

Nom : El Moustepha Med

id: 22179

-Data Mining : Extraction de modèles et connaissances à partir de grandes bases de données.

-Big Data : Enormes ensembles de données caractérisés par le volume, la vitesse et la variété.

-Relation entre Data Mining et Big Data : Big Data fournit les données massives nécessaires pour le Data Mining, qui en extrait des informations utiles.

-Relation entre Data Mining et Analyse des données : Le Data Mining est une partie de l'analyse des données, se concentrant sur la découverte de modèles cachés.

-Relation entre Data Mining et Intelligence artificielle : Le Data Mining utilise des algorithmes d'IA pour extraire des connaissances à partir des données, améliorant les systèmes d'IA.

-Méthodes et techniques de Data Mining pour l'exploration des données : Classification, clustering, association, régression.

Type de modèle

- 1.Description de concepts / classification : Identifier les caractéristiques des objets.
- 2.Analyse d'association : Découvrir les relations entre éléments.
- 3.Classification et prédiction : Catégoriser et prédire.
- 4.Analyse de clustering : Regrouper selon similitude.
- 5.Analyse d'évolution et de déviation : Suivre les changements dans le temps.

Importance de Python en science des données :

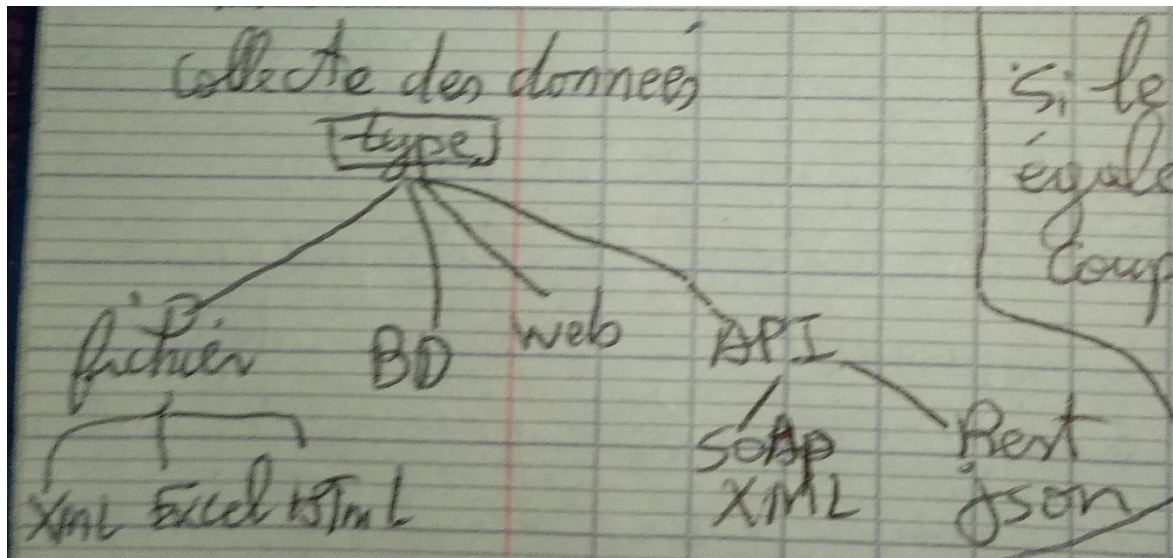
- 1.Polyvalence : Facile à apprendre avec une structure simple.
- 2.Écosystème riche : Bibliothèques comme Pandas, NumPy, Matplotlib.
- 3.Communauté active : Soutien important de la communauté.
- 4.Intégration facile : S'intègre avec d'autres technologies.
- 5.Utilisation industrielle : Norme pour l'analyse de données en entreprise.
- 6.Puissance et flexibilité : Gère efficacement les tâches complexes.

Vue d'ensemble du pipeline d'exploration de données :

1. Sélection : Choix des données pertinentes dans un ensemble de données plus large.
2. Prétraitement : Préparation des données sélectionnées pour l'analyse.
3. Transformation : Conversion des données prétraitées dans un format adapté au Data Mining.
4. Data Mining : Analyse des données transformées pour découvrir des motifs.
5. Interprétation/Évaluation : Interprétation et évaluation des motifs découverts pour extraire des connaissances significatives.

Collecte des données avec Pandas

1. Utilisation de Pandas : Importer des données depuis des fichiers CSV et Excel.
2. Sources de données :



-Fichiers : CSV, TXT, XLSX, JSON, XML, HTML.

-Bases de données : SQL, NoSQL.

-Web : Sites web, APIs.

L'objectif est d'apprendre à utiliser des outils Python pour la collecte efficace des données.

Outils Python pour la collecte de données

1. Pandas : Manipulation et analyse des données. Lecture des fichiers CSV, Excel, SQL.
2. BeautifulSoup : Analyse HTML et XML pour le scraping web.

3. Requests : Récupération de données via APIs et services web.
4. Connecteurs SQL : mysql-connector-python, psycopg2, sqlite3.
5. Yahoo Finance : Extraction de données financières (prix, historiques).

Collecte des données avec Python

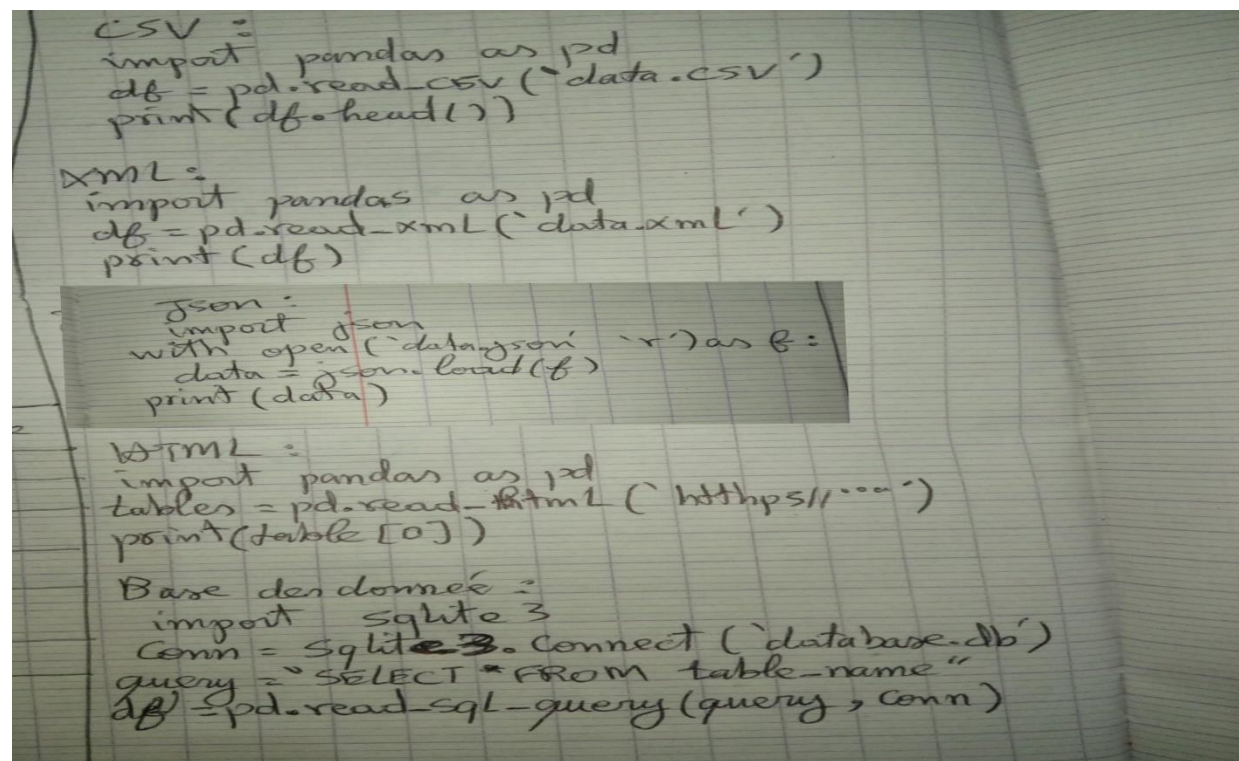
1. Pourquoi utiliser différents formats de fichiers ?

- Assurer la compatibilité avec divers outils et applications.

2. Formats populaires et outils Python associés :

- CSV : Lecture simple et largement supportée (`pandas.read_csv()`).
- TXT : Pour textes simples (`pandas.read_csv()` avec délimiteurs personnalisés).
- XLSX : Pour données complexes avec feuilles multiples (`pandas.read_excel()`).
- JSON : Données structurées, faciles à lire (`pandas.read_json()`).
- XML : Pour données hiérarchiques (`pandas.read_xml()`).
- HTML : Extraction de tableaux web (`pandas.read_html()`).

Exemple d'extraction de données via Python :



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

XML :
import pandas as pd
df = pd.read_xml('data.xml')
print(df)

JSON :
import json
with open('data.json', 'r') as f:
    data = json.load(f)
print(data)

HTML :
import pandas as pd
tables = pd.read_html('https://...')
print(table[0])

Base des données :
import sqlite3
Conn = sqlite3.connect('database.db')
query = "SELECT * FROM table-name"
df = pd.read_sql_query(query, Conn)
```

Règle d'association dans DM :

Règle d'association

- 1- Support = $\frac{\text{nombre de Tx } X}{\text{Somme Transact}}$
- 2- Confiance :

$$C = \frac{\text{Support}(X \cup Y)}{\text{Support } X}$$
- 3- lift :

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

Exemple

Trans	Article
1	pain, lait
2	pain, beurre, œuf, jus
3	lait, jus, beurre, thé
4	pain, lait, jus, beurre
5	pain, lait, jus, thé

$S = \frac{\text{nombre de transaction } X}{\text{Somme Transact}}$
 $C = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$

élément	Trans	Support	Fréquence
pain	1, 2, 4, 5	$\frac{4}{5}$	80%
lait	1, 3, 4, 5	$\frac{4}{5}$	80%
beurre	2, 3, 4	$\frac{3}{5}$	60%
œuf	2	$\frac{1}{5}$	20%
jus	2, 3, 4, 5	$\frac{4}{5}$	80%
thé	3, 5	$\frac{2}{5}$	40%

$\{ \text{lait, pain} \} \rightarrow \text{thé}$
 $S = \frac{2}{5} = 60\%$
 $C = \frac{1}{3} = 33,33\%$
 $\{ \text{pain} \} \rightarrow \text{lait}$
 $S = 80\%$
 $C = \frac{3}{4} = 75\%$

élément	S	C	L
$\{pain\} \rightarrow \{lait\}$	80%	25%	$\frac{76\%}{80\%}$ $= 0,9375$

Si Lift >1 : association positive forte entre les items.

Si Lift =1 : aucune association entre les items.

Si Lift <1 : association négative entre les items.