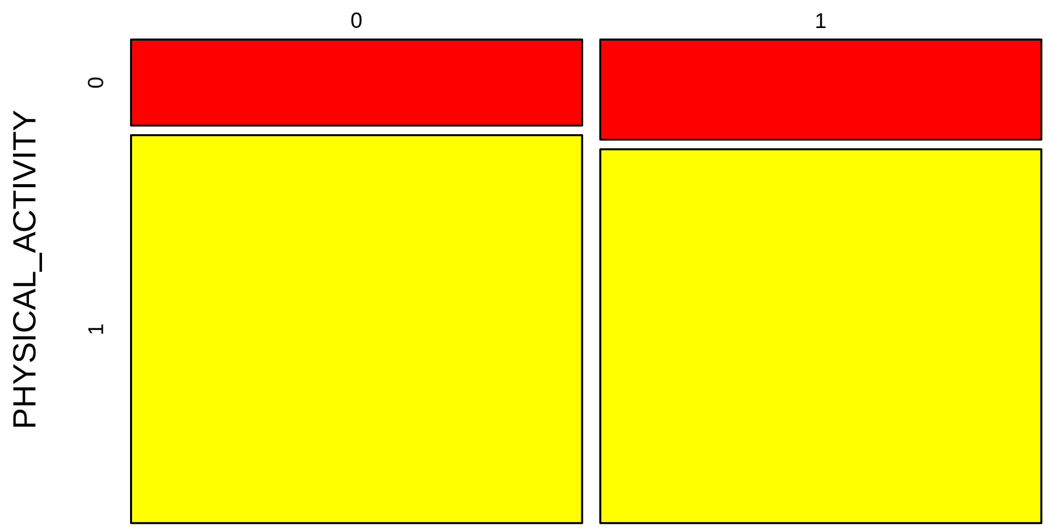

0 (bw: 1.148)
1 (bw: 1.154)
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
20
40
60
80
100
120
140
160
180
200
AP_LOW


0
1
CHOLESTEROL
1
2
3


0
1
GLUCOSE
1
2
3


0
1
SMOKE
0
1


0
1
ALCOHOL
0
1




```r
summary(tmodel)
```

```
##
## =================================== Naive Bayes ==================================
##
## - Call: naive_bayes.formula(formula = CARDIO_DISEASE ~ ., data = test,      usekernel = T)
## - Laplace: 0
## - Classes: 2
## - Samples: 13537
## - Features: 12
## - Conditional distributions:
##     - Bernoulli: 4
##     - Categorical: 2
##     - KDE: 6
## - Prior probabilities:
##     - 0: 0.5057
##     - 1: 0.4943
##
## ---------------------------------------------------------------------------------
```

```r
# ACcuracy Test
test$model_prob <- predict(tlrm, test, type = "response")

test <- test %>% mutate(model_pred = 1*(model_prob > .53) + 0,)
```

```r
test <- test %>% mutate(accurate = 1*(model_pred == CARDIO_DISEASE))

sum(test$accurate/nrow(test))
```

```
## [1] 0.5170274
```

```r
cat("The results are very similar to the train data, that's most likely because\n
of the size of the data. It resists against any skewing.")
```

```
## The results are very similar to the train data, that's most likely because
##
## of the size of the data. It resists against any skewing.
```

Strengths and Weaknesses of Naïve Bayes and Logistic Regression

```r
cat("A strength of Naive Bayes is that it doesn't require a large amount of sample\n
data, while Logistic Regression does. Logistic Regression has low bias and high\n
variance while Naive Bias has the inverse. Naive Bayes is easy to implement and\n
very fast, but independence assumptions don't always hold, it usually shows some\n
form of dependency. A major disadvantage is that Logisitic Regression assumes\n
relationships to be linear, even though it could be exponential, etc. but it's\n
very easy to extend to multiple columns or classes.")
```

```
## A strength of Naive Bayes is that it doesn't require a large amount of sample
##
## data, while Logistic Regression does. Logistic Regression has low bias and high
##
## variance while Naive Bias has the inverse. Naive Bayes is easy to implement and
##
## very fast, but independence assumptions don't always hold, it usually shows some
##
## form of dependency. A major disadvantage is that Logisitic Regression assumes
##
## relationships to be linear, even though it could be exponential, etc. but it's
##
## very easy to extend to multiple columns or classes.
```

Classification Metrics

```r
cat("Accuracy is just a way of finding the amount of correct predictions by \n
dividing the correct predictions by the number of rows. I had an accuracy of \n
about 52% which means the model was only correct about 52% of the time.")
```

```
## Accuracy is just a way of finding the amount of correct predictions by
##
## dividing the correct predictions by the number of rows. I had an accuracy of
##
## about 52% which means the model was only correct about 52% of the time.
```