

Prosit - 1 UE 1 : On sépare les appartements

Projet - Intelligence Artificielle (Machine Learning)

22 / 01 / 2025| Data preprocessing

# Mots clefs / Mots à définir :

* Jeu de données
* Secteur de recensement
* Exploration de données
* Préparation des données
* Imputation
* Statistiques descriptives
* Caractéristiques catégorielles
* Apprentissage automatique
* Modèle (Apprentissage automatique)
* Analyse de données
* Distribution de données
* K-Means

# Contexte :

Une agence immobilière a pour souhait d’utiliser l’intelligence artificielle afin d’améliorer ses propositions immobilières et d’optimiser la charge de travail lié aux visites en réduisant leur nombre.

# Problématique :

* Comment mettre en valeur les données de recensement brutes de 1990 fourni, afin de classifier les logements de Californie.

# Contraintes :

* Jeu de donnés fourni

# Livrables :

* Dataset clean
* Notebook :
  + Visualisation des données
  + K-mean clusterisation

# Généralisation :

* Data preparation
* Data visualization

# Pistes de solutions :

* Utiliser un notebook jupyter
* Utilisation de librairies pour la data preparation (pandas, seaborn ..)
* Suivre différentes phases de la data preparation :
  + Loading
  + Explorating
  + Cleaning
  + Analyzing
* Utilisation de librairie pour la data visualization (matplotlib, folium, dash …)

## Plan d’action :

### PARTIE THÉORIQUE :

1. Prendre en main les différents concepts de la data preparation.
2. Se renseigner sur les librairies les plus couramment utilisés pour ces tâches.
3. Prendre conscience des éléments permettant de valider un dataset et une data visualization.

### PARTIE PRATIQUE :

1. Réaliser un notebook jupyter pour les phases de data preparation et visualization.

# Réalisation du Plan d’action :

## Partie Théorique :

### Définitions des mots clés

**Jeu de données**  
Ensemble structuré d’observations ou d’exemples utilisés pour entraîner et évaluer des modèles d’apprentissage automatique. Peut comprendre des variables numériques, catégorielles ou textuelles, et est souvent divisé en sous-parties (entraînement, validation, test).

**Secteur de recensement**  
Zone géographique définie pour la collecte et l’analyse de données démographiques ou statistiques. On l’utilise souvent pour segmenter les données et mieux comprendre les distributions dans un contexte territorial.

**Exploration de données**  
Processus initial d’analyse visant à découvrir rapidement la structure et les caractéristiques d’un jeu de données. Comprend des résumés statistiques, des visualisations et la détection de valeurs aberrantes pour guider la suite du projet.

**Préparation des données**  
Étape cruciale consistant à nettoyer, transformer et formater les données pour qu’elles soient cohérentes et exploitables par un modèle d’apprentissage automatique. Peut inclure la gestion des valeurs manquantes, le codage des variables et le redimensionnement des caractéristiques.

**Imputation**  
Méthode pour remplacer les valeurs manquantes par des estimations pertinentes (moyenne, médiane, modèle prédictif, etc.). Réduit l’impact des données incomplètes sur la performance des modèles.

**Statistiques descriptives**  
Ensemble de mesures permettant de résumer et d’interpréter rapidement un jeu de données (moyenne, écart-type, quartiles, etc.). Sert souvent de point de départ pour comprendre la distribution et la variabilité des variables.

**Caractéristiques catégorielles**  
Variables qualitatives qui prennent un nombre fini de valeurs discrètes (couleur, type, catégorie). Leur traitement demande souvent du codage spécifique (one-hot encoding, label encoding) pour les intégrer aux modèles.

**Apprentissage automatique**  
Ensemble de techniques et d’algorithmes permettant à un système informatique d’apprendre automatiquement à partir de données. L’objectif est de généraliser et de faire des prédictions ou des classifications sur de nouvelles informations.

**Modèle (Apprentissage automatique)**  
Représentation mathématique ou algorithmique issue de l’apprentissage sur un jeu de données. Permet de réaliser des prédictions ou des décisions basées sur de nouvelles données en reproduisant les tendances détectées.

**Analyse de données**  
Processus global d’examen, de nettoyage, de transformation et de modélisation des données afin d’en extraire des informations utiles et d’appuyer la prise de décision. Englobe l’exploration, la préparation et l’application des méthodes analytiques.

**Distribution de données**  
Répartition des valeurs d’une variable dans un jeu de données. Se visualise et se décrit souvent avec des histogrammes ou des mesures comme la moyenne et l’écart-type pour comprendre sa forme et sa tendance centrale.

**K-Means**  
Algorithme de regroupement non supervisé qui partitionne les données en *k* clusters. Il assigne chaque point au cluster dont le centre (centroïde) est le plus proche, puis réajuste itérativement les centroïdes pour obtenir des groupes cohérents.

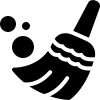
**\***

#### THEORIQUE :

La data préparation en quelques étapes :

Nous pouvons résumer la data preprocessing à l’aide des étapes suivantes :

* A black background with a black square

  Description automatically generated with medium confidence**Collecte et vérification des données** : Identifier les sources et rassembler les données brutes, puis valider leur qualité (formats, cohérence).
* **Nettoyage** : Gérer les valeurs manquantes (imputation ou suppression), corriger les incohérences et éliminer les doublons.
* A black background with a black square

  Description automatically generated with medium confidenceA black background with a black square

  Description automatically generated with medium confidence**Transformation** : Convertir les variables (encodage de variables qualitatives, standardisation ou normalisation), générer de nouvelles caractéristiques si nécessaire.
* **Sélection de variables** : Retenir les variables pertinentes, éliminer celles peu informatives ou redondantes.
* A black background with a black square

  Description automatically generated with medium confidence**Division en ensembles** : Séparer en ensembles d’entraînement, de validation et de test pour évaluer les performances du futur modèle.

Le machine learning :

Introduction — Le machine Learning

L'apprentissage automatique, ou **machine learning**, est une branche de l'intelligence artificielle qui permet aux systèmes informatiques **d'apprendre et de s'améliorer à partir de données**, sans être explicitement programmés. Il existe principalement deux types d'apprentissage en machine learning : **l'apprentissage supervisé et l'apprentissage non supervisé**. Chacun de ces types a des caractéristiques, des méthodes et des applications spécifiques.

Partie 1 — L’apprentissage supervisé

L'apprentissage supervisé utilise des ensembles de données **étiquetées** pour entraîner un modèle. Chaque donnée d'entrée est **associée** à une sortie correspondante connue. L'objectif est que le modèle apprenne la **relation entre les entrées et les sorties** afin de pouvoir **prédire** correctement les sorties pour de nouvelles entrées.

**Exemples d'applications :**

* **Classification** : Attribuer des catégories à des données, comme distinguer des emails en "spam" ou "non spam".
* **Régression** : Prédire des valeurs continues, comme estimer le prix d'un bien immobilier en fonction de ses caractéristiques.

**Algorithmes couramment utilisés :**

* Régression linéaire
* Machines à vecteurs de support (SVM)
* Arbres de décision
* Réseaux de neurones

Partie 2 — L’apprentissage non-supervisé

L'apprentissage non supervisé travaille avec des données non étiquetées. Le modèle cherche à **identifier des structures ou des motifs cachés** dans les données **sans guidance préalable**. L'objectif est de découvrir des relations ou des groupements naturels au sein des données.

**Exemples d'applications :**

* **Clustering (regroupement)** : Segmenter des clients en groupes similaires basés sur leurs comportements d'achat.
* **Réduction de dimensionnalité** : Simplifier les jeux de données en réduisant le nombre de variables tout en conservant l'essentiel de l'information.

**Algorithmes couramment utilisés :**

* K-means
* Analyse en composantes principales (PCA)
* Clustering hiérarchique

Partie 3 — Principales différences entre apprentissage supervisé et non supervisé

**Données d'entrée** :

* *Supervisé* : Utilise des données étiquetées avec des sorties connues.
* *Non supervisé* : Travaille avec des données non étiquetées sans sorties préalablement définies.

**Objectif** :

* *Supervisé* : Prédire des sorties spécifiques pour de nouvelles entrées basées sur l'apprentissage des données étiquetées.
* *Non supervisé* : Découvrir des structures ou des motifs cachés dans les données sans guidance externe.

**Complexité** :

* *Supervisé* : Nécessite un processus d'étiquetage des données, souvent coûteux en temps et en ressources.
* *Non supervisé* : Peut-être plus complexe en termes de calcul, car le modèle doit identifier des structures sans guidance préalable.

**Précision des résultats** :

* *Supervisé* : Tend à fournir des résultats plus précis grâce à l'utilisation de données étiquetées.
* *Non supervisé* : Les résultats peuvent être moins précis et nécessitent souvent une validation humaine.

#### NB – Exemple Supervisé / Non supervisé

#### Machines à Vecteurs de Support (SVM)

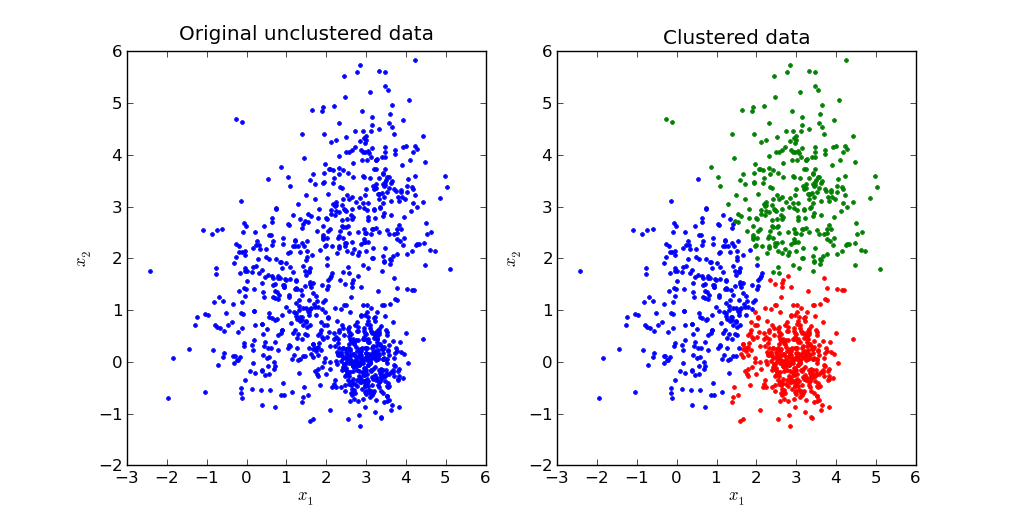
Les SVM sont des algorithmes d'apprentissage supervisé utilisés principalement pour des tâches de classification et de régression. Leur objectif est de trouver un hyperplan optimal qui sépare les différentes classes de données avec la plus grande marge possible. Cette marge est définie comme la distance entre l'hyperplan et les points de données les plus proches de chaque classe, appelés *vecteurs de support*. En maximisant cette marge, les SVM cherchent à améliorer la capacité de généralisation du modèle sur des données non vues.

#### Analyse en Composantes Principales (PCA)

L'ACP est une méthode statistique d'apprentissage non supervisé utilisée pour réduire la dimensionnalité des jeux de données tout en conservant le maximum d'information possible. Elle transforme des variables potentiellement corrélées en un ensemble de nouvelles variables décorrélées, appelées *composantes principales*. Ces composantes capturent la majeure partie de la variance présente dans les données initiales.

Partie 4 — L’algorithme K-Means

L'algorithme **K-means** est une méthode de *clustering* non supervisée largement utilisée pour partitionner un ensemble de données en *K* groupes distincts, appelés *clusters*. Chaque cluster est représenté par un centroïde, qui correspond à la moyenne des points de données qui lui sont assignés. L'objectif principal de K-means est de minimiser la variance au sein de chaque cluster, c'est-à-dire de réduire la somme des distances au carré entre les points de données et leur centroïde respectif.



**Fonctionnement de l'algorithme K-means**

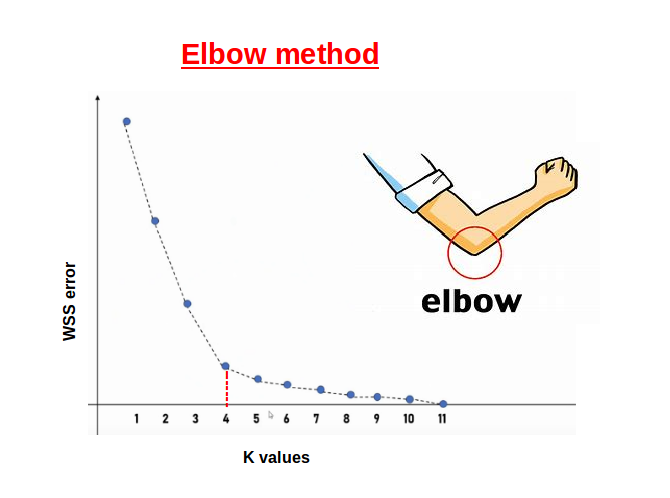
L'algorithme K-means suit généralement les étapes suivantes :

1. **Initialisation** : Choisir **aléatoirement *K* points** de données à partir de l'ensemble de données comme centroïdes initiaux.
2. **Affectation des points aux clusters** : Pour chaque point de données, calculer la **distance entre ce point et chaque centroïde**, puis assigner le point au cluster dont le centroïde est le plus proche.
3. **Mise à jour des centroïdes** : Après l'affectation de tous les points, **recalculer la position de chaque centroïde** en prenant la moyenne des points qui lui sont assignés.
4. **Itération** : Répéter les étapes 2 et 3 **jusqu'à ce que les centroïdes ne changent plus** de position de manière significative, indiquant que l'algorithme a convergé.

Il est important de noter que l'algorithme K-means peut **converger vers des solutions locales**, et le résultat final **peut dépendre des positions initiales des centroïdes**. Pour améliorer la robustesse des résultats, il est courant d'exécuter l'algorithme plusieurs fois avec des initialisations différentes et **de sélectionner la partition présentant le meilleur critère de qualité**.

**Choix du nombre de clusters (*K*)**

Déterminer le nombre optimal de clusters *K* est une étape cruciale dans l'application de l'algorithme K-means. Plusieurs méthodes peuvent être utilisées pour estimer ce nombre :

* **Méthode du coude (*Elbow method*)** : Cette méthode consiste à exécuter l'algorithme K-means pour différents nombres de clusters et à tracer la somme des distances au carré intra-cluster (inertie) en fonction de *K*. Le point où l'inertie commence à diminuer moins rapidement (formant un "coude" dans le graphique) est souvent choisi comme le nombre optimal de clusters.
* **Indice de silhouette** : Cet indice mesure la cohésion et la séparation des clusters. Une valeur élevée de l'indice de silhouette indique que les points sont bien regroupés au sein de leurs clusters et bien séparés des autres clusters.
* **Statistique du gap** : Cette méthode compare la performance de l'algorithme K-means sur les données réelles à sa performance sur des données aléatoires, permettant de déterminer si la structure de clustering observée est significative.

**Limitations de l'algorithme K-means**

Bien que K-means soit un algorithme populaire et efficace, il présente certaines limitations :

* **Sensibilité aux valeurs initiales** : Les résultats peuvent varier en fonction des centroïdes initiaux choisis. Des techniques comme K-means++ ont été développées pour améliorer l'initialisation des centroïdes.
* **Forme des clusters** : K-means suppose que les clusters ont une forme sphérique et de taille similaire. Il peut donc échouer à identifier des clusters de formes ou de densités variées.
* **Nombre de clusters fixé à l'avance** : L'algorithme nécessite de spécifier le nombre de clusters *K* avant l'exécution, ce qui peut être difficile sans connaissance préalable des données.

**Applications de l'algorithme K-means**

K-means est utilisé dans divers domaines, notamment :

* **Segmentation de la clientèle** : Identifier des groupes de clients ayant des comportements ou des caractéristiques similaires pour des stratégies de marketing ciblées.
* **Compression d'images** : Réduire le nombre de couleurs dans une image en regroupant des pixels similaires, ce qui permet de diminuer la taille des fichiers sans perte significative de qualité visuelle.
* **Analyse de données biologiques** : Classer des gènes ou des protéines en fonction de leurs expressions ou de leurs fonctions similaires.

En somme, l'algorithme K-means est un outil puissant pour le clustering, mais son application nécessite une attention particulière aux choix des paramètres et une compréhension de ses limitations pour garantir des résultats pertinents.

#### PRATIQUE :

NOTEBOOK’s

WORKSHOP