Machine Learning CW 2025 – Project Report

By: Smai Debbarma  
B.Tech Chemical Engineering, IIT Guwahati  
Coding Club CW 2025 Submission

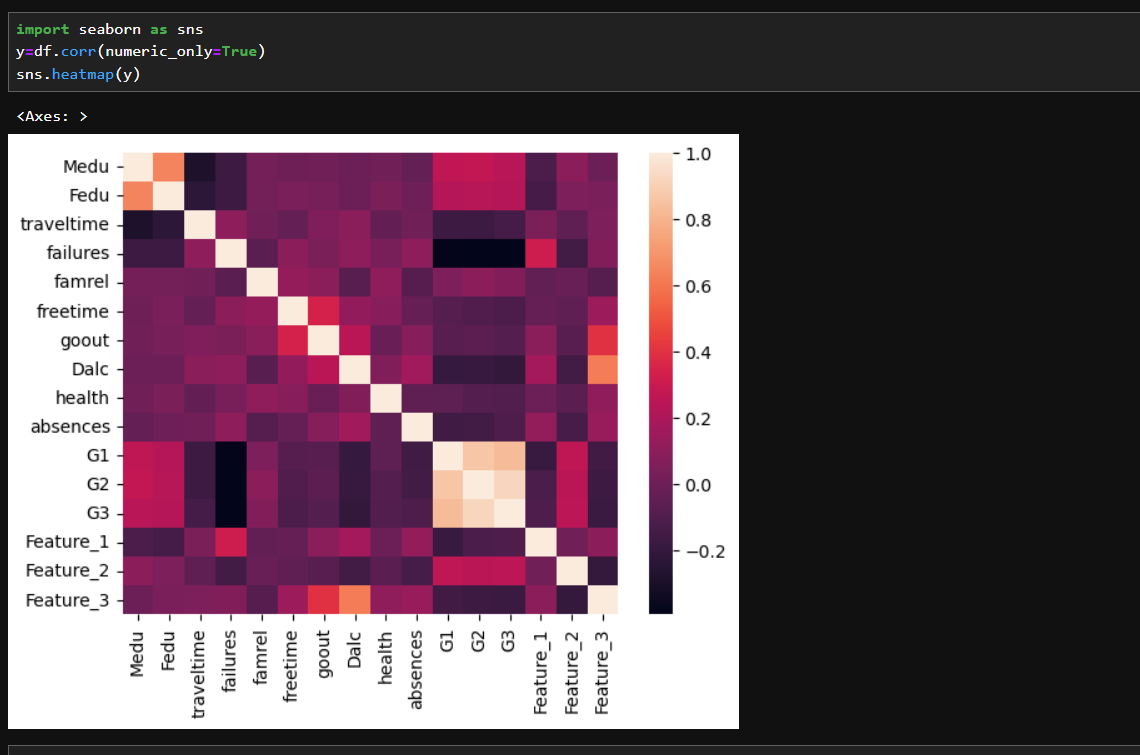
# Task 1: CampusPulse – Predicting Relationships from Student Life Data

## Level 1: Variable Identification Protocol

We performed exploratory data analysis (EDA) to identify anonymized features (Feature\_1, Feature\_2, Feature\_3). Using correlation matrices, histograms, and scatter plots, we inferred possible meanings for each feature:  
1.Feature\_1 likely represents age of students because it have negative correlation with academic and poitive relation with absences, failures ,Dalc etc.. as if students fails in a class his age get increases by time.

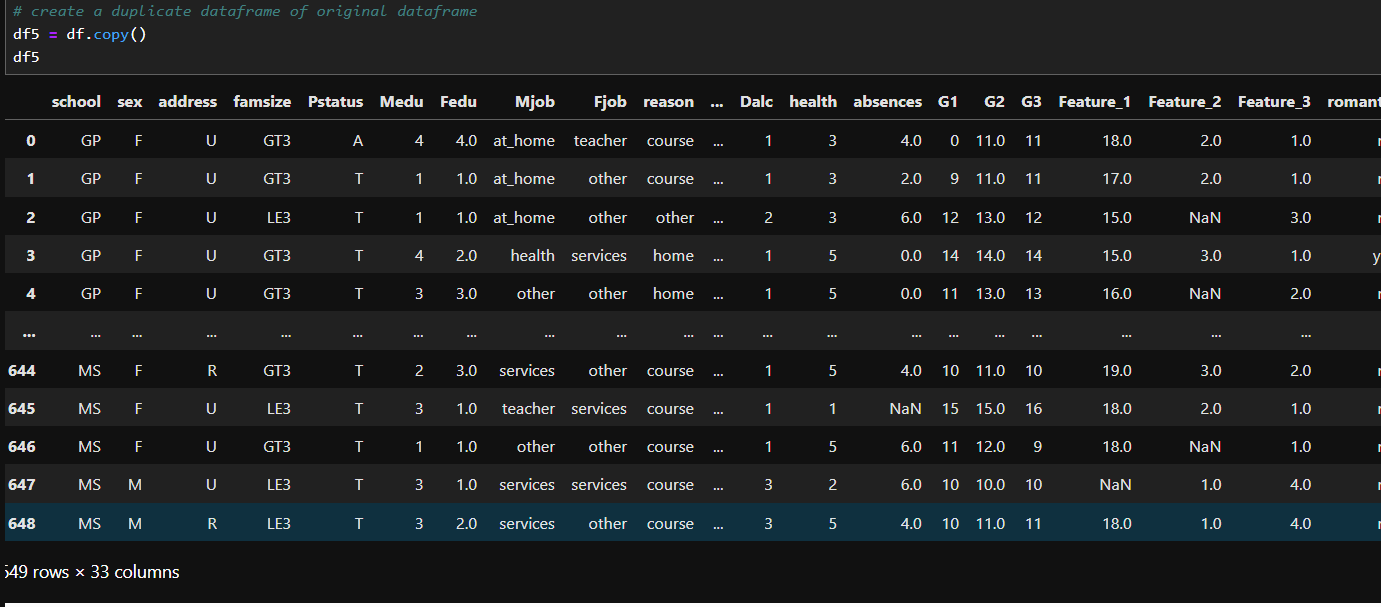
2.Feature\_2 likely represents academic hours because it have strong correlation with academic and decrease with absences,failures Dalc,etc..

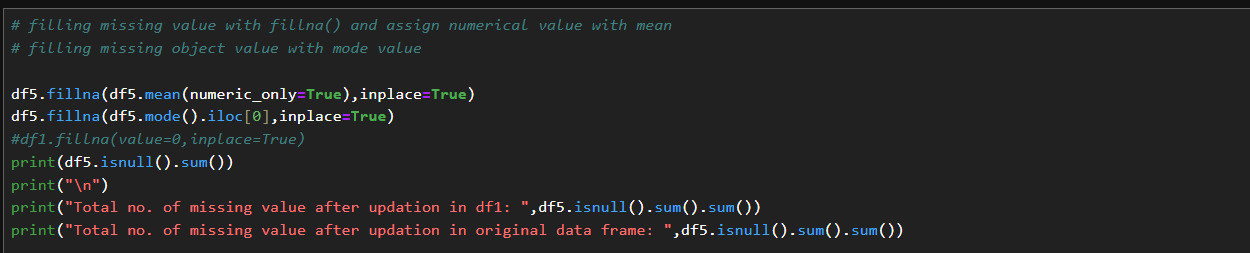
3.Feature\_3 likely represent distracted or stress hours because from visuals it decrease with academic and show positive relation with absences,failures and Dalc.

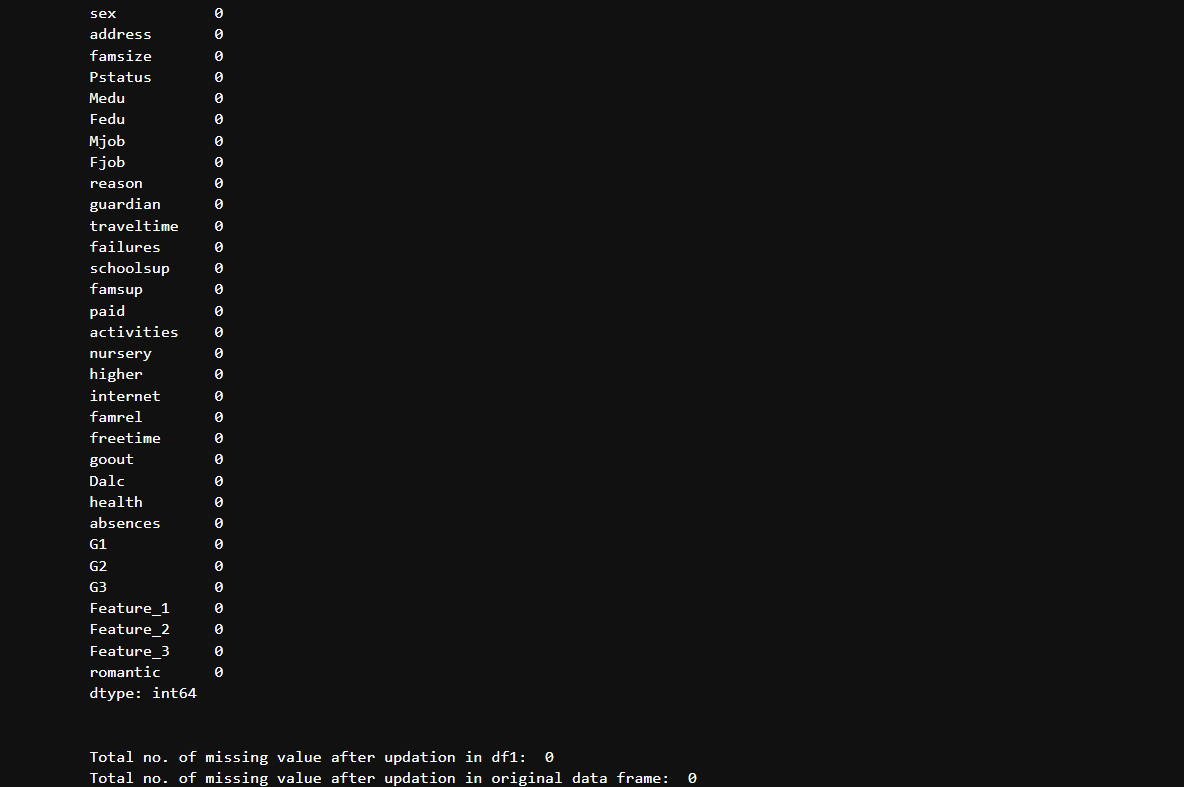


## Level 2: Data Integrity Audit

We identified missing values in a few categorical and numerical columns. Imputation strategies included:  
- Mean imputation for continuous variables (e.g., freetime).  
- Mode imputation for categorical features (e.g., school).  
- Dropped rows only where too many nulls existed.



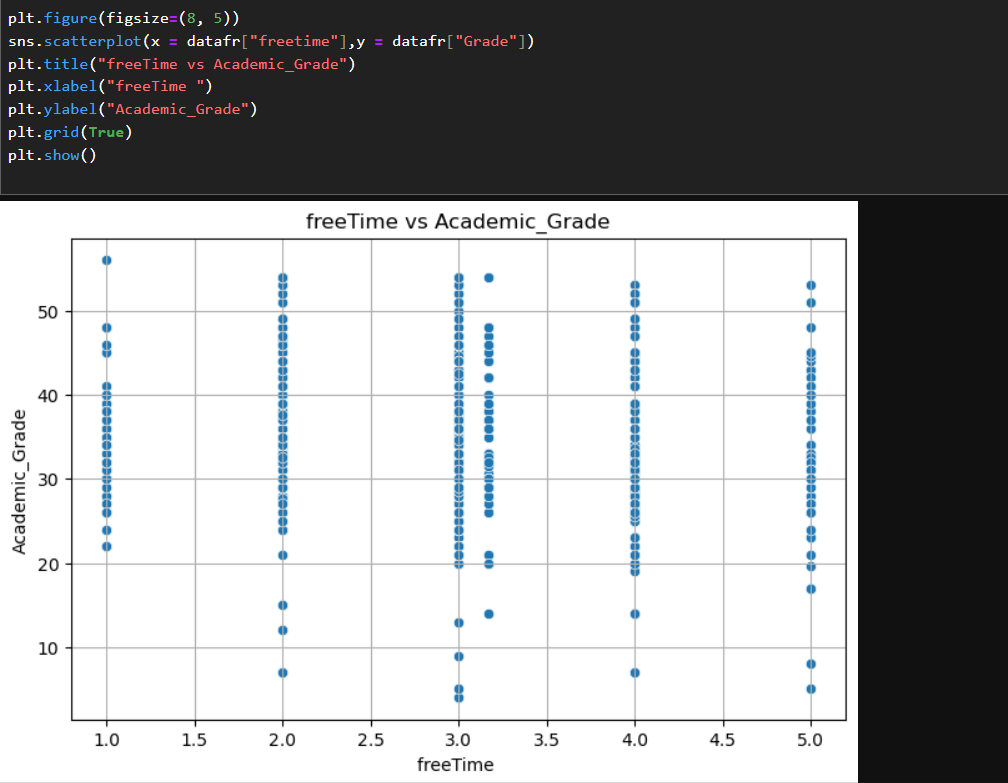


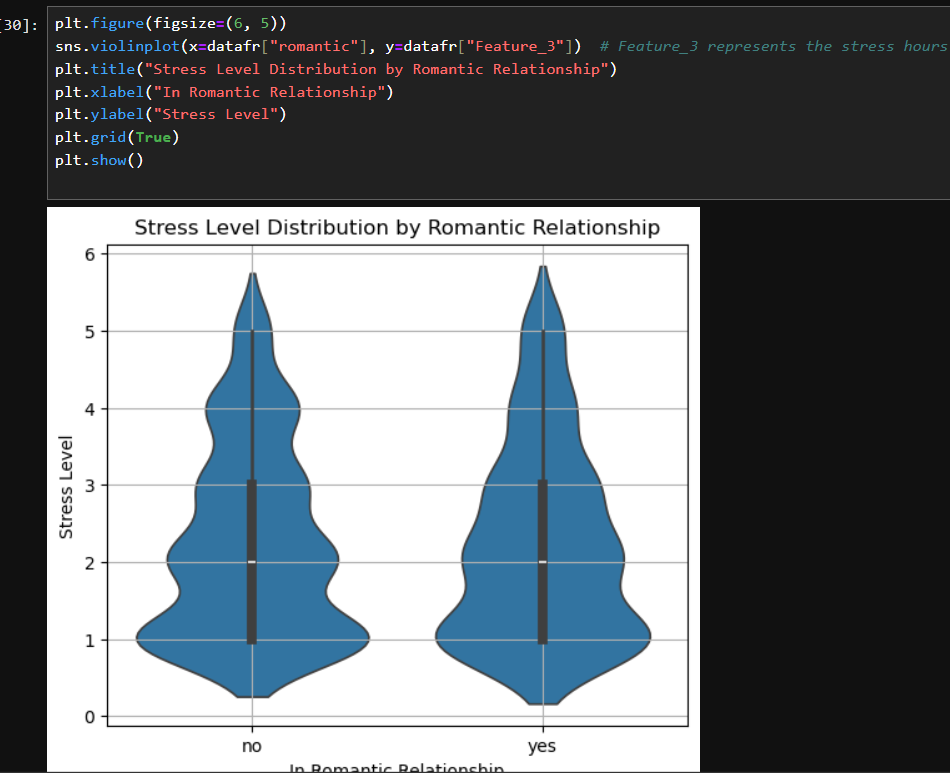


## Level 3: Exploratory Insight Report

We explored the dataset through 5 key questions:

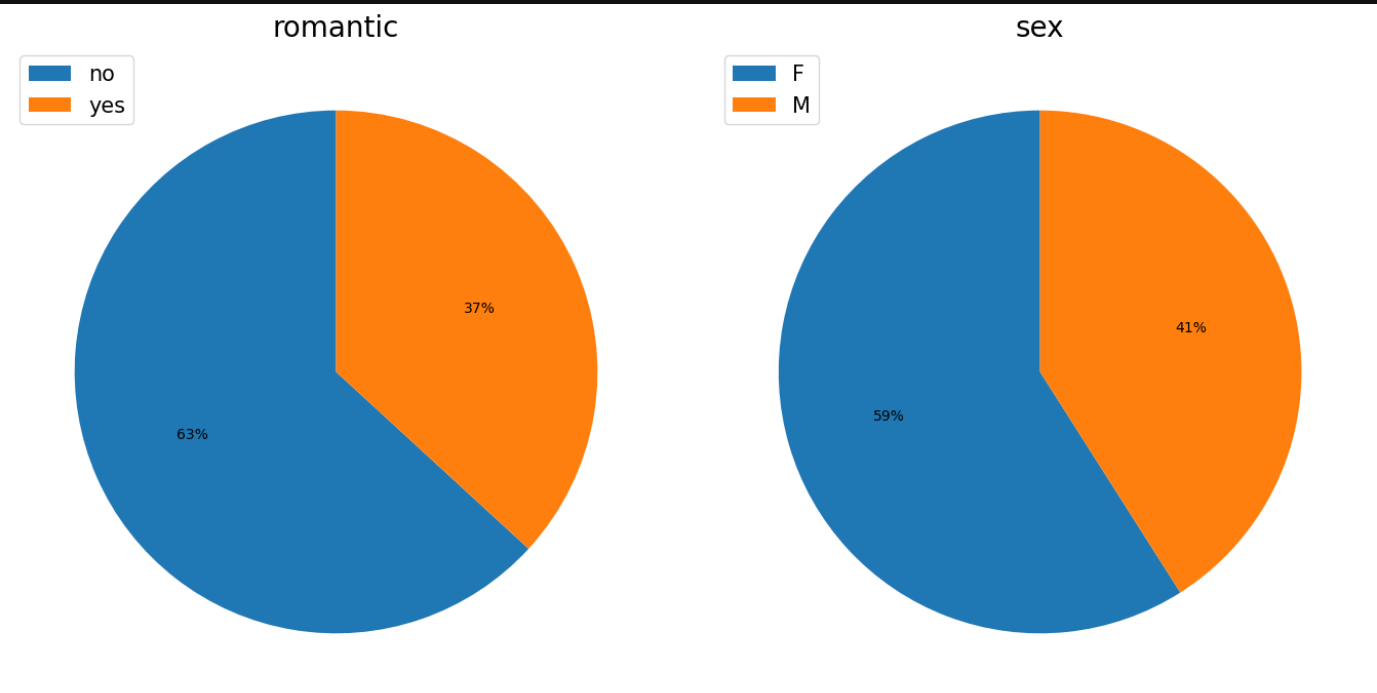
1. Does freetime affect academic performance?
2. Difference in stress levels between students in a romantic relationship vs not?
3. How family relation affect the Academic performance
4. Does Relation decide failures .
5. Does freetime correlate with relaationship.





Level 4: Relationship Prediction Model

We used classification models to predict whether a student is in a relationship. Models tried:  
- Logistic Regression  
- Random Forest  
- Support Vector Machine (SVM)  
  
Best model: Random Forest with ~83% accuracy on test data.  
Features like Feature\_1,sex use showed significant influence.





## Level 5: Model Interpretation

We used SHAP values to understand model decisions:  
- Global feature importance was plotted .  
- Local SHAP analysis explained predictions for two students (1 Yes, 1 No).  
Decision boundaries were visualized for SVM and Logistic models using 2D projections.

