Project dataset: Higher Education Students Performance Evaluation
https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation

This project aims to predict end-of-term academic performance of higher education students by analyzing a diverse range of factors captured in the "Higher Education Students Performance Evaluation" dataset. The primary objective is to identify which personal, familial, and educational habits contribute most significantly to students' cumulative GPA and expected graduation GPA, providing insights into the predictors of academic success.

## Research Questions

- Which factors are the strongest predictors of academic performance (as measured by cumulative GPA)?
- To what extent do socio-economic factors, such as family background and personal demographics, influence student performance?
- How do educational habits and behaviors, such as study hours and exam preparation, correlate with end-of-term academic outcomes?

## Methods and Approach

### Exploratory Data Analysis

- Conduct summary statistics to identify the distribution and central tendencies of key variables.
- Use visualization techniques (e.g., histograms, box plots) to explore the distribution of the GPA variable.
- Generate a correlation matrix for numerical variables to observe relationships among features and identify any high correlations that may impact the models.

### Modeling Techniques

**Regression for GPA Prediction:**

- Linear Regression: To provide a baseline predictive model and interpret the impact of each variable on GPA.
- Random Forest Regression: As a more flexible model that can capture non-linear relationships and interactions between variables, potentially improving predictive accuracy.

**Classification for GPA Performance Categories:**

- Logistic Regression: To model the probability of students belonging to different performance categories (High, Medium, Low) based on their characteristics.
- KNN: A non-parametric classifier that predicts a student's performance category by examining the categories of similar students, offering an alternative approach to logistic regression.

**Model Evaluation**

- For regression models, use Mean Squared Error to evaluate model accuracy, comparing linear regression and random forest results.
- For classification models, calculate metrics like accuracy, precision, and recall to compare logistic regression and KNN performance.
- Use cross-validation (e.g., k-fold cross-validation) to ensure robustness in model evaluation, especially for hyperparameter tuning in KNN.