3. Problem Definition and Initial Exploration

3.1 Task Definition

Having analyzed how online and offline measures of student engagement can affect our predictive power on student performance, we now aim to dive deeper into the problem. We decided to look for a new dataset, in which we could analyze more aspects of students' behavior and environment, and test the effect these had on our model performance. Ideally, we aimed to find the attributes that distinguished high-performing students, and see if these could be adopted by other students to boost their academic outcomes. Also, by being able to build a robust model, we could provide an objective way to classify students into different classes according to their level (as is done in Korean high schools). This creates for homogeneous classes, where students' necessities are taken into account, therefore possibly maximizing learning efficiency. All in all, our research project would help students understand what factors impact their performance, and schools in creating tailored programs for students to incentivize a better learning experience.

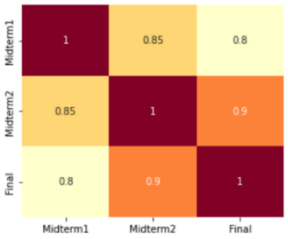
The dataset used is publicly downloadable at Kaggle[1], and includes information about student performance in 2 Portuguese high schools, both in Portuguese and math subjects. In this analysis, we will only focus on the latter. The features were collected through reports and questionnaires, and cover a wide array of aspects, ranging from social to school-related features. The Kaggle download site contains all the attributes and their descriptions.

As targets, we have also been provided the attributes 'G1', 'G2', and 'G3', which correspond to the first and second-period grades, and the final grade respectively. These are numeric values in the range 0 to 20. For this analysis, we will only be using 'G3'. This is because we are interested in predicting student performance at the beginning of the school year. Also, all of these values are highly correlated, so we can assume that student performance will be stable throughout the year. For both of the analyses in this research paper to be consistent, we will be transforming the numeric 'G3' attribute into a factor of 3 levels, 'L', 'M', and 'H'. These represent low, medium, and high achieving students respectively.

```
- Low level (0-69%)       [0  , 13]

- Medium level (70-89%)   [14 , 17]

- High level (90-100%)    [18 , 20]
```

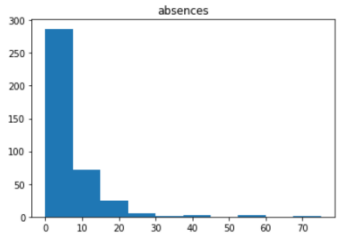


3.2 EDA

Before diving into model building, we needed to understand our data and identify possible problems. We performed a series of diagnostic plots, to check the distribution of our data and understand the information each variable was conveying. For numeric variables, we made histograms, correlation plots, and pivot tables. For categorical data, we created bar charts and used pivot tables to see how these groups affected our target variable. All of this started uncovering interesting nuances in the dataset and gave us some ideas for future data engineering. Some of the **initial questions** we wanted to answer were the following: Does more '*studytime*' relate to higher achievements? Do students receiving support of any type perform better? Does alcohol consumption affect performance? Do students that go out more often have lower grades?

[1] https://www.kaggle.com/datasets/barkhaverma/student-performance

Firstly, we analyzed the numeric variables, age, and absences. To grasp whether these 2 factors influenced student performance, we created a pivot table. Each row represented the different classes of students according to G3. The columns are the median number of absences and the average age of students in each group. We used the median for the absences column because there are many outliers in this feature, which would highly skew the group means.



G3	absences	age
H	4	16.333333
M	2	16.378049
L	4	16.806780

From the pivot table, we can see that age may be a factor that distinguishes students, where lower-achieving students tend to be older. However, we don´t know if this difference is significant, so we continue our analysis. Regarding the number of absences, no pattern seems to arise.

Next, we set out to explore the categorical variables. In the first instance, we noticed that many of the variables were very unevenly distributed. We created some pivot tables, to compare the groups across different performance levels. Due to the imbalances, we normalized the value per group, to better interpret the results.

In this step, we realized that some attributes added no value to our model. We opted to remove these variables from our model, as they seemed to only be adding noise.. We display some results here:

famsize	GT3	LE3
G3		
H	4%	5%
M	21%	20%
L	75%	75%

famsup	no	yes
G3		
H	5%	4%
M	21%	21%
L	74%	75%

paid	no	yes
G3		
H	5%	4%
M	21%	20%
L	74%	75%

However, other variables seemed to have some very discriminative values. These helped us formulate some interesting questions, and gave us some ideas as to how to reshape our features to be more informative. An example of this is shown below. If we look at the Fjob variable, we can see that the different categories don´t seem to be very different concerning the distribution between high, medium, and low students within them. However, this changes when Fjob = teacher. These students have higher prior probabilities of belonging to the higher achieving groupings. Therefore, our model may benefit from a single binary variable Fteacher (0 if the parent is not a teacher, and 1 otherwise).

schoolsup	no	yes
G3		
H	5%	0%
M	23%	6%
L	72%	94%

Dalc	1	2	3	4	5
G3					
H	6%	3%	0%	0%	0%
M	24%	15%	15%	0%	0%
L	70%	83%	85%	100%	100%

Fjob	at_home	health	other	services	teacher
G3					
H	5%	6%	5%	2%	14%
M	20%	28%	18%	22%	38%
L	75%	67%	78%	77%	48%

3.3 Feature Engineering

After EDA, we created a better understanding of our data and the relationship between variables and targets. We had formed some intuition about possible feature transformations that could be more informative.

1) Alcohol – The dataset contained 2 variables related to alcohol consumption, weekday and weekend averages. These were highly correlated, and therefore we could create a new feature, Average weekly alcohol consumption.

2) Support – In the previous data exploration, we saw that school support was the only support type that seemed to impact academic outcomes. Therefore, we create a single binary feature, *Support*, that indicated whether the student received school support (1) or otherwise (0). We can see below the new discriminative power of this variable.

3) Social Life – It seemed as if '*goout*' variable could be relevant if treated correctly. We hypothesized that it would be more informative if we measure the 'social life' by taking the ratio between '*freetime*' to '*goout*'. This way we could see how much the student was investing in going out relative to the amount of spare time he/she had. At first glance, it looked like the new variable didn´t bring great improvements.

4) Study – Similar to above, we hypothesized that study time would also be more informative when we took the ratio of it with respect to '*freetime*'. By doing this, we would be able to measure how much of the spare time students spent studying. This new feature looked very promising. There was a big distinction between the highest achieving group and the other 2.

5) Repeaters – Throughout the analysis, we noticed that age and failures seemed to be related to lower-achieving students. We decided to merge these 2 variables into a new one, Repeaters. Those students over 16 years old and with 1 or more failures were classified as so. They seemed to be more prone to belong to the 'L' group.

6) Guardian – Similar to what happened with the types of support, we saw that the guardian variable only had a significant impact on '*G3*' when its value was '*other*'. Therefore, we regrouped the values in this group into '*parent*' or '*other*'. Students where '$guardian = other$' seemed more prone to belong to the 'L' group.

7) Fteacher – As explained above, the 'Fjob' variable showed significant group differences between it being '*teacher*' versus any other value. Therefore, we created the new binary variable '*Fteacher*', indicating whether the father is employed as a teacher (1) or not (0). A positive response to this seemed to be related to better performance.

Support	0	1
G3		
H	5%	0%
M	23%	6%
L	72%	94%

Social	
G3	
H	1.075000
M	0.993699
L	1.068588

Study	
G3	
H	0.972222
M	0.733333
L	0.726328

Repeater	0	1
G3		
H	5%	2%
M	24%	2%
L	71%	96%

Guardian	other	parent
G3		
H	3%	5%
M	6%	22%
L	91%	73%

Fjob	other	teacher
G3		
H	4%	14%
M	19%	38%
L	77%	48%

4. Experimental Evaluation

4.1 Methodology

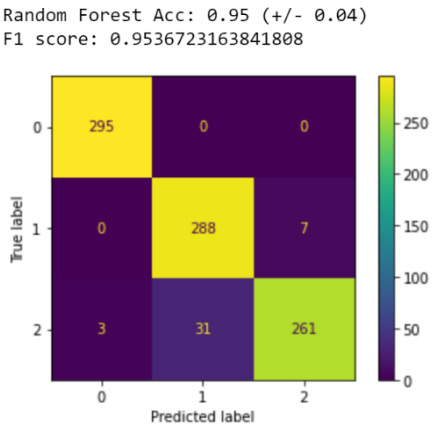
After feature engineering, we had our dataset ready to be modeled. To do this, we took into account that we were dealing with a multiclass classification problem. This impacted the algorithms we chose to compare (Decision Trees, Random Forest, Gradient Boosting, MLP, and SVMs), and the metrics we used to evaluate performance (Accuracy and F1 score). We also decided to plot the confusion matrices to see in which part our algorithms were failing most and act accordingly.

When running the first set of models, we realized that class imbalance in our target was a big problem, so we included some rebalancing techniques in the pre-processing section. These were under-sampling, over-sampling, and SMOTE. We also tested out how feature normalization affected our results, as well as different types of feature selection (Low variance filter, Stepwise Backwards Removal, Wrapper Selection, or Univariate Selection).

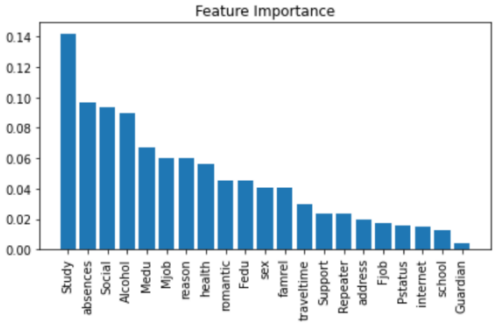
4.2 Results

Once we had the experimental setting ready, we ran each of the conditions and recorded the best achieving model, together with its accuracy and F1 score. This helped us locate the best-achieving model, a Random Forest. Oversampling was applied to the data, as well as using the Low Variance filter, and feature normalization. Model performance is displayed below. We must point out this analysis was conducted for the original data too, and the results were only slightly worse. However, this model involved more predictors and was more difficult to interpret (all features had low feature importance scores), so we kept the one with feature engineering.

Rebalancing	Feat_Sel	Norm_Features	Best Model	Accuracy	F1
None	LV	Off	SVM	0,76 (+/- 0,03)	0,76
		On	SVM	0,75 (+/- 0,03)	0,75
	1	Off	SVM	0,74 (+/- 0,01)	0,74
		On	SVM	0,74 (+/- 0,01)	0,74
	2	Off	SVM	0,75 (+/- 0,04)	0,75
		On	RF	0,74 (+/- 0,02)	0,73
	3	Off	SVM	0,75 (+/- 0,02)	0,75
		On	-	-	-
UnderSampling	LV	Off	RF	0,61 (+/- 0,32)	0,67
		On	GB	0,66 (+/- 0,35)	0,67
	1	Off	RF	0,61 (+/- 0,16)	0,63
		On	GB	0,57 (+/- 0,11)	0,57
	2	Off	RF	0,72 (+/- 0,24)	0,69
		On	RF	0,72 (+/- 0,24)	0,67
	3	Off	SVM	0,50 (+/- 0,08)	0,50
		On	-	-	-
OverSampling	LV	Off	RF	0,95 (+/- 0,05)	0,94
		On	RF	0,96 (+/- 0,04)	0,95
	1	Off	RF	0,93 (+/- 0,02)	0,93
		On	RF	0,93 (+/- 0,03)	0,93
	2	Off	RF	0,94 (+/- 0,04)	0,94
		On	RF	0,94 (+/- 0,03)	0,94
	3	Off	RF	0,74 (+/- 0,02)	0,75
		On	-	-	-
SMOTE	LV	Off	RF	0,89 (+/- 0,13)	0,88
		On	RF	0,90 (+/- 0,14)	0,91
	1	Off	RF	0,83 (+/- 0,14)	0,83
		On	RF	0,80 (+/- 0,17)	0,79
	2	Off	RF	0,90 (+/- 0,15)	0,91
		On	RF	0,88 (+/- 0,16)	0,89
	3	Off	RF	0,74 (+/- 0,15)	0,74
		On	-	-	-



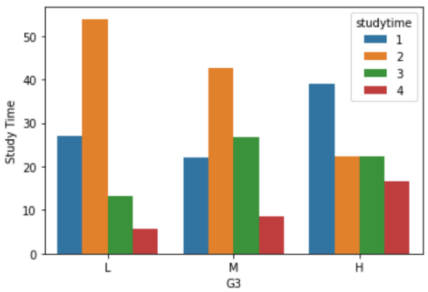
One of the advantages of using Random Forests is the fact that we can derive the feature importance's, to better understand how our model is taking decisions. For the model above, the results were very interesting, as it turned out many of the important features for classification were those that we had engineered (Study, Social, Alcohol…).



4.3 Conclusions

The results above give us a very interesting insight into which variables seem to be relevant to determining student success. For instance, we have collected evidence supporting some of the common heuristics (alcohol consumption and absences being detrimental to performance), while also challenging others (receiving support will improve performance). Now we are finally in a place to answer the questions we posed during EDA.

- *Does more 'studytime' relate to higher achievements?*



Although the proportion of students dedicating more time to studying is higher in the 'H' group than in the others, we can see how the majority of higher-achieving students don´t employ much time studying. However, when considering the ratio between time spent studying and free time, it becomes very apparent that this feature *is* discriminative between groups.

Therefore, we can conclude that study time itself is not informative. Rather, we should look at the proportion of free time spent studying. This is higher in students from group 'H'. We can therefore infer that even though from the above graph it looks like high achieving students spend less time studying, the reality is that they just have less time, but spend more of it studying. For example, we see the social life ratio is the highest in the 'H' group too. We can hypothesize this is making them more 'well-rounded' students and boosting their performance.

- *Do students receiving support of any type perform better?*

To answer this question, we created a new categorical variable, taking values from 0 to 4. It represents the type of support given to a student (0- No support, 1- family support, 2-school support, 3- paid support, 4- combined support). We can look at the distribution:

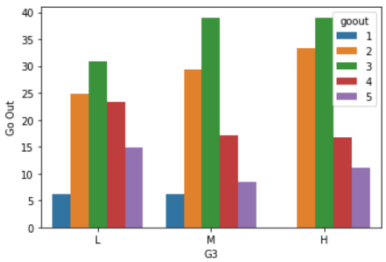
Support	0	1	2	3	4
G3					
H	6%	5%	0%	5%	4%
M	22%	25%	10%	23%	18%
L	72%	70%	90%	72%	78%

As explained before, only school support seemed to impact the student distribution across 'G3' groups. This explains our feature engineering choices above. We can conclude that receiving school support is associated with lower-achieving students. Regarding the rest of the support types, not enough evidence is available.

- *Does alcohol consumption affect performance?*

As seen in the visualizations above, it seems that those students with higher weekly alcohol consumption tend to perform worse.

- *Do students that go out more often have lower grades?*



It looks like the distribution between groups is similar. However, 'H' students seem to go our more, on average. This ties in with our conclusions regarding studytime. It suggests that higher-achieving students have a better balance between social, scholarly, and other activities.