Student Grade Prediction

TECHNICAL REPORT

Abstract:

This project revolves around the performance of Portuguese students from two different schools and their grades related to school performance. The key courses that the students are being graded on are math and the Portuguese language because in the past year's Portuguese students have been obtaining the lowest grades compared to the rest of Europe. And it was collected by using the school report and the questionnaires. Important note: The target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade, while G1 and G2 correspond to the 1st and the 2nd-period grade.

Introduction:

The universities are prestigious places of higher education, student retention in these universities is a matter of higher concern. It has also been found that students have dropped out from the universities during their first year due to the lack of support in the undergrad course. Due to this reason, the first year of undergraduate students is referred to as a "make a break year". Without getting any support for the course domain and its complexity, it may demotivate a student and cause withdrawal from the course. So, there is a great need to develop an appropriate solution to guide students retention at higher education institutions. Early grade prediction is one of the solutions that can monitor the student's progress in the course at the university and it will improve the students learning process based on the predicted grades.

Using data mining we can improve the process of the students. Different models can be developed to predict the student grades in the course in which they have enrolled in. By this, we can provide valuable information to student's retention in those courses. Through this information, we can identify students who are at risk and the instructor can suggest special attention to those students. Through this information, we can help in predicting the student grades in different courses to increase their performance in a better way and also to improve the student's retention rate at the universities. Using the various libraries such as Tableau, Seaborn, Matplotlib, SKlearn to represent the data using the different attributes graphically, to analyze the dataset for predicting the Final Grade(G3).

Technical detail:

Tableau Visualization:

Our goal with Tableau is to create graphs that can help us answer what aspects of the student lives are going to help raise grades. We are using Tableau because we can create simple and easy-to-read graphs at a glance. We later go into more details when we discuss our findings in the Data Preparation & Plotting sections of this report.

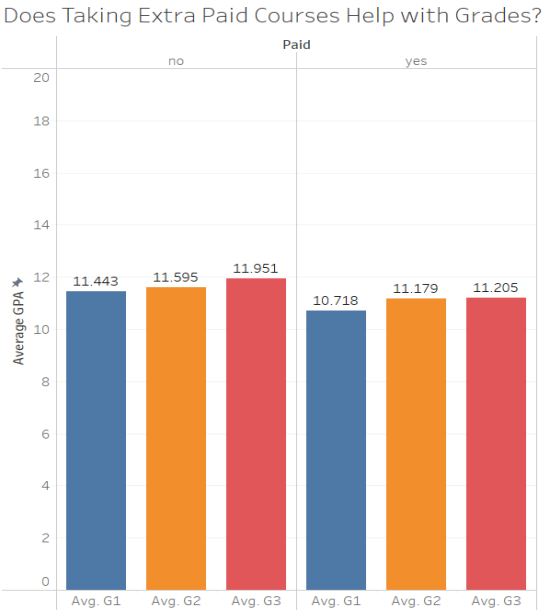


Figure 1.1

In Figure 1.1, we are using Tableau to visualize the effect paid courses will have on the student's grades over the course of the school year. Through our findings, we are able to see that students who do not take paid courses are scoring higher which is surprising. Even though the difference is minuscule, we can see that these extra courses are not the solution to fixing the high failure rate amongst the students.

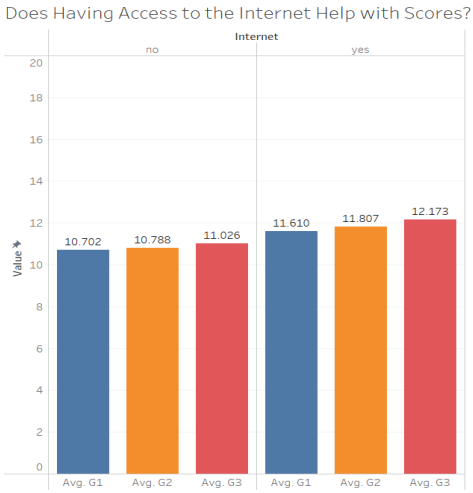
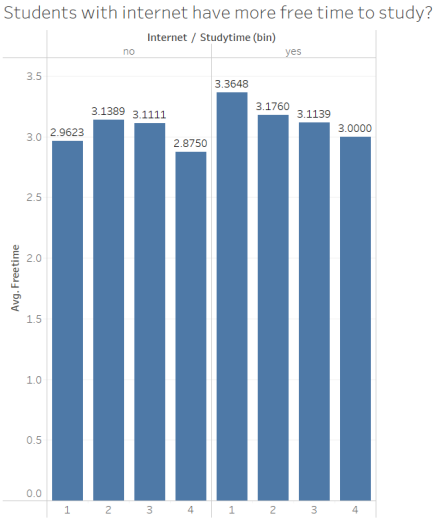


Figure 1.2

In Figure 1.2, we wanted to see if students with internet access would perform any better than those without it. Our findings concluded that on average, students with internet access performed slightly higher than those without. We believe this is the case because the students have more free time on average compared to those who do not. With this graph, we also asked if these same students on average have more free time to study, and our findings are shown below in Figure 1.3.



Data Preparation:

```
[ ] df = pd.read_csv('/content/drive/MyDrive/student/student-mat.csv', sep=";")

[ ] df
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	20	U	LE3	A	2	2	services	services	course	other	1	2	2	no
391	MS	M	17	U	LE3	T	3	1	services	services	course	mother	2	1	0	no
392	MS	M	21	R	GT3	T	1	1	other	other	course	other	1	1	3	no
393	MS	M	18	R	LE3	T	3	2	services	other	course	mother	3	1	0	no
394	MS	M	19	U	LE3	T	1	1	other	at_home	course	father	1	1	0	no

Figure 2.1: Classifying our dataset as "df" then outputting the dataset.

This outputted rows 0-4 and 390-394 of our dataset. We have rows school, sex, age, address, famsize, Fstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime.. and more. This holds the basis to our results.

```
[ ] df.columns

Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
       'Walc', 'health', 'absences', 'G1', 'G2', 'G3'],
      dtype='object')
```

Figure 2.2: Columns of our dataset

```
# 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 - 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)

# 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)

Additional note: there are several (382) students that belong to both datasets .
These students can be identified by searching for identical attributes
that characterize each student, as shown in the annexed R file.
```

Figure 2.3: Notes from the dataset provider regarding what each column means and correlates to

Plotting:

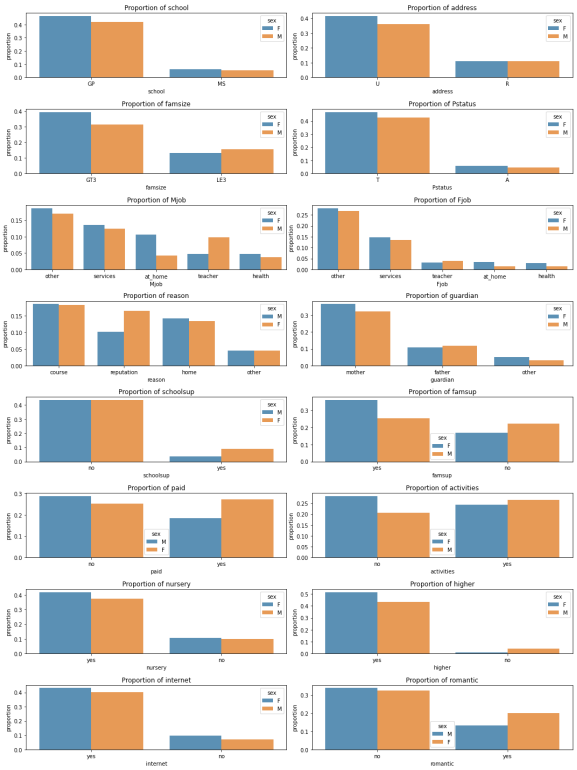


Fig:3 Illustrate the multiple portions of Data Set

This figure illustrates multiple proportions of different columns of the dataset, along with how many males/females there are for each category.

Proportion of school: which let us know the breakdown between male and female in each school. We can also see that GP school has a lot more students than MS.

Proportion of address: Shows us whether or not students live in an urban or a rural area. We can see that most students live in an urban area

Proportions of M job: What the student's mother's job does - we have other services, at_home, teacher, and health. Mainly we fall in other & services

Proportions of Job: What the students fathers job do - we have other services,

at_home, teacher, and health. Mainly we fall in other & services.
Proportion of paid: Whether or not paid tutoring helps grades
Proportion of activities: Whether or not students take activities
Proportion of romantic: If the students are in a romantic relationship or not - the majority of them do not

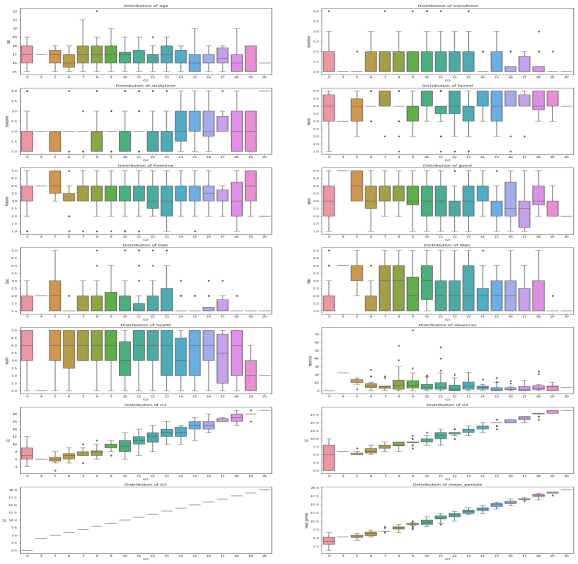


Fig:3.1 Final Grades

This is a distribution of each attribute based on their final grade value, so it gives us a more spread out view. For example, here on family relationships. We have the distribution of family relationships based on their final grade value and we see that poor family relationships output poor grades, with some outlier expectations. With the outliers mostly being in range of below average grades, with some exceptions are still in the high grade value.

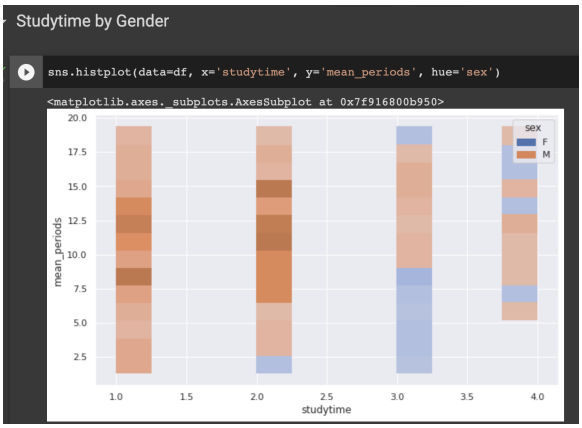


Fig:3.2

This plot illustrates that boys spend a smaller amount of time studying compared to girls.

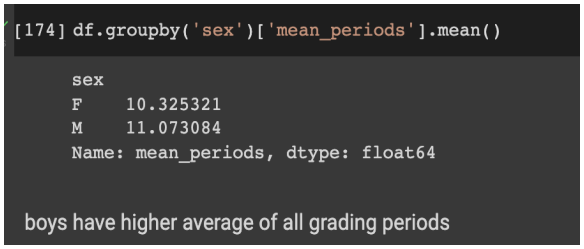


Fig: 3.3

Although boys spend less time studying than girls.. boys have a higher grades

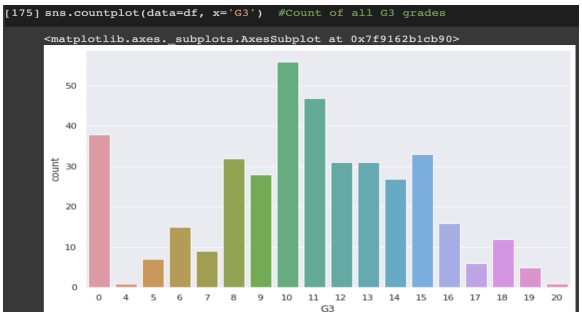


Fig:3.4: Count of all G3 Grades

We have the count of all g3 grades.. as you can see a big majority of them land in this 10-15 area.. and also actually a big portion being 0.

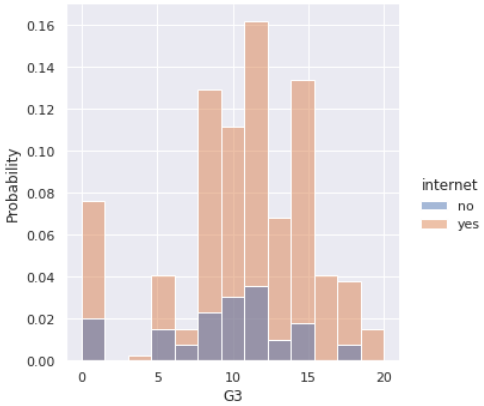


Fig: 3.5 Figure illustrating that Internet improves your marks

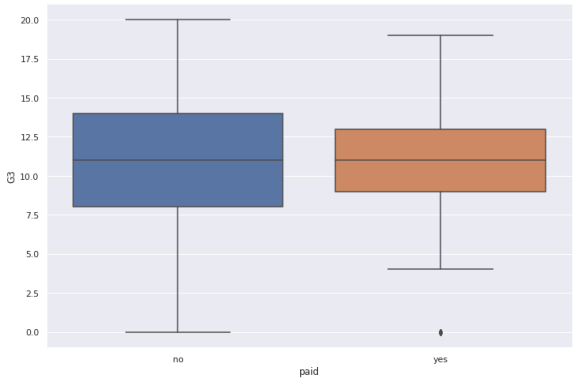


Fig:3.6 Students who don't take paid classes have better marks

Correlations:

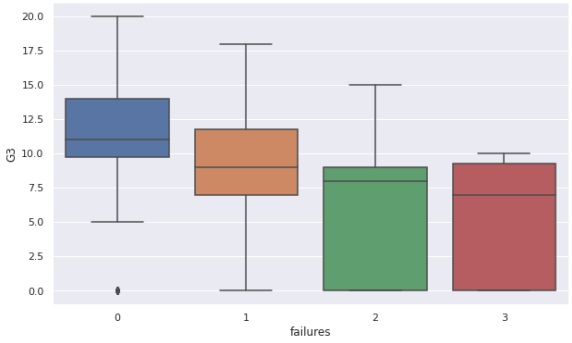


Fig:4.1 Students with less failures score more

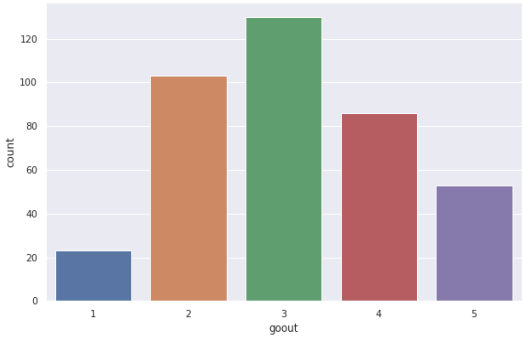


Fig:4.2

Moving on to going out, we actually see that students who go out on a scale of 3 from 1-5 score the best, compared to those who never go out or go out too much..

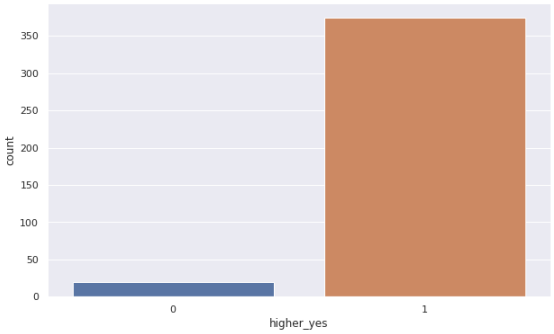


Fig:4.3 Students who want to take higher education score better

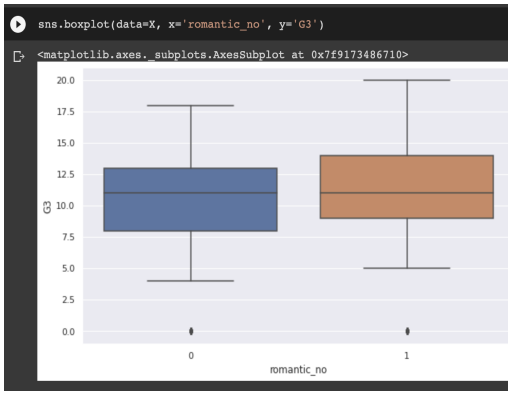


Fig: 4.4 Students who do not have romantic relationships score better

Data Mining Models:

```
[53] X_train = X_train.drop('G3', axis=1)
     X_test = X_test.drop('G3', axis=1)

[54] models_df =  pd.DataFrame(columns=['mae', 'rmse'])

[55] def make_report(models_df, model, X_test, y_test, name):
         report = pd.DataFrame(columns={'mae'}, data=[0])

         report['mae'] = mean_absolute_error(y_test, model.predict(X_test).round())
         report['rmse'] = np.sqrt(mean_squared_error(y_test, model.predict(X_test).round()))

         report.index = [name]
         models_df = models_df.append(report)
         return models_df
```

Figure 5.1

Figure 5.1 gives us our data preparation of our data mining models to report two different errors, Mean Absolute Error (mae) & Root Mean Squared Error (rsme).

```
models_df
```

	mae	rmse
Linear_regression	3.327731	4.338609
RandomForest	3.831933	4.868006
Decision Tree	4.563025	5.870450
Decision_Tree_GridSearchCV	3.436975	4.543903

Figure 5.2: Table representing four different data mining models and their respective errors

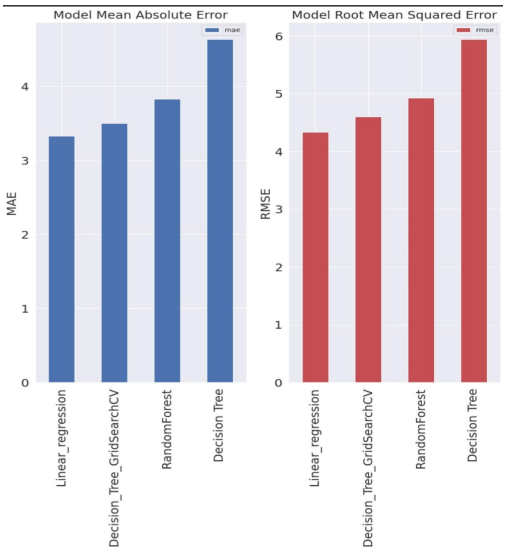


Figure 5.3

This figure shows the results in figure 5.2 in bar graph form. Here we see that linear regression gives us the best outcome in terms of lower errors.

```
models_df
```

	mae	rmse
Linear_regression	3.327731	4.338609
RandomForest	3.831933	4.868006
Decision Tree	4.563025	5.870450
Decision_Tree_GridSearchCV	3.436975	4.543903
Linear_regression_all_scaled	3.605042	4.666166
Linear_regression_all_unscaled	3.285714	4.227759

Figure 5.4:

We then run linear regression through a scaled and unscaled output but we are given different outputs for each.

	mae	rmse
Linear_regression	3.327731	4.338609
RandomForest	3.831933	4.868006
Decision Tree	4.563025	5.870450
Decision_Tree_GridSearchCV	3.436975	4.543903
Linear_regression_all_scaled	3.605042	4.666166
Linear_regression_all_unscaled	3.285714	4.227759
Linear_regression_all_unscaled_same_seed	3.605042	4.666166

Figure 5.5:

We then test the unscaled linear regression by matching its random seed state to the previous models and we now receive the same answer

```
[ ] plt.figure(figsize=(25, 14))

    # Root mean squared error
    ax = plt.subplot(1, 3, 1)
    models_df.sort_values('mae', ascending =
    plt.title('Model Mean Absolute Error', fo
    plt.ylabel('MAE', fontsize=20)

    # Median absolute percentage error
    ax = plt.subplot(1, 3, 2)
    models_df.sort_values('rmse', ascending =
    plt.title('Model Root Mean Squared Error'
    plt.ylabel('RMSE',fontsize=20)

    plt.show()
```

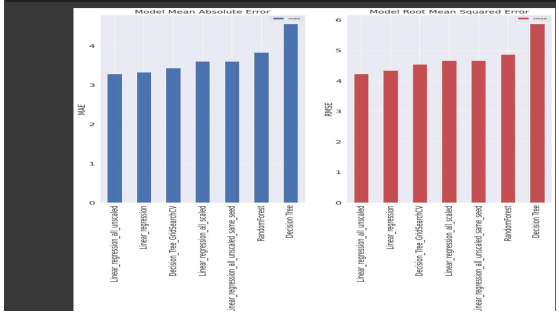


Figure 5.6:

This figure shows the results from figure 5.5 in bar graph format. Now we see that while linear regression is the best model.

Discussion & Conclusion:

Throughout our findings, we have narrowed down some aspects of the students' day to day life that can be changed to help lower the failure rate and increase the overall grades that are presented in the G1, G2, G3 periods. Some of these aspects include:

- Parents having a higher education.
    - We noticed that parents with higher education tend to have children who have some of the highest grades amongst their peers. This can be related to the fact that the parents are good role models for their children.
- Children wanting a higher education.
    - Children who want to pursue higher education on their own accord tend to do better than those around them. They have a drive and a passion for grades, which can tie into the previous aspect of parents having a good education.
- Students tend to go out of their homes less than 3 times a week, and those who do not have a relationship.
    - Students who limit their social interactions and outgoings tend to have more free time to study. These are the same students who tend to have internet access which allows them to access more information and pursue higher education. Not having a relationship at such a young age will help students focus more on their studies.

These are of course just the stepping stones to helping this population bounce back in the changing world. A lot of other factors can still go into and affect the outcome of the failure rate but with the aspects described above, the failure rate should be lowered sooner than later.