Student Performance Analysis Project

Master's Level Data Science Exercise

Project Overview

Objective

This project aims to analyze student performance using a synthetic dataset that captures complex interactions between individual, academic, and socio-economic characteristics. You will apply advanced statistical modeling techniques to understand the factors influencing academic success and graduation probability.

Dataset

A comprehensive synthetic dataset (`student_performance_dataset.csv`) is provided, capturing multidimensional aspects of student performance.

Data Dictionary

Student Performance Dataset (`student_performance_dataset.csv`)

Demographic Variables:

- `student_id`: Unique identifier for each student
- `name`: Student's name
- `age`: Student's age
- `gender`: Student's gender identity
- `major`: Academic major
- `learning_style`: Predominant learning style
- `parents_education_level`: Highest education level of parents

Academic Performance Variables:

- `high_school_gpa`: High school academic performance

- `study_hours_per_week`: Average weekly study time
- `english_proficiency`: English language skills score
- `math_proficiency`: Mathematics skills score
- `science_proficiency`: Scientific reasoning skills score
- `final_gpa`: Cumulative college GPA
- `graduation_within_4_years`: Binary indicator of timely graduation
- `academic_performance_description`: Qualitative performance assessment

Contextual Variables:

- `family_income`: Household income
- `personal_essay`: Student's reflective writing sample

Analysis Tasks

Task 1: Exploratory Data Analysis (20%)

1. Comprehensive exploratory analysis
    - Descriptive statistics for numerical variables
    - Distribution analysis of performance metrics
    - Correlation matrix between academic and demographic variables
2. Visualization Requirements:
    - Performance variations across majors
    - Study hours and GPA relationship
    - Learning style performance distributions

Task 2: Regression Modeling (40%)

Objective: Develop regression models to predict academic performance

1. Linear Regression: Predicting Final GPA
    - Dependent Variable: `final_gpa`
    - Independent Variables:
        - `high_school_gpa`
        - `study_hours_per_week`
        - `parents_education_level`
        - Subject proficiency scores
        - Demographic factors

    Requirements:

- Implement multiple linear regression
- Check and address multicollinearity
- Interpret coefficients and statistical significance

2. Regularized Regression

   - Apply Ridge/Lasso regression
   - Compare model performance metrics
   - Discuss feature importance

Task 3: Binary Outcome Prediction (40%)

Objective: Predict graduation probability

- Dependent Variable: `graduation_within_4_years`
- Independent Variables:
  - Academic performance metrics
  - Socioeconomic indicators
  - Learning style
  - Subject proficiencies

Requirements:

- Implement both Logit and Probit models
- Compare model performance using:
  - Accuracy
  - AUC-ROC
  - Confusion matrix
- Interpret marginal effects
- Discuss model selection criteria

Submission Requirements

1. Comprehensive analysis report (max 15 pages)
2. Fully documented code
3. Detailed result interpretations
4. Discussion of limitations and potential improvements

Evaluation Criteria

- Technical Complexity (40%)
- Statistical Rigor (30%)
- Visualization Quality (15%)
- Interpretation Depth (15%)

Bonus Challenges

1. Explore machine learning extensions
2. Develop advanced feature engineering techniques
3. Create predictive models for academic interventions

References