**STAT 413 (Statistical Learning)**

First Project Report

**Analyzing Student Performance Dataset**

horizontal line

# 

**For dr. Ali Duman**

**By Group 6:**

**Omer Alabs / 201769570**

**Ahmed Alturfi / 201743590**

**Omar Al Bumadrah / 201636140**

**Abdullah Mohmmed / 201740190**

# Introduction

The data science study is based on a lot of branches starting from analyzing the data to building models and predicting based on the model, using our background in data science gained by the CX program we intend to perform analysis and study on the chosen data. As the starting point we developed a plan for the project to follow throughout the term, also we explore the chosen dataset and describe the dataset itself and each attribute in the dataset and discuss other related work on the dataset.

# Data description

https://www.kaggle.com/larsen0966/student-performance-data-set

This data is collected from two different Portuguese schools regarding students performance in the mathematics course and Portuguese language course. This dataset addresses many attributes that may be helpful in analyzing student performance. The dataset consists of 35 columns and 649 rows.

**Columns (attributes) definition:**

1. **school** - student's school Gabriel Pereira or Mousinho da Silveira.
2. **address** - student's home address type (binary: 'U' - urban or 'R' - rural).
3. **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3).
4. **Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart).
5. **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education).
6. **Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other').
7. **guardian** - student's guardian (nominal: 'mother', 'father' or 'other')
8. **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).
9. **failures** - number of past class failures (numeric: n if 1<=n<3, else 4).
10. **schoolsup** - extra educational support (binary: yes or no).

# Potential Questions and Problems

## Classification:

* 1. How to select the variables?
     1. Apply Chi-Squared test (contingency tables) for categorical input variables.
     2. Apply ANOVA correlation coefficient (linear) for numerical input variables.
  2. How to evaluate the model?
     1. Calculate the accuracy of the model, ie.
     2. Calculate the value of F1: , if the value of F1 is close to 1, the model is good.
  3. ***By tackling the above statistical questions, we try to answer the following classification questions:***
     1. Which students wish to continue higher education?
     2. Which students pay for extra education?
     3. Which students receive financial family support?
     4. Which students are engaged in a romantic relationship?
     5. What is the type of student's home address?

## Regression:

* 1. Is there a relationship between the final grade and other student attributes?
     1. Fit a multiple regression model of all variables (attributes).
     2. Test the hypothesis H0: ***βX1 = βX2 = βXi= 0. , i=1,2..n***
     3. Use F-statistic and corresponding p-value. If the p-value is too small, then the predictor is significant.
  2. Which variables are associated with the final grade?
     1. Examine the p-values associated with each predictor’s t-statistic
  3. How strong is the association between each variable in MLR and final grade?
     1. Calculate the 95% confidence intervals for the coefficients in an MLR model using all variables as predictors.
     2. If the confidence intervals are narrow and far from zero, then there is evidence that these attributes are related to the final grade.
  4. Is there a linear relationship between input variables and the response variable?
     1. Observe residual plots and check if no pattern displays then relationships are linear.
  5. Are the selected input variables collinear?
     1. Check the correlation Matrix.
     2. Compute the variance inflation factor (VIF).
     3. If the VIF value exceeds 5 or 10, then this indicates a problematic amount of collinearity.
     4. Drop or combine the problematic variables.
  6. How to evaluate the model?
     1. Check residual standard error ( RSE ), Calculating the error percentage by .
     2. Check R2, as the value is closer to 1, the model is better.
     3. Check the p-value of the whole model.
  7. ***By tackling the above statistical questions, we try to answer the following questions:***
     1. Does internet access affect the final grade of the student?
     2. Does the romantic relationship affect the final grade of the student?
     3. Does alcohol consumption affect the final grade of the student?
     4. Does the quality of family relationships affect the final grade of the student?
     5. Does the grade of the first and second periods determine the final grade?

# Reviews

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.

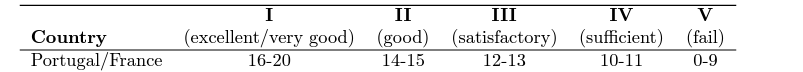
Finally, she displayed all the predictions for the test data.

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:

def five\_level\_classification():

bins = pd.IntervalIndex.from\_tuples(

[(0, 9.5), (9.5, 11.5), (11.5, 13.5), (13.5, 15.5), (15.5, 20)], closed='right')

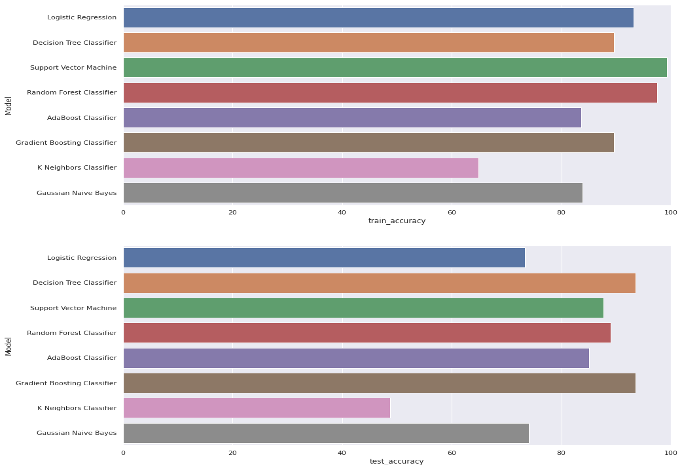
levels = ['fail', 'sufficient', 'satisfactory', 'good', 'excellent']

new\_column = 'annual\_grades\_evaluation'

students\_grades\_df[new\_column] = np.array(levels)[

pd.cut(students\_grades\_df['annual\_grades\_avg'], bins=bins).cat.codes]

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

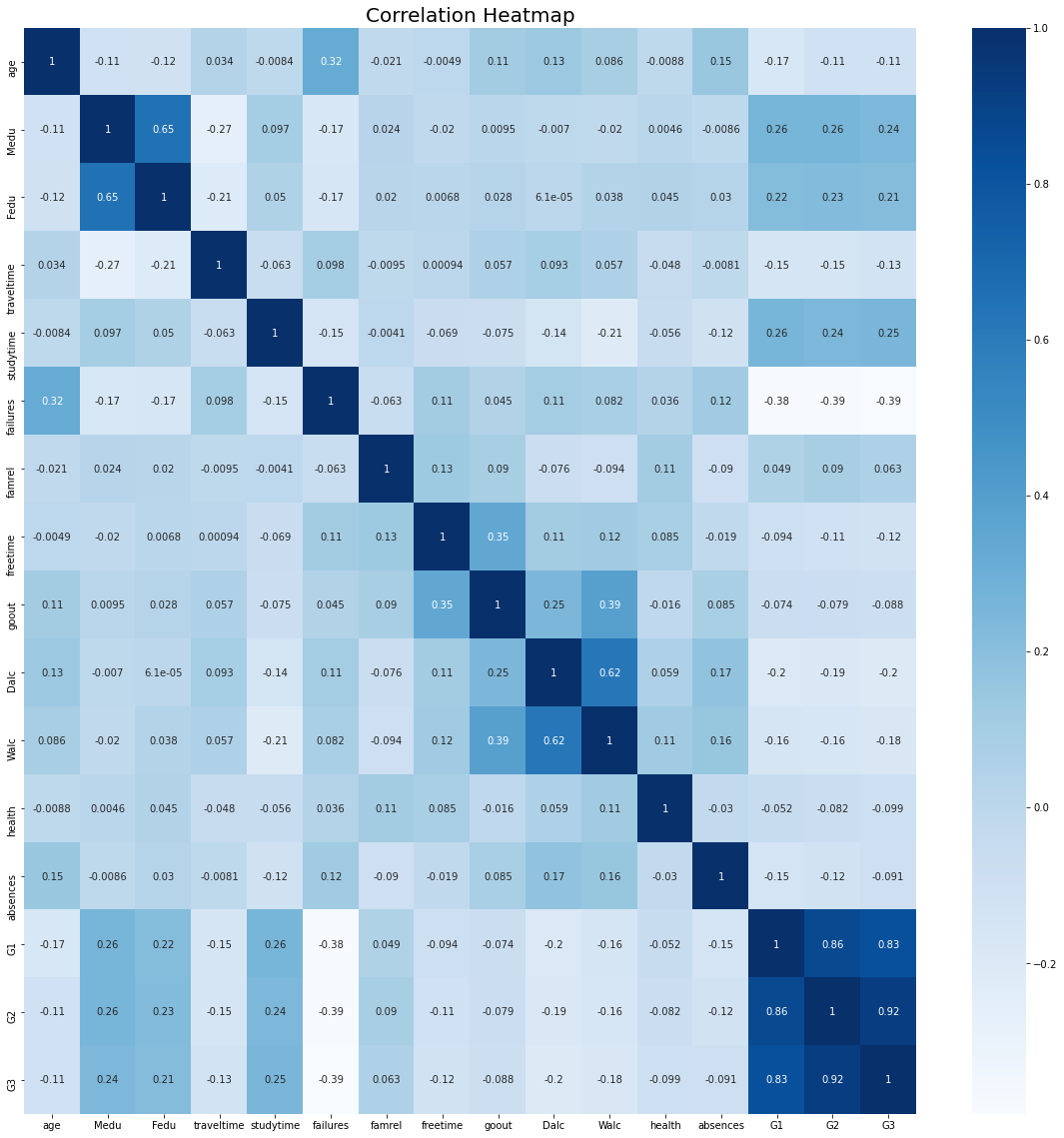
weighted avg 0.54 0.73 0.62 209

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

# Methods

1. **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.



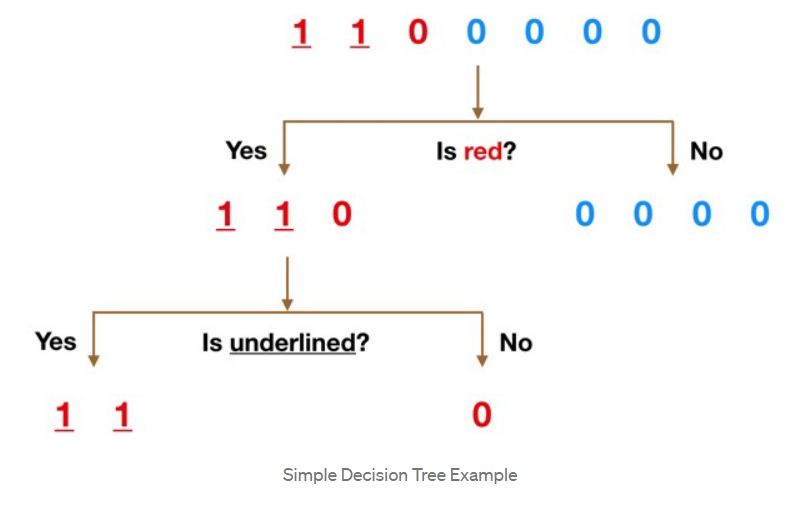
1. **Support Vector Machine**

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'.

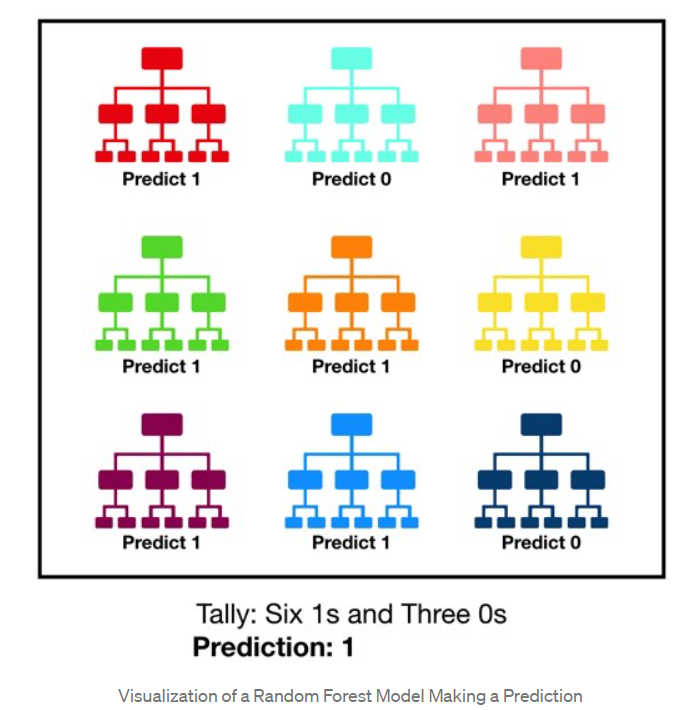
1. **Classification Using Decision Tree**

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:



1. **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 below figure:



According to **towardsdatascience** 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.

# 

# Tentative Schedule

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Week(s)** | **Assigned to** | **Phase** |
| **Problem Definition** | **1 - 4** | **All members** | **1** |
| **Data Gathering** | **2 - 6** | **All members** |
| **Submit First Report** | **6** | **Omer Alabas** |
| **Data Preprocessing and Cleaning** | **7** | **Omer Alabas**  **Omar Albumadrah** | **2** |
| **Exploratory Data Analysis (EDA)** | **7** | **Mohammed Abdullah**  **Ahmad Alturfi** |
| **Statistical Modelling** | **8 - 14** | **All members** |
| **Submit Progress Report** | **10** | **Omar Albumadrah** |
| **Model Evaluation and Improvement** | **10 - 15** | **All members** | **3** |
| **Submit Final Report** | **16** | **Mohammed Abdullah** |

*\*Note: some weeks are overlapping due to the flexible nature of some tasks.*

# 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.