CampusPulse Initiative – Task 1 Report
Machine Learning Coding Week 2025 – IIT Guwahati
Author: Ritesh Raj Singh
Roll Number: 240101079

🔍 Level 1: Variable Identification Protocol

- **Feature_1 → Age**
    - Values ranged from 15 to 22
    - Histogram centered at 16–18, aligns with college age
    - Correlated with `failures` (more failures, later entry age)
- **Feature_2 → Study Time**
    - Values ranged from 1 to 4
    - Positively correlated with G1, G2, G3
    - Indicates more study time, better grades
- **Feature_3 → Extrovertedness**
    - Correlated with `Dalc` (alcohol) and `goout`, negatively with grades
    - Suggests a behavioral trait related to social activity

🧹 Level 2: Data Integrity Audit

Columns with missing values:

- `famsize`: dropped rows
- `Fedu`: median imputation

- `traveltime`, `higher`, `freetime`, `absences`: filled with contextually logical defaults (e.g., 0)
- `G2`, `age`, `studytime`, `extrovert`: filled using median

All strategies were justified with context or distribution-based logic.

📊 Level 3: Exploratory Insight Report

1. **Absences vs Romantic Status**
    - Box plot showed higher absences among romantically involved students.
2. **Family Relationship Quality**
    - Romantic students showed lower average family relationship scores.
3. **Mother's Education vs Paid Classes**
    - Higher maternal education correlated with more paid class participation.
4. **Address vs Alcohol Consumption**
    - Urban students had more low-alcohol consumption (rating = 1), rural students showed more variation.
5. **Travel Time vs Final Grade (G3)**
    - Slight downward trend in grades with higher travel time.

🤖 Level 4: Predictive Modeling

- Categorical variables were label encoded
- Split dataset into training and test (23% test)
- Models used:

    - **Logistic Regression**
        - Accuracy: 67%
    - **Random Forest Classifier**
        - Accuracy: 60%

Tried both full-feature and reduced-feature sets to optimize performance.

🧠 Level 5: Model Reasoning and SHAP

- **SHAP Summary Plot** used for feature importance.
    - Most influential features: age, G2, absences
- **SHAP Force Plots** created for individual "Yes" and "No" predictions to explain model behavior.

Interpretation showed that age and academic performance were influential in predicting romantic relationships.

Conclusion:
This task emphasized a full ML pipeline—cleaning, exploring, modeling, and interpreting. The predictive model (logistic regression) reached a reasonable accuracy (~67%) and offered transparent explanations through SHAP visualizations.
