Predicting Student Academic Performance from Demographic and Behavioural Data

Shrijaa Venkatasubramanian Subashini, Rahiq Raees, Christine Chow & Jiro Amato

2025-12-04

Table of contents

Summary

We built a linear regression model utilizing Ridge Regression to predict a student's final grade (G3) in a Portuguese language course using demographic data, behavioural insights, and previous assessment scores. Our final model achieved strong performance on unseen test data, with an R^2 score of 0.857 and a Mean Absolute Error (MAE) of 0.776. On average, model predictions deviated from actual grades by less than one grade point on the 0-20 scale, indicating a high level of accuracy. Although the model demonstrated reliability in predicting grades within the passing range (scores exceeding 10), residual analysis revealed reduced precision for students with lower scores or those receiving a zero, frequently overestimating their performance. Given this limitation concerning failing students, we conclude that this model showcases adequate performance to serve as a decision-support tool for educators to project final grades midway through the term.

Introduction

Student academic success is influenced by a complex combination of personal, social, and educational factors. Understanding the determinants of performance and being able to reliably predict student outcomes has important implications for supporting learning, designing interventions, and improving educational strategies (Ma et al. 2000; Pritchard and Wilson 2003). Despite decades of pedagogical research, identifying at-risk students early remains challenging because performance is shaped by many interacting variables such as family background, study habits, school engagement, and socio-economic conditions.

Here we ask whether a machine learning algorithm can predict a student's final grade based on their demographic attributes, family characteristics, school-related behaviours, and past academic performance. Answering this question is valuable because traditional methods of assessing student progress often rely on subjective teacher evaluations or mid-course assessments, which may miss early warning signs or overlook external factors influencing learning (Johora et al. 2025). If a machine learning model can accurately and consistently predict final grades, this could help educators identify students in need of additional support earlier, design more targeted interventions, and ultimately contribute to improved educational outcomes.

Methods

Data

The data set used in this project is the Student Performance dataset created by Paulo Cortez from the University of Minho, Portugal (Cortez and Silva 2008). It was sourced from the UCI Machine Learning Repository (Cortez 2008) and can be found here. The dataset contains information on 649 students from two Portuguese secondary schools, with data collected through school reports and questionnaires. Each row represents a student with 32 features including demographic information (age, sex, family size), educational background (parental education, past failures, study time), social factors (going out, romantic relationships, alcohol consumption), and school-related features (absences, extra support, desire for higher education). The dataset also includes grades from the first period (G1), second period (G2), and final grade (G3), with G3 serving as the target variable for prediction.

Analysis

A linear regression model with Ridge regularization was used to predict the final grade (G3). All available features from the dataset were utilized, including student demographics, family background, and prior assessment scores (G1 and G2). Data was split with 70% into the training set and 30% into the test set. The hyperparameter alpha was tuned using 10-fold cross-validation with Randomized Search, selecting Mean Absolute Error (MAE) as the evaluation metric to prioritize interpretability and robustness against outliers. Preprocessing involved standardizing numeric features—applying Robust Scaler specifically to skewed distributions like absences—and transforming categorical variables via One-Hot Encoding. The Python programming language (Van Rossum and Drake 2009) and the following Python packages were used to perform the analysis: Pandas (McKinney 2010), NumPy (Harris et al. 2020), Altair (VanderPlas 2018), and scikit-learn (Pedregosa et al. 2011). The code used to perform the analysis and create this report can be found here: https://github.com/jiroamato/student_grade_predictor.

Results & Discussion

To understand the predictive power of our features, we examined correlations between all predictors and the target variable (Figure 1). Prior assessment grades (G1 and G2) showed the strongest correlation with final grades, which aligns with educational research suggesting past performance is the most reliable predictor of future outcomes. Other notable correlations included past failures (failures), study time (studytime) and parental education levels (Fedu and Medu), though these correlations were considerably smaller.

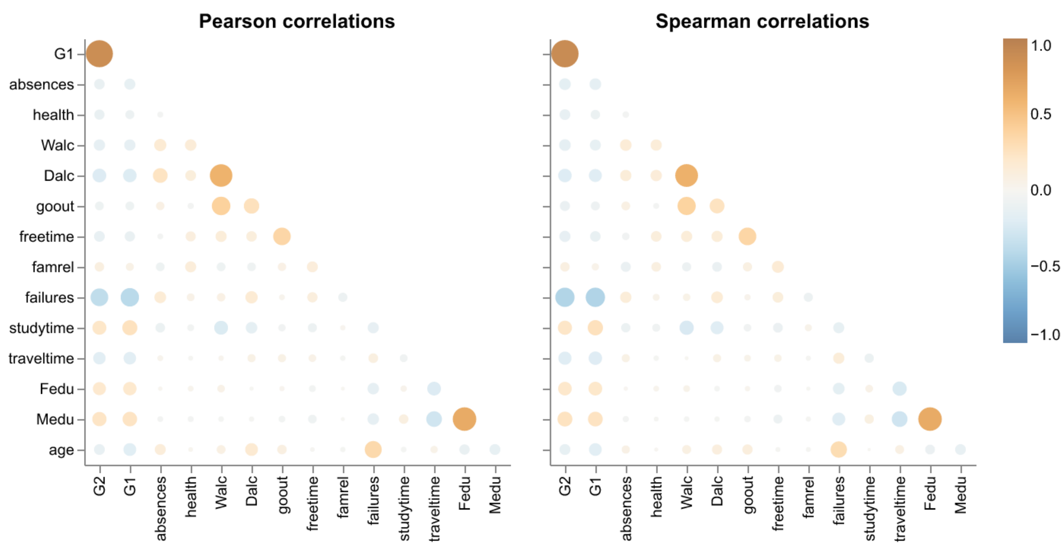

Figure 1: Pairwise correlations between all features in the dataset.

We also examined the distribution of our target variable to understand the grade landscape (Figure 2). The distribution showed that 83% of students achieved passing grades (10 or above), while 17% fell below the passing threshold. This imbalance is important context for interpreting model performance, particularly regarding predictions for lower-performing students.

To optimize model performance, we tuned the regularization parameter alpha using 10-fold cross-validation (Figure 3). The optimal alpha value was found to be 38.203, which balanced

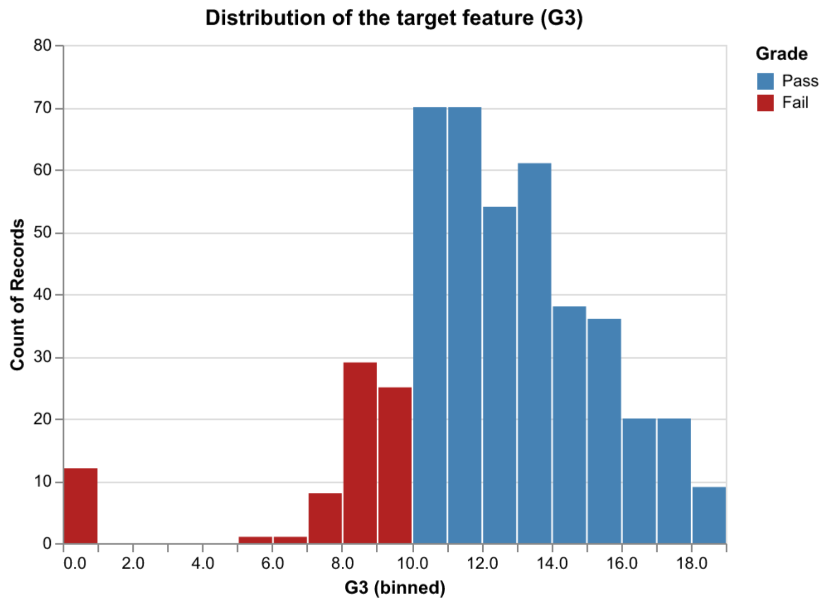

Figure 2: Distribution of the target variable (G3) showing final grade frequencies.

the trade-off between model complexity and generalization ability while yielding the minimum cross-validation MAE of 0.825.

Our prediction model performed well on test data, achieving a final R^2 of 0.857 and MAE of 0.776 (Table 1). The R^2 indicates that our model explains approximately 86% of the variance in student final grades, while the MAE confirms predictions are accurate to within less than one grade point on average.

Table 1: Model performance metrics on test data.

MAE	RMSE	R2
0.776	1.207	0.857

To better understand prediction accuracy across the grade spectrum, we examined residuals (prediction errors) against predicted values (Figure 4). The model performs well for students predicted to score 8 or above, with errors distributed evenly around zero. However, for students with lower actual grades, particularly those receiving zeros, the model consistently overestimates performance. This pattern suggests that the factors leading to very poor academic outcomes may not be fully captured by the features in our dataset.

Examination of model coefficients revealed that G2 and G1 (second and first period grades) were by far the strongest predictors of final grades (Table 2). This finding is intuitive—students who perform well early in the term tend to continue performing well. Among other features, past failures (failures) contributed the most negatively to predicted grades, while the reason for choosing the school because of its courses (reason_course) and choosing the school for other reasons (reason_other) showed modest positive and negative effects respectively.

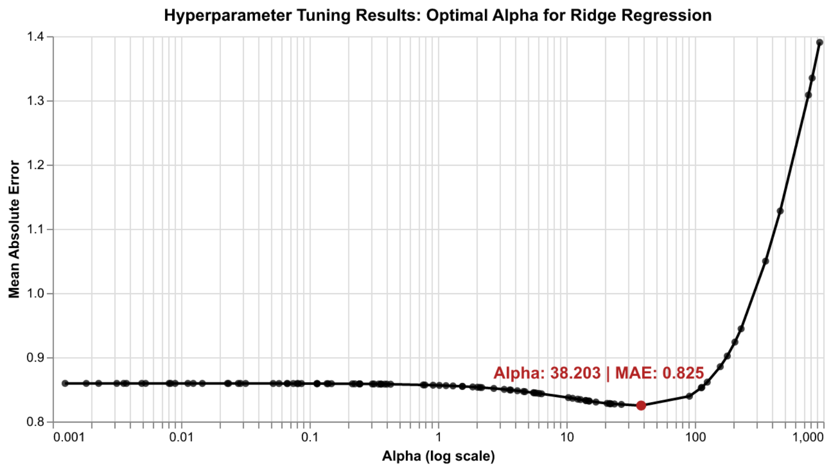

Figure 3: Hyperparameter Tuning of Alpha.

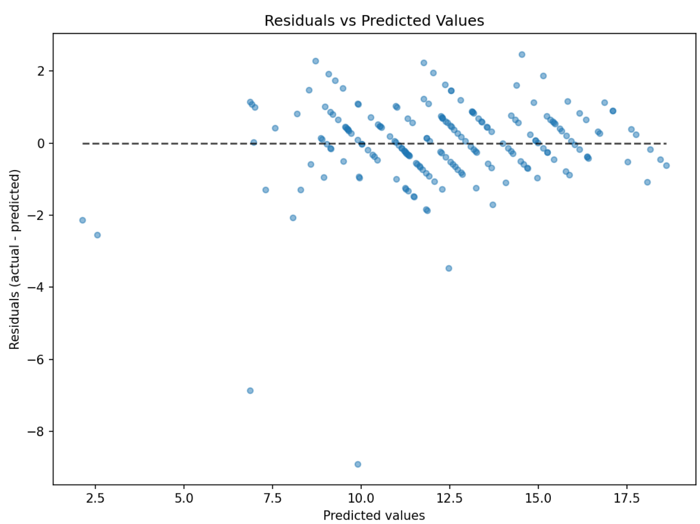

Figure 4: Residual analysis showing prediction errors across the range of predicted values.

Table 2: Top 5 model coefficients by magnitude.

Feature	Coefficient
G2	2.046
G1	0.705
failures	-0.232
reason_course	0.231
reason_other	-0.211

The strong performance of this model suggests it could serve as a useful decision-support tool for educators. By inputting early assessment scores and student characteristics, teachers could identify students at risk of poor final outcomes and implement targeted interventions. However, the model's reduced accuracy for failing students represents a significant limitation. Students most in need of early intervention are precisely those for whom predictions are least reliable.

Several directions could be explored to improve the model further. First, additional features capturing engagement metrics (such as assignment completion rates or class participation) might help predict performance for at-risk students. Second, a model excluding G1 and G2 features could be valuable for even earlier prediction—before any formal assessments occur—though this would likely come at the cost of accuracy. Finally, exploring non-linear models such as random forests or gradient boosting might capture complex interactions between features that linear regression cannot represent.

References

Cortez, Paulo. 2008. "Student Performance." UCI Machine Learning Repository.

Cortez, Paulo, and Alice Silva. 2008. "Using Data Mining to Predict Secondary School Student Performance." In *Proceedings of 5th Future Business Technology Conference*, 5–12. EUROSIS. https://api.semanticscholar.org/CorpusID:16621299.

Harris, Charles R, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Johora, Fatema Tuz, Md Nahid Hasan, Aditya Rajbongshi, Md Ashrafuzzaman, and Farzana Akter. 2025. "An Explainable AI-Based Approach for Predicting Undergraduate Students Academic Performance." *Array* 26: 100384. https://doi.org/https://doi.org/10.1016/j.array.2025.100384.

Ma, Yiming, Bing Liu, Ching Kian Wong, Philip S. Yu, and Shuik Ming Lee. 2000. "Targeting the Right Students Using Data Mining." In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 457–64. KDD '00. New York, NY, USA: ACM. https://doi.org/10.1145/347090.347184.

McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, =51–56.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30.

Pritchard, Mary E., and Gregory S. Wilson. 2003. “Using Emotional and Social Factors to Predict Student Success.” *Journal of College Student Development* 44 (1): 18–28. https://doi.org/10.1353/csd.2003.0008.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

VanderPlas, Jake. 2018. “Altair: Interactive Statistical Visualizations for Python.” *Journal of Open Source Software* 3 (7825, 32): 1057. https://doi.org/10.21105/joss.01057.