**Problem Statement:**

To study the performance of students pursuing mathematics and Portuguese language course on various socio-economic factors such as gender, age, location or address, alcohol consumption, failures, study time hours, internet connectivity, etc.

**Methodology:**

We started with importing the required libraries and datasets. We did exploratory data analysis by analyzing their datatypes, how many features along with data are provided, finding whether there are any null values. We tried to interpret the variables using visualization libraries – matplotlib and seaborn. Checked distribution of all types of variables – continuous, binary, nominal and ordinal. Finally, we did data pre-processing by creating dummy variables for all the categorical features – ordinal and nominal; standardizing them, and finally built the regression models using train-test-split. The models used are linear regression and random forest regression to find the better model by checking the required metrices – R^2 (coefficient of determination), MAE (mean absolute error) and MSE (mean squared error).

**Action points:**

1. Given two excel sheets for Math course and Portuguese language course, we have merged the data records after importing it into python. New Dataset has 1044 rows and 33 columns. Each column represents factors.
2. Out of these 33 variables, we have segregated into four types:

Continuous: Age, Absences, G1, G2, G3

Ordinal: Mother’s education, Father education, Travel time, Study time, Failures, Family relationship health, Free time, Go out with friends, Daily alcoholism, Weekly alcoholism, Health status of the student.

Nominal: Mother’s job, Father’s job, Reason for choosing the school, Guardian.

Binary: School, Sex, Address, Family size, Parent’s status, Educational support from school, Family support, Paid, Activities, Nursery, Higher, Internet, Romantic relationship status.

1. Checked the distribution and correlation of each factors and used heatmap to show the matrix.
2. We discovered there are no null values which data is clean and structured.
3. We checked the influence of each factors on the student’s grade keeping G3 as our target variable.
4. Created dummy variables for nominal variables.
5. Standardized variables.
6. Built two models using Linear Regression and Random Forest Algorithms.

Training size: 60%. Test size: 40%

1: Linear Regression:

**Accuracy (R2):** 79.29%

**MAE:** 1.053

**MSE:** 3.133

2: Random Forest:

**Accuracy (R2):** 82.83%

**MAE:** 0.959

**MSE:** 2.597

**Both models are delivering decent accuracy rates, but random forest has a better accuracy rate.**