Polytech'Lille
Département GIS

Classification Automatique
TP 3 : Classification sur des données qualitatives

Le fichier *breast_ cancer.csv* contient les informations relatives au cancer du sein pour une population données. les différentes variables du fichier sont les suivantes :
Class, age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat

1. Charger les données, nommées et analyser les différentes variables du fichier.

2. Réaliser une ACM à partir du jeu de données, en considérant comme illustratif le dégré de malinité deg-malig sans prendre en compte la variable 'class'. Commenter.

3. On se propose de réaliser à partir des axes factorielles résultant de l'ACM effectutée précédemment, une segmentation du jeu de données initiale, en considérant la variable 'class' comment étant illustrative.
   - determiner le nombre de classe optimal. On utilisera pour cela la silhouette de la segmentation .
   - proposer une segmentation de notre jeu de données.

   Décrire la segmentation obtenue.

4. A partir de la méthode `kmodes` du package `klaR`, proposer une segmentation des différents individus de la base, avec comme variable illustrative 'class'. Décrire la segmentation retenue

5. Calculer les indices de Rand des deux partitions obtenues, ainsi que celui de chacune des partitions avec la répartion des données faite par la variable 'class'. Conclure