Applied Statistics for Data Science

Student Task: Exploring Multinomial Logistic Regression

Objective

In this task, you will explore a real-world dataset and use a **multinomial logistic regression model** to predict student performance categories. You will carry out basic data exploration, fit the model, evaluate its accuracy, and submit your work via GitHub.

Dataset

- Dataset: **Student Performance Data Set**
- Source: `https://archive.ics.uci.edu/ml/datasets/student+performance`
- File: `student-por.csv`
- Target variable: `G3` (final grade, numeric from 0–20)

Step 1: Recode the Target Variable Recode `G3` into 3 categories:

- **Low** (0–9), **Medium** (10–14), **High** (15–20)

Instructions

Use either **R** or **Python** and complete the following tasks:

Task 1. Explore the data

- Show the number of observations in each class (Low, Medium, High)
- Create one plot to compare a numeric predictor (e.g., `studytime`, `absences`) across the three categories

Task 2. Fit the multinomial logistic regression model

- Use 3–4 predictors of your choice
- State the baseline category used in the model

Task 3. Interpret one coefficient

- Choose one non-baseline class and one predictor
- Interpret the coefficient in simple language (e.g., how it affects the odds)

Task 4. Evaluate model accuracy

- Create a confusion matrix
- Report the overall classification accuracy

Task 5. Write a short summary

- In 3–5 sentences, describe what you learned about the data and how well the model performed

Submission Instructions

- Upload your code and any related files (e.g., CSVs, plots, or notebooks) to a public GitHub repository.
- Include a brief `README.md` that explains:
  - The dataset and goal of your analysis
  - How to run your code (R script, RMarkdown, or Jupyter notebook)