Project Report: Machine Learning CW 2025

Submitted by - Maithilee Sagare

---

1. Campus Pulse

1.1 Problem Statement

The objective of the CampusPulse project is to analyze anonymized student survey data and uncover hidden factors influencing student life on campus, including predicting whether a student is likely to be in a romantic relationship based on lifestyle and academic behaviors.

1.2 Dataset Description

The dataset provided for the CampusPulse task consists of real, anonymized survey responses from students at IIT Guwahati. The data aims to capture a holistic view of student life by collecting information across three major dimensions:

1. **Behavioral Patterns**
   Includes data on daily routines, such as sleep duration, screen time, and physical activity.
   Captures social engagement levels, frequency of club participation, and time spent on social media.

2. **Academic Attributes**
   Includes GPA, hours spent studying per week, class attendance, and academic satisfaction.
   These features help gauge a student's academic focus and performance trends.

3. **Lifestyle and Wellbeing Factors**
   Captures subjective responses such as stress levels, mental health status, and work-life balance.
   Some features include quantitative self-ratings, while others reflect categorical choices.

Anonymized Features

To protect student privacy, three survey responses have been renamed as: Feature_1, Feature_2, Feature_3.
These anonymized variables are integral to the analysis, and the task requires identifying their likely meanings using **exploratory data analysis (EDA)** and **statistical correlations**.

Target Variable

In_a_relationship: A binary categorical variable with values:

"Yes" — The student reported being in a romantic relationship

"No" — The student reported not being in one

This variable serves as the **label** for the classification.

Data Quality Notes:

The dataset includes **missing values** in several columns, likely due to skipped or incomplete responses in the original survey.
A **data cleaning and imputation strategy** is necessary before building any predictive models.

1.3 Level - wise Approach

LEVEL 1

Here the objective was to identify the anonymized features by inferring their nature and correlation with other features using EDA tools.
Approach- I used **correlation heatmap** first to understand how the anonymized features relate to known variables like grades, absences, alcohol, etc. **Pearson correlation** was used to measure linear relationships.

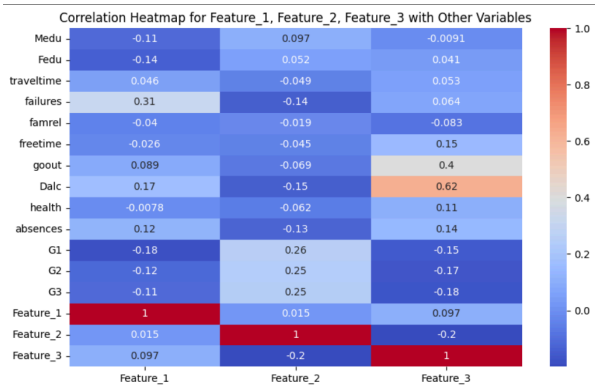
From the heatmap, positive correlations of the anonymized features:

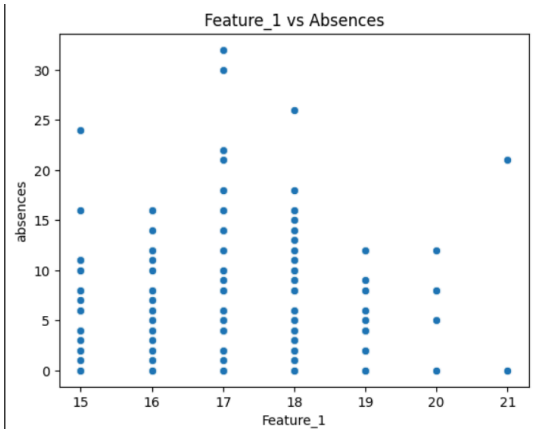
Feature_1 Correlations: failures: +0.31, absences: +0.12

Feature_2 Correlations: G1: +0.26, G2: +0.25, G3: +0.25

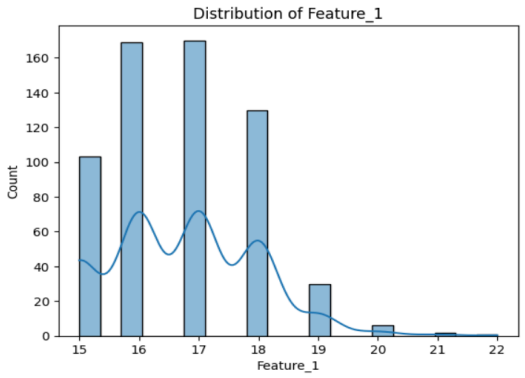
Feature_3 Correlations: Dalc: +0.62, goout: +0.40

From the correlation heatmap , I identified **which known variables are most strongly correlated** with each anonymized feature. To understand these relationships visually and support our feature identification, we now plot **targeted scatterplots** — one feature at a time — against the most relevant variables.





Scatter plots allow us to visually explore and confirm the nature of the relationships highlighted by the correlation heatmap. By plotting each anonymized feature against its most strongly correlated known variable, we can observe the pattern, strength, and any potential outliers in the data, providing deeper insight beyond correlation coefficients.



Histograms reveal the distribution of each feature—showing its shape, spread, and presence of outliers. Understanding whether a feature is symmetric, skewed, or has multiple modes helps in selecting appropriate imputation methods and informs further analysis decisions.

RESULTS –

Feature_1 is most likely a disciplinary or academic risk score
Feature_2 is most likely an academic score related to G1
Feature_3 is most likely related to outgoing nature or lifestyle of student.

LEVEL 2

The objective was to find and handle missing values of the data by applying appropriate imputation strategies.
Number of missing values were detected using `df.isnull().sum()`
These values were imputed using the following functions:

```
df['feature'].fillna(df['feature'].mode()[0])
df['Feature'].fillna(df['Feature'].median())
df['Feature'].fillna(df['Feature'].mean())
```

For categorical and nominal data mode was used, for ordinal discrete data median was used and for numerical symmetric data mean was used to impute the values.

LEVEL 3
Objective: Ask at least 5 insightful questions about the student data.
For each:
-Build a meaningful plot (bar chart, violin plot, scatterplot, etc.)
-Write a brief interpretation of the insight.

Questions were framed based on the relation of features using correlation heatmap.

Q1. Does parental education affect student's academic performance?
Q2. Does weekday alcohol consumption affect academic performance?
Q3. How does free time influence going out?
Q4. Does living lifestyle affect student performance?
Q5. Does travel time reduce student's free time?

Box plot and Violin plot are used to interpret these questions
- Violin Plot shows the distribution shape of grades (`G3`) for each level of weekday alcohol consumption (`Dalc`) and `failures`. It helps us understand density, skewness, and whether grades cluster or spread out.
- Box Plot overlays key statistics: median, quartiles, and outliers, providing a quick summary of central tendency and variability.

LEVEL 4

Objective: Apply classification techniques to model relationship likelihood and assess their performance. Think critically about what the models reveal, and what they don't.

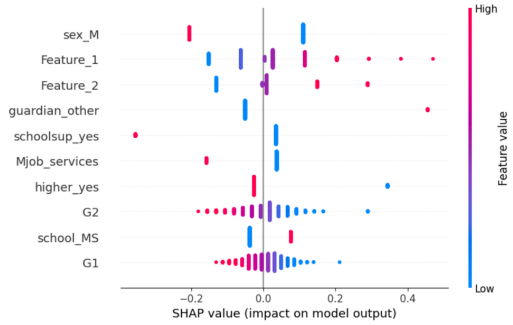
Here, I considered three models - logistic regression, random forest and naive bayes as the target variable was dichotomous.

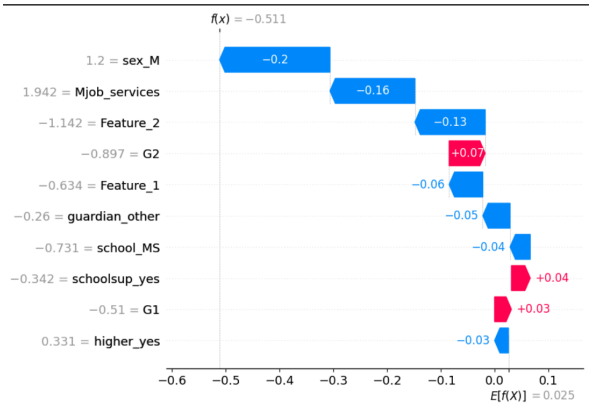
After evaluating the performance I found that logistic regression was best suitable for this problem statement. The goal was to predict and understand the feature influence on romantic relationship. Naive bayes assumes feautre independence which is why it was not suitable. Logistic regression provides coefficients which helps explain which factors increase or decrease the likelihood of being in a relationship.

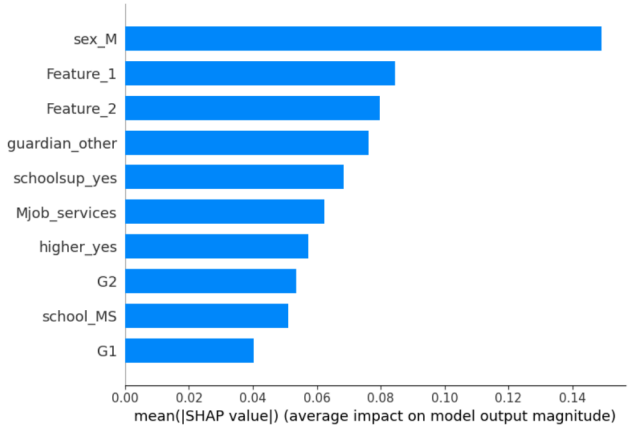
It is often used as a benchmark model in binary classification due to reliable performance. A random forest model often has a risk of overfitting and requires pruning. It also has a slow training speed and low interpretability.

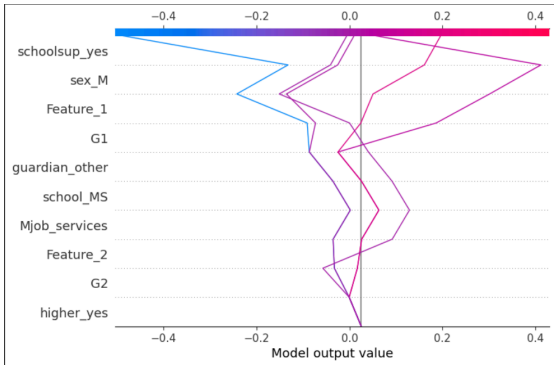
The accuracy of the model is 62%, with F1-score of 0.69 and ROC-AUC of 0.65

LEVEL 5









Interpreting the model's predictions was a crucial step in understanding the underlying patterns driving relationship outcomes. By leveraging interpretability techniques such as feature importance analysis and SHAP (SHapley Additive exPlanations) values, the model's decision-making process was made transparent. This allowed for the identification of the most influential features that contributed to predicting relationship status, providing valuable insights into the key factors affecting the outcome. For instance, features like sex, feature1, etc emerged as significant predictors, influencing the model's confidence in classification. Visualizations such as SHAP summary plots and dependence plots were used to illustrate how individual features affected predictions across different samples, enhancing interpretability for both technical and non-technical stakeholders. This interpretability not only boosted trust in the model but also facilitated actionable insights that could guide further analysis or decision-making processes in real-world applications.