Les données décrivent des informations sociaux-économiques des pays pour l'année 2007. Elles sont fournies par The World Bank Group (http://www.worldbank.org/).

Trois jeux sont fournis: A), B) et C), les jeux B) et C) sont des sous-ensembles de A).

A) countries2007_all.csv
209 pays 48 attributs

B) countries2007_noMissing1.csv
87 pays 32 attributs
Il a été obtenu à partir du jeu A), en supprimant 16 attributs contenant énormément de valeurs manquantes, puis en supprimant les pays pour lesquels parmi les 32 attributs restant il manquait encore au moins une valeur.

C) countries2007_noMissing2.csv
148 pays 20 attributs
Il a aussi été obtenu à partir du jeu A), en supprimant les mêmes 16 attributs que précédemment, puis en supprimant encore 12 autres attributs ayant eux aussi beaucoup de valeurs manquantes, et enfin en supprimant les pays pour lesquels parmi les 20 attributs restant il manquait encore au moins une valeur.

La liste des attributs des trois jeux est donnée dans countries2007_structure.txt, et un descriptif de ces attributs se trouve dans someDefinitions.txt.

OBJECTIF: trouver des combinaisons de traitements qui permettent d'obtenir des modèles et des résultats qui soient cohérents ou surprenants au regard de votre connaissance du domaine.

Quelques pistes:

- voir les opérations et les combinaisons proposées pour les jeux Iris et Bears (notamment Bears).

- utiliser la description des attributs someDefinitions.txt pour choisir des sous-ensembles d'attributs pouvant être intéressants.

- ne pas sous-estimer les composants de visualisation ...

- sélectionner vos propres sous-ensembles de données: les jeux B) et C) sont fournis pour pouvoir faire directement quelques essais, mais attention leur contenu n'a pas été choisi avec un soin particulier. Il est donc tout à fait possible (voire recommandé) de travailler sur des sous-ensembles différents du jeux A). ATTENTION: la gestion des valeurs manquantes dans KNIME n'est pas documentée, et de nombreux algorithmes ne les gèrent pas correctement (tout en fournissant quand même un résultat ... mais qui n'a alors sans doute pas de sens). Il semble donc plus raisonnable dans l'état actuel de ne pas utiliser les composants KNIME, notamment ceux de clustering et de classification, sur des jeux contenant des valeurs manquantes. Pour remplacer des valeurs manquantes ou supprimer des lignes qui en contiennent voir le composant Missing Value.