الجـــمهوريـــة الجزائريـــة الديمـــقراطيـــــة الشـعبيـة
ⵜⴰⵣⴰⵡⴰ ⵜⴰⵎⴰⵔⴰⵢⵜ ⵜⴰⵣⵣⴰⵢⵜ ⵜⴰⵖⵡⵔⵉⵜ
République Algérienne Démocratique et Populaire
وزارة التعليـــم العـــالي والبـــحث العلمـــي
ⴰⵖⵏⵣⵎ ⵏ ⵜⵓⵎⵎⴰⵏ ⵜⵏⵣⵣ ⴰ ⵜⵙⴰⵣ ⵜⵓⵙⵙⵏⵙ
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE

المدرسة الوطنية العليا للإعلام الآلي
ⴰⵖⵓⵔⴰⵙ ⴰⵖⵏⵏⴰⵍ ⵜⵏⵣⵣ ⵏ ⵜⵓⵙⵙⵏⵙⵏⵜ
École nationale Supérieure d'Informatique

Rapport TP BDA

2ème année Cycle Supérieur (2CS)

Option : Systèmes d'Information et Technologies (SIT 2)

Thème :

Pipeline d'Ingestion de Données avec Scikit-learn et Pandas sur des Données Agronomiques

Réalisé par :

- *ABDELKEBIR Achraf*

Proposé par :

- Mme AMROUCHE Karima

Table des matières

Chapitre 1 : Introduction

1.1 Contexte et Objectif du TP

L'objectif de ce travail pratique est de concevoir un pipeline d'ingestion de données en utilisant les bibliothèques **Pandas** et **Scikit-learn**. Ce pipeline a pour rôle de nettoyer, encoder et normaliser des données agronomiques afin de les préparer pour l'entraînement de modèles de machine learning.

Les données utilisées concernent principalement les types de cultures et les quantités d'eau utilisées, avec une distinction entre eau salée et eau douce. Une fois les données préparées, plusieurs modèles d'apprentissage automatique sont entraînés et comparés pour analyser leurs performances respectives.

1.2 Importance du Pipeline d'Ingestion des Données

Dans le domaine de l'apprentissage automatique, la qualité des données joue un rôle fondamental dans la performance des modèles. Un pipeline d'ingestion bien structuré permet de :

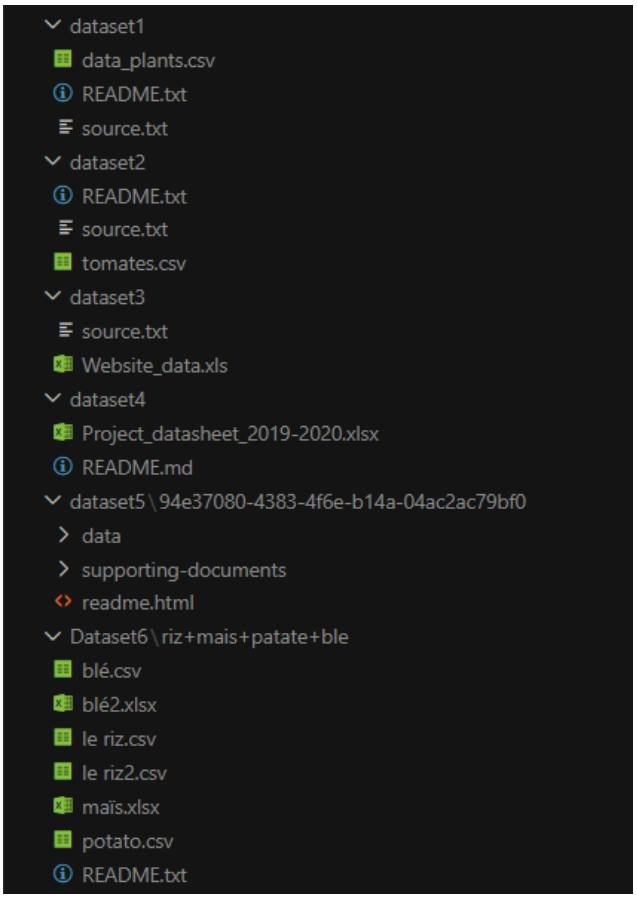
- Automatiser le prétraitement des données,
- Assurer la reproductibilité des transformations appliquées,
- Faciliter l'intégration des données brutes dans un modèle d'apprentissage,
- Réduire le risque d'erreurs manuelles et améliorer l'efficacité globale du processus.

1.3 Description des Outils Utilisés

Pour la réalisation de ce travail, plusieurs outils ont été utilisés :

- **Python** : Langage de programmation principal pour le développement du pipeline et l'entraînement des modèles.
- **Jupyter Notebook** : Environnement interactif pour l'exploration et la manipulation des données.
- **VS Code** : Éditeur de code utilisé pour structurer et organiser le projet.
- **Bibliothèques Python (Pandas, Scikit-learn)** : Utilisées pour le traitement et la préparation des données ainsi que pour l'entraînement des modèles.
- **Excel** : Employé pour l'inspection et la vérification initiale des données.

Chapitre 2 : Exploration des Données



2.1 Dataset 1

Données sur les besoins en eau des cultures dans différentes conditions.

2.1.1. Structure générale

- Total de colonnes: 15
- Variables numériques: 10 (water req, Min Temp, Max Temp, Humidity, Wind, Sun, Rad, Rain, altitude, latitude, longitude)
- Variables catégorielles: 4 (month, crop, soil, city)

2.1.2. Colonnes:

- `water req` (float): Besoin en eau de la culture
- `month` (string): Mois de l'année
- `Min Temp` (float): Température minimale
- `Max Temp` (float): Température maximale
- `Humidity` (float): Taux d'humidité

- `Wind` (float): Vitesse du vent
- `Sun` (float): Ensoleillement
- `Rad` (float): Radiation
- `Rain` (float): Précipitations
- `altitude` (float): Altitude du lieu
- `latitude` (float): Latitude du lieu
- `longitude` (float): Longitude du lieu
- `crop` (string): Type de culture (potato, wheat, rice)
- `soil` (string): Type de sol (red loamy, red sandy, BLACK CLAY)
- `city` (string): Ville

2.1.3. Problèmes de Qualité des Données

1. Valeurs Manquantes :

- Certaines colonnes comme **month, crop, soil, city** contiennent des valeurs vides.
- La colonne **water req** est parfois absente sur certaines lignes.
- Certaines valeurs météorologiques (ex. Min Temp, Max Temp, Humidity, Wind, etc.) sont parfois absentes.

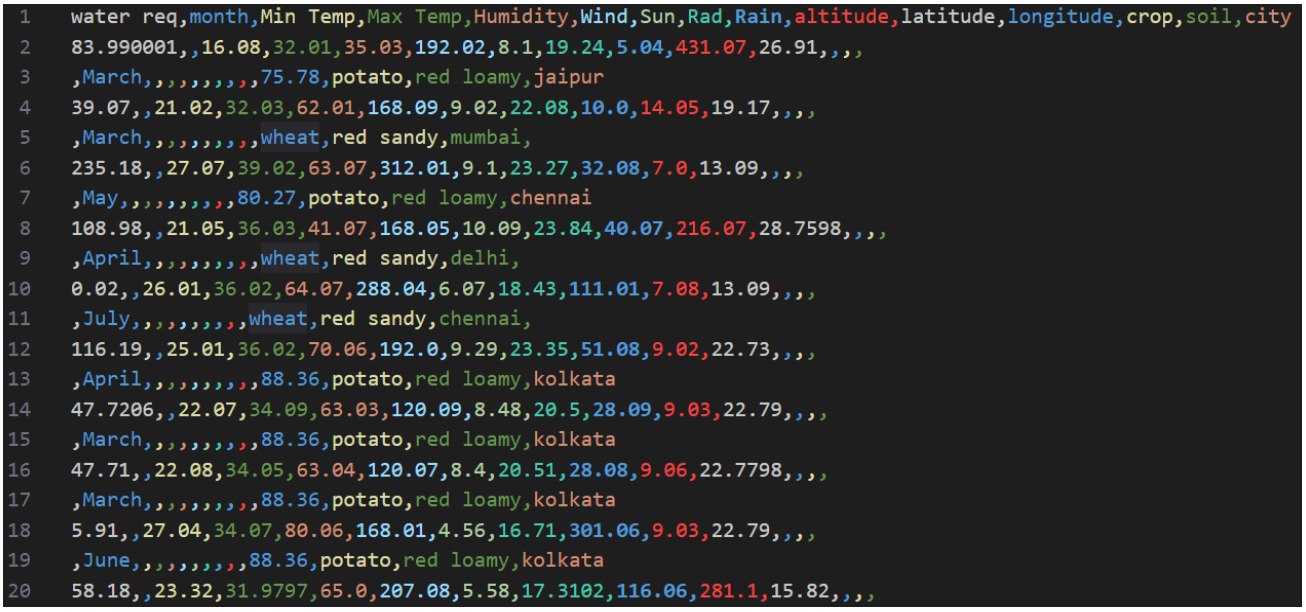
2. Incohérences :

- **Format des nombres décimaux** :
  - Certaines valeurs numériques utilisent un nombre variable de décimales (ex. *73.08* vs *73.0802*).
  - Certaines valeurs incluent des erreurs d'arrondi ou des approximations.
- **Casse des valeurs catégorielles** :
  - Les noms de mois et de villes ont des différences de casse ("may" vs "May", "july" vs "July").
  - Les types de sols ne sont pas toujours écrits de manière homogène ("BLACK CLAY" vs "Black Clay").

3. Structure de Données Mal Formée :

- Chaque enregistrement est réparti sur **deux lignes** au lieu d'une seule.
- Les colonnes de mesures sont sur la première ligne, et les colonnes catégorielles sont sur la seconde.
- Cette séparation empêche une lecture et un traitement direct du fichier.
- Certaines lignes contenant `"wheat"` dans la colonne `longitude` nécessitent un décalage vers la droite pour positionner correctement les valeurs catégorielles.

data_plants.csv:



```
water req,month,Min Temp,Max Temp,Humidity,Wind,Sun,Rad,Rain,altitude,latitude,longitude,crop,soil,city
83.990001,,16.08,32.01,35.03,192.02,8.1,19.24,5.04,431.07,26.91,,,,
,March,,,,,,,,,75.78,potato,red loamy,jaipur
39.07,,21.02,32.03,62.01,168.09,9.02,22.08,10.0,14.05,19.17,,,,
,March,,,,,,,,wheat,red sandy,mumbai,
235.18,,27.07,39.02,63.07,312.01,9.1,23.27,32.08,7.0,13.09,,,,
,May,,,,,,,,80.27,potato,red loamy,chennai
108.98,,21.05,36.03,41.07,168.05,10.09,23.84,40.07,216.07,28.7598,,,,
,April,,,,,,,,wheat,red sandy,delhi,
0.02,,26.01,36.02,64.07,288.04,6.07,18.43,111.01,7.08,13.09,,,,
,July,,,,,,,,wheat,red sandy,chennai,
116.19,,25.01,36.02,70.06,192.0,9.29,23.35,51.08,9.02,22.73,,,,
,April,,,,,,,,88.36,potato,red loamy,kolkata
47.7206,,22.07,34.09,63.03,120.09,8.48,20.5,28.09,9.03,22.79,,,,
,March,,,,,,,,88.36,potato,red loamy,kolkata
47.71,,22.08,34.05,63.04,120.07,8.4,20.51,28.08,9.06,22.7798,,,,
,March,,,,,,,,88.36,potato,red loamy,kolkata
5.91,,27.04,34.07,80.06,168.01,4.56,16.71,301.06,9.03,22.79,,,,
,June,,,,,,,,88.36,potato,red loamy,kolkata
58.18,,23.32,31.9797,65.0,207.08,5.58,17.3102,116.06,281.1,15.82,,,,
```

2.2 Dataset 2

Données sur les besoins en eau des plants de tomates durant une période de simulation.

2.2.1. Structure générale

- **Total de colonnes**: 4
- **Variables numériques**: 3 (*simulation_id, time, water, hour*)
- **Variables catégorielles**: 0 (toutes les colonnes sont numériques)

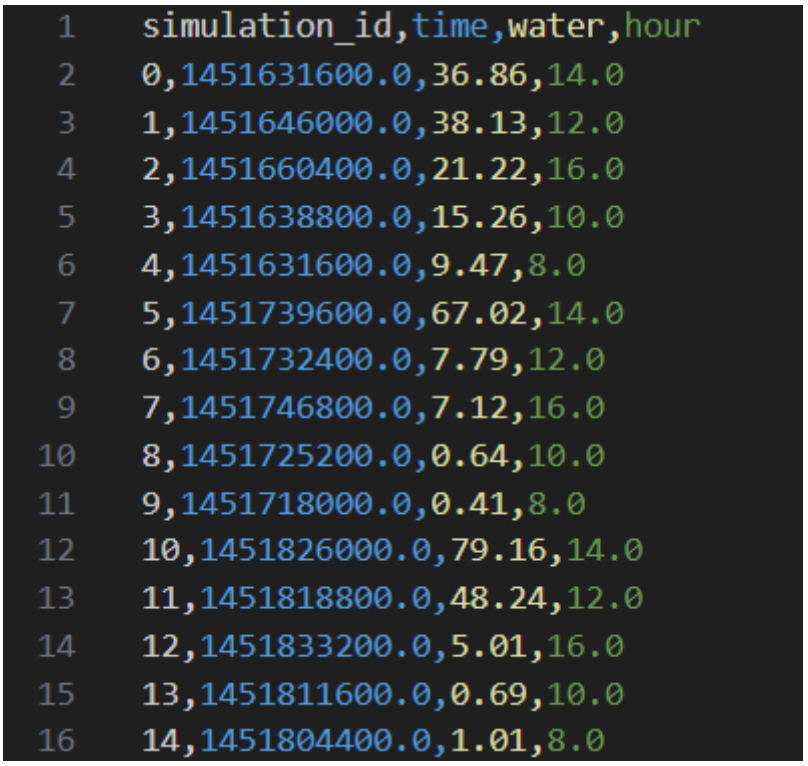
2.2.2. Colonnes

- **simulation_id (int)**: Identifiant unique de la simulation.
- **time (float)**: Horodatage Unix indiquant le moment de la collecte des données.
- **water (float)**: Quantité d'eau consommée (en unités de mesure non spécifiées).
- **hour (float)**: Heure de la journée à laquelle l'échantillon a été prélevé.

2.2.3. Problèmes de Qualité des Données

1. **Incohérences** :
    - **Format des timestamps** : La colonne *time* est en format Unix Timestamp et nécessite une conversion en format lisible (ex. *YYYY-MM-DD HH:MM:SS*).
2. **Structure de Données Mal Formée** :
    - **Manque d'informations contextuelles** : Absence d'attributs supplémentaires comme la température, l'humidité ou le type de sol, qui pourraient être utiles pour l'analyse des besoins en eau.
    - **Heure en variable numérique** : La colonne *hour* pourrait être catégorisée (matin, après-midi, soir).

tomates.csv:



```
simulation_id,time,water,hour
0,1451631600.0,36.86,14.0
1,1451646000.0,38.13,12.0
2,1451660400.0,21.22,16.0
3,1451638800.0,15.26,10.0
4,1451631600.0,9.47,8.0
5,1451739600.0,67.02,14.0
6,1451732400.0,7.79,12.0
7,1451746800.0,7.12,16.0
8,1451725200.0,0.64,10.0
9,1451718000.0,0.41,8.0
10,1451826000.0,79.16,14.0
11,1451818800.0,48.24,12.0
12,1451833200.0,5.01,16.0
13,1451811600.0,0.69,10.0
14,1451804400.0,1.01,8.0
```

2.3 Dataset 3

Données sur la période de croissance et les besoins en eau des cultures.

2.3.1. Structure générale

- **Total de colonnes**: 3
- **Variables numériques**: 2 (*Total growing period (days), Crop water need (mm/total growing period)*)
- **Variables catégorielles**: 1 (*Crop*)

2.3.2. Colonnes

- **Crop (string)**: Nom de la culture agricole.
- **Total growing period (days) (range)**: Durée de la période de croissance en jours (exprimée sous forme de plage).
- **Crop water need (mm/total growing period) (range)**: Besoins en eau totaux sur la période de croissance (exprimés sous forme de plage).

2.3.3. Problèmes de Qualité des Données

1. **Valeurs Manquantes** :
    - Certaines lignes ne contiennent pas de valeurs pour *Crop water need* (ex. *Peanut/Groundnut*).

- Des cultures comme *Peanut/Groundnut, Pea, Pepper, Potato, Radish...* ont un *Total growing period* mais pas de *Crop water need*.

2. **Incohérences** :
   - **Colonnes dupliquées** : La colonne *Crop* est répétée deux fois (besoin de fusionner les données correctement).
   - **Format des valeurs numériques** : Les plages (*ex. 100-365 jours pour Alfalfa, 800-1600 mm pour Alfalfa*) nécessitent un traitement spécial (extraction des valeurs min/max, calcul d'une moyenne).
   - **Alignement incorrect des cultures** : Certains noms de cultures sont mal alignés avec les besoins en eau (ex. *Bean dry* semble être aligné avec *Cabbage*).
3. **Structure de Données Mal Formée** :
   - Les cultures sont réparties sur deux colonnes différentes, nécessitant un regroupement correct des informations.
   - Les plages de valeurs doivent être séparées en valeurs minimales et maximales pour un meilleur traitement des données.

Website_data.xls:

Crop	Total growing period (days)	Crop	Crop water need (mm/total growing period)
Alfalfa	100-365	Alfalfa	800-1600
Banana	300-365	Banana	1200-2200
Barley/Oats/Whe	120-150	Barley/Oats/Wheat	450-650
Bean green	75-90	Bean	300-500
Bean dry	95-110	Cabbage	350-500
Cabbage	120-140	Citrus	900-1200
Carrot	100-150	Cotton	700-1300
Citrus	240-365	Maize	500-800
Cotton	180-195	Melon	400-600
Cucumber	105-130	Onion	350-550
Eggplant	130-140	Peanut	500-700
Flax	150-220	Pea	350-500
Grain/small	150-165.	Pepper	600-900
Lentil	150-170	Potato	500-700
Lettuce	75-140	Rice (paddy)	450-700
Maize sweet	80-110	Sorghum/Millet	450-650

2.4 Dataset 4

Données sur l'irrigation des cultures pour la période 2019-2020.

2.4.1. Structure générale

- **Total de colonnes** : 7
- **Variables numériques** : 7
  (CropType, CropDays, Soil Moisture, Soil Temperature, Temperature, Humidity, Irrigation (Y/N))

- **Variables catégorielles** : 0
  (Remarque : Les codes présents dans CropType et Irrigation (Y/N) représentent des informations catégorielles via des codes numériques.)

2.4.2. Colonnes

- **CropType (int)** : Code représentant le type de culture.
  Mapping : Paddy = 1, Ground Nuts = 2
- **CropDays (int)** : Nombre de jours de croissance de la culture.
- **Soil Moisture (float)** : Taux d'humidité du sol.
- **Soil Temperature (float)** : Température du sol en °C.
- **Temperature (float)** : Température ambiante en °C.
- **Humidity (float)** : Taux d'humidité de l'air (en %).
- **Irrigation (Y/N) (int)** : Indicateur d'irrigation (1 pour Oui, 0 pour Non).
  Mapping Irrigation : Yes (Y) = 1, No (N) = 0

2.4.3. Problèmes de qualité des données

Structure de données mal formée

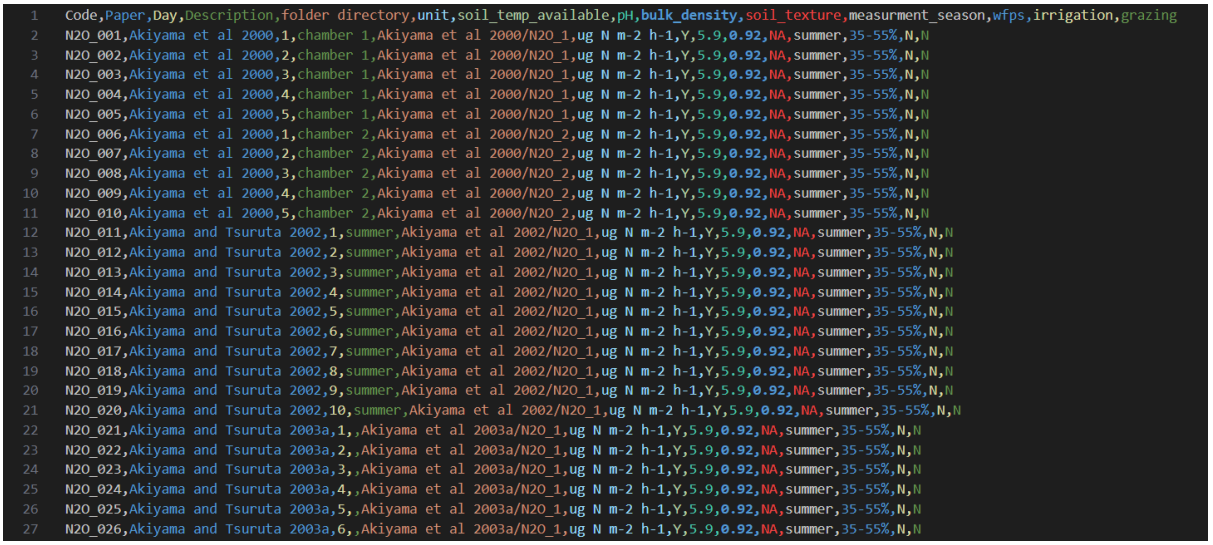
Le fichier Excel contient, en plus des données principales, des lignes d'en-tête supplémentaires (mappings pour *CropType* et *Irrigation*) qui doivent être filtrées pour obtenir un jeu de données propre.

Project_datasheet_2019-2020.xlsx:

	A	B	C	D	E	F	G	H	I	J	K
1	CropType	CropDays	Soil Moisture	Soil Temperature	Temperature	Humidity	Irrigation(Y/N)				
2	1	1	630	19	33	52	1				
3	1	2	460	20	31	50	0				
4	1	3	230	22	33	48	0			CropType	Code
5	1	4	140	24	28	62	0			Paddy	1
6	1	5	720	18	28	63	1			Ground Nuts	2
7	1	6	510	20	27	64	0				
8	1	7	327	21	29	61	0			Irrigation	Code
9	1	8	189	23	30	52	0			Yes (Y)	1
10	1	9	110	24	31	54	0			No (N)	0
11	1	10	750	18	29	56	1				
12	1	11	528	20	28	57	0				
13	1	12	321	21	29	54	0				
14	1	13	147	24	27	56	0				
15	1	14	621	19	26	58	1				
16	1	15	426	21	24	57	0				
17	1	16	254	22	23	56	0				
18	1	17	118	24	23	54	0				
19	1	18	693	19	25	62	1				
20	1	19	480	20	28	64	0				
21	1	20	326	21	29	63	0				
22	1	21	211	22	32	67	0				
23	1	22	126	24	33	64	0				

2.5 Dataset 5

286 fichiers csv pour NO2
285 fichiers csv pour Soil temp
dataset_info.csv:



```
Code,Paper,Day,Description,folder directory,unit,soil_temp_available,pH,bulk_density,soil_texture,measurment_season,wfps,irrigation,grazing
N2O_001,Akiyama et al 2000,1,chamber 1,Akiyama et al 2000/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_002,Akiyama et al 2000,2,chamber 1,Akiyama et al 2000/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_003,Akiyama et al 2000,3,chamber 1,Akiyama et al 2000/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_004,Akiyama et al 2000,4,chamber 1,Akiyama et al 2000/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_005,Akiyama et al 2000,5,chamber 1,Akiyama et al 2000/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_006,Akiyama et al 2000,1,chamber 2,Akiyama et al 2000/N2O_2,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_007,Akiyama et al 2000,2,chamber 2,Akiyama et al 2000/N2O_2,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_008,Akiyama et al 2000,3,chamber 2,Akiyama et al 2000/N2O_2,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_009,Akiyama et al 2000,4,chamber 2,Akiyama et al 2000/N2O_2,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_010,Akiyama et al 2000,5,chamber 2,Akiyama et al 2000/N2O_2,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_011,Akiyama and Tsuruta 2002,1,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_012,Akiyama and Tsuruta 2002,2,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_013,Akiyama and Tsuruta 2002,3,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_014,Akiyama and Tsuruta 2002,4,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_015,Akiyama and Tsuruta 2002,5,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_016,Akiyama and Tsuruta 2002,6,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_017,Akiyama and Tsuruta 2002,7,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_018,Akiyama and Tsuruta 2002,8,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_019,Akiyama and Tsuruta 2002,9,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_020,Akiyama and Tsuruta 2002,10,summer,Akiyama et al 2002/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_021,Akiyama and Tsuruta 2003a,1,,Akiyama et al 2003a/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_022,Akiyama and Tsuruta 2003a,2,,Akiyama et al 2003a/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_023,Akiyama and Tsuruta 2003a,3,,Akiyama et al 2003a/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_024,Akiyama and Tsuruta 2003a,4,,Akiyama et al 2003a/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_025,Akiyama and Tsuruta 2003a,5,,Akiyama et al 2003a/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
N2O_026,Akiyama and Tsuruta 2003a,6,,Akiyama et al 2003a/N2O_1,ug N m-2 h-1,Y,5.9,0.92,NA,summer,35-55%,N,N
```

diurnal_pattern_categorisation_data.csv:

```
data_code,max_hour,sec_max_hour,min_hour,sec_min_hour,cum_early,cum_late,cum_mid,cum_all,threshold,cum_early_percent,cum_late_percent,cum_mid_perce
1,20.245,16.038,4.123,7.951,8.333161075,10.23425649,6.458985081,16.66480016,0.505880865,0.500045665,0.614124166,0.38758251,N,Y,L,N,Y,Y,N,N
2,16.331,12.182,4.123,8.243,6.313319623,8.907238744,6.382615847,11.66285364,0.506799561,0.541318602,0.763727216,0.54726022,Y,Y,L,N,N,N,Y,N
3,16.175,12.088,3.997,20.097,8.514727132,9.815037158,7.248083246,12.88794505,0.506564228,0.660673761,0.761567273,0.56239247,Y,Y,L,N,N,N,Y,N
4,3.985,8.031,12.027,15.968,7.439354597,6.892273155,3.565333089,17.99123251,0.50686378,0.413498886,0.383090661,0.198170586,N,N,N,Y,Y,Y,N,Y
5,8.053,16.018,20.03,12.035,8.695295623,7.498508739,5.62951885,13.0646709,0.508560773,0.665557953,0.573953129,0.430896338,Y,Y,E,N,Y,Y,Y,N
6,20.052,16.088,4.052,8.015,6.291534301,8.673187761,5.248021828,13.90821206,0.504646957,0.452361114,0.623601921,0.377332601,N,Y,L,N,Y,Y,N,N
7,16.238,12.089,4.03,8.15,5.721242231,8.395326694,5.372811663,11.81145924,0.506799561,0.484380644,0.710778111,0.454881277,Y,Y,L,N,N,Y,Y,N
8,16.09,12.023,4.013,8.065,4.530592831,6.702627692,4.286415569,7.775275517,0.506735357,0.582692256,0.862043754,0.551287933,Y,Y,L,N,N,N,Y,N
9,19.933,7.946,16.011,12.029,4.587154998,3.765464061,1.777069516,13.0036901,0.50502925,0.352757945,0.289568886,0.136658864,N,N,N,Y,Y,Y,N,Y
10,16.02,20.017,4.016,12.04,4.229215226,6.834251923,3.603432885,10.41654293,0.508690123,0.406009485,0.656095978,0.345933666,N,Y,L,N,N,Y,N,N
11,13.892,17.844,9.94,5.716,4.923280628,5.567737429,4.810286256,7.49833119,0.590987442,0.656583512,0.742530209,0.641514243,Y,N,L,Y,N,N,N,N
12,13.736,10.098,5.921,18.048,8.065508948,9.536508527,7.101939686,11.59790701,0.597728631,0.695427972,0.822261165,0.612346666,Y,Y,L,Y,N,N,Y,N
13,18.019,9.616,13.69,6.051,5.955328145,7.277100377,5.107456795,11.237722,0.599041534,0.529940867,0.6475601,0.454492182,Y,N,L,Y,Y,Y,N,Y
14,13.926,9.766,5.947,18.17,7.409198472,9.408381761,6.929425752,10.51871874,0.599011631,0.704382221,0.894441803,0.658770894,Y,Y,L,Y,N,N,Y,N
15,14.281,10.003,6.046,18.345,8.307030872,10.94141552,8.12239453,13.04497514,0.600030002,0.636799287,0.838745601,0.622645459,Y,Y,L,Y,N,N,Y,N
16,10.167,13.805,5.99,18.117,7.310687055,8.669347203,6.57718324,9.90070345,0.601775237,0.7384077,0.875629418,0.66431474,Y,Y,L,Y,N,N,Y,N
17,10.087,18.151,6.352,14.246,6.209941691,7.828393529,6.078393891,9.318178021,0.596510414,0.666433038,0.840120624,0.652315708,Y,Y,L,Y,N,N,Y,N
18,10.408,14.312,6.249,18.726,6.088570315,7.66755329,5.899360699,10.5032155,0.591511806,0.579686318,0.730019611,0.561671871,Y,Y,L,Y,N,Y,Y,N
19,13.954,17.59,5.505,9.462,7.777461602,8.459410233,7.387736651,11.780673,0.596836765,0.660188225,0.718075294,0.627106503,Y,Y,L,Y,N,N,Y,N
20,9.722,18.04,5.647,13.626,5.497153472,6.188372508,4.797941377,7.290870055,0.594000594,0.753977705,0.848783816,0.658075283,Y,Y,L,Y,N,N,Y,N
21,10.176,14.483,18.212,6.253,10.2660264,8.090176702,8.615495979,12.43094273,0.598414202,0.825844558,0.650809587,0.693068592,Y,Y,E,Y,N,N,Y,N
22,10.215,14.277,18.225,6.259,8.659857113,7.946208846,7.653549175,10.93601012,0.596688379,0.79186623,0.7266095,0.699848399,Y,Y,E,Y,N,N,Y,N
23,10.171,6.039,17.984,14.018,8.556054565,5.972128861,7.512760368,9.994414274,0.603075686,0.856083641,0.597546659,0.751695913,Y,N,E,Y,N,N,N,N
24,10.018,6.286,18.039,13.851,7.816202315,6.500281541,6.654042818,9.528044097,0.61769702,0.820336497,0.682226223,0.698363982,Y,Y,E,Y,N,N,Y,N
25,6.288,10.368,17.817,13.989,9.406490718,6.237451428,7.751480586,11.1723665,0.610873549,0.841942548,0.558292769,0.693808298,Y,N,E,Y,Y,N,N,N
26,6.025,10.183,17.818,14.015,8.730433694,5.373079201,7.124979087,10.87358898,0.602167804,0.802902676,0.494140362,0.655255509,Y,N,E,Y,Y,N,N,N
27,17.733,5.819,9.898,13.908,4.852029665,5.434744207,3.607827287,9.955804958,0.596214041,0.487356842,0.54588697,0.362384288,Y,N,N,Y,N,Y,N,N
28,10.089,14.067,18.038,5.872,9.816520219,9.054586598,7.195092102,12.53004371,0.601262652,0.783438625,0.722630089,0.574227215,Y,Y,E,Y,N,Y,Y,N
29,14.035,5.959,18.129,9.939,7.709723915,7.181421503,5.021838544,11.94668278,0.597966912,0.645344323,0.601122641,0.420354222,Y,Y,E,Y,N,Y,Y,N
30,5.84,10.048,18.115,14.254,9.836954316,6.894519279,6.28013701,11.12005977,0.601172286,0.884613439,0.620007394,0.564757498,Y,Y,E,Y,Y,Y,Y,N
```

2.6 Dataset 6

Données sur les besoins en eau de différentes cultures (riz, maïs, pomme de terre et blé).

1. Structure générale

- **Total de fichiers** : 6 fichiers fusionnés
- **Fichiers sélectionnés pour la fusion** : 4
  - `blé.csv` *(contient les mêmes données que `blé2.xlsx`, seul `blé.csv` est conservé)*
  - `le riz2.csv` *(contient toutes les informations de `le riz.csv` avec des données supplémentaires, seul `le riz2.csv` est conservé)*
  - `maïs.xlsx`

    - `potato.csv`
- **Colonnes communes** : 14
- **Variables numériques** : 10
  (*water req, Min Temp, Max Temp, Humidity, Wind, Sun, Rad, Rain, altitude, latitude, longitude*)
- **Variables catégorielles** : 4
  (*month, crop, soil, city*)

2. Colonnes

- **water req (float)** : Besoin en eau de la culture
- **month (string)** : Mois de l'année
- **Min Temp (float)** : Température minimale
- **Max Temp (float)** : Température maximale
- **Humidity (float)** : Taux d'humidité
- **Wind (float)** : Vitesse du vent
- **Sun (float)** : Ensoleillement
- **Rad (float)** : Radiation
- **Rain (float)** : Précipitations
- **altitude (float)** : Altitude du lieu
- **latitude (float)** : Latitude du lieu
- **longitude (float)** : Longitude du lieu
- **crop (string)** : Type de culture (*riz, blé, maïs, patate*)
- **soil (string)** : Type de sol
- **city (string)** : Ville

3. Problèmes de Qualité des Données

1. **Valeurs Manquantes** :
    - pas de valeurs manquante
2. **Incohérences** :
    - **Casse des valeurs catégorielles** :
        - Les noms de mois et de villes ont des différences de casse (*"may" vs "May"*, *"july" vs "July"*).
        - Les types de sols ne sont pas toujours écrits de manière homogène (*"BLACK CLAY" vs "Black Clay"*).

blé.csv:

```
water req,month,Min Temp,Max Temp,Humidity,Wind,Sun,Rad,rain,altitude,latitude,crop,soil,city
46.8,March,16,32,35,192,8,19.2,5,431,26.91,wheat,red sandy,jaipur
181.3,April,22,37,26,240,9,22.6,4,431,26.91,wheat,red sandy,jaipur
327.5,May,26,41,27,288,9,23.4,18,431,26.91,wheat,red sandy,jaipur
187.5,June,27,39,46,312,7,20.6,51,431,26.91,wheat,red sandy,jaipur
0,July,25,34,73,264,5,17.5,213,431,26.91,wheat,red sandy,jaipur
51.1,March,24,34,70,216,9.5,23.5,2,7,13.08,wheat,red sandy,chennai
143.12,April,26,36,70,264,9.7,24.5,10,7,13.08,wheat,red sandy,chennai
236.6,May,27,39,63,312,9,23.2,32,7,13.08,wheat,red sandy,chennai
141.5,June,27,38,61,312,6.8,19.6,56,7,13.08,wheat,red sandy,chennai
0,July,26,36,64,288,6,18.4,111,7,13.08,wheat,red sandy,chennai
53.6,March,22.5,37.5,30,121,9.7,23.5,1.8,281,15.8,wheat,red sandy,kurnool
149.5,April,26,39.3,34,138,9.2,23.6,10,281,15.8,wheat,red sandy,kurnool
247.5,May,27.2,40,37,225,8.3,22.2,44,281,15.8,wheat,red sandy,kurnool
140.7,June,25,35.6,54,354,5.8,18.3,71.6,281,15.8,wheat,red sandy,kurnool
0,July,23.8,32.5,64,363,4.4,16.2,96,281,15.8,wheat,red sandy,kurnool
23.7,March,15,29,57,168,11,22.9,20,216,28.7,wheat,red sandy,delhi
108.9,April,21,36,41,168,10,23.8,40,216,28.7,wheat,red sandy,delhi
220.8,May,26,39,42,192,8,21.9,30,216,28.7,wheat,red sandy,delhi
117.9,June,28,39,54,216,7,20.6,80,216,28.7,wheat,red sandy,delhi
0,July,27,35,73,168,5,17.5,170,216,28.7,wheat,red sandy,delhi
16.7,March,22,34,63,120,8.4,20.5,28,9,22.7,wheat,red sandy,kolkata
84.4,April,25,36,70,192,9.2,23.3,51,9,22.7,wheat,red sandy,kolkata
```

le riz2.csv:

```
water req,month,Min Temp,Max Temp,Humidity,Wind,Sun,Rad,rain,altitude,latitude,longitude,crop,soil,CITY
479.4,may,26,41,27,288,9,23.4,18,431,26.91,75.78,rice,BLACK CLAY,jaipur
224.2,june,27,39,46,312,7,20.6,51,431,26.91,75.78,rice,BLACK CLAY,jaipur
47.8,july,25,34,73,264,5,17.5,213,431,26.91,75.78,rice ,BLACK CLAY,jaipur
17.9,august,24,33,81,240,5,16.9,227,431,26.91,75.78,rice,BLACK CLAY,jaipur
85.7,september,23,34,67,192,8,19.7,80,431,26.91,75.78,rice,BLACK CLAY,jaipur
416.5,may,27,39,63,312,9,23.2,32,7,13.08,80.27,rice,BLACK CLAY,chennai
176.5,june,27,38,61,312,6.8,19.6,56,7,13.08,80.27,rice ,BLACK CLAY,chennai
132.1,july,26,36,64,288,6,18.4,111,7,13.08,80.27,rice,BLACK CLAY,chennai
90.3,august,25,35,70,264,6.2,18.9,154,7,13.08,80.27,rice,BLACK CLAY,chennai
65.6,september,25,35,72,240,6.6,19.2,121,7,13.08,80.27,rice,BLACK CLAY,chennai
425.5,may,27.2,40,37,225,8.3,22.2,44.4,281,15.8,78.06,rice,BLACK CLAY,kurnool
178.8,june,25,35.6,54,354,5.8,18.3,71.6,281,15.8,78.06,rice,BLACK CLAY,kurnool
135.4,july,23.8,32.5,64,363,4.4,16.2,96,281,15.8,78.06,rice,BLACK CLAY,kurnool
123.2,august,23.5,32.1,63,302,4.9,16.9,98.7,281,15.8,78.06,rice,BLACK CLAY,kurnool
58.1,september,23.3,31.9,65,207,5.5,17.3,116,281,15.8,78.06,rice,BLACK CLAY,kurnool
411.3,may,26,39,42,192,8,21.9,30,216,28.7,77.01,rice,BLACK CLAY,delhi
155.7,june,28,39,54,216,7,20.6,80,216,28.7,77.01,rice,BLACK CLAY,delhi
47.7,july,27,35,73,168,5,17.5,170,216,28.7,77.01,rice,BLACK CLAY,delhi
10.8,august,26,34,77,144,4,15.3,200,216,28.7,77.01,rice,BLACK CLAY,delhi
31.8,september,25,34,72,120,6,16.7,110,216,28.7,77.01,rice,BLACK CLAY,delhi
281,may,26,36,73,192,8.4,22.6,126,9,22.7,88.36,rice,BLACK CLAY,kolkata
5.1,june,27,34,80,168,4.5,16.7,301,9,22.7,88.36,rice,BLACK CLAY,kolkata
0,july,26,33,84,144,4.1,16.1,375,9,22.7,88.36,rice,BLACK CLAY,kolkata
```

mais.xlsx:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	water req	month	Min Temp	Max Temp	Humidity	Wind	Sun	Rad	rain	altitude	latitude	longitude	crop	soil	CITY
2	44.40	October	19	34	43	168	9	18.60	12	431	26.91	75.78	maize	red lomy sand	jaipur
3	95.20	November	13	29	47	120	9	16.10	8	431	26.91	75.78	maize	red lomy sand	jaipur
4	100.40	December	9	24	51	120	8	13.80	3	431	26.91	75.78	maize	red lomy sand	jaipur
5	63	January	8	22	56	144	8	14.60	6	431	26.91	75.78	maize	red lomy sand	jaipur
6	2.60	February	11	26	46	192	9	17.90	10	431	26.91	75.78	maize	red lomy sand	jaipur
7	0	October	24	33	78	168	6.30	17.70	241	7	13.08	80.27	maize	red lomy sand	chennai
8	0	November	23	31	82	168	6.10	16.10	329	7	13.08	80.27	maize	red lomy sand	chennai
9	28.40	December	22	30	79	168	6.60	16	123	7	13.08	80.27	maize	red lomy sand	chennai
10	85.30	January	21	30	75	168	8.70	19.40	15	7	13.08	80.27	maize	red lomy sand	chennai
11	3.40	February	21	32	71	192	9.60	22.20	3	7	13.08	80.27	maize	red lomy sand	chennai
12	2.10	October	22.40	32.40	61	95	8.70	20.70	89.70	281	15.80	78.06	maize	red lomy sand	kurnool
13	86.30	November	19.20	31	56	78	7.70	17.60	23.60	281	15.80	78.06	maize	red lomy sand	kurnool
14	121.70	December	16.60	30.30	51	69	8.40	17.70	1.80	281	15.80	78.06	maize	red lomy sand	kurnool
15	99.40	January	17	31.30	47	104	8.80	18.70	0	281	15.80	78.06	maize	red lomy sand	kurnool
16	3.30	February	19.30	34.30	37	112	9.30	21.20	4.50	281	15.80	78.06	maize	red lomy sand	kurnool
17	29.40	October	19	33	63	72	7	15.70	10	216	28.70	77.01	maize	red lomy sand	delhi
18	65	November	13	28	68	48	9	15.50	10	216	28.70	77.01	maize	red lomy sand	delhi
19	70.80	December	9	22	71	96	10	15.30	10	216	28.70	77.01	maize	red lomy sand	delhi
20	40.30	January	8	21	73	120	10	16.10	20	216	28.70	77.01	maize	red lomy sand	delhi
21	2	February	10	23	66	144	10	18.60	30	216	28.70	77.01	maize	red lomy sand	delhi
22	0	October	24	32	78	72	6.30	16	141	9	22.70	88.36	maize	red lomy sand	kolkata
23	59.70	November	19	30	72	48	8.10	16.30	26	9	22.70	88.36	maize	red lomy sand	kolkata
24	88.70	December	14	27	72	72	8.40	15.60	7	9	22.70	88.36	maize	red lomy sand	kolkata
25	52	January	13	26	71	72	8.40	16.30	14	9	22.70	88.36	maize	red lomy sand	kolkata
26	2.20	February	17	30	66	72	8.50	18.40	22	9	22.70	88.36	maize	red lomy sand	kolkata

potato.csv:

```
water req,month,Min Temp,Max Temp,Humidity,Wind,Sun,Rad,Rain,altitude,latitude,longitude,crop,soil,city
73.7,March,21,32,62,168,9,22,10,14,19.07,72.82,potato,red loamy,mumbai
184.2,April,24,33,28,192,10,24.7,10,14,19.07,72.82,potato,red loamy,mumbai
201.7,May,27,33,69,240,10,25,10,14,19.07,72.82,potato,red loamy,mumbai
11.3,June,26,32,77,288,5,17.3,560,14,19.07,72.82,potato,red loamy,mumbai
0,July,25,30,83,336,2,12.8,640,14,19.07,72.82,potato,red loamy,mumbai
54.6,March,15,29,57,168,11,22.9,20,216,28.7,77.01,potato,red loamy,delhi
142.9,April,21,36,41,168,10,23.8,40,216,28.7,77.01,potato,red loamy,delhi
221.1,May,26,39,42,192,8,21.9,30,216,28.7,77.01,potato,red loamy,delhi
142,June,28,39,54,216,7,20.6,80,216,28.7,77.01,potato,red loamy,delhi
0,July,27,35,73,168,5,17.5,170,216,28.7,77.01,potato,red loamy,delhi
47.7,March,22,34,63,120,8.4,20.5,28,9,22.7,88.36,potato,red loamy,kolkata
116.1,April,25,36,70,192,9.2,23.3,51,9,22.7,88.36,potato,red loamy,kolkata
97.8,May,26,36,73,192,8.4,22.6,126,9,22.7,88.36,potato,red loamy,kolkata
5.9,June,27,34,80,168,4.5,16.7,301,9,22.7,88.36,potato,red loamy,kolkata
0,July,26,33,84,144,4.1,16.1,375,9,22.7,88.36,potato,red loamy,kolkata
93,March,22.5,37.5,30,121,9.7,23.5,1.8,281,15.8,78.06,potato,red loamy,kurnool
185.7,April,26,39.3,34,138,9.2,23.6,10,281,15.8,78.06,potato,red loamy,kurnool
246.4,May,27.2,40,37,225,8.3,22.2,44.4,281,15.8,78.06,potato,red loamy,kurnool
167.5,June,25,35.6,54,354,5.8,18.3,71.6,281,15.8,78.06,potato,red loamy,kurnool
16.4,July,23.8,32.5,64,363,4.4,16.2,96,281,15.8,78.06,potato,red loamy,kurnool
83.9,March,16,32,35,192,8,19.2,5,431,26.91,75.78,potato,red loamy,jaipur
221.7,April,22,37,26,240,9,22.6,4,431,26.91,75.78,potato,red loamy,jaipur
325.6,May,26,41,27,288,9,23.4,18,431,26.91,75.78,potato,red loamy,jaipur
216.4,June,27,39,46,312,7,20.6,51,431,26.91,75.78,potato,red loamy,jaipur
```

Chapitre 3 : Prétraitement des Données

3.1 Dataset 1

- **Fusion des lignes** : Regroupement des enregistrements répartis sur deux lignes en une seule ligne complète.
- **Gestion des valeurs manquantes** : Imputation des valeurs numériques par la médiane et des valeurs catégorielles par la modalité la plus fréquente.
- **Standardisation des valeurs catégorielles** : Conversion des noms de mois, villes et types de sols en majuscules.
- **Normalisation des variables numériques** : Application d'un **Min-Max Scaling** sur les variables continues.
- **Encodage des variables catégorielles** : Transformation des variables (mois, culture, sol, ville) en valeurs numériques via **Label Encoding**.
- **Feature Engineering** : Création de nouvelles variables comme les saisons et des ratios climatiques (ex. Température Max / Température Min, Humidité / Radiation).

Output\dataset_1_preprocessed.csv:

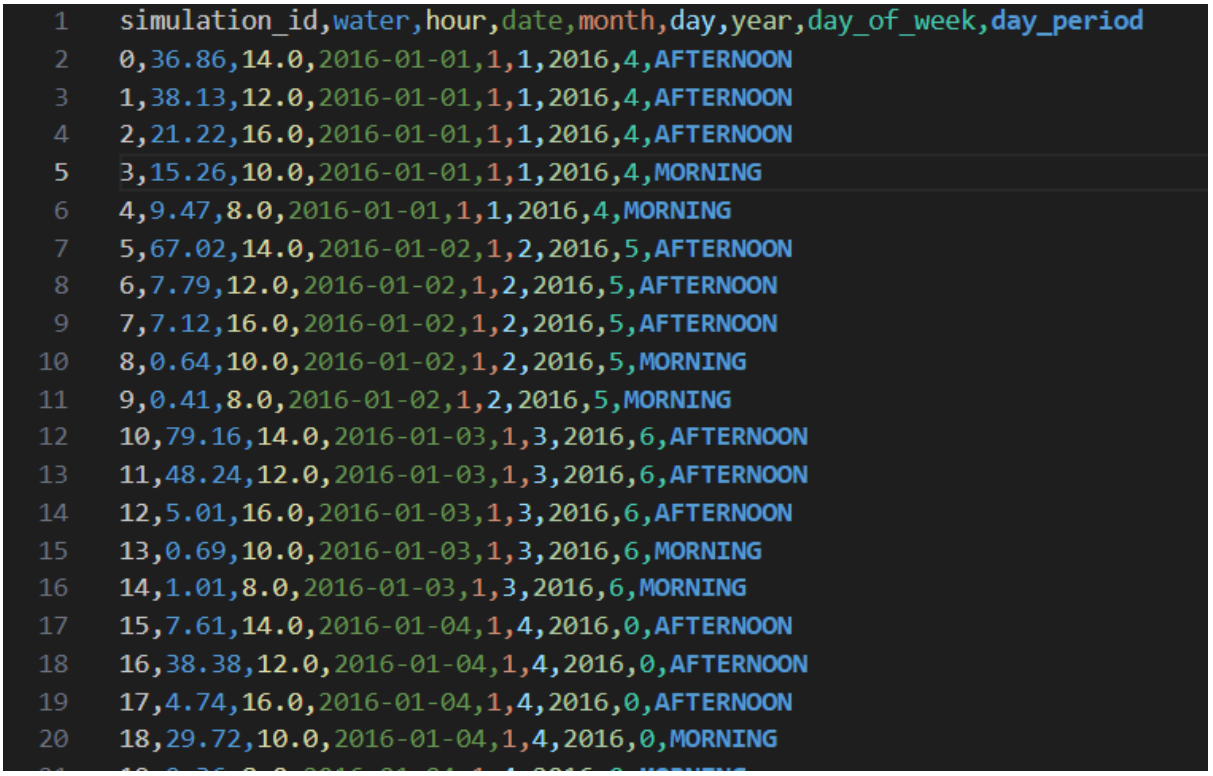
```
water req,month,Min Temp,Max Temp,Humidity,Wind,Sun,Rad,Rain,altitude,latitude,longitude,crop,soil,city,temp_ratio,humidity_rad_ratio,season
83.990001,MARCH,16.08,32.01,35.03,192.02,8.1,19.24,5.04,431.07,26.91,75.78,POTATO,RED LOAMY,JAIPUR,1.9906716417910448,1.8206860706860708,SPRING
39.07,MARCH,21.02,32.03,62.01,168.09,9.02,22.08,10.0,14.05,19.17,78.74286746987951,WHEAT,RED SANDY,MUMBAI,1.5237868696479544,2.8084239130434785,SP
235.18,MAY,27.07,39.02,63.07,312.01,9.1,23.27,32.08,7.0,13.09,80.27,POTATO,RED LOAMY,CHENNAI,1.4414480975249355,2.7103566824237215,SPRING
108.98,APRIL,21.05,36.03,41.07,168.05,10.09,23.84,40.07,216.07,28.7598,78.74286746987951,WHEAT,RED SANDY,DELHI,1.7116389548693587,1.72273489932885
0.02,JULY,26.01,36.02,64.07,288.04,6.07,18.43,111.01,7.08,13.09,78.74286746987951,WHEAT,RED SANDY,CHENNAI,1.3848519800076895,3.4763971785132934,SU
116.19,APRIL,25.01,36.02,70.06,192.0,9.29,23.35,51.08,9.02,22.73,88.36,POTATO,RED LOAMY,KOLKATA,1.4402239104358256,3.0004282655246253,SPRING
47.7206,MARCH,22.07,34.09,63.03,120.09,8.48,20.5,28.09,9.03,22.79,88.36,POTATO,RED LOAMY,KOLKATA,1.5446307204349798,3.0746341463414635,SPRING
47.71,MARCH,22.08,34.05,63.04,120.07,8.4,20.51,28.08,9.06,22.7798,88.36,POTATO,RED LOAMY,KOLKATA,1.5421195652173914,3.0736226231106776,SPRING
5.91,JUNE,27.04,34.07,80.06,168.01,4.56,16.71,301.06,9.03,22.79,88.36,POTATO,RED LOAMY,KOLKATA,1.2599852071005917,4.79114302812687,SUMMER
58.18,SEPTEMBER,23.32,31.9797,65.0,207.08,5.58,17.3102,116.06,281.1,15.82,78.06,RICE,BLACK CLAY,KURNOOL,1.3713421955403087,3.755011496112119,AUTUM
221.72,APRIL,22.08,37.07,26.04,240.01,9.07,22.62,4.02,431.07,26.95,75.78,POTATO,RED LOAMY,JAIPUR,1.678894927536232,1.1511936339522546,SPRING
47.7206,MARCH,22.1,34.0,63.08,120.05,8.41,20.6,28.08,9.03,22.71,88.36,POTATO,RED LOAMY,KOLKATA,1.5384615384615383,3.062135922330097,SPRING
221.79,APRIL,22.05,37.03,26.03,240.09,9.06,22.6502,4.06,431.02,26.91,75.78,POTATO,RED LOAMY,JAIPUR,1.6793650793650794,1.1492172254549629,SPRING
185.7099998,APRIL,26.02,39.33,34.08,138.1,9.28,23.6502,10.07,281.08,15.9,78.06,POTATO,RED LOAMY,KURNOOL,1.5115295926210608,1.4410026130857243,SPRI
85.77,SEPTEMBER,23.06,34.05,67.05,192.03,8.0,19.7798,80.09,431.07,26.96,75.78,RICE,BLACK CLAY,JAIPUR,1.476582827406765,3.3898219395544946,AUTUMN
141.58,JUNE,27.08,38.04,61.05,312.01,6.89,19.67,56.06,7.07,13.12,78.74286746987951,WHEAT,RED SANDY,CHENNAI,1.4047267355982276,3.103711235383833,SU
97.88,MAY,26.07,36.08,73.08,192.04,8.48,22.6103,126.1,9.08,22.72,88.36,POTATO,RED LOAMY,KOLKATA,1.3839662447257384,3.232155256675055,SPRING
0.02,AUGUST,25.08,30.05,83.02,288.02,3.07,14.1399,520.02,14.02,19.17,72.82,RICE,BLACK CLAY,MUMBAI,1.1981658692185009,5.8713286515463325,SUMMER
93.01,MARCH,22.57,37.56,30.06,121.1,9.78,23.54,1.84,281.02,15.88,78.06,POTATO,RED LOAMY,KURNOOL,1.6641559592379265,1.2769753610875105,SPRING
17.97,AUGUST,24.03,33.1,81.08,240.03,5.07,16.99,227.03,431.09,27.0,75.78,RICE,BLACK CLAY,JAIPUR,1.377444860590928,4.772218952324898,SUMMER
236.65,MAY,27.04,39.09,63.09,312.05,9.04,23.2,32.02,7.09,13.16,78.74286746987951,WHEAT,RED SANDY,CHENNAI,1.4456360946745563,2.719396551724138,SPRI
116.1,APRIL,25.01,36.08,70.09,192.08,9.29,23.39,51.03,9.02,22.75,88.36,POTATO,RED LOAMY,KOLKATA,1.442622950819672,2.996579734929457,SPRING
89.13,MARCH,24.02,34.07,70.1,216.08,9.55,23.51,2.02,7.05,13.13,80.27,POTATO,RED LOAMY,CHENNAI,1.4184013322231475,2.9817099106763076,SPRING
83.990001,MARCH,16.0,32.08,35.06,192.02,8.08,19.21,5.04,431.05,26.99,75.78,POTATO,RED LOAMY,JAIPUR,2.005,1.8250910983862572,SPRING
281.04,MAY,26.07,36.02,73.03,192.08,8.46,22.62,126.05,9.0,22.77,88.36,RICE,BLACK CLAY,KOLKATA,1.3816647487533564,3.2285587975243146,SPRING
97.81,MAY,26.08,36.07,73.08,192.01,8.4501,22.69,126.04,9.01,22.73,88.36,POTATO,RED LOAMY,KOLKATA,1.3830521472392638,3.2208021154693696,SPRING
73.76,MARCH,21.07,32.0,62.08,168.09,9.04,22.06,10.02,14.06,19.0802,72.82,POTATO,RED LOAMY,MUMBAI,1.5187470336971998,2.814143245693563,SPRING
135.49,JULY,23.82,32.56,64.0,363.01,4.42,16.22,96.04,281.09,15.81,78.06,RICE,BLACK CLAY,KURNOOL,1.3669185558354324,3.9457459926017266,SUMMER
58.13,SEPTEMBER,23.3302,31.91,65.05,207.1,5.53,17.39,116.04,281.01,15.83,78.06,RICE,BLACK CLAY,KURNOOL,1.367755098541804,3.740655549166187,AUTUMN
220.84,MAY,26.09,39.09,42.07,192.03,8.05,21.99,30.08,216.07,28.79,78.74286746987951,WHEAT,RED SANDY,DELHI,1.4982752012265237,1.913142337426103,SPR
0.04,JULY,25.05,30.04,83.09,336.05,2.06,12.88,640.07,14.03,19.16,78.74286746987951,WHEAT,RED SANDY,MUMBAI,1.1992015968063872,6.451086956521739,SUM
```

3.2 Dataset 2

- **Conversion des timestamps** : Transformation des timestamps Unix en format datetime et extraction des attributs temporels (jour, mois, année, jour de la semaine).
- **Catégorisation des heures** : Regroupement des heures en périodes de la journée (matin, après-midi, soir, nuit).
- **Gestion des valeurs manquantes** : Imputation des valeurs numériques par la médiane et des valeurs catégorielles par la modalité la plus fréquente.
- **Normalisation des variables numériques** : Application d'un **Min-Max Scaling** sur les valeurs continues (besoin en eau, heure).

- **Encodage des variables catégorielles** : Transformation des périodes de la journée en valeurs numériques via **Label Encoding**.

Output\dataset_2_preprocessed.csv:



```
simulation_id,water,hour,date,month,day,year,day_of_week,day_period
0,36.86,14.0,2016-01-01,1,1,2016,4,AFTERNOON
1,38.13,12.0,2016-01-01,1,1,2016,4,AFTERNOON
2,21.22,16.0,2016-01-01,1,1,2016,4,AFTERNOON
3,15.26,10.0,2016-01-01,1,1,2016,4,MORNING
4,9.47,8.0,2016-01-01,1,1,2016,4,MORNING
5,67.02,14.0,2016-01-02,1,2,2016,5,AFTERNOON
6,7.79,12.0,2016-01-02,1,2,2016,5,AFTERNOON
7,7.12,16.0,2016-01-02,1,2,2016,5,AFTERNOON
8,0.64,10.0,2016-01-02,1,2,2016,5,MORNING
9,0.41,8.0,2016-01-02,1,2,2016,5,MORNING
10,79.16,14.0,2016-01-03,1,3,2016,6,AFTERNOON
11,48.24,12.0,2016-01-03,1,3,2016,6,AFTERNOON
12,5.01,16.0,2016-01-03,1,3,2016,6,AFTERNOON
13,0.69,10.0,2016-01-03,1,3,2016,6,MORNING
14,1.01,8.0,2016-01-03,1,3,2016,6,MORNING
15,7.61,14.0,2016-01-04,1,4,2016,0,AFTERNOON
16,38.38,12.0,2016-01-04,1,4,2016,0,AFTERNOON
17,4.74,16.0,2016-01-04,1,4,2016,0,AFTERNOON
18,29.72,10.0,2016-01-04,1,4,2016,0,MORNING
```

3.3 Dataset 3

- **Séparation des données :** Extraction des informations de période de croissance et des besoins en eau des cultures en deux jeux de données distincts.
- **Nettoyage des noms de culture :** Standardisation et séparation des noms de culture multiples en entrées individuelles.
- **Traitement des valeurs en intervalle :** Conversion des plages de valeurs en colonnes min/max et calcul des moyennes.
- **Gestion des valeurs manquantes :** Remplacement des valeurs absentes par des médianes et des valeurs par défaut.
- **Fusion des jeux de données :** Regroupement des données de période et de besoins en eau sur la base des noms de culture.
- **Tri et structuration :** Réorganisation des données pour faciliter l'entraînement des modèles.

Output\dataset_3_processed.csv:

```csv
Crop,Total_growing_period_min,Total_growing_period_max,Total_growing_period_avg,Crop_wate
ALFALFA,100.0,365.0,232.5,800.0,1600.0,1200.0
BANANA,300.0,365.0,332.5,1200.0,2200.0,1700.0
BARLEY,120.0,150.0,135.0,450.0,650.0,550.0
BEAN DRY,95.0,110.0,102.5,350.0,500.0,425.0
BEAN GREEN,75.0,90.0,82.5,300.0,500.0,400.0
CABBAGE,120.0,140.0,130.0,900.0,1200.0,1050.0
CARROT,100.0,150.0,125.0,700.0,1300.0,1000.0
CITRUS,240.0,365.0,302.5,500.0,800.0,650.0
COTTON,180.0,195.0,187.5,400.0,600.0,500.0
CUCUMBER,105.0,130.0,117.5,350.0,550.0,450.0
EGGPLANT,130.0,140.0,135.0,500.0,700.0,600.0
FLAX,150.0,220.0,185.0,350.0,500.0,425.0
GRAIN,150.0,165.0,157.5,600.0,900.0,750.0
GROUNDNUT,130.0,140.0,135.0,,,
LENTIL,150.0,170.0,160.0,500.0,700.0,600.0
LETTUCE,75.0,140.0,107.5,450.0,700.0,575.0
MAIZE GRAIN,125.0,180.0,152.5,450.0,700.0,575.0
MAIZE SWEET,80.0,110.0,95.0,450.0,650.0,550.0
MELON,120.0,160.0,140.0,550.0,750.0,650.0
MILLET,105.0,140.0,122.5,1500.0,2500.0,2000.0
OATS,120.0,150.0,135.0,450.0,650.0,550.0
ONION DRY,150.0,210.0,180.0,400.0,800.0,600.0
ONION GREEN,70.0,95.0,82.5,600.0,1000.0,800.0
PEA,90.0,100.0,95.0,,,
PEANUT,130.0,140.0,135.0,,,
PEPPER,120.0,210.0,165.0,,,
POTATO,105.0,145.0,125.0,,,
RADISH,35.0,45.0,40.0,,,
RICE,90.0,150.0,120.0,,,
```

3.4 Dataset 4

- **Validation des données** : Vérification des colonnes requises et des types de données.
- **Nettoyage des noms de colonnes** : Suppression des espaces inutiles et standardisation.
- **Conversion des types de culture** : Mapping des valeurs numériques vers leurs équivalents textuels.
- **Gestion des valeurs manquantes** : Vérification et correction des entrées manquantes dans les données d'irrigation.
- **Exportation des données traitées** : Sauvegarde des données prétraitées dans un fichier CSV pour une utilisation ultérieure.

Output\dataset_4_processed.csv:

```
CropType,CropDays,Soil Moisture,Soil Temperature,Temperature,Humidity,Irrigation(Y/N)
Paddy,1,630,19,33,52,1
Paddy,2,460,20,31,50,0
Paddy,3,230,22,33,48,0
Paddy,4,140,24,28,62,0
Paddy,5,720,18,28,63,1
Paddy,6,510,20,27,64,0
Paddy,7,327,21,29,61,0
Paddy,8,189,23,30,52,0
Paddy,9,110,24,31,54,0
Paddy,10,750,18,29,56,1
Paddy,11,528,20,28,57,0
Paddy,12,321,21,29,54,0
Paddy,13,147,24,27,56,0
Paddy,14,621,19,26,58,1
Paddy,15,426,21,24,57,0
Paddy,16,254,22,23,56,0
Paddy,17,118,24,23,54,0
Paddy,18,693,19,25,62,1
Paddy,19,480,20,28,64,0
Paddy,20,326,21,29,63,0
```

3.5 Dataset 5

- **Le Dataset 5 n'a pas fourni d'informations pertinentes pouvant contribuer efficacement à l'entraînement du modèle. Par conséquent, il n'a pas été pris en compte dans le processus de traitement des données.**

3.3 Dataset 6

- **Lecture et fusion des fichiers** : Regroupement des données de différentes cultures (blé, riz, maïs, pomme de terre).
- **Standardisation des colonnes** : Conversion des noms de colonnes en minuscules et suppression des doublons.
- **Gestion des valeurs manquantes** : Correction et conversion des valeurs manquantes en types appropriés.
- **Ajout de nouvelles variables** : Création de nouvelles caractéristiques comme le ratio de température, l'humidité par rapport à la radiation et la saisonnalité.
- **Réorganisation des données** : Structuration des colonnes pour améliorer la lisibilité et la cohérence des données finales.

Output\dataset_6_processed.csv:

```
water req,month,min temp,max temp,humidity,wind,sun,rad,rain,altitude,latitude,longitude,crop,soil,city,temp_ratio,humidity_rad_ratio,season
46.8,MARCH,16.0,32.0,35,192,8.0,19.2,5.0,431,26.91,,WHEAT,RED SANDY,JAIPUR,2.0,1.8229166666666667,SPRING
181.3,APRIL,22.0,37.0,26,240,9.0,22.6,4.0,431,26.91,,WHEAT,RED SANDY,JAIPUR,1.6818181818181819,1.1504424778761062,SPRING
327.5,MAY,26.0,41.0,27,288,9.0,23.4,18.0,431,26.91,,WHEAT,RED SANDY,JAIPUR,1.5769230769230769,1.153846153846154,SPRING
187.5,JUNE,27.0,39.0,46,312,7.0,20.6,51.0,431,26.91,,WHEAT,RED SANDY,JAIPUR,1.4444444444444444,2.233009708737864,SUMMER
0.0,JULY,25.0,34.0,73,264,5.0,17.5,213.0,431,26.91,,WHEAT,RED SANDY,JAIPUR,1.36,4.171428571428572,SUMMER
51.1,MARCH,24.0,34.0,70,216,9.5,23.5,2.0,7,13.08,,WHEAT,RED SANDY,CHENNAI,1.4166666666666667,2.978723404255319,SPRING
143.12,APRIL,26.0,36.0,70,264,9.7,24.5,10.0,7,13.08,,WHEAT,RED SANDY,CHENNAI,1.3846153846153846,2.857142857142857,SPRING
236.6,MAY,27.0,39.0,63,312,9.0,23.2,32.0,7,13.08,,WHEAT,RED SANDY,CHENNAI,1.4444444444444444,2.7155172413793105,SPRING
141.5,JUNE,27.0,38.0,61,312,6.8,19.6,56.0,7,13.08,,WHEAT,RED SANDY,CHENNAI,1.4074074074074074,3.1122448979591835,SUMMER
0.0,JULY,26.0,36.0,64,288,6.0,18.4,111.0,7,13.08,,WHEAT,RED SANDY,CHENNAI,1.3846153846153846,3.4782608695652177,SUMMER
53.6,MARCH,22.5,37.5,30,121,9.7,23.5,1.8,281,15.8,,WHEAT,RED SANDY,KURNOOL,1.6666666666666667,1.2765957446808511,SPRING
149.5,APRIL,26.0,39.3,34,138,9.2,23.6,10.0,281,15.8,,WHEAT,RED SANDY,KURNOOL,1.5115384615384615,1.4406779661016949,SPRING
247.5,MAY,27.2,40.0,37,225,8.3,22.2,44.0,281,15.8,,WHEAT,RED SANDY,KURNOOL,1.4705882352941178,1.6666666666666667,SPRING
140.7,JUNE,25.0,35.6,54,354,5.8,18.3,71.6,281,15.8,,WHEAT,RED SANDY,KURNOOL,1.4240000000000002,2.9508196721311473,SUMMER
0.0,JULY,23.8,32.5,64,363,4.4,16.2,96.0,281,15.8,,WHEAT,RED SANDY,KURNOOL,1.365546218487395,3.9506172839506175,SUMMER
23.7,MARCH,15.0,29.0,57,168,11.0,22.9,20.0,216,28.7,,WHEAT,RED SANDY,DELHI,1.9333333333333333,2.4890829694323147,SPRING
108.9,APRIL,21.0,36.0,41,168,10.0,23.8,40.0,216,28.7,,WHEAT,RED SANDY,DELHI,1.7142857142857142,1.722689075630252,SPRING
220.8,MAY,26.0,39.0,42,192,8.0,21.9,30.0,216,28.7,,WHEAT,RED SANDY,DELHI,1.5,1.9178082191780823,SPRING
117.9,JUNE,28.0,39.0,54,216,7.0,20.6,80.0,216,28.7,,WHEAT,RED SANDY,DELHI,1.3928571428571428,2.6213592233009706,SUMMER
0.0,JULY,27.0,35.0,73,168,5.0,17.5,170.0,216,28.7,,WHEAT,RED SANDY,DELHI,1.2962962962962963,4.171428571428572,SUMMER
16.7,MARCH,22.0,34.0,63,120,8.4,20.5,28.0,9,22.7,,WHEAT,RED SANDY,KOLKATA,1.5454545454545454,3.073170731707317,SPRING
84.4,APRIL,25.0,36.0,70,192,9.2,23.3,51.0,9,22.7,,WHEAT,RED SANDY,KOLKATA,1.44,3.004291845493562,SPRING
96.6,MAY,26.0,36.0,73,192,8.4,22.6,126.0,9,22.7,,WHEAT,RED SANDY,KOLKATA,1.3846153846153846,3.230088495575221,SPRING
5.4,JUNE,27.0,34.0,80,168,4.5,16.7,301.0,9,22.7,,WHEAT,RED SANDY,KOLKATA,1.2592592592592593,4.790419161676647,SUMMER
0.0,JULY,26.0,33.0,84,144,4.1,16.1,375.0,9,22.7,,WHEAT,RED SANDY,KOLKATA,1.2692307692307692,5.217391304347825,SUMMER
39.0,MARCH,21.0,32.0,62,168,9.0,22.0,10.0,14,19.07,,WHEAT,RED SANDY,MUMBAI,1.5238095238095237,2.8181818181818183,SPRING
146.1,APRIL,24.0,33.0,28,192,10.0,24.7,10.0,14,19.07,,WHEAT,RED SANDY,MUMBAI,1.375,1.1336032388663968,SPRING
200.1,MAY,27.0,33.0,69,240,10.0,25.0,10.0,14,19.07,,WHEAT,RED SANDY,MUMBAI,1.2222222222222223,2.76,SPRING
10.6,JUNE,26.0,32.0,77,288,5.0,17.3,560.0,14,19.07,,WHEAT,RED SANDY,MUMBAI,1.2307692307692308,4.4508670520231215,SUMMER
0.0,JULY,25.0,30.0,83,336,2.0,12.8,640.0,14,19.07,,WHEAT,RED SANDY,MUMBAI,1.2,6.484375,SUMMER
```

Chapitre 4 : Fusion des Données

4.1 Stratégie de Fusion des Données

- Les datasets ont été fusionnés en fonction des variables communes telles que le type de culture, la période de croissance et les facteurs environnementaux.
- Un nettoyage préalable a permis d'éliminer les doublons et de standardiser les noms de colonnes.

Output\dataset_final_joined.csv:

```
water_req,month,min_temp,max_temp,humidity,wind,sun,rad,rain,altitude,longitude,soil,city,temp_ratio,humidity_rad_ratio,season,crop,total_growing_
83.990001,MARCH,16.08,32.01,35.03,192.02,8.1,19.24,5.04,431.07,75.78,RED LOAMY,JAIPUR,1.9906716417910448,1.8206860706860708,SPRING,POTATO,125.0,,,
39.07,MARCH,21.02,32.03,62.01,168.09,9.02,22.08,10.0,14.05,78.74286746987951,RED SANDY,MUMBAI,1.5237868696479544,2.8084239130434785,SPRING,WHEAT,1
235.18,MAY,27.07,39.02,63.07,312.01,9.1,23.27,32.08,7.0,80.27,RED LOAMY,CHENNAI,1.4414480975249355,2.710356682423721,SPRING,POTATO,125.0,,,
108.98,APRIL,21.05,36.03,41.07,168.05,10.09,23.84,40.07,216.07,78.74286746987951,RED SANDY,DELHI,1.711638954869359,1.7227348993288591,SPRING,WHEAT
0.02,JULY,26.01,36.02,64.07,288.04,6.07,18.43,111.01,7.08,78.74286746987951,RED SANDY,CHENNAI,1.3848519800076895,3.476397178513293,SUMMER,WHEAT,13
116.19,APRIL,25.01,36.02,70.06,192.0,9.29,23.35,51.08,9.02,88.36,RED LOAMY,KOLKATA,1.4402239104358256,3.0004282655246253,SPRING,POTATO,125.0,,,
47.7206,MARCH,22.07,34.09,63.03,120.09,8.48,20.5,28.09,9.03,88.36,RED LOAMY,KOLKATA,1.5446307204349798,3.0746341463414635,SPRING,POTATO,125.0,,,
47.71,MARCH,22.08,34.05,63.04,120.07,8.4,20.51,28.08,9.06,88.36,RED LOAMY,KOLKATA,1.5421195652173914,3.0736226231106776,SPRING,POTATO,125.0,,,
5.91,JUNE,27.04,34.07,80.06,168.01,4.56,16.71,301.06,9.03,88.36,RED LOAMY,KOLKATA,1.2599852071005917,4.79114302812687,SUMMER,POTATO,125.0,,,
58.18,SEPTEMBER,23.32,31.9797,65.0,207.08,5.58,17.3102,116.06,281.1,78.06,BLACK CLAY,KURNOOL,1.371342195540309,3.755011496112119,AUTUMN,RICE,120.0
221.72,APRIL,22.08,37.07,26.04,240.01,9.07,22.62,4.02,431.07,75.78,RED LOAMY,JAIPUR,1.678894927536232,1.1511936339522546,SPRING,POTATO,125.0,,,
47.7206,MARCH,22.1,34.0,63.08,120.05,8.41,20.6,28.08,9.03,88.36,RED LOAMY,KOLKATA,1.5384615384615383,3.062135922330097,SPRING,POTATO,125.0,,,
221.79,APRIL,22.05,37.03,26.03,240.09,9.06,22.6502,4.06,431.02,75.78,RED LOAMY,JAIPUR,1.6793650793650794,1.1492172254549629,SPRING,POTATO,125.0,,,
185.7099998,APRIL,26.02,39.33,34.08,138.1,9.28,23.6502,10.07,281.08,78.06,RED LOAMY,KURNOOL,1.5115295926210608,1.4410026130857243,SPRING,POTATO,12
85.77,SEPTEMBER,23.06,34.05,67.05,192.03,8.0,19.7798,80.09,431.07,75.78,BLACK CLAY,JAIPUR,1.476582827406765,3.389821939554494,AUTUMN,RICE,120.0,,,
141.58,JUNE,27.08,38.04,61.05,312.01,6.89,19.67,56.06,7.07,78.74286746987951,RED SANDY,CHENNAI,1.4047267355982276,3.103711235383833,SUMMER,WHEAT,1
97.88,MAY,26.07,36.08,73.08,192.04,8.48,22.6103,126.1,9.08,88.36,RED LOAMY,KOLKATA,1.3839662447257384,3.232155256675055,SPRING,POTATO,125.0,,,
0.02,AUGUST,25.08,30.05,83.02,288.02,3.07,14.1399,520.02,14.02,72.82,BLACK CLAY,MUMBAI,1.1981658692185009,5.871328651546332 5,SUMMER,RICE,120.0,,,
93.01,MARCH,22.57,37.56,30.06,121.1,9.78,23.54,1.84,281.02,78.06,RED LOAMY,KURNOOL,1.6641559592379265,1.2769753610875103,SPRING,POTATO,125.0,,,
17.97,AUGUST,24.03,33.1,81.08,240.03,5.07,16.99,227.03,431.09,75.78,BLACK CLAY,JAIPUR,1.377444860590928,4.772218952324898,SUMMER,RICE,120.0,,,
236.65,MAY,27.04,39.09,63.09,312.05,9.04,23.2,32.02,7.09,78.74286746987951,RED SANDY,CHENNAI,1.4456360946745563,2.719396551724138,SPRING,WHEAT,135
116.1,APRIL,25.01,36.08,70.09,192.08,9.29,23.39,51.03,9.02,88.36,RED LOAMY,KOLKATA,1.442622950819672,2.996579734929457,SPRING,POTATO,125.0,,,
89.13,MARCH,24.02,34.07,70.1,216.08,9.55,23.51,2.02,7.05,80.27,RED LOAMY,CHENNAI,1.4184013322231477,2.981709910676308,SPRING,POTATO,125.0,,,
83.990001,MARCH,16.0,32.08,35.06,192.02,8.08,19.21,5.04,431.05,75.78,RED LOAMY,JAIPUR,2.005,1.8250910983862567,SPRING,POTATO,125.0,,,
281.04,MAY,26.07,36.02,73.03,192.08,8.46,22.62,126.05,9.0,88.36,BLACK CLAY,KOLKATA,1.3816647487533564,3.228558797524314,SPRING,RICE,120.0,,,
97.81,MAY,26.08,36.07,73.08,192.01,8.4501,22.69,126.04,9.01,88.36,RED LOAMY,KOLKATA,1.3830521472392638,3.2208021154693696,SPRING,POTATO,125.0,,,
73.76,MARCH,21.07,32.0,62.08,168.09,9.04,22.06,10.02,14.06,72.82,RED LOAMY,MUMBAI,1.5187470336971998,2.814143245693563,SPRING,POTATO,125.0,,,
135.49,JULY,23.82,32.56,64.0,363.01,4.42,16.22,96.04,281.09,78.06,BLACK CLAY,KURNOOL,1.3669185558354324,3.945745992601727,SUMMER,RICE,120.0,,,
58.13,SEPTEMBER,23.3302,31.91,65.05,207.1,5.53,17.39,116.04,281.01,78.06,BLACK CLAY,KURNOOL,1.367755098541804,3.740655549166187,AUTUMN,RICE,120.0,
220.84,MAY,26.09,39.09,42.07,192.03,8.05,21.99,30.08,216.07,78.74286746987951,RED SANDY,DELHI,1.4982752012265237,1.913142337426103,SPRING,WHEAT,13
```

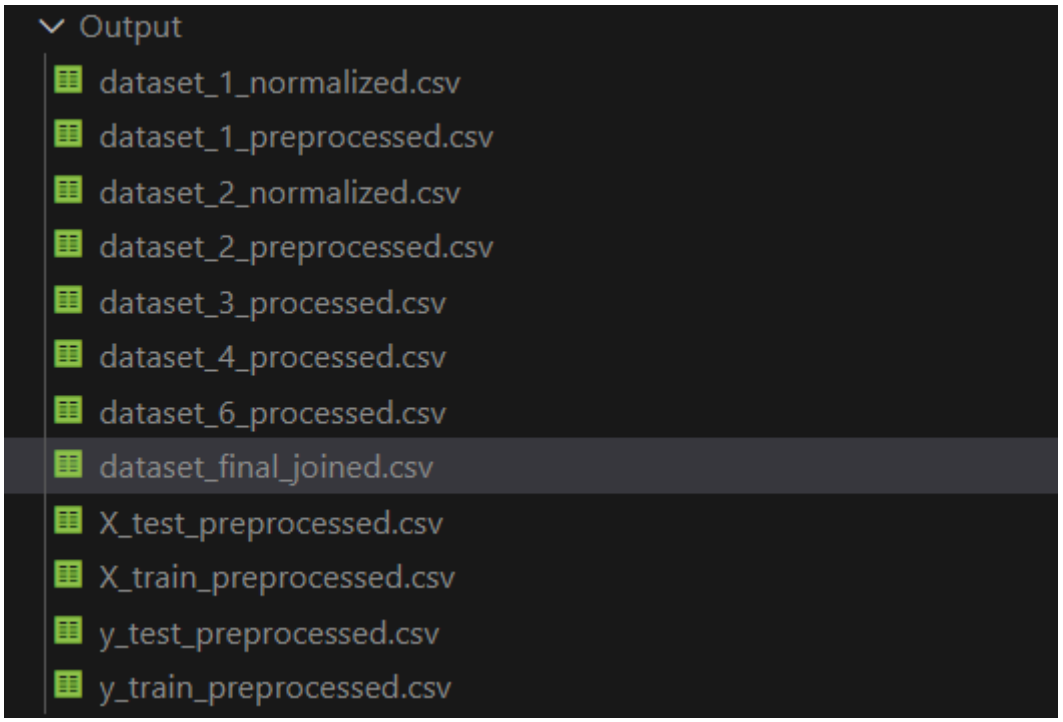
4.2 Justification des Choix Retenus

- La suppression des colonnes `crop_water_need_min`, `crop_water_need_max` et `crop_water_need_avg` a été effectuée en raison du nombre élevé de valeurs manquantes.
- Les valeurs catégorielles (`mois`, `sol`, `ville`, `saison`, `culture`) ont été encodées via **Label Encoding** pour leur intégration dans les modèles.
- Les variables numériques ont été normalisées à l'aide d'un **StandardScaler** pour assurer une meilleure convergence des modèles d'apprentissage.

4.3 Préparation des Données pour l'Entraînement

- Séparation des jeux de données en ensembles d'entraînement (80%) et de test (20%).
- Sauvegarde des datasets prétraités (`X_train_preprocessed.csv`, `X_test_preprocessed.csv`, `y_train_preprocessed.csv`, `y_test_preprocessed.csv`).

- Vérification de l’équilibre entre les différentes catégories et variables pour éviter les biais dans l’apprentissage.



Chapitre 5 : Modélisation

- Trois modèles ont été testés : **Random Forest**, **Neural Network (MLPRegressor - Scikit-learn)** et **Deep Neural Network (TensorFlow/Keras)**.
- Optimisation des hyperparamètres via **GridSearchCV** pour Random Forest et MLPRegressor.
- Construction et entraînement d'un réseau de neurones profond avec **TensorFlow/Keras**.
- Utilisation de **Mean Squared Error (MSE)** et **Mean Absolute Error (MAE)** comme métriques de performance.

```python
# 1. Random Forest Model
rf_pipeline = Pipeline([
    ('rf', RandomForestRegressor(random_state=42))
])

rf_params = {
    'rf__n_estimators': [100, 200, 300],
    'rf__max_depth': [10, 20, 30],
    'rf__min_samples_split': [2, 5],
    'rf__min_samples_leaf': [1, 2]
}

rf_grid = GridSearchCV(
    rf_pipeline,
    rf_params,
    cv=5,
    scoring='neg_mean_squared_error',
    n_jobs=-1
)

print("Training Random Forest model...")
rf_grid.fit(X_train, y_train)
models['random_forest'] = rf_grid.best_estimator_
```

```python
# 2. Neural Network (Scikit-learn MLPRegressor)
mlp_pipeline = Pipeline([
    ('mlp', MLPRegressor(random_state=42, max_iter=1000))
])

mlp_params = {
    'mlp__hidden_layer_sizes': [(50,), (100,), (50, 25)],
    'mlp__activation': ['relu', 'tanh'],
    'mlp__learning_rate_init': [0.001, 0.01]
}

mlp_grid = GridSearchCV(
    mlp_pipeline,
    mlp_params,
    cv=5,
    scoring='neg_mean_squared_error',
    n_jobs=-1
)

print("Training Neural Network (MLPRegressor) model...")
mlp_grid.fit(X_train, y_train)
models['neural_network_sklearn'] = mlp_grid.best_estimator_
```

```python
# 3. Deep Neural Network (TensorFlow)
def create_deep_model(input_dim):
    model = Sequential([
        Dense(128, activation='relu', input_dim=input_dim),
        Dropout(0.3),
        Dense(64, activation='relu'),
        Dropout(0.2),
        Dense(32, activation='relu'),
        Dense(1)
    ])
    model.compile(optimizer=Adam(learning_rate=0.001),
                  loss='mse',
                  metrics=['mae'])
    return model

print("Training Deep Neural Network (TensorFlow) model...")
dnn_model = create_deep_model(X_train.shape[1])
dnn_model.fit(
    X_train,
    y_train,
    epochs=100,
    batch_size=32,
    validation_split=0.2,
    verbose=0
)
models['deep_neural_network'] = dnn_model
```

Chapitre 6 : Comparaison des Modèles

- **Random Forest** : MSE = 0.0037, RMSE = 0.0612, MAE = 0.0098, R^2 = 0.9962.
- **MLPRegressor (Neural Network - Scikit-learn)** : MSE = 0.0008, RMSE = 0.0287, MAE = 0.0159, R^2 = 0.9992.
- **Deep Neural Network (TensorFlow/Keras)** : MSE = 0.0155, RMSE = 0.1246, MAE = 0.0871, R^2 = 0.9843.
- Le modèle **MLPRegressor** a obtenu la meilleure performance avec un **R^2 de 0.9992**.

```
Model Evaluation Results:
--------------------------------------------------------------------------------

random_forest:
MSE: 0.0037
RMSE: 0.0612
MAE: 0.0098
R² Score: 0.9962

neural_network_sklearn:
MSE: 0.0008
RMSE: 0.0287
MAE: 0.0159
R² Score: 0.9992

deep_neural_network:
MSE: 0.0155
RMSE: 0.1246
MAE: 0.0871
R² Score: 0.9843
```

- Exemple de prédictions comparées aux valeurs réelles pour les premiers échantillons de test.

```python
# Example of making predictions with the best model
# Find best model based on R² score
best_model_name = max(results.items(), key=lambda x: x[1]['r2'])[0]
best_model = models[best_model_name]

print(f"\nBest performing model: {best_model_name}")

# Example prediction using first few samples from test set
sample_features = X_test[10:15]

# Make predictions based on model type
if isinstance(best_model, tf.keras.Model):
    scaled_preds = best_model.predict(sample_features, verbose=0)
else:
    scaled_preds = best_model.predict(sample_features).reshape(-1, 1)

# Inverse transform predictions and actual values
predictions = y_scaler.inverse_transform(scaled_preds)
actual_values = y_scaler.inverse_transform(y_test[10:15].reshape(-1, 1))

print("\nSample Predictions (first 5 test samples):")
print("Predicted water requirements:", predictions.ravel())
print("Actual water requirements:", actual_values.ravel())
```

8] ✓ 0.0s

```
Best performing model: neural_network_sklearn

Sample Predictions (first 5 test samples):
Predicted water requirements: [219.08947948  50.85579111 185.82848669  10.39267439   0.40965692]
Actual water requirements: [2.2113e+02 5.1130e+01 1.8570e+02 1.0970e+01 6.0000e-02]
```

Chapitre 5 : Conclusion

Ce projet a permis de développer un pipeline d'ingestion et de modélisation des données agronomiques. Parmi les modèles testés, **MLPRegressor** a obtenu les meilleures performances avec un **R^2 de 0.9992**. Des améliorations futures pourraient inclure l'optimisation des hyperparamètres et l'intégration de nouvelles données pour affiner les prédictions.