DTI5126 / IAI 5120 / SYS 5170

Applied Data Science

Student Performance Analysis

Group #10

October 3rd, 2025

Problem Statement

Students' academic performance is shaped by a combination of variant and invariant factors. Variant factors are those within a student's reasonable control, such as study habits and alcohol consumption patterns, while invariant factors lie outside their control, including parental education level, family dynamics, and personal health status. Students facing negative invariant factors are at a higher risk of underperforming for reasons beyond their influence, and therefore stand to benefit from targeted support and attention from educators. Given the resource constraints commonly present in educational settings, and to minimize the barriers to access assistance faced by students in resource-constrained circumstances, any support must be delivered as efficiently as possible, focusing on interventions that are proven to enhance student outcomes.

Our group's goal is to identify the invariant factors that contribute to negative academic performance and the variant factors that promote positive outcomes. By doing so, we aim to recognize at-risk students early in their educational journey and assist them in a resource-effective, evidence-based manner. Success will be measured by the accuracy of predictions in identifying at-risk students based on negative invariant factors, and by the degree of improvement following the implementation of positive variant interventions.

Dataset

We are using a dataset provided by the course showcasing the records of 382 students from two secondary schools in Portugal. The dataset focuses on student performance in two subjects: Mathematics (mat) and Portuguese language (por). The records were collected using the schools reports as well as questionnaires and have no missing values. The dataset contains different variant and invariant attributes that may impact the student performance, such as demographic details and social factors. We verified the dataset using UCI Machine Learning Repository.[1]

[1] **[1]** D. Dua and C. Graff, *Student Performance*, UCI Machine Learning Repository, Univ. of California, Irvine, CA, USA, 2019. [Online]. Available: https://archive.ics.uci.edu/dataset/320/student+performance

Intended methods

Our analysis will focus on distinguishing the impact of invariant factors (e.g., parental education, family background, health status) and variant factors (e.g., study time, alcohol consumption, absences) on student performance. To achieve this, we will apply a combination of supervised and unsupervised learning techniques that allow both prediction of outcomes and exploration of hidden student groupings.

Supervised Learning

We will use supervised learning to predict student performance outcomes, measured by final grades (continuous values) and categorical classifications such as pass/fail or performance levels.

Two predictive approaches will be implemented:

Random Forest Regressor: This method will be used to predict final grades. Its ability to model non-linear interactions across categorical and numerical variables makes it well-suited for analyzing this educational data. It also provides feature importance scores that indicate which invariant and variant factors most strongly influence student outcomes.

Random Forest Classifier and K-Nearest Neighbors (KNN): These models will be applied to categorical outcomes (e.g., identifying students at risk of failing vs. succeeding). The Random Forest Classifier provides us with strong predictive accuracy and feature importance rankings, while KNN will identify risk by comparing each student to peers with similar profiles. Together, these methods will enable a clear comparison of model accuracy and interpretability.

Unsupervised Learning

To complement predictive modeling, we will apply clustering to uncover naturally emerging student profiles:

K-Means Clustering: This technique will be used to segment students into groups defined by shared invariant and variant attributes (e.g., family background, study time, extracurricular activities). The goal is to identify meaningful clusters such as “high-effort/low-resource” or “low-effort/high-resource” profiles. These clusters reveal distinct patterns of risk and resilience, providing us a basis for designing targeted interventions and directing limited resources where they are most effective.

This mixed-method strategy produces actionable findings for our project highlighting at-risk students and the interventions most likely to help.

Group Members and Roles

Name	Contribution
Adjmal Younoussa	Contributed towards problem statement & dataset description
Maria Arias	Contributed towards problem statement & intended methods
Juan Hiedra Primera	Contributed towards problem statement & intended methods
Wenkang Liu	Contributed towards problem statement & dataset description
Nouf Almudlej	Contributed towards dataset description & overall document structure
Amal Hussein	Contributed towards problem statement & intended methods