# Predict Students' Dropout and Academic Success

Grupo 42

Alexandre Morais

Carlos Costa

Nuno Ramos

# Problem Definition

- Develop and evaluate machine learning models capable of learning from labeled data to make accurate predictions regarding a specific target variable, focusing on supervised learning techniques in the context of classification problems.

- The dataset used is from "Early prediction of student's performance in higher education: a case study" which looks to predict students' dropout and academic success, making use of 4424 instances and 36 different features.

# Related Work

- **Ensemble Learning Approach** by Shivam Singh. A project on GitHub applied ensemble methods (Voting Classifier combining RF, SVM, and k-NN) using the same dataset as us.
- **Comparative Study** by Hanah Kim. Compared Neural Networks vs. Random Forests, finding RF slightly better (92.05 % vs. 91.71 %) in predicting students' drop out in the U.S.

# Methodology

- Some of the tools and libraries used are as follows:
  - Python
  - Anaconda
  - Jupyter Labs
  - NumPy
  - SciPy
  - Pandas
  - Scikit-Learn
  - MatPlotLib
  - Seaborn

# Methodology

- The algorithms that would be used:
  - Decision Trees
  - Neural Networks
  - K-nearest Neighbors
  - Support Vector Machines
- As for the metrics used:
  - Accuracy
  - Precision
  - Recall
  - F1- core
  - Confusion Matrixes
  - Training and testing times

# Work Progress

- So far it was implemented:
  - Initial Exploratory Data Analysis
  - Data loading
  - Preprocessing