

Clustering K-Means



Tamim EL AHMAD et Mickaël ADIOKO

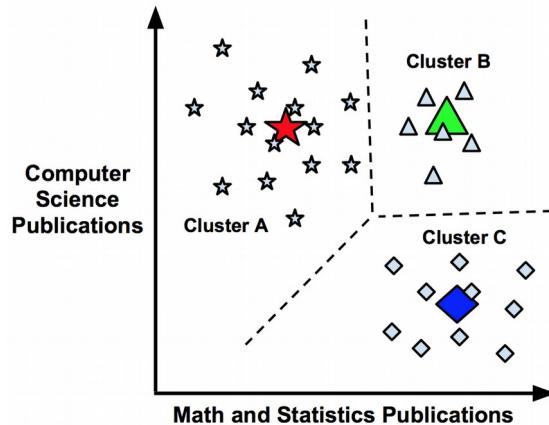
Plan

Introduction : K-Means ?

- 1) Choix du set de données
- 2) Quelques exemples de clustering et mise en évidence de la sensibilité à l'initialisation
- 3) Choix du nombre K de *clusters*

Conclusion

Introduction : K-Means ?



But : minimiser la fonction de perte quadratique : $J = \sum_{i=1}^K \sum_{X_j \in C_i} \|X_j - \mu_i\|^2$

Autre critère d'évaluation, l'indice de Davies-Bouldin : $DB = \frac{1}{K} \sum_{i=1}^K \max_{j \neq i} \left(\frac{S(C_i) + S(C_j)}{\|\mu_j - \mu_i\|} \right)$

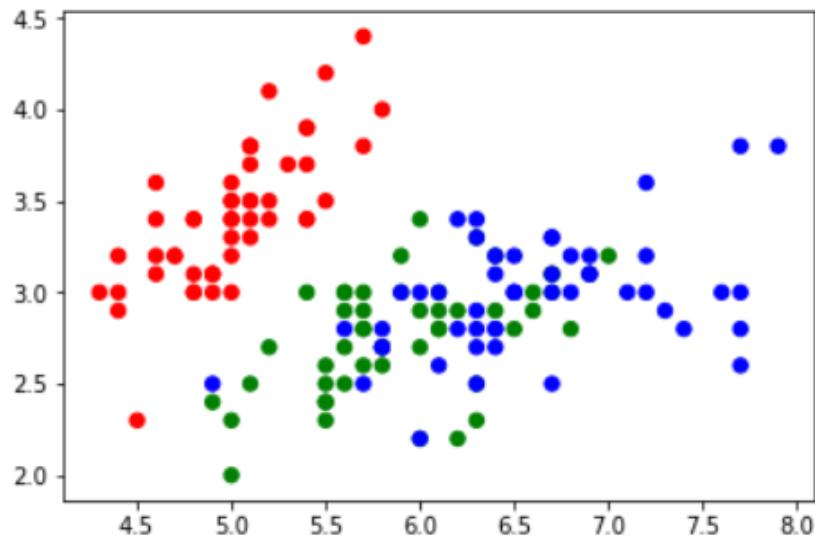
$$\text{où } S(C_i) = \frac{1}{|C_i|} \sum_{X \in C_i} \|X - \mu_i\|$$

1) Choix du set de données

```
Entrée [57]: # On regarde le set de données et ses différents attributs

print(iris)
print(iris.data)
print(iris.feature_names)
print(iris.target)
print(iris.target_names)
print(iris.data.shape)

[6.9 3.1 5.4 2.1]
[6.7 3.1 5.6 2.4]
[6.9 3.1 5.1 2.3]
[5.8 2.7 5.1 1.9]
[6.8 3.2 5.9 2.3]
[6.7 3.3 5.7 2.5]
[6.7 3. 5.2 2.3]
[6.3 2.5 5. 1.9]
[6.5 3. 5.2 2. ]
[6.2 3.4 5.4 2.3]
[5.9 3. 5.1 1.8]]
['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]
['setosa' 'versicolor' 'virginica']
(150, 4)
```

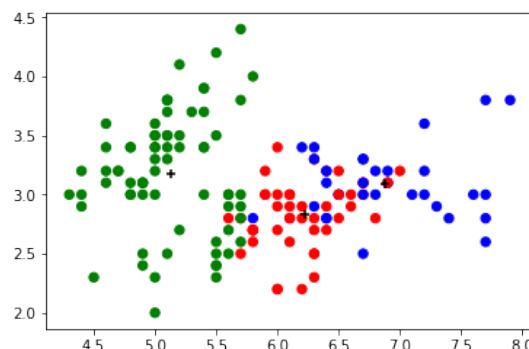


Épaisseur des sépales en fonction de la longueur des sépales,
chaque type de fleur est représenté par une couleur,
ici R : *Setosa*, V : *Versicolour*, B : *Virginica*

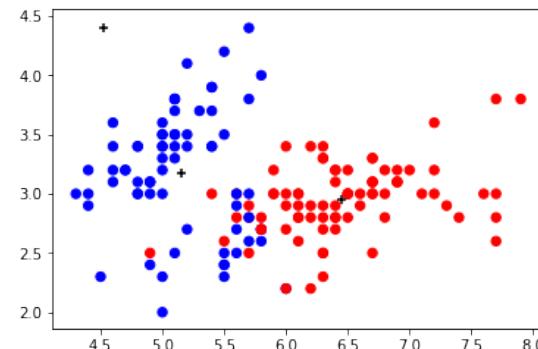
2) Quelques exemples de clustering et mise en évidence de la sensibilité à l'initialisation

Dans toute la partie 2, K=3.

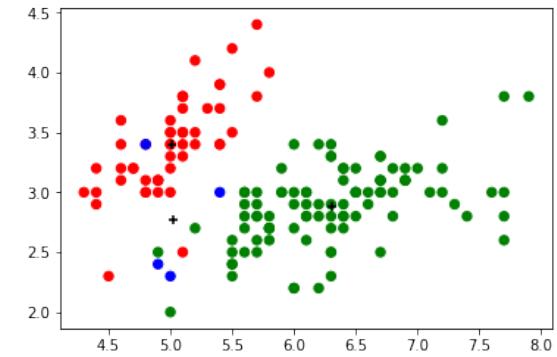
a) Initialisation aléatoire



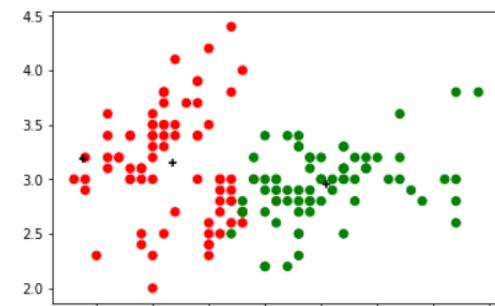
0 : S : 71, Ver : 47, Vir : 32



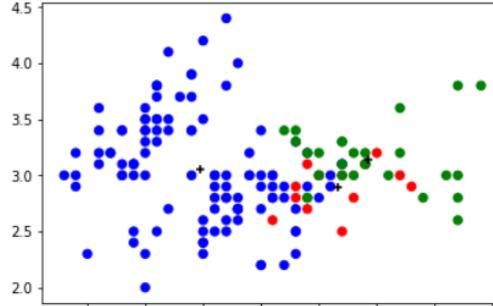
1 : S : 70, Ver et Vir : 80



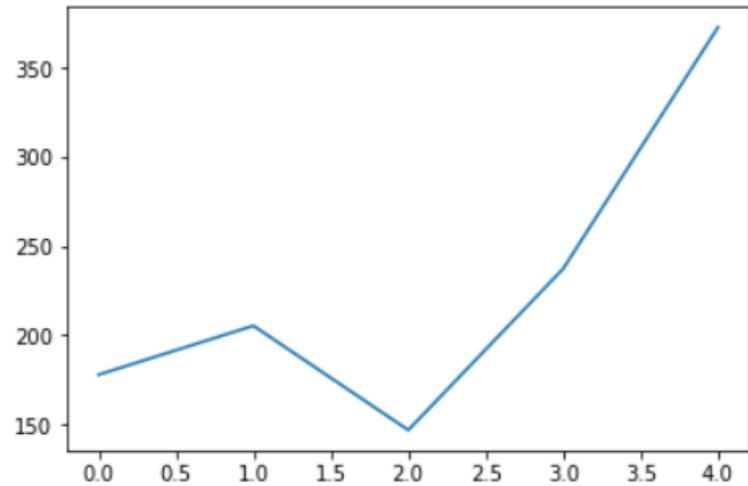
2 : S : 50, Ver : 4, Vir : 96



4 : S : 77, Ver et Vir: 73

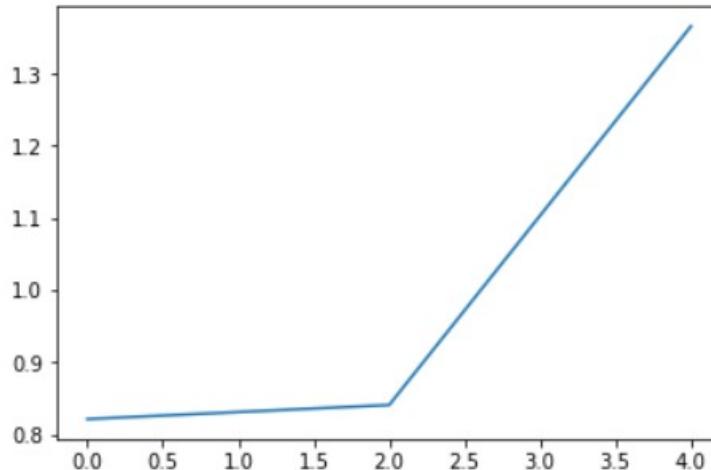


5 : S : 109, Ver : 13, Vir : 28

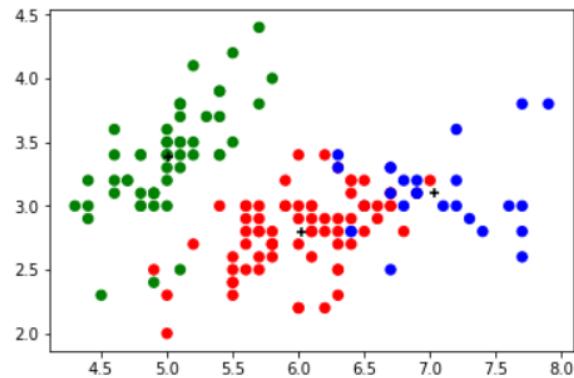


Erreurs quadratiques finales de ces 5 partitionnements

Valeurs de Davies-Bouldin pour les *clusterings* 0, 2 et 4

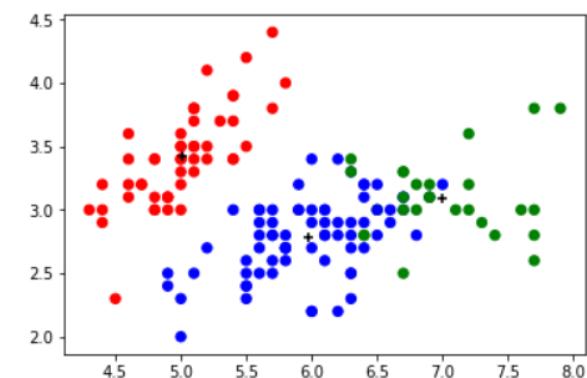


b) *Global K-Means*, initialisation par le mal classé et approche incrémentale



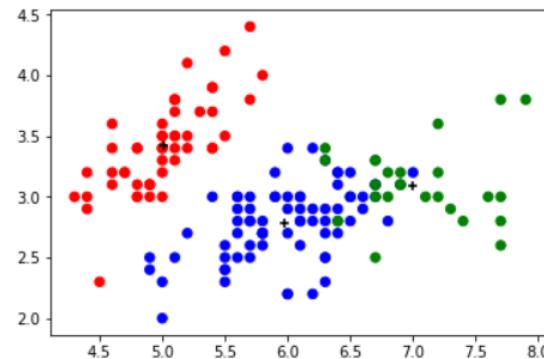
S : 52, Ver : 72, Vir : 26

Global K-Means



S : 50, Ver : 72, Vir : 28

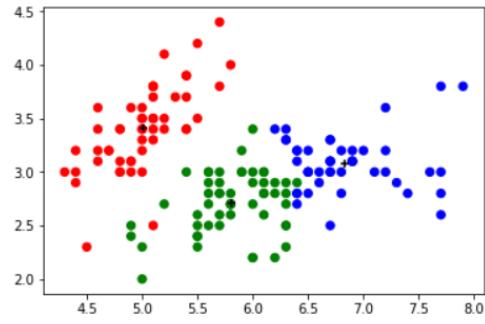
Initialisation par le mal classé



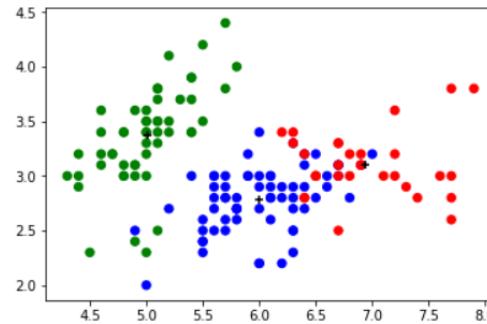
S : 50, Ver : 72, Vir : 28

Approche incrémentale

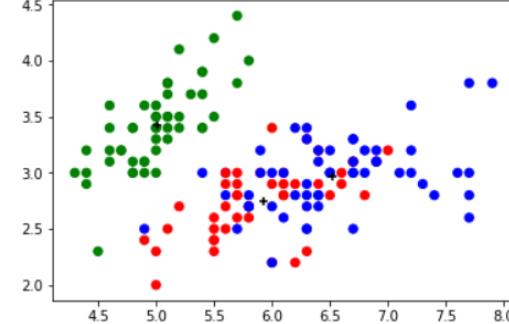
c) *K-Means ++*



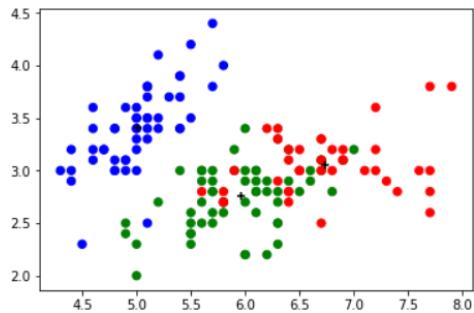
0 : S : 51, Ver : 54, Vir : 45



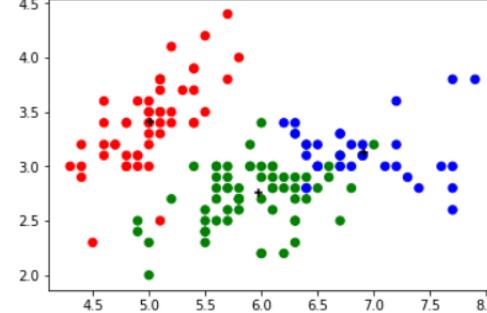
1 : S : 53, Ver : 66, Vir : 31



2 : S : 50, Ver : 44, Vir : 56



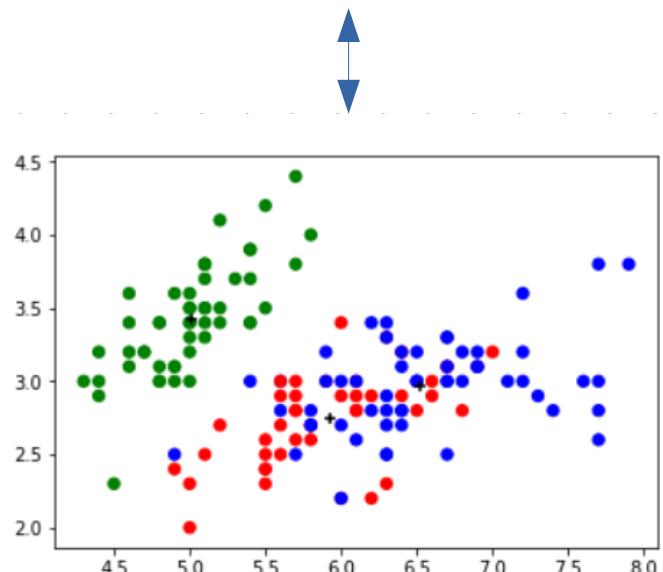
3 : S : 51, Ver : 59, Vir : 40



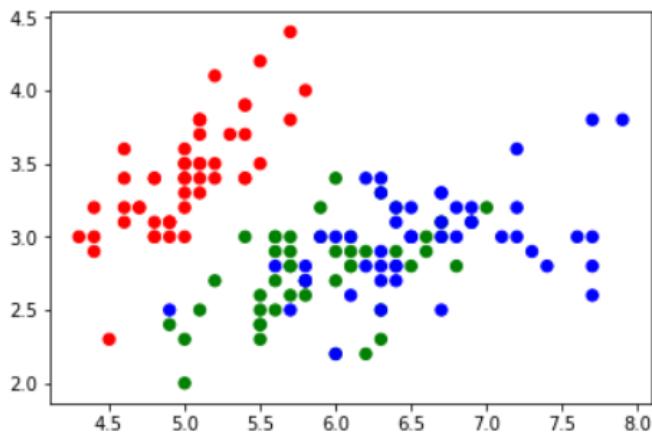
4 : S : 51, Ver : 67, Vir : 32

```
Entrée [79]: print(clusters_size2_plusplus)
print(err_quadr(iris.data, clusters2_plusplus, centers2_plusplus))
print(DB(iris.data,clusters2_plusplus,clusters_size2_plusplus,centers2_plusplus))

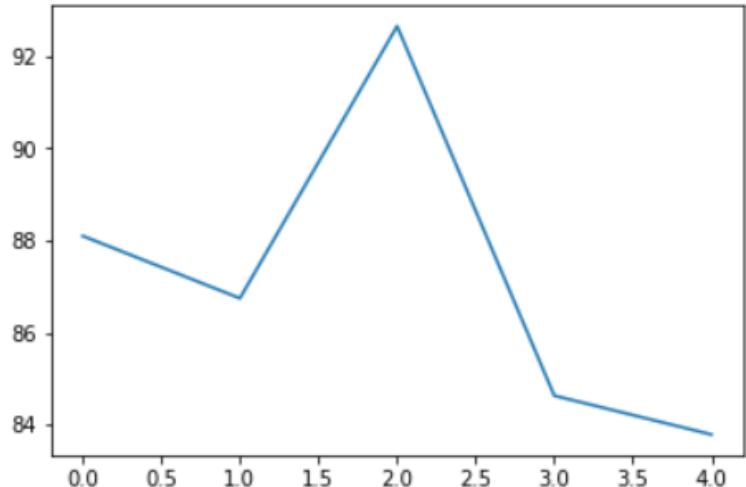
print(clusters2_plusplus)|
```



Clustering 2 obtenu par K-Means ++

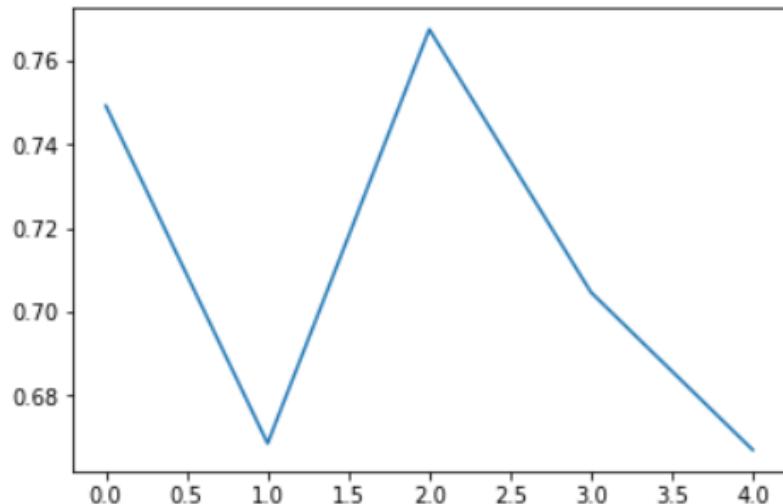


Vraie répartition des classes

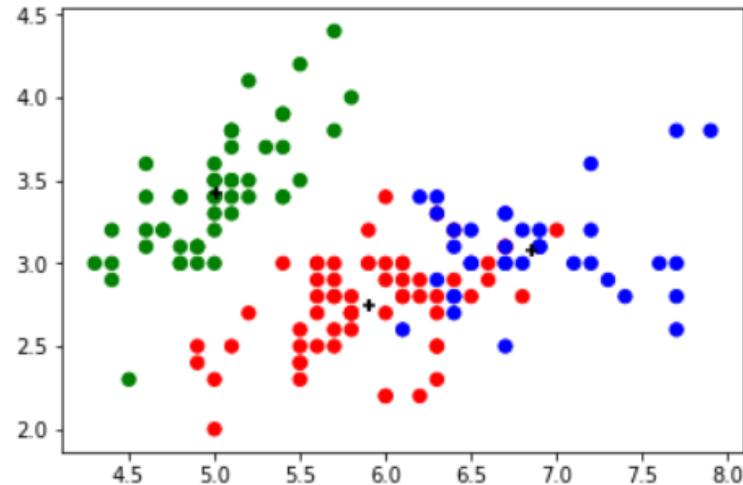


Erreurs quadratiques finales de ces 5 partitionnements obtenus par K-Means ++ : le candidat 2 est le plus mauvais !

Valeurs de Davies-Bouldin pour les *clusterings* obtenus par K-Means ++, là encore : le candidat 2 est le plus mauvais !



K-Means ++ de Scikit-Learn



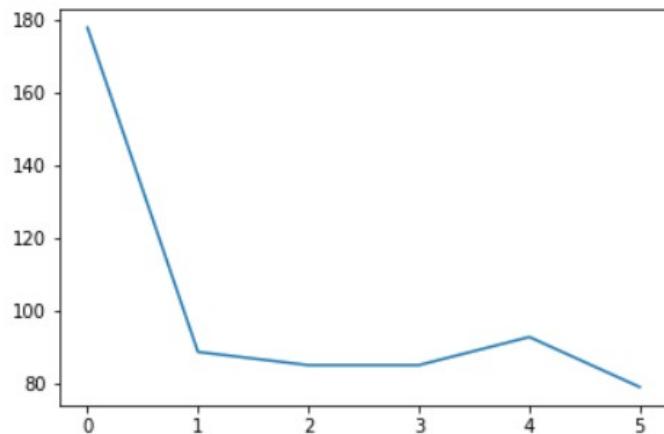
S : 50, Ver : 62, Vir : 38

d) Comparaison des méthodes

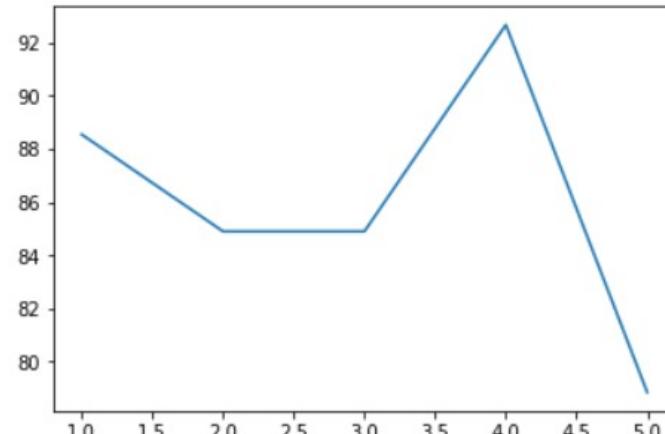
Erreur quadratique

Abscisses :

- 0 : partitionnement aléatoire 0 (le meilleur)
- 1 : *Global K-Means*
- 2 : Initialisation par le mal classé
- 3 : Approche incrémentale
- 4 : *K-Means ++* (partitionnement 2, le meilleur)
- 5 : *K-Means ++* (celui de *Scikit-Learn*)

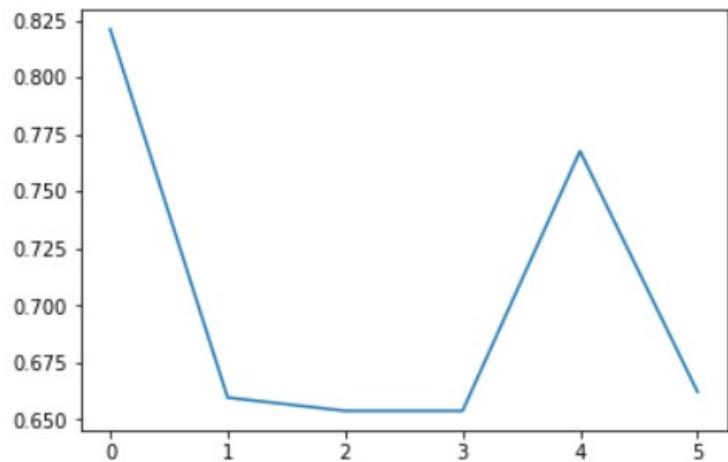


Avec le partitionnement 0

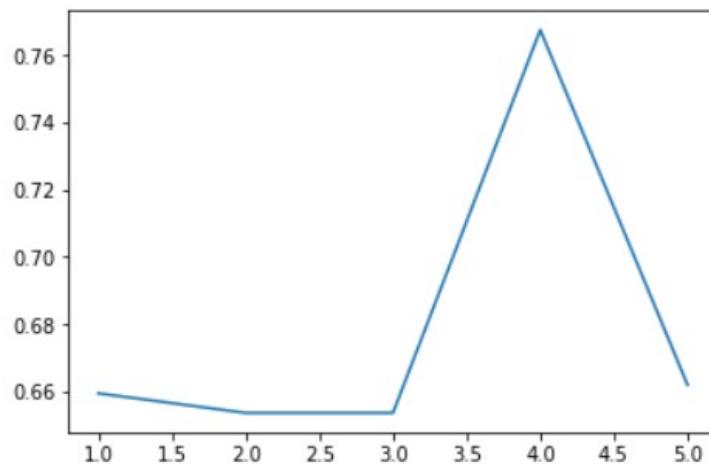


Sans le partitionnement 0

Indice de Davies-Bouldin



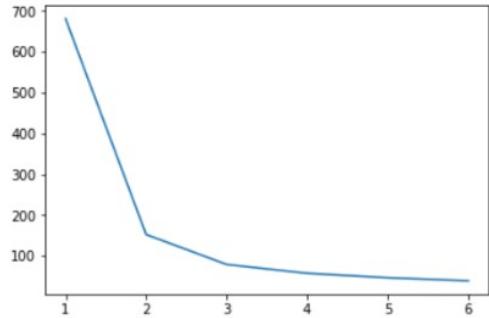
Avec le partitionnement 0



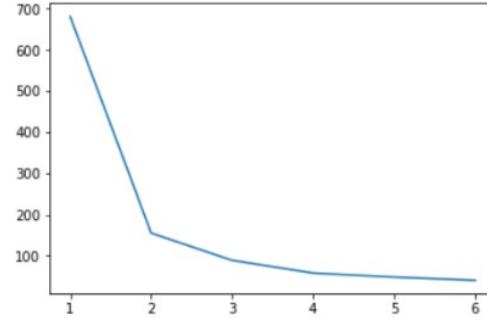
Sans le partitionnement 0

3) Choix du nombre K de *clusters*

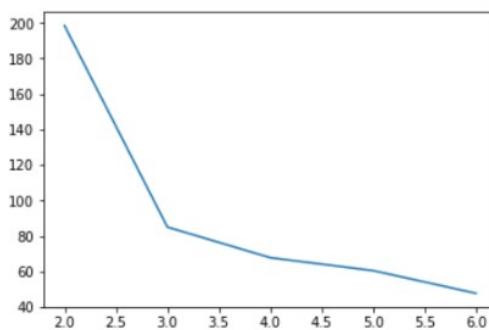
a) Erreur quadratique en fonction de K



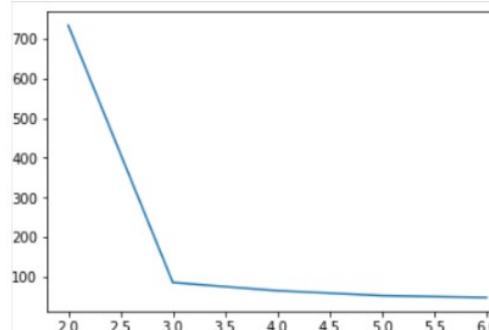
K-Means++



Global K-Means

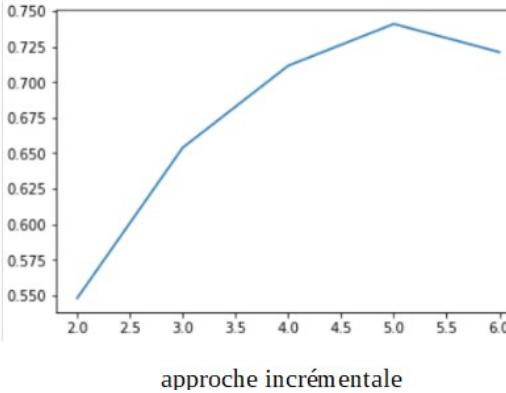
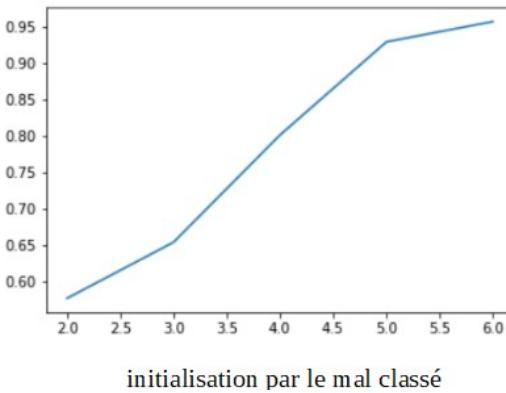
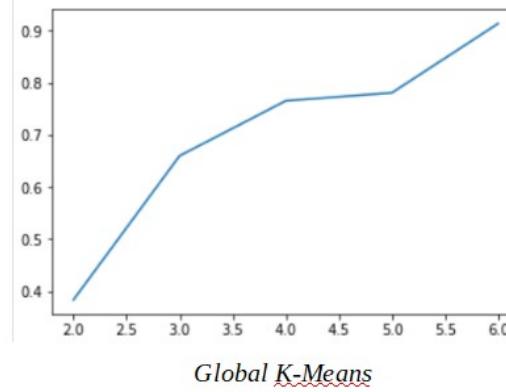
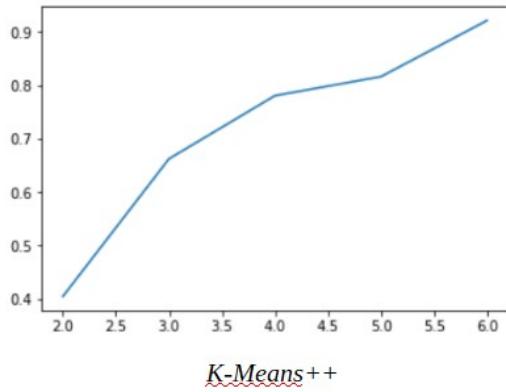


initialisation par le mal classé



approche incrémentale

b) Indice de Davies-Bouldin en fonction de K



Conclusion

Avantages	Inconvénients
Facile à comprendre et à implémenter	Très dépendant de l'initialisation
Rapide et faible coût en calculs	Le nombre de classe doit être fixé au départ
Applicable à des données de grandes tailles	Adapté à des ensembles de données dont les classes sont supposées sphériques
Permet de se faire une bonne idée du nombre de classes	