

UNSUPERVISED LEARNING 101

APPRENTISSAGE SANS SUPERVISION



Dans cette section, nous aborderons les bases de **l'apprentissage non supervisé**, notamment les concepts clés, les techniques et les applications, ainsi que les domaines dans lesquels il peut être utilisé dans le cadre du processus de science des données.

SUJETS ABORDÉS :

Unsupervised Learning

Techniques & Applications

Data Science Workflow

OBJECTIFS DE CETTE SECTION :

- Présenter les bases de l'apprentissage non supervisé
- Passer en revue la terminologie et les concepts clés
- Comprendre les différentes techniques et applications de l'apprentissage non supervisé
- Revoir le flux de travail de la science des données et identifier la place qu'y occupe l'apprentissage non supervisé



APPRENTISSAGE SANS SUPERVISION

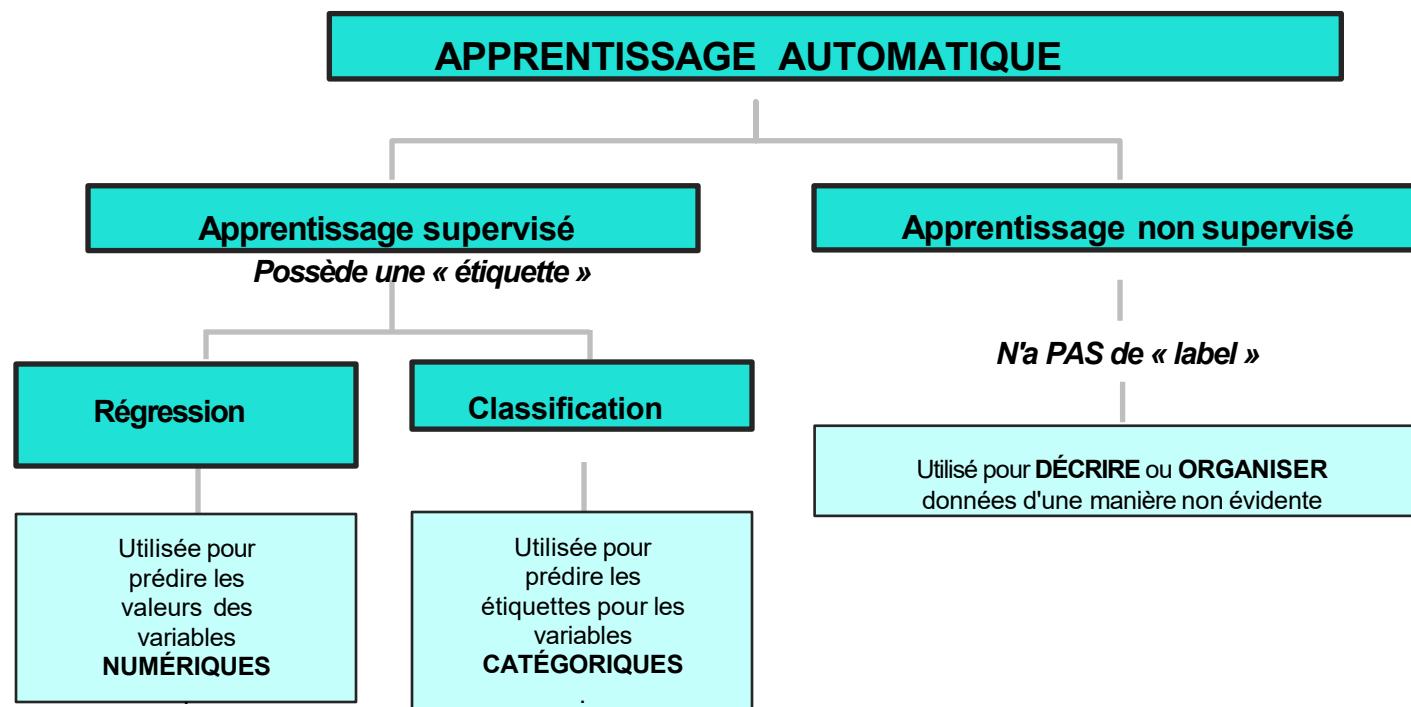
Unsupervised
Learning

Techniques &
Applications

Data Science
Workflow

L'apprentissage non supervisé consiste à **trouver des informations et des modèles cachés** dans les données

- Contrairement à la régression ou à la classification, nous ne nous soucions pas de diviser nos données en ensembles d'entraînement/de test et faire des prédictions, nous nous intéressons uniquement à *la compréhension* des relations dans nos données





APPRENTISSAGE SANS SUPERVISION

Unsupervised Learning

Techniques & Applications

Data Science Workflow

EXEMPLE

Regroupement des clients en fonction de leur comportement d'écoute

Chaque ligne représente un client

Ce sont des caractéristiques (ce qui entre dans le modèle)

Customer	Music Streaming Hours	Podcast Listening Hours
Aria	46	9
Accord	38	10
Harmonie	44	17
Mélodie	19	50
Reed	7	44
Alto	16	52
Rock	5	19
Piper	10	11
Allegra	17	9

← Notez qu'il n'y a PAS de cible



Comment pouvons-nous segmenter Ces clients ?



APPRENTISSAGE SANS SUPERVISION

Unsupervised Learning

Techniques & Applications

Data Science Workflow

EXEMPLE

Regroupement des clients en fonction de leur comportement d'écoute

Chaque ligne
représente un
client

Ce sont des caractéristiques (ce qui entre dans le modèle)

Customer	Music Streaming Hours	Podcast Listening Hours
Aria	46	9
Accord	38	10
Harmonie	44	17
Mélodie	19	50
Reed	7	44
Alto	16	52
Rock	5	19
Piper	10	11
Allegra	17	9

← Notez qu'il n'y a PAS de cible

Groupe 1
Amateurs de musique

Groupe 2
Amateurs de podcasts

Groupe 3
Auditeurs occasionnels



APPRENTISSAGE SANS SUPERVISION

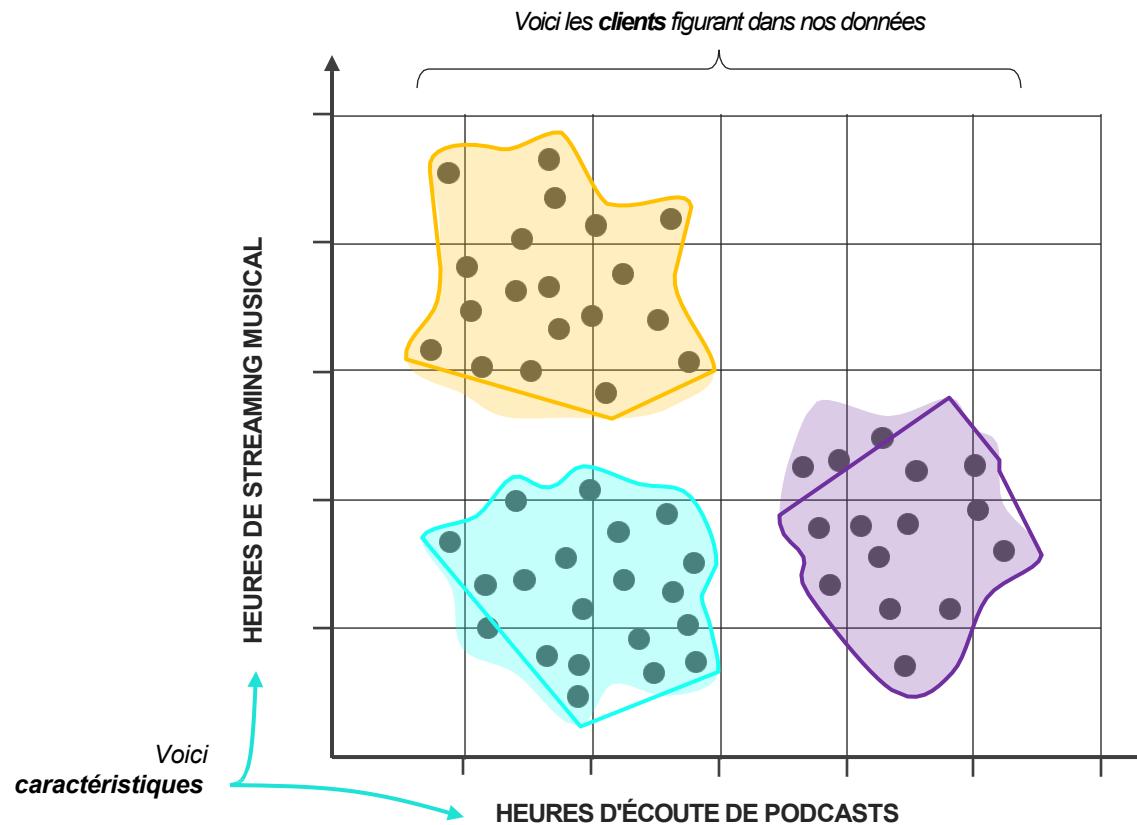
Unsupervised
Learning

Techniques &
Applications

Data Science
Workflow

EXEMPLE

Regroupement des clients en fonction de leur comportement d'écoute



*Nous pouvons clairement segmenter nos clients en **trois groupes** :*

- Amateurs de musique
- Les amateurs de podcasts
- Auditeurs occasionnels



TECHNIQUES D'APPRENTISSAGE AUTONOME

Unsupervised Learning

Techniques & Applications

Data Science Workflow

Il existe deux catégories populaires de **techniques** d'apprentissage non supervisé :



Clustering

Identifier des groupes (*ou clusters*) de points de données qui sont similaires entre eux mais distincts des autres groupes

Techniques courantes :

- Regroupement par la méthode des k-moyennes
- Clustering hiérarchique
- DBSCAN (regroupement basé sur la densité)

Applications :

- Clustering / segmentation
- Détection d'anomalies
- Recommandations



Dimensionality Reduction

Réduire le nombre de colonnes (*ou dimensions*) dans un ensemble de données tout en perdant le moins d'informations possible

Techniques courantes :

- ACP (analyse en composantes principales)
- t-SNE (t-Stochastic Neighbor Embedding)
- SVD (décomposition en valeurs singulières)

Applications :

- Extraction de caractéristiques
- Visualisation des données
- Recommandations



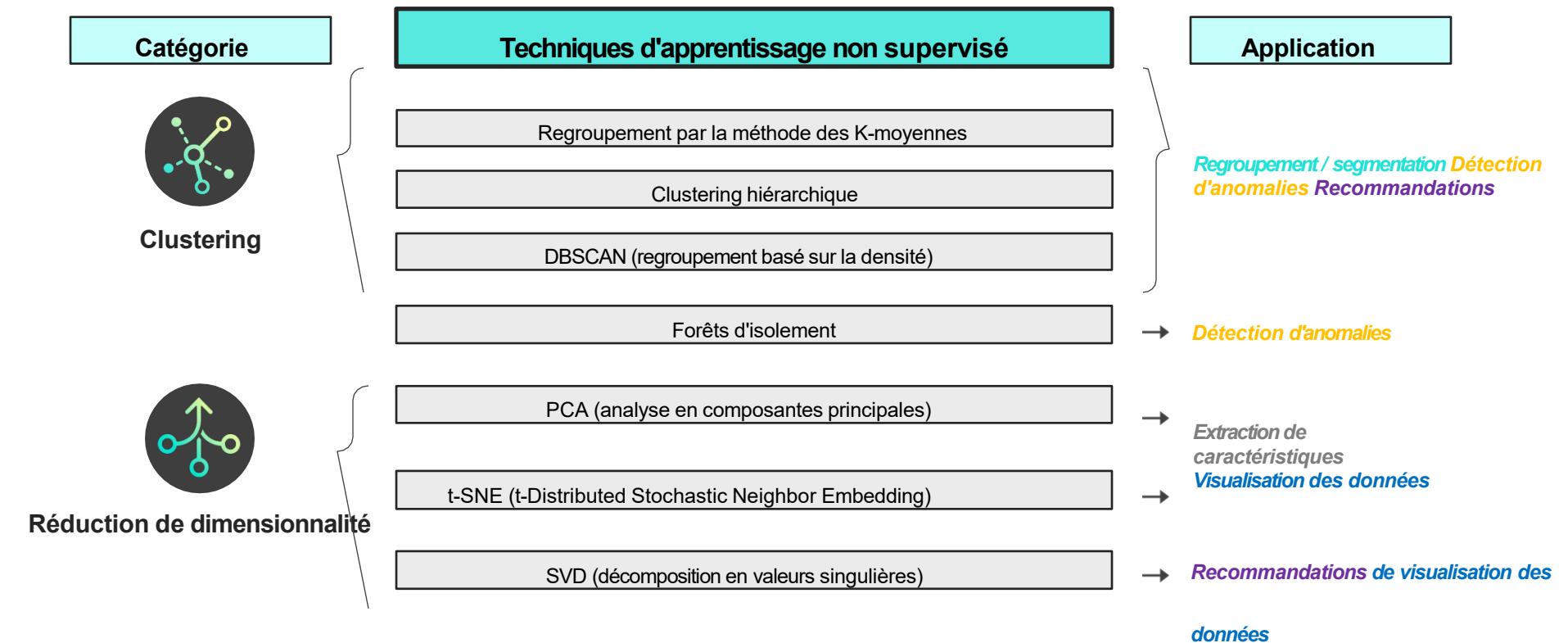
Techniques d'apprentissage non supervisé

Unsupervised Learning

Techniques & Applications

Data Science Workflow

Nous aborderons **les techniques et applications** d'apprentissage non supervisé dans l'ordre suivant :



POINTS CLÉS



L'apprentissage non supervisé est utilisé pour **trouver des modèles et des relations** dans les données

- *Il n'y a pas de prédictions ni d'étiquettes avec l'apprentissage non supervisé : nous essayons simplement de mieux comprendre la structure, l'organisation et les relations non évidentes entre les points de données*



L'apprentissage non supervisé repose sur un **état d'esprit différent** de celui de l'apprentissage supervisé

- *Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé ne nécessite pas de diviser les données en un ensemble d'entraînement et un ensemble de test, et l'évaluation repose sur une expertise approfondie du domaine en plus des métriques*



Il existe **de multiples applications** pour **les techniques** d'apprentissage non supervisé

- *Si les deux principales catégories de techniques d'apprentissage non supervisé relèvent du clustering et de la réduction de dimensionnalité, ces techniques peuvent être appliquées à la segmentation, à la détection d'anomalies, aux recommandations, etc.*



Ces techniques peuvent être utilisées à **plusieurs étapes** du processus de science des données

- *Outre l'utilisation de techniques d'apprentissage non supervisé lors de l'étape de modélisation du workflow de science des données, certaines techniques peuvent également être utilisées lors des phases de nettoyage, d'exploration et d'ingénierie des caractéristiques des données*

CLUSTERING

CLUSTERING



Dans cette section, nous présenterons les principes fondamentaux **du regroupement** et comparerons trois techniques de regroupement courantes : le regroupement par la méthode des K-moyennes, le regroupement hiérarchique et la méthode DBSCAN.

SUJETS ABORDÉS :

Clustering Basics

K-Means Clustering

Hierarchical Clustering

DBSCAN

Comparing Models

OBJECTIFS DE CETTE SECTION :

- Apprendre le fonctionnement fondamental des modèles de clustering
- Utiliser Python pour appliquer différents modèles de clustering et interpréter leurs résultats
- Comparer et contraster les techniques de clustering courantes



LES BASES DU CLUSTERING

Clustering Basics

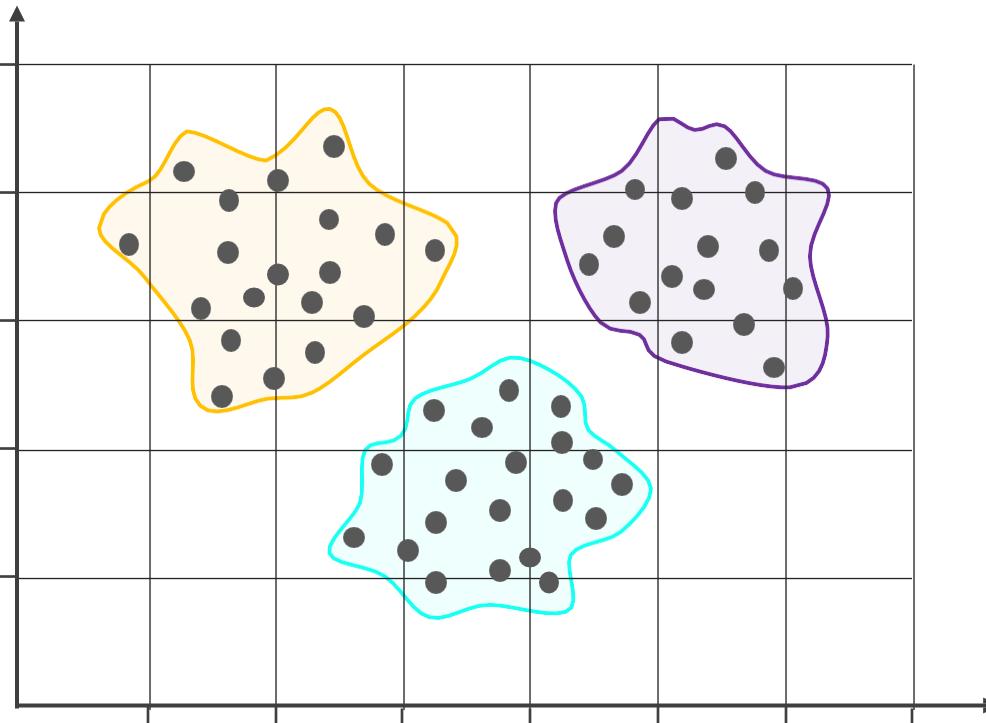
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Le regroupement vous permet de trouver des concentrations ou des groupes d'observations qui sont similaires entre eux, mais distincts des autres groupes.





FLUX DE TRAVAIL DU CLUSTERING

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Le workflow général du clustering comprend les étapes suivantes :

Data Prep

Préparez vos données pour les saisir dans un modèle ML.

- Table unique, données non nulles et numériques
- Ingénierie, sélection et mise à l'échelle des caractéristiques

Modeling

Appliquez un algorithme de clustering

- Regroupement par la méthode des k-moyennes
- Clustering hiérarchique
- DBSCAN

Tuning

Évaluez et ajustez le modèle à l'aide de métriques et de votre intuition

- Métriques (c'est-à-dire *inertie*)
- Visualisation des données
- Interpréter les résultats

Selection

Sélectionnez les meilleurs résultats et identifiez les informations pertinentes

- Objectif commercial
- Expertise dans le domaine



N'oubliez pas qu'il n'existe pas de « bonne » réponse ni de mesure d'optimisation unique en matière de regroupement ; les meilleurs résultats sont celles qui vous aident à répondre à la question posée et à prendre des décisions commerciales pratiques, fondées sur des données



CLUSTERING K-MEANS

Clustering Basics

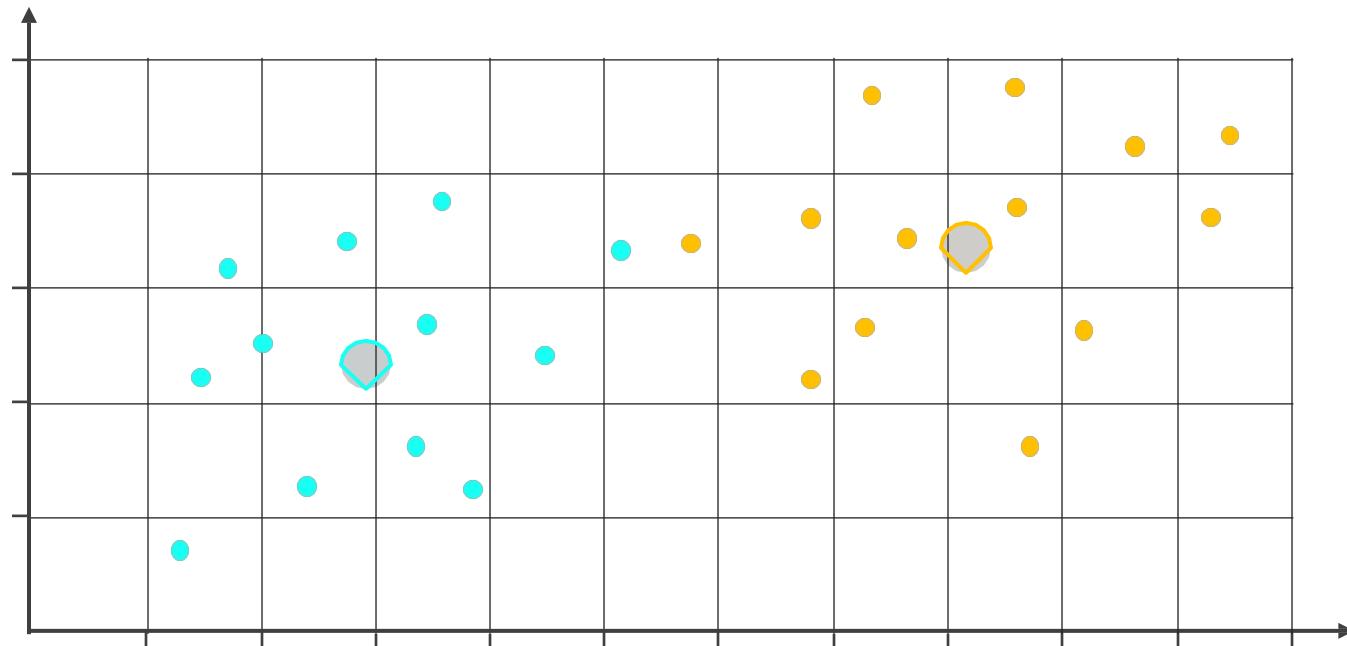
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Le regroupement par la méthode des k-moyennes est un algorithme populaire qui attribue chaque observation dans un ensemble de données à un cluster spécifique, où « K » représente le nombre de clusters





CLUSTERING K-MEANS

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Le regroupement par la méthode des K-moyennes est un algorithme populaire qui attribue chaque observation d'un ensemble de données à un cluster spécifique, où « K » représente le nombre de clusters

Voici comment cela fonctionne :

1. Sélectionnez « K » emplacements arbitraires dans un nuage de points comme centres de cluster (ou **centroïdes**),
et attribuez chaque observation à un cluster en fonction du centroïde le plus proche.
2. Recalquez et déplacez chaque centroïde vers la moyenne des observations qui lui sont attribuées, puis réattribuez chaque observation à son *nouveau* centroïde le plus proche.
3. Répétez le processus jusqu'à ce que les observations ne changent plus de cluster.

Exemples d'utilisation :

- Identification de segments de clientèle pour des campagnes marketing ciblées
- Regrouper les emplacements des magasins en fonction de facteurs tels que les ventes, les notes, la taille, etc.



CLUSTERING K-MEANS

Clustering Basics

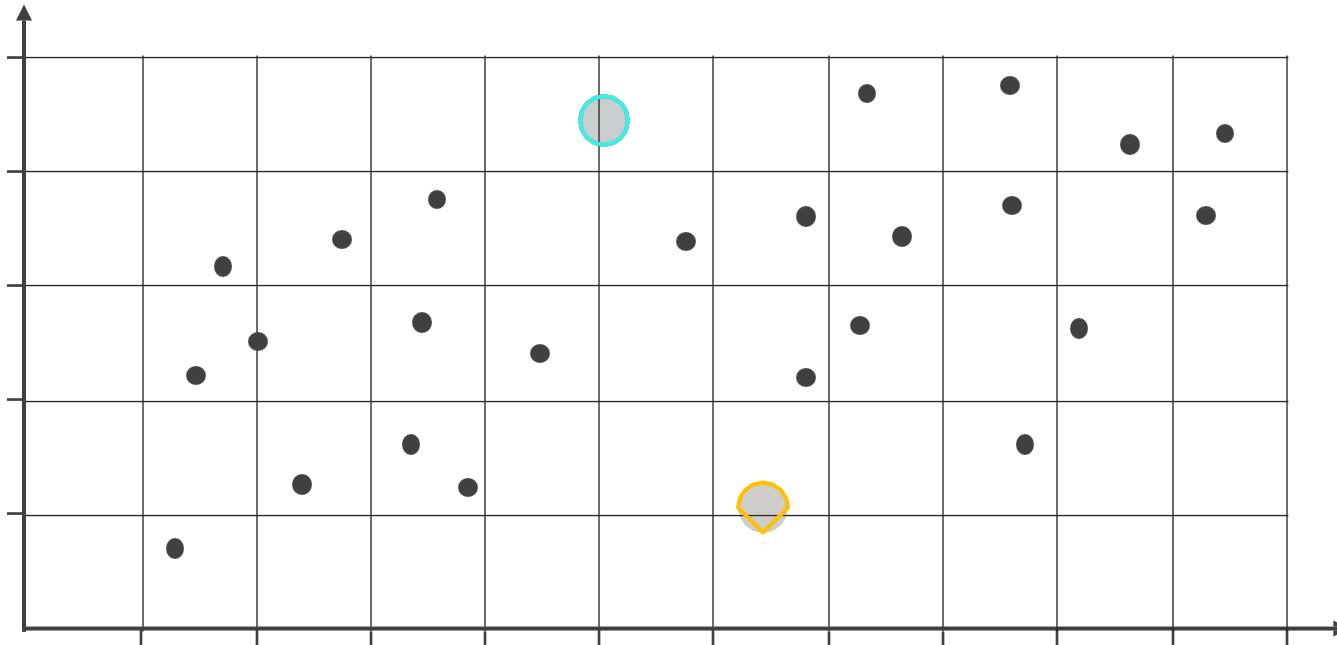
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : Déterminez ce qui vous semble être un nombre approprié de clusters (*dans ce cas, 2*), et sélectionnez des emplacements arbitraires comme centroïdes initiaux





CLUSTERING K-MEANS

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Comment mesurer la proximité ?

Pour affecter chaque point au centroïde le plus proche, on utilise la **distance euclidienne** :

Formule de la distance euclidienne

$$d(\text{point}, \text{centroïde}) = \sqrt{[(x_1 - x_2)^2 + (y_1 - y_2)^2]}$$

Exemple de calcul :

Si un point est en (3, 4) et un centroïde en (0, 0) :

$$d = \sqrt{(3-0)^2 + (4-0)^2} = \sqrt{9 + 16} = \sqrt{25} = 5$$

Objectif : Calculer la distance de chaque point vers les 2 centroïdes, puis affecter le point au centroïde le plus proche.



CLUSTERING K-MEANS

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Calcul des distances pour un point

Point analysé : (5, 6)

Distance vers Centroïde 1 (cyan)

Position : (2, 3)

$$\begin{aligned} d_1 &= \sqrt{(5-2)^2 + (6-3)^2} \\ d_1 &= \sqrt{9 + 9} = \sqrt{18} = 4.24 \end{aligned}$$

Distance vers Centroïde 2 (jaune)

Position : (8, 5)

$$\begin{aligned} d_2 &= \sqrt{(5-8)^2 + (6-5)^2} \\ d_2 &= \sqrt{9 + 1} = \sqrt{10} = 3.16 \end{aligned}$$

Décision : $d_2 (3.16) < d_1 (4.24) \rightarrow$ Le point est affecté au **Cluster 2 (jaune)**

Règle : On répète ce calcul pour chaque point et on l'affecte au centroïde dont la distance est minimale.



CLUSTERING K-MEANS

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 2 : Affecter chaque point au cluster le plus proche

Pour chaque point : Cluster = argmin_k d(point, centroïde_k)

On choisit le cluster k qui minimise la distance

Procédure :

- 1 Calculer la distance du point vers tous les centroïdes
- 2 Identifier le centroïde avec la distance minimale
- 3 Affecter le point au cluster correspondant

Résultat → Tous les points noirs sont colorés selon leur cluster



CLUSTERING K-MEANS

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 2 : Attribuer chaque observation à un cluster, en fonction du centroïde le plus proche

Après le calcul des distances, chaque point noir a été coloré selon son affectation :

Points cyan

Plus proches du **centroïde cyan**
→ Affectés au Cluster 1

Points jaunes

Plus proches du **centroïde jaune**
→ Affectés au Cluster 2

Visualisation : Les cercles autour des centroïdes montrent leur zone d'influence

Prochaine étape → Recalculer les centroïdes comme moyenne des points affectés



CLUSTERING K-MEANS

Clustering Basics

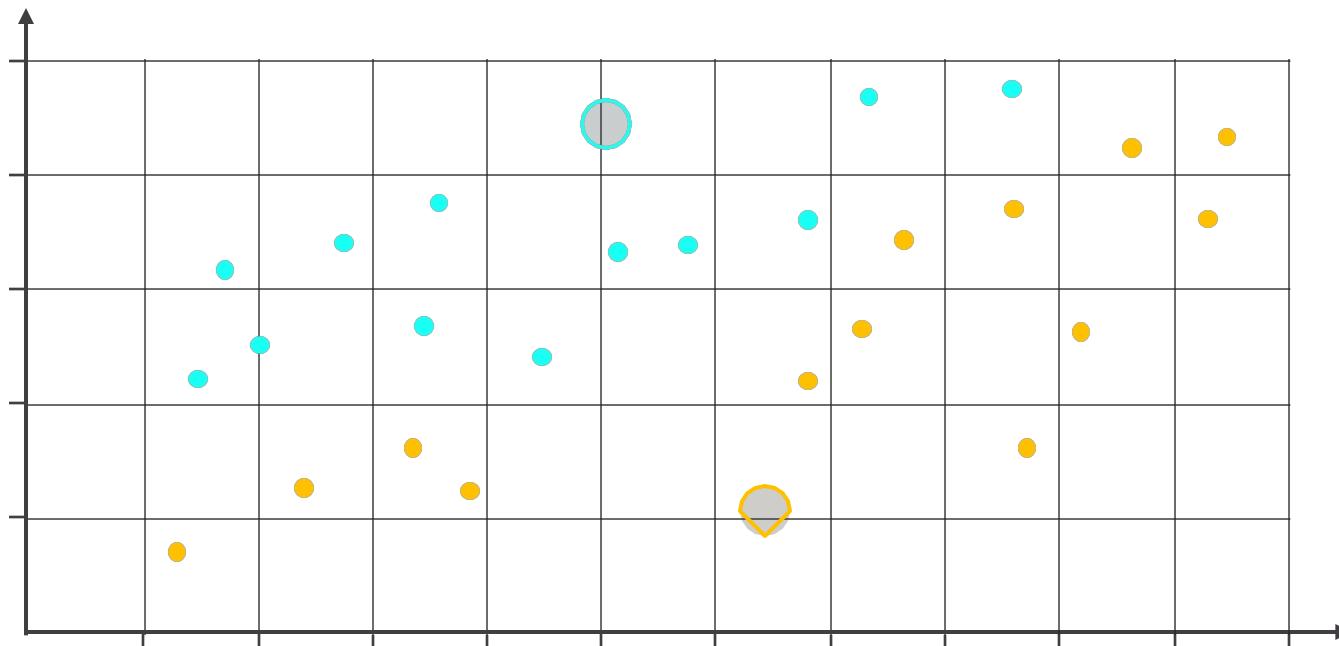
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 2 : Attribuer chaque observation à un cluster, en fonction du centroïde le plus proche





CLUSTERING K-MEANS

Clustering Basics

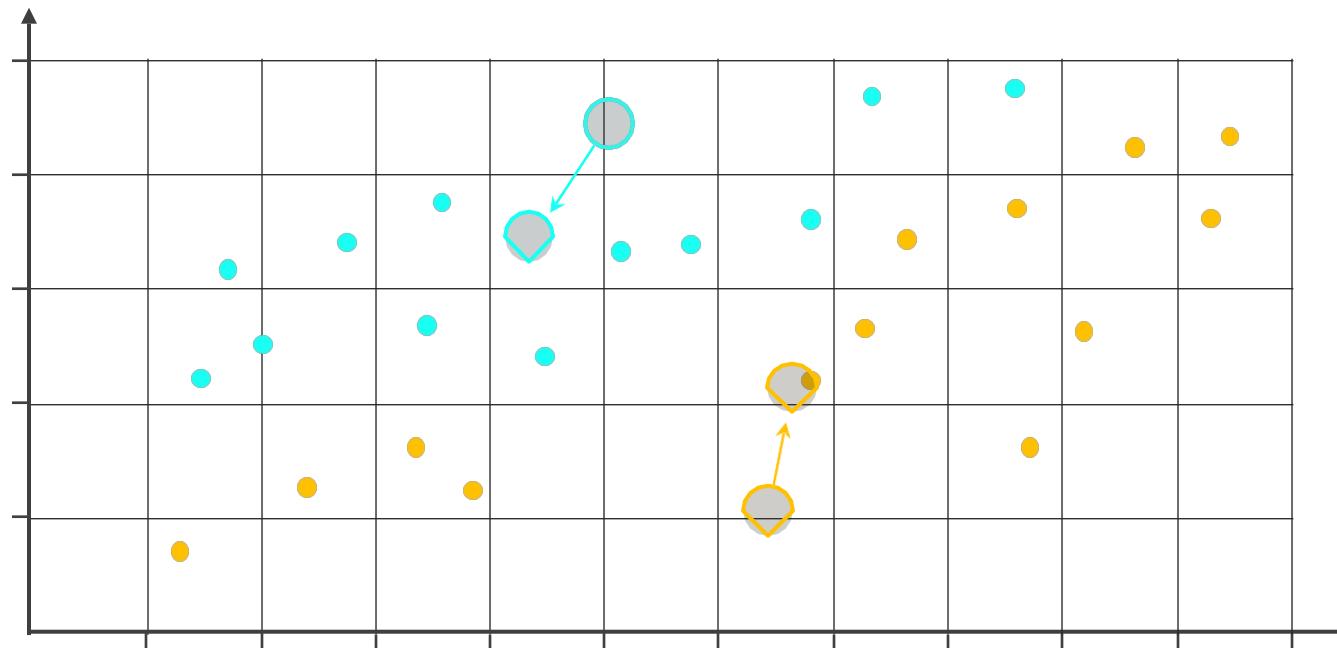
K-Means Clustering

Hierarchical Clustering

DBSCAN

Comparing Models

ÉTAPE 3 : Déplacer chaque centroïde vers la moyenne des observations qui lui sont attribuées, et réattribuer chaque observation au nouveau centroïde le plus proche.





Clustering K-Means

Clustering Basics

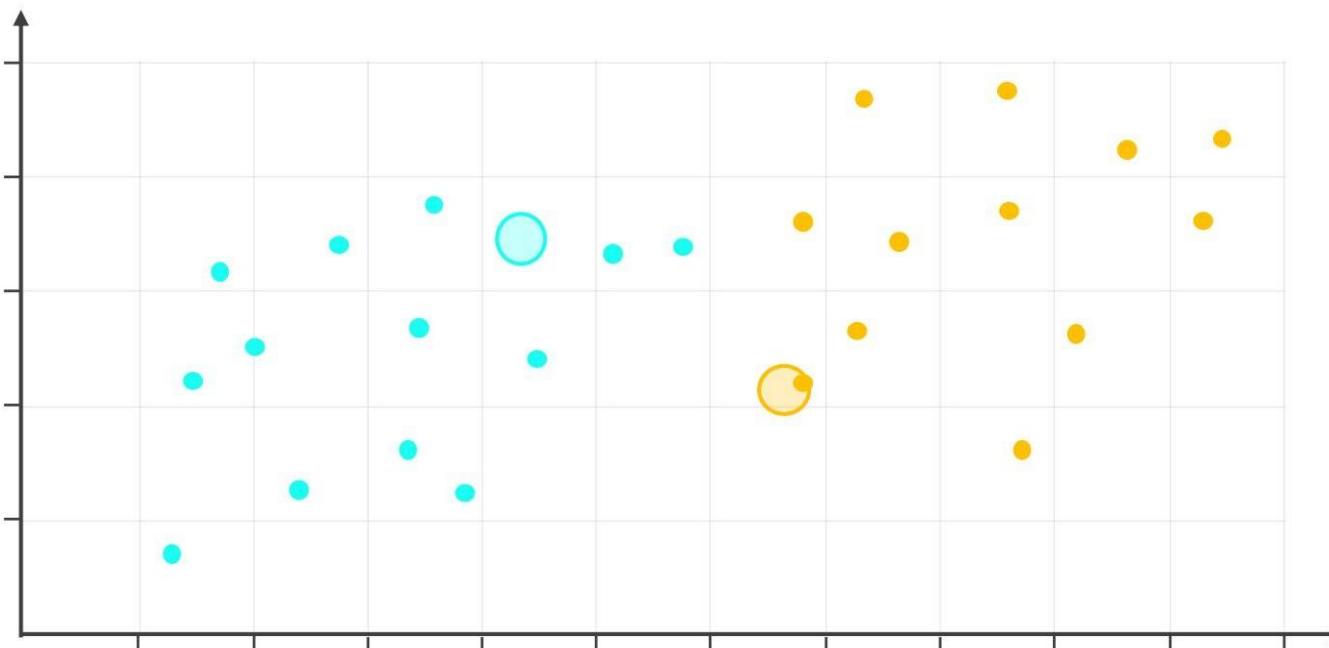
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Déplacer chaque centroïde vers la moyenne des observations qui lui sont assignées, puis réassigner chaque observation au nouveau centroïde le plus proche.





Clustering K-Means

Clustering Basics

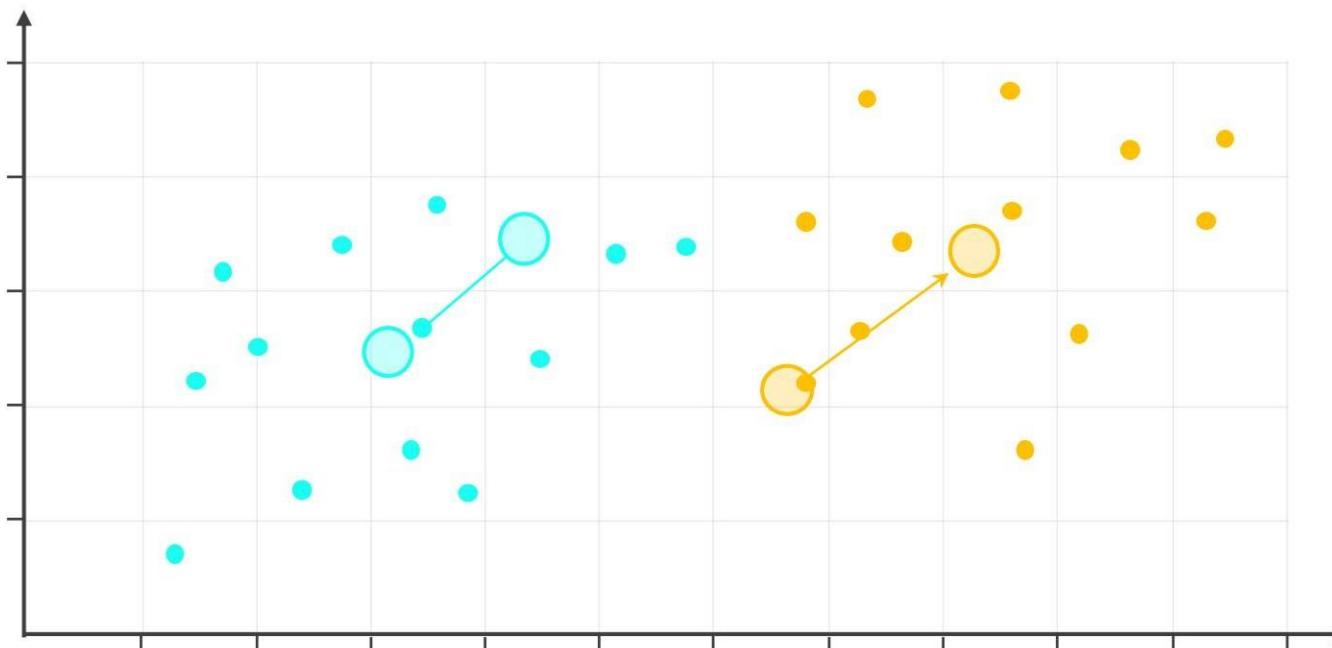
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Continuez à déplacer chaque centroïde vers la moyenne de ses observations assignées, jusqu'à ce que les groupes ne changent plus.





Clustering K-Means

Clustering Basics

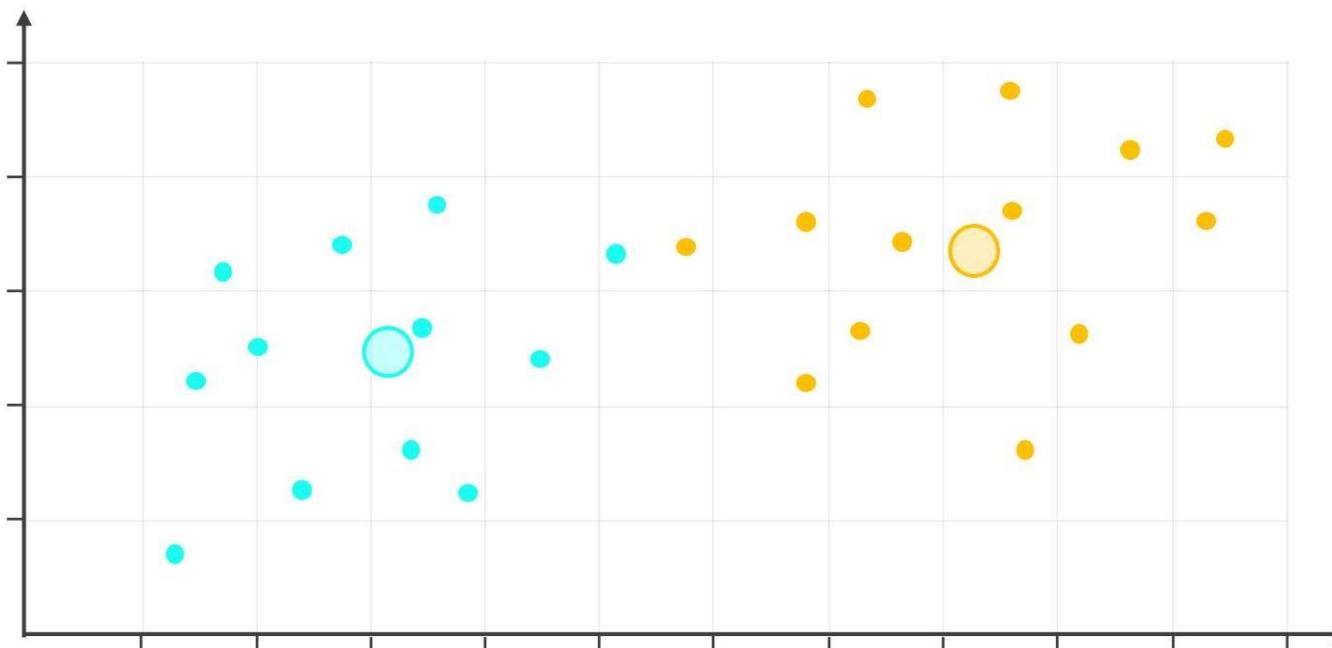
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Continuez à déplacer chaque centroïde vers la moyenne de ses observations assignées, jusqu'à ce que les groupes ne changent plus.





Clustering K-Means

Clustering Basics

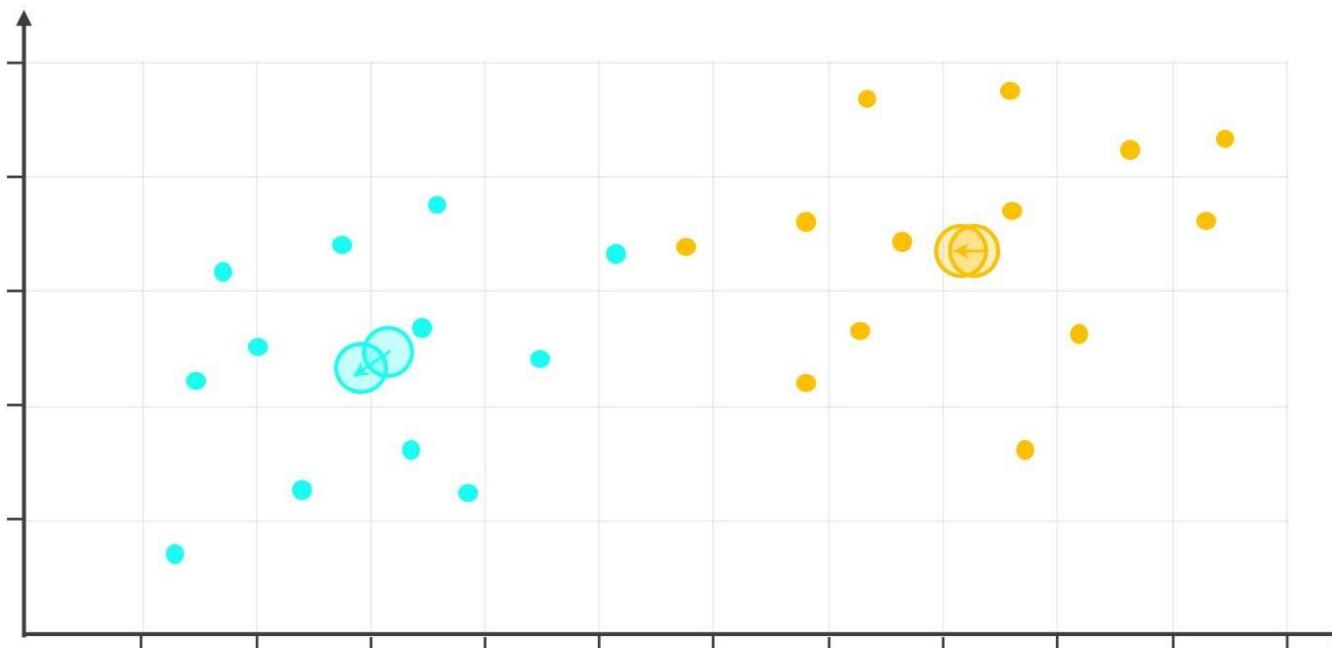
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Continuez à déplacer chaque centroïde vers la moyenne de ses observations assignées, jusqu'à ce que les groupes ne changent plus.





Clustering K-Means

Clustering Basics

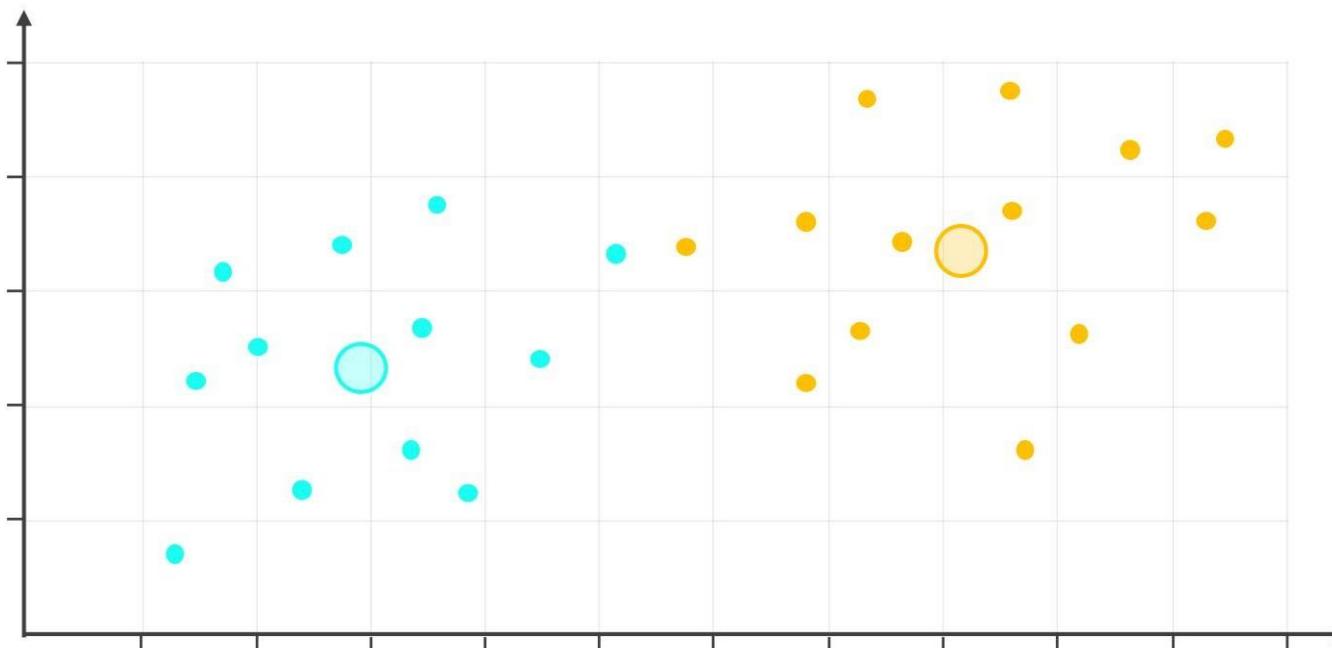
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Continuez à déplacer chaque centroïde vers la moyenne de ses observations assignées, jusqu'à ce que les groupes ne changent plus.





Clustering K-means en Python

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

```
from sklearn.cluster import KMeans  
  
kmeans = KMeans(n_clusters=2, n_init='auto', random_state=42)
```

Le nombre « k » de clusters
à identifier
(la valeur par défaut est 8)

Le nombre de modèles à ajuster avec
différents centroïdes initiaux, donnant le
meilleur résultat
(« auto » correspond à un seul modèle)

Définir une valeur pour random_state garantit
les mêmes résultats à chaque fois que le
modèle est ajusté.



CONSEIL DE PRO : Il est généralement conseillé de commencer avec 2 groupes et d'en ajouter progressivement, en comparant les résultats.



Clustering K-means en Python

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

EXEMPLE

Regrouper les lycéens selon leurs préférences en matière de divertissement afin que la bibliothèque locale puisse utiliser des publicités ciblées pour inciter les adolescents à lire davantage.

`df.head(10)`

	name	books	tv_shows	video_games
0	Aaliyah	0.5	4.6	4.9
1	Abigail	0.0	4.5	4.8
2	Addison	0.5	4.5	5.0
3	Adeline	3.5	4.5	6.6
4	Alana	2.8	3.8	5.6
5	Alexander	5.8	4.6	6.9
6	Alivia	4.2	4.5	6.7
7	Amara	3.2	4.5	5.6
8	Amelia	0.0	4.6	4.9
9	Annabelle	4.0	4.1	6.0



Certains élèves ne consacrent pratiquement aucun temps à la lecture chaque semaine



Les étudiants préfèrent passer la plupart de leur temps à jouer aux jeux vidéo.



Clustering K-means en Python

EXEMPLE

Regrouper les lycéens selon leurs préférences en matière de divertissement afin que la bibliothèque locale puisse utiliser des publicités ciblées pour inciter les adolescents à lire davantage.

Clustering Basics

K-Means Clustering

Hierarchical Clustering

DBSCAN

Comparing Models

Une fois le modèle ajusté, vous pouvez visualiser le cluster auquel chaque ligne a été assignée à l'aide de l' attribut `.labels` .



VISUALISATION DU CLUSTERING K-MEANS

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

```
# import plotting libraries
import matplotlib.pyplot as plt
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D

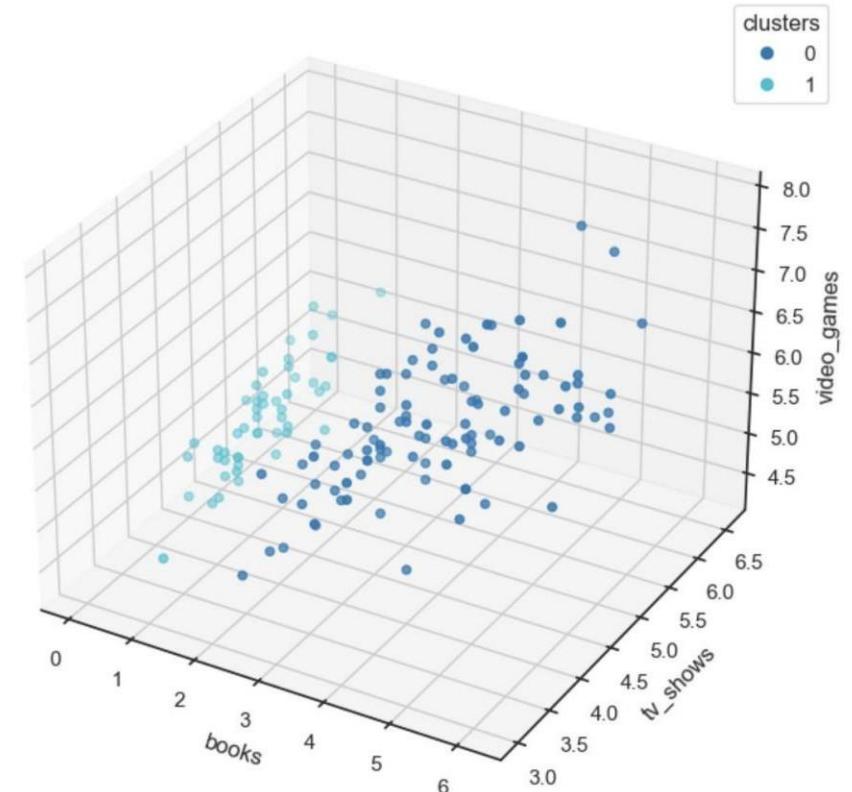
# combine the data and cluster labels
cluster_labels = pd.Series(kmeans.labels_, name='cluster')

# create a clean dataframe
df = pd.concat([data, cluster_labels], axis=1)

# create a 3d scatter plot
fig = plt.figure(figsize=(8, 6))
ax = Axes3D(fig)
fig.add_axes(ax)

# specify the data and labels
sc = ax.scatter(df['books'], df['tv_shows'], df['video_games'],
                 c=df['cluster'], cmap='tab10')
ax.set_xlabel('books')
ax.set_ylabel('tv_shows')
ax.set_zlabel('video_games')

# add a legend
plt.legend(*sc.legend_elements(), title='clusters',
           bbox_to_anchor=(1.05, 1));
```





INTERPRÉTATION DU CLUSTERING K-MEANS

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Vous pouvez interpréter les résultats d'un modèle K-Means en utilisant `.cluster_centers_` attribut et votre intuition

```
# view the column names  
data.columns
```

```
Index(['books', 'tv_shows', 'video_games'], dtype='object')
```

```
# view the cluster centers  
kmeans.cluster_centers_
```

```
array([[4.192, 4.314, 6.262],  
       [0.596, 5.13 , 5.006]])
```

Les étudiants du premier groupe dépensent en moyenne :

- 4,2 heures de lecture de livres
- 4,3 heures passées à regarder des émissions de télévision
- 6,3 heures passées à jouer aux jeux vidéo



Ces étudiants consomment une bonne quantité de chaque type de contenu, on pourrait donc les qualifier de « consommateurs de tous types de divertissement ».

Les étudiants du deuxième groupe dépensent en moyenne :

- 0,6 heure de lecture de livres
- 5,1 heures passées à regarder des émissions de télévision
- 5 heures passées à jouer aux jeux vidéo



Ces élèves ne lisent pas beaucoup de livres, on pourrait donc les qualifier de « non-lecteurs ».



VISUALISATION DES CENTRES DE GROUPEMENT

Clustering Basics

K-Means Clustering

Hierarchical Clustering

DBSCAN

Comparing Models

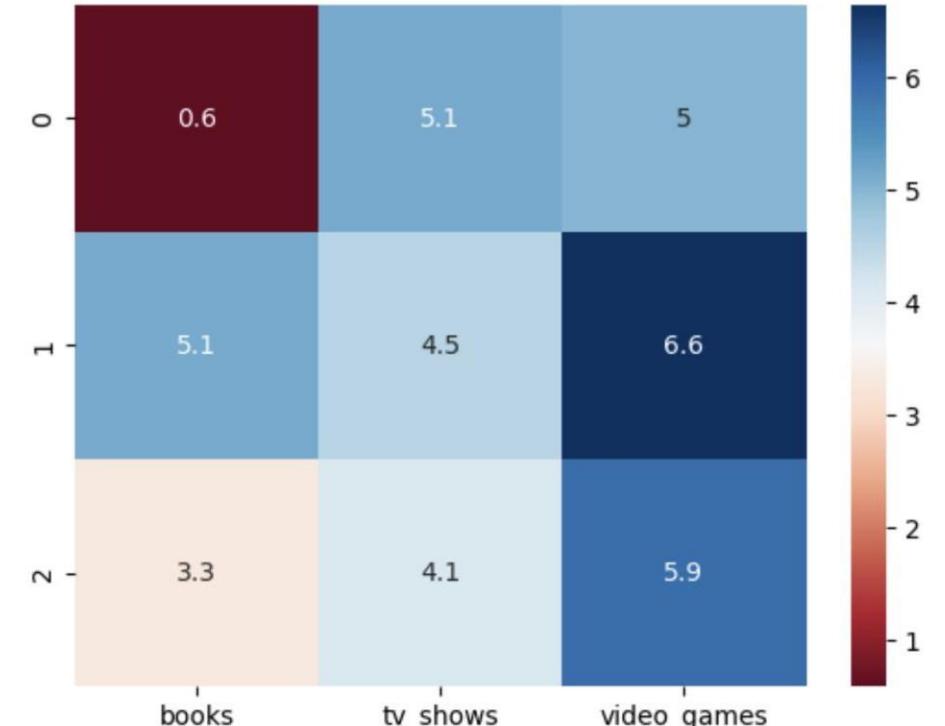
```
# view the cluster centers  
kmeans3.cluster_centers_
```

```
array([[0.596       , 5.13        , 5.006      ],  
       [5.14375    , 4.52708333, 6.63958333],  
       [3.31346154, 4.11730769, 5.91346154]])
```

```
# view the cluster centers in a dataframe  
cluster_centers3 = pd.DataFrame(kmeans3.cluster_centers_,  
                                  columns=data.columns)  
cluster_centers3
```

	books	tv_shows	video_games
0	0.596000	5.130000	5.006000
1	5.143750	4.527083	6.639583
2	3.313462	4.117308	5.913462

```
# view the cluster centers in a heatmap  
import seaborn as sns  
sns.heatmap(cluster_centers3, cmap='RdBu', annot=True);
```



Les élèves du groupe 0 ne lisent pas beaucoup de livres
Les élèves du groupe 1 consomment beaucoup de divertissements
Les élèves du groupe 2 préfèrent les jeux vidéo aux livres



INERTIE



Comment savoir quel est le « bon » nombre de clusters (K) ?

Clustering Basics

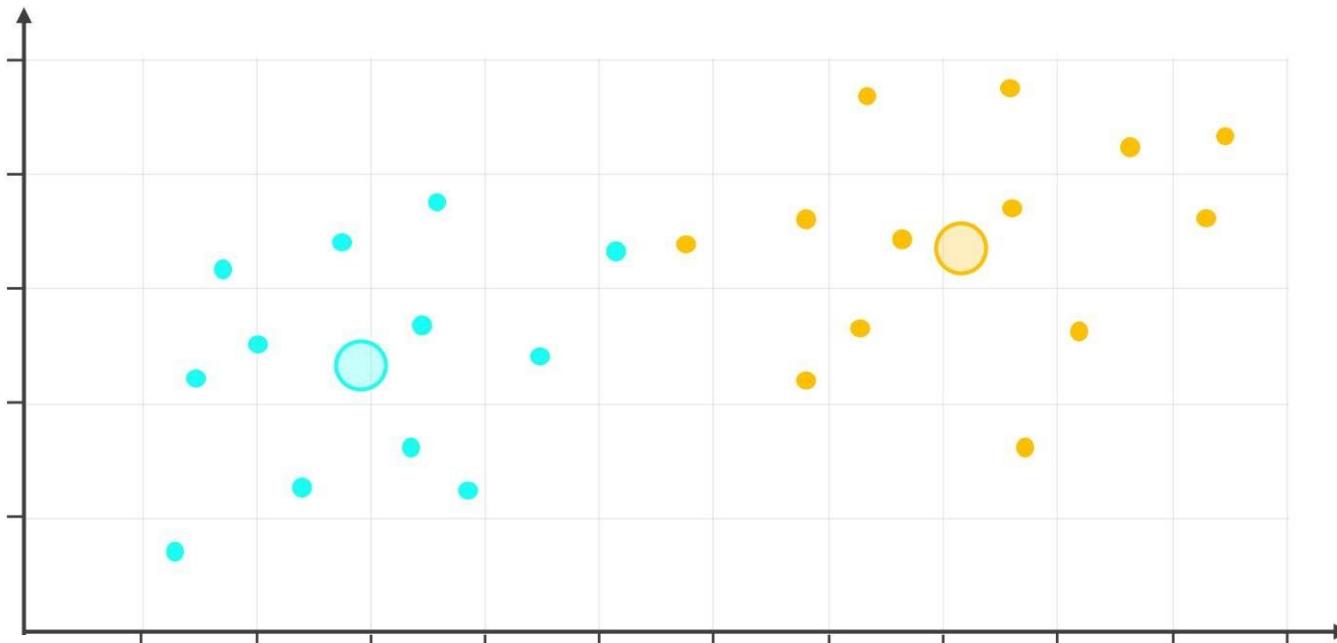
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

- Bien qu'il n'existe pas de nombre « idéal » ou « inadéquat » de grappes, vous pouvez utiliser l' inertie (également appelée somme des carrés intra-grappe ou WCSS) pour vous aider à prendre votre décision.





INERTIE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

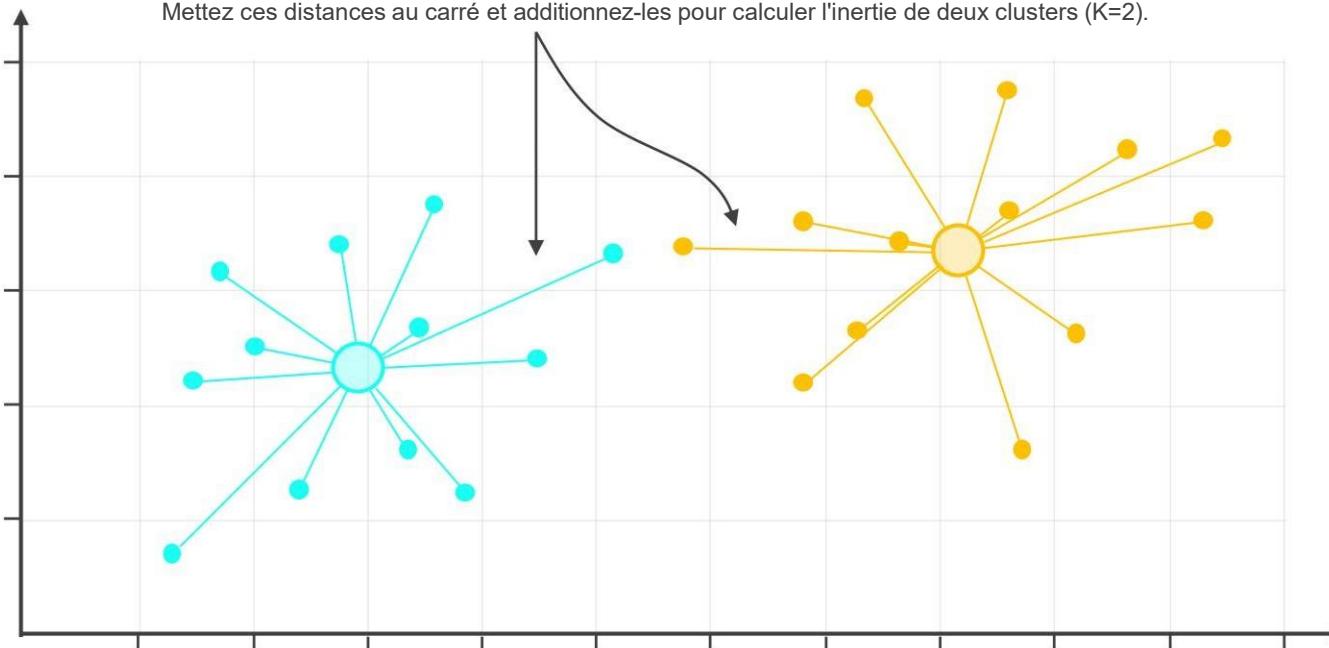
Comparing
Models



Comment savoir quel est le « bon » nombre de clusters (K) ?

- Bien qu'il n'existe pas de nombre « idéal » ou « inadéquat » de grappes, vous pouvez utiliser l' inertie (également appelée somme des carrés intra-grappe ou WCSS) pour vous aider à prendre votre décision.

Mettez ces distances au carré et additionnez-les pour calculer l'inertie de deux clusters (K=2).





INERTIE

Comment calculer l'inertie (WCSS) pour K=2 ?

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Données de départ : 6 points en 2D

Cluster 1 (cyan) :

A = (1, 2)

B = (2, 3)

C = (1, 3)

Cluster 2 (jaune) :

D = (7, 8)

E = (8, 7)

F = (9, 8)

$$\text{WCSS} = \sum_{x \in C_1} d(x, \mu_1)^2 + \sum_{x \in C_2} d(x, \mu_2)^2$$

où μ_i est le centroïde du cluster i



INERTIE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : CENTROÏDES

Centroïde μ_1 du cluster 1 (cyan)

Points : A=(1,2), B=(2,3), C=(1,3)

$$\mu_1 = ((1+2+1)/3, (2+3+3)/3) = (1.33, 2.67)$$

Centroïde μ_2 du cluster 2 (jaune)

Points : D=(7,8), E=(8,7), F=(9,8)

$$\mu_2 = ((7+8+9)/3, (8+7+8)/3) = (8, 7.67)$$

Note : Le centroïde est la moyenne des coordonnées de tous les points du cluster.



INERTIE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 2 : DISTANCES AU CARRÉ

Cluster 1 (cyan)

Centroïde $\mu_1 = (1.33, 2.67)$

Point A = (1, 2)

$$d(A, \mu_1)^2 = (1-1.33)^2 + (2-2.67)^2 = 0.11 + 0.45 = \mathbf{0.56}$$

Point B = (2, 3)

$$d(B, \mu_1)^2 = (2-1.33)^2 + (3-2.67)^2 = 0.45 + 0.11 = \mathbf{0.56}$$

Point C = (1, 3)

$$d(C, \mu_1)^2 = (1-1.33)^2 + (3-2.67)^2 = 0.11 + 0.11 = \mathbf{0.22}$$

Somme Cluster 1 : $0.56 + 0.56 + 0.22 = 1.34$



INERTIE

ÉTAPE 2 : DISTANCES AU CARRÉ

Cluster 2 (jaune)

Centroïde $\mu_2 = (8, 7.67)$

Point D = (7, 8)

$$d(D, \mu_2)^2 = (7-8)^2 + (8-7.67)^2 = 1 + 0.11 = 1.11$$

Point E = (8, 7)

$$d(E, \mu_2)^2 = (8-8)^2 + (7-7.67)^2 = 0 + 0.45 = 0.45$$

Point F = (9, 8)

$$d(F, \mu_2)^2 = (9-8)^2 + (8-7.67)^2 = 1 + 0.11 = 1.11$$

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Somme Cluster 2 : $1.11 + 0.45 + 1.11 = 2.67$



INERTIE

ÉTAPE 3 : INERTIE TOTALE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

$$\text{WCSS} = \sum_{x \in C_1} d(x, \mu_1)^2 + \sum_{x \in C_2} d(x, \mu_2)^2$$

Cluster 1

1.34

+

Cluster 2

2.67

Inertie totale (WCSS) pour K=2

4.01

Interprétation : Plus l'inertie est faible, plus les points sont proches de leurs centroïdes.

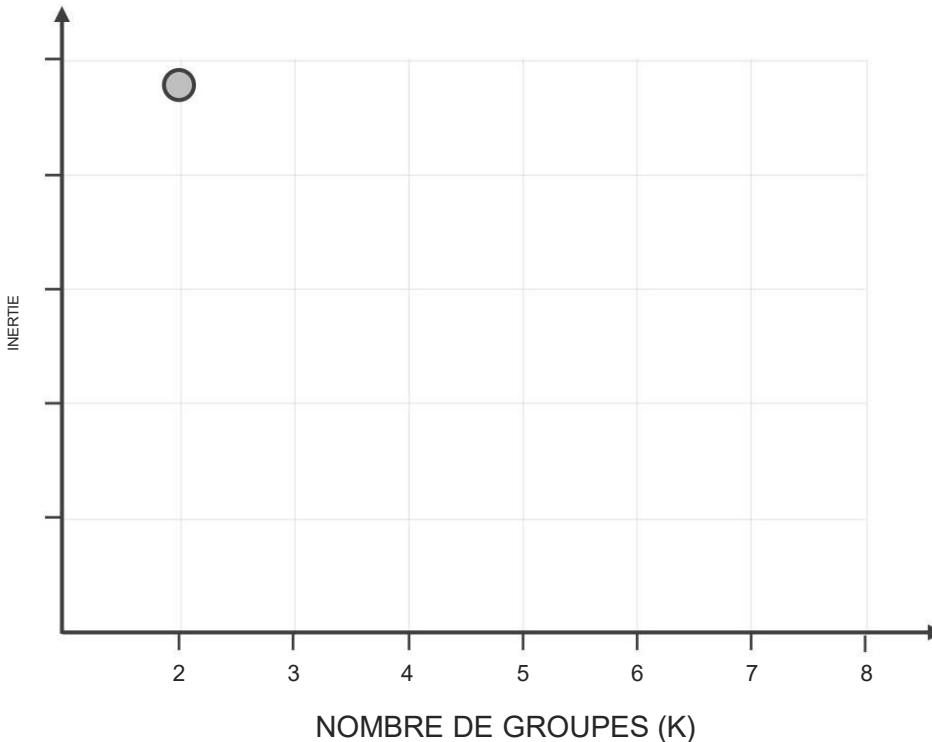


INERTIE



Comment savoir quel est le « bon » nombre de clusters (K) ?

- Bien qu'il n'existe pas de nombre « idéal » ou « inadéquat » de grappes, vous pouvez utiliser l' inertie (également appelée somme des carrés intra-grappe ou WCSS) pour vous aider à prendre votre décision.



Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models



INERTIE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

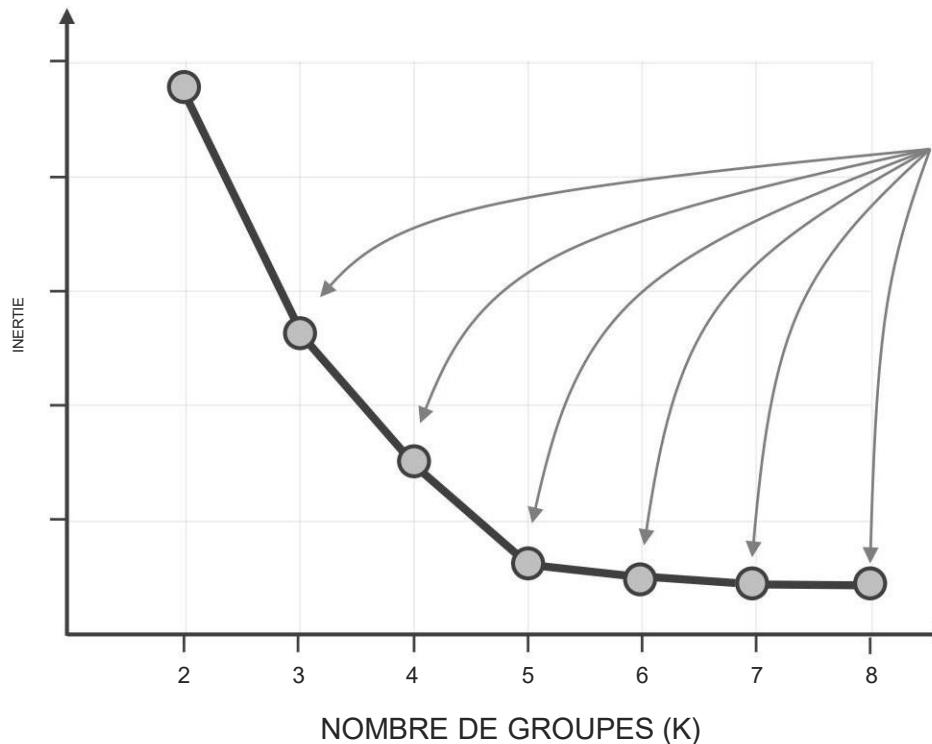
DBSCAN

Comparing
Models



Comment savoir quel est le « bon » nombre de clusters (K) ?

- Bien qu'il n'existe pas de nombre « idéal » ou « inadéquat » de grappes, vous pouvez utiliser l' inertie (également appelée somme des carrés intra-grappe ou WCSS) pour vous aider à prendre votre décision.



Relancez le modèle avec des clusters supplémentaires et tracez l'inertie pour chaque valeur de K.



INERTIE



Comment savoir quel est le « bon » nombre de clusters (K) ?

Clustering Basics

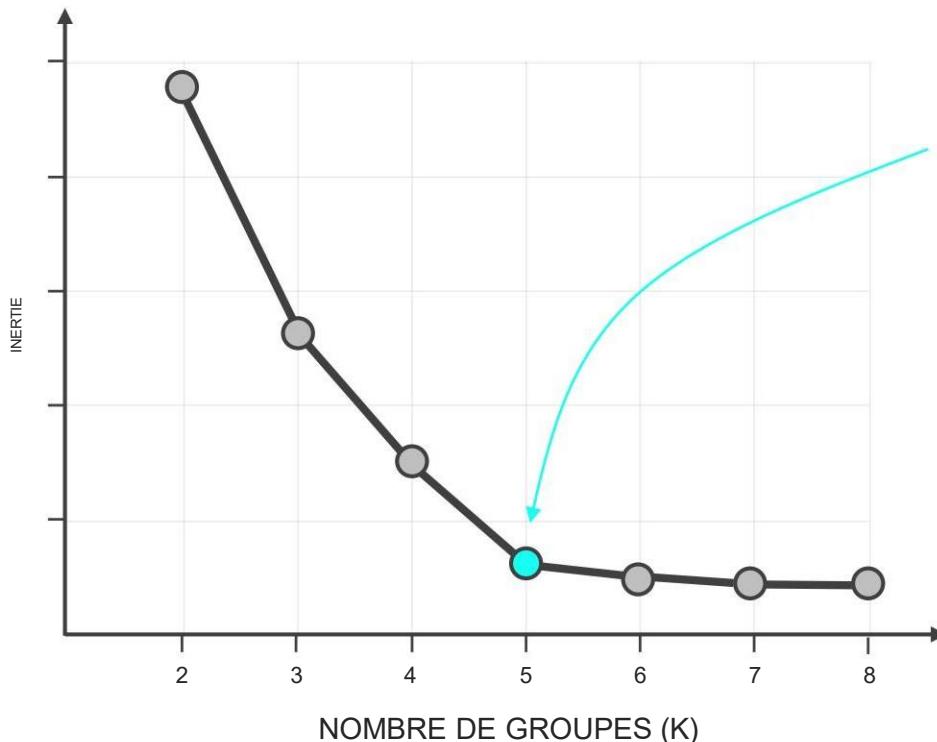
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

- Bien qu'il n'existe pas de nombre « idéal » ou « inadéquat » de grappes, vous pouvez utiliser l'inertie (également appelée somme des carrés intra-grappe ou WCSS) pour vous aider à prendre votre décision.



Recherchez un « coude » ou un point d'infexion, où l'ajout d'un autre groupe a un impact relativement faible sur l'inertie (dans ce cas où K=5).



CONSEIL DE PRO : Considérez ceci comme une ligne directrice, et non comme une règle stricte.



REPRÉSENTATION DE L'INERTIE EN PYTHON

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

```
# fit k-means models for 2-15 clusters, note the inertia scores
inertia_values = []

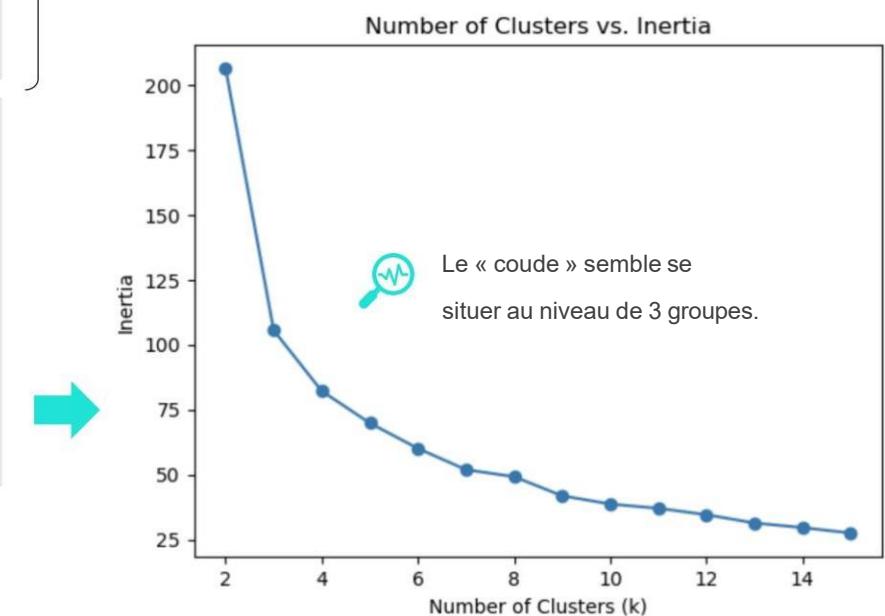
for k in range(2, 16):
    kmeans = KMeans(n_clusters=k, n_init=10, random_state=30)
    kmeans.fit(data)
    inertia_values.append(kmeans.inertia_)

# plot the inertia values
import matplotlib.pyplot as plt

# turn the list into a series for plotting
inertia_series = pd.Series(inertia_values, index=range(2, 16))

# plot the data
inertia_series.plot(marker='o')
plt.xlabel("Number of Clusters (k)")
plt.ylabel("Inertia")
plt.title("Number of Clusters vs. Inertia");
```

Ajustez les modèles en utilisant 2 à 15 clusters et ajoutez leurs valeurs d'inertie à une liste.





RÉGLAGE D'UN MODÈLE K-MEANS

Une partie du processus de clustering consiste à revenir sur différentes étapes de préparation et de modélisation des données afin d' optimiser un modèle avant de sélectionner le meilleur.

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Préparation des données

- Nettoyage des données : suppression des valeurs aberrantes, etc.
- Ingénierie des fonctionnalités : Création de fonctionnalités pertinentes, etc.
- Sélection des fonctionnalités : Plus de fonctionnalités ne signifie pas un meilleur modèle !
- Mise à l'échelle : K-Means étant un algorithme basé sur la distance, il est conseillé de mettre les données à l'échelle.

Modélisation

- Essayer un nombre différent de clusters : si vos clusters sont très différents à chaque exécution, L'algorithme K-Means avec ce nombre spécifique de clusters n'est peut-être pas le plus adapté à vos données ;essayez donc d'utiliser un nombre de clusters différent.
- Essayer d'autres modèles de clustering : le modèle K-Means fonctionne mieux lorsque les clusters sont majoritairement composés de clusters. De forme circulaire, mais des algorithmes comme le clustering hiérarchique (à suivre !) peuvent résoudre ce problème.



CHOISIR LE MEILLEUR MODÈLE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Il n'existe pas de modèle de clustering idéal , mais un bon modèle comporte des clusters pertinents, qui capturent des tendances et contribuent à résoudre le problème métier.

Vous pouvez explorer les groupes à l'aide d'approches basées sur les données, telles que :

- Comparaison des affectations de clusters pour l'ensemble de données et les lignes de données individuelles
- Comparaison des métriques des différents modèles – inertie, score de silhouette (à venir !), etc.
- Tester les modèles de clustering sur des données non vues (nous y reviendrons à la fin de cette section !)



CHOISIR LE MEILLEUR MODÈLE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

EXEMPLE

Regrouper les lycéens selon leurs préférences en matière de divertissement afin que la bibliothèque locale puisse utiliser des publicités ciblées pour inciter les adolescents à lire davantage.

```
# view the number of students in each cluster for model 1
model1_names.value_counts()
```

```
model1_clusters
Prefer Video Games to Books    52
Non-Readers                      50
Entertainment Enthusiasts        48
Name: count, dtype: int64
```

```
# view the number of students in each cluster for model 2
model2_names.value_counts()
```

```
model2_clusters
Typical Students                52
Less Entertainment (Few Books)   50
Less Screens                      36
Entertainment Enthusiasts (Many Video Games) 12
Name: count, dtype: int64
```

```
# compare the cluster assignments and means of both models
(cluster_names.groupby(['model1_clusters', 'model2_clusters'])
[[ 'books', 'tv_shows', 'video_games']]
.mean())
```

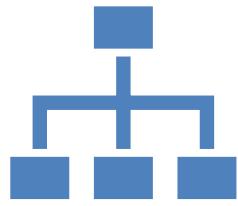
model1_clusters		model2_clusters		
Entertainment Enthusiasts	Non-Readers	Entertainment Enthusiasts (Many Video Games)	Less Screens	Typical Students
Prefer Video Games to Books		5.125000	4.691667	7.475000
		0.596000	5.130000	5.006000



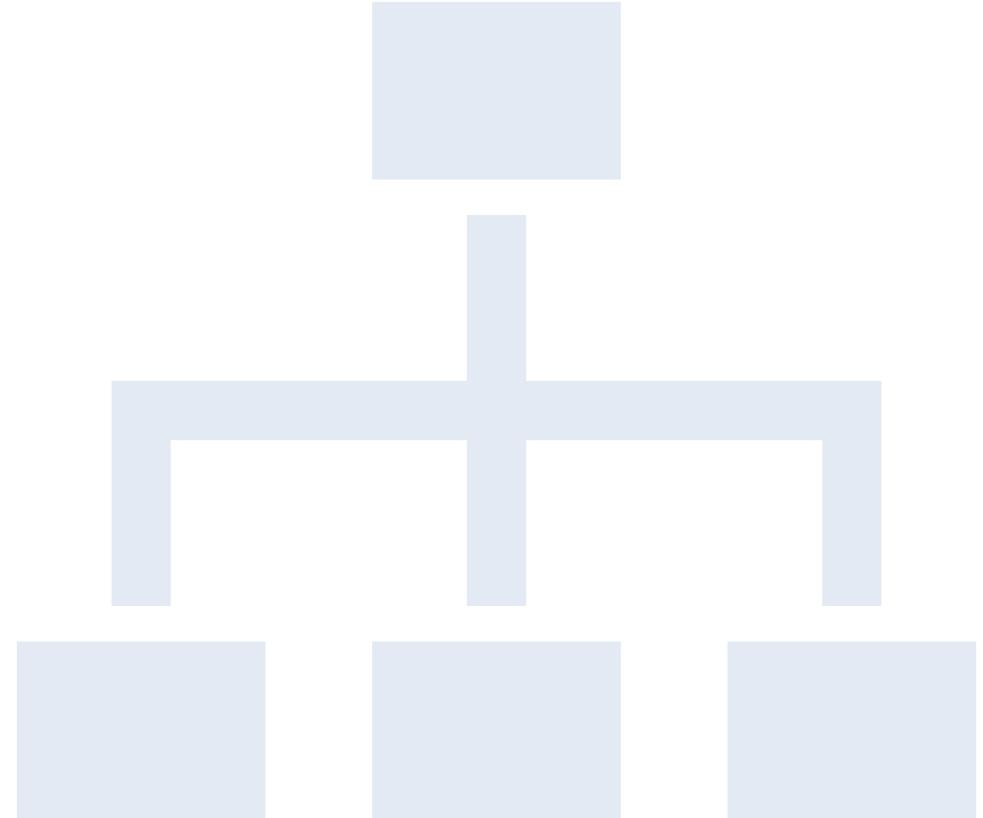
Le modèle 1 comporte 3 groupes de taille égale, tandis que le modèle 2 comporte des groupes plus spécifiques.



Les groupes « moins d'écrans » et « passionnés de divertissement » sont assez similaires et pourraient éventuellement être regroupés en un seul cluster.



CLUSTERING HIÉRARCHIQUE





CLUSTERING HIÉRARCHIQUE

Clustering Basics

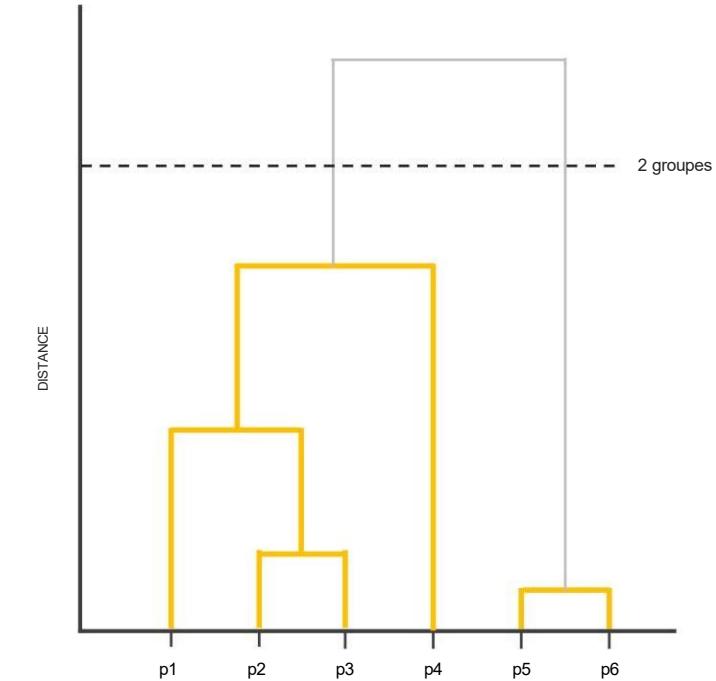
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Le clustering hiérarchique est une technique de clustering qui crée des clusters en regroupant les points de données similaires.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Le clustering hiérarchique est une technique de clustering qui crée des clusters en regroupant des points de données similaires*.

Voici comment ça fonctionne :

1. Sur un nuage de points, trouvez les deux points les plus proches et regroupez-les en un cluster.
2. Trouvez ensuite les deux points ou groupes les plus proches et regroupez-les en un seul groupe.
3. Répétez le processus de combinaison des paires de points ou groupes les plus proches jusqu'à ce que vous finissiez par n'avoir qu'un seul groupe.

Ce processus est visualisé à l'aide d'un diagramme arborescent appelé dendrogramme, qui illustre la relation hiérarchique entre les groupes.

*Ce type de clustering est connu sous le nom de clustering agglomératif ou « ascendant » (par opposition au clustering divisif ou « descendant », qui est beaucoup moins fréquent).



CLUSTERING HIÉRARCHIQUE

Clustering Basics

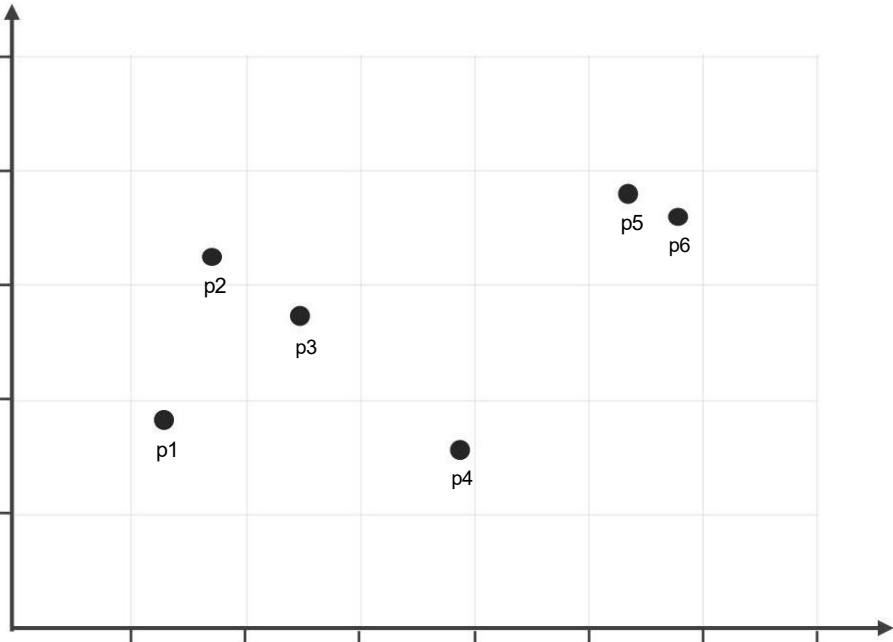
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : Trouvez les deux points les plus proches et regroupez-les en un cluster.



Comment définissez-vous « le plus proche » ?



CLUSTERING HIÉRARCHIQUE

Clustering Basics

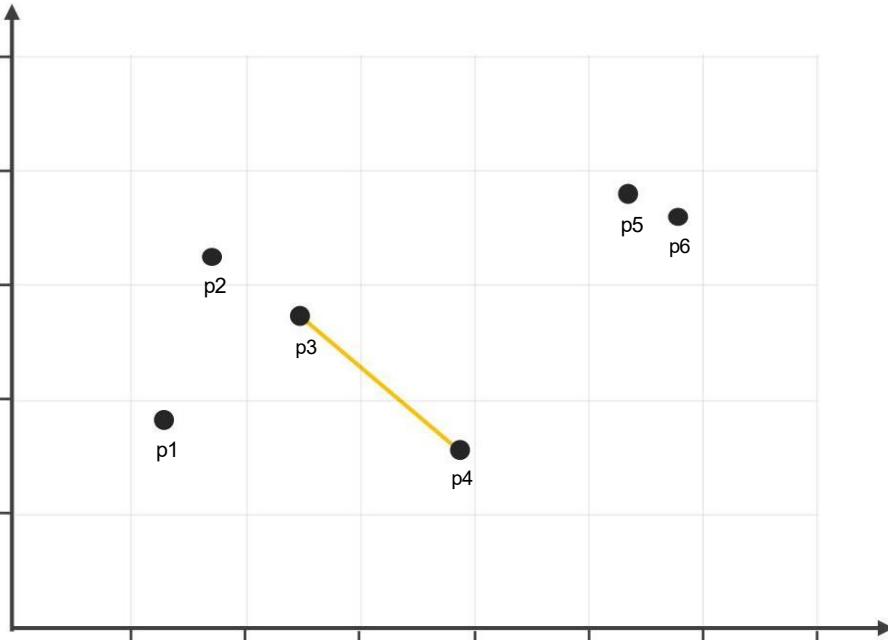
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : Trouvez les deux points les plus proches et regroupez-les en un cluster.



Comment définissez-vous « le plus proche » ?

- Le plus souvent, le
- On utilise la distance euclidienne



CLUSTERING HIÉRARCHIQUE

Clustering Basics

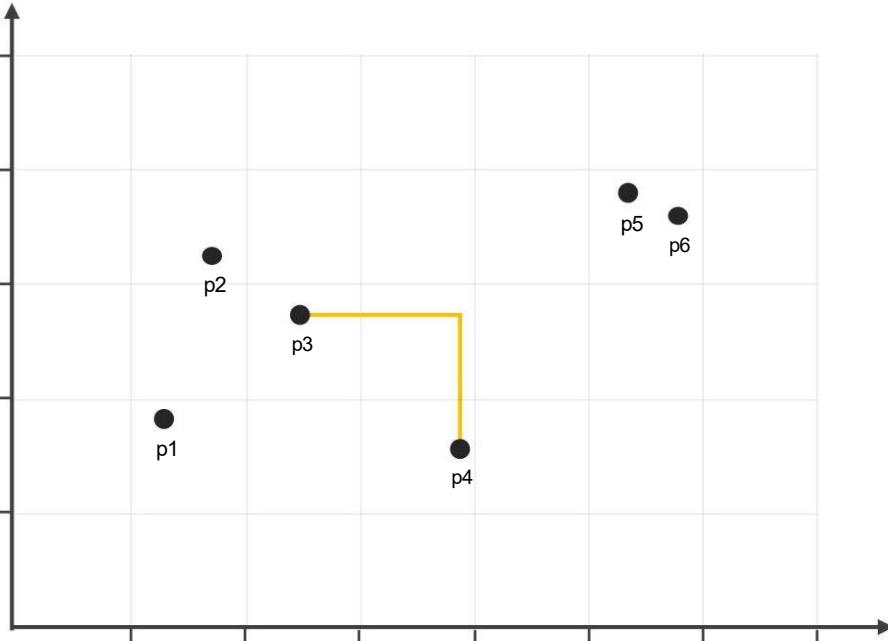
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : Trouvez les deux points les plus proches et regroupez-les en un cluster.



Comment définissez-vous « le plus proche » ?

- Le plus souvent, le
On utilise la distance euclidienne
- Sinon, il y a
distance de Manhattan



CLUSTERING HIÉRARCHIQUE

Clustering Basics

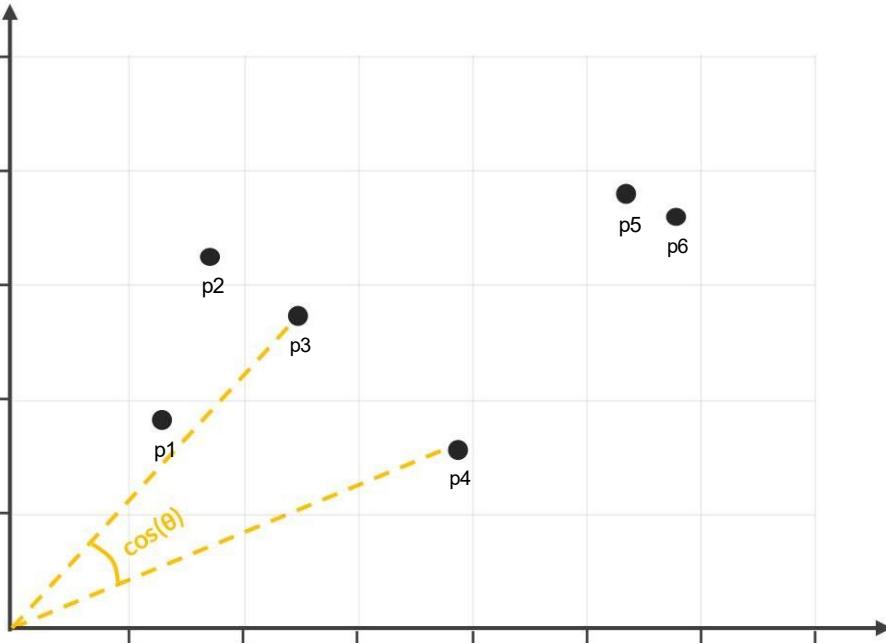
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : Trouvez les deux points les plus proches et regroupez-les en un cluster.



Comment définissez-vous « le plus proche » ?

- Le plus souvent, le
On utilise la distance euclidienne
- Sinon, il y a
distance de Manhattan
- Et la distance cosinus



CLUSTERING HIÉRARCHIQUE

Clustering Basics

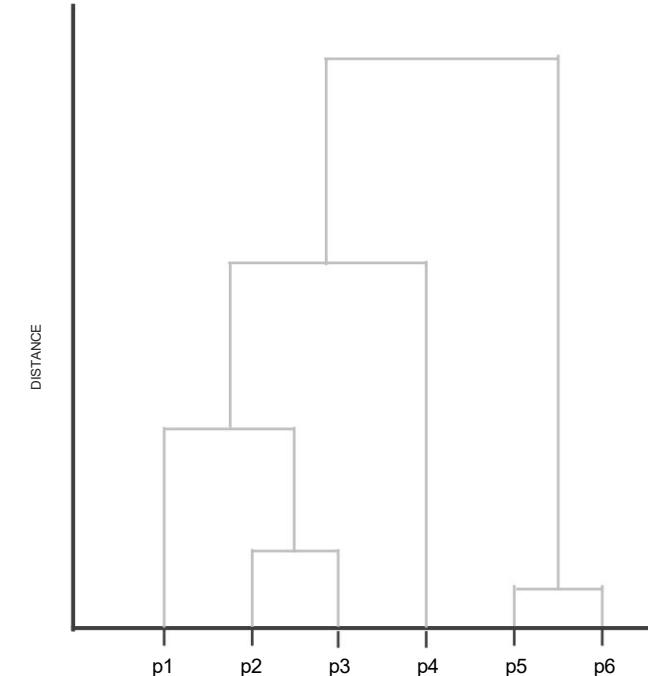
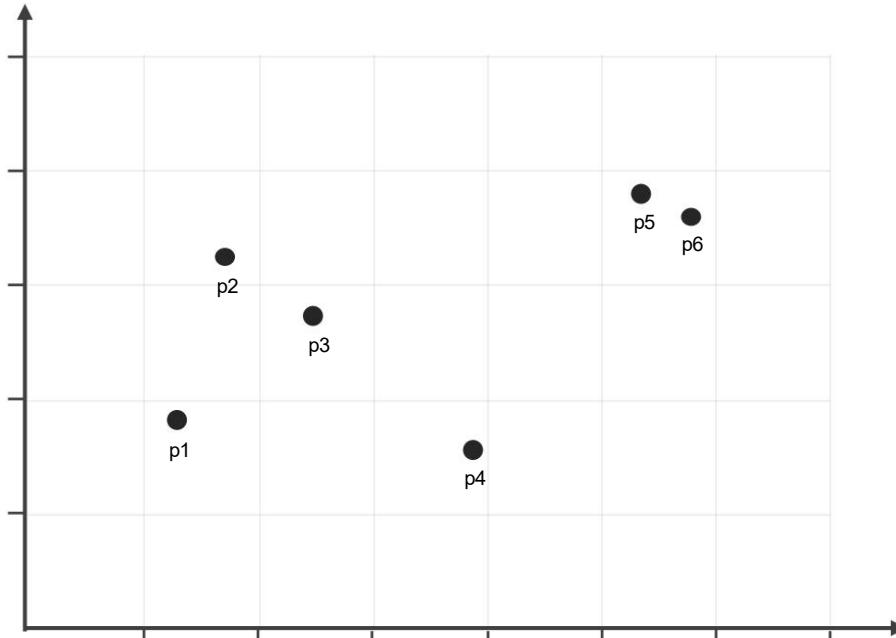
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : Trouvez les deux points les plus proches et regroupez-les en un cluster.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

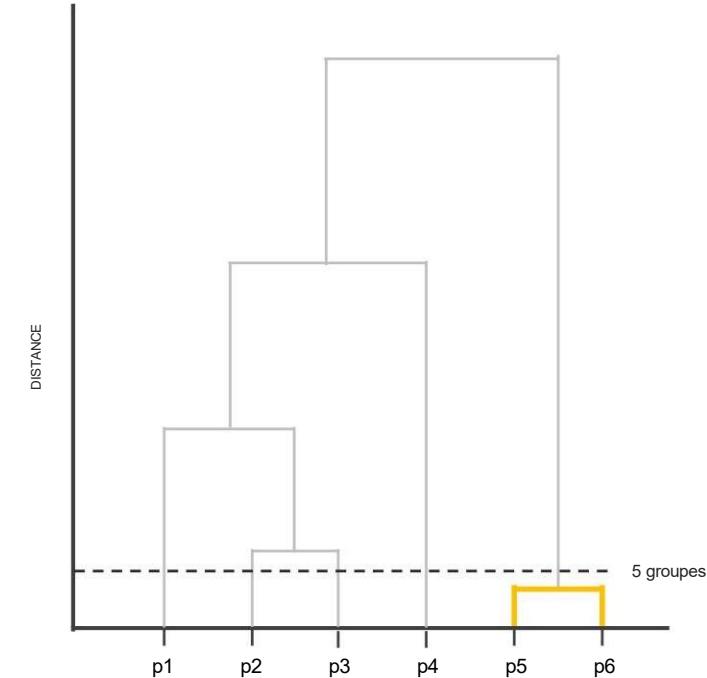
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : Trouvez les deux points les plus proches et regroupez-les en un cluster.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

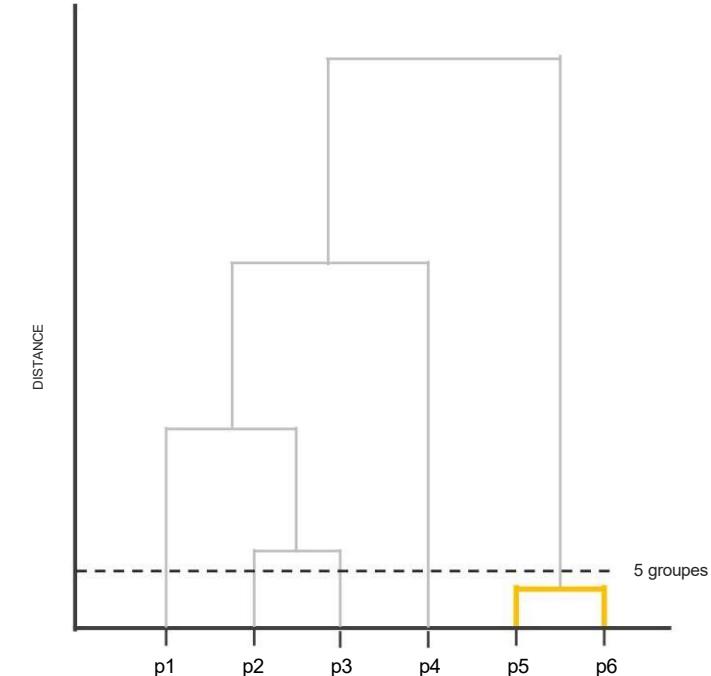
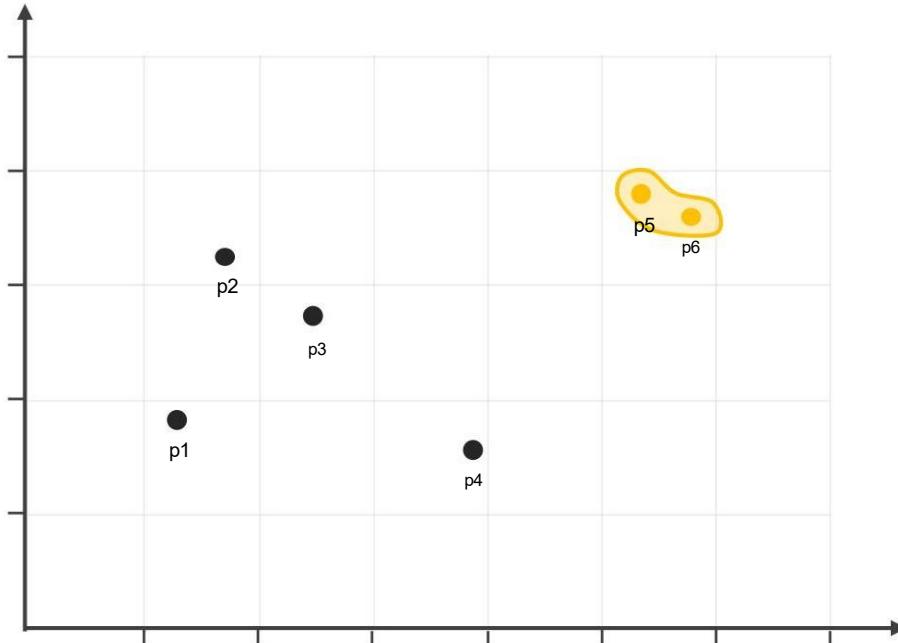
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 2 : Trouvez les deux points/groupes les plus proches et regroupez-les.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

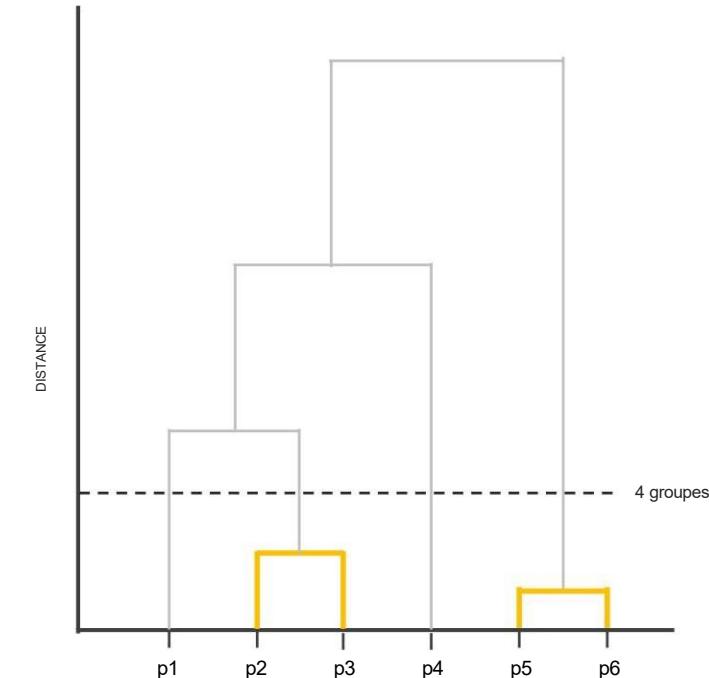
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 2 : Trouvez les deux points/groupes les plus proches et regroupez-les.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

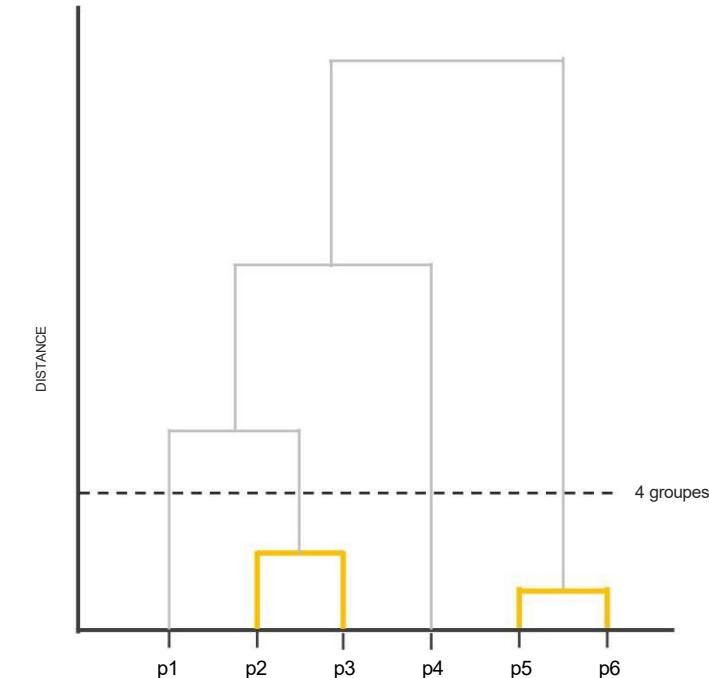
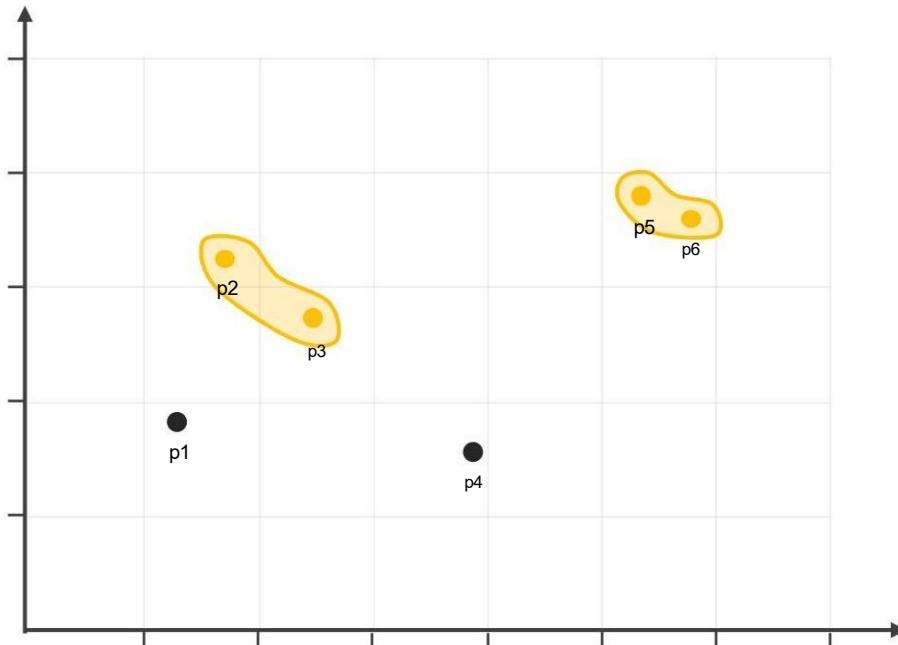
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

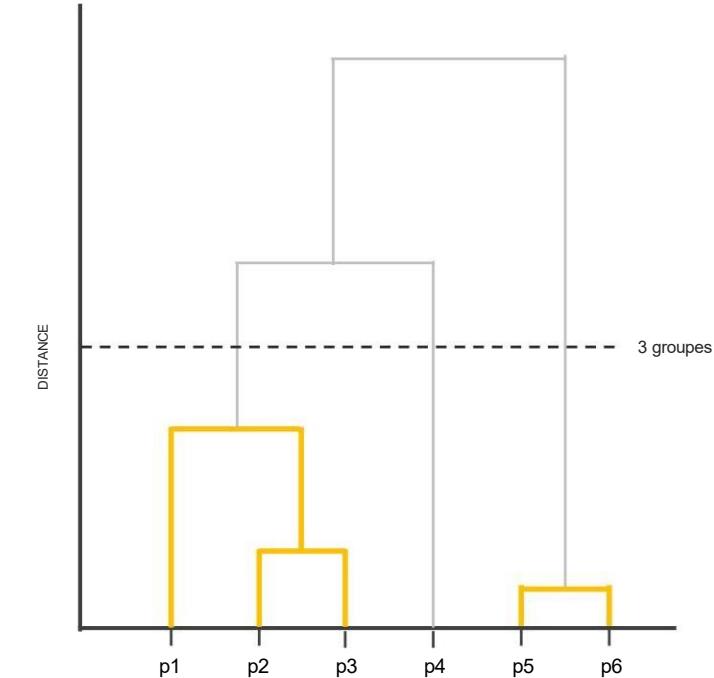
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.



Comment définir la distance entre les clusters ?



CLUSTERING HIÉRARCHIQUE

Clustering Basics

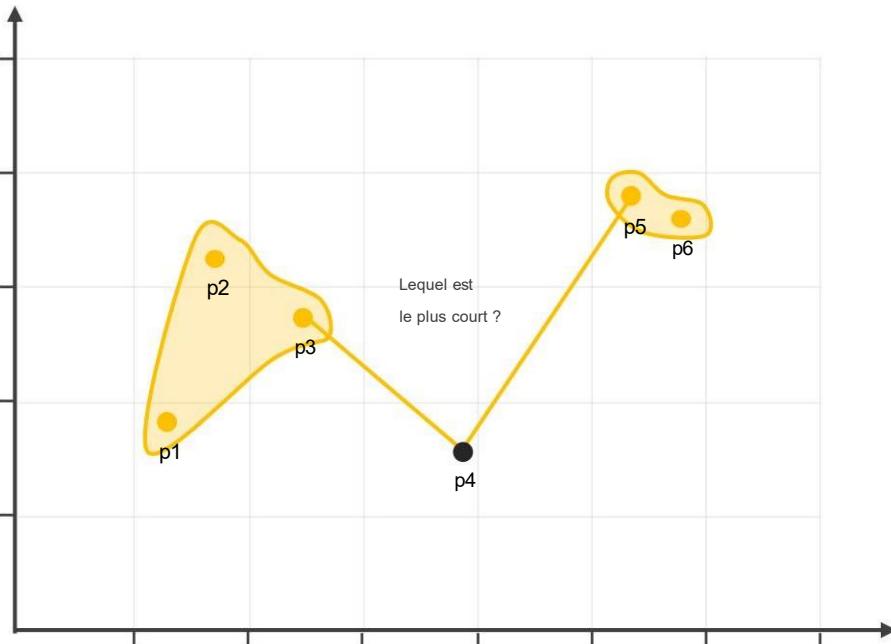
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.



Comment définir la distance entre les clusters ?

- Liaison simple (la plus proche)



CLUSTERING HIÉRARCHIQUE

Clustering Basics

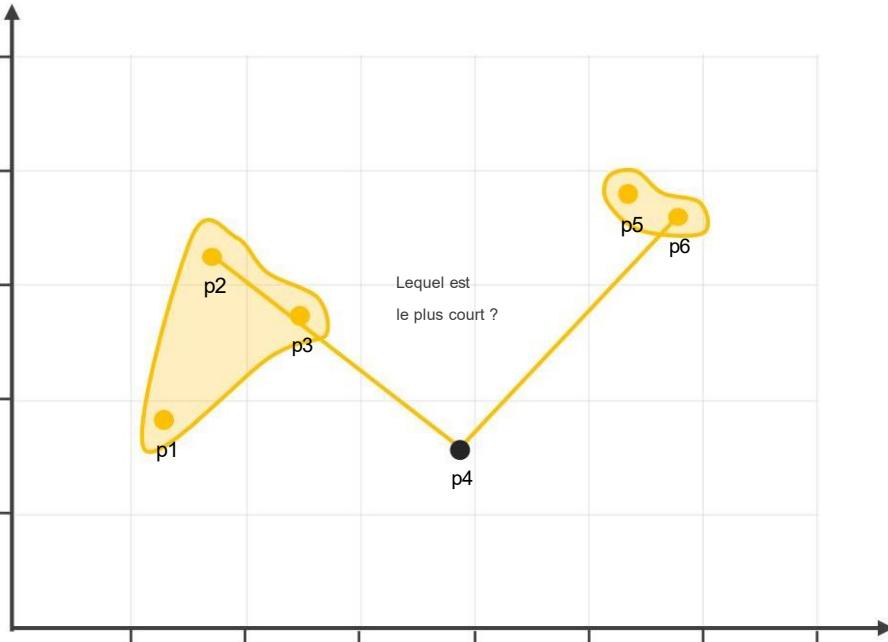
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.



Comment définir la distance entre les clusters ?

- Liaison simple (la plus proche)
- Liaison complète (la plus éloignée)



CLUSTERING HIÉRARCHIQUE

Clustering Basics

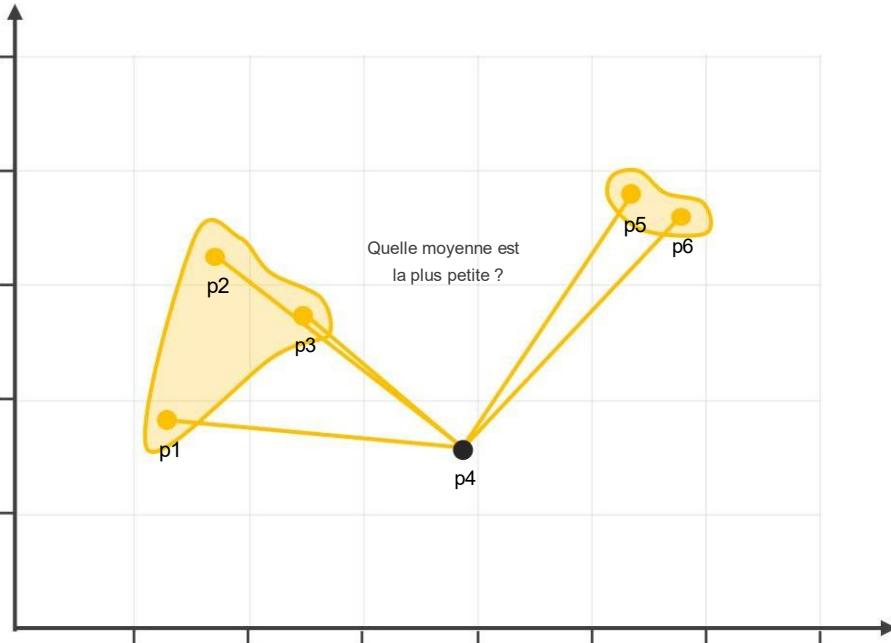
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.



Comment définir la distance entre les clusters ?

- Liaison simple (la plus proche)
- Liaison complète (la plus éloignée)
- Liaison moyenne (toutes les paires)



CLUSTERING HIÉRARCHIQUE

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.



Comment définir la distance entre les clusters ?

- Liaison simple (la plus proche)
- Liaison complète (la plus éloignée)
- Liaison moyenne (toutes les paires)
- Méthode de Ward (variance)



CLUSTERING HIÉRARCHIQUE

Clustering Basics

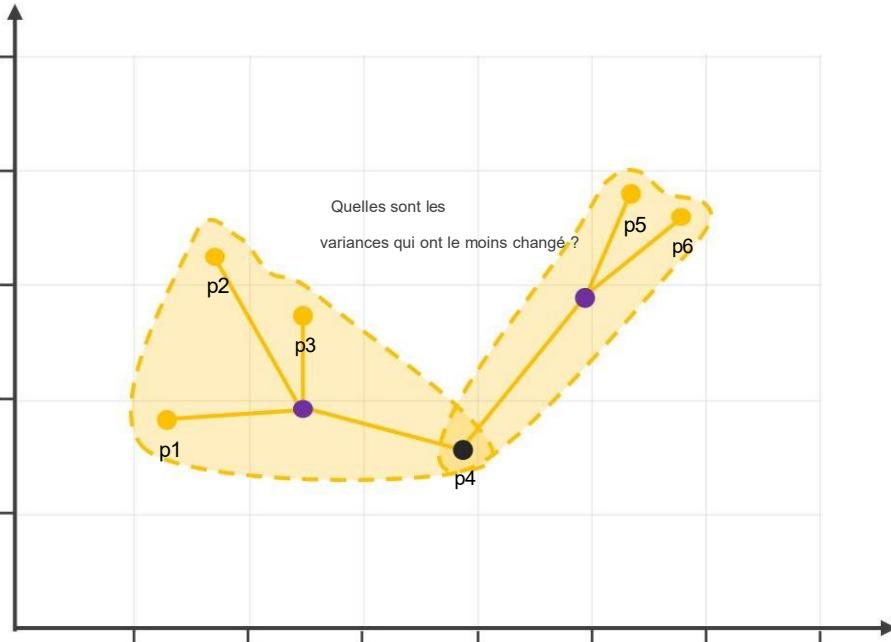
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.



Comment définir la distance entre les clusters ?

- Liaison simple (la plus proche)
- Liaison complète (la plus éloignée)
- Liaison moyenne (toutes les paires)
- Méthode de Ward (variance)



CLUSTERING HIÉRARCHIQUE

LA MÉTHODE DE WARD

La méthode de Ward est une approche de clustering hiérarchique qui cherche à créer des groupes homogènes. À chaque étape, elle fusionne les deux clusters qui entraînent la plus petite augmentation de la variance totale.

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Comment fonctionne l'algorithme ?

1. Au départ, chaque observation constitue son propre cluster
2. L'algorithme calcule pour toutes les paires : quelle augmentation de variance si on les fusionne ?
3. Il fusionne la paire qui minimise cette augmentation
4. Le processus se répète jusqu'à obtenir un seul cluster

Pourquoi Ward est-elle si populaire ?

Cette méthode produit généralement des clusters **bien séparés et équilibrés**, ce qui facilite l'interprétation. Elle est efficace pour identifier des groupes naturels.

⚠ Point important

Ward est très sensible aux échelles. Il est impératif de standardiser vos données avant d'appliquer cette méthode.



Sur un dendrogramme, la hauteur représente l'augmentation de variance. Coupez là où vous voyez un grand saut.



CLUSTERING HIÉRARCHIQUE

Clustering Basics

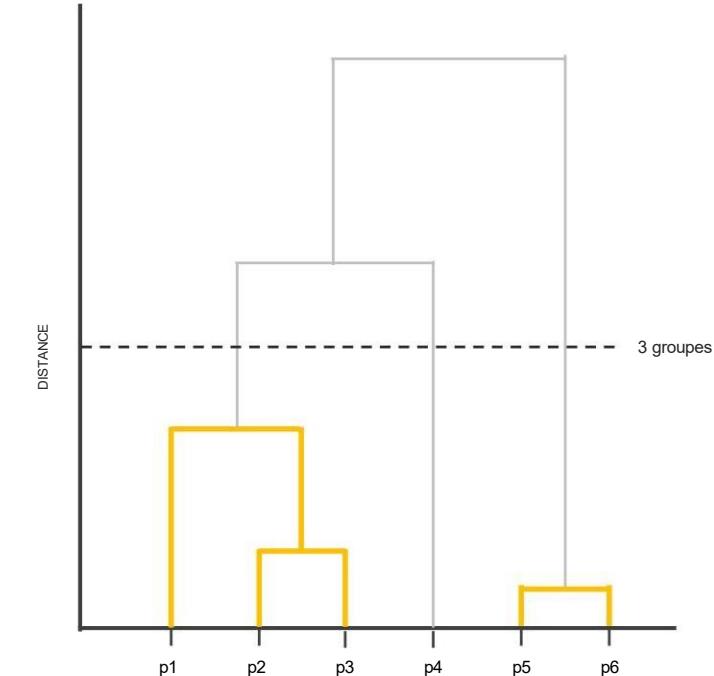
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

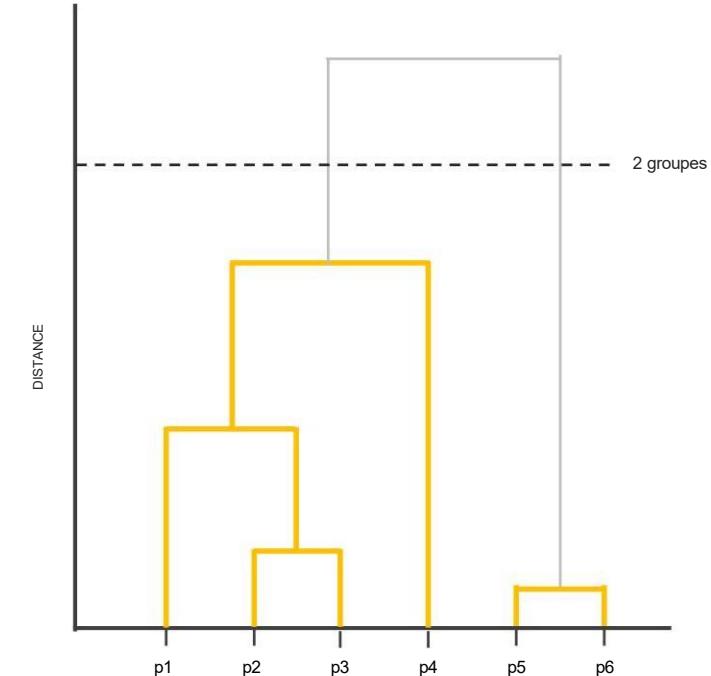
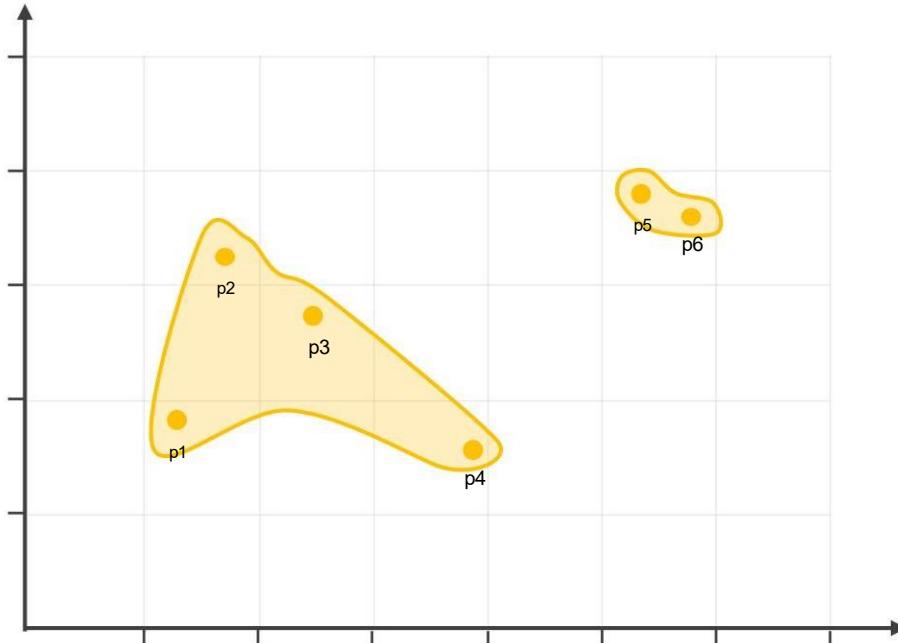
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.





CLUSTERING HIÉRARCHIQUE

Clustering Basics

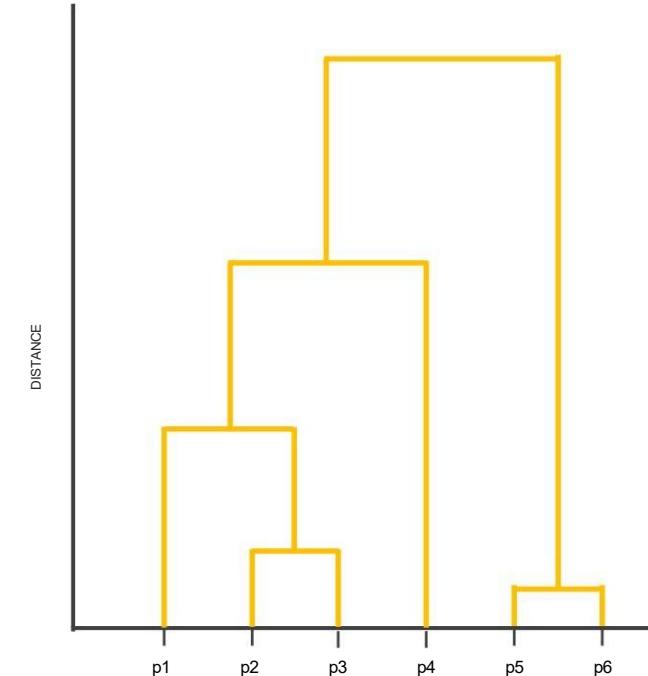
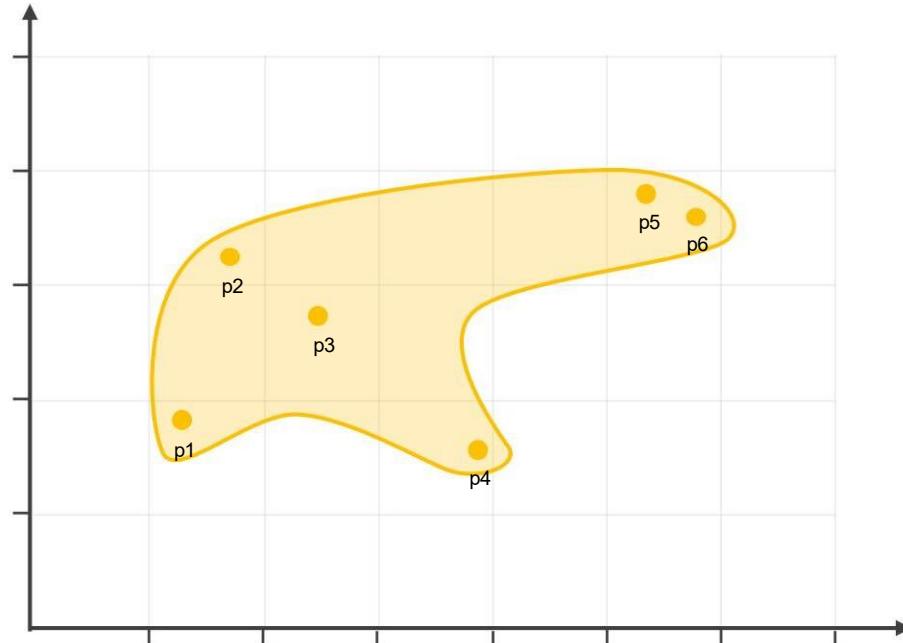
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Répétez le processus jusqu'à ce que tous les points fassent partie du même groupe.





DENDROGRAMMES EN PYTHON

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

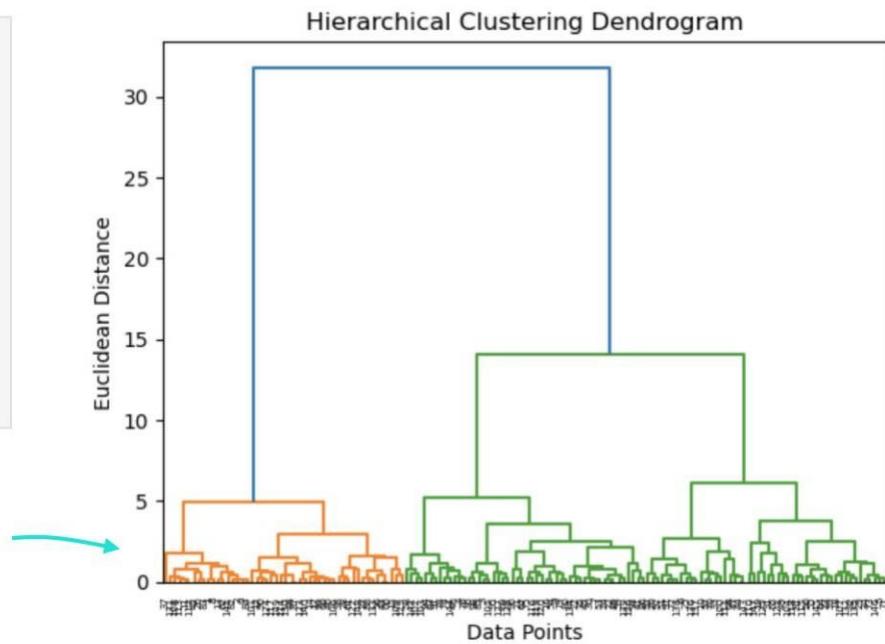
La fonction `dendrogram()` de la bibliothèque `scipy` en Python permet de visualiser les clusters de classification hiérarchique.

```
# visualize the clusters with a dendrogram
from scipy.cluster.hierarchy import linkage, dendrogram
import matplotlib.pyplot as plt

linkage_matrix = linkage(data, method='ward')
dendrogram_info = dendrogram(linkage_matrix)

plt.title("Hierarchical Clustering Dendrogram")
plt.xlabel("Data Points")
plt.ylabel("Euclidean Distance");
```

Ces couleurs sont définies par défaut, mais vous pouvez définir l'argument `color_threshold` pour les mettre à jour pour un nombre spécifique de clusters.



Il semble y avoir 3 groupes dans le dendrogramme, donc mettons à jour le seuil de couleur à 3.



CLUSTERING AGLOMÉRANT EN PYTHON

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

```
from sklearn.cluster import AgglomerativeClustering  
  
agg = AgglomerativeClustering(n_clusters=2, metric='euclidean', linkage='ward')
```

Le nombre de
clusters à identifier
(la valeur par défaut est 2)

La méthode utilisée pour
mesurer la distance entre les points
(par défaut : « euclidienne »)

La méthode utilisée pour mesurer
la distance entre les groupes
(par défaut : « ward »)

Autres distances :

- « Manhattan »
- « cosinus »
- « précalculé »

Autres méthodes :

- "célibataire"
- "complet"
- "moyenne"



CONSEIL DE PRO : Bien que vous
puissiez ajuster ces paramètres, les
valeurs par défaut sont de loin les plus courantes.



CLUSTERING AGLOMÉRANT EN PYTHON

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

La fonction AgglomerativeClustering() du module cluster de scikit-learn permet d'effectuer un clustering hiérarchique en Python.

```
# import agglomerative clustering from sklearn
from sklearn.cluster import AgglomerativeClustering
```

```
# fit an agglomerative clustering model with 3 clusters
agg = AgglomerativeClustering(3)
agg.fit(data)
```

▼ AgglomerativeClustering

AgglomerativeClustering(n_clusters=3)

Une fois le modèle ajusté, vous pouvez visualiser les affectations de cluster à l'aide de l' attribut .labels_

```
# view the cluster assignments
agg.labels_
```

```
array([1, 1, 1, 2, 2, 0, 0, 2, 1, 2, 0, 1, 1, 1, 1, 1, 1, 2, 1, 0, 0, 0, 2,
       2, 0, 2, 0, 1, 1, 1, 0, 2, 0, 0, 0, 0, 0, 2, 2, 1, 1, 0, 2, 2, 1, 1, 1, 2, 0, 2, 1, 1,
       0, 2, 1, 0, 1, 0, 0, 0, 0, 2, 2, 0, 1, 0, 2, 2, 1, 1, 1, 1, 2, 0, 0, 0, 2, 2, 1, 1, 1,
       1, 2, 0, 2, 0, 0, 2, 0, 2, 2, 2, 0, 2, 2, 2, 0, 1, 1, 2, 2, 0, 2, 1, 1, 1, 2, 2, 1, 1,
       1, 2, 1, 2, 0, 2, 2, 2, 2, 0, 2, 1, 0, 1, 2, 2, 0, 2, 1, 0, 1, 2, 2, 1, 1, 0, 1, 1, 1,
       0, 0, 1, 1, 1, 0, 2, 1, 0, 2, 2, 0, 1, 1, 1, 0, 2, 2, 2, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
       0, 1, 2, 0, 0, 0, 2, 1, 1, 2, 0, 2, 0, 0, 1, 1, 1, 0, 2, 2, 2, 1, 1, 1, 0, 1, 1, 1, 1])
```



CARTES DE GROUPEMENT EN PYTHON

Clustering Basics

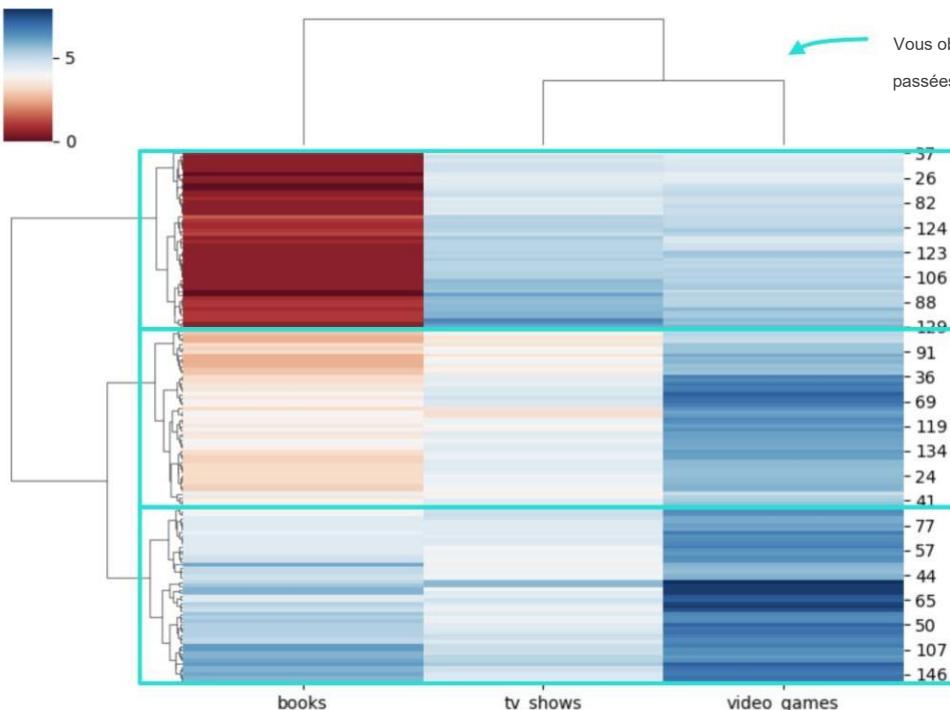
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

```
# create a cluster map
sns.clustermap(data, method='ward', cmap='RdBu', figsize=(8, 6), xticklabels=data.columns)
plt.show()
```



Vous obtenez également des liens entre les différentes fonctionnalités (les heures passées devant la télévision et les jeux vidéo sont plus étroitement liées !).

Ces élèves ne lisent pas beaucoup de livres

Ces étudiants consomment une quantité moyenne de chaque type de divertissement.

Ces élèves jouent beaucoup aux jeux vidéo



Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

TRAVAUX PRATIQUES EN PYTHON



DBSCAN

Clustering Basics

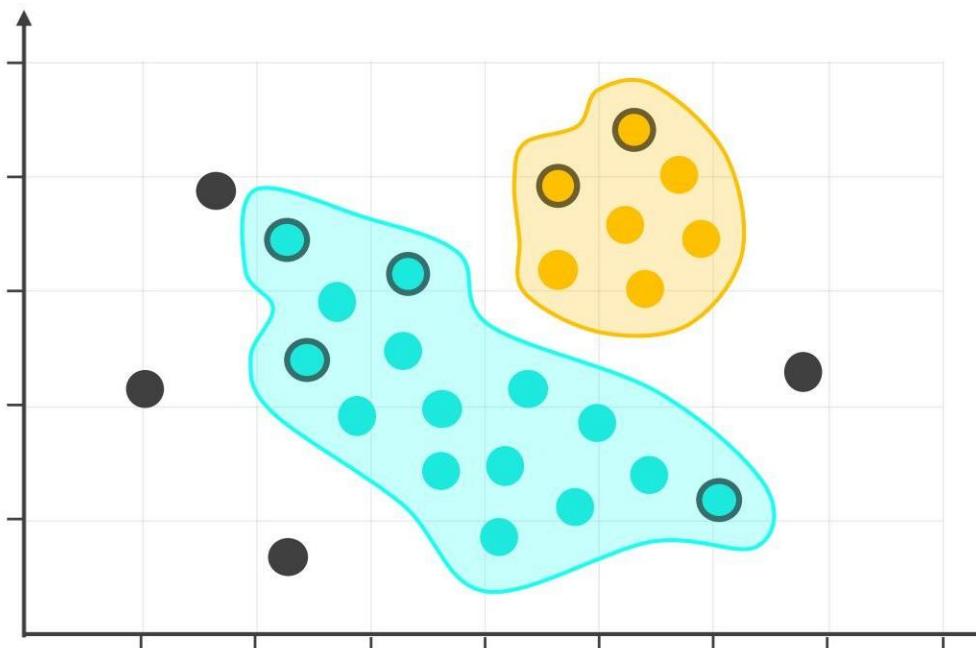
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est une technique de clustering qui identifie des clusters en fonction de la densité des points de données.





DBSCAN

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est une technique de clustering qui identifie des clusters en fonction de la densité des points de données.

Voici comment ça fonctionne* :

1. Sélectionnez un rayon (eps) et un nombre minimal de points (min_samples)
2. Sur un nuage de points, étiquetez chaque point comme suit :
 - Point central – possède le nombre minimal de points dans son rayon (dans une région dense)
 - Point frontière – ne possède pas le nombre minimal de points dans son rayon, mais possède au moins un point central dans son rayon (à la périphérie des groupes).
 - Point de bruit (valeur aberrante) – ne possède pas de point central dans son rayon (points isolés)

DBSCAN permet de traiter des groupes de formes irrégulières et d'identifier les valeurs aberrantes.

*Ceci est un résumé sommaire des étapes ;des étapes plus détaillées sont incluses à la fin de la section DBSCAN



DBSCAN

Clustering Basics

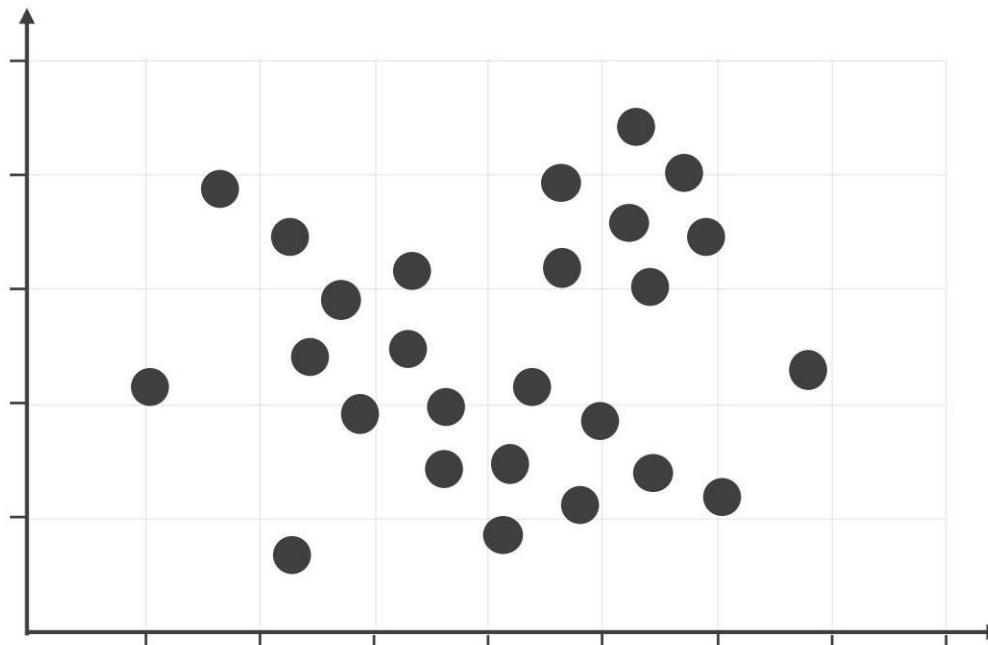
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 1 : Sélectionnez un rayon (eps) et un nombre minimal de points (min_samples)



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

K-Means
Clustering

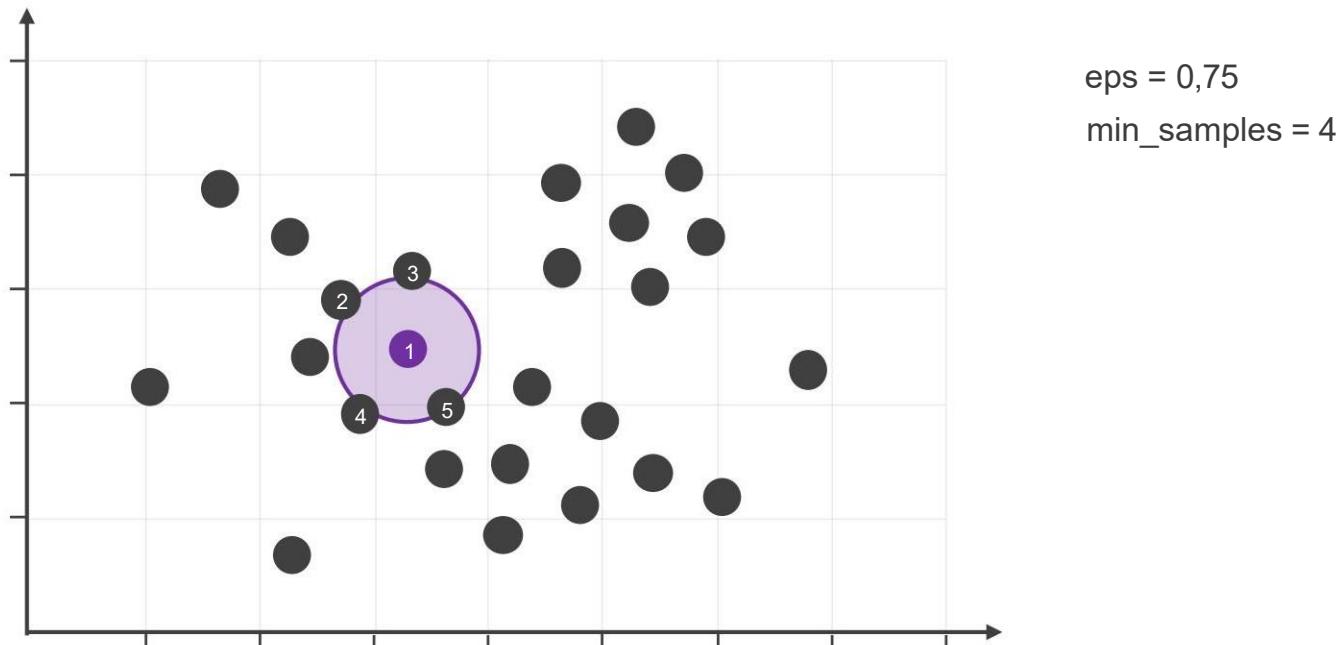
Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 2 : Sélectionnez un point au hasard et comptez les points situés dans son rayon.

- Si la valeur est supérieure ou égale à « `min_samples` », alors créez un cluster, désignez le point comme point central et marquez les points situés dans son rayon comme voisins.





DBSCAN

Clustering Basics

K-Means
Clustering

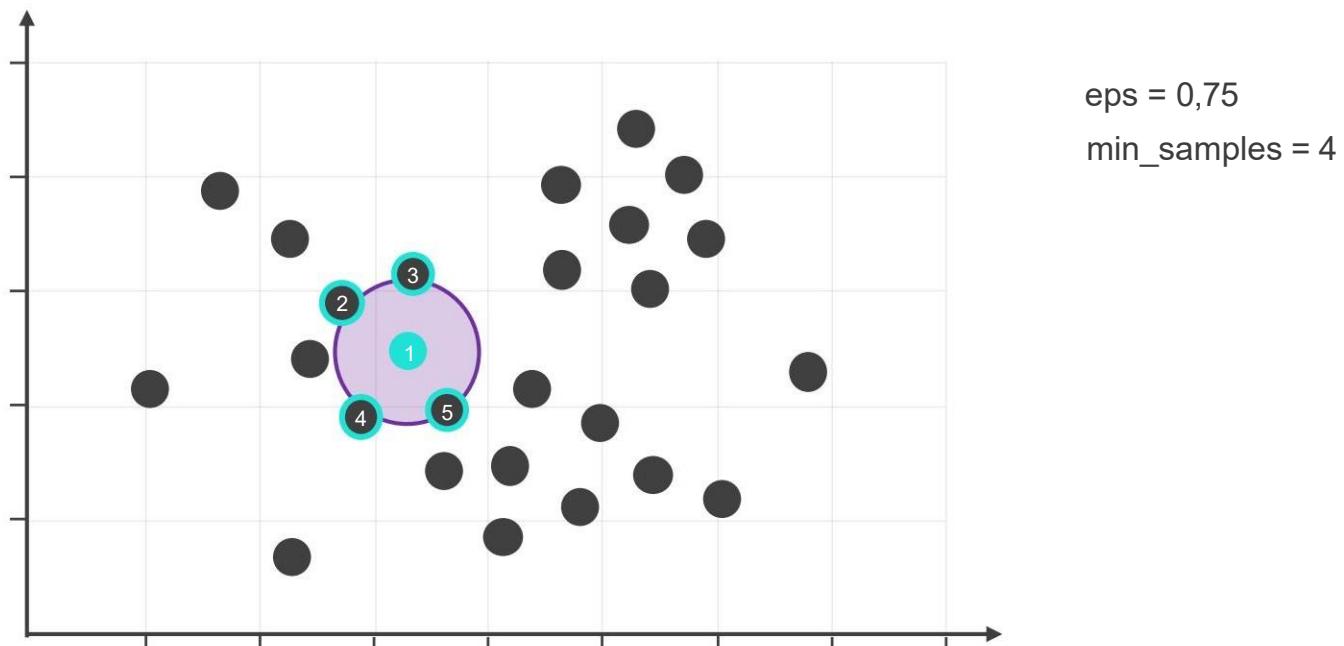
Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 2 : Sélectionnez un point au hasard et comptez les points situés dans son rayon.

- Si la valeur est supérieure ou égale à « min_samples », alors créez un cluster, désignez le point comme point central et marquez les points situés dans son rayon comme voisins.





DBSCAN

Clustering Basics

K-Means
Clustering

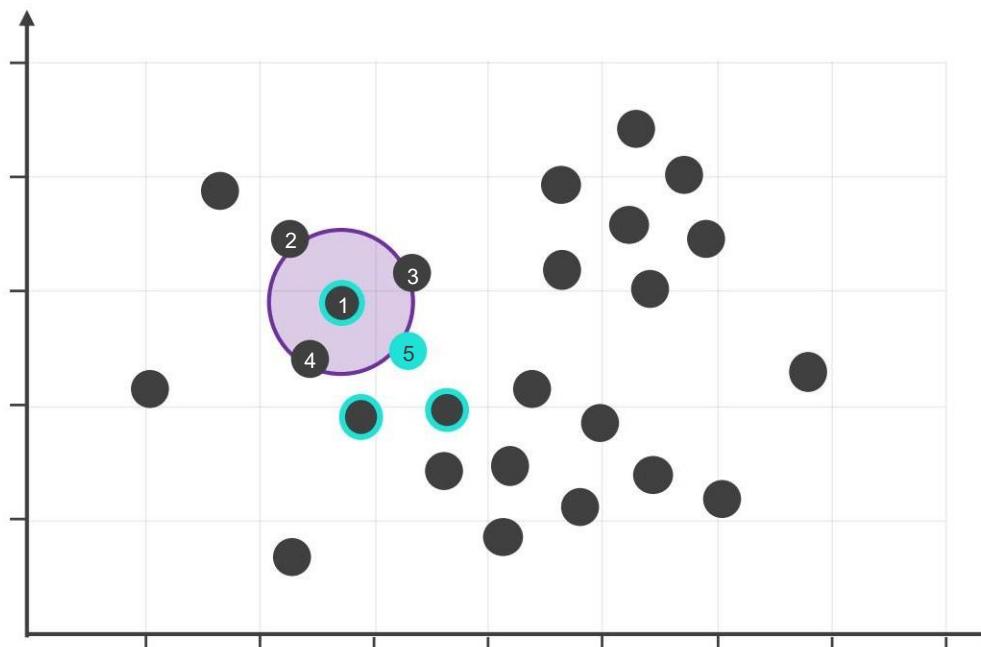
Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Déplacez-vous vers un voisin et comptez les points situés dans son rayon d'action.

- Si sa valeur est supérieure ou égale à « min_samples », étiquetez-la comme point central et marquez ses voisins.



eps = 0,75

min_samples = 4



DBSCAN

Clustering Basics

K-Means
Clustering

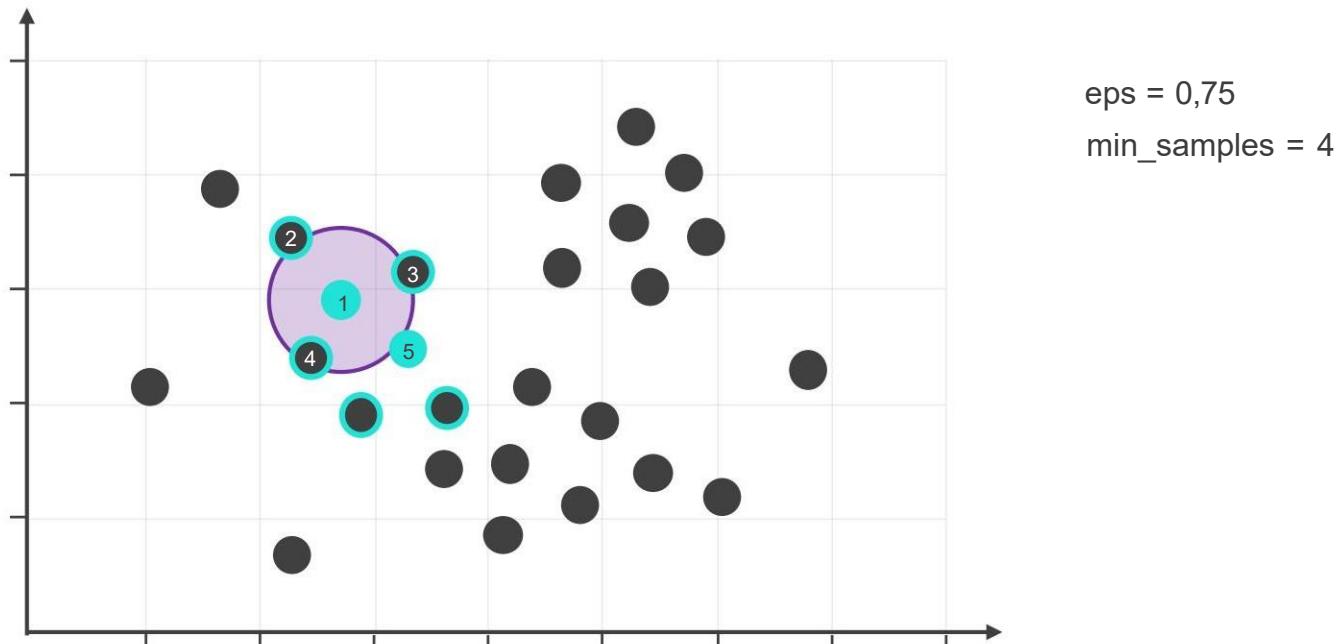
Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Déplacez-vous vers un voisin et comptez les points situés dans son rayon d'action.

- Si sa valeur est supérieure ou égale à « min_samples », étiquetez-la comme point central et marquez ses voisins.





DBSCAN

Clustering Basics

K-Means
Clustering

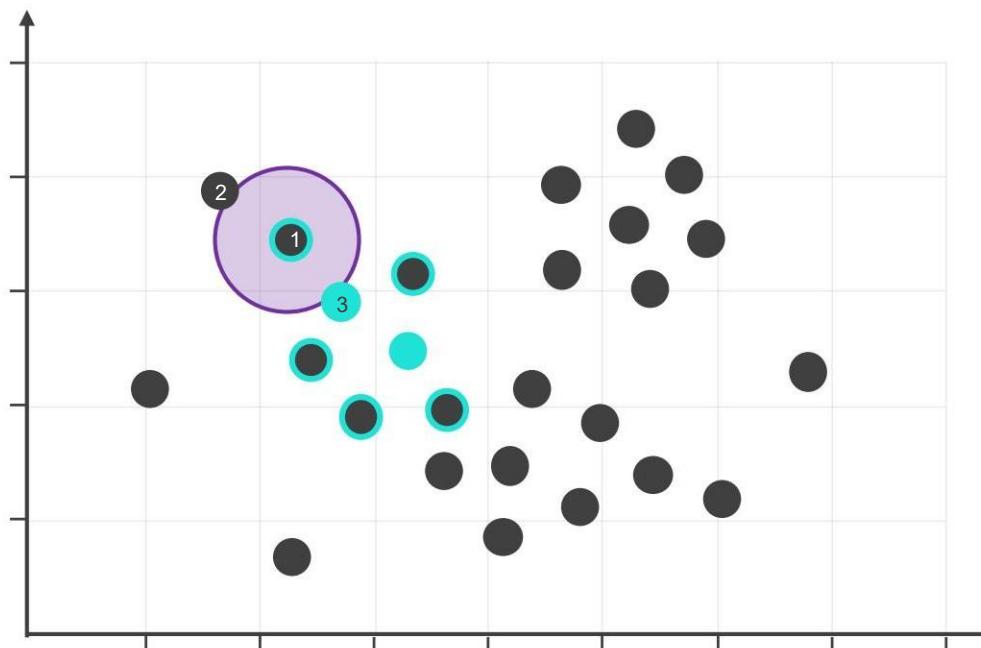
Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Déplacez-vous vers un voisin et comptez les points situés dans son rayon d'action.

- Si le nombre d'échantillons est inférieur à « `min_samples` », mais qu'au moins un de ces échantillons est un point central, qualifiez-le de point frontière.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

K-Means
Clustering

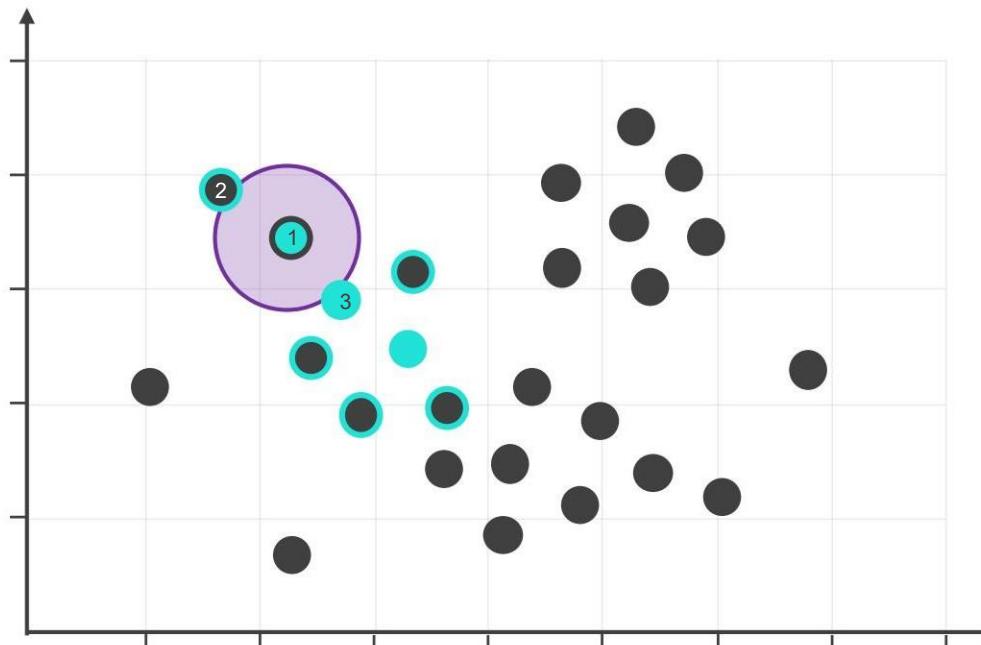
Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Déplacez-vous vers un voisin et comptez les points situés dans son rayon d'action.

- Si le nombre d'échantillons est inférieur à « `min_samples` », mais qu'au moins un de ces échantillons est un point central, qualifiez-le de point frontière.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

K-Means
Clustering

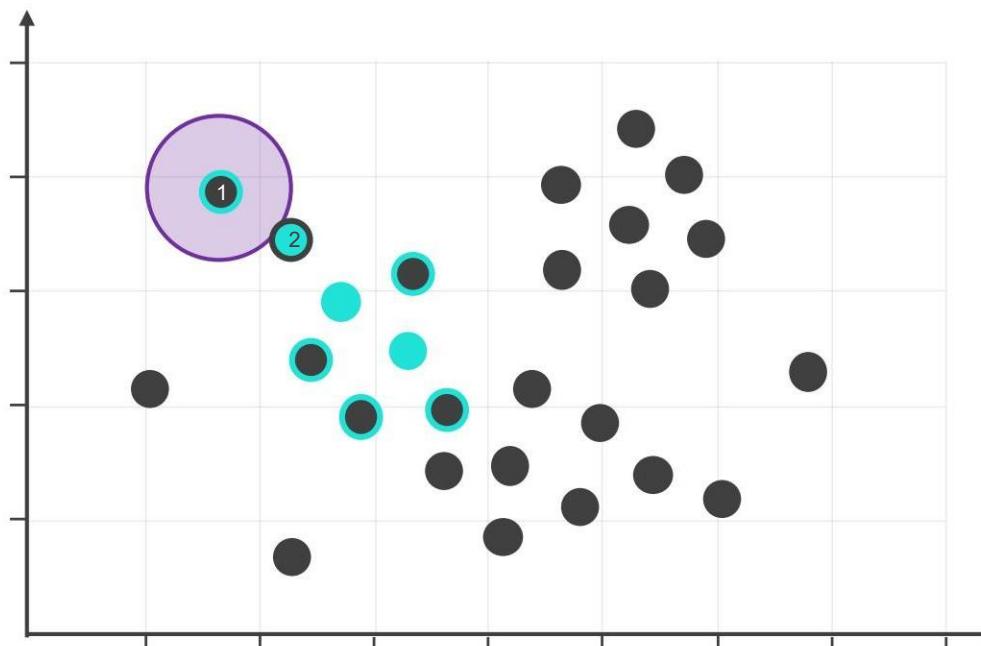
Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Déplacez-vous vers un voisin et comptez les points situés dans son rayon d'action.

- Si le nombre d'échantillons est inférieur à « `min_samples` » et qu'aucun d'entre eux n'est un point central, considérez-le comme un point de bruit.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

K-Means
Clustering

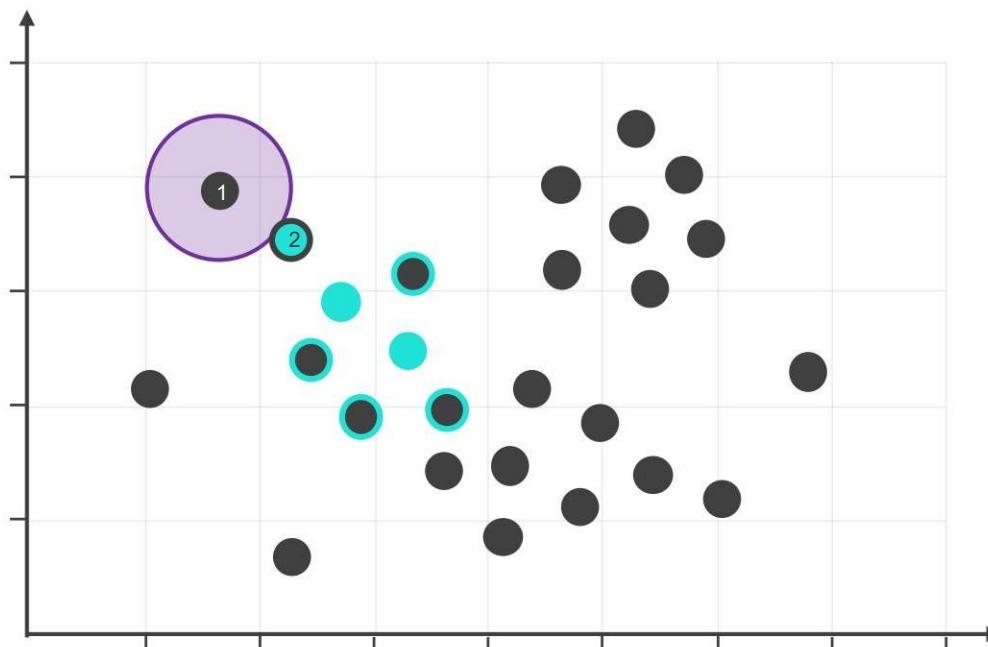
Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 3 : Déplacezvous vers un voisin et comptez les points situés dans son rayon d'action.

- Si le nombre d'échantillons est inférieur à « min_samples » et qu'aucun d'entre eux n'est un point central, considérezle comme un point de bruit.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

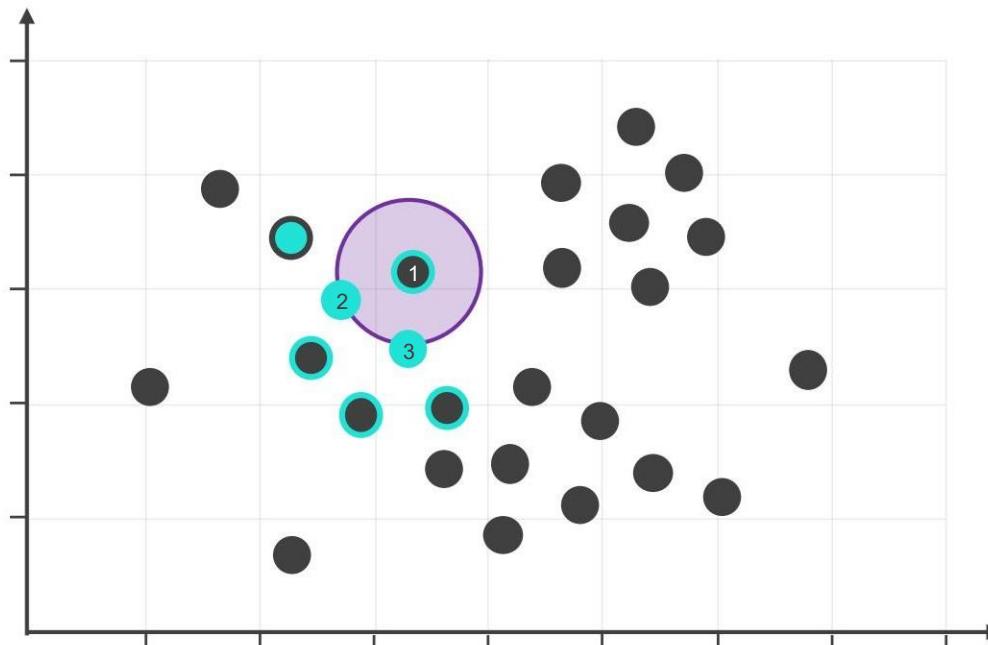
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

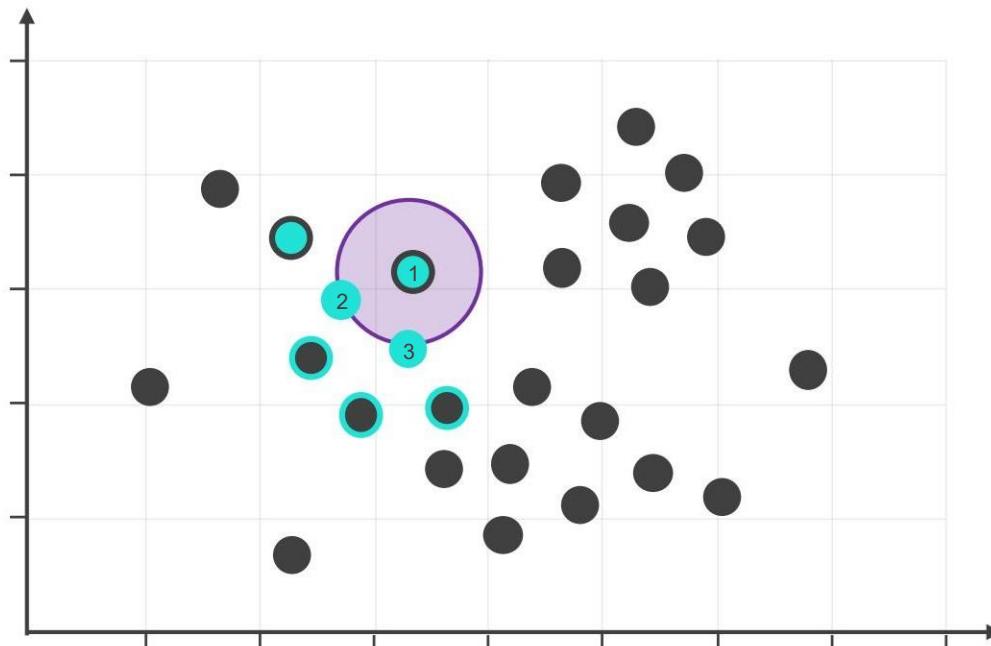
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

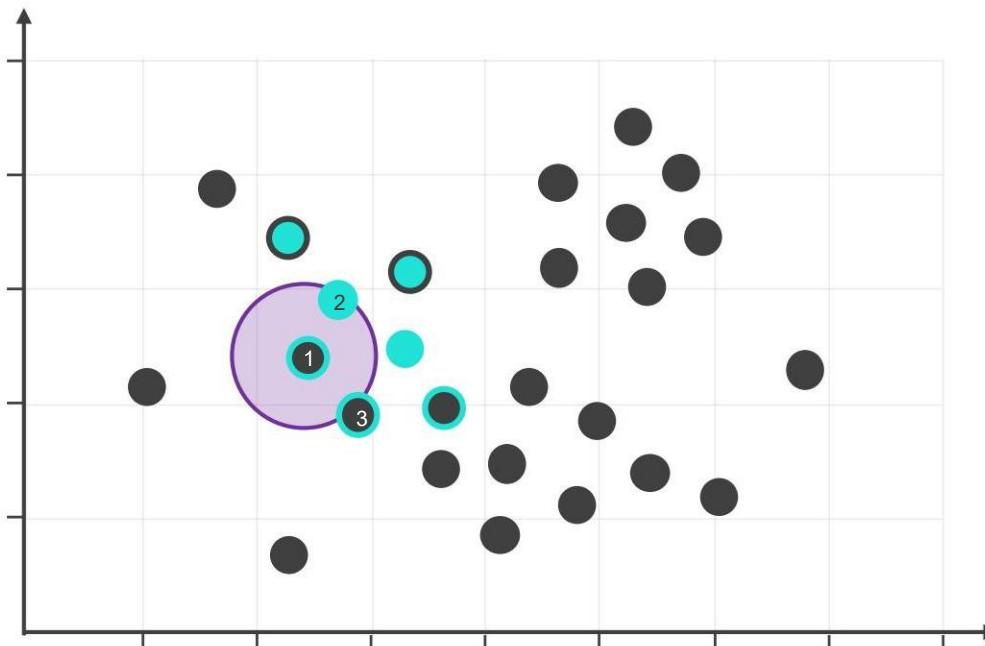
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

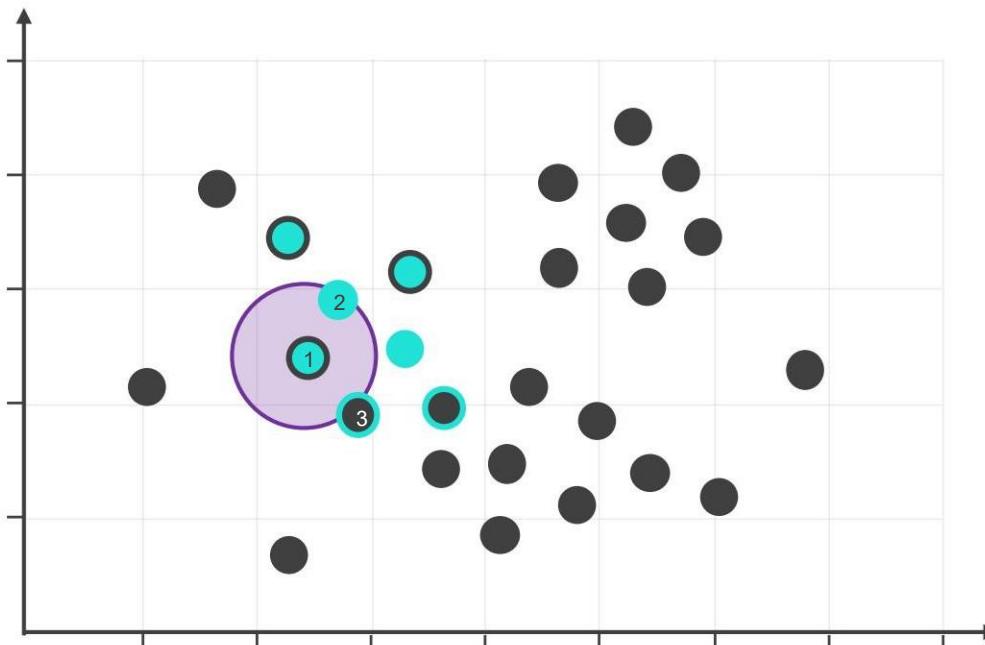
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

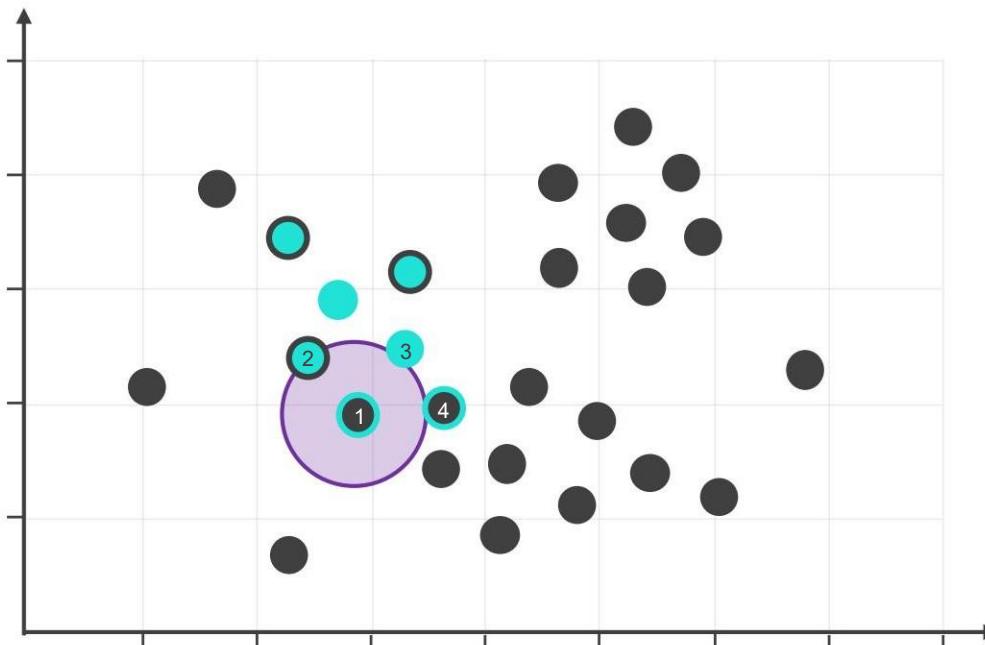
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

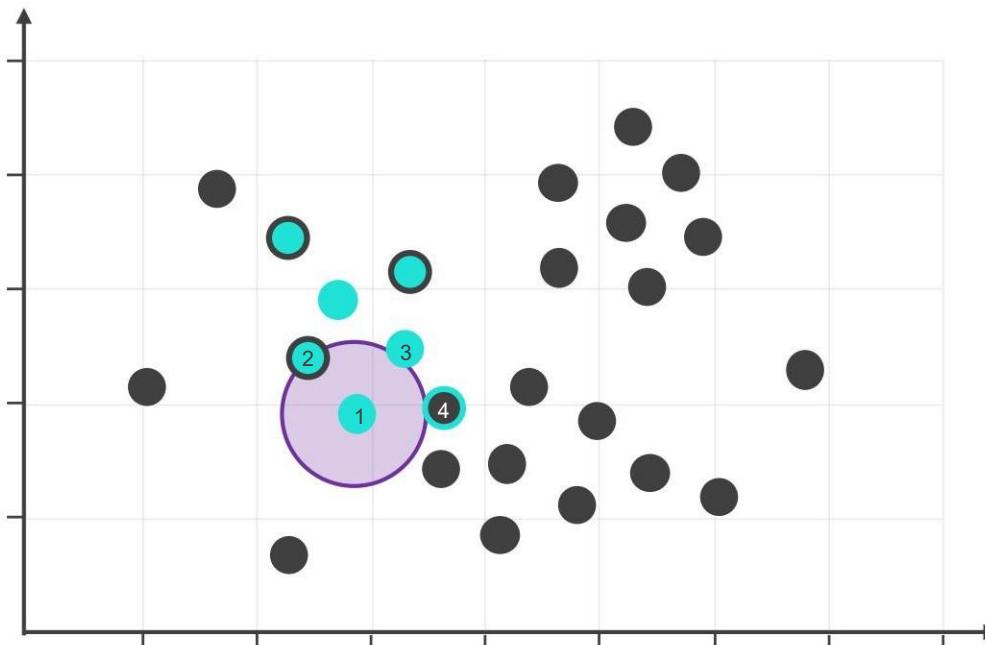
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

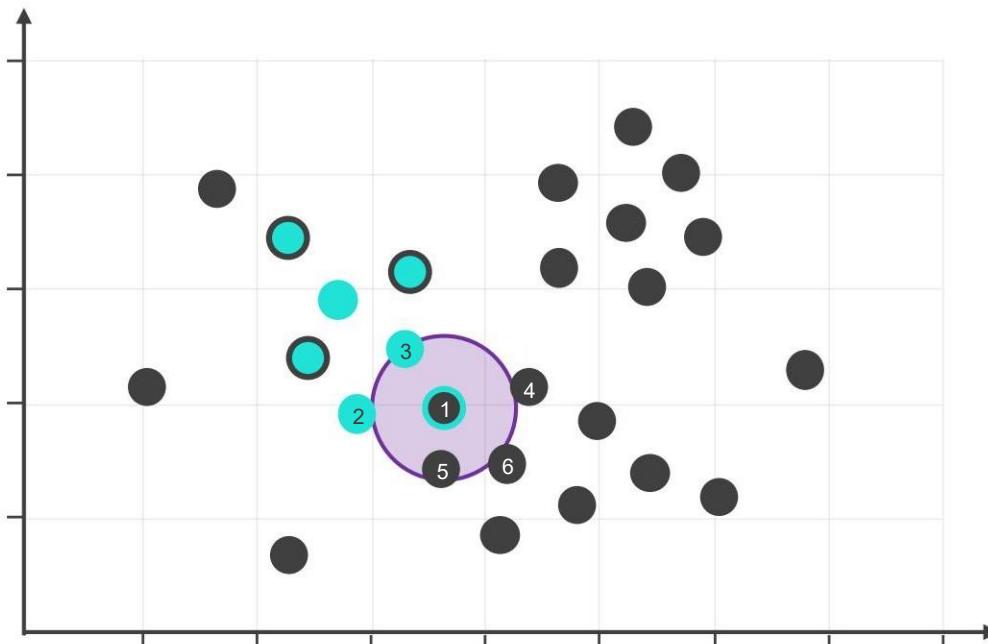
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

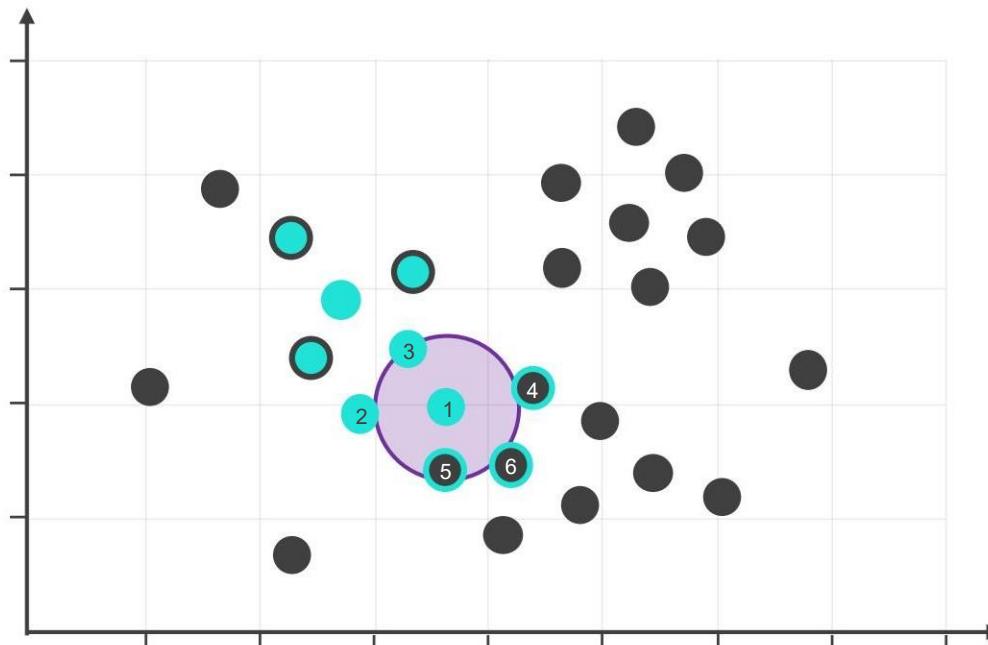
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 6



DBSCAN

Clustering Basics

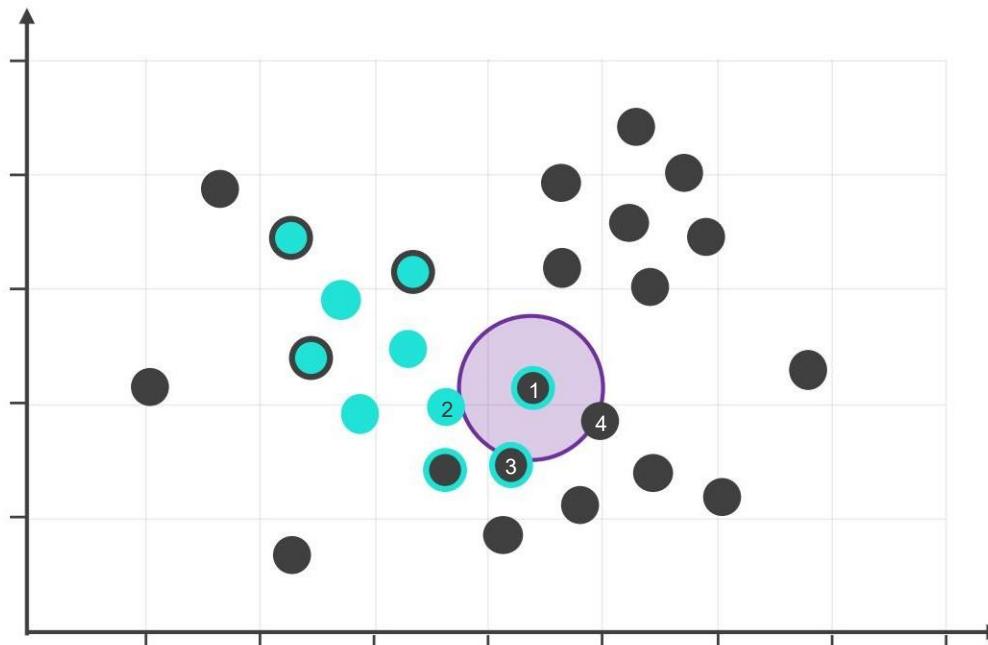
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

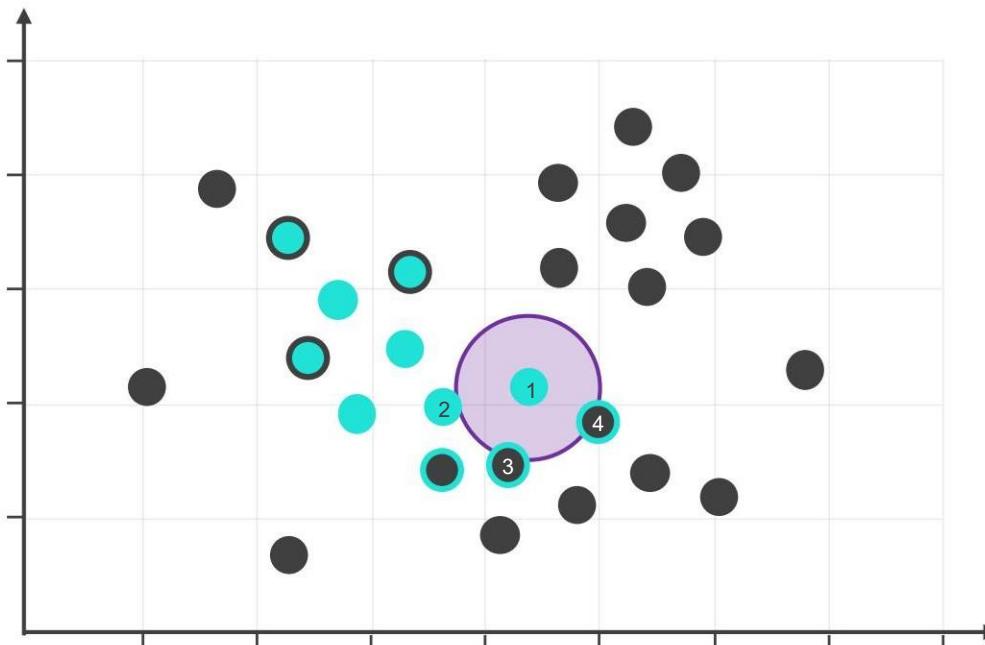
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

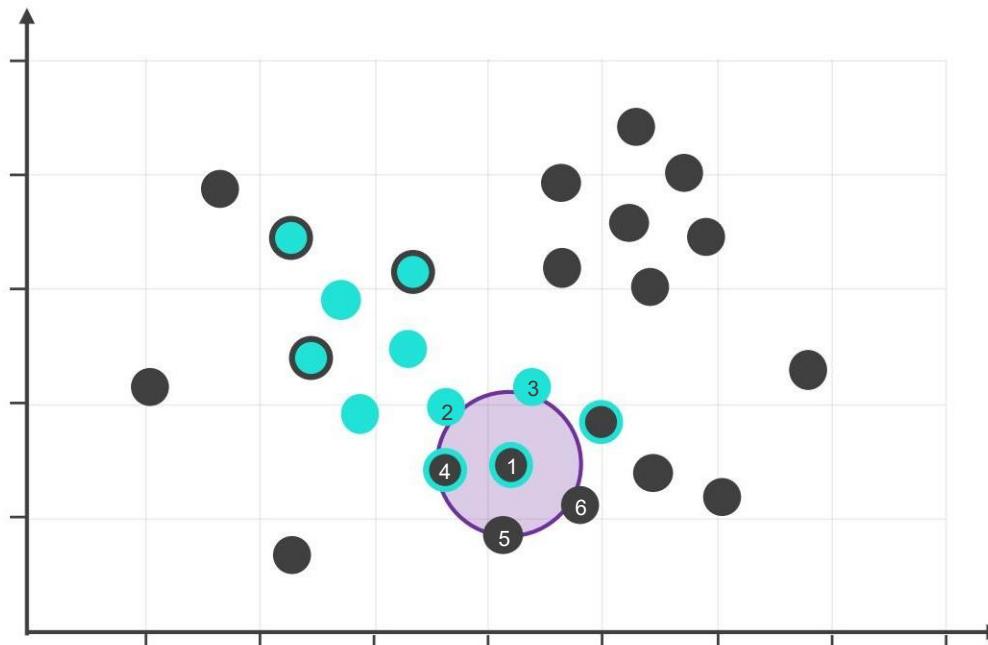
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

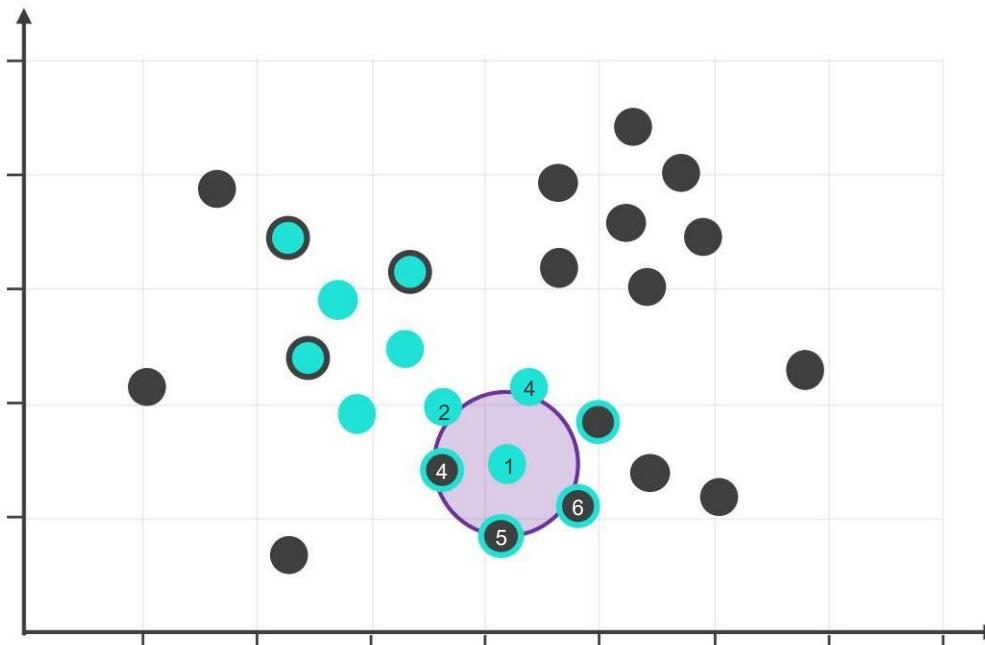
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

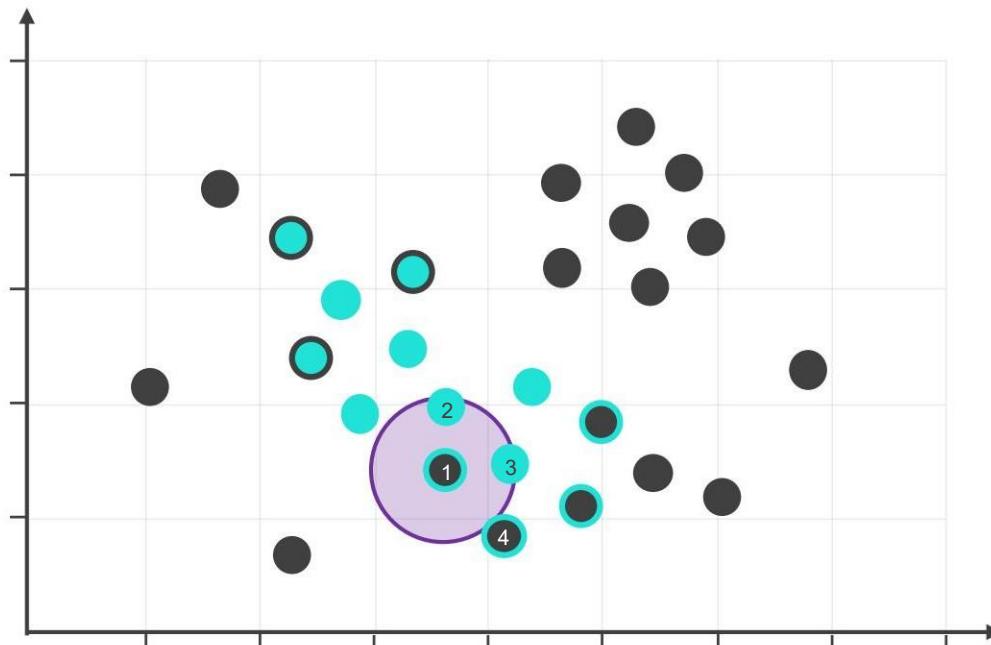
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

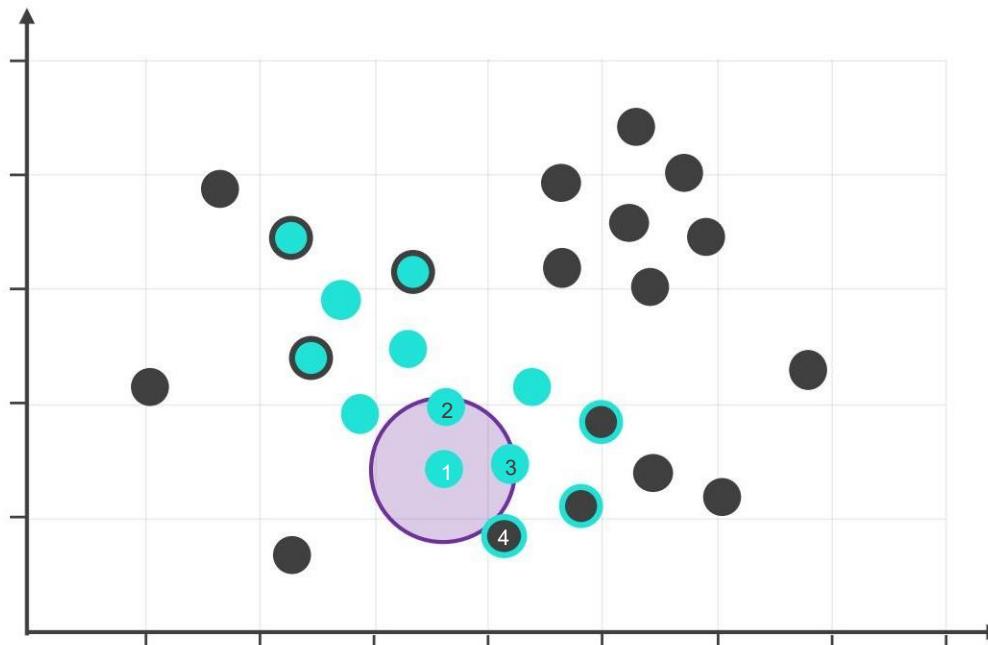
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

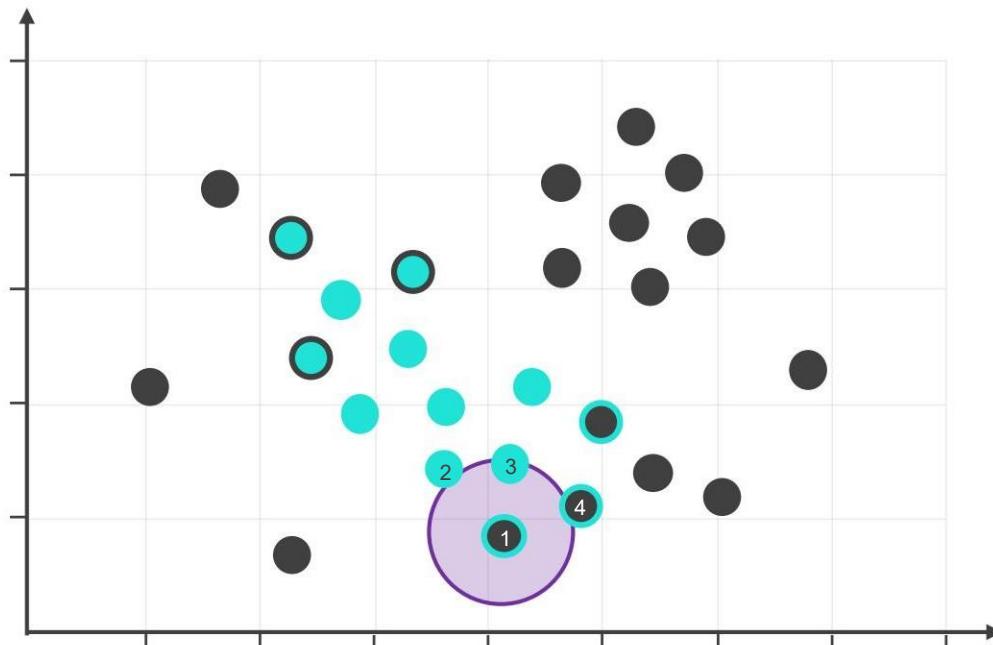
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

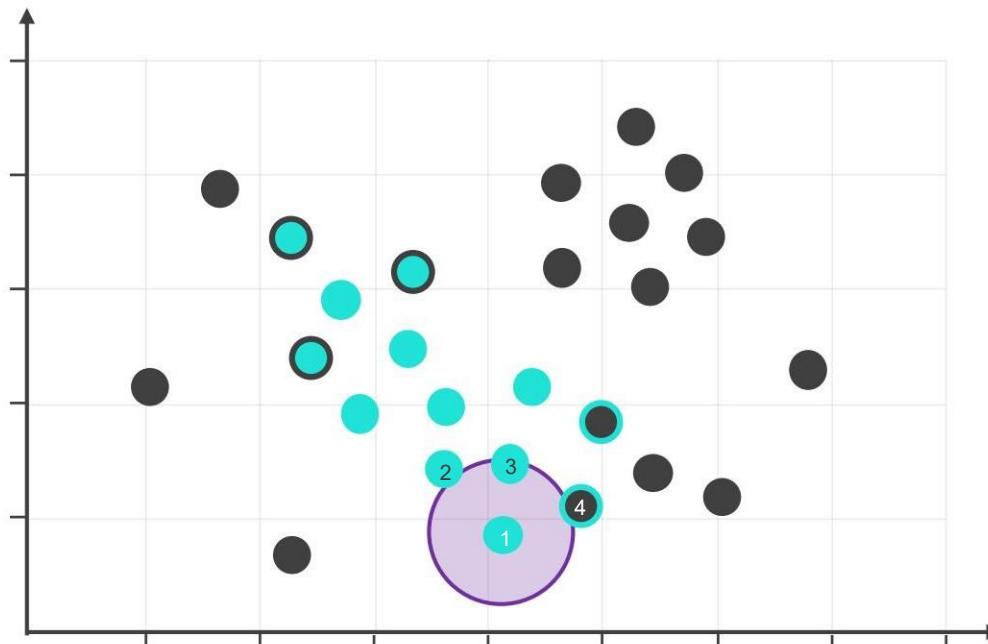
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

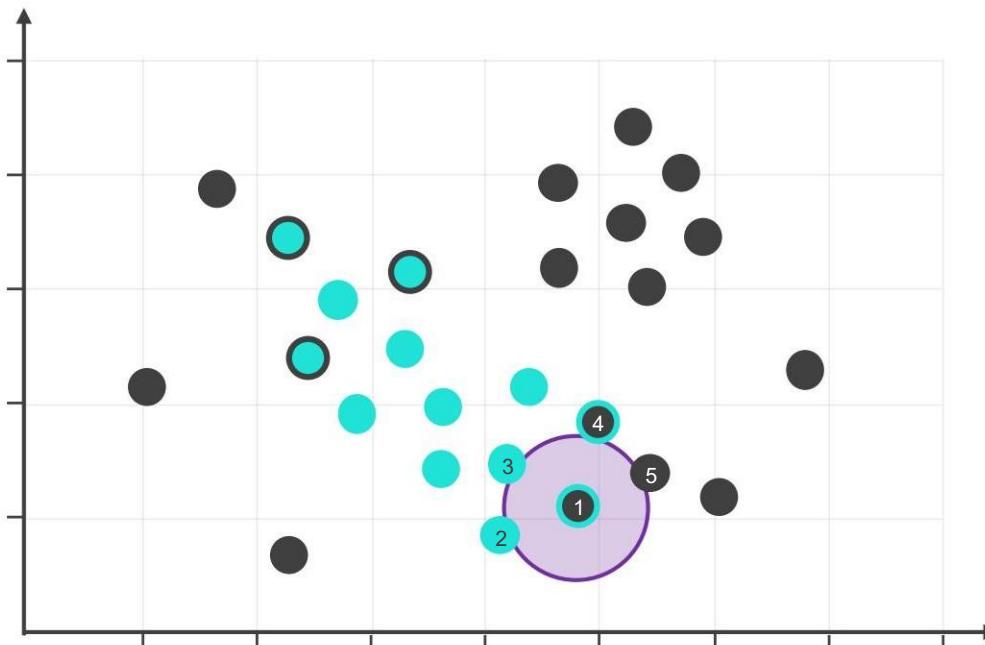
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

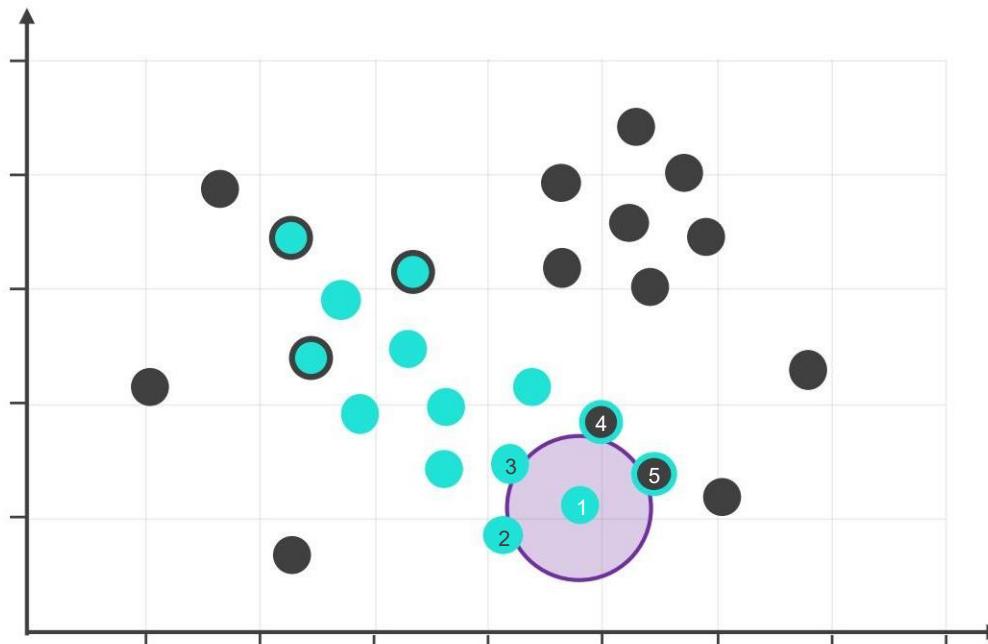
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

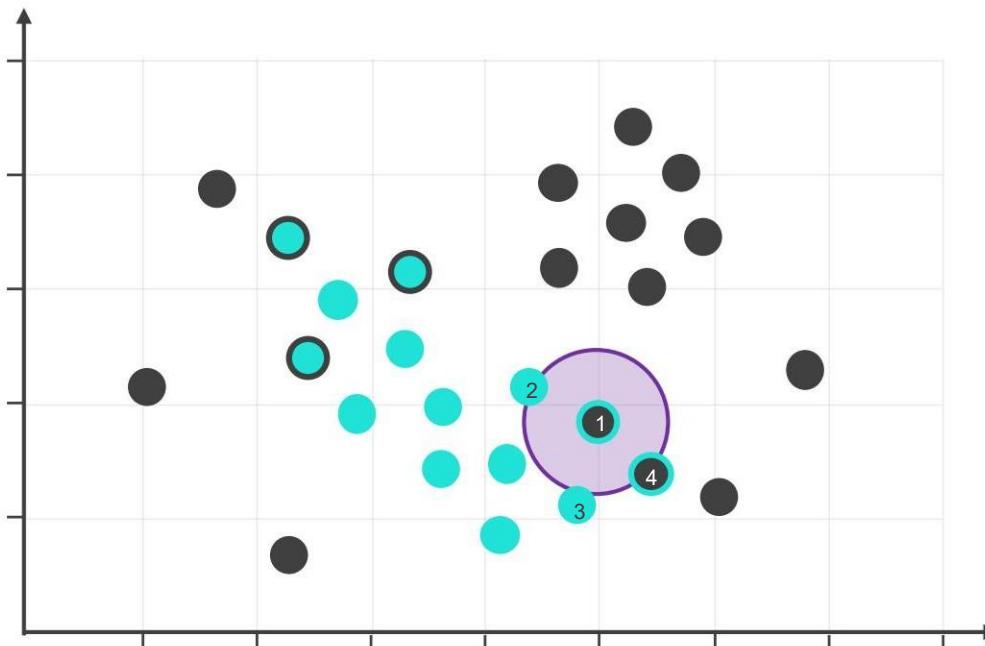
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

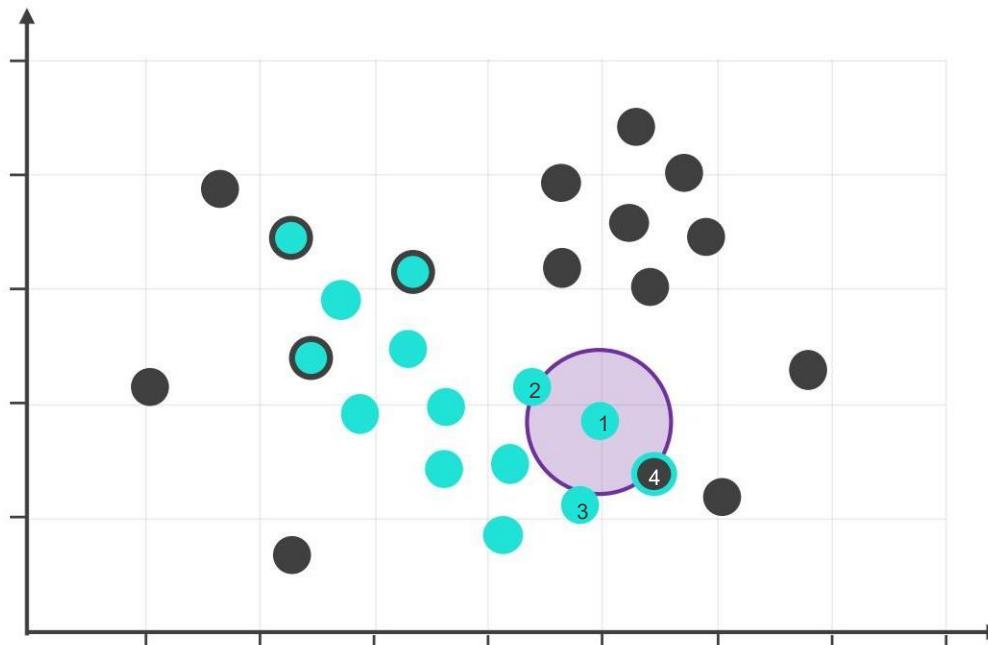
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

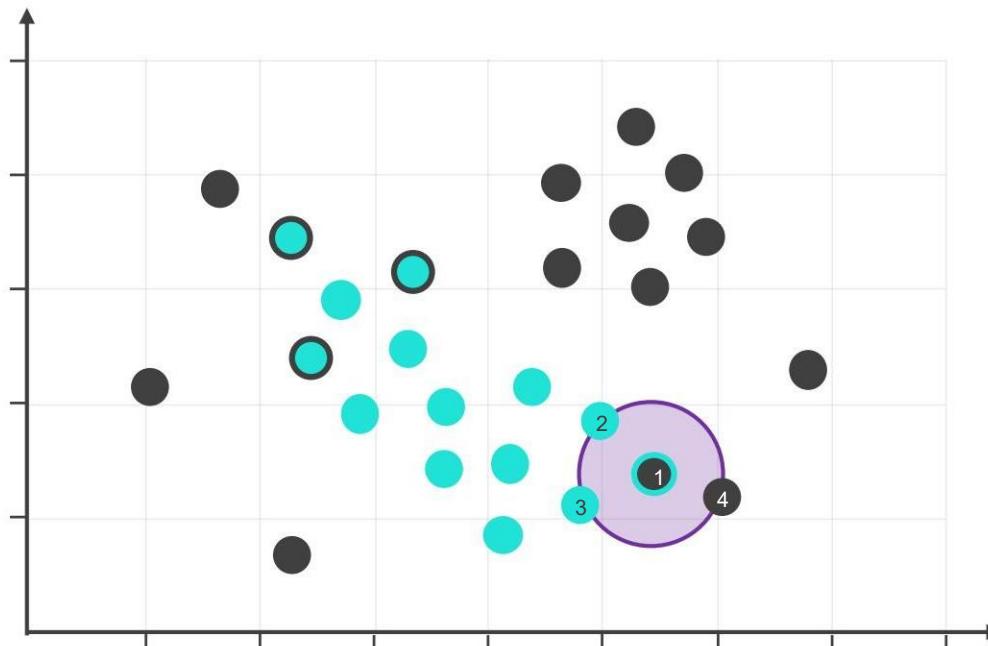
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

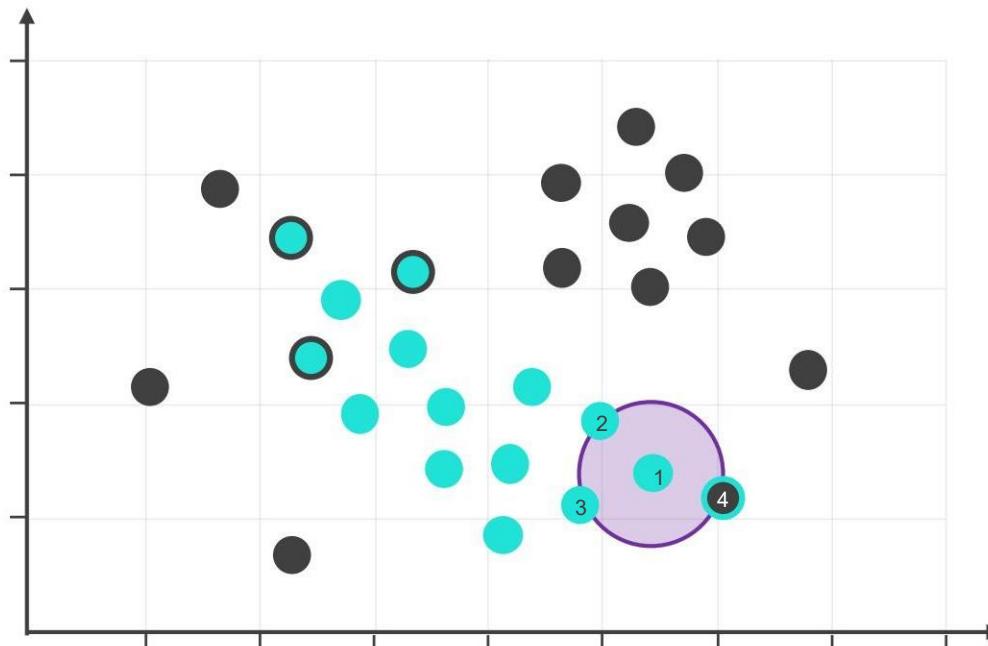
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

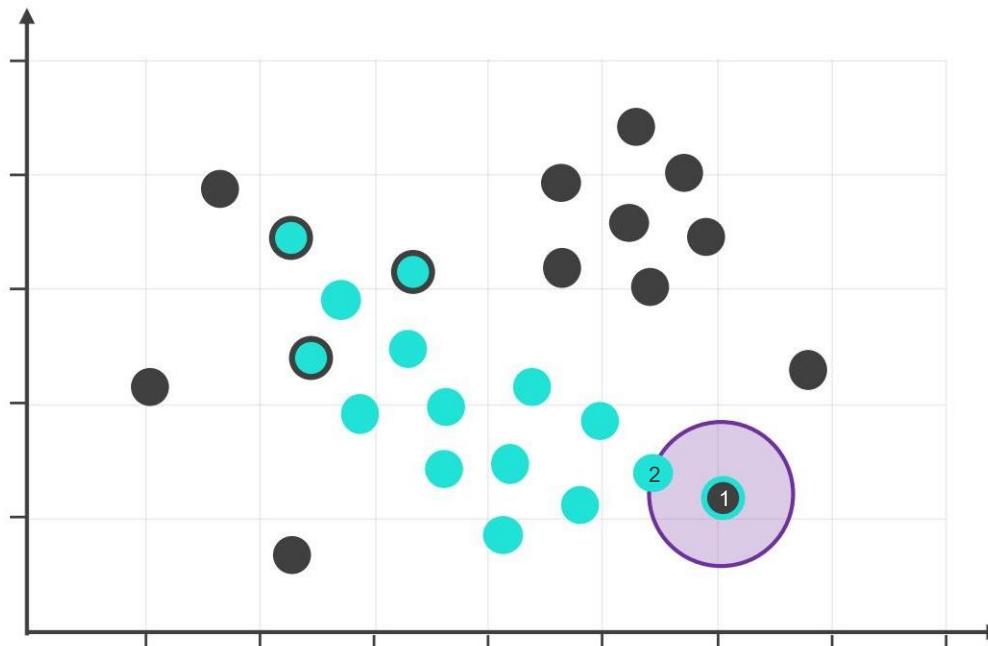
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

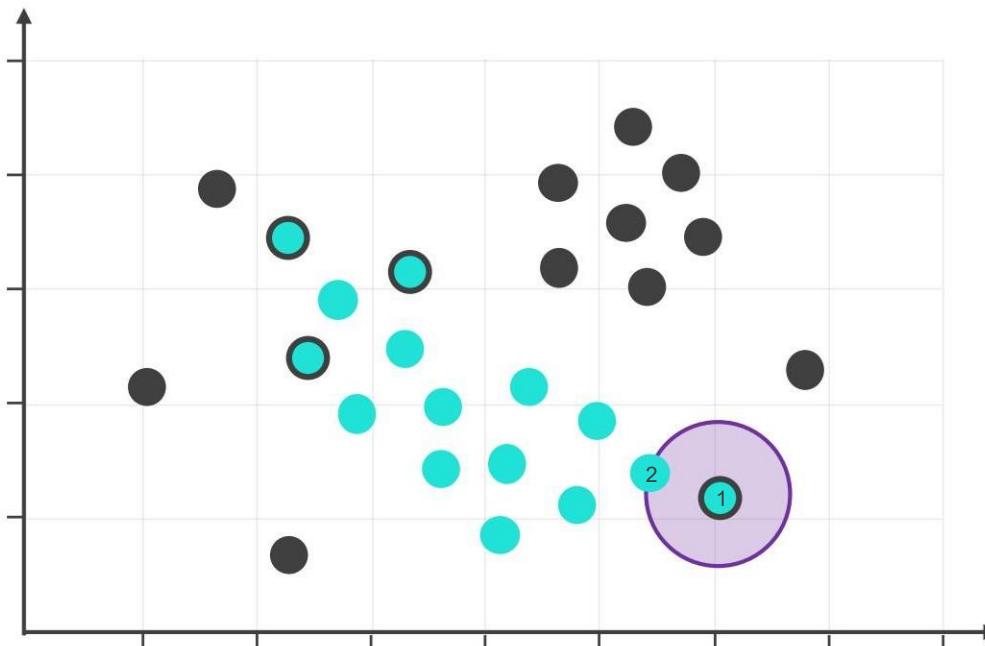
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 4 : Passez à un autre voisin et poursuivez ce processus jusqu'à ce que tous les points du groupe soient étiquetés.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

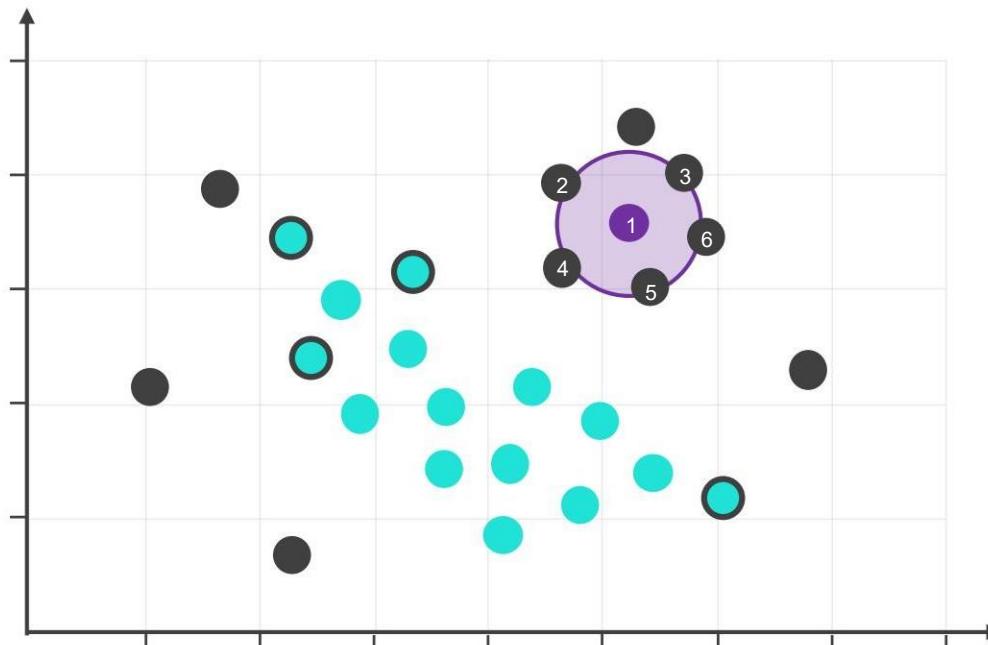
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

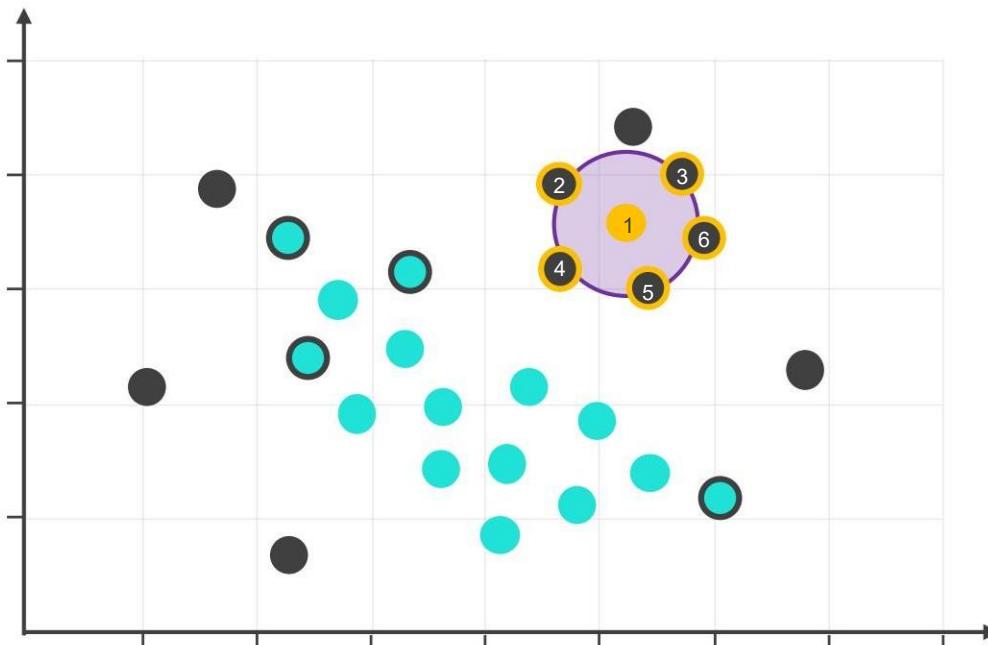
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

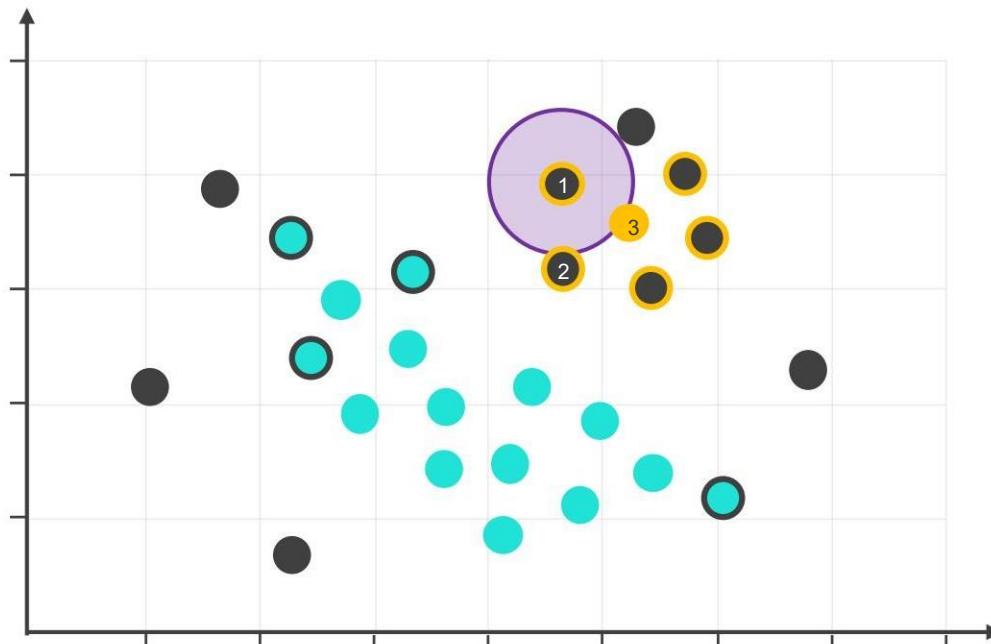
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

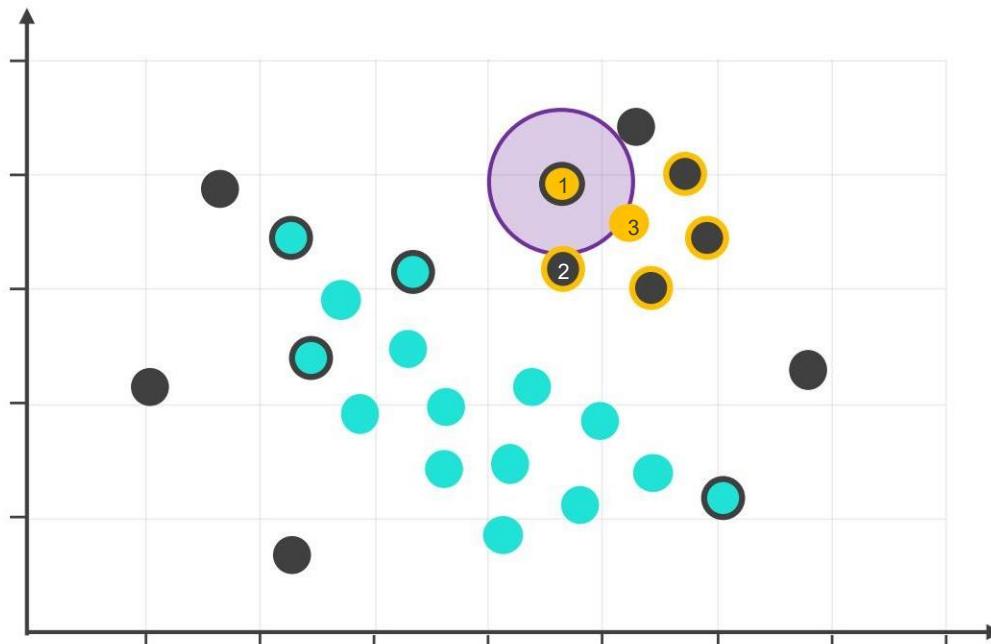
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

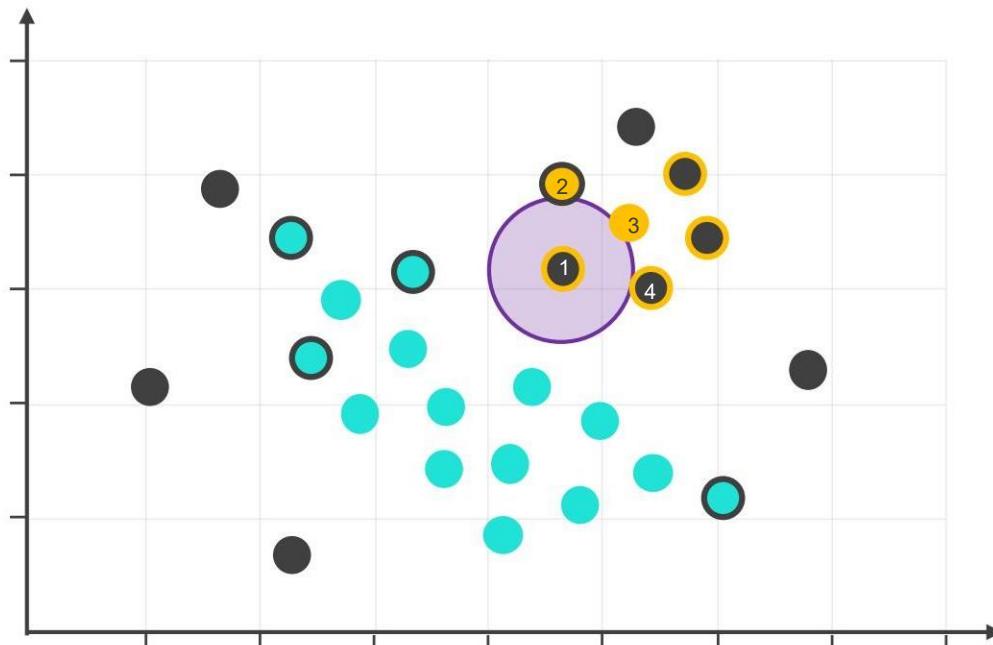
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

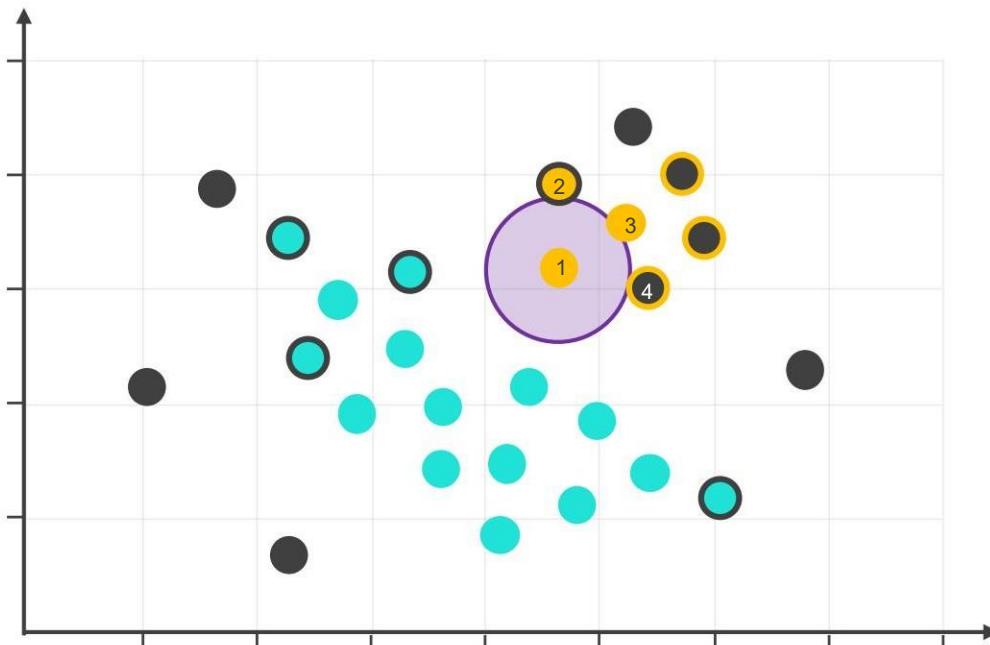
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

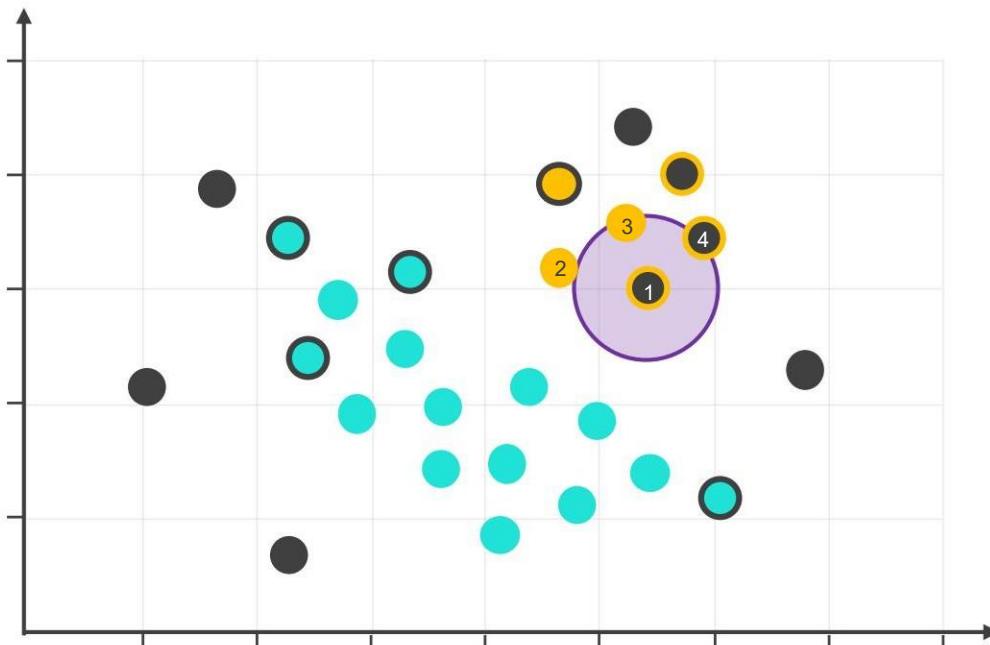
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

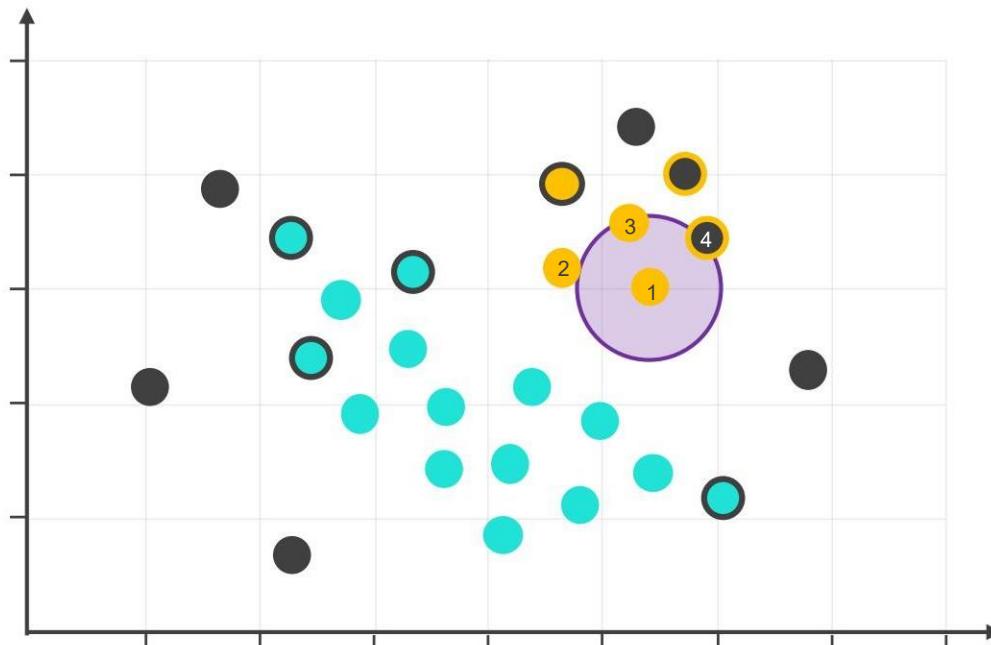
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

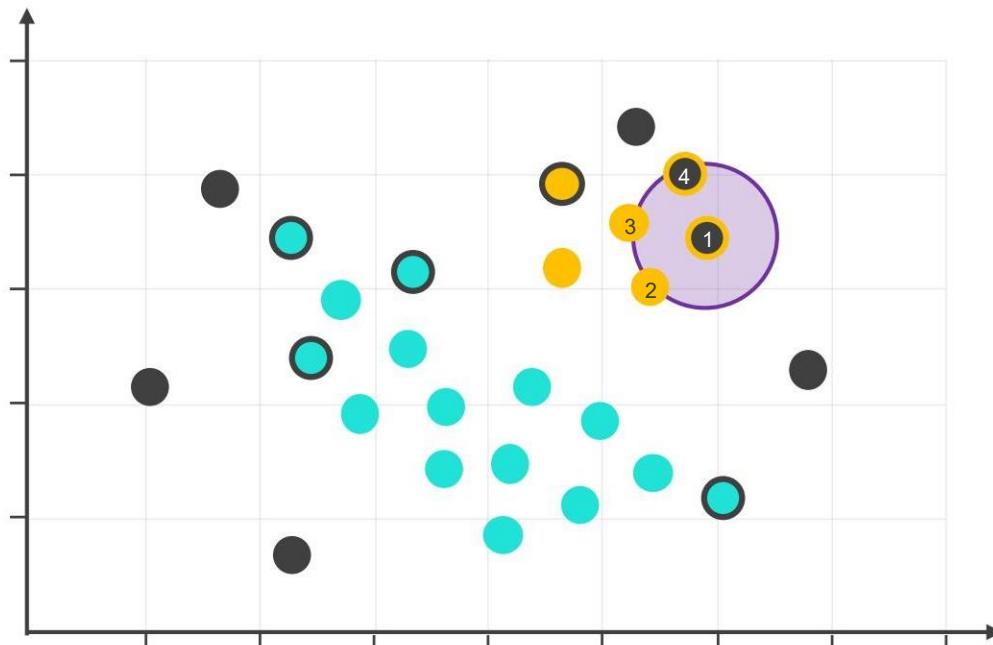
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

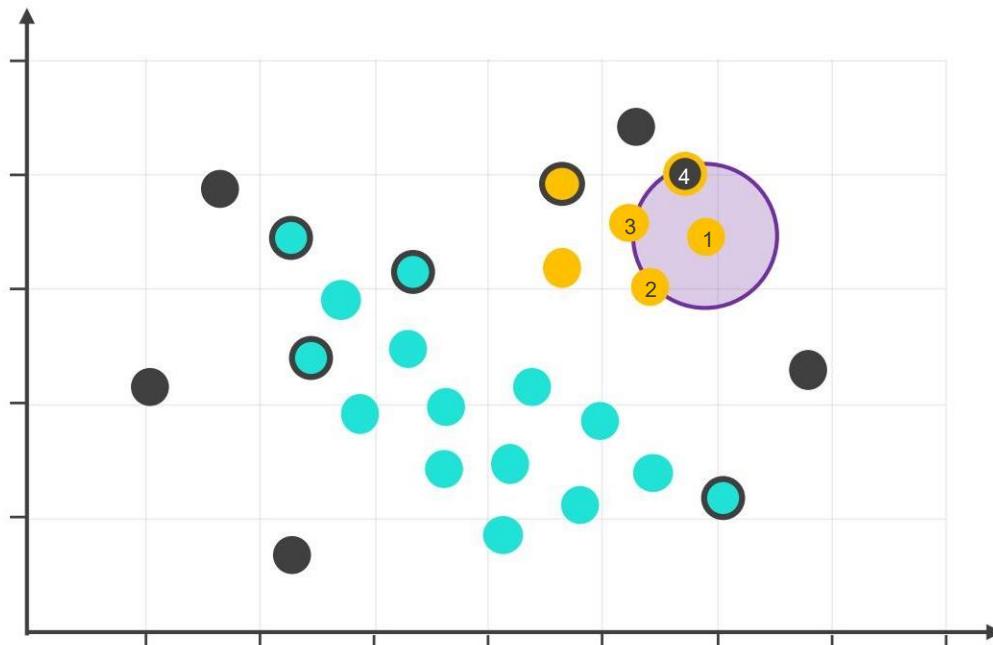
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

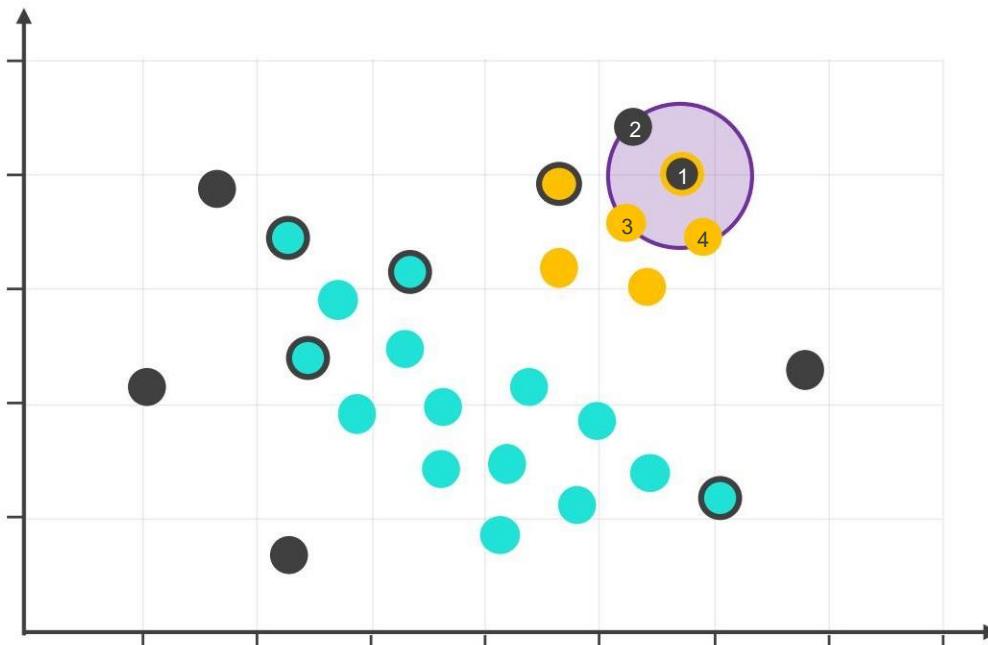
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

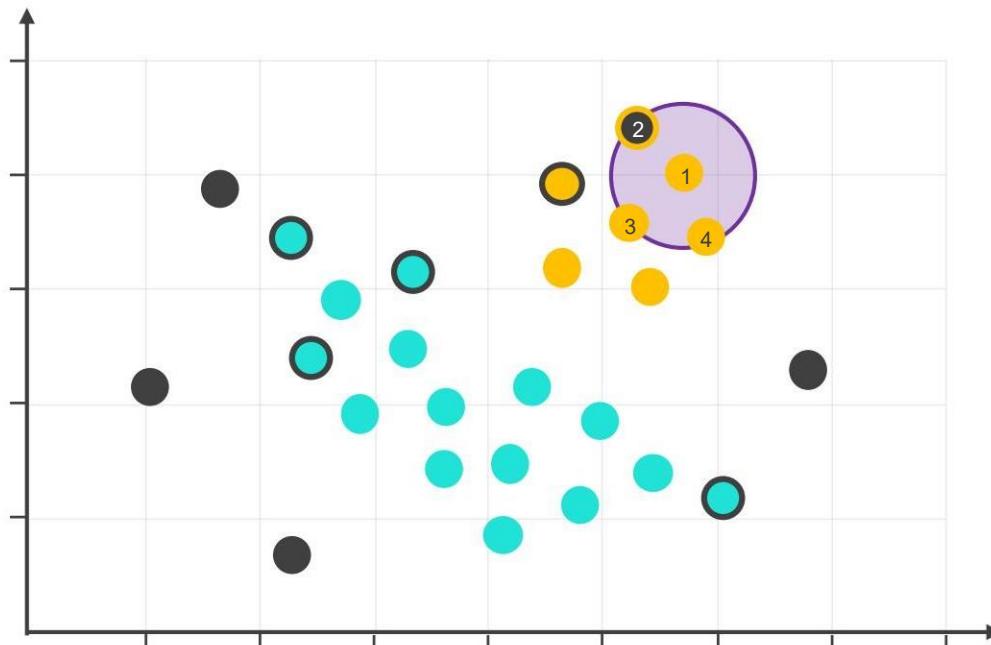
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

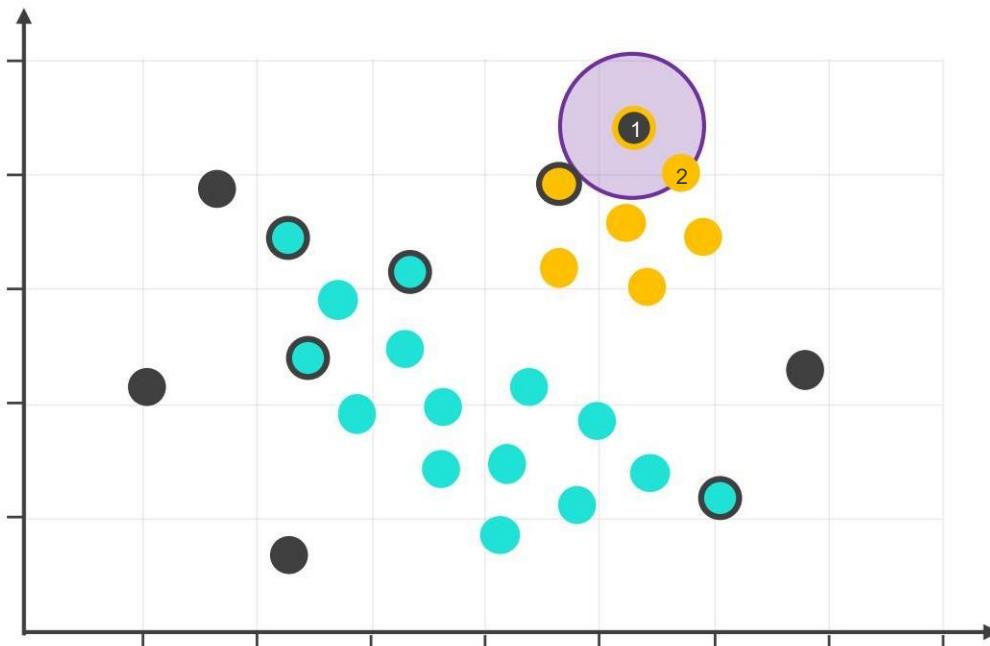
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

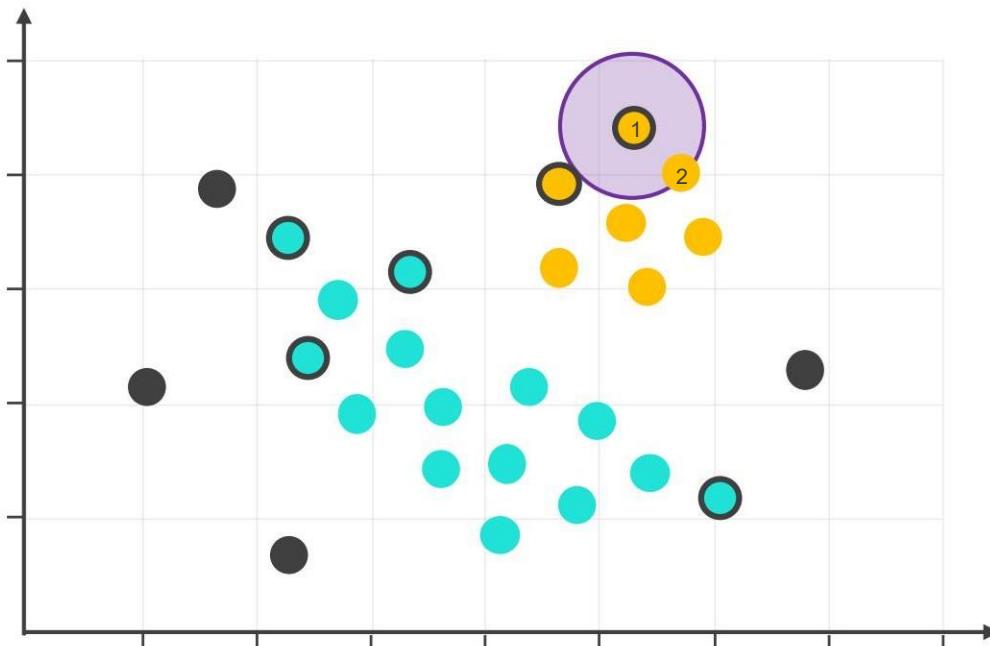
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

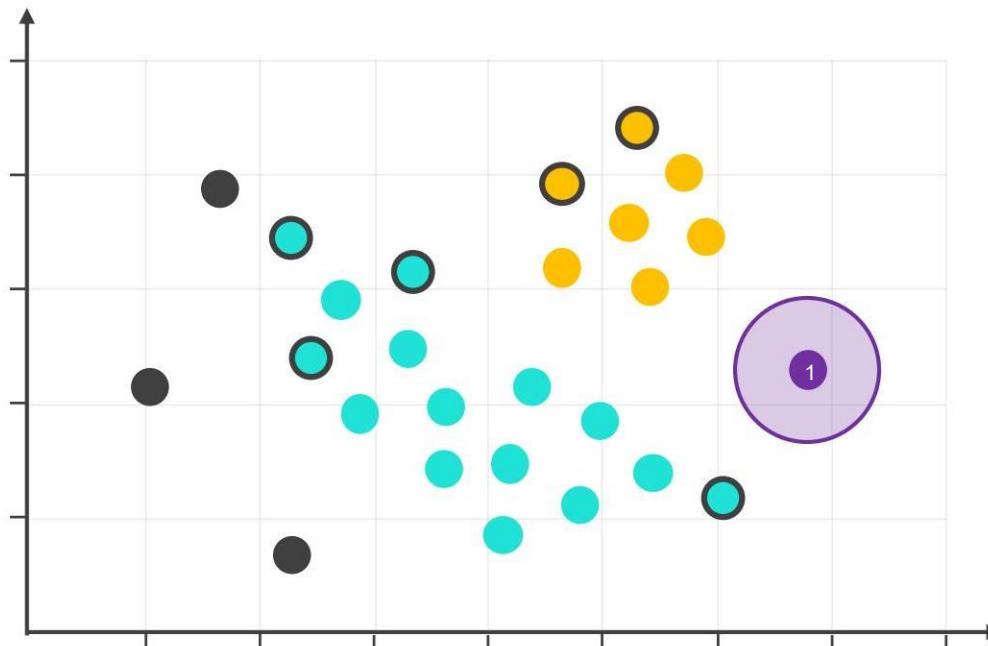
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

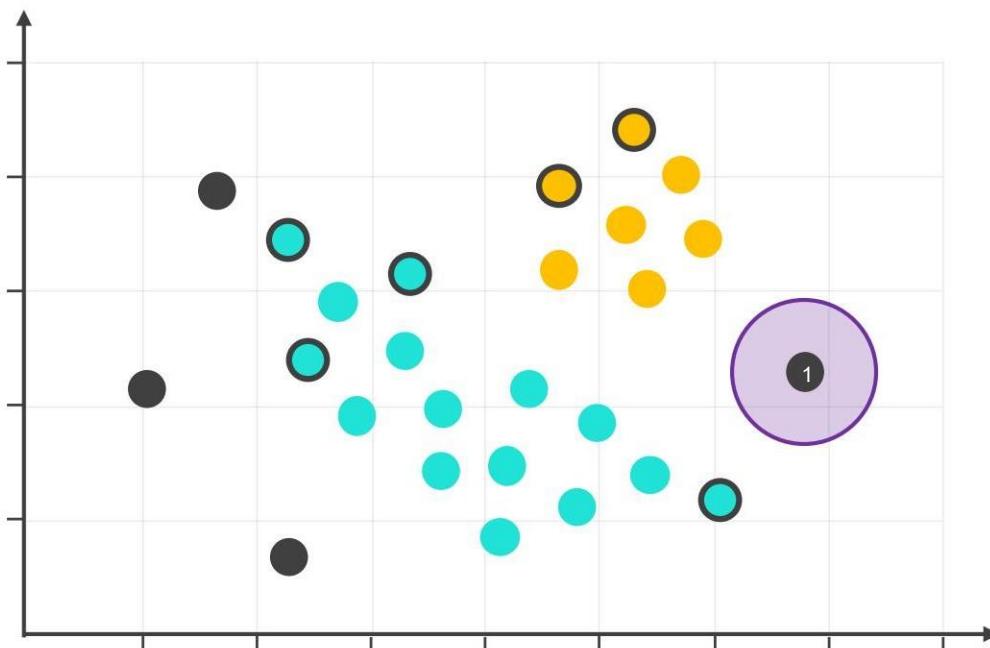
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

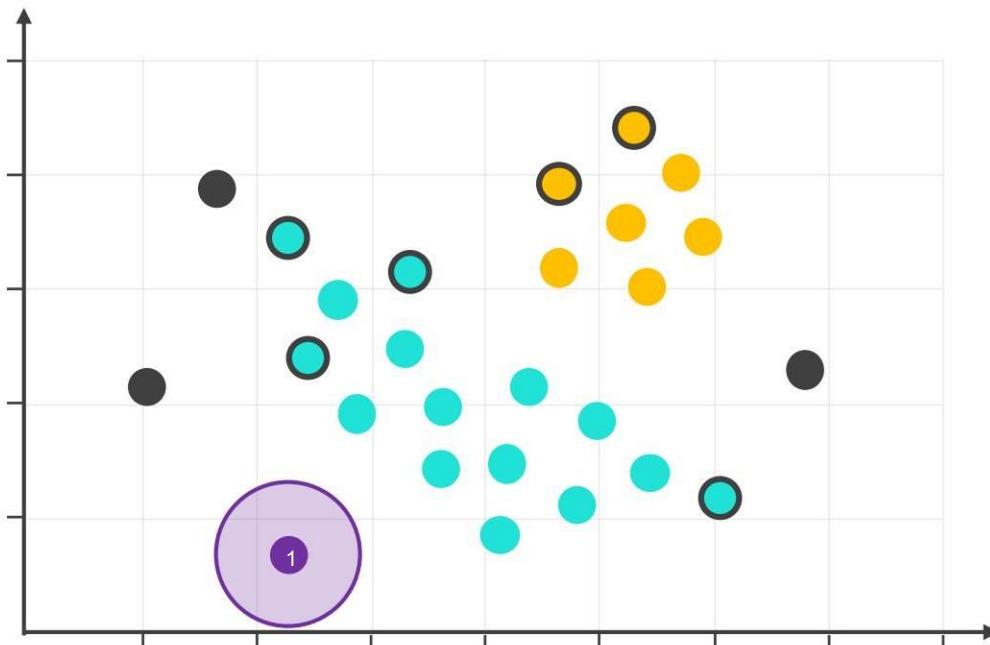
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

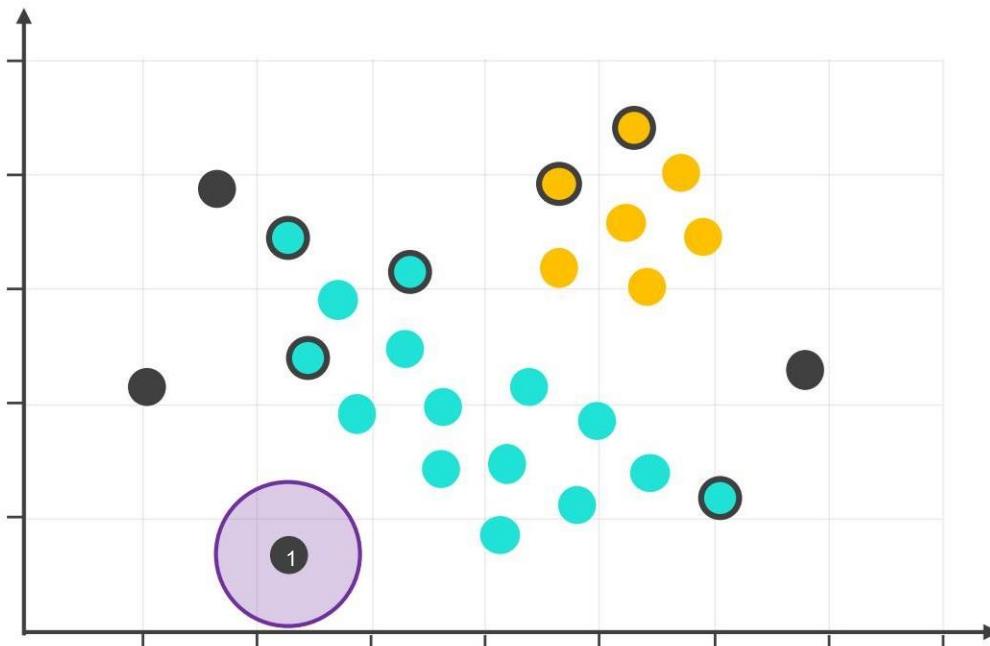
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

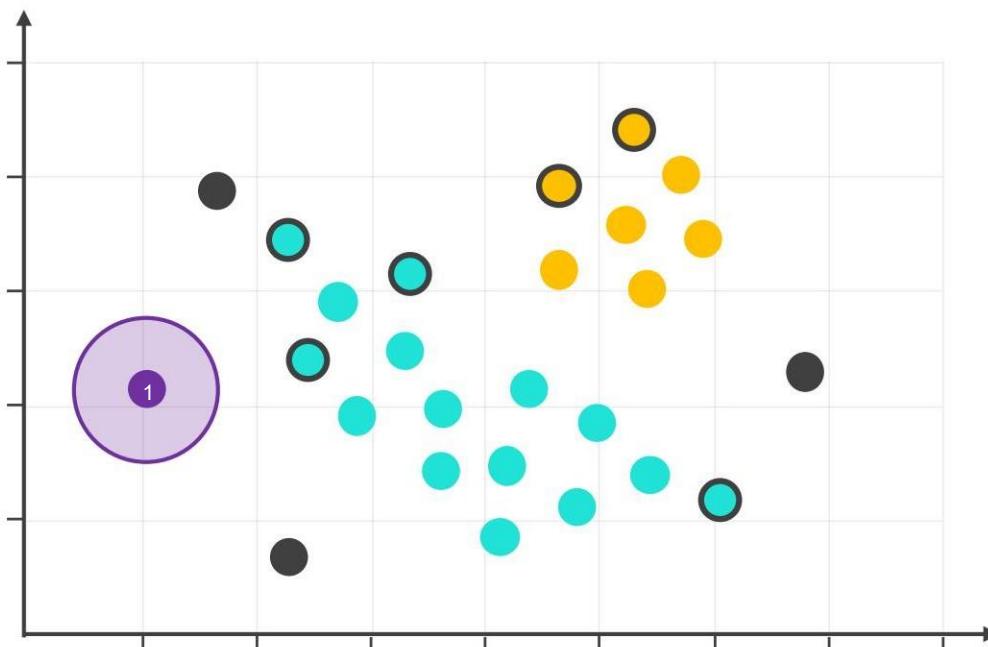
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

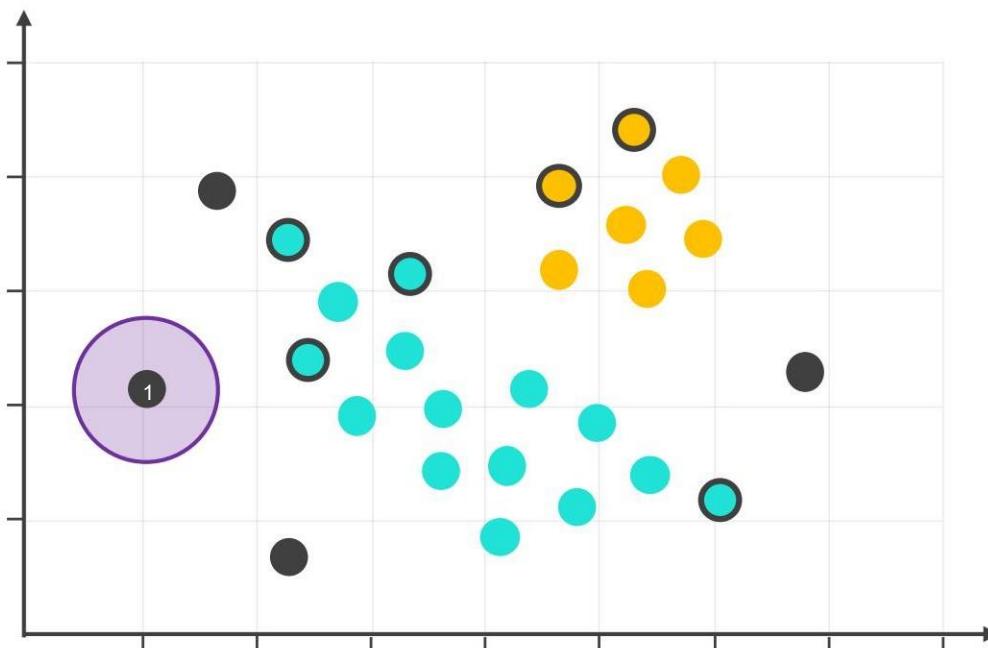
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



eps = 0,75
min_samples = 4



DBSCAN

Clustering Basics

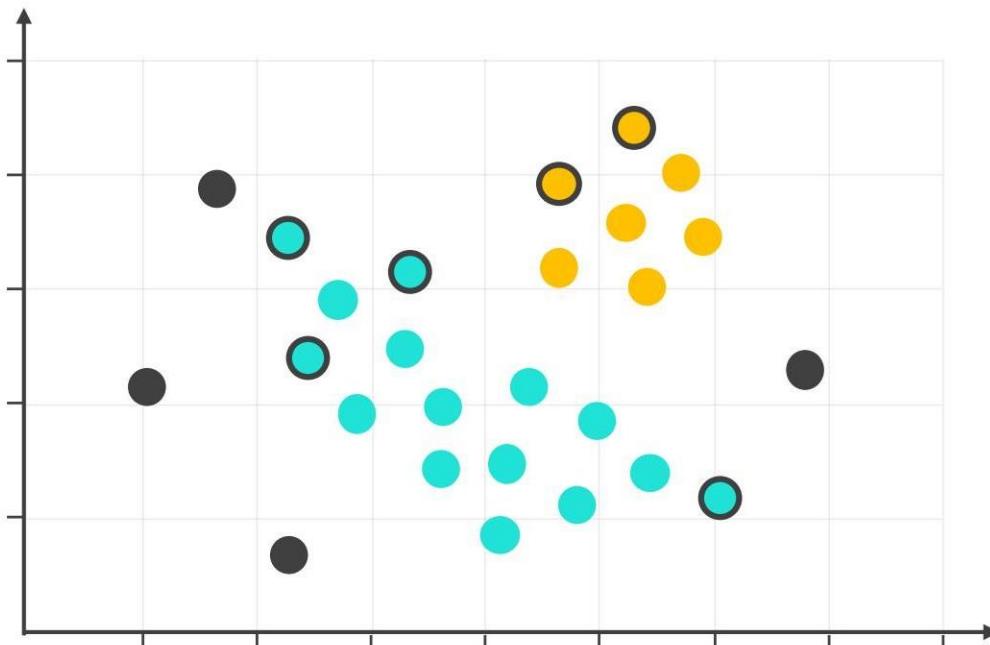
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



$\text{eps} = 0,75$
 $\text{min_samples} = 4$



DBSCAN

Clustering Basics

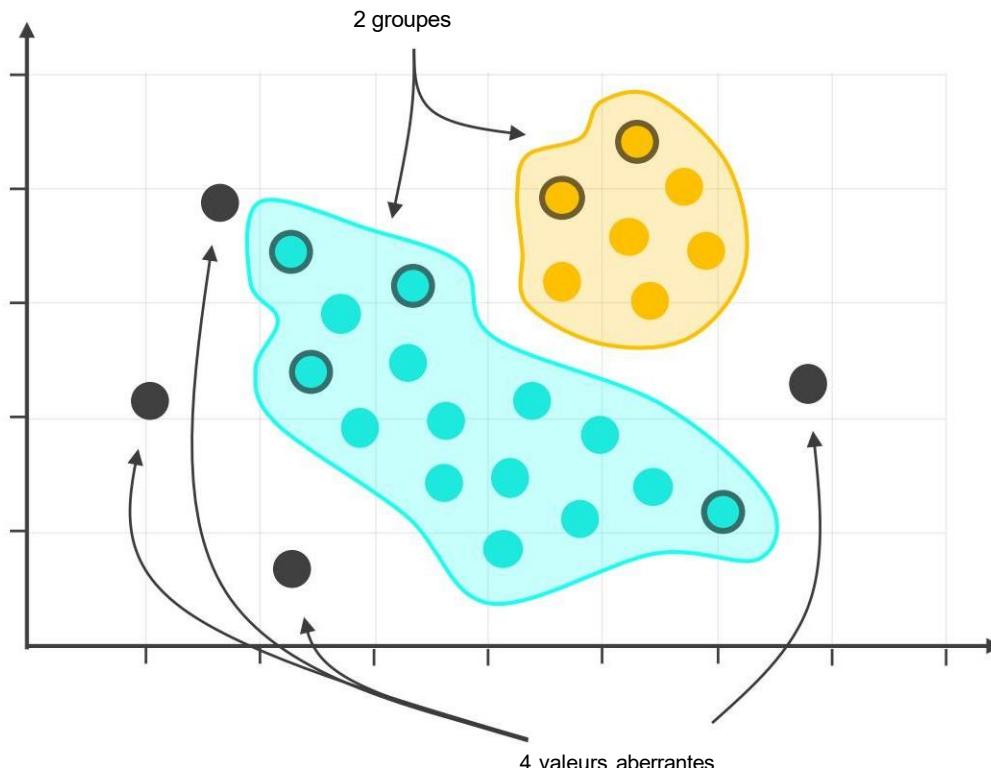
K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

ÉTAPE 5 : Passez à un autre point aléatoire et répétez les mêmes étapes.



eps = 0,75
min_samples = 4



ÉTAPES DÉTAILLÉES DE DBSCAN

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Voici quelques étapes plus détaillées du fonctionnement de DBSCAN :

1. Sélectionnez un rayon (eps) et un nombre minimal de points (min_samples)
2. Dans un nuage de points, sélectionnez un point au hasard et comptez les points situés dans son rayon.
 - Si la valeur est supérieure ou égale à « min_samples », alors créez un cluster, désignez le point comme point central et marquez les points situés dans son rayon comme voisins.
 - Si la valeur est inférieure à « min_samples », étiquetez le point comme point de bruit et passez à l'étape 5.
3. Déplacezvous vers un voisin et comptez les points situés dans son rayon d'action.
 - Si sa valeur est supérieure ou égale à « min_samples », étiquetezla comme point central et marquez ses voisins.
 - Si le nombre d'échantillons est inférieur à « min_samples », mais qu'au moins un de ces échantillons est un point central, qualifiezle de point frontière.
 - Si le nombre d'échantillons est inférieur à « min_samples » et qu'aucun de ces échantillons n'est un point central, considérezle comme un point de bruit.
4. Continuez avec un autre voisin jusqu'à ce que tous les points soient étiquetés à l'intérieur du groupe.
5. Passez à un autre point aléatoire et répétez les mêmes étapes.



DBSCAN EN PYTHON

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

```
from sklearn.cluster import DBSCAN  
  
dbscan = DBSCAN(eps=0.5, min_samples=5)
```



Le rayon, ou distance maximale entre points voisins

(valeur par défaut : 0,5)

Nombre minimal de points dans le rayon requis

pour devenir un point central, y compris le point
luimême

(la valeur par défaut est 5)



Vous aurez peut-être remarqué l'absence du paramètre `random_state`, même si DBSCAN démarre à un point aléatoire.

Cela s'explique par le fait que l'implémentation de DBSCAN par `sklearn` commence au premier point de l'ensemble de données et se poursuit à partir de là. Vous pourriez tenter d'obtenir des étiquettes différentes en mélangeant les données, mais les étiquettes de DBSCAN ne varient pas autant que celles du clustering KMean, ce qui rend ce problème moins important.



DBSCAN EN PYTHON

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

```
# import dbSCAN from sklearn
from sklearn.cluster import DBSCAN
```

```
# fit a dbSCAN model
dbSCAN = DBSCAN(eps=0.5, min_samples=5)
dbSCAN.fit(data)
```

▼ DBSCAN
DBSCAN()

```
# view the cluster assignments
dbSCAN.labels_
```

```
array([ 0, -1,  0,  1,  1,  1,  1,  1,  0,  1, -1,  0,  0,  0,  0, -1,  1,
       0,  1, -1,  1, -1,  1,  1,  1,  1,  0,  0,  0, -1,  1, -1,  1,  1,
       1,  1,  1, -1,  0,  1,  1,  1, -1,  0, -1,  1,  0, -1,  0, -1,  1,
       1, -1,  1,  1,  0,  1, -1,  1,  0,  0,  0,  1,  1, -1,  0, -1,
      -1,  1,  1,  1,  1,  1,  1,  1,  1,  1, -1,  1,  1,  1,  0,  0,  0,
       1,  1,  0,  0,  1,  0,  1,  1,  1,  1,  1, -1,  1,  1,  0,  1,  0,
      -1,  1, -1,  1,  0,  1,  0, -1, -1, -1,  0,  0,  0,  1,  1,  0,  1,
       1,  1, -1, -1,  0,  0, -1, -1,  1,  1,  0,  0,  0,  1,  0,  1,  1,
       1,  1,  1,  1,  0,  0,  1, -1, -1,  1,  1,  0,  1,  0,  1,  1,  1])
```

Vous pouvez consulter les affectations de cluster à l'aide de l' attribut .labels_
(Les valeurs 1 représentent des points de bruit)



Il y a trop de points de bruit (1) ici,
ajustons eps et min_samples pour
mieux visualiser les clusters.



SCORE DE LA SILHOUETTE

Clustering Basics

K-Means
Clustering

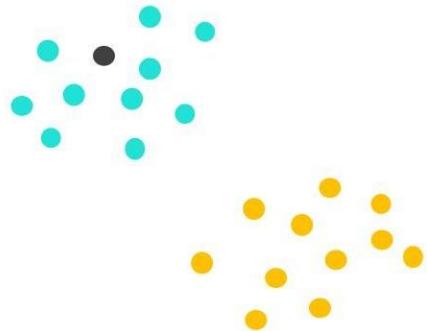
Hierarchical
Clustering

DBSCAN

Comparing
Models

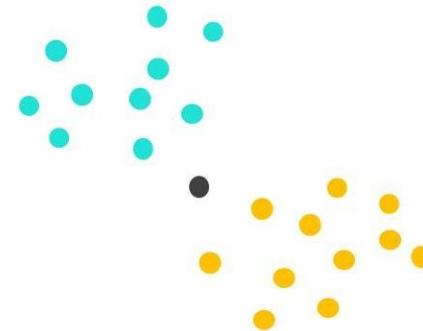
Bien que l'intuition joue un rôle important dans la comparaison des modèles de clustering, vous pouvez également utiliser des métriques comme le score de silhouette pour faciliter cette comparaison.

- Les scores varient de -1 à +1, les valeurs les plus élevées indiquant que les points de données correspondent fortement à leur propre groupe et faiblement aux autres groupes (ce qui est une bonne chose !).



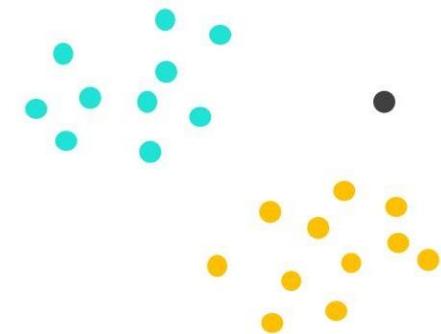
Score de silhouette = positif

Ce point de données s'intègre bien à son groupe et est éloigné des autres groupes.



Score de silhouette ~ zéro

Le point de données se situe près de la limite entre deux groupes.



Score de silhouette = négatif

Ce point de données ne s'intègre pas bien dans son propre groupe et pourrait appartenir à un groupe voisin.



SCORE DE LA SILHOUETTE

Clustering Basics

K-Means Clustering

Hierarchical Clustering

DBSCAN

Comparing Models

La formule suivante est utilisée pour calculer le score de silhouette pour chaque point de données :

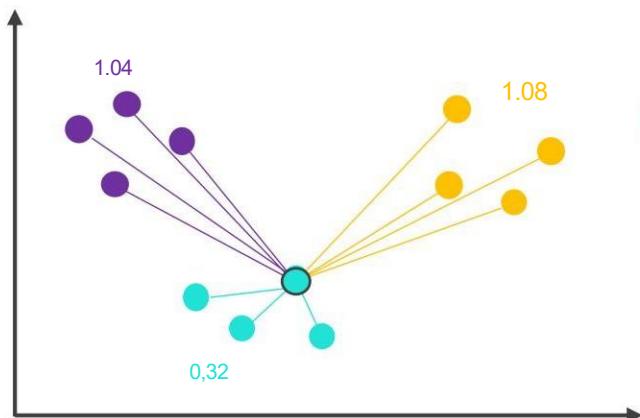
Distance moyenne minimale aux points appartenant à un autre groupe

Distance moyenne aux autres points du même groupe

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Score de silhouette pour le ième point de données

Le dénominateur normalise la différence de sorte que le score de silhouette se retrouve entre 1 et -1.



$$\frac{1.04 - 0.32}{1.04} = 0.69$$



Ce point de données s'intègre parfaitement à son groupe !



Pour calculer le **score de silhouette global** d'un modèle donné, faites la moyenne de tous les scores de silhouette.



SCORE DE SILHOUETTE EN PYTHON

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Vous pouvez utiliser la fonction silhouette_score() de scikitlearn pour calculer le score de silhouette des résultats d'un modèle de clustering en Python.

```
from sklearn.metrics import silhouette_score  
  
silhouette_score(data, labels, metric='euclidean', sample_size=None)
```

0.3419620



Les données sur lesquelles
vous ajustez le modèle
(requis)

Les étiquettes
générées par votre modèle
(requis)

Calcul de distance utilisé pour
calculer le score de silhouette
(facultatif, la valeur par défaut est « euclidien »)

Taille du sousensemble aléatoire de
données utilisé pour calculer le score
(facultatif, par défaut : Aucun)



RÉCAPITULATIF DES ALGORITHMES DE CLUSTERING

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Chaque algorithme de clustering présente des avantages et des inconvénients à prendre en compte :

Model	Pros	Cons	In Practice
Kmoyennes Clustering	Facile à comprendre et à interpréter S'adapte bien aux grands ensembles de données	Le nombre de clusters doit être spécifié. Des centroïdes initiaux différents conduisent à résultats différents Suppose que les amas sont approximativement sphériques	Modèle de clustering populaire Généralement le premier choix en matière de clustering
Hiérarchique Clustering	Il n'est pas nécessaire de prédéfinir « k » à l'avance. Peut traiter des ensembles de données complexes	Ne s'adapte pas bien aux grands ensembles de données Les points de données peuvent se connecter à des valeurs aberrantes, ce qui fausse les calculs de distance.	Bon pour la visualisation Le dendrogramme vous permet d'explorer visuellement les groupes.
DBSCAN	Il n'est pas nécessaire de prédéfinir « k » à l'avance. Peut traiter des ensembles de données complexes	Ne s'adapte pas bien aux grands ensembles de données Le réglage des hyperparamètres est un défi	Idéal pour les valeurs aberrantes Identifie et traite efficacement les points de bruit dans les données



RÉCAPITULATIF DES ALGORITHMES DE CLUSTERING

Clustering Basics

K-Means
Clustering

Hiérarchique
Clustering

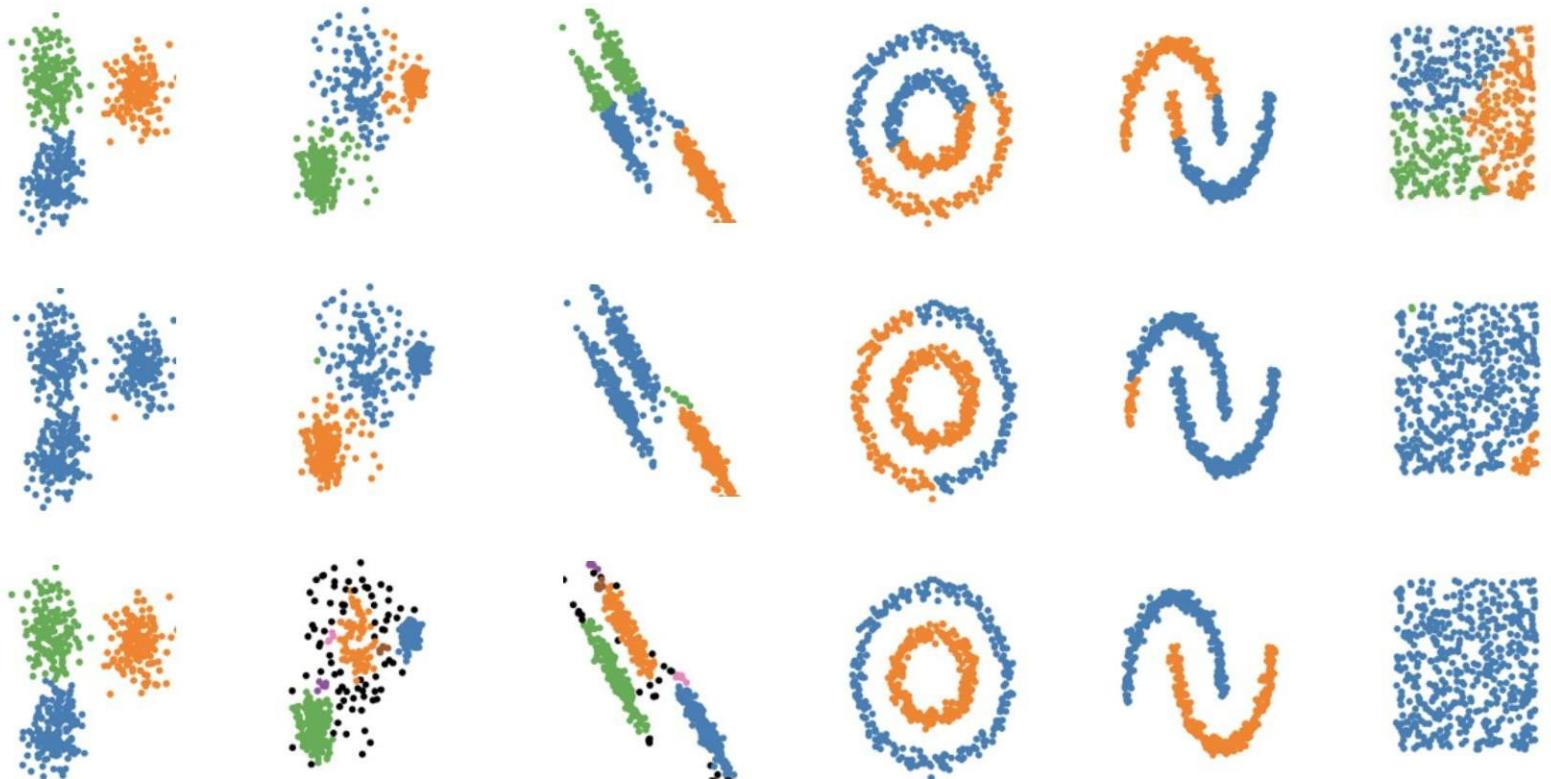
DBSCAN

Comparing
Models

Kmoyennes
Clustering

Hiérarchique
Clustering

DBSCAN



Voici comment chaque algorithme de clustering se compare sur différents ensembles de données :



PROCHAINES ÉTAPES DU GROUPEMENT

Clustering Basics

K-Means
Clustering

Hierarchical
Clustering

DBSCAN

Comparing
Models

Une fois vos modèles initiaux ajustés et interprétés, les prochaines étapes potentielles sont les suivantes :

11. Comparer les modèles de clustering

- Score de silhouette : Plus le score est élevé et proche de 1, mieux les groupes sont définis.
- Intuition : Au final, l'objectif principal est de répondre à votre question métier, donc vous
Vous souhaitez sélectionner les groupes qui vous semblent les plus pertinents pour prendre des décisions ?

22. Étiqueter les données non vues

- Préparation des données : Avant d'étiqueter des données inconnues, vous devez appliquer les mêmes transformations que celles appliquées à l'ensemble de données d'origine (mise à l'échelle des caractéristiques, etc.) lors de la création des clusters.
- Clustering KMeans : La méthode `predict` permet d'assigner des points de données non observés à un cluster.
- Clustering hiérarchique et DBSCAN : Bien qu'il n'existe pas de méthode .predict pour ces techniques, vous pouvez étiqueter les points de données inconnus en réajustant les modèles (vous devrez peut-être mettre à jour les paramètres).

POINTS CLÉS À RETENIR



Le clustering permet de trouver des groupes d'observations similaires les unes aux autres.

Les techniques de clustering courantes incluent le clustering Kmeans, le clustering hiérarchique et DBSCAN.



En pratique, suivez le flux de travail de clustering

- 1) Préparez vos données pour le clustering (granularité des lignes correcte, valeurs non nulles et numériques, mise à l'échelle des caractéristiques, etc.)
- 2) Commencez par le clustering KMeans, puis affinez et sélectionnez le « meilleur » modèle
- 3) Pour de meilleurs résultats, essayez le clustering hiérarchique ou DBSCAN, puis affinez et sélectionnez le modèle le plus adapté.

4) Comparez les modèles à l'aide des scores de silhouette et de votre intuition, et choisissez celui qui répond le mieux à votre question métier.