

Module : Analyse de données

Exercice 1 (10Pts) : Nous disposons d'un fichier de 250 employés d'une entreprise donnant pour chacun d'eux :

- | | |
|--|--|
| 1) Le nom et prénom | 6) Nombres d'enfants |
| 2) La date de naissance | 7) Nombres de pièces du logement |
| 3) Le lieu de naissance | 8) Diplôme |
| 4) Le sexe (H ou F) | 9) catégorie (grade) dans l'entreprise |
| 5) Situation familiale (célibataire, marié, divorcé, veuf) | 10) Ancienneté (nombre d'années) |
| | 11) Salaire |

- 1) Parmi ces variables choisissez celles pouvant être utilisées pour effectuer une analyse en composantes principales sur ces employés. Décrivez les calculs nécessaires pour représenter les individus sur le meilleur plan principale.
- 2) Les responsables de l'entreprise souhaitent étudier les relations existantes entre les variables, situation familiale (célibataire, marié, divorcé, veuf) et catégorie (cadre supérieur, technicien, ouvrier, autre). Quelle méthode doit-on utiliser. Expliquez le processus.

Exercice 2 (10Pts) : 100 étudiants ont été notés sur 5 matières. Nous souhaitons regrouper ces étudiants en des classes homogènes.

- 1) Proposer une méthode de classification non hiérarchique qui permettra de classer ces étudiants en exactement K classes. Expliquer le principe.
- 2) Nous souhaitons d'autres parts faire une classification hiérarchique ascendante sur ces étudiants. Quel indice de dissimilarité peut-on utiliser ? Proposez une méthode pour effectuer cette classification.