DSCI-510 Final Project - Progress Report
Stephen Rosario

Project Scope Update

My initial goal was to explore how socio-academic factors influence student performance using the UCI Student Performance dataset. I successfully merged the math and Portuguese datasets (1044 rows total) and conducted exploratory data analysis, hypothesis testing, and machine learning modeling. The project scope has since expanded to include predictive modeling using a Random Forest Classifier to categorize students into performance tiers: *poor*, *average*, and *excellent*.

Progress Summary

- Conducted exploratory visualizations: distribution of final grades, grade vs. study time, correlation heatmap
- Performed t-tests on G1, G2, and G3 comparing students with vs. without home internet - all showed statistically significant differences
- Built and evaluated a Random Forest Classifier:
  - **Accuracy:** 0.57
  - **Top features:** absences, failures, age, free time
  - Generated and visualized confusion matrix + classification report

Data Sources

- **Local Data:** student-mat.csv and student-por.csv from the UCI Machine Learning Repository
- **API Data:** World Bank API — indicators used:
  - Government Spending on Education (% of GDP): SE.XPD.TOTL.GD.ZS
  - Tertiary School Enrollment (% gross): SE.TER.ENRR

Issues / Difficulties

- Took extra care to ensure .env, data/, and results/ directories were excluded from Git
- Classification warnings due to class imbalance (some categories underrepresented)
- Model accuracy is moderate — further improvements may involve hyperparameter tuning or ensemble models
- Next steps: experiment with additional features, improve class balance, and test other ML algorithms