group memebers

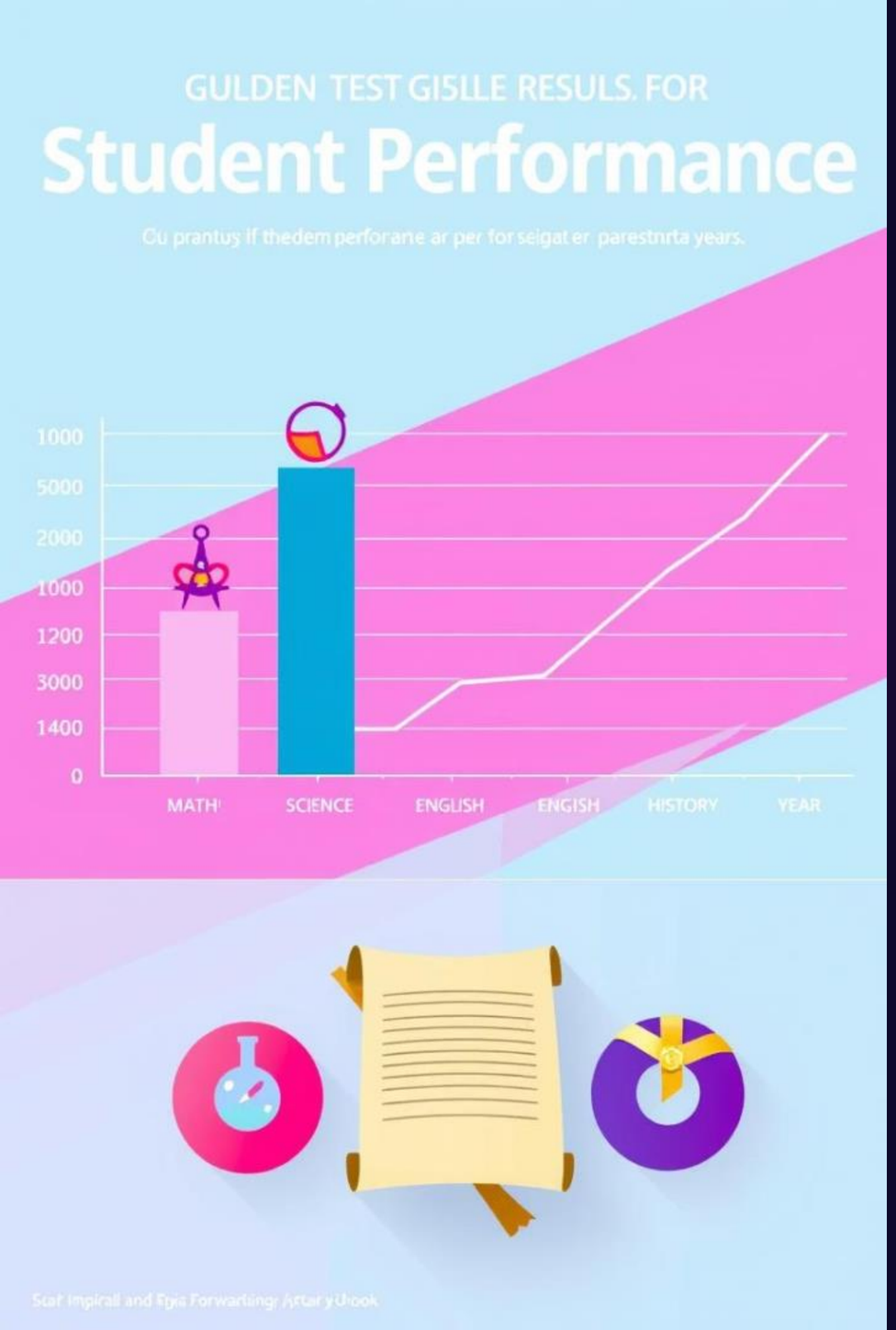
1. Furtuna G/Slassie mit/ur/046/12
2. Betelhem Haftu mit/ur/159/12
3. Meles Mengesha mit/ur/ 298/12
4. Ayda Angesom mit/ur/ 158/12
5. Yorkabel Ngatu mit/ur/190/12
6. Meles H/Slassie mit/ur/101/12
7. Makda Haile mit/ur/231/12
8. Yemane G/Michael mit/ur/273/12
9. Michiale Adhanom mit/ur/166/12

Regression Report: Predicting Student Final Grades (G3)

This presentation details a regression project aimed at predicting the final grades of Portuguese secondary school students. Using the Student Performance Dataset, we apply a linear regression model to understand and anticipate outcomes based on demographic, social, and academic factors. This approach helps educators identify students who may need additional support early in the academic year.

We focus on the target variable G3 (final grade), exploring data characteristics, preprocessing techniques, model training, and evaluation results. Future work directives are provided to further improve predictive accuracy.

M **by Meles Haileselassie**



Introduction: Project Overview and Goals

Project Aim

Predict final grade (G3) using diverse student data including social and academic attributes.

Importance

Facilitates early identification of students at risk for poor outcomes to enhance intervention strategies.

Scope

Involves data from Portuguese secondary school students with multiple features influencing academic success.

Data Exploration and Characteristics

Dataset Details

- Source: UCI Machine Learning Repository
- 649 student records
- 32 features: categorical & numerical
- No missing values detected

Visual Insights

- Histograms reveal distribution of numeric attributes
- Strong positive correlation among G1, G2, and G3 grades
- Outliers noted in variables such as absences

Methodology: Preprocessing and Data Splitting

Data Encoding & Scaling

OneHotEncoder applied to categorical variables, dropping the first to prevent multicollinearity; numerical features standardized using StandardScaler.

Train-Test Split

Dataset split into 80% training and 20% testing subsets, with a fixed random seed to ensure result reproducibility during evaluation.

Target Variable

The final grade (G3), a continuous numeric variable, is the prediction target for regression modeling.

Model Training: Linear Regression Application

Model Choice

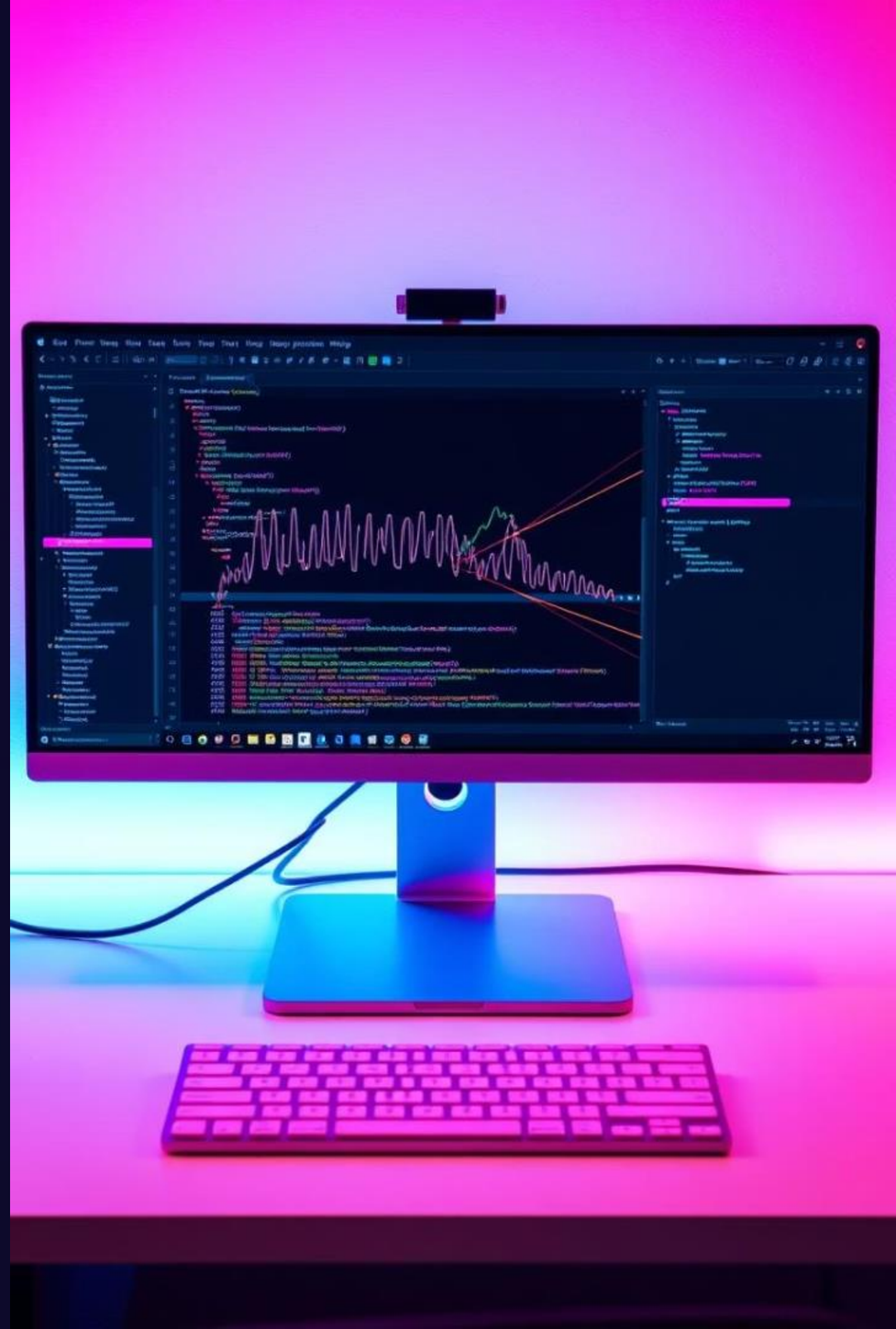
Baseline linear regression model implemented using scikit-learn for interpretability and simplicity.

Pipeline Integration

Preprocessing steps and model training combined into a single pipeline to streamline the workflow and enhance reproducibility.

Hyperparameters

Default settings used for baseline; represents a starting point before tuning more complex parameters.



Results and Model Evaluation

Evaluation Metrics

- RMSE: ~1.21
- MAE: ~0.77
- R^2: ~0.85

Model Interpretation

A strong R^2 score indicates ~81% variance explained. Average prediction errors of ~1.6 points are reasonable on a 0–20 grading scale.

Residual analysis reveals mostly linear patterns with some underestimation for high-performing students, common for simple linear models.

Conclusion, Challenges, and Future Work

1

Summary of Findings

Model effectively captures linear relationships; previous term grades G1 and G2 are strong predictors of final performance.

2

Challenges

High number of categorical variables requires robust encoding, and some features contribute little, complicating the model.

3

Future Research

- Explore nonlinear models like Random Forest and Gradient Boosting
- Implement feature selection or dimensionality reduction techniques
- Adopt cross-validation and hyperparameter tuning with GridSearchCV