1. Introduction

The objective of Task 1 in the CampusPulse project is to analyze student data and build predictive models based on survey results. This includes exploratory data analysis, preprocessing, model training, and evaluation. The dataset consists of 649 records and 33 features, covering demographic, academic, and behavioral information[1].

2. Step-by-Step Approach

2.1 Initial Inspection

The dataset was loaded and inspected using `.info()` and `.describe()` methods.

- **Rows and Columns:** 649 rows × 33 columns.
- **Data Types:** Both categorical (e.g., school, sex, address) and numerical (e.g., grades, absences, health) features.
- **Missing Values:** Several columns contain missing values, notably `famsize`, `Fedu`, `traveltime`, `higher`, `freetime`, `absences`, and some engineered features (`Feature_1`, `Feature_2`, `Feature_3`) [1].
- **Key Features Identified:**
  - Demographics: school, sex, address, family size, parental status, parental education/jobs
  - Academic: G1, G2, G3 (grades for three periods), absences
  - Behavioral: daily alcohol consumption (Dalc), health, romantic status

Visualizations such as histograms and heatmaps were created to understand feature distributions and relationships. Based on these, the anonymous features were inferred as Age, Daily hours of study, and Partying frequency[2].

2.2 Data Preprocessing

- **Missing Value Handling:**
  - Numerical features were imputed with mean or median values.
  - Categorical features were imputed with the mode.
  - Ratings and engineered features were imputed with the median where appropriate[2].
- **Encoding and Scaling:**
  - Categorical variables were encoded for modeling.
  - Numerical features were standardized for uniformity.

- **Data Consistency:**
  - Outliers and inconsistencies were checked and addressed.
  - Ensured all records were validated after imputation.

2.3 Exploratory Data Analysis

- Explored the data by asking and answering five meaningful questions to reveal insights about the dataset.
- Plotted a variety of graphs and visualizations to analyze trends and relationships, selecting the most informative and visually clear ones.
- Observed that some popular stereotypes were not supported by the data.
- Key findings included strong correlations among grade columns (G1, G2, G3), and moderate relationships between behavioral features (like alcohol consumption) and academic performance[2].

2.4 Model Building

- **Feature Scaling:**
  - Standard scaler was used to scale features for modeling.
- **Model Experiments:**
  - Logistic Regression: Used as a baseline, but yielded low accuracy.
  - Random Forest: Tried with and without hyperparameter tuning (GridSearchCV); performance was limited.
  - Naïve Bayes: Provided the best accuracy and F1 score (final pick, ~72% accuracy).
  - KNN: Tested after level 5, but did not outperform Naïve Bayes[2].
- **Evaluation:**
  - Confusion matrices and classification reports were generated for each model.
  - Hyperparameters such as max depth and number of estimators were tuned for Random Forest.

2.5 Model Analysis

- SHAP plots were generated for the best model to interpret feature importance.
- Decision boundaries were visualized for all models.
- Key findings included the importance of early grades (G1, G2), health, and absences as predictors for final grade (G3), with behavioral features like daily alcohol consumption and family support also contributing[2].

3. Bonus Task

- Decision boundaries were initially classified based on visualizations, and corrections were made after further analysis[2].

4. Conclusions and Recommendations

- **Data Quality:** Addressing missing values is crucial for robust modeling.
- **Predictive Insights:** Early academic performance and health are strong indicators of final results.
- **Modeling:** Naïve Bayes provided the best trade-off between accuracy and interpretability for this dataset.
- **Next Steps:** Further analysis could explore more complex models or additional engineered features to improve predictions[1] [2].