Project: Student Final Grade Prediction modelling – Machine learning

Student: Akeem Asiru

School: Eastern University

Project Summary & Problem Framing

Problem statement: Given a set of data that contains records of 395 Portuguese students with 35 attributes, create a regression model to predict a student's final grade performance.

Objective: To identify students in Portuguese secondary school that have poor academic performance and assist them in terms of counselling, guidance and support to improve their performance. This project aims at creating a regression model that predicts students' performance based on their personal and past academic information.

Goal: To predict a student's final grade performance by creating a regression model

Key Action:

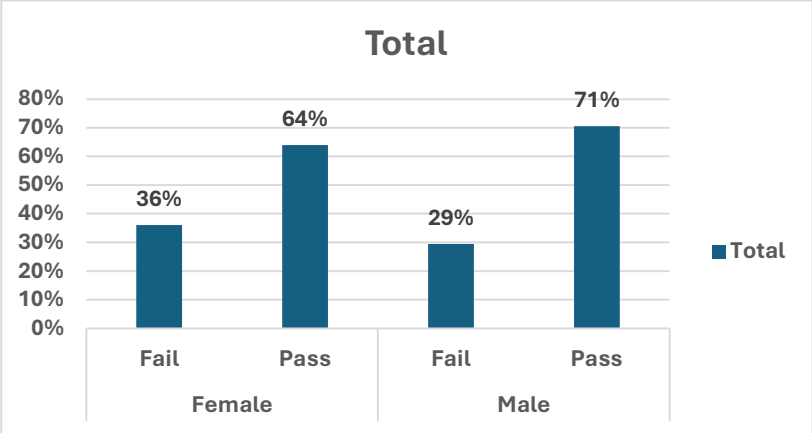
- Cleaning and preparation of the data to make sure it is suitable for analysis.
- Exploration of the dataset to understand the relevance and relationships among the attributes.
- Creating three regression models to choose the best performing model.
- Evaluation of the model's performance using relevant metrics

Linear regression was seen to be the best models (lowest rmse) out of the three models used i: e Lasso, Svm and Linear regression models. Multicollinearity relationship was established between some variable and this was taken into consideration during the analysis.

Data Overview

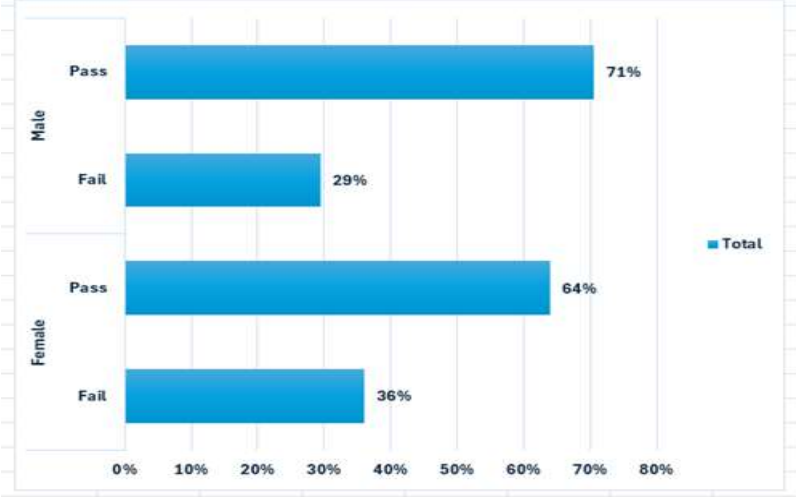
The data set consists of 35 features of 395 students from two different Portuguese schools i: e 88.3% of the students from Gabriel Pereira school and 11.64% of the student from Mousinho da Silveria.

Figure 1: Percentage of Pass/Failure in each School



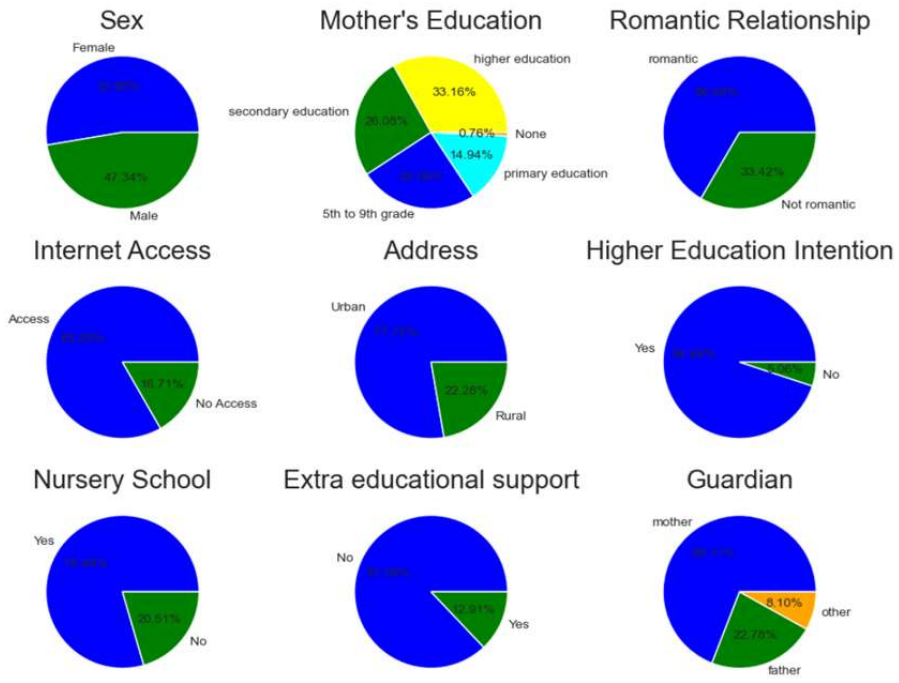
Observation: Irrespective of the size or number of students in the school the failure/pass rate is almost the same. It will be great to know if the school attended affects the final grade performance.

Figure 2: Percentage of pass/failure rate on Gender



Observation: The proportion of Female that fails are more than that of Male. It will be great also to know from our findings if gender affects the grade performance

Figure 3: Comparison of important student attributes



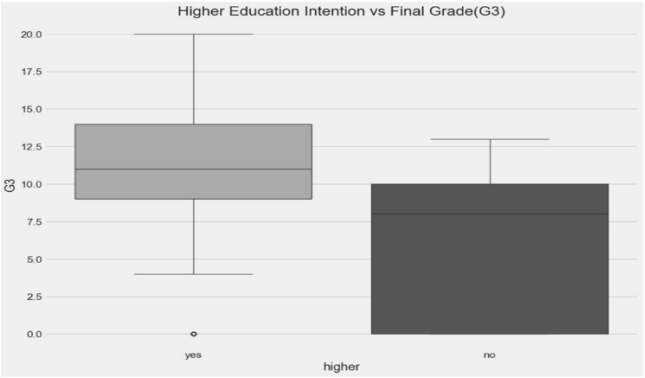
Observation:

- The number female and male students is almost the same
- Students that have extra educational support are many
- The students that have their mother as guardian are many among the students
- The students that attend nursery school are more than a quarter of the students
- Almost a quarter students have no access to the internet
- Majority of the students live in the Urban Area
- Few students have no intention of going to higher institutions
- The number of mothers who have no education is insignificant.
- 66.58% of the students are already involved in romantic relationships
- 16% of the students have no internet access connection.

Key Attributes of the students from the data set

1. Higher Institution Intention Feature: The proportion of students who have intention of going for higher education is 94%. Does this have any impact on the final grade performance of the students?

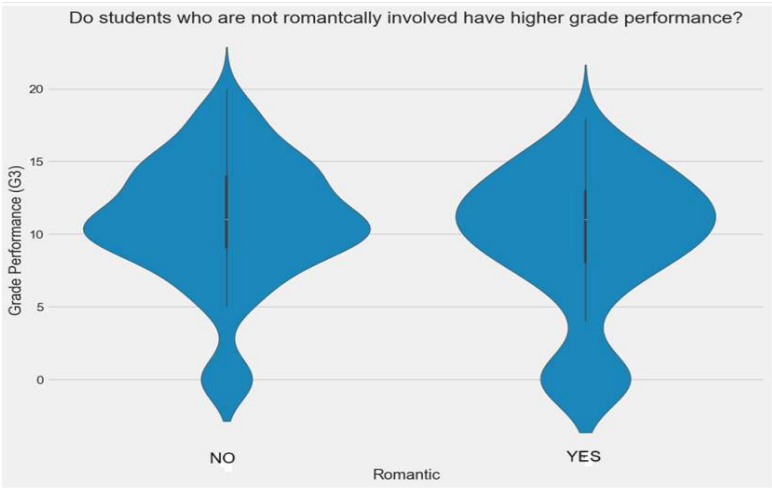
Figure 4: Student's Higher Education Intention vs Final Grade



Observation: Students that have intention of going to higher educational institution have Higher Grade performance than those students who do not have such intention of going to higher educational institution.

2. **Romantic Attributes**: It is helpful to know if the grade of those students that have romantic relationship affects their studies.

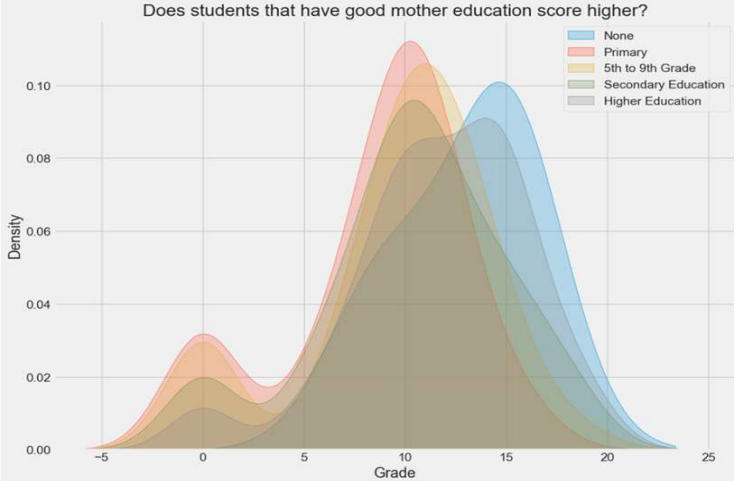
Figure 5: Romantically exposed students vs Final Grade Performance



Observation: It is observed that the students who are not yet in romantic relationships score higher final grade performance.

3. **Mother Education Attributes**: Since the mother is assumed to be student's home teacher, we would like to know if her education contributes to student's success in school.

Figure 6: Mother education vs Final Grade Performance

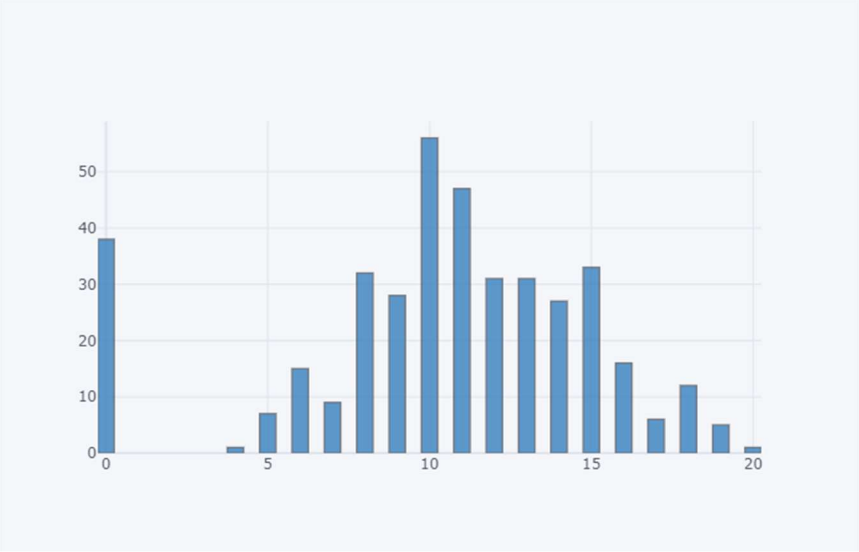


Observation: Good grade performance irrespective of Mother's Education

The target variable

The attributes we want to predict from other features in this dataset is the final grade performance, G3.

Figure 7: Final Grade (G3)

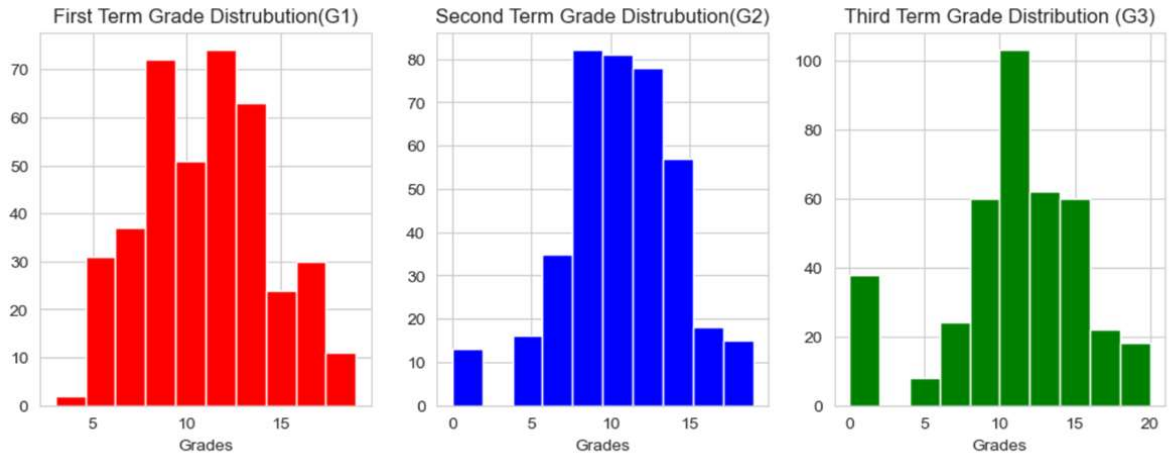


It is observed that the final grade performance of the students is normally distributed around the average but there are few students with extreme low scores (outlier) or failure.

Why model both with and without the G1/G2 grades?

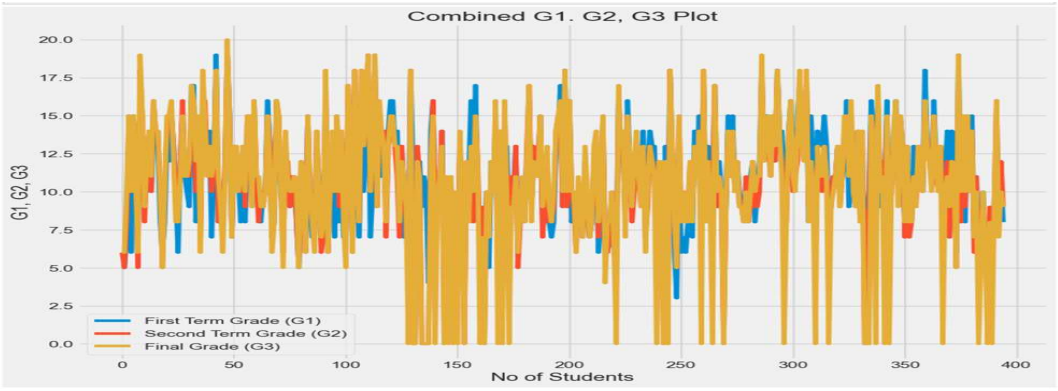
Our target variable (G3) has a very strong correlation with the first term grade (G1) and second term grade (G2) attributes. The variables plot is shown below:

Figure 6: Grades



When the variables are combined in a plot, the individual effects are difficult to isolate as they vary together. The combined effects are plotted below:

Figure 8: combined G1, G2, G3 (shows multicollinearity)



This plot shows that G1, G2 and G3 attributes have a very strong correlation which results in perfect multicollinearity.

Therefore, model with and without G1/G2: Two models will be created. One with G1/G2 attributes and the other without G1/G2 attributes.

Reason: Since G1/G2 attributes have perfect multicollinearity with our target variable G3, inclusion of both will increase the standard error of the coefficient of other variables. In other words, it will lead to misleading, vague, imprecise and erroneous results for our model. Therefore, we need two models to see the effect of this multicollinearity by including the variables G1/G2 only. While the model without G1/G2 is expected to be the model that is reliable.

Methodology: It is observed from the above analysis that Final Grade Performance, G3, has a linear relationship with the variables. It is either the variable has effect of increasing or decreasing Final grade performance. Therefore, a linear model will be used.

Regression is a data analysis method that attempts to predict the value of our unknown outcome attribute which is to be determined by using other known attributes called independent variables in our data set. It established a mathematical relationship or equation that relates the unknown outcome attributes to the know independent attributes in our dataset.

Key Results

Three models are used i: e linear regression, support vector machine (SVM) regression and Lasso regression models.

The metrics used to measure the performance of the models used are Root Mean Square (RMSE) and Mean Absolute Error (MAE).

RMSE explains the distribution of the error of the model prediction. The lower the error the better the model prediction. Out of our three models, the model with the lowest RMSE has a better fit.

R Squared – measures the proportion of the outcome variable that can be explained by all the independent variables. The higher the better.

	SVM Regression	Lasso Regression	Linear Regression
RMSE with Grades (G1/G2)	1.9719	2.1691	1.8773
RMSE without (G1/G2)	4.3454	4.3865	4.3413

The model with the lowest RMSE is Linear Regression.

I will implement only model without Grade. We need to know that the result of model with Grades is misleading as we have strong multicollinearity as explained above.