ML Task 1

CampusPulse Initiative

Name: Aryan Chakravorty

\

Table of Contents

Level 1: Variable Identification Protocol

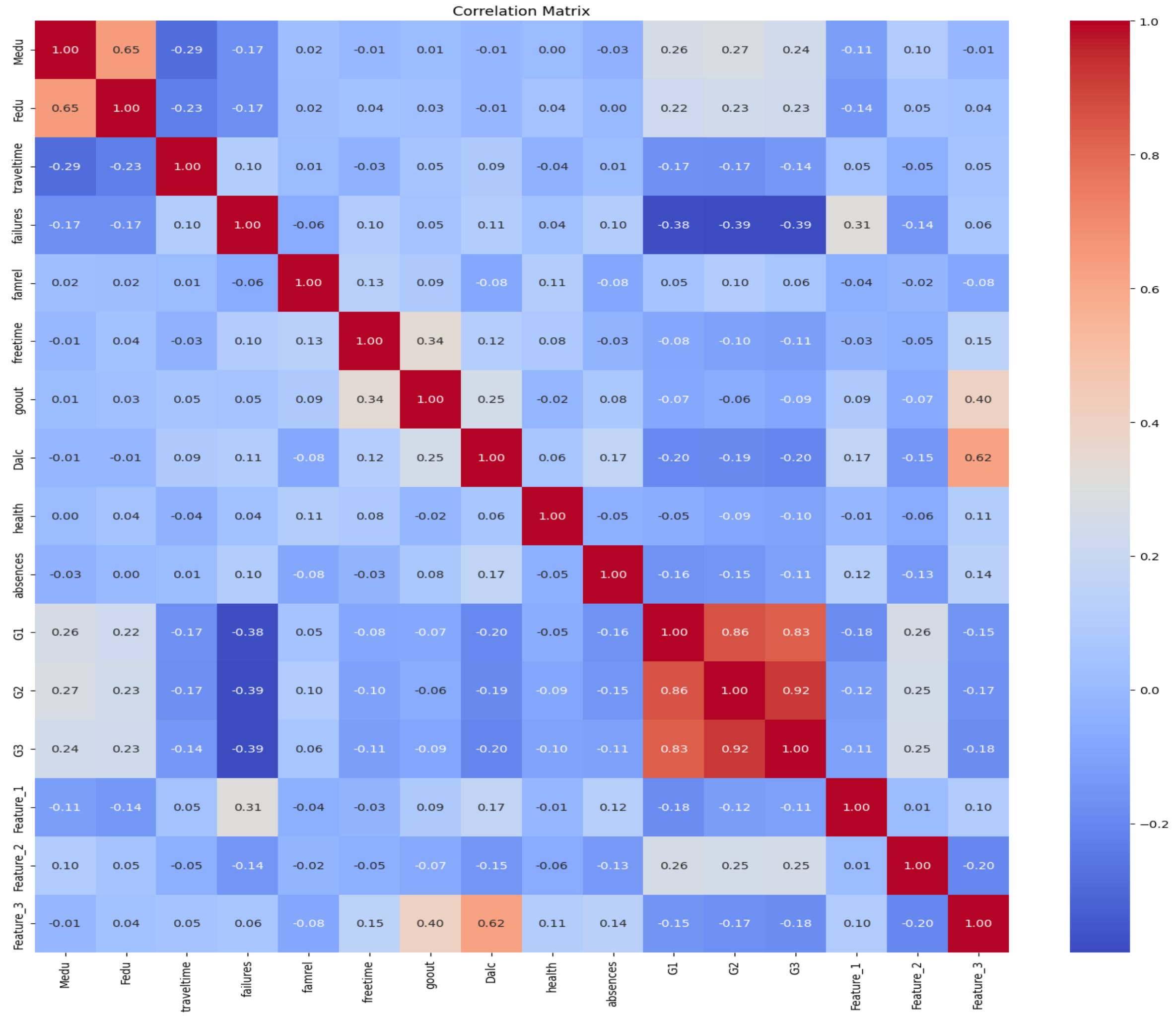
The objective was to identify three anonymized survey variables (Feature_1, Feature_2, Feature_3) based on their statistical patterns using Exploratory Data Analysis (EDA), without revealing raw labels.

Approach:

To identify the likely identities of the anonymized features, the following EDA steps were conducted:

- **Correlation Heatmap:**
  Computed and visualized the Pearson correlation matrix for all features. Strong correlations between known features and the anonymous ones (e.g., 0.62 between Feature_3 and Dalc, and 0.40 between Feature_3 and goout) provided clues.

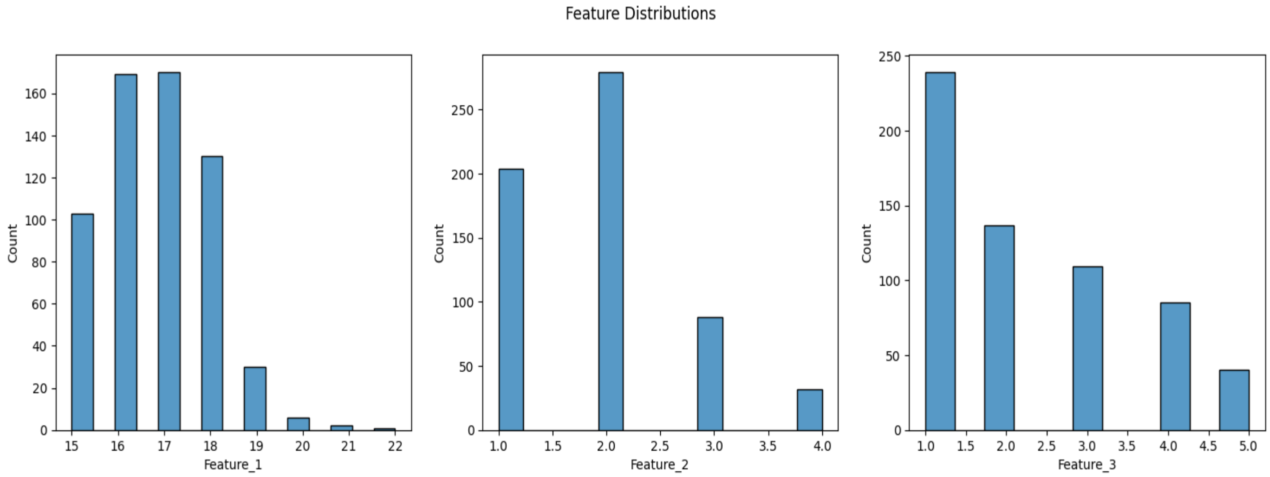
Image: Correlation Matrix



	Medu	Fedu	traveltime	failures	famrel	freetime	goout	Dalc	health	absences	G1	G2	G3	Feature_1	Feature_2	Feature_3
Medu	1.00	0.65	-0.29	-0.17	0.02	-0.01	0.01	-0.01	0.00	-0.03	0.26	0.27	0.24	-0.11	0.10	-0.01
Fedu	0.65	1.00	-0.23	-0.17	0.02	0.04	0.03	-0.01	0.04	0.00	0.22	0.23	0.23	-0.14	0.05	0.04
traveltime	-0.29	-0.23	1.00	0.10	0.01	-0.03	0.05	0.09	-0.04	0.01	-0.17	-0.17	-0.14	0.05	-0.05	0.05
failures	-0.17	-0.17	0.10	1.00	-0.06	0.10	0.05	0.11	0.04	0.10	-0.38	-0.39	-0.39	0.31	-0.14	0.06
famrel	0.02	0.02	0.01	-0.06	1.00	0.13	0.09	-0.08	0.11	-0.08	0.05	0.10	0.06	-0.04	-0.02	-0.08
freetime	-0.01	0.04	-0.03	0.10	0.13	1.00	0.34	0.12	0.08	-0.03	-0.08	-0.10	-0.11	-0.03	-0.05	0.15
goout	0.01	0.03	0.05	0.05	0.09	0.34	1.00	0.25	-0.02	0.08	-0.07	-0.06	-0.09	0.09	-0.07	0.40
Dalc	-0.01	-0.01	0.09	0.11	-0.08	0.12	0.25	1.00	0.06	0.17	-0.20	-0.19	-0.20	0.17	-0.15	0.62
health	0.00	0.04	-0.04	0.04	0.11	0.08	-0.02	0.06	1.00	-0.05	-0.05	-0.09	-0.10	-0.01	-0.06	0.11
absences	-0.03	0.00	0.01	0.10	-0.08	-0.03	0.08	0.17	-0.05	1.00	-0.16	-0.15	-0.11	0.12	-0.13	0.14
G1	0.26	0.22	-0.17	-0.38	0.05	-0.08	-0.07	-0.20	-0.05	-0.16	1.00	0.86	0.83	-0.18	0.26	-0.15
G2	0.27	0.23	-0.17	-0.39	0.10	-0.10	-0.06	-0.19	-0.09	-0.15	0.86	1.00	0.92	-0.12	0.25	-0.17
G3	0.24	0.23	-0.14	-0.39	0.06	-0.11	-0.09	-0.20	-0.10	-0.11	0.83	0.92	1.00	-0.11	0.25	-0.18
Feature_1	-0.11	-0.14	0.05	0.31	-0.04	-0.03	0.09	0.17	-0.01	0.12	-0.18	-0.12	-0.11	1.00	0.01	0.10
Feature_2	0.10	0.05	-0.05	-0.14	-0.02	-0.05	-0.07	-0.15	-0.06	-0.13	0.26	0.25	0.25	0.01	1.00	-0.20
Feature_3	-0.01	0.04	0.05	0.06	-0.08	0.15	0.40	0.62	0.11	0.14	-0.15	-0.17	-0.18	0.10	-0.20	1.00

Histograms:
Plotted the distribution of each feature. Uniform or bimodal distributions were flagged for categorical traits. Skewed or bell-shaped curves hinted at quantitative variables.

<u>Image: Feature distribution</u>



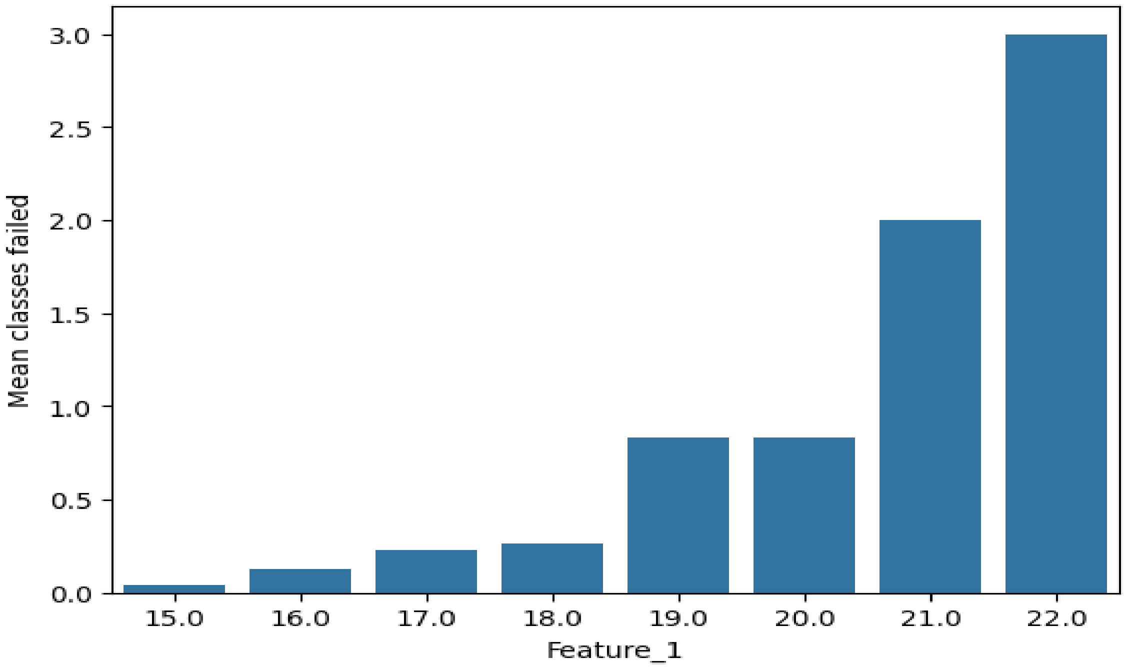
Feature 1:

Observations:
The histogram peaks around ages 15–18, which is consistent with a student population.

Inference:
Feature_1 likely represents age.

Justification:
The mean number of failed classes increases with age, suggesting older students may have backlogs.

Image: Feature 1 vs Mean Classes Failed

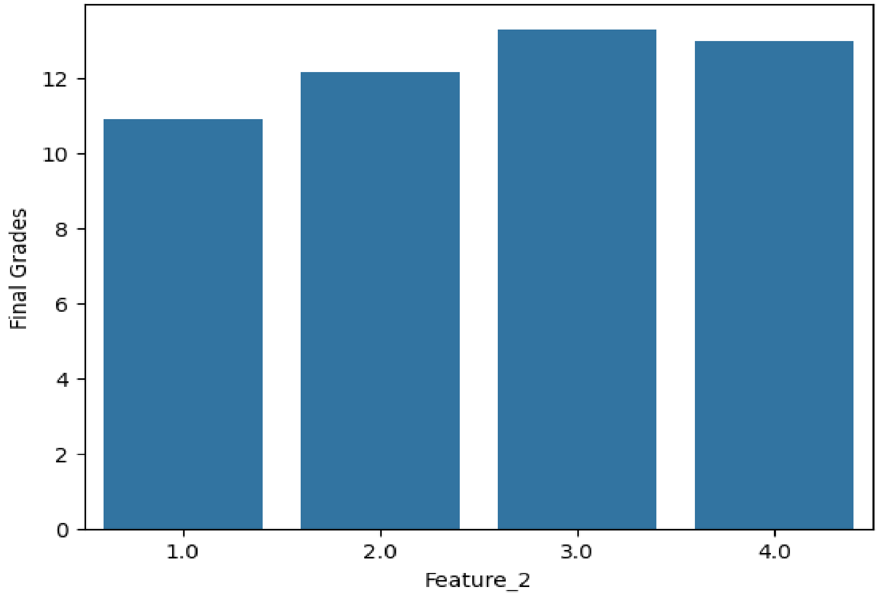


Feature 2

Observations:

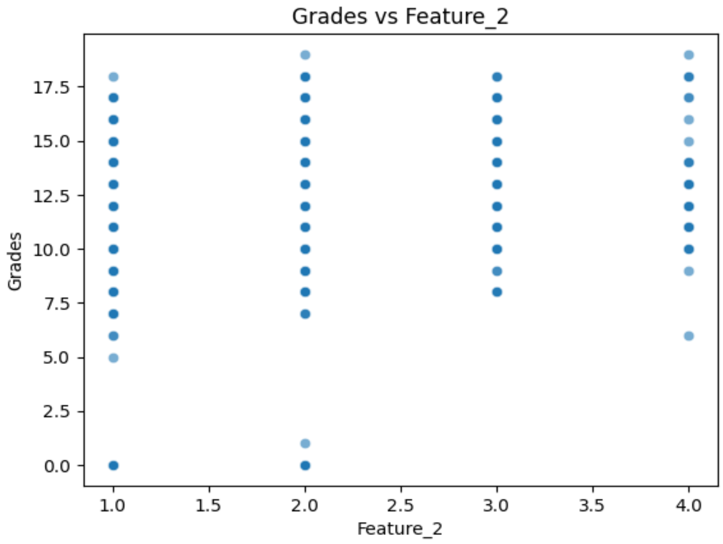
- Feature_2 shows a positive correlation with final grades, and a negative correlation with failures, Dalc, and absences.
- The mean final grade increases steadily with higher values of Feature_2. However there is a slight dip between levels 3 and 4 — possibly due to burnout or mental fatigue from overstudying.

Image: Feature_2 vs Final Grades



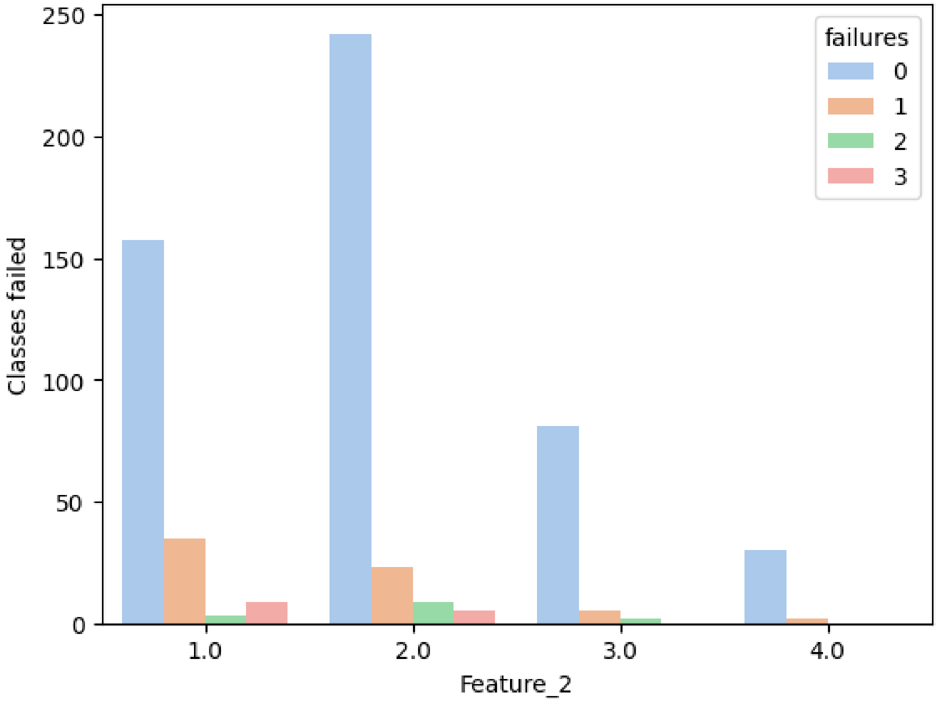
- Some of the highest-performing students fall within the highest range of Feature_2, reinforcing its association with academic success.

Image: Grades vs Feature 2



- Additionally, there is a clear and significant drop in the number of failed classes as Feature_2 increases, further supporting its role as a positive academic behavior indicator.

Image: Count plot of Classes failed vs Feature_2



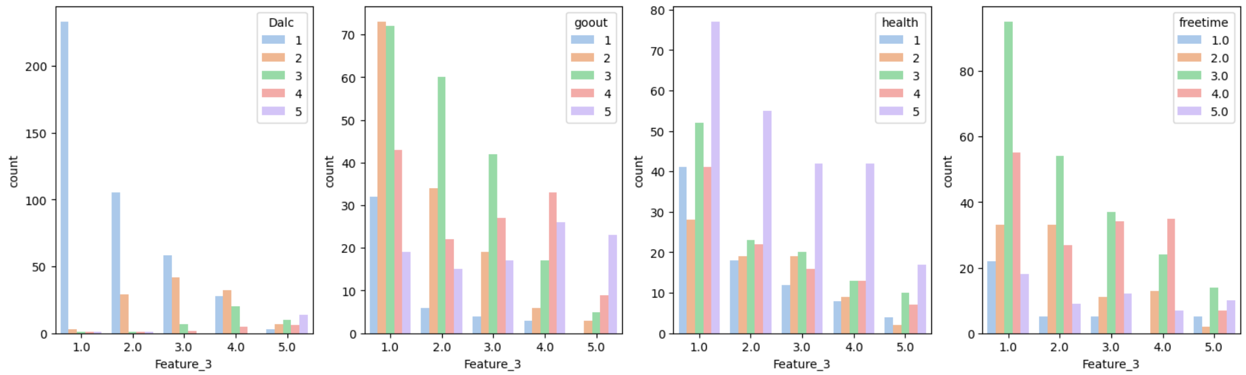
Inference:
This pattern suggests that Feature_2 could represent study hours after school(range 1-4).

Feature 3

Observations:

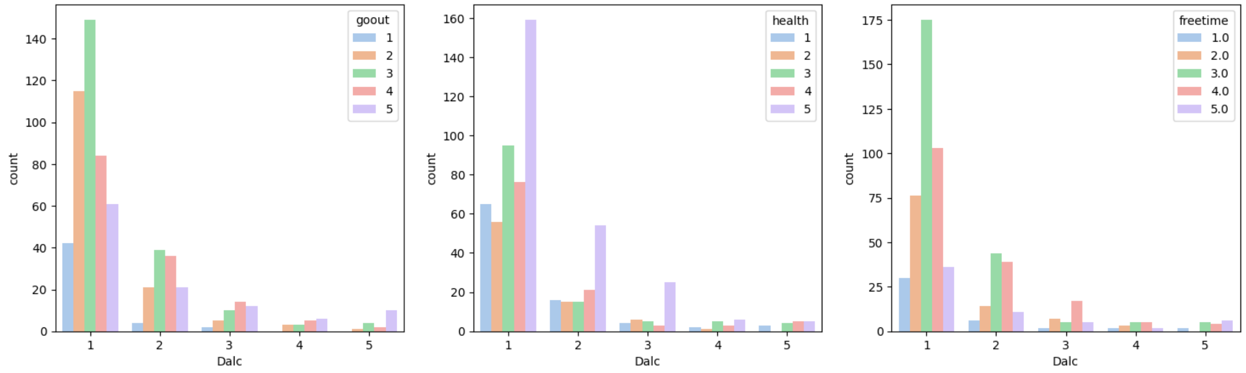
- Feature_3 is strongly correlated with 'goout' and 'Dalc'.
- As Feature_3 increases, both Dalc and goout consistently rise, suggesting that students with higher Feature_3 values tend to engage more in substance use and socializing.
- Additionally, higher Feature_3 values are associated with lower counts of students reporting better health scores(5), indicating a potential negative impact on well-being.

Image: Feature_3 vs Dalc,goout,health,freetime



- Feature-3 and dalc exhibit a consistent pattern across various features.

Image: Dalc vs goout,health,freetime



Inference:

Feature_3 could represent weekday cigarrette consumption (range 1 to 5).

Level 2: Data Integrity Audit

- Features containing null values were identified.

- Next, features were categorized into categorical and numerical types

Imputation Strategy:

- For binary categorical features such as higher and famsize, mode imputation was used, as it retains the dominant class and avoids introducing bias or artificial variation through random imputation.
- For numerical features, mean imputation is more suitable when the distribution is approximately normal (Gaussian), while median imputation is preferred for skewed distributions (either right or left-skewed).
- For features like **Feature_1**, **Feature_2**, **Feature_3**, **freetime**, **absences**, **Fedu**, and **traveltime**, median imputation was applied. Most of these features are ordinal or categorical in nature, making the use of mean inappropriate. Median imputation better preserves the ordinal structure of such variables.
- For **G2**, which followed a Gaussian-like distribution, mean imputation was used as it best preserved the overall shape of the distribution

Image: G2 histogram before imputation

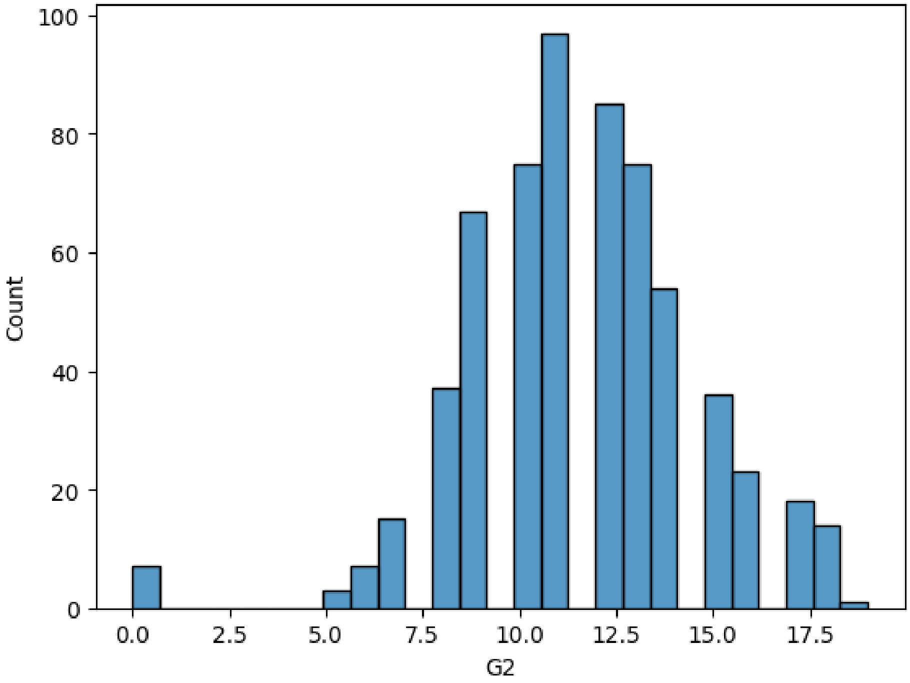
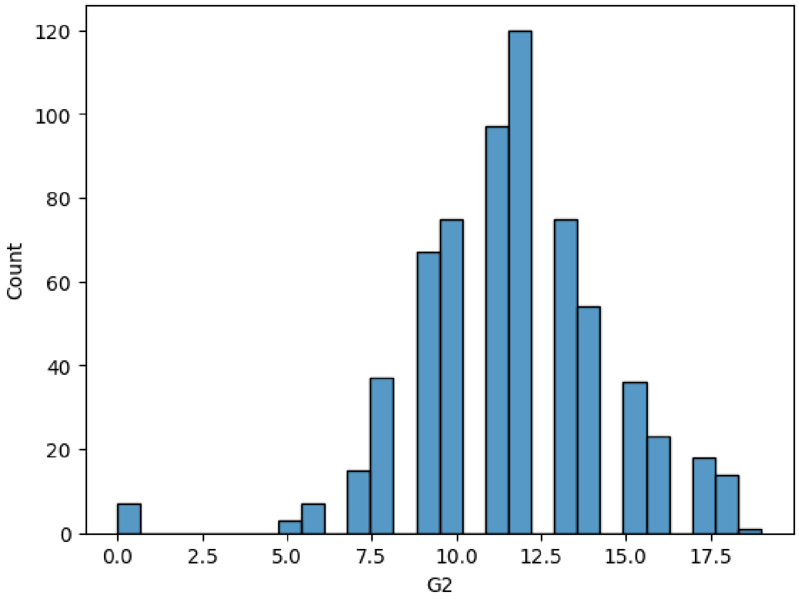


Image: G2 histogram after imputation

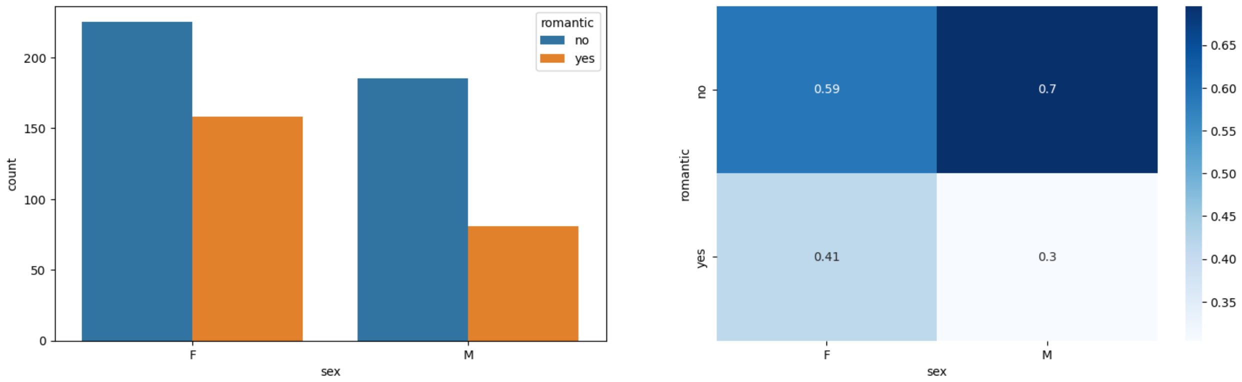


Level 3: Exploratory Data Analysis on Social & Academic Factors

An exploratory data analysis was conducted to examine how social factors (such as gender, relationships, and family) and academic factors (such as study time and tuition) influence outcomes like grades and behavior.

Question 1: How does gender affect romantic relationship status?

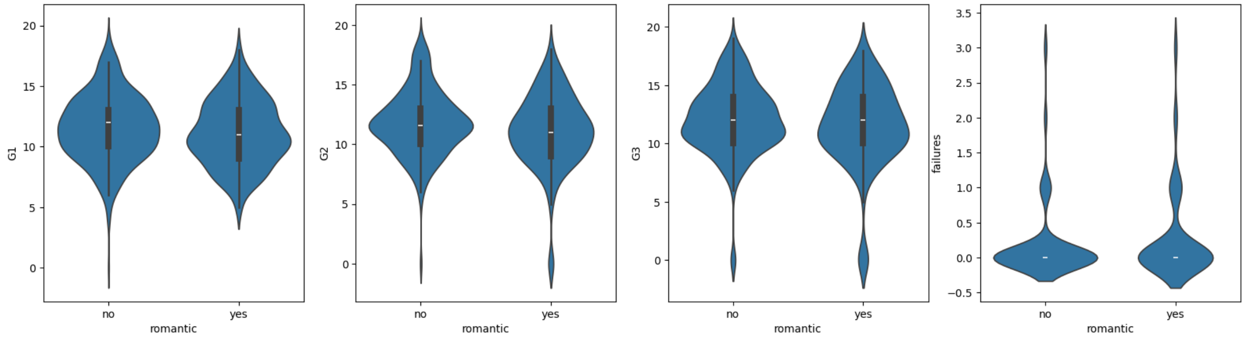
Image: Countplot and heatmap of sex vs relationship status



- Female students are more likely to be in a romantic relationship than male students.
- Male and Female students have a higher proportion of not being in a relationship than being in one.

Question 2: Does relationship status affect grades?

Image: Violinplots of academic performance vs relationship status

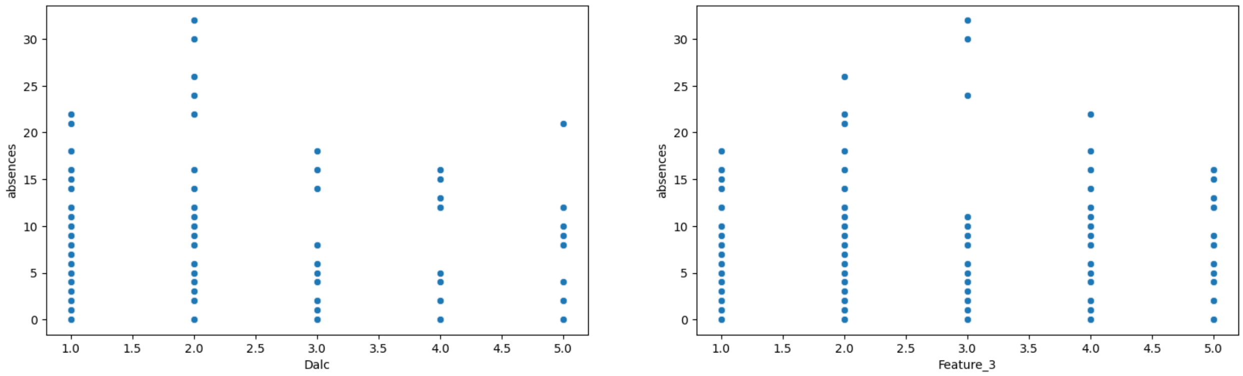


Based on the violin plots for all three grades and number of failures, romantic relationship status does not significantly affect academic performance or failure rates. The grade distributions and outlier behaviors are almost identical across both groups.

Romantic relationships may not directly interfere with or enhance academic focus for most students. Instead, academic performance is likely influenced more by study habits, motivation, support systems, and time management. Additionally, students capable of balancing relationships and responsibilities may maintain similar grades to their single peers.

Question 3: How do absences correlate with Dalc and Feature_3?

Image: Scatterplot of absences vs substance use

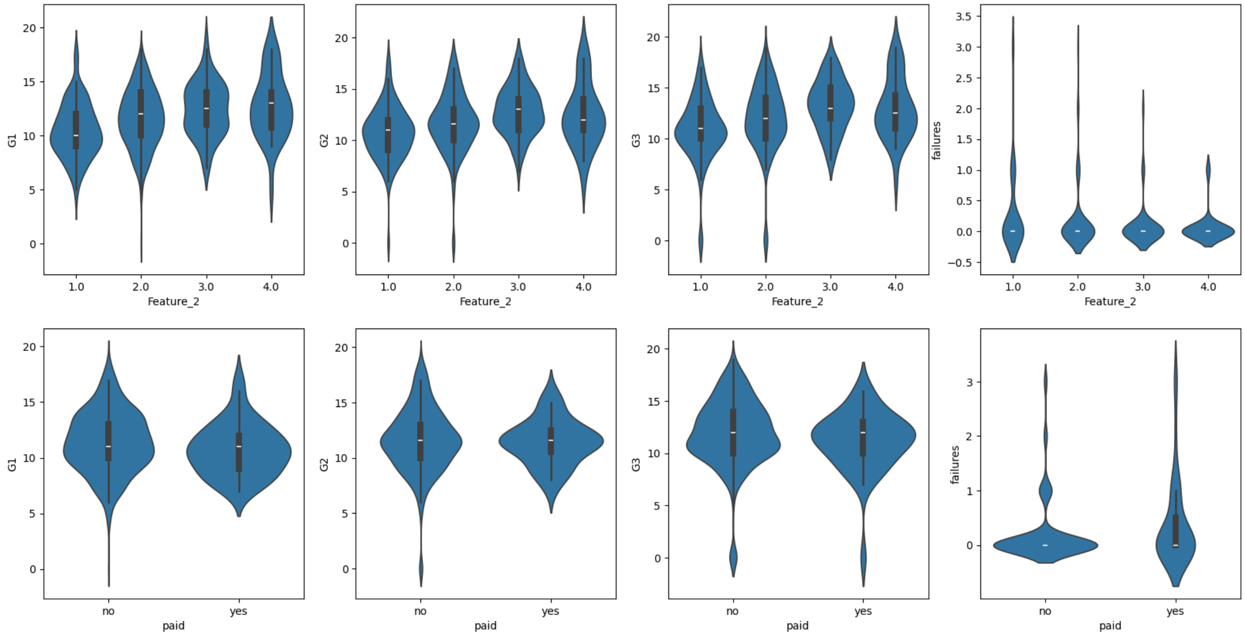


Students with moderate alcohol consumption (Dalc = 2 or 3) exhibit a wider spread in absences, including some of the highest recorded values. A similar trend is observed among students with moderate cigarette usage, indicating that regular (but not extreme) substance users may be more likely to skip school or face lifestyle-related disruptions.

Interestingly, students with very high levels of alcohol (Dalc = 4 or 5) or cigarette use (Feature_3 = 5) do not consistently show high absenteeism. This might be attributed to fewer data points in these extreme categories or unique behavioral patterns among these students.

Question 4: Do study time and tuition affect grades?

Image: Violinplots of Academic performance vs study_hours and tutions



1. Study time:

Across all graphs, higher levels of studytime are associated with a slight increase in median grades.

Additionally, as studytime increases, the number of outliers decreases, indicating more consistent academic performance among students who dedicate more time to studying.

Notably, failures are more concentrated near zero in the higher studytime categories, with fewer outliers, suggesting a clear link between increased study effort and reduced course failures.

Given that these are school students, the academic rigor may not be extremely high, and therefore, increased study hours may not always translate into significantly higher grades. However, it is evident that students who invest more time in studying are considerably less likely to fail, highlighting the value of consistent academic effort even if it doesn't drastically impact top-end performance. The highest grades however are acheived by those who study the most hours.

2. Paid:

Paid tutoring doesn’t change the median grades or overall spread, but it does cut down on very low scores and failures. In other words, tuition raises the performance floor rather than the ceiling—helping at-risk.

Question 5: How does family relationship affect romantic status?

Image: Countplot of relationship status vs family relationship

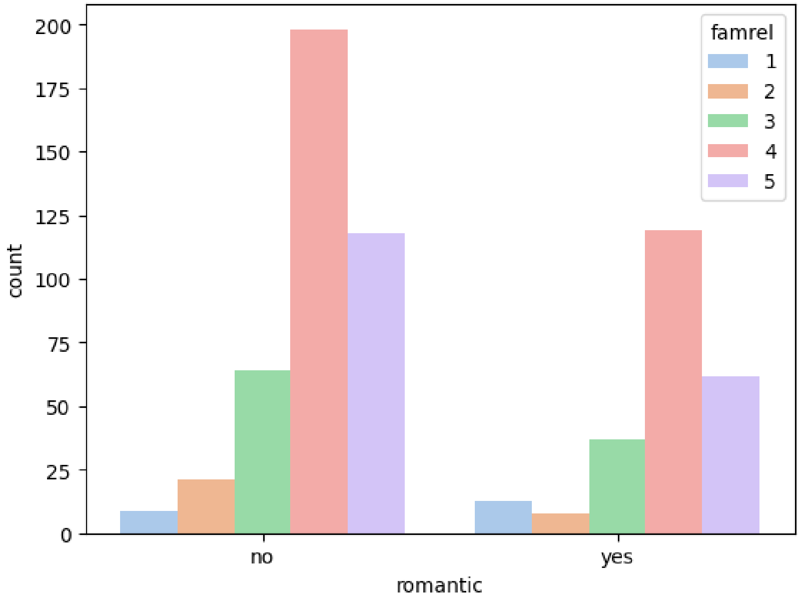
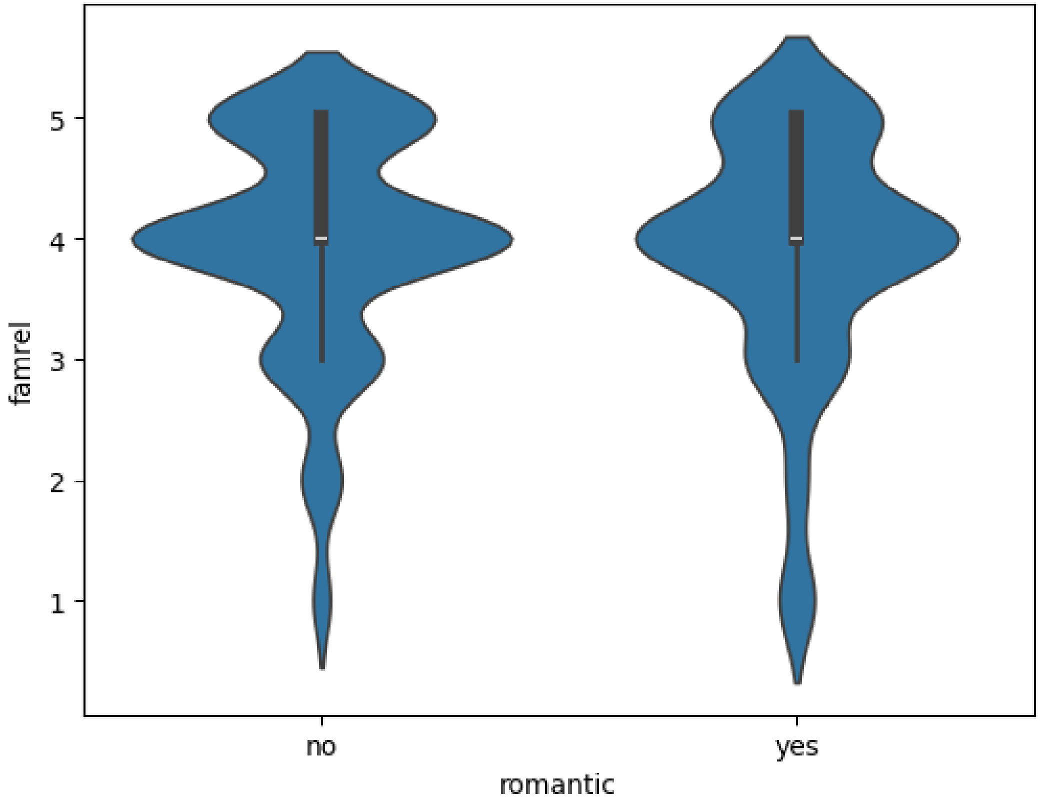


Image: Violinplot of relationship status vs family relationship



Famrel shows only a slight difference between students who are and aren’t in a romantic relationship. Both groups share the same median rating and similar interquartile ranges, indicating largely overlapping experiences .The violin plot reveals that students in relationships exhibit a marginally heavier lower tail—more instances of weaker family ties—whereas very low famrel scores are less common among singles .

Level 4: Relationship Prediction Model

Preprocessing Steps:

Initially, the data preprocessing involved dividing features into ordinal and nominal categories, followed by applying **OneHotEncoder** to nominal features and **LabelEncoder** to ordinal features. However, the model predictions were not very accurate. Subsequently**, LabelEncoder** was applied to all columns, which improved the accuracy across all models.Using **LabelEncoder** for all columns keeps the feature space compact and may better suit certain models (such as tree-based models), enabling them to learn patterns more effectively and enhance accuracy.

OneHotEncoder was also tested on all columns, but this did not yield a satisfactory accuracy score.

Next, standardization and normalization techniques were applied using **MinMaxScaler, PCA,RobustScaler** and **Normalizer**. Among these, **Normalizer** produced the best accuracy results for most models. The Normalizer scales each sample to unit norm, benefiting models that rely on distance or direction (such as KNN and Logistic Regression). It preserves the relative importance of features within each row, thereby improving model stability and accuracy compared to **MinMaxScaler** or **PCA**.

Finally, the data was split into training and test sets.

Models Fitting:

Seven different classification models were evaluated:

1. **Logistic Regression**:
   Initially, the model was trained without parameter tuning, yielding an accuracy of 57%, which was unsatisfactory. Hyperparameter tuning was then performed using GridSearchCV to find the best parameters, improving accuracy to 61%. The limited performance may be attributed to possible underfitting, weak feature-target correlation (the highest correlation being just 0.177), and potential class imbalance. Additionally, Logistic Regression assumes linearity, which might not fully capture the complexity of the data.

2. **Decision Trees and Random Forests**:
   Decision Tree and Random Forest classifiers were applied along with hyperparameter tuning.

   Decision Trees achieved an accuracy of 54%.

   Random Forests achieved an accuracy of 56%.
   These relatively low scores could be due to underfitting or weak relationships between the features and the target variable (relationship status).

3. **K-Nearest Neighbors (KNN)**:
   KNN yielded an accuracy of 59%. Its lower accuracy may be caused by sensitivity to irrelevant or noisy features, as KNN treats all features equally when calculating distances. It also struggles with high-dimensional data.

4. **Naive Bayes**:
   Various Naive Bayes classifiers were tested, achieving the highest accuracy scores among all models:

- Gaussian Naive Bayes: 69%
- Multinomial Naive Bayes: 66%
- Bernoulli Naive Bayes: 73%

Naive Bayes models performed well due to their assumption of feature independence, particularly suitable for binary or categorical data. Bernoulli Naive Bayes fits binary features such as relationship status naturally. These models are robust to irrelevant features, effective with small or imbalanced datasets, and are simple and fast to train.

Further improvements in accuracy could be achieved through hyperparameter tuning and enhanced feature engineering.

Level 5: Model Reasoning & Interpretation

Decision Boundaries

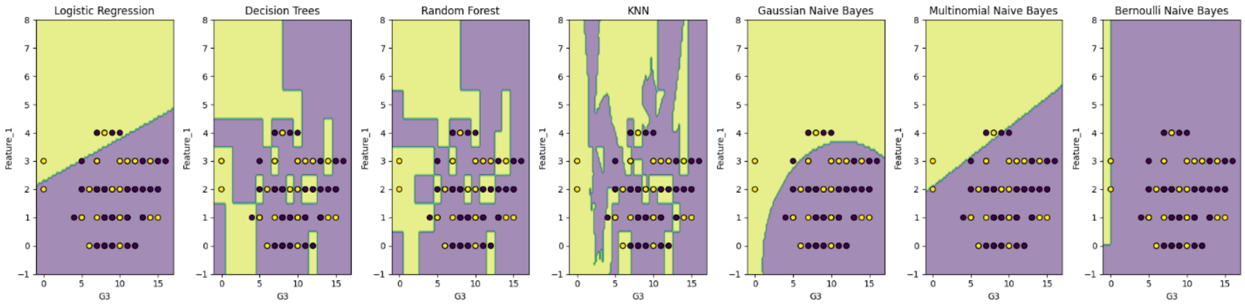
First, decision boundary plots were created using **G3** and **Feature1** as features, since these had some of the highest correlations with romantic status.

As expected:

- Logistic Regression produced a straight decision boundary.
- Decision Trees generated piecewise constant boundaries aligned with feature axes, resulting in rectangular, step-like regions.
- Random Forests produced more complex, nonlinear but still axis-aligned boundaries.
- K-Nearest Neighbors (KNN) produced highly irregular, nonlinear boundaries.
- Gaussian Naive Bayes resulted in quadratic or elliptical boundaries.
- Multinomial Naive Bayes produced linear boundaries dependent on feature distributions.

Bernoulli Naive Bayes created piecewise linear boundaries that could become complex depending on feature combinations.

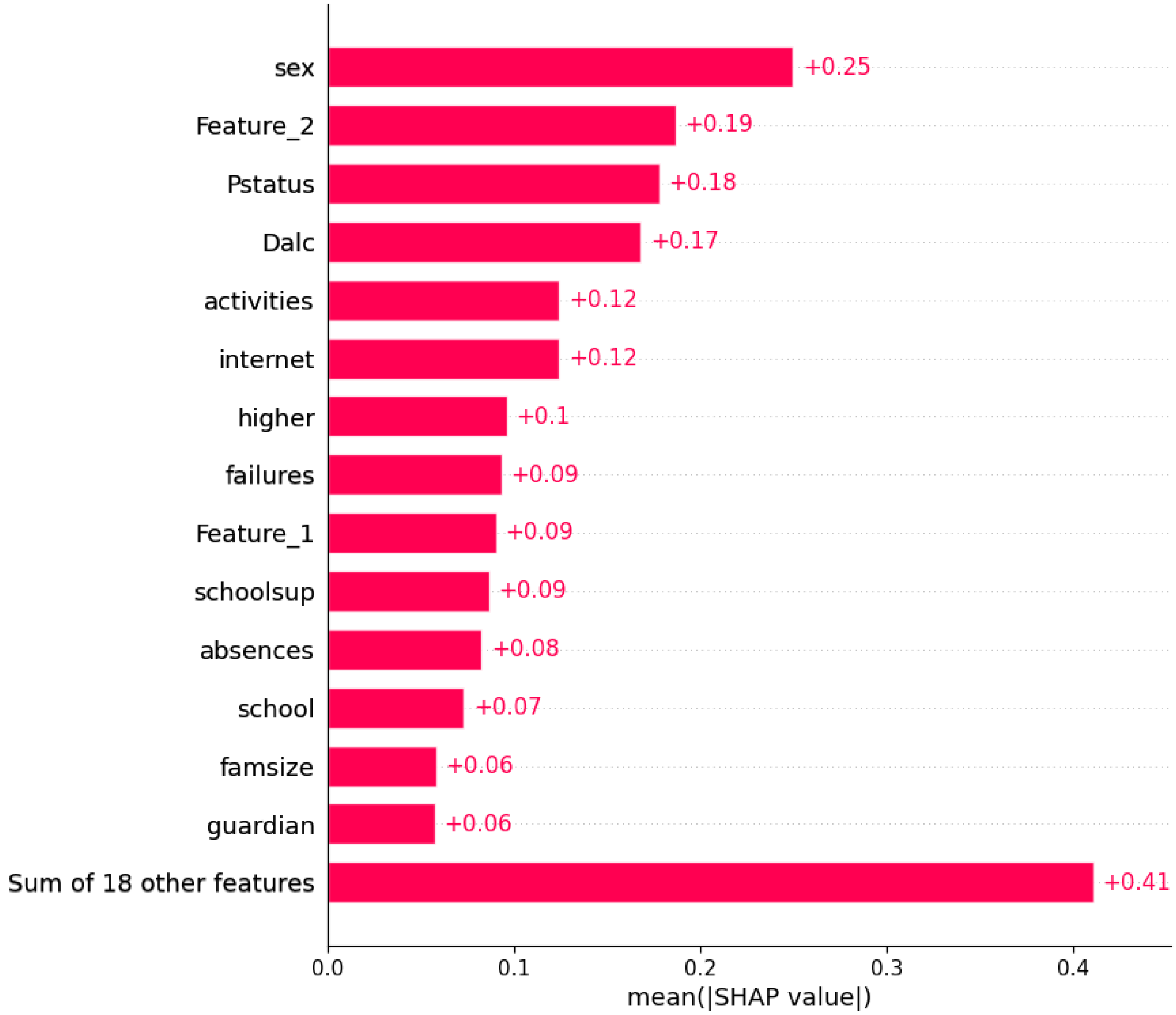
Image: Decision Boundaries of several models



SHAP Interpretations

- Mean SHAP values of different features were plotted to understand each feature's global importance

Image: Mean |Shap Values| in descending order for features



- Additionally, local SHAP values were visualized for two students—one in a relationship and one who was not—to analyze individual feature contributions to the predictions.

<u>Image: Local Shap values for a Student not in a romantic relationship</u>

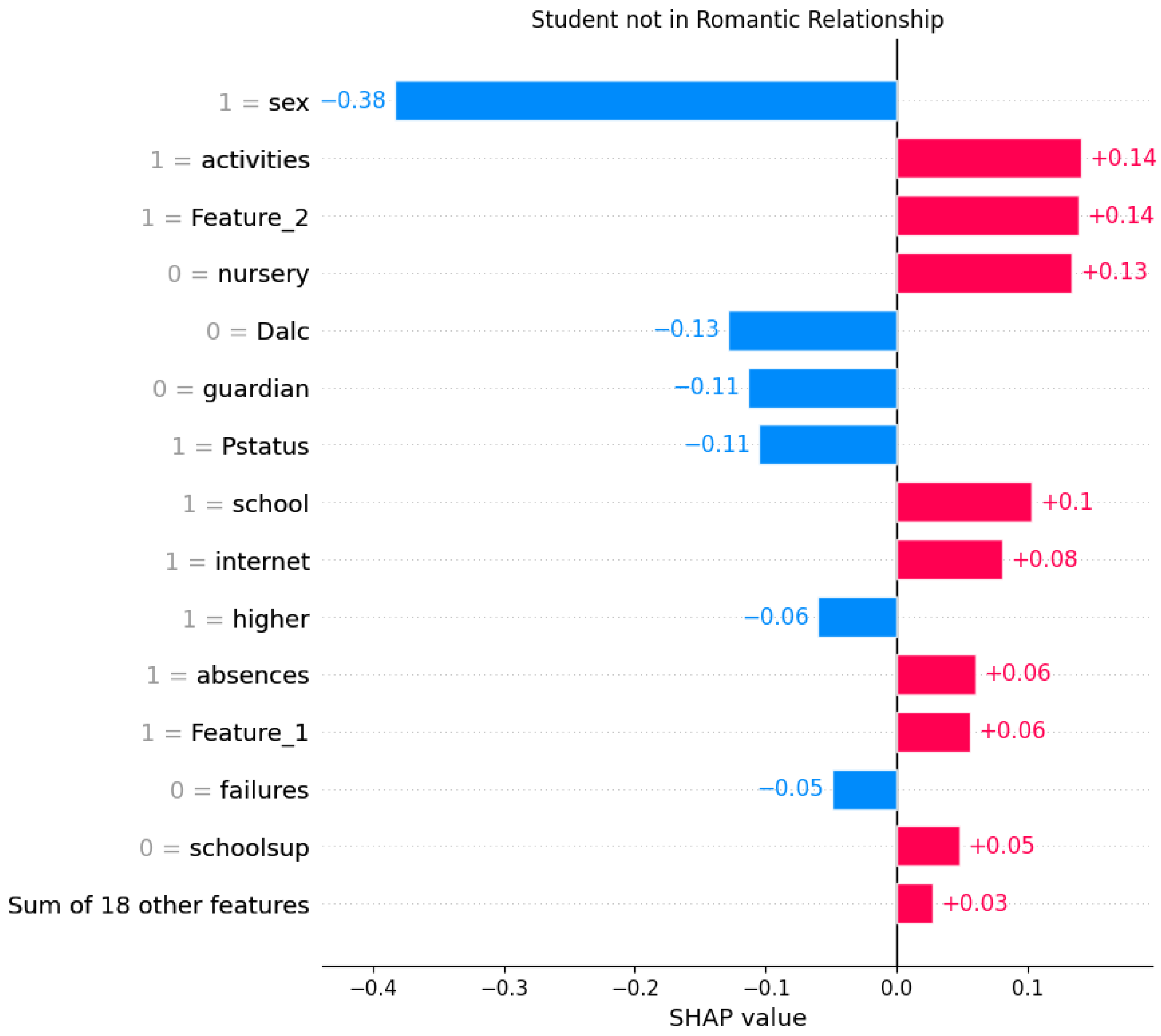
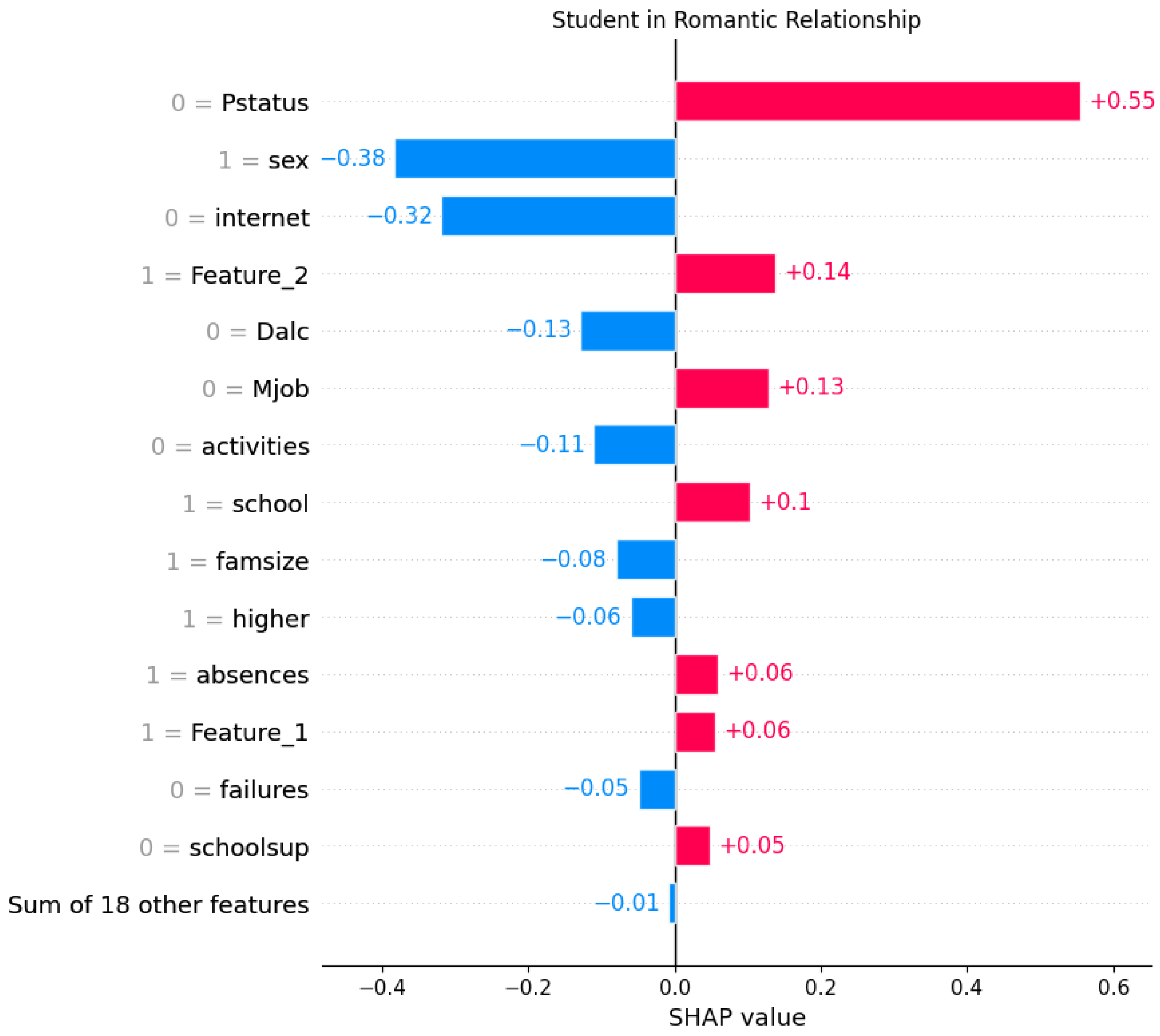


Image: Local Shap values for a Student in a romantic relationship



Key Influencing Features:

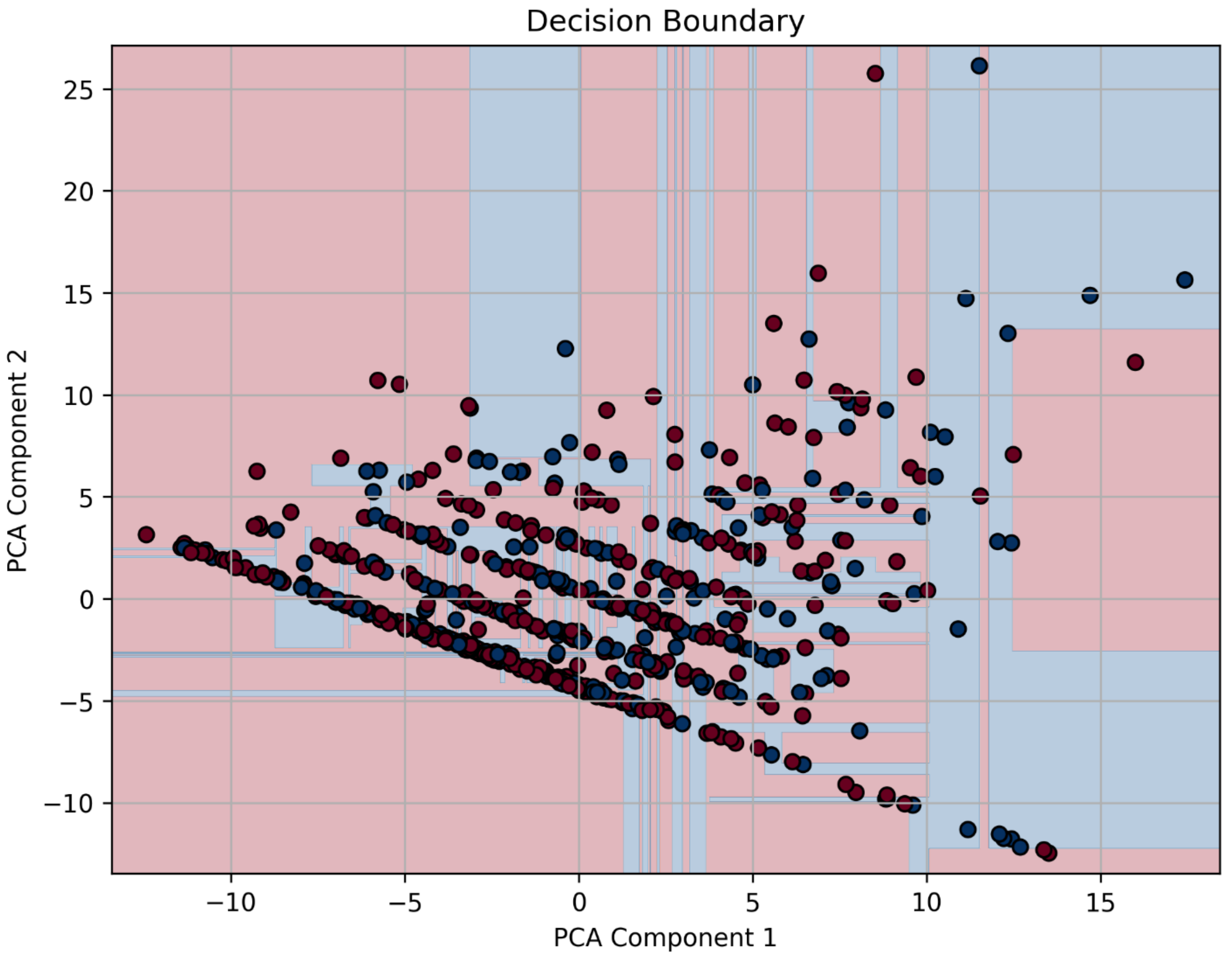
The key drivers of relationship prediction are as follows:

- Feature-2 (Study Hours): Students who study more tend to be slightly more likely to be in relationships, possibly indicating better time management or maturity.
- Sex: Gender plays a significant role, with females being more likely to be in relationships in this dataset.
- Parental Cohabitation (Pstatus): Students from stable home environments show a higher likelihood of being in relationships.
- Social Behavior: Socially active students—those who frequently go out, drink, or participate in activities—are more likely to be in relationships.

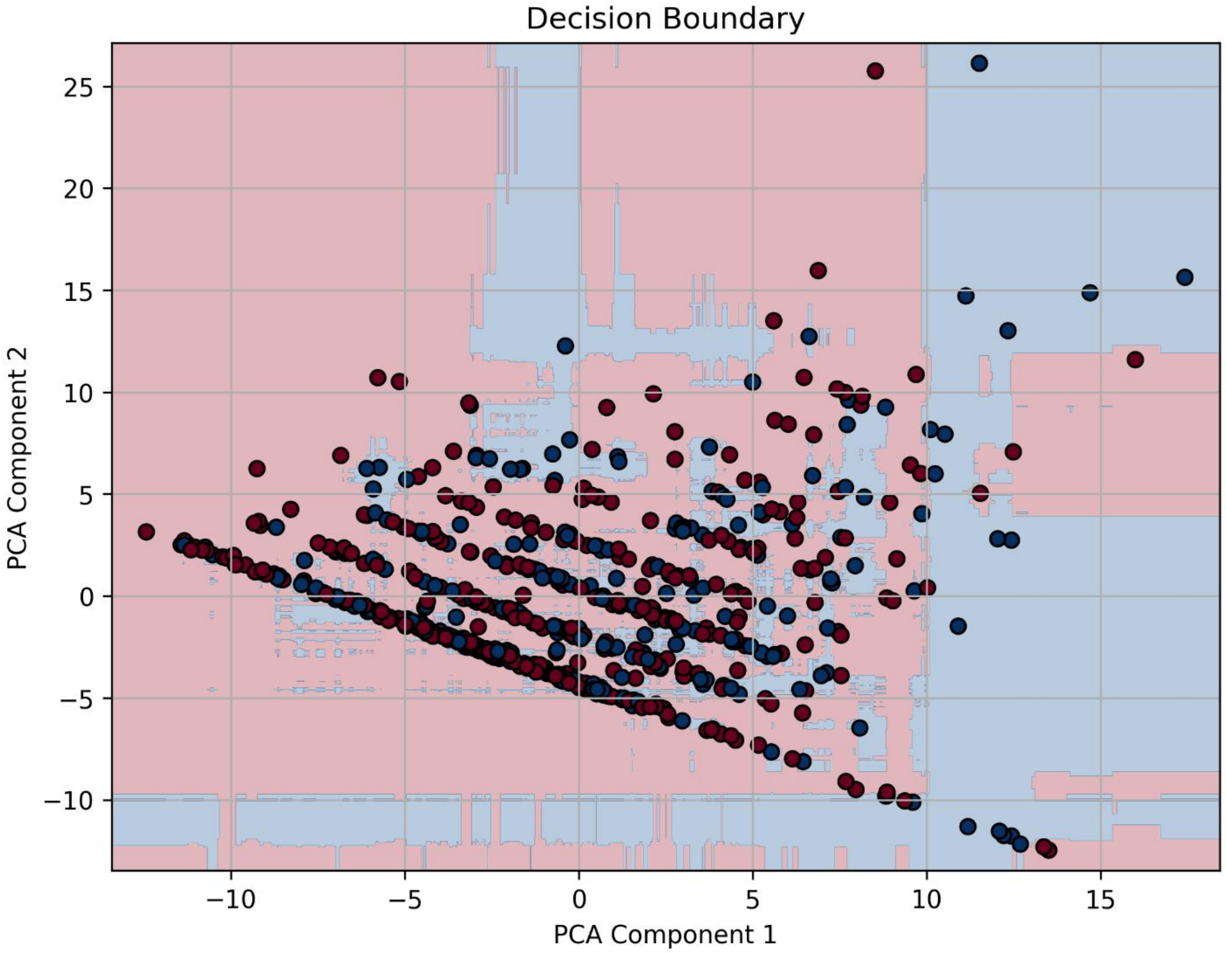
In summary, students who are socially active and come from stable home environments are more likely to be in relationships.

Bonus Task

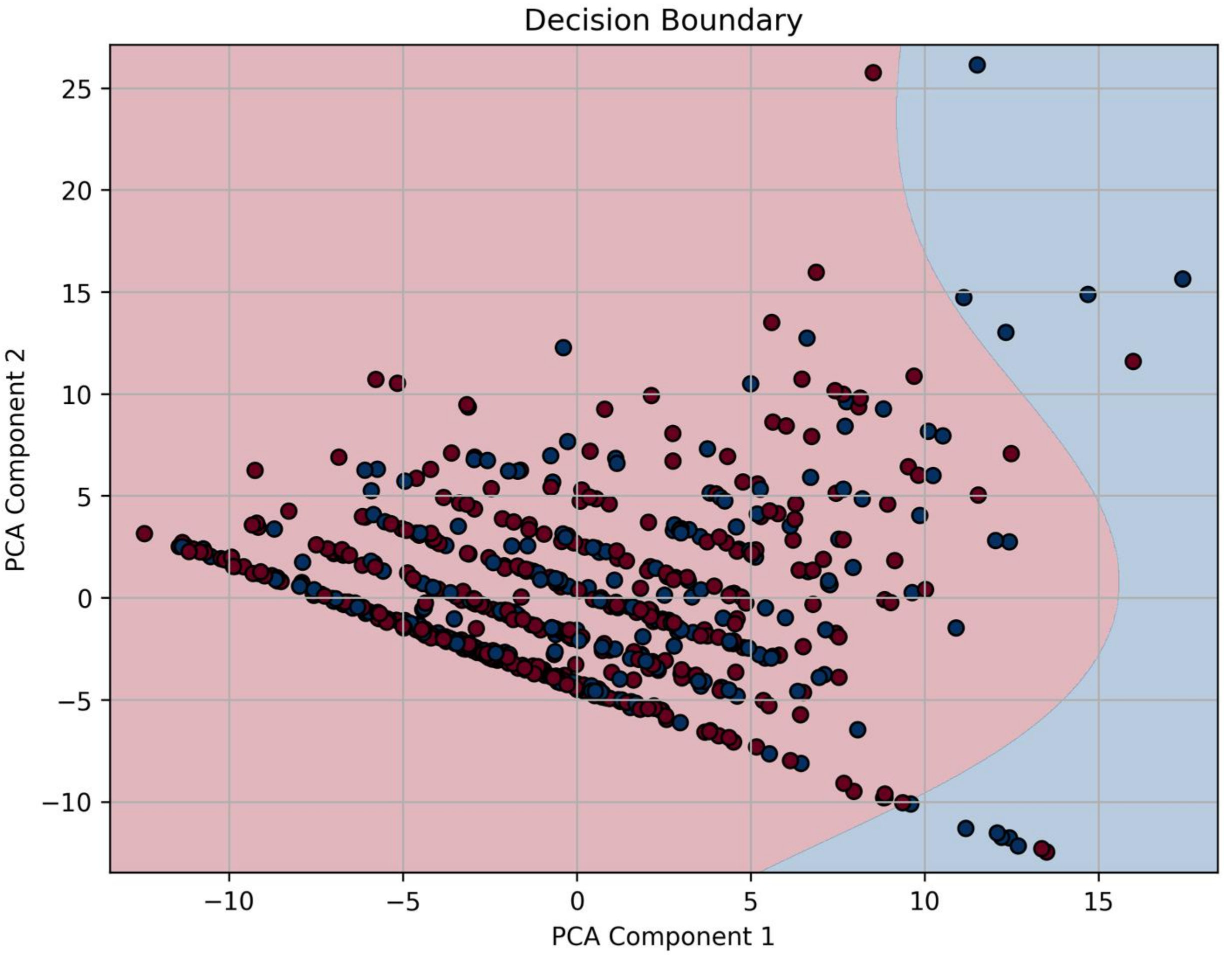
1. **Picture 1: Decision Tree**
   The decision boundary exhibits sharp, rectangular regions aligned with the feature axes. The boundaries are piecewise constant, forming distinct step-like areas that reflect the hierarchical splitting nature of decision trees.



2. **Picture 2: Random Forest**
   The boundary is smoother compared to a single decision tree but still consists of axis-aligned rectangular cuts. The ensemble averaging of multiple trees leads to less abrupt transitions between classes.

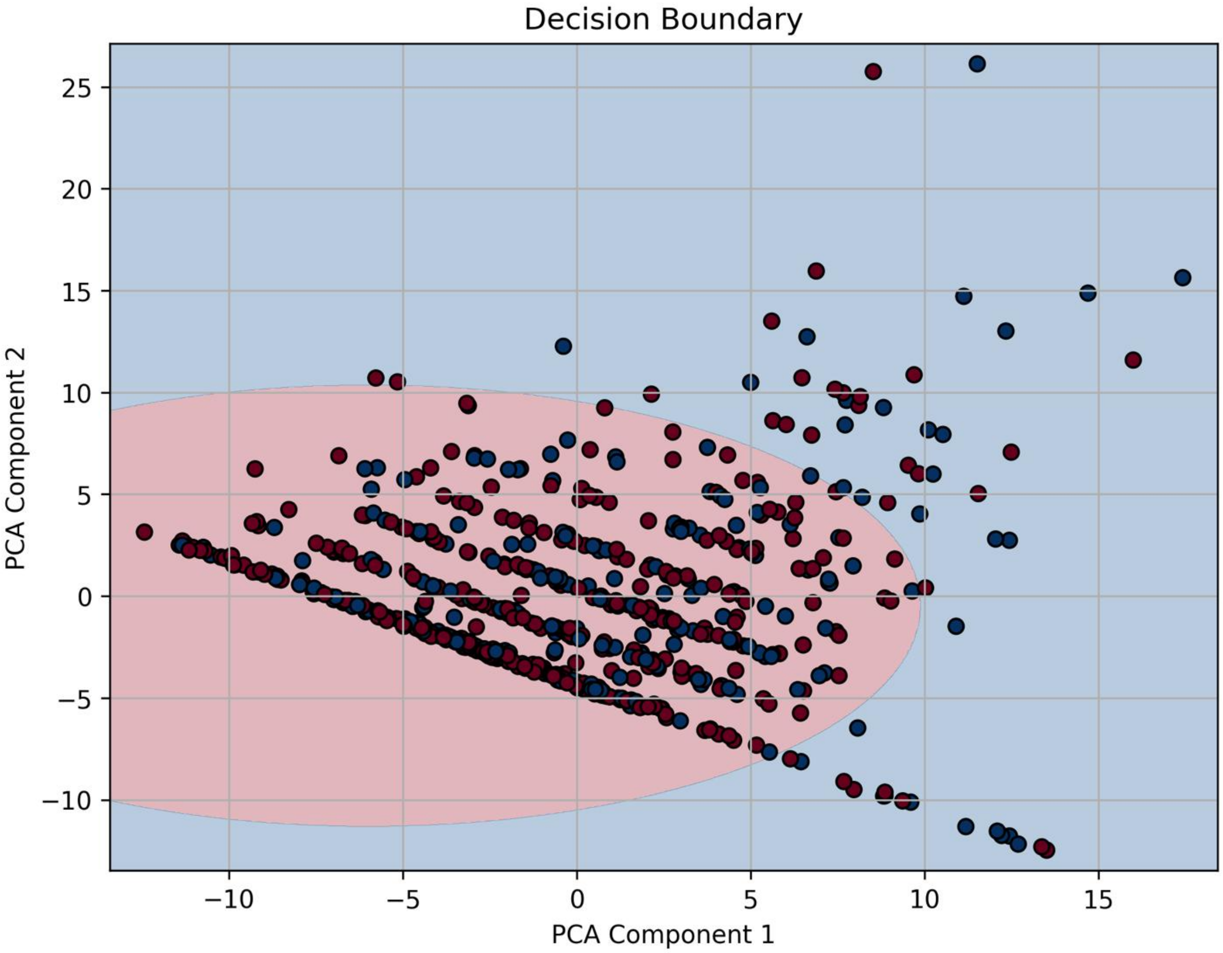


3. **Picture 3: Logistic Regression:**
   Logistic Regression normally gives linear boundaries, but with PolynomialFeatures or kernel methods, it can model curved decision boundaries similar to Gaussian Naive Bayes—especially in PCA space. This can't be Multinomial or Bernoulli Naive Bayes, as those typically produce blocky, axis-aligned boundaries suited for discrete or binary features—not the smooth, curved boundary seen here.



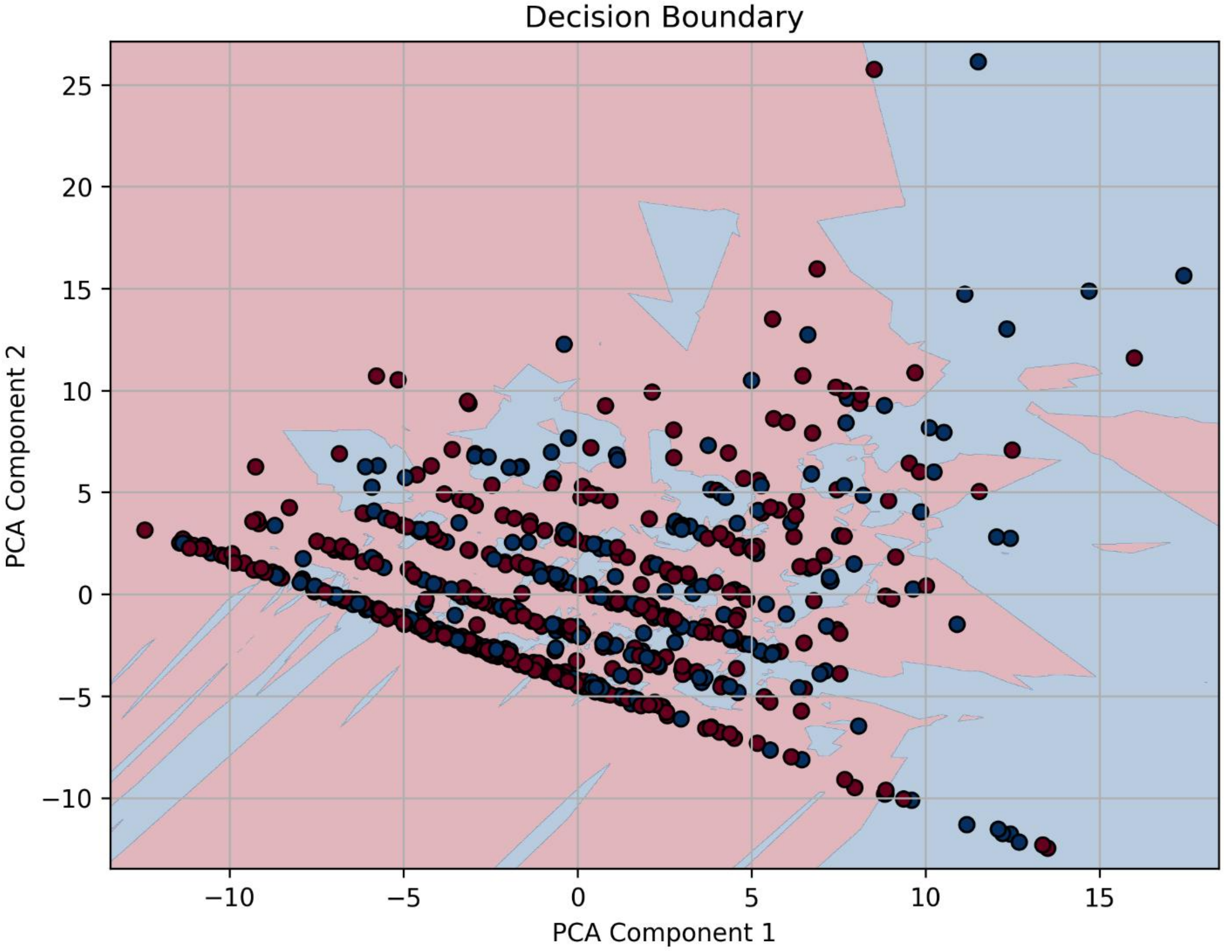
4. **Picture 4: Gaussian NB**:

   The boundaries are elliptical and quadratic in shape, reflecting the assumption of normally distributed features. This allows the model to capture more continuous and smooth class separations.



5. **Picture 5: KNN**:

The decision boundary is highly irregular and nonlinear, closely following the data points. This results in flexible, complex shapes that adapt to local variations in the data distribution.


Decision Boundary
PCA Component 2
PCA Component 1