esprit
Se former autrement
HONORIS UNITED UNIVERSITIES

Chapitre 2:

PRÉTRAITEMENT DES DONNÉES

Principes fondamentaux du Machine Learning

Objectifs

L'objectif de ce chapitre est de montrer l'importance du prétraitement des données dans un projet de Machine Learning. Nous allons :

- Expliquer les concepts de base liés aux données et leur nécessité de préparation.
- Présenter deux techniques essentielles de prétraitement :

Objectifs

L'objectif de ce chapitre est de montrer l'importance du prétraitement des données dans un projet de Machine Learning. Nous allons :

- Expliquer les concepts de base liés aux données et leur nécessité de préparation.
- Présenter deux techniques essentielles de prétraitement :
  - **Nettoyage des données** : Comment identifier et gérer les valeurs manquantes, les valeurs aberrantes et les erreurs.

Objectifs

L'objectif de ce chapitre est de montrer l'importance du prétraitement des données dans un projet de Machine Learning. Nous allons :

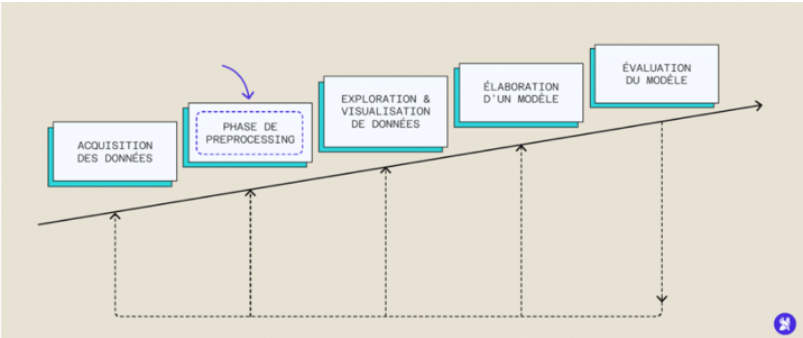
- Expliquer les concepts de base liés aux données et leur nécessité de préparation.
- Présenter deux techniques essentielles de prétraitement :
  - **Nettoyage des données** : Comment identifier et gérer les valeurs manquantes, les valeurs aberrantes et les erreurs.
  - **Transformation des données** : Comment ajuster les données pour qu'elles soient prêtes à être analysées.

Plan

1. Introduction
2. Notions de Bases
3. Les techniques de prétraitement des données
   1. Data cleaning
   2. Data transformation

1. Introduction

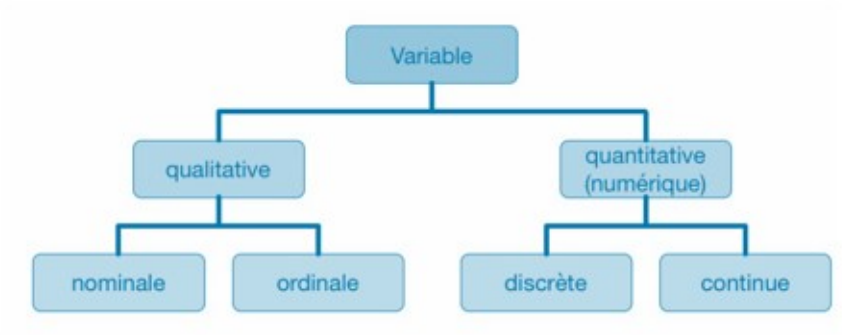
L'apprentissage automatique nécessite de grandes quantités de données , mais les données brutes provenant de diverses sources (audio, vidéo, texte, etc.) ne sont pas directement exploitables. Le prétraitement des données est une étape cruciale qui consiste à nettoyer, transformer et formater ces données afin qu'elles puissent être utilisées efficacement par les algorithmes. Bien que cette phase soit essentielle pour la réussite du projet, elle est souvent la plus longue et la plus complexe.



2. Notions de base

Types de variables

Les variables présentes dans le dataset peuvent être de différents types :



- **Les variables qualitatives** (ou variables catégorielles) qui expriment une caractéristique :
  - **Ordinales** (avec une hiérarchie ou un ordre) :
    - « un peu, beaucoup », « passionnément, à la folie »
    - « primaire, secondaire, universitaire »
    - « insatisfait, neutre, satisfait »
  - **Nominales (sans ordre)** :
    - « femme, homme »
    - « bleu, vert, marron »
    - « Tunisie, Algérie, Maroc »
- **Les variables quantitatives** qui ont une valeur numérique :
  - **Discrètes** : seulement certaines valeurs sont possibles (souvent des entiers).
  - **Continues** : n'importe quelle valeur est possible dans un intervalle.

Exemple: Pour les exemples suivants, pensez-vous que ce sont des variables **qualitatives**, **quantitatives discrètes** ou **quantitatives continues** ?

- Le nombre d'appels reçus par un centre d'appels en une heure :

Exemple: Pour les exemples suivants, pensez-vous que ce sont des variables **qualitatives**, **quantitatives discrètes** ou **quantitatives continues** ?

- Le nombre d'appels reçus par un centre d'appels en une heure : **Quantitative discrète**
- Poids :

Exemple: Pour les exemples suivants, pensez-vous que ce sont des variables **qualitatives**, **quantitatives discrètes** ou **quantitatives continues** ?

- Le nombre d'appels reçus par un centre d'appels en une heure : **Quantitative discrète**
- Poids : **Quantitative continue**
- Lancer d'un dé :

Exemple: Pour les exemples suivants, pensez-vous que ce sont des variables **qualitatives**, **quantitatives discrètes** ou **quantitatives continues** ?

- Le nombre d'appels reçus par un centre d'appels en une heure : **Quantitative discrète**
- Poids : **Quantitative continue**
- Lancer d'un dé : **Quantitative discrète**
- Quantité de chocolat consommée en une année :

Exemple: Pour les exemples suivants, pensez-vous que ce sont des variables **qualitatives**, **quantitatives discrètes** ou **quantitatives continues** ?

- Le nombre d'appels reçus par un centre d'appels en une heure : **Quantitative discrète**
- Poids : **Quantitative continue**
- Lancer d'un dé : **Quantitative discrète**
- Quantité de chocolat consommée en une année : **Quantitative continue**
- Être gaucher ou droitier :

Exemple: Pour les exemples suivants, pensez-vous que ce sont des variables **qualitatives**, **quantitatives discrètes** ou **quantitatives continues** ?

- Le nombre d'appels reçus par un centre d'appels en une heure : **Quantitative discrète**
- Poids : **Quantitative continue**
- Lancer d'un dé : **Quantitative discrète**
- Quantité de chocolat consommée en une année : **Quantitative continue**
- Être gaucher ou droitier : **Qualitative nominale**

3. Les techniques de prétraitement des données

Data Cleaning

Gestion des valeurs manquantes

Traitement des valeurs aberrantes

Suppression des doublons

Data Reduction

Réduire la taille de l'ensemble de données tout en préservant les informations importantes

Data Transformation

Normalisation

Standardisation

Encodage des variables

Feature Engineering

Créer des nouvelles variables à partir des données disponibles

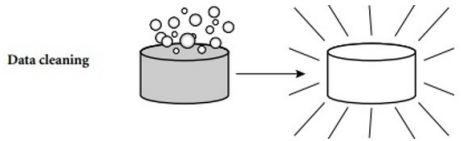
Feature selection

Identifier les variables les plus pertinentes

Dans ce chapitre, on se limitera au **Data Cleaning** et au **Data Transformation**

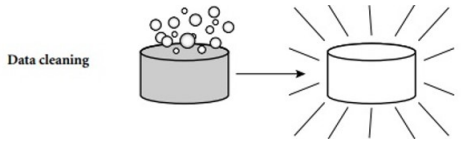
Data Cleaning

Le **nettoyage des données** (ou data cleaning) est une étape cruciale dans le processus de prétraitement des données, visant à améliorer la qualité des données avant de les utiliser dans des analyses ou des modèles d'apprentissage automatique.



Data Cleaning

Le **nettoyage des données** (ou data cleaning) est une étape cruciale dans le processus de prétraitement des données, visant à améliorer la qualité des données avant de les utiliser dans des analyses ou des modèles d'apprentissage automatique.



Les principales techniques associées au nettoyage des données :

- **Gestion des valeurs manquantes**
- **Suppression des doublons**
- **Traitement des valeurs aberrantes** (également appelées *outliers*)

Gestion des valeurs manquantes:

Les valeurs manquantes font référence aux données qui n'ont pas été collectées ou qui sont absentes dans un ensemble de données. Cela peut se produire pour diverses raisons, comme un participant qui n'a pas répondu à une question, des erreurs de mesure, ou des échantillons non disponibles.

La gestion des valeurs manquantes est essentielle dans l'analyse des données. Deux méthodes couramment utilisées sont la suppression et l'imputation des valeurs manquantes.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

Valeurs manquantes

Gestion des valeurs manquantes: Suppression

La suppression des valeurs manquantes consiste à retirer les observations ou les caractéristiques contenant des données incomplètes. Deux approches principales :

Gestion des valeurs manquantes: Suppression

La suppression des valeurs manquantes consiste à retirer les observations ou les caractéristiques contenant des données incomplètes. Deux approches principales :

Suppression par ligne :

- Retire toute ligne où au moins une valeur est manquante.
- Idéal si le nombre de valeurs manquantes est faible.

Gestion des valeurs manquantes: Suppression

La suppression des valeurs manquantes consiste à retirer les observations ou les caractéristiques contenant des données incomplètes. Deux approches principales :

Suppression par ligne :

- Retire toute ligne où au moins une valeur est manquante.
- Idéal si le nombre de valeurs manquantes est faible.

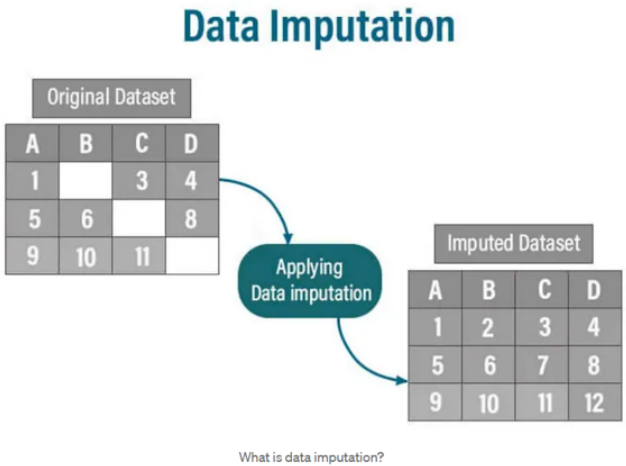
Suppression par colonne :

- Élimine les colonnes entières si une proportion significative de données est manquante (ex. $> 30\%$).
- Utile pour supprimer les caractéristiques peu informatives.

Avantage : Méthode simple et rapide.
Inconvénient : Peut entraîner une perte de données précieuses.

Gestion des valeurs manquantes: Imputation

L'imputation consiste à remplacer les valeurs manquantes par des valeurs estimées ou prédites fondées sur les données disponibles



Méthodes d'imputation	Moyenne	Médiane	Mode	Autres méthodes
Définition	Remplacer les valeurs manquantes par la moyenne des valeurs présentes dans la même colonne. ⚠️ Sensible aux valeurs aberrantes !!!	Les valeurs manquantes sont remplacées par la médiane de la colonne, une méthode plus robuste que la moyenne en présence de valeurs aberrantes.	Les valeurs manquantes sont remplacées par la valeur la plus fréquente dans la colonne. ⚠️ Sensible aux valeurs aberrantes !!!	Imputation par des algorithmes de machine Learning, exemple: méthode des k plus proches voisins (KNN)
Formule	$\frac{1}{n}\sum_{i=1}^{n} x_i$ avec x_i : les valeurs non manquantes présentes dans la même colonne, n: nombre des valeurs non manquantes	Si n est impaire alors $x_{(\frac{n+1}{2})}$ Si n est paire alors $(x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)})/2$ *avec* x_i : la valeur non manquantes d'ordre i		
Type de variables	Variables numériques continues	- Variables numériques continues - Variables ordinales (transformées d'abord en valeurs numériques)	- Variables qualitatives (s'applique directement à ces variables sans conversion en valeurs numériques).	

Exemple:

➢ Imputation par mode

Satisfaction
Bon
Mauvais
Moyen
NaN
Bon
NaN
Bon
Mauvais

Satisfaction
Bon
Mauvais
Moyen
Bon
Bon
Bon
Bon
Mauvais

➢ Imputation par median:

Âge
25
30
NaN
28
35
NaN

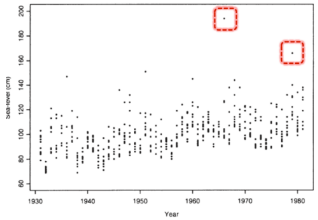
Âge
25
30
29
28
35
29

Remarque: Pour calculer la médiane, les valeurs non manquantes doivent être triées dans un ordre croissant

Gestion des valeurs aberrantes

Qu'est-ce qu'une valeur aberrante (ou extrême) ?

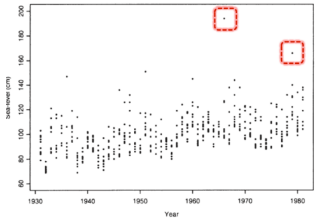
Une valeur extrême (outlier) est une observation "qui semble dévier notoirement des autres éléments de l'échantillon auquel elle appartient" (Grubbs, 1969)



Gestion des valeurs aberrantes

Qu'est-ce qu'une valeur aberrante (ou extrême) ?

Une valeur extrême (outlier) est une observation "qui semble dévier notoirement des autres éléments de l'échantillon auquel elle appartient" (Grubbs, 1969)



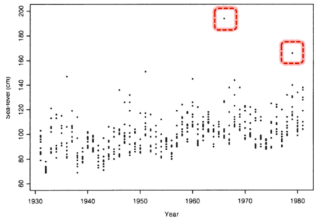
Quelles sont les causes des valeurs extrême?

Les valeurs extrêmes peuvent être des erreurs humaines (ex: erreurs de saisie), erreurs d'instrument (ex: erreurs de mesure), erreurs de traitement des données (ex: manipulation des données) et erreurs d'échantillonnage (ex: extraction des données de mauvaises sources.)

Gestion des valeurs aberrantes

Qu'est-ce qu'une valeur aberrante (ou extrême) ?

Une valeur extrême (outlier) est une observation "qui semble dévier notoirement des autres éléments de l'échantillon auquel elle appartient" (Grubbs, 1969)



Quelles sont les causes des valeurs extrême?

Les valeurs extrêmes peuvent être des erreurs humaines (ex: erreurs de saisie), erreurs d'instrument (ex: erreurs de mesure), erreurs de traitement des données (ex: manipulation des données) et erreurs d'échantillonnage (ex: extraction des données de mauvaises sources.)

Solution $\rightarrow$ Supprimer ou imputer selon le cas

Les valeurs extrêmes peuvent être légitimes:

- **Variations naturelles** : Ex. Revenus très élevés (personne très riche).
- **Cas rares mais réels** : Ex. Phénomènes météorologiques extrêmes.

Solution: Conserver, mais utiliser des méthodes comme la transformation logarithmique pour réduire leur impact sur certaines analyses.

Exemple : Revenus avant (1000, 2000, 10 000) $\rightarrow$ après ln(x): (6.9, 7.6, 9.2)

Les valeurs extrêmes peuvent être légitimes:

- **Variations naturelles** : Ex. Revenus très élevés (personne très riche).
- **Cas rares mais réels** : Ex. Phénomènes météorologiques extrêmes.

Solution: Conserver, mais utiliser des méthodes comme la transformation logarithmique pour réduire leur impact sur certaines analyses.

Exemple : Revenus avant (1000, 2000, 10 000) $\rightarrow$ après ln(x): (6.9, 7.6, 9.2)
Remarque: Une valeur extrême n'est pas automatiquement une erreur. Il est important d'examiner le contexte de vos données pour déterminer si une valeur aberrante est une observation légitime ou le résultat d'une erreur

Exemple

Données brutes (non triées):

$$5, 7, 8, 12, 15, 16, 18, 22, 24, 30, 100$$

Étape 1 : Trier les données en ordre croissant

$\Rightarrow$ 5, 7, 8, 12, 15, 16, 18, 22, 24, 30, 100

Étape 2 : Calcul des quartiles

$\Rightarrow$ 1er quartile (Q1) = 8 ; 2ème quartile (Q2) = médiane = 16 ; 3ème quartile (Q3) = 24

Étape 3 : Calcul de l'intervalle interquartile (IQR) $\Rightarrow$ IQR = Q3 - Q1 = 24 - 8 = 16

Étape 4: Détection des valeurs aberrantes

Borne inférieure :

$\Rightarrow$ Q1 - 1.5 $\times$ IQR = 8 - (1.5 $\times$ 16) = 8 - 24 = -16

$\Rightarrow$ Toute valeur **inférieure à -16** est une valeur aberrante. (Aucune ici)

Borne supérieure :

$\Rightarrow$ Q3 + 1.5 $\times$ IQR = 24 + (1.5 $\times$ 16) = 24 + 24 = 48

$\Rightarrow$ Toute valeur **supérieure à 48** est une valeur aberrante.

Ici on a 100 > 48 $\Rightarrow$ Valeur aberrante détectée.

Data Transformation

Définition:
La transformation des données (Data transformation) est un processus essentiel qui permet de préparer des données brutes pour l'analyse. Cela inclut la modification de la structure et du format des données afin de les rendre exploitables.
Différentes techniques de transformation des données peuvent être appliquées aux variables quantitatives et qualitatives :

1. Pour les variables quantitatives:
   Normalisation
   Standardisation
2. Pour les variables qualitatives:
   One-Hot Encoding
   Label Encoding

Normalisation:

Ajuste les valeurs pour qu'elles se situent sur une échelle commune entre 0 et 1. Cette méthode est souvent utilisée dans les algorithmes de Machine Learning qui reposent sur des distances.

Formule:

$$\text{Valeur normalisée} = \frac{X - x\,min}{x\,max - x\,min}$$

Exemple:

Âge
15
22
35
40
60

Valeur Normalisée
0
0.156
0.444
0.556
1

min =15

max = 60

Valeur normalisée de l'âge 22 : $\frac{22-15}{60-15} = 0{,}156$

Standarisation:

Transforme les données pour qu'elles aient une moyenne de 0 et un écart-type de 1, ce qui permet de centrer et d'homogénéiser les données.

Formule: Valeur standardisée$= \frac{X-\mu}{\sigma}$, avec μ la moyenne et σ l'écart-type.

Salaire
30000
45000
60000
75000
90000

Valeur Standardisée (Z)
-1.41
-0.7
0
0.7
1,41

$\mu = 60000, \quad \sigma = 21213.2$

Valeur standardisée pour le salaire 30000 :

$$\frac{30000 - 60000}{21213.2} = -1.41$$

Rappel : Lécart-type σ d'une population est calculé comme suit :

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

avec N: Nombre total d'observations dans la population, x_i: valeurs individuelles et μ: moyenne de la population.

Encodage: "One-Hot-Encoding"

Transforme chaque catégorie d'une variable en une nouvelle colonne binaire. Chaque colonne représente une catégorie, avec des valeurs de 1 (présence de la catégorie) ou 0 (absence de la catégorie).

Idéal pour les variables nominales.

Exemple:

Couleur
Rouge
Vert
Bleu
Rouge
Bleu

One-Hot Encoding

Couleur_Rouge	Couleur_Vert	Couleur_Bleu
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1

Encodage: "Label Encoding"

Consiste à convertir chaque catégorie d'une variable catégorielle en un nombre entier unique.

Idéal pour les variables ordinales.

- **Exemple:**

Catégorie	Label Encoding
Insatisfait	0
Satisfait	1
Très Satisfait	2

Insatisfait (0) : Indique le niveau le plus bas de satisfaction.

Satisfait (1) : Indique un niveau de satisfaction intermédiaire.

Très Satisfait (2) : Indique le niveau le plus élevé de satisfaction.

Remarque: Pour les variables nominales, le label encoding n'est généralement pas recommandé, car il peut introduire une interprétation erronée des données. Les variables nominales sont des catégories sans ordre naturel.