[ANLY 699-91- O-2019/Late Fall - Applied Project in Analytics](https://harrisburgu.instructure.com/courses/2025)

Predicting SUICIDAL factors IN Youth BEHAVIOR

Harrisburg University of Science and Technology

PRIYA SINGH

PROFESSOR

**Ali Motamedi, PhD**

Contents

[ABSTRACT 3](#_Toc32760121)

[INTRODUCTION 3](#_Toc32760122)

[1. BACKGROUND: 3](#_Toc32760123)

[2. Statement of Problem: 4](#_Toc32760124)

[3. Purpose of the Study: 5](#_Toc32760125)

[LITERATURE REVIEW 6](#_Toc32760126)

[RESEARCH OBJECTIVE 12](#_Toc32760127)

[METHOD AND DESIGN 13](#_Toc32760128)

[1. Descriptive analysis: 15](#_Toc32760129)

[2. Exploratory data analysis: 16](#_Toc32760130)

[3. Predictive analytics: 23](#_Toc32760131)

[SIGNIFICANCE AND CONCLUSION 31](#_Toc32760132)

[REFERENCES 32](#_Toc32760133)

[APPENDIX 34](#_Toc32760134)

[1. CODE 34](#_Toc32760135)

ABSTRACT

The main objective of this research is to investigate the primary factors that can lead youth to attempt suicide or make a suicide plan. Research samples consist of 36,667 number of youth who attempted suicide from the year 1991 to 2017 from all of the states in United. This research is going to apply regression and various machine learning algorithms to these data in order to study the influence of various factors on attempting suicide by youth.

INTRODUCTION

1. BACKGROUND:

Youth behaviors and risk associated with it is the most leading issue nowadays. Academicians, social scientists, and researchers have conducted exhaustive studies on a wide range of topics that can affect adolescent risk behavior. Often, adolescents are involved in these risky behaviors, which can impact their studies and forced to drop out of college.

In one of the research of (CDC Releases 2017 Youth Risk Behavior Survey (YRBS) Results. (2019, February 21)), which is a morbidity and mortality report of the 2017 Youth Risk Behavior Surveillance Survey(YRBS). The statistics show that the percentage of adolescents who had been physically forced to have sex has not improved. While the percentage of youth who had sex decreased from the year 2007 to 2017, the study shows that there have been many improvements; despite them, there are more indicators to be needed to protect the student from risky behaviors. Our research would help us in analyzing more indicators for helping adolescents.

In one of the previous study,(Khubchandani, J et al., 2017), the study emphasizes the Hispanic males and females and their predictive behavior. The Hispanic students who had git good grades are not involved in related violent behavior. Hispanic males and females tend to be involved in Tobacco, alcohol, drug use, and violent behavior, which can lead to weapon carrying behavior. The study says that the violent behavior in Hispanic females in the year 2001 to 2015 had significantly reduced and suicide attempt has increased whereas Hispanic males’ violent behavior and suicide attempt have shown a decreasing trend in the year 2001 to 2015. This research would help in carrying out future analyses on the other races and comparing the behaviors in each race and gender. Most of the researchers have used the Youth Risk Behavior Surveillance Survey data, which was collected by the Center for Disease Control and Prevention. The survey started in the year 1991 for students in grades 9-12. The survey from 1991 to 2017 will be used to analyze the trends over the year. The survey was obtained from multiple resources like a national school-based survey conducted by CDC, state schools, tribal, and urban schools. Researchers in their course of study has analyzed a wide range of topics like Dietary related risks, weapon carrying risks, and suicidal ideation among adolescents.

1. Statement of Problem:

Youth Risk Behavior Surveillance has been used in many kinds of research to understand adolescents to understand their behaviors. The drug use and alcohol use in adolescents can lead adolescents to criminal activities and may lead them as a vulnerable child. These behaviors affect other youth to get involved in risky behaviors in school. (Shopowich, S., 2019)This paper has studied the significant problems Youth and adolescents are facing. In this research, various factors were studied to find the relationship between adolescent sexual activity and HIV testing for the 2015 survey. This research says that HIV testing was more common in students who had taught than those who had not taught in school. The study had done for the year 2015 YRBSS survey. The analysis emphasizes that whether the victims who had sexual assault had to get tested for HIV if they had sex education before the sexual assault. This study also focussed on implementing more sex education in schools and HIV testing. Our study focusses on analyzing the factors and predicting which can influence youth risk and at which age they are vulnerable. The substance use, alcohol use, and sexual activity are the most common things that adolescents are involved in while they are in school. Often, these behaviors are not noticed when a kid goes to school. This research would also help in understanding the physical activity, related dietary behaviors, and self – injuries among adolescents and how are they contribute to the youth's risky behaviors.

The study (Kann, L et al.,2018), which had analyzed Youth Risk Behavior Surveillance - the United States, 2017. The article focusses on analyzing the survey of the year 2017. The research reports that the study has been done to understand the leading cause of mortality and morbidity among youth and adolescents. The most critical finding in the research is that health-related risk behaviors are significantly higher in sexual minority students than non-sexual minority students. This research will be a guide to understand the significant areas and find the relationship between the variables.

1. Purpose of the Study:

The purpose of this research is to explore adolescent’s health risks to demographic variables, sexual activity, substance use, and alcohol. The research to be focused on predicting the factors can influence lead youth to attempt suicide. By finding the driving factors, adolescents to be educated about the side effects of substance use and alcohol among males and females. The predictive algorithms will be used for finding the driving factors.

LITERATURE REVIEW

There are much research has been done on analyzing youth risk behaviors. The research has been done on a variety of factors and analyzing the trends. The Center for Disease Control and Protection has been working on the Youth Risk Behavior Surveillance Survey since 1991. The questionnaire includes state, national, territorial, tribal government, and local-based surveys of students in grades 9-12th every two years. The questionnaire includes dietary related, alcohol and substance use, minority sexual risks, and weapon carrying the risk. This study focusses on overall risks that can impact youth risks. (Kann, L. et al.,2018) focussed on analyzing the overall survey of the year 2017. The research will help understand what has driven factors in the year 2017 that can lead youth in mortality and morbidity. The result of this research that health-related risk behaviors are significantly higher in sexual minority students than non-sexual minority students.

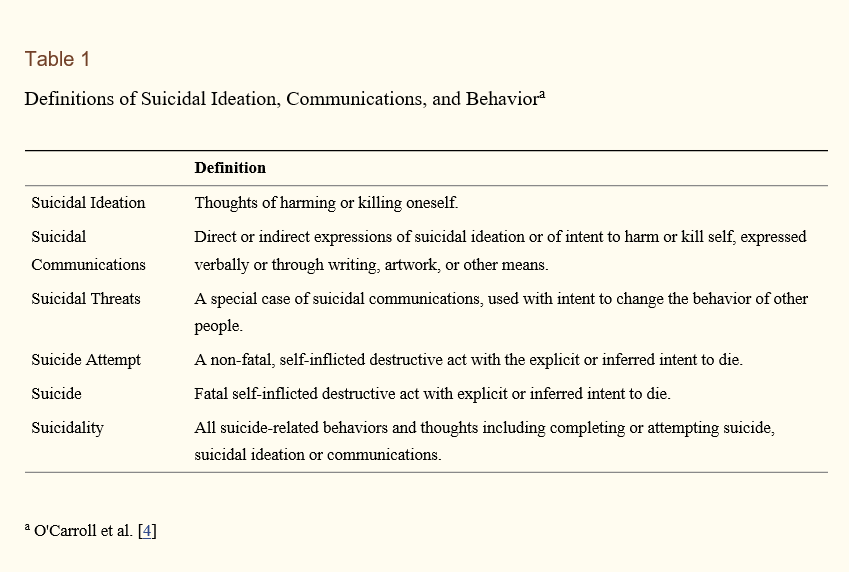
(2018, June) The CDC published a report on how to analyze the YRBS sexual minority data. The reports say that sexual identity and sex of sexual contacts may both be used to identify the sexual minority youth. The question was included in the 1995 survey. The YRBS uses a multi-stage cluster sample design. The YRBS referred to use various statistical packages like SAS, SPSS, R, etc. Identity and Experiences of Violence Victimization (2019, April 8) also focussed on analyzing the disparities in transgender youth who had experienced sexual violence, substance use, suicide risk. In the year 2017, 63.8% of transgender youth report that no protection was used during their last sexual intercourse. This study will help in getting more in-depth and understanding the trends from 1991 – 2017.

The sexual assaults are one of the significant youth risks and which have to be studied in order to understand the factors that can lead to sexual assaults. A recent study done by (Shopowich S et al.,2019) was focused on females' sexual assault and HIV testing. The study had focused primarily on females that how HIV testing is more common in students who had eduction about HIV testing. A recent study by (Fish, J. N, et al.:, 2018) discussed the trends in alcohol-related disparities among heterosexuals and sexual minority youth. The study differentiated how these activities differ between them. The research results showed that there was a decline in the percentage of alcohol-related disparities among heterosexuals, but there is no consistency in the percentage of sexual minority youth. The percentage was high for the bi-sexual girls who had an alcohol use than for the heterosexual girls.

Often transgender females become a victim of sexual violence. Many factors can lead them to suicide risks and substance use. There was another study in 2017 by Johns, M. M et al., (2019). This study focused on studying transgender females in schools who had been a victim of sexual violence. The research result showed that transgender females and males had reported sexual violence than cisgender males and females. Though there had been many limitations in the data, which was reported in many studies. There can be uncertainty that the high-risk students might not have participated.

Alcohol and substance abuse at an early age can often lead to youth and adolescents in criminal activity. The current research is still going on to find out the factors which can lead youth into risky behaviors. One of the research by (Baiden, P et al.,2019) studied the association between age at first alcohol use and suicidal ideation among high school students. The research aimed to reduce suicidal ideation by improving the health initiatives, which will help in providing better growth development among adolescents. The research listed some of the limitations. The prior medical history records questions, family history and community factors that can cause suicidal ideation in adolescents — future research to be focused on adding specific questions related to their family history. The research should also be focused on analyzing the criminal records of the youth who had been involved in substance use and alcohol. There are also certain diseases like cancer, diabetes, obesity, kidney and liver failure, etc. The research study by (Shield, Parry, & Rehm, n.d.) studied alcohol consumption as a risk factor for many chronic diseases and conditions. The average volume of alcohol can lead to morbidity and mortality due to chronic diseases. There are more than twenty-four types of chronic diseases that are caused by alcohol consumption. The research studied that alcohol consumption can also lead to cardiovascular diseases. However, there is also a protective impact. However, in general, if the adolescent is more involved in the early stages, it can impact their body parts and functions.

The other common concern of alcohol use when the youth is under the influence they can cause self-injuries to themselves. The youth involved in alcohol consumption due to various reasons like victims of violence, depression, anxiety. The suicide-related behaviors can also depend upon race and age. This leads to suicide ideation and suicide-related behaviors in youth. The research Cash, S. J., & Bridge, J. A. (2009) studied that suicide is the third leading death among young adults in the United States. The research was focused on understanding the relationships between psychopathology, substance use, child abuse, bullying, internet use, and youth suicidal behavior. The below table1 O'Carroll et al. [[4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2885157/#R4)] lists all the types of suicidal behaviors. The results showed that 14.5% of the high school students in 9th to 12th grade reported suicide ideation, and 6.9% of the youth reported at least one suicide attempt in the previous year. Also, suicide and suicide-related behaviors rate increase with the age of the boys and girls. The girl's suicide rates are more than boys.



The other common concern in youth in alcohol consumption that can lead to self-injuries to themselves. It can be caused due to depression, anxiety, bullied by other youth, etc. The youth also face sexual assaults in their communities or school. Some of these factors can influence youth to consume alcohol at an early age. The study by (Key, K et al.,2016) studied that addressed the research on self-injuries focused mostly on Caucasian Females and other populations. The research also focused on analyzing gender, culture, and diversity-related self-injuries behaviors among them. The research used logistic regression to determine the non-suicidal self-injuries behaviors and aggressive behavior that can lead to suicide attempts. The research finding described by the author was that the males try to burn themselves by use hitting and burn themselves, and females tend to cut. Limitation listed by the authors that there was a lack of linking self-injury with the suicidal attempt. The future research discussed in the article is to study the self-injury in adolescent males as one group and overall behavioral analysis of gender.

Teen dating violence is one of the significant issues in today's society. Teen dating violence is an intimate partner crime. Teen dating violence is of four types. They are listed as 1. Physical violence, 2. Sexual violence 3. Psychological aggression, 4. Stalking. Dating violence usually takes place electronically or in person by texting, posting sexual pictures of the partner without consent. The CDC results ("Preventing Teen Dating Violence," 2019) indicate that teen dating violence affects millions of youth in the U.S. each year. The results showed that 1 out of 11 females students and 1 out of 15 males had experienced physical dating violence. Teen dating violence can lead to a negative impact on the victims teen. They can go in depression, anxiety, engaging in unhealthy behaviors, alcohol use, substance use, theft, bullying and suicide ideation. Teen dating violence can be controlled by managing feelings, healthy communications in teenagers. Friends, family and community, creating a protective environment can work together in controlling the unhealthy behaviors.

The study by (Hall, C et al., 2015) analyzed the youth risk behaviors in “pretty little liars.” The American drama series had been analyzed by using content analysis to access risky behaviors. Some of the behaviors like bullying in school, alcohol consumption, illegal behaviors, trespassing, and sexual activity were analyzed. This research had focused on increasing health and education efforts in order to encourage students to improve their skills. The author pointed out that the shows like “Pretty Little Liars” could impact the adolescents for good. The research can also help parents in knowing what kinds of risky behaviors that can impact their child. Limitations listed in the research were that content analysis was a limited research methodology. The second was the use of coding instruments for analyzing the health risk behaviors among teens can be limited. Future research focused on the correlations between specific media content and healthy behaviors.

A recent study by (Baiden, P et al., 2019) discussed the association between teen dating violence and suicidal behaviors among adolescent high school students. The relationship between teen dating violence and suicidal behaviors among adolescents age 14-18 was studied. Often, team dating violence can lead to mental health problems among adolescents. The research used logistic regression to analyze the relationship between suicidal ideation, suicidal attempt, and teen dating violence as an explanatory variable. The study was able to find an association between teen dating violence and three types of suicidal behaviors. Future studies to be explored by learning all the factors that can influence suicidal behaviors.

There is also another research by Rosario M et al.,2014) which studied the sexual disparities in cancer-related risk behaviors when the adolescent is consuming tobacco, drug use, alcohol, and physical activity. The research used statistics to compare heterosexual and sexual minorities data. The article results showed that the sexual minority youth faced more cancer-related risks due to substance use and alcohol consumption than heterosexuals. This can be due to bullying and forced sex. Future research should also work on asking questions about sexual minority and their current physical health, which can help in analyzing the driving factors. The research should also include data from other states of the United States.

The youth risk also faces problems in their health due to nutrition, related dietary behaviors, physical activity, obesity, and overweight. Unhealthy food habits and alcohol consumption cause these at an early age. The early intervention in their food habits will help them in reducing dietary-related diseases. The study by CDC ("Nutrition, Physical Activity, & Obesity Data & Statistics | Adolescent and School Health | CDC," 2018) showed that 14.8% of the students in the year 2017 survey had reported obesity. 15.6% had reported that they were overweight. The youth also reported that 7.2% of the youth did not eat vegetables, 5.6% had drunk juices before the survey. The recent research by Burns, R. D. (2019) reported the association of energy-related behaviors with adolescent weight loss intent. Obesity is a significant problem in the United States. In this research, the aim was to analyze the associations between specific sedentary behavior, physical activity, and dietary-related variables and adolescent weight loss intent. The study results show that meeting daily physical activity and healthy diet consumption can lead to lower odds of weight loss intent. This research focusses that adolescents who intend to lose weight should be aware of health-related behaviors. The logistic regression was used to understand and analyze various factors like BMI, race, age, and other types of questions in the survey.

The drug use, substance use, alcohol consumption related articles have been studies which can impact risky youth behaviors. The youth risk behavior surveillance survey had six topics which monitor the youth adolescent risk. The research used various techniques to analyze the survey. They used various techniques, which are logistic regression, statistical tests, descriptive and exploratory analysis. The future research to be more focused on predicting the factors, analyzing the accuracy of the model, supervised classification model, cluster analysis, etc. The factors which are least important to be ignored and analyzing the more correlated factors.

RESEARCH OBJECTIVE

The quantitative research would have to be used to analyze various categorical and numeric factors and their dependency on each of them. The dataset of (Centers for Disease Control and Prevention (CDC), 1991-2017), which monitors the Youth Risk Behavior Surveillance Survey, would be referred to analyze the data and answer research questions. Previous analysis of the Center for Disease Control and Prevention for Adolescent and youth health behavior will be used as a guide to research. The leading research questions that will be focused on this research is: What are the factors contributing to adolescent’s behavior leading to suicide?

EXPLORATORY QUESTIONS

The below are sub-questions that were explored in this study?

* How does an unhealthy diet lead adolescents to suicidal ideation?
* Which gender has involved in suicide attempts?
* At which age and academic grade, adolescents have thought or have attempted suicide?
* How does exposure to drug abuse and alcohol, among other factors, contribute to higher rates of violence and suicide attempt among the adolescent population in the United States?

METHOD AND DESIGN

In order to find out the key determinants which can lead adolescent in making suicide plan or attempting to suicide. In this research, the main aim is to design a regression model and find out the driving factors. The regression model is the most commonly used statistical method and used as a predictive model. It is an ideal technique to explore the relationship between the various response factors and independent variables. Rstudio and jupyter used in analyzing the data. The Rstudio used for descriptive and exploratory research and Jupyter for machine learning algorithms in python. Below are the independent variables used for analyzing the research questions which were filtered out from the original data.

1. Demographic Variables:

* Age
* Gender
* Race
* Sexual identity

1. Dietary related behaviors:

* Green salad eating
* No soda drinking
* No milk drinking
* Fruit-eating
* Breakfast eating

1. Alcohol and substance use:

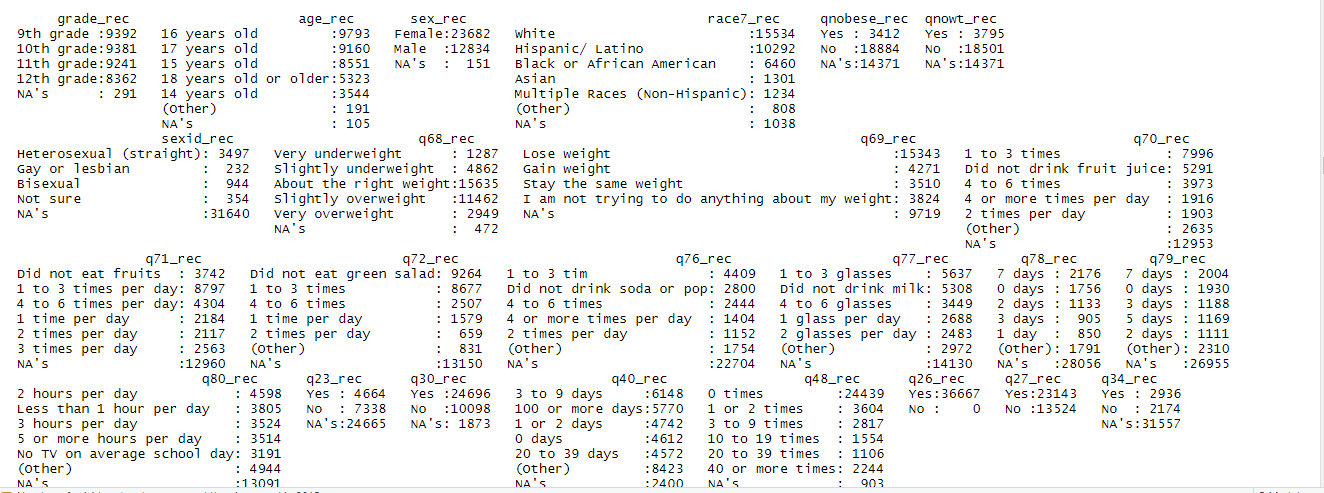
* Ever alcohol use
* Electronic vapor product use
* Ever marijuana use
* Ever cocaine use

1. Weight-related questions:

* Perception of weight
* Weight loss

1. Suicide plan
2. School grade
3. Year
4. Ever bullying in school
5. Descriptive analysis:

In order to analyze the data, the first step is the descriptive statistics of the data. The data had the total number of 20k records consolidated from all the years from 1991 to 2017. Some of the states had not participated in the survey, such as Minnesota, Oregon, Washington, and Wyoming. The question “During the past 12 months, did you ever seriously consider attempting suicide?” asked in the survey were filtered as “Yes” who tried suiciding for analyzing the group and performed regression models on by using suicide plan(q27) as a dependent variable too. The group who had tried attempting suicide in or made a suicide plan in the past 12 months the dependent variable. The total number of records was 36,667 of all the years.



1. Exploratory data analysis:

The exploratory data analysis is carried out to explore the variables statistics. In Figure 1, the percentage of suicide attempts in the past 15 years is not consistent. The suicide attempts increased alternatively after every two years of the survey. The year 2000 reported 7.17% where the youth had attempted suicide, and a tiny percentage of the youth made a suicide plan. In recent year 2017, the percentage of youth attempting suicide decreased.

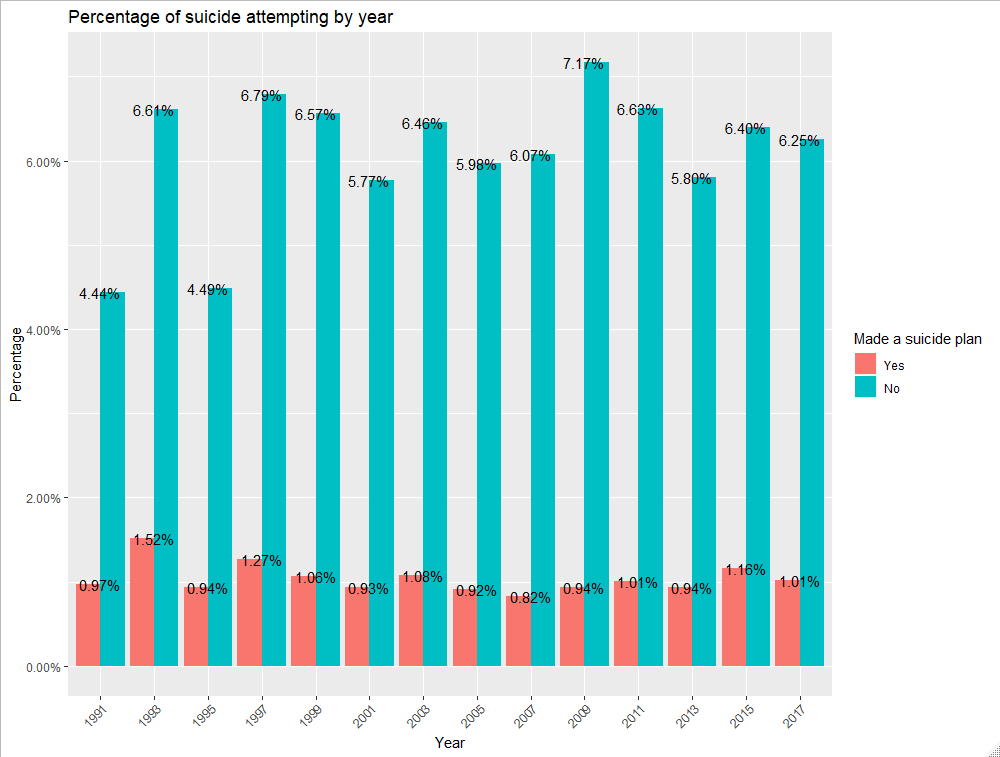


Figure 1

In Figure 2, the percentage of youth between 15 to 17 age is more attempting suicide than younger ages, and female’s percentage was higher in 15 – 17 years of age group.

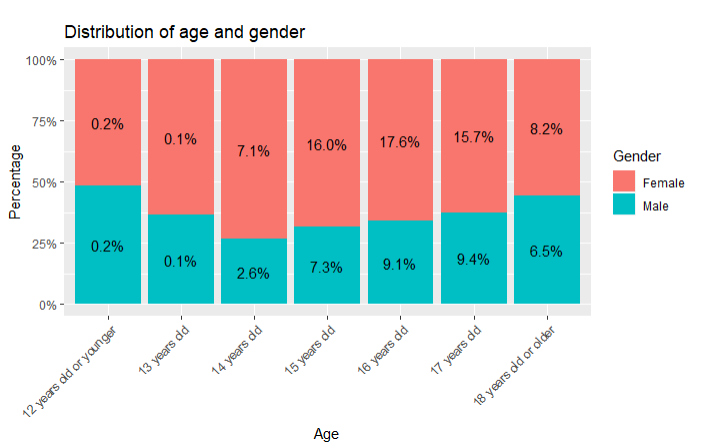


Figure 2

In figure 3, the percentage of females who tried attempting suicide was 64.9%, which more than males 35%.

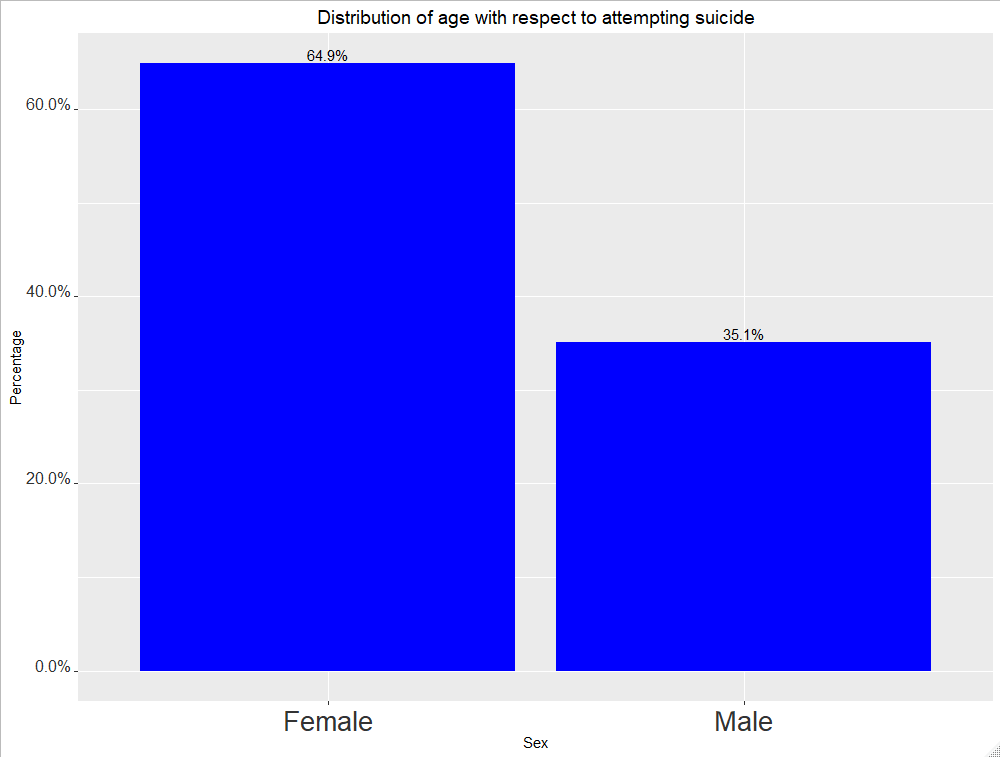


Figure 3

In figure 4, the percentage of youth who tried attempting suicide remains equally distributed in the grades.

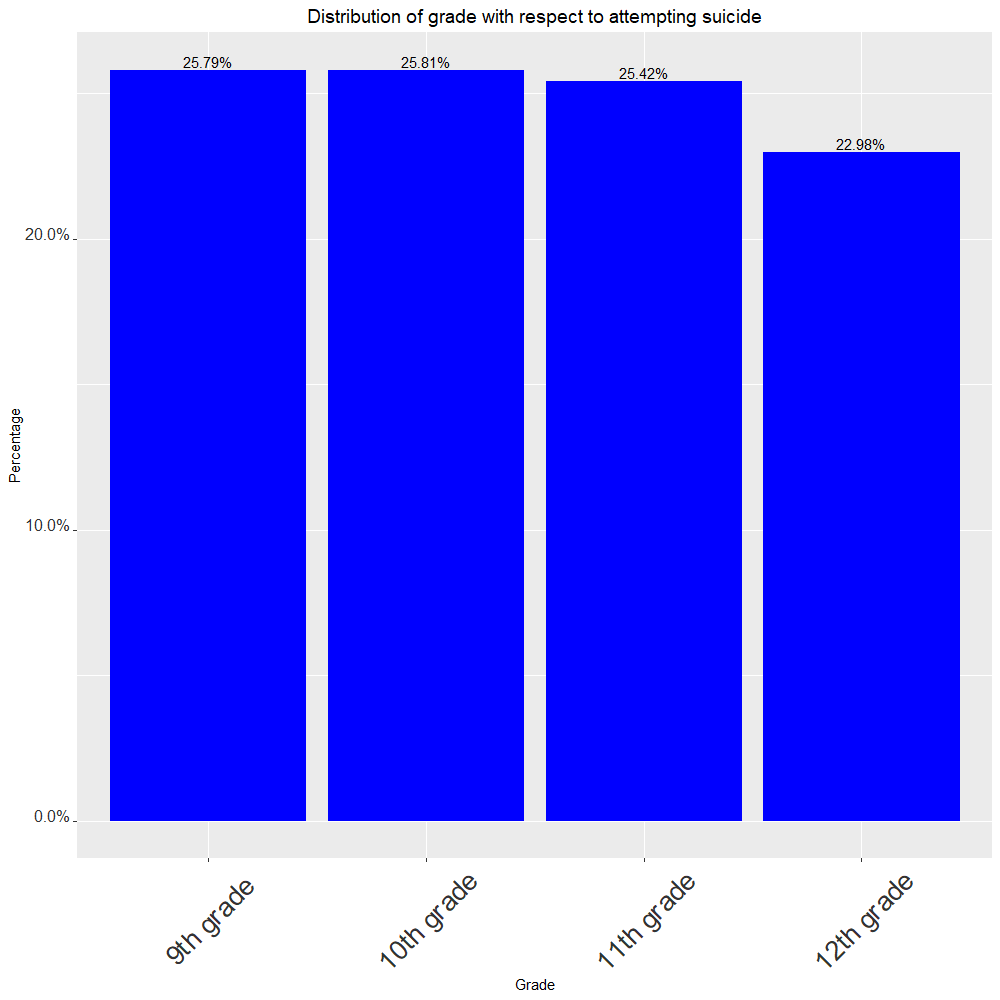


Figure 4

In figure 5, 44% of the youth is white, who had tried attempting suicide, and 29% of the youth were Hispanic/Latino.

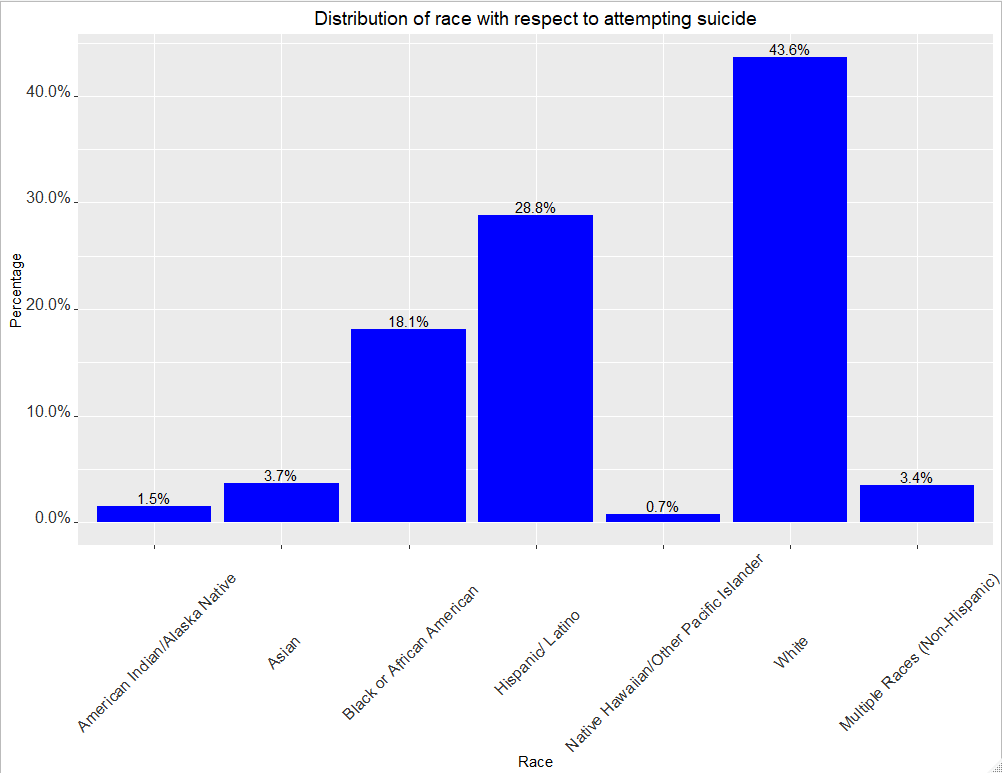


Figure 5

In figure 6, 43% of the youth think they have the right weight and 32% of them think they are slightly overweight.

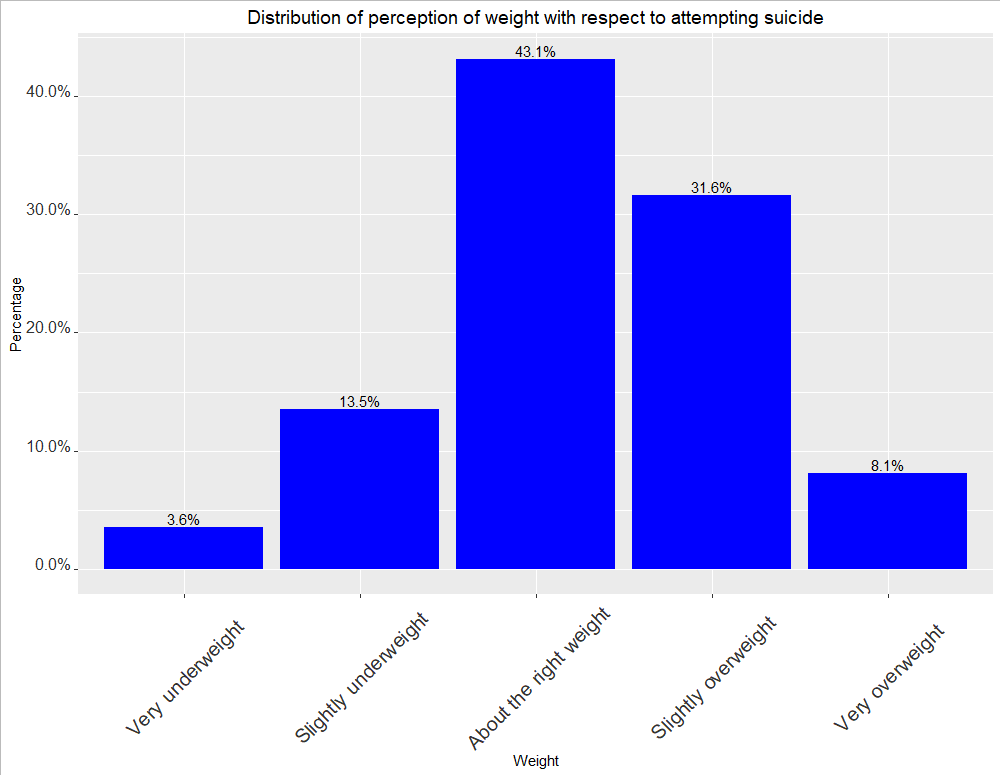


Figure 6

In figure 7, 57% of the youth trying to lose weight who had attempted suicide.

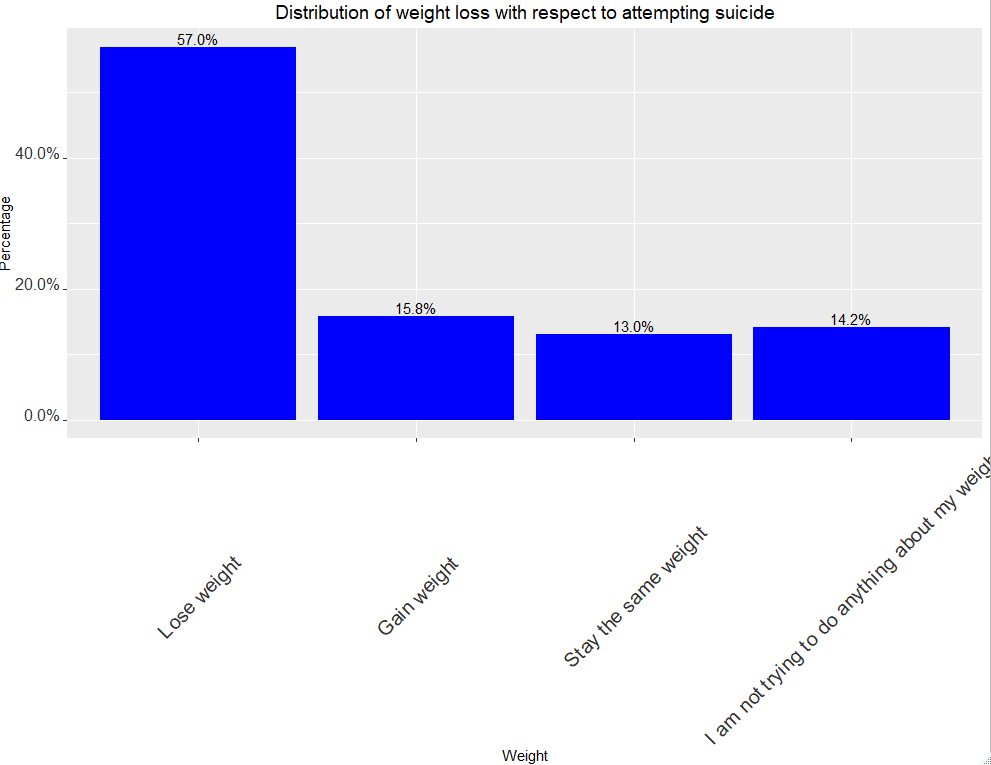


Figure 7

In figure 8, 34% of the youth drank less fruit juice (1 to 3 times per week), and 22% did not drink fruit juice past seven days.

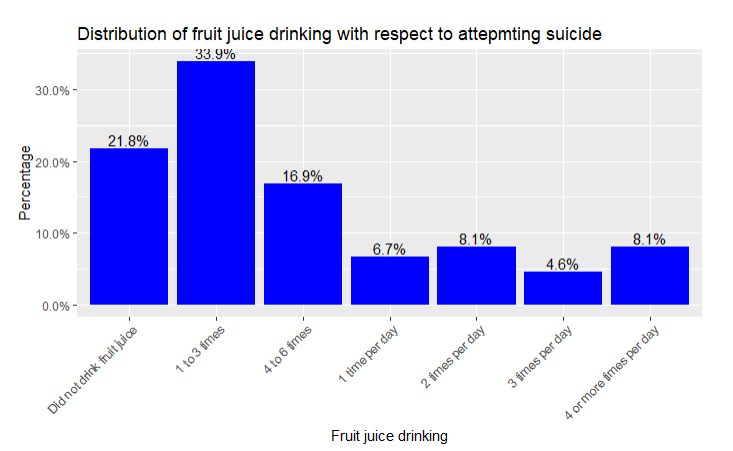


Figure 8

In figure 9, 37% of the youth had 1 to 3 times fruit-eating per week.

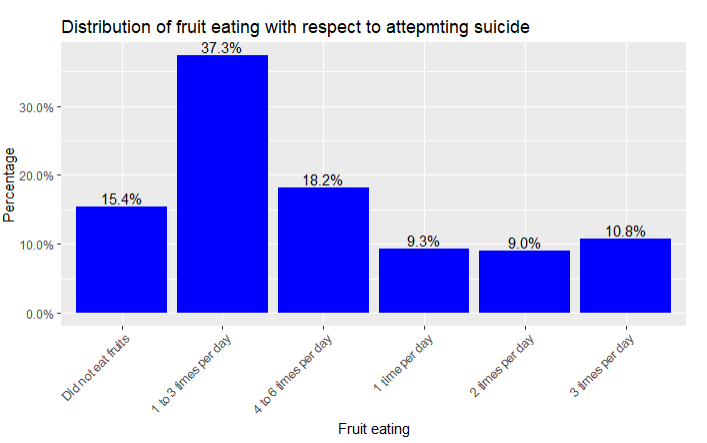


Figure 9

In figure 10, 41% of the youth did not eat salad in the past seven days, who attempted suicide.

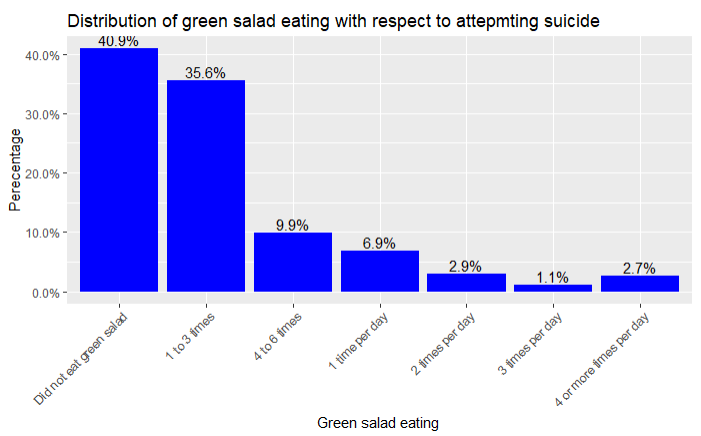


Figure 10

In figure 11, most of the youth who attempted suicide had less soda intake.

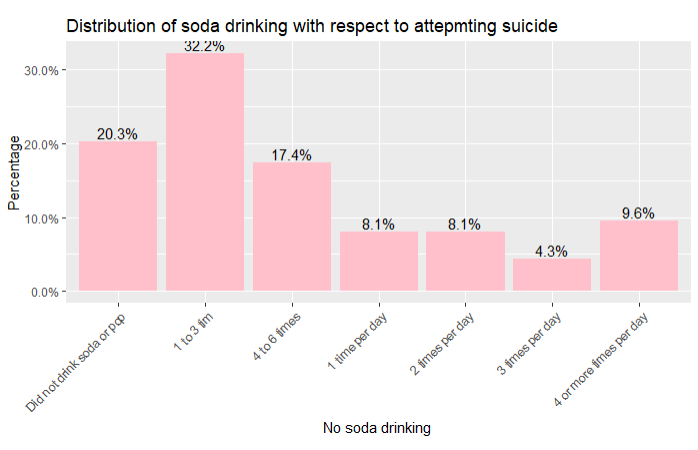


Figure 11

In figure 12, the youth percentage who attempted suicide had less or no milk drinking habit in the past seven days.

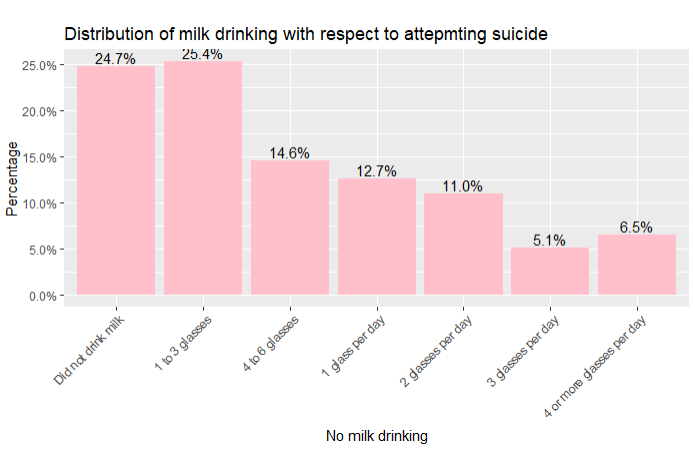


Figure12

In figure 13, 39% of the youth who attempted suicide were bullied in the school.



Figure 13

In figure 14, the youth percentage in alcohol use is less who had attempted suicide.

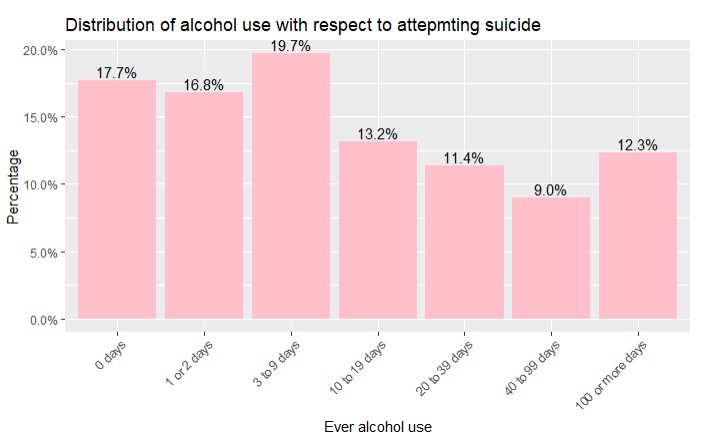


Figure 14

In figure 15, 25% of youth who were heterosexual and who had tried attempting suicide more bullied in school.

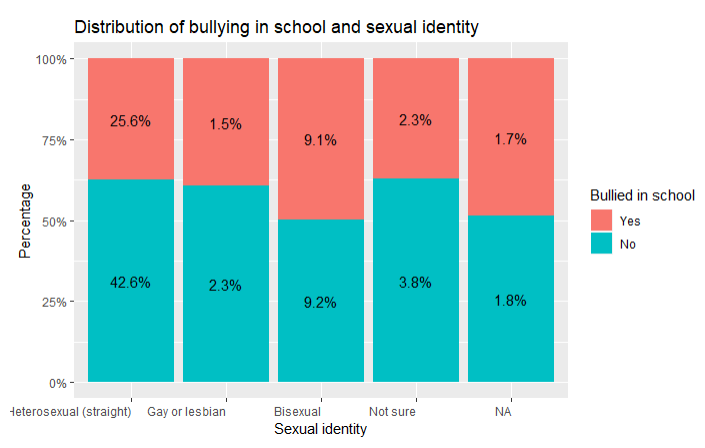


Figure 15

In figure 16, most of the youth who had good grades and attempted suicide also made a suicide plan. While the lower grades did not make a suicide plan.

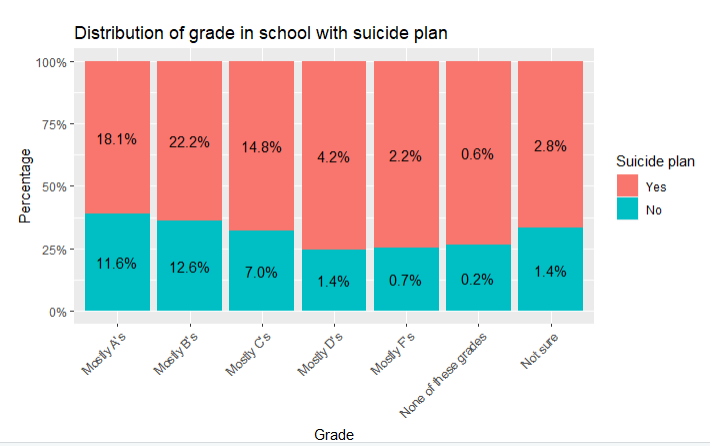


Figure 16

In figure 17, 56% of the youth who attempted suicide also had electronic vapor, which is higher than those who did not have used. It seems that suicide ideation has some relationship with dietary habits. Also, females are attempting to suiciding more than males, which can be due to bullying in school or forced sexual intercourse.

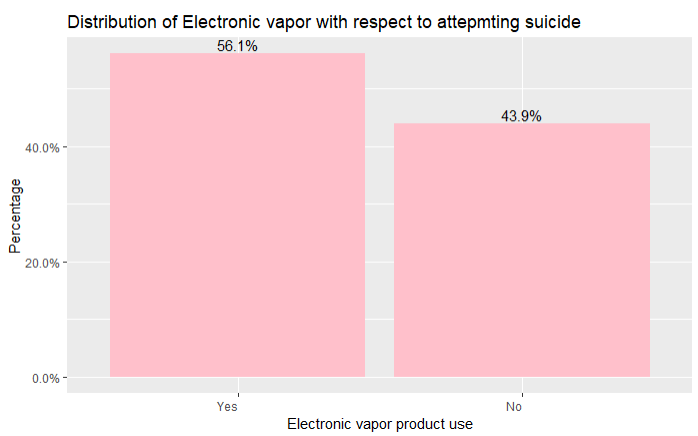


Figure 17

1. Predictive analytics:

The linear regression is used for understanding the relationship between the dependent and independent factors. In this study, we use a combination of statistical tests like multiple linear regression using R and machine learning algorithms in Python. Exploratory data analysis is also used to represent the trends and findings.

* The variables are coded as below for dietary habits, substance use, and alcohol:

A. Did not eat

B. 1 to 3 times

C. 4 to 6 times

D. 1 time per day

E. 2 times per day

F. 3 times per day

G. 4 or more times per day

* Model 1: The first research question analyzed the relationship between suicide attempts and dietary habits. In this model (figure 18), the suicide plan is the dependent variable, and milk drinking, fruit-eating, salad eating, soda drinking, and fruit juice drinking are the independent variables.

In the above result, the suicide plan shows a significant relationship with soda drinking, and they had four or more times per day. Another significant relationship of suicide plan is with the drinking milk, which had only 1 to 3 glasses, 4 to 6 glasses and one glass per day. There is also the very least significant relationship between suicide plan and the fruit juice drinking who had only 1 to 3 times per week. The R squared is 0.0106 which means that there is very little correlation between the data. Another model to test the correlation will be selecting the significant variables. In the below result, the suicide plan shows a strong correlation with dietary habits.

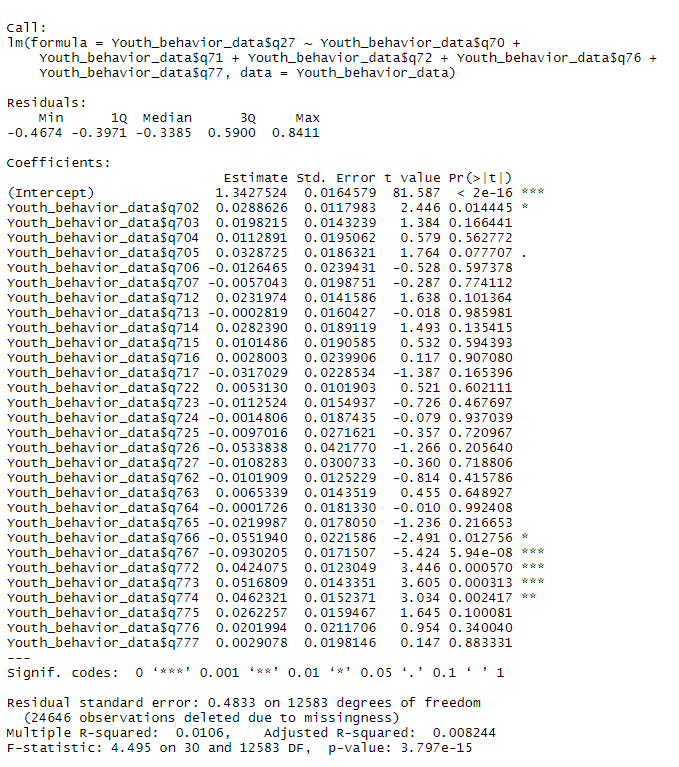


Figure 18

* Model 2: In model 2, the research question was that which gender has involved in suicide attempts?

In the below formula of linear regression, the suicide plan has a strong correlation with gender and age.

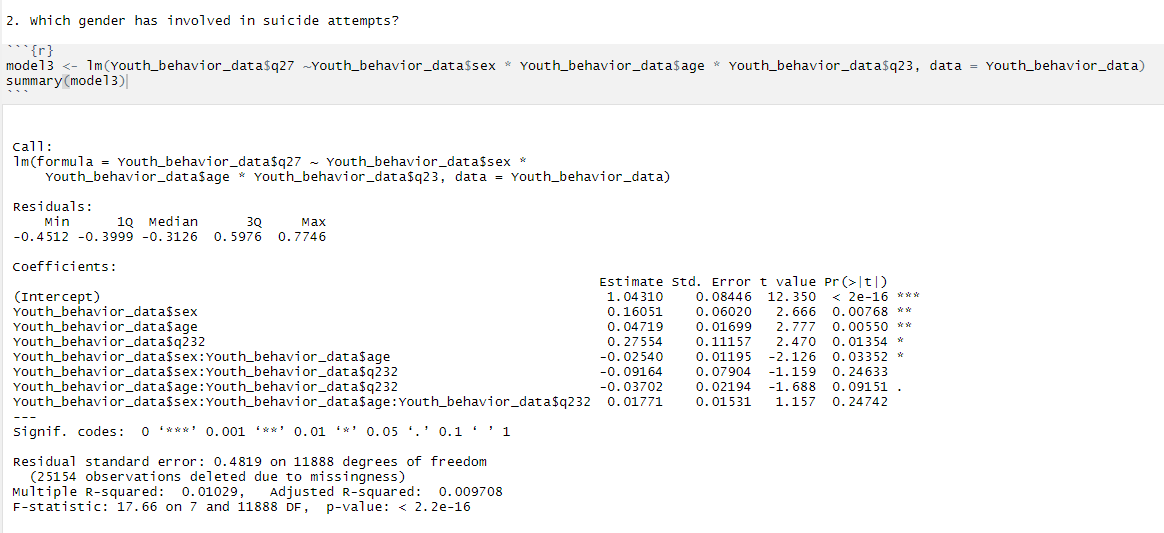


Figure 19

* Model 3: How does exposure to drug abuse and alcohol, among other factors, contribute to higher rates of violence and suicide attempt among the adolescent population in the United States?

Various factors can lead to suicide, such as drug abuse and alcohol use. In the below regression model, the use of cocaine, alcohol, and marijuana used as an independent factor and suicide plan as a dependent factor. The multiple regression in Figure 20 showed that the suicide plan has a strong correlation with youth having alcohol for 100 or more days and had marijuana for 3 to 9 times per week. There is also a moderately significant relationship between suicide plans and youth who had at least one drink of alcohol for 40 to 99 days in whole life. There is also a very least significant relationship between suicide plans, and youth had marijuana 10 to 19 times in the past 30 days. The other question is used in the analysis is how many times youth used any form of cocaine, including powder, crack, or freebase. The variable has some significant relationship with the suicide plan, where the youth had more than 40 or more times in their life.

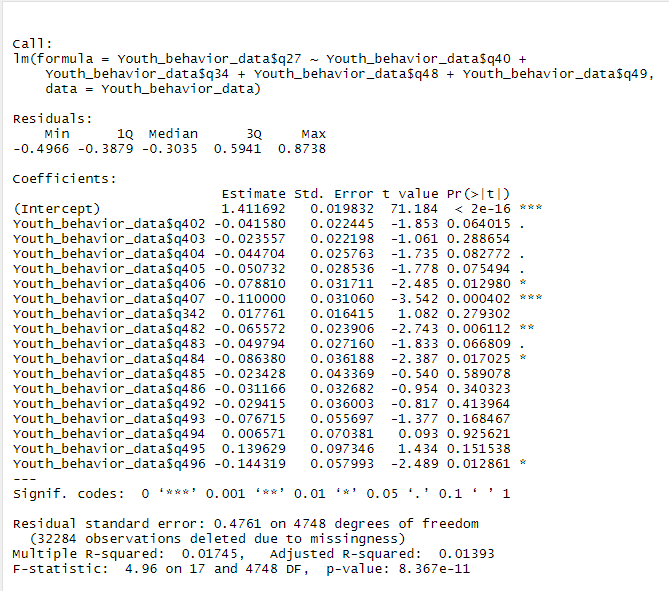
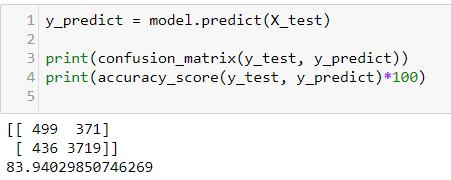


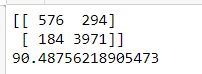
Figure 20

1. Machine learning algorithms: Machine learning algorithms like the random forest, decision tree, and logistic regression are studied to find accuracy in the model. In the first step, data is preprocessed and null values are removed for modeling using dropna() function in python. The target variable as X which is the youth who had attempted suicide in the past year is set. And the rest of the independent variables as Y. In the next step, the data is split by using train\_test\_split function into train and test data with a random state of 52.

* Decision Tree classifier: The decision tree is a supervised machine learning where the data is continuously split according to a specific parameter. The sklearn library used and imported tree function and decision tree classifier. The next model is built using the decision tree classifier. The model is then fit on x train and y train. In the next step, the accuracy is measured using the confusion matrix and accuracy score from sklearn .metrics. The model has an accuracy of 83% which is not good accuracy for predicting the data. The confusion, the matrix, is a summary of prediction resulted from classification problems. The 499 observation is genuinely positive, 3719 is negative.



* Random forest: Random forest fits several decision tree classifiers which improves the accuracy and control overfitting. The random forest classifier function is imported. The model is fit in x train and y train. The accuracy score was 90%, which is better than the decision tree classifier.



* Variable score evaluation: The variable score in Figure 21 evaluates the scores of the variables by using the feature selection library of sklearn. The cocaine use question (q49) has the highest score following by current marijuana use(q47), current alcohol use(q40), and ever alcohol use(q40).

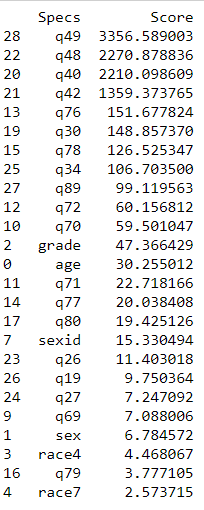


Figure 21

* Feature importance: In figure 22, the feature importance is one of the core concepts in machine learning which hugely impacts the performance of the model. The irrelevant features can negatively impact model performance. The target as y and independent variables are set. The other trees classifier function is imported and feature importances\_ is used as the inbuilt class feature importance of the tree-based classifiers. In figure 17, the most crucial feature the suicide plan made by youth (q27), gender, fruit-eating (q71), no milk drinking (q77), breakfast eating(q78), physical activity>=5(q79), bullying at school(q23), ever alcohol use(q40), television watching (q80). In the steps, these selected features will be selected, and the model will be applied.

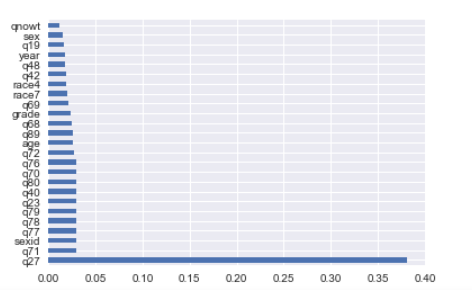


Figure 22

Figure 23 shows the correlation between the factors. There is not much correlation between the suicide attempt with the other factors other than the youth who had made a suicide plan anytime in their mind. The correlation between these two is .60, which is moderately correlated with the suicide attempt by youth.

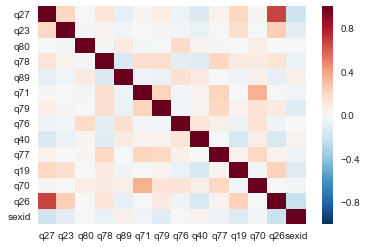
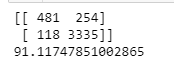
****

Figure 23

* Logistic regression: The logistic regression uses a binary dependent variable to estimate the parameters of the logistic model. The logistic regression is imported from sklearn. Linear model library. The model is fitted on x train and y train, which is then predicted by using the x test. The accuracy of the model is 91% which good accuracy compared to the decision tree and random forest classifier.



SIGNIFICANCE AND CONCLUSION

In the above analysis, youth tends to attempt suicide more when they make a suicide plan. Most of the youth who have attempted suicide are females than males. The suicidal pattern was more correlated with dietary habits. The youth do not have an excellent dietary habit, and most of them are having fewer fruit juices, salads and milk. The analysis also focused on analyzing whether the use of substance and drugs affects the youth. The youth who have made a suicide plan has a robust correlation in youth having more alcohol and had marijuana with youth having alcohol for 100 or more days and had marijuana in moderate quantity. There is also a moderately significant relationship between suicide plans and youth who had at least one drink of alcohol for 40 to 99 days in whole life. There is also a very least significant relationship between suicide plan and youth had marijuana 10 to 19 times in the past 30 days.

The youth between 15 to 17 years of age are more involved in attempting suicide and making suicide plans. The percentage of whites and Hispanic is more than any other race. 25% of the youth who had attempted suicide were also bullied in school.

In the year 2015, there was a rise in suicide attempts by youth. The youth who were most heterosexuals attempted suicide. The above factors can lead to attempting suicide. There may be other factors like family background, family violence, parents' employment can affect youth behavior.

These factors can be taken into serious consideration when understanding youth behaviors. Our study can also be a useful resource of reference for further researches.

REFERENCES

Shopowich, S., Kelly, S., & Mahat, G. (2019). Female Sexual Assault, HIV Testing, and Education: A Secondary Analysis of the 2015 Youth Risk Behavior Survey. Pediatric Nursing, (45(1)th ed., Vols. 31 - 35).

Shopowich, J., & Price, J. H. (2017). Violence Related Behaviors and Weapon Carrying Among Hispanic Adolescents: Results from the National Youth Risk Behavior Survey, 2001–2015. Journal of Community Health, 43(2), 391–399

Centers for Disease Control and Prevention (CDC). 1991-2017 High School Youth Risk Behavior Survey Data. Available at <http://nccd.cdc.gov/youthonline/>. Accessed on [September 21, 2019].

Kann, L., McManus, T., Harris, W. A., Shanklin, S. L., Flint, K. H., Queen, B., … Ethier, K. A. (2018). Youth Risk Behavior Surveillance - United States, 2017. Morbidity And Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002), 67(8), 1–114.

Centers for Disease Control and Prevention. 1991-2017 Youth Risk Behavior Survey Questionnaire. Available at: www.cdc.gov/yrbs. Accessed on [September 28, 2019].

CDC Releases 2017 Youth Risk Behavior Survey (YRBS) Results. (2019, February 21).

Transgender Identity and Experiences of Violence Victimization ... (2019, April 8).

Shield, K. D., Parry, C., & Rehm, J. (n.d.). Publications | National Institute on Alcohol Abuse and Alcoholism | Measuring the Burden � Alcohol�s Evolving Impact on Individuals, Families, and Society.

Preventing Teen Dating Violence. (2019, May 7).

Cash, S. J., & Bridge, J. A. (2009). Epidemiology of youth suicide and suicidal behavior. *Current opinion in pediatrics*, *21*(5), 613–619. doi:10.1097/MOP.0b013e32833063e1

O'Carroll PW, Berman AL, Maris RW, et al. Beyond the Tower of Babel: a nomenclature for suicidology. Suicide Life Threat Behav. 1996;26(3):237–252. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/8897663)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Suicide+Life+Threat+Behav&title=Beyond+the+Tower+of+Babel:+a+nomenclature+for+suicidology&author=PW+O'Carroll&author=AL+Berman&author=RW+Maris&volume=26&issue=3&publication_year=1996&pages=237-252&pmid=8897663&)]

Nutrition, Physical Activity, & Obesity Data & Statistics | Adolescent and School Health | CDC. (2018, November 6).

Burns, R. D. (2019). Energy Balance-Related Factors Associating with Adolescent Weight Loss Intent: Evidence from the 2017 National Youth Risk Behavior Survey. BMC Public Health, 19(1).

Rosario, M., Corliss, H. L., Everett, B. G., Reisner, S. L., Austin, S. B., Buchting, F. O., & Birkett, M. (2014). Sexual Orientation Disparities in Cancer-Related Risk Behaviors of Tobacco, Alcohol, Sexual Behaviors, and Diet and Physical Activity: Pooled Youth Risk Behavior Surveys. American Journal of Public Health, 104(2), 245–254

Baiden, P., Mengo, C., & Small, E. (2019). History of Physical Teen Dating Violence and Its Association With Suicidal Behaviors Among Adolescent High School Students: Results From the 2015 Youth Risk Behavior Survey. Journal Of Interpersonal Violence, 886260519860087.

Kemal, S., Sheehan, K., & Feinglass, J. (2018). Gun carrying among freshmen and sophomores in Chicago, New York City, and Los Angeles public schools: the Youth Risk Behavior Survey, 2007-2013.

Phillips, G., Turner, B., Salamanca, P., Birkett, M., Hatzenbuehler, M. L., Newcomb, M. E., … Mustanski, B. (2017). Victimization as a mediator of alcohol use disparities between sexual minority subgroups and sexual majority youth using the 2015 National Youth Risk Behavior Survey. Drug and Alcohol Dependence, 178, 355-362

Key, K., Byrd, R. J., & La Guardia, A. C. (2016). Adolescent Non-Suicidal Self-Injury: Analysis of the Youth Risk Behavior Survey Trends. The Professional Counselor, 6(1), 61-75.

Hall, C., West, J., & Herbert, P. C. (2015). Pretty Risky Behavior: A Content Analysis of Youth Risk Behaviors in “Pretty Little Liars.” Health Educator, 47(2), 2–11.

APPENDIX

1. CODE

library(rmarkdown)

library(readr) ## after every package installed

library(readxl)

library(plyr)

library(dplyr)

library(ggplot2)

library(leaflet)

library(plotly)

library(gridExtra)

library(knitr)

library(shiny)

library(scales)

library(cluster)

library(tidyverse)

library(odbc)

library(forcats)

library(DBI)

library(wordcloud)

library(tidyverse)

library(Amelia)

## Loading required package: Rcpp

library(fuzzyjoin)

library(stringr)

Youth\_behavior\_data <- read\_excel("C:/Users/prath/Documents/anly 699/Youth\_data.xlsx")

Descriptive data analysis of the variables

summary (Youth\_behavior\_data)

Data preprocessing

Converting numeric data into factors

Coding the variables to plot graphs

## Recoding Youth\_behavior\_data$age into ## Recoding Youth\_behavior\_data$grade into Youth\_behavior\_data$grade\_rec

Youth\_behavior\_data$grade\_rec <- as.character(Youth\_behavior\_data$grade)

Youth\_behavior\_data$grade\_rec <- fct\_recode(Youth\_behavior\_data$grade\_rec,

"11th grade" = "3",

"10th grade" = "2",

"9th grade" = "1",

"12th grade" = "4")

Youth\_behavior\_data$age\_rec <- as.character(Youth\_behavior\_data$age)

Youth\_behavior\_data$age\_rec <- fct\_recode(Youth\_behavior\_data$age\_rec,

"16 years old" = "5",

"12 years old or younger" = "1",

"14 years old" = "3",

"15 years old" = "4",

"17 years old" = "6",

"18 years old or older" = "7",

"13 years old" = "2")

## Recoding Youth\_behavior\_data$sex into Youth\_behavior\_data$sex\_rec

Youth\_behavior\_data$sex\_rec <- as.character(Youth\_behavior\_data$sex)

Youth\_behavior\_data$sex\_rec <- fct\_recode(Youth\_behavior\_data$sex\_rec,

"Male" = "2",

"Female" = "1")

## Recoding Youth\_behavior\_data$race7 into Youth\_behavior\_data$race7\_rec

Youth\_behavior\_data$race7\_rec <- as.character(Youth\_behavior\_data$race7)

Youth\_behavior\_data$race7\_rec <- fct\_recode(Youth\_behavior\_data$race7\_rec,

"White" = "6",

"Black or African American" = "3",

"Asian" = "2",

"Multiple Races (Non-Hispanic)" = "7",

"Hispanic/ Latino" = "4",

"Native Hawaiian/Other Pacific Islander" = "5",

"American Indian/Alaska Native" = "1")

## Recoding Youth\_behavior\_data$qnobese into Youth\_behavior\_data$qnobese\_rec

Youth\_behavior\_data$qnobese\_rec <- as.character(Youth\_behavior\_data$qnobese)

Youth\_behavior\_data$qnobese\_rec <- fct\_recode(Youth\_behavior\_data$qnobese\_rec,

"No" = "2",

"Yes" = "1")

## Recoding Youth\_behavior\_data$qnowt into Youth\_behavior\_data$qnowt\_rec

Youth\_behavior\_data$qnowt\_rec <- as.character(Youth\_behavior\_data$qnowt)

Youth\_behavior\_data$qnowt\_rec <- fct\_recode(Youth\_behavior\_data$qnowt\_rec,

"No" = "2",

"Yes" = "1")

## Recoding Youth\_behavior\_data$sexid into Youth\_behavior\_data$sexid\_rec

Youth\_behavior\_data$sexid\_rec <- as.character(Youth\_behavior\_data$sexid)

Youth\_behavior\_data$sexid\_rec <- fct\_recode(Youth\_behavior\_data$sexid\_rec,

"Heterosexual (straight)" = "1",

"Not sure" = "4",

"Gay or lesbian" = "2",

"Bisexual" = "3")

## Recoding Youth\_behavior\_data$q68 into Youth\_behavior\_data$q68\_rec

Youth\_behavior\_data$q68\_rec <- fct\_recode(Youth\_behavior\_data$q68,

"Slightly underweight" = "2",

"Slightly overweight" = "4",

"About the right weight" = "3",

"Very overweight" = "5",

"Very underweight" = "1")

## Recoding Youth\_behavior\_data$q69 into Youth\_behavior\_data$q69\_rec

Youth\_behavior\_data$q69\_rec <- fct\_recode(Youth\_behavior\_data$q69,

"Gain weight" = "2",

"Lose weight" = "1",

"I am not trying to do anything about my weight" = "4",

"Stay the same weight" = "3")

## Recoding Youth\_behavior\_data$q70 into Youth\_behavior\_data$q70\_rec

Youth\_behavior\_data$q70\_rec <- fct\_recode(Youth\_behavior\_data$q70,

"1 to 3 times" = "2",

"4 to 6 times" = "3",

"Did not drink fruit juice" = "1",

"2 times per day" = "5",

"1 time per day" = "4",

"4 or more times per day" = "7",

"3 times per day" = "6")

## Recoding Youth\_behavior\_data$q71 into Youth\_behavior\_data$q71\_rec

Youth\_behavior\_data$q71\_rec <- fct\_recode(Youth\_behavior\_data$q71,

"4 to 6 times per day" = "3",

"3 times per day" = "7",

"1 to 3 times per day" = "2",

"3 times per day" = "6",

"2 times per day" = "5",

"1 time per day" = "4",

"Did not eat fruits" = "1")

## Recoding Youth\_behavior\_data$q72 into Youth\_behavior\_data$q72\_rec

Youth\_behavior\_data$q72\_rec <- fct\_recode(Youth\_behavior\_data$q72,

"1 to 3 times" = "2",

"Did not eat green salad" = "1",

"3 times per day" = "6",

"4 to 6 times" = "3",

"1 time per day" = "4",

"4 or more times per day" = "7",

"2 times per day" = "5")

## Recoding Youth\_behavior\_data$q76 into Youth\_behavior\_data$q76\_rec

Youth\_behavior\_data$q76\_rec <- fct\_recode(Youth\_behavior\_data$q76,

"4 to 6 times" = "3",

"1 to 3 tim" = "2",

"2 times per day" = "5",

"Did not drink soda or pop" = "1",

"4 or more times per day" = "7",

"3 times per day" = "6",

"1 time per day" = "4")

## Recoding Youth\_behavior\_data$q77 into Youth\_behavior\_data$q77\_rec

Youth\_behavior\_data$q77\_rec <- fct\_recode(Youth\_behavior\_data$q77,

"4 to 6 glasses" = "3",

"4 or more glasses per day" = "7",

"1 to 3 glasses" = "2",

"1 glass per day" = "4",

"3 glasses per day" = "6",

"Did not drink milk" = "1",

"2 glasses per day" = "5")

## Recoding Youth\_behavior\_data$q78 into Youth\_behavior\_data$q78\_rec

Youth\_behavior\_data$q78\_rec <- fct\_recode(Youth\_behavior\_data$q78,

"7 days" = "8",

"4 days" = "5",

"5 days" = "6",

"3 days" = "4",

"0 days" = "1",

"2 days" = "3",

"6 days" = "7",

"1 day" = "2")

## Recoding Youth\_behavior\_data$q78 into Youth\_behavior\_data$q78\_rec

Youth\_behavior\_data$q79\_rec <- fct\_recode(Youth\_behavior\_data$q79,

"7 days" = "8",

"4 days" = "5",

"5 days" = "6",

"3 days" = "4",

"0 days" = "1",

"2 days" = "3",

"6 days" = "7",

"1 day" = "2")

## Recoding Youth\_behavior\_data$q80 into Youth\_behavior\_data$q80\_rec

Youth\_behavior\_data$q80\_rec <- fct\_recode(Youth\_behavior\_data$q80,

"Less than 1 hour per day" = "2",

"2 hours per day" = "4",

"3 hours per day" = "5",

"No TV on average school day" = "1",

"1 hour per day" = "3",

"5 or more hours per day" = "7",

"4 hours per day" = "6")

## Recoding Youth\_behavior\_data$q23 into Youth\_behavior\_data$q23\_rec

Youth\_behavior\_data$q23\_rec <- fct\_recode(Youth\_behavior\_data$q23,

"No" = "2",

"Yes" = "1")

## Recoding Youth\_behavior\_data$q30 into Youth\_behavior\_data$q30\_rec

Youth\_behavior\_data$q30\_rec <- fct\_recode(Youth\_behavior\_data$q30,

"No" = "2",

"Yes" = "1")

## Recoding Youth\_behavior\_data$q40 into Youth\_behavior\_data$q40\_rec

Youth\_behavior\_data$q40\_rec <- fct\_recode(Youth\_behavior\_data$q40,

"0 days" = "1",

"1 or 2 days" = "2",

"3 to 9 days" = "3",

"100 or more days" = "7",

"20 to 39 days" = "5",

"10 to 19 days" = "4",

"40 to 99 days" = "6")

## Recoding Youth\_behavior\_data$q48 into Youth\_behavior\_data$q48\_rec

Youth\_behavior\_data$q48\_rec <- fct\_recode(Youth\_behavior\_data$q48,

"0 times" = "1",

"40 or more times" = "6",

"3 to 9 times" = "3",

"1 or 2 times" = "2",

"20 to 39 times" = "5",

"10 to 19 times" = "4")

## Recoding Youth\_behavior\_data$q26 into Youth\_behavior\_data$q26\_rec

Youth\_behavior\_data$q26\_rec <- fct\_recode(Youth\_behavior\_data$q26,

"No" = "2",

"Yes" = "1")

## Recoding Youth\_behavior\_data$q27 into Youth\_behavior\_data$q27\_rec

Youth\_behavior\_data$q27\_rec <- fct\_recode(Youth\_behavior\_data$q27,

"No" = "2",

"Yes" = "1")

## Recoding Youth\_behavior\_data$q34 into Youth\_behavior\_data$q34\_rec

Youth\_behavior\_data$q34\_rec <- fct\_recode(Youth\_behavior\_data$q34,

"No" = "2",

"Yes" = "1")

## Recoding Youth\_behavior\_data$q19 into Youth\_behavior\_data$q19\_rec

Youth\_behavior\_data$q19\_rec <- fct\_recode(Youth\_behavior\_data$q19,

"No" = "2",

"Yes" = "1")

Youth\_behavior\_data$year <- as.factor(Youth\_behavior\_data$year)

## Recoding Youth\_behavior\_data$q89 into Youth\_behavior\_data$q89\_rec

Youth\_behavior\_data$q89\_rec <- fct\_recode(Youth\_behavior\_data$q89,

"Mostly B's" = "2",

"Mostly C's" = "3",

"Not sure" = "7",

"Mostly D's" = "4",

"Mostly A's" = "1",

"Mostly F's" = "5",

"None of these grades" = "6")

## Recoding Youth\_behavior\_data$q49 into Youth\_behavior\_data$q49\_rec

Youth\_behavior\_data$q49\_rec <- fct\_recode(Youth\_behavior\_data$q49,

"0 times" = "1",

"1 or 2 times" = "2",

"40 or more times" = "6",

"20 to 39 times" = "5",

"3 to 9 times" = "3",

"10 to 19 times" = "4")

Youth\_behavior\_data <- Youth\_behavior\_data %>% filter(q26\_rec == "Yes")

summary(Youth\_behavior\_data)

describe(Youth\_behavior\_data)

Exploratory data analysis

Demographic analysis:

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$sex\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$age\_rec, fill = Youth\_behavior\_data$sex\_rec)) +geom\_bar(position = "fill") + geom\_text(aes(label=scales::percent(..count../sum(..count..))), stat='count',position=position\_fill(vjust = 0.5))+scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("Age") + ylab("Percentage")+ labs(fill="Gender")+ggtitle("Distribution of age and gender")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$sex\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$sex\_rec)) + geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "blue") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) + scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(hjust = 1)) + ylab("Percentage") + xlab("Sex") + ggtitle("Distribution of age with respect to attempting suicide") + theme(plot.title = element\_text(hjust = 0.5, size = 14), # Center title position and size plot.subtitle = element\_text(hjust = 0.5),axis.text.x = element\_text(color = "grey20", size = 20, hjust = .5, vjust = .5, face = "plain"),axis.text.y = element\_text(color = "grey20", size = 12, angle = 0, hjust = 1, vjust = 0, face = "plain")) # Center subtitl

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$grade\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$grade\_rec)) + geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "blue") +geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) + scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + ylab("Percentage") + xlab("Age") + xlab("Grade") + ggtitle("Distribution of grade with respect to attempting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$race7\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$race7\_rec)) + geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "blue") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) + scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + ylab("Percentage") + xlab("Age") + xlab("Race") + ggtitle("Distribution of race with respect to attempting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q68\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q68\_rec)) + geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "blue") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + ylab("Percentage") + xlab("Weight") + ggtitle("Distribution of perception of weight with respect to attempting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q70\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q70\_rec)) + geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "blue") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) + scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("Fruit juice drinking") + ylab("Percentage") + ggtitle("Distribution of fruit juice drinking with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q71\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q71\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "blue") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("Fruit eating") + ylab("Percentage") + ggtitle("Distribution of fruit eating with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q72\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q72\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "blue") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) + scale\_y\_continuous(labels = percent)+ xlab("Green salad eating ") + ylab("Perecentage") + theme(axis.text.x = element\_text(angle = 45, hjust = 1), axis.text.y = element\_text(angle = 0)) + ggtitle("Distribution of green salad eating with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q76\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q76\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "pink") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) + scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("No soda drinking ") + ylab("Percentage") + ggtitle("Distribution of soda drinking with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q77\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q77\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "pink") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) + scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("No milk drinking") + ylab("Percentage") + ggtitle("Distribution of milk drinking with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q23\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q23\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "pink") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(hjust = 1)) + xlab("Bullying at school") + ylab("Percentage") + ggtitle("Distribution of Bullying at school with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q40\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q40\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "pink") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45,hjust = 1)) + xlab("Ever alcohol use") + ylab("Percentage") + ggtitle("Distribution of alcohol use with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q34\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q34\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "pink") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(hjust = 1)) + xlab("Electronic vapor product use") + ylab("Percentage") + ggtitle("Distribution of Electronic vapor with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$sexid\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$sexid\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), fill = "pink") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) + scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(hjust = 1)) + xlab("Sex\_id") + ylab("Percentage") + ggtitle("Distribution of sexual identity with respect to attepmting suicide")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q23\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$sexid\_rec, fill = Youth\_behavior\_data$q23\_rec)) +geom\_bar(position = "fill") + geom\_text(aes(label=scales::percent(..count../sum(..count..))), stat='count',position=position\_fill(vjust = 0.5)) + scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(hjust = 1)) + xlab("Sexual identity") + ylab("Percentage")+ labs(fill="Bullied in school")+ggtitle("Distribution of bullying in school and sexual identity")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q89\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$q89\_rec, fill = Youth\_behavior\_data$q27\_rec)) +geom\_bar(position = "fill") + geom\_text(aes(label=scales::percent(..count../sum(..count..))), stat='count',position=position\_fill(vjust = 0.5))+scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45,hjust = 1)) + xlab("Grade") + ylab("Percentage")+ labs(fill="Suicide playn") + ggtitle("Distribution of grade in school with suicide plan")

Youth\_behavior\_data <- Youth\_behavior\_data[!is.na(Youth\_behavior\_data$q27\_rec), ]

ggplot(Youth\_behavior\_data, aes(x = Youth\_behavior\_data$year, fill = Youth\_behavior\_data$q27\_rec)) +geom\_bar(aes(y = (..count..)/sum(..count..)), position = "dodge") + geom\_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count") +scale\_y\_continuous(labels = percent)+ theme(axis.text.x = element\_text(angle = 45,hjust = 1)) + xlab("Year") + ylab("Percentage") + labs(fill = "Made a suicide plan ") + ggtitle("Percentage of suicide attempting by year")

Models

Find out the youth who had attempted suiciding and correlation between dietary food habits

model3 <- lm(Youth\_behavior\_data$q27 ~Youth\_behavior\_data$q70+Youth\_behavior\_data$q76+Youth\_behavior\_data$q77, data = Youth\_behavior\_data)

summary(model3)

Which gender has involved in suicide attempts?

model3 <- lm(Youth\_behavior\_data$q27 ~Youth\_behavior\_data$sex \* Youth\_behavior\_data$age \* Youth\_behavior\_data$q23, data = Youth\_behavior\_data)

summary(model3)

At which age and academic grade, adolescents have thought or have attempted suicide? Find out the factors that lead females to suicide more than males.

model4 <- lm(Youth\_behavior\_data$q27~Youth\_behavior\_data$grade\*Youth\_behavior\_data$age)

summary(model4)

library(MASS)

chisq.test(Youth\_behavior\_data$q27, Youth\_behavior\_data$sex)

Youth\_analysis <- Youth\_behavior\_data %>% filter(sex\_rec == "Female")

model5 <- glm(Youth\_analysis$sex~Youth\_analysis$grade\*Youth\_analysis$age)

summary(model5)

female\_youth <- Youth\_behavior\_data %>% filter(sex\_rec == "Female")

female\_youth <- Youth\_behavior\_data %>% filter(sex\_rec == "Female")

female\_youth <- female\_youth %>% filter(grade == "3"|grade == "4")

model6 <- lm(female\_youth$q27 ~ female\_youth$q23 \* female\_youth$grade, data = female\_youth)

summary(model6)

How does exposure to drug abuse and alcohol, among other factors, contribute to higher rates of violence and suicide attempt among the adolescent population in the United States?

model8 <- lm(Youth\_behavior\_data$q27~Youth\_behavior\_data$q40+Youth\_behavior\_data$q34+Youth\_behavior\_data$q48+Youth\_behavior\_data$q49, data = Youth\_behavior\_data)

summary(model8)

## Importing data

import os

import pandas as pd

import numpy as np

from sklearn.metrics import r2\_score

from sklearn.metrics import mean\_squared\_error

import matplotlib.pyplot as plt

import seaborn as sns

get\_ipython().magic('matplotlib inline')

from sklearn.model\_selection import train\_test\_split

from sklearn.tree import DecisionTreeRegressor

from sklearn.model\_selection import cross\_val\_score

from sklearn.ensemble import RandomForestRegressor

import statsmodels.api as sm

from sklearn.linear\_model import LinearRegression

# ## Import data using pandas

input\_file = "C:/Users/prath/Documents/anly 699/Youth\_data\_new.csv"

youth\_data = pd.read\_csv(input\_file)

youth\_data.head()

# ## Analyze number of NA values in data

youth\_data.isnull().sum()

# ## Remove NA for applying machine learning algorithms

new = youth\_data.dropna()

new.head(10)

## Export the file and import again with cleaned data.

export\_csv = new.to\_csv (r'C:\Users\prath\Documents\anly 699\export\_dataframe.csv', index = None, header=True)

input\_file = "C:/Users/prath/Documents/anly 699/export\_dataframe.csv"

youth\_data = pd.read\_csv(input\_file)

# ## Set the target varible as q26 which is attempting suicide by youth

target = youth\_data['q26']

target.value\_counts()

# ## Split the data into train and test using train\_test\_split

from sklearn.model\_selection import train\_test\_split

y = target

X = youth\_data.drop(['q26'],axis=1)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.30, random\_state=52)

# ## Import decsion tree classifier for modeling

from sklearn import tree

from sklearn.tree import DecisionTreeClassifier

model = tree.DecisionTreeClassifier()

model = model.fit(X\_train, y\_train)

from IPython.display import SVG

from graphviz import Source

from IPython.display import display

graph = Source(tree.export\_graphviz(model, out\_file=None , feature\_names=X.columns, class\_names=['default', 'no default'] , filled = True))

display(SVG(graph.pipe(format='svg')))

# ## Import library for confusion matrix and accuracy score

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import accuracy\_score

# ## By using predict function on X\_test will help in predicting and measure the accuracy score and confusion matrix

y\_predict = model.predict(X\_test)

print(confusion\_matrix(y\_test, y\_predict))

print(accuracy\_score(y\_test, y\_predict)\*100)

y\_predict[:10]

# ## Random forest

# ## Import random forest classifier for fitting the model on X\_train and y\_train.

# Part 2 Random forest

from sklearn.ensemble import RandomForestClassifier

clf = RandomForestClassifier()

clf.fit(X\_train, y\_train)

y\_predict = clf.predict(X\_test)

print(confusion\_matrix(y\_test, y\_predict))

print(accuracy\_score(y\_test, y\_predict)\*100)

# ## Variable score evaluation

import numpy as np

from sklearn.feature\_selection import SelectKBest

from sklearn.feature\_selection import chi2

X = youth\_data.iloc[:,] #independent columns

y = youth\_data.iloc[:,-1] #target column i.e price range

#apply SelectKBest class to extract top 10 best features

bestfeatures = SelectKBest(score\_func=chi2, k=10)

fit = bestfeatures.fit(X,y)

dfscores = pd.DataFrame(fit.scores\_)

dfcolumns = pd.DataFrame(X.columns)

#concat two dataframes for better visualization

featureScores = pd.concat([dfcolumns,dfscores],axis=1)

featureScores.columns = ['Specs','Score'] #naming the dataframe columns

print(featureScores.nlargest(25,'Score')) #print 10 best features

target = youth\_data['q26']

y = target

X = youth\_data.drop(['q26'],axis=1) #target column i.e price range

from sklearn.ensemble import ExtraTreesClassifier

import matplotlib.pyplot as plt

model = ExtraTreesClassifier()

model.fit(X,y)

print(model.feature\_importances\_) #use inbuilt class feature\_importances of tree based classifiers

#plot graph of feature importances for better visualization

feat\_importances = pd.Series(model.feature\_importances\_, index=X.columns)

feat\_importances.nlargest(12).plot(kind='barh')

plt.show()

continuous\_varaible = ['q27','q23','q80','q78','q89','q71','q79','q76','q40','q77','q19','q70','q26','sexid']

##Correlation analysis for continuous variables

#Correlation plot

youth\_data\_corr = youth\_data.loc[:,continuous\_varaible]

#Setting the width of the plot

f, ax = plt.subplots(figsize=(10, 10))

#Generate correlation matrix

correlation = youth\_data\_corr.corr()

#Plot using seaborn library

## Correlation analysis

get\_ipython().magic('matplotlib inline')

sns.heatmap(youth\_data\_corr[continuous\_varaible].corr())

youth\_data\_corr[continuous\_varaible].corr()

# ## Correlation between factor

f, ax = plt.subplots(figsize=(10, 10))

#Generate correlation matrix

correlation = youth\_data.corr()

#Plot using seaborn library

## Correlation analysis

youth\_data = youth\_data.loc[:,]

get\_ipython().magic('matplotlib inline')

sns.heatmap(youth\_data.corr())

corr = youth\_data.corr()

print(corr)

# Create correlation matrix

corr\_matrix = corr.corr().abs()

# Select upper triangle of correlation matrix

upper = corr\_matrix.where(np.triu(np.ones(corr\_matrix.shape), k=1).astype(np.bool))

# Find index of feature columns with correlation greater than 0.60

to\_drop = [column for column in upper.columns if any(upper[column] > 0.6)]

Xnew = corr.drop(to\_drop, axis=1)

import pandas as pd

import numpy as np

from sklearn import preprocessing

import matplotlib.pyplot as plt

plt.rc("font", size=14)

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

import seaborn as sns

sns.set(style="white")

sns.set(style="whitegrid", color\_codes=True)

sns.heatmap(youth\_data,yticklabels=False,cbar=False)

youth\_data.drop('year',axis=1,inplace=True)

youth\_data.info()

from sklearn.linear\_model import LogisticRegression

logmodel = LogisticRegression()

logmodel.fit(X\_train,y\_train)

predictions = logmodel.predict(X\_test)

from sklearn.metrics import classification\_report

print(classification\_report(y\_test,predictions))

print("Accuracy:",metrics.accuracy\_score(y\_test, predictions))

# import the class

from sklearn.linear\_model import LogisticRegression

# instantiate the model (using the default parameters)

logreg = LogisticRegression()

# fit the model with data

logreg.fit(X\_train,y\_train)

y\_pred=logreg.predict(X\_test)

# import the metrics class

from sklearn import metrics

cnf\_matrix = metrics.confusion\_matrix(y\_test, y\_pred)

cnf\_matrix

print(confusion\_matrix(y\_test, y\_pred))

print(accuracy\_score(y\_test, y\_pred)\*100)

# import required modules

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

get\_ipython().magic('matplotlib inline')

class\_names=[0,1] # name of classes

fig, ax = plt.subplots()

tick\_marks = np.arange(len(class\_names))

plt.xticks(tick\_marks, class\_names)

plt.yticks(tick\_marks, class\_names)

# create heatmap

sns.heatmap(pd.DataFrame(cnf\_matrix), annot=True, cmap="YlGnBu" ,fmt='g')

ax.xaxis.set\_label\_position("top")

plt.tight\_layout()

plt.title('Confusion matrix', y=1.1)

plt.ylabel('Actual label')

plt.xlabel('Predicted label')

y\_pred