

projet_ando

December 14, 2023

1 Projet ANDO

1.1 Projection Orthogonale

Définition :

Soit \mathcal{E} un espace vectoriel de dimension p . Si \mathcal{D} est une droite vectorielle engendrée par le vecteur \vec{a} qui passe par un point Q de \mathbb{R}^p , l'ensemble des vecteurs orthogonaux à \mathcal{D} est un hyperplan appelé hyperplan normal à \mathcal{D} et défini par :

$$\mathcal{D}^\perp = \{ \vec{h} \in \mathbb{R}^p \mid (\vec{h} \cdot \vec{a}) = 0 \}$$

Si x est un point arbitraire de \mathbb{R}^p et si on note \vec{x} le vecteur associé qui va de Q à ce point, on peut toujours le décomposer de la façon suivante :

$$\vec{x} = \vec{x}_{\mathcal{D}} + \vec{x}_{\perp} \text{ avec } \vec{x}_{\mathcal{D}} = \frac{(\vec{x} \cdot \vec{a})}{\|\vec{a}\|^2} \vec{a}$$

Si on note $x_{\mathcal{D}}$ la projection du point sur la droite \mathcal{D} et si on note x_i la i ème composante du point x , on obtient alors les coordonnées du point $x_{\mathcal{D}}$:

$$\forall i \in [1; p], x_{\mathcal{D}_i} = Q_i + \frac{\sum_{k=1}^p (x_k - Q_k) * a_k}{\|\vec{a}\|^2} * a_i$$

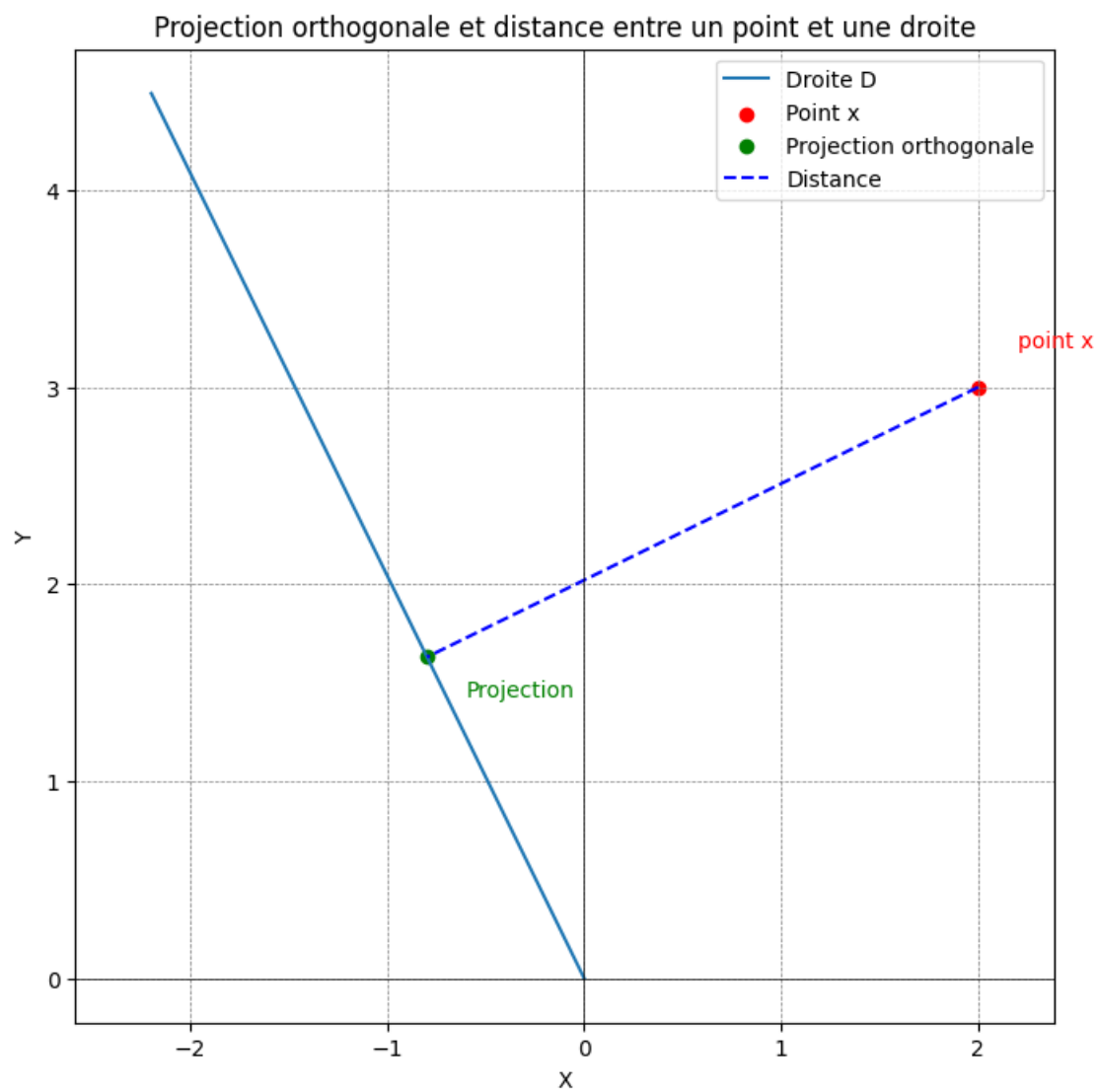
Pour avoir la distance entre le point x et la droite \mathcal{D} , on a besoin de:

$$\|\vec{x}_{\perp}\| = \|(x_1 - x_{\mathcal{D}_1}, x_2 - x_{\mathcal{D}_2}, \dots, x_p - x_{\mathcal{D}_p})\|$$

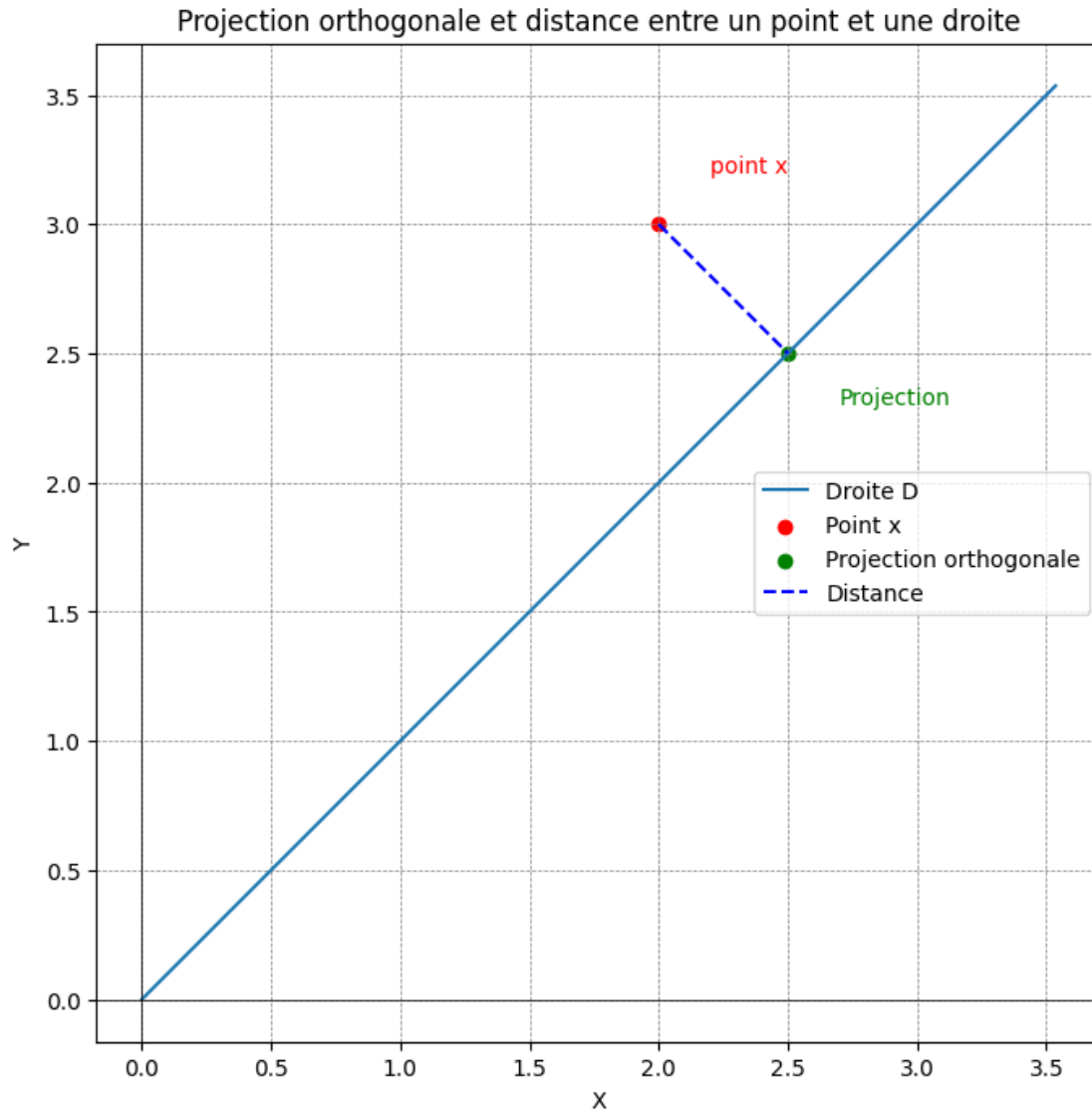
Dans la suite de ce document, on choisira de représenter une droite dans \mathbb{R}^p par un de ses vecteurs directeurs unitaires, noté \vec{u} .

Coordonnées de la projection orthogonale: [-0.79802776 1.63075237]

Distance entre le point et la droite: 3.1150920360380336



Distance entre le point et la droite: 3.1150920360380336



(Droite(Passe par le point Point([0 0]), vecteur directeur [0.70710678 0.70710678]), 'Indice de la liste: 1')

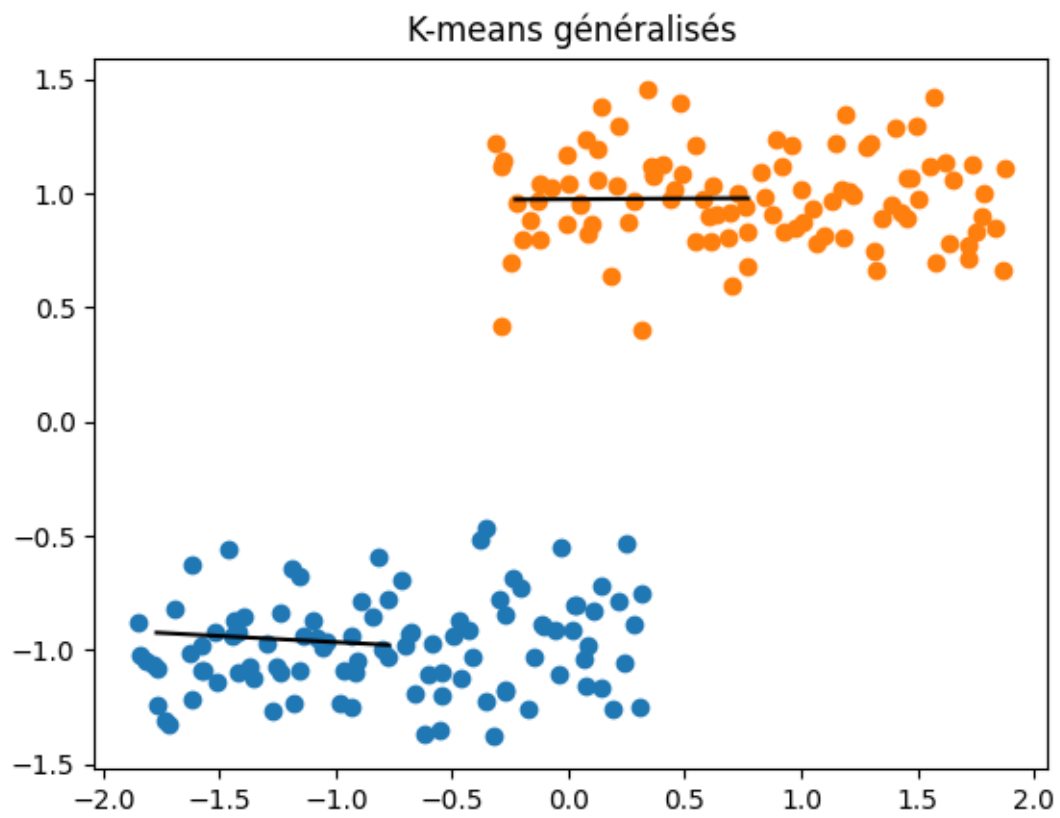
Le programme a bien déterminé que D2 est la droite la plus proche du point.

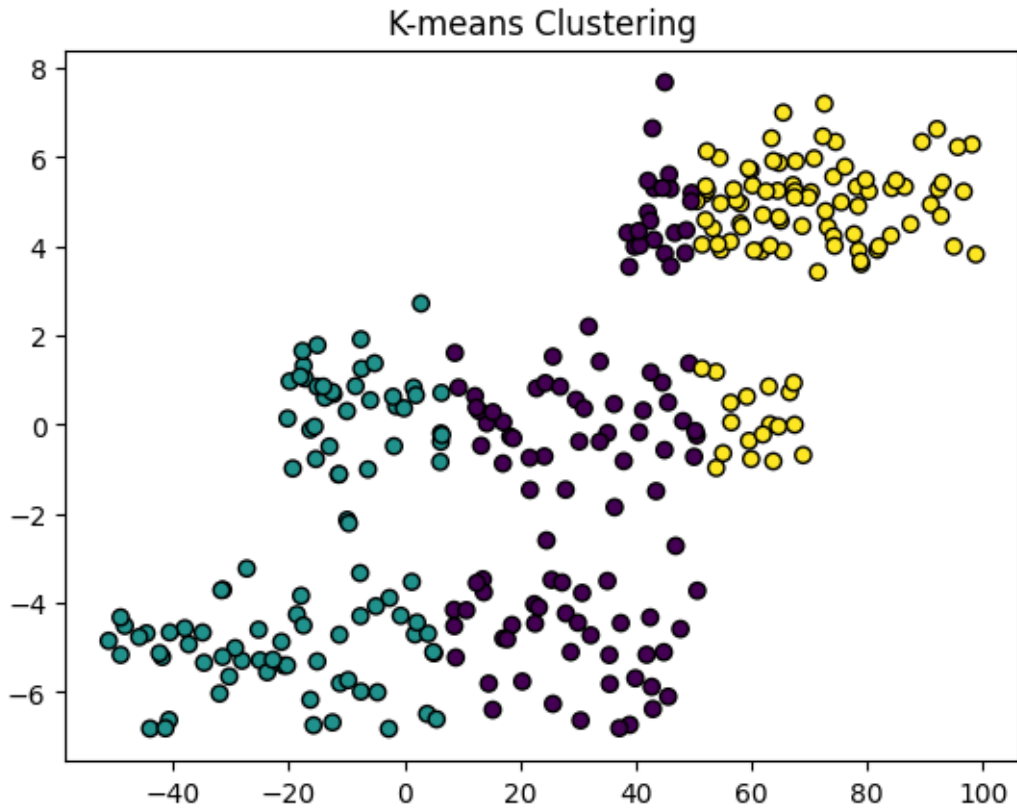
2 Les nuées dynamiques ou les kmeans généralisés

3 Exemples test

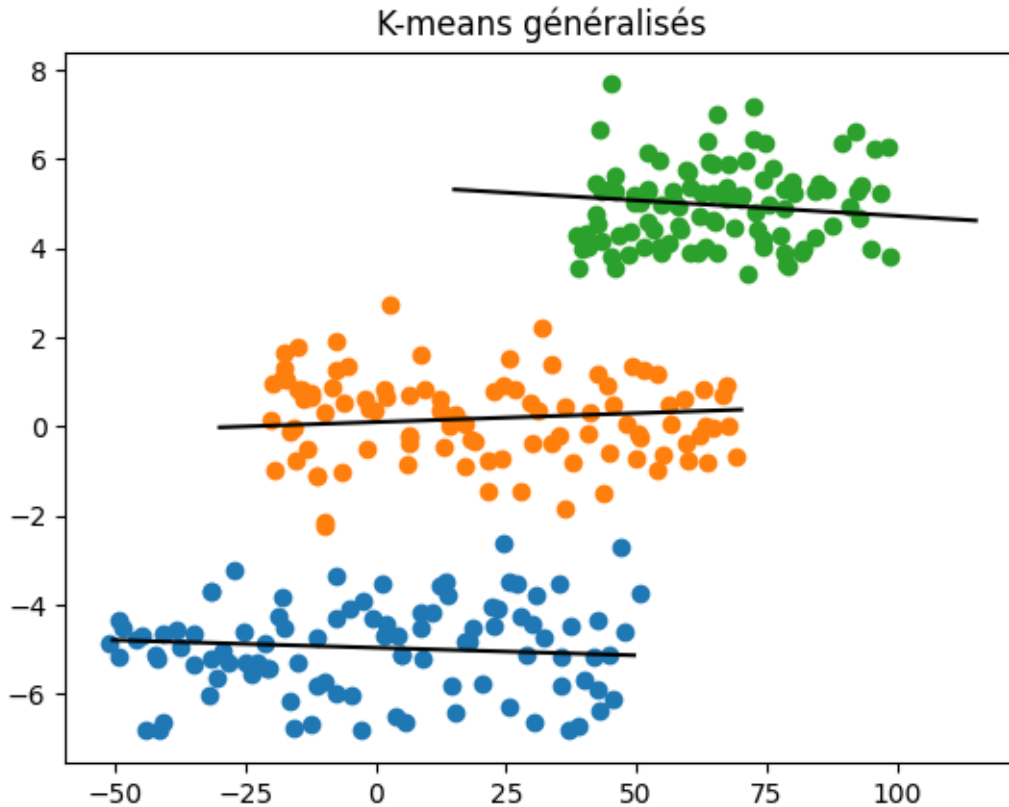
Classe 1 - Représentant : Droite(Passe par le point Point([-0.77206539 -0.97905988]), vecteur directeur [-0.99854272 0.05396694])

Classe 2 - Représentant : Droite(Passe par le point Point([0.77206539 0.97905988]), vecteur directeur [-0.9999846 -0.0055501])





Classe 1 - Représentant : Droite(Passe par le point Point([-0.43633677
-4.95844662]), vecteur directeur [-0.99999421 0.00340228])
 Classe 2 - Représentant : Droite(Passe par le point Point([20.12179571
0.18095954]), vecteur directeur [-0.99999208 -0.00398019])
 Classe 3 - Représentant : Droite(Passe par le point Point([65.09419249
4.96912639]), vecteur directeur [-0.99997579 0.00695876])

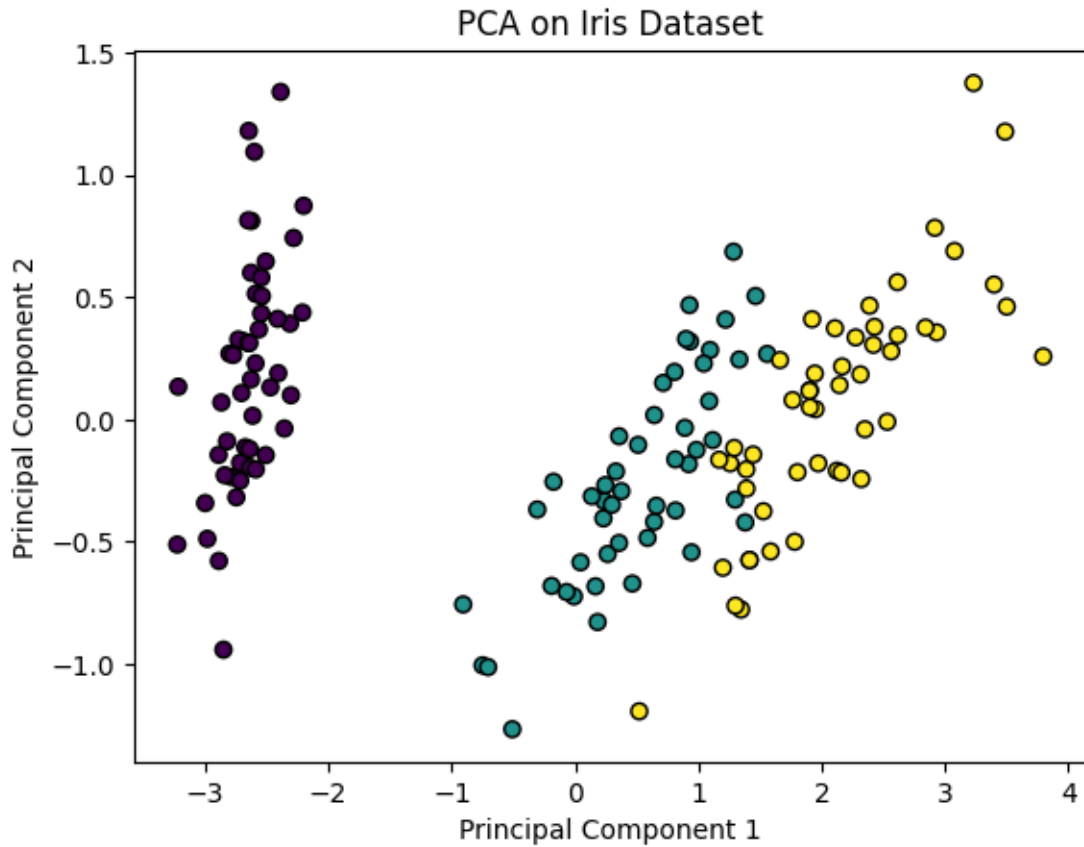


On remarque que pour ce dataset, la méthode de K-means généralisée est plus efficace que la méthode de K-means classique. En effet, la méthode classique ne permet pas de séparer les 3 nuées de points.

4 Application au dataset iris

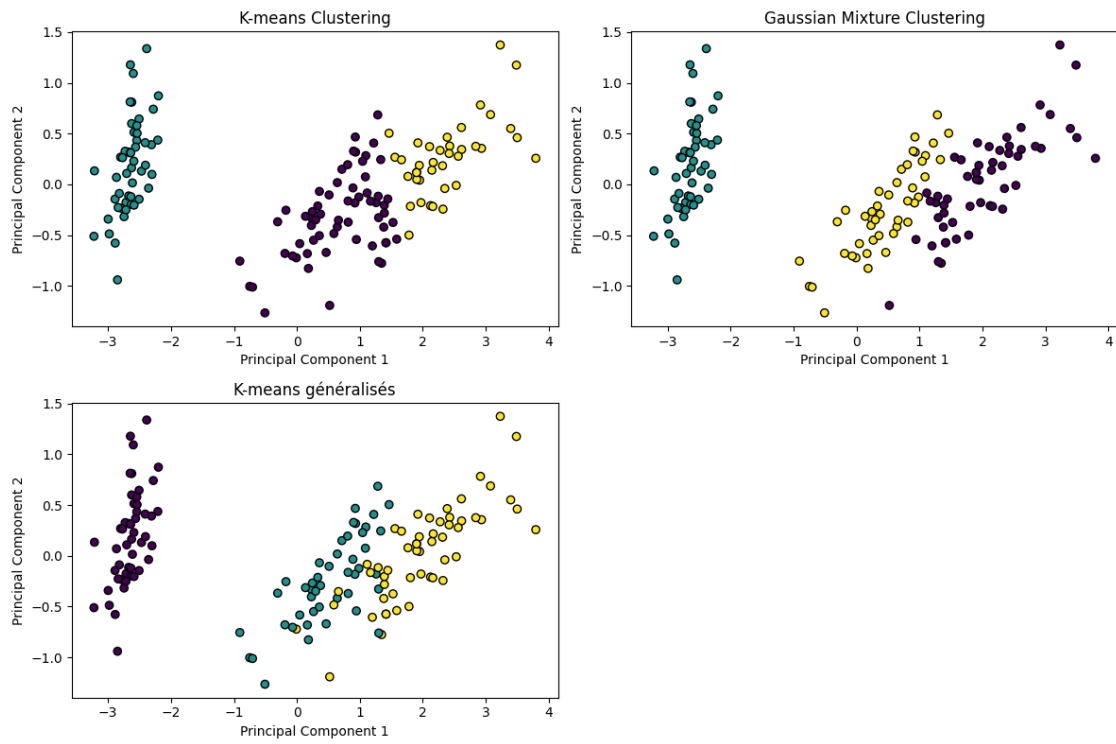
Nous allons dans cette partie tester notre algorithme des K-means généralisés sur le dataset *iris* et le comparer à deux autres modèles de clustering classiques, à savoir les K-means classiques et un modèle de mélange gaussien.

Dans un premier lieu, voici une représentation du dataset en 2 dimensions à l'aide d'une Analyse en Composantes Principales, avec chaque cluster représenté d'une couleur différente :

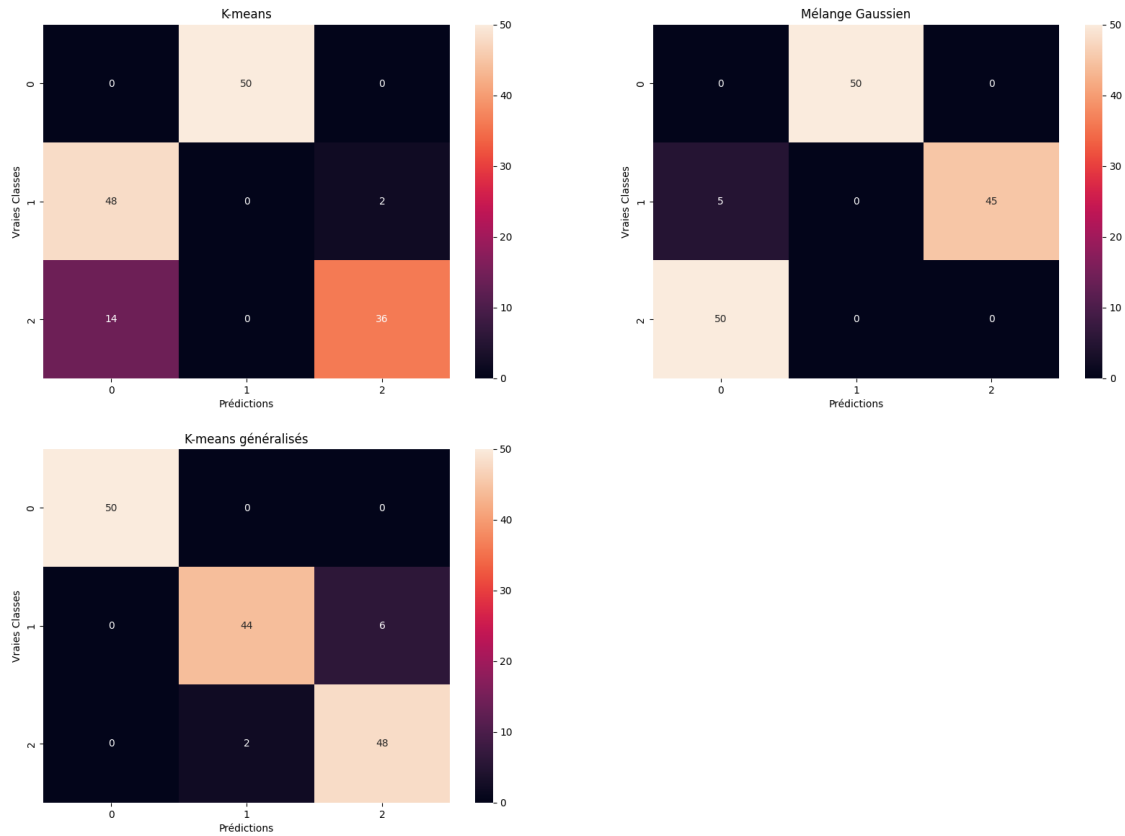


Nous observons qu'il y a exactement 3 clusters, ayant chacun une tendance de dispersion assez linéaire, et dont 2 sont assez proches l'un de l'autre. Voyons donc maintenant ce que les différentes méthodes de clustering donnent comme résultats.

Voici une visualisation des différents clustering :



Et voici leur puissance prédictive respective :



On observe donc que les deux versions des K-means sont assez précises, environ 90% pour le classique, et 94% pour le généralisé.

Cependant, le modèle de mélange gaussien est beaucoup plus précis, puisque seulement 5 fleurs sont mal classées, soit une précision de 96,6% !

Ceci pourrait s'expliquer par le fait que les algorithmes de K-means sont très sensibles à l'initialisation des clusters, qui sont déterminants.