RER - CLUSTERING

# **Contexte** :

Découvrir l'apprentissage non supervisé au travers les algorithmes **k-means**, **k-medoids** et **DBSCAN**. Analyser avec les métriques d’évaluations.

# **Problématiques** :

1. Comment utiliser clustering dans l’apprentissage non supervisés ?
2. Comment construire des modèles de clustering ?
3. Comment utiliser la technique de la standardisation ?
4. Comment utiliser les métriques d’évaluations ?
5. Comment visualiser le clustering ?
6. Comment mesurer la similarité ?
7. Comment avoir un nombre de clusters optimal ?

# **Mots clés :**

* Apprentissage non-supervisé : branche de l'apprentissage automatique (ou machine learning) où l'algorithme cherche à découvrir des structures ou des modèles dans les données sans l'aide d'une étiquette ou d'un signal de sortie connu à l'avance. Contrairement à l'apprentissage supervisé, il n'y a pas de données d'entraînement pré-annotées ou de réponse attendue pour guider le processus d'apprentissage.
* Clustering : Le clustering, également appelé regroupement ou partitionnement de données, est une technique d'apprentissage non-supervisé qui consiste à diviser un ensemble de données en groupes (ou clusters) homogènes en termes de caractéristiques ou de similarité.

L'objectif du clustering est de trouver une structure sous-jacente dans les données qui permet de les organiser en groupes significatifs.

* K-means : algorithme de clustering non supervisé qui permet de regrouper des données en K groupes distincts en utilisant une approche basée sur les centres de gravité. L'algorithme fonctionne en cherchant à minimiser la variance intra-cluster, c'est-à-dire la somme des distances entre chaque point de données et le centre de gravité de son groupe.
* K-medoids : algorithme de clustering qui permet de regrouper des données en K groupes distincts en utilisant une approche basée sur les médoides. Les médoides sont des points de données qui représentent le centre de gravité d'un groupe de données. Contrairement aux centres de gravité, qui sont calculés comme la moyenne des valeurs de chaque dimension, les médoides correspondent à une valeur réelle de l'ensemble de données.
* DBSCAN : DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering non supervisé qui permet de regrouper des données en groupes denses en fonction de leur densité dans l'espace des données. L'algorithme fonctionne en cherchant des zones de l'espace des données où la densité des points est élevée, et en créant des groupes de points en fonction de ces zones.
* Similarité : En apprentissage automatique, la similarité fait référence à une mesure de ressemblance entre deux objets ou deux ensembles de données. Elle est souvent utilisée pour comparer les caractéristiques ou les attributs de ces objets et pour évaluer à quel point ils sont similaires ou différents.
* Distances :
  + Euclidienne : La distance euclidienne est une mesure de la distance entre deux points dans un espace euclidien à n dimensions. Elle est souvent utilisée en apprentissage automatique pour mesurer la similitude entre deux exemples ou deux ensembles de données.
  + Manhattan : La distance de Manhattan, également appelée distance de la ville ou distance L1, est une mesure de la distance entre deux points dans un espace à n dimensions. Elle est souvent utilisée en apprentissage automatique pour mesurer la similitude entre deux exemples ou deux ensembles de données.
* Méthode Elbow : méthode du coude, ou méthode Elbow en anglais, est une technique couramment utilisée en apprentissage automatique pour déterminer le nombre optimal de clusters dans une tâche de clustering. Cette méthode est basée sur l'observation du changement de variance expliquée par le modèle en fonction du nombre de clusters.
* Silhouette : La méthode de Silhouette est une technique d'évaluation de la qualité des clusters obtenus lors d'une tâche de clustering en apprentissage automatique. Elle permet de mesurer à quel point chaque objet dans un cluster est similaire aux autres objets de son propre cluster et à quel point il est différent des objets des autres clusters.
* Algorithme de réduction de dimension : Un algorithme de réduction de dimension est une méthode en apprentissage automatique qui permet de réduire le nombre de variables ou de dimensions dans un jeu de données, tout en conservant autant que possible l'information pertinente. L'objectif est de simplifier la représentation des données tout en minimisant la perte d'information.
* Standardisation : La standardisation est un processus en statistiques et en apprentissage automatique qui permet de transformer les données en échelle commune, de manière à ce qu'elles aient une moyenne de 0 et un écart-type de 1. Cette transformation est également appelée normalisation.

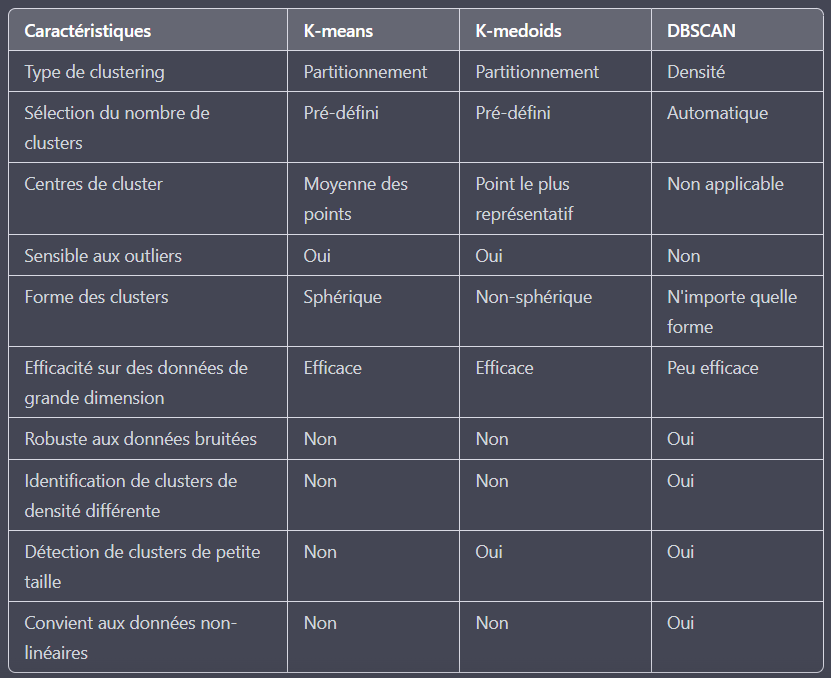
La standardisation est souvent utilisée pour prétraiter les données avant de les utiliser dans des algorithmes de machine learning, car elle peut améliorer les performances de ces algorithmes en leur permettant de mieux traiter les données. Elle peut également faciliter la comparaison des données provenant de différentes sources, qui peuvent avoir des échelles de mesure différentes.

* Ground Truth : Le terme "Ground Truth" (vérité terrain en français) fait référence à une vérité ou référence absolue dans un domaine donné, contre laquelle les résultats obtenus par une méthode ou un algorithme peuvent être comparés. Dans les domaines de l'apprentissage automatique, de la vision par ordinateur, de la classification, de la segmentation d'image ou de l'analyse de texte, le Ground Truth peut être utilisé comme une référence pour évaluer la précision d'un algorithme ou d'une méthode.
* Rand Index : Le Rand Index est une mesure de similarité entre deux ensembles de données ou partitions. Il est couramment utilisé en analyse de regroupement (clustering) pour évaluer l'accord entre différents algorithmes de clustering ou pour comparer une solution de clustering à une classification connue. Le Rand Index varie de 0 à 1, où 0 indique un accord nul entre les deux partitions et 1 indique un accord parfait.
* Adjusted Rand Index : L'Adjusted Rand Index (ARI) est une mesure de similarité entre deux ensembles de données ou partitions. Il est utilisé pour évaluer la qualité d'un algorithme de clustering ou pour comparer une solution de clustering à une classification connue.
* AMI (Adjusted Mutual Information) : L'Adjusted Mutual Information (AMI) est une mesure de similarité entre deux ensembles de données ou partitions. Elle est utilisée pour évaluer la qualité d'un algorithme de clustering ou pour comparer une solution de clustering à une classification connue.
* Inertie : En machine learning, l'inertie est une mesure de qualité pour évaluer la performance d'un algorithme de clustering. L'inertie est définie comme la somme des distances au carré entre chaque point et le centre de son cluster.
* Maximal Distance to Cluster Center : Maximal Distance to Cluster Center est une métrique de clustering qui mesure la distance maximale entre chaque point et le centroïde de son cluster. Cette mesure est utilisée pour évaluer la qualité d'un algorithme de clustering en mesurant la séparation entre les clusters.
* Average Distance to Other Center : L'Average Distance to Other Center est une métrique de clustering qui mesure la distance moyenne entre les centroïdes des clusters. Cette mesure est utilisée pour évaluer la qualité d'un algorithme de clustering en mesurant la séparation entre les clusters.
* Average Distance to Cluster Center : L'Average Distance to Cluster Center est une métrique de clustering qui mesure la distance moyenne entre chaque point et le centroïde de son cluster. Cette mesure est utilisée pour évaluer la qualité d'un algorithme de clustering en mesurant la cohérence des clusters.

# **Hypothèses :**

1. L’utilisation de l’algorithme k-means est plus simple que l’utilisations de k-medoids (Osman)
2. La standardisation s’apparente à la normalisation (Jean Paul)
3. On est obligé d’utiliser Rand index dans clustering (Tetyana)
4. Le clustering implique d’utiliser les mesures de dispersions (Étienne)
5. La distance Euclidienne est la mesure la plus utilisée dans clustering (Adeline)
6. Optimiser le nombre de cluster permet d’éviter l’over-fiting (Briand)
7. L’entropie est utilisée pour la mise en place des clusters (Seydou)
8. Plus l’entropie est faible plus l’algorithme de clustering est efficace (Axel)
9. Le clustering non-supervisé est l’unique méthode dans la classification non-supervisé (Loïc)
10. On ne peut pas utiliser le clustering pour des données avec target (Nicolas)
11. L’inertie est lié à Average Distance to Cluster Center (Aude)
12. La silhouette est une forme de dispersion des données (Solenn)
13. On utilise les métriques de classification pour faire du clustering (Adrien)

# Plan d'action :

* Explorer les ressources
* Définir et comprendre les mots clés
* Répondre aux problématiques
* Vérifier les hypothèses
* Comparaison entre k-means, k-medoids et DBSCAN.
* Faire les Workshops
* Rendu du RER