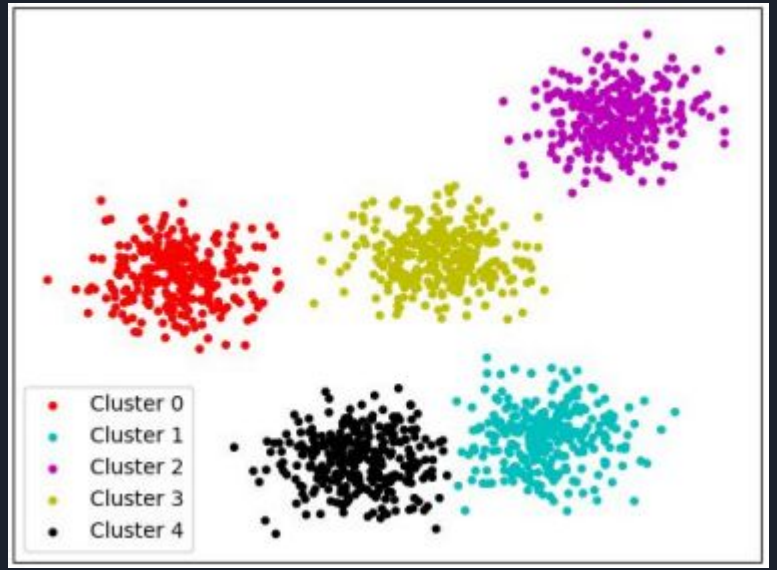
Clustering & Réduction de la dimensionnalité

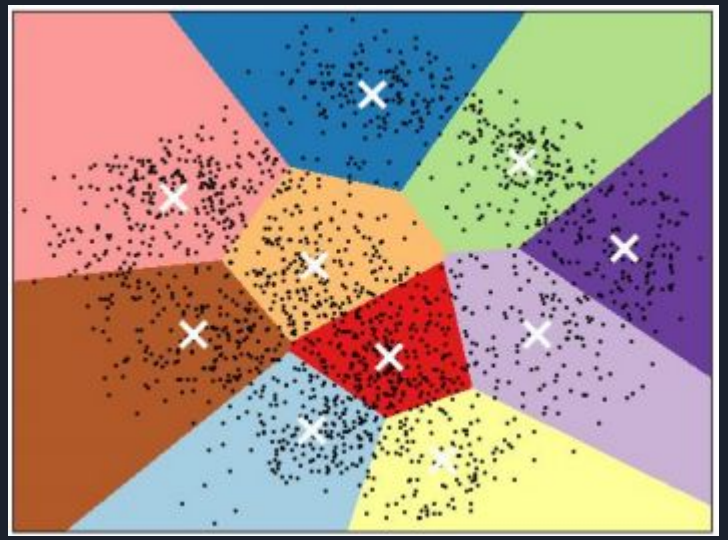
Le clustering

- Méthode d’apprentissage automatique non supervisée
- Plusieurs algorithmes utilisables



Algorithme de Kmeans

Le but est de regrouper les données en essayant de séparer les échantillons en n groupes d'égale variance, minimisant un critère connu sous le nom d'inertie ou somme des carrés intra-cluster. Il nécessite que le nombre de clusters soit spécifié. Il s'adapte bien à un grand nombre d'échantillons et a été utilisé dans une large gamme de domaines d'application dans de nombreux domaines différents.



Réduction de dimensionnalité

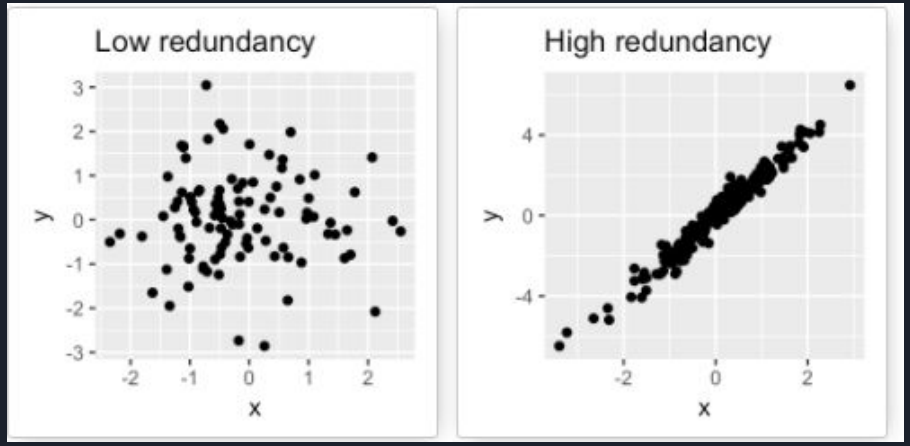
Permet de réduire le nombre de dimension d'un espace à grande dimension en un de plus petite.
Pour se faire, on effectue plusieurs étapes :

- Supprimer des dimensions (ou descripteurs)
- Combiner les variables afin d'obtenir un plus petit nombre de nouvelles variables plus expressives et/ou moins redondantes.

Analyse de composante principale

Méthode statistique qui permet d'explorer des données dites multivariées (données avec plusieurs variables). Chaque variable pourrait être considérée comme une dimension différente.

L'ACP réduit les dimensions d'une donnée multivariée à deux ou trois composantes principales, qui peuvent être visualisées graphiquement, en perdant le moins possible d'information.



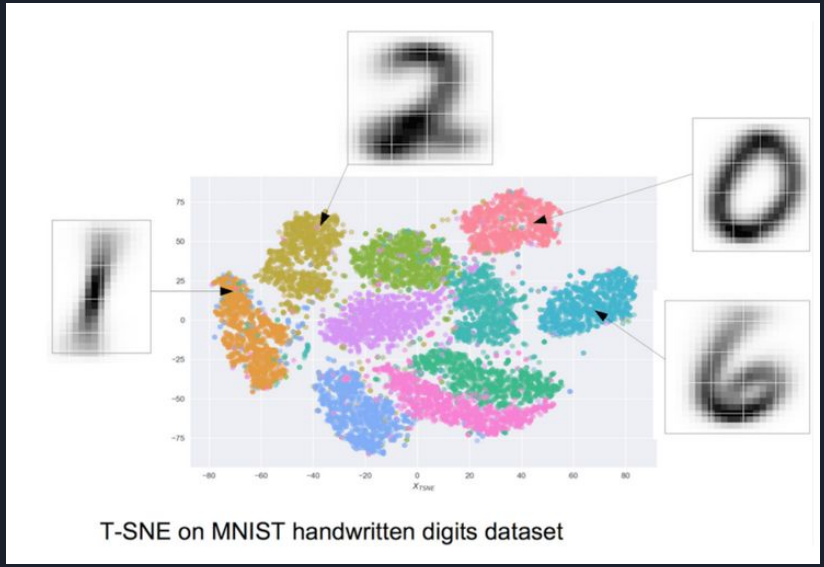
En résumé, l'analyse en composantes principales permet:

- d'identifier des "profils cachés" dans un jeu de données,
- de réduire les dimensions des données en enlevant la redondance des données,
- d'identifier les variables corrélées

t-distributed stochastic neighbor embedding (t-SNE)

T-SNE est un algorithme non-linéaire de “feature extraction” qui construit une nouvelle représentation des données de telle sorte que les données proches dans l’espace original aient une probabilité élevée d’avoir des représentations proches dans le nouvel espace.

Cependant, même si cet algorithme crée une distribution qui respecte la proximité entre les objets les plus proches, la nouvelle représentation ne respecte pas forcément les distances et les densités de distribution des données originales.



T-SNE on MNIST handwritten digits dataset