Beyond the Classroom: *Predicting Student Performance with Socio-Academic and Global Indicators*

By: Stephen Rosario

DSCI-510 Final Project

Motivation / Introduction

Goal: Analyze how socio-academic factors influence student performance.

Key Question: Which factors predict final student outcomes?

Approach:

- Merge two UCI student datasets (Math + Portuguese)
- Add macro-education indicators
- Conduct EDA, hypothesis tests, and ML modeling

Outcome: Identify strongest predictors + build a performance classifier.

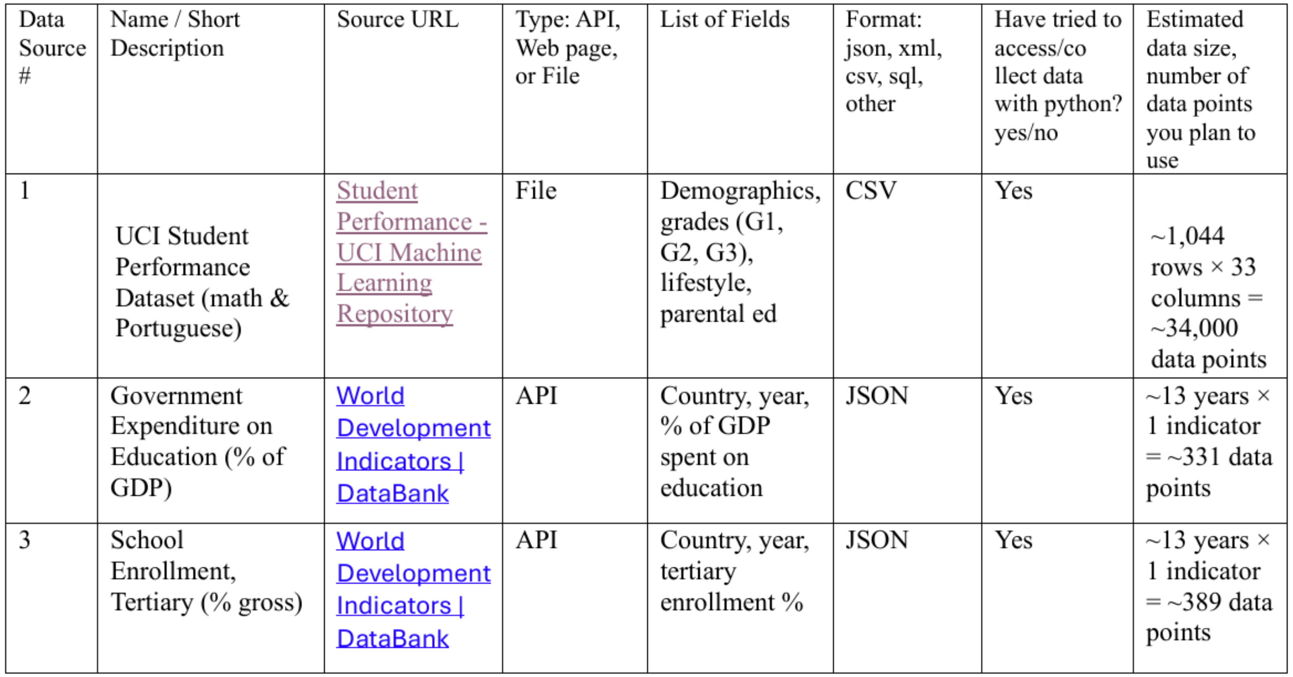
Data Sources

UCI Student Performance Dataset

- Math: 395 students | Portuguese: 649 students
- 30+ features: demographics, study habits, parental education, internet access
- Scores: G1, G2, G3 (0–20 scale)

World Bank API (30 European Countries)

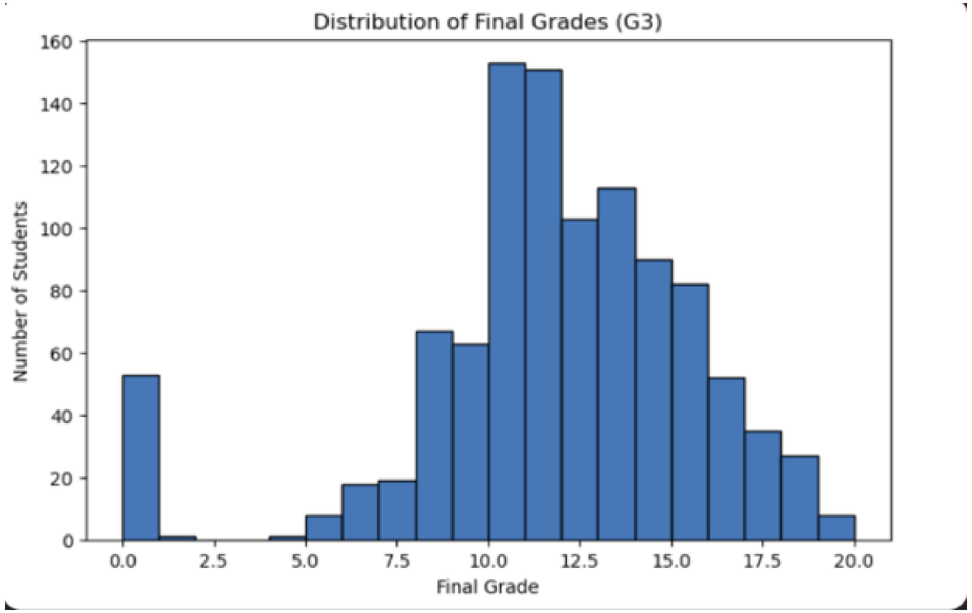
- Government education spending (% GDP)
- Tertiary enrollment (%)
- Used for adding macro-context



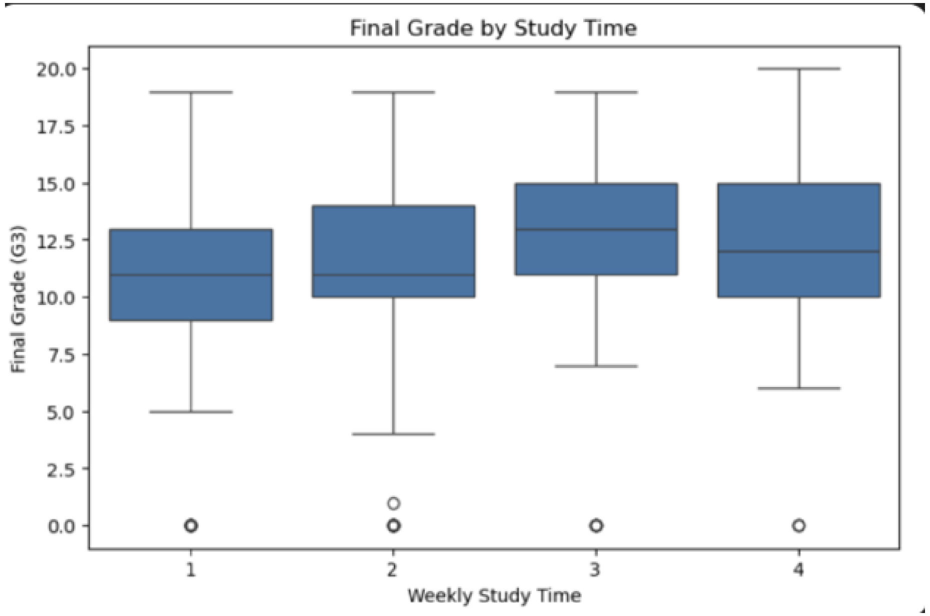
Data Source #	Name / Short Description	Source URL	Type: API, Web page, or File	List of Fields	Format: json, xml, csv, sql, other	Have tried to access/collect data with python? yes/no	Estimated data size, number of data points you plan to use
1	UCI Student Performance Dataset (math & Portuguese)	Student Performance - UCI Machine Learning Repository	File	Demographics, grades (G1, G2, G3), lifestyle, parental ed	CSV	Yes	~1,044 rows × 33 columns = ~34,000 data points
2	Government Expenditure on Education (% of GDP)	World Development Indicators | DataBank	API	Country, year, % of GDP spent on education	JSON	Yes	~13 years × 1 indicator = ~331 data points
3	School Enrollment, Tertiary (% gross)	World Development Indicators | DataBank	API	Country, year, tertiary enrollment %	JSON	Yes	~13 years × 1 indicator = ~389 data points

Student Performance Patterns – Analyzing Real Data

Histogram



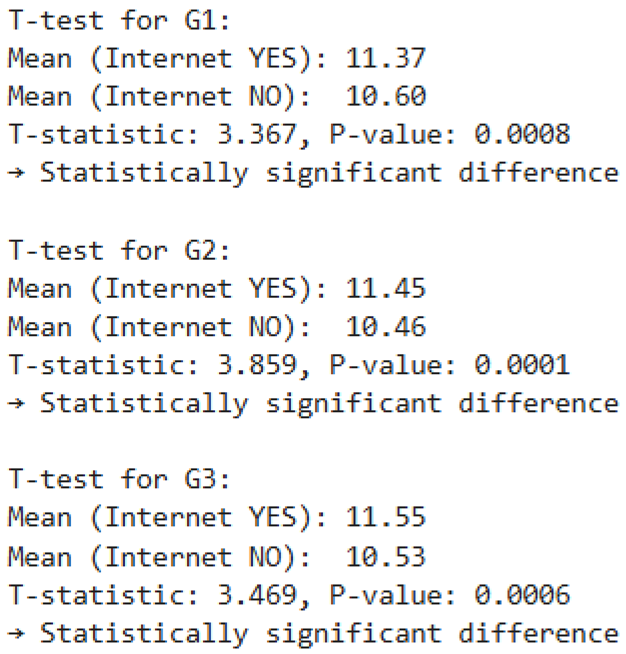
Box Plot



- Final grades cluster around 8-14
- Students studying more hours show higher median grades
- Study time is a meaningful predictor of academic outcomes

Internet Access & Statistical Testing – Analyzing Real Data

T-Test

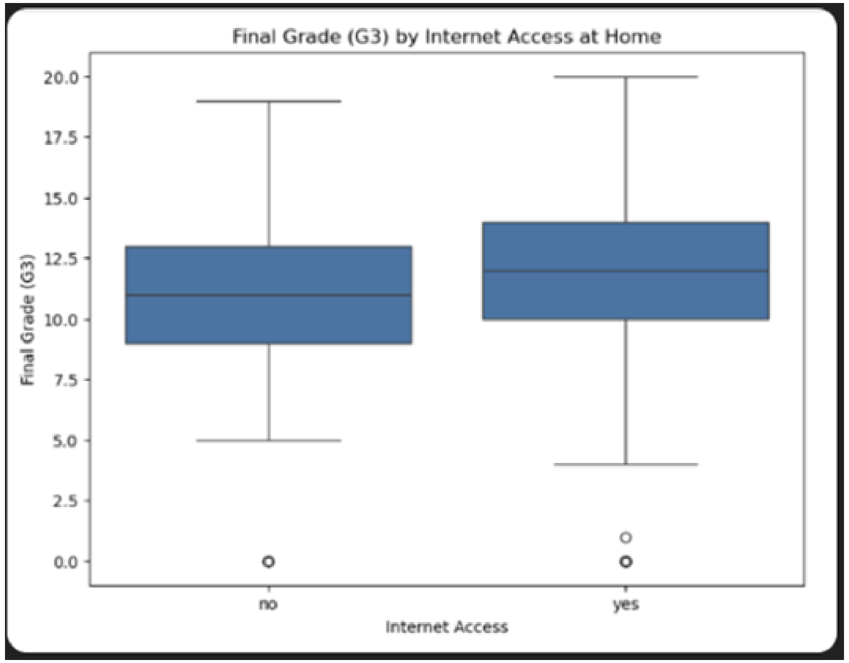


```
T-test for G1:
Mean (Internet YES): 11.37
Mean (Internet NO):  10.60
T-statistic: 3.367, P-value: 0.0008
→ Statistically significant difference

T-test for G2:
Mean (Internet YES): 11.45
Mean (Internet NO):  10.46
T-statistic: 3.859, P-value: 0.0001
→ Statistically significant difference

T-test for G3:
Mean (Internet YES): 11.55
Mean (Internet NO):  10.53
T-statistic: 3.469, P-value: 0.0006
→ Statistically significant difference
```

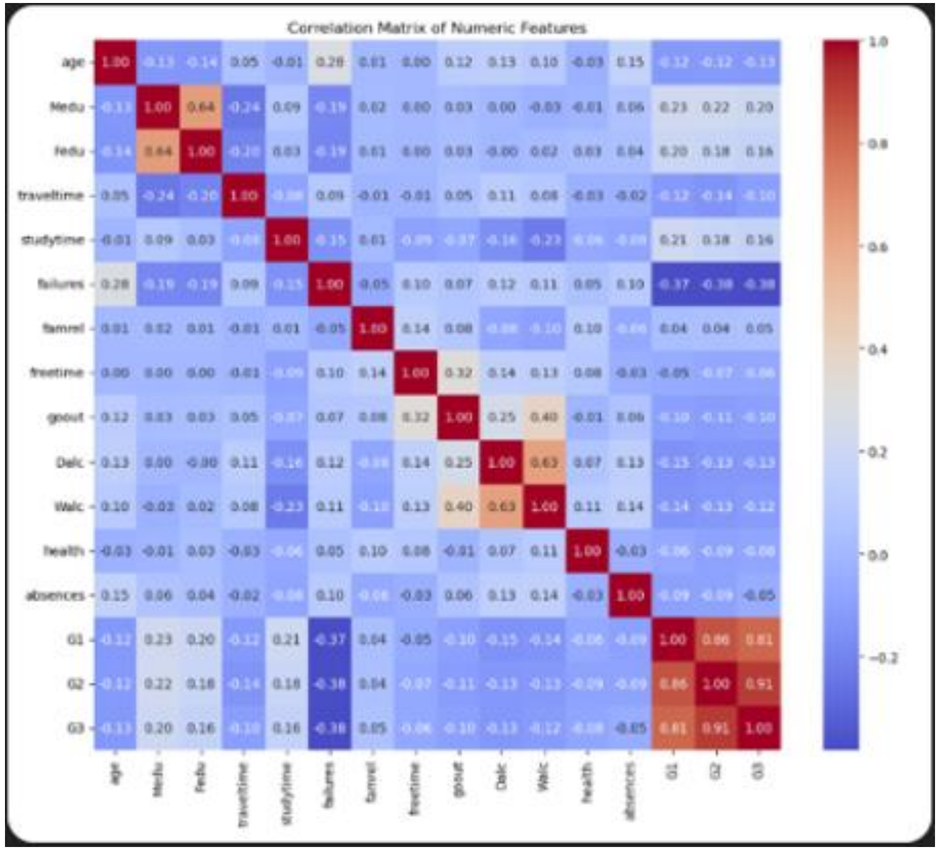
Box Plot



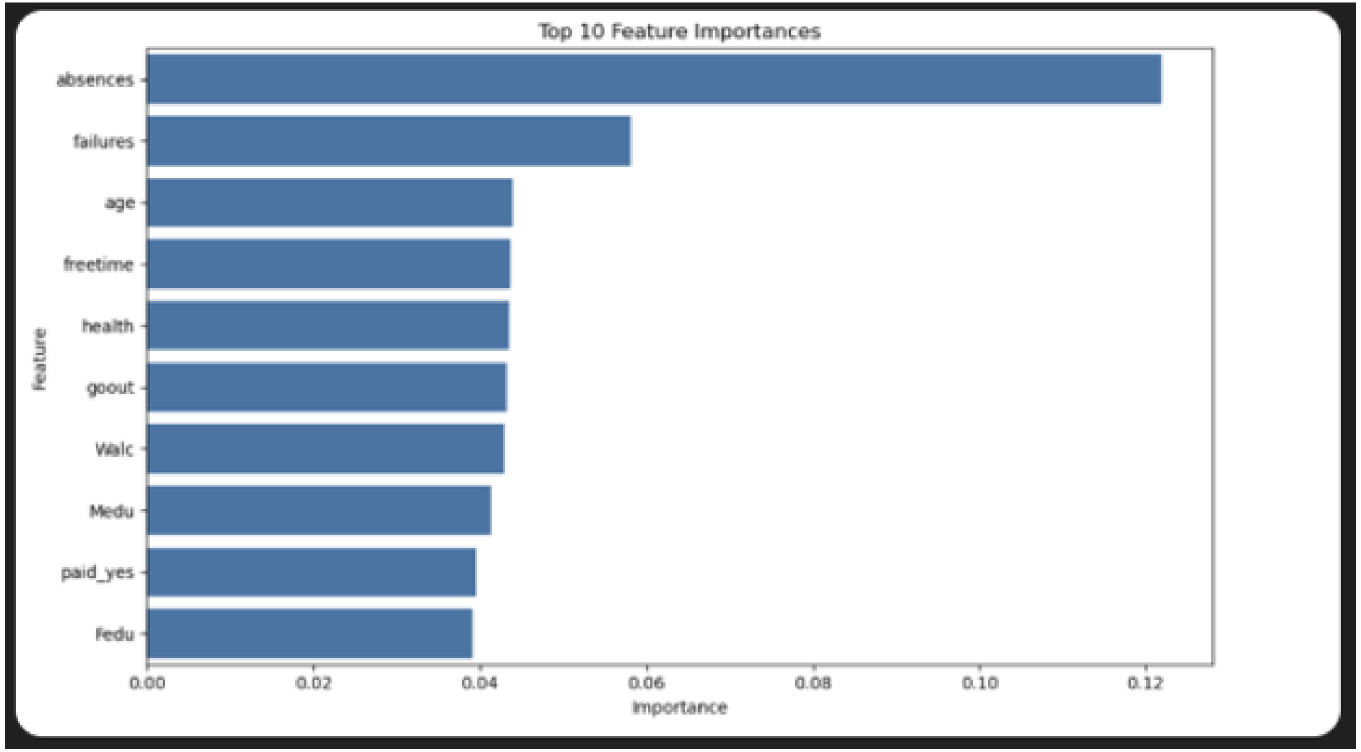
- Students with internet scored ~1 point higher on G1, G2, and G3.
- T-tests show that students with home internet consistently scored higher in G1–G3 ($p < 0.001$). Internet access is a statistically significant factor influencing student success.
- Internet access positively correlates with higher academic performance.

Correlations & Feature Importance

Correlation Matrix



Top Features Importance

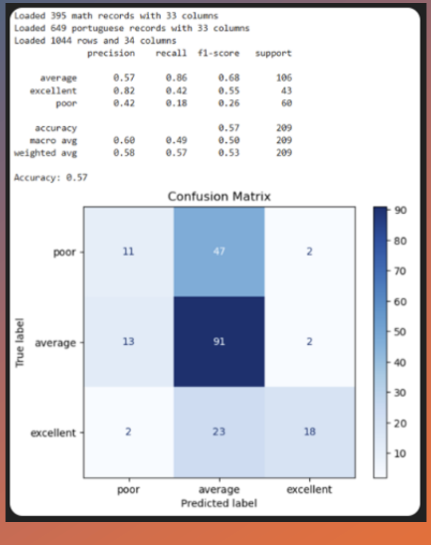


- Strongest correlations: **G1 ↔ G3**, **G2 ↔ G3**
- Past performance is the **best predictor** of final grades
- Random Forest top features: G2, G1, failures, study time, absences

ML Model Performance

- Random Forest 3-class classifier (poor / average / excellent)
- Accuracy: **57%**
- Best performance on “Average” class
- Most errors occur between “Average” ↔ “Excellent”

Model Output (Scikit-Learn)



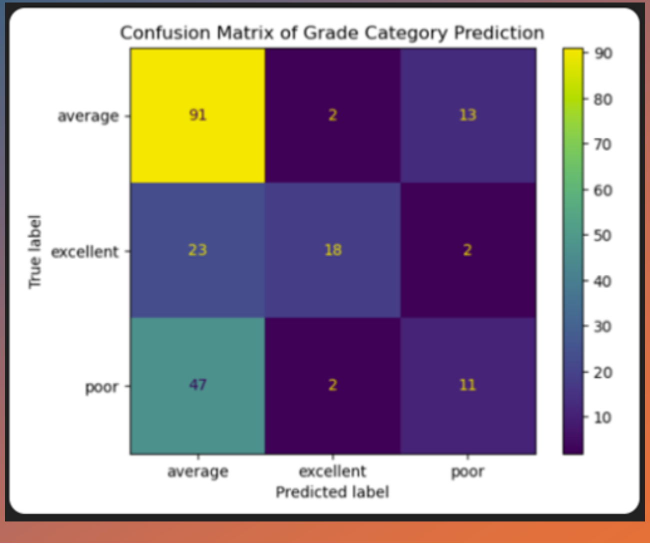
```
Loaded 395 math records with 33 columns
Loaded 649 portuguese records with 33 columns
Loaded 1044 rows and 34 columns
              precision    recall  f1-score   support

     average       0.57      0.86      0.68       106
   excellent       0.82      0.42      0.55        43
        poor       0.42      0.18      0.26        60

    accuracy                           0.57       209
   macro avg       0.60      0.49      0.50       209
weighted avg       0.58      0.57      0.53       209

Accuracy: 0.57
```

Clean Visualization (Custom Heatmap)



Project Challenges

- Merging Datasets
- Class Imbalance
- World Bank Data
- Preventing Future Leakage

Conclusion & Impact

01

Student performance is influenced most by **prior grades and study habits**.

02

Internet access shows a significant performance difference → insight for **digital equity policies**.

03

Machine Learning models can help educators identify at-risk students early.

04

Future Work: Add more countries' macro-data, test alternative ML models (XGBoost, SVM).

Thank you!

Q&A