# MEIC – ALAMEDA

# 2017/18

## Sistemas de Apoio à Decisão

## Data Science Project

### Project Goal

In this project the goal is that students apply their knowledge about data science techniques, for discovering information in two distinct problems (datasets). Students are asked to create models about data, understanding and relating those models. Additionally, students should also criticize the results achieved and discuss the difficulties faced on mining the different datasets.

### Methodology

Information discovery on both datasets has to be done in two steps: using unsupervised techniques (association rules and clustering) – **First Delivery**, and training classification models, only through decision trees (algorithms ID3, C4.5 or CART), naïve Bayes, KNN algorithm and neural networks (multilayer perceptron and RBF) – **Second Delivery**.
Students may choose the mining tools to apply (for example Weka, R, Python, SAS BI, Microsoft Analysis Services).

### *Data*

The data for the two problems are available as *.cvs* files.

### *First Delivery - Non-supervised Learning*

Non-supervised exploration has to be done through clustering and association rules. In both cases, class attributes **can't be used to explore the data**. However, these attributes may be used to assess clustering results through a comparative evaluation. Beside this, **statistical evaluation has also to be performed**, using studied indexes.

## *Second Delivery – Classification*

Supervised exploration, classification, has to be done by training classification models through KNN, Naïve Bayes, Decision Trees and Neural Networks. For this purpose, **the use of class attributes is mandatory**. Evaluation of the obtained models should be done as usual, through accuracy measures and evaluation charts, as studied in the classes.

## Deadlines

Students should register their groups and deliver the project report in two separate phases: the first delivery is on **October 23th** reporting unsupervised exploration and the second one on **November 13th** reporting classification, both via **Fénix**.

Both reports should follow the template given, containing a cover, an optional index and 4 (first delivery) or 8 (second delivery), including any appendixes. Each additional page won't be considered.

The report should describe in a succinct form the pre-processing performed, the parameters used, and the results found, their interpretation and critical analysis for each problem and technique used. Additionally, it should include a comparison of the results achieved in both problems, and the relation among the information discovered through the different techniques. The report may be written in Portuguese or English.

**Beware!!!!**

*Plagiarism is an act of fraud. It involves both stealing someone else's work and lying about it afterward. There will be made all efforts to detect plagiarism in this work.*

## Evaluation Criteria

The project will be evaluated following the listed criteria:

**First Delivery – Non-supervised mining**

a. Pre-processing (25%)
b. Association Rules (25%)
c. Clustering (25%)
d. Comparison and critical analysis (25%)

**Second Delivery – Classification**

a. Pre-processing (20%)
b. Instance Based Learning – algorithm KNN (10%)
c. Naïve Bayes (10%)
d. Decision Trees (20%)
e. Neural networks (20%)
f. Comparison and critical analysis (20%)