Group 8 ISDS 577 Report:

Substance Abuse Analytics

**ISDS 577**

Eddie Pack

Ruchika Narang

Mengci Zhang

Raymundo Venegas

Snehal Yadav

Nikhil Narkar

**TABLE OF CONTENTS**

[**I. Introduction**](#_kr66sxdj89ip) **2**

[**II. Objectives**](#_uqhtyyxyxgqo) **3**

[**III. KPIs**](#_jk09knknsleb) **3**

[**IV. Tools**](#_3ktz159uhj1n) **4**

[**V. Data Transformation**](#_njvx2njusrce) **4**

[**VI. Accuracy**](#_rditohgzn7ek) **6**

[**VI. Data Analysis**](#_px91hzp93xxq) **7**

[1. How does the D.A.R.E program help kids interpret the risk of smoking?](#_yyqjjbxkhb7u) 7

[2. How does drug use influence adolescents’ school performance?](#_grqe9a9kqwma) 13

[3. What is the likelihood of alcohol usage based on adolescent’s emotional state (depressed vs. happy) and the presence of a parent?](#_2zys9iv3irvk) 25

[**CONCLUSION AND RECOMMENDATIONS**](#_986rowrqup7b) **33**

[**VIII. Closing Statement**](#_fsaf5a5kiawc) **34**

# I. Introduction

There is a common saying, “Prevention is better than the cure.” This is especially true when it comes to drug and alcohol use which has steadily been rising over the last few years, particularly among adolescents. In response, there have been many requests for research to identify the causes and variables that dictate drug and alcohol usage. The hope is that by identifying them, the government can come up with a plan of attack to prevent the causes. Our organization, Group 8 of ISDS 577, has been tasked by the California State University of Fullerton to tackle this crucial research topic. Because our group primarily specializes in analytics, we have decided to focus our efforts in executing an analytics project which will identify the key performance indicators and associations within the datasets.

# II. Objectives

To help us get started, the university gave us a dataset from the ICPSR (Inter-University Consortium for Political and Social Research). These datasets focus on typical behaviors of 8th and 10th graders from 2012-2017 and contains over 600 variables. Our plan is to probe and analyze these datasets to try and pinpoint the causes and variables that are causing adolescents to turn to drugs, smoke, and alcohol. Our primary objective is to use the findings of our research to aid government agencies in coming up with more efficient methods of preventing drug and alcohol use amongst adolescents.

# III. KPIs

The key performance indicators will be a crucial factor in allowing us to find the causes and variables. Our first order of business was to come up with a scope. After some initial reduction of the datasets, we came up with the following KPIs to observe:

***1. How does the D.A.R.E program help kids interpret the risk of smoking?***

The D.A.R.E program is one way to help educate kids about the effects and risks of drug use, thereby preventing potential smoking. Adolescents in general are curious in nature and therefore, may not see the harm in smoking just once. The program seeks to paint a clear picture of what smoking can do and highly discourages experimentation and recreational use. Therefore, we feel that answering this question can help us see how drug education can play a crucial role in preventing smoking in adolescence.

***2. How does drug use influence adolescents’ school performance?***

It is no secret that drug use affects all aspects of the user’s life. School performance is one such factor that can be adversely affected. The student may have harder time focusing in school or forget to turn in their homework. By answering this questions, we can see how drug use correlates and contributes to poor school performance.

**3. What is the likelihood of alcohol usage based on adolescent’s emotional state (depressed vs. happy) and the presence of a parent?**

Alcohol use can happen for a variety of reasons. It can be consumed for recreational purposes or as an escape from reality from depression or stress. We want to gauge the emotional state of adolescents that consume alcohol to ascertain if these states are driving them further to consume. In addition, we wanted to reach even further and see the factor that parenting plays in those emotional states.

# IV. Tools

In order to accomplish our tasks, we relied on various software. For data preprocessing, we used Python and SPSS. To sum up, Python and SPSS helped us clean up and organize variables as well as help us verify and test our predictions through building statistical models. Once we had the pertinent data, we created a visualization of this data with the usage of SPSS, which had a built-in visualization tool that could generate geographic tools/charts. These visualizations helped bring our data to life and inform the decision makers to make the proper recommendations.

# V. Data Transformation

Once our team was approved to conduct the project, we began executing our project plan. To sum up, our plan consisted of the following tasks:

· Data reduction, eliminate unneeded variables

· Remove missing data (null values)

· Aggregate all data into a database/CSV file

Because the dataset was fragmented, we needed a thorough cleansing of the data in order to be able to conduct our research. By transforming the data to our liking, we were able to streamline the process of probing through irrelevant data and achieve our desired results.

Our first attempt to work and analyze the data is to reduce the number of dimension and variables. For this purpose, one of our group members used Python and ran a conditional script that deleted variables that had 50 percent or more missing/null values. By doing this, we were able to narrow down our variables to approximately 156. Second, another analytics staff member used SPSS for implementing dimension reduction. This was done by merging year after year of data and eliminating any variables that were introduced in subsequent later years. As a result, now our data had only the variables which were present for all 5 years. The group then went through each individual variable, named the variable, and voted on which relevant variables to keep for our target questions. By consolidating and removing variables introduced later in the survey, we can eliminate the following variables. In total, we excluded 105 variables from the data set.

Most of the variable data sets are of the nominal type,

|  |  |  |  |
| --- | --- | --- | --- |
| V7588 | V7611 | V7645 | V7616 |
| V7589 | V7612 | V7646 | V7617 |
| V7590 | V7613 | V7647 | V7588 |
| V7591 | V7614 | V7648 | V7619 |
| V7592 | V7615 | V7649 | V7620 |
| V7593 | V7616 | V7650 | V7621 |
| V7610 | V7617 | V7651 | V7622 |
| V7612 | V7588 | V7652 | V7623 |
| V7613 | V7589 | V7653 | V7624 |
| V7615 | V7590 | V7654 | V7625 |
| V7616 | V7591 | V7655 | V7626 |
| V7617 | V7592 | V7656 | V7627 |
| V7618 | V7593 | V7657 | V7628 |
| V7637 | V7610 | V7658 | V7629 |
| V7638 | V7611 | V7659 | V7630 |
| V7640 | V7612 | V7661 | V7631 |
| V7642 | V7613 | V7662 | V7632 |
| V7643 | V7614 | V7663 | V7633 |
| V7644 | V7615 | V7664 | V7618 |
| V7669 | V7593 | V7665 | V7635 |
| V7670 | V7610 | V7666 | V7636 |
| V7671 | V7611 | V7667 | V7637 |
| V7672 | V7612 | V7668 | V7638 |
| V7673 | V7613 | V7591 | V7616 |
| V7589 | V7614 | V7592 | V7617 |
| V7590 | V7615 | V7593 | V7634 |
|  |  | V7610 |  |

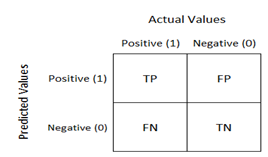
On the other hand, the data transformation forced us to look back and refine our initial target questions. For instance, one of our initial questions was to analyze the influence adolescent get from peers on drug trend. Unfortunately, we could not find enough data in our variable dataset to help us do a conclusive research. Ultimately, the group worked in a reverse manner, where we used the variables that survived SPSS and Python cleanse to come up with the final target questions. By doing this, we ensured that we had the realistic target questions with relevant variables that had enough data to help us make our assertions.

# VI. Accuracy

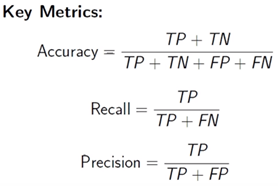
Accuracy is one metric for evaluating classification models. In other words, accuracy is the fraction of predictions our model got right.

For multinomial logistic regression model, we can determine the accuracy from the confusion matrix. Along with accuracy, confusion matrix provides different key matrices such as precision of the model and sensitivity of the model.

**Confusion Matrix**



The confusion matrix for a multi-class classification problem can help you determine mistake patterns. (“Simple guide to confusion matrix terminology”. 2014)



**1.Accuracy**

Accuracy is the fraction of predictions our model got right.

**2. Recall or Sensitivity**

Recall is nothing but sensitivity of the model. The proportion of observed positives that were predicted to be positive.

**3. Precision**

Precision provides the proportion of correct positive classifications from that are predicted as positive.

**Misclassification error:**

Misclassification of categorical variables creates problems for analysis and interpretation, leading to biased estimates if the misclassification is ignored. When all classes, groups, or categories of a variable have the same error rate or probability of being misclassified then it is said to be misclassification.

Misclassification rate tells you how often the model is wrong.

# VI. Data Analysis

Now that we had our target questions and variables, we had to determine the type of analysis we wanted to conduct. First, we determined that we could not perform linear regression as the variables and questions did fit this style of modeling. Also, we found that our variables were in general classified and coded, which means the answers within the variables had multiple layers that linear regression could not make use of.

We will instead use multinomial logistic regression because our variables are based on categories and are multi-faceted in nature. In other words, these variables are not a matter of yes or no answers which would be more suited for linear regressions. To show a preview of how we could display these variables in a visual format, a staff member used SPSS to display a graph/tree of the variable, and conduct a sample analysis based on those findings.

Now, we will describe the complete analysis approach that we followed based on the KPIs which we mentioned earlier.

## **1. How does the D.A.R.E program help kids interpret the risk of smoking?**

**INTRODUCTION**

For our project, we conduct our analysis on the “Monitoring the Future: A Continuing Study of American youth (8th- 10th-grade Survey), an ongoing study of the behaviors, attitudes, and value of American secondary school students.

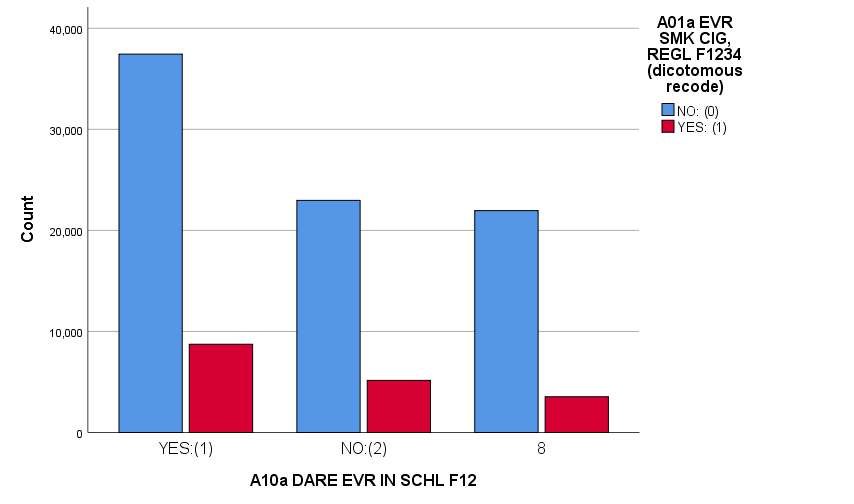
In our analysis, we first took a look at the relationship between a student has had drug education through the [DARE](https://dare.org/) program at their school and its relation to whether a student has smoked in their lifetime. More specifically, we want to answer the question of whether if a student who has taken DARE education at their school is more (or less) likely to smoke. Conversely, we will also look to see if whether if a student who has NOT taken DARE education at their school is more (or less likely to have smoked). We will focus on two variables for this analysis:

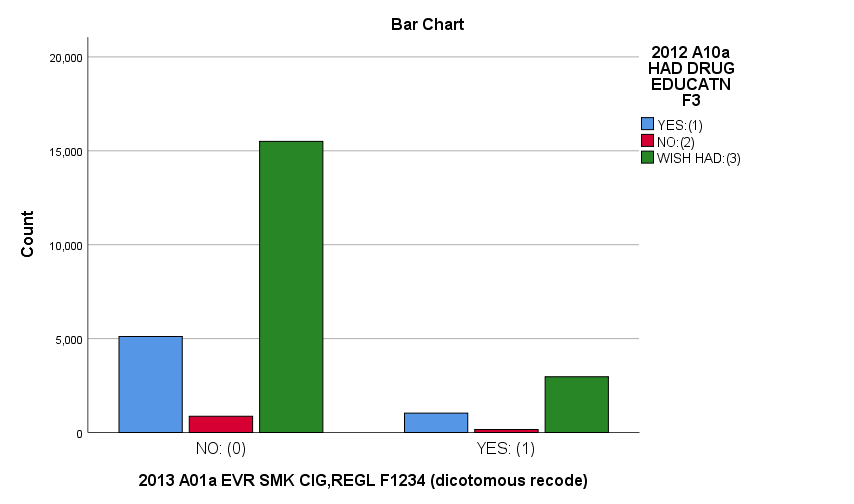
* V7482 (DARE EVR in SCHL)
* V7101D (EVR SMK, CIG,REGL-Dichotomous recode)

**METHOD**

For our analysis, we have taken in data from 2012 through 2012 and selected the dichotomous version of whether a student has smoked. In addition, since the variables we are looking at are not nominal (not continuous) we decide to analyze our data via a) cross tabulations, b) logistic regression and c) tree analysis. We conduct the customary analysis via graphs to help us understand our data better.

The DARE (Drug Abuse Resistance Education) is substance abuse prevention education program that is taught to students starting in the 4th grade to help them make better choices regarding drugs, smoking, alcohol, and other self-harming items. The program was started in 1983 and is implemented to about 75% of the nation’s school districts and around the world. The DARE program is taught nationwide to approximately 26 Million US kids and is taught in classrooms by police offers.



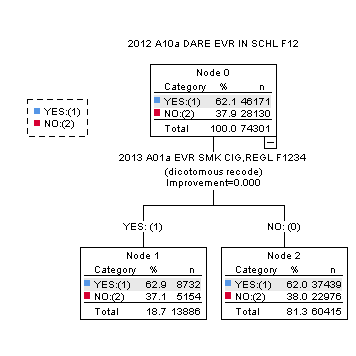


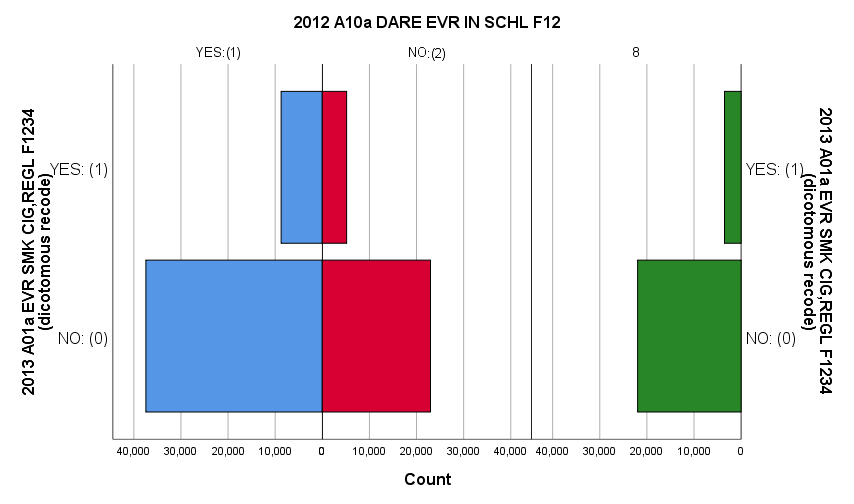
**DESCRIPTIVE RESULTS**

We first analyze our data by generating a simple graph to visualize our buckets of student smokers and if they’ve taken DARE in school. At first glance, we immediately observe:

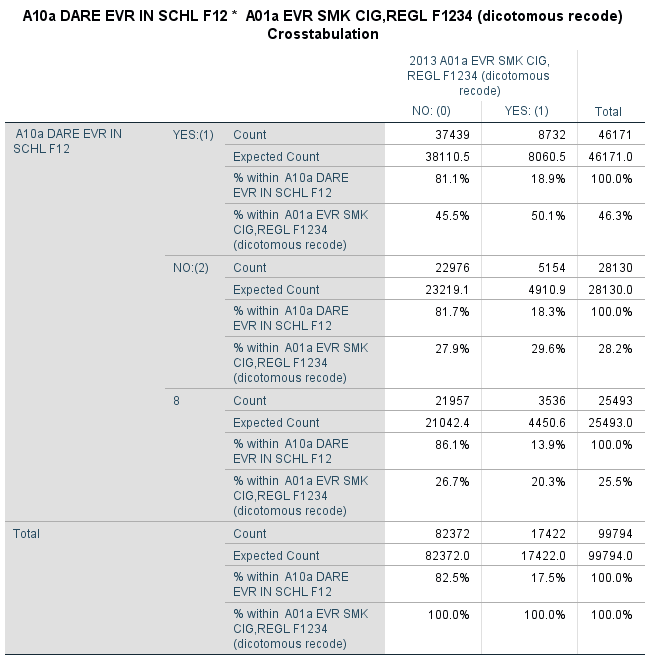
* The majority of school-aged kids surveyed are non-smokers, irrespective of DARE education.
* Most of the kids surveyed had taken the DARE program.
* An interesting observation is that that of the kids who did NOT have DARE in school and kids who don’t know if DARE was taught, the number of smokers/non-smokers is roughly the same. The third level to our DARE variable is interesting as it acts as a holdout for which to compare against our known students.

Our second analysis is a classification tree to help see the split adolescents who have taken DARE Education and not. As can be seen in the following slide, we can see a visual breakdown of our cross-tabulation table above.



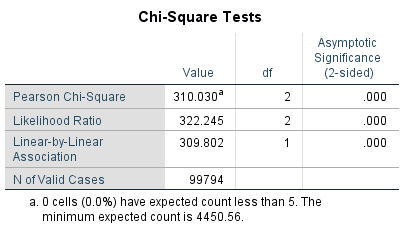


For our next descriptive analysis, we conduct a cross tabulation to determine the association between our two variables.



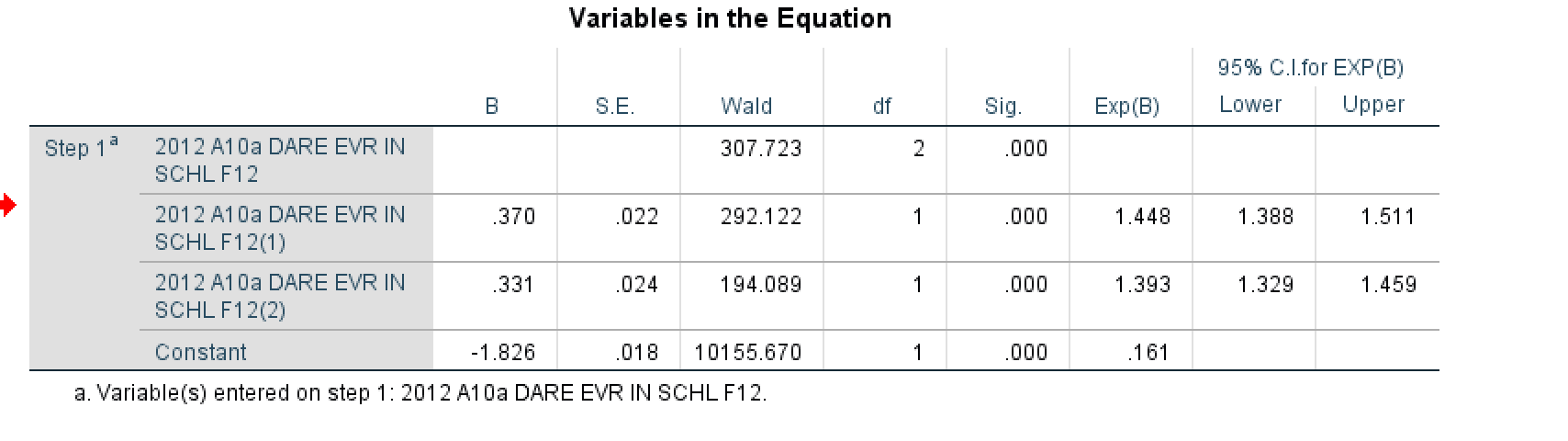
From this analysis, we show:

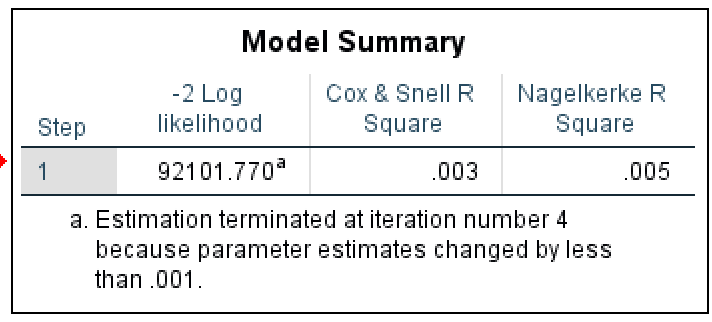
* Of all the students surveyed who have taken DARE at school, 81% have not smoked where 19% have smoked.
* Of all the students surveyed who have not taken DARE at school, 81% have not smoked where 18% had smoked.
* Of all the students surveyed who did not know DARE was offered at their school, 86% had not smoked where 14% had smoked.



Finally, we run a binary logistic regression due to the nominal nature of our data, we designate V7101D (Ever smoked cigarettes regularly) as our dependent variable (DV) and V7482 (DARE Ever in School) as our independent variable (IV). We see that for students who have smoked cigarettes, they are 1.4X more likely to have smoked if they have taken DARE education in their school. We also see that students who have smoked cigarettes to have been 1.39X more likely to have smoked cigarettes if they had not taken DARE education – a negligible difference of .05%. The difference between those who have smoked and taken and not taken DARE education is small enough to consider this significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Categorical Variables Codings | | | | |
|  | | Frequency | Parameter coding | |
| (1) | (2) |
| 2012 A10a DARE EVR IN SCHL F12 | YES:(1) | 46171 | 1.000 | .000 |
| NO:(2) | 28130 | .000 | 1.000 |
| 8 | 25493 | .000 | .000 |





**CONCLUSION AND RECOMMENDATIONS**

In closing, the DARE program has room for improvement; schools and administrators can do a better job of promoting the DARE program. Our analysis found a sizeable number of students (approximately 30,000) were unaware of their school offered (or did not offer) the DARE program.

The DARE program can be updated and refined to address some of the shortcomings and threats to student adoption of smoking. Many kids are looking at smokeless tobacco products such as “vaping” as an alternative to smoking thinking it is safer than traditional cigarettes.

In addition, there are numerous factors which contribute to student smoking cigarettes, education is only a small part of the equation, and our analysis shows that DARE education only explained .005 of the reason why students smoke.

## **2. How does drug use influence adolescents’ school performance?**

**VARIABLE SELECTION:**

For the second question, we performed two analysis. A) How adolescents’ drug use would influence their performance in school and B) How adolescents’ drug use and demographic (family) situation would influence their performance in school. As shown in the figure below, we’ve narrowed down 18 variables related to the question as potential independent variables and one dependent variable: V7234 for analyzing respondents’ school performance.

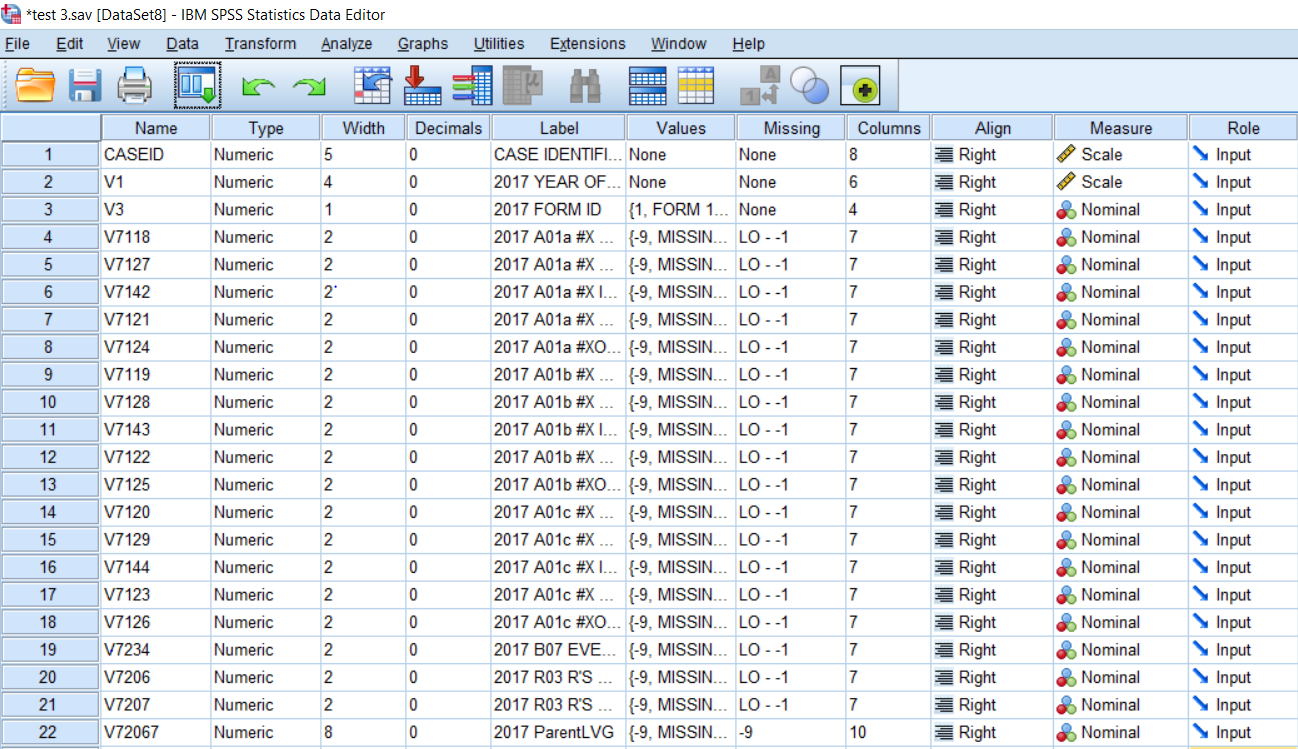
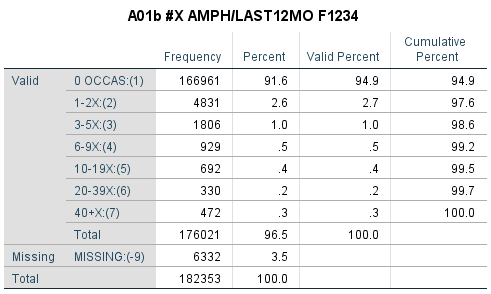
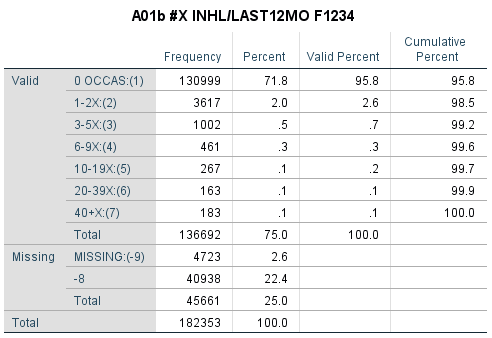


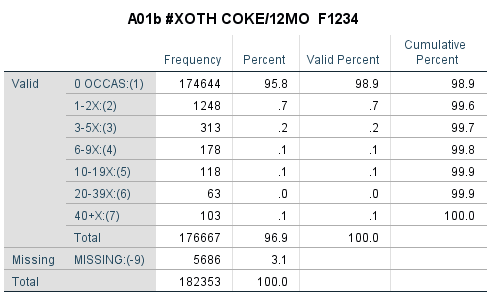
Figure 1. Potential Independent Variables

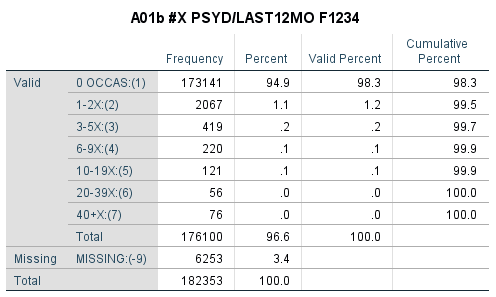
First, we located the most relevant independent variables for our further analysis. Out of our 18 potential independent variables, 15 variables are about the number of occasions of respondents’ use of different kinds of drugs during different time ranges: 30 days, 12 months and life time which include hallucinogens (V7119, V7118, V7120), crack (V7122, V7121, V7123), cocaine (V7125, V7124, V7126), amphetamines (V7128, V7127, V7129) and breathing the contents of aerosol spray cans to get high (V7143, V7142, V7144). Since the dependent variable asks respondents whether they’ve been suspended or expelled from school, we chose ‘life time’ as the most appropriate time range for analyzing the influence of adolescents’ drug use on their school performance because ’30 days’ and ’12 months’ are too short to be considered as effective factors.

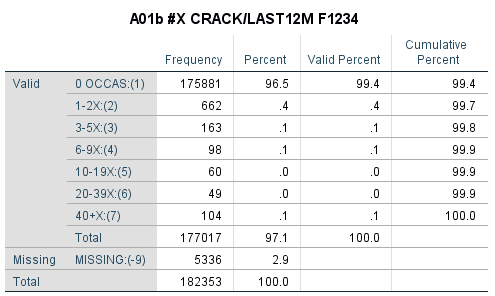
Then we ran Frequency Analysis on the 5 potential independent variables with time range of ‘life time’, and chose the top 1 drug with the most frequencies of using the drugs on more than 10 occasions during that time range, which is amphetamines. So finally, we used ‘amphetamines’ (V7127) as the final independent variable.







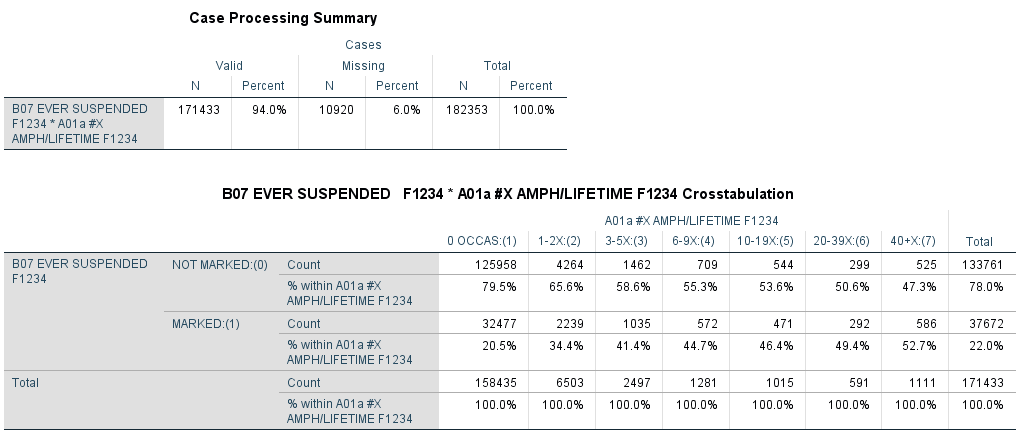




**A) How adolescents’ drug use would influence their performance in school**

We ran a cross tabulation to get the basic idea of how these the variables V7127 and V7234 are related. The dependent variable has 6% missing values, which is feasible for analyzing. It is shown in the table below that for those respondents who have never been suspended or expelled from school before, the percentage of them taking amphetamines in more occasions is decreasing, and on the other hand, for those who have been suspended or expelled from school before, the percentage of them taking amphetamines in more occasions is increasing.

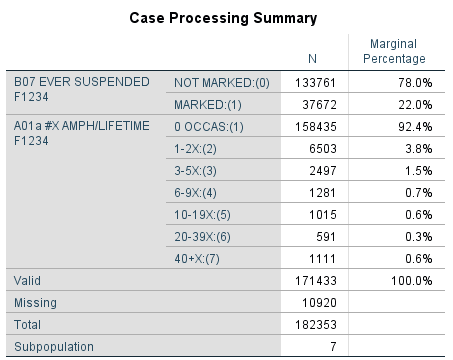
In general, it can be summarized from the table that students who have taken amphetamines more often tend to have a bad performance in school. The research of Hawkins, Catalano, and Miller (1992) also indicates that: “A low level of commitment to education and higher truancy rates appear to be related to substance use among adolescents. Cognitive and behavioral problems experienced by alcohol- and drug-using youth may interfere with their academic performance and also present obstacles to learning for their classmates” ([Bureau of Justice Statistics](https://www.bjs.gov/), 1992). Therefore, we would create two models: multinomial logistic regression model and classification tree model to verify the negative influence that drugs have on adolescents.

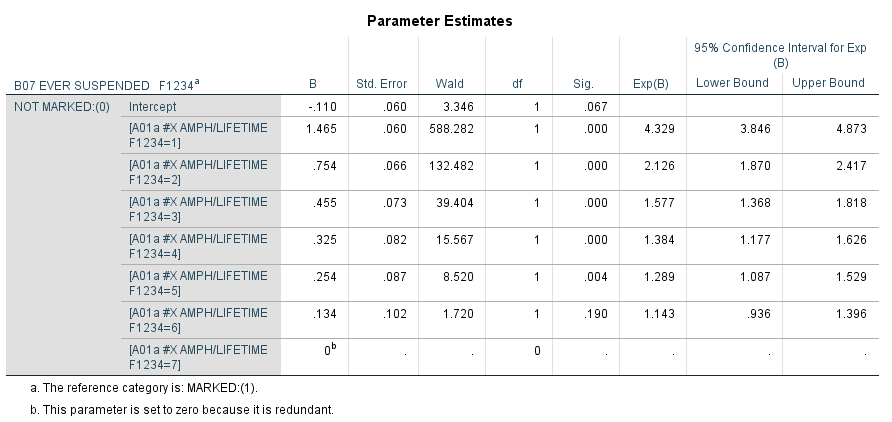
****

1. **Multinomial Logistic Regression model:**

In the first model, we ran multinomial logistic regression on V7234 as dependent variable and V7127 as an independent variable. V7234 asks respondents whether they’ve been suspended or expelled from school and V7127 asks about the number of occasions have they taken amphetamines without a doctor’s order in their lifetime. The reference category is ‘marked as has been suspended or expelled from school’ for V7234 and ‘have taken amphetamines without a doctor’s order on over 40 occasions in lifetime’ for V7127.

From the model, we can infer that for the students who have never taken amphetamines in their lifetime, the possibility of them not being suspended or expelled from school is 332.9% higher than the possibility of those who have taken amphetamines on over 40 occasions in their lifetime. And for the students who have taken amphetamines on 10 to 19 occasions in their lifetime, the possibility of them not being suspended or expelled from school is 28.9% higher than the possibility of those who have taken amphetamines on over 40 occasions in their lifetime. It can be concluded that students who have not taken any amphetamines in their lifetime tend to have a higher possibility of not being suspended or expelled from school, and the more often they have taken the drug, the lower the possibility would be.

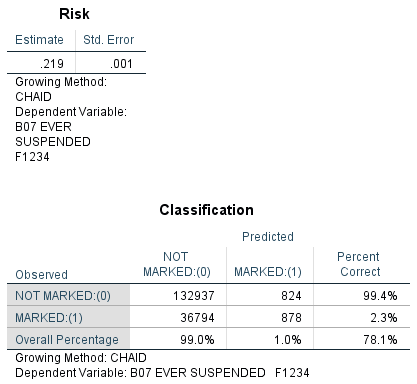


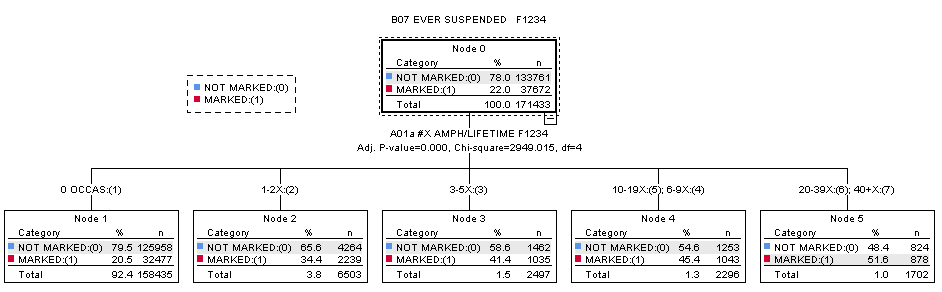


1. **Classification tree model:**

In the second model, we ran a classification tree on the same variables: V7234 as dependent variable and V7127 as an independent variable. It’s shown in the model that if students have never taken amphetamines in their lifetime, they are more likely not to be suspended or expelled from school under the possibility of 79.5%. On the other hand, for those students who have taken amphetamines in their lifetime on over 20 occasions, they are more likely to be suspended or expelled from school with the possibility of 51.6%.

The overall accuracy of this predicting model is 78.1%, which represents that based on the number of occasions that the respondents have taken amphetamines on in their lifetime, the accuracy of this model predicting their possibility of being suspended or expelled from school is 78.1%.





**Conclusion**:

From the two models, we can conclude that for students who have never taken amphetamines in their lifetime, the possibility of them not being suspended or expelled from school is higher than those who have taken the drug before. And the more occasions where students have taken amphetamines, the higher the possibility of them having poor performance in school is, which would result in being suspended or expelled from school.

**B) How adolescents’ drug use and the demographic (family) situation would influence their performance in school**

**METHOD**:

Due to the categorical nature of our dataset, we interpreted our results of Multinomial logistic regression.

Also, to make our research efficient and straightforward, we also recoded the variables, V7206, and V7207 to V72067 where 1= "STAY WITH BOTH PARENTS" 2= "STAY WITH ONE PARENT (EITHER MOTHER OR FATHER)" 3= "STAY WITH NO PARENT."

**DEPENDENT VARIABLE:**

1)V7234: 2017 B07 EVER SUSPENDED F1234

Ever been suspended or expelled from school?

1="No" 2="Yes, one time" 3="Yes, two or more times".

Categories 2 and 3 are combined in this dataset.

**INDEPENDENT VARIABLES:**

1. V7127: 2017 A01a #X AMPH/LIFETIME F1234

On how many number of occasions (if any) have you taken amphetamines [or other prescription

stimulant drugs] on your own -- that is, without a doctor telling you to take them in your lifetime?

1="0 Occasions" 2="1-2 Occasions" 3="3-5 Occasions" 4="6-9 Occasions" 5="10-19 Occasions" 6="20-39 Occasions"7="40 or More"

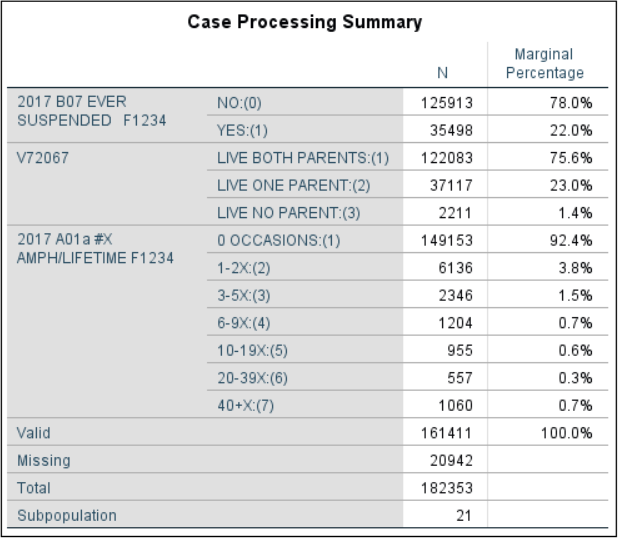
From the previous frequency analysis on the 5 potential independent variables related to drug usage with a time range of ‘lifetime’ we arrived on a conclusion of using Amphetamines as one of our independent variables for analysis due to its higher frequency rate

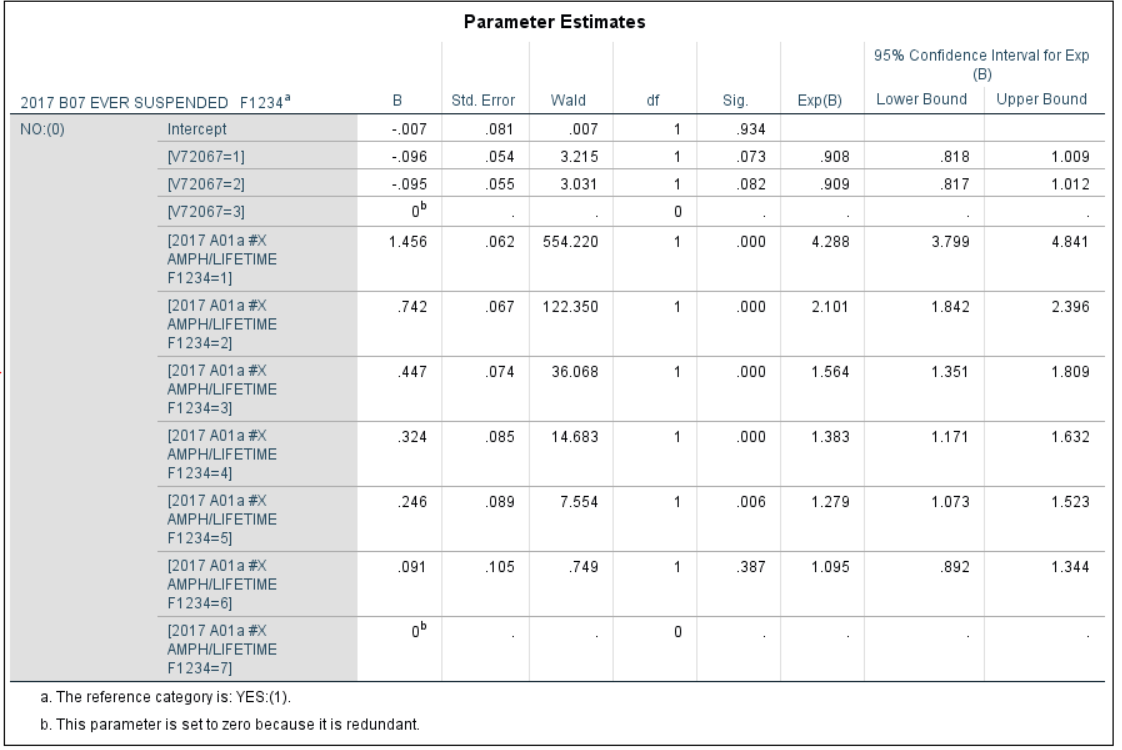
1. V72067: 2017 ParentLVG

1= "STAY WITH BOTH PARENTS" 2= "STAY WITH ONE PARENT (EITHER MOTHER OR FATHER)" 3= "STAY WITH NO PARENT."

**REFERENCE CATEGORY:**

The reference category is ‘marked as has been suspended or expelled from school’ for V7234, ‘have taken amphetamines without a doctor’s order on over 40 occasions in a lifetime’ for V7127 and ‘stay with no parent’ for V72067.





**INTERPRETATIONS**

**DRUG IMPACT:**

* From the model, we inferred that, holding demographic(family) situation constant the odds for the students who have never taken amphetamines in their lifetime, the possibility of them having not been suspended or expelled from school is 4.288 times or 328.8% higher than the possibility of those who have taken amphetamines on over 40 occasions in their lifetime.
* Similarly, holding demographic(family) situation constant the odds for the students who have taken amphetamines on 10 to 19 occasions in their lifetime, the possibility of them having not been suspended or expelled from school is 1.279 times or 27.9% higher than the possibility of those who have taken amphetamines on over 40 occasions in their lifetime.
* Also, holding demographic(family) situation constant the odds for the students who have never taken amphetamines in their lifetime, the possibility of them having not been suspended or expelled from school is 2.040 times or 104.0% higher than the possibility of those who have taken amphetamines on 1-2 occasions in their lifetime.
* Additionally, holding demographic(family) situation constant the odds for the students who have taken amphetamines on 6-9 occasions in their lifetime, the possibility of them having not been suspended or expelled from school is 1.091 times or 9.1% higher than the possibility of those who have taken amphetamines on 10-19 occasions in their lifetime.
* From this, we concluded that when there are a greater number of occasions where students take amphetamines, there is a higher possibility of them having poor performance in school which would result in being suspended or expelled from school.

**DEMOGRAPHIC IMPACT:**

* Now holding Amphetamines usage constant, the odds for the students who stay with both parents, the possibility of them having not been suspended or expelled from school is 0.092 times or 9.2% higher than the possibility of those who stay with no parents.
* Also, on holding Amphetamines usage constant, the odds for the students who stay with single parents, the possibility of them having not been suspended or expelled from school is 0.091 times or 9.1% higher than the possibility of those who stay with no parents.
* From this, we concluded that, when students stay in the right household environment with their parents, the odds of these particular students getting suspended is lower as compared to students staying with no parent.

**ACCURACY OF THE MODEL:**

Accuracy is one metric for evaluating classification models. In other words, accuracy is the fraction of predictions our model got right.

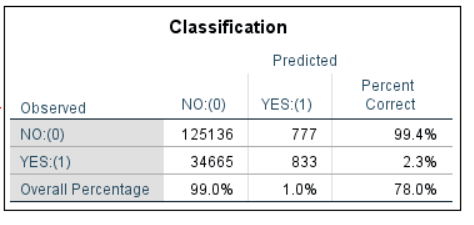
**Possible Classification Outcomes:**

TP: True Positives: Observations where the actual and predicted number of students who consume Amphetamines over a lifetime

TN: True Negatives: Observations where the actual and predicted number of students who did not consume Amphetamines over a lifetime

FP: False Positives: Observations where the actual number of students who consumed Amphetamines but predicted the number of students who did not consume Amphetamines over a lifetime

FN: False Negatives: Observations where the actual number of students who did not consume Amphetamines over a lifetime but predicted the number of students who consumed Amphetamines over a lifetime



The confusion matrix for a multi-class classification problem can help you determine mistake patterns (Shetty, 2018). The preceding confusion matrix shows that of 125,913 observations of students who did not consume Amphetamines over a lifetime, the model correctly classified 125136 as not having Amphetamines (True positives), and incorrectly classified 777 as not having Amphetamines over a lifetime (False negatives).

1. **Accuracy**

Accuracy = (125136 +833) / (125136+833+34665+777) = 0.7804 ~ 78%

That means the above model is 78% accurate

2. **Recall or Sensitivity**

In this case, it provides the percentage of the correctly predicted number of students who did not consume amphetamine over a lifetime from the pool of the actual number of students who did not consume amphetamine over a lifetime.

Sensitivity = 125136 / (125136+777) = 0.9938 ~ 99%

3. **Precision**

It gives the percentage of the correctly predicted number of students who consumed Amphetamines over a lifetime from the pool of total predicted number of students.

Precision = 125136 / (125136 + 34665) = 0.7831 ~ 78%

From above calculations, implemented multinomial regression model is 68% precise to provide correct prediction.

**Misclassification Rate** = (FP+ FN) / (TP+TN+FP+FN) = (777+ 34665) / (125136+833+34665+777) = 0.22

This shows misclassification rate is equal to 22%. So, 22% of the time the model misclassified the number of students who did not consume Amphetamines over lifetime.

**Conclusion**:

Now, look at the performance result. As seen in the above calculations, the model’s overall accuracy on the testing data is 78%. The model has a class recall of 99% for the “true” class implying that it can pick out students who did not consume Amphetamines over a lifetime with 99% accuracy. Also, this model has a misclassification rate as 22% which is less.

**OVERALL RECOMMENDATIONS:**

As the drug with the most usage frequency, amphetamines are influencing students’ school performance in a negative way. Therefore, we would recommend schools and departments concerned to strengthen the supervision of the usage of amphetamines among students and increase the awareness of the bad influence of this drug among school students. Also, the right amount of parental care and attention at home towards school children will influence their school performance in a positive way.

## **3. What is the likelihood of alcohol usage based on adolescent’s emotional state (depressed vs. happy) and the presence of a parent?**

**INTRO**

After focusing on the smoking and drug intake by adolescents, now we wanted to get some insights on the drinking habits of these adolescents. We did not look at the trend over the years since we wanted to target more on the question that do adolescents drink more when they are going through emotional changes like when they are enjoying their life or when they are depressed thinking that their life is meaningless.

For this purpose, we concentrated our attention on the below variables:

**Dependent**

* V7106D – #X DRNK/LAST12MO (Dichotomous Recode) where 1= "YES" and 0= "NO"

**Independent**

* V8502- LIFE MEANINGLESS where 1= "DISAGREE" 2="MOSTLY DISAGREE" 3="NEITHER" 4="MOSTLY AGREE" AND 5="AGREE"
* V8505- I ENJOY LIFE where 1= “DISAGREE” 2="MOSTLY DISAGREE" 3="NEITHER" 4="MOSTLY AGREE" AND 5="AGREE"

Further, to provide meaningful recommendations, we also tried to dive into the fact that is the family situation of these adolescents one of the factors responsible behind their feelings about life. We selected variables based on the from the results of the above analysis:

**Dependent (2 cases)**

* V8502- LIFE MEANINGLESS where 1= "DISAGREE" 2="MOSTLY DISAGREE" 3="NEITHER" 4="MOSTLY AGREE" AND 5="AGREE"
* V8505- I ENJOY LIFE where 1= “DISAGREE” 2="MOSTLY DISAGREE" 3="NEITHER" 4="MOSTLY AGREE" AND 5="AGREE"

**Independent**

* V72067 which we got after recoding the below variables
  + V7206- R'S HSHLD FATHER I.e. Stay only with Father, where 1= "YES" 2="NO"
  + V7207- R'S HSHLD MOTHER I.e. Stay only with Mother, where 1= "YES" 2="NO"

**METHOD**

Similar to the method used in our previous two questions of interest, we selected the dichotomous version of whether a student drank in the last 12 months. We utilized 12 months data over 30 days or once in a lifetime to get better understanding on the nature of adolescents drinking behavior.

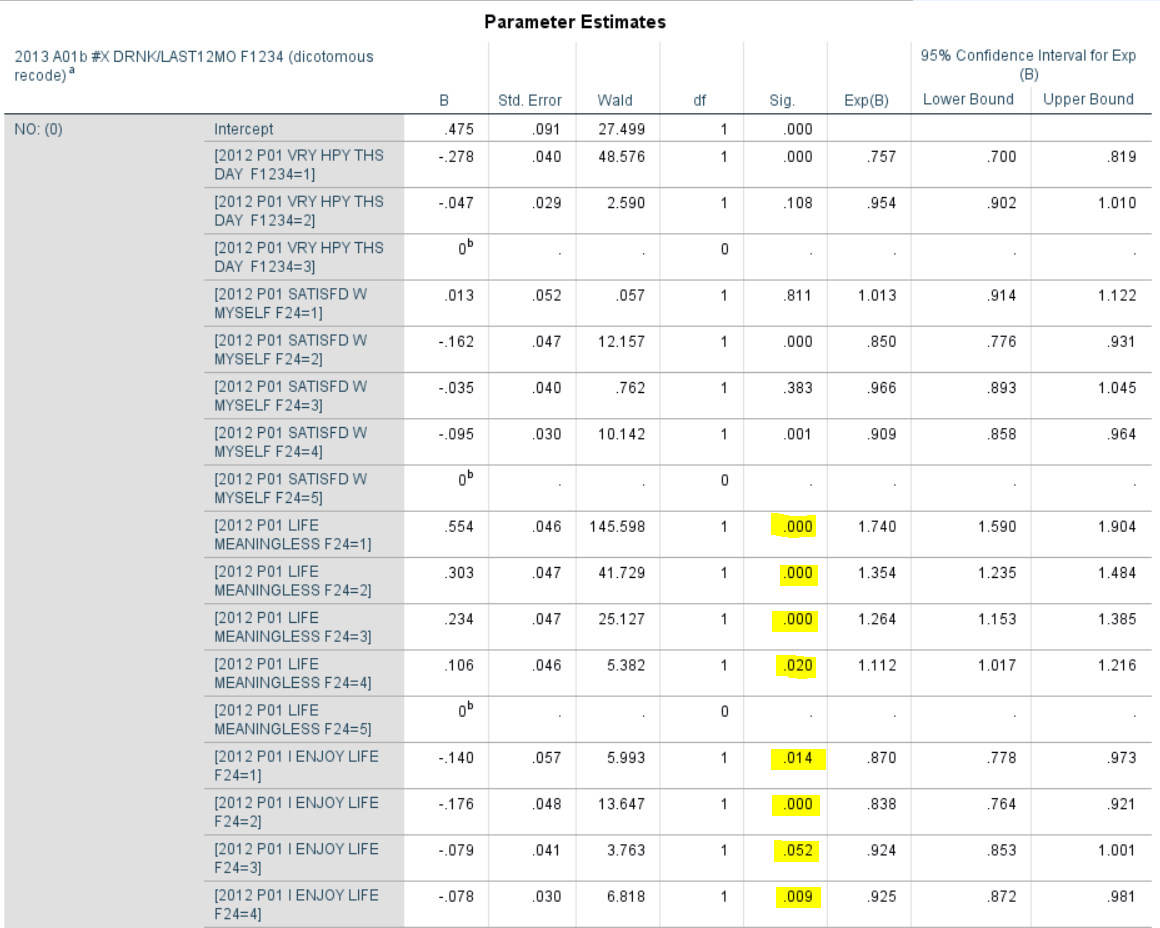
Again, due to categorical nature of our dataset, we interpreted our results of a) Multinomial logistic regression and b) Tree analysis.

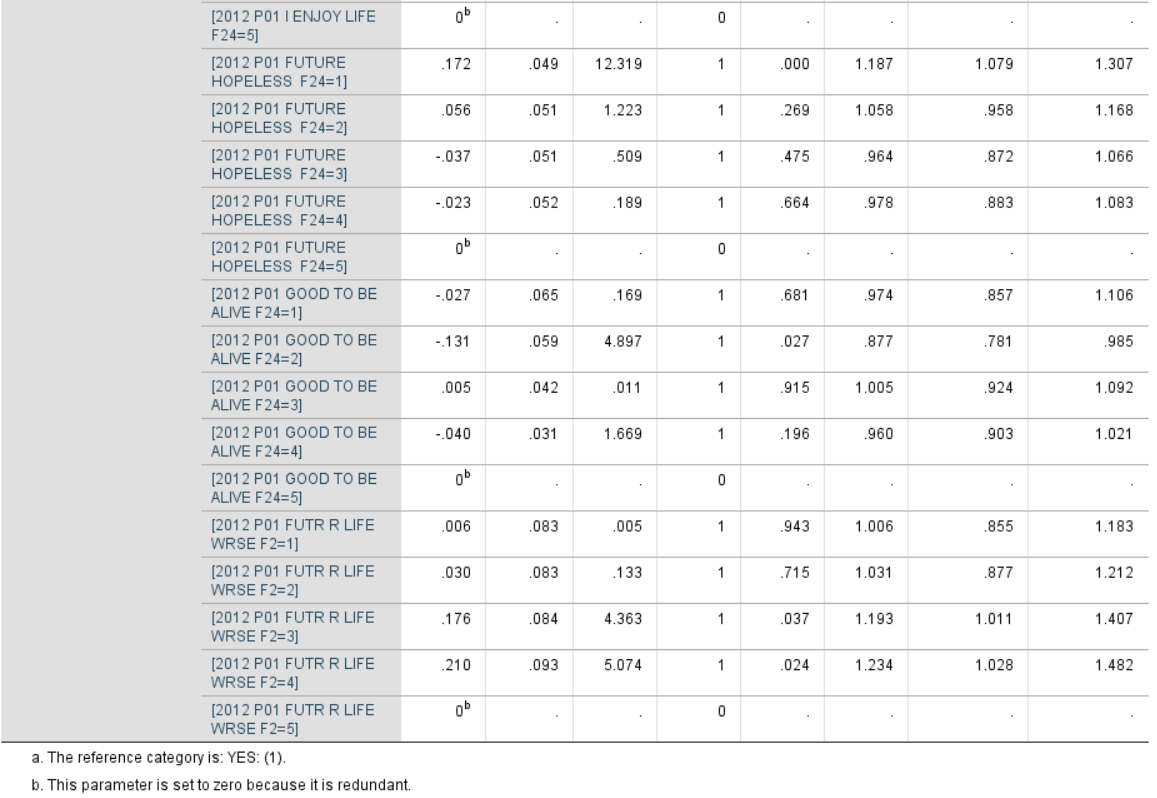
To make our research simple and efficient, we also recoded the variables, V7206 and V7207 to V72067 where 1= "STAY WITH BOTH PARENTS" 2= "STAY WITH ONE PARENT (EITHER MOTHER OR FATHER)" 3= "STAY WITH NO PARENT"

**ANALYSIS**

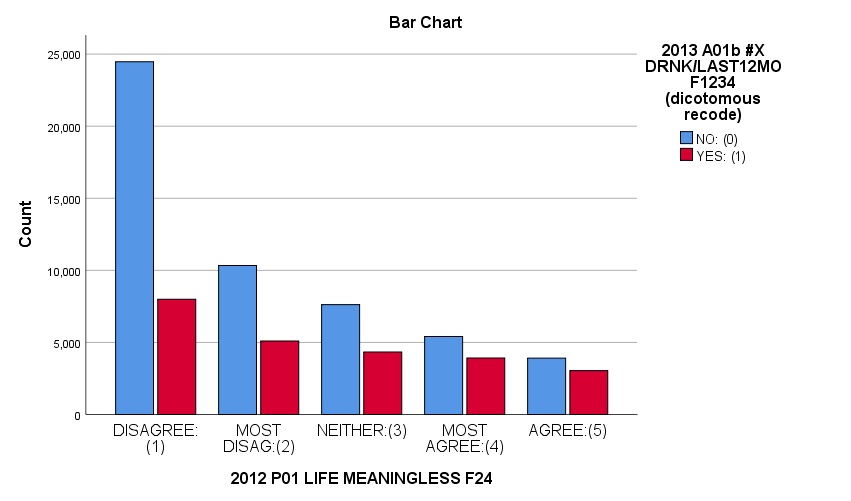
Emotional Changes and Drinking among adolescents

As the starting point, we tried to study the significance of all the variables available for emotional changes in the student’s life to the drinking variable. We ran multinomial logistic regression with drinking (V7106D) as the dependent variable and the remaining as the independent variables (V7302, V8512, V8502, V8505, V8509, V8514, and V8536) to get the pivot.

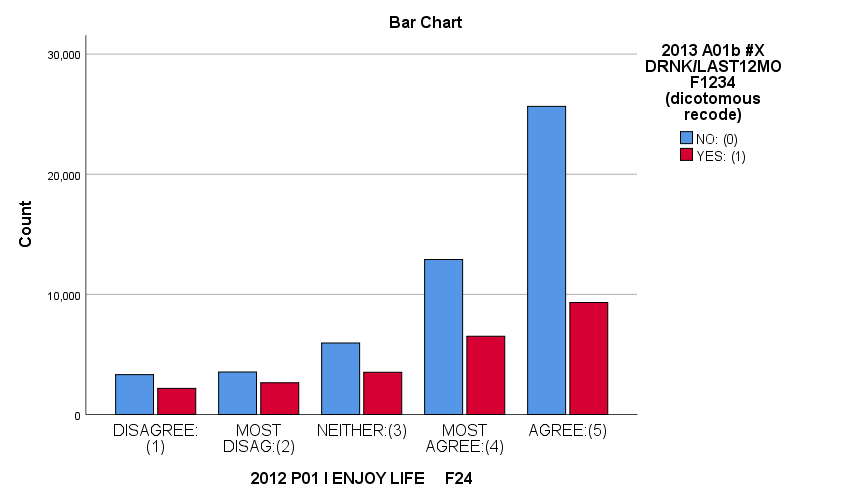
As is evident from the below table, we observed that the significant values of only the variables V8502 (Life Meaningless) and V8505 (Enjoy Life) are <0.05. Therefore, our analysis began using these 2 variables.



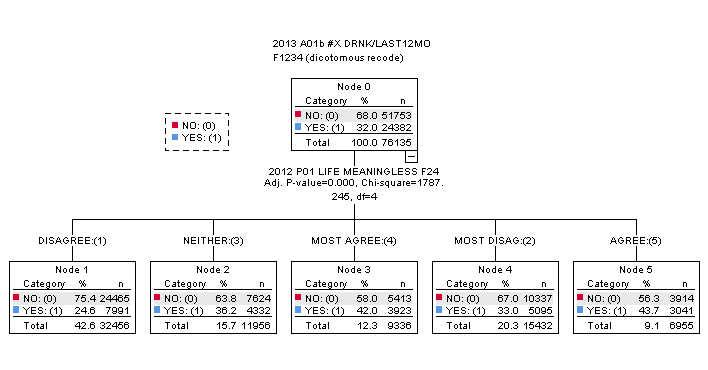
From the “red bars” of the below graph, it can be interpreted that the adolescents drink both when they are happy and unhappy about their life. However, we can see that the there are huge differences among “not drinking” levels as the emotions approach more to meaningless life. For instance, the “blue bars” depicting that the student does not drink goes extremely low as the emotions of the students in terms of life meaningless move from “disagree” to “mostly disagree” and so on till “agree”, I.e., he/she is enjoying life less.

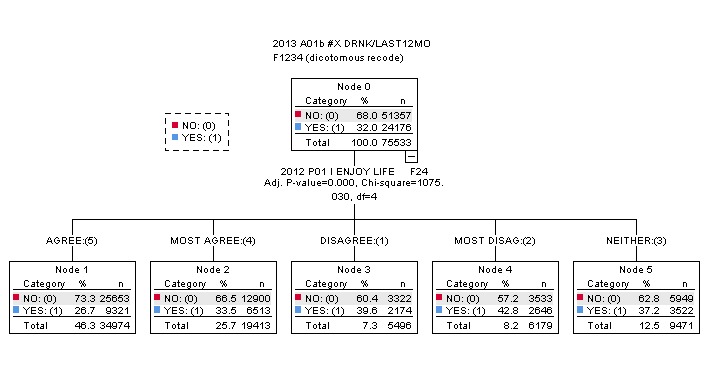


A similar trend is observed in the below graph. There is not much difference in levels of “yes” drinking. But, as the emotions within an adolescent goes from agreeing to enjoying life against disagreeing to enjoying life, I.e., life becoming more meaningless, the “no” to drinking is decreasing which means that they tend to drink more.



It is evident from the below classification trees as well that the “yes” to drinking decreases as the students go more towards agreeing to enjoy life and increases as they are morose due to agreeing to their meaningless life.

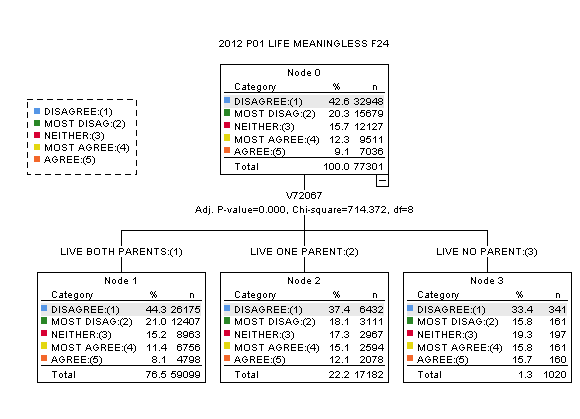


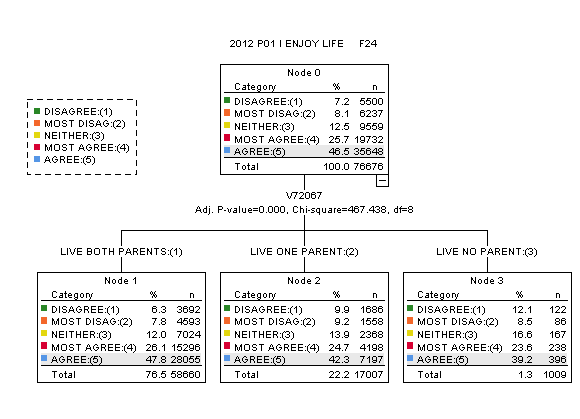


As a result, it could be the case that students favor drinking more when they are not happy with their life and think that their life is almost meaningless. Hence, it would be a good idea to see if their family situations are a factor of their life being meaningless so that we could recommend suggestions which would help in reducing drinking among adolescents.

**Family Situations (Parenting) and Adolescent Emotions**

From the below 2 classification trees we observed that the percentage of adolescents who “agree” that their “life is meaningless,” and that they are not enjoying their life is the most for those living with “no parent.” If we take a closer look at the direction of “life meaningless,” it increases from staying with “both parents” to staying with “one parent” to staying with “no parent.” In other words, the bent of adolescents close to “enjoying life” is decreasing from staying with “both parents” to staying with “one parent” to staying with “no parent.” The p-value is also extremely lower than 0.05 which is a good observation for our results.





**Accuracy of the model:**

1. **Accuracy of the multinomial logistic regression model**

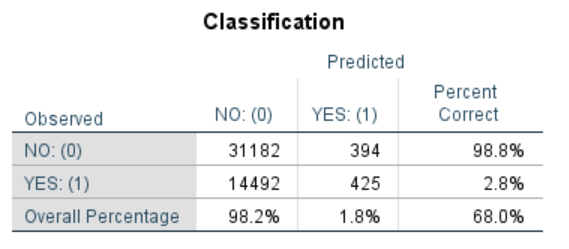
**Possible Classification Outcomes:**

TP: True Positives: The observations where the actual number of students who did not consume alcohol in the last 12 months are the same as the predicted number of students who did not drink alcohol.

TN: True Negatives: The observations where the actual number of students who consumed alcohol in the last 12 months are the same as our predicted number of students who drink alcohol.

FP: False Positives: The observations where the actually adolescents did not consume alcohol in last 12 months. However, our prediction says that they did.

FN: False Negatives: The observations where the actually adolescents consumed alcohol in last 12 months. However, our prediction says that they did not.



The preceding confusion matrix shows that of 31,576 observations of students who actually did not consume alcohol in last 12 months, the model correctly classified 31,182 as not having alcohol (True positives), and incorrectly classified 394 as having alcohol in the last 12 months while they actually did not (False positives).

1. **Accuracy**

Accuracy = (31182 + 425) / (31182+425+14492+394) = 0.6798 ~ 68%

That means above model is 68% accurate.

1. **Recall or Sensitivity**

In this case, it provides the percentage of Correctly predicted number of students who did not consume alcohol in the last 12 months from the pool of actual number of students who did not consume alcohol in the last 12 months.

Sensitivity = 31182 / (31182+394) = 0.9875 ~ 99%

1. **Precision**

It gives the percentage of correctly predicted number of students who consumed alcohol in the last 12 months from the pool of total predicted number of students.

Precision = 31182 / (31182 + 14492) = 0.6827 ~ 68%

From above calculations, implemented multinomial regression model is 68% precise to provide correct prediction.

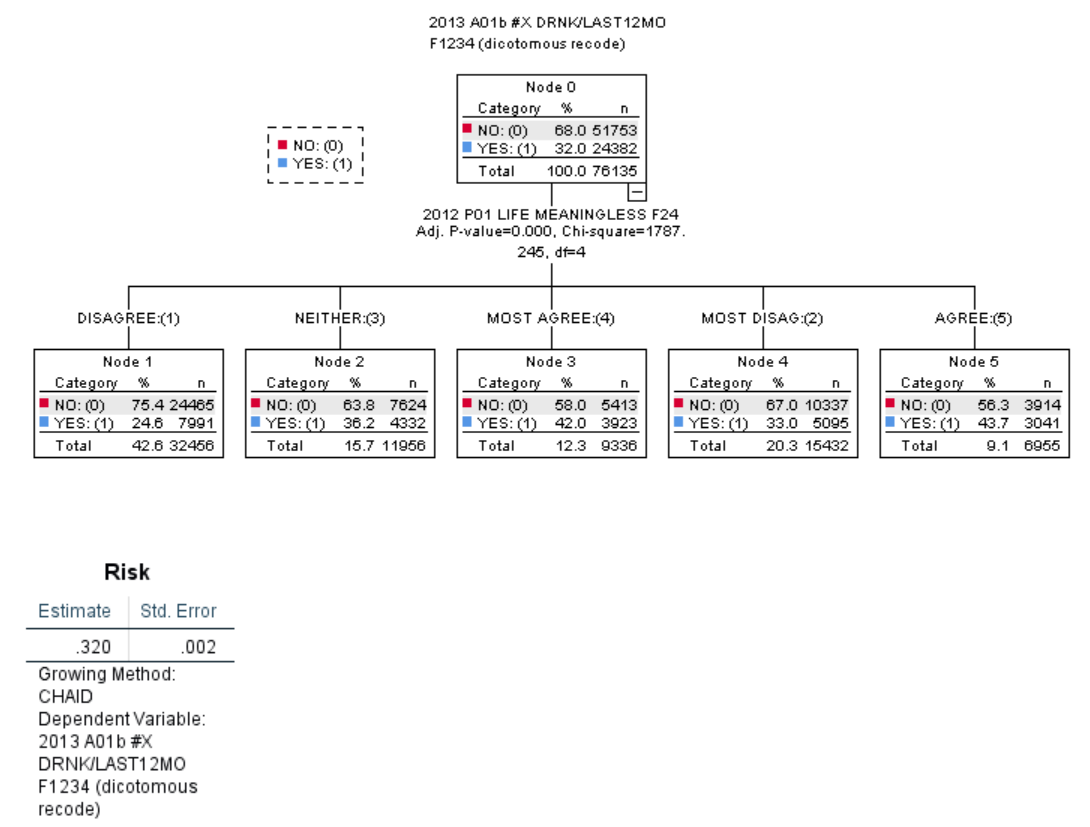
**Misclassification Rate** = (FP+ FN) / (TP+TN+FP+FN) = (394+ 14492) / (31182+425+14492+394) = 0.32

This shows misclassification rate is equal to 32%. So, 32% of the time the model misclassified the number of students who did not consume alcohol in the last 12 months.

**Conclusion:**

Now let us look at the performance result. As seen in above calculations, the model’s overall accuracy on the testing data is 68%. The model has a class recall of 99% for the “true” class implying that it is able to pick out students who did not consume alcohol in last 12 months with 99% accuracy. Also, this model has misclassification rate as 32% which is less.

**B. Accuracy of the classification tree**



Above risk estimates provides clear estimation of misclassification and it is affected by the misclassification cost matrix that we have supplied. This error is around 32% which is acceptable misclassification. Accuracy of the model is (1- misclassification error) which is equal to 68%.

**CONCLUSION AND RECOMMENDATIONS**

Evidence suggests that family environments constitute the basic ecology where children’s behavior is manifested, learned, encouraged, and suppressed. Parents’ roles in the family environment have primarily been to prepare children for adulthood through rules and discipline. During adolescence, however, the influence of peers also serves as an important socialization agent. Despite this new sphere of influence, research has clearly demonstrated that parenting accounts for more variance in externalizing behaviors in adolescence than any other one factor (Hoskins, 2014).

We also identified something similar from our dataset that one of the factors why adolescents feel that their life has no meaning is that their family situation is not correct. Their parents do not pay attention towards them and they stay with either their mother or father or no on. This causes depression in their mind. As a result, they drink more to overcome it. If we keep ourselves in their shoes, we would also feel dejected in such a situation and believe that our life is meaningless and has no colors. Not only this, these adolescents might also move towards taking more drugs or smoking as they grow in an environment where they do not enjoy their life.

Adolescence can be a confusing time of change for teens and parents alike. Therefore, we would like to recommend that parents should spend more time with their children and show positive attention towards them. Parents should ensure that their children are not aware of the bad circumstances between them. Also, both parents should together stay with their children since an adolescent requires care and love of both parents, not just one.

Moreover, parents should discuss with their children what behavior is acceptable and unacceptable at home, at school and elsewhere and what is good for them and what not. For instance, parents can teach their children that drinking, smoking, or taking drugs is something that would harm them. From our study, we believe these recommendations will help in reducing not only drinking habits among the adolescents but also the intake of drugs and smoke.

# **VIII. Closing Statement**

After doing extensive analyses on the effects that substance abuse has on adolescents, we believe that education and parental involvement will play a huge factor in helping curb this rising trend. It is therefore, our recommendation that governments and research agencies invest in more preventative measures that can also function as a cure. By doing so, we can foster an environment for adolescents to excel in their studies and redirect them from a destructive lifestyle.

References

Bureau of Justice Statistics. 1992 (December). Drugs, Crime, and the Justice System. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

“Confusion Matrix.” Confusion Matrix - an Overview | ScienceDirect Topics, [www.sciencedirect.com/topics/computer-science/confusion-matrix](http://www.sciencedirect.com/topics/computer-science/confusion-matrix).

Hoskins, Donna Hancock. “Consequences of Parenting on Adolescent Outcomes.” MDPI, Multidisciplinary Digital Publishing Institute, 18 Sept. 2014, [www.mdpi.com/2075-4698/4/3/506/htm](http://www.mdpi.com/2075-4698/4/3/506/htm).

Miech-Richard A.- Johnston- Bachman-Jerald G.- O'Malley-Patrick M.- Schulenberg-John E.- Lloyd D.. “Monitoring the Future: A Continuing Study Of American Youth (8th- and 10th-grade Surveys)”, 2017 - <https://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/37183>

Xiang Lee, Min. “The Two Approaches to Data Analysis: A Beginner’s Guide”, 4 Apr. 2018, <https://medium.com/@minxianglee/the-two-approaches-to-data-analysis-a-beginners-guide-df7059c4416e>

“Simple guide to confusion matrix terminology”. 25 Mar. 2014, <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

“An overview of correlation measures between categorical and continuous variables”, Sep. 13 2018, <https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365>

Badreesh Shetty. “The Supervised Machine Learning: Classification” 12 Dec. 2018,

<https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d>