TÉCNICO LISBOA

# Data Science 2020 Project

## Deadline – December 4th 2020 @ (23:59)

## PROJECT GOAL

Critical application of data science techniques to discover information in two distinct problems. Students are asked to explore the datasets and, in accordance with their findings, adequately select and learn models for the available data, as well as assess and relate those models. Additionally, students should be able to criticize the results achieved, hypothesize causes for the limited performance of the learned models, and identify opportunities to improve the mining process.

## DATA

The datasets for analysis in this project are:

- **Heart Failure Prediction –** target = DEATH_EVENT
  https://www.kaggle.com/andrewmvd/heart-failure-clinical-data
- **QSAR oral toxicity Data Set –** target = last variable
  https://archive.ics.uci.edu/ml/datasets/QSAR+oral+toxicity

## METHODOLOGY

Information discovery on both datasets must be done using *data profiling*, *data preparation*, ***unsupervised techniques*** (pattern mining and clustering), and ***classification techniques***, including naïve Bayes, kNN, decision trees, random forests and XGBoost.

### *Development*

Students may choose the mining tool to apply, between **python** (using *scikit-learn*), **R** and any other language. Other business intelligence platform may be used but discouraged, since they are not prepared to deliver the charts required.

### Data Profiling

The students should perform a statistical analysis of the datasets in advance and summarize relevant implications in the report, such as the underlying distributions and hypothesize feature dependency.

### Data Preparation

In accordance with the properties of the input dataset and the behavior of the target learning algorithm, the students are allowed to apply preprocessing techniques when needed or under a solid conjecture of its potential impact on learning.

### Unsupervised Learning

Unsupervised exploration must be done through clustering and association rule mining. Class variables **cannot be used** to cluster the data, only to plot and evaluate the results, if desired. However, cohesion and separability have to be evaluated.

### Classification

Supervised exploration must be done via the application of *kNN, Naïve Bayes, Decision Trees, Random Forests* and *XGBoost*. For this purpose, the use of class variables is mandatory. Evaluation of the obtained models should be done as usual, through confidence measures and evaluation charts, as. A thorough comparison of the adequacy of the models should be present taking into consideration the adequacy of their behavior against the properties of each dataset and their observed performance.

### Report

The report file should be named **`report_X.pdf`** (replacing X by the team number) and submitted through Fénix. It should follow the template, with at most **10 pages** including any appendix. Each additional page won't be considered.

The report may be written in Portuguese or English. It should describe the majority of experiments made over the data, from their profile to the discovered models. Beside the placed choices, preparation performed, applied parameterizations and found results for each dataset, their interpretation and critical analysis are mandatory. Additionally, it should include a comparison of the results achieved in both problems, and the relation among the information discovered through the different techniques.

### Delivery

The project has to be delivered through Fenix system, after enrolling the team. Only one report per team has to be submitted.

The submission deadline is the December 4$^{th}$ at 23:59.

## EXCELLENCE

A project that applies the suggested data mining techniques over the given datasets and provides a clear and *sound analysis of the collected results is not necessarily an excelling project*.

Excelling projects have three major characteristics.

*First*, they show an acute understanding of the data characteristics and their impact on the discovery, formulating hypothesis to explain differences in performance.

Second, robust assessments go beyond simple performance indicators, studying different and adequate parameters, and deriving trends from the experiments.

Third, poor results are not acceptable, and there is always something that we can learn from the data.

### *Plagiarism*

Plagiarism is an act of fraud. We will apply state-of-the-art software to detect plagiarism. Students involved in projects with evidence of plagiarism will be reported to the IST pedagogical council in accordance with IST regulations.

## EVALUATION CRITERIA

The project will be evaluated as a *whole*. Nevertheless, we provide below a decomposition of the total project score for the purpose of guidance and prioritization:

1. **Data profiling** (**5%**)
2. **Data preparation** (**20%**)
3. **Unsupervised**
    a. Association Rules (**5%**)
    b. Clustering (**10%**)
4. **Classification**
    a. Naïve Bayes (**2%**)
    b. KNN(**3%**)
    c. Decision Trees (**5%**)
    d. Random Forests (**5%**)
    e. xGBoost (**5%**)
5. **Evaluation and critical analysis (30%)**
6. **Improvement strategies (10%)**