Introduction

Predicting student performance is a vital area of research in educational data mining, offering actionable insights for educators and policymakers to foster academic success. This project explores multiple machine learning techniques, including Decision Trees, Ridge and Lasso Regression, and Gradient Boosting, to model students' grades based on diverse predictors such as parental education, study habits, extracurricular involvement, and school support.

Decision Trees

Decision Trees were employed for their interpretability and ability to handle nonlinear interactions. Evaluated with Mean Squared Error (MSE) and R^2 scores, a depth of three yielded the best performance by balancing predictive accuracy and complexity, highlighting the importance of avoiding overfitting through regularization strategies.

Ridge and Lasso Regression

Ridge and Lasso Regression addressed challenges of multicollinearity and feature selection. Ridge Regression's L2 penalty stabilized the model by shrinking coefficients, while Lasso's L1 penalty effectively performed feature selection by setting coefficients of irrelevant features to zero. These techniques demonstrated moderate predictive capabilities, with Lasso showing a slight edge by isolating significant predictors.

Gradient Boosting Regression

Gradient Boosting Regression captured complex relationships through iterative refinement. Experimentation with various learning rates revealed a marginal improvement over linear models, though the relatively low R^2 values across all models emphasized the need for richer datasets and advanced features to capture the complexities influencing academic performance.

Literature Review

A literature review further contextualizes the methodology, drawing from Martins et al. (2021) and Goyal & Kumar (2023), who also explored machine learning techniques in educational settings. These studies underscore the importance of addressing data imbalances, with methods like SMOTE, and emphasize the efficacy of ensemble methods such as Gradient Boosting in similar domains.

This project not only investigates the predictive capabilities of established algorithms but also highlights their limitations and potential enhancements. By integrating these methods, this study aims to contribute to the growing field of educational analytics, paving the way for more nuanced and impactful applications in student performance prediction.