PREDICTION OF STUDENT'S PERFORMANCE

Depend on consumption of alcohol and other metrics

Final Project Submission of
Data Analysis and Knowledge Discovery

Ahsan Bilal
ahsan.bilal@est.fib.upc.edu

Table of Contents

Motivation

The background of this project is to perform the Data Mining Process on real world dataset and apply the different statistical techniques. For this reason, "Student Alcohol Consumption Dataset" used available in UCI Public Machine Learning Repository[1]. The motivation of this project is to predict how students performance in terms of grades effect on the consumption of alcohol in weekend and weekdays depends on other variables. This model can use to analyse how the consumption of alcohol depending on other parameters are affecting the final grades of students. The complete analysis is performed in R-Studio with the help of different ML packages. The R-Script is also available in the *src* directory inside the submission.

Methodology

To perform machine learning algorithm on particular dataset the CRISP Data-Mining Process is well known process methodology are used. In this project, the modeling is not directly applied to the data but after performing through various steps of Data Preparations which need 70% of time of building to prepare data for one classification or clustering algorithm. Different algorithms are designed for different types of variables (numerical, categorical, ordinal, nominal etc).

CRISP-DM Model

CRISP-DM model is a very popular methodology that provides a structured approach to planning a data mining project. CRISP stands for Cross-Industry Process for data mining [2]. The six data mining processes are; Business Understanding, Data Understanding, Data Preparation, Data Modeling, Data Evaluation, Data Deployment.

Business Understanding

The first stage is to understand the problem to be solved, and what we need to accomplish in this stage is to figure out from a business perspective what you are trying to do, and creativity often plays in massive role in this business understanding stage, the specific steps in this stage are; determine business objectives, assess the situation which means a detailed fact-finding exercise, data mining goals, produce the project plan.

Data Understanding:

Now let's shift gears to the second phase which is the data understanding stage, this stage requires the data-scientist to acquire the data and before data scientist does anything with the data, first he

[1] Student Alcohol Consumption Dataset
https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION
[2] CRISP
https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

needs to understand the strengths and limitations of the data because rarely will the data exactly match the problem that he is trying to solve. It is also important to recognize that data costs money and some data may not be available, so the data scientist needs to evaluate the costs and benefits of all the different potential data sources and what one may find that the initial solution paths that already start to diverge in the second stage as the data comes up to the surface, in this stage the specific steps are; collect initial data, describe the data, explore the data, verify the quality of the data.

Data Preparation:

In the third stage of the CRISP model is the data preparation, and essentially analytics technologies often require the data be in a form that is different from how the data was initially provided and so conversion may be necessary. Some common data preparation examples include; convert data into tabular formats, removing or inferring missing values, converting data to different types, and scaling numerical values. Actually this phase is one of the most time consuming stages of these six stages model. In this stage the specific steps are; select the data, clean the data, construct the data, integrate the data.

Modeling:

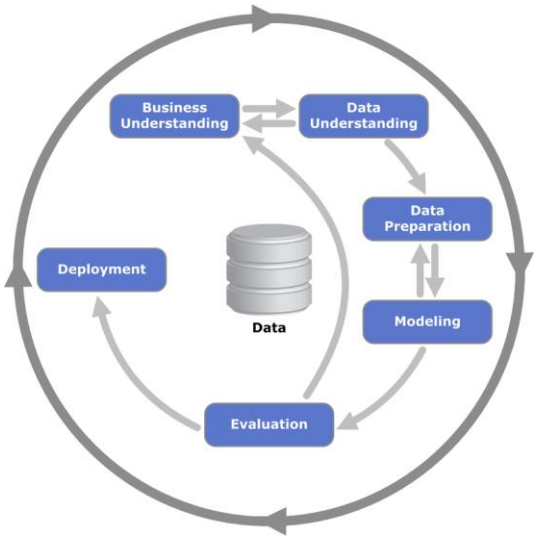
In the fourth stage of CRISP model, we encounter modeling and this is the primary stage of data mining technologies, applied to the data with output being some sort of model or pattern. Modeling stage is iterative with the data preparation stage so as data comes to the light, new models are built, as models are built you may go back to the preparation stage if you realise that the models are not very related. In this stage, the specific steps are; select the modeling technique, generate tests for model robustness, build the model, and assess the model.

Evaluation:

Now we move to the fifth stage of the CRISP-DM model, which is the Evaluation stage. The purpose of this stage is to assess the data mining results, both from quantitative and qualitative perspectives and have a keen attention to detail, determine if the results are both justifiable and feasible. Basic purpose of this stage that if the model satisfies the original business goals, and so it has a direct link with first stage of the models, i.e, Business Understanding, that you may go back to the first stage if you don't have meaningful results. In this stage, specific steps are; evaluate the results, review the process, determine the next steps.

Deployment:

In the very last stage which is Deployment, where the results of the data mining are put into use.



Data Understanding:

Our dataset contains following variables and we are going to analyze which field or factor is more important for us analysis.As this dataset is originally generated in University of Portugal. Data is gathered from 2 different course in Mathematics and Portuguese to analyze individually the performance in each subject.

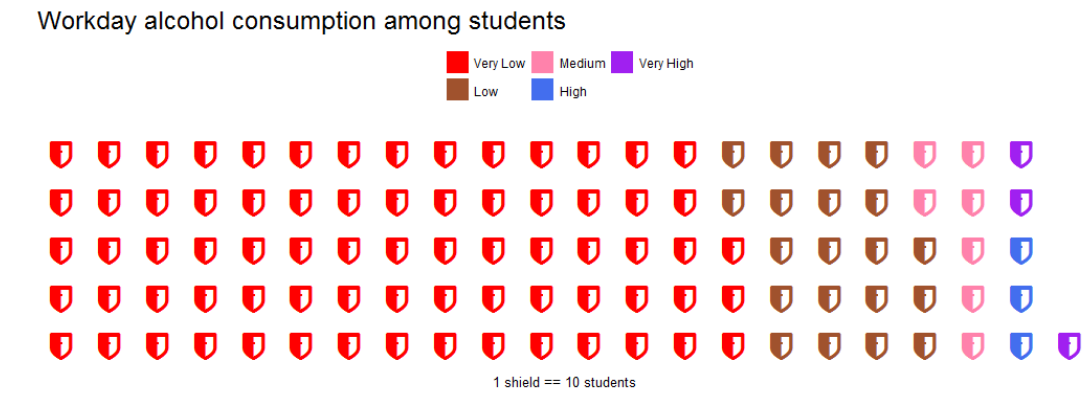
Data statistics

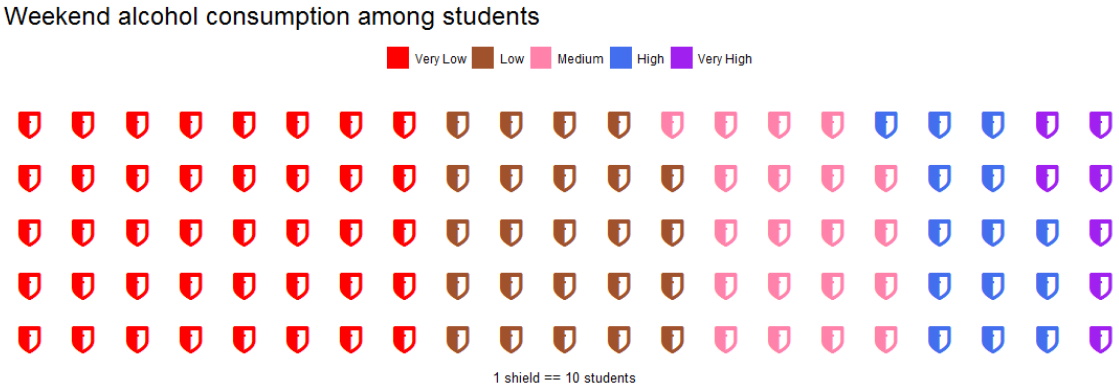
Number of Variables	33
Number of Observation (Mathematics**)**	395
Number of Observation (Portuguese**)**	649
Total number of Observation	1044

Data Analysis

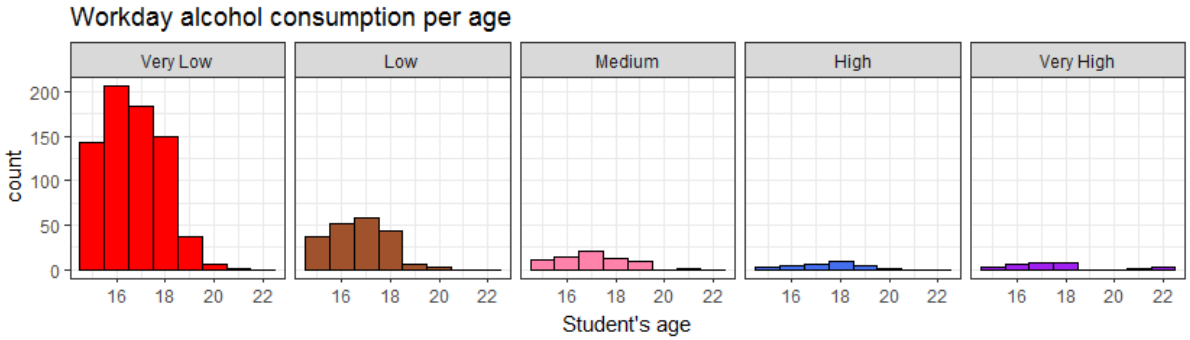
On our current dataset we performed different analysis as follows

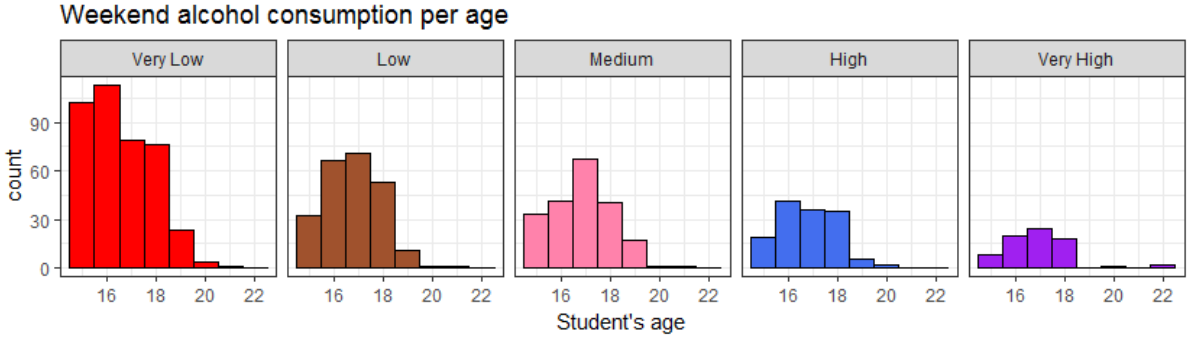
Analysis 1: Alcohol consumption among students



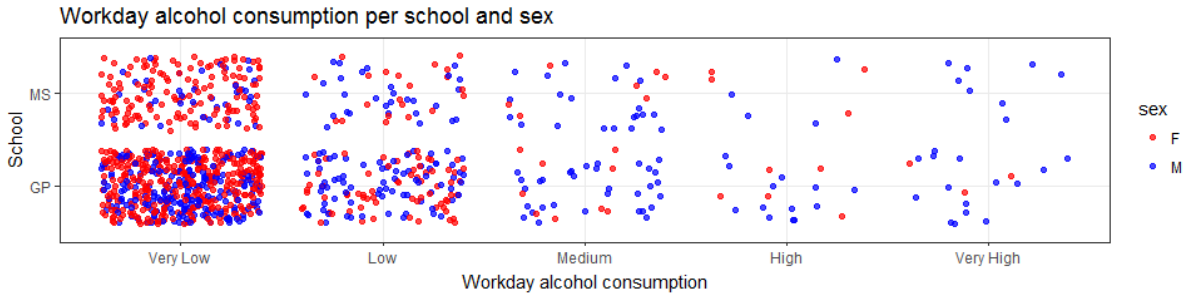


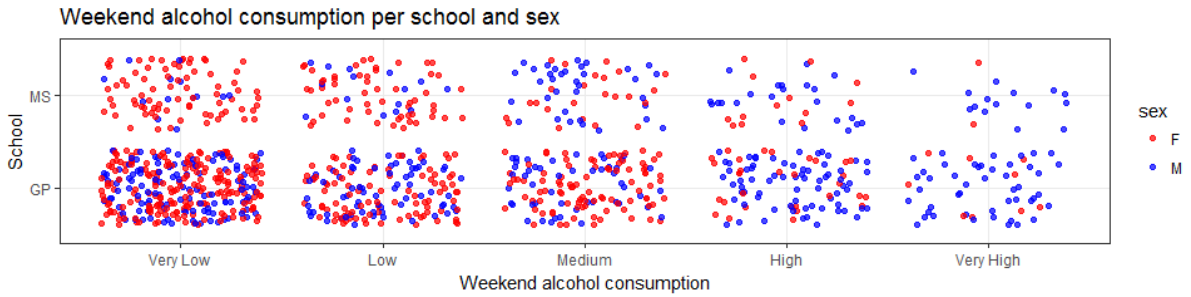
Analysis 2: Alcohol consumption per age



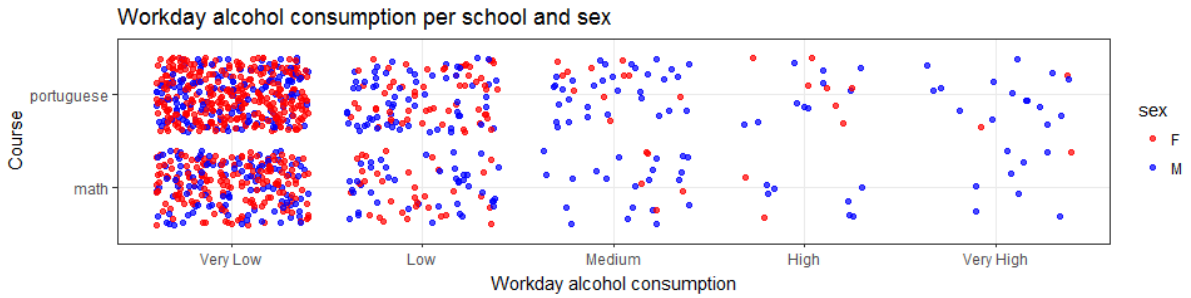


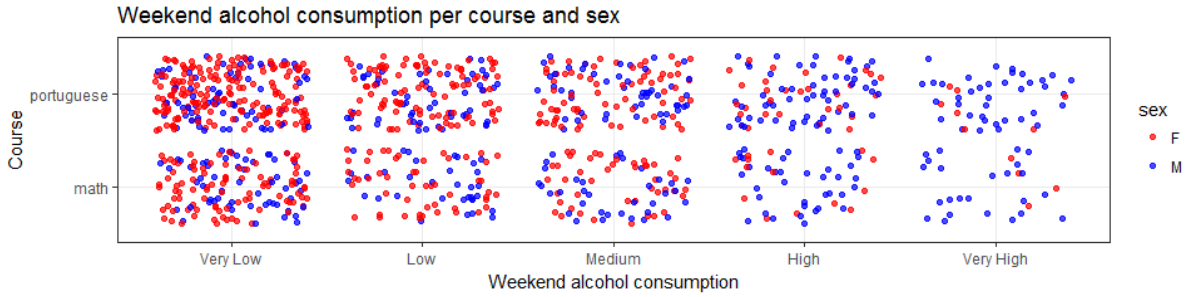
Alcohol consumption per school



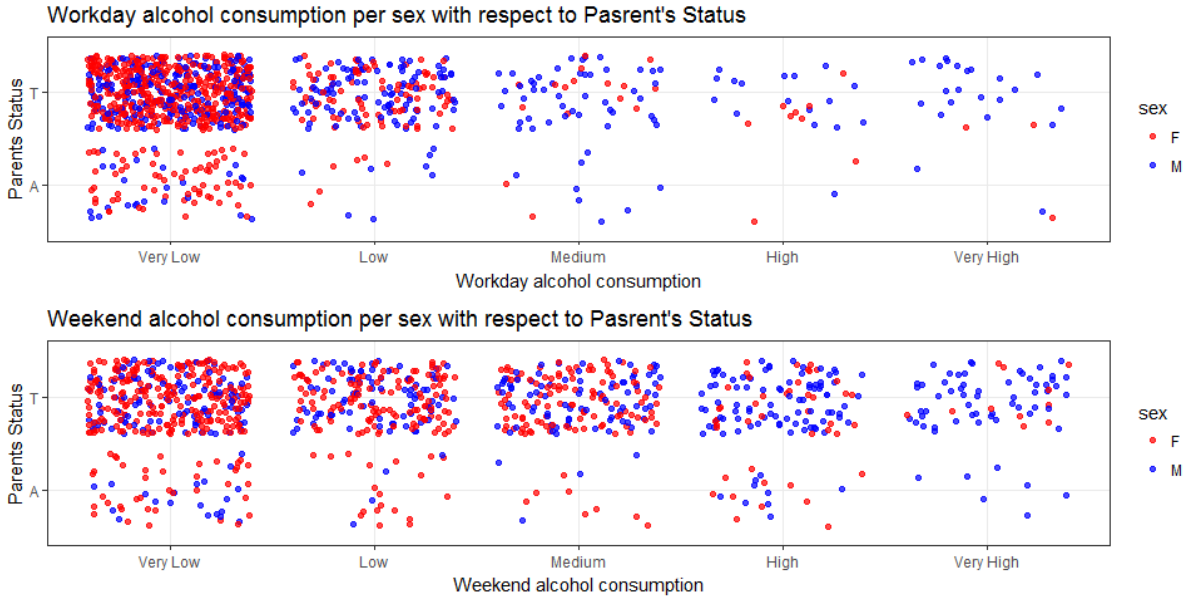


Alcohol consumption per course

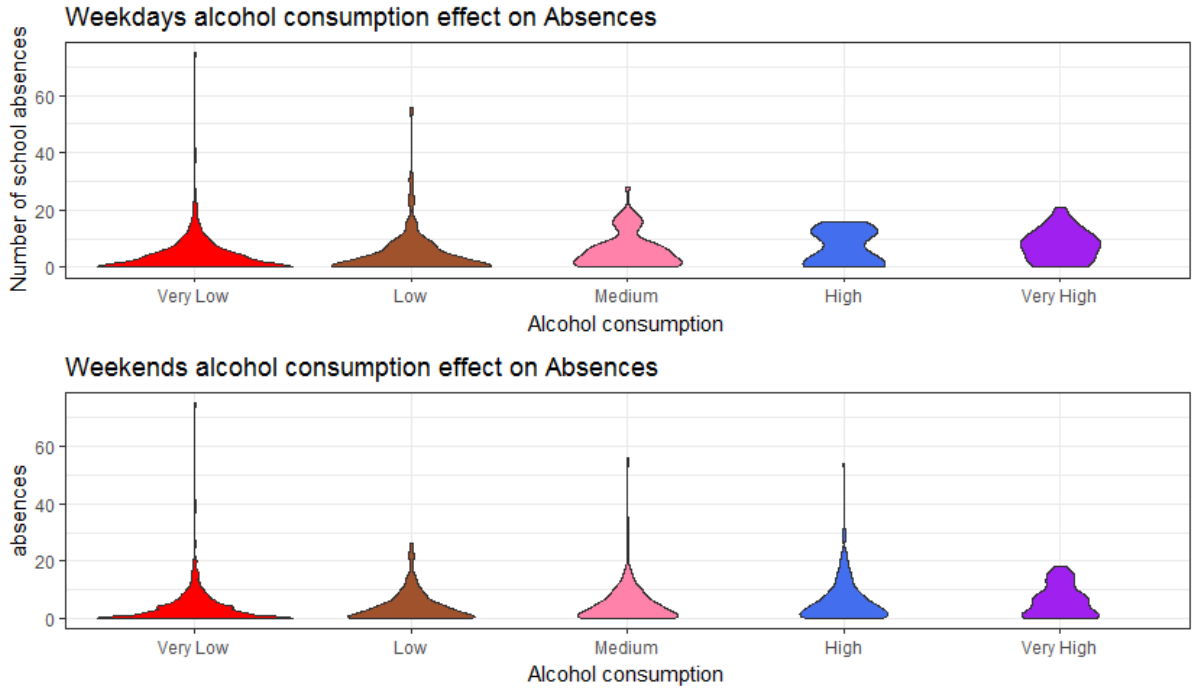




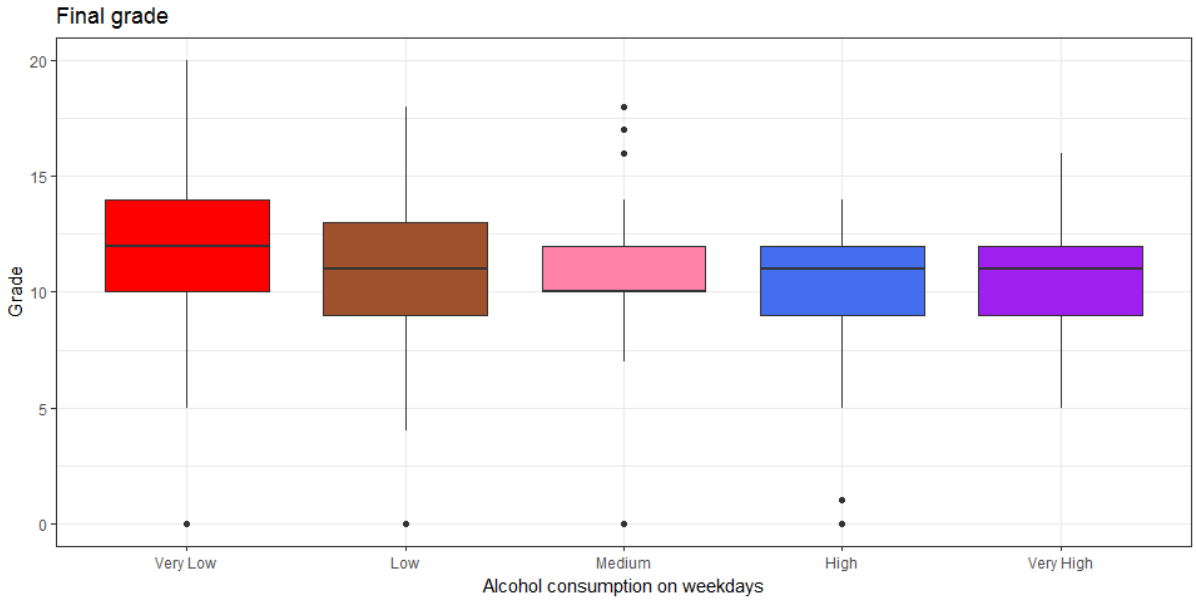
Effect of parents status on students alcohol consumption

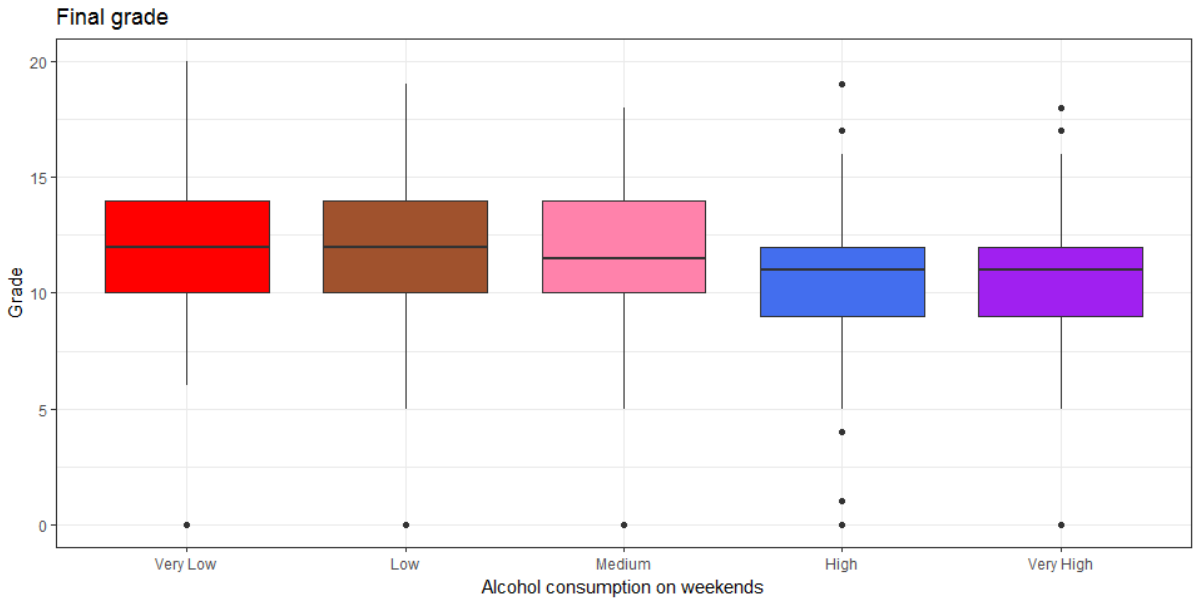


Alcohol consumption effect on students absences



Alcohol consumption effect on final grade





Data Preparation

Data preparation indeed most important step in order to achieve good accuracy of model. In this document we tried to cover most important Data Preparation steps as follows.

1. Features Analysis

In first step we categorise the type of features as follows:

Numerical Variable:	**Categorical Variable:**	**Nominal Variable:**
age, absences, Dalc, Walc, G1, G2, G3	school, sex, address, famsize, Pstatus, Medu, Fedu, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, health, math_instance	Mjob ,Fjob, reason, guardian

Complete Description of Dataset can be found in Annex.

As the given dataset available in two files. First we combine both the dataset and add one more variable name "course" which contains ("maths", "Portuguese") to make it distinguishable.

As we have to predict the performance of students, we consider only the G3 (Final Grade) of student and excluding the G1 (Grade of First Period) and G2 (Grade of Second Period). Also we convert the G3 score points (0 to 20) to the Grades system according to the following portugal grade conversion chart [3].

Scale	**U.S. Grade**
20	A+
18 - 19.99	A+
16 - 17.99	A
14 - 15.99	B
10 - 13.99	C
7 - 9.99	F
1 - 6.99	F

[3] http://www.foreigncredits.com/resources/Grade-Conversion/

2. Missing Values

There is no missing value found in our dataset

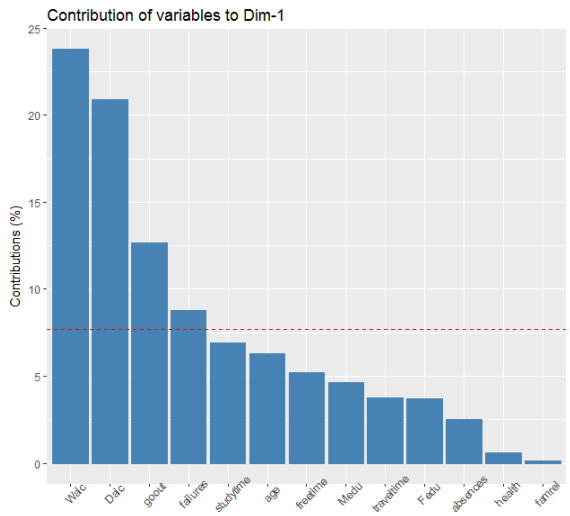
3. Feature Selection

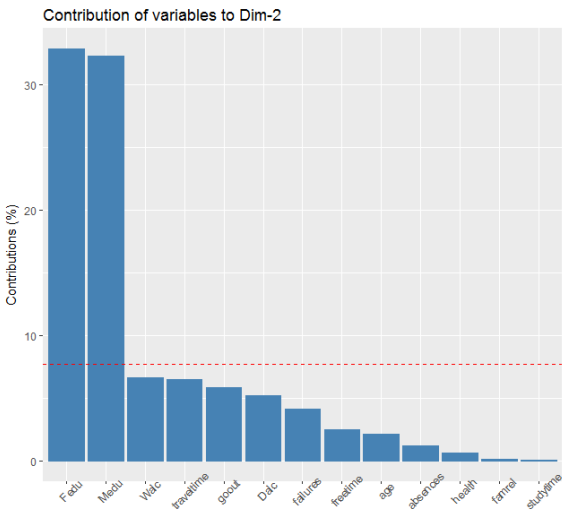
As currently we have one variable to predict which depends on other 30 variables excluding G1 and G2 because as discussed in above section we only considering G3. In this section we are going to select most significant variables for our model. So that we can reduce the least significant variable and reduce from 30 variables to 5 to 10. Currently we have two types of variable numerical and categorical here we apply the PCA Principal Component Analysis on numeric variables because its design only for numerical variable.

PCA for Numeric Variable

Principal component analysis (PCA), which is a technique often used to make data easy to explore and visualize highly correlated variables and reduce the variables when you have large number of variables. Otherwise with a large number of variables, the matrix may be too large to study and interpret properly, and there would be too many pairwise correlations between the variables to consider. For interpretation of each component, we must compute the correlations between the original data for each variable and each principal component.

In following PCA Contribution Plot its shows the contribution in percentage of each variable and red dotted line represent the all the variable which contribution down to these red dotted line is most significant means we can ignore these variables.





Thus from Dimension1 ***Walc, Dalc, Gout and failures*** most contributing in Dimension 1 these variables are contributing almost 66% and in Dimension 2 Fedu and Medu both contributing to 60%. Hence ***Walc, Dalc, Gout ,failures , Fedu , Medu*** are important numerical variables.

Correlation Matrix

Dimension1

```
> model.desc$Dim.1
$quanti
           correlation       p.value
Walc         0.7384502 1.565420e-180
Dalc         0.6919795 1.247898e-149
goout        0.5383844  1.628187e-79
failures     0.4486949  7.473746e-53
age          0.3796194  3.986954e-37
freetime     0.3458513  1.061527e-30
traveltime   0.2938147  3.108224e-22
absences     0.2402094  3.615797e-15
health       0.1190630  1.151824e-04
Fedu        -0.2904760  9.549105e-22
Medu        -0.3262776  2.552883e-27
studytime   -0.3973090  8.285174e-41
```

Dimension2

```
> model.desc$Dim.2
$quanti
           correlation       p.value
Fedu         0.7823427 1.760168e-216
Medu         0.7755391 2.230617e-210
Walc         0.3511172  1.186396e-31
goout        0.3291003  8.598224e-28
Dalc         0.3101836  1.017844e-24
freetime     0.2160817  1.701568e-12
absences     0.1467459  1.920399e-06
health       0.1070628  5.299865e-04
age         -0.1970542  1.340875e-10
failures    -0.2779616  5.620484e-20
traveltime  -0.3472169  6.037674e-31
```

After checking the Correlation Matrix it's also clear that the ***Walc, Dalc, Gout and failures*** are highly correlated with dimension 1 and **Fedu and Medu** are highly correlated with dimension 2.

PCA Assumption: We have to consider top 5 numerical variable

Thus Final Numeric Variables are ***Medu, Fedu, Dalc, Walc and goout***

MCA for Categorical Variable

For categorical variable we cannot use PCA and indeed we have almost 26 categorical variables and we really need here to reduce the dimensions. MCA finds the contribution of each category of variables as shown in this figure.

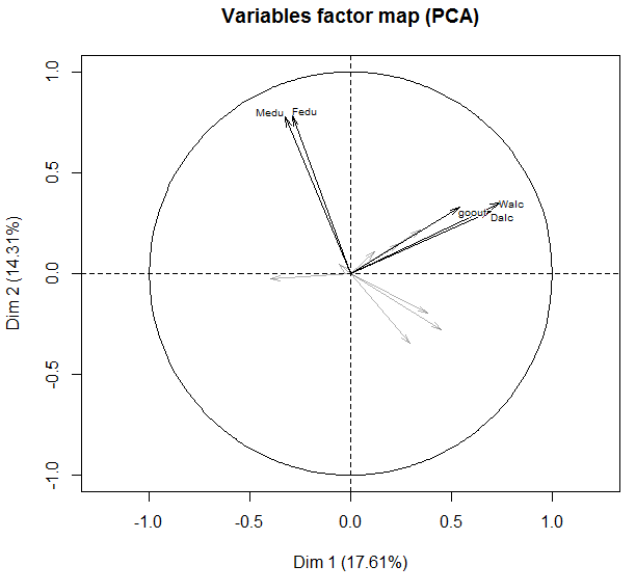
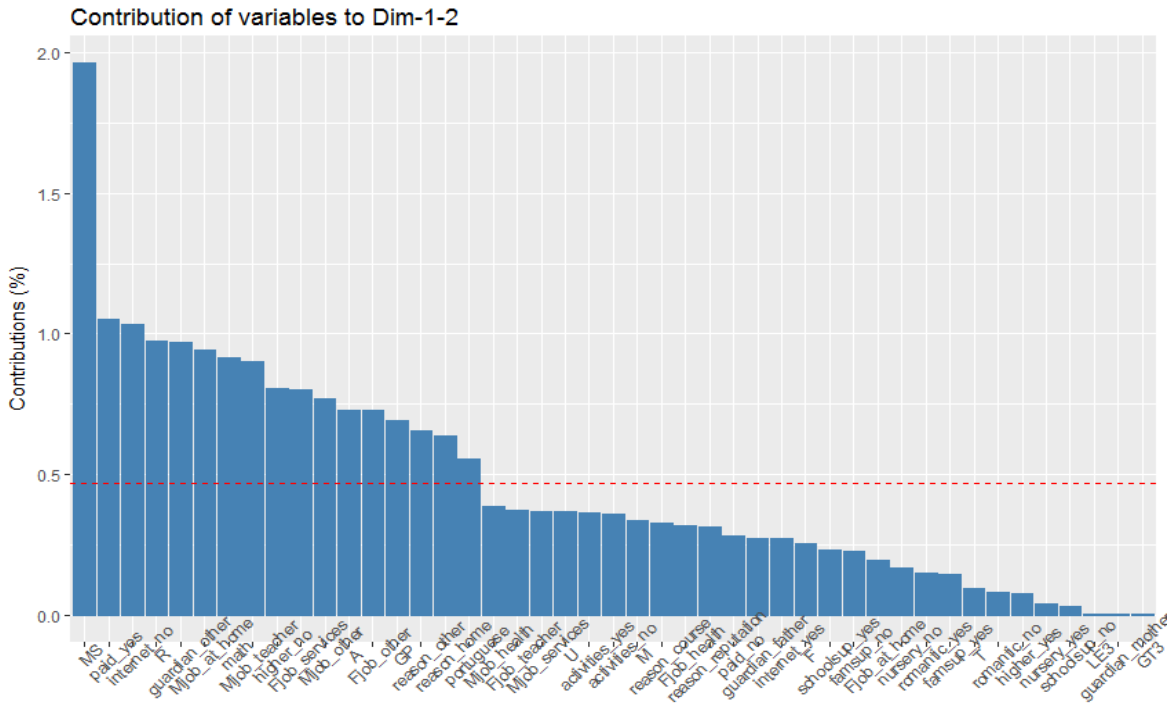
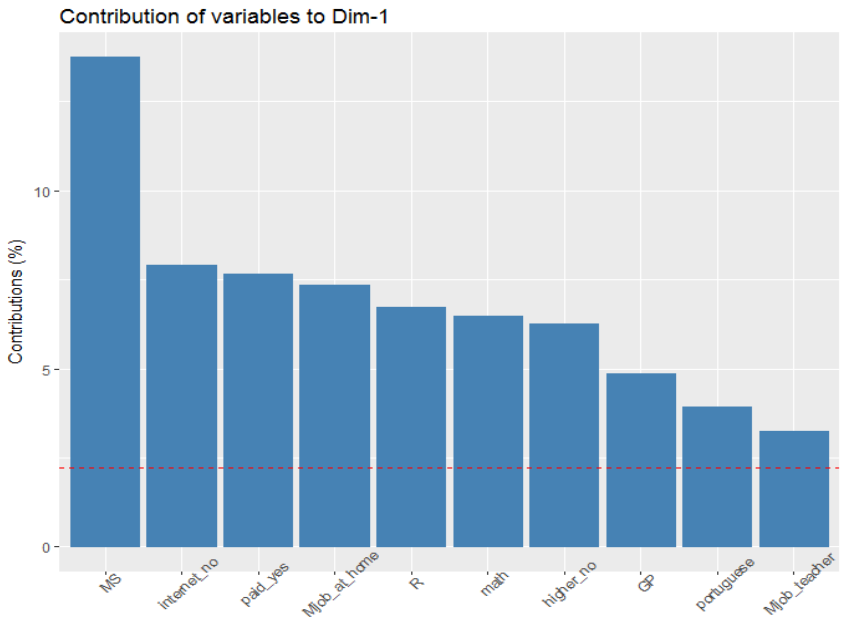


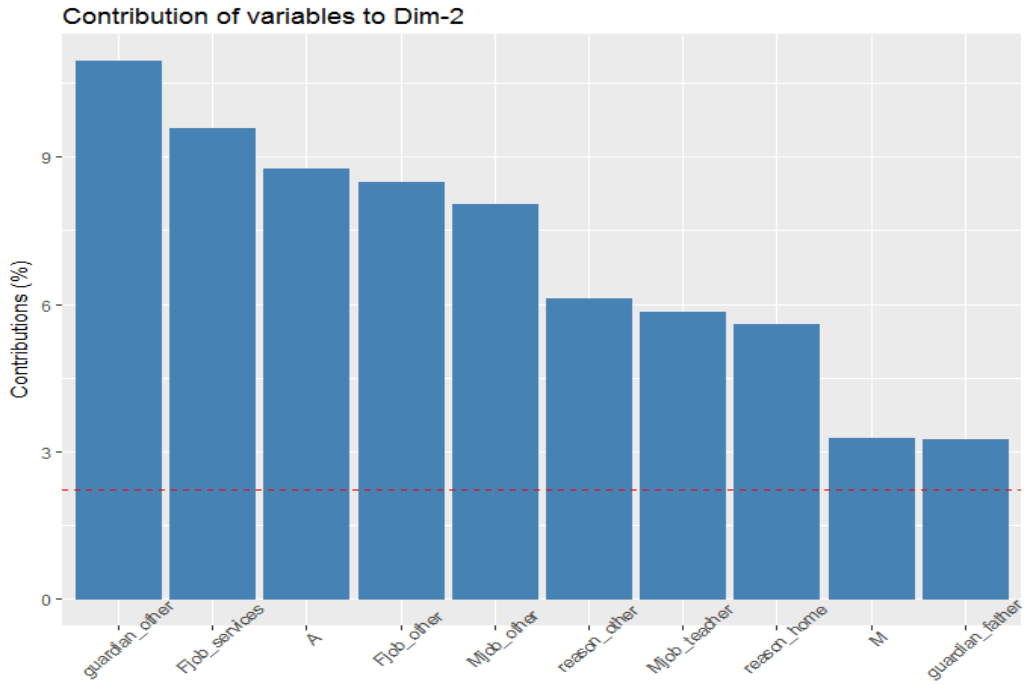
Fig: 1PCA Map



To zoom this visualization in terms of top 10 categories per dimension we can identify the most significant variables as shown in following figure.



Selection in Dimension -1: School ,internet, paid, Mjob, address, course, higher



Selection in Dimension- 2: **Guardian, FJob, PStatus, Mjob, reason, sex** thus final Feature Selection School ,internet, paid, address, course, higher ,guardian, FJob, PStatus, Mjob, reason, sex Medu, Fedu, Dalc, Walc and gout.

4. Data Sampling

We divide the dataset in testing and training dataset. We divide the 70% of data for the model to train and 30% for evaluating the model.

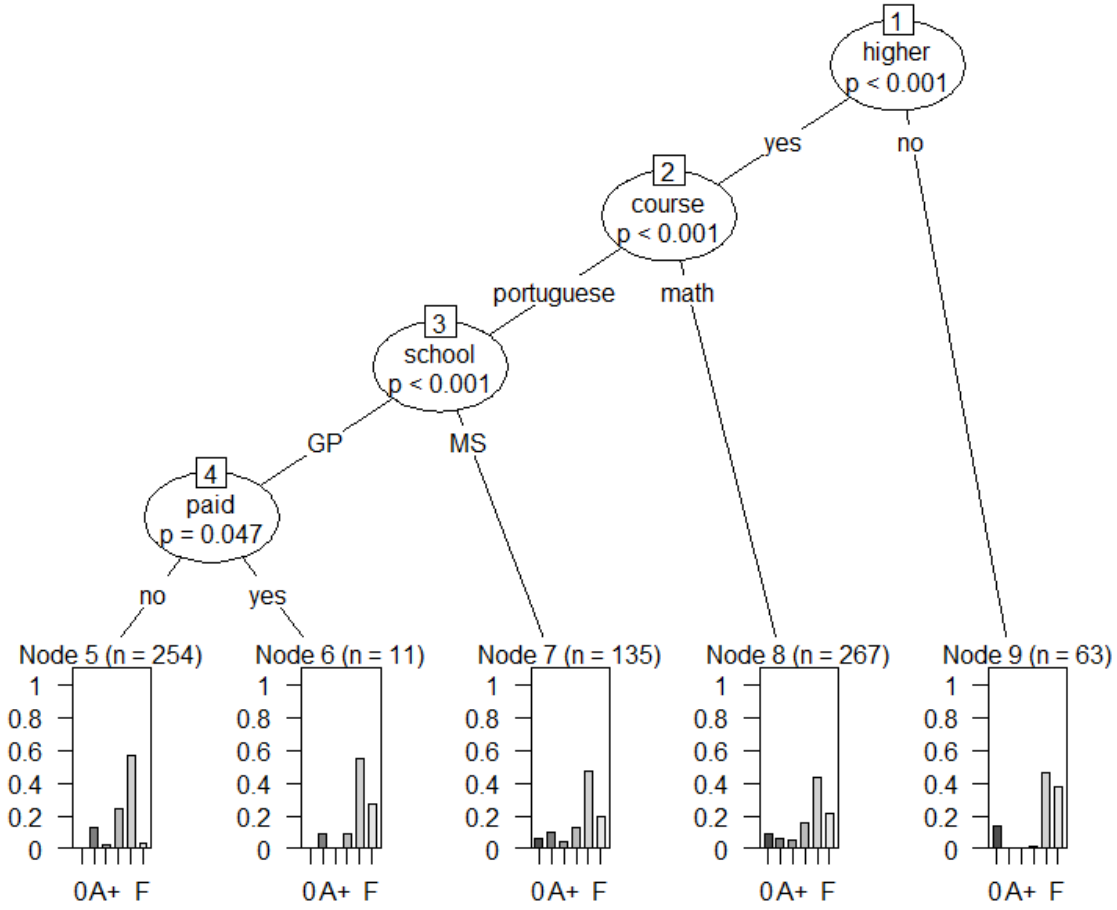
Data Modeling

After data preparation step, we are now ready to build the model using these final features

Predictor	G3
Features	school, internet ,paid, address, course, higher , guardian, Fjob, Pstatus, Mjob,, reason, sex, Medu, Fedu, Dalc, Walc , goout

Decision Tree

First we apply Decision Tree algorithm (ctree) available Party Package. This is our build decision tree.



Naive Bayes:

One more classification techniques Naïve Bayes is used using ***e1071*** library package in R.

Model Evaluation & Selection

Hence we checked 2 classification algorithm decision Tree and Naive Bayes which are two well know classification algorithm. We need to evaluate the accuracy of our model based on test dataset. Evaluation is performed using confusion matrix available in caret package of R.

Decision Tree	Naive Bayes

```
Confusion Matrix and Statistics

          Reference
Prediction   0   A  A+   B   C   F
         0   0   0   0   0   0   0
         A   0   0   0   0   0   0
         A+  0   0   0   0   0   0
         B   0   0   0   0   0   0
         C  11  23   9  51 161  59
         F   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.5127
                 95% CI : (0.456, 0.5693)
    No Information Rate : 0.5127
    P-Value [Acc > NIR] : 0.5227

                  Kappa : 0
 Mcnemar's Test P-Value : NA
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   A  A+   B   C   F
         0   0   0   0   0   1   1
         A   1   1   1   5   2   1
         A+  0   1   1   5   0   0
         B   2  13   3  20  31  12
         C   7   8   4  19 105  38
         F   1   0   0   2  22   7

Overall Statistics

               Accuracy : 0.4268
                 95% CI : (0.3714, 0.4835)
    No Information Rate : 0.5127
    P-Value [Acc > NIR] : 0.9991

                  Kappa : 0.1043
 Mcnemar's Test P-Value : NA
```

After analysing this evaluation, we found that Decision Tree showed 51% of accuracy while the naive bayes only 42%. In this case we select Decision Tree model and use this for prediction.

Conclusion

In this project, I learned so many new techniques for building the model, for analysing the data, and to visualize data with different plots. As far as summarize the accuracy of model, the accuracy 50% is not supposed to be very good model there could be many of the reason possible that data have some skewness which does not build very good model. Also the distribution of grades is dominating to 'F' grades. We cannot judge at that beginning to apply different methods in data preparation because of limited expertise in this domain. Thus this process is iteration based as explain in CRISP-DM Model we have to iterate the process from the beginning to play with data in order to achieve the best accuracy at least greater than 80%. Because the deployment of solution needs very good accuracy which can use for classification or clustering algorithm.

ANNEX
DATA SET DESCRIPTION

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 Ã¢â‚¬â€œ 5th to 9th grade, 3 Ã¢â‚¬â€œ secondary education or 4 Ã¢â‚¬â€œ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 Ã¢â‚¬â€œ 5th to 9th grade, 3 Ã¢â‚¬â€œ secondary education or 4 Ã¢â‚¬â€œ 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)