Faculté d'informatique
Departement IA & SD

DATA MINING

TP1: Préparation des données

BENADJAL.N
Phd student in computer science
LAB: MOVEP
Email: nesrinebenadjel63@gmail.com

INGSOFT4

2025/2026

Préparation des données

Pas de données de qualité, pas de résultats de qualité.

Dans le cadre de notre premier TP, nous allons nous concentrer sur la préparation des données.

1-Chargement des données :

Pour charge un ensemble de données, on peut utiliser la fonction **read_csv().** Mais pour pouvoir l'utiliser, il faut d'abord importer la bibliothèque **Pandas.**
Pour affiche les 5 premières lignes, on peut utiliser la fonction **head().**

Préparation des données

2-Compréhension des Données:

Pour obtenir un résumé des colonnes, de leurs types de données (int, object, float) et du nombre de valeurs non-nulles, on peut utiliser la fonction **info().**
Fournir les statistiques de base pour les colonnes numériques (moyenne, min, max), on peut utiliser la fonction **describe().**

3-Nettoyage des Données:

Pour identifier et supprimer les lignes de données identiques, on peut utiliser la fonction : **drop_duplicates().**
Compter le nombre de valeurs manquantes (NaN) par colonne, on peut utiliser, **isnull().sum().**

Préparation des données

4-Traitement des Données Manquantes

- Si une ligne contient beaucoup de valeurs manquantes, nous pouvons supprimer la ligne entière **(dropna(axis=0))**.
- Si une colonne donnée contient beaucoup de valeurs manquantes, nous pouvons choisir de supprimer la colonne **(dropna(['collonne1','collonne2'],axis=1)).**
- Remplacement par une valeur arbitraire **(fillna(0)).**
- Remplacement par la moyenne **(fillna(data["collonne3"].mean()).**
- Remplacement par la valeur avant **(fillna(method='ffill')).**
- Remplacement avec la valeur suivante **(fillna(method='bfill')).**

Préparation des données

5-le typage des données:

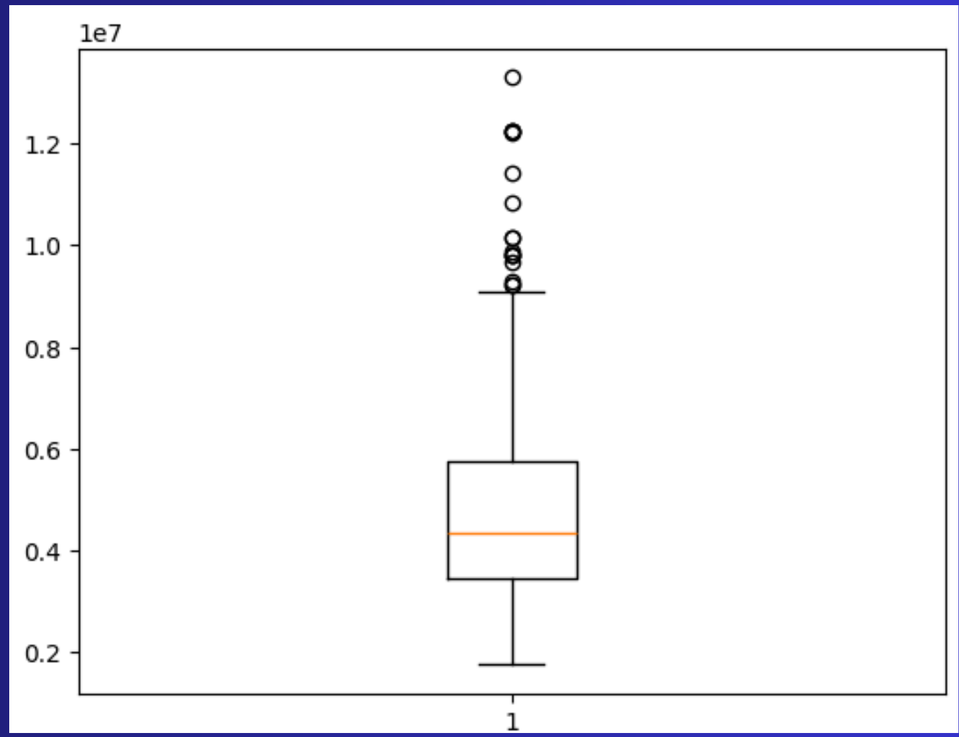
Les algorithmes d'apprentissage automatique travaillent principalement avec des types numériques.
Lorsqu'une variable qui ne contient que des valeurs entières (25.0), est stockée par défaut comme un nombre flottant (float64), cela engendre une surconsommation de mémoire. En effet, la taille d'un float64 est supérieure à celle d'un entier int32 ou int64.
Pr ouréduire la consummation mémoire du jeu de données et améliorer l'efficacité des traitements, on peut utiliser la fonction **astype('int64')**.

Préparation des données

6-Traitement des outliers:

Une outlier est une observation qui se trouve à une distance anormale des autres valeurs d'un échantillon aléatoire d'une population.

Un diagramme en boîte (aussi appelé diagramme à moustaches) est une manière visuelle de représenter la distribution d'un ensemble de données. Il montre les quartiles (et la médiane), et met également en évidence les valeurs aberrantes.



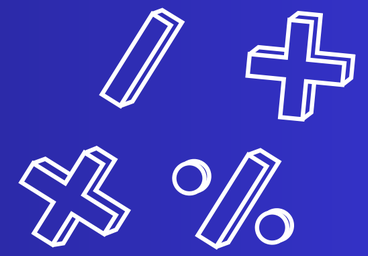
Préparation des données

Stratégies:

Suppression:
Si l'outlier est clairement dû à une erreur de mesure ou de saisie et que la quantité de données supprimées reste faible.

Transformation:
est une opération mathématique appliquée à une variable (colonne) dans le but de changer sa distribution ou sa relation avec d'autres variables, sans altérer l'information relative entre les observations.

Préparation des données



Normalisation (Min-Max Scaling) :

$$X_{new} = (X - X_{min}) / (X_{max} - X_{min})$$

Standardisation (Z-score Scaling) :

$$X_{std} = (X - \mu) / \sigma$$

Transformation des données catégorielles:

consiste à convertir les variables qualitatives (étiquettes, catégories, chaînes de caractères) en un format numérique que les algorithmes d'apprentissage automatique (Machine Learning) peuvent comprendre et traiter.