Sujet TP Fouille de données

Exercice 1 :

On s'intéresse aux informations statistiques sur les 50 états américains dans les années 1970 disponibles dans le fichier usData.csv. Ces données sont associées aux attributs suivants :

- ‘State’: nom de l'état
- ‘Population’: population estimate as of July 1, 1975
- ‘Income’: per capita income (1974) *(moyenne des revenus individuel par habitant)*
- ‘Illiteracy’: illiteracy (1970, percent of population) *(taux d'analphabétisme)*
- ‘Life Exp’: life expectancy in years (1969-71) *(espérance de vie)*
- ‘Murder’: murder and non-negligent manslaughter rate per 100,000 population (1976) *(taux de mort violente pour 100 000 habitants)*
- ‘HS Grad’: percent high-school graduates (1970) *(pourcentage de diplômés du supérieurs)*
- ‘Frost’: mean number of days with minimum temperature below freezing (1931-1960) in capital or large city *(nombre de jours moyen de gel)*
- ‘Area’: land area in square miles *(superficie de l'état)*

1. Réaliser une étude de ces données en vous focalisant sur les attributs *Income* et *HS Grad*. Comparer et analyser les résultats des algorithmes de clustering SimpleKMeans, EM et Canopy.

2. Réaliser une fouille de ces données permettant de mettre en évidence une autre structuration des informations mettant en évidence une relation entre attributs.

Exercice 2 :

On s'intéresse maintenant aux données issues du fichier labor.arff disponible avec weka. Les informations contenues dans ce fichier résument les conditions de travail assurées par des conventions collectives au sein d'entreprises canadienne à la fin des années 80. Chaque convention est évaluée positivement ou négativement par la classe.

Réaliser une classification de ces données à l'aide de l'algorithme J48. Tester les différents paramètres afin de produire l'arbre de décision qui vous semble le plus pertinent, le présenter et expliquer le choix réalisé.

Attendus (en binôme) :

Rapport au format **pdf** présentant les résultats obtenus.