Factors that Influence Youth Drug Use

4/13/23

Ava Delanty

Abstract

This project investigates factors associated with drug and alcohol use in youth groups, using a subset of the 2020 National Survey on Drug Consumption and Health dataset. The study uses decision trees, boosting, bagging, and random forest to predict cigarette, marijuana, and alcohol use in youth. The results indicate that factors such as the youth's grade level, number of school absences, race, health, and metropolitan area size were significant predictors of cigarette use. Marijuana use was predicted by the youth's perception of their close friends, parents, and peers' attitudes towards drug use, as well as their race and health. Grade level and school absences were significant predictors of alcohol use. The study emphasizes the importance of considering demographic and environmental factors when studying drug and alcohol use in youth.

Introduction

Multiple factors can influence what correlates to youth drug and alcohol use. In this report, we will be specifically looking at cigarette, marijuana, and alcohol use in the youth population. To determine what factors influence the youth we will use decision trees and the ensemble methods of boosting and bagging. There will be three classifications binary (e.g. has or has not used cigarettes), multi-class (e.g. number of days of marijuana use in the past year), and regression (e.g. alcohol frequency in the past year). The goal is to perform accurate machine learning models and predictions using decision trees to select which response variables within the 2020 National Survey on Drug Use and Health (NSDUH) dataset influence youth cigarette, marijuana, and alcohol use.

The original dataset from NSDUH has around 33,000 responses and around 3000 questions about drug use and health. In order to conduct concise modeling, the three predicting variables and a subset of the data will be used. This includes categories of basic demographics (e.g. sex, race, household income), youth-specific demographics(e.g. parental presence in household and school attendance), and youth experience questions (e.g. education around drugs, parental involvement, and influence from parents and peers about drugs/alcohol). The importance of these findings will help lead to better ways society and the government can reduce youth drug and substance usage.

Theoretical Background

Decision Trees: Fitting classification trees, Fitting regression trees, and Pruning

Decision trees, a machine learning approach used for classification and regression tasks, will be the method used throughout this study. Decision trees split the data into subsets based on the values of selected features until a decision is reached. Classification trees specifically have the goal to assign a class label to each input instance. Regression trees are used where the goal is to predict a continuous numerical value. For fitting a classification tree, the algorithm selects a

feature that splits the data into subsets with the greatest homogeneity in class labels. The process is repeated recursively. For fitting a regression tree instead of class labels, it splits data based on the feature that produces the greatest reduction in the variance of the target variable.

Decision trees, however, have a tendency to overfit the training data, which increases complexity and results in subpar generalization performance on potentially unseen data. Pruning is one of the techniques employed in this study as a result. In order to solve this issue, pruning is performed to remove specific tree components that do not improve the model's overall accuracy. Overall, pruning is an excellent technique that will be utilized to determine the factors that affect young people's drug and alcohol use.

Random Forest and Bagging

Bagging with random forest will be one of the ensemble approaches used. By creating numerous iterations of a model utilizing diverse random subsets of the training data, the technique called bagging or bootstrap aggregating is used to enhance a model's performance. Each model is independently trained to produce a forecast. The forecasts from all the models are combined to produce the final projection. Bagging increases the accuracy and performance of the model by merging the outputs of numerous models to reduce the variance of the predictions. A version of bagging called random forest adds more randomization to the feature selection process, improving performance. One of the models in this study will combine bagging and the random forest function.

Boosting

Boosting will be the final ensemble technique used. By creating a series of more powerful models that concentrate on the incorrectly classified instances, boosting increases a model's accuracy. The goal is to train a series of models, with each one attempting to fix the mistakes made by the one before it. A weighted subset of the training data is used to train the base model after each instance in the training data is given a weight depending on its misclassification rate. The procedure is repeated until the stopping condition is satisfied, with the next models being trained on the instances of the prior model that were incorrectly classified with higher weights. Two of the models in this study will use the boosting method.

Methodology

Data Processing/Cleaning

Using the subset data frame from YouthParse.r that was provided on Canvas, three subsets from that data frame were created to perform the three classification problems: has or has not used cigarettes, the number of marijuana use in the past year, and alcohol frequency in the past year. The variables pulled out to be used as the predictor variables were tobflag (any tobacco ever used), mrjydays (number of days of marijuana in the past year), and iralcfy (alcohol

frequency past year). All of the subset datasets were cleaned by omitting missing values, converting categorical data to factors, and renaming variable label attributes for better reading.

For the **binary classification**, the predictor variable was tobflag split into two categories has or has not used cigarettes. The response variables used were all variables from the demographic columns. The following are the demographic variables:

1) Binary sex (male or female)
2) Race (7 categories)
3) Overall Health (4 categories)
4) Now going to school (yes or no)
5) What grade is now/will be in (11 categories and a blank/skip option)
6) How many days were skipped school in the past month (1-30, or blank/skip)
7) Mother in the household (yes or no)
8) Father in the household (yes or no)
9) Total family income (4 categories)
10) Got government assistance (yes or no)
11) Poverty level (4 categories)
12) Population density (greater than 1 million, less or can't be determined)
13) metropolitan size status (large metro, small metro, nonmetro)

For the **multi-class classification**, the predictor variable was mrjydays or the recorded number of days used marijuana in the past year split into the categories: 1-11 days, 12-49 days, 50-99 days, 100-299 days, 300-365 days, Non-user or No past year use.

The response variables were from youth experience questions and demographic questions. The questions for youth experiences included the following:

1) How youth felt about going to school in the past year
2) The teacher let the youth know doing a good job in the past year
3) Grade average for the last grading period completed
4) Students in youth grade that use marijuana
5) Parents limit time out on school nights in the past year
6) How youth thinks about how parents feel about youth trying marijuana
7) How youth feels about peers trying marijuana
8) How do youth think close friends feel about trying marijuana
9) Who youth talks with about serious problems
10) The youth talked with the parent about the danger of tobacco/alcohol/drug
11) Youths see alcohol or drug prevention messages outside of school
12) If the youth had any drug or alcohol education in school

Lastly, for the **regression analysis**, the predictor variable used was iralcfy or alcohol frequency in the past year split into three variables never used alcohol, used alcohol in the past year, and did not use alcohol in the past year. The response variables were the demographic columns as seen above in the binary classification.

Models Created

For the **binary classification**, the training set was split into 1000 observations of the 3,426 rows of data. The remainder went into the testing set. After plotting the decision tree, the pruning and cross-validation technique was used to see if that eliminated any variables. Additionally compared both test error rates on the pruned and unpruned trees, and the training error rates. Boosting was then used to see which model overall performed the best and gave us which variables influenced cigarette use the most.

For the **multi-class classification**, the training set was split into 1000 observations of the 4,269 rows of data. After creating a decision tree, the pruning and cross-validation technique was used to see if that eliminated any variables. Additionally compared both test error rates on the pruned and unpruned trees, and the training error rates. Random forest and bagging with the training and test split even (50/50) were created to compare MSEs, how many variables to use when splitting each tree, and how well the models perform on the test set.

For the **regression analysis**, the training set was split into 1000 observations. After plotting the decision tree, the pruning and cross-validation technique was used to see if that eliminated any variables. Boosting was then used to see which model overall performed the best and gave us which variables influenced alcohol frequency use the most.

Results

Binary Classification:

Problem: Demographic factors that influence if youth have used or have not used cigarettes.

Decision trees:

Results gave a training error rate of 0.095 or 9%. The decision tree that was produced used the variables for what grade they are in, days of missed school, family income, race, overall health, and metropolitan size. Estimating the test error gave an accuracy of 89% in the test data. After cross-validating and finding the optimal K, a pruned tree was created that overall produced a much more concise tree with fewer variables. The variables used were what grade they are in, race, overall health, and metropolitan size. The pruned tree training error rate was 0.095 or 9%, indicating that either the pruned or unpruned tree resulted in the same errors. Additionally, the test error gave an accuracy of 89% of the test data which results in the pruning tree being the better one to use since it has fewer variables.

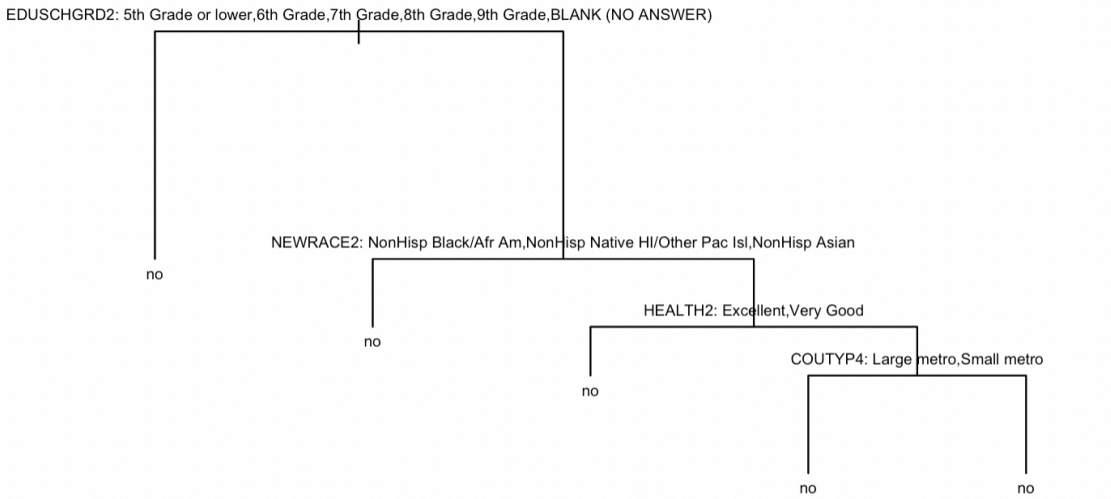

Figure 1: Pruned Decision Tree for Binary Classification

Boosting:

Boosting was then performed to see which variables influence cigarette use the most. The most important variables were grade level and the number of days of school that were skipped in the past month. The MSE was 1.07 which was a lot bigger than the MSE of the pruned tree 0.424. This indicates that pruning might be the better-performing model. In conclusion, the most important variables that influence whether youth have used cigarettes or not are what grade they are in, days of school skipped, race, overall health, and metropolitan size.

Multi-Class Classification:

Problem: Youth experience questions and demographics that influence the number of days using marijuana in the past year.

Decision Trees:

Results gave a training error rate of 0.096 or 9%. The decision tree that was produced used 10 different response variables both from demographics (4 variables) and youth experiences (6 variables). The variables were how youth think close friends feel about trying marijuana, students in youth grade that use marijuana, how youth feels about peers trying marijuana, race, what grade youth is in, how youth thinks about how parents feel about youth trying marijuana, overall health, who youth talks with about serious problems, if parents limit time out on school nights in the past year and metropolitan size. Estimating the test error gave an accuracy of 86% in the test data.

After cross-validating and finding the optimal K, 5 variables were used in the pruned tree with 3 from youth experiences and 2 from demographics. The variables were how youth think close friends feel about trying marijuana, students in youth grade that use marijuana, how youth thinks about how parents feel about youth trying marijuana, race, and overall health. The pruned tree training error rate was 0.103 or 10%. Additionally, the test error gave an accuracy of 86% of the test data which therefore the pruned tree will be better for simplicity of the number of variables used.

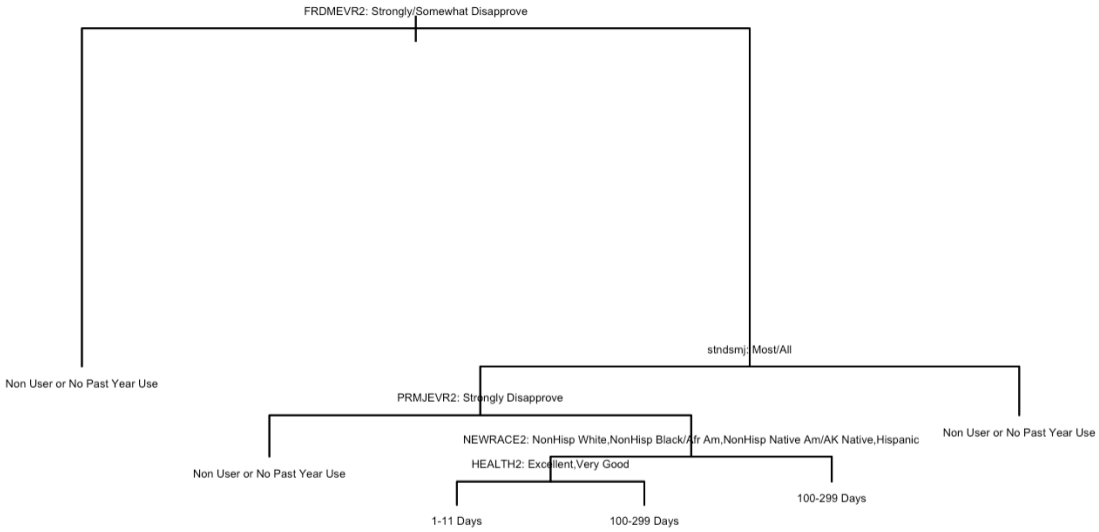

Figure 2: Pruned Decision Tree for Multi-Class Classification

Bagging and Random Forest:

After using the ensemble method bagging the model produced an MSE of 2.19 using all variables at the splitting of each tree. Growing a random forest uses a smaller value of the mtry argument which in this case was lowered to 12. The test set MSE was 1.99 which indicates that random forests yielded an improvement over bagging. Based on the mean decrease accuracy plot the results indicate that across all of the trees considered in the random forest, the two most important variables were how youth think close friends feel about trying marijuana and how youth feels about peers trying marijuana.

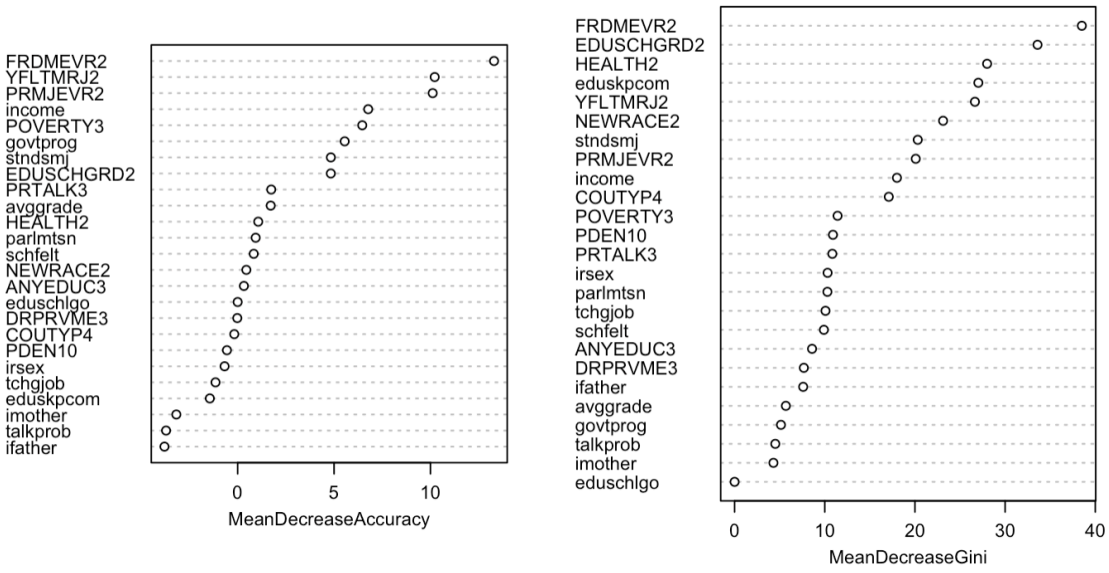

Figure 3: Random Forest Variable Importance Plot

Regression Analysis:

Problem: Demographics that influence youths' alcohol frequency in the past year.

Decision Trees:

Results gave a residual mean deviance of 26%. The decision tree that was produced used the variables of what grade they are in and days of missed school. After cross-validating and finding the optimal K, a pruned tree was created that produced a similar tree but with fewer nodes and the same variables. The pruned tree had a slightly larger residual mean deviance of 26.6% than the unpruned tree. The MSE associated with the regression tree was .29 and the square root of the MSE was around 0.54 indicating that this model leads to test predictions that are on average .54 or close to 1 which was using alcohol in the past year.

Boosting:

Boosting was then performed to see which variables influence alcohol use the most. The most important variables were grade level and race. The MSE was .39 which indicates that boosting was the better-performing model than decision tree making. In conclusion, the most important variables that influence youths' alcohol frequency are what grade they are in and their race.

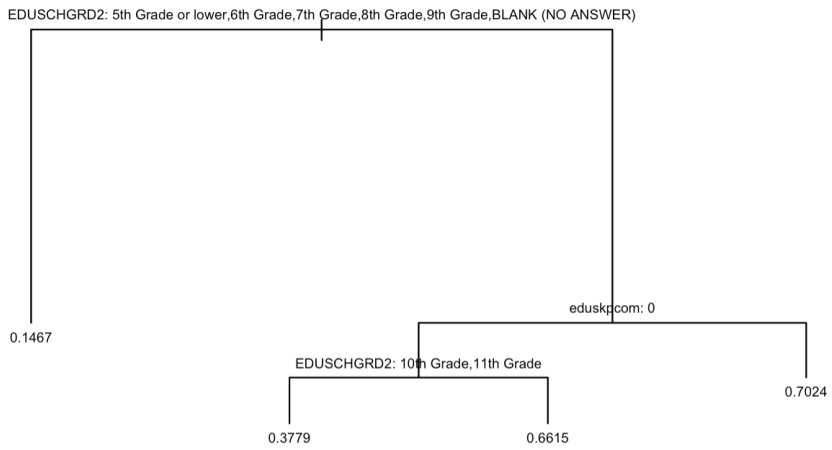

Figure 4: Unpruned Decision tree for Regression

Discussion

The results from the three classification problems in this project showcase how the variables can predict youth drug and alcohol use. One noticeable pruned tree model from Figure 1 shows a binary classification of cigarette use with three levels of nodes splitting the data based on demographics. The majority class was “not used cigarettes” or 0 with a proportion of .905 and the minority class of “used cigarettes” or 1 with a proportion of 0.095. The first split occurred on the variable of what grade the youth is in. If the participant completed 5th grade or lower, 6th, 7th, 8th, or 9th grade, or provided no answer they were assigned to the left branch classified as 0. Otherwise, if the participant was in 10th, 11th, or 12th grade they were assigned to the right branch.

The next level and split occurred on the variable race/ethnicity where if the participant is non-Hispanic Black/African American, non-Hispanic Native Hawaiian/Other Pacific Islander, or non-Hispanic Asian, they are assigned to the left branch and classified as 0. Otherwise, if the participant is non-Hispanic White, non-Hispanic Native American/Alaska Native, non-Hispanic multiracial, or Hispanic, they were assigned to the right branch for further splitting. The next split occurred on the variable which represents the participant's health status. If the participant rated their health as excellent or very good, they are assigned to the left branch and classified as 0. Otherwise, if the participant rates their health as good, fair, or poor, they are assigned to the right branch for further splitting on the variable metropolitan size.

One noteworthy end node in this tree model was node 31 which represents participants who live in non-metropolitan areas and have completed 10th grade, 11th grade, or 12th grade, are non-Hispanic White, non-Hispanic Native American/Alaska Native, non-Hispanic

multiracial, or Hispanic, and rate their health as good, fair, or poor. The data in this end node consists of 33 observations with a deviance of 45.72, and the majority class is 0 with a proportion of 0.51515, while the minority class is 1 with a proportion of 0.48485. This node suggests that this group has a higher risk of having used cigarettes than other groups. The path to this node involves a split based on education level, race/ethnicity, and health status. It can be assumed based on this model that most of this group is older, in worse health, and not from a large metropolitan area.

Some of the variables used throughout this project are the same information but coded into either binary, ordinal, or numerical variables. The predictions and modeling will change, for example with binary variables there are only two categories and the model will assume that the effect of the variable is the same regardless of the specific category. An example of this can be using sex from the demographics data. This variable is appropriate for regression and classification models. Ordinal categorical variables, allow the model to capture the effect of the variable and can be used for regression and classification models. However, the distance between the categories is not equal. An example of this can be seen with the variable health where excellent or good health is likely to have a different effect on the outcome variable compared to youth that has poor health. Lastly, numerical variables are continuous and can take on any value which allows the model to capture the effect of the variable more accurately. For example one of the variables used was the number of days of skipped school. Numerical variables are good for regression analysis and provide good predictions overall.

The findings suggest that certain variables such as the grade level of youth, the number of school absences, race/ethnicity, overall health, and metropolitan size are highly significant in predicting cigarette use. Similarly, for predicting marijuana use among youth, the crucial variables are the perception of the youth regarding their friends' and peers' attitudes towards trying marijuana, as well as their thoughts on their parent's perspective. Additionally, demographic variables such as race and overall health are contributing factors in influencing marijuana use. Finally, the variables of the grade level of youth and the frequency of school absences were found to be predictive of alcohol use.

This may indicate that education around drugs, schooling, parental involvement, health history, and influences from parents and peers about drugs are the biggest factors in youth drug use. Further analysis is needed for better accuracy and understanding of how these variables play a role, but this can be interpreted as better education needs to be put in place for the youth. As a data scientist, it is important that when identifying these groups as being at a higher risk of drug or alcohol use, interventions and tailored policies can be developed to address these groups' needs to reduce the use in youth groups. In order to discuss these findings in an ethical way some good starting points would be to ensure the protection of an individual's privacy, avoid bias, and use the data for good intentions.

Conclusion

The results from this study indicate that demographics, education, views, and perspectives on drugs influence youth drug and alcohol usage. The binary, multi-class, and regression models were made which showed that cigarette, marijuana, and alcohol use relied heavily on demographics related to race/ethnicity, health, education, and schooling. Whereas with marijuana use more factors played in regarding youths' friends' and peers' attitudes towards trying marijuana, as well as their thoughts on their parent's perspective. Going forward creating new policies and providing more education to parents and youth to address these issues will be important to reduce substance use.

References

National Survey on Drug Use and Health 2020 (NSDUH-2020-DS0001) | SAMHDA. (n.d.). https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001

2020 NSDUH Public Use File CodeBook (NSDUH-2020-DS001) | SAMHDA. (n.d.).

https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH-2020/NSDUH-2020-datasets/NSDUH-2020-DS0001/NSDUH-2020-DS0001-info/NSDUH-2020-DS0001-info-codebook.pdf

YouthParse.r and youth_data.Rdata from Canvas

Appendix

Loading data set from YouthParse.r:

```
load('~/DATA 5322/youth_data.rdata')

#Omitting NAs

df <- na.omit(df)
```

Pulling binary variable for whether youth used cigs or not:

```
tobflag <- df$tobflag
```

Cleaning the demographic data frame and omitting an NAs:

```
#selecting the demographic columns

df_demog <- df %>% select(demographic_cols)

# cleaning

df_demo_clean <- df_demog%>%

  mutate(irsex = factor(irsex, levels = c(1,2),labels = c('male','female')),

        NEWRACE2 = factor(NEWRACE2, levels = c(1,2,3,4,5,6,7), labels = c("NonHisp White","NonHisp Black/Afr Am","NonHisp Native Am/AK Native","NonHisp Native HI/Other Pac Isl","NonHisp Asian","NonHisp more than one race","Hispanic")),

        HEALTH2 = factor(HEALTH2, levels = c(1,2,3,4), labels = c("Excellent","Very Good","Good","Fair/Poor"),ordered=TRUE),

        eduschlgo = factor(eduschlgo, levels = c(1,2),labels = c("Yes","No")),

        EDUSCHGRD2 = factor(EDUSCHGRD2, levels = c(1,2,3,4,5,6,7,8,9,10,11,98,99), labels = c("5th Grade or lower","6th Grade","7th Grade","8th Grade","9th Grade","10th Grade","11th Grade","12th Grade","College or university/1st year","College or university/2nd Year, 3rd year"," College or university/4th Year, 5th or higher year","BLANK (NO ANSWER)","LEGITIMATE SKIP"),ordered=TRUE),

        eduskpcom = factor(eduskpcom, levels = c(0:30,94,97,98,99),labels = c(0:30, "No Answer", "No Answer", "No Answer", "No Answer")),

        imother = factor(imother, levels = c(1,2,3,4), labels = c("Yes","No","Don't know","Over 18")),

        ifather = factor(ifather, levels = c(1,2,3,4), labels = c("Yes","No","Don't know","Over 18")),

        income = factor(income, levels = c(1,2,3,4), labels = c("Less than $20,000","$20,000 - $49,999","$50,000 - $74,999","$75,000 or More"),ordered=TRUE),

        govtprog = factor(govtprog, levels = c(1,2), labels =c("Yes","No")),
```

POVERTY3 = factor(POVERTY3, levels = c(1,2,3), labels = c("Living in Poverty","Income Up to 2X Fed Pov Thresh"," Income More Than 2X Fed Pov Thresh"),ordered=TRUE),

PDEN10 = factor(PDEN10, levels = c(1,2,3), labels = c(">1M people","<1M people","Can't be determined")),

COUTYP4 = factor(COUTYP4, levels = c(1,2,3), labels = c("Large metro","Small metro","Nonmetro")))

Cleaning youth experiences data frame:

```
#selecting the youth experience data

df_youth <- df %>% select(schfelt:rlgfrnd) # use all youth questions, start with schfelt and go through rlgfrnd

# selecting specific columns

df_youth_m <- subset(df_youth, select = c("schfelt", "tchgjob","avggrade","stndsmj","parlmtsn","PRMJEVR2","YFLTMRJ2","FRDMEVR2","talkprob","PRTALK3","DRPRVME3","ANYEDUC3"))

# cleaning

df_youth_clean_m <- df_youth_m %>%
  mutate(schfelt = factor(schfelt, levels = c(1,2),
                          labels = c("Liked A Lot/Kind of Liked","Didn't Like Very Much/Hated"))) %>%
  mutate(tchgjob = factor(tchgjob, levels = c(1,2),
                          labels = c("Always/Sometimes","Seldom/Never"))) %>%
  mutate(avggrade = factor(avggrade, levels = c(1,2),
                           labels = c("D or Lower","A, B, or C"))) %>%
  mutate(stndsmj = factor(stndsmj, levels = c(1,2),
                          labels = c("Most/All","None/Few"))) %>%
  mutate(parlmtsn = factor(parlmtsn, levels = c(1,2),
                          labels = c("Always/Sometimes","Seldom/Never"))) %>%
  mutate(PRMJEVR2 = factor(PRMJEVR2, levels = c(1,2),
                          labels = c("Strongly Disapprove","Somewhat Disapprove or Neither"))) %>%
  mutate(YFLTMRJ2 = factor(YFLTMRJ2, levels = c(1,2),
                          labels = c("Strongly/Somewhat Disapprove","Neither Approve Nor Disapprove"))) %>%
```

```
  mutate(FRDMEVR2 = factor(FRDMEVR2, levels = c(1,2),
                           labels = c("Strongly/Somewhat Disapprove","Neither Approve Nor
Disapprove"))) %>%
  mutate(talkprob = factor(talkprob, levels = c(1,2),
                           labels = c("No one","Someone"))) %>%
  mutate(PRTALK3 = factor(PRTALK3, levels = c(1,2),
                          labels = c("Yes","No"))) %>%
  mutate(DRPRVME3 = factor(DRPRVME3, levels = c(1,2),
                           labels = c("Yes","No"))) %>%
  mutate(ANYEDUC3 = factor(ANYEDUC3, levels = c(1,2),
                           labels = c("Yes","No")))
```

Binary Classification Problem:

Combing cigarette and cleaned demographics into a dataset:

```
df.cig2 <- cbind(tobflag,df_demo_clean)
#removing any Nas from data set
df.cig2 <- na.omit(df.cig2)
```

Decision Trees:

```
#training and test data with cigs and demographics
set.seed(1)
train <- sample(1:nrow(df.cig2), 1000)
cig.train <- df.cig2[train,]
cig.test <- df.cig2[-train,]
tree.cig <- tree(tobflag ~ ., df.cig2, subset = train)
tree.cig
summary(tree.cig)
```

Plotted Tree:

```
plot(tree.cig)
text(tree.cig, pretty = 0)
```

Accuracy and Matrix table:

```
set.seed(1)
test <- df.cig2$tobflag[-train]
```

```
tree.pred <- predict(tree.cig, cig.test,
   type = "class")
table <- table(tree.pred, test)
table
mean(tree.pred == test)
accuracy_Test <- sum(diag(table)) / sum(table)
print(paste('Accuracy for test', accuracy_Test))
```

Pruning tree:

```
set.seed(123)
tree.cig <- tree(tobflag ~ ., df.cig2, subset = train)
cv.cig <- cv.tree(tree.cig,K= 5,FUN = prune.tree)
optimal_K <- which.min(cv.cig$dev)
pruned_tree <- prune.tree(tree.cig, best = optimal_K)
```

Plotting cross-validation results:

```
par(mfrow = c(1, 2))
plot(cv.cig$size, cv.cig$dev, type = "b", xlab = "Number of terminal nodes", ylab =
"Cross-validation error rate", main = "Cross-validation results")
abline(v = optimal_K, lty = 2)
pruned_tree
```

Plotting pruned tree:

```
plot(pruned_tree, main = "Pruned Decision Tree")
text(pruned_tree, cex = 0.6, pretty = 0)
```

Accuracy and matrix table:

```
set.seed(1)
test <- df.cig2$tobflag[-train]
tree.pred <- predict(pruned_tree, cig.test,
   type = "class")
table <- table(tree.pred, test)
table
accuracy_Test <- sum(diag(table)) / sum(table)
print(paste('Accuracy for test', accuracy_Test))
```

MSE:

```
yhat <- predict(pruned_tree, cig.test)
yhat.boost.num <- as.numeric(as.character(yhat))
cig.test.num <- as.numeric(as.character(test))
mean((yhat.boost.num - cig.test.num)^2)
```

Boosting:

```
set.seed(1)
train <- sample(1:nrow(df.cig2), 1000)
cig.train <- df.cig2[train,]
cig.test <- df.cig2[-train,]
boost.cig <- gbm(tobflag ~ .- eduschlgo, data = df.cig2[train, ],
   distribution = "gaussian", n.trees = 100,
   interaction.depth = 4)
summary(boost.cig)
```

MSE:

```
yhat.boost <- predict(boost.cig,
   newdata = df.cig2[-train, ], n.trees = 50)
yhat.boost.num <- as.numeric(as.character(yhat.boost))
cig.test.num <- as.numeric(as.character(test))
mean((yhat.boost.num - cig.test.num)^2)
```

Multi-Class Classificaiton:

Pull multi-class variable to create a new factor variable:

```
mrjydays <- factor(df$mrjydays,
                   levels = c(1, 2, 3, 4, 5, 6),
                   labels = c("1-11 Days", "12-49 Days", "50-99 Days", "100-299 Days", "300-365
Days", "Non User or No Past Year Use"))
# convert the new variable to a factor
mrjydays <- factor(mrjydays)
```

Combining marijuana, demographics and youth experiences into a data set:

```
df.ma2 <- cbind(mrjydays,df_youth_clean_m, df_demo_clean)
#removing any Nas from data set
```

```
df.ma2 <- na.omit(df.ma2)
```

Decision Tree:

```
#training and test data with marj and youth experiences
set.seed(1)
train <- sample(1:nrow(df.ma2), 1000)
ma.train <- df.ma2[train,]
ma.test <- df.ma2[-train,]
tree.ma2 <- tree(mrjydays ~ ., ma.train)
summary(tree.ma2)
```

Plotting tree:

```
plot(tree.ma2)
text(tree.ma2, cex = 0.6,pretty = 0)
```

Accuracy and Matrix table:

```
set.seed(1)
test <- df.ma2$mrjydays[-train]
tree.pred <- predict(tree.ma2, ma.test,
   type = "class")
table <- table(tree.pred, test)
table
mean(tree.pred == test)
accuracy_Test <- sum(diag(table)) / sum(table)
print(paste('Accuracy for test', accuracy_Test))
```

Pruned Tree:

```
set.seed(123)
tree.ma2 <- tree(mrjydays ~ ., ma.train)
cv.ma <- cv.tree(tree.ma2, FUN = prune.misclass)
optimal_K <- which.min(cv.ma$dev)
pruned.ma <- prune.misclass(tree.ma2, best = optimal_K)
```

Plotting the pruned tree:

```
plot(pruned.ma)
text(pruned.ma,cex = 0.4, pretty = 0)
```

```
summary(pruned.ma)
```

Accuracy and Matrix table:

```
set.seed(1)
test <- df.ma2$mrjydays[-train]
tree.pred <- predict(pruned.ma, ma.test,
   type = "class")
table <- table(tree.pred, test)
table
mean(tree.pred == test)
accuracy_Test <- sum(diag(table)) / sum(table)
print(paste('Accuracy for test', accuracy_Test))
```

Bagging:

```
set.seed(1)
#train <- sample(1:nrow(df.ma), 1000)
train <- sample(1:nrow(df.ma2), nrow(df.ma2) / 2)
test <- df.ma2[-train,"mrjydays"]
bag.ma <- randomForest(mrjydays ~ ., df.ma2, subset = train, mtry = 25, importance = TRUE)
bag.ma
```

MSE:

```
yhat.bag <- predict(bag.ma, newdata = df.ma2[-train, ])
yhat.bag.num <- as.numeric(yhat.bag)
test.num <- as.numeric(test)
# Check for any non-numeric values in yhat.bag.num or test.num
any(!is.na(yhat.bag.num) & !is.numeric(yhat.bag.num))
any(!is.na(test.num) & !is.numeric(test.num))
# Calculate the mean squared error
mean((yhat.bag.num - test.num)^2)
importance(bag.ma)
varImpPlot(bag.ma)
```

Random Forest:

```
set.seed(1)
```

```
rf.ma <- randomForest(mrjydays ~ ., df.ma2, subset = train, mtry = 12, importance = TRUE)
yhat.rf <- predict(rf.ma, newdata = df.ma2[-train, ])
yhat.rf.num <- as.numeric(yhat.rf)
test.num2 <- as.numeric(test)
# Check for any non-numeric values in yhat.bag.num or test.num
any(!is.na(yhat.rf.num) & !is.numeric(yhat.rf.num))
any(!is.na(test.num2) & !is.numeric(test.num2))
# Calculate the mean squared error
mean((yhat.rf.num - test.num2)^2)
importance(rf.ma)
```

Plot variance importance plot:

```
var_imp <- varImpPlot(rf.ma, main = "Variable Importance Plot for Marijuana use",cex = 0.7)
```

Regression Problem:

Pulling quantitative variable to factor with 3 levels:

```
iralcfy<- factor(ifelse(df$iralcfy == 991, "Never Used Alcohol",
                        ifelse(df$iralcfy == 993, "Did Not Use Alcohol Past Year",
                               "Used Alcohol Past Year")),
                        levels = c("Never Used Alcohol", "Used Alcohol Past Year", "Did Not Use
Alcohol Past Year"))
# setting it up for a regression analysis 0,1,2
iralcfy <- ifelse(iralcfy == "Never Used Alcohol", 0,
                  ifelse(iralcfy == "Used Alcohol Past Year", 1,
                         ifelse(iralcfy == "Did Not Use Alcohol Past Year", 2, NA)))
```

Combining alcohol and demographics into a data set:

```
df.al <- cbind(iralcfy,df_demo_clean)
#removing any Nas from data set
df.al <- na.omit(df.al)
```

Decision Tree:

```
set.seed(1)
train <- sample(1:nrow(df.al), 1000)
#train <- sample(1:nrow(df.al), nrow(df.al) / 2)
```

```
tree.al <- tree(iralcfy ~ ., df.al, subset = train)
summary(tree.al)
tree.al
```

Plotting tree:

```
plot(tree.al)
text(tree.al,cex = 0.6, pretty = 0)
```

Cross-validation and Pruning:

```
set.seed(7)
tree.al <- tree(iralcfy ~ ., df.al, subset = train)
cv.al <- cv.tree(tree.al,K=5,FUN = prune.tree)
optimal_K <- which.min(cv.ma$dev)
pruned.ma <- prune.tree(tree.al, best = optimal_K)
```

Plotting the pruned tree:

```
plot(pruned.ma)
text(pruned.ma,cex = 0.6, pretty = 0)
summary(pruned.ma)
```

MSE:

```
yhat <- predict(tree.al, newdata = df.al[-train, ])
al.test <- df.al[-train, "iralcfy"]
plot(yhat, al.test)
abline(0, 1)
mean((yhat - al.test)^2)
sqrt(.2970049)
```

Boosting:

```
set.seed(1)
boost.al <- gbm(iralcfy ~ .-eduschlgo, data = df.al[train, ],
   distribution = "gaussian", n.trees = 2000,
   interaction.depth = 4)
summary(boost.al)
```

MSE:

```
yhat.boost <- predict(boost.al,
```

```
    newdata = df.al[-train, ], n.trees = 2000)
mean((yhat.boost - al.test)^2)
```