final\_project\_p8106

Michelle Lui

4/27/2021

## Import dataset

#import dataset  
data("NHANES")  
  
#clean dataset - limit to 2011/12 and only include biological predictors, dropped some biological predictors due to excess missing data  
nhanes = NHANES %>% janitor::clean\_names() %>% filter(survey\_yr == "2011\_12") %>% select(gender, age, race1, weight, height, bmi, pulse:bp\_dia3, direct\_chol:urine\_flow1, diabetes) %>% drop\_na()  
  
summary(nhanes)

## gender age race1 weight height   
## female:1735 Min. : 8.00 Black : 392 Min. : 23.00 Min. :124.5   
## male :1781 1st Qu.:25.00 Hispanic: 218 1st Qu.: 62.90 1st Qu.:160.3   
## Median :42.00 Mexican : 315 Median : 77.10 Median :167.8   
## Mean :41.81 White :2319 Mean : 78.25 Mean :167.5   
## 3rd Qu.:57.00 Other : 272 3rd Qu.: 91.10 3rd Qu.:175.4   
## Max. :80.00 Max. :188.50 Max. :199.9   
## bmi pulse bp\_sys\_ave bp\_dia\_ave   
## Min. :12.90 Min. : 40.00 Min. : 81 Min. : 0.00   
## 1st Qu.:22.90 1st Qu.: 64.00 1st Qu.:107 1st Qu.: 62.00   
## Median :26.60 Median : 72.00 Median :117 Median : 69.00   
## Mean :27.63 Mean : 73.41 Mean :119 Mean : 68.44   
## 3rd Qu.:31.30 3rd Qu.: 82.00 3rd Qu.:128 3rd Qu.: 77.00   
## Max. :69.00 Max. :136.00 Max. :209 Max. :116.00   
## bp\_sys1 bp\_dia1 bp\_sys2 bp\_dia2   
## Min. : 74.0 Min. : 0.0 Min. : 82.0 Min. : 0.00   
## 1st Qu.:108.0 1st Qu.: 62.0 1st Qu.:108.0 1st Qu.: 62.00   
## Median :118.0 Median : 70.0 Median :118.0 Median : 70.00   
## Mean :119.5 Mean : 69.2 Mean :119.2 Mean : 68.55   
## 3rd Qu.:128.0 3rd Qu.: 78.0 3rd Qu.:128.0 3rd Qu.: 78.00   
## Max. :212.0 Max. :110.0 Max. :208.0 Max. :116.00   
## bp\_sys3 bp\_dia3 direct\_chol tot\_chol   
## Min. : 78.0 Min. : 0.00 Min. :0.470 Min. : 1.530   
## 1st Qu.:106.0 1st Qu.: 62.00 1st Qu.:1.110 1st Qu.: 4.110   
## Median :116.0 Median : 70.00 Median :1.290 Median : 4.780   
## Mean :118.7 Mean : 68.33 Mean :1.365 Mean : 4.858   
## 3rd Qu.:128.0 3rd Qu.: 78.00 3rd Qu.:1.580 3rd Qu.: 5.530   
## Max. :210.0 Max. :116.00 Max. :4.030 Max. :10.290   
## urine\_vol1 urine\_flow1 diabetes   
## Min. : 1 Min. : 0.0110 No :3223   
## 1st Qu.: 47 1st Qu.: 0.3995 Yes: 293   
## Median : 88 Median : 0.6800   
## Mean :113 Mean : 0.9642   
## 3rd Qu.:156 3rd Qu.: 1.2140   
## Max. :446 Max. :10.1430

#create a feature plot to better visualize the data  
theme1 <- transparentTheme(trans = .4)   
trellis.par.set(theme1)  
#featurePlot(x = nhanes[, 1:19],  
# y = nhanes$diabetes,  
# scales = list(x = list(relation = "free"), y = list(relation = "free")),  
# plot = "density", pch = "|", auto.key = list(columns = 2))

## Create Data Partition

set.seed(2)  
  
train\_rows = createDataPartition(y = nhanes$diabetes,p = 0.7,list = FALSE)  
train = nhanes[train\_rows, ]

## Warning: The `i` argument of ``[`()` can't be a matrix as of tibble 3.0.0.  
## Convert to a vector.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

test = nhanes[-train\_rows, ]  
  
x = model.matrix(diabetes ~ ., train)[ ,-1]  
y = train$diabetes  
  
x2 = model.matrix(diabetes ~ ., test)[ ,-1]  
y2 = test$diabetes  
  
control1 = trainControl(method = "cv", selectionFunction = "best", sampling = "down")