Rapport TP Data Augmentation

Thème :

Variations mensuelles de la consommation d'eau et des paramètres environnementaux pour diverses cultures agricoles en Inde.

Objectifs :

- Exploration de la Data Augmentation pour enrichir le jeu de données initial en générant de nouvelles données à partir des observations existantes.
- Analyse comparative entre les données d'origine et les données augmentées afin d'évaluer la fidélité et la diversité des données générées.
- Évaluation de l'efficacité de la Data Augmentation dans la résolution du défi du manque de données, en examinant son impact sur la performance des modèles d'apprentissage automatique.

Réalisé par :
- Tagzirt Elissa
- Messar Cylia

Encadré par :
M. Ait Ali Yahia Yacine

2023/2024

Sommaire

1. Introduction

Dans le cadre de ce travail pratique réalisé dans le module FASI (File d'attente et simulation) de la 4ème année à l'École Supérieure d'Informatique (ESI), spécialité systèmes informatiques (SIQ), notre objectif était d'implémenter un modèle d'augmentation de données dédié à l'agriculture. Nous avons travaillé avec huit bases de données, portant des informations sur diverses cultures :

1) Arachide
2) Riz
3) Pomme de terre
4) Tomate
5) Maïs
6) Blé

Chacune des bases de données se compose principalement d'un ensemble d'observations détaillant les facteurs environnementaux pour différents mois et pour différentes cultures. Ces observations comprennent notamment les variations des besoins en eau selon les mois et les cultures, ainsi que d'autres paramètres tels que le rayonnement solaire, les précipitations, la température, et d'autres encore.

L'importance de cette pratique réside dans sa capacité à fournir des ensembles de données riches en observations, offrant ainsi des opportunités pour des analyses statistiques approfondies ainsi que pour le développement de modèles de machine learning. De plus, ces données peuvent être utilisées pour des prédictions futures, permettant ainsi la création de stratégies visant à optimiser la consommation en eau et d'autres ressources.

2. Description des paramètres environnementaux

Paramètre	Unité
Water Requirement (Besoin en eau)	(L/m²)
Month (Mois)	Les noms des mois
Min Temp (Température minimale)	(°C)
Max Temp (Température maximale)	(°C)
Humidity (Humidité)	(%)
Wind (Vent)	(km/h)
Sun (Ensoleillement)	(h/j)
Rad (Rayonnement solaire)	(W/m²)
Rain (Précipitations)	(mm)
Altitude (Altitude)	(m)
Latitude (Latitude)	(°)
Longitude	(°)
Crop (Culture)	Les cultures
Soil (Type de sol)	Les types de sol
City (Ville)	Les noms des cités
Soil Moisture (Humidité du Sol)	(%)
Soil Temperature (Température du sol)	(°C)
Temperature (Température de l'air)	(°C)
Humidity (Humidité relative de l'air)	(%)
Irrigation(Y/N)	

3. Augmentation de données

L'augmentation des données est le processus qui consiste à générer artificiellement de nouvelles données à partir de données existantes, principalement pour entraîner de nouveaux modèles de machine learning (ML). Les modèles ML nécessitent des jeux de données volumineux et variés pour la formation initiale, mais il peut être difficile de trouver des jeux de données du monde réel suffisamment diversifiés en raison des silos de données, des réglementations et d'autres limitations. L'augmentation des données augmente artificiellement le jeu de données en apportant de légères modifications aux données d'origine. Les solutions d'intelligence artificielle (IA) générative sont désormais utilisées pour une augmentation rapide et de haute qualité des données dans divers secteurs.

4. Description des outils utilisés

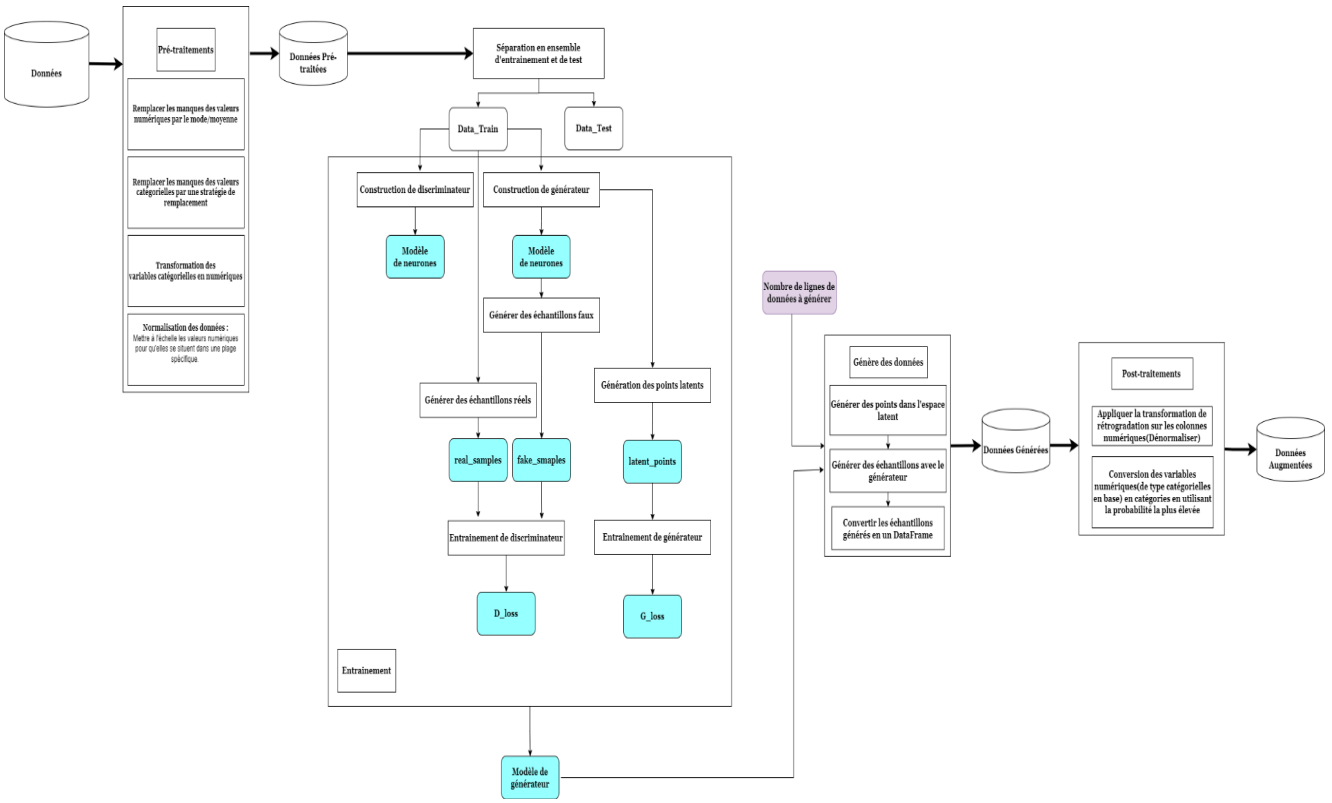
4.1. Pandas

Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

4.2. GAN

En intelligence artificielle, les réseaux antagonistes génératifs (RAG) parfois aussi appelés réseaux adverses génératifs (en anglais generative adversarial networks ou GANs) sont une classe d'algorithmes d'apprentissage non supervisé. Ces algorithmes ont été introduits par Goodfellow et al. 2014. Ils permettent de générer des images ou des données avec un fort degré de réalisme.

Un GAN est un modèle génératif où deux réseaux de neurones sont placés en compétition dans un scénario de théorie des jeux. Le premier réseau est le générateur, il génère un échantillon, tandis que son adversaire, le discriminateur essaie de détecter si un échantillon est réel ou bien s'il est le résultat du générateur. Ainsi, le générateur est entraîné avec comme but de tromper le discriminateur.



La perte du discriminateur (d_loss) est calculée comme une moyenne entre la perte sur les échantillons réels et la perte sur les échantillons générés.

La perte du générateur (g_loss) est calculée à partir de la sortie du modèle combiné (combined), où le discriminateur est gelé et le générateur est entraîné seul.

"latent_points" fait référence aux points latents dans l'espace latent d'un modèle génératif, comme un GAN (Generative Adversarial Network).

Dans un GAN, le générateur prend souvent des points aléatoires de cet espace latent en entrée et les utilise pour générer de nouvelles données. L'espace latent est un espace de représentation abstrait où chaque point représente un vecteur latent, qui peut être interprété comme une sorte de code latent ou de représentation cachée.

5. Qualité des données générées

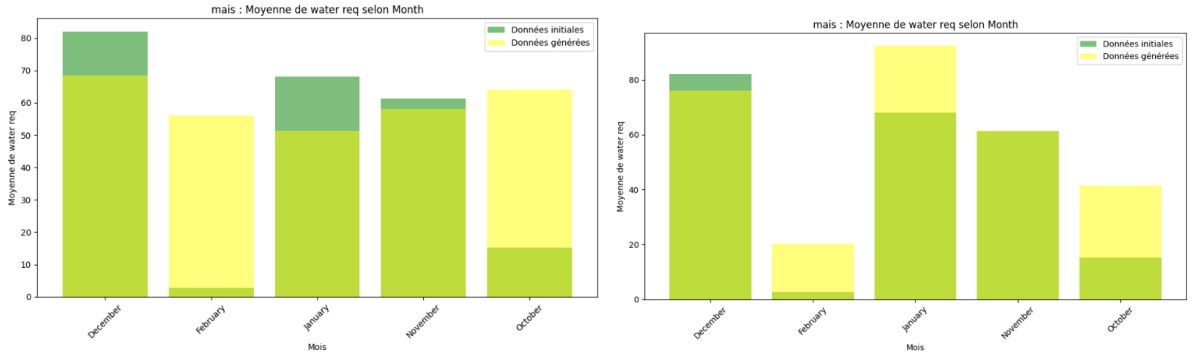
La qualité élevée des données générées par GAN revient à plusieurs paramètres clés tels que la profondeur de modèle de neurones génératifs, le nombre d'itérations d'entraînement et la richesse de la base de données initiale.

5.1. Profondeur du modèle de neurons génératifs

La profondeur du modèle se réfère à la complexité et à la taille du réseau de neurones utilisé pour générer de nouvelles données. Un modèle plus profond peut capturer des motifs plus complexes dans les données et produire des échantillons de meilleure qualité.

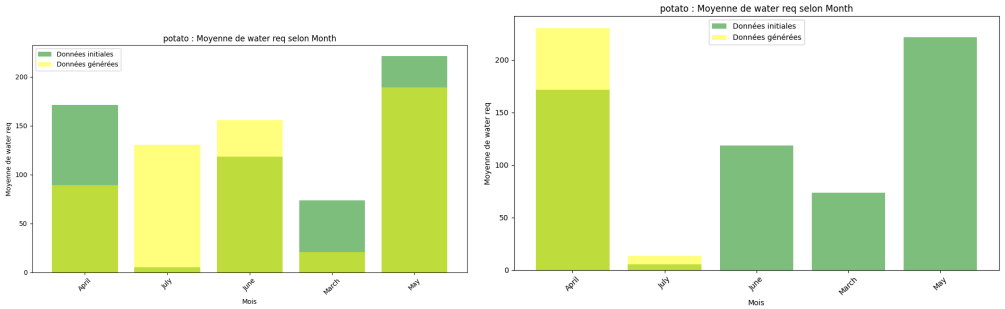
5.2. Nombre d'itérations d'entraînement

Le nombre d'itérations d'entraînement fait référence au nombre de fois que le générateur et le discriminateur sont mis à jour lors du processus d'apprentissage. Un nombre suffisant d'itérations est nécessaire pour permettre au modèle de converger vers une distribution de données réaliste.



5.3. Richesse de la base de données initiales

La richesse de la base de données initiale fait référence à la diversité et à la quantité des données disponibles pour l'entraînement du GAN. Une base de données plus riche et plus diversifiée fournit au modèle plus d'informations pour apprendre les caractéristiques essentielles des données et produire des échantillons de meilleure qualité.



6. Conclusion

Cette étude représente un premier pas dans l'exploration d'une pratique puissante et de haute qualité pour l'augmentation des données. Nous avons réalisé un travail qui fournit des observations proches de la réalité, fidèles à la distribution des

données initiales. Il est important de noter que cela dépend étroitement de paramètres tels que le nombre d'itérations d'entraînement et la qualité de la base de données initiale. En augmentant le nombre d'itérations d'entraînement, les données générées deviennent de plus en plus fidèles aux données de base, améliorant ainsi la capacité du modèle à généraliser et à produire des résultats précis. Ce travail ouvre la voie à des recherches plus approfondies dans le domaine de l'augmentation des données et de son application dans divers domaines, y compris l'agriculture et d'autres secteurs.

7. Webographie

- Compréhension de principe de GAN
- GitHub Tutorial
- Conditional Gan In Tensorflow
- Github Gan
- Youtube Tutorial