Analyse de Données
Exercices

Pour ces exercices, vous allons utiliser l'outil libre Weka (Waikato Environment for Knowledge Analysis).

Pour ceux que ça pourrait intéresser, le livre suivant, sur l'analyse de données, introduit cet outil :
Ian H. Witten, Eibe Frank, Mark A. Hall
Data Mining : Practical Machine Learning Tools and Techniques
Morgan Kaufmann Series in Data Management Systems

1. Tout d'abord, téléchargez et installez l'outil.
2. Démarrez *Weka* et appuyez sur le bouton EXPLORER.
3. Vous allez utiliser comme données le contenu du fichier **arff** suivant qui décrit l'incidence du diabète (la classe) pour une population d'Amérindiens Pima en fonction de huit autres attributs. Du commentaire d'entête du fichier :

```
% 1. Title: Pima Indians Diabetes Database
% 5. Number of Instances: 768
% 6. Number of Attributes: 8 plus class
% 7. For Each Attribute: (all numeric-valued)
%    1. Number of times pregnant
%    2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
%    3. Diastolic blood pressure (mm Hg)
%    4. Triceps skin fold thickness (mm)
%    5. 2-Hour serum insulin (mu U/ml)
%    6. Body mass index (weight in kg/(height in m)^2)
%    7. Diabetes pedigree function
%    8. Age (years)
%    9. Class variable (0 or 1)
% 8. Missing Attribute Values: None
% 9. Class Distribution: (class value 1 is interpreted as "tested positive for
%    diabetes")
%    Class Value  Number of instances
%    0            500
%    1            268
```

Téléchargez le fichier http://storm.cis.fordham.edu/ gweiss/data-mining/weka-data/diabetes.arff et sauvegardez-le sur le bureau.

4. Pour charger ce fichier dans l'outil, utilisez le bouton Open File.
   Le panneau **Current Relation** devrait maintenant décrire : le nom de la relation, nombre d'instances (nombre de cas) et le nombre d'attributs.
5. Vous pouvez visualiser le contenu du fichier en procédant comme suit :
   — Fenêtre principale -> Tools -> ArffViewer
   — Une nouvelle fenêtre s'ouvre et vous devez choisir le même fichier avec File -> Open
   — Pour que le contenu soit plus facile à lire faire View -> Optimal Column Width (all).

   Vous pouvez maintenant vérifier que le contenu est bien cohérent avec ce qu'on a vu dans le panneau **Current Relation** (nom de la relation, nombre d'instances (une par ligne), nombre d'attributs (colonnes)).
   Vous pouvez maintenant fermer cette fenêtre.
6. Au niveau de la deuxième fenêtre (Explorer), observez les choses suivantes.
   Le panneau **Attributes** contient la liste des attributs. Si vous cliquez sur un attribut (ne pas cocher, seulement cliquer sur le nom de l'attribut !) on voit apparaître dans le partie de droite de la fenêtre les statistiques (min, max, moyenne, écart-type) et en dessous l'histogramme pour la classe (par défaut, ceci peut être changé avec le menu déroulant).
   En choisissant l'attribut `class`, vous pouvez vérifier que la couleur bleue représente `tested_negative` et le rouge `tested_positive` ! De plus, il y a plus de cas négatifs que positifs.
7. Vous allez maintenant explorer plus en détail deux fonctionnalités de l'outil : la sélection des attributs et la génération d'un arbre de décision.
   (a) Tout d'abord, vous n'allez conserver que les deux attributs `preg` et `class`. Ceci vous permettra d'explorer cette méthode simple de transformation de la relation et puis de construire un arbre de décision largement simplifié qui sera d'autant plus facile à interpréter.
   Dans le panneau **Attributes** cochez les cases `preg` et `class`. En appuyant sur le bouton Invert, vous pouvez inverser la sélection et puis en actionnant le bouton Remove enlever tous les autres attributs. Effectuez ces opérations et assurez-vous qu'il ne reste que les attributs `preg` et `class`.
   (b) Vous allez maintenant construire un arbre de décision pour cette relation simplifiée. Dans l'onglet *classify*, bouton choose choisissez `trees` et `j48`. Appuyez sur start pour faire construire l'arbre.
   Dans le panneau **Result list** sélectionnez, avec le bouton de droite, visualize tree. Une fenêtre contenant l'arbre de décision devrait apparaître. Si jamais l'arbre n'est pas lisible, choisissez fit to screen dans le menu du bouton de droite de cette fenêtre. Observez cet arbre et comparez avec l'histogramme de l'onglet *Preprocess* pour vous assurer que ce résultat semble raisonnable.
   (c) S'il vous reste du temps, revenez en arrière avec le bouton Undo et refaites la construction de l'arbre avec d'autres ensembles d'attributs.