Portuguese University Slots

# Project Setup

Universities in Portugal (as anywhere else) only have a limited number of slots each year.

*“The access to higher education is made through a national online process, where the students enter the university by priority based on their grades.” -* [*Sir Winston Wikipedia (2020)*](https://en.wikipedia.org/wiki/Education_in_Portugal#Access_to_higher_education)

This year, universities want to be more prepared, and get an estimate on the grades of students, to possibly accommodate for the right amount of required university slots. For this reason, they conducted a student survey in the previous year, collecting a lot of personal information, as well as first, second and third period grades. They collected this data for portuguese classes, and for mathematics classes. If students apply to a mathematics/portuguese study place, the mean of these three grades influences their chances of being accepted, however, no strict threshold or fixed number of seats is given.

Therefore, it is your task, not necessarily to predict how many students will be accepted into the university this year, but how well we can estimate the grades of students, based on the survey data from the previous year. Ideally, the university would get reliable estimates as early as possible, meaning that they are interested in the precision of estimates before and after the first and second period of the final school year.

The dataset(s) can be found at the following address:

<https://archive.ics.uci.edu/ml/machine-learning-databases/00320/student.zip>

# Task

* You are not directly creating an application to deliver certain predictions, but rather making a proof-of-concept (POC) as to how well such predictions can be made.
* As this is a highly analytical task, we are not just interested in raw precision scores, but also insights into the influential factors of the data for the predictions. After all, reliability of a system rests on two pillars, an automatic evaluation, and the manual understanding and interpretability of the system.
* Evaluate the predictability of each of the grades at different stages, i.e. with and without knowing all/some of the previous grades
* Evaluate the predictability of the university criterion (mean grade over the three periods) in 3 different stages:
  1. Before the first period ends
  2. After having the grades from the first period, but before the second period ends
  3. After having the grades from the first two periods, but before the third period ends
* Choose one of the classes (Mathematics or Portuguese) to restrict your evaluation to
  1. Mathematics: student-mat.csv
  2. Portuguese: student-por.csv

## Optional Extra Tasks

* If time permits, evaluate both classes (Math. and Port.) and compare differences in influential factors for either prediction.
* If time and your models permit it, describe a few prototypical student profiles that are un-/likely to get accepted into the university.

# Dataset Attributes

## Survey Data

**school** - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

**sex** - student's sex (binary: 'F' - female or 'M' - male)

**age** - student's age (numeric: from 15 to 22)

**address** - student's home address type (binary: 'U' - urban or 'R' - rural)

**famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

**Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

**Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

**Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

**Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

**Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

**reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

**guardian** - student's guardian (nominal: 'mother', 'father' or 'other')

**traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

**studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

**failures** - number of past class failures (numeric: n if 1<=n<3, else 4)

**schoolsup** - extra educational support (binary: yes or no)

**famsup** - family educational support (binary: yes or no)

**paid** - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

**activities** - extra-curricular activities (binary: yes or no)

**nursery** - attended nursery school (binary: yes or no)

**higher** - wants to take higher education (binary: yes or no)

**internet** - Internet access at home (binary: yes or no)

**romantic** - with a romantic relationship (binary: yes or no)

**famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

**freetime** - free time after school (numeric: from 1 - very low to 5 - very high)

**goout** - going out with friends (numeric: from 1 - very low to 5 - very high)

**Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

**Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

**health** - current health status (numeric: from 1 - very bad to 5 - very good)

**absences** - number of school absences (numeric: from 0 to 93)

## Grades

**G1** - first period grade (numeric: from 0 to 20)

**G2** - second period grade (numeric: from 0 to 20)

**G3** - final grade (numeric: from 0 to 20)

# Project Guideline

* Collect your results in a readable notebook (on google colab)
* Present your findings in a **15 minute** presentation
  + If you made certain decisions to alter the input data, make a clear argument to support these decisions
  + The university administration is not very akin to machine learning. Describe to them how your (different layers of) predictions are being derived in a comprehensible way, to create trust in your approach.
  + Also make clear why your approach may be the “correct” or “best” approach.

# 

# Hints

* You have a variety of tools in your toolbox by now, make use of them
  + Explore your data
  + Use different encodings
  + Select valuable features
  + Extract new features
* If you are using complex models (which you should, considering that we want to learn about ensemble methods), comparison to a weak baseline model is sometimes an eye-opener for a certain target audience.
* Consider different metrics for evaluating your predictions, in favour of explainability.
* Right now, the grades are numerical, however, they are integers on a scale between 0 and 20. Thus, you could view this as a classification task, but then you have 21 target classes. Perhaps there is a way to reduce the number of classes to predict.