# ACADEMIC PERFORMANCE PREDICTION

MADE BY: SHREYA SINGH RAGHUVANSHI

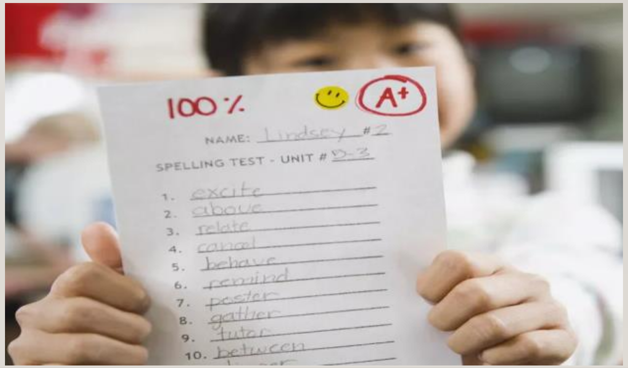
# Introduction

Academic performance prediction is the process of using data to predict how well a student is likely to perform academically.

I have made this project  with the help of Machine Learning. The project is based on a Linear Regression Model.

Machine learning is a branch of Artificial Intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Analyzing academic qualities and guiding to elevate performance



# PROBLEM STATEMENT

The purpose of this project is to develop a linear regression model to predict academic performance . The model will be used to identify  factors affecting the model to predict academic performance with a reasonable degree of accuracy. Some of its advantages include:

- **1. Early Intervention:** Predicting academic performance can help identify students who may be at risk of falling behind or experiencing difficulties in their studies
- **2. Resource Allocation:** Predictive models can assist educational institutions in allocating their resources more effectively.
- **3. Personalized Learning:** Enabling personalized learning experiences by understanding their strengths, weaknesses, and learning preferences, educators can tailor instruction to meet individual needs.

# Methodology

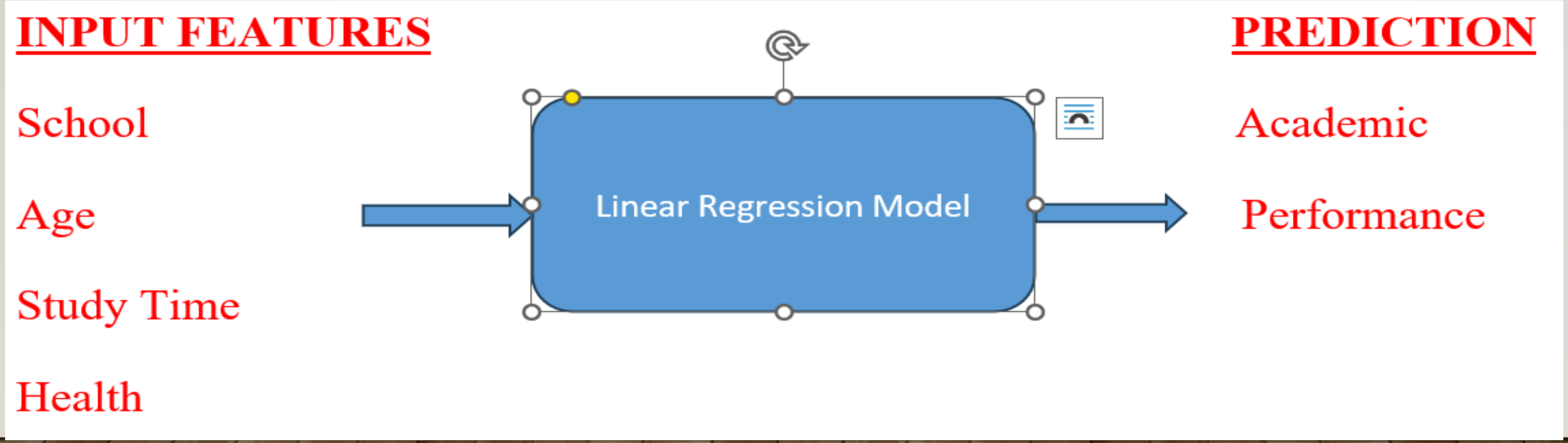
Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

A linear regression line has an equation of the form :

$$Y = a + bX$$

where ***X*** is the explanatory variable and ***Y*** is the dependent variable.

The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).



**INPUT FEATURES**

School

Age

Study Time

Health

Linear Regression Model

**PREDICTION**

Academic

Performance

## 1: Dataset

The database that has been used in the project includes students' achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features  and it was collected by using school reports and questionnaires.

## 2: Environment

Kaggle Code

## 3: Import Libraries

The two significant libraries that have been used are:

a) sklearn.preprocessing – It provides scaling, encoding, and imputation to prepare the data.

b)LabelEncoder - LabelEncoder is a class from the sklearn.preprocessing module It assigns a unique integer to each unique category in the data.

## 4: Exploratory Data Analysis and Visualization

a) We checked the dataset for null values with the help of a heatmap.

b) Then we analyzed the distribution of data with the help of histograms, scatterplot and heatmap.

c) Afterwards we created a correlation matrix to show the correlation between different variables.

d) At last we used label encoder algorithm to convert categorical features into numerical features.

## 5: Create Training and Testing Data

We split the data into two sets: a training set(80%) and a test set(20%). The training set will be used to train the machine learning model, and the test set will be used to evaluate the model's performance.

We used train_test_split() function which gives ouput as tuple of four arrays:

a)x_Train:The training data

b)x_Test:The test data

c)y_train: The labels for the training data.

d)y_Test: The labels for the test data.

## 6: Train the model using scikit-learn

We created a linear regression model  and used fit() method to fit the model to the training data.

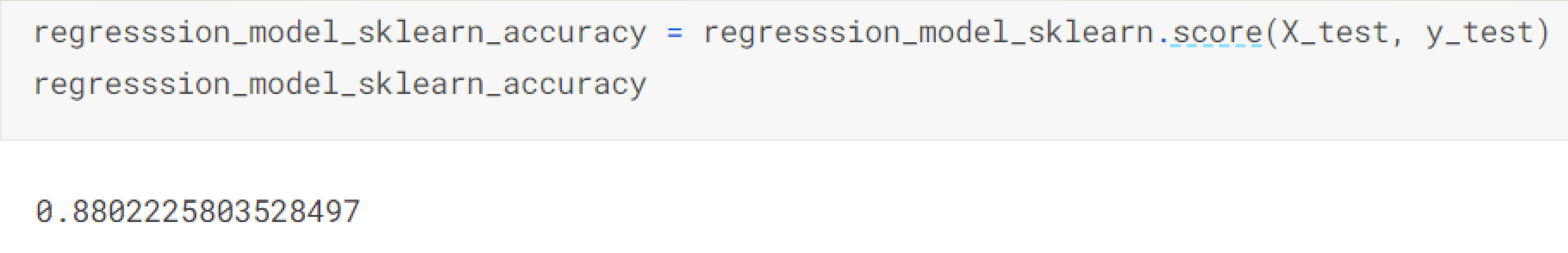
We chose the G3 column as dependent variable(response) and the rest as independent variables(predictors).

# Results and Discussion

## 1: Accuracy of the Model

**regresssion_model_sklearn.score(X_test, y_test)**: This line uses the score() method of the linear regression model to calculate the R-squared value. The X_test parameter represents the feature matrix of the testing data, and y_test represents the corresponding true target values.

**regresssion_model_sklearn_accuracy**: This variable stores the R-squared value, which is a measure of how well the linear regression model fits the testing data. It indicates the proportion of the variance in the target variable that can be explained by the model.



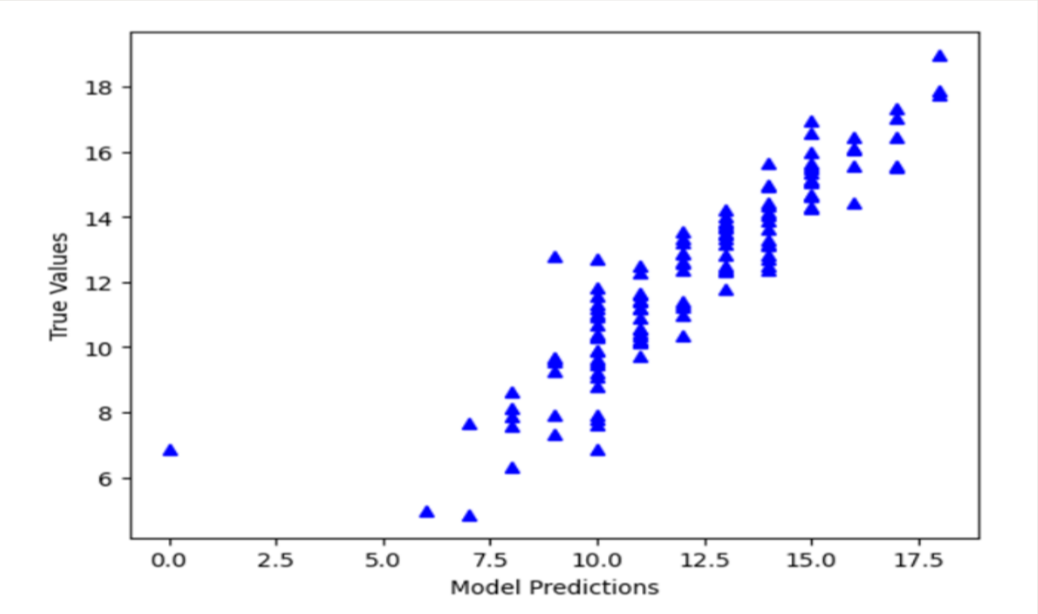
```
regresssion_model_sklearn_accuracy = regresssion_model_sklearn.score(X_test, y_test)
regresssion_model_sklearn_accuracy
```

```
0.8802225803528497
```

## 2: Plot the Results

Each data point on the plot represents an instance in the testing set, where the x-coordinate corresponds to the predicted value and the y-coordinate corresponds to the true value.

The points on the plot shows how well the model predictions match the true values. The closer the points are to the line y=x, the better the model predictions are.



# Conclusion and Future Work

We have explored the Academic Performance Project to learn regression using label encoding technique and visualized the results through different plots.

There are several other machine learning algorithms that can be used to predict academic performance.

**a) Decision Trees**: Versatile algorithms that can handle both categorical and continuous data. These can be used to predict academic performance by considering features such as demographics, study habits etc.

**b) Random Forests**: Ensemble learning technique that can handle high-dimensional data. It can capture non-linear relationships, and handle missing values effectively.

**c) Neural Networks**: Neural networks can handle complex data and capture intricate patterns. They can be used to predict academic performance by considering diverse features and leveraging the hierarchical learning capabilities of neural networks.