# Aprendizagem Automática

Trabalho Prático 2 - 2022/2023

## Dataset - Student dropout

## 1. Objetivo

Utilizando informação do histórico académico dum conjunto de alunos (curso, ECTS matriculados e concluídos e notas médias ao longo de vários semestres), construir um modelo preditivo que responda à pergunta: "quais os alunos em risco de abandonar os estudos?"

## 2. Descrição do trabalho

O trabalho foi desenhado como um desafio para o qual todos os alunos terão acesso ao mesmo conjunto de dados. Os grupos poderão usar os métodos que quiserem, devendo ter em atenção o modo como o trabalho será testado. O relatório é uma parte integrante e importante do trabalho pelo que não o devem descurar. Será no relatório que irão descrever como procederam e as opções que tomaram para propor um classificador e maximizar o seu desempenho.

O trabalho será desenvolvido na linguagem Python com as bibliotecas sklearn e pandas. Além destas pode incluir outras bibliotecas que considerem adequadas para manipulação e visualização de dados (Matplotlib, Seaborn, etc …).

As métricas de desempenho a serem utilizadas deverão ser a **precisão** e a **cobertura**, do ponto de vista de classificação binária em que a classe positiva é a deteção de “insucesso académico” (indicado como 1 na coluna “Failure” do dataset).

O objetivo principal é maximizar a deteção de casos de insucesso (ou seja, que todos alunos em risco de insucesso sejam classificados na classe positiva) minimizando os Falsos Negativos. Pretende-se portanto maximizar a cobertura. No entanto, para evitar que se lance a suspeita de insucesso demasiadas vezes sobre casos que não estão em risco, pretende-se garantir um nível mínimo de precisão. Estes objetivos primários podem ser resumidos como:

- **Maximizar a cobertura.**
- **Garantindo um mínimo de 70% de precisão.**

Como objetivo secundário deve propor um modelo alternativo simplificado que tenha no máximo 2 atributos, os quais poderão ser alguns dos atributos indicados no dataset, ou transformação dos atributos existentes no dataset. Por exemplo, poderia criar um atributo que fosse média-global calculada a partir das médias em cada semestre. Mas também poderia selecionar diretamente dois dos atributos. Pode portanto criar esses dois atributos do modo que quiser, desde que use apenas a informação disponível nos atributos do dataset.

Estes objetivos secundários podem ser resumidos como:

- **Maximizar a cobertura.**
- **Garantindo um mínimo de 70% de precisão.**
- **Usar apenas dois atributos**

O relatório, para além de outra informação que ache relevante, deve incluir a seguinte informação:

- Indicação dos requisitos em termos de bibliotecas python necessárias ao funcionamento do trabalho proposto.
- Apresentação e análise do conjunto de dados
- Descrição do conjunto de experiências realizadas e que consideraram mais relevantes (incluindo atributos utilizados, algoritmos e parâmetros testados, etc) para fundamentar as decisões tomadas.
- Indicação do desempenho alcançado nas diversas experiências realizadas.
- Discussão de resultados e conclusões.

# 3. Datasets

O conjunto de dados a que terão acesso para usar em todas as experiências está no ficheiro "dropout-trabalho2.csv". O dataset tem 2110 instâncias, e tem os seguintes atributos:

['Id', 'Program', 'Y0s1_enrol', 'Y0s2_enrol', 'Y1s1_enrol',
'Y1s1_complete', 'Y1s1_grade', 'Y1s2_enrol', 'Y1s2_complete',
'Y1s2_grade', 'Y2s1_enrol', 'Y2s1_complete', 'Y2s1_grade', 'Y2s2_enrol',
'Y2s2_complete', 'Y2s2_grade', 'Y3s1_enrol', 'Y3s1_complete',
'Y3s1_grade', 'Y3s2_enrol', 'Y3s2_complete', 'Y3s2_grade', 'Y4s1_enrol',
'Y4s1_complete', 'Y4s1_grade', 'Y4s2_enrol', 'Y4s2_complete',
'Y4s2_grade', 'Rest_enrol', 'Rest_complete', 'Rest_grade', 'Failure']

Id - é um identificador único do aluno
Program - tem valores de 0 a 3 e indica a licenciatura a que o aluno pertence
YNsX_enrol - número de ECTS inscritos, há N anos atrás, no semestre X
YNsX_grade - classificação média há N anos atrás, no semestre X
YNsX_complete - número de ECTS em UCs com aprovação, há N anos atrás, no semestre X
RestXXX - ECTS inscritos, aprovados e classificação média há mais de 4 anos atrás
Failure - classe binária, indicação da classe 0 para sucesso; 1 para Insucesso (classe positiva)

O conjunto de teste "dropout-teste.csv" que será usado na apresentação, terá 528 instâncias, foi extraído dum dataset original e é uma extração aleatória de 20%, constituindo o conjunto em "dropout-trabalho2.csv" os restantes 80%.

# 4. Condições Gerais

- O trabalho deverá ser efetuado por grupos até 3 elementos. O trabalho será discutido em dia e horário a anunciar.
- O trabalho será testado de acordo com o ficheiro **teste.py** com o qual será testado usando um ficheiro de dados de teste ("dropout-teste.csv") que estará disponível apenas na apresentação do trabalho.
- O upload do trabalho deve ser efetuado através do moodle através de um ficheiro **`.zip`** com um nome com **`num1_num2`** em que `num1` e `num2` são os números dos alunos que compõem o grupo. O ficheiro submetido deve incluir:
    - um ficheiro **trabalho2.py** com todo o código necessário à definição do modelo proposto.
    - Relatório em **PDF** de acordo com o solicitado acima.
- Será aplicado o código de conduta do Departamento de Informática. Em caso de fraude, para além reprovação à disciplina, a situação será reportada.

# 5. Ficheiros disponíveis

- Enunciado do trabalho trabalho2.pdf
- dropout-trabalho2.csv (dataset principal)
- dropout-teste-modelo.csv (modelo de dataset de teste com apenas 1 linha de dados)
- teste.py (modelo do ficheiro de teste)