REPORT: EVALUATION OF TEEN DRUG CONSUMPTION

ABSTRACT

Using decision trees and ensemble approaches, this research investigates the factors linked to teen drug use. The National Survey on Drug Use and Health, which collects thorough data on respondents' demographics, experiences as kids, drug use, and other relevant variables, provided the data utilized in the paper. With one example of each, the report discusses three different problem categories: binary classification, multi-class classification, and regression. This report contains a thorough examination of the one tree model's flow, how predictions change with different data kinds, the factors that frequently serve as reliable predictors of drug use, and the ethical issues raised by these conclusions.

INTRODUCTION

The study of the statistics on the youth substance dataset is presented in this report. The major goal of this investigation is to identify the ages at which young people become hooked to various substances, including marijuana, alcohol, and cigarettes. And what are factors causing such substance being used by the youth.

In the United States, juvenile drug use is a serious public health issue. This paper employs decision trees and ensemble approaches to analyze data from the National Survey on Drug usage and Health in order to determine the factors that affect juvenile drug usage. The dataset contains a range of demographic and behavioral factors, such as the frequency of alcohol, marijuana, and cigarette use, the age at which first use occurred, parental participation, drug education, and more. The three sorts of issues that are examined in this study are regression, multi-class classification, and binary classification.

This study examines binary classification, multi-class classification, and regression to assess characteristics associated with teen drug use using decision trees and ensemble techniques on data from the National Survey on Drug Use and Health. The goal of the study is to determine the ages at which young people become addicted to various substances, such as marijuana, alcohol, and cigarettes, as well as the variables that lead to young individuals using such substances. The one-tree model's flow, how predictions alter with different data types, the variables that frequently serve as accurate predictors of drug use, and ethical concerns are all included in the paper.

THEORETICAL BACKGROUND

In the field of machine learning, decision trees are frequently employed as a technique for creating prediction models. The trees are made up of nodes, which stand in for variables or features, and branches, which show several possible values for that variable. The goal is to partition the data at each node in such a way that the homogeneity of the generated groups is maximized.When attempting to predict the class of a given observation based on a set of input data, decision trees are very helpful.

A common strategy for improving the accuracy and stability of decision tree models is ensemble methods. Bagging and boosting are the two ensemble techniques that are most frequently utilized. Using different subsets of the training data, several decision trees are built to perform bagging. The predictions from each tree are then averaged. Contrarily, boosting produces a series of trees in which each succeeding tree makes an effort to fix the mistakes committed by the prior branch by looking at decreasing the error(in the current tree) to the lateral(next tree). In situations where individual trees may not perform well owing to noise or other issues, ensemble approaches can produce more reliable and accurate models by aggregating the predictions of numerous trees(this can be done using 'ntrees' parameter).

METHODOLOGY:

An assessment of teen drug use Before anything else the data was cleaned by removing replies from children and young adults under the age of 18 and choosing particular characteristics pertaining to substance use, demographics, and adolescent experiences. Variable labels were cleaned for better reference, and categorical data were transformed into factors. Column names for youth experiences, demographics, and substance use were designed for three vectors.

The work examines whether or not a person has smoked cigarettes in the binary categorization problem. Given that smoking is one of the major causes of avoidable deaths in the globe, this is a significant concern. A decision tree was constructed to address this issue using data from the National Survey on Drug Use and Health, which included details on various demographics, juvenile experiences, and drug use data from the National Survey on Drug Use and Health, which included details on various demographics, juvenile experiences, and drug use, a decision tree was constructed to address this issue.

It was difficult to discriminate between rare, occasional, and frequent alcohol intake for the multi-class classification task. As it may significantly affect a person's health, social life, and future chances, this is a crucial issue. To improve the model's accuracy and stability, ensemble technique boosting is used.

COMPUTATIONAL RESULTS:

Let's check at what early ages people were addicted to several drugs:

Cigarette:

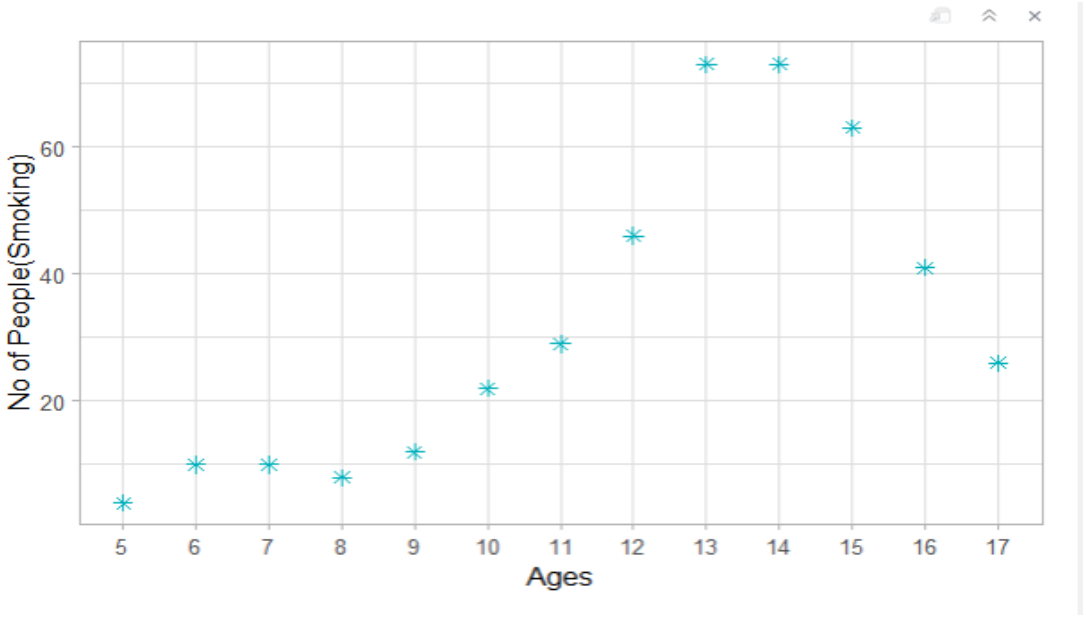

Fig 5.1: Early ages of Cigarette Smoking

Alcohol:

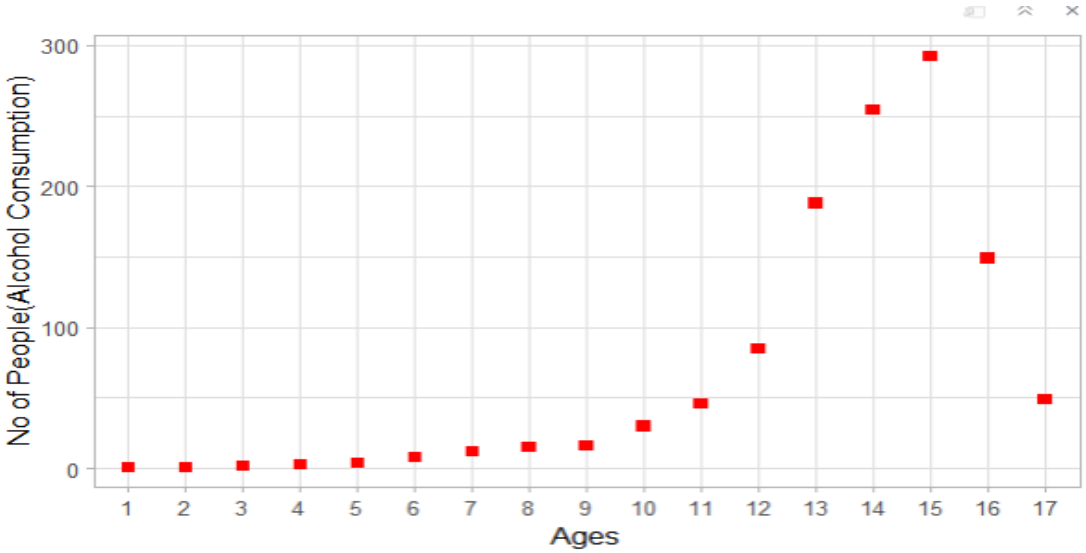

Fig 5.2: Early ages of Alcohol consumption

In my data analysis and above plots I express , people who were first exposed to smoking were usually between the ages of 5 and 17. Ages 13 and 14 in particular are the most vulnerable to developing an addiction. Similar to this, persons who start drinking alcohol early are more likely to be between the ages of 13 and 15. It's interesting to note that marijuana addiction affects people of all ages, beginning as early

as age 10 and lasting until age 17, when addiction rates tend to peak. In the end, the evidence shows that adolescents are more prone than adults and elderly citizens to develop early addictions to alcohol, cigarettes, tobacco, and marijuana.

Let's check the consumption of Alcohol and Marijuana based in gender and then go for decision trees for the factors influencing youth to move towards these drugs.

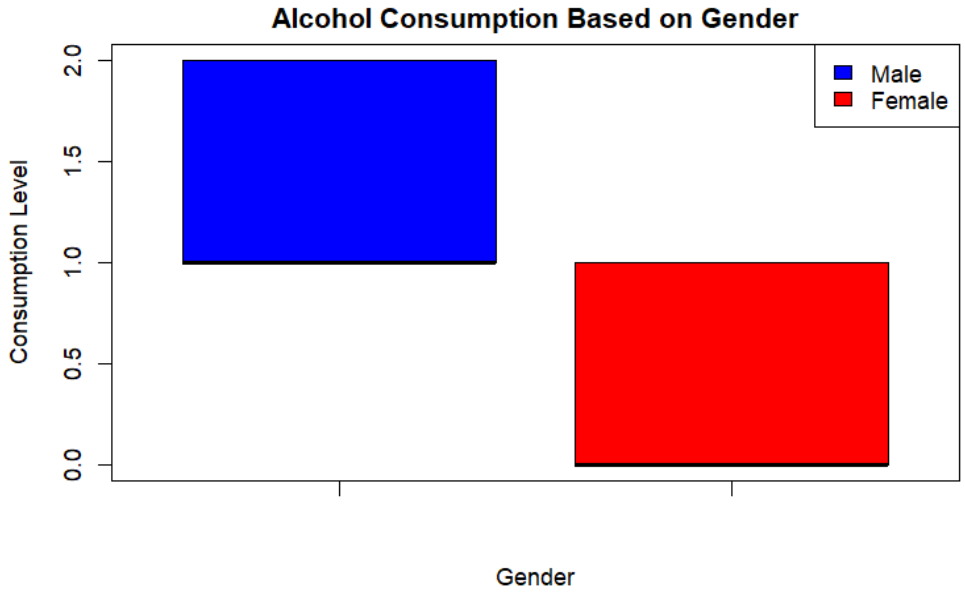

Fig 5.3: Gender based alcohol consumption

The aforementioned plot makes it very evident that marijuana use is mostly carried out by men, whilst female use is quite insignificant. Unexpectedly, alcohol consumption rates across the sexes are essentially similar, even though they do differ significantly, with female consumption being lower and male consumption being higher

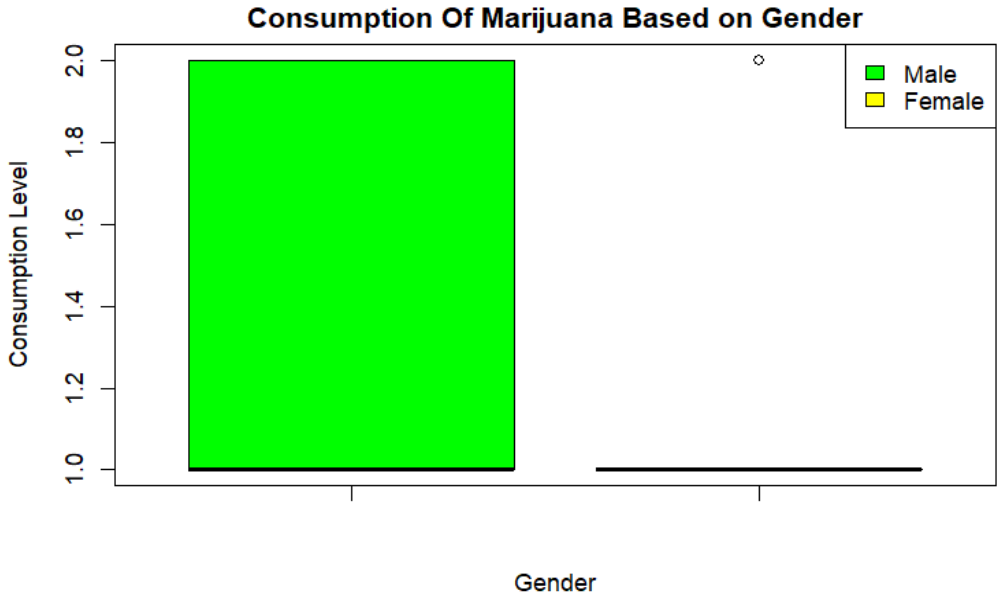

Fig 5.4: Gender based Marijuana Consumption

. We are unable to definitively confirm the first claim, though, because the plot is a box plot. According to the statistics, information may have been gathered from men more frequently than from women, which may explain some of the observed gender-based disparities in marijuana use.

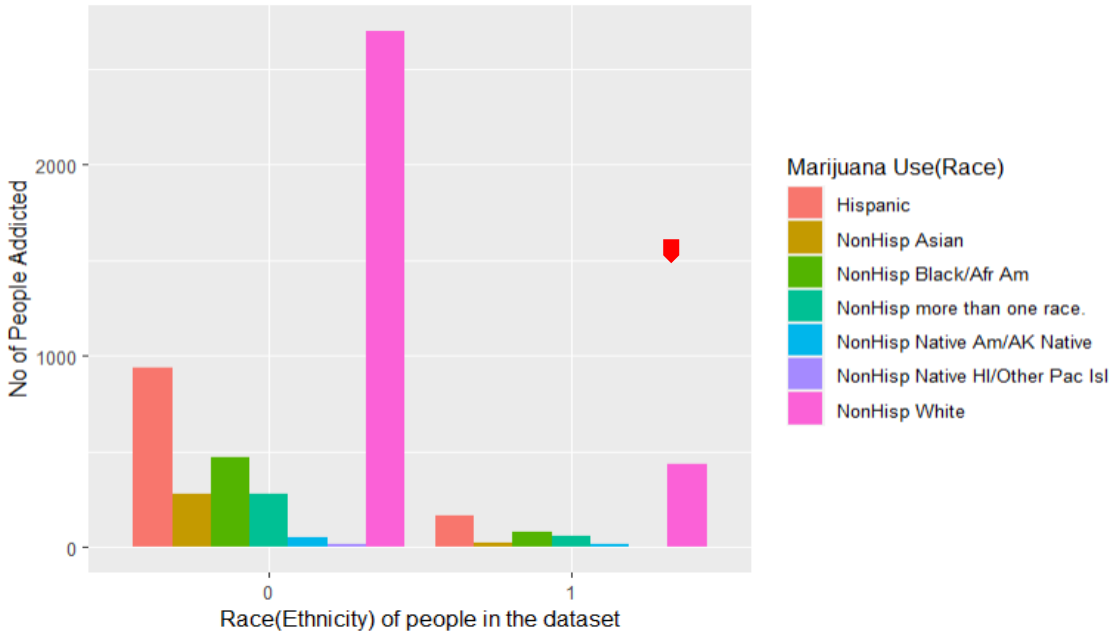

Fig 5.5: Marijuana Consumption based on Ethnic Categories

Data from seven categories or types are displayed on the graph, with type 1 having the most data and type 4 having the least. It also illustrates the connection between marijuana use and race, with the unexpected finding that marijuana use is most prevalent among the group with the least amount of data.

BINARY CLASSIFICATION (DECISION TREE):

Our Response Variable: Alcohol used (ever/never) :

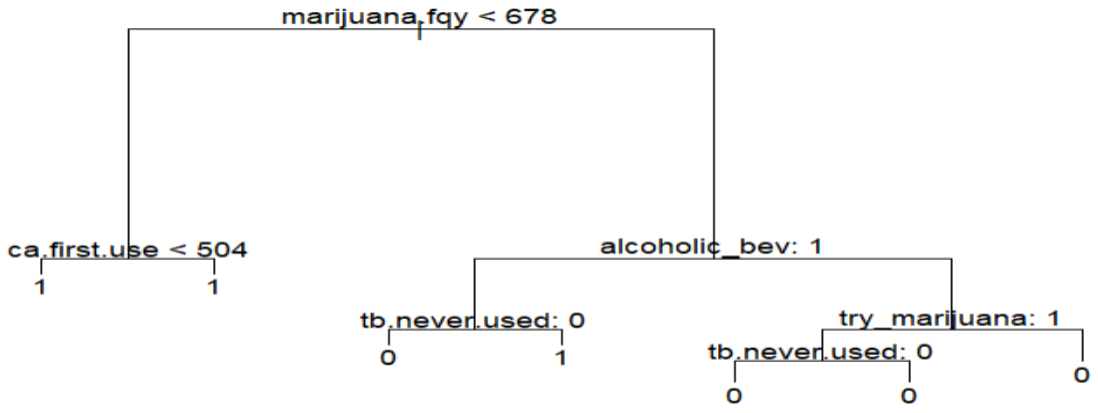

Fig 6.1: Decision Tree for predicting Alcohol Intake

The model predicts that a person will not have used tobacco in their lifetime with a high degree of confidence (yval=0.84310) if they have used marijuana for fewer than 5.5 days. Depending on whether a person has never used tobacco or not, the tree is divided if they have used marijuana for more than 5.5 days. The model predicts, with low confidence (yval=0.18710), that a person who has never smoked tobacco (node 6) has not smoked much marijuana in the last year. The model predicts with high confidence (yval=0.62620) that they have used marijuana more in the last year if they have used cigarettes in the past.

Depending on whether the person has used marijuana before or not, further divides are further established between remaining nodes. we can also tell that people who have used marijuana before and have also consumed alcoholic drinks are more likely than those who haven't to have used marijuana in the previous year. On the other hand, it demonstrates that people who have never used marijuana are more likely than those who have to have used it in the previous year.

Our Response Variable: Marijuana used (ever/never) :

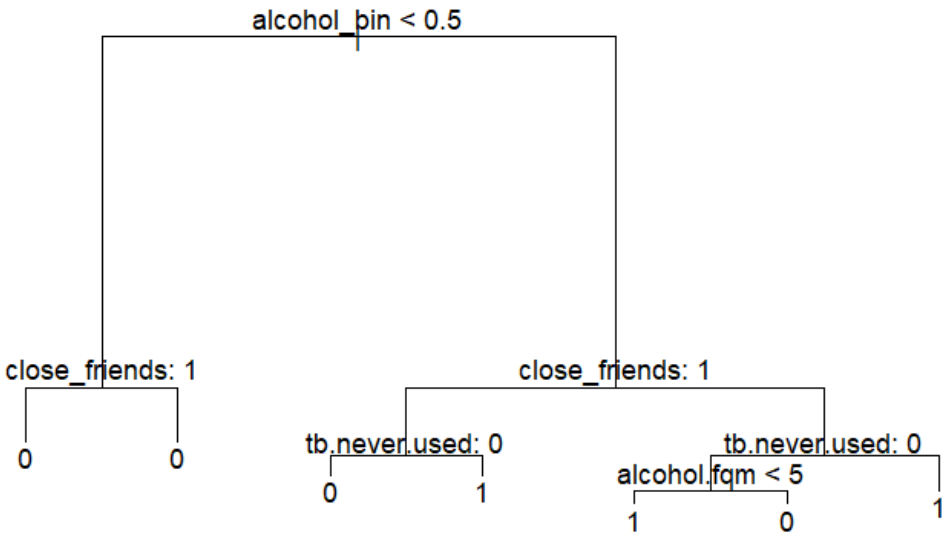

Fig 6.2: Decision Tree for finding out Marijuana Intake

This decision tree model makes predictions about marijuana use based on a person's behavior with tobacco and alcohol. The tree has a number of nodes, each of which represents a split on a different variable. The variable "alcohol_bin," which denotes whether a person has tried alcohol or not, is where the first split occurs. The model suggests that individuals are less likely to use marijuana if they haven't. If not, the model further divides based on the proportion of close friends that support marijuana use. The

model predicts that a person is unlikely to use marijuana if they don't have any of these friends. Otherwise, the model takes into account their behavior in terms of tobacco usage and how often they drink alcohol while making its prediction.

In conclusion, this decision tree makes predictions about marijuana use based on a few behavioral indicators. The model takes into account a person's intake of alcohol, the proportion of their close friends who support marijuana use, as well as their behavior regarding tobacco use and alcohol consumption. Based on the combination of these behavioral elements, the tree splits based on these variables and eventually predicts whether a person is likely to use marijuana or not.

Lets use ensemble methods to find the a boosted regression tree for the above fitted decision tree.

Missing values are eliminated in this case, and the dataset is divided into training and testing data. Then, in order to forecast alcohol intake, a boosted tree model is constructed using the gbm package with a Bernoulli distribution and 1000 trees. To assess the model's accuracy, the mean squared error (MSE) is determined. The average squared difference between the anticipated and actual values of alcohol intake in the test data is shown by the MSE value of 10.47.

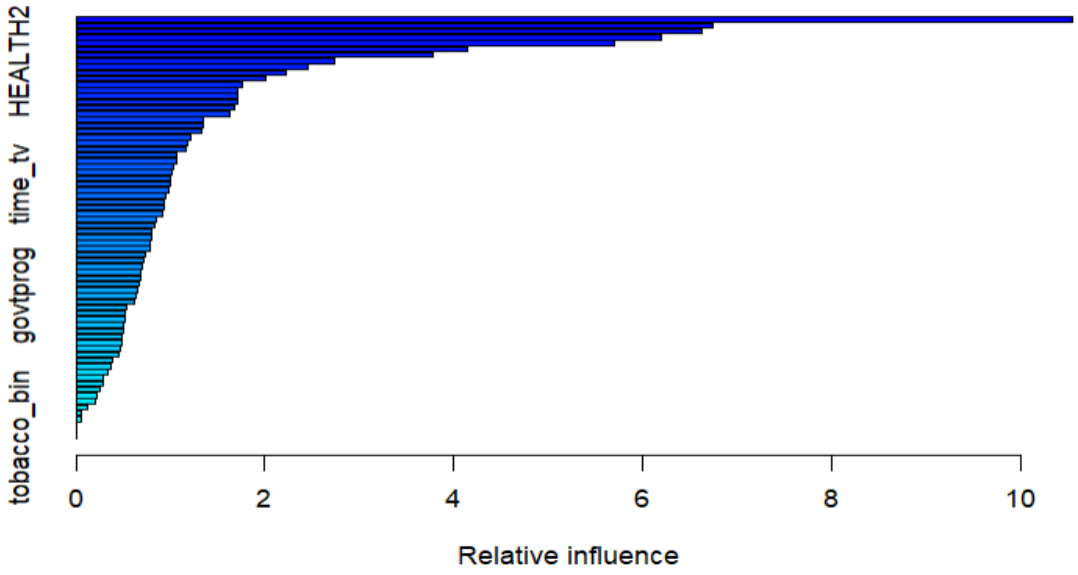

Fig 6.3: Factors affecting the alcohol consumption among people.

According to the enhanced tree above, parents who limit television viewing, a person's race, and the child's grade all have an impact on how much alcohol a person consumes.

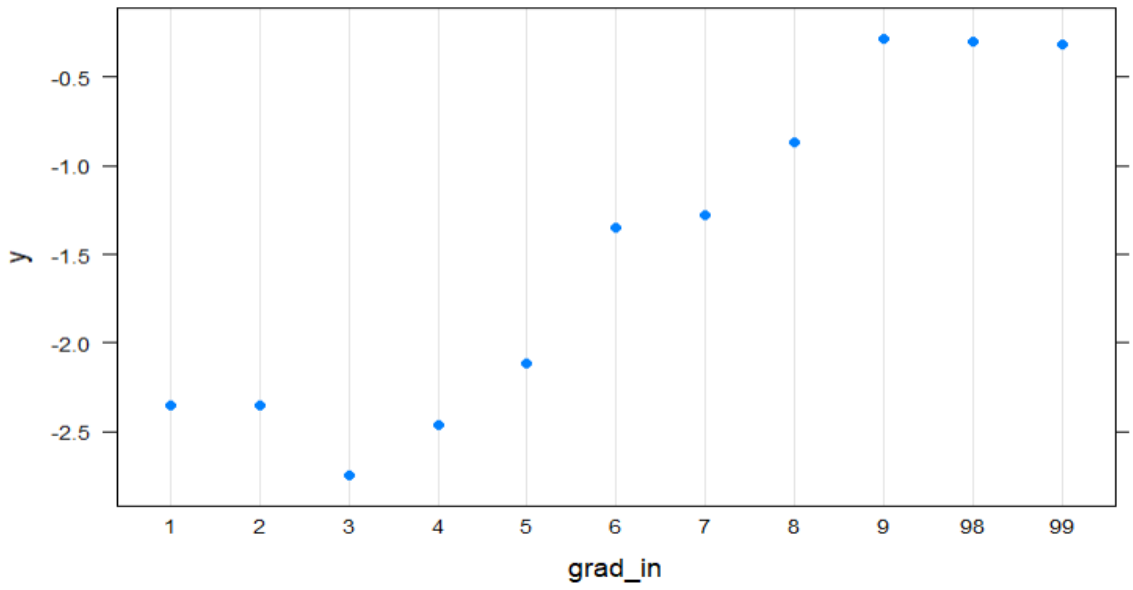

Fig 6.4: Children's Grade as one of most affecting factors.

As the child grade keeps on moving to higher grades, his risk for influence of alcohol consumption will be high.

MULTI-CLASS CLASSIFICATION (DECISION TREE):

Here, our un-pruned tree has lot many branches and its hard to interpret the information and so we will prune the tree and then interpret the model. And the pruned tree is:

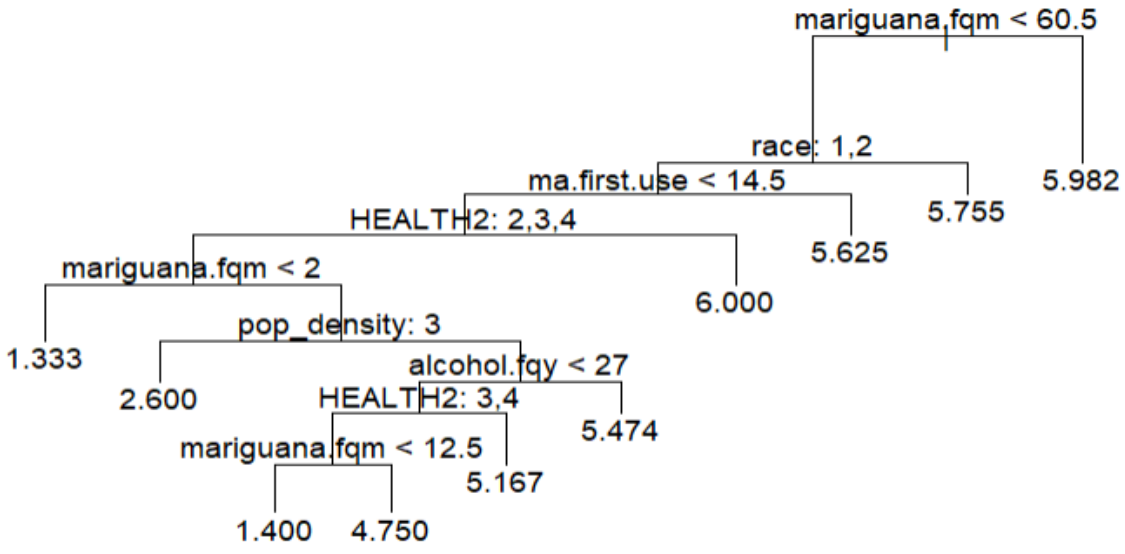

Fig 7.1: Pruned Decision Tree for precious tree (Fig 7.1)

The input variables with the best predictive power are used as the basis for the divides. Since it is located near the top of the tree and has the greatest deviance reduction, the split on "mariguana.fqm," for instance, has the best predictive power.

The summary does not specifically state what percentage each component has on the response variable. However, by examining the splits in the tree, we can determine which input factors are the most significant predictors. The most frequent divide is on "marijuana.fqm," followed by "race," "health2," "alcohol.fqy," "pop_density," "ma.first.use," "time_tv," "par_disapprove," "tb.never.used," and "gender." This shows that "mariguana.fqm," "race," and "health2" are the three most crucial predictors.

REGRESSION:

Models that predict continuous numerical outcomes can be created using regression trees, a sort of decision tree technique. Regression trees can be used to gauge the weight or significance of each predictor variable in your model. This gauges how much the variable aids in forecasting the result. This enables us to comprehend which predictors are most crucial in predicting the outcome variable, and we may utilize this knowledge to ascertain how many smokes you will consume. Regression trees can be used to create models that successfully identify key predictors of an event and properly predict result variables.

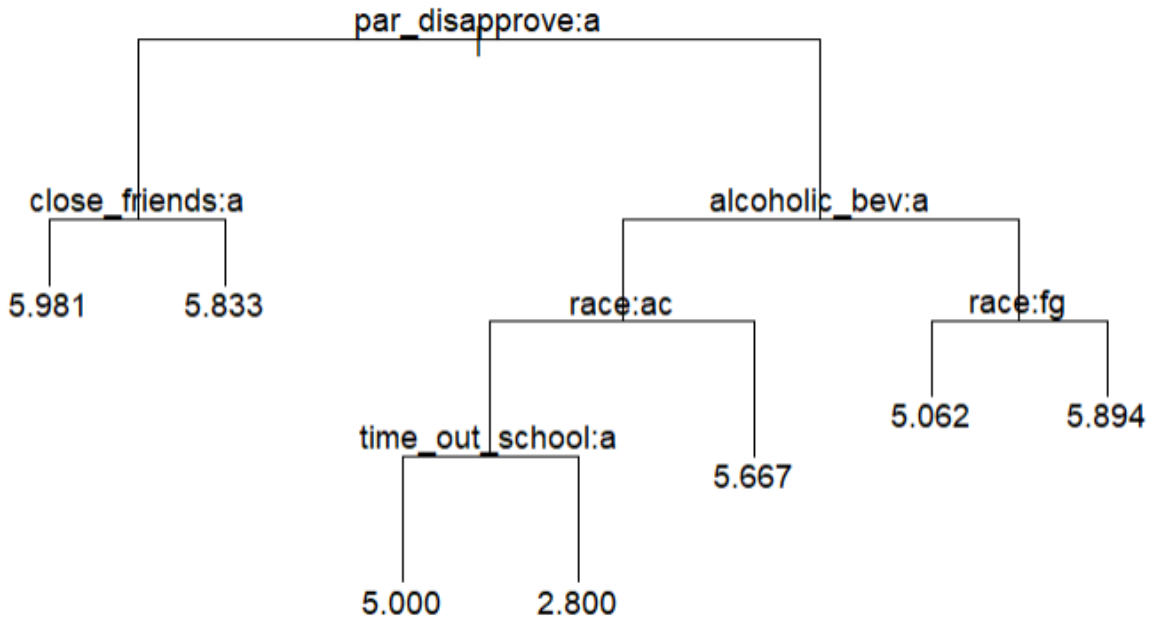

Fig 8.1: Cigarette Usage per month; Regression tree;

The decision tree model for predicting the likelihood of teen substance use based on numerous demographic and behavioral factors is summarized in the tree. The split variables in the tree, which has six layers and fifteen nodes, include parental disapproval of substance use, the frequency of alcohol intake, race, absence from school, and approval of close friends' use of drugs. According to the model, the most significant predictor of youth substance use is the substance use of close friends, with those who disapprove of close friends' use being less likely to use themselves. The model also suggests that parental figures' disapproval plays a significant role in lowering the risk of young people using drugs.

Overall, the decision tree suggests that demographic factors such as race and time spent out of school are less predictive of youth substance use than behavioral factors such as frequency of alcohol consumption and the approval of close friends' substance use. The model suggests that targeted interventions focused on changing peer influence and increasing parental involvement could be effective in reducing youth substance use. However, this model is not the best, it should be validated on additional data(I mean more data can be collected, without any bias) to ensure its generalizability and reliability.

Our Test M.S.E of the regression tree model is found to be 0.37 which is really good pertaining to our data as we don't have numerous data to really judge this situation by saying that our model is best but factors what we chose in our model can be proclaimed as they are best predictors in determining the no of cigarettes used by a person in any given month. Hence, these are the factors strongly influencing our response variable,

CONCLUSION:

Using decision trees and ensemble techniques, this study investigates the aspects related to teen drug use. The National Survey on Drug Use and Health, which gathers detailed information on respondents' demographics, experiences as children, drug use, and other pertinent variables, provided the data used in this research. The three problem types covered in the report are binary classification, multi-class classification, and regression. In the field of machine learning, decision trees are frequently used to build prediction models, and ensemble techniques like bagging and boosting help decision tree models be more accurate and stable. The binary categorization test explores whether a person has ever smoked cigarettes, whereas the multi-class classification job deals with the topic of infrequent, sporadic, and regular alcohol consumption.

It was found that most smokers started off while they were between the ages of 5 and 17, with ages 13 and 14 being the most susceptible to being addicted. In a similar vein, those who begin consuming alcohol young are more likely to be between the ages of 13 and 15. The study demonstrates that marijuana addiction affects individuals of all ages, starting as early as age 10 and continuing until age 17 when addiction rates often peak.

Teenagers are more likely than adults and senior persons to become addicted to alcohol, cigarettes, nicotine, and marijuana at a young age.

And Men use marijuana more frequently than women, who use it much less frequently. However, despite the fact that they do differ noticeably, with female consumption being lower and male consumption being larger, alcohol consumption rates between the sexes are essentially identical. The study provides a decision tree for the binary classification problem that accurately predicts a person's alcohol consumption based on their age, gender, and marijuana use. Overall, the research offers insightful information about the factors contributing to teen drug use that may be used to create successful preventive and intervention plans.

Discussion:

In this report, decision trees and ensemble techniques are used to investigate the factors linked to teen drug use. This analysis makes use of information on demographics, drug use, and pertinent variables gathered from the National Survey on Drug Use and Health.

This report covers specific drug-related problems like smoking, drinking, and marijuana addiction while concentrating on three problem categories. It illustrates the variations in drug usage between men and women and shows that adolescents are more likely to develop early addictions. A decision tree for the binary classification problem, which predicts alcohol usage based on variables including age, gender, and marijuana use, is also included in the study.

Overall, the report offers insightful information about teen drug use and the potential of ensemble techniques and decision trees to forecast drug usage, assisting in the creation of successful prevention measures.

References:

- Substance Abuse and Mental Health Services Administration. (2020). National Survey on Drug Use and Health: 2020 codebook [Data file and codebook]. Retrieved from https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001-nid18785

- Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag. Retrieved from https://ggplot2.tidyverse.org

- Gupta, A. (2022, March 21). A comprehensive guide on ggplot2 in R. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2022/03/a-comprehensive-guide-on-ggplot2-in-r

- Education and Employment Accounting, Finance, IT, Digital Marketing, and Business Analytics. (n.d.). package for R trees. R-tree package was retrieved from https://www.educba.com

Appendix(CODE- .Rmd):

````
---
title: "Project"
author: "Gokula"
date: "2023-04-04"
output: html_document
---
```{r}
library(dplyr)
library(tree)
library(ISLR2)
library(ggplot2)
library(randomForest)
library(tidyverse)

```

```{r}
project_df <- df
project_df
```
```{r}

#youth_experience_cols
```

```{r}
#substance_cols
```
```{r}
#demographic_cols
project_df = as.data.frame(project_df)
```

```{r}
project_df <- project_df %>%
  rename(
    ca.first.use = ircigage,
    ta.first.use = irsmklsstry,
    aa.first.use = iralcage,
    ma.first.use = irmjage )
```


```{r}
project_df <- project_df %>%
  rename(
    alcohol.fqy = iralcfy,
    marijuana.fqy = irmjfy,
    cigarette.fqm = ircigfm,
    tobacco.fqm = IRSMKLSS30N,
    alcohol.fqm = iralcfm,
    mariguana.fqm = irmjfm
  )
```

```{r}
project_df <- project_df %>%
  rename(
    mj.never.used = mrjflag,
    al.never.used = alcflag,
    tb.never.used = tobflag
  )
```

```{r}
project_df <- project_df %>%
  rename(
    al.days.fy = alcydays,
    mj.days.fy = mrjydays,
    al.days.month = alcmdays,
    mj.days.month = mrjmdays,
````
````

```
    cg.days.month = cigmdays,
    tb.days.month = smklsmdays
  )
```

```{r}
project_df <- project_df %>%
  rename(
    mother_household = imother,
    father_household = ifather,
    gender = irsex,
    race = NEWRACE2
  )
```

```{r}
project_df <- project_df %>%
  rename(
  school_skip = eduskpcom,
  family_income = income,
  poverty_level = POVERTY3
  )
```

```{r}

project_df <- project_df %>%
  rename(
  grad_in = EDUSCHGRD2,
  youth_activity = YTHACT2,
  par_disapprove = PRPKCIG2,
  pop_density = PDEN10
  )
```


```{r}
project_df <- project_df %>% rename(
  time_out_school = parlmtsn,
  time_tv = PRLMTTV2
)
```

```{r}
project_df['marijuana_bin'] <- df$mrjflag
project_df['alcohol_bin'] <- df$alcflag
project_df['tobacco_bin'] <- df$tobflag
```

```{r}
project_df <- project_df %>% rename(
  alcoholic_bev = stndalc,
  parent_opinion = PRALDLY2,
  try_marijuana = YFLTMRJ2,
  close_friends = FRDMEVR2
)
```

```{r}
project_df
```


#lets check at what early ages people addicted to which drug

```{r}
cigarette_age <- table(project_df$ca.first.use)
cadf <- as.data.frame(cigarette_age)
ggplot(cadf, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=8, color = "#00AFBB")+
  xlab("Ages")+
  ylab("No of People(Smoking)")+
  theme_light()
```

As we see that most of the people never used cigarette and also we know that people who were first exposed to drinking

cigarette the first time are from 5-17 aged people.

Now Lets Have a clearly look into the 5-17 aged and see at which age people are getting addicited to cigarette.

```{r}
cigarette_age <- table(project_df$ca.first.use)
cadf <- as.data.frame(cigarette_age)
cadf <- cadf[(1:nrow(cadf)-1),]
ggplot(cadf, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=8, color = "#00AFBB")+
  xlab("Ages")+
  ylab("No of People(Smoking)")+
  theme_light()
```

Now we can see that people with ages 13 & 14 are more prone to cigarette addiction at intital stages.

```{r}
nrow(project_df)
```

#lets continue the same thing with alcohol and marijuana

```{r}
alcohol_age <- table(project_df$aa.first.use)
cadf <- as.data.frame(alcohol_age)
ggplot(cadf, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=8, color = "#00AFBB")+
  xlab("Ages")+
  ylab("No of People(Alcohol Consumption)")+
  theme_light()
```

```{r}
alcohol_age <- table(project_df$aa.first.use)
cadf <- as.data.frame(alcohol_age)
cadf <- cadf[(1:nrow(cadf)-1),]
ggplot(cadf, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=15, color = "red")+
  xlab("Ages")+
  ylab("No of People(Alcohol Consumption)")+
  theme_light()
```

Here, we see that people get addicted to alcohol consumption at very early stages fall into this age bin [13-15] respectively.

```{r}
marijuana_age <- table(project_df$ma.first.use)
cadf <- as.data.frame(marijuana_age)
ggplot(cadf, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=8, color = "#00AFBB")+
  xlab("Ages")+
  ylab("No of People(Marijuana Consumption)")+
  theme_light()
```

By the above to my surprise i could see people of several ages are addicted to the marijuana lets investigate deeper into it.

```{r}
marijuana_age <- table(project_df$ma.first.use)
cadf <- as.data.frame(marijuana_age)
cadf <- cadf[(1:nrow(cadf)-1),]
ggplot(cadf, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=8, color = "#00AFBB")+
  xlab("Ages")+
  ylab("No of People(Marijuana Consumption)")+
  theme_light()
```

We can see that people are starting to use the marijuana slowly starting from 10 year old to 17 year old this is the period usually people start getting addicted to this drug.

And the same conclusion goes here the people in ages [13-17] are addicted most.

```{r}
tobacco_age <- table(project_df$ta.first.use)
cadf <- as.data.frame(tobacco_age)
ggplot(cadf, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=8, color = "#00AFBB")+
  xlab("Ages")+
  ylab("No of People(Tobacco Consumption)")+
  theme_light()
```

```{r}
tobacco_age <- table(project_df$ta.first.use)
cadf <- as.data.frame(tobacco_age)
cadf <- cadf[(1:nrow(cadf)-1),]
ggplot(cadf, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=8, color = "#00AFBB")+
  xlab("Ages")+
  ylab("No of People(Tobacco Consumption)")+
  theme_light()
```

Here we see that people aged 15 years are mostly affected to tobacco at initial stages.

Conclusion 1:

* Mostly Teenagers are addicted to all four i.e., alcohol, cigarette, tobacco, and marijuana in their early stages.

Lets check the All these Four consumptions last year:

```{r}
alcohol_fqy <- table(project_df$alcohol.fqy)
afqy <- prop.table(alcohol_fqy)
afqy <- data.frame(age=names(afqy), percent=as.numeric(afqy))
ggplot(afqy, aes(x=age, y=percent)) +
  geom_point(size=2, shape=3, color = "#00AFBB")+
  xlab("No of times Consumed")+
  ylab("Percentage of People(alcohol Consumption Past Year)")+
  theme_light()
```

As usual we can see that people who didnt at all consume alcohol is more and people who didnt consume last year is also slightly
higher.

```{r}
fun <- function(x_lower,x_upper){
alcohol_fqy <- table(project_df$alcohol.fqy)
afqy <- prop.table(alcohol_fqy)
afqy <- data.frame(age=names(afqy), percent=as.numeric(afqy))
afqy <- afqy[c(x_lower:x_upper),]
ggplot(afqy, aes(x=age, y=percent)) +
  geom_point(size=2, shape=3, color = "#00AFBB")+
  xlab("No of times consumed")+
  ylab("Percentage of People(alcohol Consumption Past Year)")+
  theme_light()
}
fun(1,25)
fun(26,50)
fun(51,75)
fun(76,100)
```

It seems like mostly the people who doesn't smoke in past year are about almost 77% of the picture and the rest all accumulated will account for 33% and most among the 33% was found to be that, people have consumed alcohol like around 60 times majorly among the no of times people drinking the alcohol.

Now will check the same for marijuana based on the previous month.

```{r}
marijuana_fqm <- table(project_df$mariguana.fqm)
maqm <- as.data.frame(marijuana_fqm)
```

```
maq <- maqm[(1:(nrow(maqm)-2)),]
ggplot(maq, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=1, color = "#00AFBB")+
  xlab("No of times")+
  ylab("Marijuana consumption Past Month")+
  theme_light()
maq2 <- maqm[((nrow(maqm)-1):nrow(maqm)),]
ggplot(maq2, aes(x=Var1, y=Freq)) +
  geom_point(size=2, shape=1, color = "#00AFBB")+
  xlab("No of times")+
  ylab("Marijuana consumption Past Month(No of times)")+
  theme_light()
```

```{r}
names(project_df)
```

```{r}
#Training & testing
set.seed(1)
project_df <- na.omit(project_df)
train <- sample(1:nrow(project_df), 0.60 * nrow(project_df))
project_df.train <- project_df[train,]
project_df.test <- project_df[-train,]
```

Decision Tree

Alcohol

```{r}
set.seed(2)
tree_project_01 <- tree(alcohol_bin ~ . - -marijuana.fqy -yflmjmo -marijuana_bin -mj.never.used -ma.first.use -al.days.month -aa.first.use - al.never.used -alcohol.fqy - al.days.fy - alcohol.fqm, project_df,subset = train,minsize = 10)
```

```{r}
summary(tree_project_01)
```

```{r}
tree_project_01
```

```{r}
plot(tree_project_01)
text(tree_project_01, pretty = 0)
```

Lets use ensemble methods to find the a boosted regression tree for the above fitted decision tree.

```{r}
set.seed(2)
project_df <- na.omit(project_df)
project_df$alcohol_bin <- as.numeric(as.character(project_df$alcohol_bin))
train <- sample(1:nrow(project_df), 0.45 * nrow(project_df))
project_df.train <- project_df[train,]
project_df.test <- project_df[-train,]
```

```{r}
library(gbm)
set.seed(1)
boost.tree.01 <- gbm(alcohol_bin ~ . -marijuana_bin -mj.never.used -ma.first.use -al.days.month -aa.first.use -al.never.used -alcohol.fqy - al.days.fy - alcohol.fqm, data = project_df.train,
   distribution = "bernoulli", n.trees = 1000,
   interaction.depth = 3)
```
```
```

#our formula for regression tree
```{r}
cig.linear$call
```

SOME MORE PLOTS
```{r}
boxplot(project_df$gender,project_df$alcohol_bin,
      main="Alcohol Consumption Based on Gender",
      xlab="Gender",ylab="Consumption Level",col=c("blue","red")
      )
legend("topright", border="black", fill = c("blue","red"), c("Male","Female"))


boxplot(project_df$gender,project_df$marijuana_bin,main="Consumption Of Marijuana Based on Gender",xlab="Gender",ylab="Consumption Level",
      col=c("green","yellow"))
legend("topright", border="black", fill = c("green","yellow"), c("Male","Female"))
```
```{r}
plot(project_df$race,project_df$marijuana_bin)
```


```{r}
plot(project_df$schfelt,project_df$alcohol.fqy)
```

```{r}

```