Risky Behaviors

How do risky behaviors predict student academic achievement?

by Andrea Chamorro, Mariam Khan, Janani Shankaran; Georgetown University's McCourt School of Public Policy

May 8, 2017

Introduction

For several decades, the academic performance of students has been a major concern. Many studies have discovered that academic success has been strongly linked with health-related factors. According to the Centers for Disease Control and the 2009 National Youth Risk Behavior Survey (YRBS), there is a negative association between health-risk behaviors and academic achievement among high school students. In other words, students with higher grades are less likely to engage in health-risk behaviors than students with lower grades. Similarly, students who do not engage in health-risk behaviors are more likely to receive higher grades than students who engage in health-risk behaviors. It should be noted these associations do not prove causation.

The objective of this study is to build upon the CDC research, in order to better understand how certain behaviors can impact the grades of students. These results can encourage schools to promote health and safety among students, which would help them to establish lifelong healthy behaviors.

Data

The City of Somerville's Youth Risk Behavior Survey is an annual student survey conducted at Somerville High School. Students at Somerville High School were surveyed every two years, from 2002 to 2014. The dataset includes a total of 8,003 student survey responses. The dataset can be accessed here: http://bit.ly/2nRvJYa

```r
setwd("C:/Users/anchamorro/Desktop/_MPP/Intro to Data Science/Class project")
df<- read.csv("Somerville_High_School_YRBS_Raw_Data_2002-2014.csv")

###Libraries###
library(plyr)
library(Hmisc)
library(MASS)
library(pscl)
library(randomForest)
library(VIM)
library(rpart)
library(rpart.plot)
library(ggplot2)
library(gridExtra)
library(plotROC)

meanf1 <- function(actual, predicted){
  #Mean F1 score function
  #actual = a vector of actual labels
  #predicted = predicted labels
  classes <- unique(actual)
  results <- data.frame()
  for(k in classes){
    results <- rbind(results,
                     data.frame(class.name = k,
                                weight = sum(actual == k)/length(actual),
                                precision = sum(predicted == k & actual == k)/sum(predict
ed == k),
                                recall = sum(predicted == k & actual == k)/sum(actual ==
k)))
  }
  results$score <- results$weight * 2 * (results$precision * results$recall) / (results$p
recision + results$recall)
  return(sum(results$score))
}
```

Descriptive statistics

In terms of demographic characteristics, roughly 40.1 percent of students are White, 15.0 percent are Black, 24.4 percent are Hispanic, 8.7 percent are Asian, and 11.8 percent identify as Other race. In addition, 219 observations exhibit missingness in race. Approximately 52.4 percent of the sample is female, while the average age of students is 16.25.

Among the risky behaviors, students are most likely to have engaged in sexual activity or consumed alcohol. Some of the variables exhibit considerable missingness, including variables related to hurting oneself, gang affiliation, altercations, and drug use.

Research Strategy

First, the original dataset comes with 219 variables. Based on the CDC report, we have narrowed it down to 29 variables, which focuses specifically on risky behaviors, such as gang affiliation, gun possession, and drug and alcohol use.

Second, many of the categorical variables were recoded into dummy variables. For example, the variable, chew_30, examines how many days a student has chewed tobacco in the past 30 days. Respondents had the option to choose 0 days, 1 or 2 days, 3 to 5 days, 6 to 9 days, 10 to 19 days, 20 to 29 days or All 30 days. In this case, chew_30 would be recoded as a dummy variable, in that respondents who have never smoked a cigarette, would be coded as a 0, while who respondents have smoked a cigarette at least once, would be coded as a 1.

In addition, the race variables were also recoded into dummy variables. For example, if the respondent identified as Asian, it would be coded as a 1 and likewise, if the respondent did not identify as Asian, it would be coded as 0.

Furthermore, the dependent variable, skl_gra, which provides the grades of students, is also recoded. The variable is recoded on a scale from 1-5, in which a 5 corresponds to â□□mostly Aâ□□sâ□□, while a 1 corresponds to â□□mostly Eâ□□s/Fâ□□s.â□□

```
##Grades
df$grades[df$skl_gra =="Mostly A's"] <- 5
df$grades[df$skl_gra =="Mostly B's"] <- 4
df$grades[df$skl_gra =="Mostly C's"] <- 3
df$grades[df$skl_gra =="Mostly D's"] <- 2
df$grades[df$skl_gra =="Mostly E's or F's"] <- 1


##Grades option 2
df$grades2[df$skl_gra =="Mostly A's"] <- "Mostly A's"
df$grades2[df$skl_gra =="Mostly B's"] <- "Mostly B's"
df$grades2[df$skl_gra =="Mostly C's"] <- "Mostly C's"
df$grades2[df$skl_gra =="Mostly D's"] <- "Mostly D's"
df$grades2[df$skl_gra =="Mostly E's or F's"] <- "Mostly E's or F's"


##Ingang
df$ingang[df$gang =="Yes"] <- 1
df$ingang[df$gang =="No"] <- 0


##schoolaltercation
df$schoolaltercation[df$fit_skl =="0 times"] <- 0
df$schoolaltercation[df$fit_skl =="1 time"] <- 1
df$schoolaltercation[df$fit_skl =="2 or 3 times"] <- 1
df$schoolaltercation[df$fit_skl =="4 or 5 times"] <- 1
df$schoolaltercation[df$fit_skl =="6 or 7 times"] <- 1
df$schoolaltercation[df$fit_skl =="8 or 9 times"] <- 1
df$schoolaltercation[df$fit_skl =="10 or 11 times"] <- 1
df$schoolaltercation[df$fit_skl =="12 or more times"] <- 1


##outsidealtercation
df$outsidealtercation[df$fit_out =="0 times"] <- 0
df$outsidealtercation[df$fit_out =="1 time"] <- 1
df$outsidealtercation[df$fit_out =="2 or 3 times"] <- 1
df$outsidealtercation[df$fit_out =="4 or 5 times"] <- 1
df$outsidealtercation[df$fit_out =="6 or 7 times"] <- 1
df$outsidealtercation[df$fit_out =="8 or 9 times"] <- 1
df$outsidealtercation[df$fit_out =="10 or 11 times"] <- 1
df$outsidealtercation[df$fit_out =="12 or more times"] <- 1

##schoolweapon
df$schoolweapon[df$weap_skl =="0 days"] <- 0
df$schoolweapon[df$weap_skl =="1 day"] <- 1
df$schoolweapon[df$weap_skl =="2 or 3 days"] <- 1
df$schoolweapon[df$weap_skl=="4 or 5 days"] <- 1
df$schoolweapon[df$weap_skl =="6 or more days"] <- 1
```

```
##outsideweapon
df$outsideweapon[df$weap_out =="0 days"] <- 0
df$outsideweapon[df$weap_out =="1 day"] <- 1
df$outsideweapon[df$weap_out =="2 or 3 days"] <- 1
df$outsideweapon[df$weap_out=="4 or 5 days"] <- 1
df$outsideweapon[df$weap_out =="6 or more days"] <- 1

##hurtingself
df$hurtingself[df$hurtself =="0 times"] <- 0
df$hurtingself[df$hurtself =="1 or 2 times"] <- 1
df$hurtingself[df$hurtself =="3 to 5 times"] <- 1
df$hurtingself[df$hurtself =="6 to 9 times"] <- 1
df$hurtingself[df$hurtself =="10 to 19 times"] <- 1
df$hurtingself[df$hurtself =="20 or more times"] <- 1

##CigUse
df$ciguse[df$cig_30 =="0 days"] <- 0
df$ciguse[df$cig_30 =="1 or 2 days"] <- 1
df$ciguse[df$cig_30 =="3 to 5 days"] <- 1
df$ciguse[df$cig_30 =="6 to 9 days"] <- 1
df$ciguse[df$cig_30 =="10 to 19 days"] <- 1
df$ciguse[df$cig_30 =="20 to 29 days"] <- 1
df$ciguse[df$cig_30 =="All 30 days"] <- 1


##Tobacco
df$tobacco[df$chew_30 =="0 days"] <- 0
df$tobacco[df$chew_30 =="1 or 2 days"] <- 1
df$tobacco[df$chew_30 =="3 to 5 days"] <- 1
df$tobacco[df$chew_30 =="6 to 9 days"] <- 1
df$tobacco[df$chew_30 =="10 to 19 days"] <- 1
df$tobacco[df$chew_30 =="20 to 29 days"] <- 1
df$tobacco[df$chew_30 =="All 30 days"] <- 1

##Ectasy
df$ecstasy[df$x_30 =="0 times"] <- 0
df$ecstasy[df$x_30 =="1 or 2 times"] <- 1
df$ecstasy[df$x_30 =="3 to 9 times"] <- 1
df$ecstasy[df$x_30 =="10 to 19 times"] <- 1
df$ecstasy[df$x_30 =="20 to 39 times"] <- 1
df$ecstasy[df$x_30 =="40 or more times"] <- 1

##Oxy
df$oxy[df$oxy_30 =="0 times"] <- 0
df$oxy[df$oxy_30 =="1 or 2 times"] <- 1
df$oxy[df$oxy_30 =="3 to 9 times"] <- 1
df$oxy[df$oxy_30 =="10 to 19 times"] <- 1
```

```
df$oxy[df$oxy_30 =="20 to 39 times"] <- 1
df$oxy[df$oxy_30 =="40 or more times"] <- 1

##Other
df$otherdrug[df$oth_30 =="0 times"] <- 0
df$otherdrug[df$oth_30 =="1 or 2 times"] <- 1
df$otherdrug[df$oth_30 =="3 to 9 times"] <- 1
df$otherdrug[df$oth_30 =="10 to 19 times"] <- 1
df$otherdrug[df$oth_30 =="20 to 39 times"] <- 1
df$otherdrug[df$oth_30 =="40 or more times"] <- 1

##Sexual
df$sexual[df$sex_ever =="No"] <- 0
df$sexual[df$sex_ever =="Yes"] <- 1

##Pregnancy
df$pregnancy[df$pregnant =="No"] <- 0
df$pregnancy[df$pregnant =="I have never had sexual intercourse"] <- 0
df$pregnancy[df$pregnant =="Yes"] <- 1

##Age
#Note that age variable is left and right censored
df$age2[df$age=="13 years old or younger"] <- 13
df$age2[df$age=="14 years old"] <- 14
df$age2[df$age=="15 years old"] <- 15
df$age2[df$age=="16 years old"] <- 16
df$age2[df$age=="17 years old"] <- 17
df$age2[df$age=="18 years old or older"] <- 18

##Race

#Race = White
df$white[df$race=="White"] <- 1
df$white[df$race=="American Indian or Alaska Native"] <- 0
df$white[df$race=="Asian or other Pacific Islander"] <- 0
df$white[df$race=="Black"] <- 0
df$white[df$race=="Hispanic or Latino"] <- 0
df$white[df$race=="Other"] <- 0

#Race = Black
df$black[df$race=="White"] <- 0
df$black[df$race=="American Indian or Alaska Native"] <- 0
df$black[df$race=="Asian or other Pacific Islander"] <- 0
df$black[df$race=="Black"] <- 1
df$black[df$race=="Hispanic or Latino"] <- 0
df$black[df$race=="Other"] <- 0

#Race = Asian
```

```r
df$asian[df$race=="White"] <- 0
df$asian[df$race=="American Indian or Alaska Native"] <- 0
df$asian[df$race=="Asian or other Pacific Islander"] <- 1
df$asian[df$race=="Black"] <- 0
df$asian[df$race=="Hispanic or Latino"] <- 0
df$asian[df$race=="Other"] <- 0

#Race = Hispanic
df$hispanic[df$race=="White"] <- 0
df$hispanic[df$race=="American Indian or Alaska Native"] <- 0
df$hispanic[df$race=="Asian or other Pacific Islander"] <- 0
df$hispanic[df$race=="Black"] <- 0
df$hispanic[df$race=="Hispanic or Latino"] <- 1
df$hispanic[df$race=="Other"] <- 0

#Race = Other
df$otherrace[df$race=="White"] <- 0
df$otherrace[df$race=="American Indian or Alaska Native"] <- 1
df$otherrace[df$race=="Asian or other Pacific Islander"] <- 0
df$otherrace[df$race=="Black"] <- 0
df$otherrace[df$race=="Hispanic or Latino"] <- 0
df$otherrace[df$race=="Other"] <- 1

##Gender
df$female[df$GENDER=="Male"] <- 0
df$female[df$GENDER=="Female"] <- 1

##Alcohol
df$alcohol[df$alc_30 =="0 days"] <- 0
df$alcohol[df$alc_30 =="1 or 2 days"] <- 1
df$alcohol[df$alc_30 =="3 to 5 days"] <- 1
df$alcohol[df$alc_30 =="6 to 9 days"] <- 1
df$alcohol[df$alc_30 =="10 to 19 days"] <- 1
df$alcohol[df$alc_30 =="20 to 29 days"] <- 1
df$alcohol[df$alc_30 =="All 30 days"] <- 1

##Marijuana
df$marijuana[df$pot_30 =="0 times"] <- 0
df$marijuana[df$pot_30 =="1 or 2 times"] <- 1
df$marijuana[df$pot_30 =="3 to 9 times"] <- 1
df$marijuana[df$pot_30 =="10 to 19 times"] <- 1
df$marijuana[df$pot_30 =="20 to 39 times"] <- 1
df$marijuana[df$pot_30 =="40 or more times"] <- 1

##Heroin
df$heroin[df$her_30 =="0 times"] <- 0
df$heroin[df$her_30 =="1 or 2 times"] <- 1
df$heroin[df$her_30 =="3 to 9 times"] <- 1
```

```
df$heroin[df$her_30 =="10 to 19 times"] <- 1
df$heroin[df$her_30 =="20 to 39 times"] <- 1
df$heroin[df$her_30 =="40 or more times"] <- 1


##Meth
describe(df$meth_30)
```

```
## df$meth_30
##        n  missing distinct
##     8003        0        7
##
## (1513, 0.189), 0 times (6438, 0.804), 1 or 2 times (28, 0.003), 10 to 19
## times (5, 0.001), 20 to 39 times (7, 0.001), 3 to 9 times (10, 0.001), 40
## or more times (2, 0.000)
```

```
df$meth[df$meth_30 =="0 times"] <- 0
df$meth[df$meth_30 =="1 or 2 times"] <- 1
df$meth[df$meth_30 =="3 to 9 times"] <- 1
df$meth[df$meth_30 =="10 to 19 times"] <- 1
df$meth[df$meth_30 =="20 to 39 times"] <- 1
df$meth[df$meth_30 =="40 or more times"] <- 1
```

Third, the dataset is divided into a 70-15-15 partition.

```
###New data frame
df2 <- df[c(1:3, 12, 194:219)]

#Remove observations where grades2=NA
df2 <- subset(df2, !is.na(grades2))

#Summary statistics
summary(df2)
```

```
##   survey          year            id                      skl_gra   
##      :1315   Min.   :2002   Min.   :   2.0   Mostly B's       :2958  
##  SH04:1293   1st Qu.:2004   1st Qu.: 330.0   Mostly C's       :2035  
##  SH06: 935   Median :2008   Median : 666.5   Mostly A's       :1444  
##  SH08:1007   Mean   :2007   Mean   :1187.3   Mostly D's       : 639  
##  SH10: 917   3rd Qu.:2010   3rd Qu.:1306.8   Mostly E's or F's: 178  
##  SH12: 876   Max.   :2014   Max.   :9999.0                    :   0  
##  SH14: 911                  NA's   :1328     (Other)          :   0  
##      grades        grades2             ingang       schoolaltercation
##  Min.   :1.000   Length:7254        Min.   :0.0000   Min.   :0.0000   
##  1st Qu.:3.000   Class :character   1st Qu.:0.0000   1st Qu.:0.0000   
##  Median :4.000   Mode  :character   Median :0.0000   Median :0.0000   
##  Mean   :3.669                      Mean   :0.0405   Mean   :0.1054   
##  3rd Qu.:4.000                      3rd Qu.:0.0000   3rd Qu.:0.0000   
##  Max.   :5.000                      Max.   :1.0000   Max.   :1.0000   
##                                     NA's   :1497     NA's   :1344     
##  outsidealtercation  schoolweapon    outsideweapon     hurtingself   
##  Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000  
##  Median :0.0000     Median :0.0000   Median :0.0000   Median :0.000  
##  Mean   :0.1931     Mean   :0.0465   Mean   :0.1051   Mean   :0.133  
##  3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.000  
##  Max.   :1.0000     Max.   :1.0000   Max.   :1.0000   Max.   :1.000  
##  NA's   :1351       NA's   :1342     NA's   :1347     NA's   :3571   
##      ciguse          tobacco           ecstasy            oxy        
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000  
##  Mean   :0.1412   Mean   :0.02159   Mean   :0.0227   Mean   :0.0158  
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  
##  NA's   :80       NA's   :120       NA's   :1355     NA's   :1372    
##    otherdrug         sexual         pregnancy           age2      
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :13.00  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:15.00  
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :16.00  
##  Mean   :0.0212   Mean   :0.4686   Mean   :0.0501   Mean   :16.26  
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:17.00  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :18.00  
##  NA's   :1366     NA's   :271      NA's   :364      NA's   :22     
##      white            black            asian            hispanic     
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000  
##  Mean   :0.4131   Mean   :0.1477   Mean   :0.08884   Mean   :0.2354  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000  
##  NA's   :174      NA's   :174      NA's   :174       NA's   :174     
##    otherrace          female          alcohol         marijuana
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1148   Mean   :0.5251   Mean   :0.3577   Mean   :0.2091
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :174      NA's   :53       NA's   :77       NA's   :101
##      heroin            meth
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.0049   Mean   :0.0085
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
##  NA's   :1347     NA's   :1350
```

```
###Partition###
library(dplyr)

#Option 1
dftrain <- df[sample(nrow(df),
                     size = round(0.7*nrow(df)),
                     replace = F),]
dftest <- anti_join(df, dftrain, by = "id")
dfval <- dftest[sample(nrow(dftest),
                       size = round(0.5*nrow(dftest)),
                       replace = F),]
dftest <- anti_join(dftest, dfval, by = "id")

#Option 2
set.seed(100)
rand <- runif(nrow(df2))
train <- df2[rand > 0.3,]
validate <- df2[rand > 0.15 & rand <= 0.3,]
test <- df2[rand <= 0.15,]
```

Finally, Decision Trees, Random Forest, and Ordered Logistic Regression will be used in predicting the actual grades of students. The Mean-F1 and the AUC value will be taken into consideration, when determining the optimal technique.

Methodology

Decision tree

For our decision tree analysis, we tested attribute values for each input feature using the information gain entropy measure. We were able to calculate results for the default, zero, and the optimal CP-values. Also, we conducted a variable of importance test on all of our variables of interest. We found that sex, alcohol, marijuana, cigarette, pregnancy, and chewing tobacco were some of the variables that were most important. Unfortunately, the decision tree results yielded a Mean-F1 score of 1 for all measures in our sample. We made sure to remove any variables that would be result in multicollinearity, but the Mean-F1 score was still 1. Just looking at the predicted values shows this is clearly not accurate. Therefore, we could not determine which measure produced the most accurate results. In general, decision trees tend to overfit predictive models.

```
#Train
fittingall <- rpart(grades2 ~ ciguse + tobacco + ingang + hurtingself
+ schoolaltercation + schoolweapon + outsidealtercation + outsideweapon + schoolweapon
+ outsideweapon + ecstasy + oxy + otherdrug + sexual + pregnancy + age2
+ white + asian + hispanic + otherrace + female + alcohol + marijuana
+ heroin + meth, method = "class", data = dftrain)
fittingall$variable.importance
```

```
##    sexual   alcohol    female marijuana      age2    ciguse pregnancy
## 59.080049 15.523193 15.400005 13.199821 11.821116  9.242428  5.923324
```

```
#Predict values for train
predict.opt.train <- predict(fit.opt, dftrain, type='class')
predict.0.train <- predict(fit.0, dftrain, type='class')
predict.train <- predict(fit, dftrain, type='class')

input.train <- rbind(data.frame(model = "optimal", d = dftrain$grades2,  m = predict.opt.
train),
                     data.frame(model = "CP = 0", d = dftrain$grades2,  m = predict.0.tra
in),
                     data.frame(model = "default", d =  dftrain$grades2,  m = predict.tra
in))

input.trainopt <- rbind(data.frame(model = "optimal", d = dftrain$grades2, m = predict.op
t.train))

input.train0 <-rbind( data.frame(model = "CP = 0", d = dftrain$grades2,  m = predict.0.tr
ain))

input.traindef <-rbind( data.frame(model = "default", d = dftrain$grades2,  m = predict.t
rain))

#Predict values for test
predict.opt.test <- predict(fit.opt, dftest, type='class')
predict.0.test <- predict(fit.0, dftest, type='class')
predict.test <- predict(fit, dftest, type='class')

input.test <- rbind(data.frame(model = "optimal", d = dftest$grades2, m = predict.opt.tes
t),
                    data.frame(model = "CP = 0", d = dftest$grades2,  m =
predict.0.test),
                    data.frame(model = "default", d = dftest$grades2,  m = predict.test))

input.testopt <- rbind(data.frame(model = "optimal", d = dftest$grades2, m = predict.opt.
test))

input.test0 <-rbind(data.frame(model = "CP = 0", d = dftest$grades2,  m =
predict.0.test))

input.testdef <-rbind(data.frame(model = "default", d = dftest$grades2,  m =
predict.test))


#Predict values for val
predict.opt.val <- predict(fit.opt, dfval, type='class')
predict.0.val <- predict(fit.0, dfval, type='class')
predict.val <- predict(fit, dfval, type='class')



input.val <- rbind(data.frame(model = "optimal", d = dfval$grades2, m = predict.opt.val),
                   data.frame(model = "CP = 0", d = dfval$grades2,  m = predict.0.val),
                   data.frame(model = "default", d = dfval$grades2,  m = predict.val))

input.valopt <- rbind(data.frame(model = "optimal", d = dfval$grades2, m = predict.opt.va
l))

input.val0 <-rbind(data.frame(model = "CP = 0", d = dfval$grades2,  m = predict.0.val))

input.valdef <-rbind(data.frame(model = "default", d = dfval$grades2,  m = predict.val))

#meanf1

#FYI meanf1 is w/o  NaNs, but all are wrongly giving 1

meanf1(is.nan(input.val$d), is.nan(input.val$m))
```

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

```
meanf1(is.nan(input.test$d), is.nan(input.test$m))
```

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

```
meanf1(is.nan(input.train$d), is.nan(input.train$m))
```

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

```
meanf1(is.nan(input.traindef$d), is.nan(input.traindef$m))
```

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

```
meanf1(is.nan(input.valdef$d), is.nan(input.valdef$m))
```

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

```
meanf1(is.nan(input.testdef$d), is.nan(input.testdef$m))
```

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

Random Forest

Two random forest models were analyzed in this study. One of the model utilized complete observations and the other model imputed missing values using KNN through the VIM library. When using only complete observations, the data dropped to approximately 3,000 observations. This yielded a relatively high OOB error of 56.25 percent. In contrast, when the model imputed missing values, there were roughly 7,000 observations. It should be noted that the missing values of the dependent variable or the demographic factors were not imputed. Similarly, the OOB error was still high, at 56.7 percent. While both models had low overall predictability power, it was discovered that age has the greatest importance in both models. Furthermore, gender and marijuana were relatively important in both models.

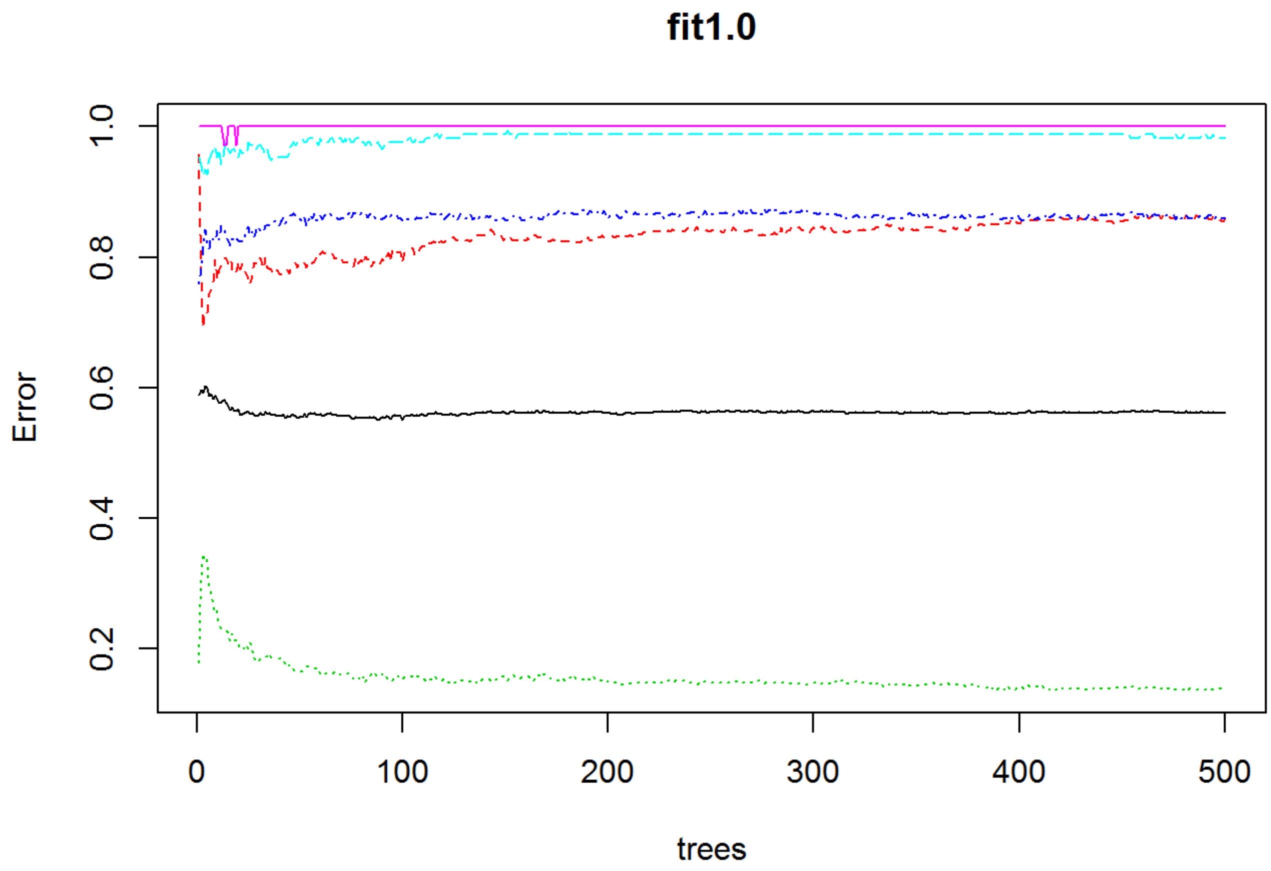
```
#Create new dataframe with recoded variables and dependent variable
df2 <- df[c(1:3, 12, 194:219)]

#First iteration (RF1): include only observations with complete data
df2 <- df2[complete.cases(df2),]

#RF1: 70-15-15 partition
set.seed(100)
rand <- runif(nrow(df2))
train <- df2[rand > 0.3,]
validate <- df2[rand > 0.15 & rand <= 0.3,]
test <- df2[rand <= 0.15,]

#RF1: Include all variables
train$grades2 <- factor(train$grades2)
fit1.0 <- randomForest(grades2 ~ ingang + schoolaltercation + outsidealtercation
                       + schoolweapon + outsideweapon + hurtingself + ciguse + tobacco
                       + ecstasy + oxy + otherdrug + sexual + pregnancy + age2 + white
                       + black + asian + hispanic + otherrace + female + alcohol + mariju
ana
                       + heroin + meth, data = train)

#RF1: Diagnostics
fit1.0
```

```
##
## Call:
##  randomForest(formula = grades2 ~ ingang + schoolaltercation +      outsidealtercation
+ schoolweapon + outsideweapon + hurtingself +      ciguse + tobacco + ecstasy + oxy + ot
herdrug + sexual + pregnancy +      age2 + white + black + asian + hispanic + otherrace +
female +      alcohol + marijuana + heroin + meth, data = train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 56.25%
## Confusion matrix:
##                   Mostly A's Mostly B's Mostly C's Mostly D's
## Mostly A's                74        437         17          0
## Mostly B's                68        819         64          1
## Mostly C's                24        426         73          4
## Mostly D's                 1        134         34          3
## Mostly E's or F's          0         30          5          0
##                   Mostly E's or F's class.error
## Mostly A's                        0   0.8598485
## Mostly B's                        1   0.1406086
## Mostly C's                        0   0.8614801
## Mostly D's                        0   0.9825581
## Mostly E's or F's                 0   1.0000000
```

```
print(importance(fit1.0, type = 2))
```

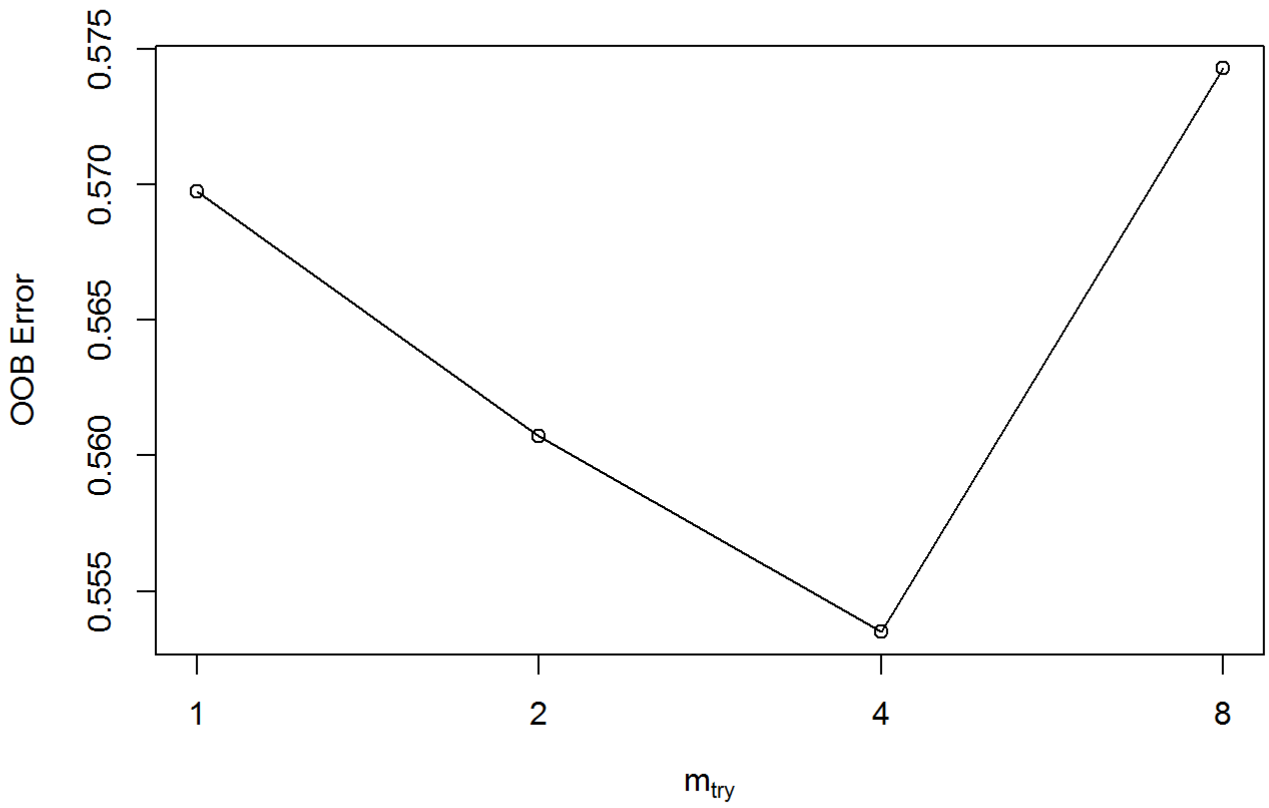
```
##                     MeanDecreaseGini
## ingang                    8.0899669
## schoolaltercation        15.8328246
## outsidealtercation       19.7355463
## schoolweapon              8.5137819
## outsideweapon            16.4697612
## hurtingself              17.2952644
## ciguse                   17.2435198
## tobacco                   6.3945486
## ecstasy                   5.4286638
## oxy                       3.6088664
## otherdrug                 7.3631570
## sexual                   21.6739620
## pregnancy                11.6166314
## age2                     50.8022042
## white                    17.4942090
## black                    12.3162063
## asian                    17.2533251
## hispanic                 15.2282547
## otherrace                12.5266322
## female                   22.9466788
## alcohol                  20.2750109
## marijuana                22.8066280
## heroin                    0.5572619
## meth                      1.7857303
```

```
plot(fit1.0)
```



```
#RF1: OOB error = 56.25%, tune model
fittune1.0 <- tuneRF(train[,-(1:6)], train$grades2, ntreeTry = 500, mtryStart = 1, stepFa
ctor = 2,
                     improve = 0.001, trace = TRUE, plot = TRUE)
```

```
## mtry = 1  OOB error = 56.98%
## Searching left ...
## Searching right ...
## mtry = 2     OOB error = 56.07%
## 0.01584786 0.001
## mtry = 4     OOB error = 55.35%
## 0.01288245 0.001
## mtry = 8     OOB error = 57.43%
## -0.03752039 0.001
```



```
fittune1.0
```

```
##       mtry  OOBError
## 1.OOB    1 0.5697517
## 2.OOB    2 0.5607223
## 4.OOB    4 0.5534989
## 8.OOB    8 0.5742664
```

```
#RF1: Four variables per split minimizes OOB error; tuning model
fit1.1 <- randomForest(grades2 ~ ingang + schoolaltercation + outsidealtercation
                       + schoolweapon + outsideweapon + hurtingself + ciguse + tobacco
                       + ecstasy + oxy + otherdrug + sexual + pregnancy + age2 + white
                       + black + asian + hispanic + otherrace + female + alcohol + mariju
ana
                       + heroin + meth, data = train, mtry=4)
fit1.1
```

```
##
## Call:
##  randomForest(formula = grades2 ~ ingang + schoolaltercation +      outsidealtercation
+ schoolweapon + outsideweapon + hurtingself +      ciguse + tobacco + ecstasy + oxy + ot
herdrug + sexual + pregnancy +      age2 + white + black + asian + hispanic + otherrace +
female +      alcohol + marijuana + heroin + meth, data = train, mtry = 4)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 55.62%
## Confusion matrix:
##                    Mostly A's Mostly B's Mostly C's Mostly D's
## Mostly A's                 89        420         19          0
## Mostly B's                 69        818         65          1
## Mostly C's                 21        430         73          3
## Mostly D's                  1        134         34          3
## Mostly E's or F's           0         31          4          0
##                    Mostly E's or F's class.error
## Mostly A's                         0   0.8314394
## Mostly B's                         0   0.1416579
## Mostly C's                         0   0.8614801
## Mostly D's                         0   0.9825581
## Mostly E's or F's                  0   1.0000000
```

```
#RF1: Unfortunately, the OOB error is still fairly high, but we will test the model anywa
y
pred.rf.train <- predict(fit1.1, train, type='prob')
pred.rf.test <- predict(fit1.1, test, type='prob')
input.rf <- rbind(data.frame(model = "train", d = train$grades2, m = pred.rf.train),
                  data.frame(model = "test", d = test$grades2, m = pred.rf.test))

#RF1: Plot ROC for grade = Mostly A's; resulting plot switches axis of Mostly A's and not
 A's; test AUC = 0.6766
a <- input.rf
a$d <- as.factor(a$d)
revalue(a$d, c("Mostly B's" = "Not A's")) -> a$d
revalue(a$d, c("Mostly C's" = "Not A's")) -> a$d
revalue(a$d, c("Mostly D's" = "Not A's")) -> a$d
revalue(a$d, c("Mostly E's or F's" = "Not A's")) -> a$d
roc.rf <- ggplot(a, aes(d = d, model = model, m = m.Mostly.A.s, colour = model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf)
```

```
##   PANEL group       AUC
## 1     1     1 0.2112472
## 2     1     2 0.3234421
```

```
#RF1: Plot ROC for grade = Mostly B's; resulting plot switches axis of Mostly B's and not B's; test AUC = 0.301
b <- input.rf
b$d <- as.factor(b$d)
revalue(b$d, c("Mostly A's" = "Not B's")) -> b$d
revalue(b$d, c("Mostly C's" = "Not B's")) -> b$d
revalue(b$d, c("Mostly D's" = "Not B's")) -> b$d
revalue(b$d, c("Mostly E's or F's" = "Not B's")) -> b$d
roc.rf2 <- ggplot(b, aes(d = d, model = model, m = m.Mostly.B.s, colour = model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf2)
```

```
##   PANEL group       AUC
## 1     1     1 0.3009447
## 2     1     2 0.4517517
```

```
#RF1: Plot ROC for grade = Mostly C's; resulting plot switches axis of Mostly C's and not C's; test AUC = 0.236
c <- input.rf
c$d <- as.factor(c$d)
revalue(c$d, c("Mostly A's" = "Not C's")) -> c$d
revalue(c$d, c("Mostly B's" = "Not C's")) -> c$d
revalue(c$d, c("Mostly D's" = "Not C's")) -> c$d
revalue(c$d, c("Mostly E's or F's" = "Not C's")) -> c$d
roc.rf3 <- ggplot(c, aes(d = d, model = model, m = m.Mostly.C.s, colour = model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf3)
```

```
##   PANEL group      AUC
## 1     1     1 0.235664
## 2     1     2 0.406936
```

```
#RF1: Plot ROC for grade = Mostly D's; resulting plot switches axis of Mostly D's and not
 D's; test AUC = 0.188
d <- input.rf
d$d <- as.factor(d$d)
revalue(d$d, c("Mostly A's" = "Not D's")) -> d$d
revalue(d$d, c("Mostly B's" = "Not D's")) -> d$d
revalue(d$d, c("Mostly C's" = "Not D's")) -> d$d
revalue(d$d, c("Mostly E's or F's" = "Not D's")) -> d$d
roc.rf4 <- ggplot(d, aes(d = d, model = model, m = m.Mostly.D.s, colour = model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf4)
```

```
##   PANEL group       AUC
## 1     1     1 0.1875847
## 2     1     2 0.2495748
```

```
#RF1: Plot ROC for grade = Mostly E's or F's; resulting plot switches axis; test AUC = 0.
142
e <- input.rf
e$d <- as.factor(e$d)
revalue(e$d, c("Mostly A's" = "Not E's")) -> e$d
revalue(e$d, c("Mostly B's" = "Not E's")) -> e$d
revalue(e$d, c("Mostly C's" = "Not E's")) -> e$d
revalue(e$d, c("Mostly D's" = "Not E's")) -> e$d
roc.rf5 <- ggplot(e, aes(d = d, model = model, m = m.Mostly.E.s.or.F.s, colour = model))
+
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf5)
```

```
##   PANEL group       AUC
## 1     1     1 0.1423657
## 2     1     2 0.2904328
```

```
#RF1: Predict activity for validate sample; only 43.5% were correctly classified using RF
1
validate$gradepred <- predict(fit1.1, validate, type='class')
validate$correct[validate$grades2 == validate$gradepred] <- 1
validate$correct[validate$grades2 != validate$gradepred] <- 0
mean(validate$correct)
```

```
## [1] 0.4352442
```

```
#RF1: Variable importance; age has the most importance, followed by marijuana use, gende
r, sexual activity and alcohol use
fit1.1$importance
```

```
##                     MeanDecreaseGini
## ingang                      8.169895
## schoolaltercation          16.017567
## outsidealtercation         19.747189
## schoolweapon                8.482969
## outsideweapon              16.629662
## hurtingself                16.826139
## ciguse                     17.581044
## tobacco                     6.541737
## ecstasy                     5.052518
## oxy                         3.555228
## otherdrug                   7.653669
## sexual                     21.507228
## pregnancy                  11.928904
## age2                       49.969373
## white                      17.040204
## black                      11.713651
## asian                      17.080971
## hispanic                   14.786111
## otherrace                  12.591051
## female                     22.675758
## alcohol                    20.400265
## marijuana                  22.882414
## heroin                      0.640794
## meth                        1.642547
```

```
#RF1: Using only complete observations, RF provides low predictability power, possibly be
cause sample is too small

#Second iteration (RF2): impute missing data on independent variables
df3 <- df[c(1:3, 12, 194:219)]

#RF2: Include only observations without missing values for grade, race, gender and age
df3 <- df3[!is.na(df3[,6]),]
df3 <- df3[!is.na(df3[,20]),]
df3 <- df3[!is.na(df3[,21]),]
df3 <- df3[!is.na(df3[,26]),]

#RF2: View summary of NA values
summary(df3)
```

```
##   survey          year            id                     skl_gra
##      :1285   Min.   :2002   Min.   :   2   Mostly B's       :2857
##  SH04:1269   1st Qu.:2004   1st Qu.: 329   Mostly C's       :1979
##  SH06: 907   Median :2008   Median : 670   Mostly A's       :1395
##  SH08: 965   Mean   :2007   Mean   :1194   Mostly D's       : 617
##  SH10: 884   3rd Qu.:2010   3rd Qu.:1330   Mostly E's or F's: 170
##  SH12: 840   Max.   :2014   Max.   :9999                    :   0
##  SH14: 868                  NA's   :1298   (Other)          :   0
##      grades        grades2             ingang       schoolaltercation
##  Min.   :1.000   Length:7018        Min.   :0.0000   Min.   :0.0000
##  1st Qu.:3.000   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :4.000   Mode  :character   Median :0.0000   Median :0.0000
##  Mean   :3.668                      Mean   :0.0406   Mean   :0.1051
##  3rd Qu.:4.000                      3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :5.000                      Max.   :1.0000   Max.   :1.0000
##                                     NA's   :1458     NA's   :1309
##  outsidealtercation  schoolweapon    outsideweapon     hurtingself
##  Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000     Median :0.0000   Median :0.0000   Median :0.000
##  Mean   :0.1928     Mean   :0.0463   Mean   :0.1045   Mean   :0.131
##  3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.000
##  Max.   :1.0000     Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##  NA's   :1317       NA's   :1310     NA's   :1316     NA's   :3487
##      ciguse          tobacco           ecstasy            oxy
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000
##  Mean   :0.1409   Mean   :0.02144   Mean   :0.0221   Mean   :0.0155
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##  NA's   :78       NA's   :114       NA's   :1324     NA's   :1340
##    otherdrug          sexual         pregnancy          age2
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :13.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:15.00
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :16.00
##  Mean   :0.0213   Mean   :0.4698   Mean   :0.0508   Mean   :16.26
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:17.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :18.00
##  NA's   :1335     NA's   :258      NA's   :347
##      white            black            asian            hispanic
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.4139   Mean   :0.1463   Mean   :0.08906   Mean   :0.2355
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##
```

```
##    otherrace          female          alcohol         marijuana     
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000  
##  Mean   :0.1151   Mean   :0.5222   Mean   :0.3585   Mean   :0.2088  
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##                                    NA's   :75       NA's   :98      
##      heroin            meth        
##  Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000  
##  Mean   :0.0049   Mean   :0.0084  
##  3rd Qu.:0.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.0000  
##  NA's   :1315     NA's   :1319    
```

```
#RF2: Remove additional columns, impute values; some warnings appear (NAs introduced by c
oercion)
df4 <- df3[-c(1:5)]
#It should be noted that this following code may take 5 to 10 minutes
df5 <- kNN(df4, variable = c(2:14, 22:25), k=5)
summary(df5)
```

```
##    grades2              ingang       schoolaltercation outsidealtercation
##  Length:7018        Min.   :0.0000   Min.   :0.0000    Min.   :0.0000
##  Class :character   1st Qu.:0.0000   1st Qu.:0.0000    1st Qu.:0.0000
##  Mode  :character   Median :0.0000   Median :0.0000    Median :0.0000
##                     Mean   :0.1512   Mean   :0.2254    Mean   :0.2915
##                     3rd Qu.:0.0000   3rd Qu.:0.0000    3rd Qu.:1.0000
##                     Max.   :1.0000   Max.   :1.0000    Max.   :1.0000
##   schoolweapon    outsideweapon     hurtingself        ciguse
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.000   Median :0.0000
##  Mean   :0.1603   Mean   :0.2082   Mean   :0.443   Mean   :0.1425
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##     tobacco           ecstasy            oxy            otherdrug
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.02223   Mean   :0.1254   Mean   :0.07125   Mean   :0.1146
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##      sexual         pregnancy           age2           white
##  Min.   :0.0000   Min.   :0.0000   Min.   :13.00   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:15.00   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :16.00   Median :0.0000
##  Mean   :0.4887   Mean   :0.0721   Mean   :16.26   Mean   :0.4139
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:17.00   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :18.00   Max.   :1.0000
##      black            asian             hispanic        otherrace
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000
##  Mean   :0.1463   Mean   :0.08906   Mean   :0.2355   Mean   :0.1151
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##      female          alcohol         marijuana          heroin
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   :0.5222   Mean   :0.3641   Mean   :0.2135   Mean   :0.06996
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##       meth           2_imp           3_imp           4_imp
##  Min.   :0.00000   Mode :logical   Mode :logical   Mode :logical
##  1st Qu.:0.00000   FALSE:5560      FALSE:5709      FALSE:5701
##  Median :0.00000   TRUE :1458      TRUE :1309      TRUE :1317
##  Mean   :0.08222   NA's :0         NA's :0         NA's :0
##  3rd Qu.:0.00000
##  Max.   :1.00000
##    5_imp           6_imp           7_imp           8_imp
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:5708      FALSE:5702      FALSE:3531      FALSE:6940
##  TRUE :1310      TRUE :1316      TRUE :3487      TRUE :78
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##    9_imp           10_imp          11_imp          12_imp
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:6904      FALSE:5694      FALSE:5678      FALSE:5683
##  TRUE :114       TRUE :1324      TRUE :1340      TRUE :1335
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##    13_imp          14_imp          22_imp          23_imp
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:6760      FALSE:6671      FALSE:6943      FALSE:6920
##  TRUE :258       TRUE :347       TRUE :75        TRUE :98
##  NA's :0         NA's :0         NA's :0         NA's :0
##
##
##    24_imp          25_imp
##  Mode :logical   Mode :logical
##  FALSE:5703      FALSE:5699
##  TRUE :1315      TRUE :1319
##  NA's :0         NA's :0
##
##
```

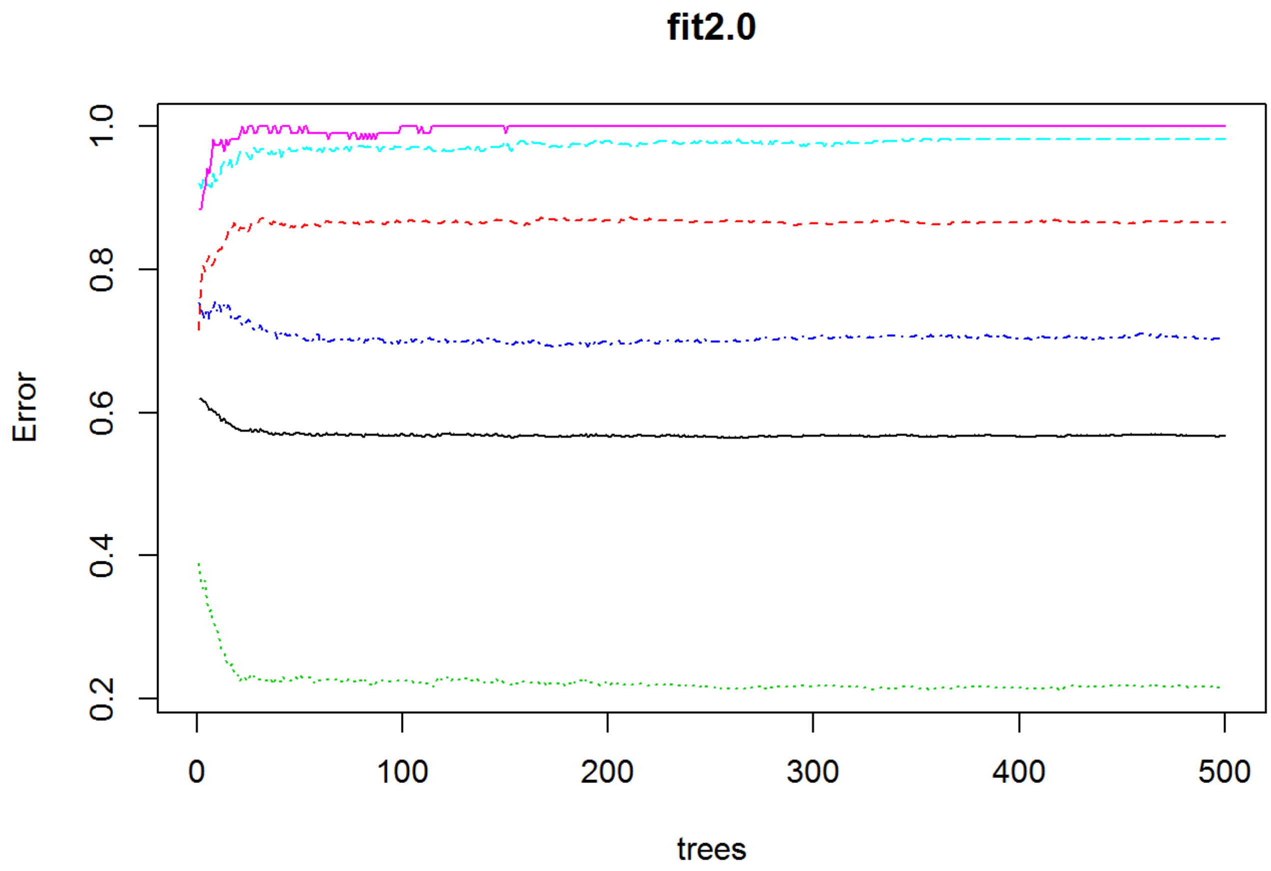
```
#RF2: Create new data frame of only variables
df6 <- df5[c(1:25)]

#RF2: 70-15-15 partition
set.seed(100)
rand <- runif(nrow(df6))
train2 <- df6[rand > 0.3,]
validate2 <- df6[rand > 0.15 & rand <= 0.3,]
test2 <- df6[rand <= 0.15,]

#RF2: Include all variables
train2$grades2 <- factor(train2$grades2)
fit2.0 <- randomForest(grades2 ~ ingang + schoolaltercation + outsidealtercation
                       + schoolweapon + outsideweapon + hurtingself + ciguse + tobacco
                       + ecstasy + oxy + otherdrug + sexual + pregnancy + age2 + white
                       + black + asian + hispanic + otherrace + female + alcohol + mariju
ana
                       + heroin + meth, data = train2)

#RF2: Diagnostics
fit2.0
```

```
##
## Call:
##  randomForest(formula = grades2 ~ ingang + schoolaltercation +      outsidealtercation
+ schoolweapon + outsideweapon + hurtingself +      ciguse + tobacco + ecstasy + oxy + ot
herdrug + sexual + pregnancy +      age2 + white + black + asian + hispanic + otherrace +
female +      alcohol + marijuana + heroin + meth, data = train2)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 56.7%
## Confusion matrix:
##                   Mostly A's Mostly B's Mostly C's Mostly D's
## Mostly A's               129        736         94          6
## Mostly B's                88       1580        337          8
## Mostly C's                25        909        407         22
## Mostly D's                 6        245        178          8
## Mostly E's or F's          1         56         52          6
##                   Mostly E's or F's class.error
## Mostly A's                        0   0.8663212
## Mostly B's                        5   0.2170466
## Mostly C's                        4   0.7022677
## Mostly D's                        3   0.9818182
## Mostly E's or F's                 0   1.0000000
```

```
print(importance(fit2.0, type = 2))
```

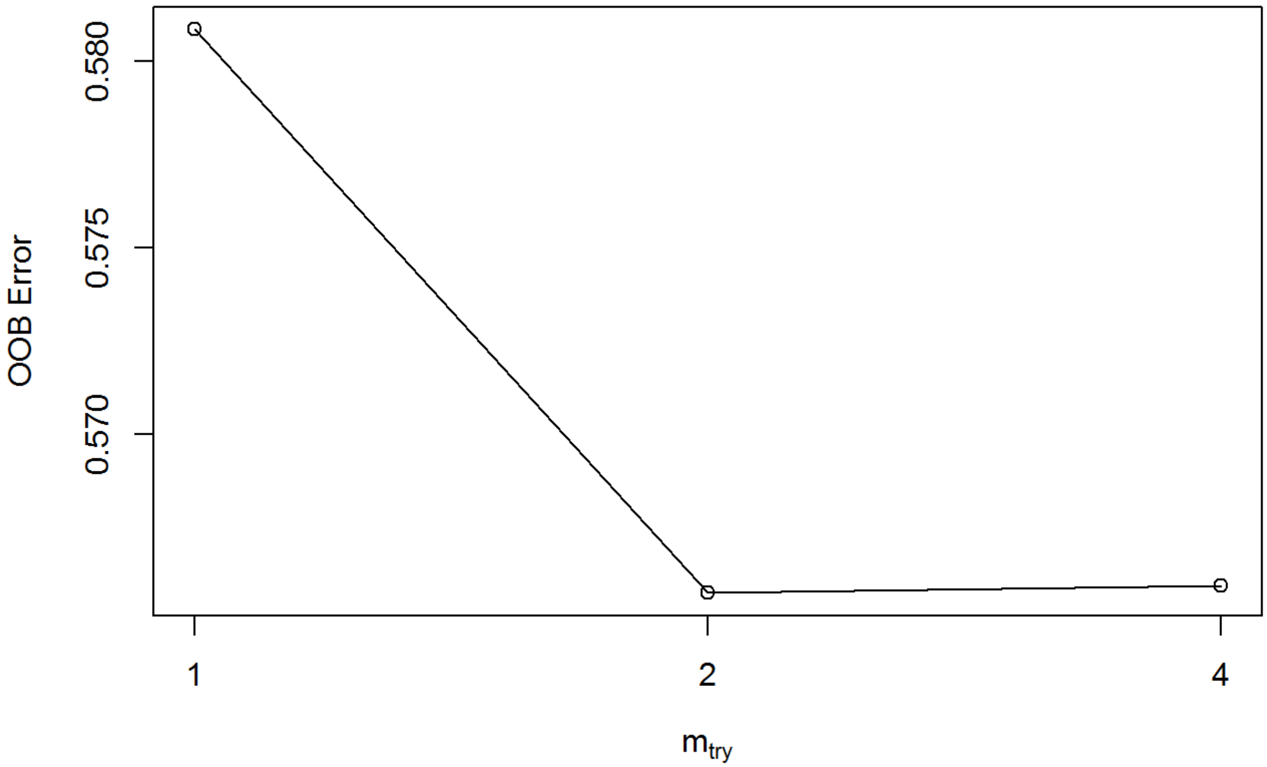
```
##                    MeanDecreaseGini
## ingang                     24.48160
## schoolaltercation          34.23191
## outsidealtercation         37.09968
## schoolweapon               22.51555
## outsideweapon              31.40988
## hurtingself                42.02763
## ciguse                     43.71706
## tobacco                    17.59558
## ecstasy                    18.98888
## oxy                        17.80505
## otherdrug                  17.90843
## sexual                     50.04320
## pregnancy                  27.14882
## age2                      115.85805
## white                      35.26027
## black                      28.42008
## asian                      36.27921
## hispanic                   28.48058
## otherrace                  24.72503
## female                     50.48644
## alcohol                    42.54319
## marijuana                  41.75054
## heroin                     10.43022
## meth                       13.93494
```

```
plot(fit2.0)
```



```
#RF2: OOB error = 56.7%, tune model
fittune2.0 <- tuneRF(train2[c(2:25)], train2$grades2, ntreeTry = 500, mtryStart = 1, step
Factor = 2,
                     improve = 0.001, trace = TRUE, plot = TRUE)
```

```
## mtry = 1  OOB error = 58.08%
## Searching left ...
## Searching right ...
## mtry = 2     OOB error = 56.57%
## 0.02597403 0.001
## mtry = 4     OOB error = 56.6%
## -0.0003603604 0.001
```



```
fittune2.0
```

```
##       mtry  OOBError
## 1.OOB    1 0.5808359
## 2.OOB    2 0.5657492
## 4.OOB    4 0.5659531
```

```
#RF2: Two variables per split minimizes OOB error; tuning model
fit2.1 <- randomForest(grades2 ~ ingang + schoolaltercation + outsidealtercation
                       + schoolweapon + outsideweapon + hurtingself + ciguse + tobacco
                       + ecstasy + oxy + otherdrug + sexual + pregnancy + age2 + white
                       + black + asian + hispanic + otherrace + female + alcohol + mariju
ana
                       + heroin + meth, data = train2, mtry=2)
fit2.1
```

```
##
## Call:
##  randomForest(formula = grades2 ~ ingang + schoolaltercation +      outsidealtercation
+ schoolweapon + outsideweapon + hurtingself +      ciguse + tobacco + ecstasy + oxy + ot
herdrug + sexual + pregnancy +      age2 + white + black + asian + hispanic + otherrace +
female +      alcohol + marijuana + heroin + meth, data = train2, mtry = 2)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 56.62%
## Confusion matrix:
##                   Mostly A's Mostly B's Mostly C's Mostly D's
## Mostly A's                57        853         55          0
## Mostly B's                26       1809        183          0
## Mostly C's                 9       1096        262          0
## Mostly D's                 2        305        133          0
## Mostly E's or F's          0         67         48          0
##                   Mostly E's or F's class.error
## Mostly A's                        0   0.9409326
## Mostly B's                        0   0.1035679
## Mostly C's                        0   0.8083394
## Mostly D's                        0   1.0000000
## Mostly E's or F's                 0   1.0000000
```

```
#RF2: Unfortunately, the OOB error is still fairly high, but we will test the model anywa
y
pred.rf.train2 <- predict(fit2.1, train2, type='prob')
pred.rf.test2 <- predict(fit2.1, test2, type='prob')
input.rf2 <- rbind(data.frame(model = "train", d = train2$grades2, m = pred.rf.train2),
                   data.frame(model = "test", d = test2$grades2, m = pred.rf.test2))

#RF2: Plot ROC for grade = Mostly A's; resulting plot switches axis of Mostly A's and not
 A's; test AUC = 0.611
a2 <- input.rf2
a2$d <- as.factor(a2$d)
revalue(a2$d, c("Mostly B's" = "Not A's")) -> a2$d
revalue(a2$d, c("Mostly C's" = "Not A's")) -> a2$d
revalue(a2$d, c("Mostly D's" = "Not A's")) -> a2$d
revalue(a2$d, c("Mostly E's or F's" = "Not A's")) -> a2$d
roc.rf6 <- ggplot(a2, aes(d = d, model = model, m = m.Mostly.A.s, colour = model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf6)
```

```
##   PANEL group       AUC
## 1     1     1 0.2860848
## 2     1     2 0.4005472
```

```
#RF2: Plot ROC for grade = Mostly B's; resulting plot switches axis of Mostly B's and not B's; test AUC = 0.366
b2 <- input.rf2
b2$d <- as.factor(b2$d)
revalue(b2$d, c("Mostly A's" = "Not B's")) -> b2$d
revalue(b2$d, c("Mostly C's" = "Not B's")) -> b2$d
revalue(b2$d, c("Mostly D's" = "Not B's")) -> b2$d
revalue(b2$d, c("Mostly E's or F's" = "Not B's")) -> b2$d
roc.rf7 <- ggplot(b2, aes(d = d, model = model, m = m.Mostly.B.s, colour = model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf7)
```

```
##   PANEL group       AUC
## 1     1     1 0.3654196
## 2     1     2 0.4237809
```

```
#RF2: Plot ROC for grade = Mostly C's; resulting plot switches axis of Mostly C's and not C's; test AUC = 0.315
c2 <- input.rf2
c2$d <- as.factor(c2$d)
revalue(c2$d, c("Mostly A's" = "Not C's")) -> c2$d
revalue(c2$d, c("Mostly B's" = "Not C's")) -> c2$d
revalue(c2$d, c("Mostly D's" = "Not C's")) -> c2$d
revalue(c2$d, c("Mostly E's or F's" = "Not C's")) -> c2$d
roc.rf8 <- ggplot(c2, aes(d = d, model = model, m = m.Mostly.C.s, colour = model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf8)
```

```
##   PANEL group       AUC
## 1     1     1 0.3127046
## 2     1     2 0.3812522
```

```
#RF2: Plot ROC for grade = Mostly D's; resulting plot switches axis of Mostly D's and not
 D's; test AUC = 0.237
d2 <- input.rf2
d2$d <- as.factor(d2$d)
revalue(d2$d, c("Mostly A's" = "Not D's")) -> d2$d
revalue(d2$d, c("Mostly B's" = "Not D's")) -> d2$d
revalue(d2$d, c("Mostly C's" = "Not D's")) -> d2$d
revalue(d2$d, c("Mostly E's or F's" = "Not D's")) -> d2$d
roc.rf9 <- ggplot(d2, aes(d = d, model = model, m = m.Mostly.D.s, colour = model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf9)
```

```
##   PANEL group       AUC
## 1     1     1 0.2416194
## 2     1     2 0.3249516
```

```
#RF2: Plot ROC for grade = Mostly E's or F's; resulting plot switches axis; test AUC = 0.
106
e2 <- input.rf2
e2$d <- as.factor(e2$d)
revalue(e2$d, c("Mostly A's" = "Not E's")) -> e2$d
revalue(e2$d, c("Mostly B's" = "Not E's")) -> e2$d
revalue(e2$d, c("Mostly C's" = "Not E's")) -> e2$d
revalue(e2$d, c("Mostly D's" = "Not E's")) -> e2$d
roc.rf10 <- ggplot(e2, aes(d = d, model = model, m = m.Mostly.E.s.or.F.s, colour =
model)) +
  geom_roc(show.legend = TRUE) + style_roc() + ggtitle("Train")
calc_auc(roc.rf10)
```

```
##   PANEL group       AUC
## 1     1     1 0.1021303
## 2     1     2 0.3260611
```

```
#RF2: Predict activity for validate sample; only 42.7% were correctly classified using RF
2 via imputation
validate2$gradepred <- predict(fit2.1, validate2, type='class')
validate2$correct[validate2$grades2 == validate2$gradepred] <- 1
validate2$correct[validate2$grades2 != validate2$gradepred] <- 0
mean(validate2$correct)
```

```
## [1] 0.426306
```

```
#RF2: Variable importance; age has the most importance, followed by sexual activity, gend
er, cigarette use, being Asian, and marijuana use.
fit2.1$importance
```

```
##                    MeanDecreaseGini
## ingang                    11.278866
## schoolaltercation         16.359434
## outsidealtercation        16.839263
## schoolweapon              10.570157
## outsideweapon             13.019746
## hurtingself               13.449536
## ciguse                    23.260606
## tobacco                    7.816593
## ecstasy                    8.958019
## oxy                        9.088635
## otherdrug                  9.639591
## sexual                    29.001844
## pregnancy                 12.770381
## age2                      34.212960
## white                     14.957538
## black                     13.660365
## asian                     22.725731
## hispanic                  13.797576
## otherrace                  9.470836
## female                    24.425412
## alcohol                   18.162452
## marijuana                 19.953078
## heroin                     6.041695
## meth                       7.963536
```

```
#RF2: Even after imputation, RF provides low predictability power

#Concluding remarks: Both RF models demonstrate that age has the greatest importance, whi
le risky behaviors such as sexual activity and marijuana use are also important.
```

Ordered Logistic Regression

Since the dependent variable had ranked categorical responses, an ordered logistic regression was conducted. When analyzing the different models, the Mean-F1 score and the pseudo R-squared value were taken into consideration. After analyzing several models, it was determined that Model 1 was optimal. Unfortunately, the model had a low Mean-F1 score of 0.742487. However, the McFadden's pseudo R-squared value was 0.3381522. A value between 0.2 and 0.4 indicates that the model is a good fit. It should be noted that in many of the models, pregnancy, sexual activity, in-school altercation, cigarette use, and marijuana have a negative, statistically significant association with grades.

```
##Model 1
m1<- polr(factor(grades) ~ sexual + pregnancy +schoolaltercation + outsidealtercation
          + outsideweapon +  oxy  + +alcohol + ciguse + marijuana + age2 + white + black
          + asian + hispanic + female, data = train)
summary(m1)
```

```
## Call:
## polr(formula = factor(grades) ~ sexual + pregnancy + schoolaltercation +
##     outsidealtercation + outsideweapon + oxy + +alcohol + ciguse +
##     marijuana + age2 + white + black + asian + hispanic + female,
##     data = train)
##
## Coefficients:
##                        Value Std. Error t value
## sexual             -0.31496    0.09423 -3.3426
## pregnancy          -0.63834    0.20515 -3.1116
## schoolaltercation  -0.48346    0.16453 -2.9383
## outsidealtercation -0.25938    0.12315 -2.1063
## outsideweapon      -0.21032    0.15667 -1.3425
## oxy                 0.50434    0.47761  1.0560
## alcohol            -0.07024    0.09911 -0.7087
## ciguse             -0.74288    0.14780 -5.0263
## marijuana          -0.56021    0.11927 -4.6971
## age2                0.01470    0.03400  0.4322
## white               0.78043    0.13932  5.6019
## black               0.05931    0.16545  0.3585
## asian               1.45412    0.18152  8.0106
## hispanic            0.08875    0.14297  0.6208
## female              0.46659    0.08208  5.6848
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -3.9684  0.5889    -6.7387
## 2|3 -2.0374  0.5695    -3.5773
## 3|4 -0.2860  0.5678    -0.5037
## 4|5  1.8432  0.5684     3.2430
##
## Residual Deviance: 5399.833
## AIC: 5437.833
```

```
grades<- predict(m1, test)
id <- test$id
myPredictions<- cbind.data.frame(id, grades)
meanf1(is.na(test$grades), is.na(myPredictions$grades))
```

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

```
pR2(m1)
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -2.699917e+03 -2.902245e+03  4.046575e+02  6.971455e-02  1.669733e-01
##          r2CU
##  1.800763e-01
```

```
##mean-f1: 0.742487
##Psuedo R-squared values:
# McFadden: 0.3381522
# r2ML:  0.7145596
# r2CU: 0.7325338
##Sexual, pregnancy, school altercation, outside altercation, ciguse and marijuana are st
atistically significant.

##Model 2
m2<- polr(factor(grades) ~ sexual + pregnancy +schoolaltercation + outsidealtercation
          + outsideweapon +  oxy+alcohol + ciguse + marijuana + hurtingself + age2 + whit
e + black
          + asian + hispanic + female, data = train)
summary(m2)
```

```
## Call:
## polr(formula = factor(grades) ~ sexual + pregnancy + schoolaltercation +
##     outsidealtercation + outsideweapon + oxy + alcohol + ciguse +
##     marijuana + hurtingself + age2 + white + black + asian +
##     hispanic + female, data = train)
##
## Coefficients:
##                       Value Std. Error t value
## sexual             -0.31399    0.09428 -3.3305
## pregnancy          -0.63448    0.20533 -3.0901
## schoolaltercation  -0.48210    0.16454 -2.9299
## outsidealtercation -0.25817    0.12319 -2.0957
## outsideweapon      -0.20809    0.15677 -1.3274
## oxy                 0.51190    0.47842  1.0700
## alcohol            -0.06985    0.09912 -0.7047
## ciguse             -0.73752    0.14854 -4.9652
## marijuana          -0.55916    0.11931 -4.6865
## hurtingself        -0.04490    0.12490 -0.3595
## age2                0.01388    0.03408  0.4073
## white               0.78124    0.13933  5.6070
## black               0.05728    0.16554  0.3460
## asian               1.45469    0.18153  8.0135
## hispanic            0.08791    0.14298  0.6148
## female              0.47148    0.08320  5.6667
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -3.9833  0.5904    -6.7472
## 2|3 -2.0519  0.5710    -3.5937
## 3|4 -0.3004  0.5692    -0.5278
## 4|5  1.8288  0.5698     3.2098
##
## Residual Deviance: 5399.704
## AIC: 5439.704
```

```
grades<- predict(m2, test)
id <- test$id
myPredictions<- cbind.data.frame(id, grades)
meanf1(is.na(test$grades), is.na(myPredictions$grades))
```

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

```
pR2(m2)
```

```
##           llh       llhNull            G2      McFadden          r2ML 
## -2.699852e+03 -2.902245e+03  4.047868e+02  6.973683e-02  1.670219e-01 
##          r2CU 
##  1.801288e-01
```

```
##mean-f1:0.551699
##Psuedo R-squared values:
# McFadden: 0.5891756
# r2ML: 0.9693411
# r2CU: 0.9719647
##Sexual, pregnancy, school altercation, ciguse and marijuana are statistically significant.


##Model 3
m3<- polr(factor(grades) ~ sexual + pregnancy +schoolaltercation + outsidealtercation +schoolweapon
          + outsideweapon +  oxy  + alcohol + ciguse +  marijuana + tobacco + age2 + white + black
          + asian + hispanic + female, data = train)
summary(m3)
```

```
## Call:
## polr(formula = factor(grades) ~ sexual + pregnancy + schoolaltercation +
##     outsidealtercation + schoolweapon + outsideweapon + oxy +
##     alcohol + ciguse + marijuana + tobacco + age2 + white + black +
##     asian + hispanic + female, data = train)
##
## Coefficients:
##                       Value Std. Error t value
## sexual             -0.31469    0.09424 -3.3394
## pregnancy          -0.63709    0.20526 -3.1038
## schoolaltercation  -0.49194    0.16529 -2.9762
## outsidealtercation -0.26292    0.12340 -2.1306
## schoolweapon        0.12712    0.26850  0.4734
## outsideweapon      -0.25494    0.18014 -1.4152
## oxy                 0.49561    0.47923  1.0342
## alcohol            -0.07353    0.09937 -0.7399
## ciguse             -0.74657    0.14802 -5.0438
## marijuana          -0.56310    0.11940 -4.7161
## tobacco             0.08399    0.29300  0.2866
## age2                0.01437    0.03402  0.4223
## white               0.78262    0.13948  5.6108
## black               0.06074    0.16548  0.3671
## asian               1.45536    0.18151  8.0179
## hispanic            0.09062    0.14296  0.6339
## female              0.46858    0.08244  5.6840
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -3.9719  0.5891    -6.7418
## 2|3 -2.0415  0.5698    -3.5826
## 3|4 -0.2898  0.5681    -0.5101
## 4|5  1.8399  0.5686     3.2357
##
## Residual Deviance: 5399.526
## AIC: 5441.526
```

```
grades<- predict(m3, test)
id <- test$id
myPredictions<- cbind.data.frame(id, grades)
meanf1(is.na(test$grades), is.na(myPredictions$grades))
```

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

```
pR2(m3)
```

```
##           llh       llhNull            G2      McFadden          r2ML 
## -2.699763e+03 -2.902245e+03  4.049645e+02  6.976744e-02  1.670888e-01 
##          r2CU 
##  1.802009e-01
```

```
##mean-f1: 0.7382061
##Psuedo R-squared values:
# McFadden: 0.3406277
# r2ML: 0.7186248
# r2CU: 0.7364225
##Sexual, pregnancy, school altercation, outside altercation, ciguse and marijuana are st
atistically significant.

##Model 4
m4<- polr(factor(grades) ~ sexual +pregnancy + ingang + schoolaltercation + outsidealterc
ation + schoolweapon
          + outsideweapon +  oxy  + alcohol + ciguse +  marijuana + tobacco + age2 + whit
e + black
          + asian + hispanic + female, data = train)
summary(m4)
```

```
## Call:
## polr(formula = factor(grades) ~ sexual + pregnancy + ingang +
##     schoolaltercation + outsidealtercation + schoolweapon + outsideweapon +
##     oxy + alcohol + ciguse + marijuana + tobacco + age2 + white +
##     black + asian + hispanic + female, data = train)
##
## Coefficients:
##                        Value Std. Error t value
## sexual              -0.31255    0.09429 -3.3148
## pregnancy           -0.62377    0.20621 -3.0249
## ingang              -0.18330    0.28373 -0.6460
## schoolaltercation   -0.47835    0.16662 -2.8709
## outsidealtercation  -0.25819    0.12360 -2.0890
## schoolweapon         0.14722    0.27036  0.5446
## outsideweapon       -0.24421    0.18099 -1.3493
## oxy                  0.48732    0.47861  1.0182
## alcohol             -0.07471    0.09939 -0.7516
## ciguse              -0.74934    0.14811 -5.0595
## marijuana           -0.56156    0.11944 -4.7016
## tobacco              0.09306    0.29342  0.3172
## age2                 0.01400    0.03402  0.4116
## white                0.78283    0.13948  5.6125
## black                0.06262    0.16551  0.3784
## asian                1.45790    0.18156  8.0299
## hispanic             0.09190    0.14297  0.6428
## female               0.46555    0.08257  5.6385
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -3.9811  0.5892    -6.7562
## 2|3 -2.0485  0.5699    -3.5947
## 3|4 -0.2957  0.5681    -0.5206
## 4|5  1.8337  0.5686     3.2247
##
## Residual Deviance: 5399.11
## AIC: 5443.11
```

```
grades<- predict(m4, test)
id <- test$id
myPredictions<- cbind.data.frame(id, grades)
meanf1(is.na(test$grades), is.na(myPredictions$grades))
```

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

```
pR2(m4)
```

```
##           llh       llhNull            G2      McFadden          r2ML 
## -2699.5547762 -2902.2452545   405.3809565     0.0698392     0.1672454 
##          r2CU 
##     0.1803697
```

```
##mean-f1: 0.7313642
##Psuedo R-squared values:
# McFadden: 0.3537376
# r2ML: 0.7393928
# r2CU: 0.7562858
##Sexual, pregnancy, school altercation, outside altercation, alcohol, ciguse, and mariju
ana are statistically signficant.
```

Conclusion

Overall, random forest and ordered logistic regression are preferred, however, there is low predictability power. This may be attributed to the limitations of the data. It is more than likely that there were other factors which are stronger predictors of grades, but were not captured in the model, such as household type, family stability, and the income of parents. Furthermore, the self-reported nature of the data may have decreased the strength of predictability. While limitations do exist, more research should be conducted in this area, which can better inform schools and policymakers.

Application in the Real World

The next step would be to ideally create the basis of a scoring engine. This engine could take into account of other academic, behavioral, and environmental factors which were not described in this study. Such an engine could help to support the mitigation of risky behaviors.