**Team Phoenix**

**Track B: Analytics using Python**

*By* Amey Rode

Harshal Malode

Harshil Bhorawat

Hetvi Hatadia

Karim Samnani

Vikas Verma

**About Dataset**

Attributes:

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 (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)
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)

**Analysing difference between males and females’ grades:**

* The variables 'school', 'address', 'higher', 'romantic' and 'Internet' influence both females and males’ grade in the same way. Specifically, on average the students that attend Gabriel Pereira (GP) school have better grades with respect to the Mousinho da Silveira (MS) students. The students who live in an urban zone have better grades with respect to the students that live in Rural zones. The students who plan to continue their studies have better grades with respect to those that do not, and the same thing happens with the students who have not a romantic relationship and those who have one. Lastly, students who have an Internet connection have better grades with respect to those that do not have one.
* The variables 'Pstatus' and 'guardian' influence females and males’ grade in different ways. Indeed, the females who have parents that are together have better grades with respect to those who have divorced parents, while for the males the opposite is true. Similarly, the females who have a guardian different from a parent have worse grades then the females who have a parent as the guardian, instead the males’ grades do not seem to be influenced by this variable.
* All other considered variables do not influence the grades significantly.

**Analysing how the weekend and the workday alcohol consumption influence the school performances of the students:**

On an average the grades(G3) are influenced by the alcohol consumption. In particular, an higher alcohol consumption is associated with a lower school performance.

**Analysing if there is any difference between the influence of the alcohol consumption on the males and on the females:**

Males who have a lower alcohol consumption have better grades in general, while for the females the situation is more complicated. Indeed, we have that the females with the highest workday and weekend alcohol consumption, i.e., Walc and Dalc equal to 5, perform better than the others. Probably, since the dataset is small, this is accidental, namely the considered females with a high alcohol consumption are particularly talented or particularly suited for the subject. Moreover, delating these females with the highest alcohol consumption, also for the females we observe an increase of the average grade with the degrowth of alcohol consumption.

**Analysing how the parent’s job and education influence the school performances of the students:**

Both females and males having a teacher as a parent significantly increase the grade. It is interesting to notice that for females having a parent who works into the health sector increase the grade too, while for males that is not true. Furthermore, for both females and males the average grade increases as the level of education of the parents increases. We have an exception for the males who have a father with a level of education equals to zero, but it is plausible that even in this case, as it happens for females with the highest alcohol consumption, this is due to the talent of the latter.

A single parent education influence G3. Obviously, we expect that students having both parents with high level of education have better grades with respect to those who have only one. Let us consider a plot that allow us to visualize both parents’ education level:

Our expectation is met students with both parents well educated perform better than the students with a single parent who have a high level of education.

**Analysing how the quality of family relations, the amount of free time, going out with friends and the health influence G3:**

1. On average the students with a good level of family relations perform better than those who have a poor quality of family relations.
2. For females it is clear that on average the grades tend to decrease with the increasing of free time, while for men this variable does not substantially affect the grades.
3. On average students with the poorest level of 'goout' have the worst grades, while those who have a level slightly higher have the best grades. After that all the students with a higher level of 'goout' tend to have worse grades. Thus, students with a decent level of sociability perform better than those who have a terrible or excessive level of sociability.
4. The health slightly influences the performances of the students and in particular the average grades decrease with increasing of the health level. Probably this is due to the fact that students with a poor level of health have a lower level of alcohol consumption and a lower level of sociability. We will see in a while how the health of the students influences these variables.

**Analysing how the last three variables, i.e., 'failures', 'age' and 'absences', influence the latter. We start from 'failures' and 'age’:**

The median value of G3 is not influenced by the age if the latter is included between 15 and 18, even though the distribution tend to be wider by increasing the age. Conversely, the median value of G3 decreases for ages higher than 18. With regard to the number of failures, the median value of G3 decreases with the increasing of the number of failures.

**Conclusions:**

According to the provided EDA the most important features for the prediction of G3 are 'studytime', 'higher', 'failures', 'goout' and 'Walc'. Moreover, 'failures' variable is by far the most important one.