Unknown Feature Analysis (Level 1)

From this graph it can be observed that

feature 1 is higly realted with failures and other as gaurdian and also realted to the dalc and absences

feature 2 is higly realted with the grades and also realted with review for higher academic studies

feature 3 is highly realted with the Dalc , goout and freetime.

from this graph a realation between the feature 2 and grades can been observed. The value of feature 2 is tend to increase as with increase of academic grades. The median and the maximum both tend to increase here . It can been said that the feature 2 is directly related to some academic stuff.

But the reverse slight negative trend can been observed with the feature 3 . As there is a decrement in the feature 3 with the class grades.

The feature 2 is realted in a constant manner across the atendence . It does not vary much with the attendence . And the value of attendece for the higher feature 2 is lower.

The feature 3 is inversely realted to the attendece . As the feature 3 increases the attendece generaly tend to decrease . From the above and this box plot we can say that the feature 3 is related to the negative side of academics .

Alcohol consuption tend to increase with increase in the feature 1 . generally for the higher value of feature 1 the alcohol consuption is much higher that indicates a kind of jump . This suggests that feature 1 acts as a kind of filter after with there is sudden jump of the alcohol consuption.

Feature 2 has the lesser value of the alochol consuption and is observed to be constatn.

Feature 3 is directly related to the alcohol consuption. More the feature 3 more is the alchol consuption . Again indication it as some negative paramter.

Overall including this violen plot we can observe the followin

The feature 1 is realted to values in a filter method . Its means the values suddenly increase after a certain limit of the the feature 1 . And the higher value of feature 1 have lesser family size and higher go out . Feature 1 seem to represent the age of person

Feature 2 is realted with the academic and good behaviour . Overall it seem to be some kind of academic involvement.

Feature 3 is realted to goout and alcohol consuption a lot .It seems to be related with the social exposure.

Feature 1 : Age

Feature 2 : Academic Involvement

Feature 3 : Social Exposure

Imputation making (Level 2)

Argument for selecting each of the modeling method

Famsize - if famrel<3 then LT3 else GT3 family size

Fedu - keeps the central trend on an orderly scale , strong for outlers

Travel time - catches the typical travel category. Median avoids distortive discomfort groups.

Higher - binary filler with majority course one two. Reduces variance

free time -protects the central ranking. Likert strong for scale

Absence - preserves central tendency

G2 - holds a specific character without removing the outlier

Function 1 (age) - Median is robust to outliers and preserves the central trend

Function 2 (academic exposure) -maintains the center area of the discreet scale

Function 3 (social exposure) - Fills with the majority class to keep the balance of students

Exploratory Insight Report (Level 3)

1. Does parental education influence whether a student is in a romantic relationship?

Student with lower parental education are higher in romantic relationship . This can be observed due to higher level of supervision and other engaging content provided by them to their children. This also corelates to the family values provided to them , that influence their teenage dating behaviour.

2. Do students in a romantic relationship report different family-relationship quality than those not?

Wider sections mean more students report that score and Narrow means few. Both the violen plots are nearly same but we can observe that the violen plot for the yes is more wider at the higher family relationship quality that symbolise ,students who report being in a romantic relationship tend to have slightly lower family-relationship quality scores than those who are not.

3. How do school absences vary by romantic status, and what is their relationship with final grade ? In other words, do romantic students miss more classes, and does that correlate with lower grades?

We can observe that the slope of trend line of non romantic student is higher as compared to the slope for the trend of the Romantic Student. So for the non Romantic student we have a higer final grade and a lesser absence.Hence absences reflects the clash of dating schedule and the class schedule .

4. Is there a difference in the number of absences between students who have internet at home and those who do not?

The spread line for the no is lower as compared to the student having internet access at home . And ther are a lot of outliers student having the access to the internet and have a very large number of absences at school . Hence students without internet at home tend to have fewer absences, whereas those with internet show a wider range of missed days.

5. Do students who participate in extracurricular activities have different final grades than those who do not?

The median line for the yes grade is higher reflecting that the students who participate in the extracurricular activity tend to have more involvement in academics also . Thus students who participate in extracurricular activities tend to have slightly higher final grades and exhibit less variability, suggesting a positive link between balanced involvement and academic success.

6. How does weekday alcohol use relate to the amount of free time?

Students with slightly more free time tend to drink a little more, but the relationship is not strong . Male point cluters nera the higher values for the higher free time than the females clusters suggesting male with more free time tend to drink more . Hence there is a mild positive association between after-school free time and weekday alcohol consumption.

7. How do travel time, internet access, and final grade interact? In other words, do students with longer commutes and no internet at home tend to have lower final grades compared to shorter-commute students who do have internet?

Both shorter travel time and having internet access are associated with higher and more consistent final grades . And those with the higher travel time and no internet acces have lower grades . This refrence the importance of internet in the modern era of education along with the importacne of student's time.

8. How do family educational support, extra paid classes, and first-term grade interact? In other words, among students who do or do not receive family support, does taking paid classes correspond to a higher first-term grade?

Students with famsup = yes outscore those with famsup = no neither group takes paid classes , observed from the median line.Those lacking both are clustered at the bottom .There is a thin strech of students without family support having lower final grades marks. Hence the additional family support along with the extra paid classes lead to the overall improvement of the final grades.

Relationship prediction model (Level 4)

Data Preparation

- Load the fully imputed dataset.
- Convert categorical columns (e.g., sex, address, paid, etc.) into one-hot dummy variables.
- Map the target column (romantic) to a binary flag (romantic_flag = 0/1).

Train/Test Split

- Separate features (X) from the target (y = romantic_flag).
- Stratify an 80/20 split (train_test_split(..., stratify=y)) so that "yes"/"no" proportions remain roughly the same in both sets.

Feature Scaling

- Identify numeric/ordinal columns (e.g., Medu, Fedu, traveltime, famrel, G1, G2, G3, etc.).
- Fit a StandardScaler on the training set's numeric columns.
- Transform both training and test numeric features so they each have mean ≈ 0, std ≈ 1.
- Leave one-hot dummy variables unscaled (they're already 0/1).

Train on different models

- We have trained 3 different models namely logistic regression , random forest classifier and the SVM classifier .

Make Predictions

- Predict class labels on the test set: y_pred = svc.predict(X_test_scaled).
- Predict probabilities: y_proba = svc.predict_proba(X_test_scaled)[:, 1] (used for ROC AUC and curve).

Evaluate Performance

- Compute accuracy, precision, recall, and F1-score (via accuracy_score, precision_score, recall_score, f1_score).
- Compute ROC AUC (roc_auc_score) using y_proba.
- Generate the confusion matrix (confusion_matrix).
- Plot the ROC curve (roc_curve), comparing true positive vs. false positive rates.

Compare with Other Models

- **Assemble** a table of metrics for all trained models (e.g., Logistic Regression, Random Forest, SVM).
- **Focus** on metrics beyond accuracy—particularly precision, recall, F1, and ROC AUC, since classes often are imbalanced.

	Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC AUC**
0	Logistic Regression	0.577	0.443	0.562	0.495	0.615
1	Random Forest	0.615	0.444	0.167	0.242	0.568
2	SVM	0.585	0.450	0.562	0.500	0.611

This is the final model output and base on the following we are going to use the SVM model classifier for our project.