Using Machine Learning To Predict Secondary School Student Alcohol Consumption

Parth Patel
Computer Science, NYU Tandon School of Engineering
Brooklyn, U.S.A.

Ishan Handa
Computer Science, NYU Tandon School of Engineering
Brooklyn, U.S.A.

In this project, we use Machine Learning to predict secondary school student alcohol consumption levels based on an individual student's demographic statistics. We hope to identify and predict the impact of student alcohol consumption and addiction on academic performance.

I. Introduction

For our project, we propose to predict secondary school student alcohol consumption levels based on an individual student's demographic statistics. Thus, in our case, the data instance would be a single student and their associated descriptive statistics. We envision the eventual model to potentially be deployed either annually or semi-annually using demographic and survey data from all secondary school students in the associated schools to classify them per their risk of developing unhealthy drinking habits.

As a data mining problem, this would likely fall into category of classification, with the target variable that we intend to predict using the dataset being a binning of the average alcohol consumption. The data has two fields: "Workday alcohol consumption" and "Weekend alcohol consumption" and we think that merging the two fields into a single indicator of alcohol consumption variable would be more useful as a target variable. The merging can be done by taking a weighted average of the two original fields (As workday consumption is more likely to be related to a problem we would need to provide it more weight against weekend consumption). There is currently a total of 31 features in the dataset, though more could be engineered if necessary.

Finally, on completion the model could be very valuable in providing insights about the alcohol consumption of various students and, similar to a business problem targeting churn, it could help schools optimize the use of their limited resources to provide interventions and counseling for those students who have a higher alcohol consumption than a particular threshold and are most at risk of developing dangerous alcohol consumption habits. It would also allow school administrators to target individuals who display low to medium alcohol consumption habits with seminars on alcohol abuse prevention to prevent the problems from developing.

II. Motivation

Alcohol use among college students is often discouraged and advised against. This study is an effort to understand a measured impact of alcohol consumption amongst students.

Prevalence of alcohol consumption amongst students: [1]

- *Prevalence of Drinking*: In 2013, 59.4 percent of full-time college students' ages 18–22 drank alcohol in the past month compared with 50.6 percent of other persons of the same age. [2]
- *Prevalence of Binge Drinking*: In 2013, 39 percent of college students ages 18–22 engaged in binge drinking (5 or more drinks on an occasion) in the past month compared with 33.4 percent of other persons of the same age. [3]
- *Prevalence of Heavy Drinking*: In 2013, 12.7 percent of college students ages 18–22 engaged in heavy drinking (5 or more drinks on an occasion on 5 or more occasions per month) in the past month compared with 9.3 percent of other persons of the same age. [4]

Consequences of alcohol consumption amongst students: [5]

- 1,825 college students between the ages of 18 and 24 die from alcohol-related unintentional injuries, including motor vehicle crashes. [6]
- Another student who has been drinking assaults 696,000 students between the ages of 18 and 24. [7]
- Roughly 20 percent of college students meet the criteria for an AUD. [8]
- About 1 in 4 college students report academic

consequences from drinking, including missing class, falling behind in class, doing poorly on exams or papers, and receiving lower grades overall. [9]

III. RELATED WORK

Using Data Mining to Predict Secondary School Student Alcohol Consumption - Fabio Pagnotta, Hossain Mohammad Amran, Department of Computer Science, University of Camerino [10]

IV. DATA

We use a data set about Portuguese student which was composed by Paulo Cortez and Alice Silva, University of Minho, Portugal. In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system. This study will consider data collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal. Hence, the database was built from two sources:

- school reports, based on paper sheets and including few attributes
- questionnaires, used to complement the previous information

It's a collection of features likely to be predictors of the target variable are "absences", "health", "gout (going out with friends"), "famrel (quality of family relationships"), "study time", "Father education", "Mother Education" etc.

Attribute Information:
Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

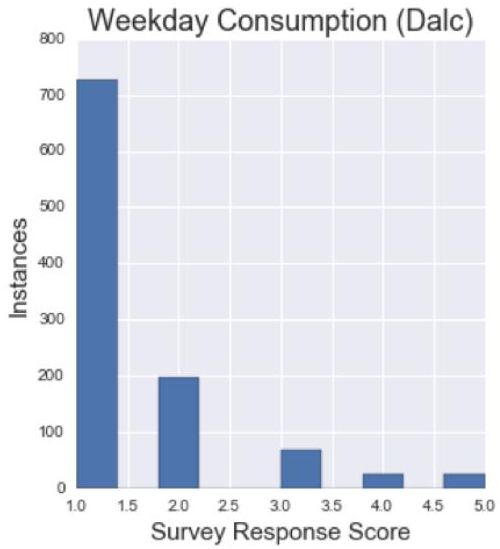
1. **school** - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. **sex** - student's sex (binary: 'F' - female or 'M' - male)
3. **age** - student's age (numeric: from 15 to 22)
4. **address** - student's home address type (binary: 'U' - urban or 'R' - rural)
5. **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. **Pstatus** - parent's cohabitation status (binary: '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)
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 (numeric: n if 1 <= n < 3, else 4)
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 (Math or Portuguese) (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. **goout** - 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)

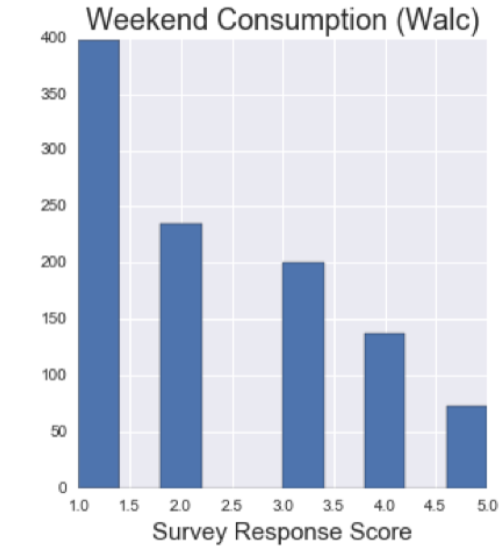
These grades are related with the course subject,

Math or Portuguese:

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)

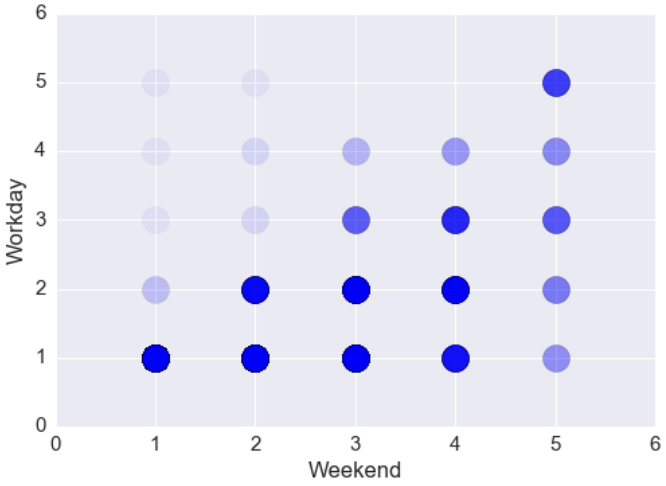
Below are the histogram plots of both the target variables Dalc and Walc.





V. ALGORITHM(S) USED

Thus, far we have performed data preparation and taken several different approaches towards generating a baseline model. We started off by creating a scatter plot, to create discrete classes by combining the workday and weekend alcohol consumption columns and get one single discrete target variable. The outcome of that scatter plot wasn't found to be that impressive as seen from the plot attached below because the raw target variables have discrete numerical values.



For the data preparation, we have taken our Student Alcohol Consumption data set from the UCI machine learning repository and used binning to convert all the categorical features into binary 0-1 features with the help of get_dummies of pandas library. For example, the sex attribute which takes 'M' or 'F' as values, will be replaced by 2 columns, sex_M and sex_F. These new columns will be binary, and will have 1 if the student is male or female respectively and 0 otherwise. In lieu of using a weighted average of the reported workday and weekend alcohol consumption (on a scale of 1-5) as the target variable, we created bins corresponding to a reported workday alcohol consumption above 1 (some alcohol consumption during the school week) and a reported weekend alcohol consumption above 3 (roughly the top 20-25% of respondents) as potential target variables.

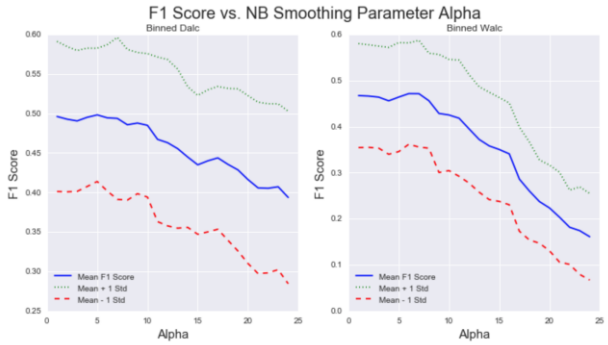
1. NAÏVE BAYES:

We started with a Naïve Bayes to come up with a baseline model. Using the BeranulliNB from sklearn we came up with the following metric scores:

1. Dalc Accuracy: 0.650717703349
2. Dalc Recall: 0.327868852459
3. Dalc Precision: 0.384615384615
4. Dalc F Score: **0.353982300885**
5. Walc Accuracy: 0.77033492823
6. Walc Recall: 0.148936170213
7. Walc Precision: 0.466666666667
8. Walc F Score : **0.225806451613**

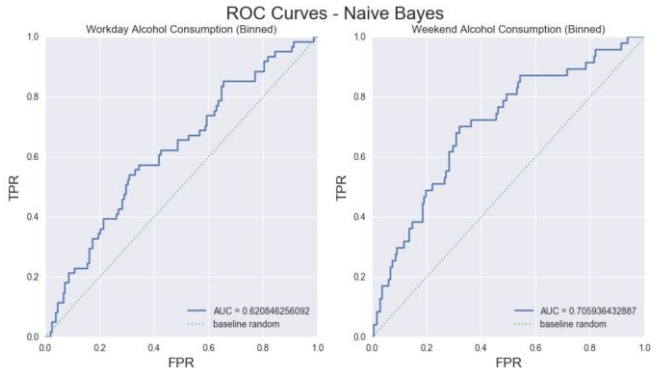
These scores were calculated using the default parameters for the sklearn algorithm.

To further tune the model, we varied the alpha parameter for the BeranulliNB function and plotted the variation in F scores.



It was seen that the maximum value for the F scores is achieved when alpha=5 for Dalc and alpha=7 for Walc.

We used these alpha values to plot train the model again and plotted ROC for the test. We achieved better scores with these parameters:



1. Dalc Accuracy: 0.617224880383
2. Dalc Recall: 0.573770491803
3. Dalc Precision: 0.393258426966
4. Dalc F Score: **0.466666666667**
5. Walc Accuracy: 0.684210526316
6. Walc Recall: 0.702127659574
7. Walc Precision: 0.388235294118
8. Walc F Score : **0.5**

2. EVALUATION METRICS

- Precision $$\text{Precision} = \frac{tp}{tp + fp}$$
- Recall $$\text{Recall} = \frac{tp}{tp + fn}$$
- Accuracy $$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$
- F score $$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Accuracy was rejected as a metric due to its inherent equal emphasis on the maximization of true positives and true negatives. As treatment would only be applied to a subset of the positive cases in the implementation of the model, greater focus should be put upon true positives, which could be tackled using either precision or recall. Maximizing precision, which is the total number of instances correctly classified as positive out of all positive classifications of the model, would be beneficial by minimizing the amount of resources wasted by applying treatments (in the form of counseling and seminars) to individuals who were unlikely to be in any danger of excessive drinking. Maximizing recall on the other hand, which is the total number of correct positive classifications out of all instances that were actually positive, would ensure that as many individuals who were truly in danger of excessive drinking as possible received the treatment that they need. F1 score was chose as the final parameter to evaluate our models on.

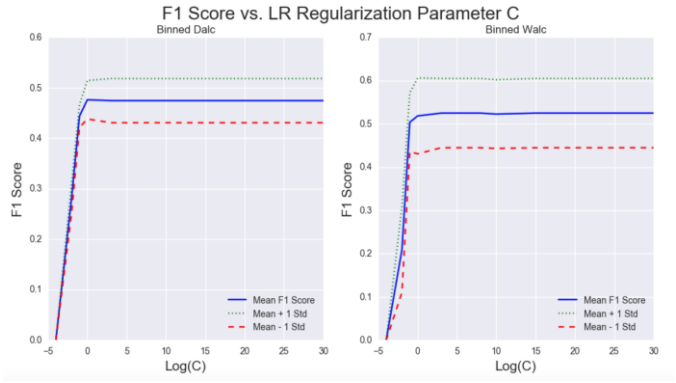
3. LOGISTIC REGRESSION

We used logistic regression from sklearn with default values and were able to generate F score values of

Dalc F score:0.5161290

Walc F Score: 0.48

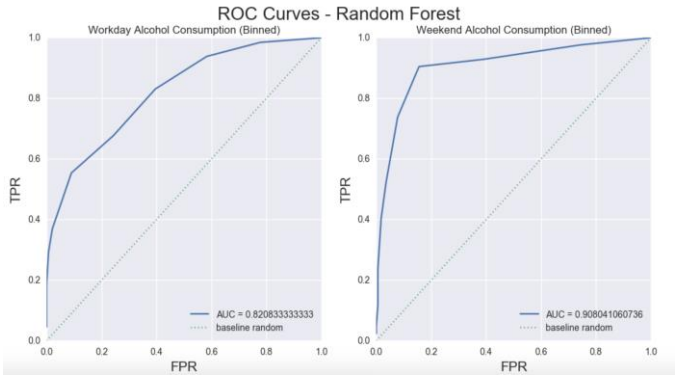
Next, we tried to vary the regularization parameters to fine tune the model but no improvements were noticed.



.

4. RANDOM FORESTS

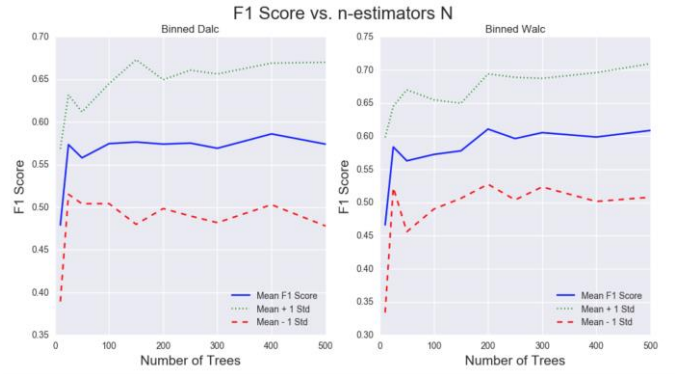
We implemented a Random Forest classifier with default hyper parameters from the sklearn library. For that we got the F score values as below:

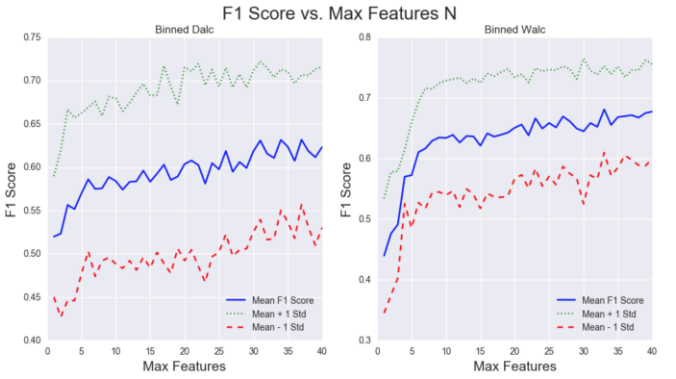


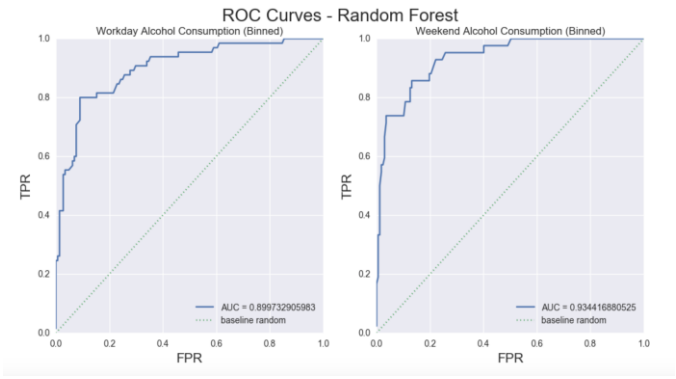
Dalc F score: 0.52173

Walc F Score: 0.54838

Next, we tried to find tune the model by varying the n_estimators and max_feature hyper parameters. We found the optimum values to be 400 and 37 repectively for Dalc target variable/. For Walc these values were 200 and 33 respectively.







.

Dalc F score: 0.67857

Walc F Score: 0.7123

VI. RESULT

As our goal is classification of instances into 0-1 bins for both target variables, possible evaluation metrics we considered were precision, recall and F-score. Accuracy was rejected as a metric due to its inherent equal emphasis on the maximization of true positives and true negatives. As treatment, would only be applied to a subset of the positive cases in the implementation of the model, greater focus should be put upon true positives, which could be tackled using either precision or recall. Maximizing precision, which is the total number of instances correctly classified as positive out of all positive classifications of the model, would be beneficial by minimizing the amount of resources wasted by applying treatments (in the form of counseling and seminars) to individuals who were unlikely to be in any danger of excessive drinking. Maximizing recall on the other hand, which is the total number of correct positive classifications out of all instances that were actually positive, would ensure that as many individuals who were truly in danger of excessive drinking as possible received the treatment that they need. As doing social good and responsibly managing a finite budget are both pressing concerns for secondary school administrations, we felt that the F1-score, which is the harmonic mean of

precision and recall, would be an appropriate metric to balance the two objectives.

	F scores	
	Weekday (Walc)	Weekend (Dalc)
Naive Bayes	0.353982300885	0.225806451613
Naive Bayes tuned	0.466666666667	0.5
Logistic Regression	0.516129032258	0.48
Random Forrest	0.521739130435	0.548387096774
Random Forrest tuned	0.678571428571	0.712328767123

VII. Code

https://github.com/iamparth/machine-learning.git

VIII. Video Link

HTTPS://YOUTU.BE/QVVDSDF1W5G

IX. Evaluation

In the data preparation stage, we tried to find clusters between the two target variables (Dalc & Walc) which weren't that useful because of which we had to take a back step and change our idea about the target variable.

We had also found some interesting data sets which could be merged with this one to find more interesting patterns.

X. Conclusion

As elaborated above, our Data Mining models are useful to improve the quality of education and enhance school resource management. School can identify students with their possibilities of having drinking problems to make best use of the resources to help them. The model also applies to other potential harmful situations, for instance, tobacco, illegal drugs, or suicide-related or violent behaviors. As the target variables change, not only school but other organizations for adults can also deploy the model to predict which cases are more possible to suffer from problems above. Since the datasets involve too much personal information, the model must be monitored and executed by authorities. The Data Mining methods are much more advanced than the original method of universal treatment due to less true negative cases. It does have the risk of false positive cases, who is likely to have or have already had drinking issues but the model fails to reveal. One possible solution is to use the model as a support to the schoolwide treatment: while making sure all students are educated, use our model to select those who need more attention.

From the above table, we conclude that the tuned Random Forest model gave us the best prediction values, with F Scores of 0.678571 and 0.712328 for Walc & Dalc respectively.

XI. References

1.https://www.collegedrinkingprevention.gov/statistics/Default.aspx

2. SAMHSA. 2013 National Survey on Drug Use and Health (NSDUH). Table 6.88B—Alcohol Use in the Past Month among Persons Aged 18 to 22, by College Enrollment Status and Demographic Characteristics: Percentages, 2012 and 2013. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabsPDFWHTML2013/Web/HTML/NSDUH-DetTabsSect6peTabs55to107-2013.htm#tab6.88b

3. SAMHSA. 2013 National Survey on Drug Use and Health (NSDUH). Table 6.89B—Binge Alcohol Use in the Past Mo nth among Persons Aged 18 to 22, by College Enrollment Status and Demographic Characteristics: Percentages, 2012 and 2013. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabsPDFWHTML2013/Web/HTML/NSDUH-DetTabsSect6peTabs55to107-2013.htm#tab6.89b

4. SAMHSA. 2013 National Survey on Drug Use and Health (NSDUH). Table 6.90B—Heavy Alcohol Use in the Past Month among Persons Aged 18 to 22, by College Enrollment Status and Demographic Characteristics: Percentages, 2012 and 2013. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabsPDFWHTML2013/Web/HTML/NSDUH-DetTabsSect6peTabs55to107-2013.htm#tab6.90b

5.https://www.collegedrinkingprevention.gov/statistics/consequences.aspx

6. Hingson, R.W.; Zha, W.; and Weitzman, E.R. Magnitude of and trends in alcohol-related mortality and morbidity among U.S. college students ages 18–24, 1998–2005. Journal of Studies on Alcohol and Drugs (Suppl. 16):12–20, 2009. PMID:19538908 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701090/

7. Hingson R, Heeren T, Winter M. et al. Magnitude of alcohol-related mortality and morbidity among U.S. college students ages 18–24: changes from 1998 to 2001. Annual Review of Public Health 26: 259–279, 2005. PMID: 15760289 http://www.ncbi.nlm.nih.gov/pubmed/15760289

8. Blanco, C.; Okuda, M.; Wright, C. et al. Mental health of college students and their non-college- attending peers: Results from the National Epidemiologic Study on Alcohol and Related Conditions. Archives of General Psychiatry 65(12):1429–1437, 2008. PMID: 19047530 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2734947/

9. Wechsler, H.; Dowdall, G.W.; Maenner, G.; et al. Changes in binge drinking and related problems among American

college students between 1993 and 1997: Results of the Harvard School of Public Health College Alcohol Study. Journal of American College Health 47(2):57–68, 1998. PMID: 9782661 http://www.tandfonline.com/doi/pdf/10.1080/07448489809595621

10.https://www.researchgate.net/publication/296695247_USING_DATA_MINING_TO_PREDICT_SECONDARY_SCHOOL_STUDENT_ALCOHOL_CONSUMPTION

11.http://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION