Using KDD and CRISP-DM data mining process to Analyse Student Alcohol Consumption

Affects of Alcohol Consumption in Academic’s

Navjot Singh Virk: x13112406

Software Development, 4th Year BSc. Honors in Computing

National College of Ireland

Dublin, Ireland

[Virksaabnavjot@gmail.com](mailto:Virksaabnavjot@gmail.com)

In this project, for analysis purpose I have used 2 datasets of Portuguese student on

2 different courses Mathematics and Portuguese Language available at <https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION>

*Abstract*

The legal age for alcohol consumption is 18 years in most countries but teenagers find ways for consumption of this fluid and sometimes this leads in frustration and negative and uncontrolled behavior in our world future generation the teenagers. For this analysis the 2 datasets used are about Portuguese students in secondary school for two different courses (Mathematics and Portuguese language). The purpose of the analysis is to identify student alcohol consumption, the factors contributing to it and show the results how alcohol may affect the students in different ways and long term results of it and the potential risks. The chosen datasets contain student performance and other attributes that will help study and find out how alcohol consumption affects student grades, behavior like their relationship with the family and other other aspects of academic life. For, the project I had the choice to select any 2 datasets. The motivation for selecting the Alcohol consumption datasets was the results of this analysis can help understand student problems that may be leading to more consumption of this fluid and possibly suggest ways to fix them and help the society. The following analysis has been done using some aspects of KDD process and CRISP-DM process model for data mining.

# Introduction

The value of education is high in everyone’s life and a well educated person contributes to the society and live a happy life and school play a major role in our lives it’s the place where everyone starts learning new things.

These days’ alcohol has become a highly regarded party and enjoyment drink and people in social gatherings crave alcohol and people that don’t drink it feel left alone and its kind of considered a taboo for people not drinking when they go out with friends. This kind of thinking have had bad impact on the today’s teenagers as drinking these days is highly regarded as cool and no drinking not so cool. Which motivates student’s teenagers to start consumption of alcohol at very early age as every student in secondary school wants to be fun and popular and no one wants to be left alone and also to not follow the whole group of people and stand for your decision takes a lot of self control on nerves which only few adults can successfully do hence the young people of the world these days are exposed to a mindset and thinking of “alcohol is life” that is regarded fun and aesthetic lifestyle but in reality its not so good.

Most countries around the world have age restrictions on alcohol consumption but young people find ways to work around. Early, exposure to alcohol in many cases leads to people becoming addicted to this fluid for life. Alcohol, not only costs money it costs life, in the past alcohol has been linked to suicide, self harm, loss of memory and many more devastating effects. It affects mood and its consumption makes a person less conscious and lead to many unwanted mistakes and people may face for life. Alcohol also contributes to bad decision making and young people taking bad steps, sexual frustration and unwanted pregnancy at very early age the time which should be consumed for education and building future.

Alcohol consumption has bad affect on students academically, like failing exams, bad behavior and relationship with their parents and health wise alcohol is deadly for young people as their bodies are delicate compared to adults and consumption of alcohol may affect their health and weaken their bodies and reduce immunity and may cause a lot of diseases and mental health problems.

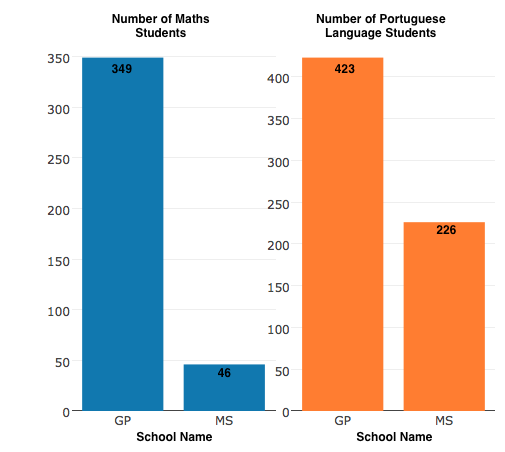
For the analysis, the chosen datasets are about Portuguese students in two schools Gabriel Pereira School and Mousinho da Silveira School and about two courses Mathematics and Portuguese language course which was composed by P. Cortez and A. Silva, at University of Minho in Portugal.

# Data

There are two datasets used in the analysis are below are some quick pointers that will help understand this report and datasets –

Mathematics course dataset contains 393 rows (student data) and 33 columns (attributes) and out of 393 student data 208 females and 187 male students & GP school has 349 students in math’s course and MS has 46 students in its math’s course.

Portuguese course dataset contains 649 student records and same no. of attributes as the other course out of which are 383 females and 266 are male students & GP school has 423 students in math’s course and MS has 226 students in its math’s course.



|  |  |
| --- | --- |
| Attribute | Description |
| school | Students School (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira) |
| sex | Student’s Sex (binary: "F" - female or "M" - male) |
| age | Student’s Age (numeric: from 15 to 22) |
| address | Student's home address type (binary: "U" - urban or "R" - rural) |
| famsize | family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3) |
| Pstatus | parent's cohabitation status (binary: "T" - living together or "A" - apart) |
| Medu | mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) |
| Fedu | Father’s education (numeric: 0 – 4) |
| Mjob | mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at home" or "other") |
| Fjob | father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at home" or "other") |
| reason | reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other") |
| guardian | student's guardian (nominal: "mother", "father" or "other") |
| 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) |
| studytime | weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| failures | number of past class failures (numeric: n if 1<=n<3, else 4) |
| schoolsup | extra educational support (binary: yes or no) |
| famsup | family educational support (binary: yes or no) |
| paid | extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) |
| activities | extra-curricular activities (binary: yes or no) |
| nursery | attended nursery school (binary: yes or no) |
| higher | wants to take higher education (binary: yes or no) |
| internet | Internet access at home (binary: yes or no) |
| romantic | with a romantic relationship (binary: yes or no) |
| famrel | quality of family relationships (numeric: from 1 - very bad to 5 - excellent) |
| freetime | free time after school (numeric: from 1 - very low to 5 - very high) |
| goout | going out with friends (numeric: from 1 - very low to 5 - very high) |
| Dalc | workday alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| Walc | weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| health | current health status (numeric: from 1 - very bad to 5 - very good) |
| absences | number of school absences (numeric: from 0 to 93) |
| G1 | first period grade (numeric: from 0 to 20) |
| G2 | second period grade (numeric: from 0 to 20) |
| G3 | final grade (numeric: from 0 to 20) |

And some analysis has been done by combining both the dataset and adding one new attribute called “subject” (course) in the combined version of the dataset and set it to numeric (1 – Math’s and 2 – Portuguese language course).

# Methodology

The two main methodologies that have been applied in this analysis are KDD (Process for useful knowledge extraction from data) and CRISP-DM (a process model for data mining) and their traces can be found in the whole analysis.

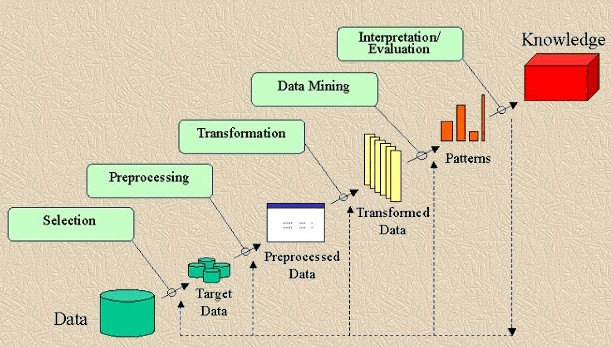


Figure: Steps of KDD Process

Image Source: <https://goo.gl/3msNdV>

The above figure shows the steps involved in KDD for extracting useful information and this analysis is a good example of KDD implementation. Following the process of KDD and little description and how they were implemented in this analysis (Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34):

1. Developing and Understanding of prior knowledge.

Thinking what I want to achieve out of this analysis and if it could be beneficial to the society and read papers on KDD and CRISP-DM for Data mining.

1. Creating a target dataset: includes selecting dataset on which the discovery is to be made.

Through research selected the alcohol consumption dataset it was not easy as there was a requirement for two datasets and most found datasets were single datasets.

1. Data Cleaning and Data Processing: removing any unnecessary data.

The found datasets that I worked with were very well formatted already as some analysis had been performed on these datasets but my analysis finds more additional information on top of it and small data cleaning and processing were conducted.

1. Data Transformation: transforming data into forms appropriate for analysis.

Used the csv file and imported the data in RStudio to start working with it.

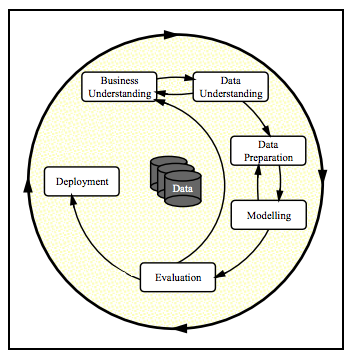
1. Data Mining: apply intelligent method and extract data patterns.

Applied various methods to draw plot with data patterns.

1. Interpreting the mined data through mining process.
2. Knowledge Representation

(as you will see in Results section)

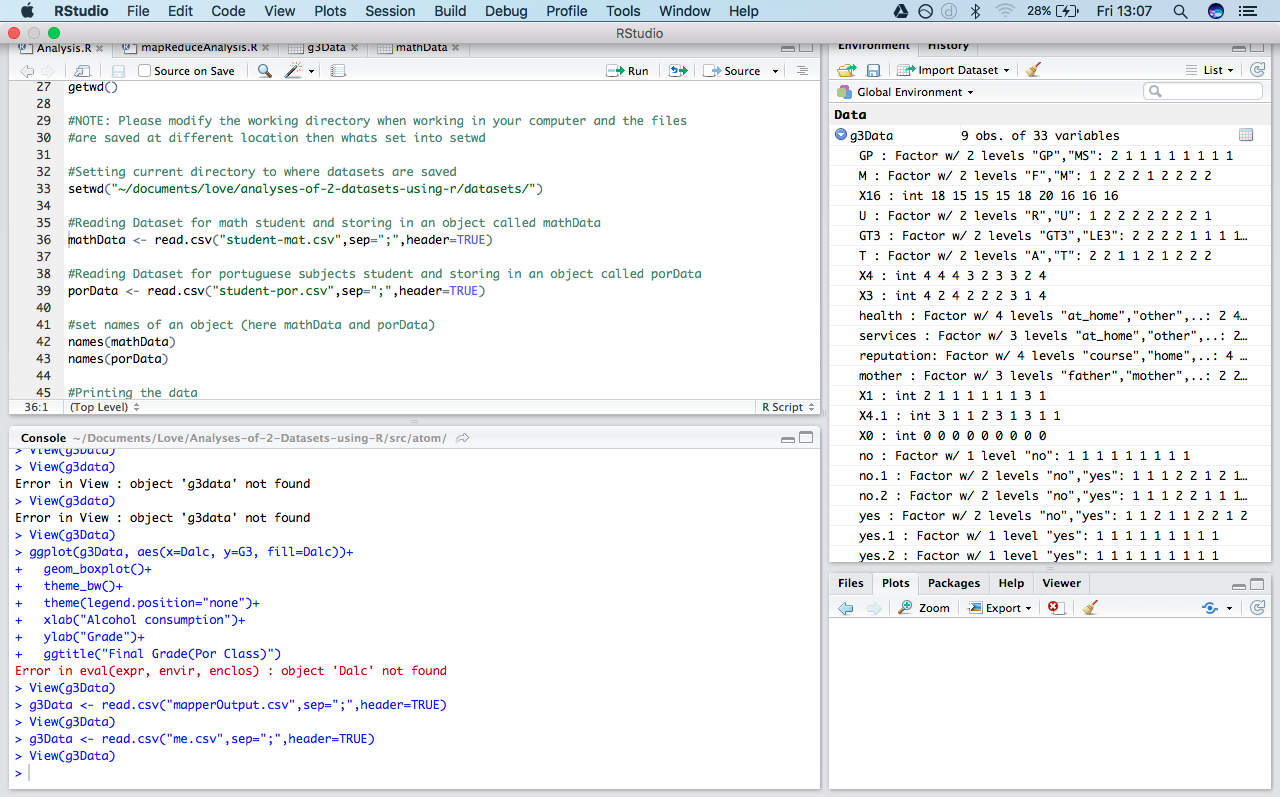
Know we will see how some CRISP-DM phases were used during analysis:



Like KDD, in CRISP-DM that the data selection is necessary and datasets were selected, data understanding was involved as can be seen in Data section of the paper, data preparation and modelling was involved to some stage and deployment (the final results of the analysis).

# Implementation and artitecture

Once the datasets were selected, RStudio was used as the work environment for analysis and first step was to setup the current working directory and then both datasets were loaded and saved as R objects. The dataset were clean and didn’t require much cleaning and then some testing was performed on the data to understand the data like the summary() function in R.



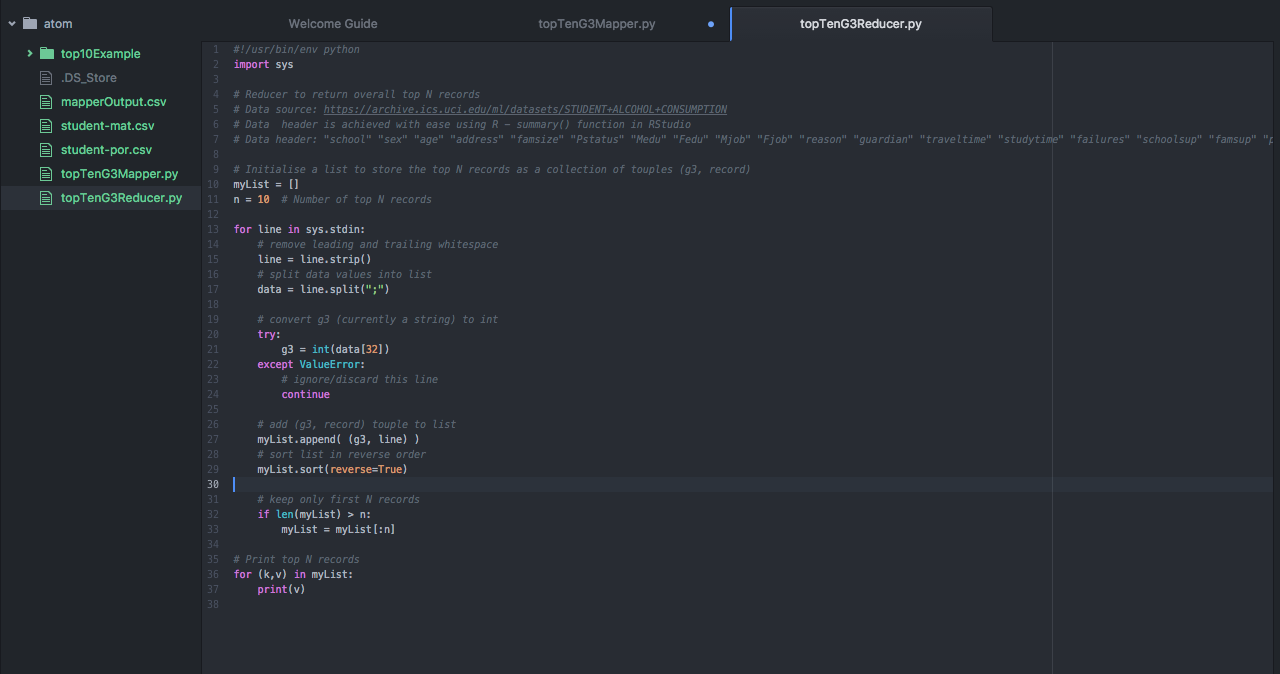
There are two types of analysis used during the analysis –

* Exploratory analysis - to find relations and connections in datasets and visualization.
* Descriptive analysis – used describe mostly the basic feature of the datasets and provide summary.

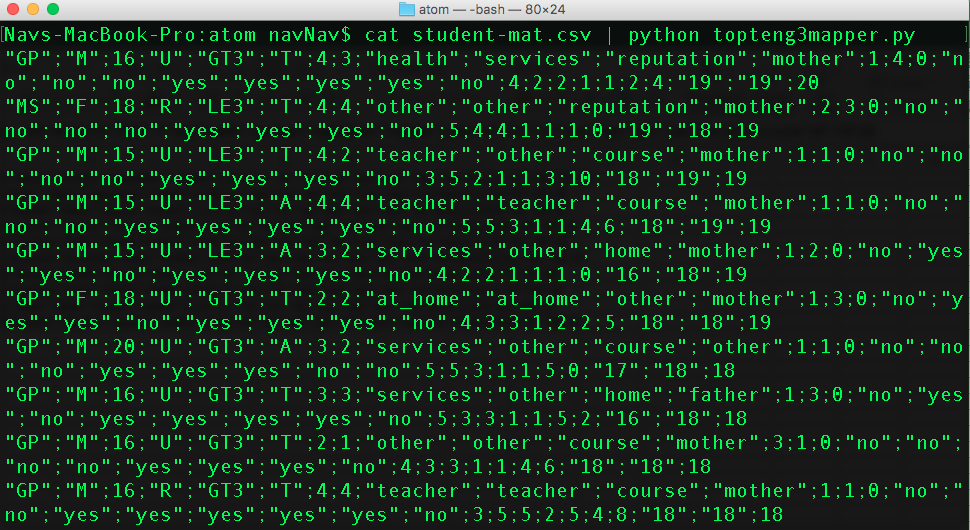
The analysis of the dataset is done in two ways –

* Individually: Firstly, both the datasets are analyzed individually and results are produced (available in results section).
* Combined: Secondly, the data from both the datasets were combined and analysis was conducted and results were produced (available in results section).

Map Reduce: The mathematics dataset went through map reduce programming paradigm and top 10 student with best G3 (Final Grade were selected) and presented the whole map reduce process was done using Atom editor for .py files and mac terminal for passing dataset through mapper and reducer and outputting the results in a CSV file. And the found result was the student’s getting highest grades were not consuming alcohol and hence had better results then other students.



Screenshot of the environment for mapper and reducer.



Screenshot of passing dataset to mapper in mac terminal.

# RESULTS

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

### For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

#### Selection (Heading 4): Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Deletion: Delete the author and affiliation lines for the second affiliation.

### For author/s of more than two affiliations: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the “Columns” icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Highlight author and affiliation lines of affiliation 1 and copy this selection.

#### Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

#### Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Column icon and select “2 Columns”. If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

# Conclusion and Future Work

here

##### Acknowledgment

Similar work - Using data mining to predict secondary school student alcohol consumption by Fabio P. & M. Hossain at University of Camerino (Department of Computer Science)

##### References

1. F. Pagnotta and M. Amran Hossain. Using data mining to predict secondary school alcohol consumption
2. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978- 9077381-39-7.
3. "Alcohol And Mental Health". Drinkaware.co.uk. N.p., 2016. Web. 30 Dec. 2016.
4. "Alcohol's Effects On The Body | National Institute On Alcohol Abuse And Alcoholism (NIAAA)". Niaaa.nih.gov. N.p., 2016. Web. 30 Dec. 2016.
5. "CRISP-DM: Towards A Standard Process Model For Data Mining". citeseerx.ist.psu.edu. N.p., 2016. Web. 30 Dec. 2016.
6. "Student Alcohol Consumption". Kaggle.com. N.p., 2016. Web. 18 Dec. 2016. "Student Alcohol Consumption | Kaggle". Kaggle.com. N.p., 2016. Web. 15 Dec. 2016.
7. "The KDD Process For Extracting Useful Knowledge From Volumes Of Data". pbworks.com. N.p., 2016. Web. 30 Dec. 2016.
8. "UCI Machine Learning Repository: STUDENT ALCOHOL CONSUMPTION Data Set". Archive.ics.uci.edu. N.p., 2016. Web. 5 Dec. 2016.
9. “Data Mining: Knowledge Descovery” | Tutorialspoint.com N.p., 2016. Web. 28 Dec. 2016.