Predicting Students Final Grade in School, by Exploring Three Different Machine Learning Models.

ABSTRACT

Although the educational level of the Portuguese population has improved in the last decades, the statistics keep Portugal at Europe's tail end due to its high student failure rates. In particular, lack of success in the core classes of Mathematics and the Portuguese language is extremely serious. The present work intends to approach student achievement in secondary education using some machine learning techniques. 3 models used for this analysis are (1) The Regression model, (2) Decision Tree and (3) Random Forest. Although student achievement is highly influenced by past evaluations, an explanatory analysis has shown that there are also other relevant features (e.g. number of absences, parent's job and education, alcohol consumption e.t.c). The choice of model is because we are interested in predicting the student overall grade using some of the features in the data set. And since it's a regression problem, the aforementioned models are good for such task.

INTRODUCTION

Education is a key factor for achieving a long term economic progress. During the last decades, the Portuguese educational level has improved, however statistics keep the Portugal at Europe's tail and due to its high student failure and dropping out rates. In this work, we will analyze recent real-world data from two Portuguese secondary schools. ('GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira). Since the objective of study was to predict student final grade and if possible to also identify the key variables that affect educational success/failure in Mathematics, survey were carried out and some of the information retrieved were several demographic, social and school related attributes (e.g. student's age, alcohol consumption, mother's education). The two core courses involved were Mathematics and Portuguese language course. The structure of the data is as follows: The Mathematics course data had 396 rows with 32 columns while the Portuguese language course consist of 650 rows with 32 columns. Each row defines the student while the columns are the attributes. We have both the categorical columns like school, sex, famsize, Psatus, Medu, Fedu et.c). while some continuous variables are age, G1, G2, G3, absence, Dalc, Walc e.t.c. Since the aim is to predict students final grade, my responds variable would be G3 (final grade), while the remaining features are the explanatory variables.

DATA PROCESSING

The data was tested for missing variables using the descriptive statistic and it was observed that there was no missing variable in the dataset. The categorical variables were converted to factors, using as.factor() function in r. After which data was split into training and testing (75% and 25%) respectively, this is because, after training the different models, in order to measure success we need to evaluate or trained model on a text set. This would enable us compare results among the 3 models, and also know the performances of each model on new data set.

METHODOLOGY

In analysis, the benefits of visualization cannot be over emphasized. For this reason, we carried out some visualization mainly to answer these 2 questions (1) what are the age distribution of the students in both courses. (2) Do Girls Perform Better in School? The answers can be seen in the result section (EDA). Furthermore, when building models in Machine Learning, it is advisable to start with a baseline/ or a simple model. Hence, the first model used was a linear regression with all the variables. From the result, we noticed that most of the variables were not significant, hence, we decided to reduce the number of features by using only the numerical variables, and that give rice to model 1, a residual plot as well as the normal Q-Q plot was plotted to visualize and see how the model fit the data. The second model we looked at was the Decision Tree, when the **decision trees was built** many of the branches where reflecting noise or outliers in the training data. Hence, a technique called pruning was introduced. **Tree pruning** methods address this problem as well as overfitting (a situation where the model cannot generalize well on new instances). The last model explored was the Random Forest. This method Random is also called random decision forests, It's an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees. After all these 3 models were explored, we then make predictions using the set aside test set. The correlation between the predictions made and the original test set were estimated. Also, the metric use for evaluation is the RMSE, To select the best model, we compere the RMSE and the Correlation of each of the models, the model with the least RMSE and highest correlation of (predicted vs actual) is chosen to be the best model.

RESULTS

Exploratory Data Analysis

Table 1: Age distribution across the two course.

Courses / Ages	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Mathematics	82	104	98	82	24	3	1	1
Portuguese language course	112	177	179	140	32	6	2	1

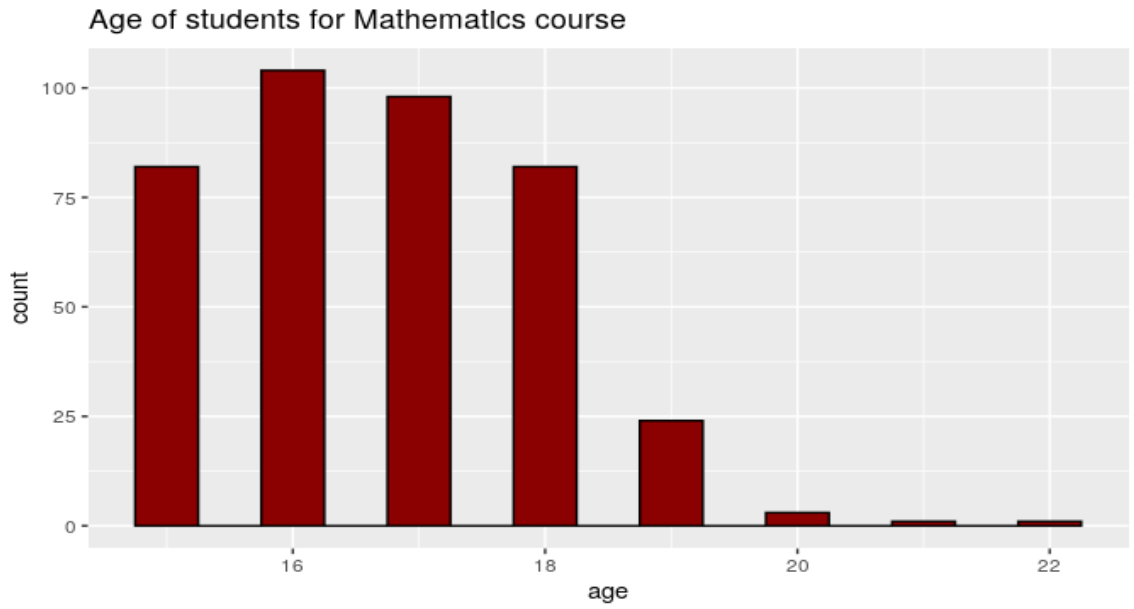


Figure 1

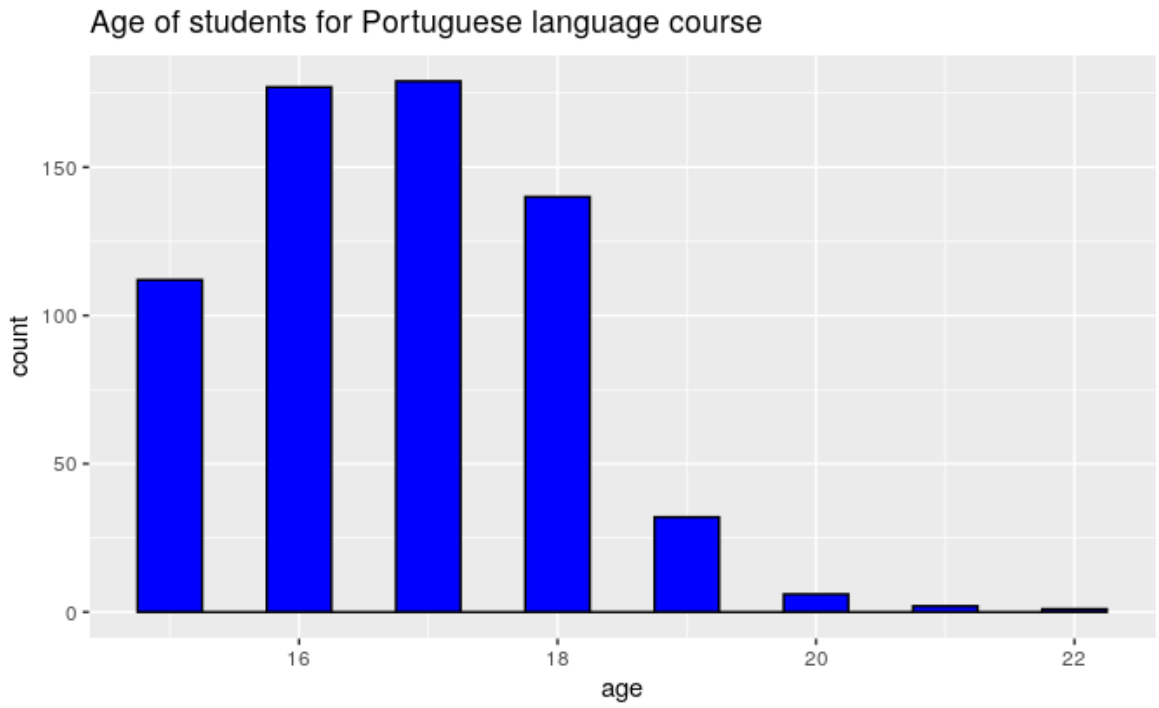


Figure 2

The table above as well as the plots shows the distribution of students age across the two courses. From the table, we can deduce that for Mathematics course, most of the students (104) are 16 years while for the Portuguese language course, most of them (179) were 17 years of age. Overall, the mean age for both course is 17 years. Since we have the same distributions for course, we can now proceed to answer the question bellow:

- Do Girls Perform Better in School?

We will explore gender differences in the classroom with respect to G3 (student final grade).

We will:

1) check the number of female and male students in the school.

2) Examine the performance in class based on gender and age by replying to the following questions:

a) Who does better at school? Do girls perform better or do boys gets better results than girls?

b) Does students performance gets better with age?

Table 2

Courses / Ages	**Males**	**Females**
Mathematics	187	208
Portuguese language course	266	383

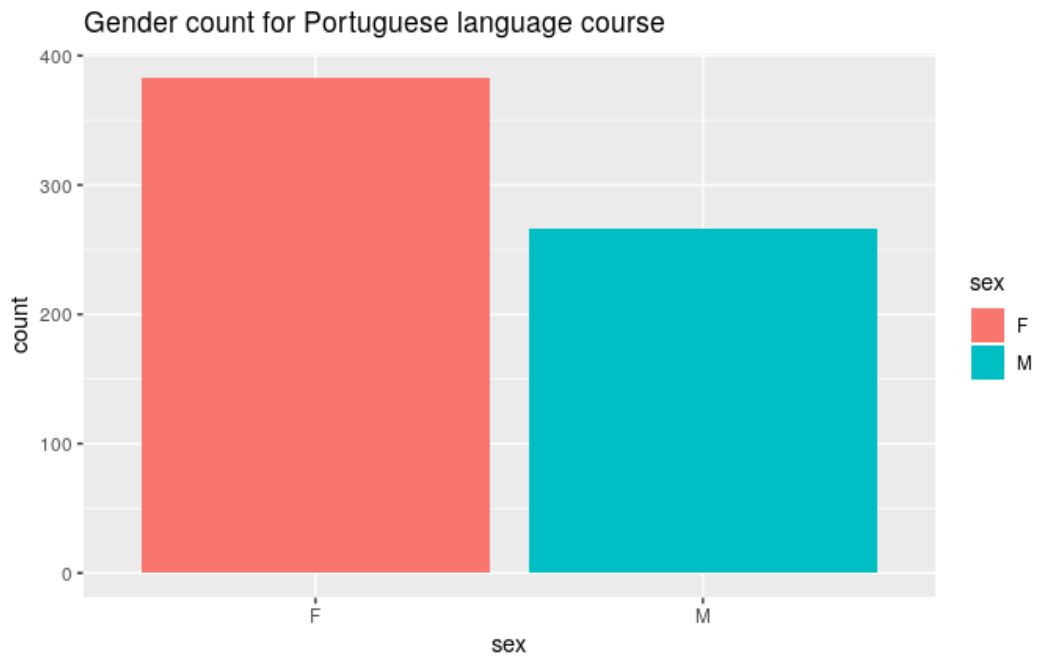


Figure 3

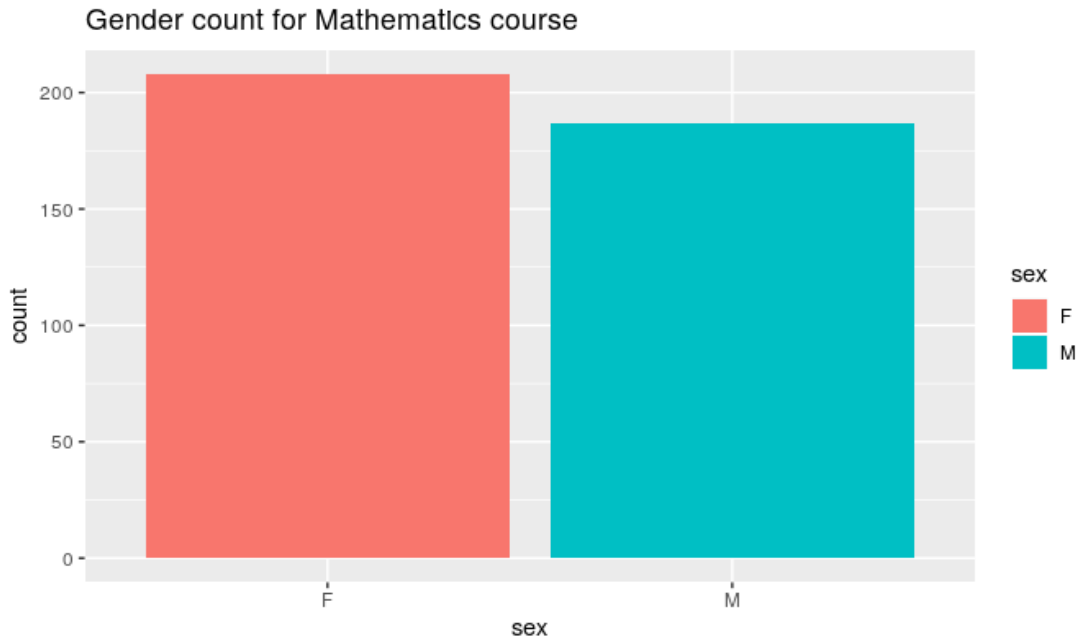


Figure 4

Table 2, Figure 3 and Figure 3 shows the count of students in each course. From the bar plots, it is obvious that we have more females and males in the two courses. However, the difference in that of Mathematics course was not as much as that of the Portugal language course.

a) Who does better at school? Do girls perform better or do boys gets better results than girls?

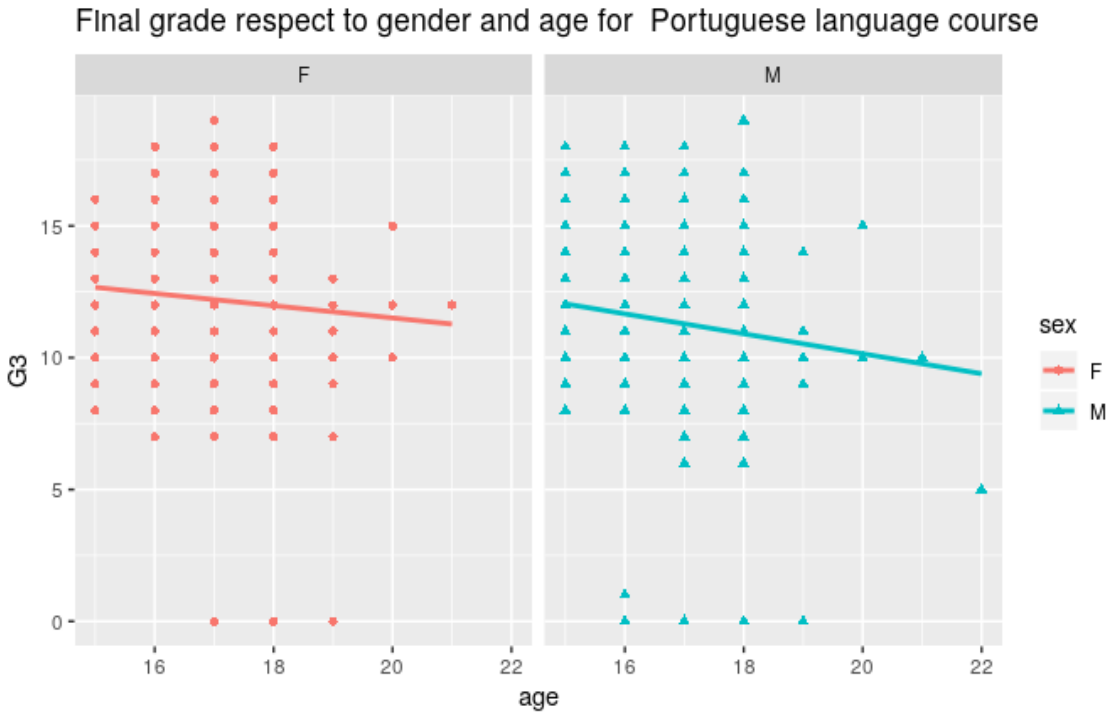


Figure 5

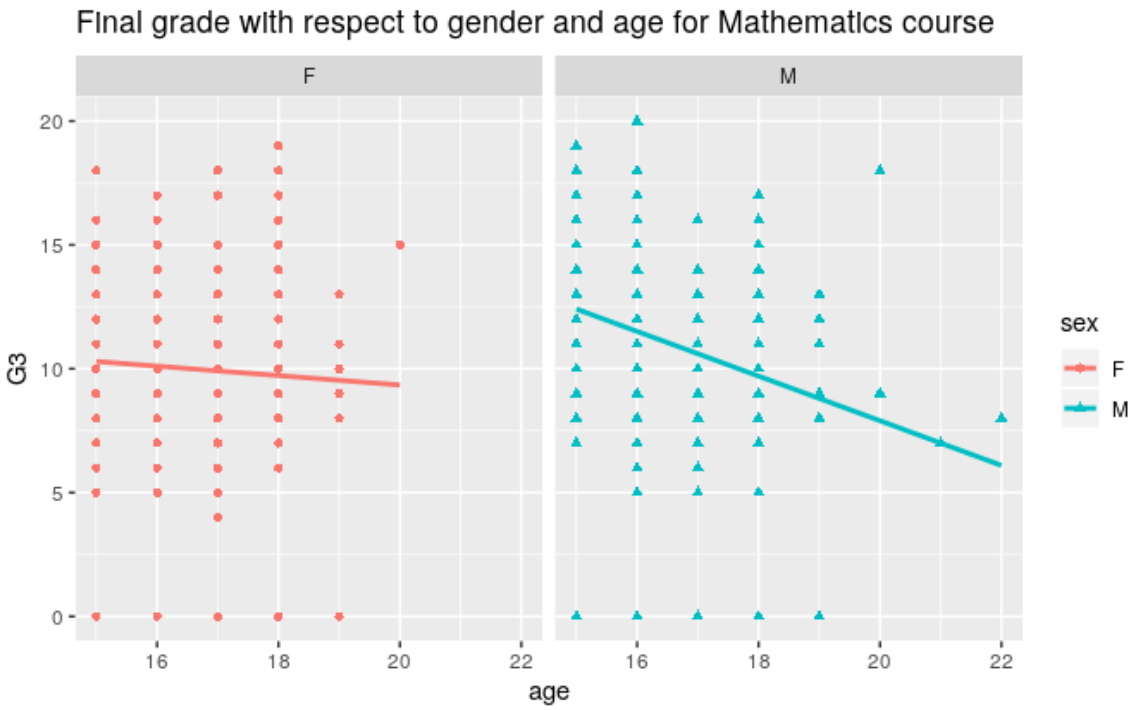


Figure 6

From Figure 5 and Figure 6, We can see that a sharp decrease in the males performance in Mathematics course. So the question that comes to our ming is: Why do male students performance decrease with age? let's explore more.

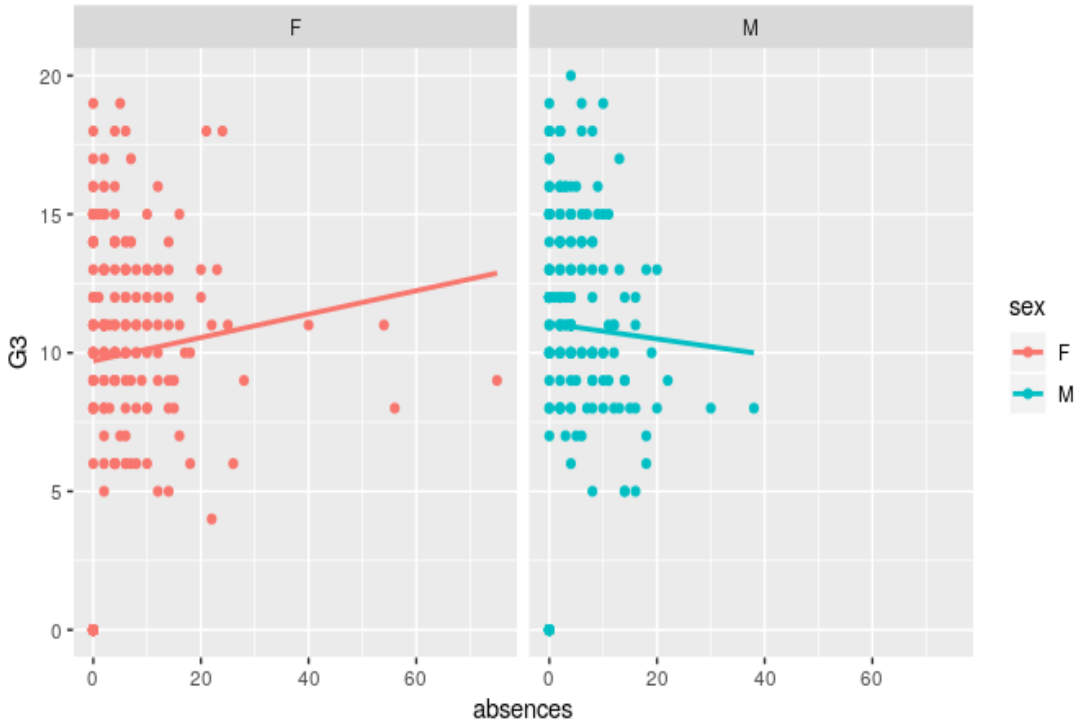


Figure 7

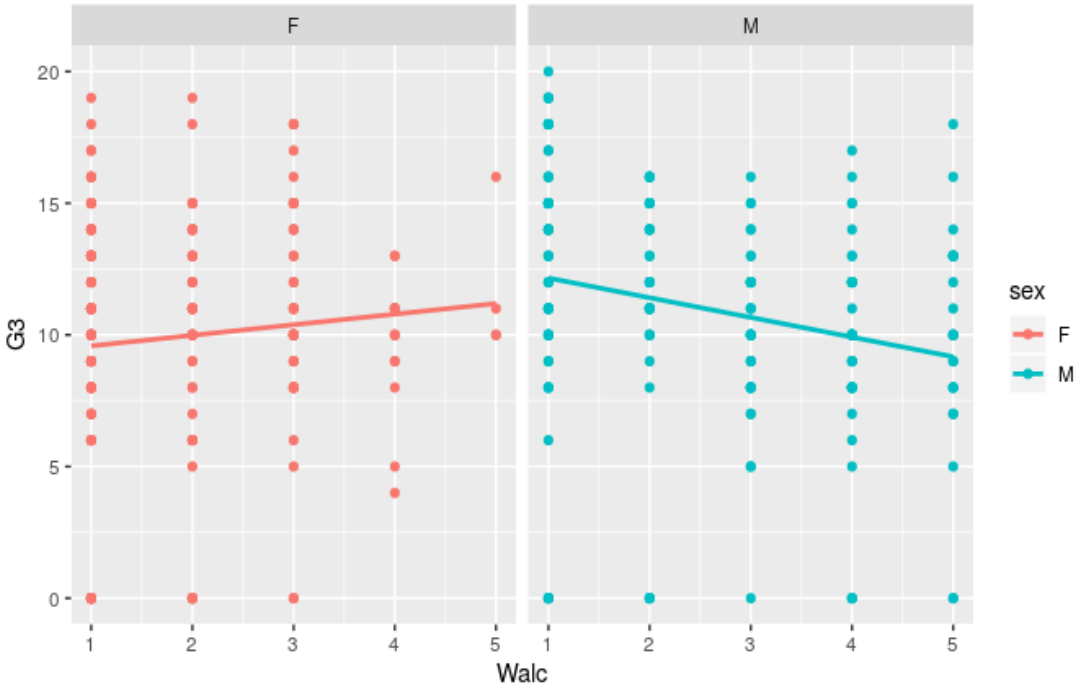


Figure 8

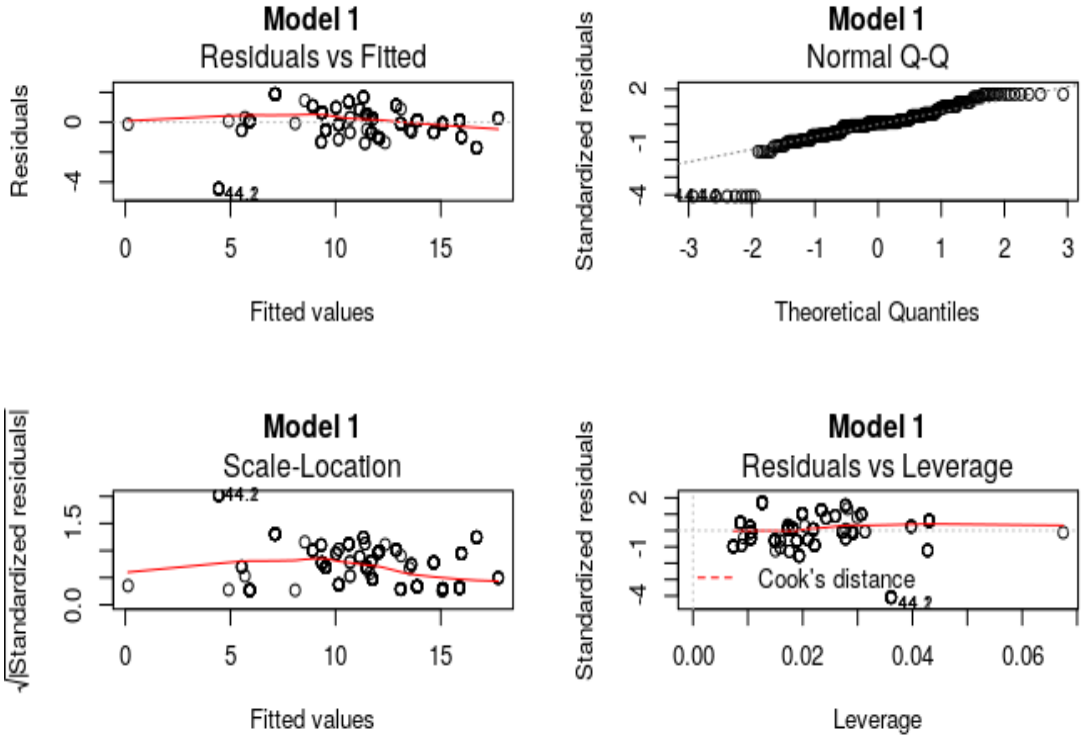
Obviously, we can observe that ***absences*** - (number of school absences) and ***Walc*** - (weekend alcohol consumption) might be the possible factors why male performance dropped in Mathematics course. On the other hand, absences - (number of school absences) and Walc -(weekend alcohol consumption) has no effect on the performance of the females.

Model 1: Multiple Linear Regression Model

```
Call:
lm(formula = G3 ~ G2 + age + Medu + Fedu + traveltime, data = training_math,
    subset = studytime + failures + famrel + freetime + goout +
        Dalc + Walc + health + absences + G1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4435 -0.5303  0.1077  0.5261  1.8758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.18484    1.20964   0.153  0.87865
G2           0.97176    0.02155  45.091  < 2e-16 ***
age         -0.08993    0.06723  -1.338  0.18204
Medu         0.19745    0.08484   2.327  0.02063 *
Fedu         0.26516    0.08171   3.245  0.00131 **
traveltime   0.27800    0.08653   3.213  0.00146 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.107 on 290 degrees of freedom
Multiple R-squared:  0.898,     Adjusted R-squared:  0.8962
F-statistic: 510.6 on 5 and 290 DF,  p-value: < 2.2e-16
```

Above is the multiple linear regression of some of the numerical variables. From the model, we notice that the variance explained by the explanatory variables is 89.8%. We can also see that G2 has a very strong significant relationship in predicting students final score, as well as the Medu, Fedu and traveltime. Find below the residuals of the model.

Residual plots of the regression model.

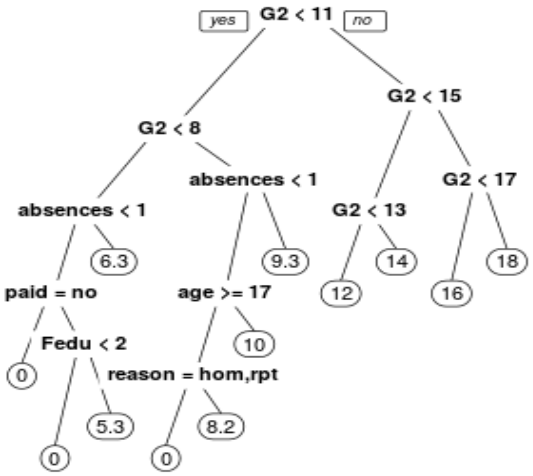


From the residual plots, one can observe that, the residual vs the fitted almost lies at 0, meaning that almost all the errors have zero mean. Also, the QQ plots shows that the data lies on the fitted line (although there are some outliers off the fitted line).

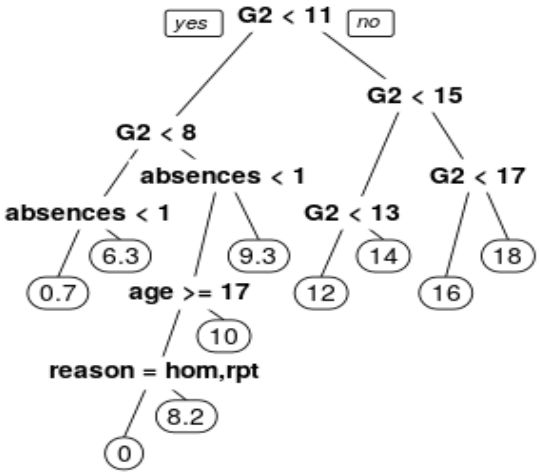
Model 2: Decision Tree Regression

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes.

Model 2 before Pruning



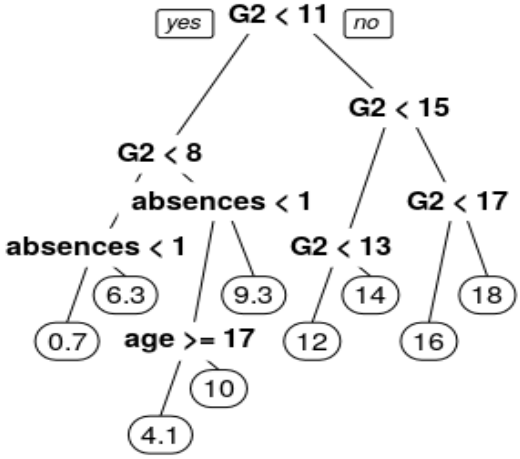
Model 2 - after Pruning



The graph on the left is the model before pruning was carried out, while that on the right is the model after pruning. (the concept on=f pruning was explained in the methodology).

Model 3 : ***Random Forest Regression***

Model 4 Random Forest



Model Evaluation

Table 3: Model Evaluation & Comparison

Model	RMSE	Correlation of actual value to predicted
Multiple Regression	1.755525	0.919
Decision Tree	1.484753	0.938
Random Forest	1.30196	0.953

After exploring the 3 models above, predictions were made using the test set. The correlation between the predictions made and the original test set were estimated. And from the table, is is observed that the Random forest model had the highest correlation compared to others. Also, in other to futher validate our choice of model, the metric use for evaluation is the RMSE, we compere the RMSE and the Correlation of each of the models, and it was observed that, the Random Forest model, had the least RMSE as well as the highest correlation of (predicted vs actual), hence, we have sufficient reason choice the Random forest as the best model. This same analysis can be reproduce this same analysis for the Portuguese language course.

REFERENCES

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

2. http://www3.dsi.uminho.pt/pcortez/student.pdf

3. https://www.kaggle.com/micahshull/r-students-scores-linreg-tree-forest-svm

4. https://www.kaggle.com/hindelya/students-grade-prediction