Decision Tree and Random Forest on Data with Missing Values

Imputation

Patricia Reynoso

May 22, 2020

Chapter 8 Exercises 8.1,

Run Models

3. Decision Tree, using c5.0,

4. Random Forest,

using training and test datasets, as described in Exercise 8.3.

3. Decision Tree, using c5.0,

```
library(dplyr)
library(NHANES)
library(tidyverse)
library(C50)
```

1. Data Readying

```
NHANESb<-NHANES%>%
  select(-ID)%>%  #Drop id variable
  mutate(SleepTrouble = as.numeric(NHANES$SleepTrouble) - 1)%>%
  drop_na(SleepTrouble)  # drop rows with missing values for sleep trouble
```

```
# Drop variables with more than 50% of missing values
NHANES2 <- NHANESb[, colMeans(is.na(NHANESb)) < 0.5]
```

```
# create a subset of numeric variables
NHANES3<- NHANES2[,sapply(NHANES2, is.numeric)]
```

```
#Imputate the mean in the missing values of numeric variables
NHANES4 <-data.frame(
  sapply(
    NHANES3,
    function(x ) ifelse(is.na(x ),
                        mean(x , na.rm = TRUE),
                        x )))
```

```
# subset of categorical variables
NHANESFACT<-NHANES2[,sapply(NHANES2, is.factor)]
```

```
#Bind the imputed numerical subset to the categorical vars subset.
NHANESCLEAN <- cbind(NHANES4, NHANESFACT)
```

```
saveRDS(NHANESCLEAN, file="NHANESCLEAN.RDS")
```

```r
nhanes <- readRDS("NHANESCLEAN.RDS")
nhanes <- nhanes%>%
            mutate(SleepTrouble = as.factor(SleepTrouble))
```

```r
#Build the train data set with 75% of the data
n<-nrow(nhanes)
test_idx <- sample.int(n, size = round(0.25*n))
train1 <- nhanes[-test_idx, ]

#Build the test data set with the remaining 25%
test1 <- nhanes[test_idx, ]
```

```r
# Define model's formula
form <- as.formula("SleepTrouble ~ BPSys1 + BPSys2 +
                DaysPhysHlthBad + DaysMentHlthBad + Depressed +
                Gender +
                HomeRooms  + HomeOwn  + HealthGen + HardDrugs +
                LittleInterest +
                PhysActive +
                Race1 +
                SurveyYr + Smoke100 + SexEver + SexNumPartnLife +
                TotChol +
                Work + Weight")
```

Model 3. Decision Tree, using c5.0.

```
model<- C5.0(formula= form, data = train1)
```

```
summary(model)
```

```
##
## Call:
## C5.0.formula(formula = form, data = train1)
##
##
## C5.0 [Release 2.07 GPL Edition]      Fri May 22 00:37:58 2020
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 5829 cases (21 attributes) from undefined.data
##
## Decision tree:
##
## DaysPhysHlthBad > 6:
## :...DaysMentHlthBad > 14:
## :   :...Smoke100 = No:
## :   :   :...HomeRooms > 9: 1 (6)
## :   :   :   HomeRooms <= 9:
## :   :   :   :...Race1 = Black: 1 (11.9/3.9)
## :   :   :       Race1 = Other: 0 (5)
## :   :   :       Race1 = Hispanic:
## :   :   :       :...HomeRooms <= 5: 1 (4/1)
## :   :   :       :   HomeRooms > 5: 0 (5)
## :   :   :       Race1 = Mexican:
## :   :   :       :...HomeOwn = Own: 1 (7/2)
## :   :   :       :   HomeOwn in {Rent,Other}: 0 (3)
## :   :   :       Race1 = White:
## :   :   :       :...TotChol <= 5.12: 0 (13.4/1)
## :   :   :           TotChol > 5.12:
## :   :   :           :...HomeRooms <= 4: 1 (11)
## :   :   :               HomeRooms > 4:
## :   :   :               :...Gender = male: 0 (6)
## :   :   :                   Gender = female:
## :   :   :                   :...Work in {Looking,NotWorking}: 1 (6.9/1)
## :   :   :                       Work = Working: 0 (3)
## :   :   Smoke100 = Yes:
## :   :   :...HardDrugs = Yes: 1 (41.9/4.9)
## :   :       HardDrugs = No:
## :   :       :...Weight > 104.2: 0 (8.4/1.2)
## :   :           Weight <= 104.2:
## :   :           :...Work = Looking: 0 (0.6)
## :   :               Work = NotWorking:
## :   :               :...Gender = female: 1 (22)
## :   :               :   Gender = male:
## :   :               :   :...BPSys1 <= 118: 0 (4.2)
## :   :               :       BPSys1 > 118: 1 (10.9/1.2)
## :   :               Work = Working:
## :   :               :...Race1 = Mexican: 1 (0)
```

```
## :   :                   Race1 in {Black,Hispanic,Other}: 0 (4)
## :   :                   Race1 = White:
## :   :                   :...SurveyYr = 2009_10: 1 (13.2/0.6)
## :   :                       SurveyYr = 2011_12:
## :   :                       :...Weight <= 68.5: 0 (3)
## :   :                           Weight > 68.5: 1 (2.6)
## :   DaysMentHlthBad <= 14:
## :   :...Work = Looking:
## :       :...SexNumPartnLife <= 4: 0 (6/1)
## :       :   SexNumPartnLife > 4:
## :       :   :...BPSys2 <= 128: 1 (13)
## :       :       BPSys2 > 128: 0 (2)
## :       Work = Working:
## :       :...HomeOwn in {Rent,Other}: 0 (97/16)
## :       :   HomeOwn = Own:
## :       :   :...Weight > 117:
## :       :       :...HealthGen in {Excellent,Vgood,Fair,Poor}: 1 (12)
## :       :       :   HealthGen = Good:
## :       :       :   :...DaysPhysHlthBad <= 27: 0 (6)
## :       :       :       DaysPhysHlthBad > 27: 1 (3)
## :       :       Weight <= 117:
## :       :       :...Depressed = Several: 1 (26.8/11.6)
## :       :           Depressed = Most: 0 (4.1/1)
## :       :           Depressed = None:
## :       :           :...HomeRooms <= 5:
## :       :               :...PhysActive = Yes: 0 (16.8/0.8)
## :       :               :   PhysActive = No:
## :       :               :   :...DaysMentHlthBad <= 5: 1 (22/5)
## :       :               :       DaysMentHlthBad > 5: 0 (3)
## :       :               HomeRooms > 5:
## :       :               :...BPSys2 > 102: 0 (84.3/7)
## :       :                   BPSys2 <= 102:
## :       :                   :...HealthGen in {Excellent,Fair,Poor}: 1 (7)
## :       :                       HealthGen in {Vgood,Good}: 0 (8/1)
## :       Work = NotWorking:
## :       :...TotChol > 6.13:
## :           :...HealthGen in {Excellent,Vgood}: 0 (4)
## :           :   HealthGen in {Good,Fair,Poor}:
## :           :   :...BPSys1 <= 158: 1 (34/5)
## :           :       BPSys1 > 158: 0 (3)
## :           TotChol <= 6.13:
## :           :...Smoke100 = No:
## :               :...Race1 = Mexican: 0 (13.4/1)
## :               :   Race1 = Hispanic:
## :               :   :...TotChol <= 4.34: 1 (4.9/0.9)
## :               :   :   TotChol > 4.34: 0 (4)
## :               :   Race1 = Other:
## :               :   :...DaysPhysHlthBad <= 22: 1 (3.9/0.9)
## :               :   :   DaysPhysHlthBad > 22: 0 (6)
## :               :   Race1 = Black:
## :               :   :...LittleInterest = Several: 0 (4.7)
## :               :   :   LittleInterest = Most: 1 (1)
## :               :   :   LittleInterest = None:
## :               :   :   :...BPSys1 <= 120: 0 (2.2)
```

```
## :               :   :       BPSys1 > 120: 1 (3)
## :               :   Race1 = White:
## :               :   :...LittleInterest = Most: 1 (4.6/0.6)
## :               :       LittleInterest in {None,Several}:
## :               :       :...TotChol > 5.66: 1 (7/1)
## :               :           TotChol <= 5.66:
## :               :           :...HomeRooms <= 4: 1 (6/2)
## :               :               HomeRooms > 4: 0 (52.3/6)
## :               Smoke100 = Yes:
## :               :...BPSys2 <= 102: 1 (19.6/3.6)
## :                   BPSys2 > 102:
## :                   :...SurveyYr = 2011_12: 1 (41.1/17.1)
## :                       SurveyYr = 2009_10:
## :                       :...HealthGen in {Excellent,Vgood}: 0 (11)
## :                           HealthGen = Poor:
## :                           :...HomeOwn in {Own,Other}: 0 (5.7/1.7)
## :                           :   HomeOwn = Rent: 1 (2.3)
## :                           HealthGen = Good:
## :                           :...Race1 in {Black,Other}: 1 (6.1/1.1)
## :                           :   Race1 in {Hispanic,Mexican}: 0 (3.5/1)
## :                           :   Race1 = White:
## :                           :   :...Weight <= 109.8: 0 (15.1/1)
## :                           :       Weight > 109.8: 1 (2)
## :                           HealthGen = Fair:
## :                           :...TotChol <= 4.5: 0 (11.5)
## :                               TotChol > 4.5:
## :                               :...SexNumPartnLife <= 10: 0 (4)
## :                                   SexNumPartnLife > 10: 1 (7/1)
## DaysPhysHlthBad <= 6:
## :...Race1 in {Hispanic,Mexican,Other}: 0 (1125/119)
##     Race1 in {Black,White}:
##     :...DaysMentHlthBad > 0:
##         :...DaysMentHlthBad > 29:
##         :   :...Smoke100 = Yes:
##         :   :   :...HomeRooms <= 2: 0 (5)
##         :   :   :   HomeRooms > 2: 1 (70.4/24.4)
##         :   :   Smoke100 = No:
##         :   :   :...LittleInterest in {None,Several}: 0 (54/13.7)
##         :   :       LittleInterest = Most:
##         :   :       :...Gender = male: 1 (9)
##         :   :           Gender = female:
##         :   :           :...BPSys2 <= 116: 1 (7.6/0.3)
##         :   :               BPSys2 > 116: 0 (8)
##         :   DaysMentHlthBad <= 29:
##         :   :...Gender = female:
##         :       :...HardDrugs = Yes:
##         :       :   :...SexNumPartnLife <= 8:
##         :       :   :   :...DaysPhysHlthBad > 4: 0 (6/1)
##         :       :   :   :   DaysPhysHlthBad <= 4:
##         :       :   :   :   :...DaysPhysHlthBad > 2: 1 (21)
##         :       :   :   :       DaysPhysHlthBad <= 2:
##         :       :   :   :       :...Depressed in {None,Most}: 0 (12/4)
##         :       :   :   :           Depressed = Several: 1 (7/1)
##         :       :   :   SexNumPartnLife > 8:
```

```
##         :       :   :   :...SexNumPartnLife > 28:
##         :       :   :       :...TotChol <= 4.16: 0 (4)
##         :       :   :       :   TotChol > 4.16: 1 (17/2)
##         :       :   :       SexNumPartnLife <= 28:
##         :       :   :       :...HealthGen in {Excellent,Fair,
##         :       :   :           :             Poor}: 0 (19.4/4)
##         :       :   :           HealthGen = Vgood:
##         :       :   :           :...LittleInterest = None: 0 (26.2/4.7)
##         :       :   :           :   LittleInterest = Most: 1 (4.1/1.4)
##         :       :   :           :   LittleInterest = Several:
##         :       :   :           :   :...SexNumPartnLife <= 20: 0 (10.4/1.9)
##         :       :   :           :       SexNumPartnLife > 20: 1 (3)
##         :       :   :           HealthGen = Good:
##         :       :   :           :...SexNumPartnLife > 15: 0 (31.3/10.1)
##         :       :   :               SexNumPartnLife <= 15:
##         :       :   :               :...Depressed in {None,Most}: 1 (4)
##         :       :   :                   Depressed = Several: [S1]
##         :       :   HardDrugs = No:
##         :       :   :...HealthGen in {Excellent,Vgood,Poor}: 0 (461.5/120.6)
##         :       :       HealthGen in {Good,Fair}:
##         :       :       :...SurveyYr = 2009_10:
##         :       :           :...HomeRooms <= 6.171381: 0 (130.1/54.2)
##         :       :           :   HomeRooms > 6.171381:
##         :       :           :   :...DaysPhysHlthBad > 3.442454: 1 (5/1)
##         :       :           :       DaysPhysHlthBad <= 3.442454:
##         :       :           :       :...Smoke100 = No: 0 (51.5/4)
##         :       :           :           Smoke100 = Yes:
##         :       :           :           :...BPSys2 > 114: 0 (7.3)
##         :       :           :               BPSys2 <= 114:
##         :       :           :               :...TotChol <= 4.89: 0 (2.9)
##         :       :           :                   TotChol > 4.89: 1 (6.4/0.3)
##         :       :           SurveyYr = 2011_12:
##         :       :           :...Depressed = Most: 1 (14.5/4.9)
##         :       :               Depressed = Several:
##         :       :               :...DaysPhysHlthBad > 3: 0 (15.3/4.6)
##         :       :               :   DaysPhysHlthBad <= 3:
##         :       :               :   :...PhysActive = Yes: 1 (24.5/3)
##         :       :               :       PhysActive = No:
##         :       :               :       :...BPSys1 <= 124: 0 (6.2)
##         :       :               :           BPSys1 > 124: 1 (9.5/0.8)
##         :       :               Depressed = None:
##         :       :               :...BPSys2 <= 120.6054:
##         :       :                   :...SexNumPartnLife <= 17: 0 (62.6/27.3)
##         :       :                   :   SexNumPartnLife > 17: 1 (7)
##         :       :                   BPSys2 > 120.6054:
##         :       :                   :...HomeRooms > 6.171381: 0 (23.6/0.2)
##         :       :                       HomeRooms <= 6.171381:
##         :       :                       :...PhysActive = Yes: 1 (5.1/1.2)
##         :       :                           PhysActive = No:
##         :       :                           :...DaysPhysHlthBad <= 3.442454: 0 (16.5/1.7)
##         :       :                               DaysPhysHlthBad > 3.442454: 1 (2.2)
##         :       Gender = male:
##         :       :...Smoke100 = No: 0 (319.9/56.3)
##         :           Smoke100 = Yes:
```

```
##         :           :...HomeOwn = Other: 0 (12.7/1.1)
##         :               HomeOwn = Rent:
##         :               :...DaysMentHlthBad <= 1: 0 (16.1)
##         :               :   DaysMentHlthBad > 1:
##         :               :   :...DaysPhysHlthBad <= 3:
##         :               :       :...TotChol <= 5.35: 0 (57.7/26)
##         :               :       :   TotChol > 5.35: 1 (19/2)
##         :               :       DaysPhysHlthBad > 3:
##         :               :       :...SexNumPartnLife <= 17: 0 (36.2/7.5)
##         :               :           SexNumPartnLife > 17: 1 (3)
##         :               HomeOwn = Own:
##         :               :...HomeRooms <= 5: 0 (50.1/3)
##         :                   HomeRooms > 5:
##         :                   :...HealthGen in {Excellent,Vgood,
##         :                       :             Poor}: 0 (81.3/11.6)
##         :                       HealthGen in {Good,Fair}:
##         :                       :...SurveyYr = 2011_12:
##         :                           :...Depressed = Most: 1 (0)
##         :                           :   Depressed = None:
##         :                           :   :...HomeRooms <= 9: 1 (19.7/2.8)
##         :                           :   :   HomeRooms > 9: 0 (2.4)
##         :                           :   Depressed = Several:
##         :                           :   :...Weight <= 81.8: 1 (8.3/2.7)
##         :                           :       Weight > 81.8: 0 (8.6/0.2)
##         :                           SurveyYr = 2009_10:
##         :                           :...Weight > 108.8: 1 (9.1/1.5)
##         :                               Weight <= 108.8:
##         :                               :...Weight > 82.2: 0 (26.6/1.5)
##         :                                   Weight <= 82.2:
##         :                                   :...DaysMentHlthBad <= 3: 1 (9/1)
##         :                                       DaysMentHlthBad > 3: 0 (10.3/1.5)
##         DaysMentHlthBad <= 0:
##         :...LittleInterest = Most:
##             :...Gender = male: 0 (14.4)
##             :   Gender = female:
##             :   :...Smoke100 = No: 1 (7.1/1)
##             :       Smoke100 = Yes: 0 (2.3)
##             LittleInterest = Several:
##             :...Work in {Looking,Working}: 0 (92.8/21.2)
##             :   Work = NotWorking:
##             :   :...HealthGen in {Excellent,Fair}: 0 (19.7/6)
##             :       HealthGen = Poor: 1 (1)
##             :       HealthGen = Vgood:
##             :       :...TotChol <= 5.04: 0 (5.6/0.1)
##             :       :   TotChol > 5.04: 1 (8.4/0.2)
##             :       HealthGen = Good:
##             :       :...DaysPhysHlthBad > 2: 1 (5.4/0.4)
##             :           DaysPhysHlthBad <= 2:
##             :           :...Gender = male: 0 (10.7)
##             :               Gender = female:
##             :               :...Weight <= 77.9: 0 (5.8)
##             :                   Weight > 77.9: 1 (6.2/0.2)
##             LittleInterest = None:
##             :...Race1 = Black: 0 (277.5/34)
```

```
##                 Race1 = White:
##                 :...Smoke100 = Yes:
##                     :...Gender = male:
##                     :   :...HealthGen in {Vgood,Poor}: 0 (158/22.7)
##                     :   :   HealthGen = Excellent:
##                     :   :   :...BPSys1 <= 104: 1 (3.4/0.4)
##                     :   :   :   BPSys1 > 104: 0 (48.5/3)
##                     :   :   HealthGen = Fair:
##                     :   :   :...Work = Looking: 1 (1)
##                     :   :   :   Work = NotWorking: 0 (23/1)
##                     :   :   :   Work = Working:
##                     :   :   :   :...BPSys2 > 116: 0 (12/1)
##                     :   :   :       BPSys2 <= 116:
##                     :   :   :       :...SexNumPartnLife <= 15.08507: 1 (6.4/0.4)
##                     :   :   :           SexNumPartnLife > 15.08507: 0 (2)
##                     :   :   HealthGen = Good:
##                     :   :   :...SexNumPartnLife <= 2: 0 (21.3)
##                     :   :       SexNumPartnLife > 2:
##                     :   :       :...BPSys2 > 148: 0 (16)
##                     :   :           BPSys2 <= 148:
##                     :   :           :...Work = Looking: 0 (7.6)
##                     :   :               Work = Working:
##                     :   :               :...TotChol <= 5.66: 0 (75.3/12)
##                     :   :               :   TotChol > 5.66:
##                     :   :               :   :...DaysPhysHlthBad <= 4: 0 (24/10)
##                     :   :               :       DaysPhysHlthBad > 4: 1 (4)
##                     :   :               Work = NotWorking:
##                     :   :               :...Weight <= 85.7: 1 (20.9/5.9)
##                     :   :                   Weight > 85.7:
##                     :   :                   :...SurveyYr = 2009_10: 0 (15.8/1)
##                     :   :                       SurveyYr = 2011_12:
##                     :   :                       :...Weight <= 98.6: 0 (5)
##                     :   :                           Weight > 98.6: 1 (7/2)
##                     :   Gender = female:
##                     :   :...HomeOwn = Other: 0 (5.1)
##                     :       HomeOwn = Rent:
##                     :       :...SurveyYr = 2009_10:
##                     :       :   :...HomeRooms <= 3: 1 (4/1)
##                     :       :   :   HomeRooms > 3: 0 (21.8/1)
##                     :       :   SurveyYr = 2011_12:
##                     :       :   :...PhysActive = No: 1 (14.1/1.1)
##                     :       :       PhysActive = Yes:
##                     :       :       :...BPSys1 <= 108: 1 (4.1/1.1)
##                     :       :           BPSys1 > 108: 0 (5)
##                     :       HomeOwn = Own:
##                     :       :...Weight > 95.8:
##                     :           :...BPSys1 <= 112: 0 (8)
##                     :           :   BPSys1 > 112:
##                     :           :   :...BPSys1 <= 134: 1 (19.4/3.4)
##                     :           :       BPSys1 > 134: 0 (3)
##                     :           Weight <= 95.8:
##                     :           :...Work = Looking:
##                     :               :...BPSys2 <= 114: 0 (3.2/0.4)
##                     :               :   BPSys2 > 114: 1 (2)
```

```
##                       :               Work = Working:
##                       :               :...SexNumPartnLife <= 27: 0 (95.4/8.4)
##                       :               :   SexNumPartnLife > 27: 1 (7/2)
##                       :               Work = NotWorking:
##                       :               :...DaysPhysHlthBad > 2: 0 (6)
##                       :                   DaysPhysHlthBad <= 2:
##                       :                   :...HealthGen in {Excellent,Fair,
##                       :                       :             Poor}: 0 (10.4/1)
##                       :                       HealthGen = Good:
##                       :                       :...HomeRooms <= 5: 0 (8)
##                       :                       :   HomeRooms > 5:
##                       :                       :   :...HomeRooms <= 10: 1 (11)
##                       :                       :       HomeRooms > 10: 0 (3)
##                       :                       HealthGen = Vgood:
##                       :                       :...SurveyYr = 2009_10: 0 (11.9/1)
##                       :                           SurveyYr = 2011_12: [S2]
##                     Smoke100 = No:
##                     :...Work = Looking:
##                         :...SexNumPartnLife <= 2: 1 (10/1)
##                         :   SexNumPartnLife > 2:
##                         :   :...Gender = male: 0 (11.9)
##                         :       Gender = female:
##                         :       :...SexNumPartnLife <= 12: 0 (9.6/1)
##                         :           SexNumPartnLife > 12: 1 (2.5)
##                         Work in {NotWorking,Working}:
##                         :...TotChol > 6.65: 0 (65/1)
##                             TotChol <= 6.65:
##                             :...HardDrugs = Yes:
##                                 :...SexNumPartnLife <= 1: 1 (5)
##                                 :   SexNumPartnLife > 1: 0 (51.8/12)
##                                 HardDrugs = No:
##                                 :...BPSys2 <= 112: 0 (234.4/22.3)
##                                     BPSys2 > 112:
##                                     :...HomeRooms > 10:
##                                         :...SexNumPartnLife > 17: 1 (6)
##                                         :   SexNumPartnLife <= 17:
##                                         :   :...SexNumPartnLife <= 2: 1 (9/2)
##                                         :       SexNumPartnLife > 2:
##                                         :       :...BPSys1 > 120: 0 (15.3)
##                                         :           BPSys1 <= 120: [S3]
##                                         HomeRooms <= 10:
##                                         :...Gender = female:
##                                             :...TotChol <= 5.56: 0 (142.8/22.1)
##                                             :   TotChol > 5.56: [S4]
##                                             Gender = male:
##                                             :...PhysActive = Yes:
##                                                 :...SurveyYr = 2011_12: 0 (87/3)
##                                                 :   SurveyYr = 2009_10: [S5]
##                                                 PhysActive = No: [S6]
##
## SubTree [S1]
##
## LittleInterest in {None,Most}: 0 (3)
## LittleInterest = Several: 1 (5/1)
```

```
##
## SubTree [S2]
##
## SexNumPartnLife <= 10: 1 (4)
## SexNumPartnLife > 10:
## :...TotChol <= 2.43: 1 (2)
##     TotChol > 2.43: 0 (14/2.4)
##
## SubTree [S3]
##
## SexNumPartnLife <= 6: 1 (3)
## SexNumPartnLife > 6: 0 (8.8/0.9)
##
## SubTree [S4]
##
## SexNumPartnLife > 12: 0 (13.9)
## SexNumPartnLife <= 12:
## :...TotChol <= 6.15: 1 (32/9)
##     TotChol > 6.15: 0 (12/1)
##
## SubTree [S5]
##
## DaysPhysHlthBad <= 1: 0 (64.5/4.9)
## DaysPhysHlthBad > 1:
## :...DaysPhysHlthBad <= 2: 1 (6/1)
##     DaysPhysHlthBad > 2: 0 (6)
##
## SubTree [S6]
##
## Work = NotWorking: 0 (22.5)
## Work = Working:
## :...TotChol > 4.11: 0 (43.6/4)
##     TotChol <= 4.11:
##     :...HealthGen in {Excellent,Fair}: 0 (2)
##         HealthGen in {Vgood,Poor}: 1 (5.9)
##         HealthGen = Good:
##         :...TotChol <= 3.57: 1 (2.8)
##             TotChol > 3.57: 0 (4)
##
##
## Evaluation on training data (5829 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##     202  847(14.5%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##    4254   130    (a): class 0
##     717   728    (b): class 1
##
```

```
##
##  Attribute usage:
##
##  100.00% DaysPhysHlthBad
##   90.72% Race1
##   80.70% DaysMentHlthBad
##   52.79% Gender
##   48.41% Smoke100
##   37.45% LittleInterest
##   36.30% Work
##   28.14% HealthGen
##   26.39% HomeRooms
##   25.29% BPSys2
##   25.01% TotChol
##   24.09% HardDrugs
##   21.36% SexNumPartnLife
##   18.91% SurveyYr
##   16.74% HomeOwn
##   11.80% Weight
##    6.93% PhysActive
##    6.21% Depressed
##    3.57% BPSys1
##
##
## Time: 0.2 secs
```

```
fitted.values.train <- predict(model, newdata = train1)
summary(fitted.values.train)
```

```
##    0    1
## 4971  858
```

```
fitted.values.test <- predict(model, newdata = test1)
summary(fitted.values.test)
```

```
##    0    1
## 1633  310
```

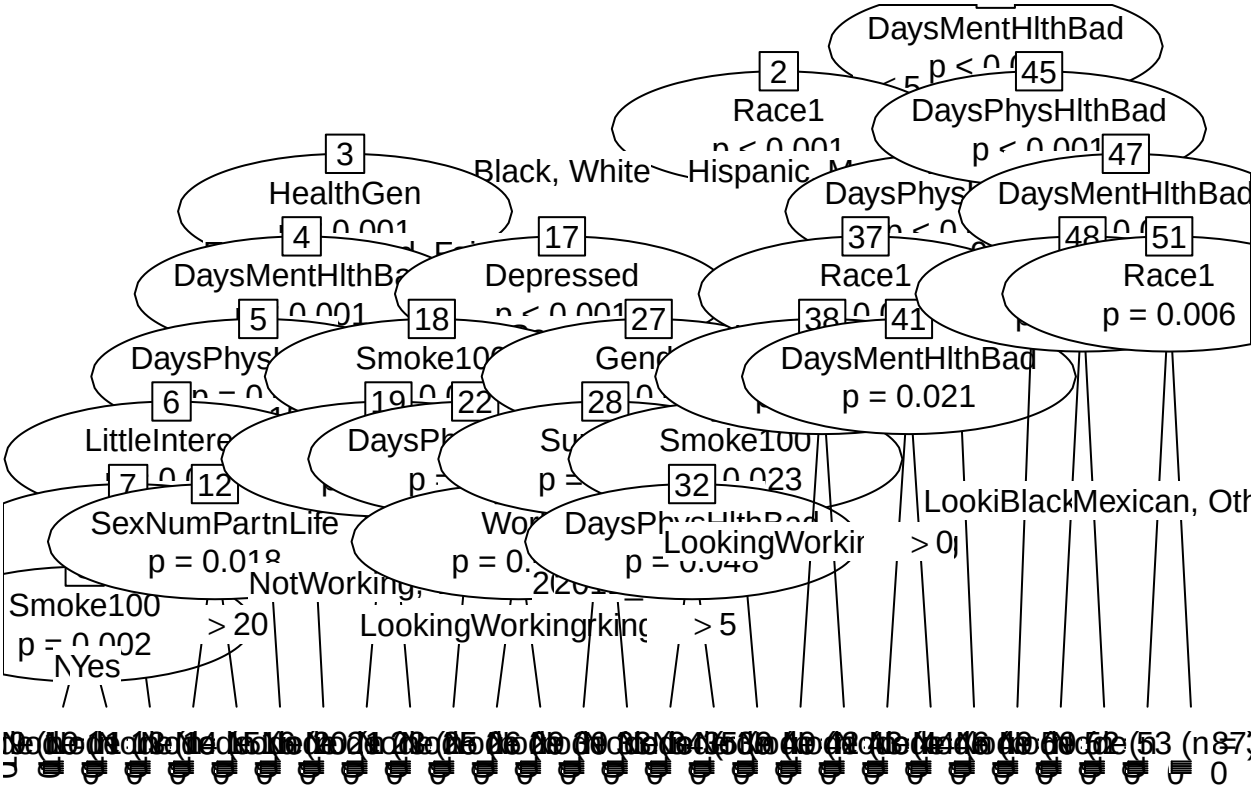
```
misClasificError_train <- mean(fitted.values.train != train1$SleepTrouble, na.rm=TRUE)
print(paste('Accuracy training data',1-misClasificError_train))
```

```
## [1] "Accuracy training data 0.854692056956596"
```

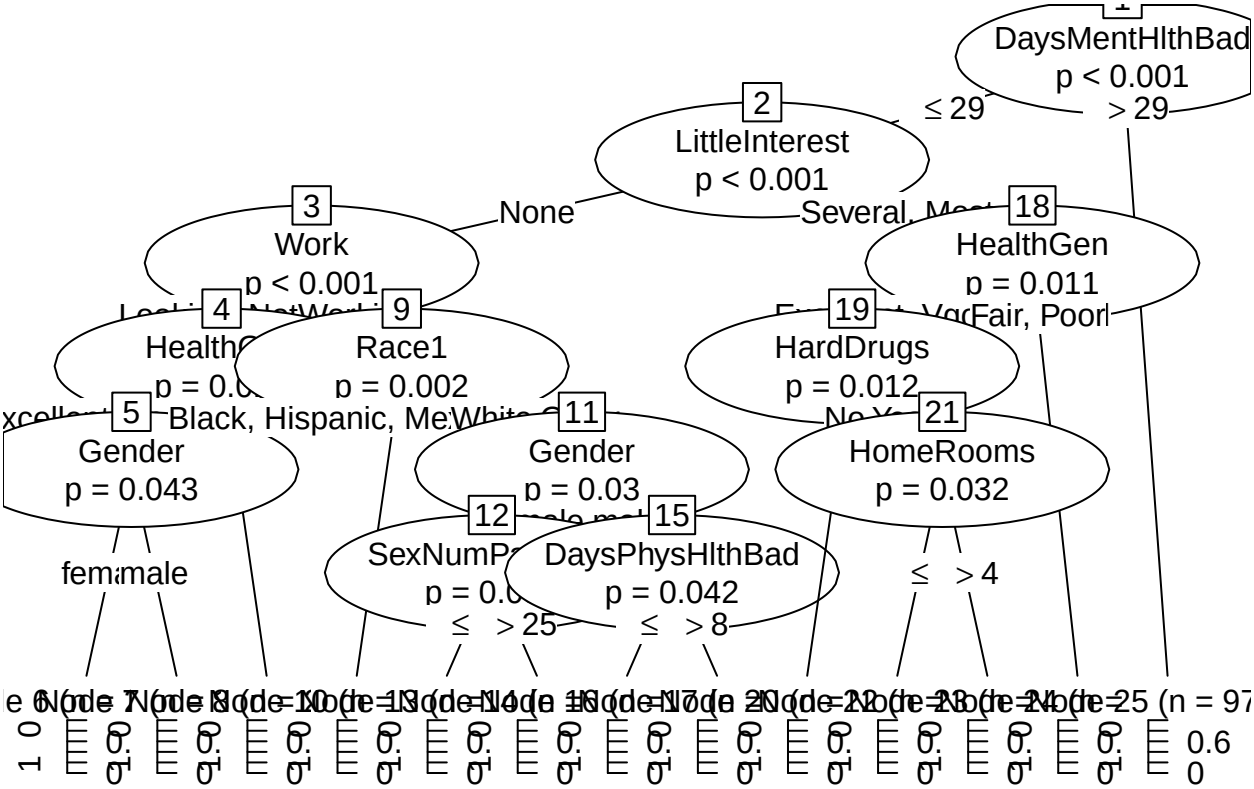
```
misClasificError_test <- mean(fitted.values.test != test1$SleepTrouble, na.rm=TRUE)
print(paste('Accuracy test data',1-misClasificError_test))
```

```
## [1] "Accuracy test data 0.778692743180649"
```

```
library(partykit)
mod <- ctree(formula = form, data=train1)
plot(mod)
```




```
library(partykit)
mod1 <- ctree(formula = form, data=test1)
plot(mod1)
```


DaysMentHlthBad
p < 0.001
≤ 29
> 29
2
LittleInterest
p < 0.001
None
3
Work
p < 0.001
18
HealthGen
p = 0.011
Fair, Poor
4
9
19
HardDrugs
p = 0.012
Race1
p = 0.002
5
Black, Hispanic, Me
11
21
Gender
p = 0.043
Gender
p = 0.03
HomeRooms
p = 0.032
12
15
DaysPhysHlthBad
p = 0.042
≤
> 4
≤
> 8
25 (n = 97

Model 3. Random Forest.

```
library(randomForest)
# Drop the rows with missing values in factor variables.
train2 <- na.omit(train1) # Drop missing values
nrow(train2)
```

```
## [1] 3175
```

```
mod_forest <- randomForest(form, data=train2, ntree=201, mtry = 3)
mod_forest
```

```
##
## Call:
##  randomForest(formula = form, data = train2, ntree = 201, mtry = 3)
##                Type of random forest: classification
##                      Number of trees: 201
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 10.2%
## Confusion matrix:
##      0   1 class.error
## 0 2331  47  0.01976451
## 1  277 520  0.34755332
```

```
sum(diag(mod_forest$confusion)) / nrow(train2)
```

```
## [1] 0.8979528
```

***The accuracy ratio is 0.89. This training data set was the result of imputation to numerical variables and the dropping of missing values in factor variables.

Next, I'll imputate both, categorical and numerical values and running Random Forest. To imputate numerical values we use the mean, and to imputate categorical values we use the function "rfImpute" from the "randomForest" package. Take a look at the accuracy results***

```
train1.na <- train1
set.seed(111)
## artificially drop some data values.
for (i in 1:4) train1.na[sample(150, sample(20)), i] <- NA
set.seed(222)
train1.imputed <- rfImpute(form, train1.na)
```

```
## ntree      OOB      1      2
##   300:  11.58%  1.82% 41.18%
## ntree      OOB      1      2
##   300:  11.77%  2.12% 41.04%
## ntree      OOB      1      2
##   300:  11.75%  2.10% 41.04%
## ntree      OOB      1      2
##   300:  12.03%  2.10% 42.15%
## ntree      OOB      1      2
##   300:  12.03%  2.24% 41.73%
```

```
set.seed(333)
train1.rf <- randomForest(form, train1.imputed)
print(train1.rf)
```

```
##
```

```
## Call:
##  randomForest(formula = form, data = train1.imputed)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 12.13%
## Confusion matrix:
##      0   1 class.error
## 0 4295  89  0.02030109
## 1  618 827  0.42768166
```

```
sum(diag(train1.rf$confusion)) / nrow(train1.imputed)
```

```
## [1] 0.8787099
```

```
#Important variables
library(tibble)
importance(mod_forest)%>%
   as.data.frame() %>%
   rownames_to_column()%>%
   arrange(desc(MeanDecreaseGini))
```

```
##            rowname MeanDecreaseGini
## 1           Weight        124.23657
## 2          TotChol        119.31426
## 3           BPSys2        105.31239
## 4           BPSys1         99.98942
## 5  SexNumPartnLife         97.49594
## 6  DaysMentHlthBad         88.92535
## 7        HomeRooms         79.22643
## 8  DaysPhysHlthBad         77.39406
## 9        HealthGen         57.19788
## 10           Race1         52.25554
## 11       Depressed         34.70834
## 12  LittleInterest         30.53060
## 13            Work         30.42944
## 14        Smoke100         25.83242
## 15         HomeOwn         25.19040
## 16          Gender         24.41139
## 17        SurveyYr         23.92057
## 18      PhysActive         20.98842
## 19       HardDrugs         18.35827
## 20         SexEver          3.28496
```

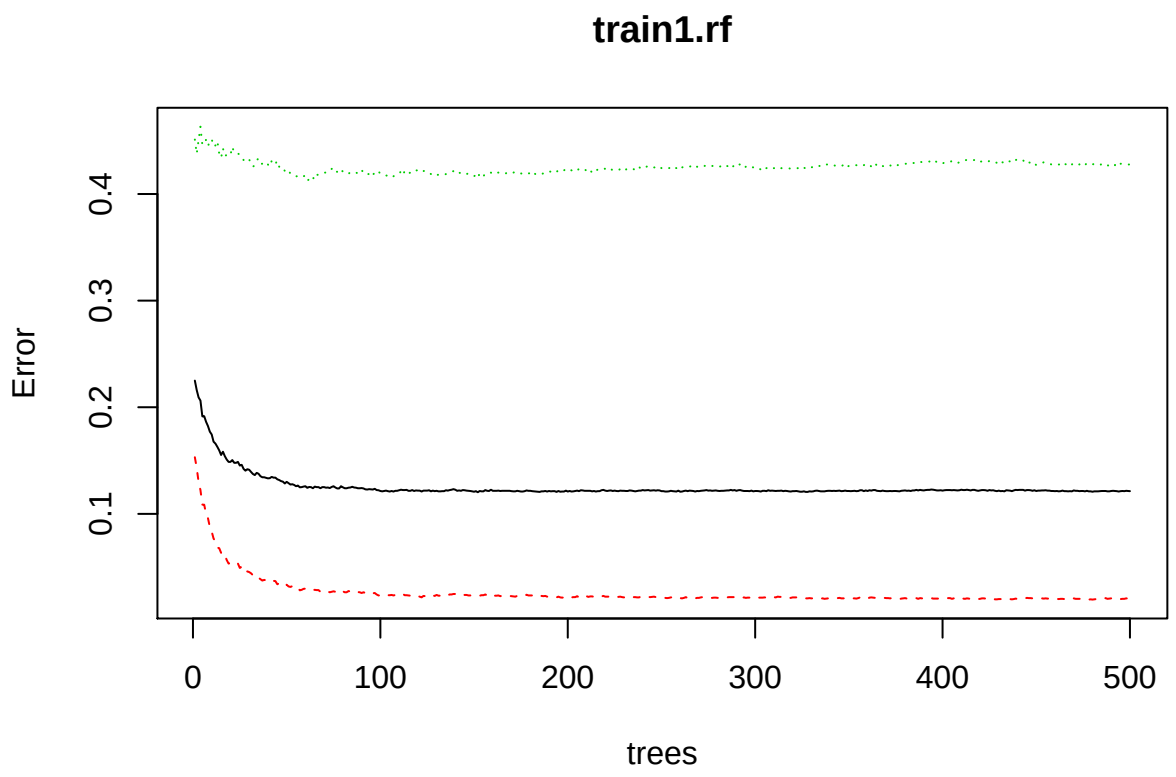
```
#predict(mod_forest)
```

```
importance(train1.rf)%>%
   as.data.frame() %>%
   rownames_to_column()%>%
   arrange(desc(MeanDecreaseGini))
```

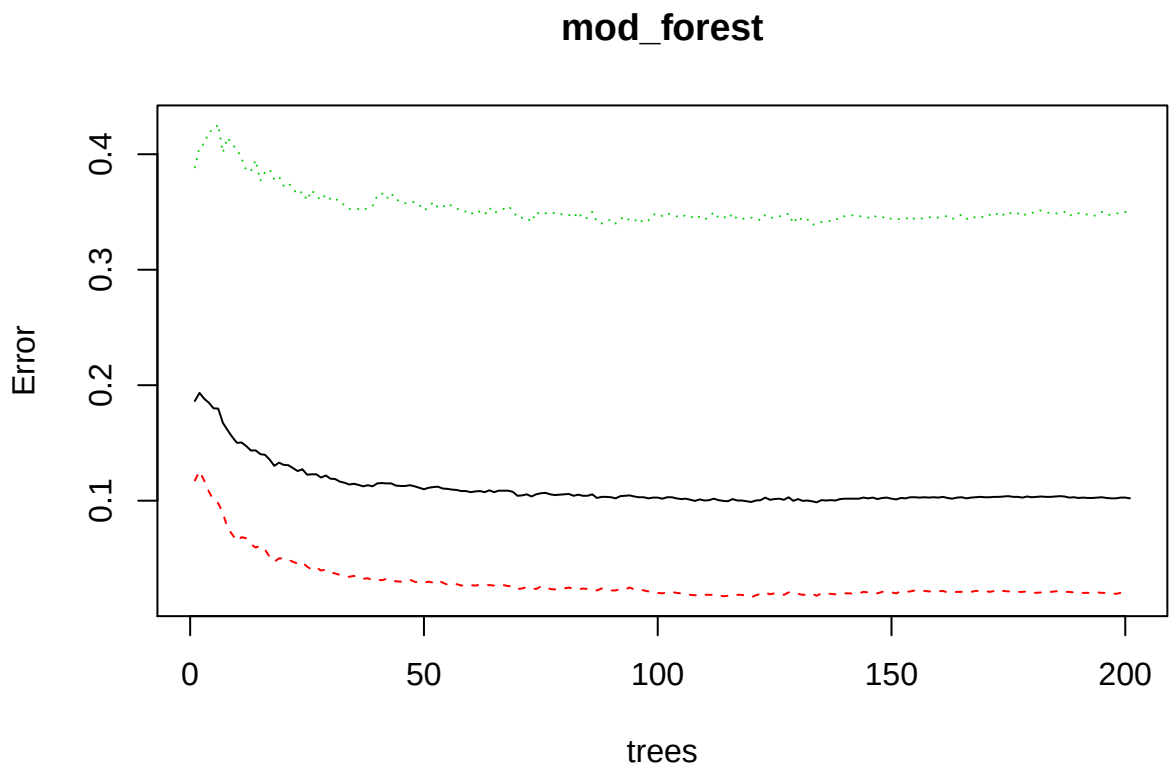
```
##            rowname MeanDecreaseGini
## 1           Weight        262.76904
## 2          TotChol        253.52611
## 3           BPSys2        201.42089
```

```
## 4           BPSys1        201.15551
## 5  SexNumPartnLife        164.20618
## 6        HomeRooms        148.46653
## 7  DaysMentHlthBad        148.24981
## 8  DaysPhysHlthBad        133.35324
## 9        HealthGen        106.09479
## 10           Race1         92.03215
## 11            Work         59.88439
## 12       Depressed         57.06796
## 13  LittleInterest         50.75014
## 14         HomeOwn         43.92478
## 15        SurveyYr         43.42683
## 16        Smoke100         42.18897
## 17          Gender         40.48254
## 18      PhysActive         37.63824
## 19       HardDrugs         34.27294
## 20         SexEver         19.96611
```

```
plot(train1.rf)
```




```
plot(mod_forest)
```




```
print(mod_forest)
```

```
##
## Call:
##  randomForest(formula = form, data = train2, ntree = 201, mtry = 3)
##                Type of random forest: classification
##                      Number of trees: 201
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 10.2%
## Confusion matrix:
##      0   1 class.error
## 0 2331  47  0.01976451
## 1  277 520  0.34755332
```

The error seems to be lowest at around 50 trees

```
test2 <- na.omit(test1)
nrow(test2)
```

```
## [1] 1068
```

```
mod_forest_test <- randomForest(form, data=test2, ntree=201, mtry = 3)
mod_forest_test
```

```
##
## Call:
##  randomForest(formula = form, data = test2, ntree = 201, mtry = 3)
```

```
##                Type of random forest: classification
##                      Number of trees: 201
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 17.7%
## Confusion matrix:
##     0   1 class.error
## 0 732  36    0.046875
## 1 153 147    0.510000
```

```
sum(diag(mod_forest_test$confusion)) / nrow(test2)
```

```
## [1] 0.8230337
```

```
plot(mod_forest_test)
```