Machine learning - Assignment 6

Alice Tivarovsky

2/27/2020

## Setup

Loading the necessary packages for data cleaning, modeling, partitioning, nhanes interface, and classification.

Comparison between Classification Trees, SVM and Logistic Regression

The posted article by Yu et al utilized NHANES data from 1999-2004 to predict diabetes and pre-diabetes using Support Vector Machines. You will conduct a similar analysis using data within the NHANES package in R. For this exercise, you will try to predict Diabetes using similar (although not all) variables. The available data is also slightly different, so you likely won’t get the same answers.

## Number 1: Loading NHANES, Restricting and Partitioning

Restrict the NHANES data to the list of 12 variables below. Partition the data into training and testing using a 70/30 split.

“Age”, “Gender”, “Race1”, “Education”, “HHIncome” (DEMO), “Weight”, “Height” (BMXBMI), “Pulse” (BPX), “Diabetes” (DIQ), “BMI” (BMXBMI), “PhysActive”, “Smoke100” (SMQ)

# Loading  
demo\_99 = nhanes\_load\_data("DEMO", "1999-2000")

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.  
## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.  
## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Downloading DEMO.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/DEMO.XPT

demo\_01 = nhanes\_load\_data("DEMO", "2001-2002")

## Downloading DEMO\_B.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/DEMO\_B.XPT

demo\_03 = nhanes\_load\_data("DEMO", "2003-2004")

## Downloading DEMO\_C.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/DEMO\_C.XPT

bmx\_99 = nhanes\_load\_data("BMX", "1999-2000")

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.  
## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Downloading BMX.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/BMX.XPT

bmx\_01 = nhanes\_load\_data("BMX", "2001-2002")

## Downloading BMX\_B.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/BMX\_B.XPT

bmx\_03 = nhanes\_load\_data("BMX", "2003-2004")

## Downloading BMX\_C.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/BMX\_C.XPT

bpx\_99 = nhanes\_load\_data("BPX", "1999-2000")

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.  
## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Downloading BPX.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/BPX.XPT

bpx\_01 = nhanes\_load\_data("BPX", "2001-2002")

## Downloading BPX\_B.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/BPX\_B.XPT

bpx\_03 = nhanes\_load\_data("BPX", "2003-2004")

## Downloading BPX\_C.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/BPX\_C.XPT

paq\_99 = nhanes\_load\_data("PAQ", "1999-2000")

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.  
## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Downloading PAQ.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/PAQ.XPT

paq\_01 = nhanes\_load\_data("PAQ", "2001-2002")

## Downloading PAQ\_B.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/PAQ\_B.XPT

paq\_03 = nhanes\_load\_data("PAQ", "2003-2004")

## Downloading PAQ\_C.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/PAQ\_C.XPT

diq\_99 = nhanes\_load\_data("DIQ", "1999-2000")

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.  
## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Downloading DIQ.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/DIQ.XPT

diq\_01 = nhanes\_load\_data("DIQ", "2001-2002")

## Downloading DIQ\_B.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/DIQ\_B.XPT

diq\_03 = nhanes\_load\_data("DIQ", "2003-2004")

## Downloading DIQ\_C.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/DIQ\_C.XPT

smq\_99 = nhanes\_load\_data("SMQ", "1999-2000")

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.  
## Cycle 1999-2000 doesn't always follow the normal naming conventions, so skipping the file suffix check.

## Downloading SMQ.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/SMQ.XPT

smq\_01 = nhanes\_load\_data("SMQ", "2001-2002")

## Downloading SMQ\_B.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/SMQ\_B.XPT

smq\_03 = nhanes\_load\_data("SMQ", "2003-2004")

## Downloading SMQ\_C.XPT to /var/folders/5j/2zqbq7jx49g71545t3lq5bfh0000gn/T//RtmpS6G2rU/SMQ\_C.XPT

# Joining  
data\_99 =   
 left\_join(demo\_99, bmx\_99, by = "SEQN") %>%   
 left\_join(bpx\_99, by = "SEQN") %>%   
 left\_join(paq\_99, by = "SEQN") %>%   
 left\_join(diq\_99, by = "SEQN") %>%   
 left\_join(smq\_99, by = "SEQN") %>%   
 select("SEQN", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "INDHHINC", "BMXBMI", "BMXHT", "BMXWT", "BPXPLS", "DIQ010", "PAQ180", "SMQ020")   
  
data\_01 =   
 left\_join(demo\_01, bmx\_01, by = "SEQN") %>%   
 left\_join(bpx\_01, by = "SEQN") %>%   
 left\_join(paq\_01, by = "SEQN") %>%   
 left\_join(diq\_01, by = "SEQN") %>%   
 left\_join(smq\_01, by = "SEQN") %>%   
 select("SEQN", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "INDHHINC", "BMXBMI", "BMXHT", "BMXWT", "BPXPLS", "DIQ010", "PAQ180", "SMQ020")   
  
data\_03 =   
 left\_join(demo\_03, bmx\_03, by = "SEQN") %>%   
 left\_join(bpx\_03, by = "SEQN") %>%   
 left\_join(paq\_03, by = "SEQN") %>%   
 left\_join(diq\_03, by = "SEQN") %>%   
 left\_join(smq\_03, by = "SEQN") %>%   
 select("SEQN", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "INDHHINC", "BMXBMI", "BMXHT", "BMXWT", "BPXPLS", "DIQ010", "PAQ180", "SMQ020")   
  
final\_data = bind\_rows(data\_99, data\_01, data\_03) %>%   
 janitor::clean\_names() %>%   
 filter(diq010 %in% c(1,2)) %>%   
 mutate(diq010 = as.factor(diq010),   
 riagendr = as.factor(riagendr),   
 ridreth1 = as.factor(ridreth1),   
 dmdeduc2 = as.factor(dmdeduc2),   
 indhhinc = as.factor(indhhinc),   
 paq180 = as.factor(paq180),   
 smq020 = as.factor(smq020)) %>%   
 drop\_na()

Partitioning:

training\_data = final\_data$diq010 %>% createDataPartition(p = 0.7, list = F)  
train\_data = final\_data[training\_data, ]  
test\_data = final\_data[-training\_data, ]  
  
  
#Store outcome   
#alc\_cons\_train = train\_data$alc\_consumption  
#alc\_cons\_test = test\_data$alc\_consumption  
  
# store matrices excluding outcome  
#train = model.matrix(alc\_consumption~., train\_data)[,-1]  
#test = model.matrix(alc\_consumption~., test\_data)[,-1]

## Number 2: Prediction Models

Construct three prediction models to predict diabetes using the 11 features from NHANES. You will use the following three algorithms to create your prediction models:

1. Classification Tree
2. Support Vector Classifier (i.e. Support Vector Machine with a linear classifier)
3. Logistic regression.

## Classification Tree

train\_control = trainControl(method = "cv", number = 10)  
grid\_2 = expand.grid(cp = seq(0.001, 0.4, by = 0.01))  
tree\_diabetes = train(diq010 ~., data = train\_data, method = "rpart", trControl = train\_control, tuneGrid = grid\_2)  
tree\_diabetes$bestTune

## cp  
## 40 0.391

tree\_diabetes

## CART   
##   
## 8244 samples  
## 12 predictor  
## 2 classes: '1', '2'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 7419, 7420, 7419, 7420, 7420, 7420, ...   
## Resampling results across tuning parameters:  
##   
## cp Accuracy Kappa   
## 0.001 0.8835512 0.118323  
## 0.011 0.9025961 0.000000  
## 0.021 0.9025961 0.000000  
## 0.031 0.9025961 0.000000  
## 0.041 0.9025961 0.000000  
## 0.051 0.9025961 0.000000  
## 0.061 0.9025961 0.000000  
## 0.071 0.9025961 0.000000  
## 0.081 0.9025961 0.000000  
## 0.091 0.9025961 0.000000  
## 0.101 0.9025961 0.000000  
## 0.111 0.9025961 0.000000  
## 0.121 0.9025961 0.000000  
## 0.131 0.9025961 0.000000  
## 0.141 0.9025961 0.000000  
## 0.151 0.9025961 0.000000  
## 0.161 0.9025961 0.000000  
## 0.171 0.9025961 0.000000  
## 0.181 0.9025961 0.000000  
## 0.191 0.9025961 0.000000  
## 0.201 0.9025961 0.000000  
## 0.211 0.9025961 0.000000  
## 0.221 0.9025961 0.000000  
## 0.231 0.9025961 0.000000  
## 0.241 0.9025961 0.000000  
## 0.251 0.9025961 0.000000  
## 0.261 0.9025961 0.000000  
## 0.271 0.9025961 0.000000  
## 0.281 0.9025961 0.000000  
## 0.291 0.9025961 0.000000  
## 0.301 0.9025961 0.000000  
## 0.311 0.9025961 0.000000  
## 0.321 0.9025961 0.000000  
## 0.331 0.9025961 0.000000  
## 0.341 0.9025961 0.000000  
## 0.351 0.9025961 0.000000  
## 0.361 0.9025961 0.000000  
## 0.371 0.9025961 0.000000  
## 0.381 0.9025961 0.000000  
## 0.391 0.9025961 0.000000  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was cp = 0.391.

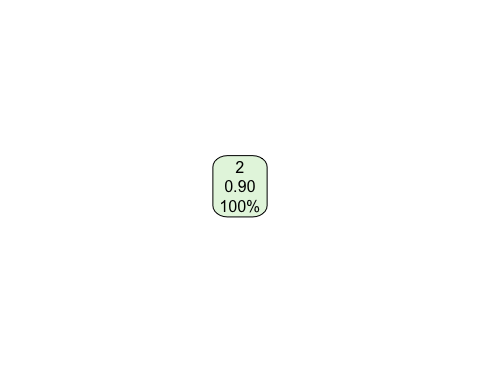
varImp(tree\_diabetes)

## rpart variable importance

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning  
## -Inf

## only 20 most important variables shown (out of 39)  
##   
## Overall  
## dmdeduc29 NaN  
## indhhinc4 NaN  
## indhhinc2 NaN  
## smq0209 NaN  
## paq1802 NaN  
## bmxwt NaN  
## paq1809 NaN  
## bmxbmi NaN  
## indhhinc9 NaN  
## indhhinc99 NaN  
## smq0202 NaN  
## dmdeduc22 NaN  
## ridreth15 NaN  
## riagendr2 NaN  
## indhhinc10 NaN  
## paq1804 NaN  
## indhhinc8 NaN  
## dmdeduc25 NaN  
## ridreth12 NaN  
## bpxpls NaN

rpart.plot(tree\_diabetes$finalModel)



## Support Vector Classifier

svm\_diabetes = svm(diq010 ~ ., data = train\_data, kernel="linear", cost=1, scale=TRUE)  
print(svm\_diabetes)

##   
## Call:  
## svm(formula = diq010 ~ ., data = train\_data, kernel = "linear",   
## cost = 1, scale = TRUE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: linear   
## cost: 1   
##   
## Number of Support Vectors: 2069