projet_ando

December 14, 2023

1 Projet ANDO

1.1 Projection Orthogonale

Définition :

Soit $\mathcal{E}$ un espace vecteuriel de dimension p. Si $\mathcal{D}$ est une droite vecteurielle 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 :

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

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 :

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

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

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

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

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

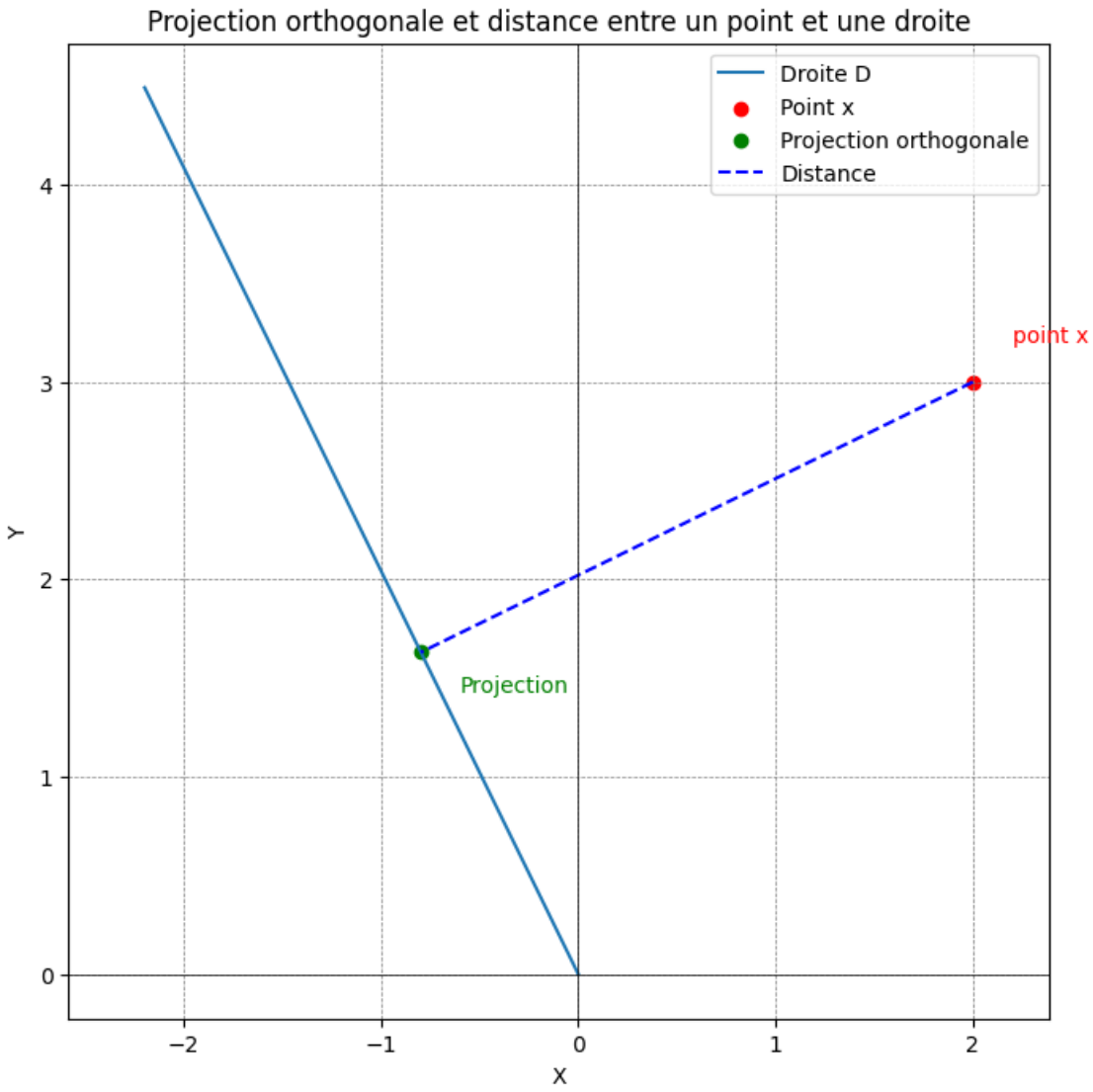
Dans la suite de ce document, on choisira de représenter une droite dans $\mathbb{R}^p$ par un de ses vecteurs directeurs unitaires, noté $\vec{u}$.

1.2 Exemple

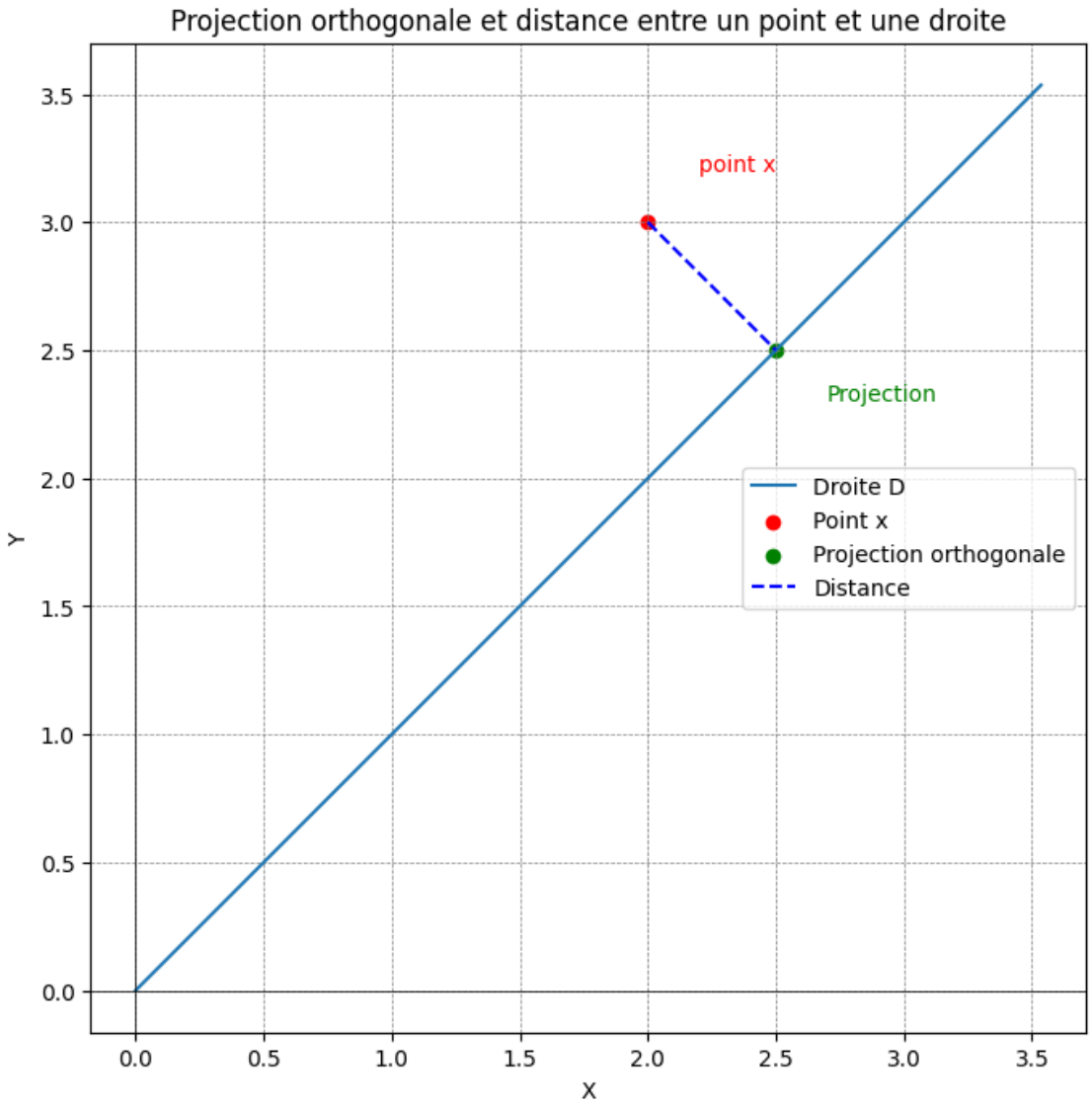
Dans l'exemple qui suit, nous allons travailler dans $\mathbb{R}^2$ et: - x est le point $(2, 3)$ - $D1$ est la droite passant par le point $(0, 0)$ et de vecteur directeur $\vec{u_1} = (-2.3, 4.7)$ - $D2$ est la droite passant par le point $(0, 0)$ et de vecteur directeur $\vec{u_2} = (1, 1)$

Traçons ces droites et le point x :

```
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
```



Déterminons maintenant la droite la plus proche du point x parmi les deux droites $D1$ et $D2$.

```
(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

L'algorithme des K-means permet de déterminer un nombre fixé de clusters dans un ensemble de points. Il utilise la notion de représentant de classe et de distance entre un point et un représentant de classe.

Dans le cas des K-means standard, le représentant de la classe est un point de l'espace, tandis que

la distance est la distance euclidienne.

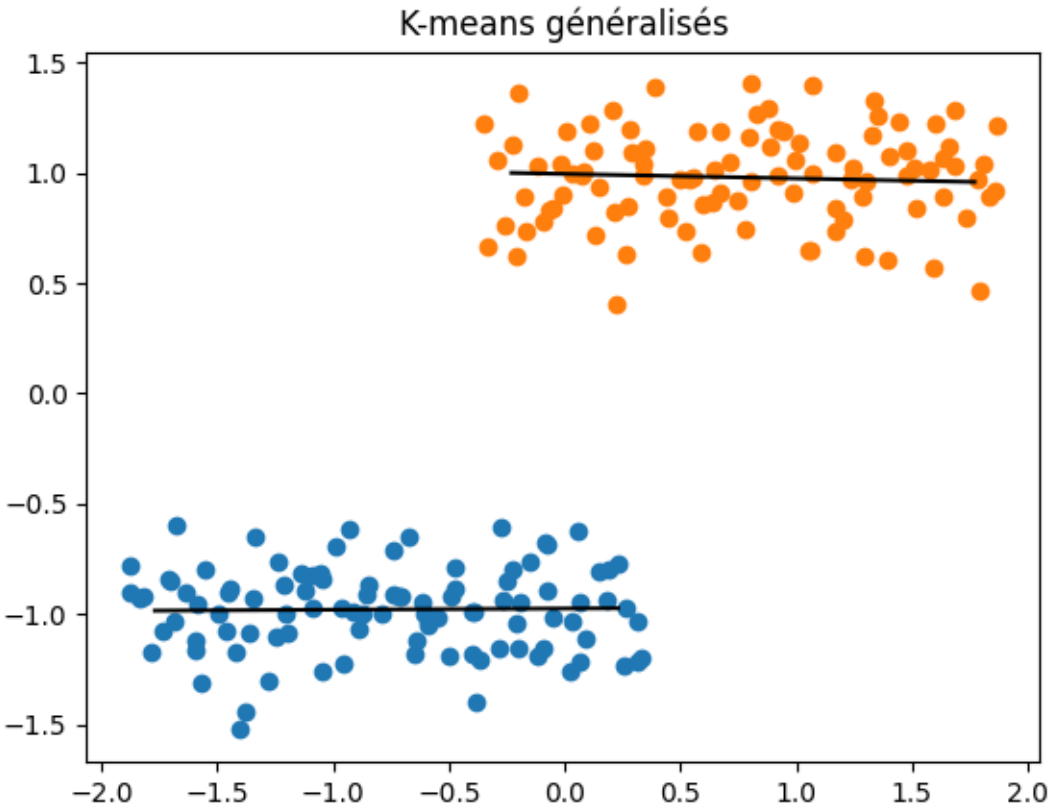
Nous allons utiliser une généralisation de cet algorithme, où le représentant de la classe est une droite et la distance est la distance entre un point et une droite, donc la norme de la projection orthogonale. On définit alors l'algorithme des K-means généralisés.

3 Exemples test

3.0.1 Exemple 1

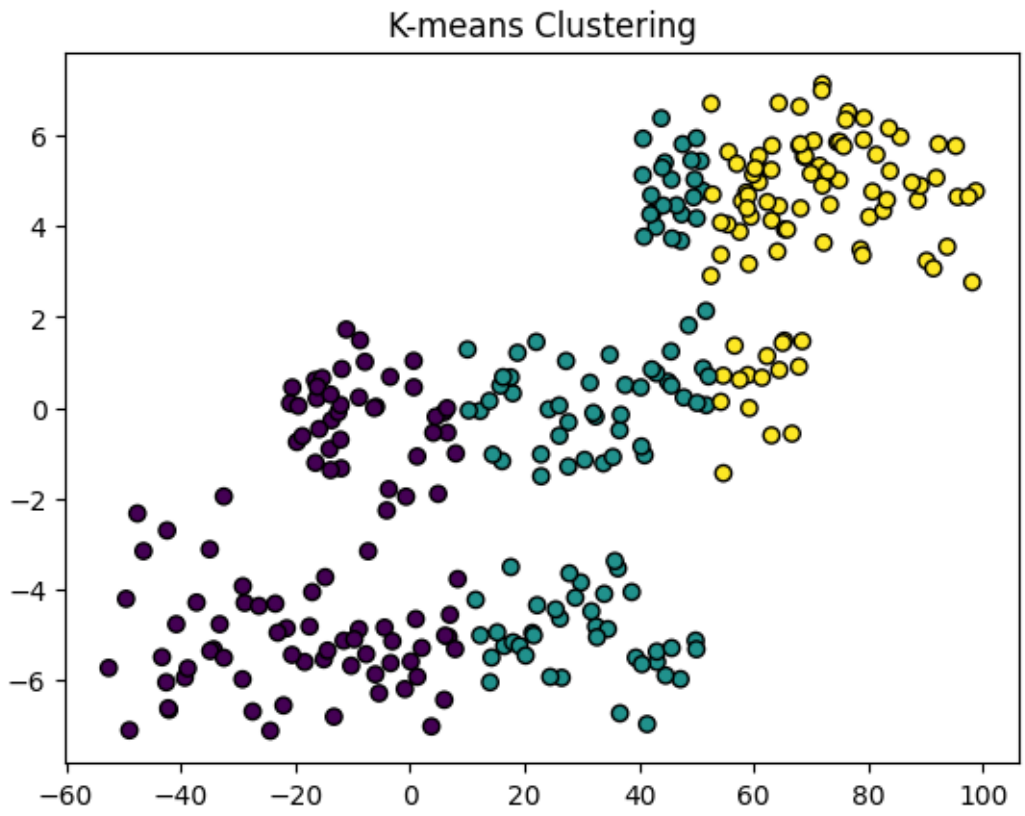
Nous allons tester l'algorithme sur un ensemble de points générés aléatoirement, suivant 2 droites dans $\mathbb{R}^2$. Il y aura donc 2 clusters:

```
Classe 1 - Représentant : Droite(Passe par le point Point([-0.76986854
-0.97945504]), vecteur directeur [-0.99998507 -0.00546505])
Classe 2 - Représentant : Droite(Passe par le point Point([0.76986854
0.97945504]), vecteur directeur [-0.99978876  0.02055299])
```

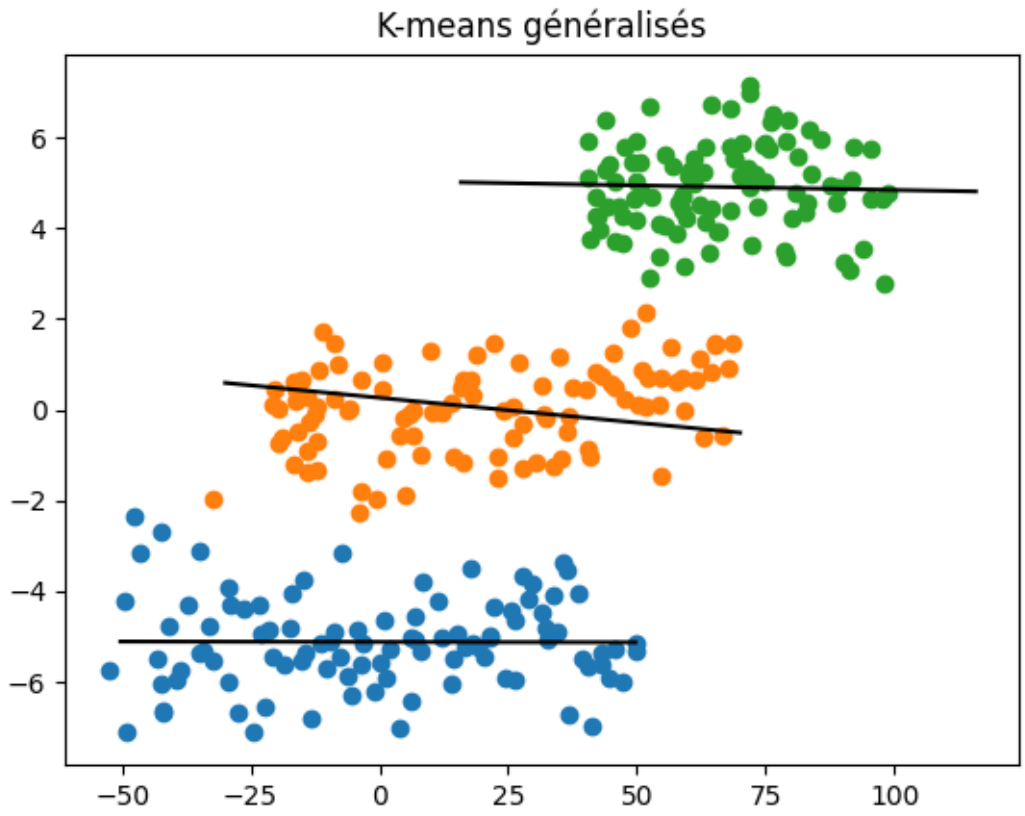


3.0.2 Exemple 2

Nous allons tester l'algorithme sur un ensemble de points générés aléatoirement, suivant 3 droites dans $\mathbb{R}^2$. Il y aura donc 3 clusters. Vérifions également qu'avec cette génération, l'algorithme des K-means standard ne fonctionne pas, et que l'algorithme des K-means généralisés fonctionne.




```
Classe 1 - Représentant : Droite(Passe par le point Point([-0.27777846
-5.12140996]), vecteur directeur [-9.99999986e-01  1.69863367e-04])
Classe 2 - Représentant : Droite(Passe par le point Point([20.05730336
0.04051682]), vecteur directeur [-0.99993977  0.01097565])
Classe 3 - Représentant : Droite(Passe par le point Point([65.82735691
4.90998824]), vecteur directeur [-0.99999797  0.00201612])
```

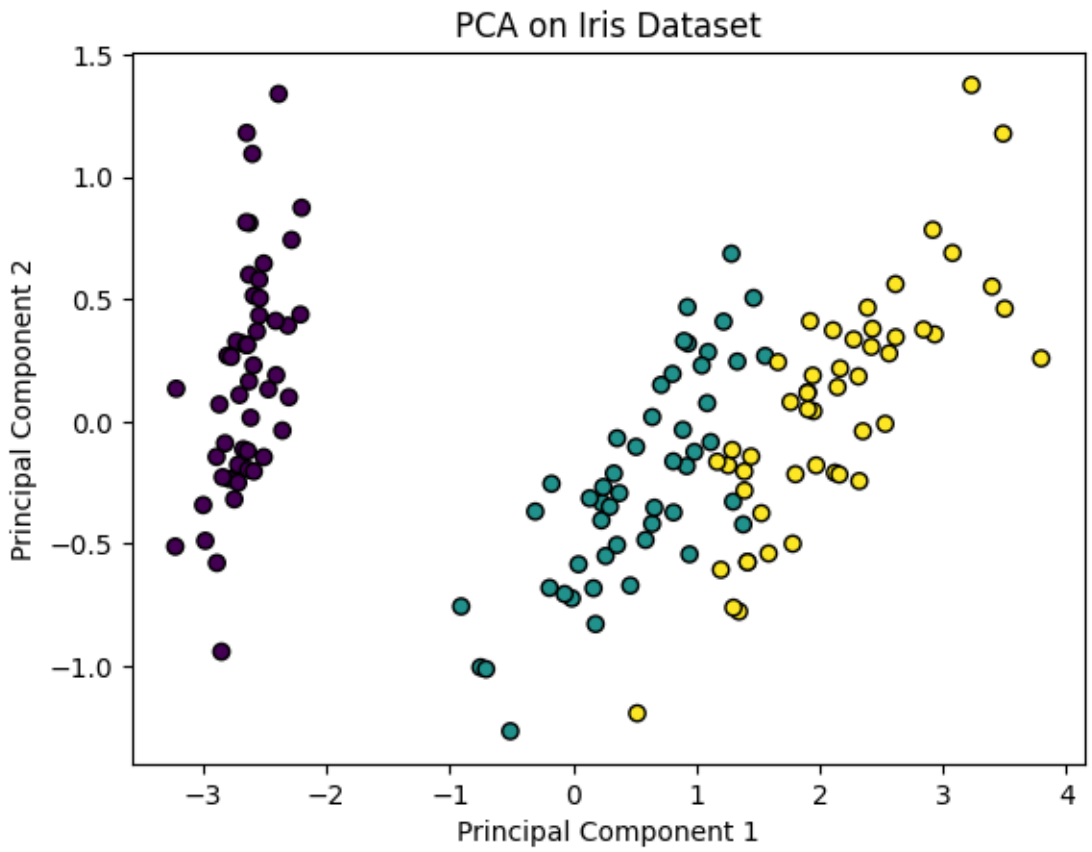


On remarque que pour ce dataset, la méthode de K-means généralisée est plus efficace que la méthode de K-means classique. En effet, la méthode classique ne permet pas de séparer les 3 nuées de points.

4 Application au dataset iris

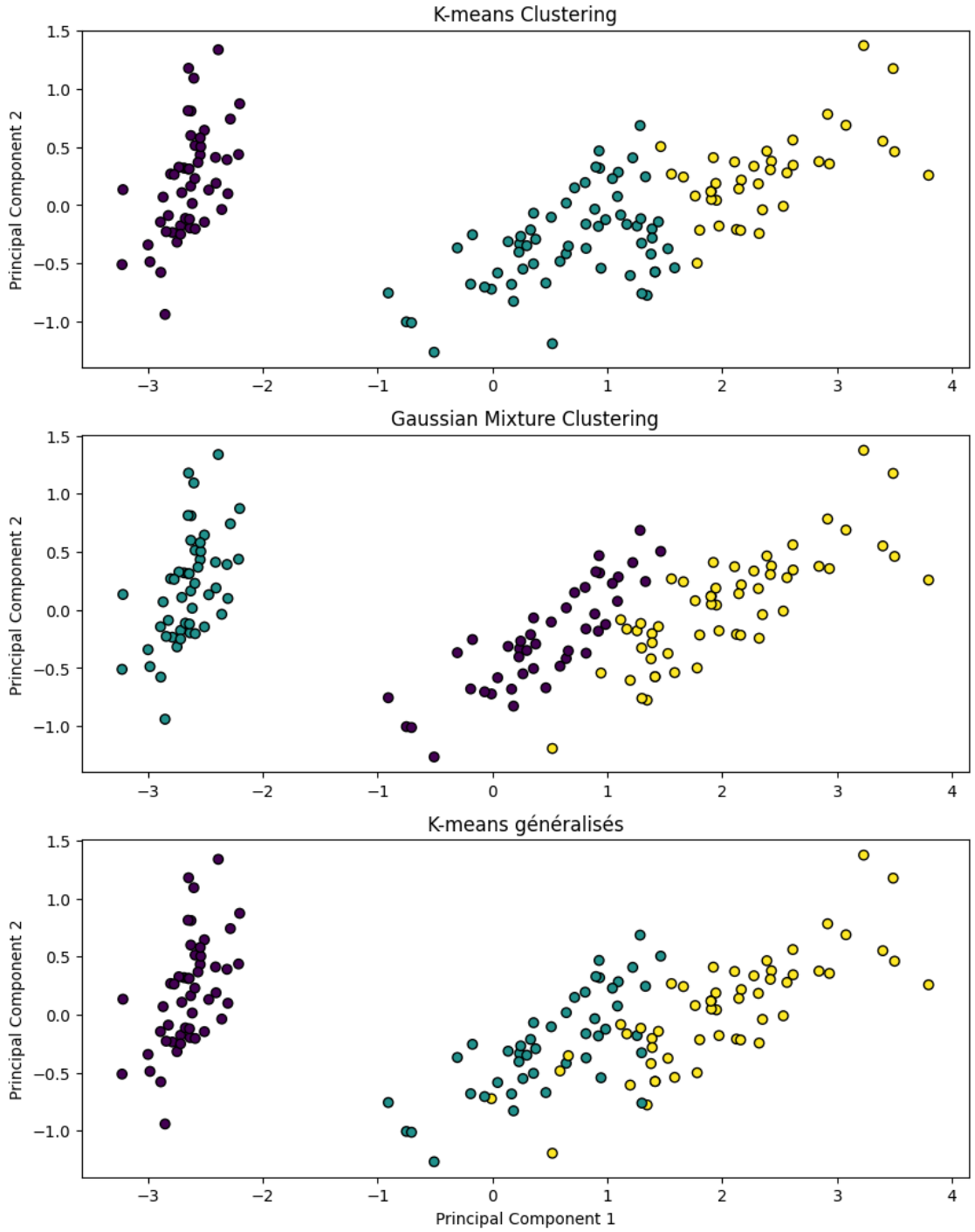
Nous allons dans cette partie tester notre algorithme des K-means généralisés sur le dataset *iris* et le comparer à deux autres modèles de clustering classiques, à savoir les K-means classiques et un modèle de mélange gaussien.

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 :

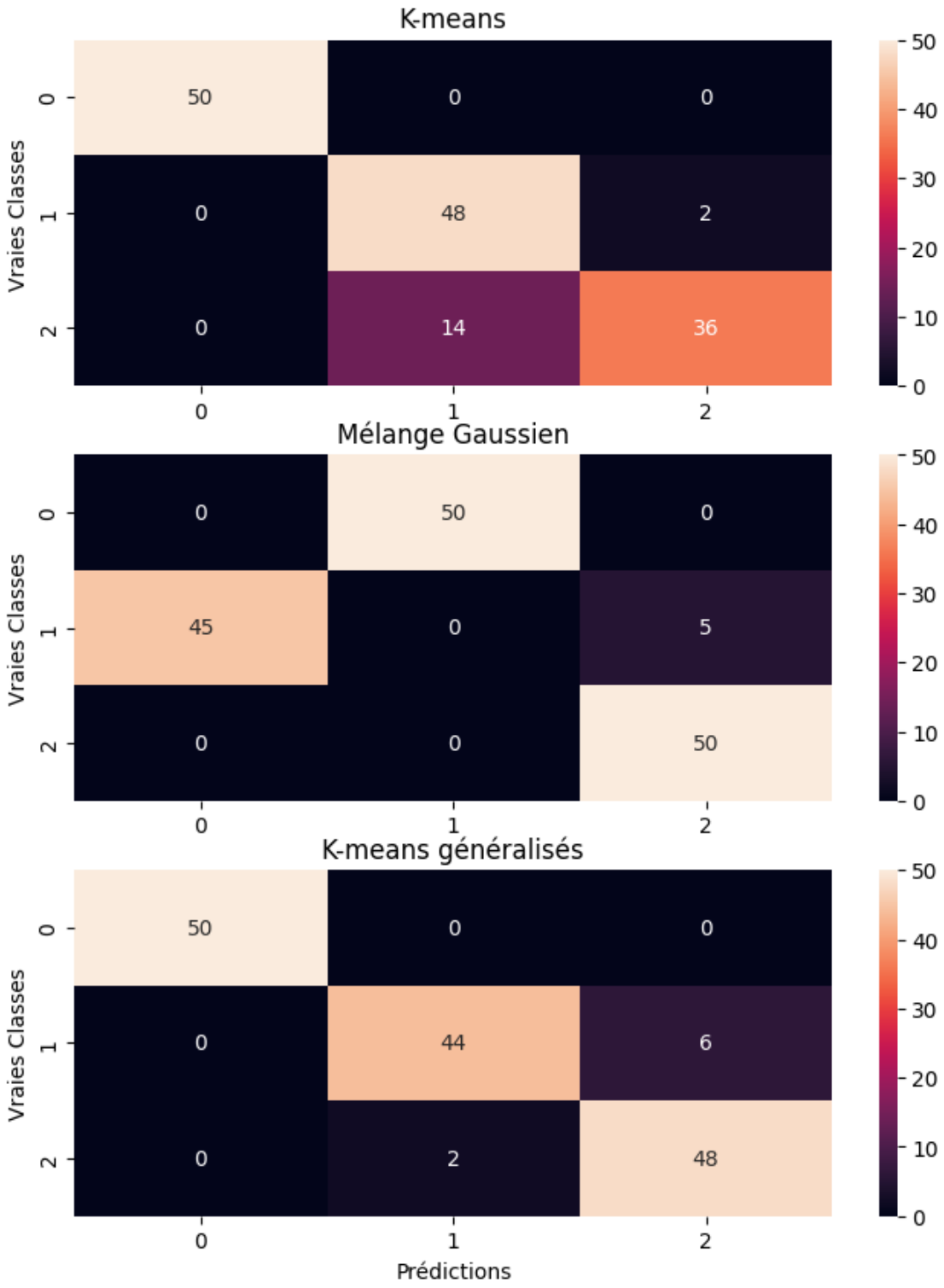


Nous observons qu'il y a exactement 3 clusters, ayant chacun une tendance de dispersion assez linéaire, et dont 2 sont assez proches l'un de l'autre. Voyons donc maintenant ce que les différentes méthodes de clustering donnent comme résultats.

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.