# RAPPORT: Projet Machine Learning



# SOMMAIRE

[SOMMAIRE 2](#_Toc1076106509)

[Jeu de Données : 3](#_Toc1583203233)

[Tache de Classification : 3](#_Toc1111221910)

[Préparation des données : 3](#_Toc46815812)

[Algorithmes de Classification: 3](#_Toc1971968501)

[Arbre de Décision ; 4](#_Toc509059278)

[Foret Aleatoire : 4](#_Toc1177366008)

[AdaBoost : 4](#_Toc795009728)

[Machine a Vecteurs de Support (SVM) 4](#_Toc106605647)

[K-Nearest Neighbors (KNN) 4](#_Toc1392375718)

[Naive Bayes : 4](#_Toc1483923934)

[Taches de Clustering 4](#_Toc13321918)

[Chargement des données et Visualisation : 5](#_Toc1470183981)

[Algorithmes de Clustering : 5](#_Toc912965590)

[K-Means Clustering : 6](#_Toc1991149879)

[Clustering Hiérarchique Agglomératif : 6](#_Toc1352635986)

[Clustering DBSCAN : 6](#_Toc1006076553)

[Clustering HDBSCAN : 6](#_Toc1866532225)

[Analyse Comparative : 6](#_Toc1952417144)

[Analyse de la classification 7](#_Toc464856850)

[Analyse du clustering 7](#_Toc1144493778)

## Jeu de Données :

Explication du jeu de données

## Tache de Classification :

### Préparation des données :

### Algorithmes de Classification:

#### Arbre de Décision ;

#### Foret Aleatoire :

#### AdaBoost :

#### Machine a Vecteurs de Support (SVM)

#### K-Nearest Neighbors (KNN)

#### Naive Bayes :

## Taches de Clustering

### Chargement des données et Visualisation :

### Algorithmes de Clustering :

*1. K-means :*

*- Fonctionnement: Divise l'ensemble de données en K clusters en minimisant la variance intra-cluster.*

*- Avantages :*

*- Simple et rapide.*

*- Efficace pour des ensembles de données de taille moyenne à grande avec des clusters de forme sphérique.*

*- Limitations :*

*- Sensible à l'initialisation des centroïdes et aux valeurs aberrantes.*

*- Nécessite de spécifier le nombre de clusters à l'avance.*

*- Applicabilité: Convient aux données avec des clusters de forme sphérique et de taille égale.*

*2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) :*

*- Fonctionnement: Identifie les clusters en se basant sur la densité des points.*

*- Avantages :*

*- Peut détecter des clusters de formes arbitraires et de densités variables.*

*- Ne nécessite pas de spécifier le nombre de clusters à l'avance.*

*- Limitations:*

*- Sensible aux valeurs des paramètres `eps` et `min\_samples`.*

*- Peut avoir du mal avec des clusters de densités différentes.*

*- Applicabilité : Utile pour détecter des clusters de formes et de densités variables.*

*3. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) :*

*- Fonctionnement : Extension de DBSCAN qui utilise une approche hiérarchique.*

*- Avantages :*

*- Peut identifier des clusters de formes et de densités variées.*

*- Robuste aux valeurs des paramètres.*

*- Limitations :*

*- Peut être plus lent que DBSCAN en raison de sa complexité.*

*- Nécessite plus de mémoire pour stocker la structure hiérarchique.*

*- Applicabilité: Recommandé lorsque le nombre de clusters n'est pas connu à l'avance et pour des données complexes avec des clusters de formes et de densités variées.*

*4. Clustering hiérarchique agglomératif :*

*- Fonctionnement: Construit une hiérarchie de clusters en fusionnant successivement les clusters les plus proches.*

*- Avantages :*

*- Peut capturer la structure hiérarchique intrinsèque des données.*

*- Ne nécessite pas de spécifier le nombre de clusters à l'avance.*

*- Limitations:*

*- Coûteux en termes de temps de calcul, surtout pour de grands ensembles de données.*

*- Sensible aux choix de la mesure de distance et de la méthode de liaison.*

*- Applicabilité: Utile pour explorer la structure hiérarchique des données sans spécifier le nombre de clusters à l'avance.*

*En résumé, le choix de l'algorithme de clustering dépendra de la structure des données, de la nature des clusters que vous recherchez et des performances spécifiques que vous recherchez en termes de temps de calcul et de précision du clustering. Il est souvent judicieux d'expérimenter plusieurs algorithmes et d'évaluer leurs performances sur votre ensemble de données spécifique pour trouver le meilleur choix.*

## Analyse Comparative :

### Analyse de la classification

### Analyse du clustering