# TP1: Discover Weka and Python

During this lab, you will discover Weka and Python for data processing. The goal is to process the same dataset via both tools, to compare their effectiveness in terms of results and understanding. For the continuation of the TPs, you will have to choose one which you prefer.

# WEKA

1. Overview

Weka is a set of tools for manipulating and analyzing data files, implementing most artificial intelligence algorithms, including decision trees and neural networks. It is written in Java, available on the web, and is based on the book Data Mining, practical machine learning tools and techniques with Java implementations, Witten & Frank Publisher: Morgan Kauffman.

1. Installation, initialization

Weka is installed in the TP rooms. If you want to use it at home:

- Download zip archive from Weka website

- Unzip it.

-That's all.. . .

- The classes are in weka.jar

- The sources in weka-src.jar

- WekaManual.pdf is a fairly advanced introduction to Weka's features

3. Graphical interface

It starts with the command java -jar $ WEKAHOME / weka.jar

After launching it, you get the window called Weka GUI Chooser: choose the Explorer. The new window that opens (Weka Knowledge Explorer) has six tabs (you can add tabs by installing other packages via the package manager tool):

- Preprocess: to choose a file, inspect and prepare the data.

- Classify: to choose, apply and test different classification algorithms: these are supervised classification algorithms.

- Cluster: to choose, apply and test segmentation algorithms.

- Associate: to apply the association rule generation algorithm.

- Select Attributes: to choose the most promising attributes.

- Visualize: to display (in two dimensions) certain attributes in functions of others.

4. ARFF file

WEKA uses (among others) the arff file format to save the data. An arff file is composed of a list of examples defined by their attribute values.

An arff file always includes three types of information:

- the name of the relation: @relation allows to give a name to the relation

- the list of attribute names and the type of values: @attribute allows to define an attribute

- the list of instances: @DATA allows defining the data. each line represents a description, by the list of values of each of its attributes. A missing value is replaced by a question mark.

- the "%" character marks the comment lines

# Python

## Présentation

Python is a computer language and for which there are libraries implementing most artificial intelligence algorithms, including decision trees and neural networks. The basic libraries for TP are numpy, scipy, pandas and scikit-learn.

## Installation, initialization

Python Anaconda and the libraries are installed in the TP rooms. If you want to use it at home:

- Use and register on Kaggle which gives access to python tools without installation via an internet connection.

1. Interface graphic

On Kaggle, you organize yourself as a notebook (Kernel). This makes it possible to have a code cell followed by its execution for an immediate result.

## Files

Absolutely, any type of file can be read via python.

If you have questions about the implementation for WEKA or Python, do not forget to consult the documentation !!

# TP1: Pré-treatment et KNN

## 1. Loading data

You will have to carry out the continuation of the TP on the two tools, in order to master them and to be able to choose the one which suits you for the rest of the TPs.

**2. Analysis of attribute distribution**

For each attribute available, display its distribution, then its distribution according to the class, if it is possible. What do you find and what measures should you take?

## Attributes processing

You must process the table to keep only the attributes most relevant to the previous questions. In addition, we found that there are missing values for some tuples in the database. Why do we have to treat them? What are the means to achieve this? Choose one and apply it in our case.

## Transformation des attributes

You will notice that some attributes are discrete, why is it interesting to turn them into categorical attributes. Find the method that allows you to move from a discrete series to a finite value set.

In the same way, several attributes are chains of characters, to find the method which makes it possible to transform them in number.

## Création d'ensemble d'apprentissage et de test

You must create two sets from the Titanic dataset to train and test your future models. Why is it important?

You are asked to set up a random system to split the tuples into two sets based on the desired split, i.e. 70-30%. Study the distribution of data and criticize the limitations of this method.

## KNN

Set up the KNN model, vary the parameters of the models. What do you notice?