1. Introduction

This project investigates the research question: ***Which socioeconomic and behavioral factors are most important for predicting student performance in Math, and how consistently do different machine learning models identify these key predictors?*** Academic achievement reflects a complex combination of family background, access to resources, daily habits, and personal circumstances. Prior research in educational data mining shows that parental education, study behaviors, school support, and socioeconomic factors meaningfully influence student outcomes (Hussain et al., 2021; Kesgin et al., 2025). Mathematics performance, in particular, is shaped by a complex interaction of several different variables. In our experience, success in math depends not only on cognitive ability but also on factors such as consistent practice, attention to detail, receptiveness to how the material is taught, familiarity with the types of exam questions, and even a student's emotional relationship with the subject: whether feel confident in it or experience anxiety around it. By applying four machine learning methods: Penalized Regression, Support Vector Regression (SVR), Gradient Boosting, and a Neural Network, this project examines whether student performance can be predicted from non-grade variables and evaluates which factors consistently emerge as the strongest predictors. These insights may help educators identify vulnerable students, allocate resources more effectively, and support equitable academic success.

2. Methods

The data used for this project come from the UCI Machine Learning Repository's *Student Performance* dataset, originally collected from two Portuguese secondary schools. The dataset includes 649 students and 33 variables encompassing demographic, socioeconomic, and behavioral information, as well as grade outcomes across two periods: G1, G2, and the final grade G3. Participants range in age from 15 to 22 ($M \approx 16.7$), with a roughly even gender distribution. The attributes provide a rich description of family structure, parental education, school support programs, study habits, health indicators, alcohol use, and school engagement, offering meaningful context for understanding how non-academic factors may influence academic performance.

During the data cleaning stage, we used a version of the dataset from Kaggle because the original UCI files used semicolon separators that created complexities and issues with reading in the file. After verifying data types and confirming the absence of missing values, we conducted exploratory data analysis (EDA). Distribution plots and a numeric correlation matrix revealed that G1 and G2, earlier period grades, were the strongest predictors of G3, which is expected since they are components of the final grade. Since the purpose of our project is to examine **non-academic** predictors, we graphed distributions of various features with G3 to understand their relationship.

By embedding all preprocessing steps, such as one-hot encoding of categorical variables and scaling where applicable, inside the cross-validation loop, each fold learned transformations using only its training portion, ensuring that no information from the validation or test data leaked into the fitted model. Model performance was assessed using R^2, MSE, RMSE, and MAE.

For penalized regression (Ridge, Lasso, and Elastic Net), we used **5-fold cross-validation** via GridSearchCV, tuning regularization strength and mixing parameters, with Elastic Net ultimately performing best. The Support Vector Regression model also used **5-fold cross-validation with GridSearchCV**. The best model used an RBF kernel with C = 1 and gamma = "auto," allowing nonlinear modeling. Gradient Boosting was tuned using **RandomizedSearchCV with 5-fold cross-validation**, exploring hyperparameters such as number of estimators, learning rate, maximum depth, and subsample rate; this model achieved the strongest predictive performance. Finally, the neural network was implemented in PyTorch using nn.Module, and evaluated with a **manual 5-fold cross-validation loop (since we used Pytorch)**, tuning hidden layer sizes and learning rates, and retraining the network from scratch in each fold to ensure no leakage of learned weights across splits.

3. Results

Across the four machine learning models, Gradient Boosting demonstrated the highest cross-validated R^2 and the lowest prediction error, outperforming other models. On the held-out test set, the Gradient Boosting model produced a slightly higher R^2, which indicates that the model explains more variance in student performance on unseen data. However, the test MSE, RMSE, and MAE were higher than in cross-validation, which means the model's individual predictions are less accurate and the errors are larger on new data.

Model	**Best Hyperparameters**	**CV R^2**
Elastic Net	alpha = 0.4281 (with optimal 1l_ratio)	0.107
Support Vector Regression	C=1, gamma="auto", kernel = "rbf"	0.106
Gradient Boosting Regressor	N_estimators =452, learning_rate=0.00655, max_depth=2, subsample=0.6	0.227
Neural Network	hidden_dim=32, learning_rate=0.01	-0.047

Beyond overall performance, the models provide insight into which socioeconomic and behavioral features are most useful for predicting final math grades, and how consistently these predictors appear across different learning algorithms. Across all four models, school support *(schoolsup_yes)* repeatedly emerged as one of the strongest predictors. In every method, students who receive additional school support tend to have lower grades, a pattern that perhaps reflects selection effects rather than causal harm, as such support is often provided to students who are already struggling academically. Another consistently strong predictor was romantic relationship status *(romantic_yes)*, which appeared as the single-largest predictor in Gradient Boosting and remained one of the most important factors in the test-set permutation analysis. This suggests that students in romantic relationships may experience competing demands on their time or attention, lowering academic performance.

Moderately predictive features included parental education, extracurricular participation, and certain school-choice motivations. Other features appeared important during model training but did not generalize well when tested. For example, Gradient Boosting assigned high importance to internet access, nursery attendance, and several parental occupation variables, yet permutation tests on the held-out set revealed that these features either contributed little or even slightly degraded predictive accuracy. These discrepancies suggest that some signals learned by the model were instance-specific patterns or nonlinear interactions that did not translate to new data, highlighting mild overfitting in the ensemble method.Overall, the model comparisons reveal that only a subset of socioeconomic and behavioral variables consistently contributed meaningful predictive signals across algorithms. School support and romantic relationships emerged as dominant factors, while parental education, extracurricular involvement, and some home-environment indicators also showed moderate but less stable influence. In contrast, many demographic and lifestyle variables provided minimal or inconsistent value, and several features identified as important during model training did not generalize to the test set.

4. Discussion

Importantly, the limited generalizability of several training-time predictors underscores how easily models can latch onto patterns that do not hold beyond the training sample, highlighting the value of test-set permutation importance in distinguishing meaningful predictors from data of model fitting.

Performance differences across the four modeling approaches further illuminate the structure of the data. Gradient Boosting achieved the strongest performance and the only positive R^2, likely because it can flexibly capture nonlinear interactions and complex threshold effects present in behavioral and socioeconomic variables. In contrast, Elastic Net struggled because linear models cannot represent the nonlinear relationships and interaction patterns inherent in student

behavior data, resulting in negative R^2 values. SVR performed moderately well but was sensitive to the choice of hyperparameters and less stable across folds, while the neural network underperformed due to the small dataset, which limited its ability to generalize complex patterns without overfitting. These challenges point to several limitations of the project: the dataset is small and tabular, restricting the expressive power of higher-capacity models; key academic predictors (G1 and G2) were intentionally excluded to focus on contextual factors; and the models were constrained by noisy, highly collinear variables that may mask true causal relationships. Despite these constraints, the project demonstrates that a small number of contextual features provide meaningful predictive value and that ensemble methods such as Gradient Boosting are best suited for capturing the nuanced relationships that characterize this educational performance data.