Handling Null Values

Initial Strategy:
To address missing values in the dataset, I first computed the **mean**, **median**, and **mode** for all relevant columns. For categorical variables, I filled missing values using the **mode**, while for numerical features, I experimented with both mean and median imputation strategies.

Refined Strategy:

- For columns like `Fedu` and `Medu`, I observed a **high correlation** through a heatmap. Thus, missing values in one were imputed using the values from the other.
- For other features, the earlier strategy (mean/median/mode imputation) was retained.

Understanding Hidden Features

Feature_1:

- **Initial Hypothesis:** Suspected this to be **age** based on `.describe()` — mean ≈ 17, min = 15, max = 22 — which matches the high school age range.
- **Validation:**
  - Plotted box plots of Feature_1 against `absences`, `health`, `traveltime`, and `freetime`.
  - Notable insight: Students aged 20–21 tend to have **higher absences**, possibly due to work/family responsibilities, supporting the assumption that Feature_1 represents **age**.

Feature_2:

- **Exploration:**
  - `.describe()` showed that it takes only 4 discrete values (1–4).
  - Countplots and bar charts with `failures`, `G1`, `G2`, and `G3` revealed:
    - A **negative correlation with failures**
    - A **positive correlation with academic performance**

- **Interpretation:** Feature_2 likely captures **study hour levels** rather than IQ, as its distribution is skewed toward lower values — not consistent with expected IQ distribution.

Feature_3:

- **Exploration:**
    - Showed strong positive correlation with `Dalc` (weekday alcohol use) and `goout` (socializing).
- **Conclusion:** Feature_3 likely represents **extroversion level** — students scoring higher are more social and drink more frequently.

Exploratory Questions Raised

1. How does **parental education** influence academic performance?
2. Are there significant differences in **urban vs. rural** student profiles?
3. What is the impact of **parental separation** on students?
4. How does being in a **romantic relationship** affect grades?
5. Do **alcohol consumption patterns** vary between students in relationships and those not?

Romantic Relationship Prediction Modeling

Problem Statement: Predict whether a student is in a romantic relationship (binary classification).

Modeling Strategy

- **Initial Consideration:** Linear Regression was discarded due to the categorical nature of the target variable.
- **Chosen Models:**

- **Gaussian Naive Bayes**
- **Logistic Regression**
- **Random Forest Classifier**

Naive Bayes Classifier

Pipeline:

1. **Preprocessing:**
   - Converted target column (`romantic`) to binary values.
   - One-hot encoding for categorical features.
   - Created **combined features** from correlated columns to reduce redundancy.
2. **Feature Selection:**
   - Retained only features with correlation > ±0.08 with the target.
   - Visualized top correlations using heatmaps.
3. **Feature Scaling and Transformation:**
   - Applied polynomial transformation (degree 2) and standardization.
4. **Model Training and Evaluation:**
   - Used **Stratified K-Fold Cross-Validation (k=5)**.
   - Evaluated using **accuracy**, **precision**, and **confusion matrix**.

Results:

- Accuracy: ~70%
- Precision: 71% (No), 67% (Yes)
- **Note:** Despite good performance, this model was not interpretable enough, so I explored tree-based and linear models for SHAP analysis.

Logistic Regression & Random Forest

Pipeline Strategy:

1. **Train-Test Split:** 80% training, 20% testing using stratified sampling.
2. **Hyperparameter Tuning:** Performed using `GridSearchCV`.
3. **Standardization:** Applied before model fitting to normalize features.
4. **Evaluation:** Accuracy, classification report, confusion matrix (visualized via heatmap).
5. **Model Comparison:** Both models achieved ~62% accuracy. Logistic Regression had slightly better interpretability and confusion matrix performance.

Model Reasoning & Interpretation

To visually understand how features impact the model:

Decision Boundary Plot:

- Constructed between **Feature_1 (age)** and **absences**.
- Followed standard steps: select features → create mesh grid → model prediction → plot with contour.

SHAP for Model Explainability

Global Interpretation Strategy:

1. Extract preprocessing components (scaler + model) from the pipeline.
2. Scale and reformat data for SHAP.
3. Use:
    - **TreeExplainer** for Random Forest
    - **LinearExplainer** for Logistic Regression
4. Generate **summary plots** to visualize global feature importance.

Local Interpretation Strategy:

1. Scale test samples.
2. Use trained model to make predictions.
3. Run SHAP on test samples.
4. Select two cases (predicted as Yes & No).
5. Use **waterfall plots** to see how each feature influenced the individual prediction.

Final Conclusion

The key drivers for romantic relationship prediction (as observed in SHAP analysis) were:

- **Feature_1 (Age)**
- **Absences**
- **Grades (G1, G2, G3)**
- **Guardian type (especially "other")**