Javier Algarra (268286) and Nil Lescure (269010)

Student Dropout Family Assistant

Problem Statement

Early school dropout is a significant challenge for families and educational centres, often emerging gradually through patterns such as frequent absences, low study time or lack of academic engagement. Families commonly struggle to understand whether certain behaviours are warning signs or not, asking questions like “Should I be worried about these absences?” or “Would studying more really reduce dropout risk?”.

This project addresses this need by developing a visual, interactive, interpretable and family-oriented system that predicts a student’s dropout risk and explains the key factors behind it. The goal is to turn complex educational data into clear insights, enabling parents, guardians and tutors to understand risk levels and explore actionable changes that could help a student stay engaged in school.

Dataset Overview

We use the student_dropout.csv dataset containing 700 rows and 34 features of students’ information. We chose the variable Dropped_Out as target, which indicates whether the student left school early. It is important to say that we have dropped the columns Grade_1, Grade_2 and Final_Grade because they are features about specific assessments that we can replicate. Leaving us with 31 features (including the Dropped_Out feature).

https://www.kaggle.com/datasets/abdullah0a/student-dropout-analysis-and-prediction-dataset/data

The dataset consists of student records containing demographic, academic, family, and lifestyle information, including variables such as age, gender, parental status, school support, family support, study habits, past failures, absences, and health indicators. The target variable is whether a student dropped out during the academic period.

The exploratory analysis of the data reveals key patterns:

- **Academic Factors:** Students with higher numbers of past failures are significantly more likely to drop out. High absenteeism strongly correlates with dropout, indicating that regular attendance is important for academic success.
- **School Selection:** The choice of school is a key factor influencing dropout risk, as it shapes access to resources, academic support, and the overall learning environment.
- **Aspirations and Engagement:** Students who aspire to higher education exhibit lower dropout probability.

Overall, the analysis highlights that dropout is influenced by a combination of academic performance, social/family support, lifestyle habits, and personal aspirations. These insights form the basis for further modeling and actionable recommendations to identify and support at-risk students.

Business Questions and Objectives

The tool is designed around the real questions families and school counselors confront when trying to understand a student's wellbeing and academic trajectory. The model and the visual interface aim not only to predict risk but also to empower families to interpret and act upon it.

Some of the key guiding questions include:

- What is the dropout probability for a specific student?
- Which factors or set of factors contribute positively or negatively to this risk, and in which scale?
- How would the risk change if the student adjusted certain behaviours or habits?

The objectives of the project are therefore to:

- Provide a validated machine learning model capable of predicting dropout risk.
- Offer an intuitive interface where users can input the student information, and get the dropout risk as a percentage.
- Deliver a What-If simulation tool to test the impact of modifiable factors such as absences, study time, school support, extracurricular activities, and lifestyle habits.
- Incorporate Explainable AI components to ensure trust, understanding and responsible use.

Methodology

The model was developed by first cleaning and preparing the dataset: removing specific grade-related columns, encoding all categorical variables, and validating numeric consistency across features. After preprocessing, the data was split into training and test sets to ensure that the model generalised well to unseen students. A gradient-boosting decision-tree model was selected because of its strong performance on tabular educational data and its ability to capture non-linear interactions between academic, behavioural and family factors. Once trained and evaluated, the final model and its encoders were exported as dropout_model.pkl and integrated directly into the application to support real-time, interpretable dropout-risk predictions.

The results generated by the model were then embedded in a multi-page Streamlit interface designed to make insights clear and actionable for families and tutors. The Data Exploration page allows users to filter students and observe contextual patterns through dynamic statistics, distributions and correlation analysis. The Dropout Predictor page provides personalised risk estimation with intuitive labels and interactive What-If buttons to test the effect of modifying individual features. The Changes Impact Simulator expands this functionality by evaluating grouped interventions and offering recommendations or scholarship suggestions based on predicted impact. Finally, the Model Explainability page incorporates SHAP visualisations to clarify both global model behaviour and the specific factors behind an individual prediction, ensuring transparency and supporting responsible, trust-based use of the system.

Conclusions

The findings reveal that dropout is less about isolated behaviours and far more about the ecosystem in which a student develops. The model consistently shows that academic difficulty is not just a symptom of disengagement but a turning point: the first failure dramatically reshapes the trajectory, signalling both capability challenges and a weakened sense of academic belonging. This creates an inflection point where timely support is disproportionately impactful.

A second major insight is the surprisingly large structural and institutional influence. The school itself acts almost like a latent variable: some environments inherently elevate risk regardless of student characteristics. This indicates that dropout is not purely an individual issue but is strongly shaped by school-level culture, expectations, and resource availability. Interventions that ignore this dimension risk placing responsibility solely on students and families when the context may be the true driver.

Motivation emerges as a third key pillar. The strong protective effect of wanting higher education suggests that aspirations buffer structural disadvantages. Students with a sense of future orientation can withstand academic setbacks better, implying that nurturing long-term goals may be as important as improving academic skills.

Interestingly, several lifestyle factors do not behave as intuition might suggest. Reducing social activity or increasing free time does not have a significant effect on the dropout risk, hinting that these variables don't influence the disengagement.

Across simulations, one consistent pattern emerges: micro-changes in everyday habits rarely shift dropout risk, whereas structural changes (school environment, access to resources) and identity-level shifts (motivation, aspirations) produce meaningful improvements. This suggests that effective dropout prevention is less about correcting behaviours and more about changing contexts and reinforcing purpose.

The explainability tools show that each student's risk is governed by a small cluster of dominant factors. This means interventions do not need to be broad or complex; they simply need to target the right levers. Dropout, therefore, is not a diffuse phenomenon but one with clear pressure points, making it possible to act with precision rather than generalised advice.

Limitations

- **Small dataset size:** With only around 700 student records, the results may not generalize to larger populations or different educational contexts.
- **High dimensionality:** The dataset contains many variables relative to the number of observations, which can increase the risk of overfitting in predictive models.