LEVEL 1

Used a histogram plot of feature 1,2,3 to see their distribution of values and then used a correlation heatmap as a multivariate analysis tool to see which factor are more correlating with the feature 1,2,3.
By analysis of that
Feature1 is likely age as its value ranges from 15to22 and also due to its fairly positive correlation with Dalc and failures
Feature2 is likely studytime as its negatively correlated to absences and failures whereas positively related to grades.
Feature3 can be extrovertedness due to its high positive correlation with Dalc and goout as extroverts are likely the ones to go out frequently and enjoy themselves.Also a fairly negative correlation with grades indicates the same.

LEVEL2

MISSING FEATURES	IMPUTATION STRATEGY	JUSTIFICATION
famsize	mode	
fedu	median	To maintain realistic distribution
traveltime	median	Same as above
freetime	median	Same as above
absences	median	Same as above
higher	mode	it captures the prevailing intention among students and is less likely to misrepresent attitudes.
g2	Average of g1 and g3	As the performance in second term would likely be around that of first and third term
Age	median	To avoid the influence of slightly older students
studytime	median	
extrovertedness	median	To ensure neutralness that doesn't distort interpersonal variability.

LEVEL3(EXPLORATORY INSIGHTS)

- How does traveltime affect grades?

Slightly downward trend is observed

- Does extrovertedness relate with absences?

a positive trend is observed

- Do students in romantic relationship show absence in periods?

Box plot shows higher absences in romantically involved people

- Which students have higher alcohol consumption?

Urban students had more low alcohol consumption whereas rural students showed variation

- How do family relationship quality influence romantic involvement?

Romantic students showed lower average family relationship score

Level4

By allowing each decision tree to vote on the final prediction, Random Forest integrates numerous decision trees into a single model. By reducing overfitting, this ensemble approach frequently produces more accurate results.
To enable the model to process numeric labels, we change the "romantic" column from "yes"/"no" to 1/0. Words are converted into numbers that algorithms can comprehend through this straightforward mapping.
A new mapping sheet is used for each column to convert each text feature (such as department or city) into an integer. In order to convert numbers back into the original words when necessary, we store these mappings for later use.
To assess performance on unseen data, the data is split into 80% training and 20% testing sets. it is fine when random_state=42 is fixed because it guarantees the same split each time.
To ensure that every feature has the same scale, we standardize inputs to mean = 0 and variance = 1. Scaling frequently increases accuracy and speeds up model convergence.
Algorithm description: Random Forest aggregates the votes of several trees it has grown on arbitrary subsets of data and features to arrive at the final prediction. Each instance's model label is determined by majority voting across trees.
Metrics for Evaluation:

Precision (TP / (TP + FP)): The percentage of predicted positives that were accurate.

Recall (TP / (TP + FN)): The percentage of real positives that the model detected.

F1-Score (2 × Precision × Recall / (Precision + Recall)): It rewards a balance between precision and recall by taking the harmonic mean of the two.

Random Forests provide robust performance and a lower chance of overfitting, making them a dependable option for classification tasks. Accurate and repeatable results are achieved with appropriate encoding, scaling, and fixed randomness.
ACCURACY=0.615
I also used logistic regression to check if higher accuracy could be achieved

In order for the model to interpret the "romantic" column as a numerical result, it is mapped from "yes"/"no" to 1/0.

By establishing a distinct mapping for every value (for example, Red→0, Blue→1) and applying it to the data, each remaining text column is converted into an integer. In order to undo the process if necessary, we save these mappings.

To assess how well the model generalises to new data, we split the data into 80% for training and 20% for testing. The split is fixed when random_state=42 is set, ensuring that the same rows always show up in both train and test.

To guarantee that all inputs are on the same scale, all numerical features are standardised to have mean = 0 and variance = 1. This standardizes the data

To find the optimal weights a logistic regression model is constructed and fitted to the scaled training data. Weighted sums are transformed into probabilities between 0 and 1 by the model using a sigmoid function.

We calculate accuracy, precision, recall, and F1-score to evaluate the trained model's performance in predicting labels on the test set. To illustrate where the model works and where it fails, a confusion matrix separates true positives and negatives from false positives and negatives.

ACCURACY=0.569(Lower accuracy)

To determine which variables most strongly push predictions towards "romantic = 1," we extract the learnt coefficients (weights) for each feature and sort them. The most significant impact on a romantic prediction is indicated by the highest positive weights. Like age , G1, dalc were the top3 features influencing relationship

LEVEL5

In level5 we plot mean shape values of features influencing romantic column

which showed G2,G1,sex,age affected the most.