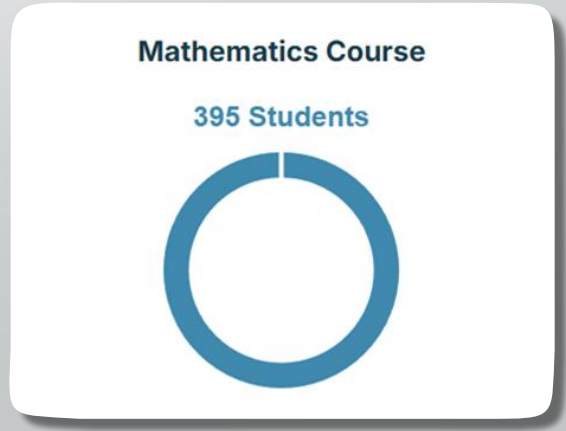
Predicting Student Performance: A Data-Driven Approach

Capstone Project – AI Tech Institute

Kresna Laksono

Introduction & Objective

- Context: High student failure rates in education are a significant concern, particularly in core subjects. Data Mining offers tools to understand and predict student achievement.
- Our Objective: To predict students' final grades (G3) using various demographic, social, and school-related attributes see which provides better insights when prior academic performance is excluded (G1 & G2 represent prior period grades).
  - This project aim to mimic the predictive process outlined by Cortez & Silva (2008).
  - Model Comparison: Will compare the performance of a Linear Regression model against a more complex Random Forest Regressor.

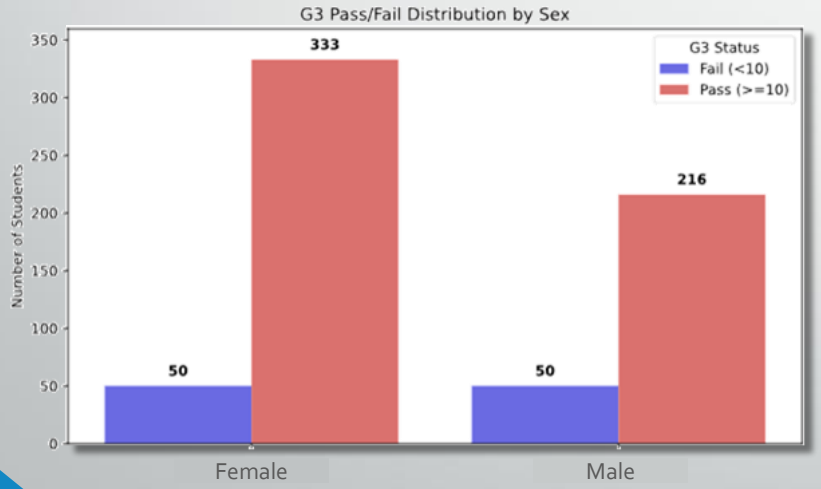


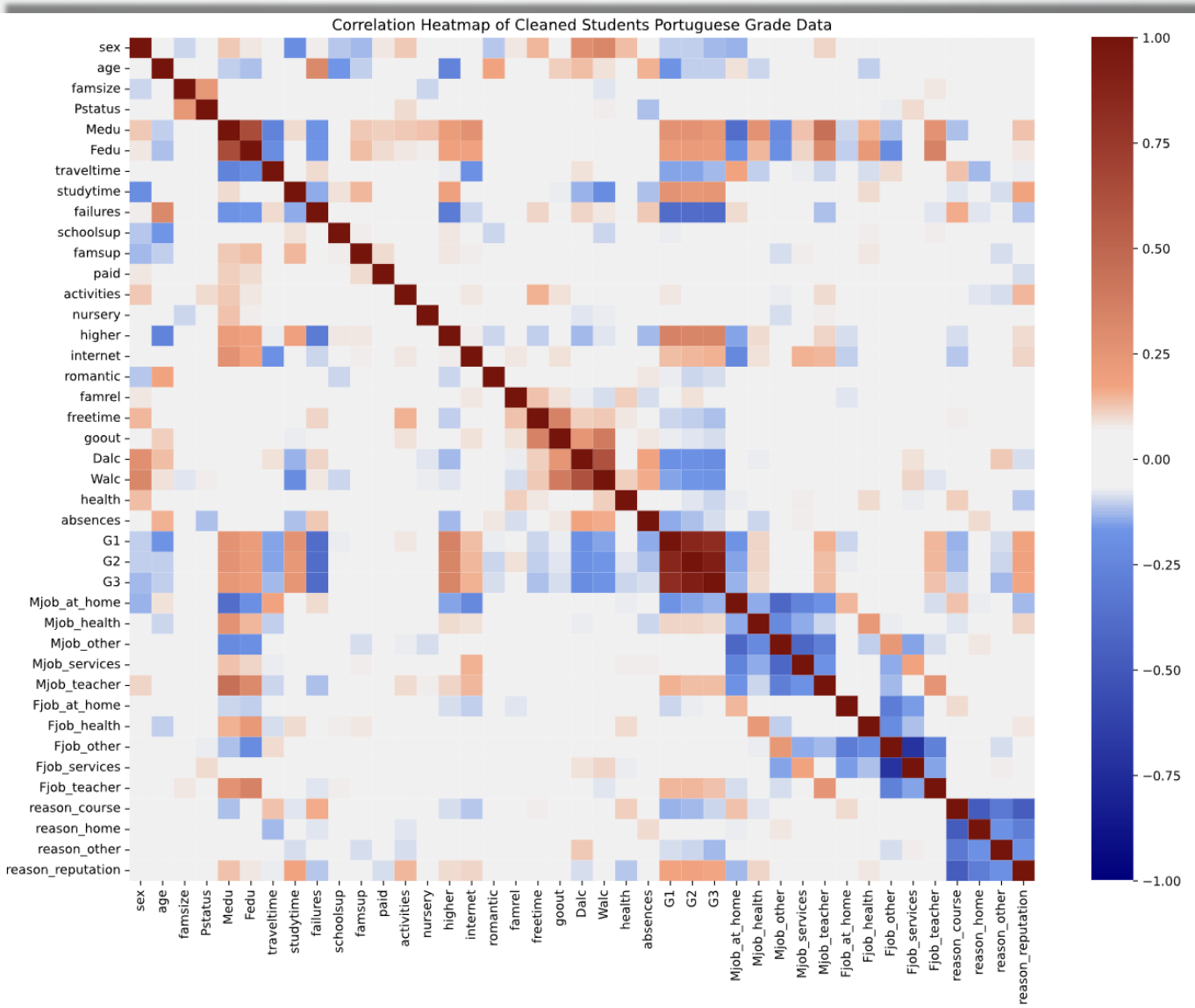
Data Preparation & Machine Learning Workflow

- Data Source: Real-world student data from Portuguese secondary schools.
  - *30+ attributes – 11 ordinal integers – 2 general integers – 9 binary strings – 8 categorical strings*
- Key Data Preparation Steps:
  - Cleaning: Handling missing values, standardizing text entries.
  - Feature Engineering: Converting all categorical data (e.g., gender, family size, parents' jobs, school reason) into numerical formats (binary or one-hot encoding).
  - Target Variable: G3 (final grade, 0-20) as our continuous prediction target.
- Simplified Workflow:

Descriptive Data

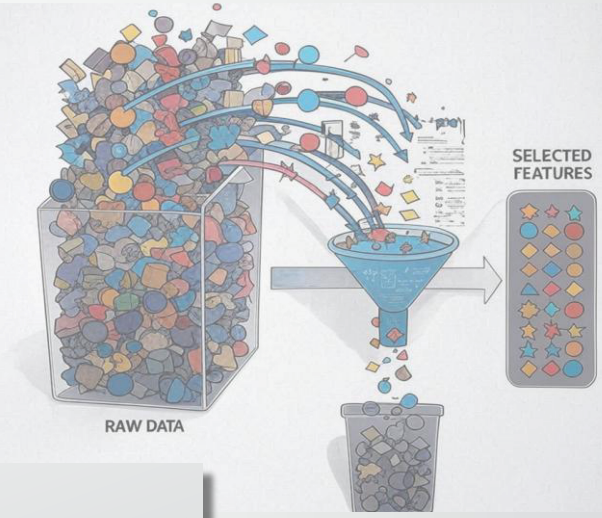
- G1 & G2 have a very high correlation could indicate multicollinearity
- ~ 30% fail portion that indicate concerning rate of students dropping out





Model Performance: Linear Regression

- Model: Linear Regression (a simple, interpretable model)
- Impact of "Feature Selection" == focusing on a smaller subset of the most relevant features led to a significant improvement in performance.



Before Selection

17.5%

R-squared (R^2)

Explained very little of the grade variation.

After Selection

29.8%

R-squared (R^2)

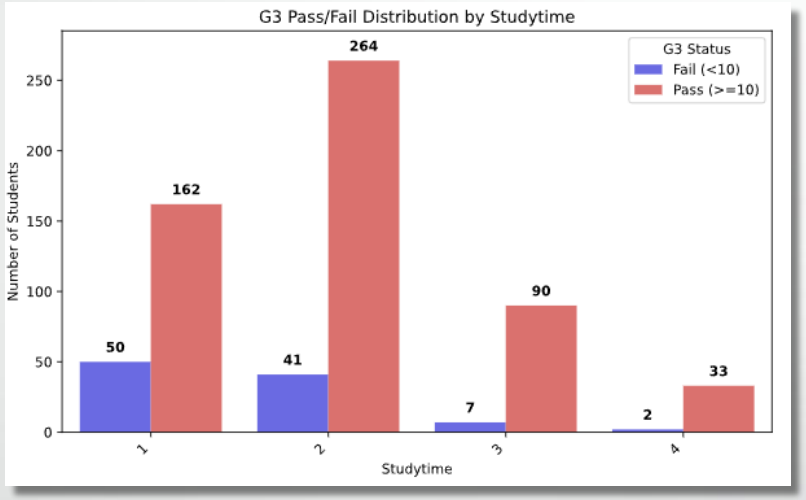
A noticeable improvement in explanatory power.

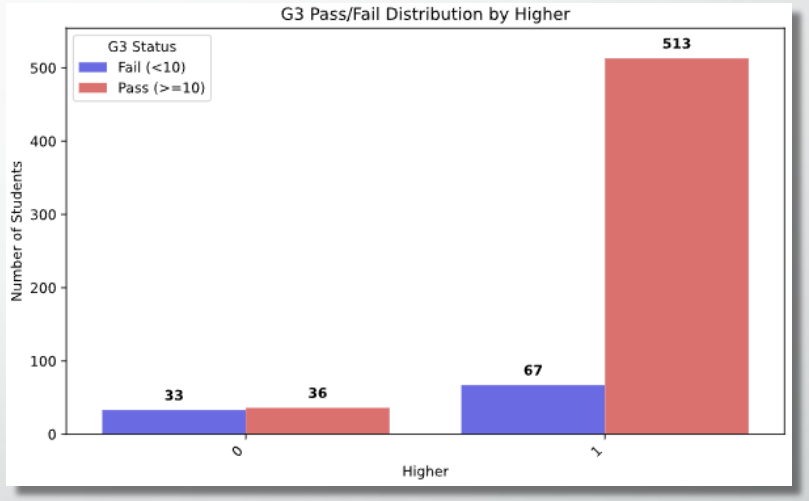
Model Performance: Random Forest Regressor

- Model: Random Forest Regressor (a more complex, ensemble tree-based model)
- Performance after Feature Selection:
  - RMSE: 2.6895 (average error of ~2.7 grade points)
- Interpretation:
  - Surprisingly, the Random Forest Regressor did not significantly outperform the tuned Linear Regression model based on R2. In fact, it performed slightly worse than the best Linear Regression model with selected features.
  - This indicates that even a more powerful, non-linear model struggles to explain the variance in G3 when direct, highly correlated predictors like G1 and G2 are explicitly excluded from the feature set. The underlying relationships might be more complex or require different features.

Conclusion & Future Recommendations

- **Conclusion**:
  - Predicting student final grades (G3) solely based on demographic, social, and school-related features (excluding prior grades G1, G2) proves challenging, yielding relatively low R-squared values (below 30%).
  - Both Linear Regression and Random Forest Regressor, despite feature selection, show limited explanatory power in this specific setup.





- **Future Recommendations**:
  - Advanced Feature Engineering: Consider creating interaction terms (e.g., studytime combined with failures) or polynomial features to capture more complex non-linear relationships.
  - Include Prior Grades (G1 & G2): The original Cortez & Silva paper highlights G1 and G2 as the most significant predictors. For substantially higher predictive R2 score, these features have to be included. This would likely drastically improve model performance albeit with further multicollinearity investigation.
  - Explore Other Models: Investigate other robust models for instance Decision Tree, Neural Network or SVM.

End of Slide

Reference:

- Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. https://repositorium.sdum.uminho.pt//bitstream/1822/8024/1/student.pdf

Source data:

- Portuguese Language CSV - UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/320/student+performance