

# Apprentissage automatique - partie 1

Mohamed Amine Remita

16 mars 2016

Adapté des cours de Mohamed Bougessa Ph.D (DIC9370)  
et Ahmed Halioui (BIF7101)

1 Introduction

2 Les données

3 Prétraitement des données

4 Apprentissage automatique  
• Apprentissage supervisé

5 Plateforme WEKA

6 Atelier

7 Lecture

# Apprentissage naturel

- La faculté d'apprendre de ces expériences passées et de s'adapter est une caractéristique essentielle des formes de vies
- Elle est essentielle dans les premières étapes de la vie pour apprendre des choses aussi fondamentales que reconnaître une voix, un visage familier, apprendre à comprendre ce qui est dit, à marcher et à parler

# Apprentissage artificiel (machine learning)

- Une tentative de comprendre et reproduire cette faculté d'apprentissage dans des systèmes artificiels
- Concevoir des algorithmes capables à partir d'un nombre important d'exemples (expériences passées) d'en assimiler la nature afin de pouvoir appliquer ce qu'ils ont appris aux cas futurs
- Objectif de l'apprentissage : déterminer la relation entre les exemple et leurs catégories pour la prédiction et la découverte des connaissances

# Apprentissage artificiel

- Un programme possède des capacités d'apprentissage si au cours du traitement d'exemples représentatifs de données il est capable de construire et d'utiliser une représentation de ce traitement en vue de son exploitation
- => Élaboration d'un modèle pour la prédition et la découverte des connaissances
- Modèle = Description formelle des relations qui existent entre l'ensemble des attributs qui décrivent les données à traiter.

# Les données

# Les données

Dans un problème d'apprentissage automatique, les informations caractérisant une étude sont présentées sous la forme d'**attributs** et d'**objets**

- **Attribut** : est un descripteur d'une entité (dimension, variable)
- **Objet** : est un ensemble d'attributs (vecteur, tuple, enregistrement)

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

# Types d'attributs

- **Attribut discret**

- Numérique discret : la valeur de l'attribut appartient à  $N$
- Catégorie/symbole : exemple  $\{\text{rouge}, \text{vert}, \text{bleu}\}$ ,  $\{A, B, C\}$
- Données binaires (booléen)

- **Attribut numérique continu**

La valeur de l'attribut peut prendre une valeur numérique (montant du compte en banque, poids, etc.)

## Prétraitement des données

# Introduction

- Le prétraitement des données est crucial dans le processus de l'apprentissage automatique car les résultats dépendent de leur qualité
- Les données réelles sont souvent :
  - Incomplètes
  - bruitées
  - incohérentes

# Principales étapes

- Intégration
- Nettoyage
- Transformation
- Réduction

# Intégration

Combiner des sources de données différentes dans une seule structure

- Déetecter et résoudre les conflits de valeurs

Dans un seul attribut on peut avoir plusieurs mesures différentes (cm et pouces)

- Gestion de la redondance

- Le même attribut peut avoir des noms différents
- Un attribut peut être déduit d'un autre
- Solution : analyse de corrélation entre les attributs

# Nettoyage

## ① Données manquantes

Données non disponible / certaines attributs n'ont pas de valeur

### ① Causes

- Mauvais fonctionnement de l'équipement
- Incohérences avec d'autres données => supprimées
- Non saisies car non ou mal comprises (considérées peu importantes au moment de la saisie)

### ② Solutions

Ces données doivent être inférées

- Ignorer le tuple  
peu efficace quand le pourcentage de valeurs manquantes est élevé
- Utiliser une constante globale
- Utiliser la moyenne de l'attribut
- Utiliser la valeur la plus probable : formule Bayésienne ou arbre de décision

# Nettoyage

## ② Données bruitées

Bruit :

- erreur ou valeur aléatoire (excessive)
- un objet qui a des caractéristiques complètement différentes du reste de l'ensemble de données.

### ① Causes

- Instrument de mesure défectueux
- Problème de saisie
- Problème de transmission

### ② Solutions

- Clustering
- Écart interquartile ( $IQR = Q3 - Q1$ ) pour identifier les valeurs aberrantes (outliers) :  
inférieur à  $Q1 - (\alpha \times IQR)$   
supérieur à  $Q3 + (\alpha \times IQR)$

# Transformation

## ① Normalisation

- **Méthode min-max**

Mise à l'échelle pour avoir un petit intervalle spécifié

$$v' = \frac{v - \min_A}{\max_A - \min_A} (new\_max_A - new\_min_A) + new\_min_A$$

- **Z-score**

Même ordre de grandeurs pour les valeurs des attributs

$$v' = \frac{v - \mu_a}{\sigma_A}$$

# Transformation

## ② Discrétisation

- Attributs numérique => attributs nominaux
- Découper le domaine de variation en un nombre fini d'intervalles
- Discrétisation supervisée
  - Découper en K domaines égaux
- Discrétisation non supervisée
  - Découper itérativement en 2 sous ensembles jusqu'à une certaine condition d'arrêt

# Réduction de la dimension

- Certains attributs contiennent de l'information non pertinente et rend l'analyse des données plus complexe
- La présence des attributs non pertinents augmente potentiellement le temps d'exécution des algorithmes
- Solution : Réduction de la dimension
  - Obtenir une représentation réduite du jeu de données, plus petit en volume mais qui produit (ou presque) les mêmes résultats analytiques
    - analyse en composantes principales (PCA) : création d'un nouvel attribut à partir des attributs originaux
    - Techniques de sélection d'attributs (feature selection)

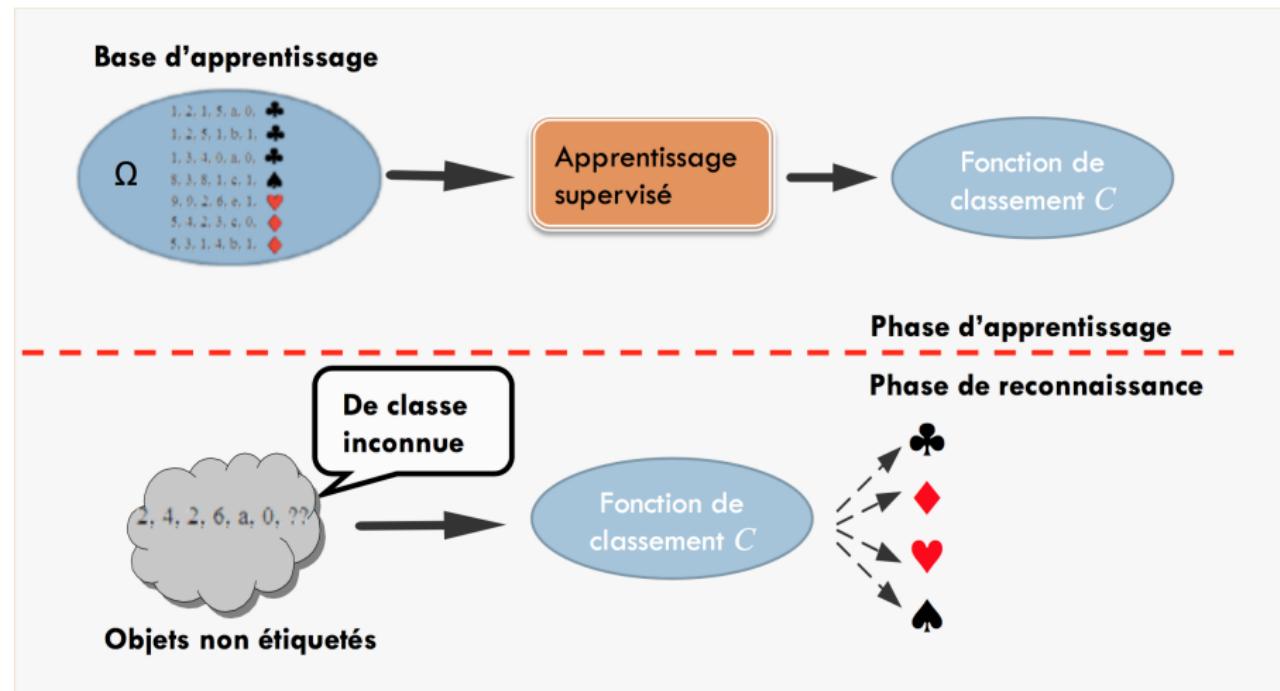
# Apprentissage automatique

- ① Si la décision est connue (apprentissage supervisé) :
  - Classification : décision / classe = catégorie
  - Régression : décision / classe = nombre continu
- ② Si la décision est inconnue (apprentissage non supervisé) :
  - Partitionnement (Clustering)

# Aprendissage supervisé

- Entrée : base de données d'apprentissage, ensemble d'exemples
- Trouver une fonction  $c : X \rightarrow Y$  qui approxime et généralise au mieux la relation entre les objets  $x_i$  et leur catégorie  $y_i$
- But :
  - Modèle de prédiction : classification de nouvelle données
  - Schéma explicatif : aide à comprendre les relations qui existent entre les entrées et les sorties
  - Régression : approximation de fonction

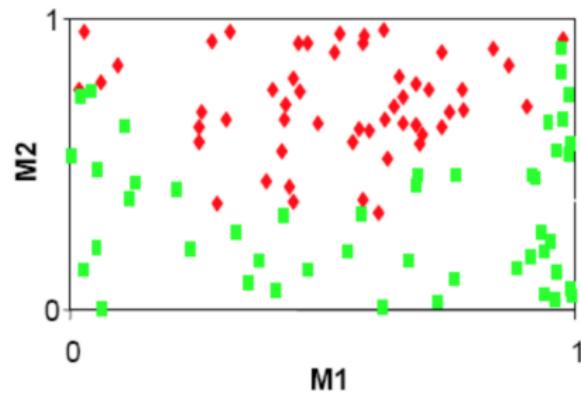
# Apprentissage supervisé



# Algorithmes d'apprentissage supervisé

Exemple illustratif : diagnostic médical à partir de deux mesures

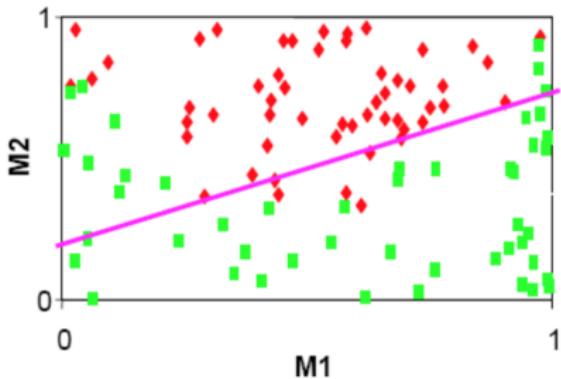
M1	M2	Y
0.52	0.18	Healthy
0.44	0.29	Disease
0.89	0.88	Healthy
0.99	0.37	Disease
...	...	...
0.95	0.47	Disease
0.29	0.09	Healthy



But : trouver un modèle qui classifie au mieux les nouvelles données

## Modèle linéaire

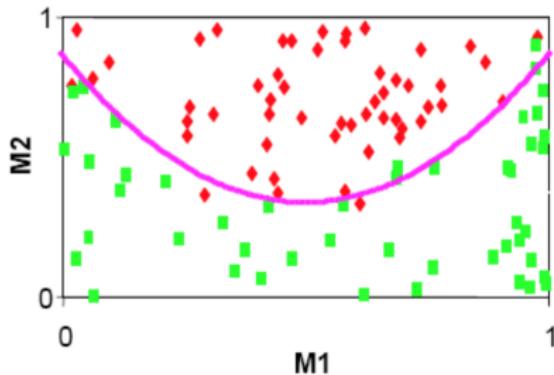
$$c(M1, M2) = \begin{cases} \text{Disease if } w_0 + w_1 * M1 + w_2 * M2 > 0 \\ \text{Normal otherwise} \end{cases}$$



- Phase d'apprentissage : à partir des données d'apprentissage, trouver les meilleures valeurs pour  $w_0$ ,  $w_1$  et  $w_2$ .
- Plusieurs alternatives pour ce modèle simple (analyse discriminante linéaire, SVM)

# Modèle quadratique

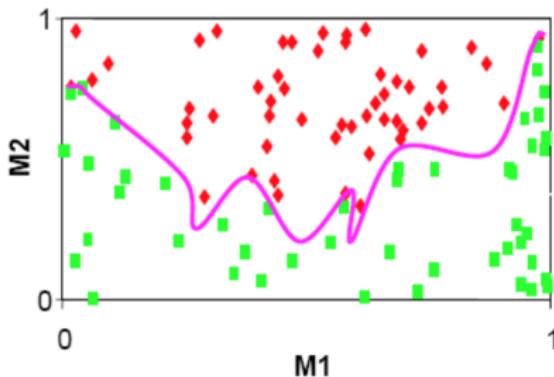
$$C(M1, M2) = \begin{cases} \text{Disease if } w_0 + w_1 * M1 + w_2 * M2 + w_3 * M1^2 + w_4 * M2^2 > 0 \\ \text{Normal otherwise} \end{cases}$$



- Phase d'apprentissage : à partir des données d'apprentissage, trouver les meilleures valeurs pour  $w_0$ ,  $w_1$ ,  $w_2$ ,  $w_3$  et  $w_4$ .
- Plusieurs alternatives pour ce modèle simple (Perceptron et SVM)

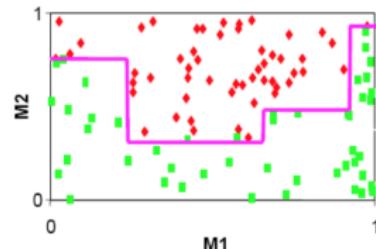
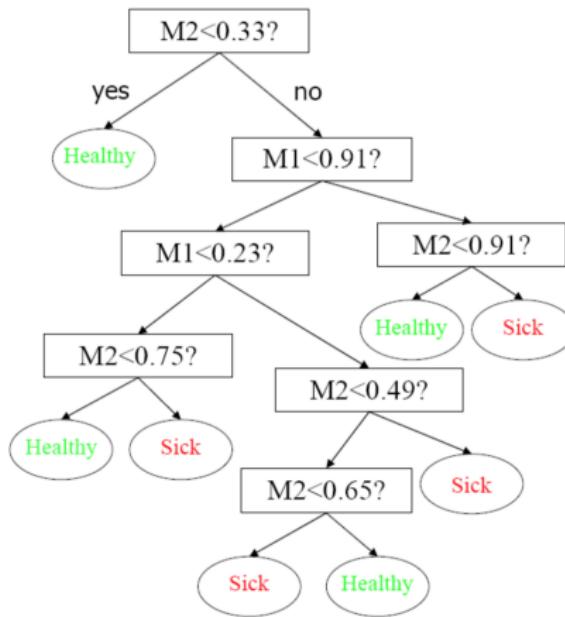
# Réseaux de neurones

$$c(M1, M2) = \begin{cases} \text{Disease if } \text{à partir de quelques fonctions complexes} \\ \text{Normal otherwise} \end{cases}$$



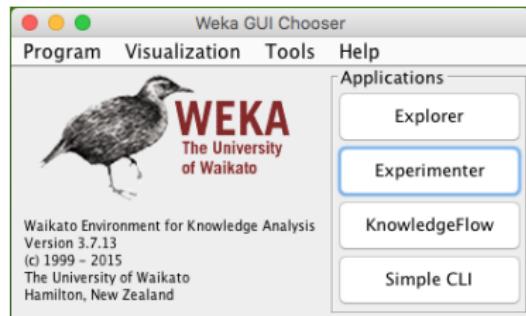
- Phase d'apprentissage : à partir des données d'apprentissage, trouver plusieurs paramètres d'une fonction complexe.

# Arbres de décision



- Phase d'apprentissage : à partir des données d'apprentissage, identifier des règles de décision qui permettent la séparation des deux classes

# Plateforme WEKA



# Introduction

- Suite de logiciels d'apprentissage automatique et d'exploration de données
- Développé à l'université Waikato de Nouvelle-Zélande
- Disponible pour différents OS
- Documentation riche et communauté large

# Que contient Weka ?

- Outils de traitement de données
  - Sélection, transformation, combinaison d'attributs, normalisation, ré-échantillonage, etc.
- Algorithmes pour l'exploration des données
  - Classification, clustering, régression
- Analyse de résultats
  - Évaluation des performances, comparaison d'algorithmes, etc.

## Formats d'entrées

- ARFF (Attribute Relation File Format)
- Autres formats : CSV, Bases de données, URL, etc.

# Formats d'entrées (ARFF)

## Exemple 1 ecoli

```
% Kenta Nakai
% Institute of Molecular and Cellular Biology
% Osaka, University
% 1-3 Yamada-oka, Suita 565 Japan
% nakai@imcb.osaka-u.ac.jp
% http://www.imcb.osaka-u.ac.jp/nakai/projC.html
% Donor: Paul Horton (paulh@cs.berkeley.edu)
% Date: September, 1996
% See also: yeast database

@relation ecoli

@attribute mcg numeric
@attribute gvh numeric
@attribute lip numeric
@attribute chg numeric
@attribute aac numeric
@attribute alm1 numeric
@attribute alm2 numeric
@attribute class {cp,im,pp,imU,om,omL,imL,imS}

@data

0.49,0.29,0.48,0.5,0.56,0.24,0.35,cp
0.07,0.4,0.48,0.5,0.54,0.35,0.44,cp
0.56,0.4,0.48,0.5,0.49,0.37,0.46,cp
0.59,0.49,0.48,0.5,0.52,0.45,0.36,cp
0.23,0.32,0.48,0.5,0.55,0.25,0.35,cp
0.67,0.39,0.48,0.5,0.36,0.38,0.46,cp
0.29,0.32,0.45,0.44,0.33,0.34,cp
```

Description de l'ensemble de données :  
**@relation**

Features / Attributs / Propriétés :  
**@attribute**

Nominaux

Numériques

Chaines de caractères

Date

PAS D'ID !!!!!!!

Début des exemples (instances) :  
**@data**

# Formats d'entrées (ARFF)

## Exemple 2 - weather

```
% Historique des observations météorologiques
% pour pouvoir jouer dehors

@relation test_play

@attribute outlook {sunny,overcast,rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {TRUE,FALSE}
@attribute play {yes,no}

@data
sunny,85,85, FALSE, no
sunny,80,90, TRUE, no
overcast,83,86, FALSE, yes
rainy,70,96, FALSE, yes
rainy,68,80, FALSE, yes
rainy,65,70, TRUE, no
overcast,64,65, TRUE, yes
sunny,72,95, FALSE, no
sunny,69,70, FALSE, yes
rainy,75,80, FALSE, yes
sunny,75,70, TRUE, yes
overcast,72,90, TRUE, yes
overcast,81,75, FALSE, yes
rainy,71,91, TRUE, no
```

- Nom de l'ensemble de données :  
**@relation**

- Features / Attributs /  
Propriétés : **@attribute**

- Nominaux

- Numériques

- Début des exemples  
(instances) :  
**@data**

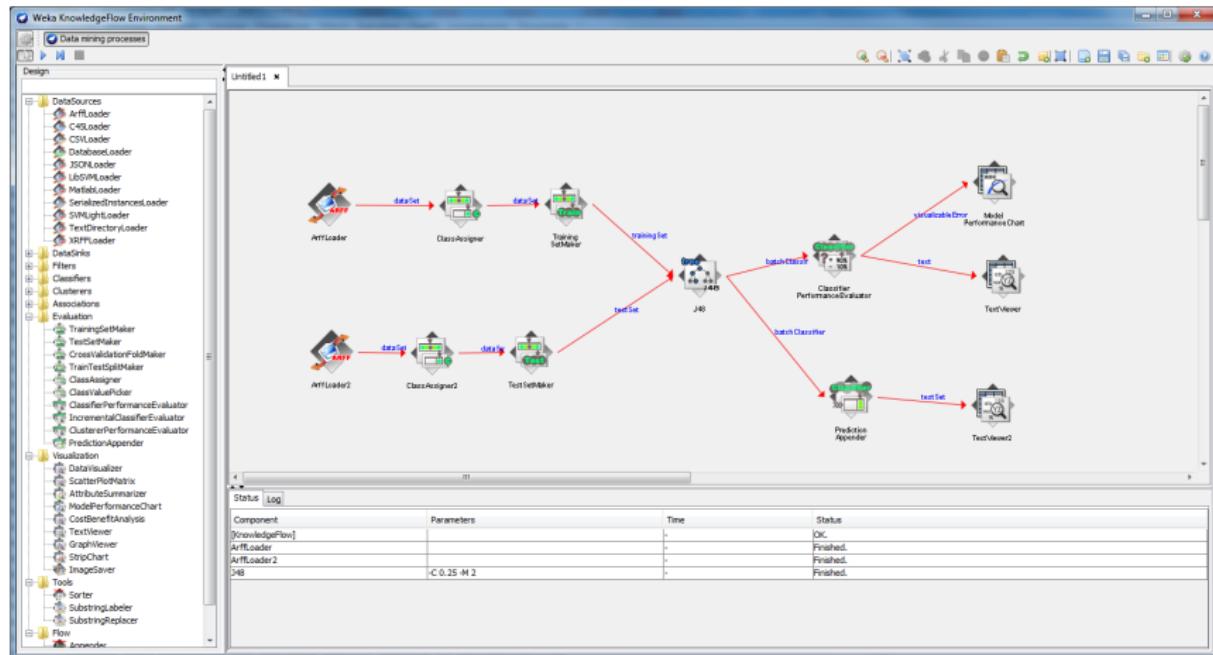
# Weka Explorer

The image shows two windows side-by-side. On the left is the "Weka GUI Chooser" window, which has a green background with a bird logo and the text "Weka Environment for Knowledge Analysis Version 3.7.13 (c) 1999 - 2015 The University of Waikato Hamilton, New Zealand". It features a sidebar titled "Applications" with options: Explorer, Experimenter, KnowledgeFlow, and Simple CLI. On the right is the "Weka Explorer" window, which has a white background. It includes a toolbar with buttons for Preprocess, Classify, Cluster, Associate, Select attributes, Visualize, and CPython Scripting. Below the toolbar are buttons for Open file..., Open URL..., Open DB..., Generate..., Undo, Edit..., and Save... A "Filter" section contains a "Choose" button and a dropdown set to "None", with an "Apply" button. The main area is divided into sections: "Current relation" (Relation: None, Instances: None), "Attributes" (All, None, Invert, Pattern, Remove), and "Selected attribute" (Name: None, Missing: None, Distinct: None, Type: None, Unique: None). At the bottom is a status bar with "Welcome to the Weka Explorer", a Log button, and a message "x 0".

# Weka Experimenter

The screenshot shows the Weka Experiment Environment window. At the top, there is a toolbar with 'Setup' (selected), 'Run', and 'Analyse' buttons. Below the toolbar, the 'Experiment Configuration Mode' is set to 'Simple'. The 'Results Destination' section includes an 'ARFF file' dropdown and a 'Filename:' input field with a 'Browse...' button. The 'Experiment Type' section shows 'Cross-validation' selected in a dropdown, and 'Classification' is chosen under 'Number of folds:'. The 'Iteration Control' section has a 'Number of repetitions:' input field and radio buttons for 'Data sets first' (selected) and 'Algorithms first'. The 'Datasets' section contains buttons for 'Add new...', 'Edit selected...', and 'Delete selected...'. A checkbox for 'Use relative ...' is also present. The 'Algorithms' section has similar buttons for adding, editing, and deleting algorithms. At the bottom, there are 'Load options...', 'Save options...', and navigation buttons for 'Up', 'Down', and 'Notes'. A status bar at the bottom displays file names: 'volte\_frog\_Screen\_AB\_28.osc', 'pmlr.pdf', 'pmlr\_tch.htm'.

# Weka Knowledge Flow Environment



# Autres plateformes pour l'apprentissage automatique

- RapidMiner (<https://rapidminer.com>)
- scikit-learn (<http://scikit-learn.org>)
- caret (<https://cran.r-project.org/web/packages/caret/>) pour R
- e1071 (<https://cran.r-project.org/web/packages/e1071/>) pour R
- OpenCV (<http://opencv.org>)

# Atelier

- Avec **WEKA Explorer** ouvrez le jeux de données **diabetes.arff**
- Combien d'attributs et d'objets dans ce jeux de données ?
- Entrainez des modèles avec différents algorithmes (J48, JRip, RandomForest, NaiveBayes, IBK, LibSVM, AdaboostM1) et en utilisant une validation croisée avec 10 itérations
- Ouvrez le jeux de données **segment-challenge.arff**
- Dans l'étape de classification fournissez le jeux de données **segment-test.arff** comme jeux de test
- Dans test options > More options > Output predictions , choisissez **PlainText**
- Entrainez quelques modèles et consultez leurs sorties
- Refaites cet exercice avec **WEKA KnowledgeFlow**

# Lecture

