# TASK 1

# CAMPUS PULSE

## OVERVIEW

This report summarizes my approach and what I learned while solving the task. The task involved analyzing real, anonymized student data to build predictive models and analyzing the data.

# Level 1: Variable Identification Protocol

## OBJECTIVE

Use EDA techniques to uncover the identity of `Feature_1`, `Feature_2`, and `Feature_3`.

## APPROACH

- I tried to make 3 histograms, one for each of the 3 features, to see the frequency of each parameter given in the rows of each column, feature_1,2, and 3.
- Then I made a Correlation matrix for all three  features collectively in a single correlation matrix, so that I can get better and proper visualization of how these 3 anonymous features are related to other parameters, so that I can get insights on what these 3 could be
- Then, after making a proper guess as to what these features are, I drew scatter plots for each one differently to show they are likely the ones that I guessed.

## INSIGHTS

### *Feature_1*

- Feature_1 shows strong positive correlation with Failures(0.31) , Dalc(0.17) and Absences(0.12) and strong negative correlation with G1(-0.18) , Fedu(-0.14) , G2(-0.12)
- With this data, I found Feature_1 has an inverse relation with the ones that have a negative correlation and a direct relation with the ones that have a positive correlation.
- This means(kind of) that if grades go up, then to some extent Feature_1 goes down, and so when failures go up the Feature_1 goes up to some extent; hence Feature_ may mean **Stress Level or Academic Stress**

### *Feature_2*

- Feature_2 shows strong positive correlation with G1(0.26) , G2(0.25) and G3(0.25) and strong negative correlation with Absences(-0.13) , Dalc(-0.15) , Failures(-0.14)

- Feature_2 may mean **weekly/Daily Study hours or Academic interest** kind of thing, because as Grades go up to some extent, Feature_2 also goes up

*Feature_3*

- Feature_3 shows strong positive correlation with goout(0.4) , Dalc(0.62)(very strong) and freetime(0.15) and strong negative correlation with G3(-0.18) , G2(-0.17) , G1(-0.15).
- Feature_3 may mean **Social activities, or Party Frequency,** as it has such a great correlation with goout, alcohol consumption, and free time.

# Level 2: Data Integrity Audit

## OBJECTIVE

Detect and fix missing/inconsistent values.

## ACTIONS

- First, I used **df.isnull().sum()** so that I can get an overview of how many and what are columns have some unfilled cells.
- So for categorical columns, I used the most frequent data, that is mode of the columns, to fill those columns, and for the numeric features, I used the mean to fill the cells in such columns
- Hence, for numeric columns, I added mean in place of ‘Nan’, and for categorical data, I imported SimpleImputer from sklearn.impute, and hence I could add mode in place of ‘Nan’

## CHECKS

- I checked again using **df.isnull().sum()** if any else nan is remaining, which was not checked for duplicate rows using **df.duplicated().sum() ,** which was also not present.

# Level 3: Exploratory Insight Report

## OBJECTIVE

Ask and explore at least 5 interesting questions about student data.

## Questions and Plots

1. Does **More Internet access ensure more marks,** or is it vice versa?
    - Box Plot of Internet access(No Internet and Internet) VS Final grade(G3)

2. Does **going out more mean less grade?**

    - Box plot of Frequency of Going Out VS Final Grade(G3)

3. **Do Health and Absences** have anything in common?

    - Scatter Plot of Health Status VS and no. of absences

4. Is there any relationship between **weekday alcohol consumption and final grades(G3)?**

    - Box plot

5. Do more **educated parents mean more grades**?

    - Grouped bar Graph between Medu(mother's education) and G3(final grades) with the colour of each bar representing the father's education

6. Does **going out more mean the student is in a relationship?**

    - Bar Graph between the frequency of going out and the number of students with colours of a graph showing not or yes in a relationship.

# Level 4: Relationship Prediction Model

## OBJECTIVE

Build and evaluate classification models to predict relationship status

## APPROACH

- Data preprocessing
- Model training
- Model Evaluation

## INSIGHTS

The model suggests that:

- Students who go out more frequently are more likely to be in a relationship
- Students with stronger family relationships are slightly less likely to be in a relationship
- Academic performance has a mixed influence on relationship status.
- Logistic Regression shows the highest accuracy that is of 63.79%.

This simple model achieves reasonable accuracy, though more sophisticated models and additional feature engineering could improve performance.

# Bonus Level

**Plot 1:Logistic Regression**
Straight-Line boundary(linear separation for binary classification)
**Plot 4:KNN**
localized boundaries(instance-based voting creates irregular shapes).
**Plot 5:Linear Regression**
Linear boundary(continous)