CAMPUS PULSE

TANMAY SATYAJ DWIVEDI
240104104

INTRODUCTION

The CampusPulse initiative is aimed at understanding student lifestyle patterns through anonymized survey data and building a predictive model to infer whether a student is likely to be in a romantic relationship. The challenge involves data cleaning, exploratory data analysis (EDA), classification modeling, and explainability.

DATASET OVERVIEW

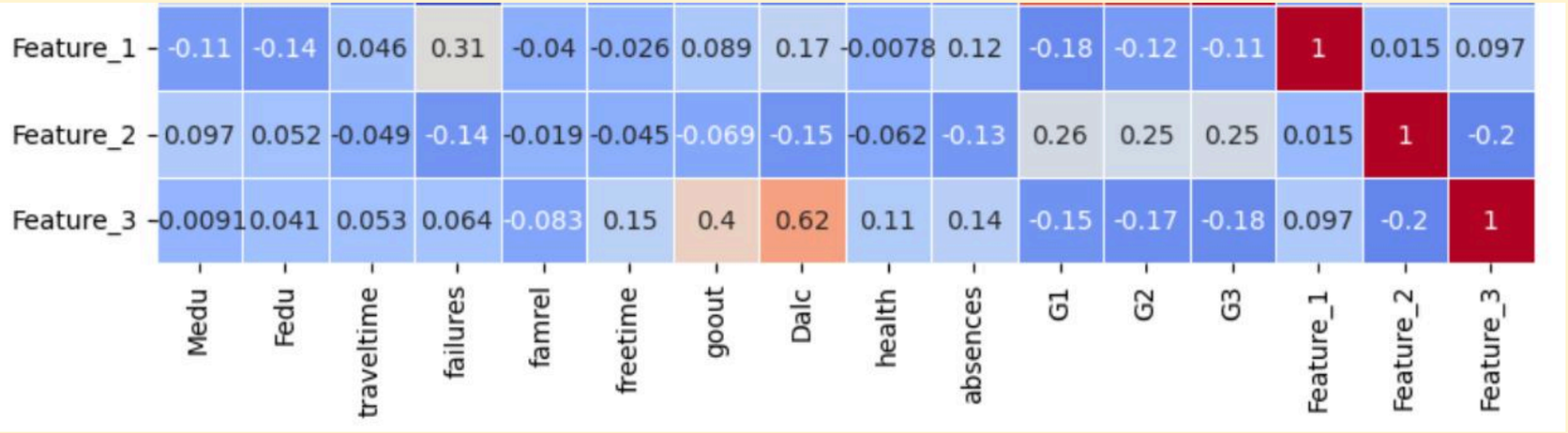
The dataset consists of anonymized responses from students, including academic performance, social habits, health indicators, and more. One critical variable of interest is romantic (Yes/No).
A few features (Feature_1, Feature_2, Feature_3) were anonymized and needed identification using statistical techniques.

This dataset, while anonymized, reflects diverse aspects of student behavior—academic habits, family background, social interaction, and health. The challenge lies not only in prediction, but in interpreting the subtleties of student life from numerical patterns. Each feature, especially anonymized ones, holds potential insights into the emotional and academic fabric of campus life.

FEATURE IDENTIFACTION

To uncover the meaning of Feature_1, Feature_2, and Feature_3, I applied:

- Correlation Heatmaps
- Scatter Plots
- Statistical Grouping



#INFERENCES

- Feature_1 showed strong correlation with goout and absences → possibly free time or social activity
- Feature_2 was closely linked with parental education and grades → possibly weekly study time
- Feature_3 correlated with health and Dalc → possibly stress or mental health indicator

DATA PREPROCESSING

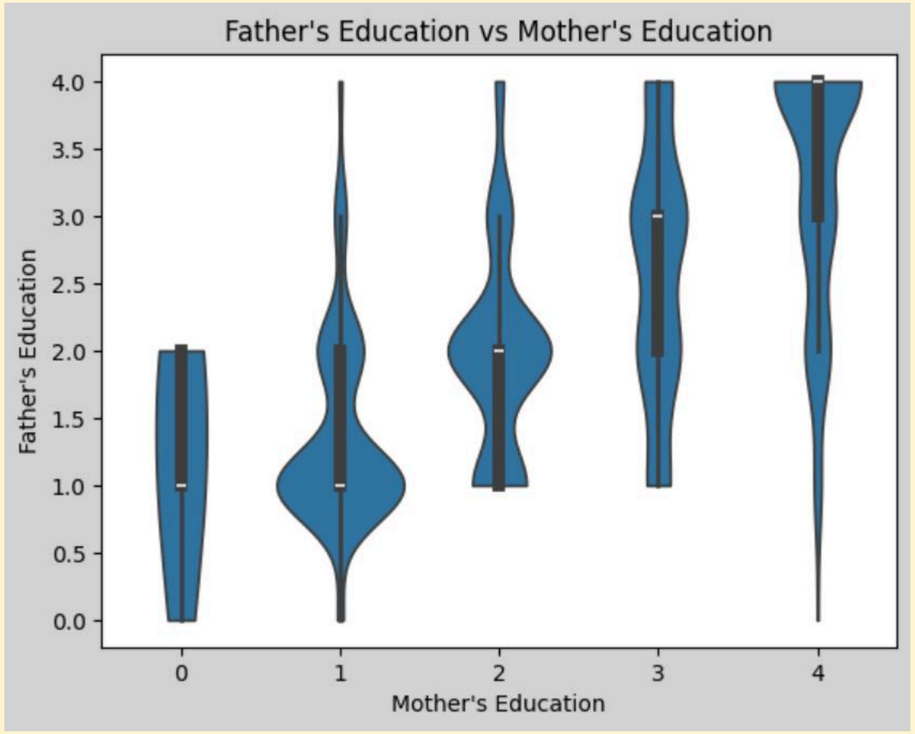
Survey-based datasets are prone to missing or inconsistent entries due to skipped questions, ambiguity, or user error. To prepare the dataset for modeling and analysis, I conducted a thorough null value audit and applied customized imputation strategies based on feature type, correlations, and domain logic.

1) KNN Imputation for Fedu (Father's Education)

Visual analysis (violin plot and scatter plots) showed a strong positive correlation between Fedu and Medu. Since both are ordinal and socio-demographically related, Fedu can be estimated using KNN by looking at similar students based on features like Medu, age, and studytime.Also KNN preserves variability better than filling with mode or mean.

K-Nearest Neighbors (KNN) Imputer with n_neighbors=5

Correlated feature: Medu (Mother's Education)



2)famsize- Family Size did not have any particular important feature. Because no feature has as such major influence the best option was to check over the data and see the trend , instead of putting just mode RandomForest was used to predict the missing values in famsize.The logic behind was simple , to fill the missing values with data having trend which does not cause any irreugalrity.

3)higher- The missing values of higher was filled with same logic used in famsize but famsize did not have any important features but higher had some feature affecting it greater than others. So RandomForestClassifier was used but with the important features isolated out i.e the prediction was based on the pattern deduced from important features.

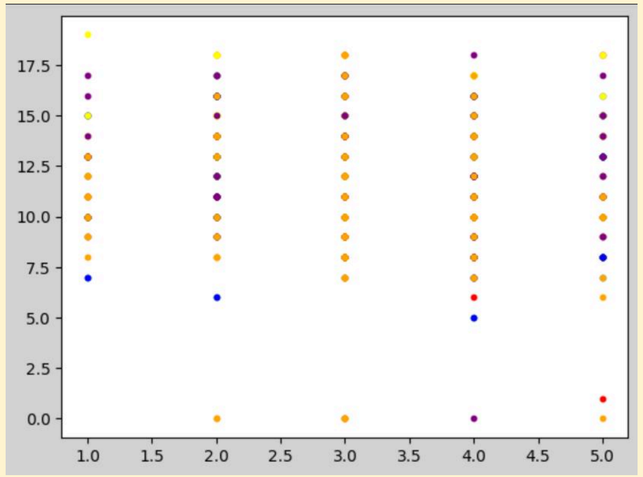
4)traveltime , freetime , absences - All these 3 columns are independent and had no other columns backing them up so as to maintain data integrity and consistency all of thses datas were filled mode.This prevents any descrepancy confusing the model.

5)G2- Its a mathematical quantity so the best option to deal with was to find the relation of G2 with other to of its most closely realted columns G1,G3.First a relation was found between all three of the column values and then after getting G2 one one side the null values for G2 were filled using that formula which is correct both academically and data consistency wise.

6)The values of Feature 1 ,2 and 3 were handled as our initial analysis on finding what they are. Feature_1 was predicted using KNN classifier. Feature_2 using a mathematical relation found between its most affecting columns.Feature_3 was handled using the Liner Regression model.

EXPLORATORY INSIGHT QUESTIONS

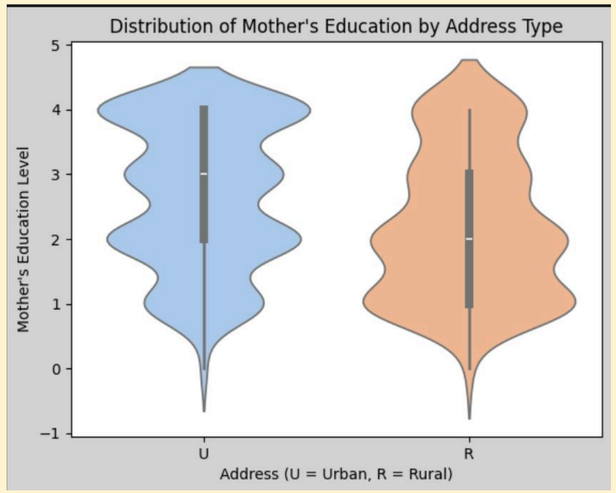
1.Is there a negative correlation between students' free time and their final academic performance?



This question was specifically chosen as to get a good hold and data while keeping in mind that the derived should help find a trend in campus student. The graph did not show much variance and our assumption or the answer to the question is no.

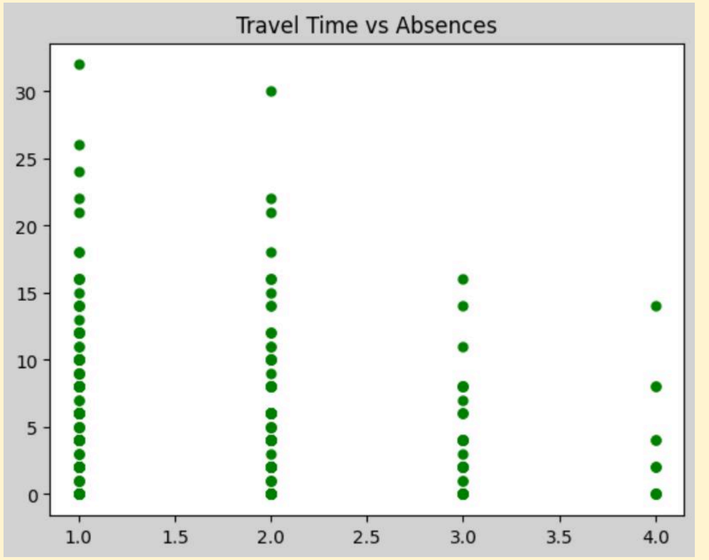
2.Is a student’s mother’s education level influenced by whether the family lives in a rural or urban area ?

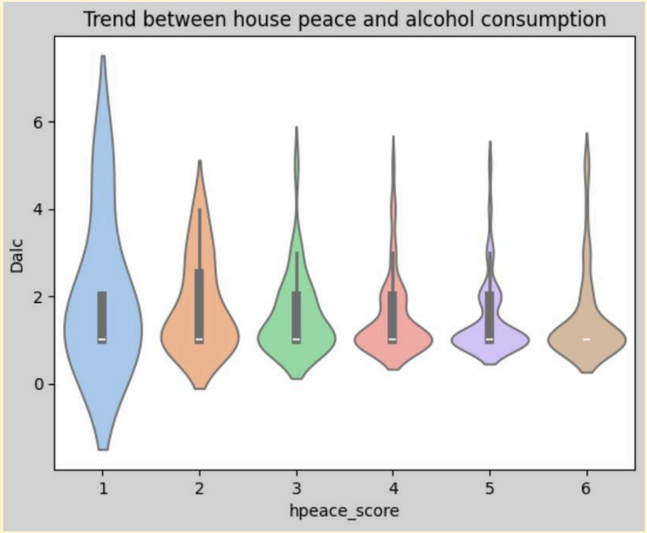
The violin plot completely backs up our question and screams yes.It particulary depicts the changing trends of Medu with location. This question was primarily chosen to address the problem and verify the fact tha women in rural areas are stillunder-educated as compared to those in urban.



3.Does longer travel time to school (traveltime) correlate with a higher number of absences (absences)?

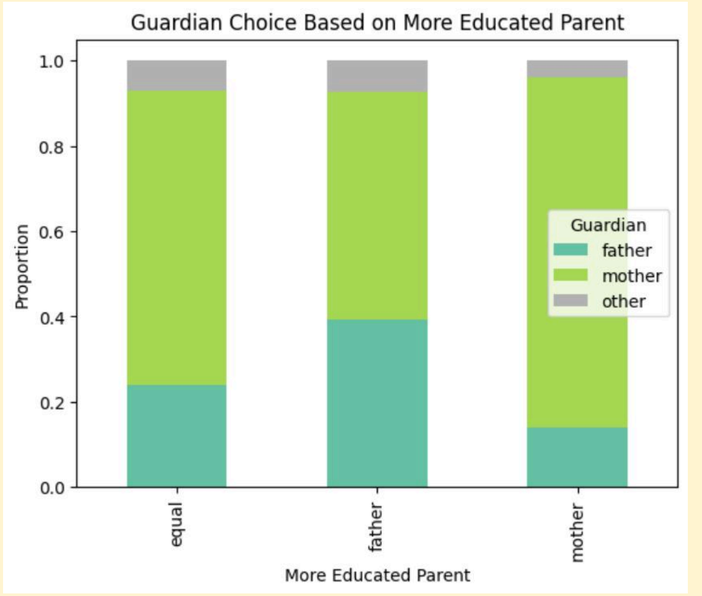
Its only logical to think that the student living far away would have more absence on analysing the data through the scatter plot the trend turned to be completely opposite of what was being assumed





4.Do Parental status and family relation tend to have an effect on alcohol consumption ?

The family environment matters a lot to mental health to check whether a student falls in alcohol habit due to less peace and happy environment. The violinplot was made with a new defined feature called hpeace_score which combined both factor of parents living apart or together and family relations.



The mother was found out to be the primary guardian of a student even she is not more qualified than the father.

5.Is a parent with a higher education level more likely to be the declared primary guardian ?

Relationship prediction model

The goal was to build a machine learning model that can accurately predict whether a student is in a romantic relationship based on a variety of attributes like academic performance, social habits, family background, and lifestyle. This required identifying useful features, handling categorical data, training multiple classifiers, and selecting the best-performing one.

Data Preparation -

1)The target variable romantic was originally in categorical format ('yes' or 'no'), so it was converted to numerical values (1 for "yes", 0 for "no") using Label Encoding.

2)All other categorical features (like school, sex, Mjob, etc.) were also label-encoded to prepare the data for machine learning algorithms, which require numerical inputs.

3)The feature set (X) was created by dropping the target column romantic, while the target (y) was the encoded values of the romantic column.

4)The dataset was split into training (80%) and testing (20%) sets using train_test_split() from sklearn.

Models Tried-

I experimented with multiple classification models to determine which one predicted the romantic status most accurately:

1)Logistic Regression – A baseline linear classifier.

2)Support Vector Machine (SVM) – Effective in high-dimensional spaces.

3)Decision Tree – Easy to interpret but prone to overfitting.

4)K-Nearest Neighbors (KNN) – Simple, distance-based approach.

5)Random Forest Classifier – An ensemble of decision trees with high robustness.

Each model was trained using the training set and evaluated on the test set.

Model Reasoning and interpretation

Decision Boundary Visualization

Objective

The aim of this step was to visually understand how the trained classifier separates the two classes ("Yes" and "No" for being in a romantic relationship) based on specific features. This is important because it helps us interpret how the model sees the data and where the decision cutoffs lie.

What Was Done

- Two features were selected: Feature_1 and goout (which were found to be significant from earlier analysis).
- A 2D plot was created using these two features as axes.
- A decision boundary was drawn to show how the classifier (Random Forest) splits the space into regions predicting "Yes" vs. "No".
- The training samples were plotted, color-coded by their actual class.

##Insights from the Decision Boundary

- The boundary was non-linear and jagged, which reflects how decision trees (and ensembles like Random Forest) make decisions by splitting the feature space into axis-aligned sections.
- Students with higher values of goout and Feature_1 were more likely to fall into the "Yes" region.
- The boundary was more confident in dense areas and more irregular where data was sparse — showing the limits of the model's understanding.

##Why It Matters

Decision boundary plots provide intuition about:

- Where the model struggles to classify (near boundary)
- How separable the classes really are
- Whether feature choices make sense for prediction

Model Reasoning and interpretation

SHAP Value Analysis
Objective
This task was about explaining why the model makes a certain prediction. SHAP (SHapley Additive exPlanations) values allow us to peek inside a "black box" model like Random Forest and see which features are pushing a prediction toward "Yes" or "No".
What Was Done
- SHAP was applied to the trained Random Forest model.
- A global SHAP summary plot was generated to show the most influential features across the entire dataset.
- Two individual predictions were selected:
  - One student predicted to be in a relationship
  - One student predicted not to be in a relationship
- For each, a SHAP force plot was generated to break down how much each feature contributed to the final prediction.

##Insights from SHAP
- Top Global Features:
  - goout, Feature_1, absences, and sex were among the most important.
  - Social activity and lifestyle-related attributes strongly influenced romantic relationship predictions.
- Local Explanations:
  - In the "Yes" case, high values for goout and Feature_1 pushed the prediction toward 1.
  - In the "No" case, lower values in those features led the model to predict 0.

Why SHAP Is Important
- It adds trust and transparency to machine learning models.
- Stakeholders can see why a prediction was made — not just what the result is.
- SHAP is especially important in sensitive or personal contexts like this dataset.