

Projet Optimisation

**La classification non
supervisée (clustering)**

01. Introduction

Présentation de l'équipe

02. Problème UKP

Algorithmes Implémentés et
présentation des interfaces

03. Problème du clustering

Présentation du problème du
clustering

04. Etat de l'art

Solutions populaires du
clustering


05. Solution

Solution proposée et
l'implémentation de l'algorithme


06. Etude comparative

Comparaison avec Kmeans sur
deux datasets


Notre Équipe




AMIEUR Nardjes
Implémentation et
tests de l'algorithme




AOUADJ Iheb Nassim
Mise en page de l'article
et présentation




ASSENINE Mohamed Sami
Algorithme k-means et
ACP



AZIZ Rezak
Calibrage des
paramètres



BOUMENDIL Anaïs
Etat de l'art et
rédaction de l'article



MEDJADJI Chaimaa
Implémentation et
tests de l'algorithme

2. Problème UKP

L'objectif est de sélectionner un sous-ensemble qui maximise le bénéfice total des objets sélectionnés et dont la somme des poids ne dépasse pas la capacité.

$$\sum_{i=1}^n x_i p_i \quad \text{sous la contrainte} \quad \sum_{i=1}^n x_i w_i \leq C$$

(p_1, p_2, \dots, p_n) : les profits

(w_1, w_2, \dots, w_n) : les poids des objets

(x_1, x_2, \dots, x_n) : le objets

```
graph LR; UKP((UKP)) --> 1[1 Méthodes exactes]; UKP --> 2[2 Algorithmes heuristiques]; UKP --> 3[3 Algorithmes méta-heuristiques];
```

UKP

1

Méthodes exactes

- Branch and bound
- Programmation dynamique

2

Algorithmes heuristiques

- Heuristique par arrondi
- Density ordered greedy
- Weighted ordered greedy
- Recherche locale

3

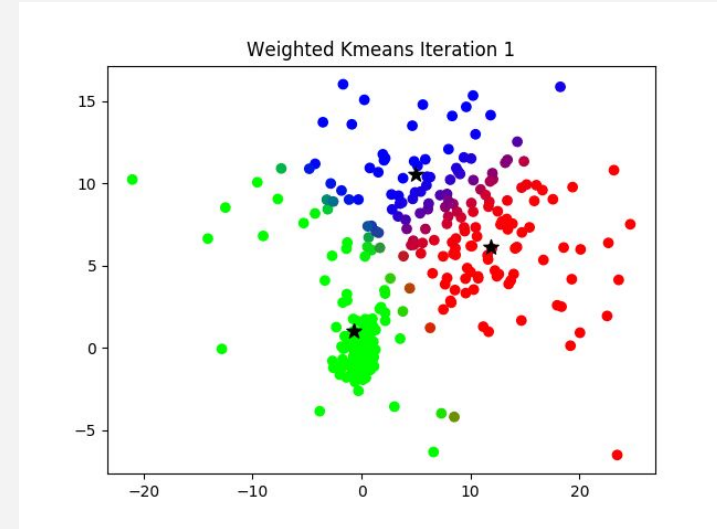
Algorithmes méta-heuristiques

- Le recuit simulé
- ALGORITHME génétique
- Colonie de fourmis

3. Problème du clustering



La classification non supervisée (Clustering) vise à découvrir des groupes de données similaires et à les regrouper en clusters



Pourquoi les méta-heuristiques dans le Clustering ?

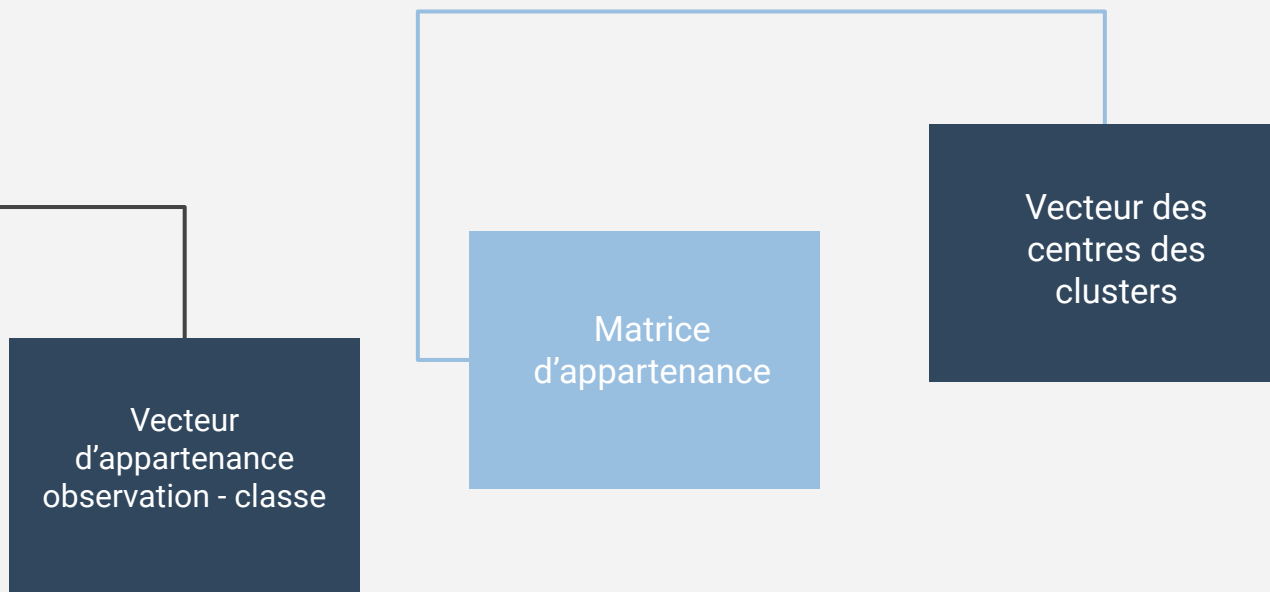
```
graph TD; A[Le clustering est un problème NP-difficile] --- B[Un bon compromis entre le temps d'exécution et la qualité de la solution]; A --- C[Puissance des métaheuristiques]
```

Un bon compromis
entre le temps
d'exécution et la qualité
de la solution

Le clustering est
un problème
NP-difficile

Puissance des
métaheuristiques

Codage de la solution du problème

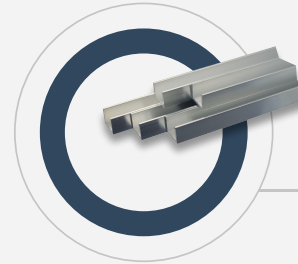


4. État de l'art



Algorithmes Génétique

Hybridation avec
l'algorithme K-means



Recuit simulé



Colonies de fourmis

Inspiré du comportement
des fourmis

5. Solution Biogeography-Based Optimisation (BBO)

01

Une méthode inspirée de la biogéographie

02

Étudie la répartition spatiale des espèces vivantes

03

Une métaheuristique à base de population

04

Manipule une population composée d'îles

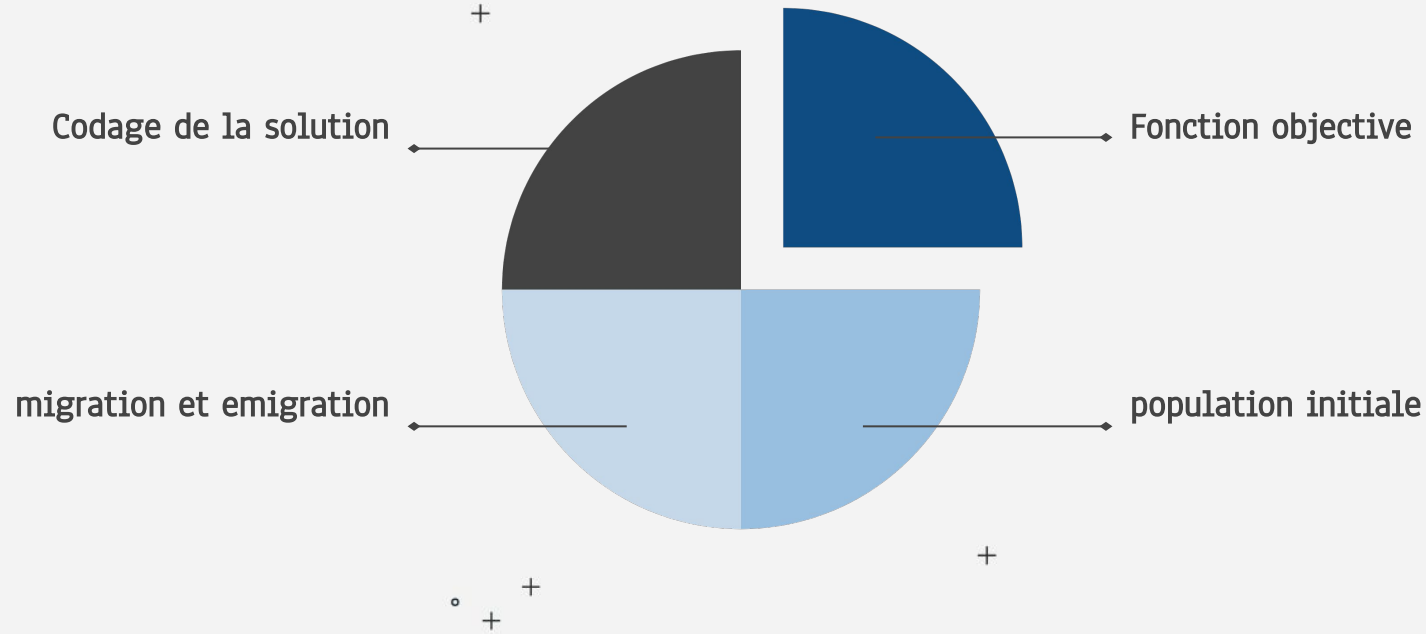
05

La fitness de chaque île est représentée par son HSI

06

Opérations d'immigration et émigration

5. Adaptation du BBO au problème du clustering



Codage de la solution

Représentation d'une population



Caractéristiques du centre du premier cluster dans l'espace des variables

Population initiale

Chaque habitat est initialisé par les coordonnées des centres des k clusters retournés par K-means suivant l'encodage préalablement décrit. Pour chaque habitat, nous faisons varier les paramètres de k-means pour différencier les solutions.

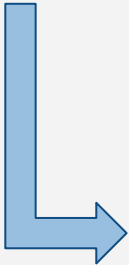
Migration et Emigration

- Descendants à partir des individus de la population.
- Le nombre et la destination des migrants est décidé aléatoirement en fonction des taux d'émigration et d'immigration .
- Le taux d'émigration μ représente les départs vers l'extérieur :
$$\mu = (\text{la taille de la population} + 1 - (\text{indice de la population})) / (\text{la taille de la population} + 1)$$
- Le taux d'immigration λ représente les arrivées venant de l'extérieur : $\lambda = 1 - \mu$

Fonction Objective

Matrice des distances

Distance entre individu 1 et le centre 1	Distance entre individu 1 et le centre 2	Distance entre individu 1 et le centre 3
Distance entre individu 2 et le centre 1	Distance entre individu 2 et le centre 2	Distance entre individu 2 et le centre 3
Distance entre individu 3 et le centre 1	Distance entre individu 3 et le centre 2	Distance entre individu 3 et le centre 3



Matrice distances minimales

Distance entre individu 1 et le centre 1
Distance entre individu 2 et le centre 3
Distance entre individu 3 et le centre 2

HSI= Distance entre individu 1 et le centre 1+Distance entre individu 2 et le centre 3+Distance entre individu 3 et le centre 2

Algorithme BBO

Mutation

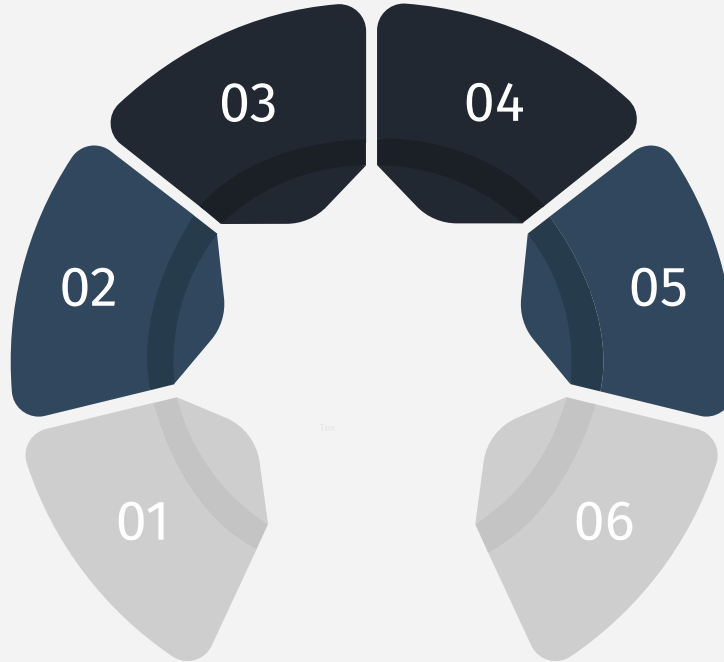
insérer des variables aléatoires au niveau des populations

Migration & émigration

Suivant les paramètres λ et μ de chaque population

Population initiale

À Partir des résultats du kmeans



Evaluation

Évaluer les populations en utilisant la fonction objective HSI

Élitisme

garder les meilleurs population avant d'effectuer remplacement

Mise à jour de la solution

a chaque génération garder la meilleur solution

Présentation des datasets



IRIS

- Classification des fleurs
- 3 classes
- 4 attributs
- 150 observations



Heart

- Classification des individus pour l'étude de cardiopathie
- 2 classes
- 14 attributs
- 303 individus

Paramètres du BBO

■
**Taille de la
population**

■
**Nombre
d'itérations**

■
Élitisme

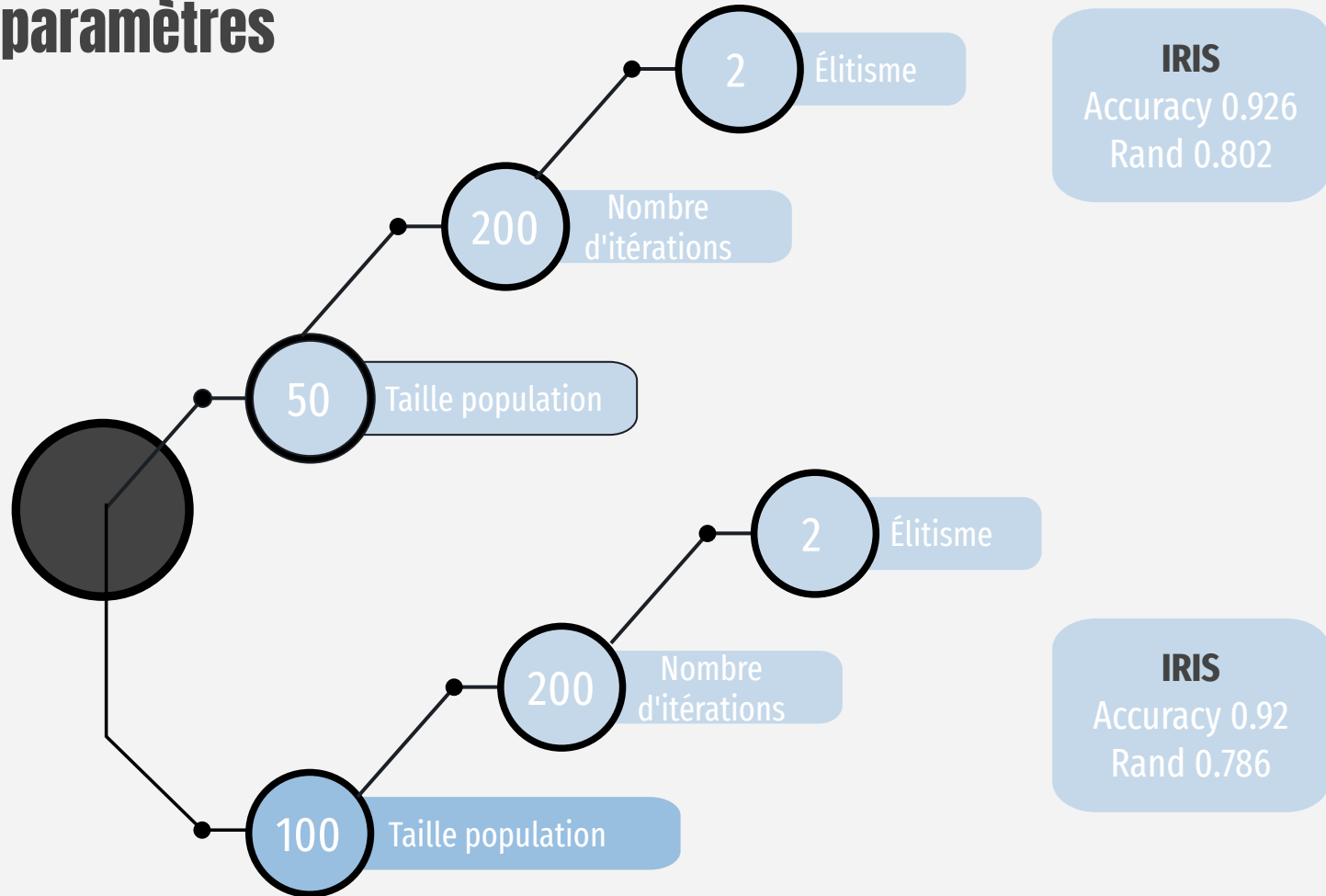
■
Mutation

Métriques utilisées

- La justesse (accuracy)
- Indice de Rand

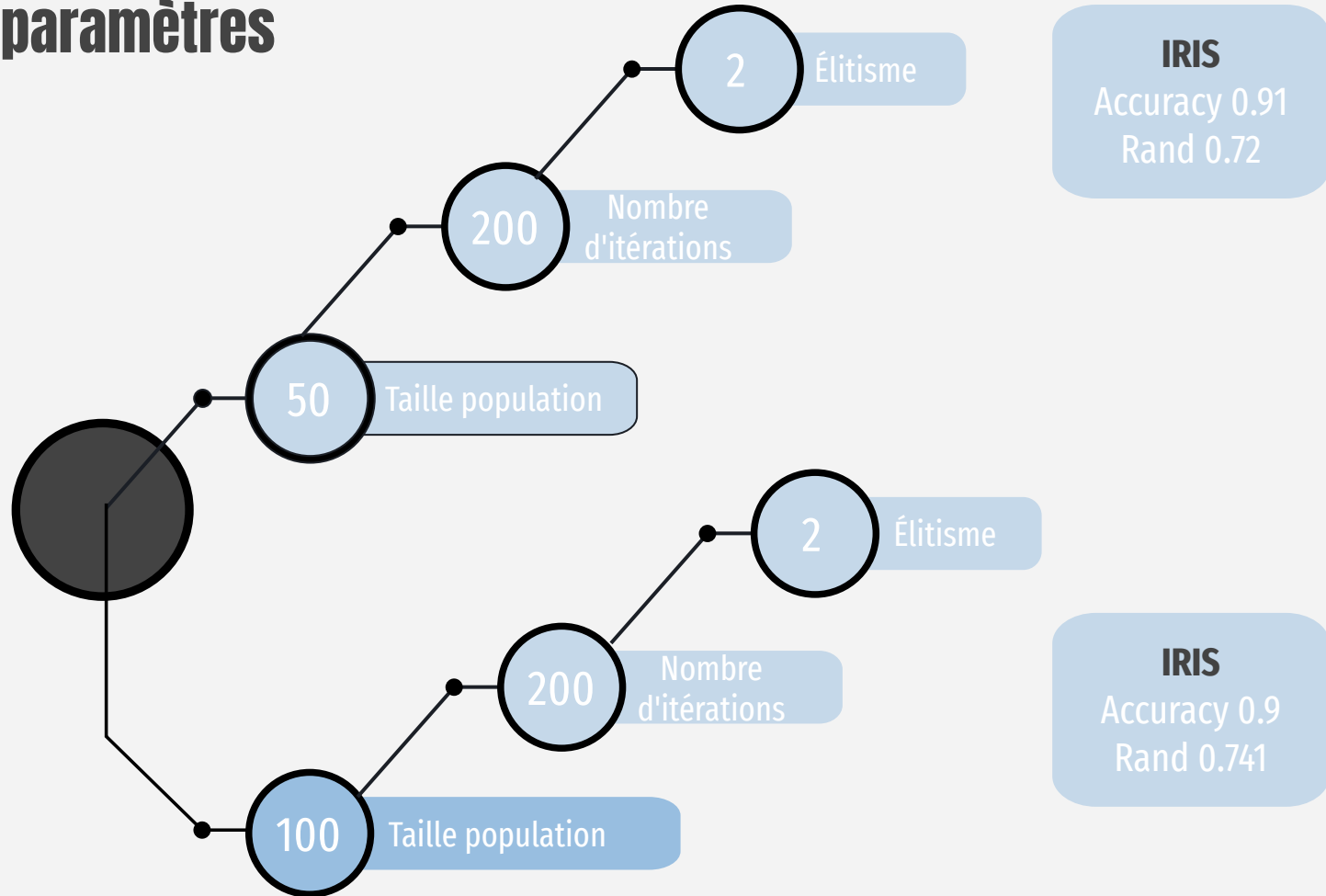
Calibrage des paramètres

Avec mutation



Calibrage des paramètres

Sans mutation

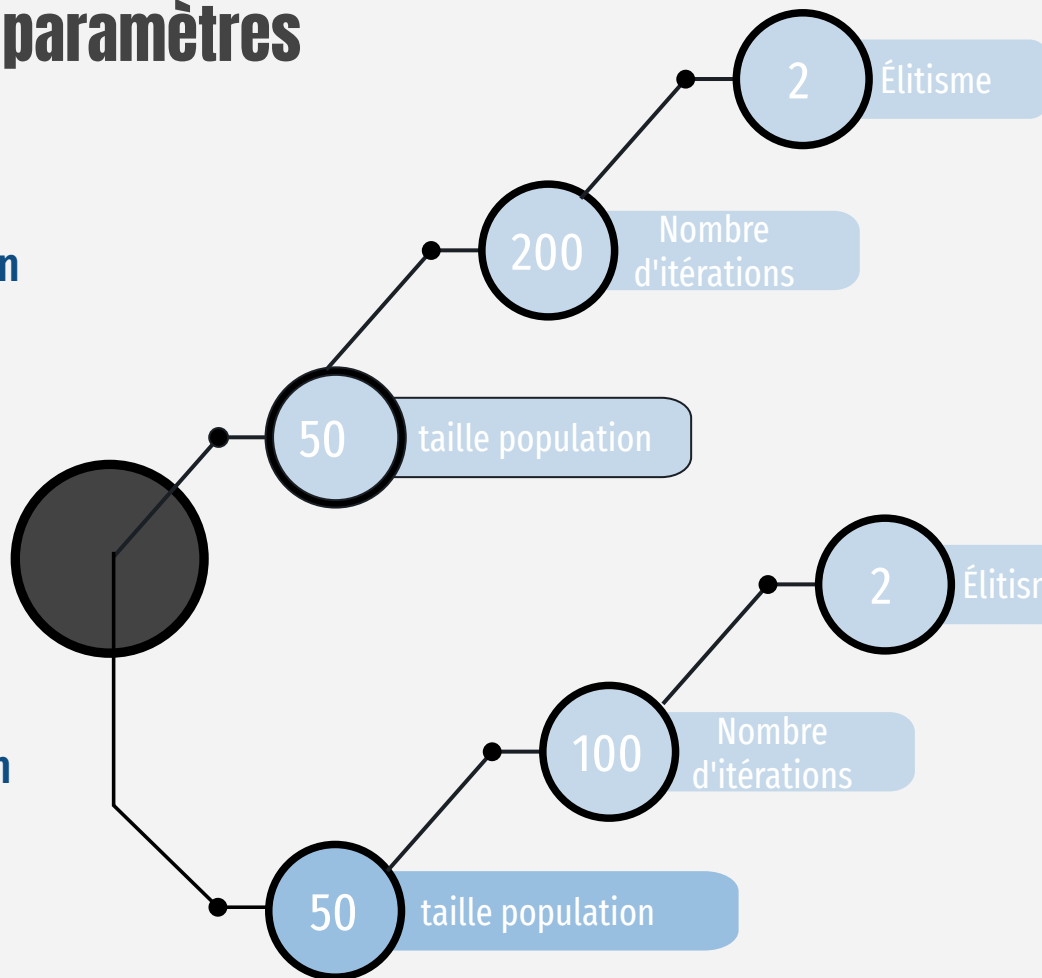


Population initiale

- Première moitié de la population initialisé par Kmeans avec :
Random State = 3 et une faible altération
- Deuxième moitié de la population initialisé par Kmeans avec :
Random State variant et une faible altération

Calibrage des paramètres

Avec mutation



Heart
Accuracy 0.7623
Rand 0.2729

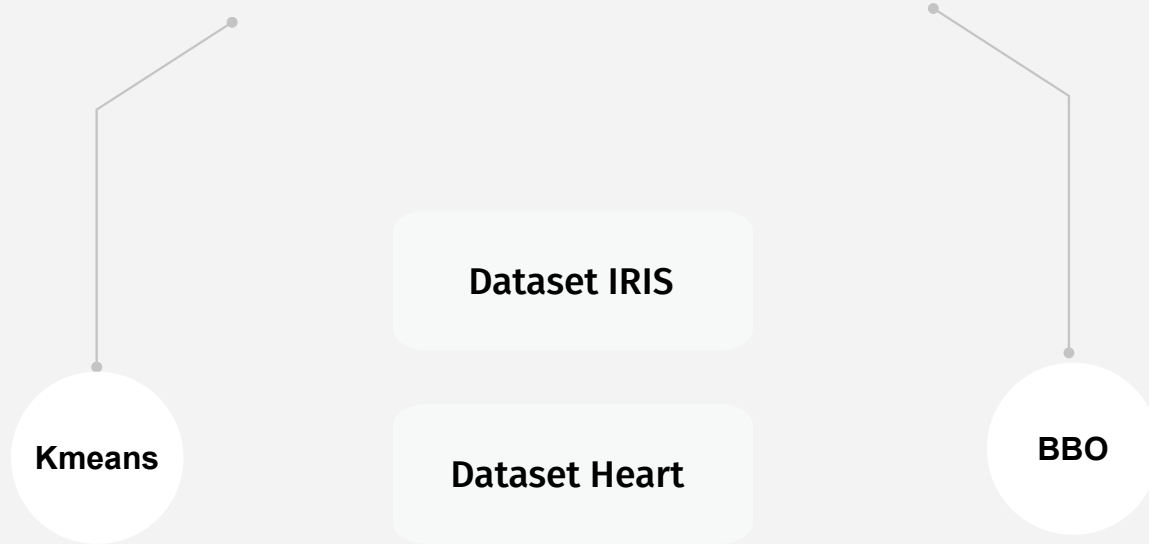
Sans mutation

Heart
Accuracy 0.84
Rand 0.47

Population initiale

- Première moitié de la population initialisé par Kmeans avec :
Random State = 8 et sans altération
- Deuxième moitié de la population initialisé par Kmeans avec :
Random State variant et sans altération

6. Étude comparative



Dataset Iris

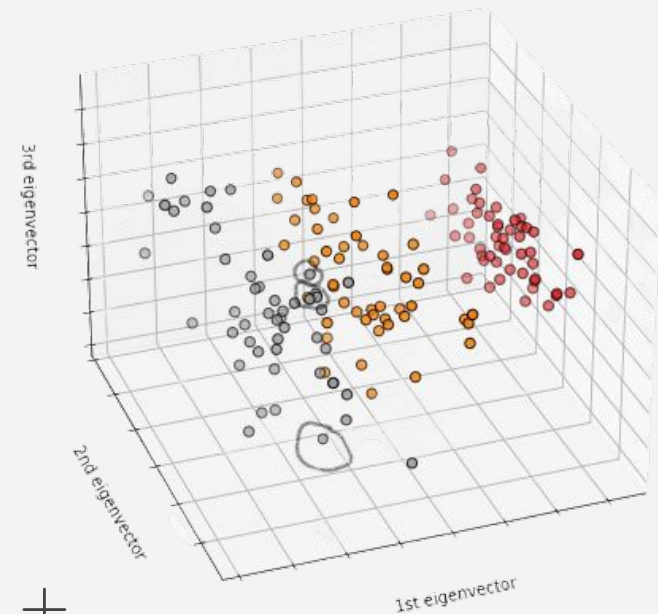
	Mal classifié	Accuracy	HSI	Rand Score
Kmeans	16	0.89333	97.204	0.7302
BBO	10	0.9333	99.823	0.7714

Dataset Heart

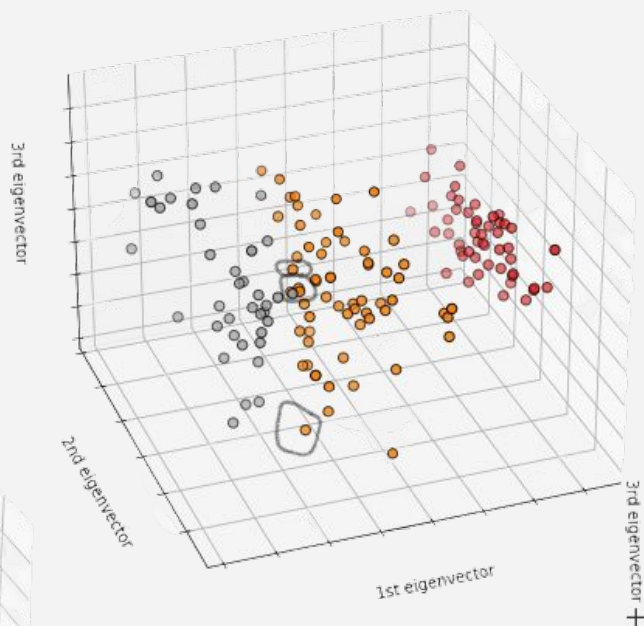
	Mal classifié	Accuracy	HSI	Rand Score
Kmeans	128	0.577	11931.44	0.0205
BBO	72	0.7623	92878.48	0.2729

Dataset IRIS

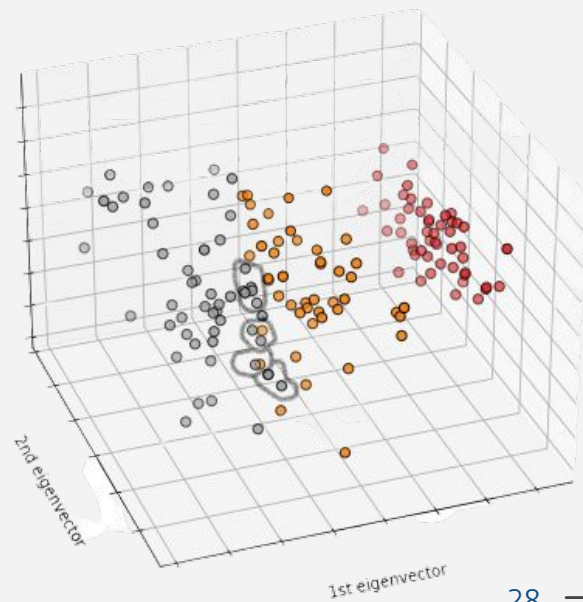
Dataset Original



Résultat Kmeans

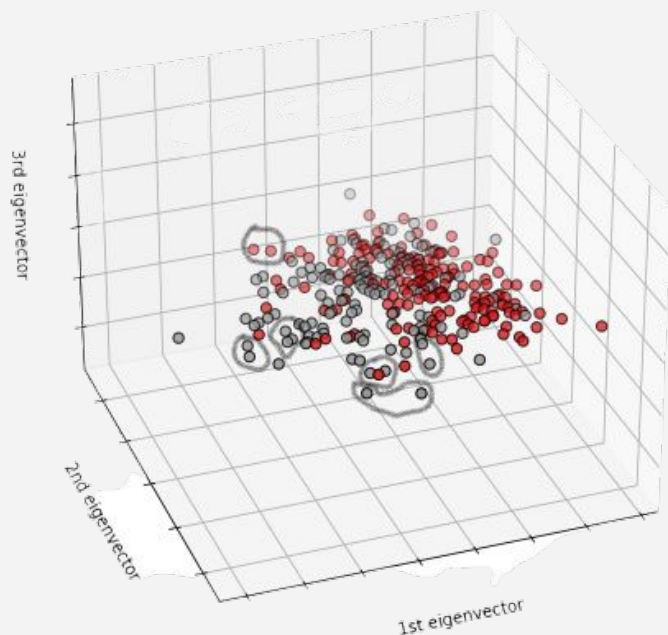


+ Résultat BBO

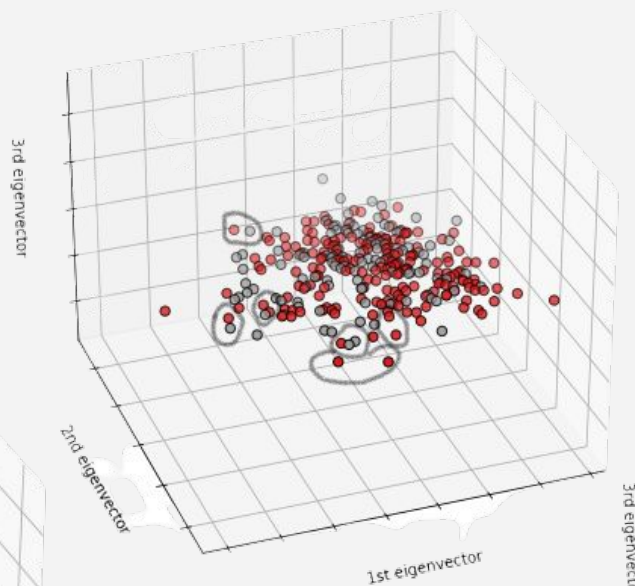


Dataset Heart

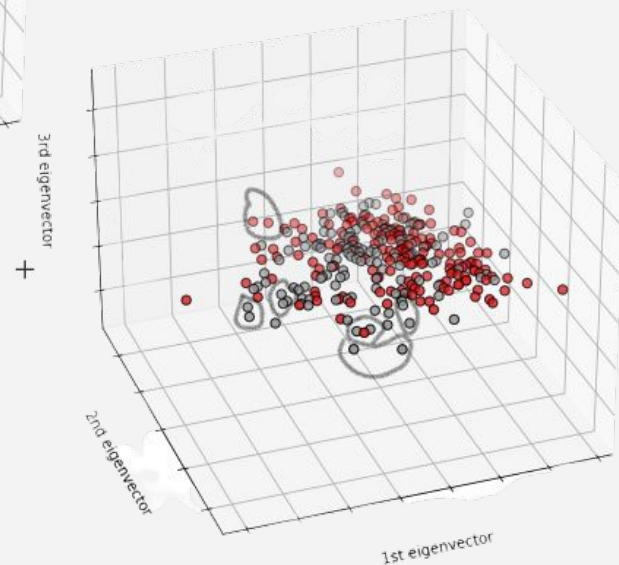
Dataset Original



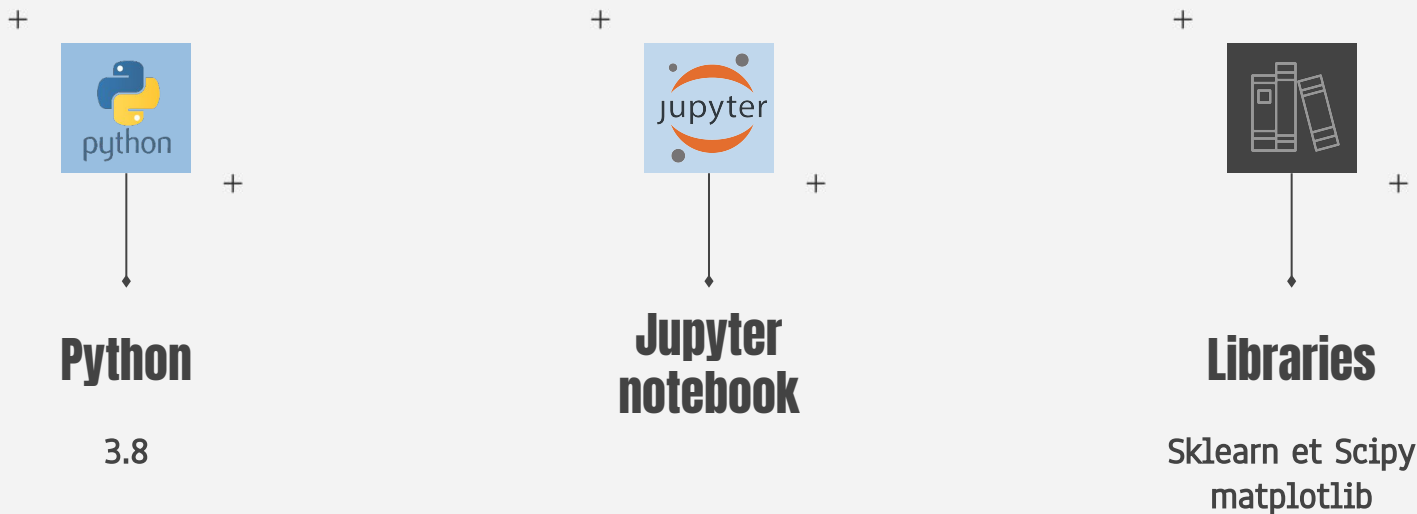
Résultat Kmeans

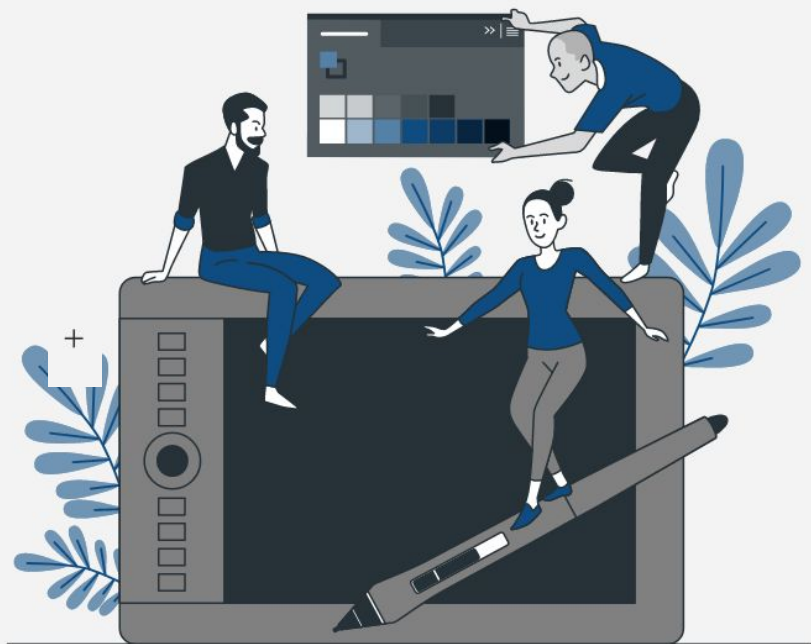


Résultat BBO



Environnement d'exécution

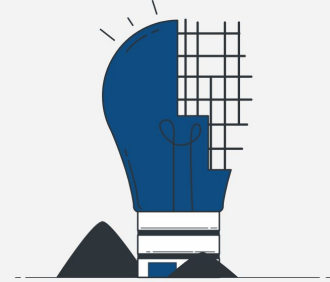




Démonstration

Références bibliographiques

- Vijay Kumar, Jitender Kumar Chhabra, Dinesh Kumar. Initializing Cluster Center for K-Means Using Biogeography Based Optimization, 2011.
- Raju Pal, Mukesh Saraswat. Data clustering using enhanced biogeography-based optimization, 2017.
- Allam Farida et Attab Saida. Hybridation de la méthode des k-means avec le recuit simulé, Université de Mouloud Mammeri, Tizi-Ouzou, 2009.
- Arbia Djamila. Métaheuristiques appliquées à la classification non supervisée de données, Université Mohamed Boudiaf, M'SILA, 2019.





Merci Pour Votre Attention !

Des questions?

