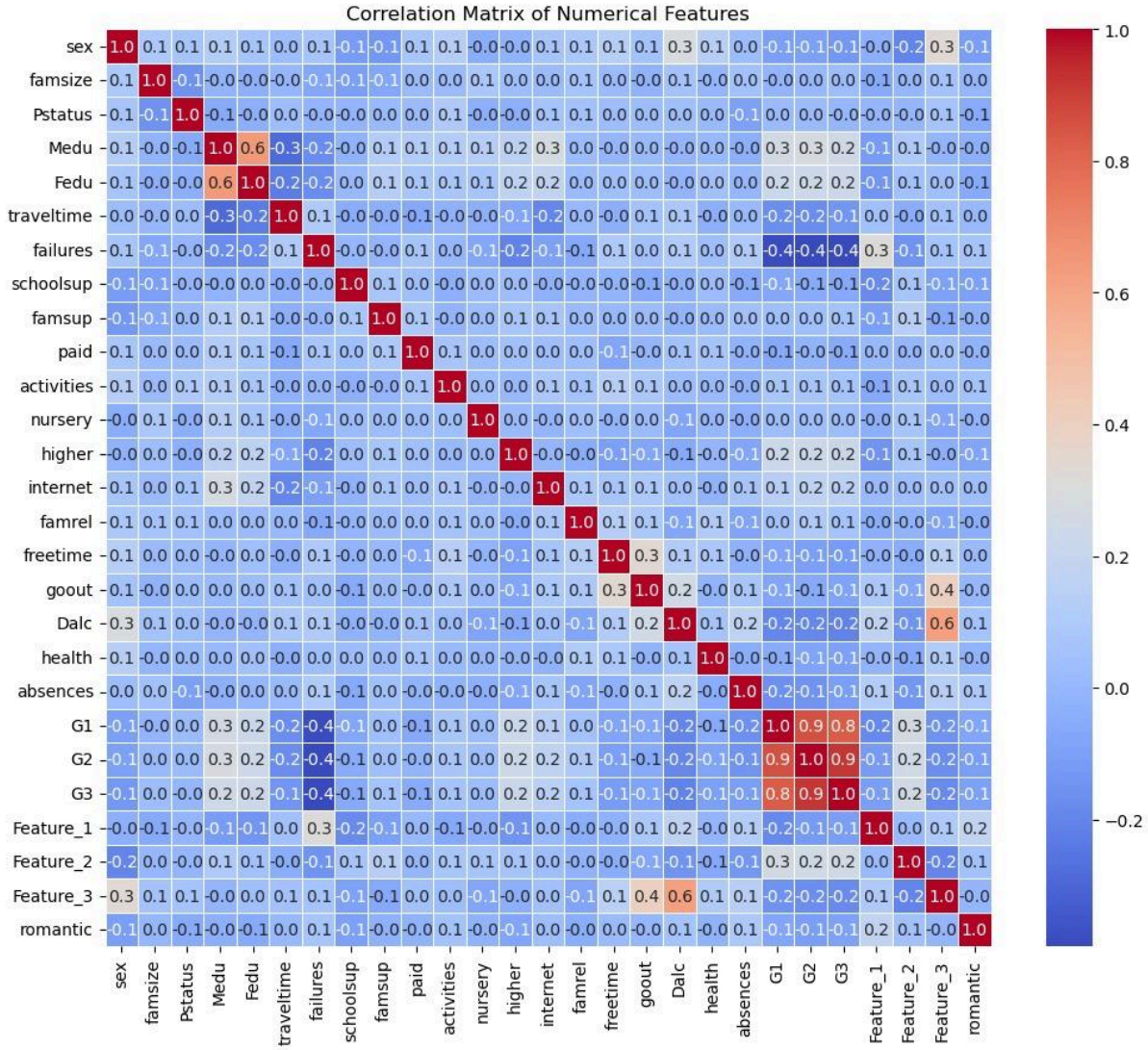
CODING WEEK - ML

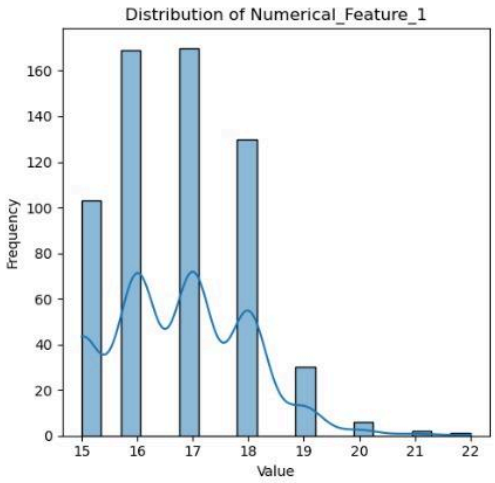
TASK 1
The CampusPulse Initiative

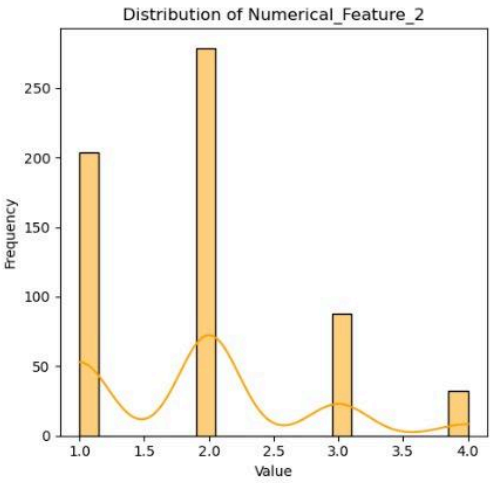
By
Sahana Athota
240102246

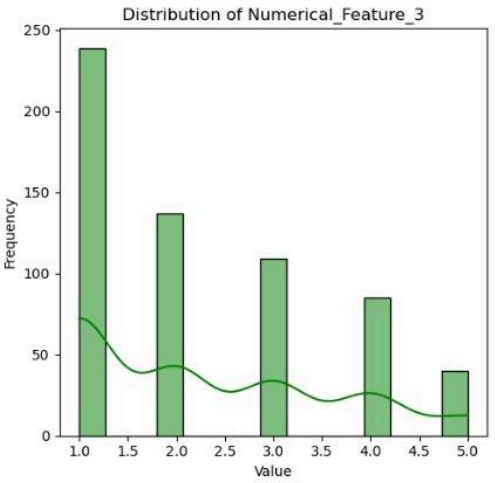
Level 1: Variable Identification Protocol

Using graphical analysis and statistical correlation, we inferred the real-world meaning of each feature as follows:









Feature 1: Age

Feature 1 spans from 15 to 22, with a high concentration in the 15-18 range before tapering off from 19 onward, suggesting that it represents age. Additionally, its strong correlation with both failures and daily alcohol consumption further supports the idea that Feature 1 is linked to age, as these factors often vary with age-related patterns.

Feature 2: Study Time

Feature 2 displayed a strong correlation with academic performance, with the corresponding graph revealing a discrete pattern where higher values were associated with better outcomes. Given that Feature 2 ranged from 1 to 4 and that G1, G2, and G3 improved as its values increased, this strongly suggests that Feature 2 represents study hours per day.

Feature 3: Weekly alcohol consumption frequency

The third feature exhibited a strong correlation with the number of outings(0.4) and daily alcohol level(0.6). Given this pattern and that 1 is the highest and 5 is the least chosen, it likely corresponds to a student's weekly alcohol consumption frequency.

.

Level 2: Data Integrity Audit

Handling missing data:

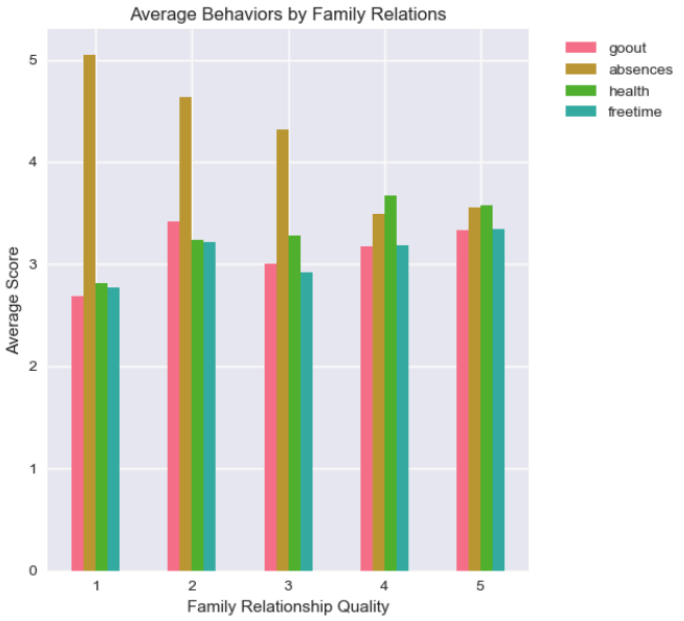
```
famsize       50
Fedu          73
traveltime    73
higher        76
freetime      45
absences      69
G2            35
Feature_1     38
Feature_2     46
Feature_3     39
```

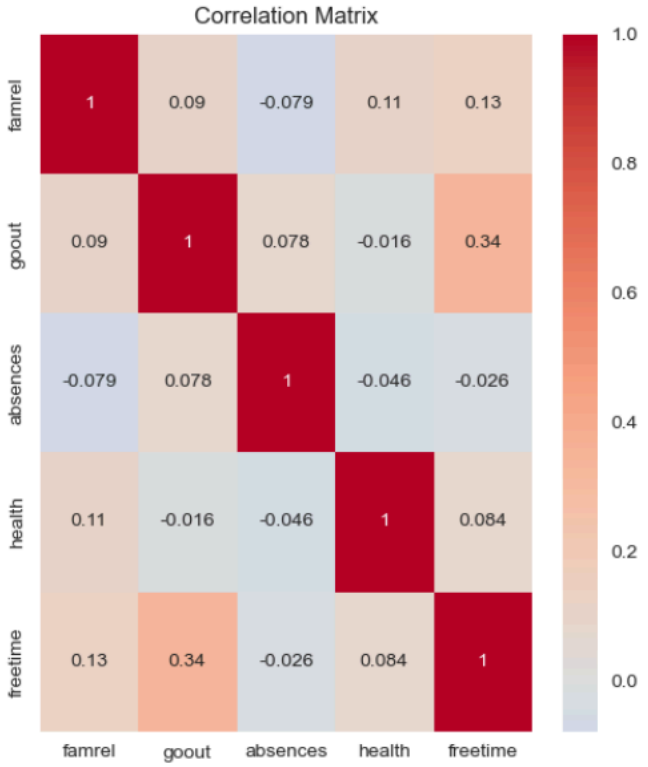
Solution:

- **Numerical Features (Fedu, traveltime, freetime, absences, Feature_1):**
  Missing values were replaced with the **median** of the respective columns. Since these features are numerical and may contain outliers, the median serves as a more robust measure compared to the mean, ensuring that extreme values do not unduly influence the imputed data.
- **Categorical/Ordinal Features (famsize, higher, G2, Feature_2, Feature_3):**
  For these features, missing values were filled using the **most frequent value** in the respective columns. As these variables are either categorical or ordinal, imputing with the mode helps preserve the distribution while minimizing artificial variations in the data.

Level 3: Exploratory Insight Report

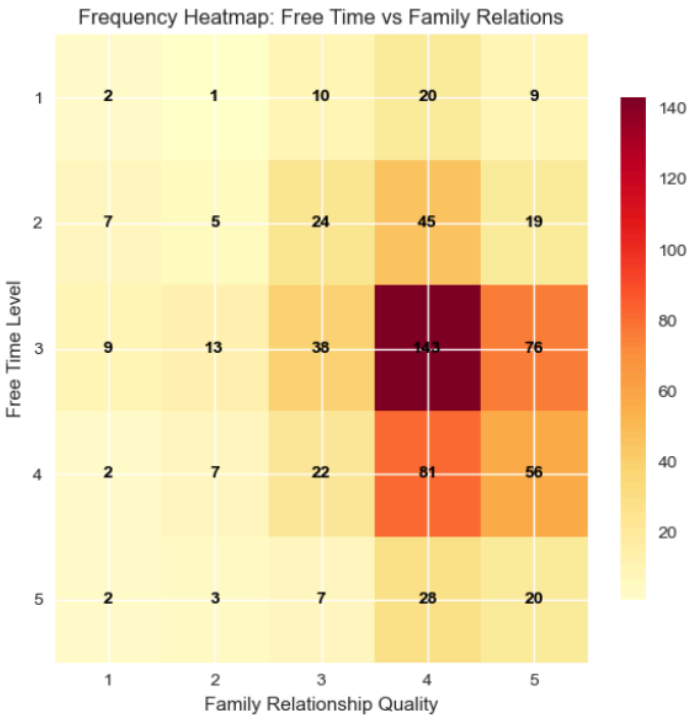
Q1. Do students with better family relationships exhibit different social behaviour?

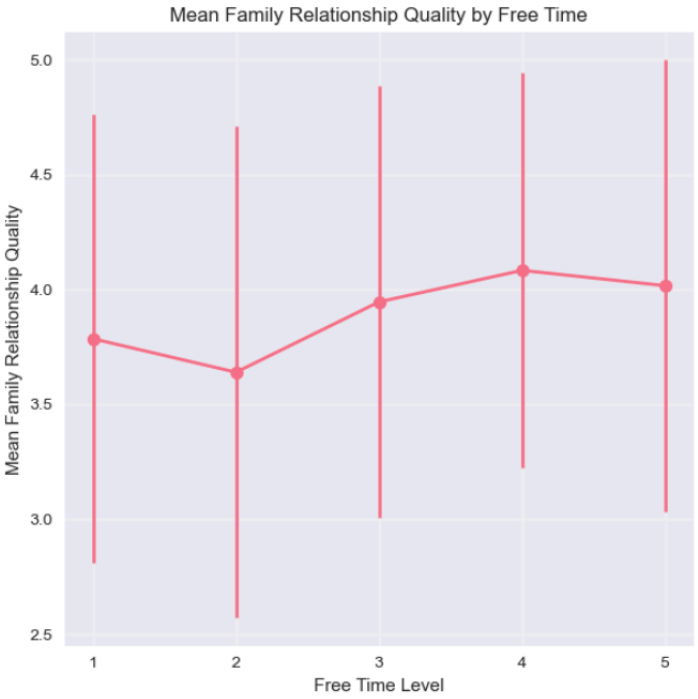




INTERPRETATION:
Students with better family relationships tend to:
• Have fewer school absences (better attendance)
• Report better health status
• This suggests family stability creates overall life balance

Q2. Is there an optimal ‘free time’ level for family relationships?

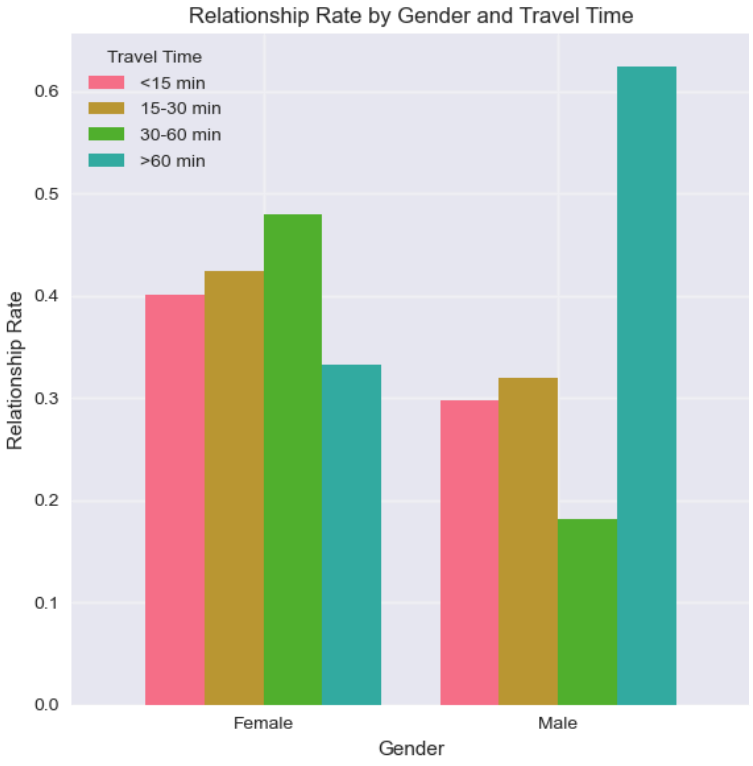


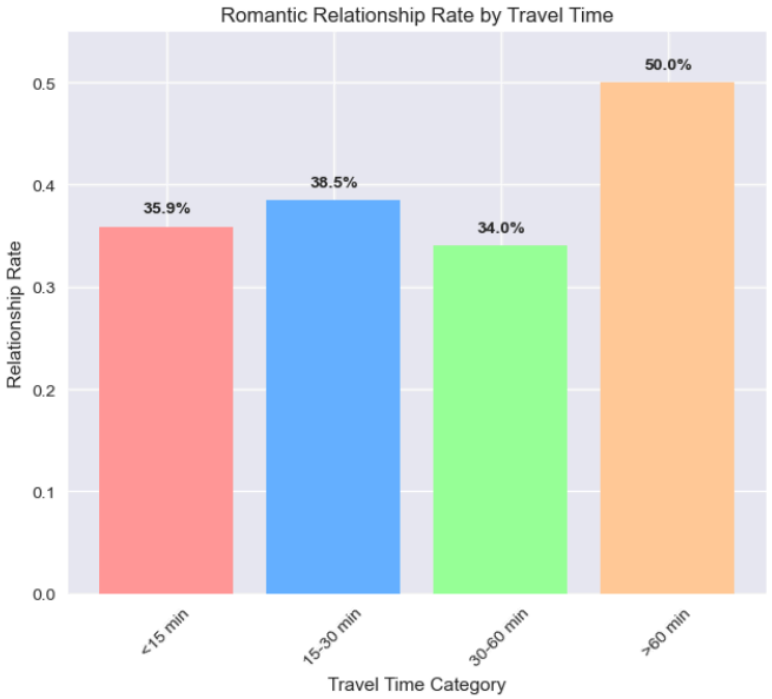


Interpretation:

- The optimal free time level for family relationships is: 4.0
- Students with free time level 4.0 have the highest mean family relationship quality (4.083)
- 81.5% of students with this free time level have good family relationships

Q3. How does travel time affect relationship formation?






```
ROMANTIC RELATIONSHIP RATES BY TRAVEL TIME
<15 min:
  Total students: 393
  In relationships: 141 (35.9%)
  Not in relationships: 252 (64.1%)

15-30 min:
  Total students: 195
  In relationships: 75 (38.5%)
  Not in relationships: 120 (61.5%)

30-60 min:
  Total students: 47
  In relationships: 16 (34.0%)
  Not in relationships: 31 (66.0%)

>60 min:
  Total students: 14
  In relationships: 7 (50.0%)
  Not in relationships: 7 (50.0%)
```

Interpretation:

- Relationship rates show a increasing trend with longer travel times
- Longer travel times may provide opportunities to meet people during commute
- Students living further may be more mature or independent.

Level 4: Relationship Prediction Model

1. Logistic Regression

Confusion matrix:
[[65 16]
[39 10]]
Accuracy: 0.5769230769230769
Classification Report:

	precision	recall	f1-score	support
0	0.62	0.80	0.70	81
1	0.38	0.20	0.27	49
accuracy			0.58	130
macro avg	0.50	0.50	0.48	130
weighted avg	0.53	0.58	0.54	130

ROC AUC Score: 0.5314

2. K Nearest Neighbours (KNN)

Confusion matrix:
[[76 5]
[47 2]]
Accuracy: 0.6
Classification Report:

	precision	recall	f1-score	support
0	0.62	0.94	0.75	81
1	0.29	0.04	0.07	49
accuracy			0.60	130
macro avg	0.45	0.49	0.41	130
weighted avg	0.49	0.60	0.49	130

ROC AUC Score: 0.4793

3. Support Vector Machine(SVM)

Confusion matrix:
[[80 1]
[49 0]]
Accuracy: 0.6153846153846154
Classification Report:

	precision	recall	f1-score	support
0	0.62	0.99	0.76	81
1	0.00	0.00	0.00	49
accuracy			0.62	130

```
   macro avg       0.31      0.49      0.38       130
weighted avg       0.39      0.62      0.47       130

ROC AUC Score: 0.5231
```

4. Kernel SVM (rbf)

```
Confusion matrix:
[[79  2]
 [48  1]]
Accuracy:  0.6153846153846154
Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.98      0.76        81
           1       0.33      0.02      0.04        49

    accuracy                           0.62       130
   macro avg       0.48      0.50      0.40       130
weighted avg       0.51      0.62      0.49       130

ROC AUC Score: 0.5276
```

5. Naive Bayes

```
Confusion matrix:
[[56 25]
 [31 18]]
Accuracy:  0.5692307692307692
Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.69      0.67        81
           1       0.42      0.37      0.39        49

    accuracy                           0.57       130
   macro avg       0.53      0.53      0.53       130
weighted avg       0.56      0.57      0.56       130

ROC AUC Score: 0.5062
```

6. Decision Tree

```
Confusion matrix:
[[47 34]
 [25 24]]
Accuracy:  0.5461538461538461
Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.58      0.61        81
           1       0.41      0.49      0.45        49

    accuracy                           0.55       130
```

```
   macro avg       0.53      0.54      0.53       130
weighted avg       0.56      0.55      0.55       130

ROC AUC Score: 0.5350
```

7. Random Forest

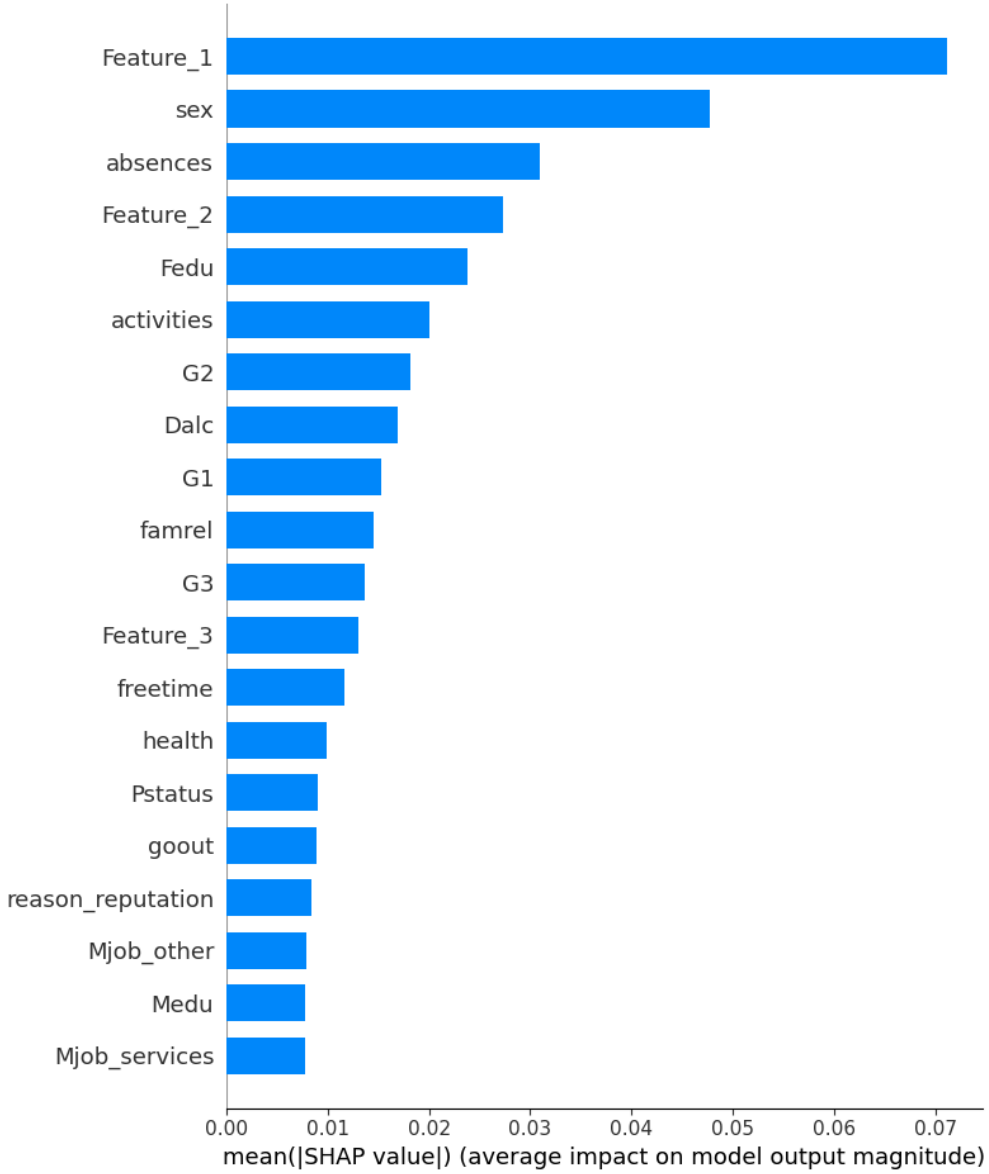
```
Confusion matrix:
[[60 21]
 [31 18]]
Accuracy:  0.6
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.74      0.70        81
           1       0.46      0.37      0.41        49

    accuracy                           0.60       130
   macro avg       0.56      0.55      0.55       130
weighted avg       0.58      0.60      0.59       130

ROC AUC Score: 0.5870
```

Prefered Model: Random Forest

Since it has a higher ROC AUC score compared to the other classifier models.

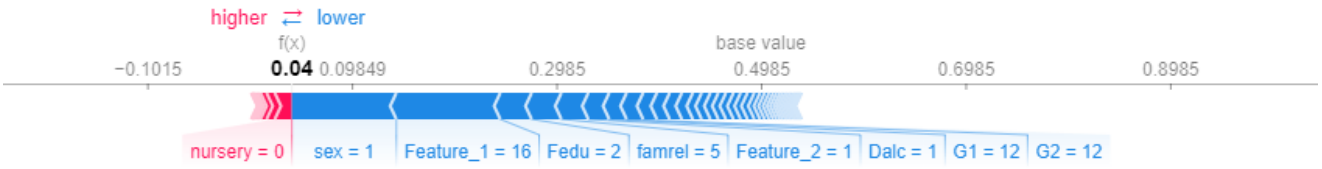
Level 5: Model Reasoning & Interpretation

Feature importance:



SHAP Force Plot:

For a student predicted ‘No’:



This SHAP force plot explains why the student is predicted to be very unlikely to be in a romantic relationship.

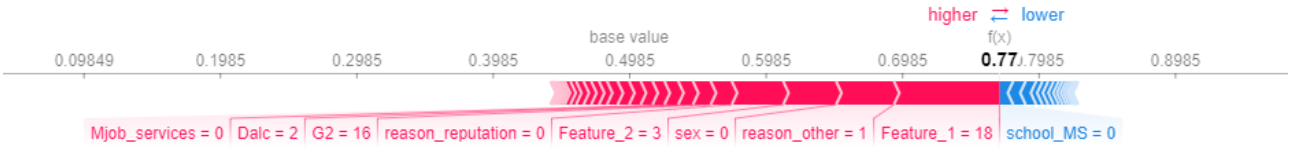
A. Prediction (f(x)): 0.04 (extremely low), significantly below the Base Value (average prediction): 0.4985.

B. Key Factors Pushing Prediction *Lower* (Blue Bars):
   - sex = 1: Being of a certain sex (male) strongly decreases the predicted likelihood.
   - Feature_1 (Age) = 16: This feature, with a value of 16, is a strong negative contributor.
   - Fedu = 2: Father's education level of 2 decreases the likelihood.
   - famrel = 5: A perfect family relationship score surprisingly pushes the prediction lower.
   - Feature_2(Study time) = 1, Dalc = 1, G1 = 12, G2 = 12: These all provide further, smaller pushes decreasing the predicted likelihood. G1 and G2 being 12 might indicate average grades, still contributing negatively.

C. Key Factor Pushing Prediction *Higher* (Red Bar):
   - nursery = 0: Not attending nursery school provides the only positive push, but its effect is minimal compared to the combined negative factors.

In short, this student's very low predicted likelihood of being in a romantic relationship is driven primarily by their sex, specific values for 'Feature_1', 'Fedu', and surprisingly, a high 'famrel' score, along with average grades.

For a student predicted ‘Yes’:



This SHAP force plot explains why a student is predicted to be likely in a romantic relationship.

A. Prediction (f(x)): 0.77 (high likelihood), significantly above the Base Value (average prediction): 0.4985.
B. Key Positive Factors (Red Bars, pushing higher):
    - Mjob_services = 0 (Mother's job not services)
    - Dalc = 2 (Moderate daily alcohol consumption)
    - G2 = 16 (High grade 2)
    - reason_reputation = 0 (Not choosing school for reputation)
    - Feature_2(Study time) = 3 (high study time)
    - sex = 0 (Likely female)
    - reason_other = 1 (Some other reason for choosing school)
    - Feature_1(Age) = 18
C. Key Negative Factor (Blue Bar, pushing lower):
    - school_MS = 0 (Not attending school 'MS')

The student's high predicted likelihood of being in a romantic relationship is driven by a strong combination of factors including high academic performance (G2), moderate alcohol consumption (Dalc), their sex (female), specific school choice reasons, with only a minor counteracting effect from their school type (MS).

What really drives relationship prediction?

Students are more likely to be predicted in a romantic relationship if they have:

- Good (but not necessarily perfect) grades: A G2 grade of 16 was a strong positive.
- Some level of social engagement: Moderate workday alcohol consumption (Dalc=2) was a positive sign, suggesting a social aspect.
- Specific background factors: being of a specific sex (female), and certain reasons for choosing their school could increase the likelihood.
- Certain unknown characteristics: Features like 'Age' and 'Study hours' when at specific values also strongly boost the prediction.

Students are less likely to be predicted in a romantic relationship if they have:

- Very low or extremely high grades: Both very poor grades (e.g., G1, G2, G3 in the low single digits) and surprisingly, perfect grades tend to decrease the predicted likelihood. This might suggest a 'sweet spot' for grades, or that extreme focus on academics (both very poor or very good) could detract from relationships.
- Perfect attendance: Having zero absences can oddly decrease the likelihood, perhaps implying less social interaction or a very rigid focus.
- Surprisingly strong family relationships: A very high family relationship score (famrel=5) can sometimes be a negative indicator, which is counter-intuitive.

In essence, the model isn't looking for extremes. It seems to favor students with good (but not necessarily perfect) academic performance, a certain level of social engagement, and specific demographic or background attributes. Conversely, students who are either struggling significantly academically or are too academically focused (perfect grades, no absences), lack social activities, are predicted to be less likely in a relationship.

Bonus Level: The Mystery Boundary Match

Plot 1:Decision Tree

- The decision boundaries are composed entirely of straight lines that are parallel to the x-axis and y-axis. This staircase pattern is a property of tree-based models.
- A large region might be split horizontally, and then one of those sub-regions might be split vertically, and so on, creating a nested structure. This reflects the hierarchical nature of how decision trees make decisions.

Plot 2: Random Forest

- Random forests combine multiple decision trees which leads to complex boundaries.
- The many small patches of blue inside red regions (and vice versa) are caused by the combined voting of many decision trees.

Plot 3: Kernel Support vector machine (SVM)

- This boundary is a smooth, flowing curve. This indicates a model capable of capturing more complex, non-linear relationships between the features. SVMs are excellent at creating smooth, curved decision boundaries. The shape of the boundary is very typical of an SVM with a non-linear kernel.

Plot 4: Naive Bayes

- The smooth, curved boundary that forms an elliptical or oval shape, encapsulating one class (red) within a larger region of the other class (blue). This kind of boundary is characteristic of models that assume the data for each class follows a Gaussian distribution.

Plot 5: K-Nearest Neighbors (KNN)

- The boundary shows many small islands of one class within the dominant region of the other. This indicates that the classification is determined very locally based on the nearest neighbors of a point, which is the principle of KNN.