Data Mining and Machine Learning
Project Proposal

Predicting Student Dropout in Higher Education Using Supervised Learning

Master's Degree in Artificial Intelligence and Data Engineering

MARTINA FABIANI

Academic year 2024-2025

Problem Description

- University student dropout is a significant issue affecting both academic institutions and students' personal and professional development. Identifying students at risk early can help reduce dropout rates and improve overall educational outcomes.
- Data Mining and Machine Learning techniques allow for the extraction of hidden patterns from large and heterogeneous educational datasets. These techniques are particularly effective for:
  - Predictive modeling of student outcomes
  - Early warning systems for academic risk
  - Supporting data-driven interventions
- This project aims to build a **supervised classification model** to predict whether a student is likely to:
  - Graduate
  - Remain enrolled
  - Drop out

The goal is not only to **improve prediction accuracy**, but also to **identify key factors** associated with academic success or failure, supporting proactive and personalized support strategies.

Dataset Description

- The dataset was obtained from a public repository and contains data from a ***Portuguese higher education institution***. It was constructed by merging information from multiple independent administrative sources, including national admission records, academic performance data, and socioeconomic indicators.
  The dataset includes students enrolled in a wide range of undergraduate programs, such as Agronomy, Design, Education, Nursing, Journalism, Management, Social Services, and Technology-related fields.
- The dataset is publicly available at the *UCI Machine Learning Repository*: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
- The dataset contains **4,424 student records** and **37 attributes**. Each record includes:
  - ✓ Demographic data (e.g., gender, age, nationality)
  - ✓ Socioeconomic data (e.g., family education, scholarship, tuition status)
  - ✓ Macroeconomic context (e.g., inflation, GDP)
  - ✓ Academic data at enrollment (e.g., course, previous qualification, application mode)
  - ✓ Academic performance in the first and second semesters

- ***Input:*** structured tabular data with mixed data types (categorical, binary, continuous)
- ***Output:*** categorical target variable with 3 classes: *Graduate, Enrolled, Dropout*

References

- Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V.
  Predicting Student Dropout and Academic Success. *Data* **2022**, *7*, 146.
  https://doi.org/10.3390/data7110146