PROJECT E8
Marek Kalda
Nele-Liis Võhma

Two Paths to Predicting Success

UNIVERSITY OF TARTU
Institute of Computer Science

01. Introduction

Academic performance is influenced by various factors, such as study hours, sleep hours, parental involvement etc. Our Kaggle dataset is synthetic and includes information about 6607 students and how their attributes as social, behavioral and environmental factors affect students performance. This project explores two paths to analyse factors influencing student grades and compare which method offers the most practical insights.

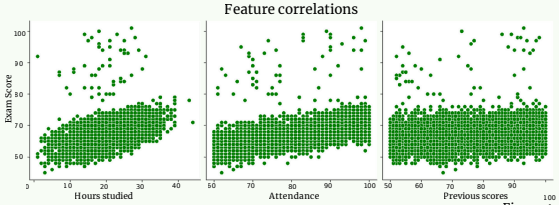

Figure 1

02. Goals

- Better understand the drivers of academic success among students.
- Build predictive tools that could identify students at risk and who need more support.
- Assess which prediction model provides more practical and actionable insights.

03. Methodology

Methods:

- Correlation analysis
- Cross-validation
- Hyper-parameter tuning
- Data scaling
- Feature importance

Prediction models:

- Linear Regression (Lasso / Ridge)
- Random Forest Classifier/ Regressor
- K Neighbors Classifier
- Decision Tree Classifier
- Logistic Regression

04. Analysis

Analysis was started by exploring the relationships between the students' final exam score and the available features in the dataset. Several variables required preprocessing. Initial correlation analysis and visual exploration indicated that student performance is influenced most strongly by attendance rate, hours studied and previous exam score which showed the clearest positive associations with the final exam score (Figure 1). Other features demonstrated minimal or no meaningful impact.

Based on these findings, both regression and classification models were developed to further assess and quantify the importance of these factors. As shown on graph on the right, these features also show most importance in training the models (Figure 4).

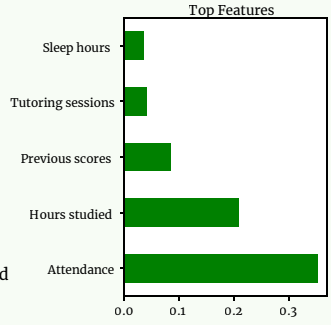

Figure 4

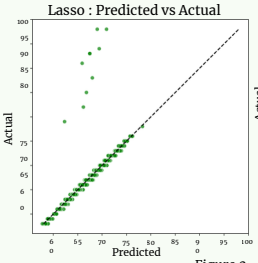

Figure 2

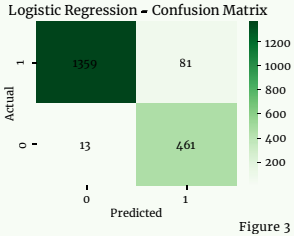

Figure 3

Data source: https://www.kaggle.com/datasets/mosapabdelghany/student-performance-factors-dataset/data
Repository: https://github.com/liisnele/IDS2025-performace-factors

05. Results

Overall, from the regression models used, Lasso Regression performed best after hyperparameter tuning and cross-validation, achieving the lowest test RMSE among the four models. Linear and Ridge were almost at the same level, meanwhile Random Forest showed signs of overfitting.

From the classification models, best recall score was achieved by Logistic Regression model, which got up to 0.973. Decision Tree and Random Forest Classifier did not get that level on recall or accuracy as the scores were 0.753 and 0.721.

Figure 2 and 3 are showing the best regression and classification models and their results. As seen, Lasso Regression predicted more people would get lower scores (Figure 2) and Logistic Regression classifies more actual low-achievers as high-achievers than high-achievers as low-achievers (Figure 3).

These results demonstrate that although both regression and classification offer valuable insights, the classification approach delivers more directly actionable outcomes for real educational decision making as the regression provides numerical estimates and classification identifies students in need of support which was the goal.