Using Data Mining to build predictive model for first period grades of Student Performance Data Set on the UCI machine learning repository

Hitesh santwani[1]

[1] Engineering Sciences (Data Science), University at Buffalo, Buffalo, United States
E-mail: hiteshka@buffalo.edu

Abstract

In this paper we are following standard CRISP-DM development methodology to build the predictive model for the first period grades of Student Performance Data Set on the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/student+performance).
This is clearly the classification problem and we approach the problem by performing multiple regression on the dataset which will be first pre-processed to get the best outcome

Keywords: classification, multiple regression, pre-processed

1. Exploratory data analysis

In this stage we will be exploring the data with help of simple statistical tools like scatterplots, bar charts, histograms, co-relation heat maps etc

This stage also involves the pre-processing of the data so that can be used ahead in the Analysis phase

1.1 Summary of data

For the description of the data columns and CRISP-DM refer to the Appendix

1.2 Missing data.

There are no missing values as shown from the figure 1

1.3 Bar Charts.

Observe the fig from 2 to 4 we can clearly see the duplicate columns with different name.

This calls for the removal of these reduntant columns as a pre processing task.

1.4 Histograms

While not much can be understood from the data in the histograms but there is certain normal distribution pattern that can be obbserved in the First, second and third period grades for both mathematics and portuguese

1.4 Correlation Analysis

Lets investigate the Co-relation heat map in figure 7 and 8
It can be easily colncluded from figure 7 that there is high co-relation between G1, G2 and G3.
While Fugure 8 does not reveal much

1.5 Boxplots

From figure 9 to 12 we can see the box plots of the Period 1 grades for both maths and portuguese vs all the attributes.

1.5 Scatterplots

Figure 13 and 14 Shows the Scatter plots for period 1 grades for mathematics and portuguese but again it does not reavl much of the correlation except it confirms again that there is high correlation between G1 and G3 for both maths and portuguese.

1.6 Scatter Plots with focus on some features (age vs first period grades)

It can be clearly seen from the figure 15 that there is some kind of negative correlation between period 1 grades and age.

By removing the outliers may be we can better fit the regression line.

1.7 Scatter Plots and histograms with focus on some features
(Extra activities vs first period grades and paid classes vs first period grades)

It can be clearly interpreted from the figures in 16 that Doing extra activities and paid classes are equally important.

1.8 Relation between first period grades and final grades

It can be clearly seen from the figure 18 that there is high correlation between first and final period grades.

1.8 Relation between first period grades and school

The studens from the school GP seems to be performing better but again they have high population.

2. pre-processing

Following are the conclusions from EDA and preprocessing requirements:
We have requirement for predicting the first period grades i.e G1 for both maths and portuguese.

- At first after merging the data from two data sources Maths and portuguese there were duplicate columns with different names that needed to be removed.
- A negative corelation between age and grades is evident we can improve the prediction by removing outliers
- A positive correlation between desire for higher education and first period grades
- First period grades and final grades are very highly correlated so again removing outliers will improve the regression line
- Failures and First period grades are negatively correlated
- High correlation between maths first period and portuguese first period grades removing outliers may help fit the line better.

3. Predictors that appear to have significant relation ship to response G1

- Age (negative correlation)
- Desire to go for higher education
- Previous failures (negative correlation)
- Second and final level grades i.e G2 and G3
- Absences are also moderately corelated
- Parents education plays a positive role
- Extra cirricular and paid classes are equally important
- Study time does plays a role

4. Suggestions to First year student based on the above analysis

- ***Achieveing good first grade is important as it will also become the motivation for good performance in the future***
- ***Target for higher education***
- ***Have a well balance between Extra cirricular activities and paid classes***
- ***Always ensure the good Study time and most importantly choose the school with good reviews***
- ***Being punctual helps a lot in Achieveing good grades.***
- ***Travel less***

5. interactions that are significant

Frome the summary of three models in the complematry R file it is clear that model 1 is a good model so the Significant interactions are the ones that are used in model 1

References

[1] Surname A, Surname B and Surname C 2015 *Journal Name* **37** 074203
[2] Surname A and Surname B 2009 *Journal Name* **23** 544

Graphs:

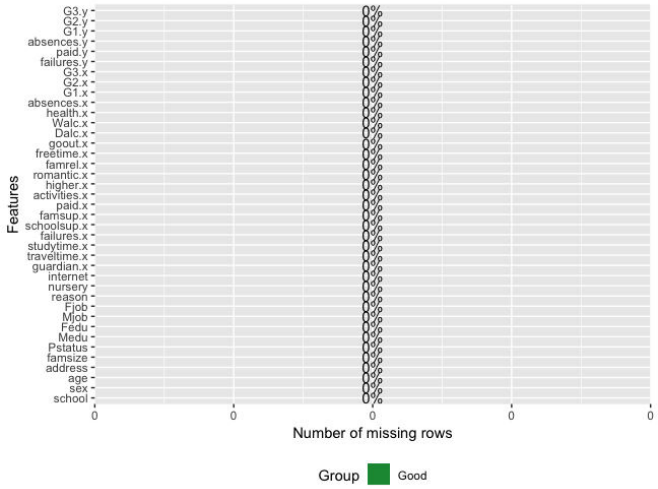


fig 1 : missing values

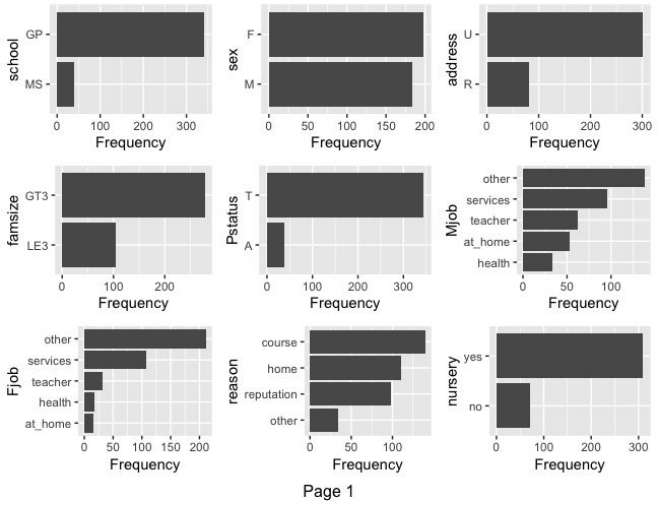


fig 2 : Bar charts

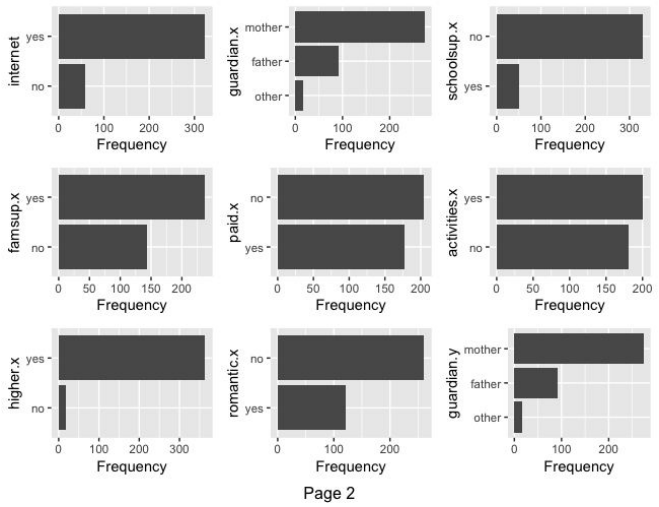


fig 3 : Bar Charts

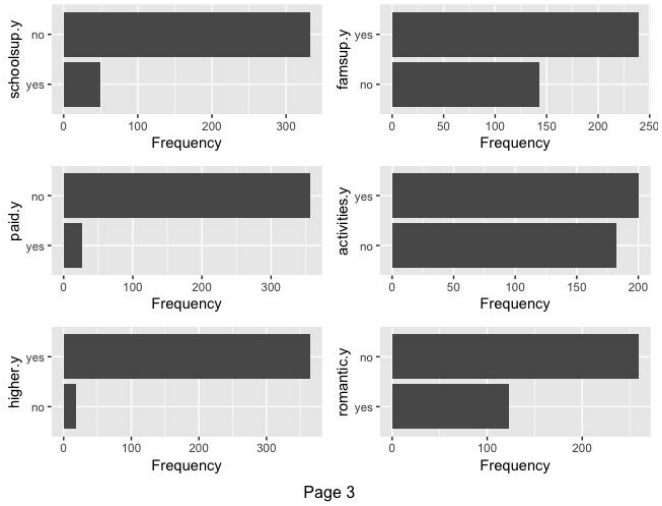


fig 4 : Bar Charts

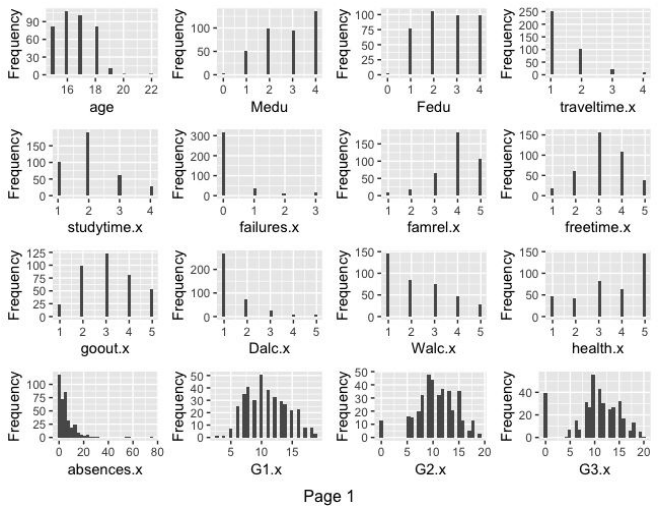


Fig 5 : Histograms

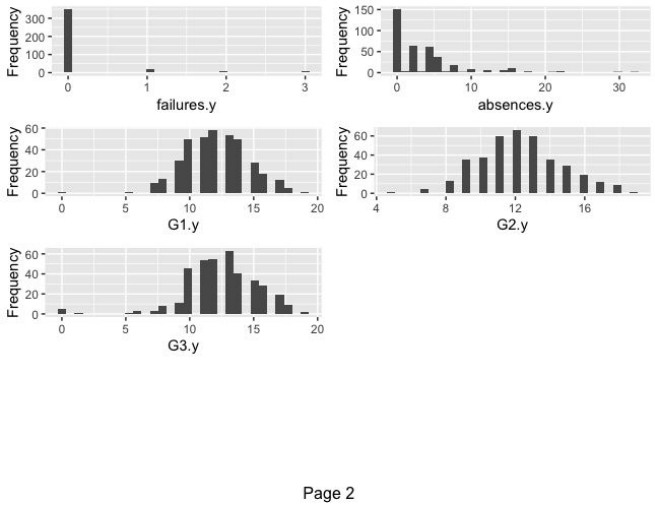


Fig 6 : Histograms

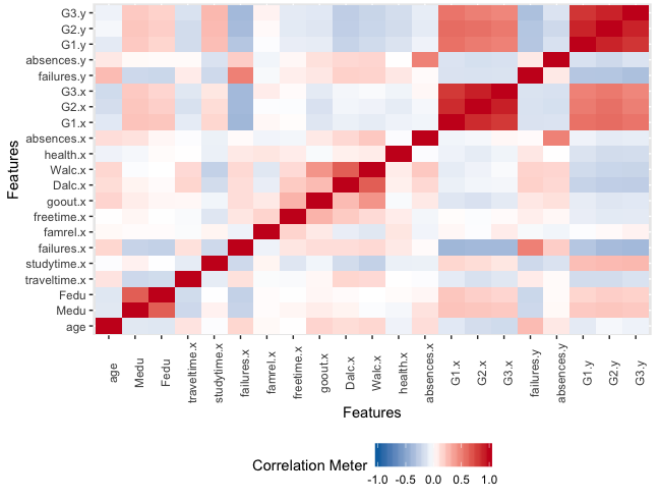


Fig 7 : Correlation heat map for continous columns

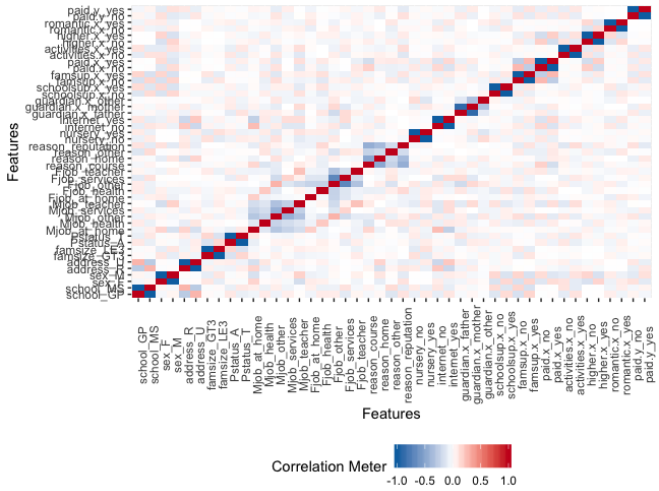


Fig 8 : correlation heat map for discrete columns

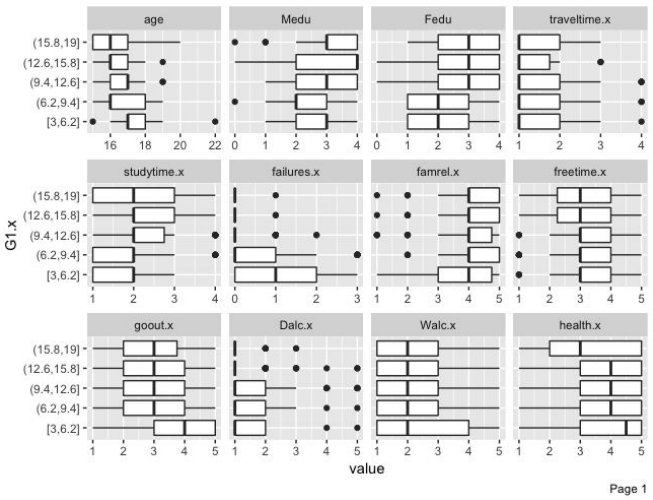


Fig 9 : Box plot Period 1 grades (Mathematics) vs all the attributes

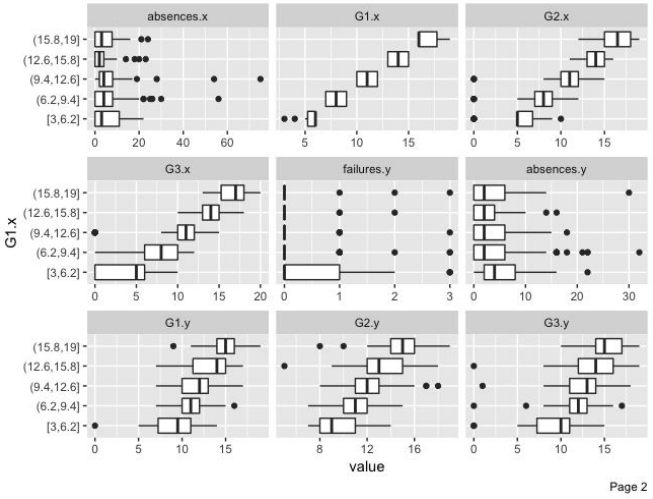


Fig 10 : Box plot Period 1 grades (Mathematics) vs all the attributes

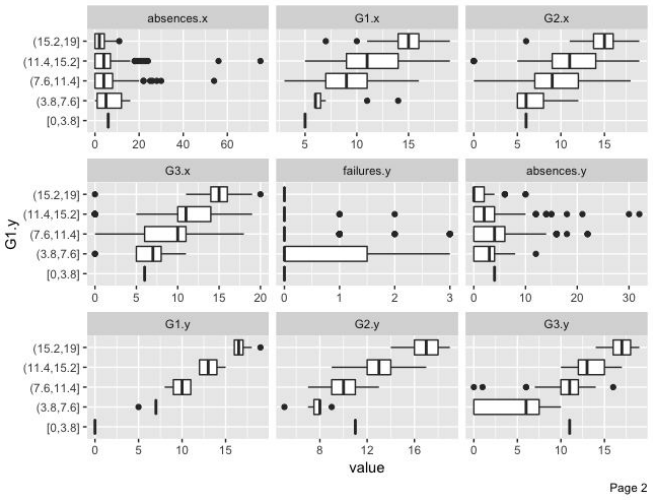


Fig 11 : Box plot Period 1 grades (Portuguese) vs all the attributes

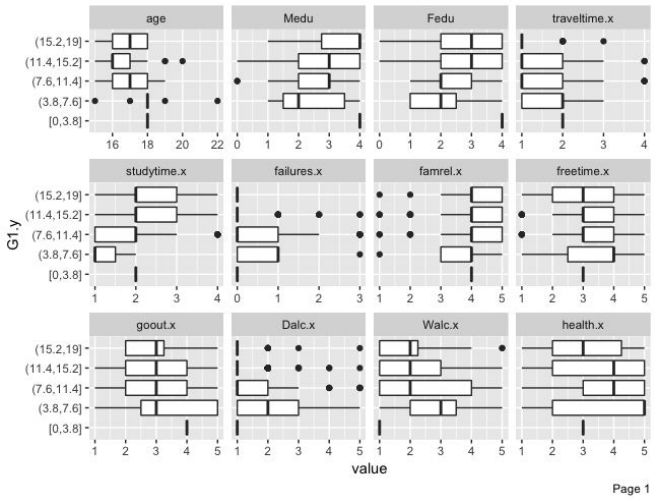
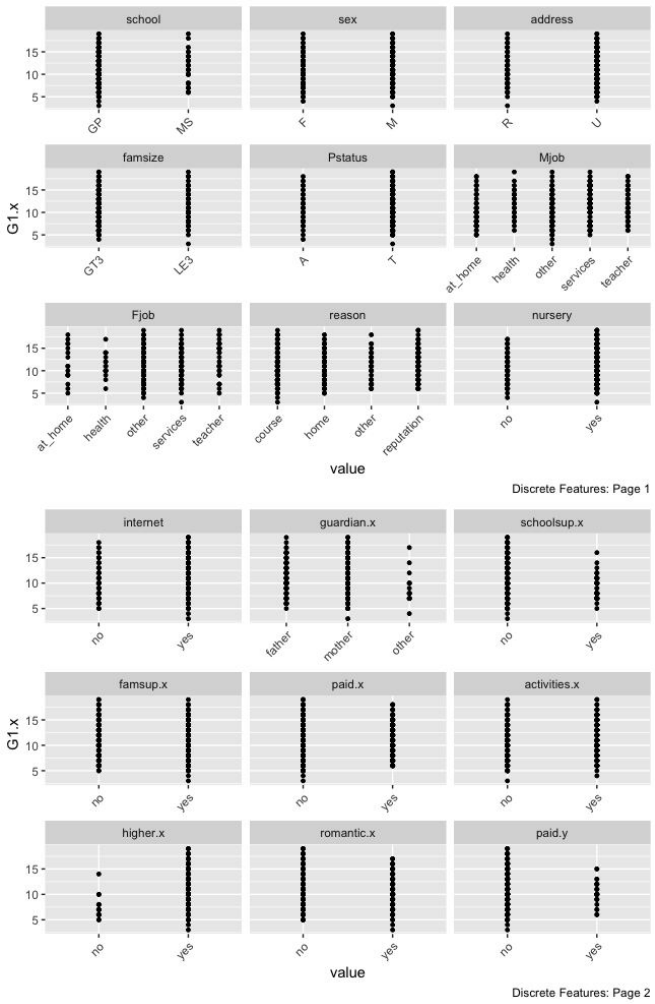


Fig 12 : Box plot Period 1 grades (Portuguese) vs all the attributes



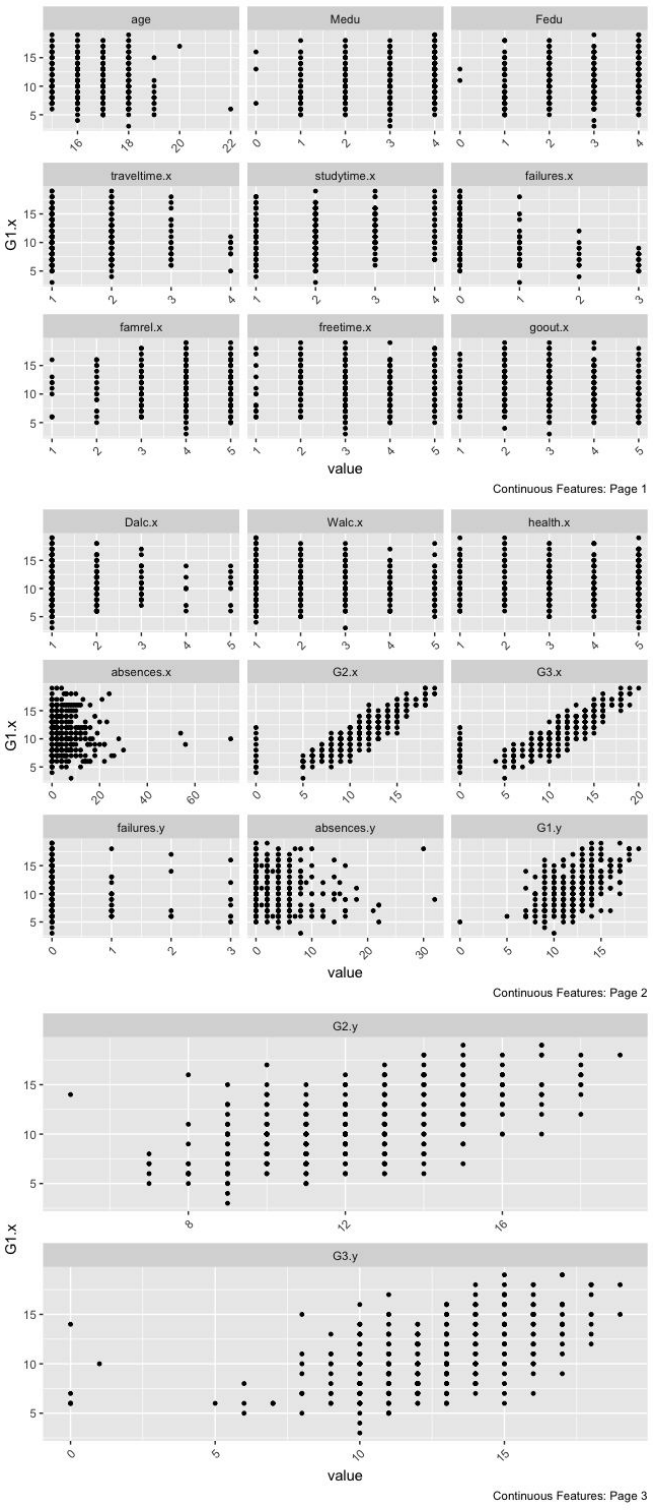
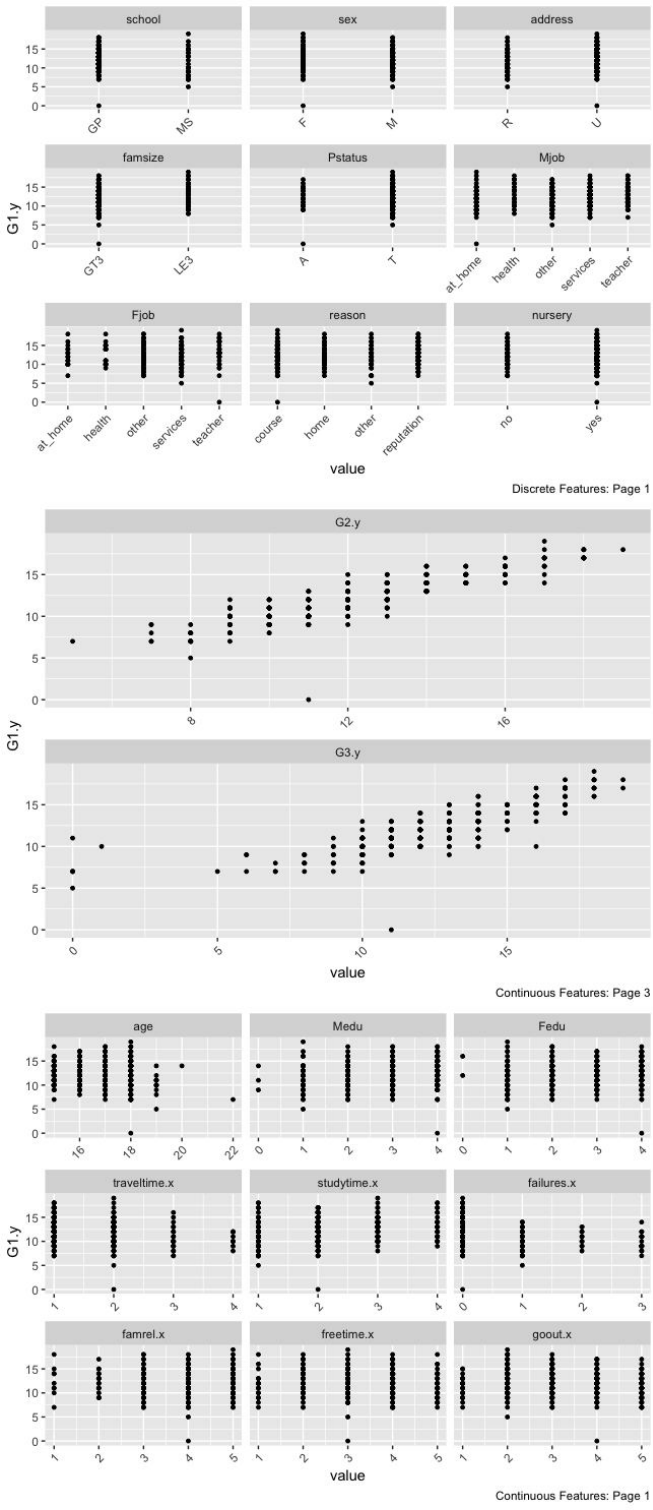


Fig 13 : Scatter plots for First period grades vs all the attributes



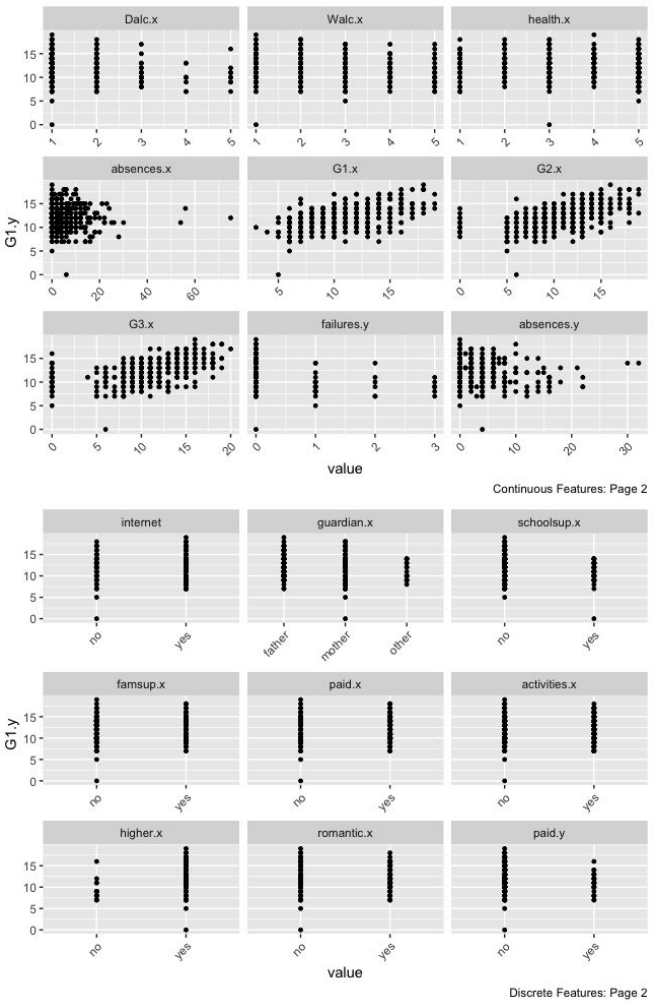
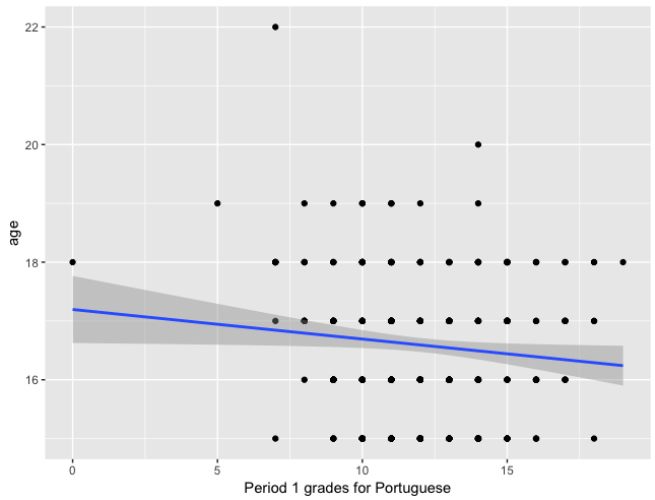
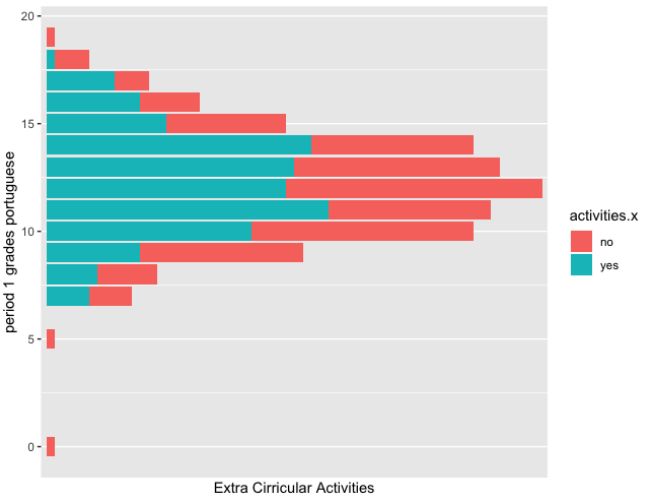
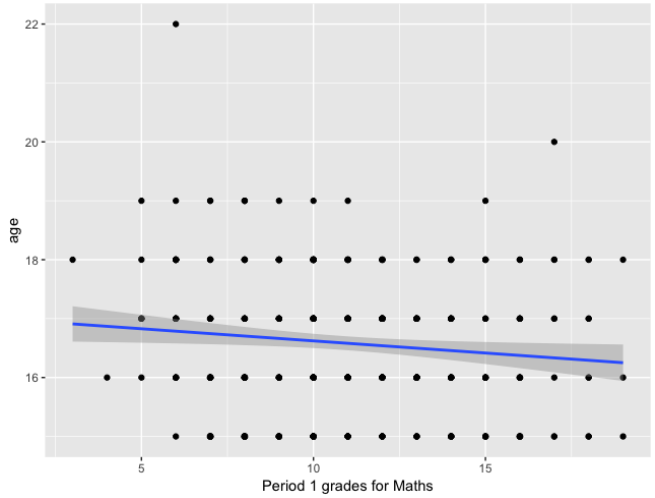


Fig 14 Scatter plot for portuguese vs all the attributes







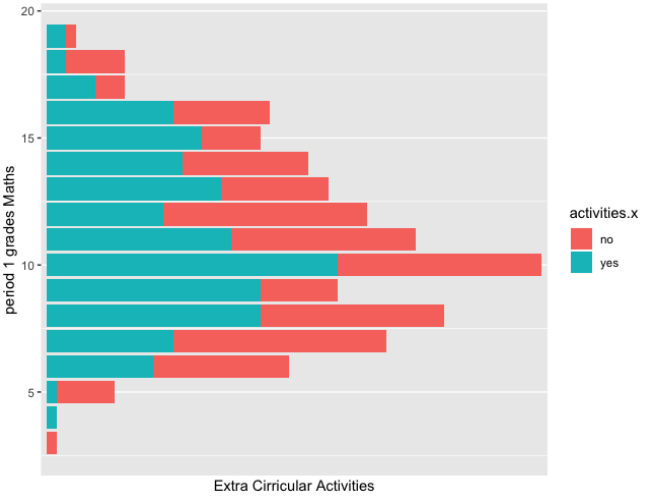
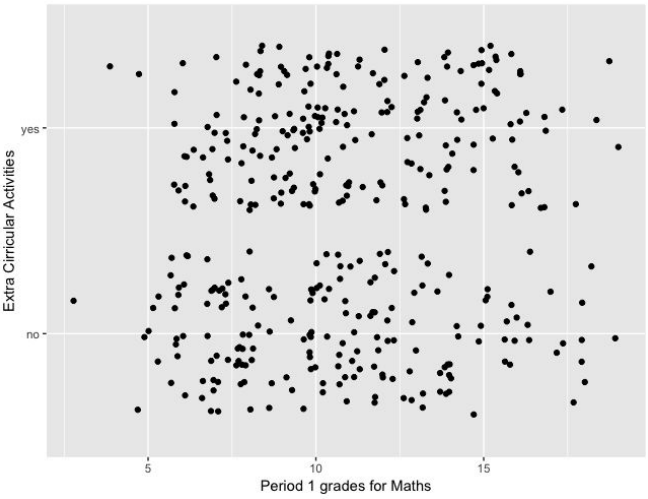
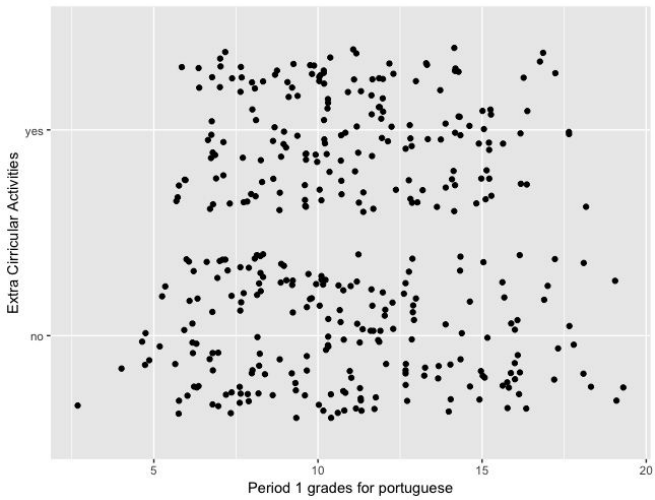
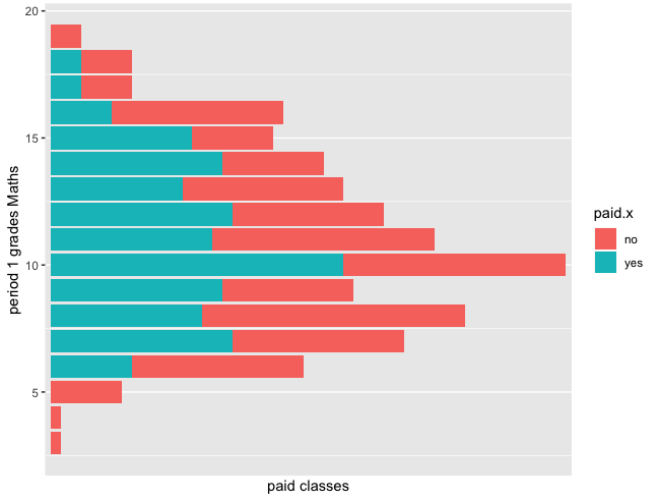


Fig 15 age vs period 1 grades for maths and portuguese respectively.







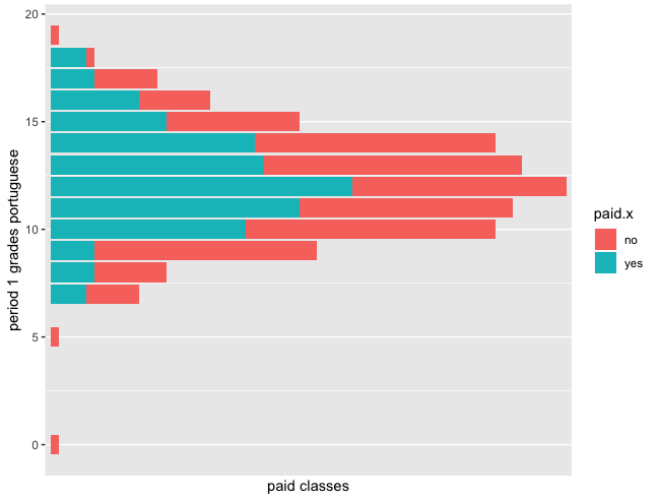


Fig 16 : Jitter and histogram for paid classes and extra cirricular activities vs first period grades.

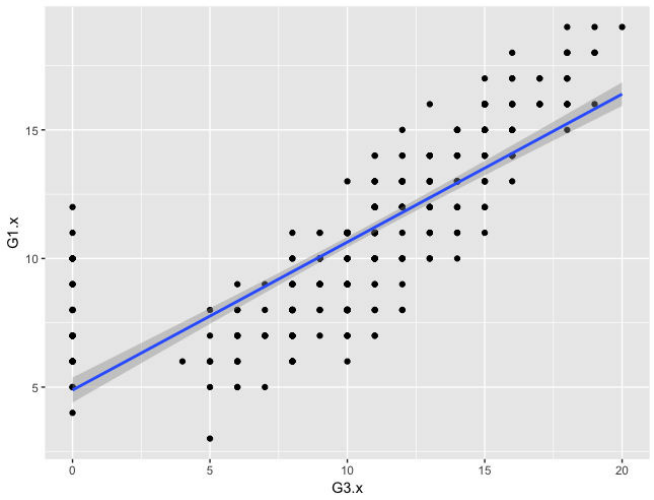


fig : 18 first period grade vs final period grade.

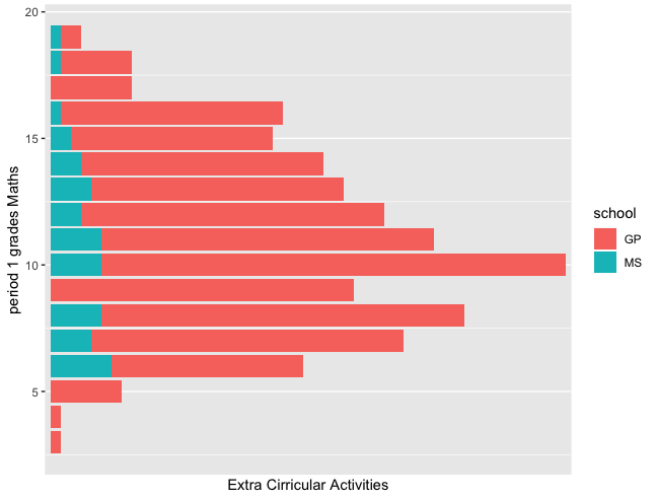
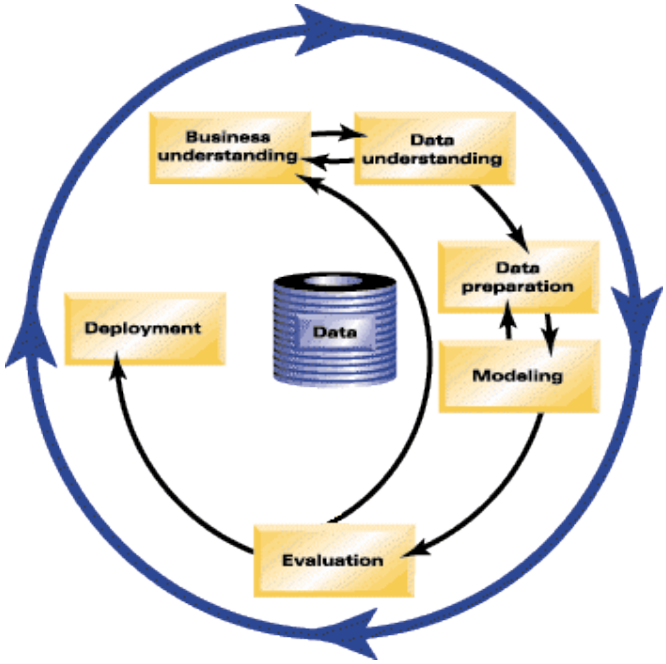


Fig 19 : First period grades vs school.

APPENDIX:



Information : This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

The above description is taken from the above mentioned source

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 â€" 5th to 9th grade, 3 â€" secondary education or 4 â€" higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€" 5th to 9th grade, 3 â€" secondary education or 4 â€" 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)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)