Machine Learning

Hackathon Project Report

Student Performance Prediction

The Dataset

Source : https://archive.ics.uci.edu/ml/datasets/Student+Performance

Columns :

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)

We will predict the final grade of the student.

31. Grade - average grade of G1,G2 and G3

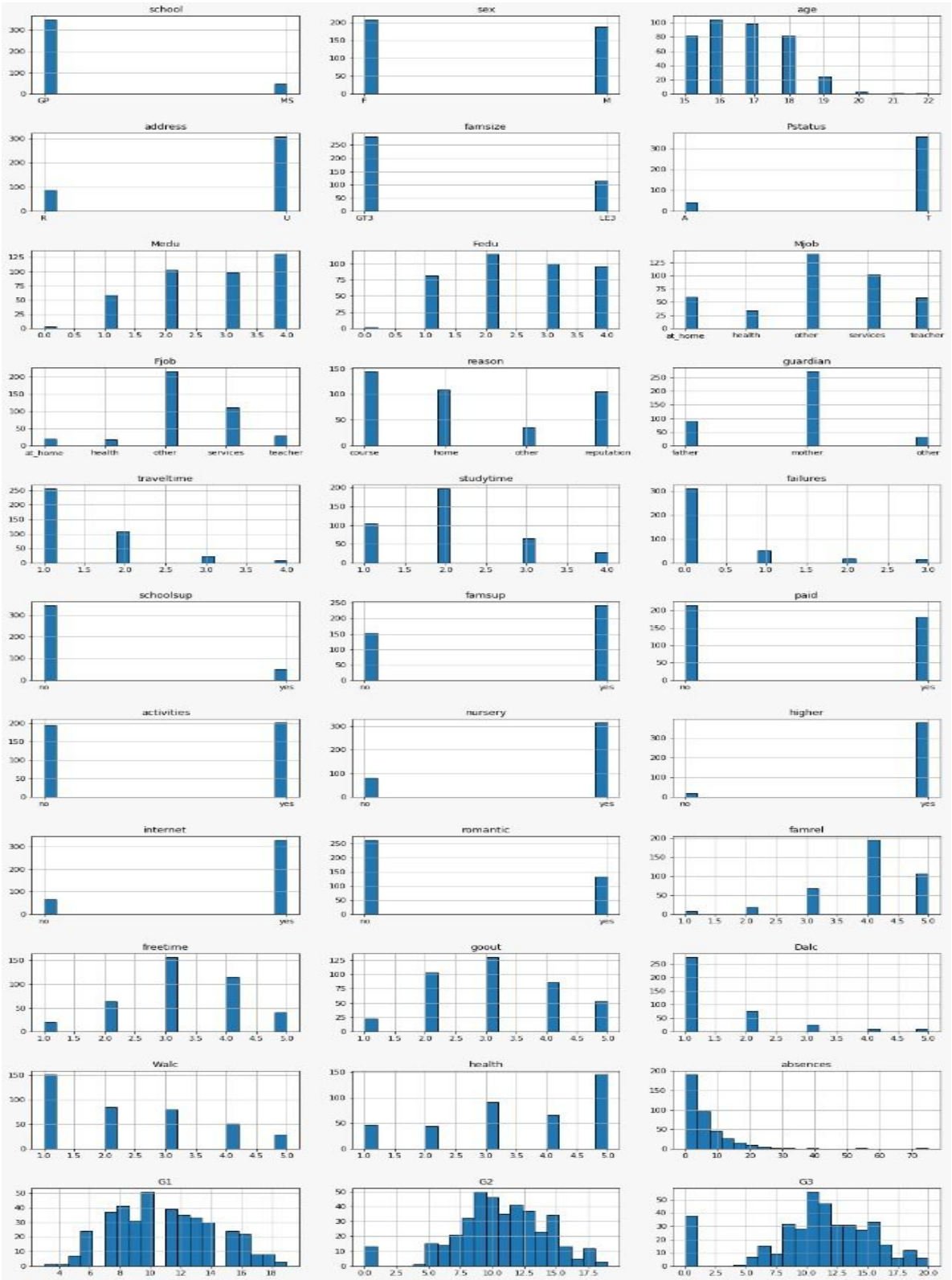
Preprocessing Data

1. View data distributions
2. Identify skewed predictors
3. Identify outliers

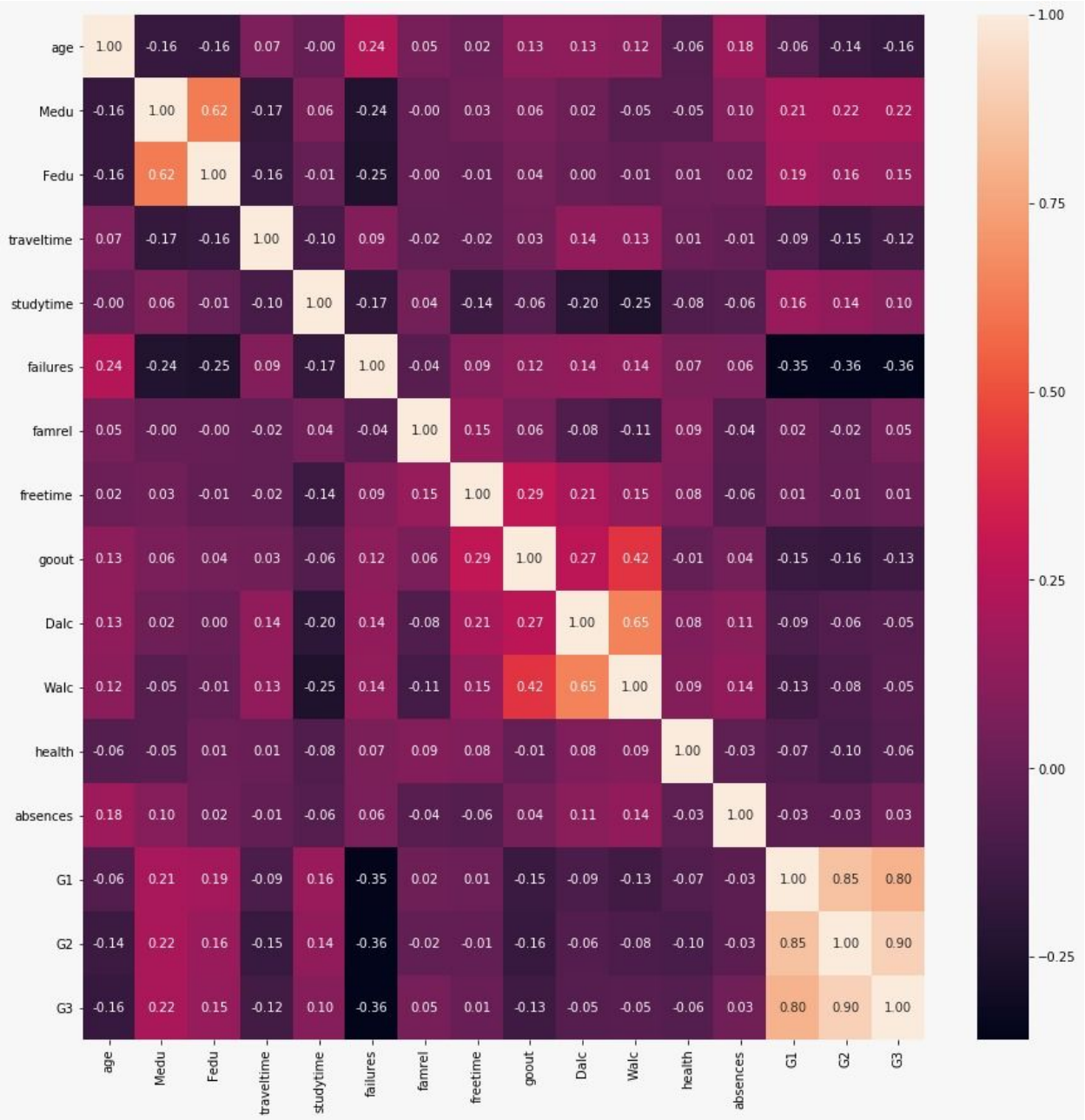
We checked how many null values are present in each column and cleaned up the data. We printed the count plots of each column value.

1. School : We saw the distribution is highly skewed as there are only two schools with 88% of schools being of the same label ‘GP’
2. Age : Age is distributed between 16-22, with only five data-points in 20-22
3. Sex : Evenly distributed
4. Family Size : Most families have more members than 3
5. Mother’s and Father’s Education : mostly evenly distributed
6. Traveltime : Most of the students live near the school they choose
7. Failures : Data is highly skewed with most of the students not failing at all.
8. Schoolsup : No for most of the students.
9. HigherEdu : Yes for most of the Students
10. Walc : Almost even distribution
11. Dalc : Highly skewed, most students not drinking on week day

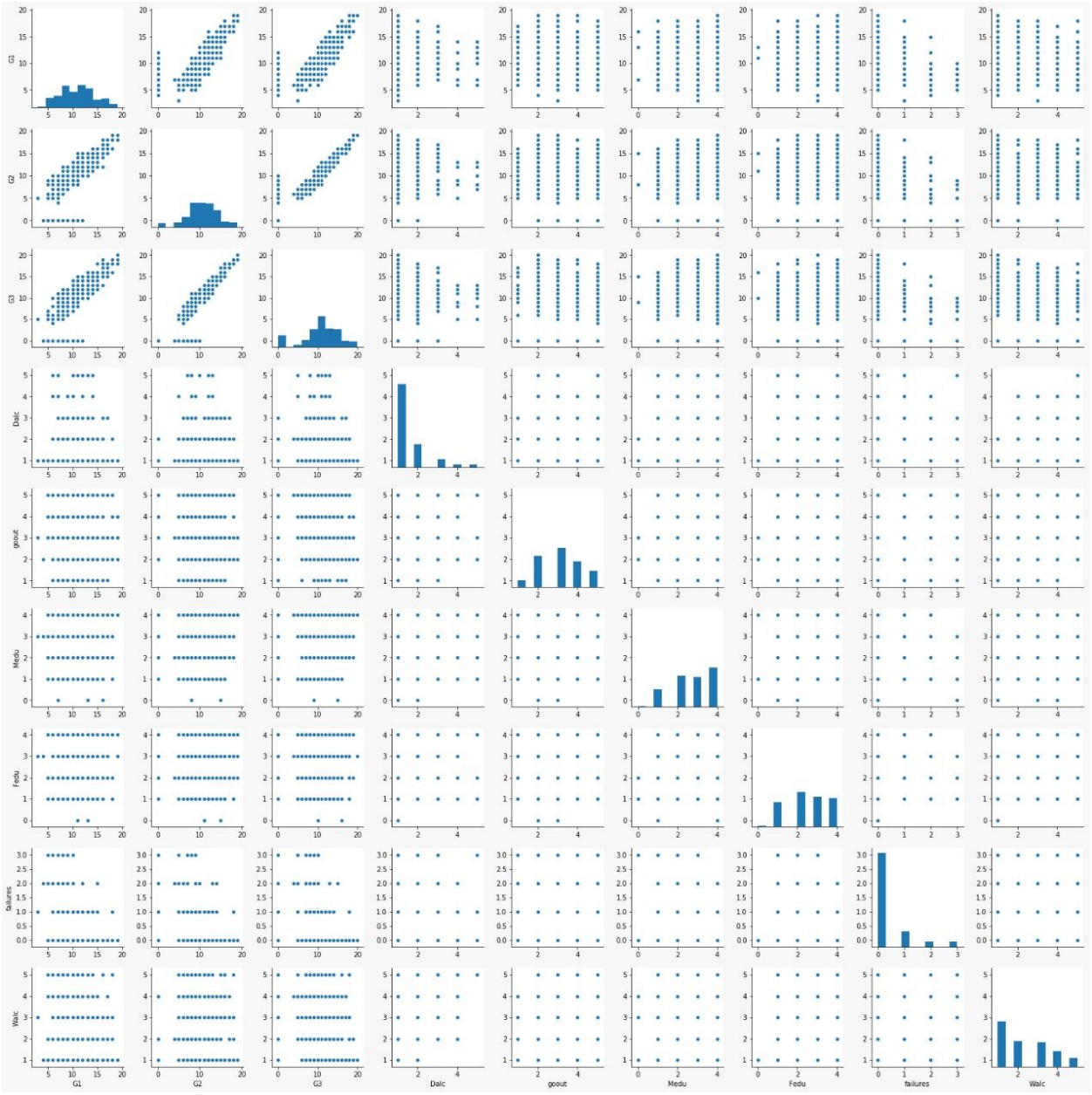
We plotted each column against the final grade ‘G3’.



We also plotted a heat map of all the columns against each other.



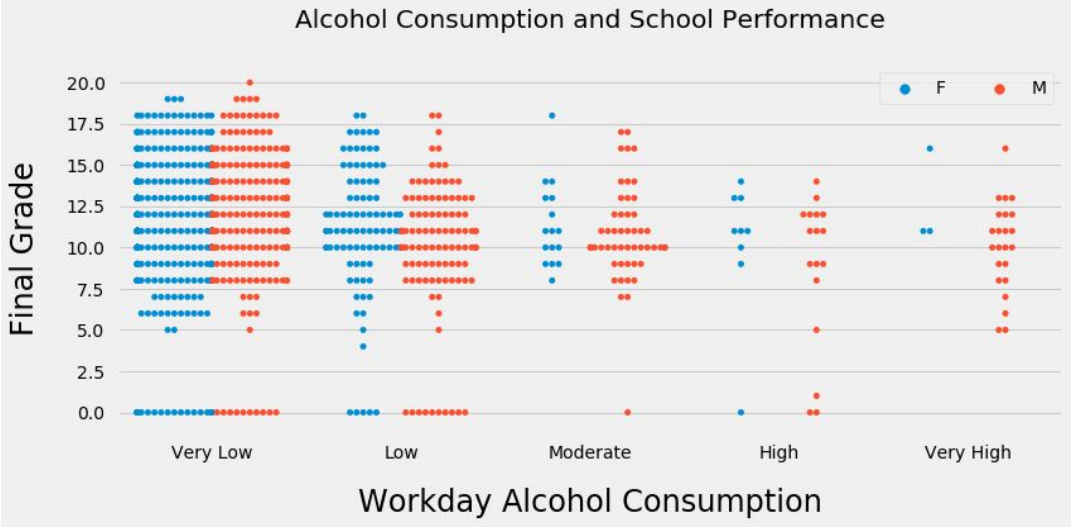
Pair Plots :

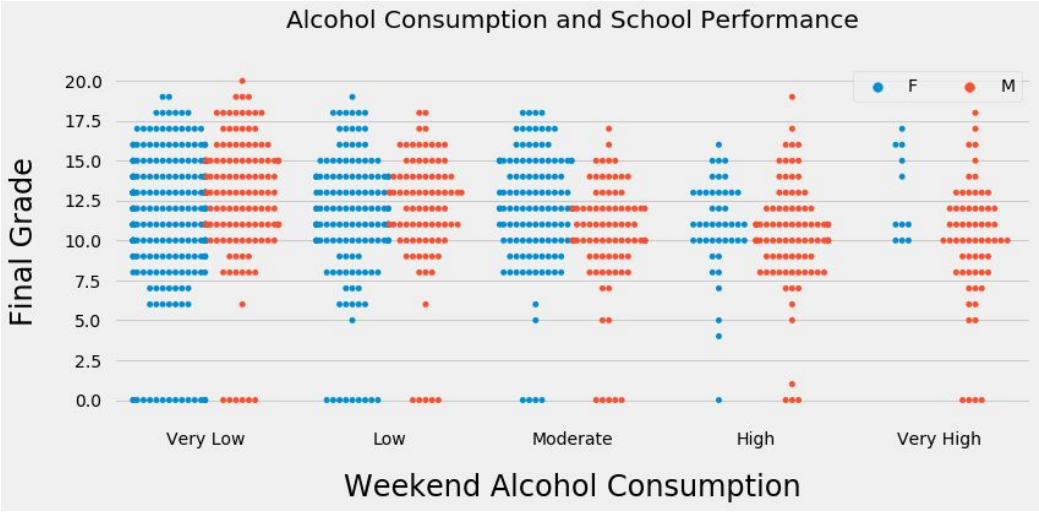


Here we observe that G3 is :

1. heavily depended on G1 and G2.
2. Very weakly depended on walc, dalc, freetime, absences and health
3. Mildly dependent on Fedu and failures.

Alcohol Consumption and Grades :





So we notice that alcohol does not affect the student grades. The grade distribution remains the same with alcohol consumption. Seeing our previous count plots we see very less number of people have very high alcohol consumption so that makes our data a bit sparse. Still we have the same distribution among the grades

Training a Model

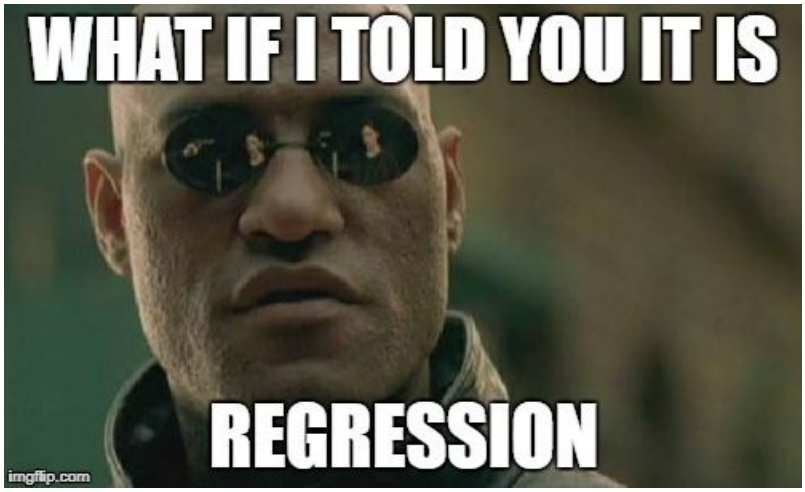
This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).

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.

Apparently, the variables G1, G2, and G3 do not show big relationships with the rest of the variables that we have. This could be due to the small sample size, or maybe we would need to collect data from some other variables. Either way, we'll have to change our approximation to this dataset if we want to extract some model.

In the correlation analysis we noticed that the grades (G1, G2 and G3) have strong correlations between them. This indicates that someway the students who have big grades on the first period (G1) use to have big grades on the second period (G2) too.

Now training the model !!



Selecting a model :

Naive Bayes (GaussianNB) :

Strengths :

- Naive Bayes performs well when inputs are independent from one another, and when there are many features and few parameters so it's hard to overfit
- Fast training speed
- Good at seperating signal from noise
- Able to handle many features

Weaknesses :

- Naive Bayes performs badly for problems where probabilities are sometimes greater than zero for a particular classification
- Simplistic nature can yield bad generalizations or suboptimal solutions
- Doesn't account for interactions between features
- There is a high bias when there is a small amount of data

Applicability : Our student data has a lot of features - 30 in total - which could influence the pass rate. Naive Bayes could be a good candidate on this basis.

Decision Trees

Strengths :

- Decision trees perform well with certain boolean functions, and when the model can be built with smaller decision trees rather than bigger trees.
- Fast training speed
- Easy visualization

Weaknesses :

- Decision trees perform badly if the tree grows quickly, and can overfit
- Trees can grow very fast
- High memory footprint
- Overfitting happens very easily
- Generalize relatively poorly

Applicability : Considering we have a large number of "yes"/"no" (i.e boolean) features in our dataset, this is well suited to a decision tree. Moreover, we only have 1044 data points, there is not a high risk of intense memory and CPU usage.

SVM

Strengths :

- Support Vector Machines perform well when there is a definite distinction between two classes
- Fast training speed
- Good for feature-intense data, e.g. text

Weaknesses :

- Support Vector Machines perform badly when the classification problem is not binary
- Doesn't take into account all data equally, so can't see overall trends in data
- A good choice in kernel is required, which is not always clear
- Slow in test phase

Applicability : Considering we have a binary classification - pass or fail - Support Vector Machines become a good candidate as a model.

We are using F1 Score :

This score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy. Especially as we have an uneven class distribution f1 score would be a more useful measure.

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Model :		Naive Bayes	Decision Tree	Logistic Regression	SVC	Random Forest
With G1,G2 :	Train	.90344	1.0	.95948	.97184	1.0
	Test	.87	.93627	.94527	.93203	.94607
Without G1,G2 :	Train	.85809	.99753	.89392	.90022	.99754
	Test	.81188	.81999	.86936	.88017	.86723

Best Model :

With G1 and G2 : Logistic regression gives best result as G1 and G2 are heavily co-related with G3. We are focusing more on without G1,G2 as it is more useful and gives a better insight to the data.

Without G1 and G2 :

Svc : Best result with testing data.

Decision Tree is overfitting the training data and giving lower accuracy with test data.

Random forest gave the best result with 50 trees and with minimal increment in accuracy after increasing more trees. It is also overfitting the training data but does well with test data as well.

GaussianNB delivers a very un-robust and poor score. Clearly, Support Vector Classifier has the most robust and strongest model going forward.

Team : Phoenix

Rishikesh Chaumal - IMT2016131
Abhinil Agarwal - IMT2016015
Saurabh Jain - IMT2016098