**STAT 413 (Statistical Learning)**

Third Project Report

**Analyzing Student Performance Dataset**



A picture containing text, sign

Description automatically generated

**For dr. Ali Duman**

**By Group 6:**

**Omer Alabas / 201769570**

**Ahmed Alturfi / 201743590**

**Omar Al Bumadrah / 201636140**

**Abdullah Mohmmed / 201740190**

Table of Contents

[Table of figures 3](#_Toc100406059)

[Introduction 4](#_Toc100406060)

[Results 5](#_Toc100406061)

[Data preprocessing 5](#_Toc100406062)

[Variable’s type 5](#_Toc100406063)

[missing values 7](#_Toc100406064)

[Outliers 8](#_Toc100406065)

[Feature creation 9](#_Toc100406066)

[Data Visualization 9](#_Toc100406067)

[Correlation matrix 9](#_Toc100406068)

[alcohol consumption vs grades 11](#_Toc100406069)

[Internet access at home 11](#_Toc100406070)

[higher education 12](#_Toc100406071)

[Parents education 13](#_Toc100406072)

[Guardian effects on student's grades 13](#_Toc100406073)

[Age effect on Grades 14](#_Toc100406074)

[Grades vs Travel Time 15](#_Toc100406075)

[Grades vs Study Time 15](#_Toc100406076)

[Relationship and Patterns 16](#_Toc100406077)

[grades distribution 16](#_Toc100406078)

[Male vs Female comparison 17](#_Toc100406079)

[Alcoholic vs non-alcoholic students 17](#_Toc100406080)

[Students in relationship vs students with no relationship 18](#_Toc100406081)

[Parents together or apart 18](#_Toc100406082)

[High absences students vs low absences students 18](#_Toc100406083)

[Tentative Schedule 19](#_Toc100406084)

[Conclusion 20](#_Toc100406085)

# Table of figures

[Figure 1- missing values 8](#_Toc100406086)

[Figure 2- Outliers 9](#_Toc100406087)

[Figure 3- Correlation matrix 10](#_Toc100406088)

[Figure 4- alcohol workday consumption 11](https://kfupmedusa-my.sharepoint.com/personal/s201636140_kfupm_edu_sa/Documents/Second%20Report.docx#_Toc100406089)

[Figure 5- alcohol weekend consumption 11](https://kfupmedusa-my.sharepoint.com/personal/s201636140_kfupm_edu_sa/Documents/Second%20Report.docx#_Toc100406090)

[Figure 6- Internet access at home 11](#_Toc100406091)

[Figure 7- higher education 12](#_Toc100406092)

[Figure 8- higher education count 12](#_Toc100406093)

[Figure 9- Parents education 13](#_Toc100406094)

[Figure 10- Guardian of the student 13](#_Toc100406095)

[Figure 11- Age effect on Grades 14](#_Toc100406096)

[Figure 12- Grades vs Travel Time 15](#_Toc100406097)

[Figure 13- Grades vs Study Time 15](#_Toc100406098)

[Figure 14- grades distribution 16](#_Toc100406099)

# Introduction

An upvote is welcomed if this Data Set is beneficial. This data is based on secondary school student achievement in two Portuguese schools. The data was acquired through school reports and surveys and includes student grades, demographics, and social and school-related aspects. However, in this project , the data will be visualized to explore the insight data. After that,it will be analyzed using different classifications and regression methods to answer how can various factors affect student grades.

# Data

This dataset addresses many attributes that may be helpful in analyzing student performance. The dataset consists of 35 columns and 649 rows. However, the table below will illustrate the data and its types:

|  |  |  |
| --- | --- | --- |
| Variables | | |
| Variable name | Description | Variable type |
| school | binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira | Categorical |
| sex | binary: 'F' - female or 'M' - male |
| address | tudent's home address type (binary: 'U' - urban or 'R' – rura)l |
| famsize | family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) |
| Pstatus | parent's cohabitation status (binary: 'T' - living together or 'A' - apart) |
| Medu | mother's education (numeric: 0 - none, 1 – primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education) |
| Fedu |
| Mjob | job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') |
| Fjob |
| reason | reason to choose this school ( close to 'home', school 'reputation', 'course' preference or 'other') |
| guardian | student's guardian ( 'mother', 'father' or 'other') |
| traveltime | home to school travel time ( 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) |
| studytime | weekly study time ( 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| failures | number of past class failures (n if 1<=n<3, else 4) |
| schoolsup | extra educational support (binary: yes or no) |
| famsup | family educational support (binary: yes or no) |
| paid | extra paid classes within the course subject (Math or Portuguese) (binary: yes or no |
| activities | extra-curricular activities (binary: yes or no) |
| nursery | attended nursery school (binary: yes or no) |
| higher | wants to take higher education (binary: yes or no |
| Variables | | |
| Variable name | Description | Variable type |
| internet | Internet access at home (binary: yes or no) Internet access at home (binary: yes or no) | Categorical |
| romantic | with a romantic relationship (binary: yes or no) |
| famrel | quality of family relationships ( from 1 - very bad to 5 - excellent) |
| freetime | free time after school (from 1 - very low to 5 - very high) |
| goout | going out with friends (from 1 - very low to 5 - very high) |
| Dalc | workday alcohol consumption ( from 1 - very low to 5 - very high) |
| Walc | weekend alcohol consumption ( from 1 - very low to 5 - very high) |
| health | current health status (from 1 - very bad to 5 - very good) |
| age | student's age (from 15 to 22) | Numeric |
| absences | number of school absences (from 0 to 93) |
| G1 | first period grade (from 0 to 20) |
| G2 | second period grade (from 0 to 20) |
| G3 | final grade ( from 0 to 20,) |
| averageG | Average of the three periods grades |

# Reviews(abdullah)

1. **ML Test by Eric Biernacki:**

<https://www.kaggle.com/ericbiernacki/ml-test>

After importing and examining the data, he started by dropping all columns that he thinks it’s not strongly related to the final grade. So, he left out with only 7 columns, 6 of them are considered as predictors which are : 'age', 'studytime', 'failures', 'freetime', 'goout', 'absences'. To predict the final grade ‘ G3 ‘.

After that, he split data into training and testing data. Then he started with Decision Tree Regressor and fitted the model with the training data, then, he calculated the mean absolute error ‘ MAE ‘ of the fitted model which was 3.181697. Moreover, to find the best value for leaf nodes, he compared the number of leaf nodes with the corresponding MAE, and he noticed there is no big difference in MAE. To clarify, when he changed the leaf nodes from 2 to 50,000, the difference in MAE was only 1.

However, he tried a different technique which is Random Forest Regressor. He fitted the model with the training data, Then he calculated the mean absolute error ‘ MAE ‘ of the fitted model which was 2.6921205.

1. **Simplified Linear Regression by Sadaf Zabeen:**

<https://www.kaggle.com/sadafzabeen/linear-regression-simplified-beginners>

First, she imported the data, then she started modelling. After observing the variables she decided to choose 'absences', 'failures', 'G1', 'G2', and 'studytime' as her predictors, while ‘G3’ is the response variable. It is important to note that her choice was not based on any statistical insights. Perhaps she had some field knowledge on the subject.

Secondly, she split her data into two datasets. One for training the model and the other for testing, the test size was 33% of the data.

After that, she fitted the model with the training data using ***sklearn*** library on python. She then tested the model accuracy with the testing data, surprisingly, the score (R2) was 86%, which is very good for a first model.

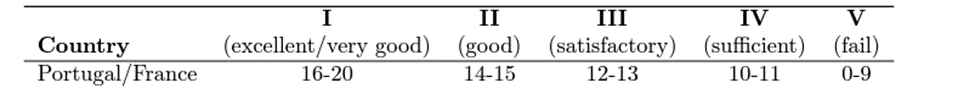
Finally, she displayed all the predictions for the test data.

1. **Annual Grades Average 5-Level Classification by Sharon Yaroshetsky:**

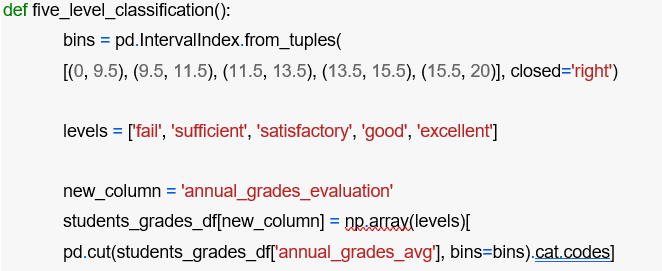
<https://www.kaggle.com/sharonyaroshetsky/annual-grades-average-5-level-classification>

On the Student Performance Dataset Yaroshevsky`s work was about classification methods, he uses a method called 5-Level Classification, his classification was based on the annual average instead of the G3 grades, using the usual ML libraries such as pandas, numby, and *sklearn* in his code he used a Library called seaborn.

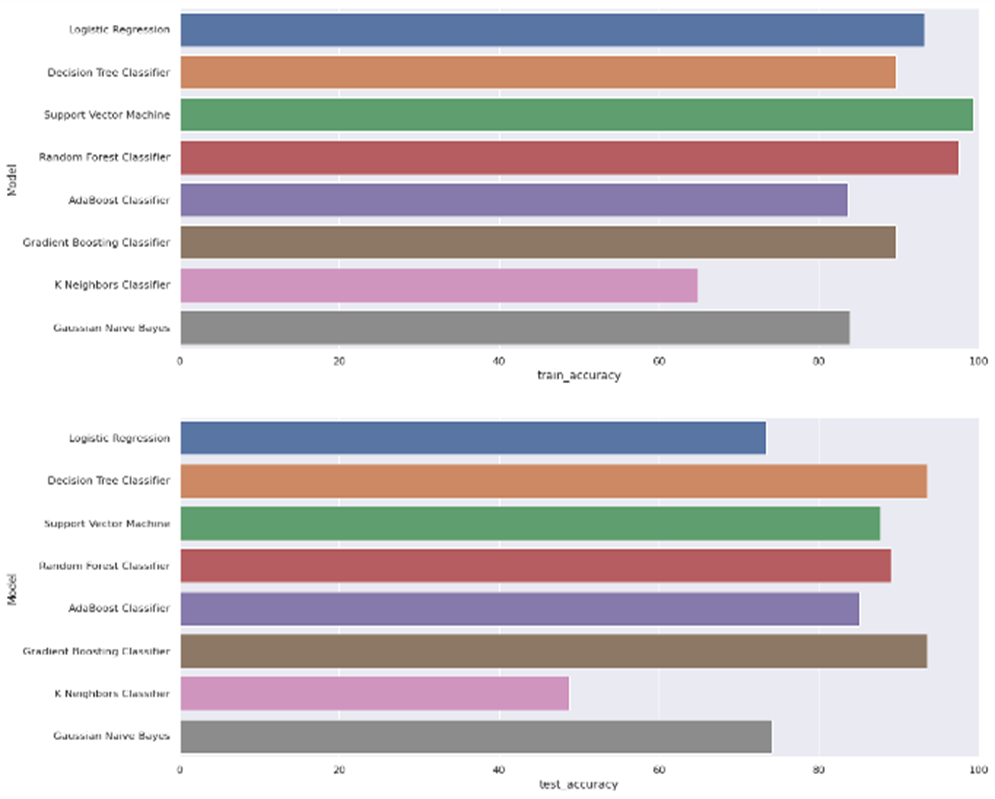
The work is based on classifying the student on the annual average grade into 5 levels is shown in the table below:



Using the code below he added a new column to calculate average grades and classify it based on the average:



in Yaroshevsky`s work, he uses different classification methods, such as Logistic Regression, Decision Tree Classifier and Support Vector Classifier. Also, he uses more advanced methods such as Random Forest Classifier, AdaBoost Classifier, and Gradient Boosting Classifier. The results of these methods are as follows:



1. **Student Performance Analysis by ARCHIT9406:**

<https://www.kaggle.com/archit9406/student-performance-analysis>

in ARCHIT9406’s work he started by adding a column called ‘total grades’ which is basically the average of G1, G2, and G3 then he classify the grades into three classes (high, low, average). Then he visualizes the result as a histogram showing the data of the grade classes.

Also, he uses ***seaborn*** library to plot a correlation heat map to explore the correlation between every two attributes.

The conclusion of his work was to discover every attribute and its relation with the output of interest which was the class of the student’s grades after analyzing the attribute relations he uses SVC classifier from ***sklearn*** library and the results were solid for it:

precision recall f1-score support

average 0.73 1.00 0.85 153

high 0.00 0.00 0.00 38

low 0.00 0.00 0.00 18

accuracy 0.73 209

macro avg 0.24 0.33 0.28 209

weighted avg 0.54 0.73 0.62 209

“ SO BY CONFUSION MATRIX AND F-SCORE, WE FIND OUT THAT RANDOM FOREST IS BEST CLASSIFIER FOR GIVEN PROBLEM.” - ARCHIT9406.

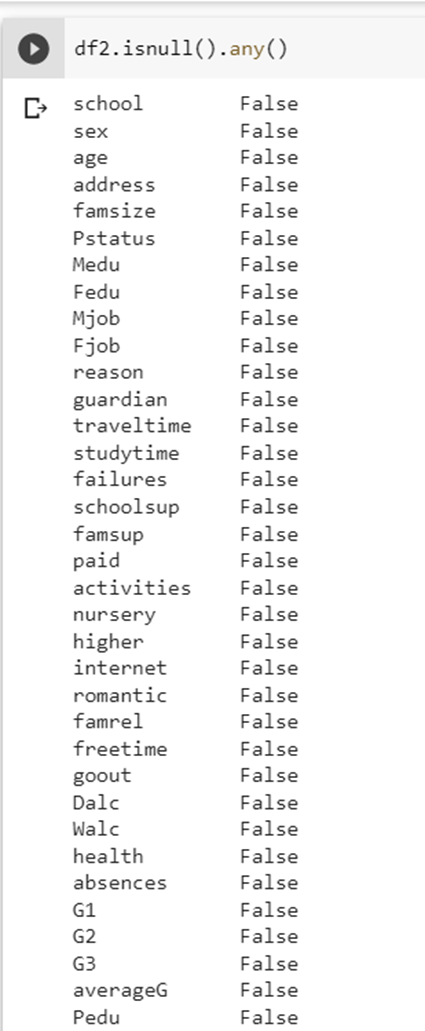
# Problem Definition(ahmed)

In our term project we intend to use a historical data to build statistical models that can answer the following questions “Is there a relationship between the final grade and other student attributes? Can we predict averageG? , can we know which student will continue their higher education? What are the factors? Can we know which student will pass the course without knowing their grades? “ to answer these question we will use several statistical modeling technique as well as python libraires , based on the dataset and using python programing language we will answer these questions

# Data Preprocessing(omar)

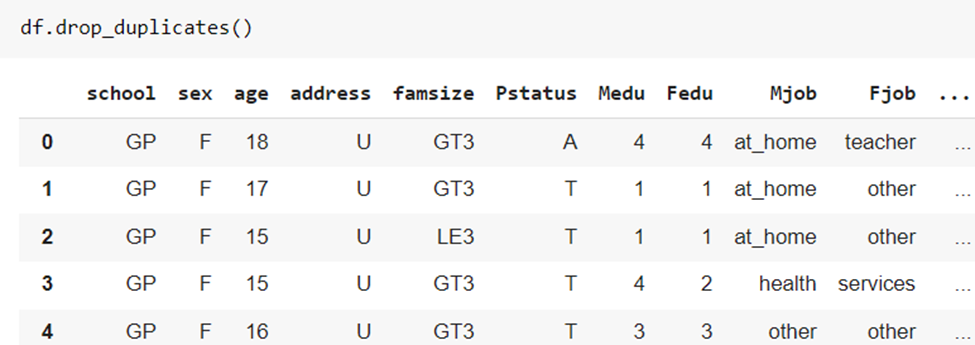
## Nulls value handling

There are no missing values as shown below:



*Figure 1- missing values*

## Dropping all duplicates



*Figure 2 dropping duplicates*

## Clearing the data from outliers



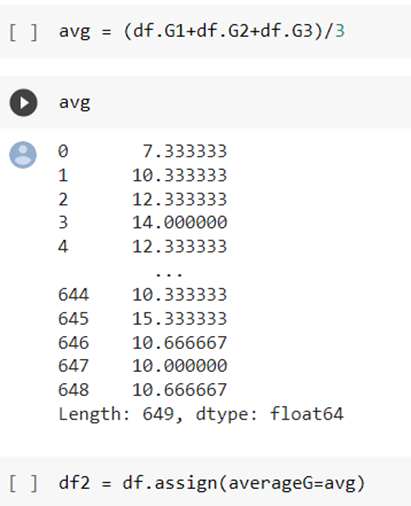
*Figure 3- Outliers*

## Feature creation

We created two more features, both are the average of other variables

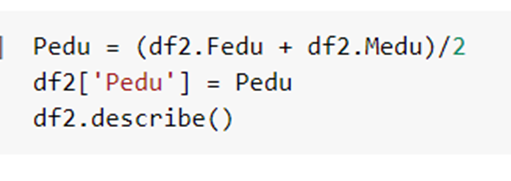
* **averageG:** which is the outcome of (G1+ G2+G3)/3
* **Pedu:** which is the outcome of (Fedu + Medu )/2 .Merging Fedu and Medu is very useful to the model since their effects are almost identical as proved in the Data Visualization section.

### average grade



*Figure 4 averageG creation*

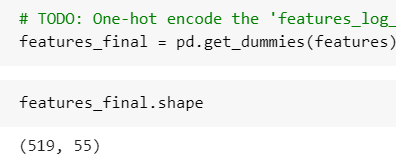
### parent’s education



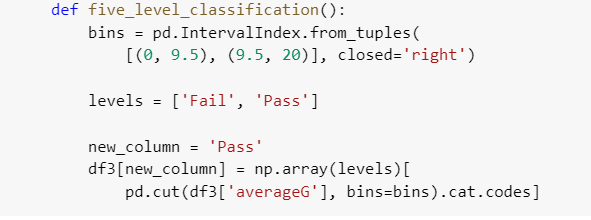
*Figure 5 Pedu creation*

**One-hot encoding**

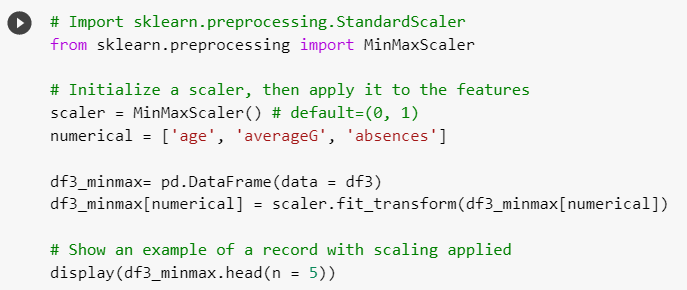
One-hot encoding creates a *"dummy"* variable for each possible category of each non-numeric feature



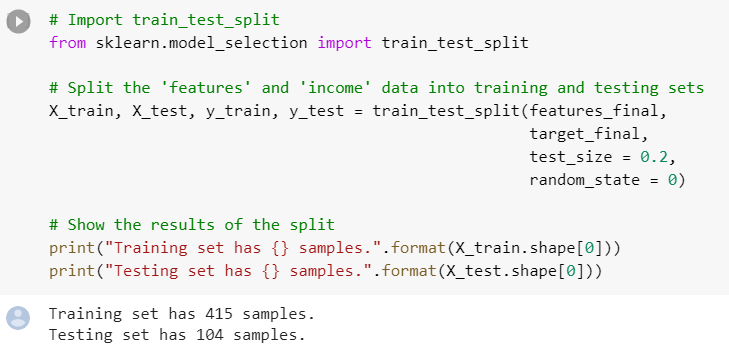
### Create the new variable 'Pass'.



### Normalizing Numerical Features

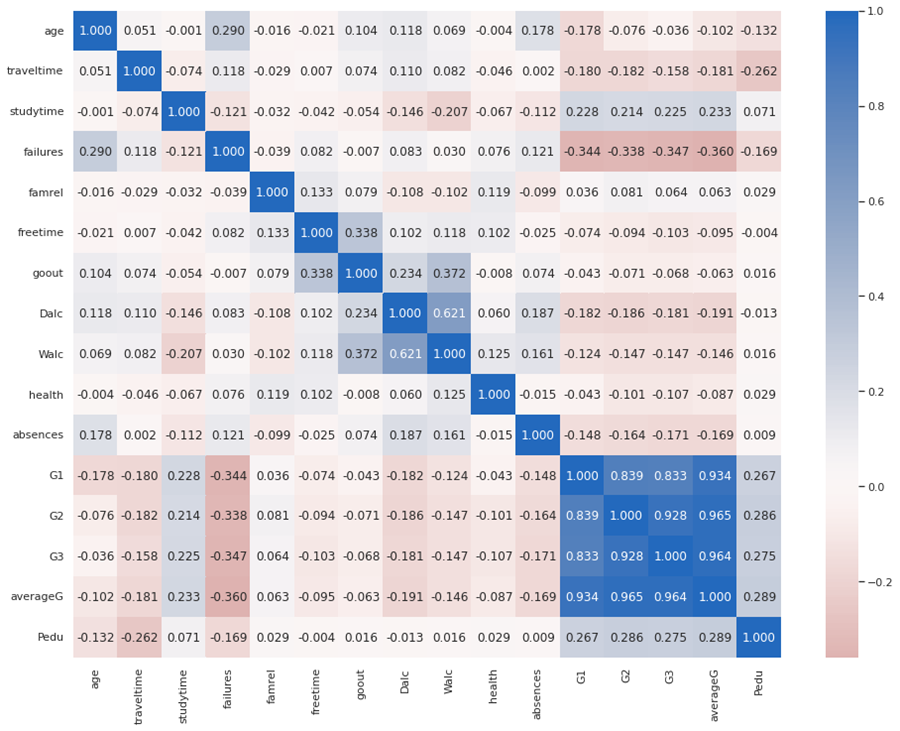


### Shuffle and Split Data



# EDA (omar)

## Correlation Matrix

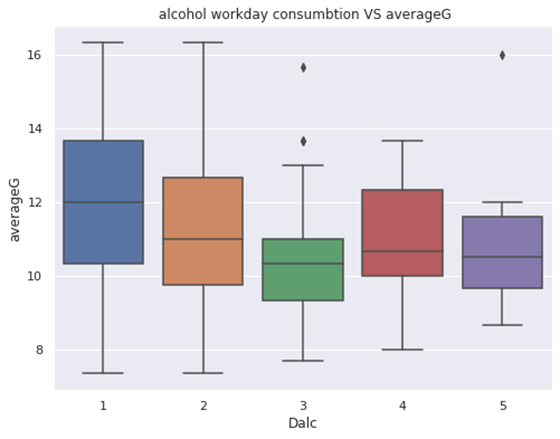


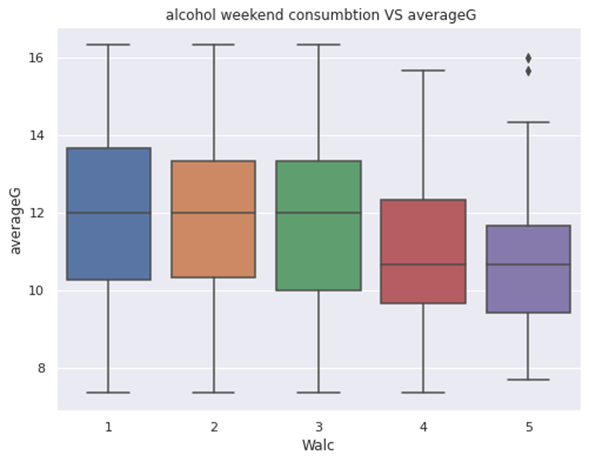
*Figure 3- Correlation matrix*

We can observe that **failures**, **Pedu**, **Dalc** and **studytime** are the variables that have a significant correlation with **averageG.**

### 

### Alcohol Consumption

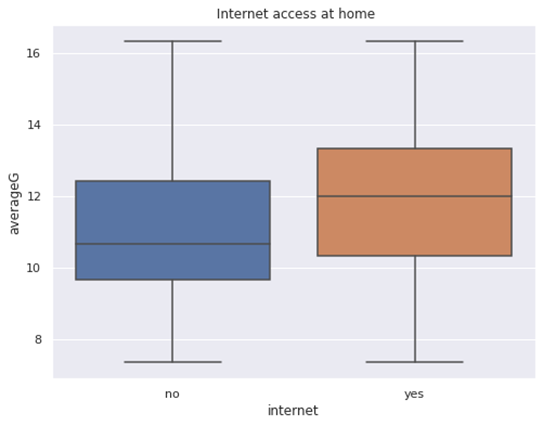




As alcohol consumption on weekends increases, the average grade decreases.

### 

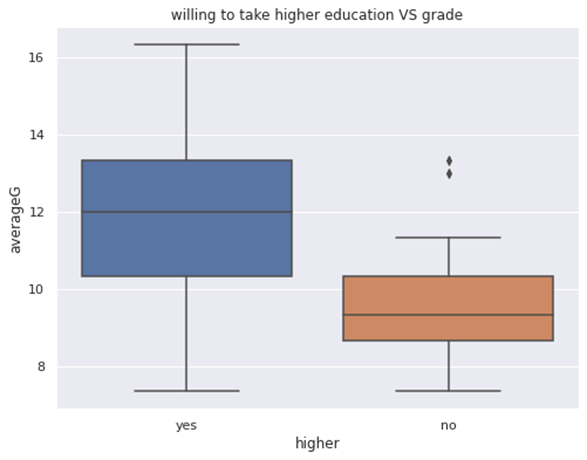
### Internet Access at Home



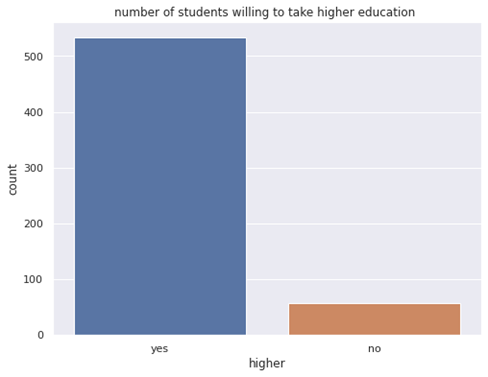
*Figure 6- Internet access at home*

We can see clearly that having internet access at home would increase your chance to get a higher grade.

### Higher Education



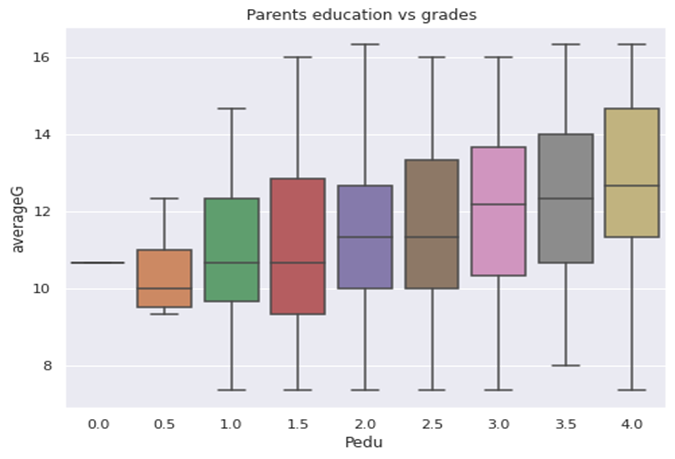
*Figure 7- higher education*



*Figure 8- higher education count*

Regardless of the small number of students who are not willing to take a higher education. It is obvious that students who want to complete their education and enter college have a higher chance to score a higher grade.

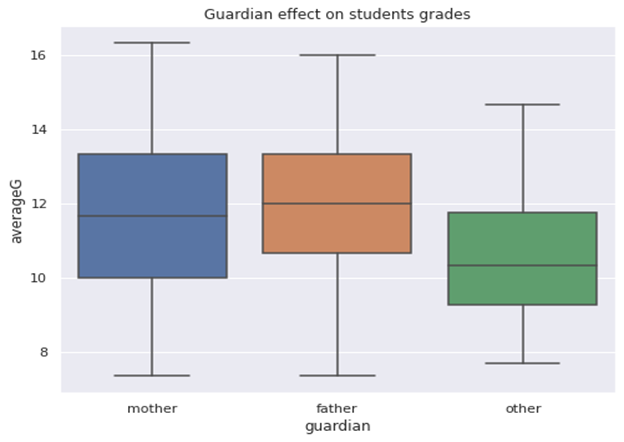
### Parents Education



*Figure 9- Parents education*

As the level of parents education increases, it reflects positively on the students' scores as it is noticeable in the boxplot graph in figure 9.

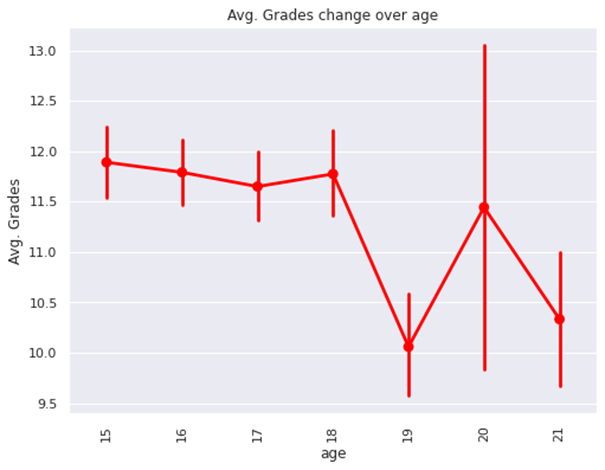
### Guardian Effects on Student's Grades



*Figure 10- Guardian of the student*

We can notice that there is almost no difference between the mother guardian and father guardian while clearly there is a significant difference when the guardian is other than the parents. It reflects negatively on the student performance on the average grade.

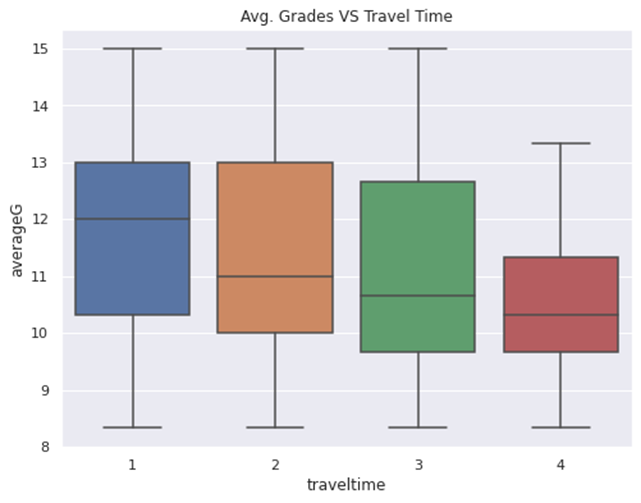
### g. **Age Effect on Grades**



*Figure 11- Age effect on Grades*

We can see that higher age corresponds to lower grades, the reason behind this could be that older students are students who repeated school, therefore will have a worse performance compared to younger students.

### a. Grades vs Travel Time

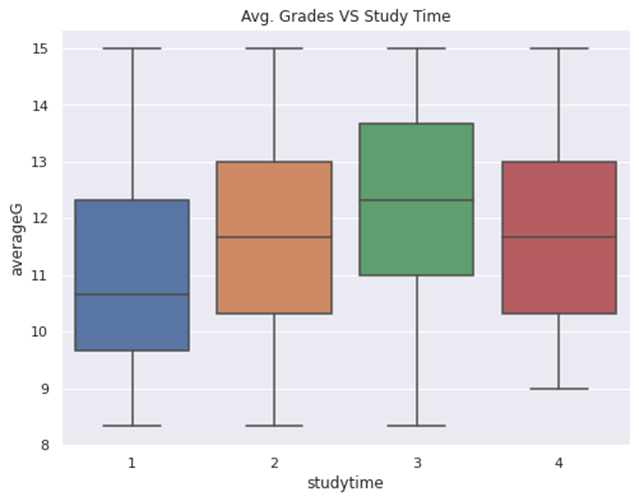


*Figure 12- Grades vs Travel Time*

The time from home to school has a strong effect on students’ performance!

### 

### I. Grades vs Study Time

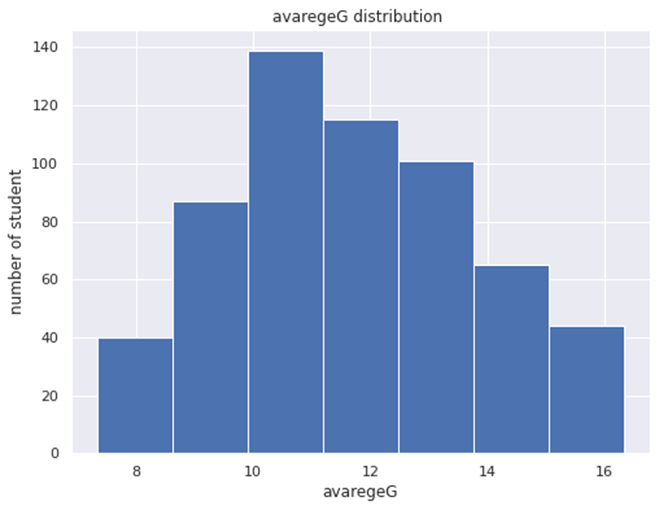


*Figure 13- Grades vs Study Time*

Interestingly, higher study time does not mean higher grades.

## **3. Relationships and Patterns**

## a. Grades Distribution



*Figure 14- grades distribution*

We can see clearly that the distribution is almost little right skewed normal since mode < median < mean.

|  |  |
| --- | --- |
| mode | 10.0 |
| median | 11.666666666666666 |
| mean | 11.677382966723075 |

### b. Male vs Female Comparison

Referring to the detailed analysis section in Collab Notebook, we can see clearly that female student on average 11.905714 has a much higher average grade than male student 11.345781.

Moreover, we can notice that female student study on average more time than male student also they are less likely to have a failure then male student.

|  |  |
| --- | --- |
| **Study time** | **Failures** |
| 2.060000 | 0.165714 |
| 1.734440 | 0.232365 |

### c. Alcoholic vs Non-Alcoholic Students

Referring to the detailed analysis section in Collab Notebook, it is clear that students who consume more alcoholic drinks on the weekend “ Walc” score 10.972222 on average, while students who consume less alcoholic drinks score a much higher grade of 11.857042 on average.

The same situation is on the weekday alcoholic consumption “ Dalc”. But since the count of students who consume alcoholic drinks during the weekdays is very small (27), we cannot conclude a statement.

### d. Students in Relationship vs Students with no Relationship

Referring to the detailed analysis section in Collab Notebook, we can see that there is almost no relationship between students in relationship and average grade 11.635945, as well as students with no relationship and average grade 11.633690. Almost no effect is noted.

### e. Parents Together or Apart

Referring to the detailed analysis section in Collab Notebook, we can see there is a little difference between students' average grade with the living situation ***together*** or ***apart***. Parents together with average grade 11.653153. Parents apart with average grade 11.849315. which means parents living apart have little positive effect on student performance.

However, the mean of the absences is significant between them, we can notice that it is 5.419355 for students whose parents are apart. While it is 3.577681 for students whose parents live together.

### f. High Absences Students vs Low Absences Students

Referring to the detailed analysis section in the Collab Notebook, we can see clearly that students with absences more than five score on average 11.226337 while the student who has absences less than five 11.847708 which is something significant to be looked at.

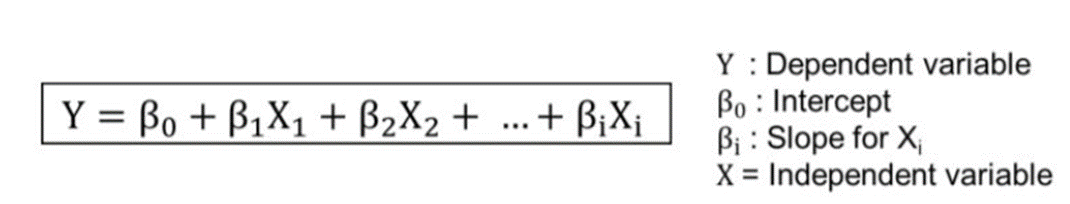
We can see clearly that students with absences more than five are more likely to have failures 0.265432, while the student who has absences less than five is less likely to have a failure 0.165501.

# Methods

* **Multiple Linear Regression Analysis**

One of the strongest techniques for prediction is the MLR which is used to predict dependent variables based on independent variables called predictors. and predictors can be quantitative or categorical. However, this method will be more efficient if there is a strong relation whatever negative or positive relation between the predictor and the output. As shown in figure 1, there is a positive correlation between G1, G2, Study time, mother education and father education with G3, also there is a negative correlation between failures with G3. Therefore, the Multiple Linear regression method is useful in this case.

- **Multiple Linear Regression equation:**



|  |  |
| --- | --- |
| **Advantages** | **disadvantages** |
| **Linear Regression is straightforward to perform and easier to interpret the output coefficients.** | **Linear regression technique outliers can have large effects on the regression, and boundaries are linear in this technique, also it could give negative prediction.** |
| **Linear Regression is susceptible to over fitting, however it can be avoided using some dimensionality reduction techniques, such as cross validation.** | **susceptible to underfitting, this occurs when the hypothesis function cannot fit the data well.** |

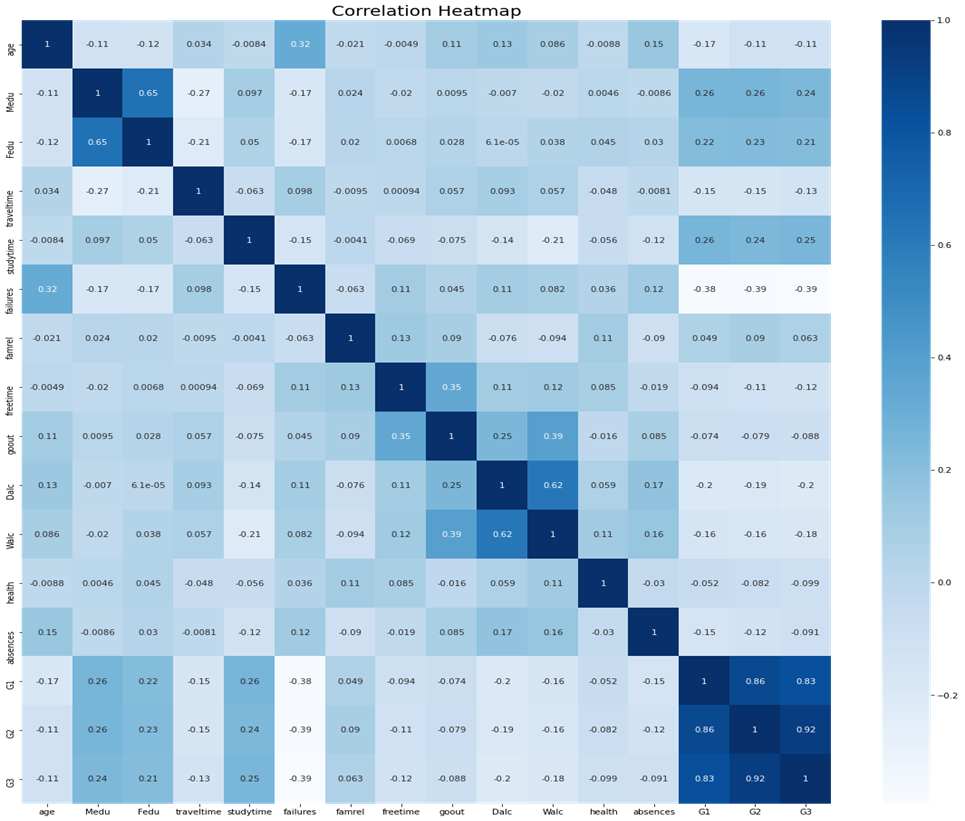


figure 1: Correlation

* **Support Vector Machine**

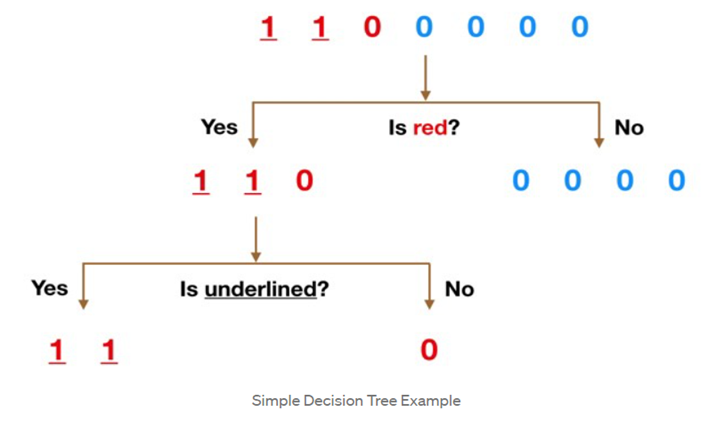
One of the most popular supervised machine learning methods that can be useful for both classification and regression. Also, it’s called widest street and it works by separating the data to the class using a hyper-plane. It also aims to maximize the margin between classes.

Since the final grade is categorical, SVM is a very efficient tool for handling boundary decisions. Therefore, SVM will be used to predict student final grades based on predictors which are : 'age', 'studytime', 'failures', 'freetime', 'goout', 'absences'.

|  |  |
| --- | --- |
| **Advantages** | **disadvantages** |
| **Effective in high dimensional spaces.** | **Not useful if the number of features is much greater than the number of samples,** |
| **Still effective in cases where number of dimensions is greater than the number of samples.** | **SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.** |
| **Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.** |  |

* **Classification Using Decision Tree**

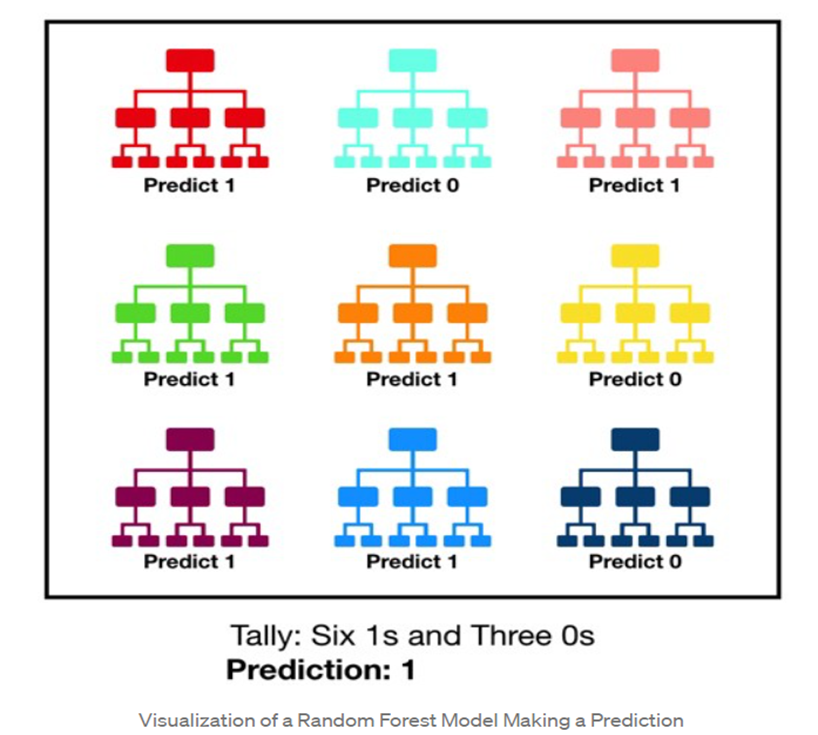
A Decision Tree is a great classification method because it does not require any statistical assumption to be used. The reason for that is we have no probabilistic model, but just a binary split. Moreover, our data contains about 20 categorical variables which we can predict or use as input variables. What we can do is given a set of information about a student, we could predict unknown extra categorical information. An example of a decision tree:



|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| **Simple to understand and to interpret. Trees can be visualized.** | **Decision-tree learners can create over-complex trees that do not generalize the data well.** |
| **Requires little data preparation.** | **Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.** |
| **Able to handle both numerical and categorical data** | **Decision tree learners create biased trees if some classes dominate.** |

* **Classification Using Random Forest**

The Random Forest classifier is fundamentally based on the Decision Tree. It consists of a large number of Decision Trees that behave as an ensemble statistical method. In the Random Forest, each tree will produce a class prediction. After that, the class with the highest votes becomes our prediction. See the below figure:



According to **towards datascience** website, Random Forest classifiers will outperform any individual consistent model. The reason for that is the large number of uncorrelated trees working as one unit and how they protect each other from their individual errors.

We will use both classification methods to answer the classification questions mentioned earlier.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| **It gives variable importance which helps in determining the variable which impacts positively.** | **The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions.** |
| **It takes care of null values.** |
| **Often machine learning models are overfitted, random forest classifiers wouldn't get overfitted.** |
| **It can be used as a regression as well as classification model.** | **Random forest is a predictive modeling tool and not a descriptive tool.** |
| **The method also handles variables fast, making it suitable for complicated tasks.** |

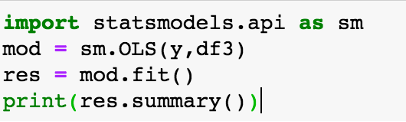
# Omer

# Results

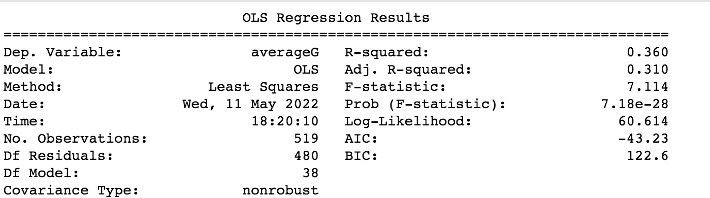
# Ahmad

## Linear Regression Modeling

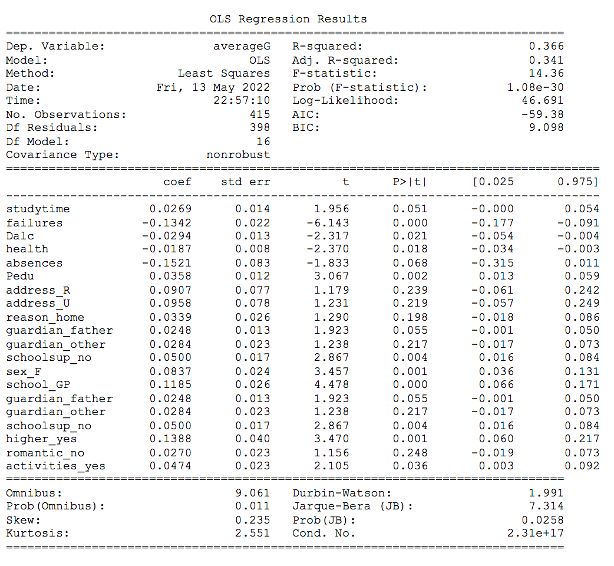
After one hot encoding the categorical data in the data set the data set became with 60 different attribute . we used “ols” function in statsmodels library in python to fit all of the 60 attribute and the response the average grade into linear regression model



The results was not good with R-squared equal to 0.36



The next step was to improve the model by variable selection method , we chose to go with backward selection Backward stepwise selection (or backward elimination) is a variable selection method which begins with a model that contains all variables under consideration (called the Full Model) then starts removing the least significant variables one after the other until a pre-specified stopping rule is reached or until no variable is left in the mode we stopped when the largest p-value in the model is less than 0.3 .  
 the final model we have after backward selection is :



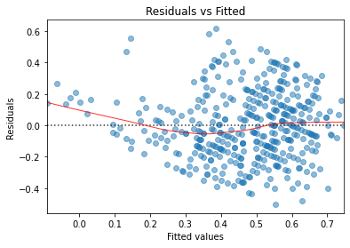
the R-squared increase form 0.36 to 0.366 for the model with 20 variable compared to the first one we used with 60 variable .

then to evaluate our linear model we run some diagnostic test to our model .

the first diagnostic test was to calculate RMSE test value was 0.21432683828646565

RMSE is a good measure of accuracy, but only to compare prediction errors of different models or model configurations for a particular variable and not between variables, as it is scale-dependent. The RMSE indicate that the linear model is producing huge error

Also we run another diagnostic test which is Residuals vs Fitted plot



The plot indicate that the absolute value of the residuals are strongly positively correlated with the fitted values, whereas no such trend is evident in the third plot. So if it were the case that, theoretically speaking, in a heteroscedastic linear model with normally distributed error , it violate one of the assumptions of the linear regression

## Methodology of model selection

We will follow the same methodology in the next two classification tasks, which goes as follows:

1. **Supervised Learning Models:** We choose which classification algorithms to use in our modeling. Let us call this pool of algorithms ***M***.
2. **Training the Models:** We train our models on the training set, then, we choose the best model out of these classifiers in ***M***.
3. **Initial Model Evaluation:** We choose the best model according to our validation defined metrics. Now we have the best model ***m***.
4. **Model Tuning:** We perform GridSearchCV on ***m*** to tune the hyperparameters and get the optimized classifier ***m\****.
5. **Feature Selection:** We perform variable selection on the best optimized classifier ***m\**** to get a higher score.

## Classification Modeling (I)

Trying to answer the following question:

* ***Q:* Which students wish to continue higher education?**

### Classification Metrics

We observed that the data in the **higher** column in imbalanced, as shown below:

Chart, bar chart

Description automatically generated

Therefore, regular classification metrics will not be sufficient to assess the model. For example, if we built a naive model that predicts ‘yes’ for all inputs, that model will get a high accuracy since most of the data is labeled as ‘yes’. However, such model will get poor precision score if we are trying to catch the ‘no’ instances. See the below formulas:

Chart, diagram

Description automatically generated

Instead, we will use the Fbeta-score as our main metric as it combines both recall and precision.

The F-beta score is the weighted harmonic mean of precision and recall, reaching its optimal value at 1 and its worst value at 0.

The ***beta*** parameter determines the weight of recall in the combined score. beta < 1 lends more weight to precision, while beta > 1 favors recall (beta -> 0 considers only precision, beta -> +∞ only recall).

Text

Description automatically generated

### Supervised Learning Models

We chose the following three classifiers, refer to the **Methods** section mentioned above to know the strengths and weaknesses of these algorithms:

1. **Decision Trees**
2. **Random Forest**
3. **Support Vector Machines**

***M* = {DT, RF, SVM}**

### Training Pipeline

Text

Description automatically generatedTo properly evaluate the performance of each model we've chosen, it's important that we create a training and predicting pipeline that allows us to quickly and effectively train models using various sizes of training data and perform predictions on the testing data.

(Code from Udacity.com)

### Training the Models

To properly evaluate the performance of each model you've chosen, it's important that you create a training and predicting pipeline that allows you to quickly and effectively train models using various sizes of training data and perform predictions on the testing data.

**Bar chart

Description automatically generated**After our training we got the following results:

Now according to the above results, the highest F-score on the testing set goes to both **SVM** and **RF**. However, we will proceed our work with the **RF** model since it is slightly higher.

Now ***m*** = **RF.**

### Model Tuning

Now we will use **GridSearchCV** on ***m***. **GridSearchCV** is a powerful **sklearn** package that performs exhaustive search over specified parameter values for an estimator. In this case we will fine tune the following parameters:

* **n\_estimators:** The number of trees in the forest.
* **min\_samples\_split:** The minimum number of samples required to split.
* **min\_samples\_leaf:** The minimum number of samples required to be at a leaf node.

Table

Description automatically generatedThe results are after model tuning:

As we can see, we got a slight increase in both accuracy and F-score. Model became ***m***\*.

### Feature Selection

Now we need to see which features are most important in our model. To check this, we will a great attribute **RF** called **feature\_importances\_.** Which represents the impurity-based feature importances.

If we plot the weights of the five most important features according to our model, we will get the following:

Chart, bar chart

Description automatically generated

**Interpreting the results:** As shown above, it turns out that the grades of the student, (either **averageG or G1**) has the most effect on whether the student will enter college or not, which makes sense. Then the number of absences is also an important factor, since discipline is crucial when studying in higher education. Followed by parents’ education, which also makes sense. In many cases the educated parents push their sons to continue their education. Lastly, we have age, a possible explanation is that older students who struggled during high school will probably not enter college.

### Final Model

We will now train the model with the five most important features; we will use both the train and test sets for this:

Text

Description automatically generated

As we can see, we improved the model further when using the relevant features.

## Classification Modeling (II)

Trying to answer the following question:

* ***Q:* Can we predict the whether the student will pass or fail without knowing the students grade?**

For this task, we created a new variable called **Pass,** which indicates whether the student passed or not.

A picture containing diagram

Description automatically generatedWe considered [0,9] /20 as a failing grade, we took this fact from the following source:

### Classification Metrics

We will use the same classification metrics as in the first classification task, F-score and accuracy.

### Supervised Learning Models

We chose the same three classifiers:

1. **Decision Trees**
2. **Random Forest**
3. **Support Vector Machines**

***M* = {DT, RF, SVM}**

### Training the Models

Will use the same training pipeline as before.

Bar chart

Description automatically generatedAfter training the three models with the training data, we got the following results:

We will continue with the highest score, and try to tune the results.

Therefore, we will continue our work with the **SVM** model.

Now ***m*** = **SVM.**

### Model Tuning

Now we will use **GridSearchCV** on ***m***:

* **C:** Regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive.
* **gamma:** Kernel coefficient for ‘rbf’, ‘poly’ and ‘sigmoid’.
* **kernel:** Specifies the kernel type to be used in the algorithm.

The results are after model tuning:

Table

Description automatically generated

As we can see, we did not get an improvement this time, the default parameters gave us the optimized model ***m***\*.

### Feature Selection

Plotting the weights of the five most important features according to our model, we will get the following:

Chart, bar chart

Description automatically generated

**Interpreting the results:** As shown above, it turns out that the number of failures of the student has the most effect on whether the student will pass the course or not, which makes sense. Then, surprisingly, the school of the student is also an important factor, which may show that quality of the two schools (**MS** and **GP**) is quite different. Followed by higher education variable, which may prove that students who are planning to enter college are more likely to pass the course. Finally, we have health, a possible explanation is that students with permanent disease are more likely to fail the course, or withdraw from the semester.

*\*Note: refer to the Collab Notebook to see the full code:*

<https://drive.google.com/file/d/1ErR7z4SOM30WD1yyiY3UmfWBf4IknhhV/view?usp=sharing>

# Conclusion and Findings

After modeling we conclude the following points:

* For the MLR, we had a poor model (R2 = 37%) due to the difficulty of the task of predicting the student’s grade without the grades themselves (**G1, G2, G3**).
* The absolute values of the residuals were strongly positively correlated with the fitted values, which is a violation of the linear regression assumption, which could also explain the poor model.
* In both classification tasks, our data was imbalanced, which is why we used F-score as main metric.
* In the first classification task, we were able to obtain a powerful model with: accuracy = 91% and F-score = 93%.
* We found that student’s grades are the most significant features to determine whether the student will continue their higher education or not. This insight suggests that the school should help low performance students to enter college and not lose hope.
* An alternative for predicting the grade itself was predicting whether the student will pass or not, which turns out to be an easier task.
* In the second classification task, we were also able to obtain a reliable model with: accuracy = 86% and F-score = 97%.
* Students who failed previously are more likely to fail again, this could suggest that the school should establish programs that help students who failed, programs that provide them with special guidance and support.
* The school should pay attention to students with special health conditions, as it is possible that such students could fail the course.

# References

<https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree>

<https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>

<https://towardsdatascience.com/how-to-best-evaluate-a-classification-model-2edb12bcc587>

<https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b>

<https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>

An Introduction to Statistical Learning: with Applications in R by Gareth James, Daniela Witten, et al.

<https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>

https://udacity.com/