# 1. Présentation de WEKA

WEKA est un ensemble de classes et d'algorithmes en Java implémentant les principaux algorithmes de datamining. Il est disponible gratuitement, dans des versions pour Unix et Windows, à l'adresse http://www.cs.waikato.ac.nz/ml/WEKA/.

Ce logiciel est développé en parallèle avec un livre : **Data Mining par I. Witten et E. Frank (éditions Morgan Kaufmann).**

WEKA peut être utilisé de plusieurs façons :

- Par l'intermédiaire d'une interface utilisateur : c'est la méthode utilisée dans nos TP.
- Sur la ligne de commande.
- Par l'utilisation des classes fournies à l'intérieur de programmes Java

- **Intérêt de WEKA**

L'intérêt de WEKA dans le cadre du cours d'Apprentissage Artificiel est multiple. Nous citons à titre indicatif :

- Pouvoir mettre en œuvre les algorithmes étudiés en cours, sans devoir réécrire tout le code correspondant ;
- Comprendre et utiliser intelligemment les différentes sorties de ces algorithmes ;
- Pouvoir programmer des agents intelligents en un temps raisonnable, pour des tâches non triviales (par exemple : les jeux) ;
- Evaluer les performances d'un algorithme ;
- Comparer les performances de deux algorithmes ;
- Etudier le rôle des paramètres d'un algorithme ;
- Combiner plusieurs algorithmes ;
- Eventuellement définir un nouvel algorithme.

# 2. Ouverture de WEKA

WEKA est installé sur vos machines sous le dossier : ~/weka-3-6/weka.jar

Après l'avoir lancé, vous obtenez la fenêtre intitulée **WEKA GUI Chooser** présentant les environnements suivants :

- **Explorer :** un environnement pour explorer des données sur WEKA
- **Experimenter :** un environnement pour réaliser des expérimentations et des essais statistiques entre des schémas d'apprentissage.
- **KnowledgeFlow :** cet environnement présente les mêmes fonctionnalités que celles de l'environnement Explorer mais avec une interface 'drag-and-drop'. Egalement, il supporte l'apprentissage incrémental.
- **Simple CLI** : une interface ligne de commandes.

- Pour ce TP, choisissez l'environnement **Explorer.** Une nouvelle fenêtre s'ouvre **WEKA Knowledge Explorer** présentant six onglets :

  - **Preprocess** : pour choisir un fichier, inspecter et préparer les données ;
  - **Classify** : pour choisir, appliquer et tester différents algorithmes de classification: dans notre cas il s'agirait d'algorithme d'arbre de décision ;
  - **Cluster** : pour choisir, appliquer et tester les algorithmes de segmentation ;
  - **Associate** : pour appliquer l'algorithme de génération de règles d'association ;
  - **Select Attributes** : pour choisir les attributs les plus prometteurs ;
  - **Visualize** : pour afficher (en deux dimensions) certains attributs en fonctions d'autres.

## 3. Les données

Dans ce tp, vous allez vous familiariser avec l'utilisation de WEKA en utilisant la base de données iris.arff que vous trouverez dans le dossier: ~/weka-3-6/data/Iris.arff. Il s'agit d'une base de données, très célèbre, comportant 150 exemples de fleurs décrites par 5 attributs à valeur continue et appartenant à 3 classes.

- Ouvrez en premier lieu le fichier Iris.arff avec un éditeur de texte pour découvrir le format ARFF -Attribute-Relation File Format-.

En ce qui concerne WEKA, il vous offre, sous l'onglet **Preprocess,** les choix suivants pour générer un fichier de données :

- **Open File** : générer le fichier à partir d'un dossier local. Plusieurs formats sont supportés par WEKA par exemple : ARFF, CSV, C 4.5, etc ;
- **Open URL**: générer le fichier à partir d'une adresse URL ;
- **Open DB**: lire le fichier à partir d'une base de données ;
- **Generate** : générer des données artificielles grâce aux générateurs de données fournis.

- Cliquez sur le bouton **'Open file… '**. Choisissez le fichier de données, 'Iris.arff'. Un certain nombre d'informations vont apparaître dans la fenêtre **WEKA Explorer** : le nombre d'exemples, le nombre d'attributs. Dans la sous-fenêtre **Selected Attribute**, vous pouvez obtenir des statistiques basiques relatives à l'attribut sélectionné : Nom, type, manquant, unique, distinct, valeur min/max, etc. Le bouton **Visualize All** permet de voir tous les histogrammes en même temps, permettant d'avoir une idée de la répartition des données.

- WEKA vous offre également la possibilité d'opérer par un prétraitement en appliquant un filtre sur les attributs. Veuillez cliquer sur le bouton **Filter** pour découvrir les filtres présentés par WEKA. Par exemple, le filtre **DiscretizeFilter** permet de discrétiser des valeurs continues.