

“Exploration Approfondie des Techniques ML : PCA, GMM et KNN dans la Classification et la Réduction de Dimensionnalité”

Réaliser par:

- DAGHMOUMI Marouan
- BENJELLOUN Abdelmajid

Encadrer par:

- Pr. El Mokhtar EN-NAIMI

TRAVAIL COLLABORATIF



Abdelmajid BENJELLOUN

Etudiant en master
intelligence artificiel et
science de données



PR. El Mokhtar EN-NAIMI

Professeur de Sciences
Informatiques à la Faculté
des Sciences et Techniques
de Tanger



Marouan DAGHMOUMI

Etudiant en master
intelligence artificiel et
science de données

TABLE DES MATIÈRES

- **MACHINE LEARNING**

- **ANALYSE EN COMPOSANTES PRINCIPALES (PCA) :**

1. DÉFINITION
2. Les étapes de l'Analyse en Composantes Principales (PCA)
3. Caractéristiques de PCA
4. Avantages et Inconvénients de PCA

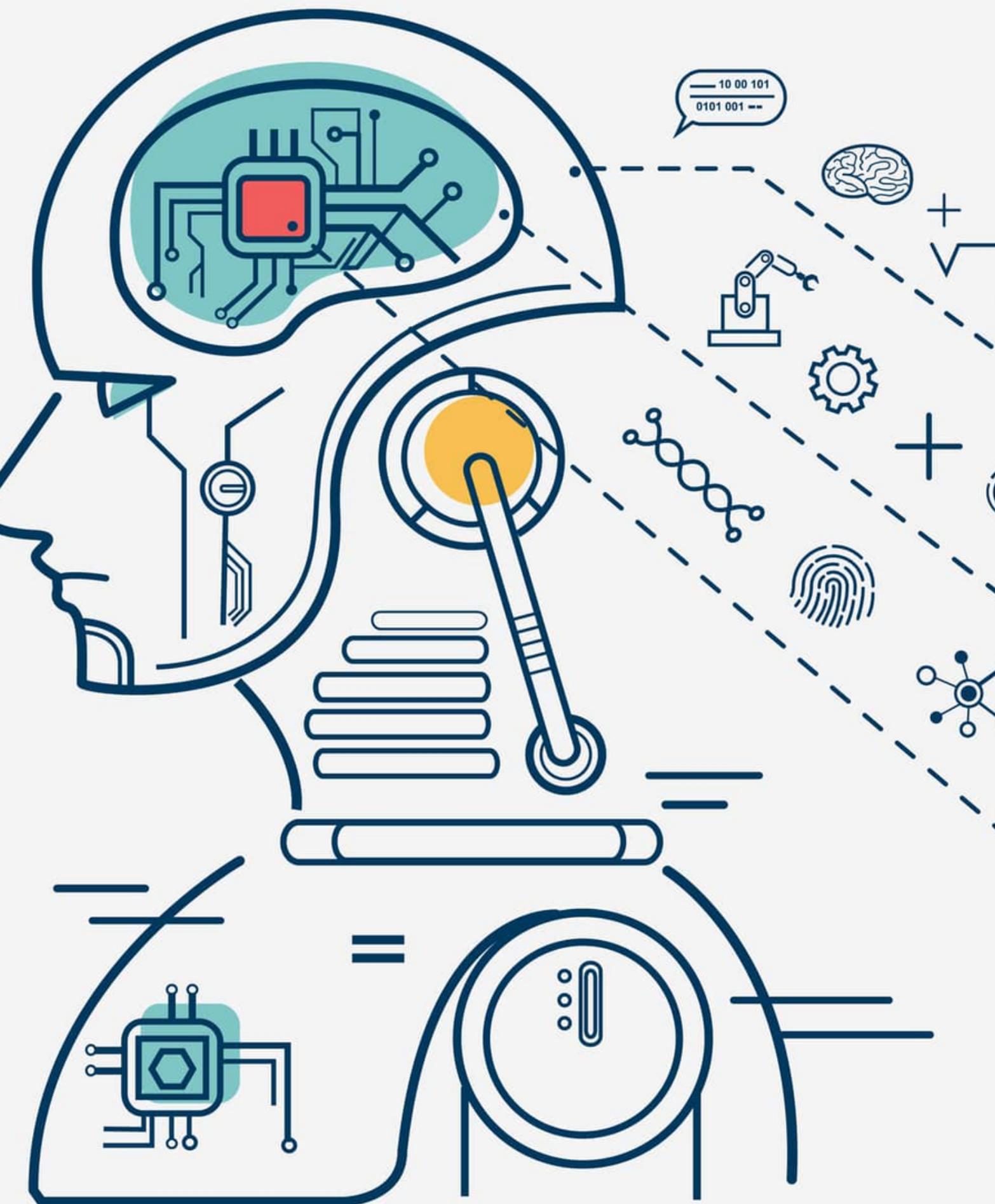
- **K- Nearest Neighbors (KNN) :**

1. DÉFINITION
2. Les Étapes de k plus proches voisins (KNN)
3. Caractéristiques de KNN
4. Avantages et Inconvénients de KNN

- **Gaussian Mixture (GMM) :**

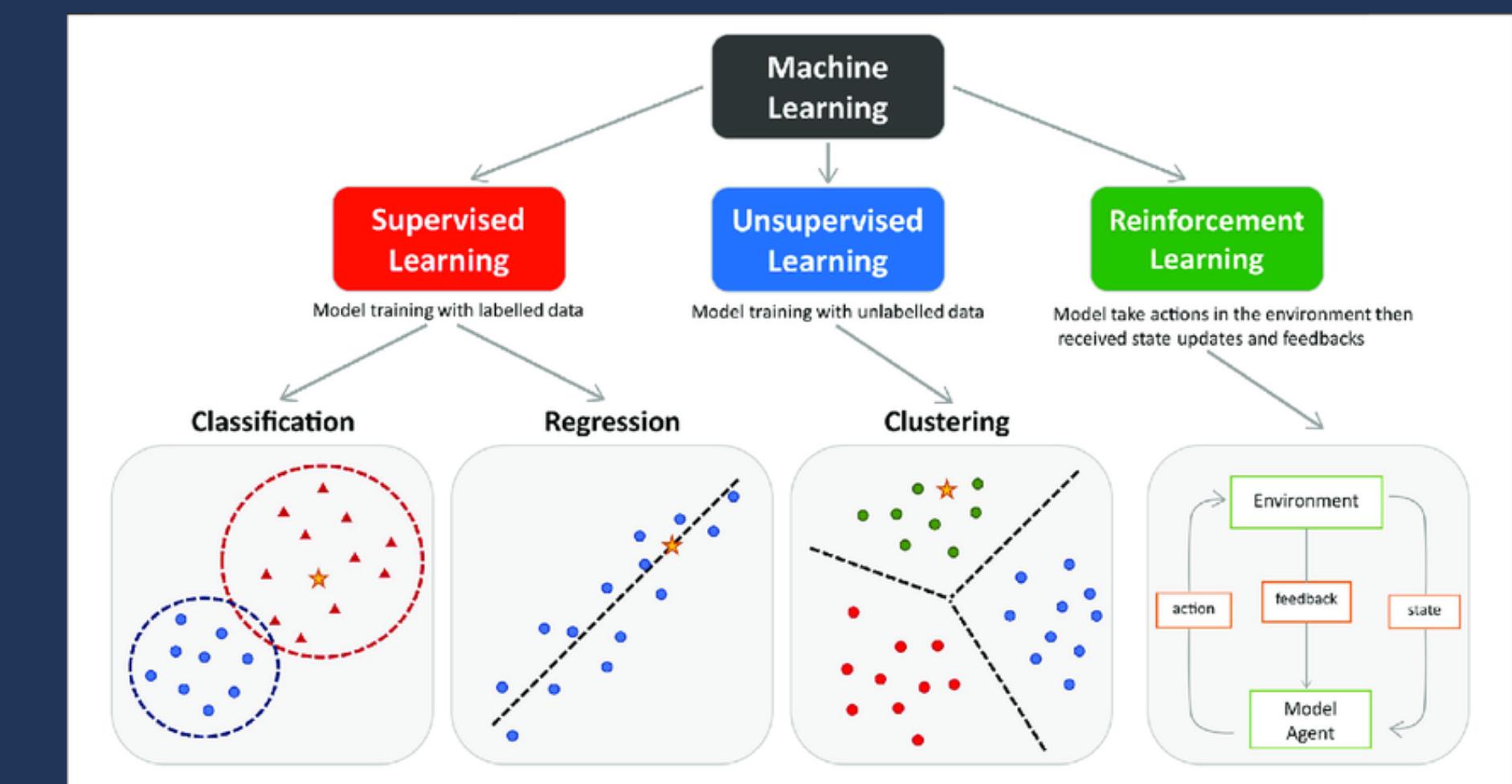
1. DÉFINITION
2. Étapes de Gaussian Mixture (GMM)
3. Caractéristiques de GMM
4. Avantages et Inconvénients de GMM





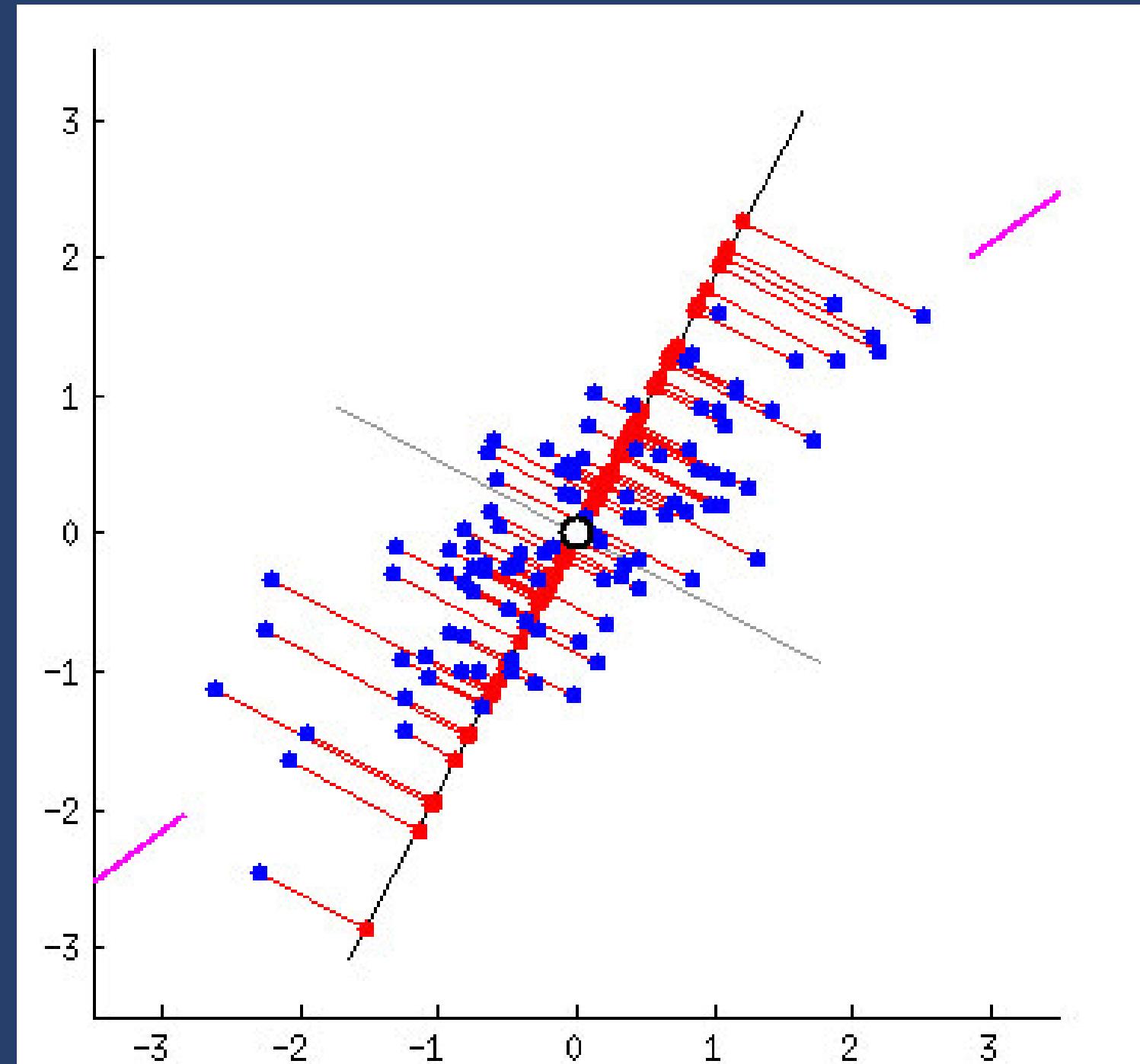
MACHINE LEARNING

L'apprentissage automatique, ou machine learning, est une discipline de l'informatique qui permet aux ordinateurs d'apprendre à partir de données et d'effectuer des tâches sans être explicitement programmés. Il se divise généralement en trois catégories principales : la supervision, la non-supervision et le renforcement.

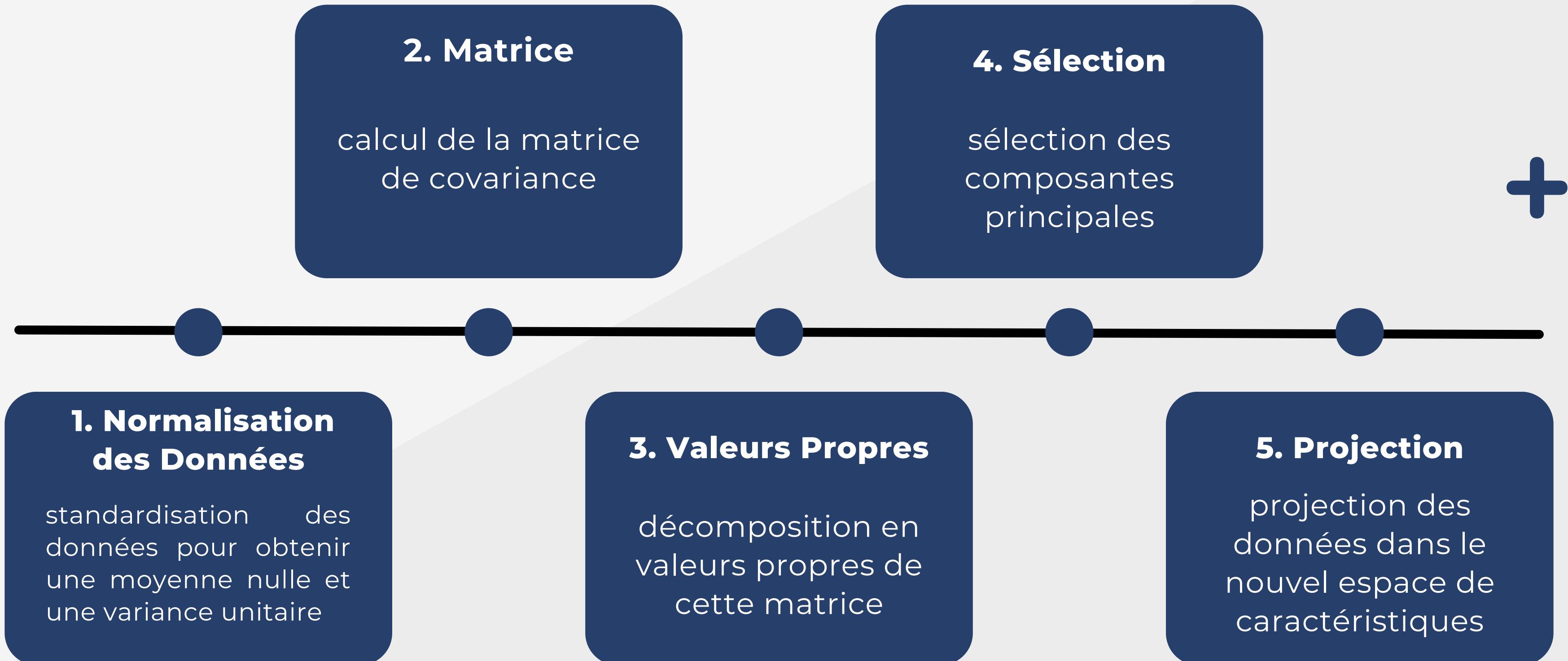


ANALYSE EN COMPOSANTES PRINCIPALES

L'Analyse en Composantes Principales (PCA) est une technique largement utilisée en statistique et en apprentissage automatique pour réduire la dimensionnalité des données tout en préservant leur structure essentielle. En d'autres termes, elle permet de simplifier la complexité des données en les projetant dans un espace de dimension inférieure tout en conservant autant d'informations que possible.



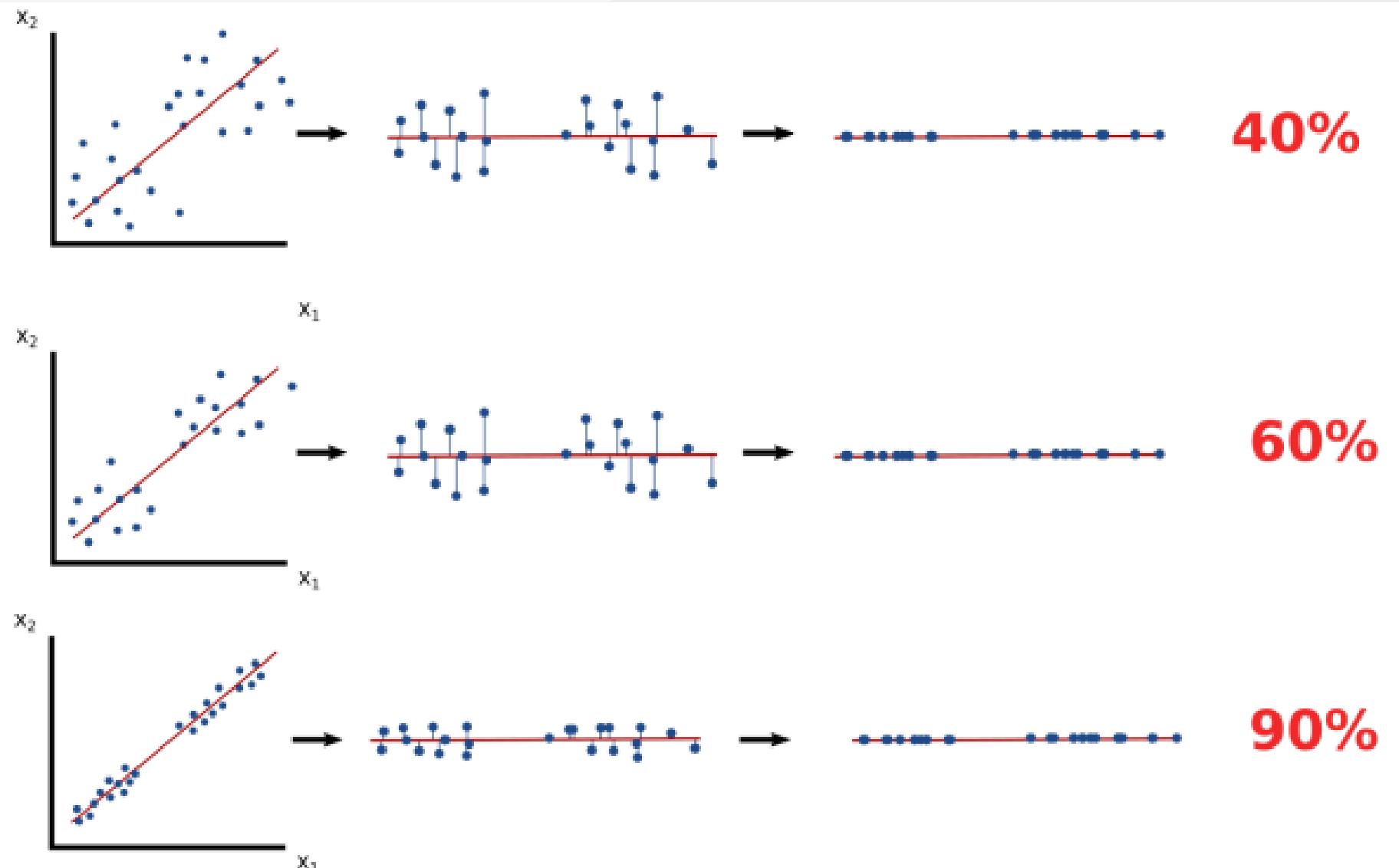
LES ÉTAPES DE L'ANALYSE EN COMPOSANTES PRINCIPALES





CARACTÉRISTIQUES DE LA PCA

- capacité à réduire la dimensionnalité des données tout en préservant leur structure
- sa facilité d'interprétation des composantes principales
- Sa robustesse aux données bruitées



Réduction de la dimensionnalité

PCA permet de réduire le nombre de variables dans un ensemble de données tout en conservant autant d'informations que possible

patterns cachés

PCA mette en évidence des relations et des patterns qui seraient difficiles à détecter dans l'ensemble de données original.

AVANTAGES DE PCA

Élimination du bruit

extrayant les composantes principales qui capturent le plus de variance, la PCA peut aider à éliminer le bruit et à mettre en évidence les structures sous-jacentes dans les données.

Facilité d'interprétation

Les composantes principales obtenues après l'application de la PCA sont des combinaisons linéaires des variables d'origine, ce qui facilite leur interprétation.

Perte d'interprétabilité

Lorsque de nombreuses variables sont combinées en composantes principales, il peut être difficile d'interpréter la signification des composantes résultantes.

Dépendance linéaire

PCA suppose que les relations entre les variables sont linéaires, ce qui peut ne pas être le cas dans tous les ensembles de données réels.

INCONVÉNIENTS DE PCA

Sensibilité aux données aberrantes

PCA est sensible aux données aberrantes, ce qui signifie qu'un petit nombre de valeurs aberrantes peut avoir un impact disproportionné sur les résultats de l'analyse.

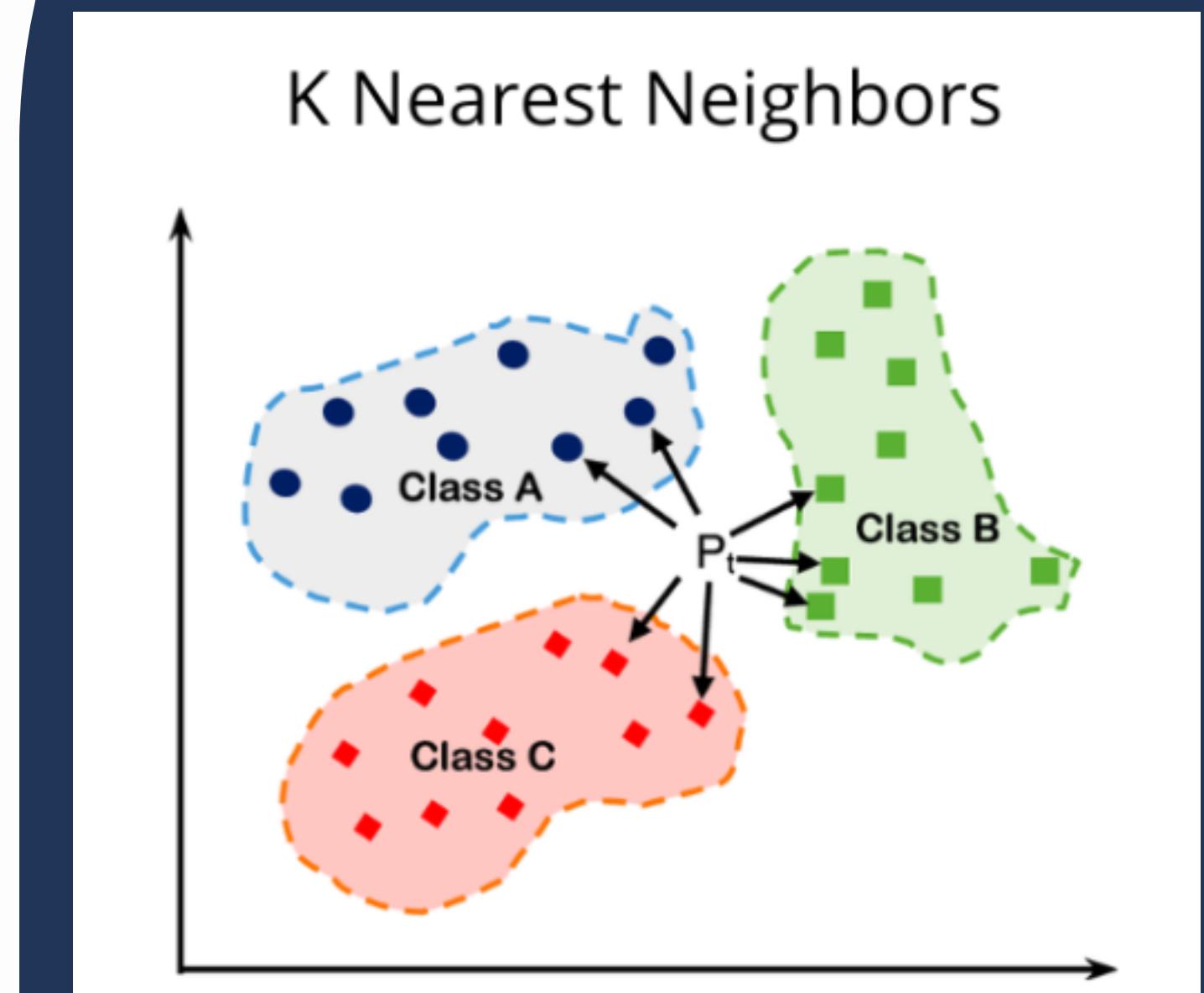
Traitement des données catégorielles

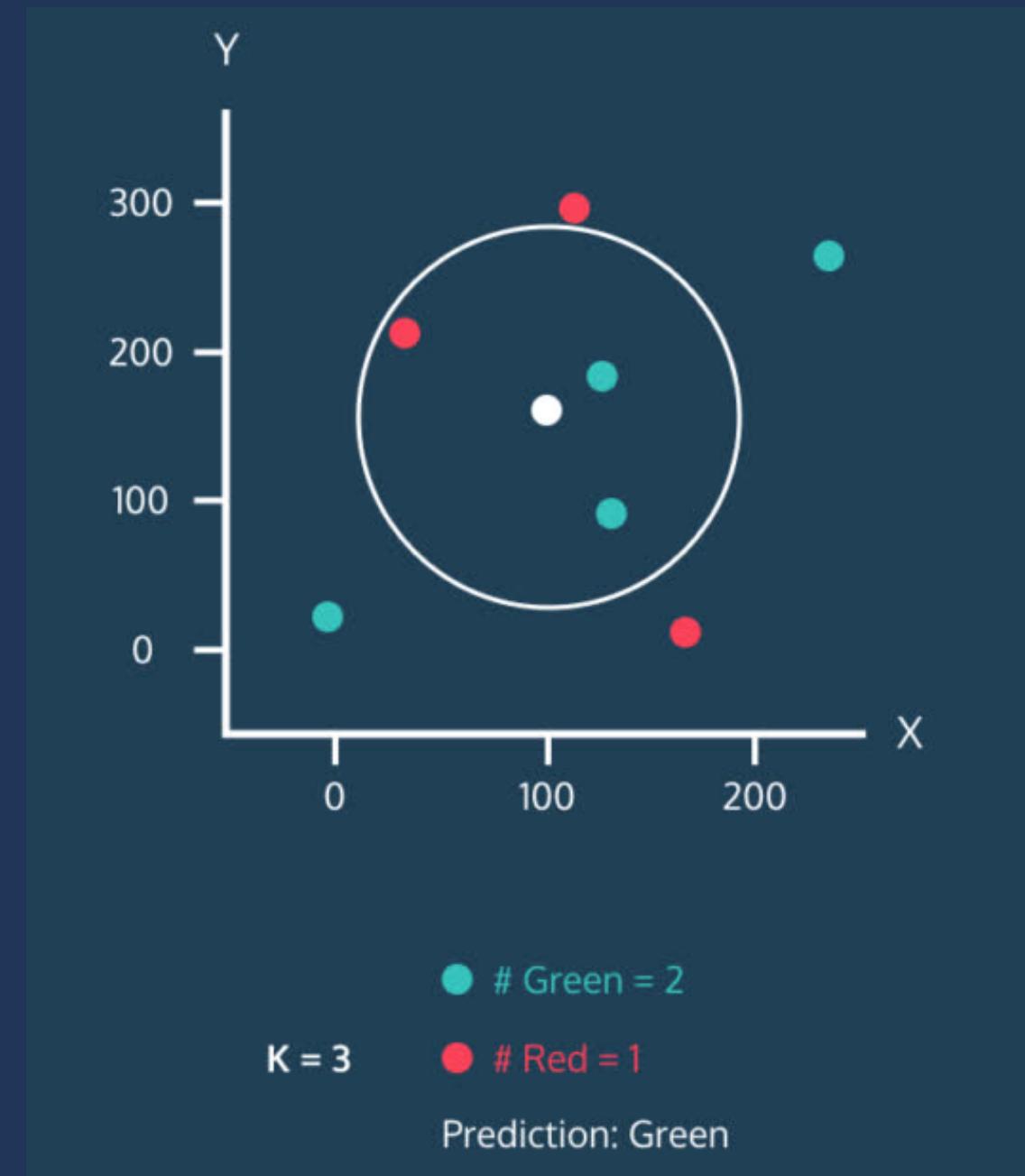
PCA est conçue pour les données numériques continues et peut ne pas être appropriée pour les données catégorielles ou binaires sans transformation préalable.

K- Nearest Neighbors (KNN)

L'algorithme des k plus proches voisins (KNN) est une méthode d'apprentissage supervisé utilisée pour la classification et la régression:

- En classification, l'algorithme attribue une étiquette de classe à une instance en fonction de la majorité des étiquettes de classe de ses voisins les plus proches.
- En régression, il prédit la valeur cible d'une instance en prenant la moyenne des valeurs cibles de ses voisins les plus proches.





01

SÉLECTION DU NOMBRE DE VOISINS (K)

Déterminez le nombre de voisins à considérer pour la prédiction K

02

CALCUL DE LA DISTANCE

Utilisez une mesure de distance comme la distance euclidienne pour calculer la distance entre le nouvel exemple et tous les exemples du jeu de données.

03

SÉLECTION DES K VOISINS LES PLUS PROCHES

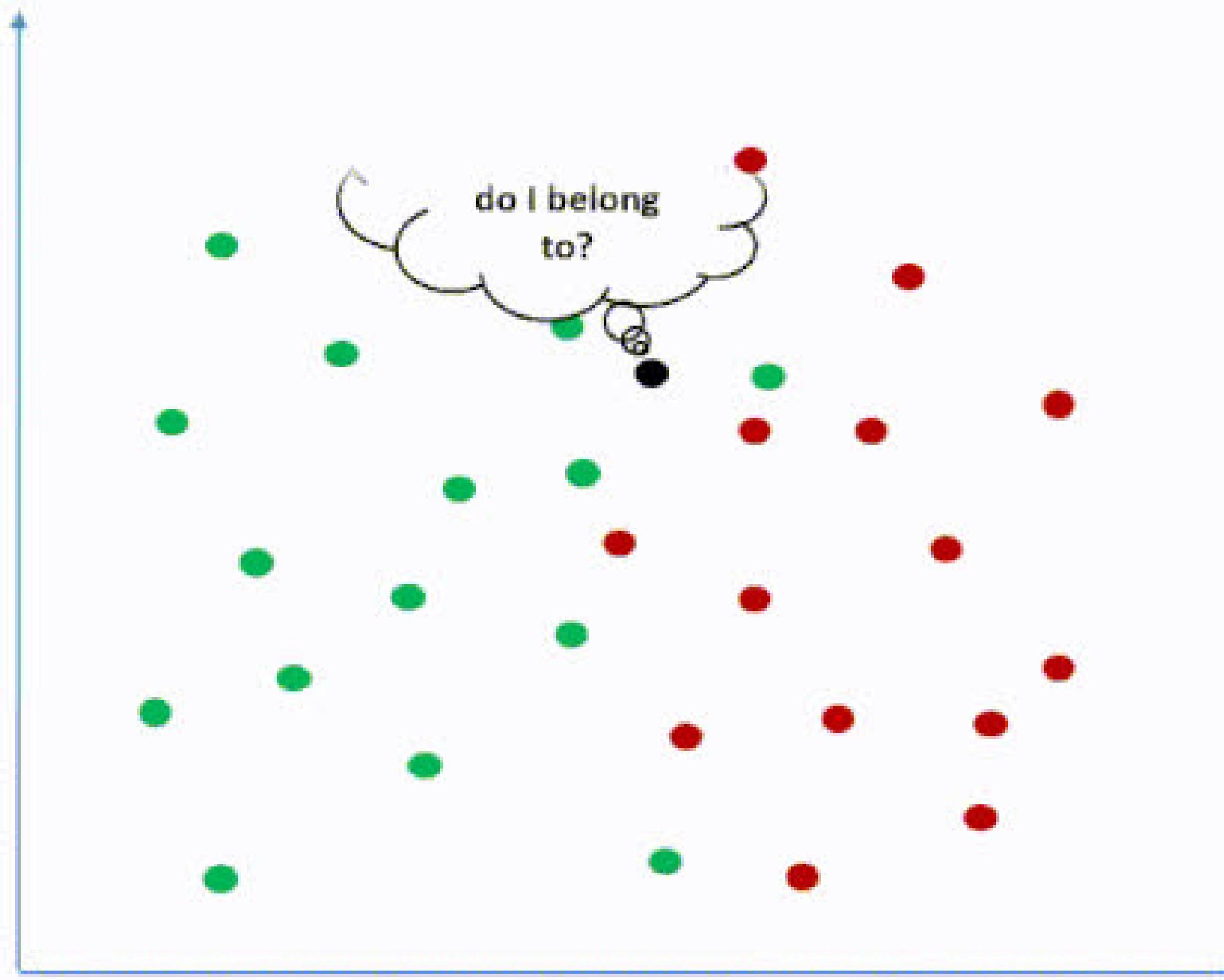
Identifiez les K voisins les plus proches en fonction des distances calculées .

04

ATTRIBUTION DE L'ÉTIQUETTE DE CLASSE MAJORITAIRE

Sélectionnez l'étiquette de classe majoritaire parmi les K voisins.

K-Nearest Neighbors Classification



CARACTÉRISTIQUES DE KNN

MÉTHODE NON-PARAMÉTRIQUE

KNN est une méthode d'apprentissage non paramétrique, ce qui signifie qu'elle ne fait aucune hypothèse explicite sur la distribution des données.

CLASSIFICATION ET RÉGRESSION

KNN peut être utilisé à la fois pour la classification et la régression, ce qui en fait un algorithme polyvalent

SIMPLICITÉ

KNN est simple à comprendre et à mettre en œuvre. Il ne nécessite pas de phase d'apprentissage coûteuse car il mémorise simplement les données d'entraînement.

LES AVANTAGES DE KNN

Facilité d'implémentation

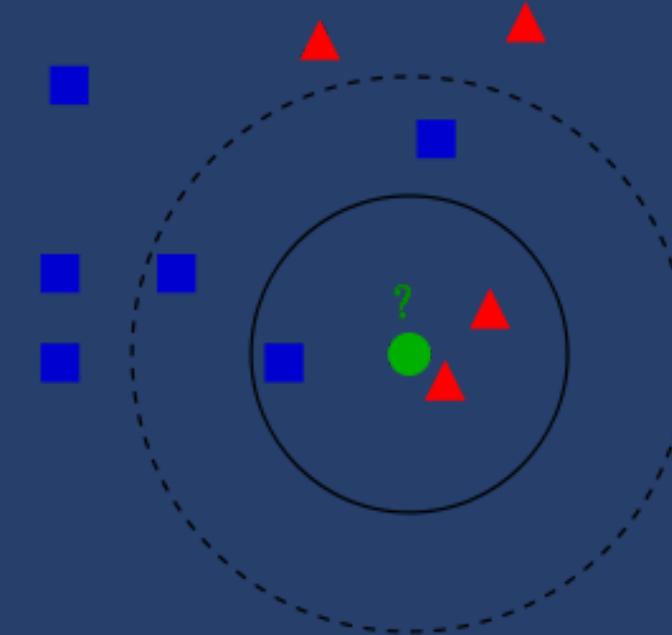
KNN est facile à comprendre et à mettre en œuvre, ce qui en fait un bon choix pour les problèmes simples où la complexité de l'algorithme n'est pas une préoccupation majeure.

Adaptabilité aux données complexes

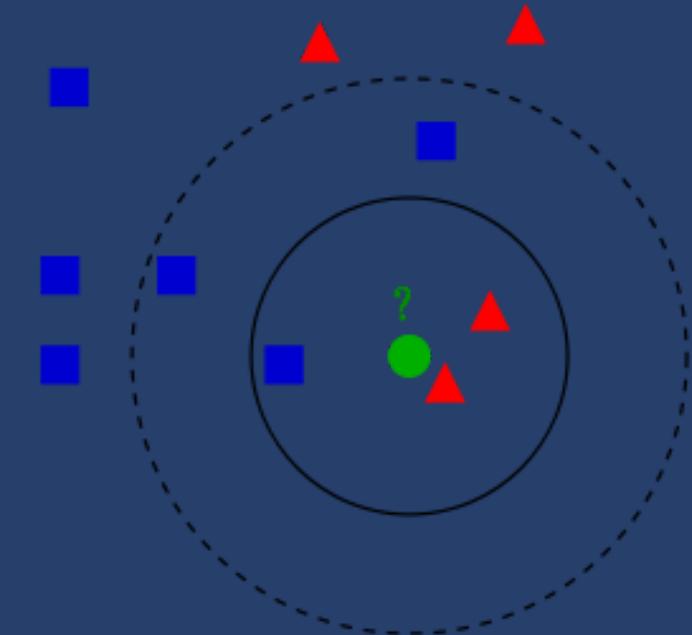
KNN peut s'adapter à des distributions de données complexes et à des frontières de décision non linéaires sans nécessiter de modifications de l'algorithme lui-même.

Utilisation dans les systèmes de recommandation

En raison de sa simplicité et de sa capacité à capturer les similitudes entre les instances, KNN est souvent utilisé dans les systèmes de recommandation pour recommander des éléments similaires à ceux que l'utilisateur a précédemment appréciés.



LES INCONVÉNIENTS DE KNN



Sensibilité à la dimensionnalité

KNN devient moins efficace à mesure que le nombre de dimensions des données augmente, en raison du phénomène de la malédiction de la dimensionnalité.

Calcul intensif

prédition lente car il faut revoir tous les exemples à chaque fois.
méthode gourmande en place mémoire

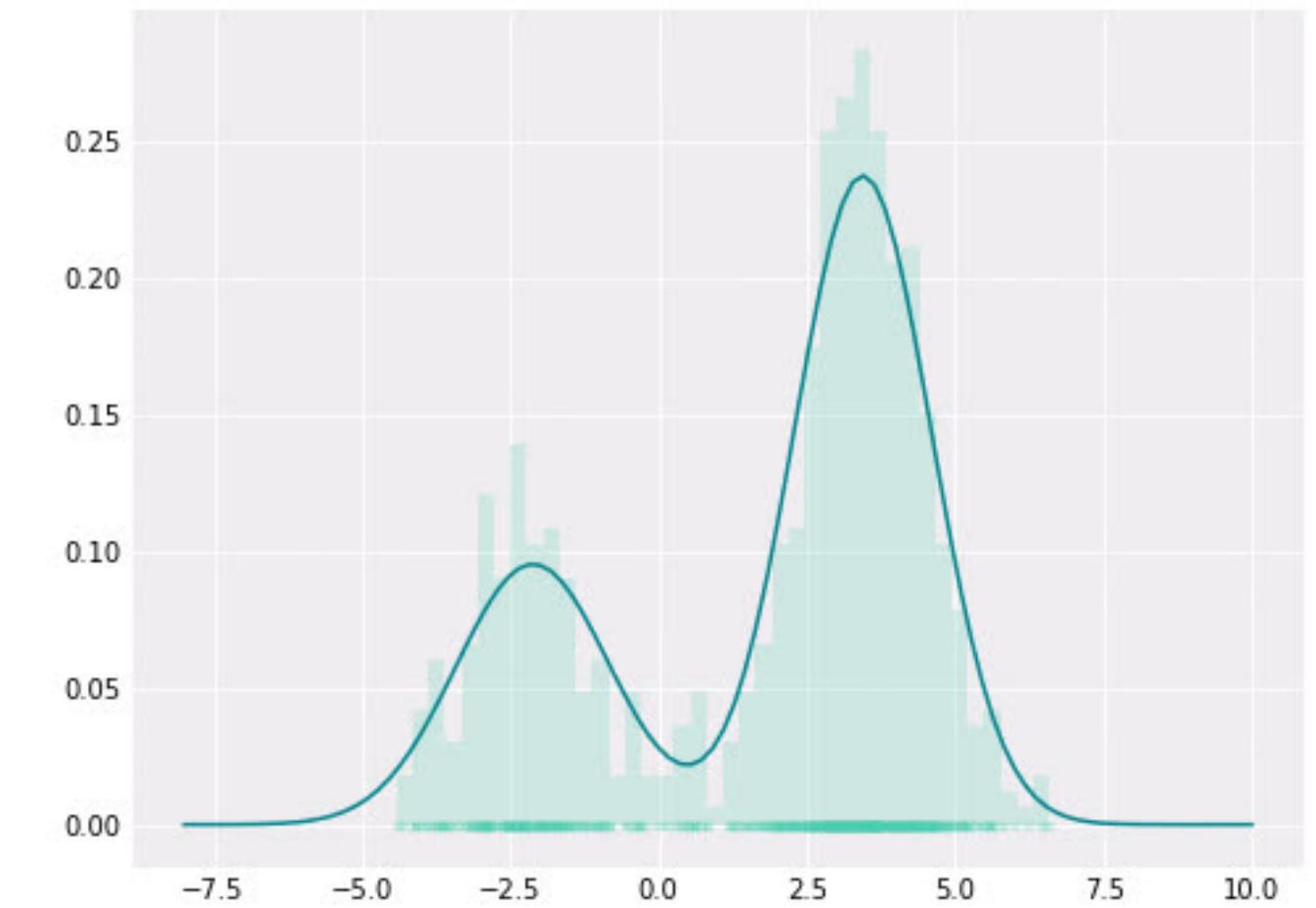
Besoin de données étiquetées

KNN est un algorithme d'apprentissage supervisé, il nécessite des données étiquetées pour fonctionner. Si les données ne sont pas étiquetées ou si les étiquettes sont incorrectes, cela peut affecter la qualité des prédictions.

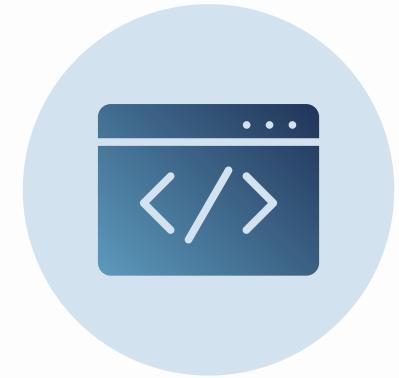
GAUSSIAN MIXTURE (GMM)

L'**algorithme Gaussian Mixture Model (GMM)** est une méthode d'apprentissage non supervisé largement utilisée pour la modélisation de données.

Gaussian Mixture Model est une méthode probabiliste qui suppose que les données sont générées à partir d'un mélange de plusieurs distributions gaussiennes. Il cherche à estimer les paramètres de ces distributions pour mieux comprendre la structure sous-jacente des données.



Caractéristiques de GMM



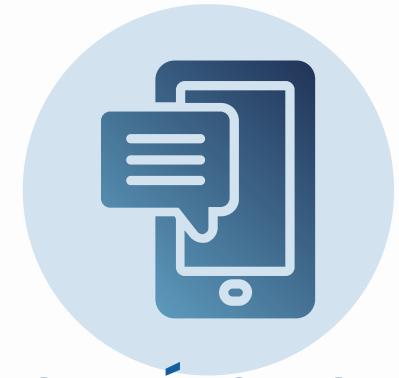
CLUSTERING SOUPLE

Contrairement à K-Means, où chaque point de données est assigné à un seul cluster, GMM permet à un point de données d'appartenir à plusieurs clusters avec des probabilités différentes, ce qui permet une segmentation plus flexible des données.



ESTIMATION DES PARAMÈTRES

GMM est capable d'estimer les paramètres du modèle, y compris les moyennes, les covariances et les poids de mélange, à partir des données d'entraînement, ce qui en fait une méthode d'apprentissage non supervisé puissante.



DONNÉES NON-LINÉAIRES

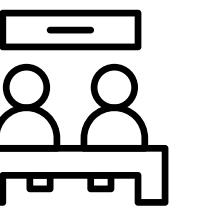
Bien que GMM soit souvent utilisé pour modéliser des données gaussiennes, il peut également être utilisé pour des données non linéaires en combinant plusieurs composantes gaussiennes pour représenter la distribution des données.

LES AVANTAGES DE GMM



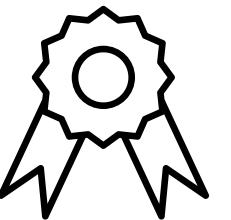
Flexibilité dans la Modélisation

GMM peut modéliser des distributions de données complexes en utilisant plusieurs composantes gaussiennes, ce qui le rend approprié pour un large éventail de données.



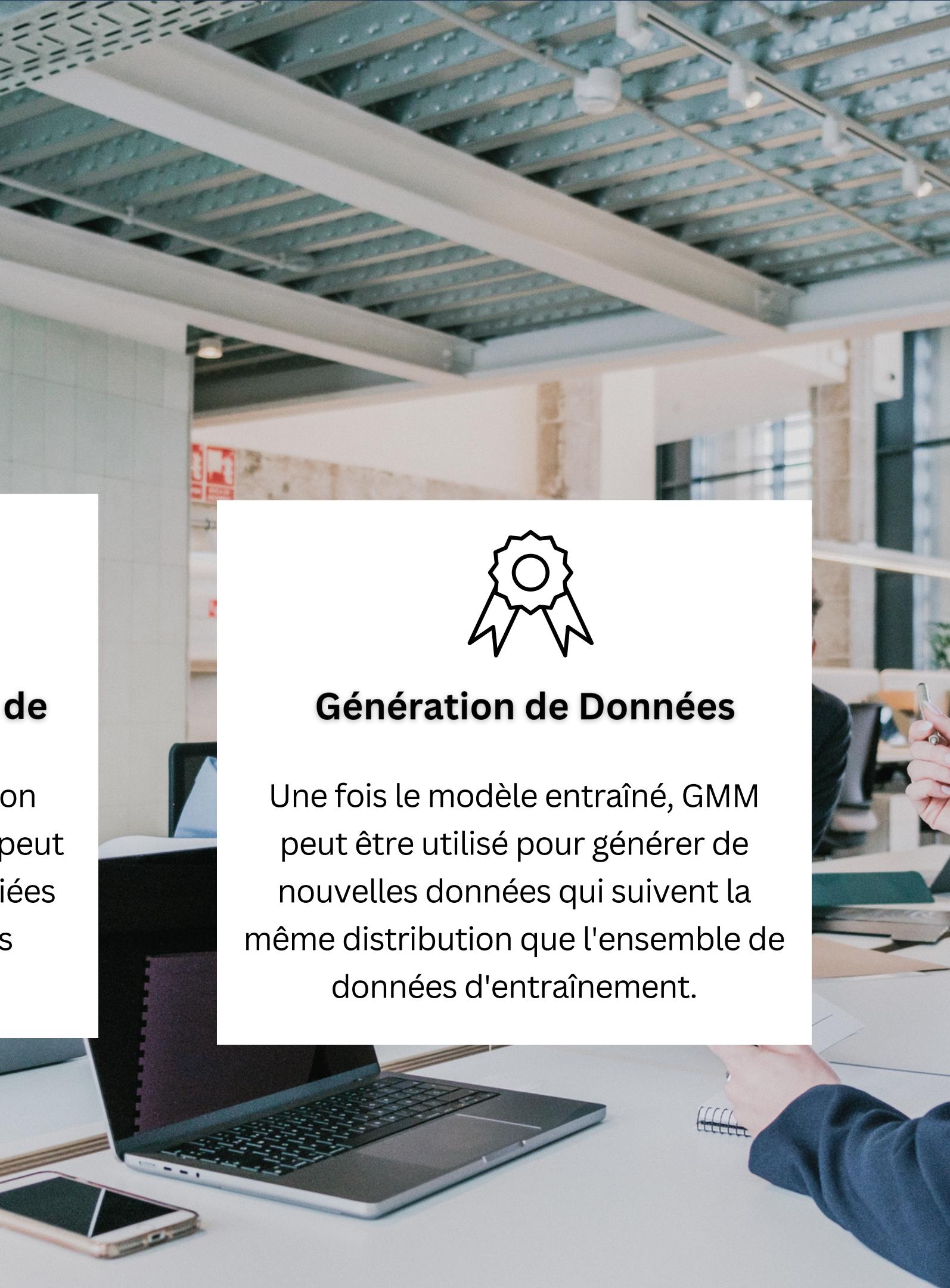
Robustesse aux Formes de Cluster

En permettant une modélisation probabiliste des clusters, GMM peut gérer des formes de cluster variées et des structures de données complexes.



Génération de Données

Une fois le modèle entraîné, GMM peut être utilisé pour générer de nouvelles données qui suivent la même distribution que l'ensemble de données d'entraînement.



LES INCONVÉNIENTS DE GMM

Sensibilité au Nombre de Composantes

GMM nécessite de spécifier le nombre de composantes gaussiennes, ce qui peut être difficile à déterminer, surtout pour des ensembles de données de grande dimension ou lorsque les clusters sont fortement chevauchants.

Temps de Calcul

L'entraînement de GMM peut être intensif en termes de calcul, en particulier pour des ensembles de données de grande taille ou avec un grand nombre de dimensions, car il implique l'estimation de plusieurs paramètres.

Initialisation Sensible

Les performances de GMM dépendent de l'initialisation des paramètres du modèle, et une mauvaise initialisation peut entraîner des résultats sous-optimaux ou des problèmes de convergence.





MERCI POUR
VOTRE
ATTENTION

