

## Résumé du cours : Data Mining

### Définition

Processus d'extraction de modèles et de connaissances cachés dans des ensembles de données à l'aide

### Types de données concernés

1. Bases de données relationnelles : Organisées sous forme de tables reliées.
2. Entrepôts de données : Grandes bases pour l'analyse décisionnelle.
3. Bases de données transactionnelles : Gèrent les transactions quotidiennes.
4. Bases de données avancées : Gèrent des données complexes ou non structurées.

### Fonctionnalités principales

1. Description et classification : Identification des caractéristiques des objets.
2. Analyse d'association : Découverte de relations entre données.
3. Classification et prédiction : Classement ou prédiction de valeurs.
4. Analyse de clustering : Regroupement basé sur la similitude.
5. Analyse d'évolution et de déviation : Suivi des changements dans le temps.

### Importance de Python dans la science des données

1. Langage polyvalent : Syntaxe intuitive, accessible à tous niveaux.
2. Écosystème riche : Bibliothèques comme Pandas, NumPy, Scikit-learn, TensorFlow.
3. Communauté active : Large support communautaire.
4. Intégration facile : Compatibilité avec d'autres technologies.
5. Utilisation industrielle : Standard dans le Big Data et l'IA.
6. Puissance et flexibilité : Gestion des données massives et visualisation.

### Pipeline d'exploration de données

1. Collecte des données : Essentielle pour des décisions éclairées.
  - Sources : Fichiers (CSV, Excel), bases de données (SQL, NoSQL), sites web, APIs.
  - Outils nécessaires : Frameworks de scraping, connecteurs de bases, librairies de fichiers.
2. Prétraitement (Preprocessing)
3. Transformation
4. Data Mining
5. Validation / Information
6. Valider l'information / Connaissance cachée

### Collecte de Données avec Python

Sources principales de données :

- Fichiers traditionnels : CSV, TXT, XLSX, JSON, XML, HTML.
- Bases de données : SQL, NoSQL.
- Web : Sites web, APIs.

Outils nécessaires pour la collecte de données :

- Frameworks de scraping web.
- Connecteurs de bases de données.

**- Librairies de manipulation de fichiers.**

**Exemple d'utilisation avec CSV :**

**Utiliser Pandas pour lire un fichier CSV :**

```
import pandas as pd  
df = pd.read_csv('data.csv')  
print(df.head())
```

## **Chapitre 2 : Règles d'association**

L'objectif est de définir une association et de trouver la relation cachée qui existe entre les éléments.

Image 1 : Illustration des règles d'association

Règle d'association:  $X \rightarrow Y$

~~X → Y~~ Voici les mesures des règles d'association:

\* Support:  $S = \frac{\text{nombre de Transactions } X \cup Y}{\text{Total}}$

\* Confiance:  $C(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$

\* Lift:  $L(X \rightarrow Y) = \frac{\text{Confiance}(X \rightarrow Y)}{\text{Support}(Y)}$