Machine Learning-Based Prediction of Student GPA from Academic Behaviors

Group 6
Đỗ Nguyễn Gia Như Nguyễn Quang Anh Vương Hồng Minh
Trần Thanh Phát Nguyễn Quang Minh

May 2025

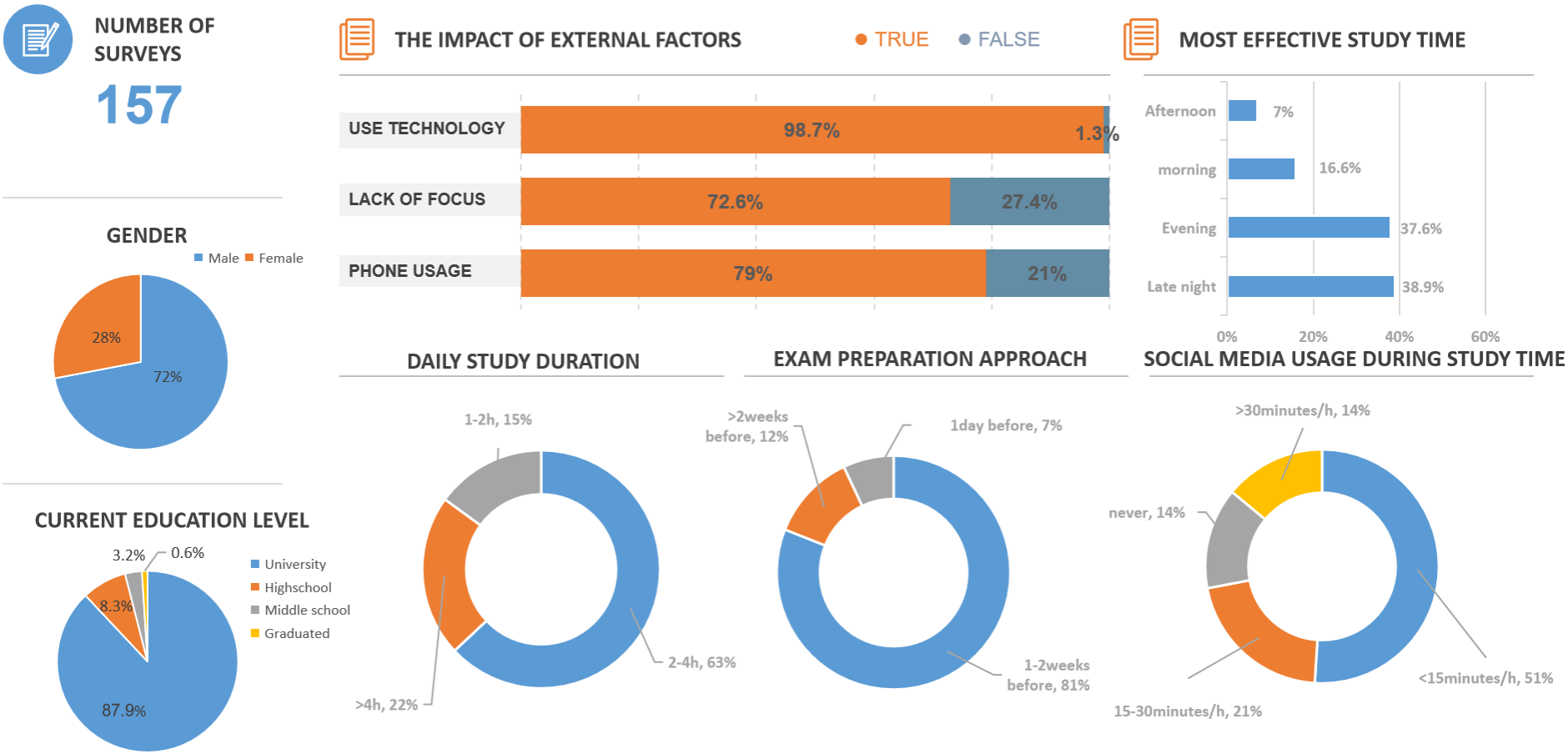
Table of content

Problem definition

Project Goals

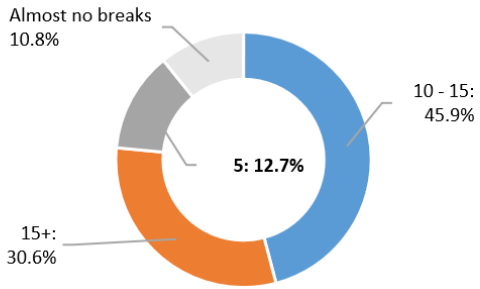
- **Predict** student academic performance
- **Analyze** learning patterns
- **Identify** at-risk students
- **Provide** early interventions

Summary Dashboard

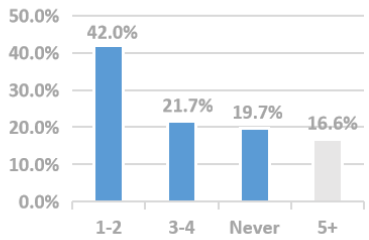


Summary Dashboard

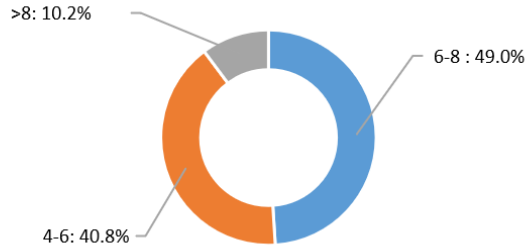
Study break frequency (min/h)



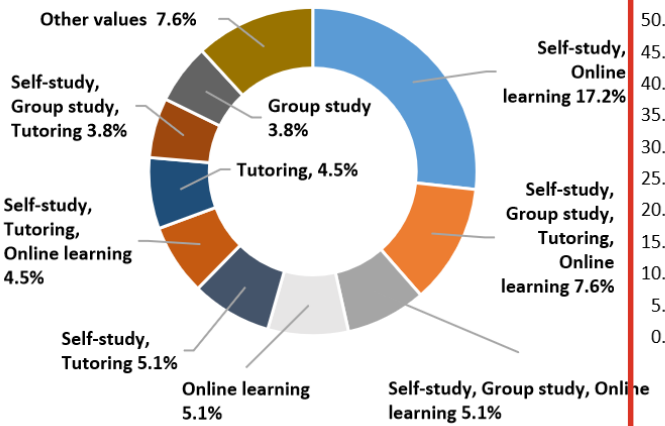
Weekly exercise frequency(times)



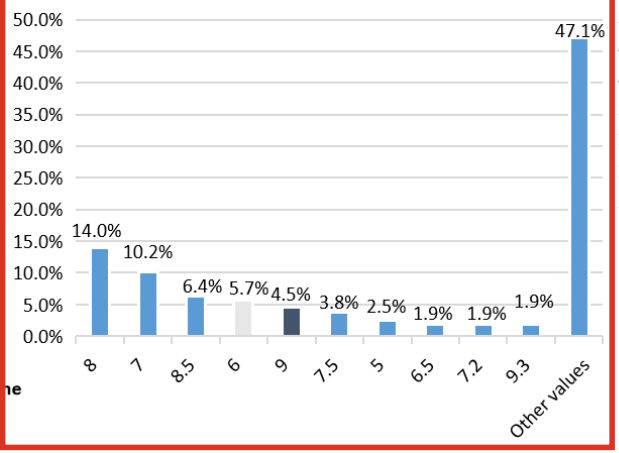
Daily sleep duration (hours)



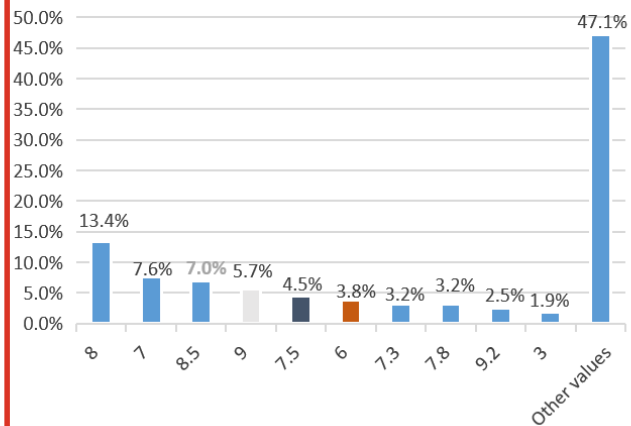
Preferred study method



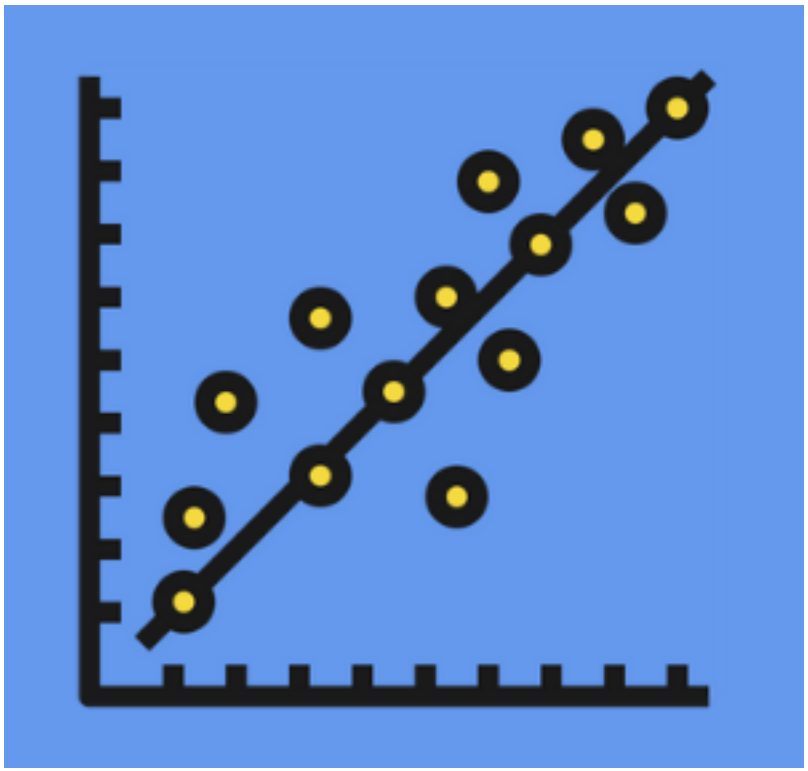
Latest semester GPA



Latest academic year GPA



Methodology



Linear Regression

with some libraries

Methodology

1. **Data preprocessing**
   - The dataset is divided into three types: **numerical**, **ordinal**, and **nominal** features.
   - Missing values are handled using `SimpleImputer`.
   - The data is split into 80% training and 20% testing sets.
2. **Training the model using Linear Regression**
   - Apply a bounded Linear Regression model with pipeline integration.
3. **Evaluating the model**
   - Use `MAE`, `MSE`, and R^2 to assess performance.
4. **Predicting output using the test set**
   - Generate predicted GPA values and compare with actual outcomes.

Result

- **Predicted & Actual GPA:**

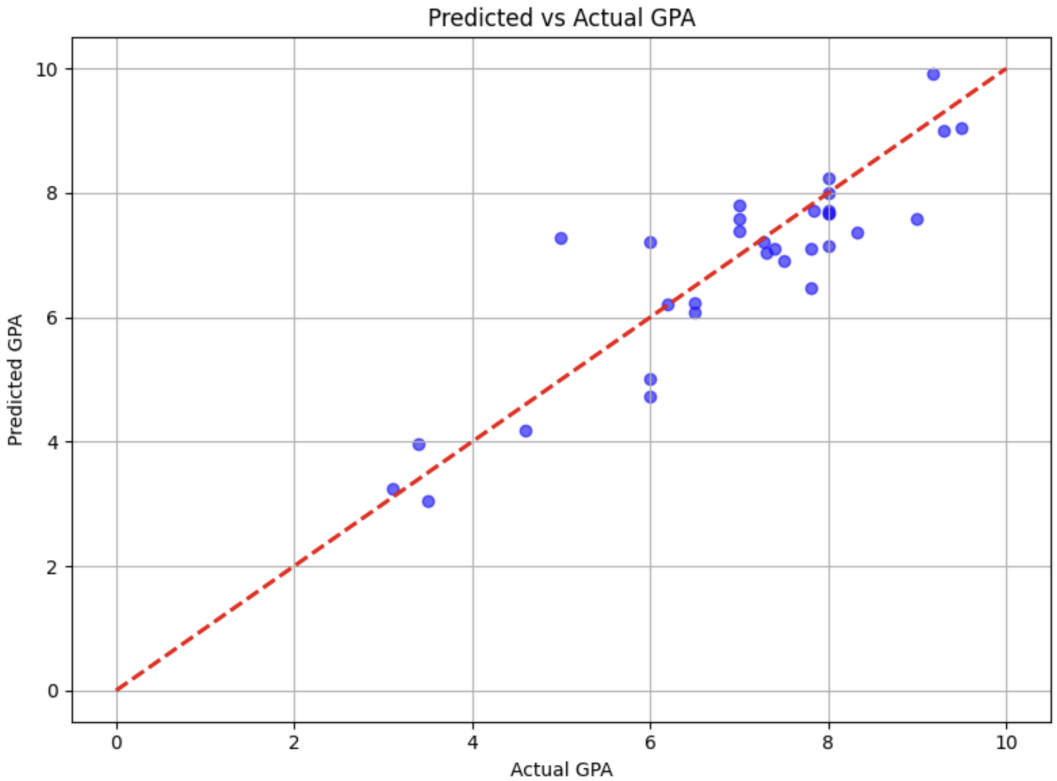
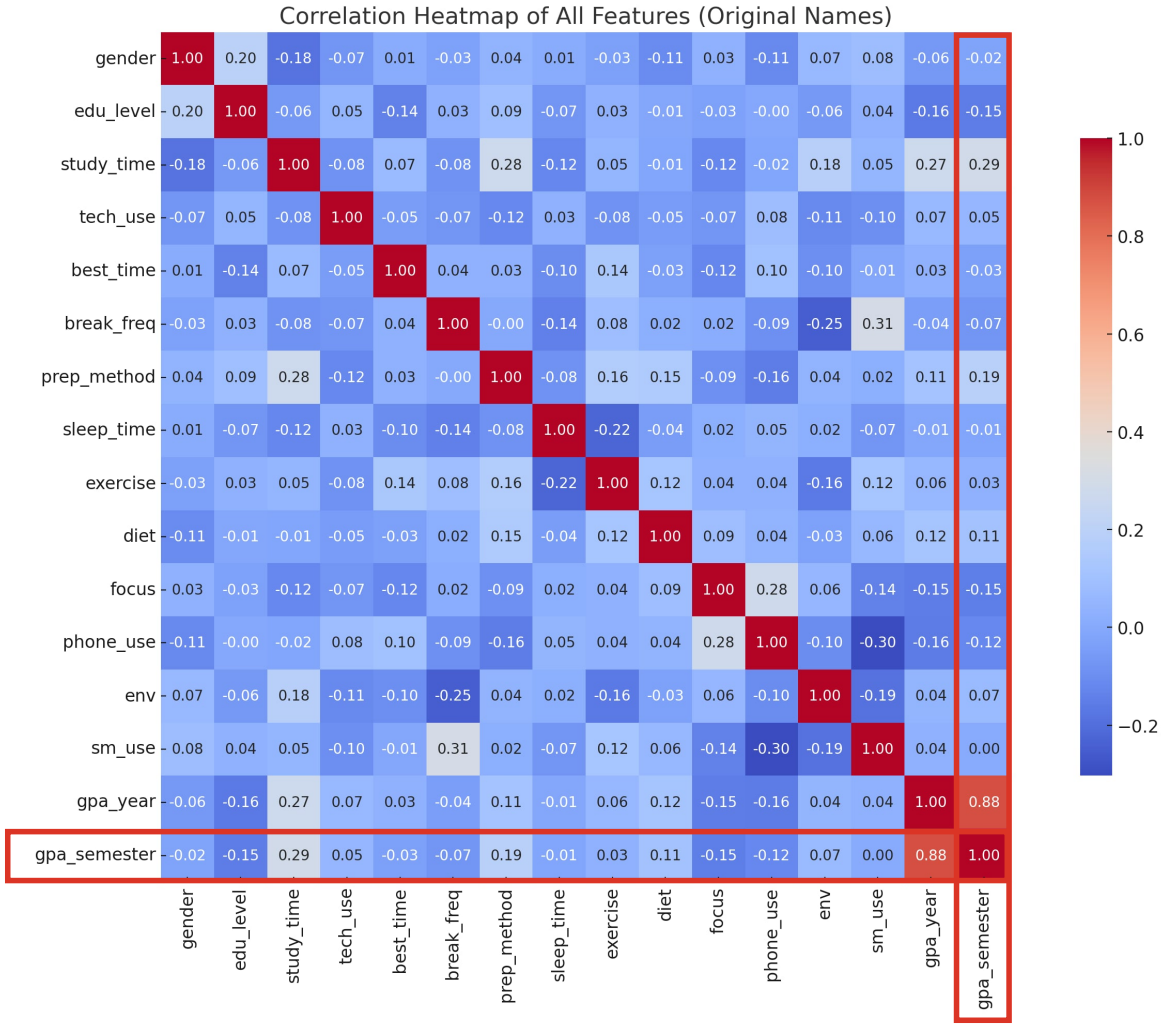


Figure 3.1: Predicted vs Actual GPA

Result

- **Correlation heatmap of all features:**



Model evaluation

Evaluate Model Results

- **Mean Absolute Error (MAE)** ≈ 0.597 — The average of absolute differences between predicted and actual GPA values.
- **Mean Squared Error (MSE)** ≈ 0.596 — Gives more weight to large errors, useful for identifying outliers.
- **Coefficient of Determination (R^2 score)** ≈ 0.755 — Indicates that the model explains about 77.5% of the variance in GPA.

$\Rightarrow$ **The model shows good predictive performance overall.**

Model evaluation

Disadvantages

- Only works well if data is simple and linear
- Can be wrong if data has outliers
- Cannot learn complex patterns
- Gets confused if features are too similar

Future work

- Try better models like Random Forest or Neural Network
- Add more useful features (class time, mental health, etc.)


DEMO
TIME!


Thank you
for listening