CODING CLUB

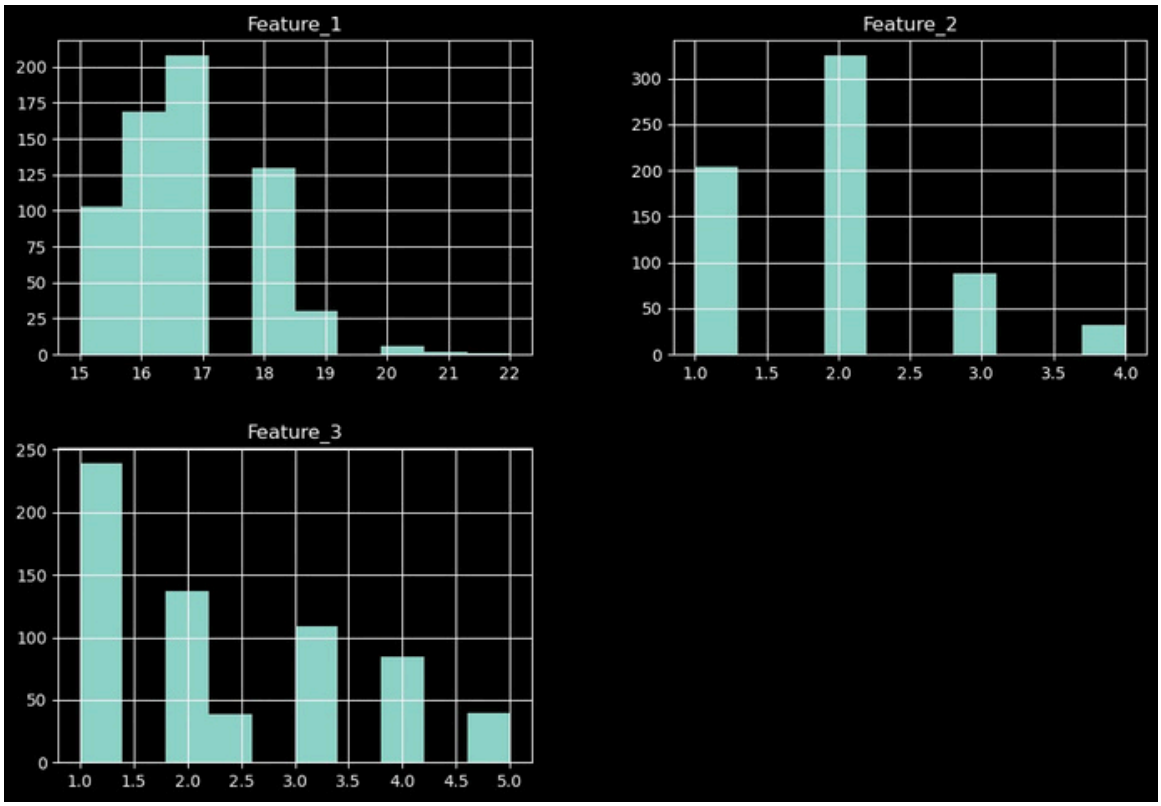
TASK 1 REPORT

PREPARED BY:
RAM TEJ DUVVURI

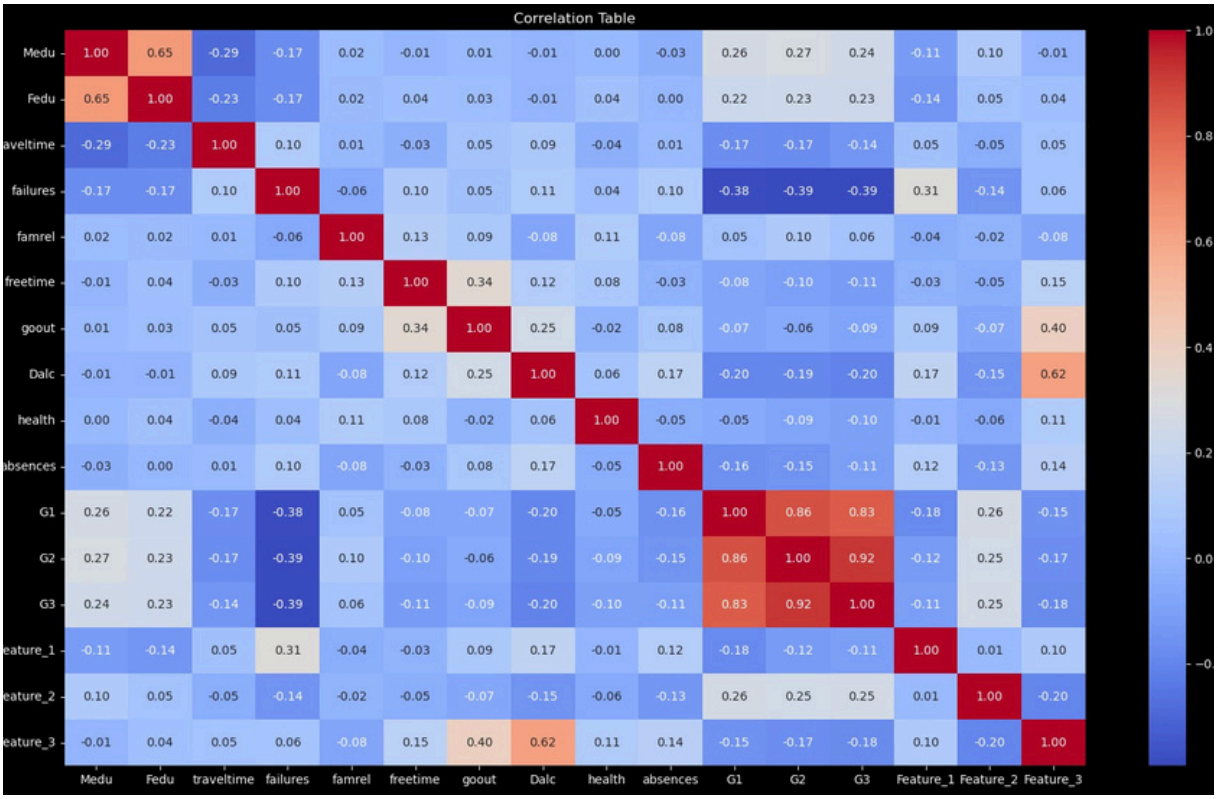
Level 1 (Variable Identification Protocol)

- First I plotted histograms of all three unknown Features:

Graph Between Feature_1 and failure :

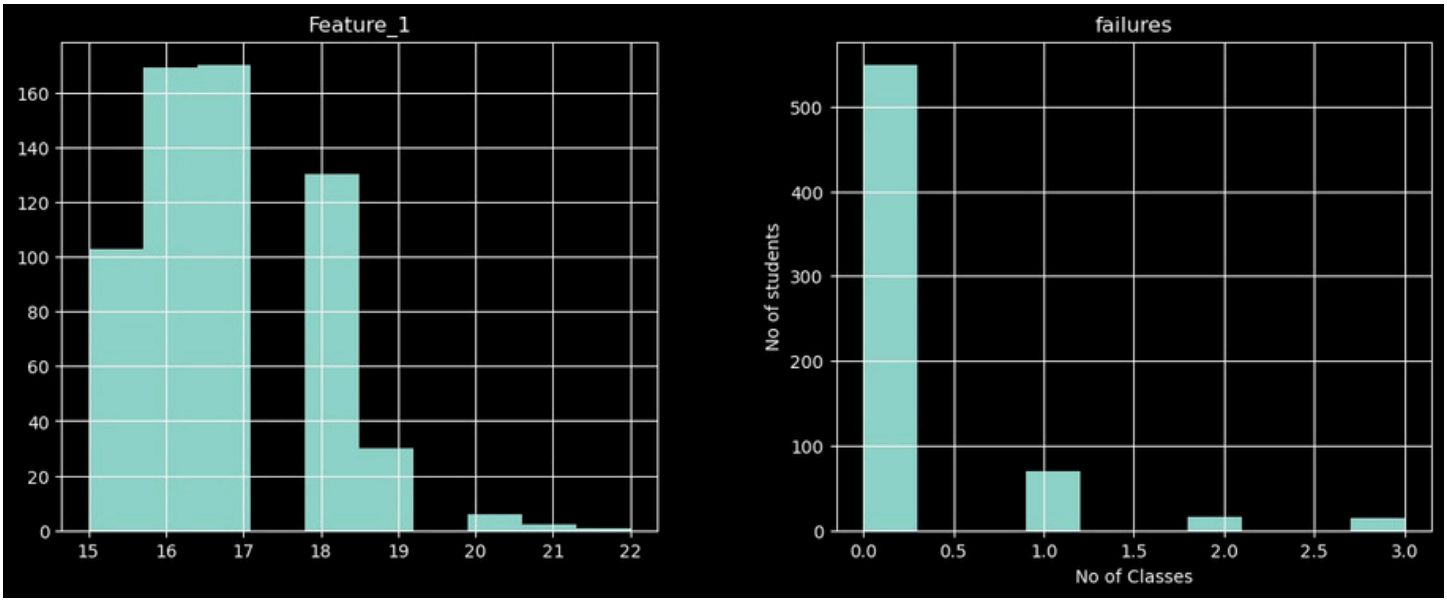


- Next I plotted a correlation table between numeric values of dataset :

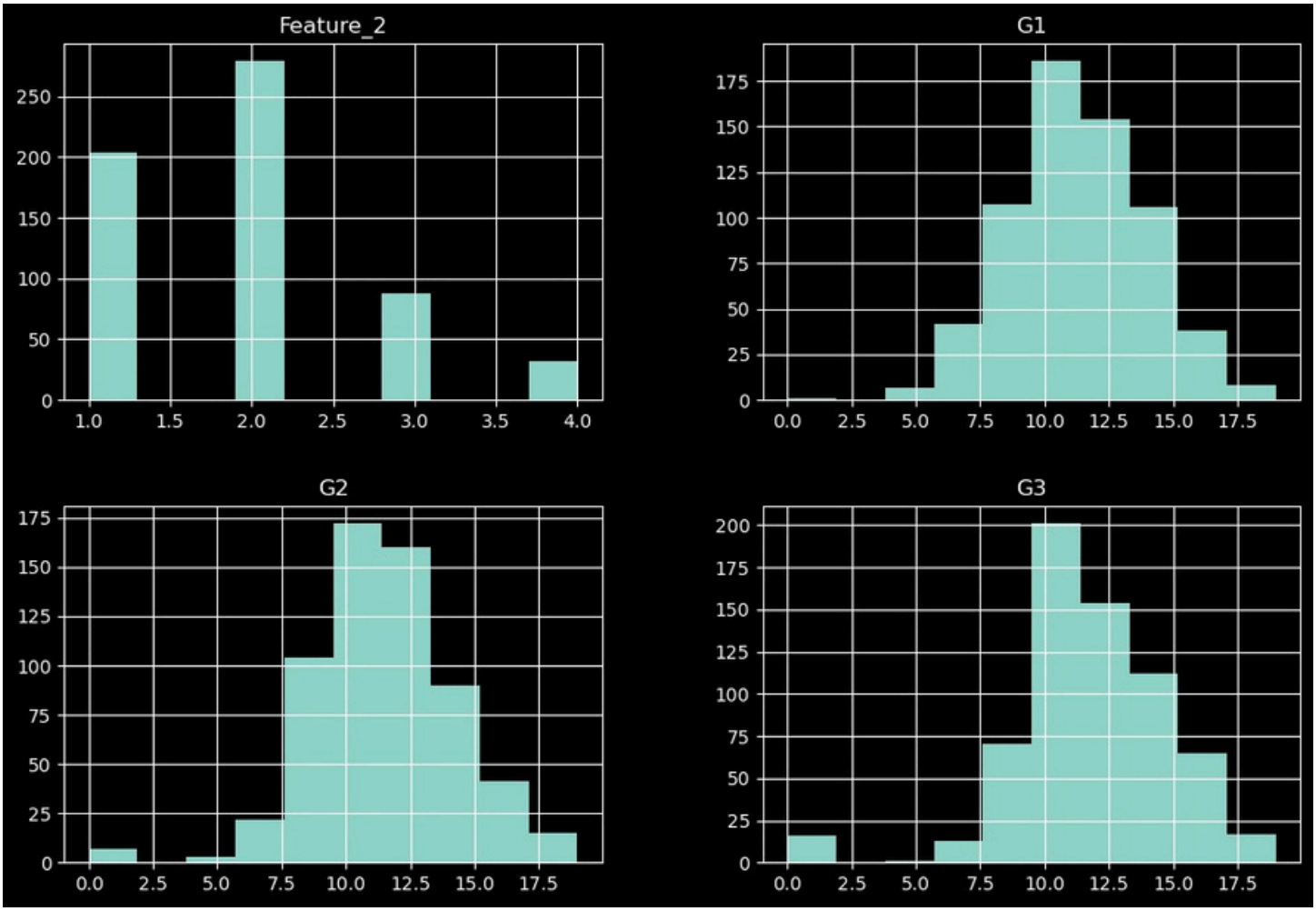


- Then I proceed to again plot histograms of Features but now side with the column that has a high correlation with that features.

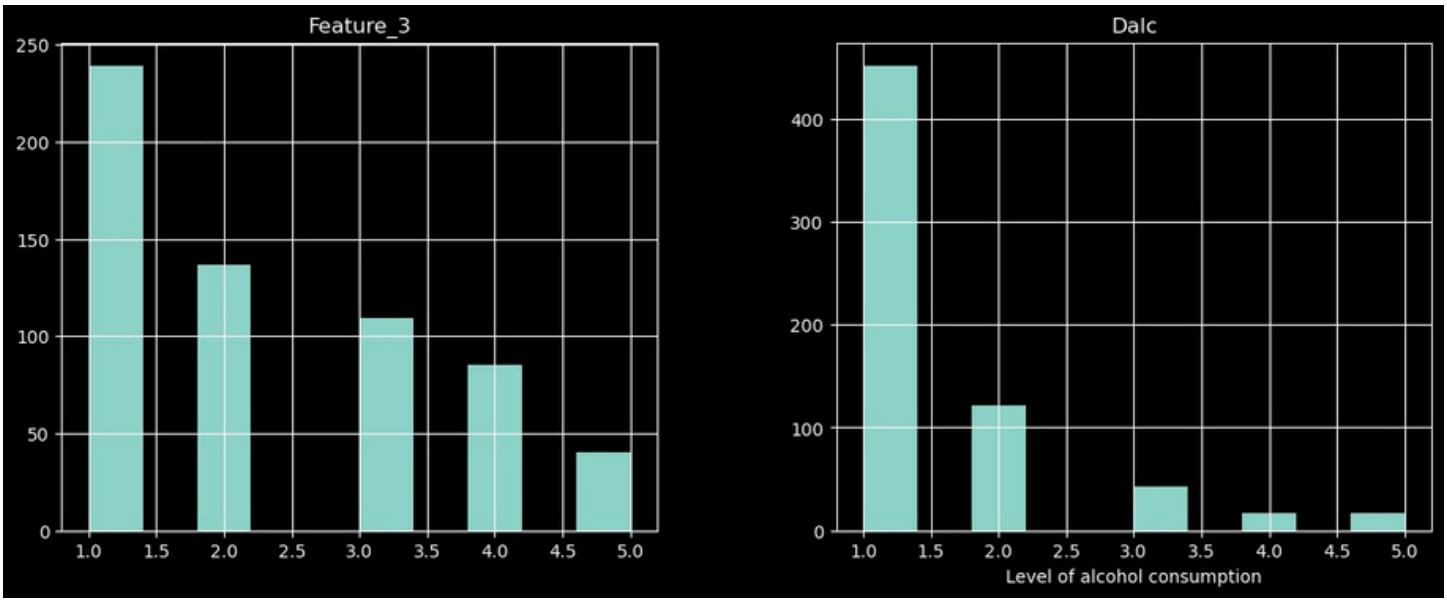
Graph Between Feature_1 and failures.



Graph Between Feature_2 and Grades.



Graph Between Feature_3 and DALC .



Conclusions:

1. Feature_1 is highly correlated to Dalc So it may indicate a **student's choice of age to start drinking** .

2. Feature_2 is highly correlated to grades So it may indicate the **study time** in intervals (for ex: 1 : < 30min, 2:<1 hour and so on) .

3. Feature_3 highly correlated to Dalc So it maybe indicating **stress levels of students**

Level 2: Data Integrity Audit

Now I am Ready to start working on my dataset but before that I have to make sure that dataset is clean for it to be used for the model. That is why I make sure to use all necessary steps used in preprocessing.

Steps I followed while preprocessing the dataset :

- I started of with filling the null values of the numeric based columns and decided to fill them with mean values of the respective columns.
- And for object based columns I decided to fill them with the mode of the respective columns .
- Next, I used a new method I learned during the project called winsorization. I applied this method to the 'absences' column because it had the most outliers compared to other columns. Winsorization works by clipping the outliers to a set threshold.
- Then I applied label Encoding to all object based columns.
- Finally, I performed scaling on the 'grades' and 'absences' columns. I applied MinMaxScaler to the 'grades' column and StandardScaler to 'absences'. I chose not to scale the other columns because their magnitudes were already within the same range.

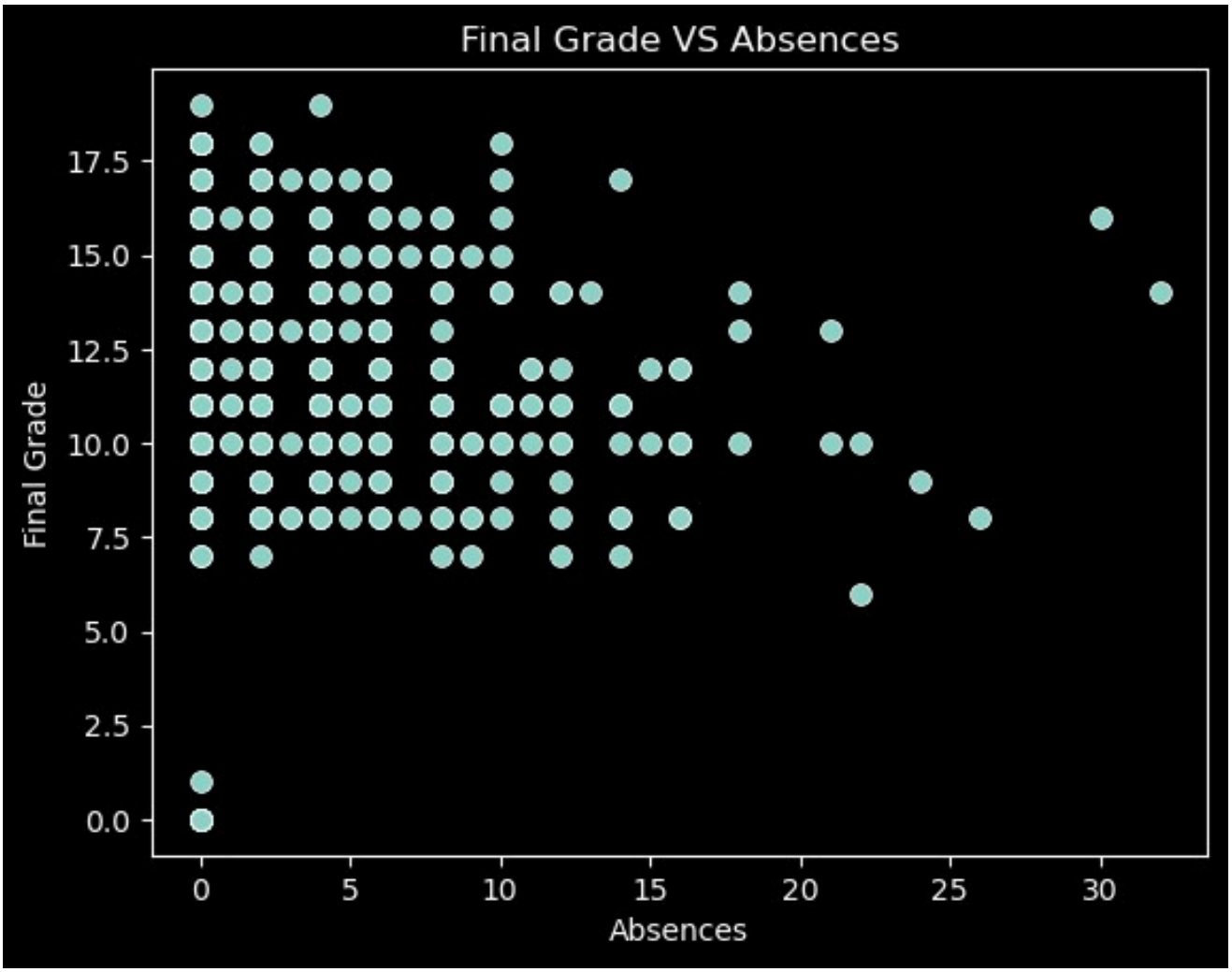
Level -3 (Exploratory Insight Report)

I decided to jump into the model But I thought why not have some look into data and have some insight. So I decided to ponder some questions.

Questions and conclusions :

Question - 1 . How do absences impact final grades(G3) ?

Graph :

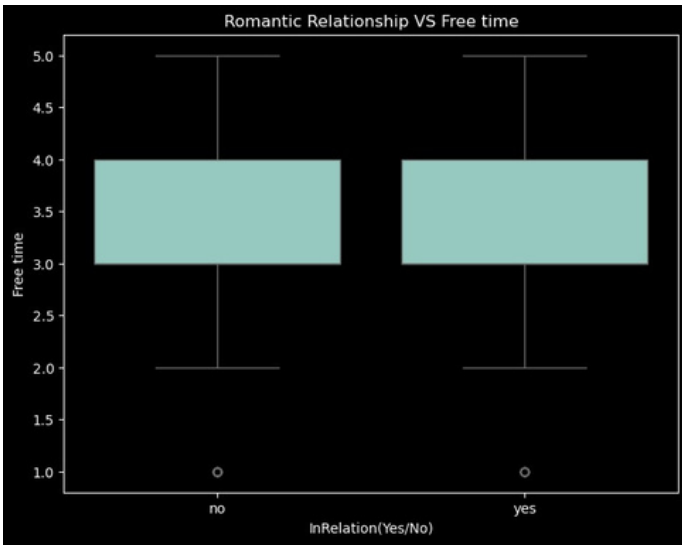


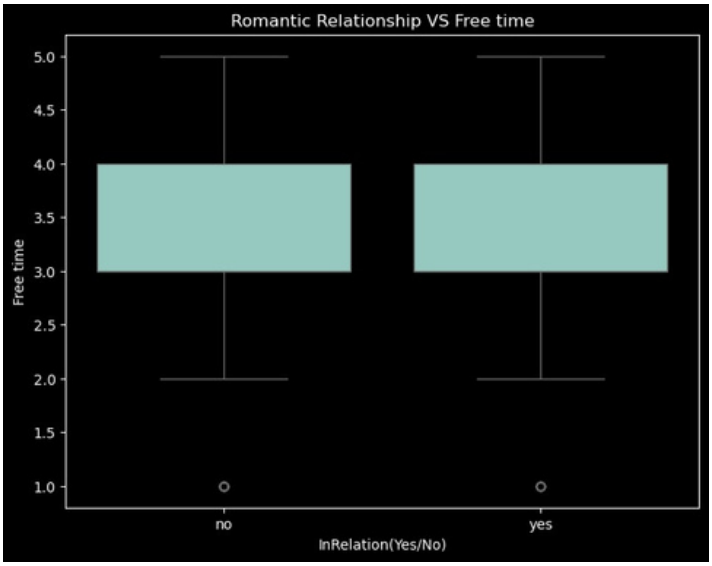
Correlation Values : **-0.106205**

Conclusion : So we can conclude that higher absences will result in less grades .

Question - 2 . Do students in romantic relationships have different study and life patterns ?

Graph :





correlation Values :

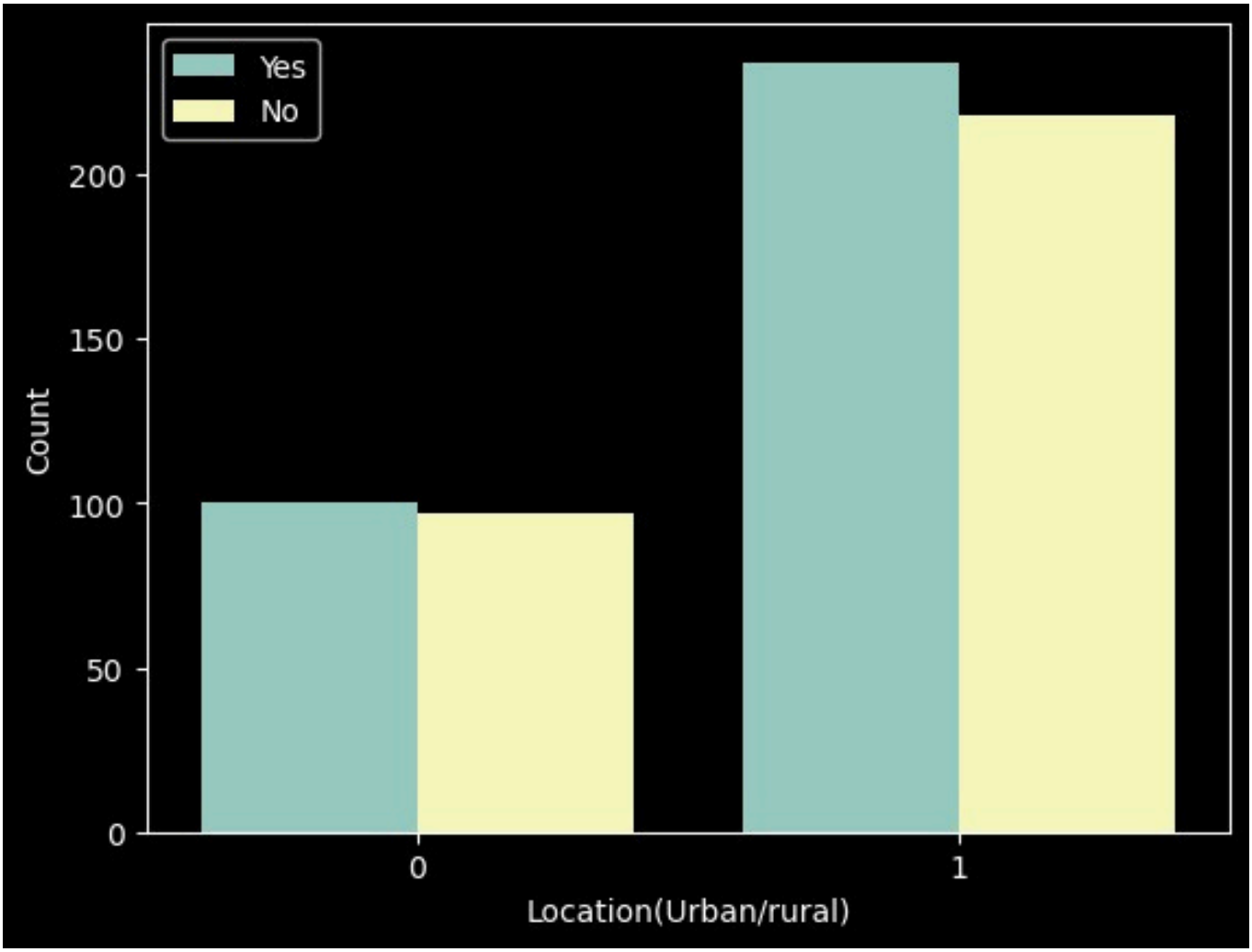
- Between freetime and in relation : **0.01624**
- Between Grades and in relation : **-0.074973**

Conclusion :

So we can conclude that students who are in relationships have more free time but also have less grades.

Question 3: Do students from Urban (U) and rural(R) areas show different levels of involvement in extracurricular activities?

Graph :



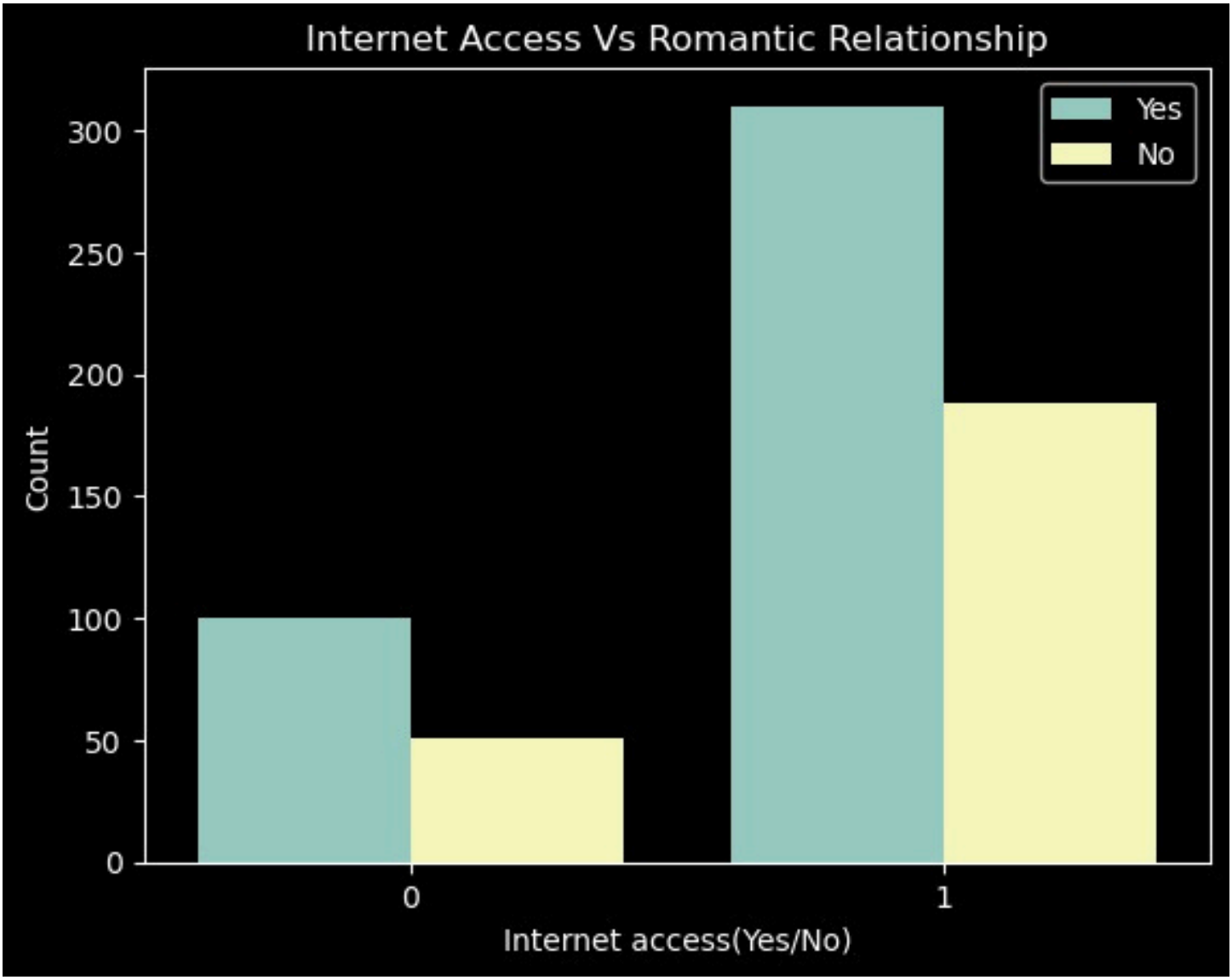
correlation Value : **-0.009278**

Conclusion :

So we can conclude that students from both backgrounds are likely to be part of extracurricular activities.

Question - 4 . Does internet access influence romantic relationships ?

Graph :



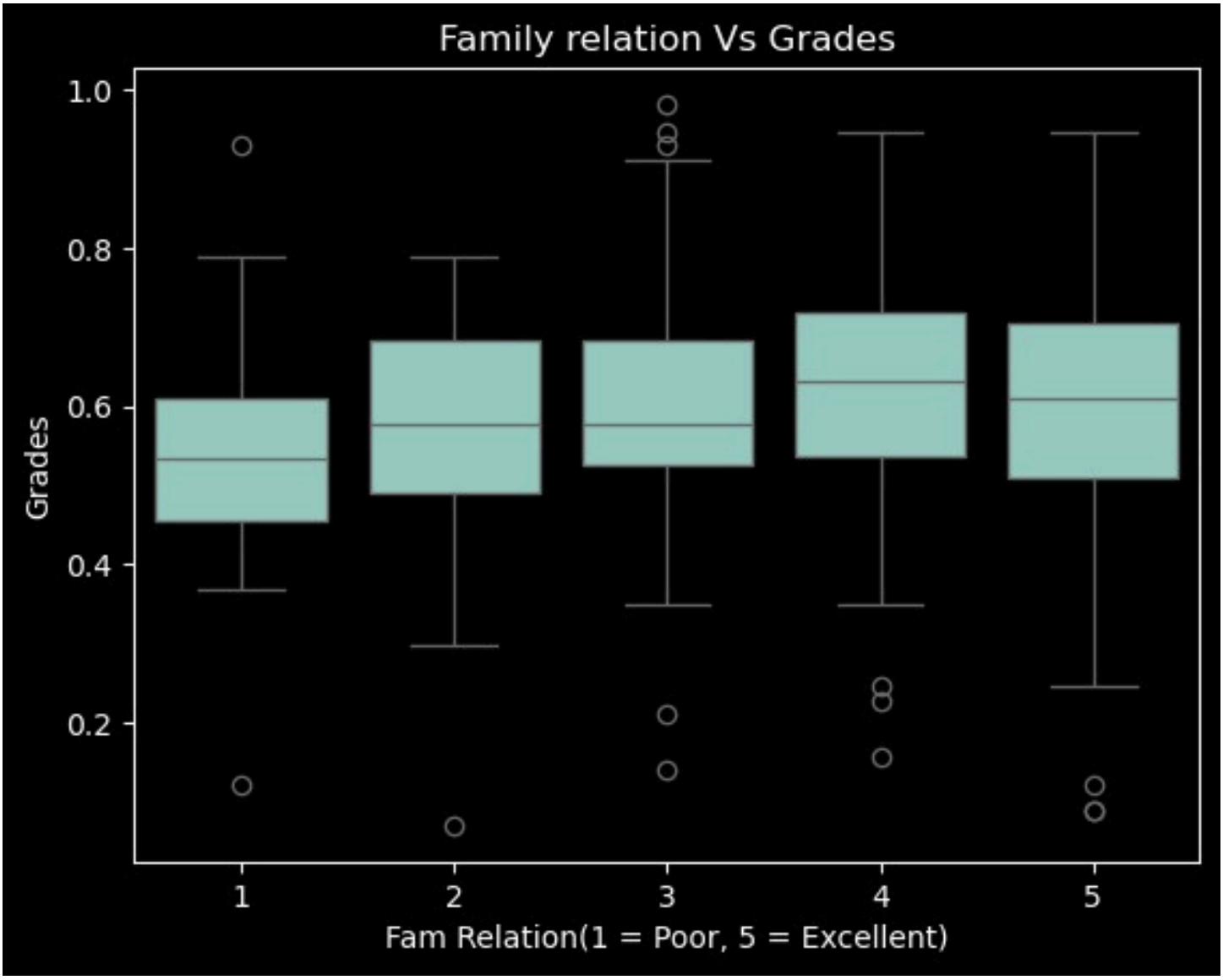
correlation Values : 0.034832

Conclusion :

So we can conclude that the internet has an influence on relationships.

Question - 5 . How Family Relations A affect Grades ?

Graph :



correlation Values : **0.074289**

Conclusion :

So that we can conclude that family relations a ect grades positively.

Level - 4 (Relationship Prediction Model)

Now I am at this stage I am all set to start training my model .

Steps i follow will training my model :

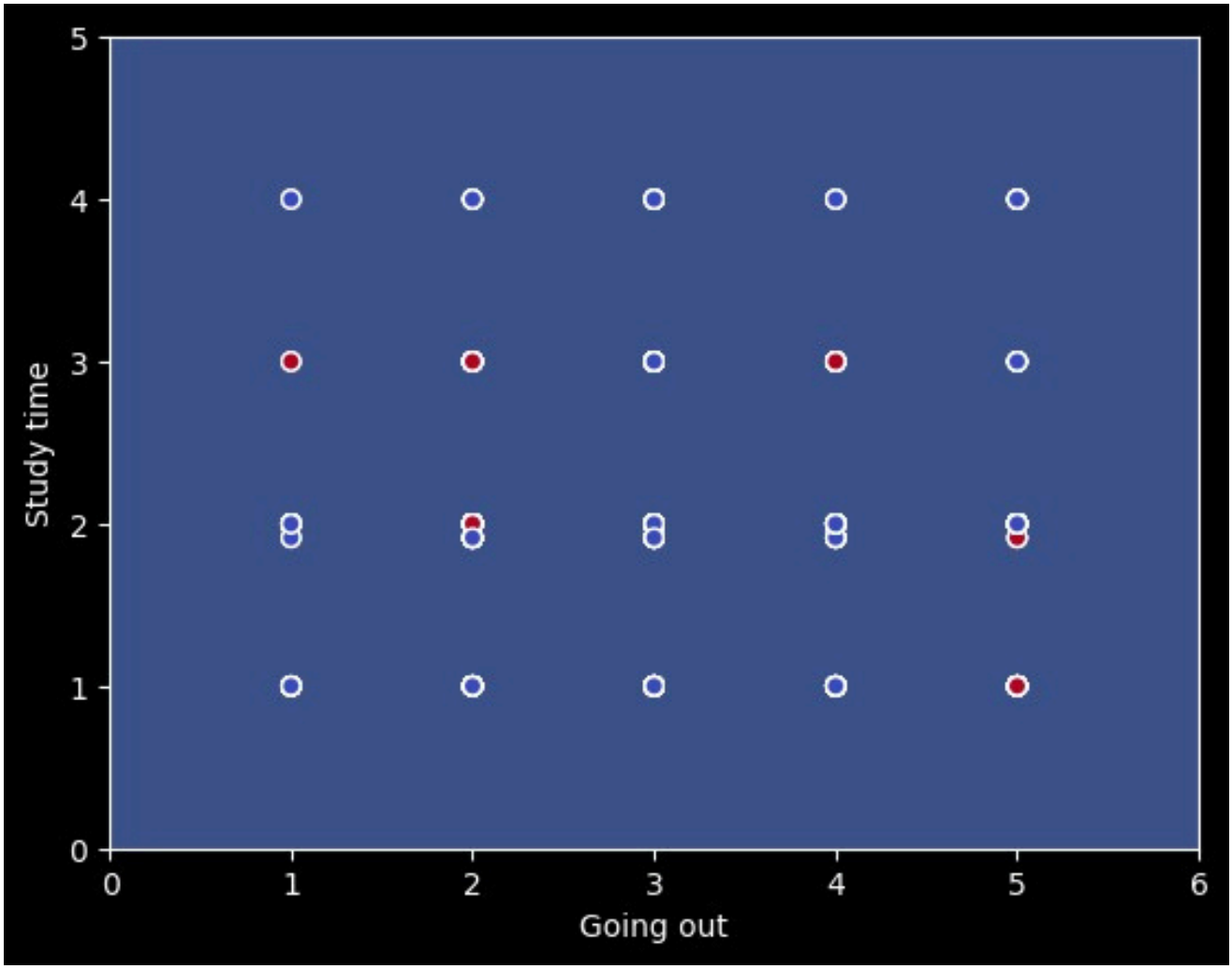
- First o I split the dataset into two sets: Train and Test sets .
- While training I understood that columns with high correlation values confuses the model so i drop grades columns and add a average grade column
- At first I tried a decision tree model. I got around 55%.
- I train my dataset using a logistic regression model. And I calculate the Accuracy of the model. It comes around 60 % and sticks with the logistic regression model .

My Thoughts on accuracy :

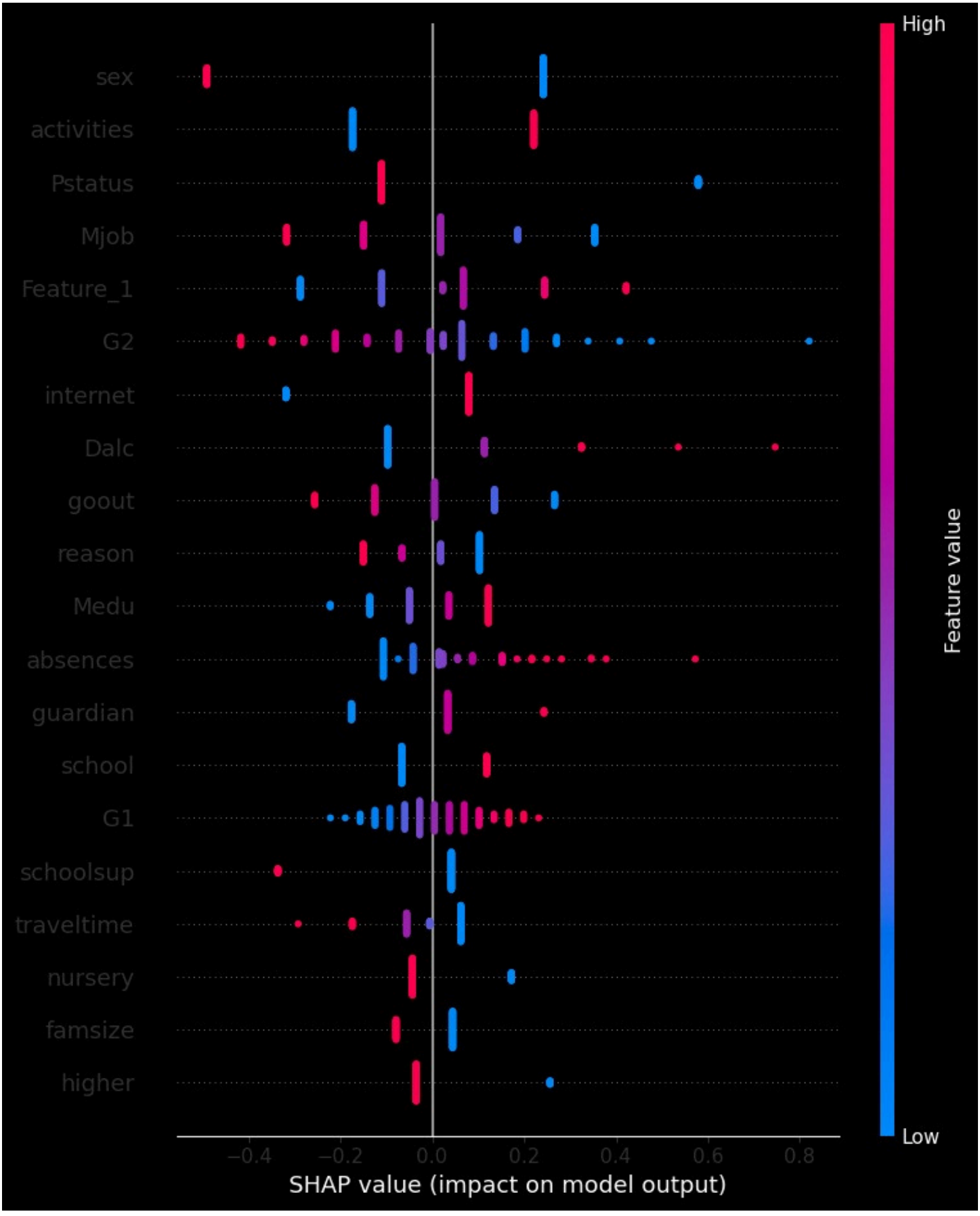
- At first try I got around 50 accuracy and then worked on column refining and got accuracy up to 59 to 60%. I think a dataset of 600 values is too small for a model to find patterns in data.
- Or it may not be But I tried my level best to increase the accuracy up to this mark .

Level -5 Model Reasoning Interpretation

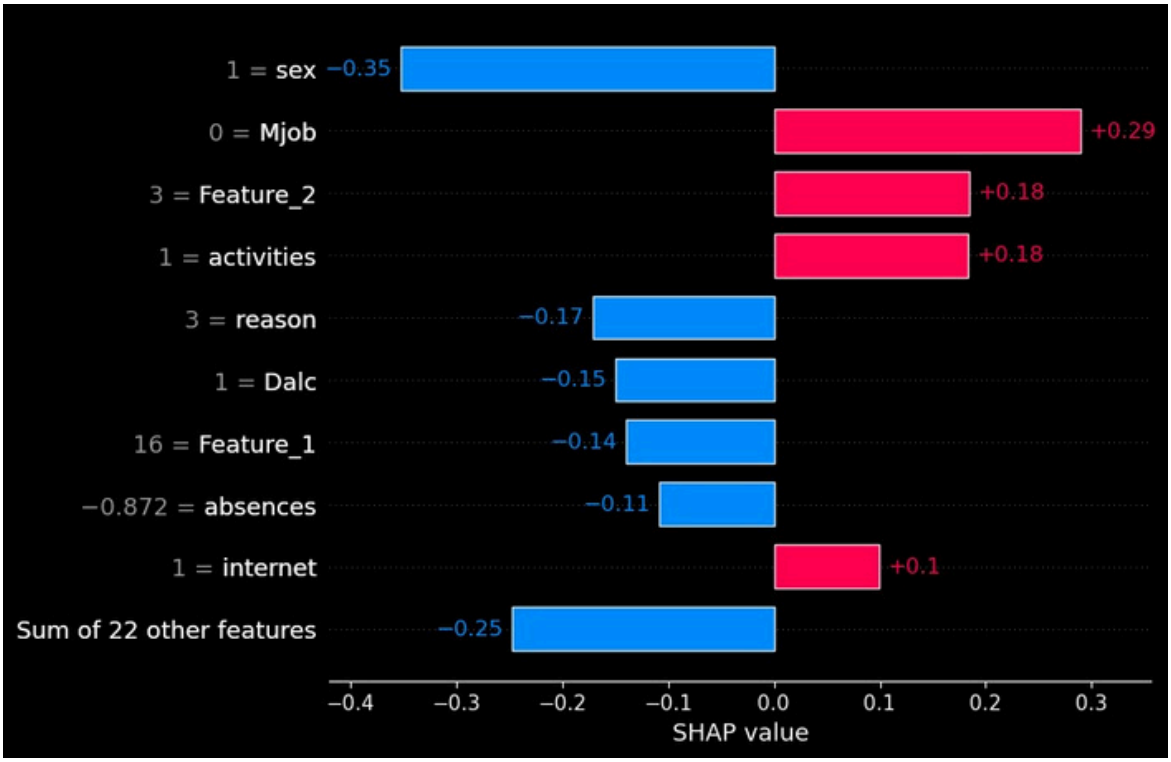
Visualizing decision boundaries using 2D feature pairs:

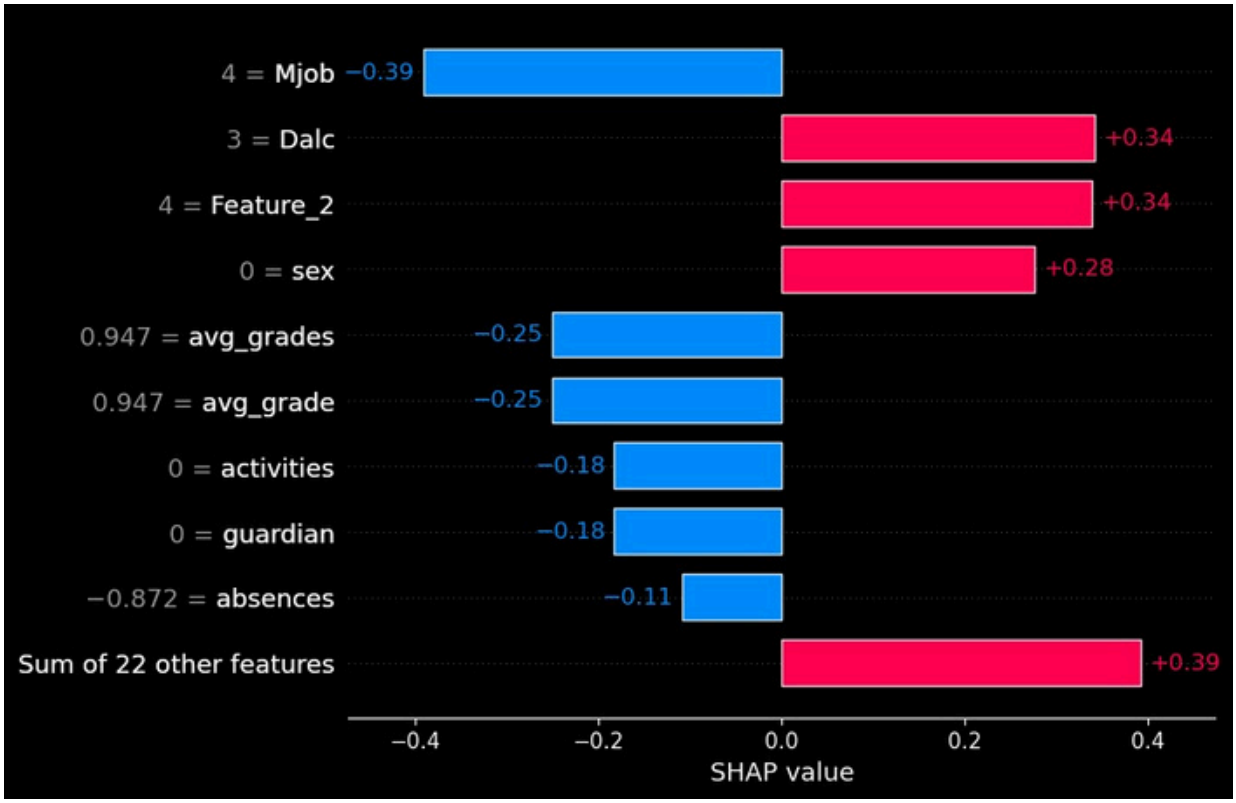


Plotting global feature importance



Generating local explanations for two students, one predicted “Yes” and one “No.





Interpreting the results in plain language

1. Students who involved in social activities tends have a relationship.

2. Students with good family relationships tend to have relationships due to strong emotional damage.

3. Students who are good tend to have less chance of relationships because of the academic pressure and less freetime .