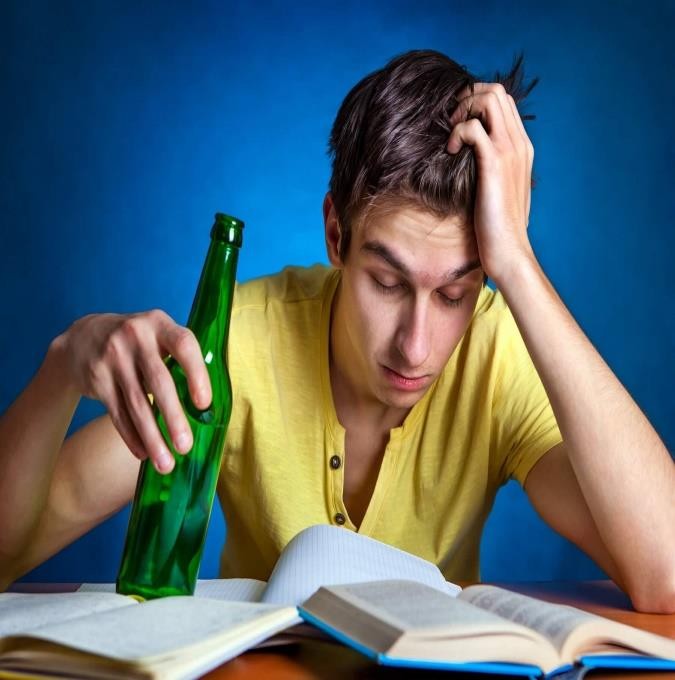
**STUDENT**

**ALCOHOL**

**CONSUMPTION**

**ST. CLAIR COLLEGE, MISSISSAUGA**

|  |  |
| --- | --- |
| **Presenters:** |  |
| Udit Arora | 0769036 |
| Manjit Singh | 0765235 |
| Komal Arora | 0763537 |
| Diksha Dalia | 0766406 |
| Sapna | 0765232 |
| **Supervisor -** | **Savita Seharawat** |

**Abstract:**

We found this dataset very amazing. We downloaded it from Kaggle.

The reason behind using this data set is finding the reason behind the consumption of alcohol among youth because in this present era students become addict of narcotic substance easily, because of the surrounding, modernization and the company of their friends. Our focus is to focus over the effect on the performance of students due to alcohol in grades the course subject, Math.

Using the attributes and performing the functions and other codes help us to find out the consumption of alcohol made by students. While performing on the datasets and finding the solutions to some questions

**Research Questions -**

**Does alcohol effect success i.e., Grades and what factors contribute in that?**

This data will help us to find that what are the factors that affect the grades of students those could be the education level of parents, how often they absence from school, travel time and study time.

**Which gender, age group consume more alcohol?**

This data will help us to find that which age group and gender are addicted more towards the alcohol.

**Which school has better at grade performance among both schools?**

From the research of that questions, we will come to a point that which school have better grades from GP and MS.

**INTRODUCTION** –

**Literature Review**

Education is the key to success, and success is usually measured in terms of student academic performance. But today, many students are addicted to alcohol, which affects their grades. P. Cortez and A. Silva. Use data mining to predict secondary school grades. A. Brito and J. Teixeira ed., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008), Portugal, Porto, April 2008, EUROSIS, ISBN 9789077381397. Alcohol is a major cause of morbidity and mortality, and heavy drinking is especially common in the United Kingdom. In 2017, 29% of men and 26% of women in the UK reported drinking, and 10% of adults reported drinking more than 5 days a week. More than 1 million hospitalizations (7% of all hospitalizations) and 24,208 deaths in the United Kingdom were due to the use of alcohol . This puts a heavy burden on medical care and costs an estimated £ 3.5 billion annually to the United Kingdom National Health Service (NHS) . However, the patterns of alcohol use and the associated harm are complex and vary widely by geographic region and duration. group. Public health surveillance is important for monitoring and understanding the epidemiology of health- related outcomes, informing about policies and resource allocation, and assessing the impact of interventions. However, there is often a delay between real-time health-related events and the availability of data about those events. Nowcasting is defined as predicting the recent past, present, and near future and has been used to correct such delays in data availability. In contrast, predictions are important in assessing long-term trends in education and attendance, health, and what students do in their leisure time.

**Description:** Dataset contains 33 columns about students which explains data about their grades, performance, how often they are using internet, their attendance record and family background like their parents are living together or apart and other factors related to their personal life.

**Attributes in the dataset:**

|  |  |  |
| --- | --- | --- |
| S. No | VAR\_NAME | VAR\_DESCRIPTION |
| 1 | School | student's school ('GP' or 'MS’) |
| 2 | Sex | student's sex ('F' - female or 'M' – male) |
| 3 | Age | student's age (15 to 22) |
| 4 | Address | student's home address type ('U' - urban or 'R' – rural) |
| 5 | Famsizeh | family size ('LE3' - less or equal to 3 or 'GT3' - greater than 3) |
| 6 | Pstatus | parent’s cohabitation status (‘T’ – living together or ‘A’ – apart) |
| 7 | Medu | mother’s education (numeric: 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)s |
| 8 | Fedu | father’s education (numeric: 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) |
| 9 | Mjob | mother’s job (nominal: ‘teacher’, ‘health’ care related, civil ‘services’ (e.g., administrative or police), ‘at home’ or ‘other’) |
| 10 | Fjob | father’s job (nominal: ‘teacher’, ‘health’ care related, civil ‘services’ (e.g., administrative or police), ‘at home’ or ‘other’) |
| 11 | Reason | reason to choose this school (nominal: close to ‘home’, school ‘reputation’, ‘course’ preference or ‘other’) |
| 12 | Guardian | student’s guardian (nominal: ‘mother’, ‘father' or 'other') |
| 13 | Traveltime | home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) |
| 14 | Studytime | weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| 15 | Failures | number of past class failures |
| 16 | Schoolsup | extra educational support (binary: yes or no) |
| 17 | Famsup | family educational support (binary: yes or no) |
| 18 | Paid | extra paid classes within the course subject (binary: yes or no) |
| 19 | Activities | extra-curricular activities (binary: yes or no) |
| 20 | Nursery | attended nursery school (binary: yes or no) |
| 21 | Higher | wants to take higher education (binary: yes or no) |
| 22 | Internet | Internet access at home (binary: yes or no) |
| 23 | Romantic | with a romantic relationship (binary: yes or no) |
| 24 | Famrel | quality of family relationships (numeric: from 1 - very bad to 5 - excellent) |
| 25 | Freetime | free time after school (numeric: from 1 - very low to 5 - very high) |
| 26 | Gout | going out with friends (numeric: from 1 - very low to 5 - very high) |
| 27 | Dalc | workday alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| 28 | Walc | weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| 29 | Health | current health status (numeric: from 1 - very bad to 5 - very good) |
| 30 | Absences | number of school absences (numeric: from 0 to 93) |
| 31 | G1 | first period grade (numeric: from 0 to 20) |
| 32 | G2 | second period grade (numeric: from 0 to 20) |
| 33 | G3 | final grade (numeric: from 0 to 20, output target) |

**Target attribute**: Grade (G3)

The numerical variables are – age, absences, G1, G2, G3

The categorical variables are – school, sex, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, health, alc.

**METHODOLOGY**

**Number of rows and columns:**

|  |  |  |
| --- | --- | --- |
| S. No | NUMBER OF ROWS | NUMBER OF COLUMNS |
| DF1 | 395 | 33 |
| DF2 | 649 | 33 |
| DF | 1044 | 33 |

**Number of NA values**:

|  |  |
| --- | --- |
| Total missing values | 0 |

**Summary of Descriptive Statistics**

The structure of the data is to calculate basic statistics, such as the min, max, mean, and median, and missing value counts. Here we use summary command to check those statistics.

The following shows basic statistics of each feature:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | age | Medu | Fedu | traveltime | studytime | failures | famrel |
| count | 1044 | 1044 | 1044 | 1044 | 1044 | 1044 | 1044 |
| mean | 16.72605 | 2.603448 | 2.387931 | 1.522989 | 1.970307 | 0.264368 | 3.935824 |
| std | 1.239975 | 1.124907 | 1.099938 | 0.731727 | 0.834353 | 0.656142 | 0.933401 |
| min | 15 | 0 | 0 | 1 | 1 | 0 | 1 |
| 25% | 16 | 2 | 1 | 1 | 1 | 0 | 4 |
| 50% | 17 | 3 | 2 | 1 | 2 | 0 | 4 |
| 75% | 18 | 4 | 3 | 2 | 2 | 0 | 5 |
| max | 22 | 4 | 4 | 4 | 4 | 3 | 5 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
| count | 1044 | 1044 | 1044 | 1044 | 1044 | 1044 | 1044 | 1044 | 1044 |
| mean | 3.201149 | 3.15613 | 1.494253 | 2.284483 | 3.543103 | 4.434866 | 11.2136 | 11.24617 | 11.34195 |
| std | 1.031507 | 1.152575 | 0.911714 | 1.285105 | 1.424703 | 6.210017 | 2.983394 | 3.285071 | 3.864796 |
| min | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 25% | 3 | 2 | 1 | 1 | 3 | 0 | 9 | 9 | 10 |
| 50% | 3 | 3 | 1 | 2 | 4 | 2 | 11 | 11 | 11 |
| 75% | 4 | 4 | 2 | 3 | 5 | 6 | 13 | 13 | 14 |
| max | 5 | 5 | 5 | 5 | 5 | 75 | 19 | 19 | 20 |

**Outliers**: Outliers are the unusual values or the data points which are significantly far from the other data points.

**REMOVE OUTLIERS**

The boxplot with the outliers. The round black circles inside the boxplot are the outliers.

Chart, box and whisker chart

Description automatically generated

The boxplot without outliers.

Chart, box and whisker chart

Description automatically generated

**Number of Records After Removing Outliers:**

After removing the outliers from the selected attributes these are the number of records that are left in the data set.

|  |  |
| --- | --- |
| No. of Records | **971** |

**Distribution of Categorical attributes:**

**1& 2 About the mother and father education:**

This Graph represent the 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

Chart, bar chart

Description automatically generated

The mother education histogram given above represent the axis, Where X axis shows the level of education, where Y axis shows the count, where majority of mothers lies within level 4 with the higher education.

The correspond graph against the mother education represent the father education where axis represent basically the level of father’s education but in this graph majority of father edu lies between level 2- 2.5

**3-Time Taken to travel from home to school**

This Graph Elucidateshome to school travel time ( 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

Chart

Description automatically generated

Travel time histogram represent the time taken to travel to school i.e 15mint to 1 hours which is represented in levels of 1,2,3,4 in the data set , where less than 600 students travel 15 mints to come to school and not more than 50 students take 1 hour to reach school.

**4&5- Time taken to study and freetime:**

Weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

According to study time histogram representation most of the student’s study for 2 to 5 hours and very less no. of student’s study for more than 10 hours.

Free time after school (from 1 - very low to 5 - very high)

In the free time histogram, it displays the data about the free time after studies. Near about 380 students get average free time after studies

Chart

Description automatically generated

.

Chart, bar chart

Description automatically generated**6- Workday alcohol consumption** ( from 1 - very low to 5 - very high)

**7- Weekend and Weekdays alcohol consumption** ( from 1 - very low to 5 - very high)

Comparision: On weekdays and weekends, majority of students drink very less alcohol. Very few number of students drink more alcohol on weekdays and weekends.

Chart, bar chart, box and whisker chart

Description automatically generated

**Factors Effecting on Final Grades:**

Now let’s have a look on graphs:

**Consumption of Alcohol in weekdays corresponds to G3**

Chart, line chart

Description automatically generatedThis line graph is describing about the students who drink alcohol in weekdays how that alcohol consumption is affecting their grades. It can be clearly seen that the students who’s alcohol consumption level is 4.0 they are scoring very less marks as compare to the students who drink a very few amount of alcohol during weekdays, as it is showing the students with the level of 1.0 they are scoring 20.0 grades in their G3.

**Consumption of Alcohol in Weekends corresponds to Final grades:**

This line graph is showing same trend as the previous one, but it is about the students who drink alcohol during weekends how that affects their grades. The students scoring 2.5 and 5.0 have same level of drinking as compare other students.Highr marks achieved by the students who drink 1.0 level of alcohol.

Chart

Description automatically generated

**G3 Vs Absence**

Chart

Description automatically generatedIt shows fluctuation in this graph between G3 and absence of the students from the class.intially, when the number of absences is almost 0 graders were 11 but after that it start fluctuating till the count of absences is 24 and when it reaches 25 it was the time when students score 7 grades which was least. As the absents from the class increased then the grades started decresed.

**Travel time VS G3**

This line graph is describing about the time taken by the students to reach home and how how that affect their studies. Because the students who just took 15 to 30 mints to reach home they are scoring good final grades as compare to the other students to whom it took more than 3 to 4 hour to reach home.

Chart, line chart

Description automatically generated

Chart

Description automatically generated**Study Time VS Final Grades:**

In this graph we can clearly seen that as the study time increase the grades in final terms are also increasing. In simple words, the more time they are spending on study the more final grades they are getting.

**Comparison between 2 school on the basis of their students final grades:**

From below graph we can seen that GP students got more marks in term 3 as compared to other school

Chart, bar chart, histogram

Description automatically generated

**Comparison between Gender On the behalf of Weekdays Alcohol:**

This graph represents the information about consumption of alcohol by male or female. In ordinary words who drink more alcohol on weekdays.

Chart

Description automatically generated

**Mother job with family support and Father job with family support**

Chart, bar chart

Description automatically generated

Here we plot 2 bar graphs one for mother job with family support and another one is father job with family support to show that how much a student receives family support from parents’ side.

**Correlation of numerical values:**

For numeric data, correlations are important to help determine that which attribute can be deleted from the dataset because if two variables are highly correlated, they are providing same information.

In this case, we see that the grades G1 and G2 are highly correlated, meaning the higher the grades in one session, the higher the grades in another session.

Now we must check the values of standard deviation of G1 and G2 attributes. The attribute with low standard deviation we can delete that one.

Chart

Description automatically generated with low confidence

**FORWARD ELIMINATION METHOD**

In this method ,we will check which are the 3 features that show a significant change in the model.

|  |  |
| --- | --- |
| S.No | Selected Attribute |
| 1 | G1 |
| 2 | G2 |
| 3 | Age |

So here we can seen that G1, G2 and Age are those three features which show a significant change in model .Further we will see the accuracy for these 3 best features.

|  |  |
| --- | --- |
| Accurracy | 0.9033245569405237 |

**SPLITTING THE DATASET:**

|  |  |
| --- | --- |
| SPLITTING SETS | PERCENTAGE |
| Test set | 40% |
| Train set | 60% |

Here we Split the data in train and test set. 40% data in test set and 60% data in the train set. And set random state 40 so every time our sample remain same.

**Accuracy train set and test set:**

Here, we will check the training and testing accuracy of linear regression.

|  |  |
| --- | --- |
| Accuracy of training set | 0.8993543899900719 |
| Accuracy of testing set | 0.904818304689573 |

**Linear Model –** G3 is considered as a target variable in this linear model which is the final grades among all the grades whereas G1,G2 and Age are independents variables. After that, finding the coefficients of independents variables. We also check the accuracy of model.

|  |  |
| --- | --- |
| Intercept | -0.854251255353077 |
| 1.Coefficient | 0.16403260169023456 |
| 2. Coefficient | 0.8103398104653706 |
| 3. Coefficient | 0.09473139508300513 |
| Accuracy | 0.904818304689573 |

**Error checking in G3**

Table

Description automatically generated

Above table have 3 columns- grades, predicted values and error. G3\_grades are actual grades of students and predicted values are prediction of grade3 and the last one is error which we got.

**LOGISTIC REGRESSION –**

Here, used the same target and independent variable. In logistic regression we split dataset 40% for testing and 60% for training. If the value of target variable is less than 8 then the students are failing and considered as 0. On the other hand, if the value is greater than 8 then the students are pass and considered as 1. And same with the G1 and G2 in independent variables. And with age variable under 17 these are underage of drinking alcohol. If they drink so we consider 0 and above 17 considered as 1. After that we split the dataset in train and test set and check the accuracy our model.

|  |  |
| --- | --- |
| Accuracy of model | 0.9484536082474226 |

**Confusion Matrix:**

The confusion matrix is used to assess the performance based on actual values and predicted values.

Chart, treemap chart

Description automatically generated

**Decision tree classifier:**

Diagram

Description automatically generatedNow let’s look at decision tree. Here, first step is basically our root node. Root node is node which start the graph. After that it split the data into two branches either true or false. If its less than 1.5 then it would be true otherwise it would be false. We can also call these branches as intermediate node which means final node are still pending. At the end of decision trees all are leaf node which means the final nodes of tree. In our decision tree we have three features which are weekdays alcohol and weekends alcohol.

**Accuracy of decision tree**

|  |  |
| --- | --- |
| Accuracy | 0.36246786632390743 |

**RANDOM FOREST REGRESSION**

Random forest regression always provide higher accuracy and we can use this for classification as well as regression tsask**.**

|  |  |
| --- | --- |
| **Accuracy** | 0.8607474665389289 |

**Accuracies of models:**

|  |  |
| --- | --- |
| Model’s name | Accuracy |
| Linear Model | 0.904818304689573 |
| Logistic Model | 0.9484536082474226 |
| Decision Tree | 0.36246786632390743 |
| Random Forest | 0.8607474665389289 |



**Conclusion: -**

All in all, in this project, we found that females drink more alcohol as compared to male students. We have dataset about two schools and after analysis it give us GP school learners consume large amount of alcohol rather than other school. Additionally, this dataset is all about the consumption of alcohol along with the factors effecting the grades of students. We also find the correlation and we found that G2 and G3 are strongly positively correlated. Then we performed four model in our project and got the accuracy of every model. According to our dataset logistic regression give more accuracy. So logistic regression is suitable for our project.

**References:**

1. World Health Organization (WHO) *Global Status Report on Alcohol and Health 2014*. Geneva: WHO; 2014, pp. 1–392. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Global+Status+Report+on+Alcohol+and+Health+2014&publication_year=2014&)]

2. Gakidou E., Afshin A., Abajobir A. A., Abate K. H., Abbafati C., Abbas K. M. et al Global, regional, and national comparative risk assessment of 84 behavioral, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 2017; 390: 1345–1422. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5614451/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/28919119)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Lancet&title=Global,+regional,+and+national+comparative+risk+assessment+of+84+behavioural,+environmental+and+occupational,+and+metabolic+risks+or+clusters+of+risks,+1990%E2%80%932016:+a+systematic+analysis+for+the+Global+Burden+of+Disease+Study+2016&volume=390&publication_year=2017&pages=1345-1422&pmid=28919119&)]

3. Office for National Statistics (ONS). Adult Drinking Habits in Great Britain: 2017. London: ONS; 2018.

4. National Health Service (NHS) Digital. Statistics on Alcohol, England, 2018. Leeds: Health and Social Care Information Centre; 2018.

5. Public Health England (PHE). Local Alcohol Profiles for England. London: PHE; 2018.

6. Home Office. Impact Assessment on a Minimum Unit Price for Alcohol. London: Home Office; Institute for Public Policy Research (IPPR) 2012. Available at [https://www.ippr.org/files/images/media/files/publication/2013/06/moving‐on‐older‐people\_June2013\_10898.pdf](https://www.ippr.org/files/images/media/files/publication/2013/06/moving-on-older-people_June2013_10898.pdf) (accessed 2 August 2019). [[Google Scholar](https://scholar.google.com/scholar?q=+Home+Office+.++Impact+Assessment+on+a+Minimum+Unit+Price+for+Alcohol+.++London:+Home+Office+;++Institute+for+Public+Policy+Research+(IPPR)++2012++Available+at++https://www.ippr.org/files/images/media/files/publication/2013/06/moving%E2%80%90on%E2%80%90older%E2%80%90people_June2013_10898.pdf++(accessed+2+August+2019).+)]

7. World Health Organization (WHO). WHO Guidelines on Ethical Issues in Public Health Surveillance. Geneva: WHO; 2017.

8. Bańbura M., Giannone D., Modugno M., Reichlin L. Chapter 4: Nowcasting and the real‐time data flow In: Elliott and Timmermann *Handbook of Economic Forecasting*, Vol. 2, Part A, 2013, pp 195–237 Elsevier B.V. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Handbook+of+Economic+Forecasting&publication_year=2013&)]

Dataset : <https://www.kaggle.com/uciml/student-alcohol-consumption>

Github : <https://github.com/KomalAroraKA/Capstone-Project-Student-Alcohol-Consumption-/upload>