Homework 7: NHANES Dataset Analysis

Julianne

April 03, 2017

Rcode available here: https://github.com/JulianneA/N741_Homework7

Pull subset of variables from NHANES Dataset

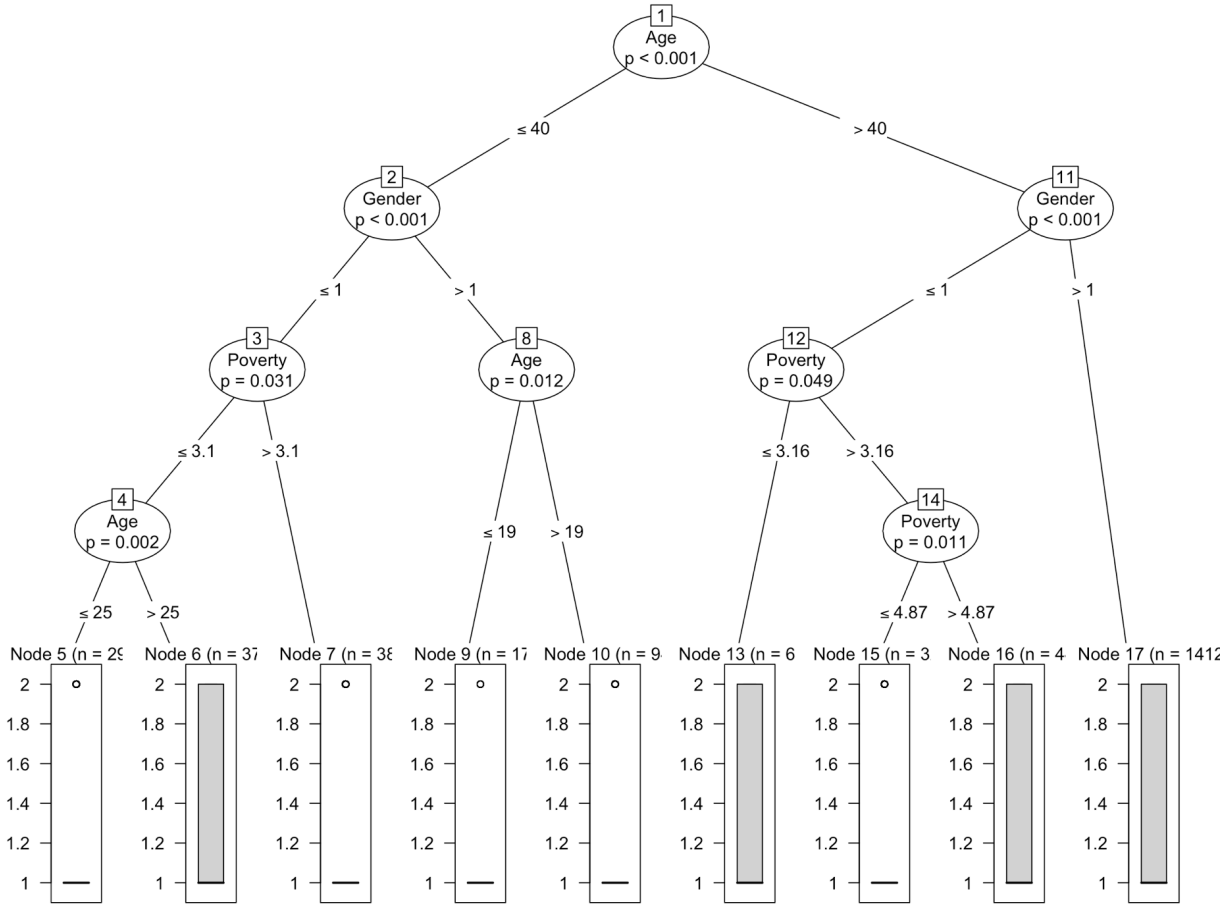
```
## [1] "tbl_df"     "tbl"        "data.frame"

## Observations: 10,000
## Variables: 4
## $ Age          <int> 34, 34, 34, 4, 49, 9, 8, 45, 45, 45, 66, 58, 54, ...
## $ Gender       <fctr> male, male, male, male, female, male, male, fema...
## $ SleepTrouble <fctr> Yes, Yes, Yes, NA, Yes, NA, NA, No, No, No, No, ...
## $ Poverty      <dbl> 1.36, 1.36, 1.36, 1.07, 1.91, 1.84, 2.33, 5.00, 5...

## Observations: 7,175
## Variables: 4
## $ Age          <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 58, 50, 3...
## $ Gender       <dbl> 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2, 2, 1, 1, 2...
## $ SleepTrouble <dbl> 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2...
## $ Poverty      <dbl> 1.36, 1.36, 1.36, 1.91, 5.00, 5.00, 5.00, 2.20, 5...
```

Run Models

```
## [1] 100

## [1] 99.34495

## [1] 98.89895

## [1] 96.58537

##
## knn.1    1    2
##     1 5336    0
##     2    0 1839

##
## knn.3    1    2
##     1 5333   44
##     2    3 1795

##
## knn.5    1    2
```

```
##     1 5334   77
##     2    2 1762

##
## knn.20    1    2
##      1 5336  245
##      2    0 1594

## Observations: 10,000
## Variables: 2
## $ Age     <dbl> 16.00000, 16.64646, 17.29293, 17.93939, 18.58586, 19.2...
## $ Poverty <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...

## Observations: 7,175
## Variables: 4
## $ Age          <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 58, 50, 3...
## $ Gender       <fctr> male, male, male, female, female, female, female...
## $ SleepTrouble <fctr> Yes, Yes, Yes, Yes, No, No, No, No, No, Yes, No,...
## $ Poverty      <dbl> 1.36, 1.36, 1.36, 1.91, 5.00, 5.00, 5.00, 2.20, 5...

## Rattle: A free graphical interface for data mining with R.
## Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract

## Classes 'tbl_df', 'tbl' and 'data.frame':    7175 obs. of  4 variables:
##  $ Age         : int  34 34 34 49 45 45 45 66 58 54 ...
##  $ Gender      : Factor w/ 2 levels "female","male": 2 2 2 1 1 1 1 2 2 2 .
..
##  $ SleepTrouble: Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 1 1 1 2 ...
##  $ Poverty     : num  1.36 1.36 1.36 1.91 5 5 5 2.2 5 2.2 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:2825] 4 6 7 14 17 20 23
27 35 39 ...
##   .. ..- attr(*, "names")= chr [1:2825] "4" "6" "7" "14" ...

## n= 5022
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 5022 1301 No (0.7409399 0.2590601) *
```

```
##
## Classification tree:
## rpart(formula = SleepTrouble ~ ., data = crs$dataset[crs$train,
##     c(crs$input, crs$target)], method = "class", parms = list(split = "inf
ormation"),
##     control = rpart.control(minsplit = 5, usesurrogate = 0, maxsurrogate =
0))
##
## Variables actually used in tree construction:
## character(0)
##
## Root node error: 1301/5022 = 0.25906
##
## n= 5022
##
##   CP nsplit rel error xerror xstd
## 1  0      0         1      0    0

##
##   Conditional inference tree with 9 terminal nodes
##
## Response:  SleepTrouble
## Inputs:  Age, Gender, Poverty
## Number of observations:  5022
##
## 1) Age <= 40; criterion = 1, statistic = 51.89
##   2) Gender == {female}; criterion = 1, statistic = 37.152
##     3) Poverty <= 3.1; criterion = 0.969, statistic = 6.547
##       4) Age <= 25; criterion = 0.998, statistic = 12.102
##         5)*  weights = 296
##       4) Age > 25
##         6)*  weights = 379
##     3) Poverty > 3.1
##       7)*  weights = 389
##   2) Gender == {male}
##     8) Age <= 19; criterion = 0.988, statistic = 8.272
##       9)*  weights = 173
##     8) Age > 19
##       10)*  weights = 940
## 1) Age > 40
##   11) Gender == {female}; criterion = 0.999, statistic = 13.04
##     12) Poverty <= 3.16; criterion = 0.951, statistic = 5.747
##       13)*  weights = 664
##     12) Poverty > 3.16
##       14) Poverty <= 4.87; criterion = 0.989, statistic = 8.421
##         15)*  weights = 327
##       14) Poverty > 4.87
##         16)*  weights = 442
##   11) Gender == {male}
##     17)*  weights = 1412
```



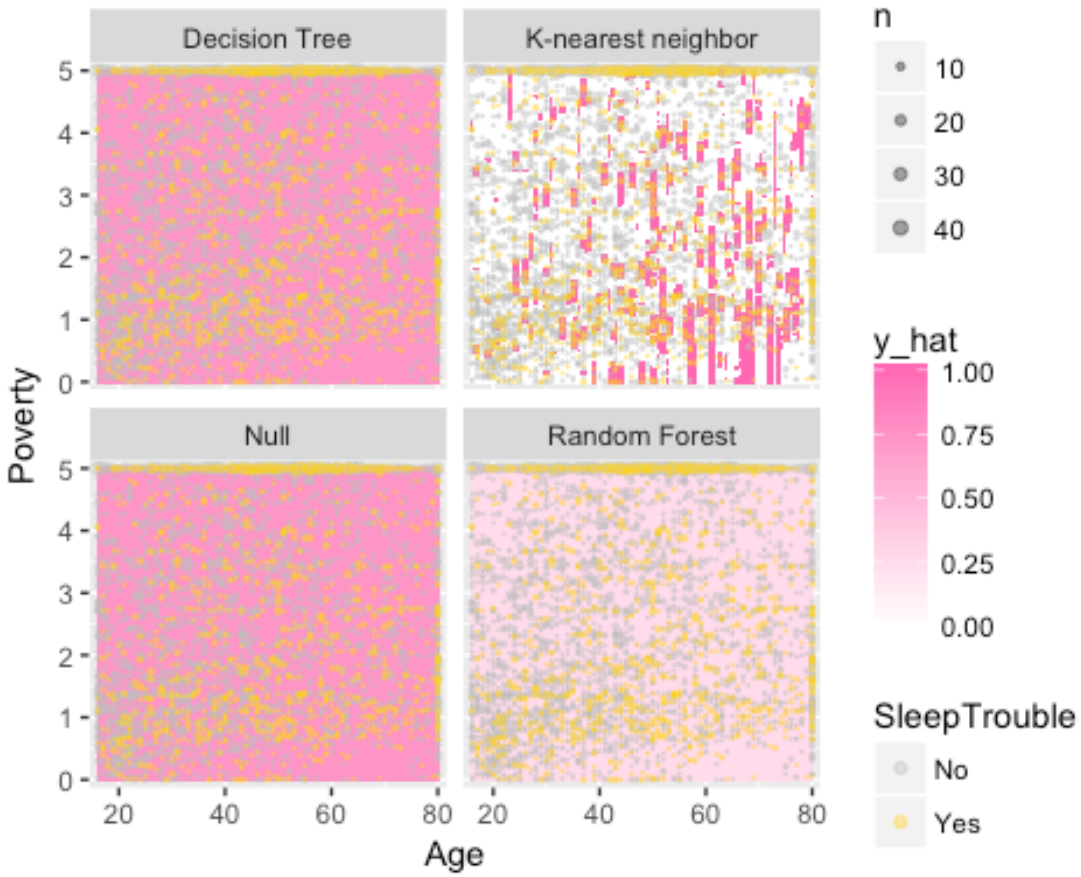
Note: SleepTroubles =2 indicates "Yes", and SleepTroubles =1 indicates "No"

Two visualizations were generated to model the Sleep Troubles variable. These two models are:

- Sleep Troubles ~ Age, Gender, Poverty
  - Used RATTLE to generate a Decision Tree for this information using the PARTY library (ctree was used instead of rpart)
- Sleep Troubles ~ Age, Poverty
  - Used Dr. Hertzberg's code as a model to generate a 2x2 visualization of the KNN, Decision Tree, Null, and Random Forest fit plots for this data.

This decision tree is the most intuitive model to grasp (at least for me). This model shows that differences in the likelihood an individual will have disturbed sleep can be separated first and foremost by age, with individuals over 40 having the highest likelihood of self-reported sleep troubles. Within this population (>40), women also are more likely to report varying degrees of sleep troubles regardless of SES, whereas male risk for sleep troubles seems to depend upon SES, with certain demographics having risk that is able to be modeled, and other groups having a range of risk. For younger individuals (<40), women

do not appear to report as troubled sleep, and for men, again, their sleep troubles seem to have altered likelihood within groups of differing SES.




```
## [1] 10000

## [1] 10000

## [1] 10000
```

Pull Second subset of variables from NHANES Dataset

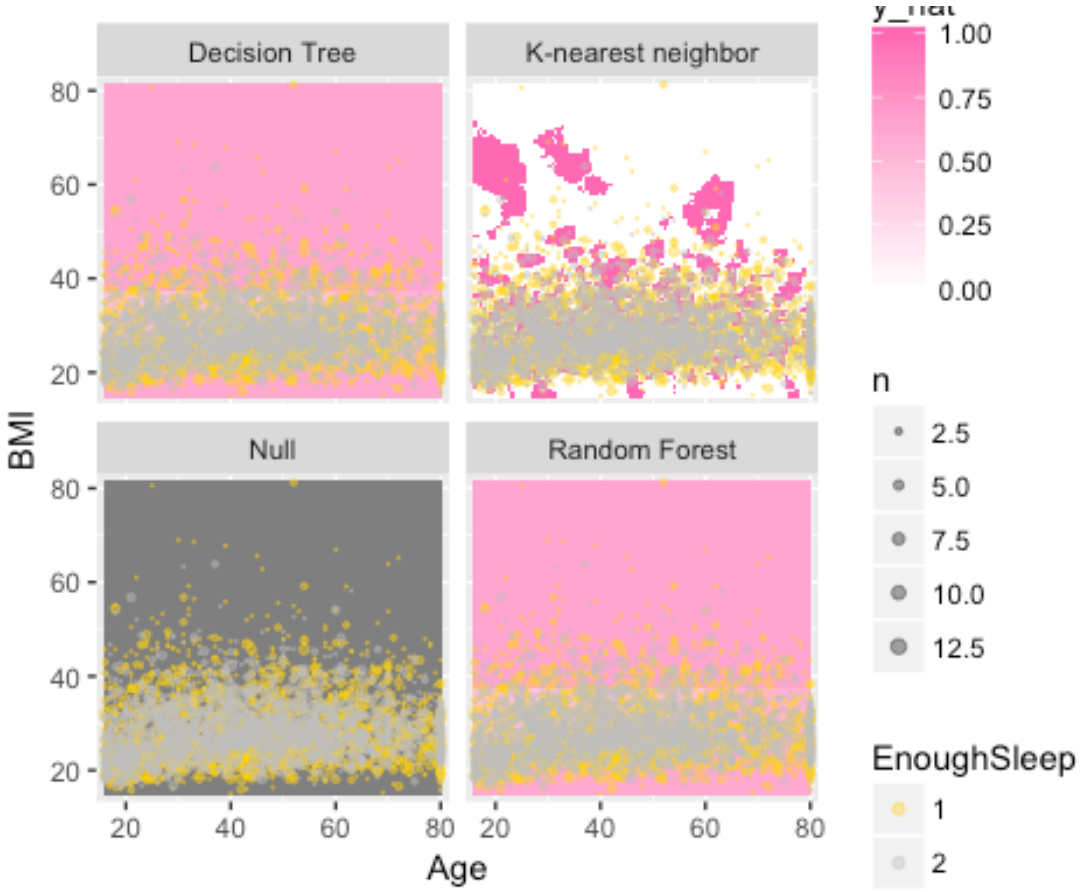
```
## Observations: 10,000
## Variables: 4
## $ Age         <int> 34, 34, 34, 4, 49, 9, 8, 45, 45, 45, 66, 58, 54, 1...
## $ Gender      <fctr> male, male, male, male, female, male, male, femal...
## $ BMI         <dbl> 32.22, 32.22, 32.22, 15.30, 30.57, 16.82, 20.64, 2...
## $ EnoughSleep <dbl> 2, 2, 2, NA, 1, NA, NA, 1, 1, 1, 1, 2, 2, NA, 2, 1...

## EnoughSleep
##     1     2  <NA>
## 48.00 29.55 22.45

## [1] "tbl_df"     "tbl"        "data.frame"
```

```
## Observations: 10,000
## Variables: 4
## $ Age         <int> 34, 34, 34, 4, 49, 9, 8, 45, 45, 45, 66, 58, 54, 1...
## $ Gender      <fctr> male, male, male, male, female, male, male, femal...
## $ BMI         <dbl> 32.22, 32.22, 32.22, 15.30, 30.57, 16.82, 20.64, 2...
## $ EnoughSleep <dbl> 2, 2, 2, NA, 1, NA, NA, 1, 1, 1, 1, 2, 2, NA, 2, 1...

## Observations: 7,683
## Variables: 4
## $ Age         <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 58, 50, 33...
## $ Gender      <dbl> 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2, 2, 2, 1, 1,...
## $ BMI         <dbl> 32.22, 32.22, 32.22, 30.57, 27.24, 27.24, 27.24, 2...
## $ EnoughSleep <dbl> 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2, 2, 1, 1,...
```

Run Models

```
## [1] 100

## [1] 97.50098

## [1] 96.48575

## [1] 89.88676

##
## knn.1    1    2
##     1 4758    0
##     2    0 2925

##
## knn.3    1    2
##     1 4689  123
##     2   69 2802

##
## knn.5    1    2
##     1 4684  196
##     2   74 2729

##
## knn.20    1    2
##      1 4618  637
##      2  140 2288

## Observations: 10,000
## Variables: 2
## $ Age <dbl> 16.00000, 16.64646, 17.29293, 17.93939, 18.58586, 19.23232...
## $ BMI <dbl> 15.02, 15.02, 15.02, 15.02, 15.02, 15.02, 15.02, 15.02, 15...

## Observations: 7,683
## Variables: 4
## $ Age         <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 58, 50, 33...
```

```
## $ Gender      <dbl> 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2, 2, 2, 1, 1,...
## $ EnoughSleep <dbl> 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2, 2, 1, 1,...
## $ BMI         <dbl> 32.22, 32.22, 32.22, 30.57, 27.24, 27.24, 27.24, 2...

## Warning: attributes are not identical across measure variables; they will
## be dropped
```

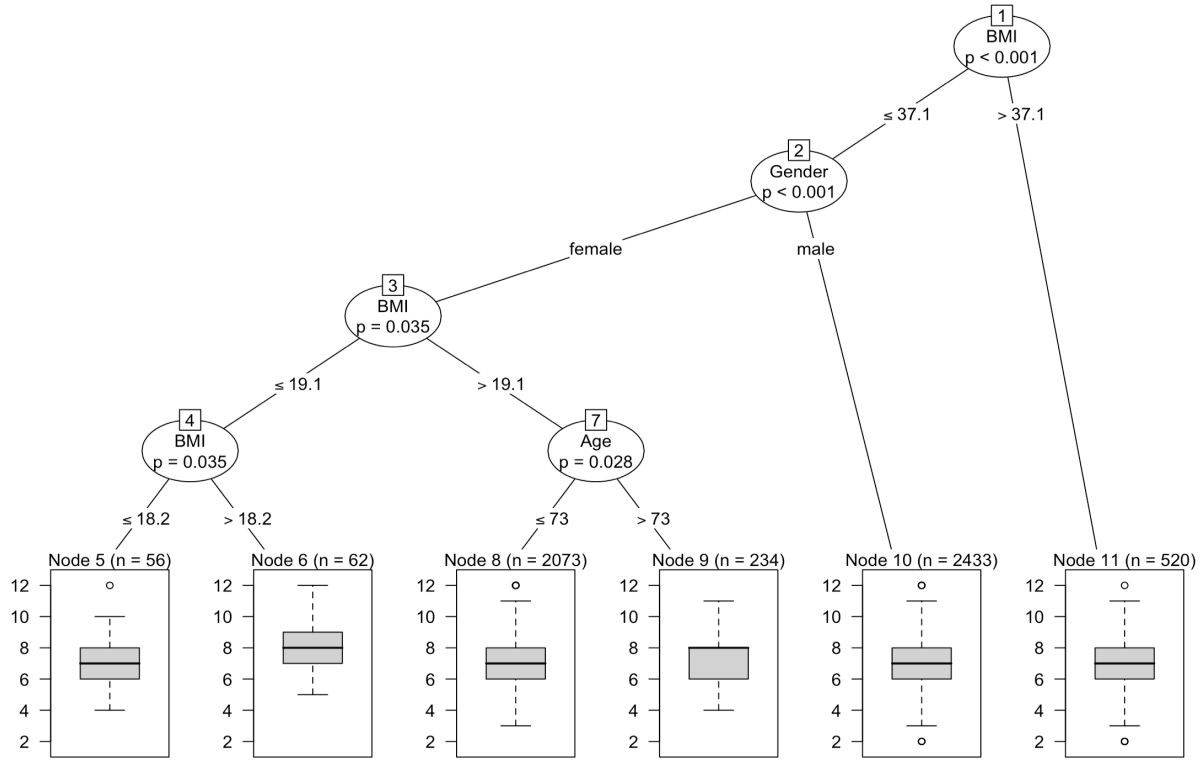



```
## [1] 10000

## [1] 10000

## [1] 10000
```

Note: EnoughSleep =1 indicates "Yes, btwn 7-9hrs per night", and EnoughSleep =2 indicates "No, less than 7 or more than 9hrs per night"

```
## 'data.frame':    7683 obs. of  4 variables:
##  $ Age          : int  34 34 34 49 45 45 45 66 58 54 ...
##  $ Gender       : Factor w/ 2 levels "female","male": 2 2 2 1 1 1 1 2 2 2
...
##  $ BMI          : num  32.2 32.2 32.2 30.6 27.2 ...
##  $ SleepHrsNight: int  4 4 4 8 8 8 8 7 5 4 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:2317] 4 6 7 14 17 27 39
```

```
40 52 53 ...
##   .. ..- attr(*, "names")= chr [1:2317] "4" "6" "7" "14" ...

##
##   Conditional inference tree with 6 terminal nodes
##
## Response:  SleepHrsNight
## Inputs:  Age, Gender, BMI
## Number of observations:  5378
##
## 1) BMI <= 37.1; criterion = 1, statistic = 25.848
##   2) Gender == {female}; criterion = 1, statistic = 24.21
##     3) BMI <= 19.1; criterion = 0.965, statistic = 6.355
##       4) BMI <= 18.2; criterion = 0.965, statistic = 6.341
##         5)*  weights = 56
##       4) BMI > 18.2
##         6)*  weights = 62
##     3) BMI > 19.1
##       7) Age <= 73; criterion = 0.972, statistic = 6.728
##         8)*  weights = 2073
##       7) Age > 73
##         9)*  weights = 234
##   2) Gender == {male}
##     10)*  weights = 2433
## 1) BMI > 37.1
##   11)*  weights = 520
```



Again, two visualizations were generated to model the variable capturing Hours of sleep per night. These two models are:

- SleepHrsNight ~ Age, BMI, Gender
  - Used RATTLE to generate a Decision Tree for this information using the PARTY library (ctree was used instead of rpart)
- SleepHrsNight ~ Age, Poverty
  - Used Dr. Hertzberg's code as a model to generate a 2x2 visualization of the KNN, Decision Tree, Null, and Random Forest fit plots for this data.
  - Turned SleepHrsNight into a categorical variable with 1 equaling "Normal Sleep Duration" (7-9 hrs) and 2 equaling "Abnormal Sleep Duration" (<7 or >9hrs/night)

This decision tree seems to be the most descriptive model. This model shows that differences in the duration of sleep can be split by BMI, with individuals with a higher BMI (37.1) reporting similar sleep patterns compared to those with a BMI below that threshold. Within the sample with a BMI<37.1, gender is the next variable upon with the sleep duration is split, followed by another split by BMI, then further grouped by BMI and Age.

Classifiers and Sleep Troubles and Duration

Neither one of these sets of classifiers seem to accurately model Sleep Trouble or Sleep Duration. For the first model (SleepTroubles), the model seems to be able to model certain demographics who report consistently untroubled sleep (young women), but does not have the ability to isolate groups that consistently report troubled sleep (as seen in the boxplots indicating that the range of responses for troubled sleep is 1-2 for many of the decision tree groups.

The Decision tree for Sleep Duration seems to be a bit better at classifying differences in sleep based on demographic measures (BMI, Age, Gender), but this is more likely due to the nature of the dependent variable (continuous instead of categorical) rather than the quality of the classifiers.