ENSA
ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES
KHOURIBGA

Université Sultan Moulay Slimane

Rapport

Master Big Data et Aide à la Décision

Module: Datamining

K-means application in java

---

4 juillet 2021

Réalisé par :

Ossama MAJALI, ossama.majali1@gmail.com

Encadré par :

M. Abelmajid DARGHAM

Année académique 2020-2021

Résumé

Ce projet consiste à réaliser une application de bureau construit avec java, c'est une application dédiée à la classification de chaque groupe de points séparément par la méthode k-means.

Abstract

This project consists in creating a desktop application built with java, it is an application dedicated to the classification of each group of points separately by the k-means method.

Table des matières

Introduction

En quelques mots, la classification automatique est la tâche qui consiste à regrouper, de façon non supervisée, un ensemble d'objets ou plus largement de données, de telle manière que les objets d'un même groupe (appelé cluster) sont plus proches (au sens d'un critère de (dis)similarité choisi) les unes aux autres que celles des autres groupes (clusters). Il s'agit d'une tâche principale dans la fouille exploratoire de données, et une technique d'analyse statistique des données très utilisée dans de nombreux domaines, y compris l'apprentissage automatique, la reconnaissance de formes, le traitement de signal et d'images, la recherche d'information, etc.

L'idée de notre algorithme est donc de découvrir des groupes au sein des données, de façon automatique.

Dans ce cadre plusieurs méthodes ont été développées, la plus populaire est celle des k moyennes (K-means), elle doit sa popularité à sa simplicité et sa capacité de traiter de larges ensembles de données.

L'algorithme k-means mis au point par McQueen en 1967, un des plus simples algorithmes d'apprentissage non supervisé, appelée algorithme des centres mobiles, il attribue chaque point dans un cluster dont le centre (centroïde) est le plus proche. Le centre est la moyenne de tous les points dans le cluster, ses coordonnées sont la moyenne arithmétique pour chaque dimension séparément de tous les Points dans le cluster c'est à dire chaque cluster est représentée par son centre de gravité.

Chapitre 1 :

La description de l'algorithme

1- Qu'est-ce que le clustering :

Le clustering est une méthode d'apprentissage non supervisé (unsupervised learning). Ainsi, on n'essaie pas d'apprendre une relation de corrélation entre un ensemble de features d'une observation et une valeur à prédire , comme c'est le cas pour l'apprentissage supervisé. L'apprentissage non supervisé va plutôt trouver des patterns dans les données. Notamment, en regroupant les choses qui se ressemblent.
En apprentissage non supervisé, les données sont représentées comme suit :

$$X = \begin{pmatrix} x_{(1,1)} & x_{(1,2)} & x_{(1,...)} & x_{(1,n)} \\ x_{(2,1)} & x_{(2,2)} & x_{(2,...)} & x_{(2,n)} \\ \cdots & \cdots & \cdots & \cdots \\ x_{(m,1)} & x_{(m,2)} & x_{(m,...)} & x_{(m,n)} \end{pmatrix}$$

Chaque ligne représente un individu (une observation). A l'issu de l'application du clustering, on retrouvera ces données regroupées par ressemblance. Le clustering va regrouper en plusieurs familles (clusters) les individus/objets en fonction de leurs caractéristiques. Ainsi, les individus se trouvant dans un même cluster sont similaires et les données se trouvant dans un autre cluster ne le sont pas.

2- Les types de clustering :

Il existe deux grands types du clustering :

- *Le clustering hiérarchique : d'agglomération **(« bottom-up »)***
- *Le clustering non-hiérarchique : de division **(« top-down »)***

Dans le premier cas, on décompose l'ensemble d'individus en une arborescence de groupes.

Dans le 2ème, on décompose l'ensemble d'individus en K groupes, les algorithmes de ce type peuvent aussi être utilisés comme algorithmes de division dans le clustering hiérarchique.

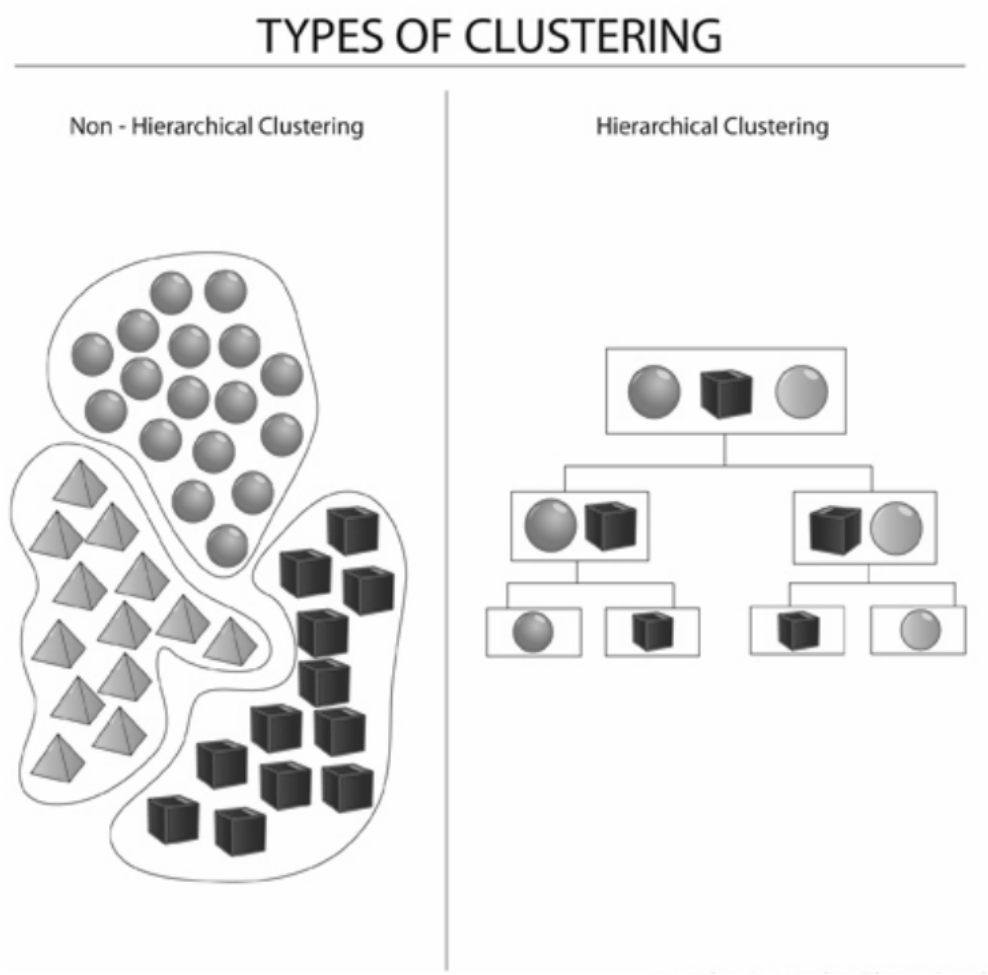


Figure 1: les deux types de clustering non-hiérarchique / hiérarchique

3- Qu'est-ce que K-means :

K-means est un algorithme non supervisé de clustering non hiérarchique. Il permet de regrouper en clusters distincts les observations du data set. Ainsi les données similaires se retrouveront dans un même cluster. Par ailleurs, une observation ne peut se retrouver que dans un cluster à la fois (exclusivité d'appartenance). Une même observation, ne pourra donc, appartenir à deux clusters différents.

L'algorithme :

- *K points sont placés dans l'espace de données d'objet représentant le groupe initial de centroïdes.*

- *Chaque objet ou point de données est affecté au k le plus proche.*
- *Une fois tous les objets affectés, les positions des k centroïdes sont recalculées.*
- *Les étapes 2 et 3 sont répétées jusqu'à ce que les positions des centroïdes ne bougent plus.*

Chapitre 2 :

Modélisation UML de l'application

1- Diagramme de cas d'utilisation :

Le diagramme de cas d'utilisation se sert de certain concept comme : Cas d'utilisation ou use case de anglais qui est l'ensemble de séquences d'actions réalisées par le système produisant un résultat observable intéressant pour un acteur particulier. Collection de scénarios reliés par un objectif utilisateur.

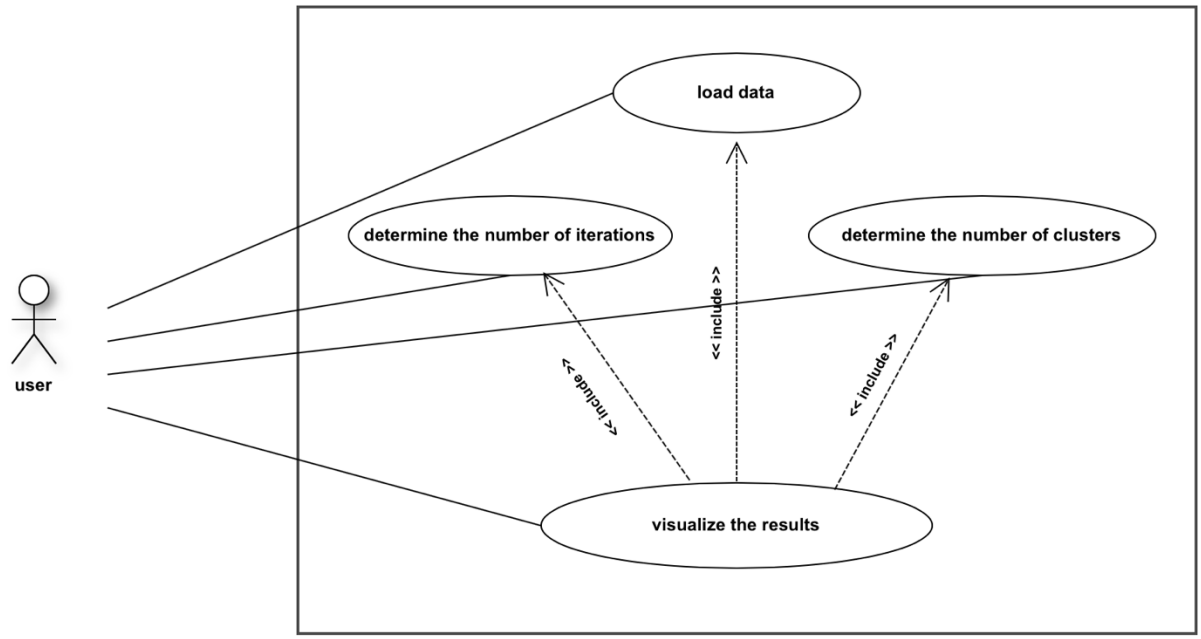


Figure 2 : Cas d'utilisation du système en étude.

La figure 2 présente les différents cas d'utilisation dont il sera question dans le présent travail.

Dans un premier temps, l'utilisateur devra télécharger les données puis il faudra sélectionner le nombre de groupes et d'itérations de l'algorithme pour visualiser les résultats à la fin.

2- Diagramme d'activité :

On peut représenter le fonctionnement de mon algorithme k-means de la façon suivante :

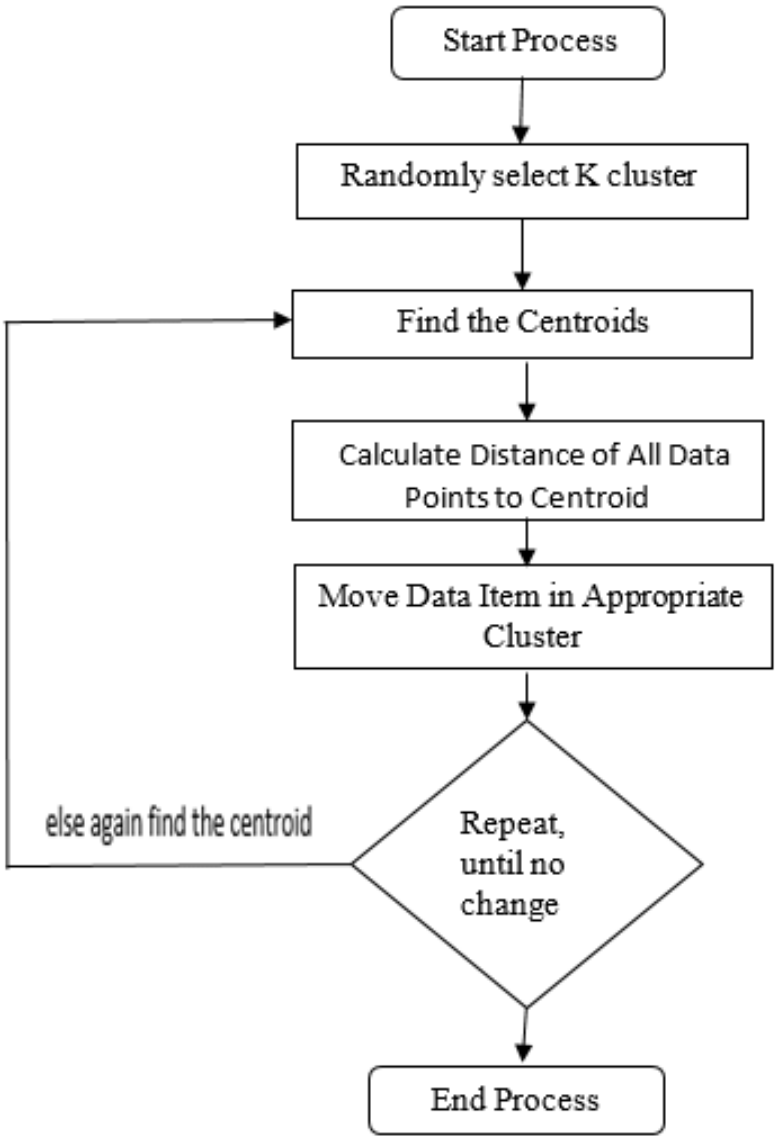
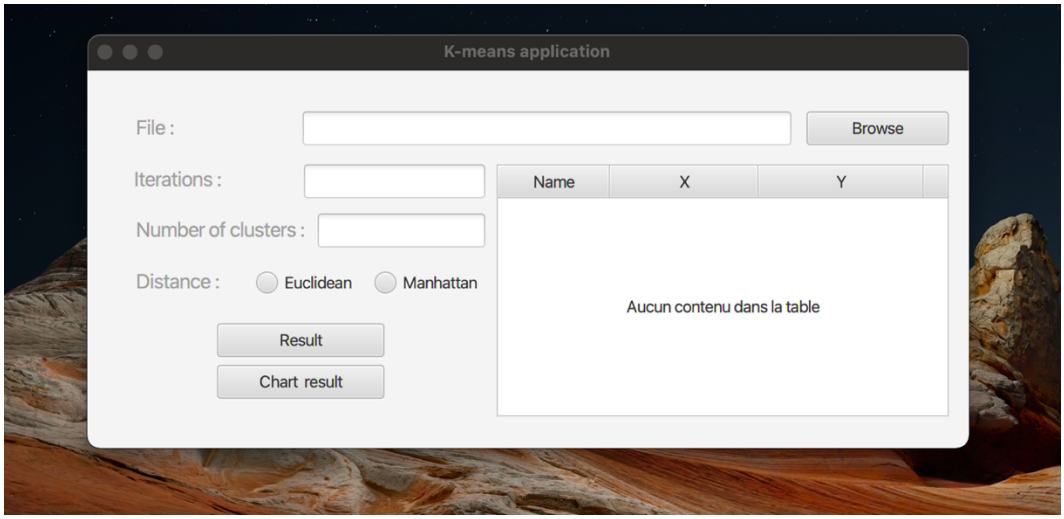


Figure 3 : diagramme d'activité correspondant à l'exécution de l'algorithme k-means

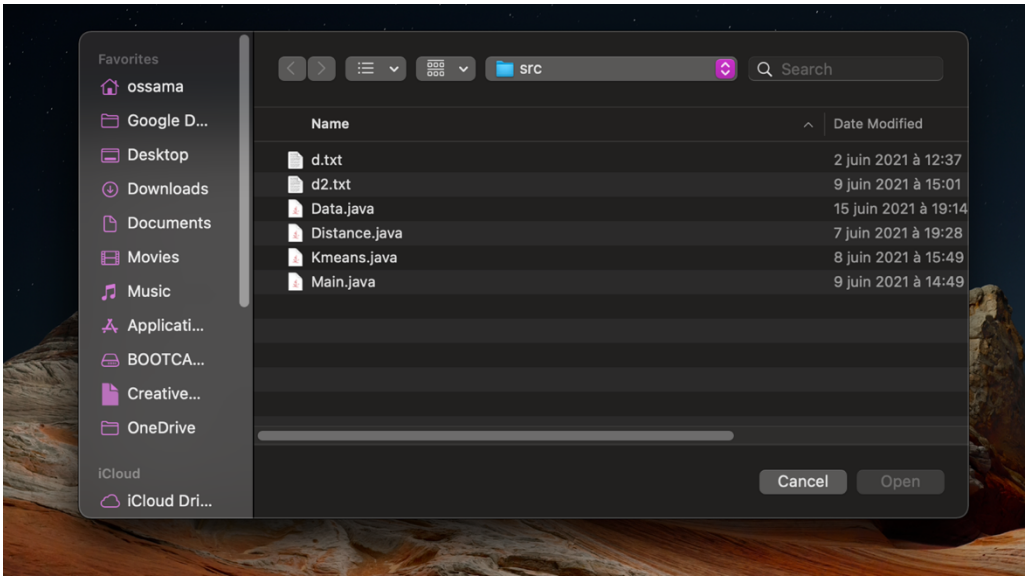
Chapitre 3 :

La description de l'interface graphique

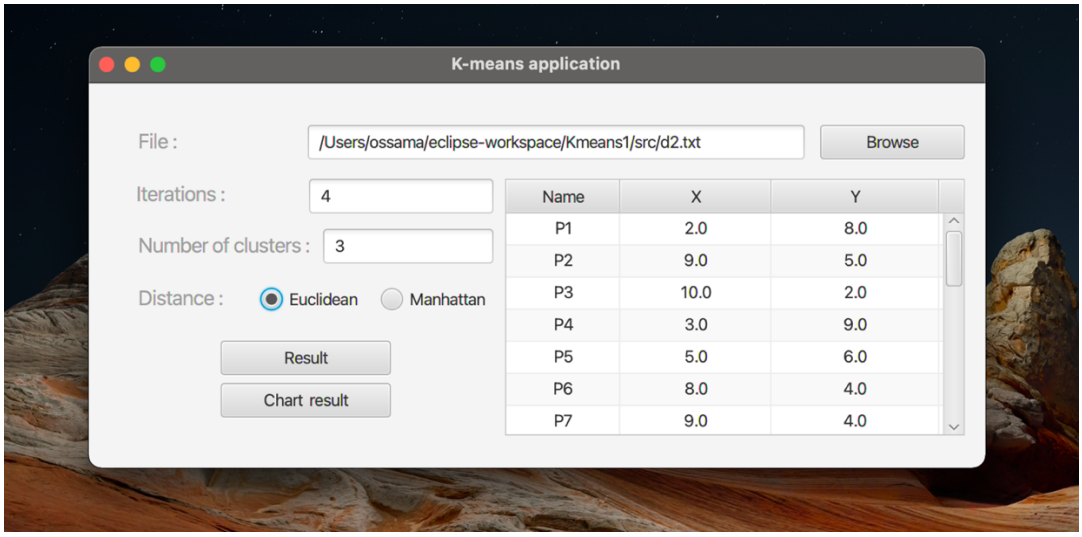
Étape 1 : Ouverture de l'application.



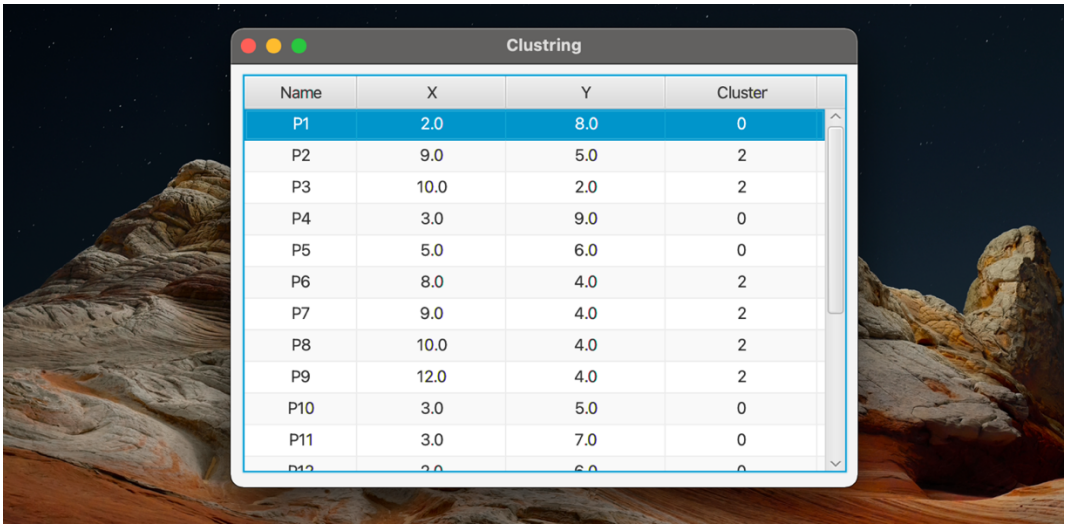
Étape 2 : Chargement des données



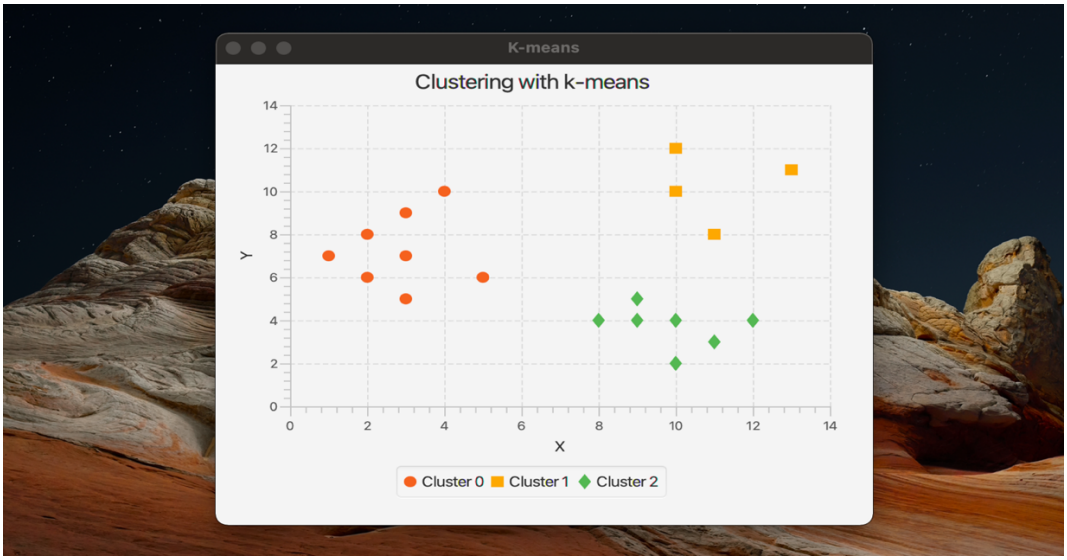
Étape 3 : Choisir le nombre des itérations et les clusters avec la distance.



Étape 4 : Résultats de l'algorithme dans une table.



Étape 5 : Résultats de l'algorithme dans une graphique.



Conclusion

Dans plusieurs domaines des sciences sociales, nous sommes amenés à constituer des groupes homogènes en leur sein et qui diffèrent suffisamment l'un de l'autre. C'est l'objet des méthodes de classification dont fait partie la méthode des k-means, cet algorithme est une version améliorée et randomisée de la méthode des nuées dynamiques. Il est actuellement l'un des plus utilisés et des plus efficaces en analyse des données. De fait, il permet de partitionner une population finie d'éléments en un nombre K (entier) de classes homogènes.

Il est utile de noter que l'algorithme k-means est très performant en termes de temps d'exécution, mais il souffre du problème de dépendance des résultats aux choix effectués lors de l'initialisation.

Le travail est archivé dans un ***Github*** *repository :* ***K-means application link***