# Data analysis project 2021-2022

## I Aim

The aim of this project is to perform a complete analysis of a real case data set. You will be asked to work in groups of 4 to 5 students and to produce a pdf report containing a detailed analysis of a data set of your choice.

Several data sets coming from various science fields will be at your disposal and you will have to pick one before mid-November. Your final report shall be sent before midnight on June 3rd .

## II Expected work

In order to produce a good report, we strongly recommend you to follows all the steps mentioned thereafter:

- First, you will have to familiarize yourself with the field from whence your data come. The idea is that you should be able to understand and explain all the variables of your set and be capable of grasping which type of data mining task is the most relevant for this field and these data.
- For most data, you will have to start with a pre-processing step. This step may include: re-formating your data, dealing with missing values, dealing with aberrant and outlier values, grouping or deleting features, normalizing your data, etc. In the case of difficult data set, you may also have to choose to work only on a subsample of your data or a subsample of the features at your disposal.
- You may want to analyze the different descriptors of your data so that you can extract interesting information about their distributions, their repartition and eventual correlations between them.
- Depending on the type of data sets, you will have to either try to find interesting structures and clusters in the data, or to highlight strong links between the variables, build models allowing to detect the different classes in your data set, or building predictive models from your data.

You will be required to provide a detailed account of these different analysis in your reports including figures, statistical results, and your personal interpretation of all your results.

## III Data sets

| **Dataset** | **Abalone (UCI)** |
|---|---|
| Field | Nature |
| Size | 4177x8 |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

| **Dataset** | **Battalia4 (Misc)** |
|---|---|
| Field | Space (artificial), game |
| Size | 14898x33 |
| Missing values | NO |
| Max students | 5 |
| Difficulty | ** |

| **Dataset** | **Astorb (Caltech)** |
|---|---|
| Field | Space |
| Size | 717962x25 |
| Missing values | YES |
| Max students | 5 |
| Difficulty | **** |

| **Dataset** | **Exoplanet (NASA)** |
|---|---|
| Field | Space |
| Size | 3388x26 |
| Missing values | YES |
| Max students | 5 |
| Difficulty | *** |

| **Dataset** | **Communities (UCI)** |
|---|---|
| Field | Social |
| Size | 1993x128 |
| Missing values | YES |
| Max students | 5 |
| Difficulty | ** |

| **Dataset** | **Forestfires (UCI)** |
|---|---|
| Field | Forestry, misc |
| Size | 517x12 |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

| **Dataset** | **Body (AMSTAT)** |
|---|---|
| Field | Biology |
| Size | 507x25 |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

| **Dataset** | **German Credit (UCI)** |
|---|---|
| Field | Bank |
| Size | 1000x20 |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

| **Dataset** | **Dermatology (UCI)** |
|---|---|
| Field | Health |
| Size | 366x34 |
| Missing values | YES |
| Max students | 5 |
| Difficulty | ** |

| **Dataset** | **Heart-disease (UCI)** |
|---|---|
| Field | Health |
| Size | 920x14* |
| Missing values | YES |
| Max students | 5 |
| Difficulty | ** |

| **Dataset** | **EColi (UCI)** |
|---|---|
| Field | Biology |
| Size | 336x8 |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

| **Dataset** | **PM10 (CMU)** |
|---|---|
| Field | Pollution |
| Size | 500x8 |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

| Dataset | SkillCraft1 (UCI/TL) |
|---|---|
| Field | Games, theorycraft |
| Size | 3395x20 |
| Missing values | NO |
| Max students | 5 |
| Difficulty | ** |

| Dataset | SHealth (Misc) |
|---|---|
| Field | Health, mobile phone |
| Size | 610x9 |
| Missing values | YES |
| Max students | 3 |
| Difficulty | *** |

| Dataset | Vertebral Column (UCI) |
|---|---|
| Field | Health |
| Size | 300x6 |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

| Dataset | WDR11 (World Bank) |
|---|---|
| Field | Economy |
| Size | 10551x145 |
| Missing values | YES |
| Max students | 5 |
| Difficulty | *** |

| Dataset | Wind (CMU) |
|---|---|
| Field | Weather |
| Size | 6574x15 |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

| Dataset | Wine quality (UCI) |
|---|---|
| Field | Food, chemistery |
| Size | 6498x12* |
| Missing values | NO |
| Max students | 4 |
| Difficulty | * |

Remarks: No more than 2 groups per data sets. The final mark will take into consideration the difficulty of the data set you picked.