Université paris 8

PARIS 8 UNIVERSITE VINCENNES - SAINT-DENIS

LIASD

LABORATOIRE PARAGRAPHE

# MÉMOIRE

pour obtenir le grade de Master délivré par

## Université Paris 8 Vincennes à Saint Denis

Mention *Informatique*

Parcours *Big data et Fouille de données*

*présenté et soutenu publiquement par*

Angélique **Delevingne**

le 8 septembre 2023

# Prédiction de potentiel de production de crédits carbone sur des projets AFOLUs

Directeur de mémoire : Gilles **Bernard**

Maître de stage : Aude **Carro**

Stage effectué à : **TotalEnergies**
2 PLACE JEAN MILLIER, 92400 COURBEVOIE

COMUE Paris Lumières
Laboratoire d'Intelligence Artificielle et Science des Données
Laboratoire Paragraphe

MASTER

# Remerciements

Je tiens tout d'abord à remercier toute l'équipe du service TENBS de TotalEnergies pour son accueil chaleureux, son soutien tout au long de mon stage et les connaissances qu'ils ont partagées avec moi durant toute cette période. Je remercie Aude Carro pour la confiance qu'elle m'a accordée ainsi que son soutien tout au long de ma mission.

Je tiens également à remercier les enseignants du Master 1 Informatique qui m'ont permis d'apprendre, de comprendre et de créer des programmes informatiques spécialisés en data science. J'ai aimé la diversité des cours et les projets que j'ai pu réaliser à l'Université Paris 8. Grâce à l'enseignement fort en pratique, j'ai pu m'adapter avec agilité à un contexte, un environnement métier et aux attentes de la mission qui m'a été confiée.

Ces connaissances et cette agilité m'ont permis d'être plus performante lors de mon stage en entreprise. Elles m'ont permis de trouver des solutions auxquelles je n'aurai peut-être pas pensé auparavant.

Un grand merci !

# Sommaire

# Introduction

J'ai effectué mon stage de Master 1 Big Data, de début juillet à fin septembre 2023 au sein de l'entreprise TotalEnergies située dans le quartier d'affaires de La Défense.

TotalEnergies est une société française multi-énergie mondialement connue. Elle produit et fournit de l'énergie telle que : pétrole, biocarburant, gaz et électricité. Elle se présente comme le 1er fournisseur alternatif d'énergie de France avec comme ambition d'être un acteur majeur de la transition énergétique.

J'ai effectué mon stage dans la branche TotalEnergies Nature Based Solutions (TENBS). Cette entité finance, développe et gère des exploitations qui séquestrent du carbone. TotalEnergies mène des actions afin de compenser les émissions de gaz à effet de serre générées par ses activités.

La mission, qui m'a été confiée, consiste à extraire, archiver, traiter et analyser les informations importantes présentes dans des projets d'Agriculture, Foresterie et Autres Usages des Terres (AFOLU). Pour extraire, j'utilise des techniques de scrapping, une méthode OCR et une analyse de texte. Puis, je stocke ces informations dans une base de données. Par la suite, je les traite afin de réaliser des analyses de benchmarking et de prédiction.

J'ai choisi de faire ce stage car c'est un sujet de stage couvrant le domaine de mes acquis : de la capture et archivage des données jusqu'à leur traitement et analyse. J'avais aussi envie de découvrir et en apprendre plus sur le domaine de l'élimination du carbone par la nature. Au cours de cette mission, j'ai été à l'écoute des besoins métiers. Pour cela, j'ai endossé différents rôles tels que :

- Analyste de besoins,
- Architecte de solution,
- Data Engineer,
- Data Analyst,
- Data Visualisation,

- Testeuse.

# Première partie

# Problématique et état de l'art

# Table des matières

# Chapitre 1

# Contexte de résolution du problème

## Sommaire

# 1 Introduction

La mission, qui m'a été confiée, consiste à extraire, archiver, traiter et analyser les informations importantes présentes dans des projets d'Agriculture, Foresterie et Autres Usages des Terres (AFOLU). Puis à les traiter afin de réaliser des analyses de benchmarking et de prédiction.

# 2 Entreprise

Crée en 1924, le groupe Total a décidé en 2021 de changer de nom pour symboliser son investissement dans les énergies propres pour devenir TotalEnergies.


Fig. 2.1 : Logo TotalEnergies

TotalEnergies est une compagnie multi-énergies mondiale de production et de fourniture d'énergies : pétrole et biocarburants, gaz naturel et gaz verts, renouvelables et électricité. Son objectif est de satisfaire la demande mondiale en énergie tout en réduisant ses émissions carbones. Elle s'est engagée pour la neutralité carbone en 2050. Pour cela l'ensemble de la société avec ses 101 milles collaborateurs s'est mobilisée pour une énergie toujours plus abordable, plus propre, plus fiable et accessible au plus grand nombre.

La société est présente dans plus de 130 pays avec 160 nationalités représentées et plus de 740 métiers.

D'après son site, TotalEnergies est présente dans ces différentes régions du monde :

- Afrique : 40 pays
- Amérique : 17 pays
- Europe : 31 pays
- Moyen-Orient : 9 pays

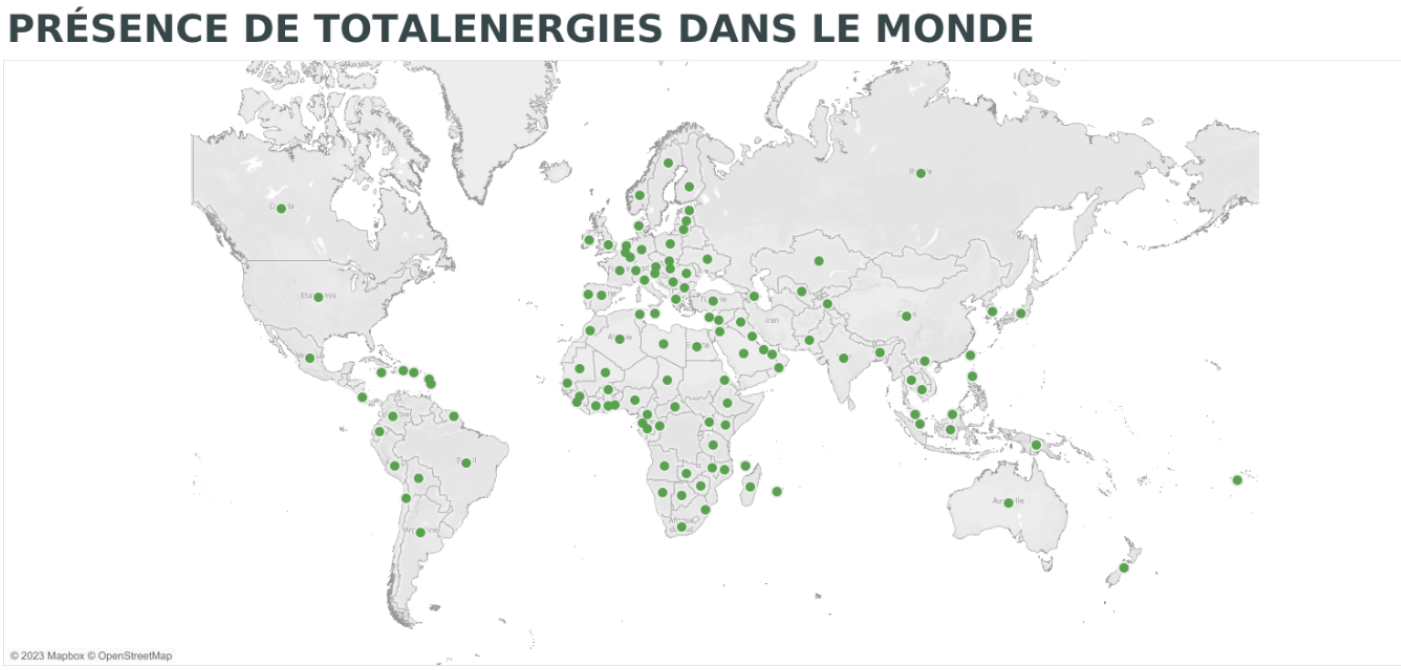


Fig. 2.2 : Présence de TotalEnergies dans les différentes régions du monde

TotalEnergies peut opérer mondialement en s'appuyant localement sur des collaborateurs qui connaissent les spécificités de chaque pays. La collaboration mondiale est source à la fois de solutions innovantes et d'opportunités nouvelles.

En 2022, l'entreprise Total a généré un chiffre d'affaires record de 281 milliards de dollars américains, après un fort ralentissement économique subit en 2020 à cause du COVID-19.

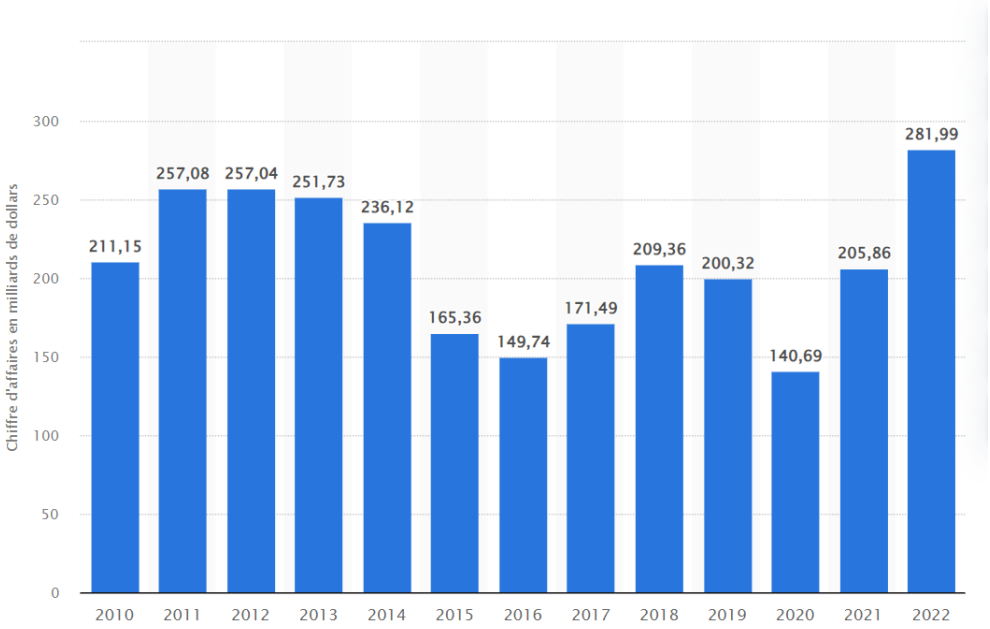


Fig. 2.3 : Chiffre d'affaire de TotalEnergies

## 3 Gouvernance

Depuis octobre 2014, le président-directeur général de TotalEnergies est Patrick Pouyanné. Sous sa responsabilité de Patrick Jean Pouyanné, le Comité exécutif (COMEX) constitue l'instance de direction de TotalEnergies. Une seconde instance, le Comité de performance de la Compagnie, a pour mission l'examen, l'analyse et le pilotage des résultats HSE (Hygiène, sécurité et environnement), financiers et opérationnels.

Chez TotalEnergies, la gouvernance s'articule autour du Conseil d'administration et de la Direction générale. Le rôle du Conseil d'administration est de définir les orientations stratégiques de TotalEnergies, avec l'appui de quatre comités (Comité d'audit, Comité de gouvernance et d'éthique, Comité des rémunérations et Comité Stratégie & RSE). Il est composé de 14 administrateurs aux profils variés, dont neuf membres indépendants. Le Conseil d'administration veille à l'application des meilleures pratiques de gouvernance et se réfère dans cette démarche au Code AFEP-MEDEF de gouvernement d'entreprise des sociétés cotées.

Lors de mon stage, le Vice-Président de Nature Based Solutions, Adrien Henry a proposé un projet de réduction carbone lors d'un COMEX. Ce projet a ainsi pu être validé et être lancé à l'issu de ce conseil.

# 4 Le service

Ma mission s'est déroulée dans le service Nature Based Solutions (TENBS). TENBS a été créé en juin 2019. Ce service finance, développe et gère des exploitations qui séquestrent du carbone et diminuent les émissions carbones. Ces exploitations tirent parti de la préservation ou de la capacité à absorber les émissions des agroécosystèmes. Ces solutions sont dites « basées sur la nature ».

Ce service travaille à compenser les émissions de gaz à effet de serre (GES) générées par les activités de TotalEnergies.

TENBS a l'ambition de compenser toutes les émissions résiduelles de Scope 1 et 2 de l'entreprise à compter de 2030, ce qui équivaut à environ 5 à 10 millions de tonnes équivalent CO2 par an.

Actuellement, TENBS fait partie de la nouvelle direction Neutralité Carbone qui a été créée en septembre 2021 dans le service d'Exploration Production (EP). Cette direction englobe également au même niveau que TENBS les activités : Carbon Footprint Reduction (CFR) et Carbon Capture and Storage (CCS).

TENBS développe des solutions basées sur la nature avec des partenaires et des communautés locales. Elle possède et développe un portefeuille diversifié d'opérations intégrées et durables produisant des crédits carbones certifiés, des revenus et des co-bénéfices à partir de chaînes de valeur enracinées localement et partagées. Lorsque l'équipe y parvient, les conditions de vie locale s'améliorent, la dégradation des espaces recule et les émissions carbones avec elle. La recherche de l'équilibre des usages rend possible une transition juste.

En 2022, TotalEnergies a noué de nouveaux partenariats et contrats avec des acteurs reconnus au Gabon, au Pérou, en Asie du Sud-Est et au Guatemala. Le budget annuel consacré à ces projets est de 100 M de dollar. Le volume cumulé de crédits carbone espéré est de 45 millions à 2030 et 69 millions sur la durée de vie des projets. La réalisation des projets déterminera les quantités finales crédit carbone obtenues.

# 5 L'équipe

L'équipe TENBS est composée de différents membres :

- Le Vice-President Nature Based Solutions,
- La Directrice adjointe,
- Des Chefs de projets,

- Des Managers selon les régions des projets,
- L'Equipe technique appelée « Expertise carbone, foresterie » dans laquelle j'ai réalisé mon stage.

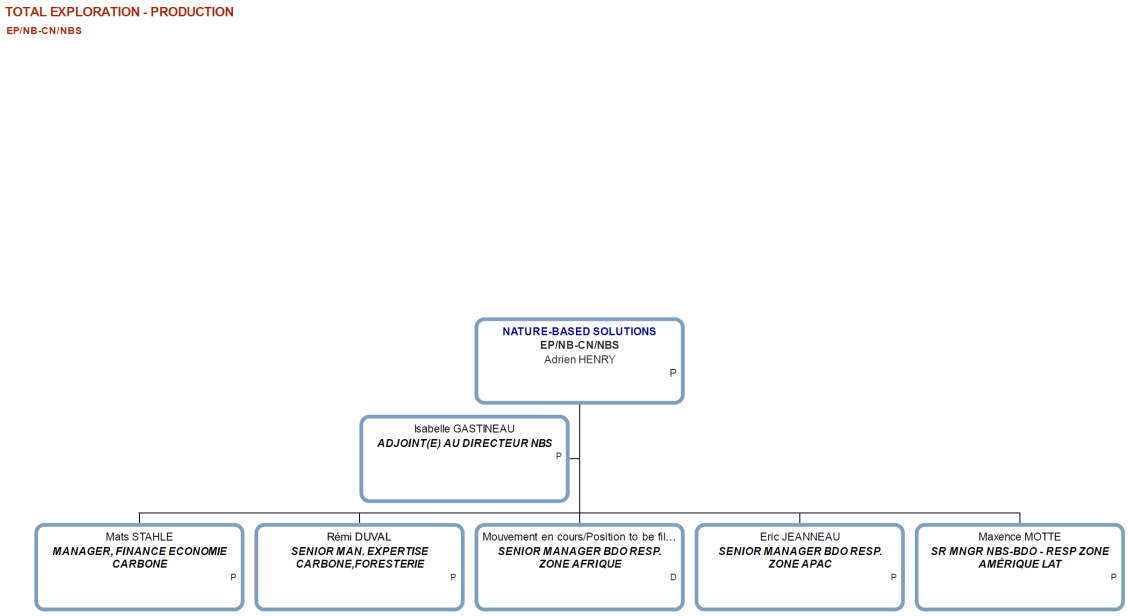


Fig. 5.1 : Organisation TENBS

L'équipe met du cœur à l'ouvrage pour trouver, analyser, investir et suivre des projets écologiques partout dans le monde qui répondent à l'objectif de la société.

Lors des réunions d'équipe, j'ai constaté combien l'équipe TENBS est très investie, réactive et dynamique pour rechercher activement tout à la fois de nouveaux projets et de nouveaux collaborateurs. De par ses projets internationaux, TENBS est amené à prospecter localement.

Lors de mon stage, j'ai travaillé plus particulièrement avec l'équipe technique. Elle est composée de de personnes exerçant les fonctions suivantes :

- Le Manager sur l'expertise carbone et foresterie,
- Les Experts en carbone, foret et agriculture
- L'analyste Environnement, Social et Gouvernance

TOTAL EXPLORATION - PRODUCTION
EP/NB-CN/NBS

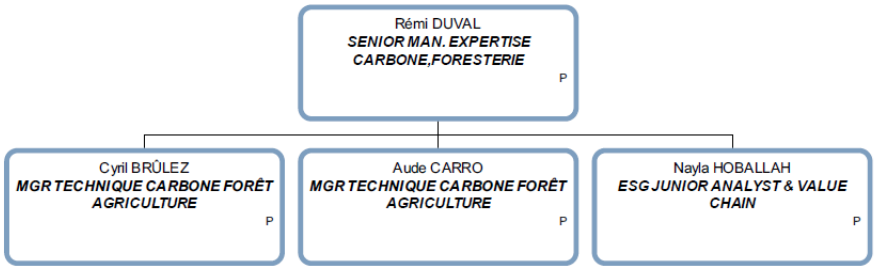


Fig. 5.2 : Organisation de l'équipe Technique

# 6 Répartition du travail

Ma tutrice de stage est Aude Carro. Elle est experte en carbone forêt et agriculture. Elle a été mon interlocuteur privilégié pour comprendre les besoins métier et cerner les données à extraire. Nous avons fait des points hebdomadaires spécifiques pour suivre l'avancée de mes travaux. De plus j'étais invitée aux réunions de l'équipe. Au sein de TENBS, Grégorie Fabre (chef de projet) avait mené un premier projet d'extraction de donnée du registre Verra. J'ai repris son code afin de l'adapter au mieux aux besoins métier. Comme mon stage a débuté début juillet, je n'ai pas pu être mise en relation avec un membre de l'équipe de développeur pour connaitre les outils internes qui auraient pu m'aider dans mon code. J'ai donc fait une méthode open source.

# 7 La démarche et Planning

La démarche suivie :


Fig. 7.1 : Etapes du stage

Mon stage s'étant déroulé en été, la présence des experts a été entrecoupée. La mission de ce stage était initialement prévue pour 6 mois. La durée du stage a été réduit à 3 mois car l'entreprise ne fait pas de contrat de Stage alternée (la rentrée M2 entraine un changement de rythme). La courte durée du stage a été une contrainte importante.

# 8 Conclusion

TENBS est un service jeune. Disposer d'une base conséquente d'informations sur les projets carbones AFOLUs mondiaux opérés et en développement constitue un atout important pour identifier les caractéristiques des projets à plus haut potentiel. Ce faisant, ils concourent à la bonne tenue des objectifs de TotalEnergies de réduction des émissions de GES en vue d'atteindre la neutralité carbone en 2050.

# Chapitre 2

# Le problème à résoudre

## Sommaire

# 1 Introduction

l'objectif métier est d'investir dans des initiatives de compensation carbone qui reposent sur la préservation et/ou l'augmentation de la capacité de stockage de carbone des agroécosystèmes. Le service TENBS cherche à explorer comment des initiatives concrètes, axées sur la réduction des émissions, peuvent être non seulement bénéfiques pour l'environnement, mais aussi viables sur le plan économique et socialement responsables. Cette quête vise à apporter des réponses à une question cruciale : comment atteindre la neutralité carbone de TotalEnergies d'ici 2050 ?

Dans cette démarche, TENBS utilise principalement le Standard Verra pour vérifier les réductions d'émissions des projets dans lesquels il investit. Le bénéfice tiré est l'obtention de crédits carbones.

Un crédit carbone est une unité de mesure représentant une réduction d'une tonne de gaz à effet de serre( tCO2), utilisée pour compenser les émissions de carbone en réduisant ou en évitant des émissions équivalentes. Un crédit carbone est calculé par les experts métier de la manière suivante : facteur d'activité par facteur d'émission. Les facteurs d'activité correspondent à la quantité d'activités responsables des émissions (exemple : couper des arbres, utiliser des carburants, pratiquer un type de culture donné). Les facteurs d'émission correspondent à la quantité d'émission par unité d'activité (tCO2e/ha d'arbre coupé, tCO2e/L de carburant utilisé, tCO2e/ha mis en agriculture d'un type donné). Les bases de ces calculs sont fournies par le Groupe Inter- Gouvernemental d'Experts sur le Climat (GIEC). Tous les standards carbones volontaires ou nationaux s'y réfèrent.

Verra a été créé dans les années 2000 pour permettre à des individus, des compagnies et/ou des gouvernements de mesurer et d'échanger sur le marché dit « volontaire » des crédits carbone non soumis à la régulation par les marchés réglementaires tels que l'European Trading Scheme (ETS). Cela signifie que ces réductions d'émissions sont réalisées sur une base volontaire et non en raison d'une obligation légale. Verra administre un site web comprenant un registre de tous les projets validés ou en cours d'enregistrement, ainsi que les réductions d'émissions vérifiées par le Standard Verra pour chaque projet.

Un projet Verra est déclaré sur le site Verra par le promoteur du projet à l'aide une description de projet à l'aide du modèle VCS (Verified Carbon Standard) Project Description (VCS PD). Le VCS PD décrit tous les détails du projet et de l'activité du projet, y compris l'emplacement, la date de début, la période de crédit du projet, la propriété des réductions d'émissions et l'estimation de la réduction d'émission.

Le promoteur du projet doit faire valider la description du projet par un organisme de validation/vérification approuvé. Verra valide/vérifie le projet présenté.

Une fois qu'un projet commence, le promoteur du projet commence à surveiller et à mesurer les réductions d'émissions et d'autres données pour produire des rapports de surveillance à l'aide du modèle de rapport de suivi VCS. Une fois le rapport de surveillance terminé, le promoteur du projet peut sélectionner un organisme de validation/vérification pour vérifier les réductions d'émissions générées et déclarées. Les promoteurs du projet peuvent choisir de terminer le projet.

Les projets AFOLUs (Agriculture, Foresterie et Autres Usages des Terres) générant des crédits carbones consistent en la plantation d'arbres, la gestion durable des forêts, des pratiques agricoles durables, l'utilisation de techniques agricoles efficaces et des mesures pour réduire les émissions de GES. Les projets AFOLUs déclarés sur Verra comportent des informations utiles pour le service TENBS.

Le projet que je développe vise à collecter un ensemble de données cibles sur les projets AFOLUs mondiaux déclarés en anglais dans le registre du Standard Verra. Cette base de connaissance va permettre de mieux comprendre comment les paramètres bio-physiques (climat, altitude, bio- région), socio-économiques (pays, type de filières d'intervention) et techniques (développeur de projet, type d'intervention, écosystèmes ou espèces protégées/plantées) influent sur l'estimation de la productivité des projets, mesurée en tCO2 par an. Cette analyse permettra d'élargir la base de connaissance mobilisée par l'équipe technique pour évaluer les projets d'investissement étudiés par TENBS.

## 2 Objectif technique

L'objectif technique de mon stage est de prédire un score de potentiel de production de crédits carbone sur des projets AFOLUs.

Données en Entrée :

Les données proviennent du registre du Standard Verra, qui répertorie les projets de réduction des émissions de carbone dans les secteurs de l'agriculture, de la foresterie et des utilisations des terres.

Données en Sortie :

Ces données sont ensuite utilisées pour développer un modèle permettant d'estimer la productivité des projets carbone en tCO2/an. Ce modèle aidera l'équipe de TENBS à évaluer l'efficacité des projets lors de la phase de due diligence.

L'objectif technique du projet se résume en plusieurs étapes :

- Comprendre les besoins métier et identifier les variables clefs décrivant les projets listés sur le registre du Standard Verra,
- Collecter les données des projets AFOLUs mondiaux publiés en anglais sur le site Verra,
- Les archiver dans une base de données,
- Nettoyer les données collectées,
- Analyser les données,
- Faire un modèle de prédiction de potentiel de production de crédits carbone sur des projets carbone,
- Présenter les données collectées, les résultats et l'outil de visualisation Tableau à l'équipe,
- Optimiser : Donner des pistes d'amélioration.

# 3 Données

## 3.1 Listes des données

Avec le service TENBS, j'ai identifié les attributs clefs décrivant les projets déclarés dans le registre du Standard Verra. J'ai ainsi pu collecter deux types donnés des projets AFOLUs :

- Les données générales,
- Les données détaillées.

### 3.1.1 Données générales

Il y a 16 colonnes représentant des données générales, et en fonction de l'évolution du projet, il peut y avoir un nombre variable (N) de mises à jour, chacune ajoutant 6 attributs supplémentaires.

| Données générales |
|---|
| ID |
| Nom de projet |
| Adhérent |
| Statuts |
| Type de projet |
| Région |
| Pays |
| Latitude |
| Longitude |
| Surface |
| Date de début de projet |
| Date de fin de projet |
| Période du projet |
| Estimation annuelle de la réduction carbone |
| Estimation totale de la réduction carbone |
| Estimation carbone productivité par an |
| Date de début de mise à jour du projet N |
| Date de fin de mise à jour du projet N |
| Période de mise à jour du projet N |
| Productivité de crédit carbone de mise à jour N |
| Quantité de crédit carbone de la mise à jour N |
| Certification de la mise à jour N |

Fig. 3.1 : Données générales

### 3.1.2 Données détaillées

Il y a 18 colonnes représentant les données détaillées.

| Données détaillées |
| --- |
| Projet de groupe |
| Écosystème |
| Générique nom Écorégion |
| Méthodologie |
| Catégorie |
| Sous-catégorie AFOLU |
| Agriculture |
| Produit forestier |
| Poisson nom commun |
| Poisson nom scientifique |
| Poisson nom générique |
| Espèce |
| Pluie annuelle |
| Altitude |
| Type de sol |
| Type de sol Dominant |
| Nom du rapport PDF |
| Répertoire du rapport |

Fig. 3.2 : Données détaillées

## 3.2 Acquisition des données

### 3.2.1 Acquisition des données générales

J'ai collecté les données générales de chaque projet grâce à l'API Verra. J'envoie une requête d'informations concernant un projet à l'API Verra et elle retourne une réponse sous forme de JSON.

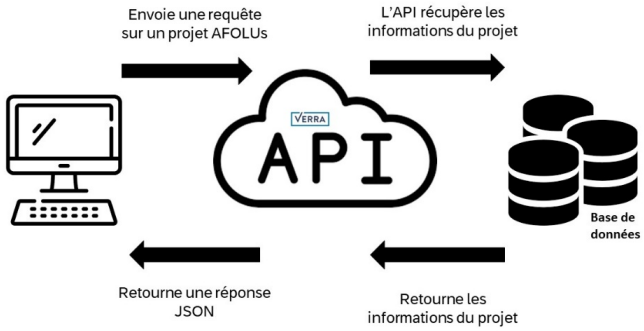


Fig. 3.3 : Schéma pour acquisition des données générales

Après la réception du JSON, je sélectionne les données générales cibles et les archives dans la base de données.

### 3.2.2 Acquisition des données détaillées

J'ai fait un programme python qui télécharge les rapports de description (Project Description ou PD). Les rapports de description me permettent d'obtenir les données détaillées. Pour cela, j'ai envoyé une requête à l'API Verra qui me renvoie tous les fichiers des différents répertoires du projet.

J'ai rencontré plusieurs problèmes pour identifier les bons les rapports de description de chaque projet :

- Les projets AFOLU n'ont pas le même statut d'avancement. Certains sont en cours de validation et d'autres ont déjà commencé depuis plusieurs années. En fonction de leurs statuts, les rapports PDF de description n'ont pas tous été rédigés,
- Les rapports de description ne sont pas tous rédigés en anglais. Il a été décidé de traiter que les projets rédigés en anglais,
- Le nom des fichiers des rapports de description n'est pas homogène. J'ai adapté mon programme pour qu'il supporte différents nommages,
- Il peut exister plusieurs rapports de description pour un même projet dans différents répertoires. Je devais prendre la version la plus récente,

- Les rapports de description peuvent être au format PDF ou DOCX.

Sur les 1066 projets AFOLUs, il y a 55 projets pour lesquels je n'ai pas trouvé de rapport de description en anglais. Pour faire l'analyse de données, j'ai gardé les projets qui ont des rapports de description en anglais soit 1012 projets.

Une fois les rapports de description téléchargés pour chaque projet, j'ai fait un programme python qui en extrait le texte brut pour son exploitation ultérieure. Le texte est sauvegardé dans un fichier .txt.

Les rapports de description ne suivent pas tous le même plan, car il en existe quatre versions différentes. Par conséquent, j'ai dû chercher des données détaillées au sein d'un texte qui n'était pas structuré.

Pour les 1012 projets j'ai pu extraire des rapports de description les données suivantes :

Une mesure selon un contexte :

- Pluviométrie annuelle : exemple "The average annual rainfall on the coast of Sindh amounts to about 200 mm".
- Altitude de la zone de projet : exemple « the Andes Mountains between 1000 and 2000 meters ».

Entités nommées :

- Type d'écosystème,
- Espèces,
- Type de sol,
- Marchandises,
- Méthodologie,
- Catégorie.

Avant d'opérer la recherche pour une entité nommée, il est essentiel de définir clairement une base connaissance. La base de connaissance contient les termes qu'on cherche à identifier dans le document. Plusieurs bases de connaissances m'ont été fournies par ma tutrice de stage sous forme de fichier Excel. Il y a 6 bases de connaissances contenant un nombre variable d'entrées :

Plus la base de connaissance est riche, plus l'extraction des informations est complète.

| | nombre d'instances dans la base de connaissance |
|---|---|
| Type d'écosystème | |
| Ecoregion | 847 |
| Generic ecosysteme name | 82 |
| Espèces | 57915 |
| Marchandises | |
| Agriculture Crop Name | 227 |
| Forest Product name | 54 |
| Fishery Nom Commun | 9970 |
| Fishery Nom Scientific | 13417 |
| Fishery genreric name | 51 |
| Type de sol | 26 |
| Les méthodologies | 41 |
| AFOLU catégories | 6 |

Fig. 3.4 : Nombre d'instances différents dans les bases de connaissances

En raison de contraintes de temps limitées pendant mon stage, j'ai dû simplifier le processus d'extraction de données en vérifiant simplement la présence ou l'absence des entités nommées dans le texte. Cependant, avec davantage de temps, il aurait été préférable d'appliquer le traitement automatique du langage naturel (NLP) pour une compréhension plus approfondie du contexte.

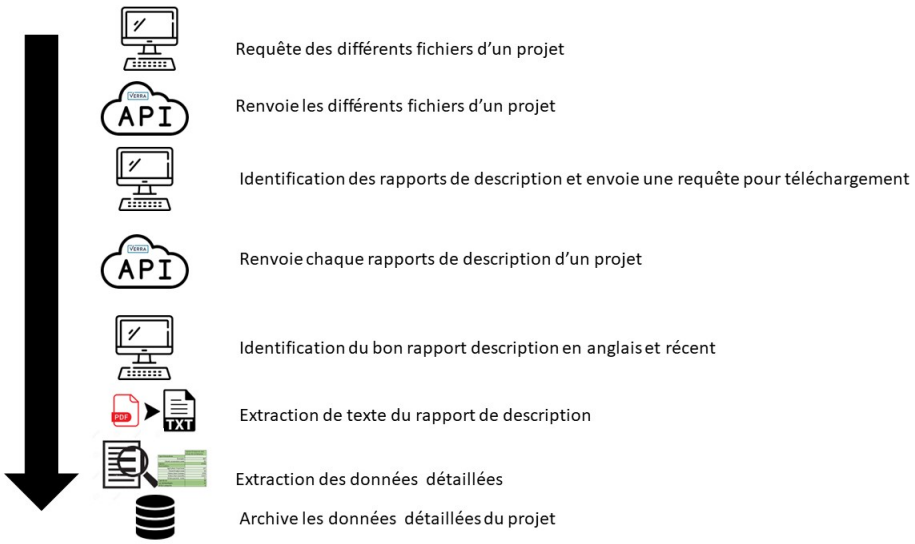


Fig. 3.5 : Schéma pour acquisition des données détaillées

Les 1012 projets comportent 18 données générales et 16 données détaillées. En fonction des mises à jour effectuées, on obtient N x 6 informations supplémentaires.

J'ai stocké l'ensemble des données dans une base de données.

Les données deviennent par la suite exploitables de multiple façon.

Une version Excel a été sortie pour le service TENBS.

# 4 Analyse des colonnes

Le tableau ci-dessous montre le nombre de données présentes et manquantes pour chaque colonne pour tous les projets AFOLUs.

| Nom de colonne | Remplis | Manquantes |
|---|---|---|
| Date de début de projet | 1066 | 0 |
| Date de fin de projet | 1066 | 0 |
| Période du projet | 1066 | 0 |
| Estimation annuelle de la réduction carbone | 1066 | 0 |
| Estimation totale de la réduction carbone | 1066 | 0 |
| Latitude | 1065 | 1 |
| Longitude | 1065 | 1 |
| Surface | 1065 | 1 |
| Estimation carbone productivité par an | 1065 | 1 |
| Catégorie AFOLU | 1061 | 5 |
| Région | 1041 | 25 |
| Nom du rapport PDF | 1012 | 54 |
| Répertoire du rapport | 1012 | 54 |
| Type de sol | 972 | 94 |
| Adhérent | 937 | 129 |
| Agriculture | 935 | 131 |
| Produit forestier | 929 | 137 |
| Générique nom Ecorégion | 921 | 145 |
| Groupe | 823 | 243 |
| Pluie annuelle | 700 | 366 |
| Sous-catégorie AFOLU | 662 | 404 |
| Espèce | 546 | 520 |
| Altitude | 531 | 535 |
| Poisson nom commun | 516 | 550 |
| Poisson nom générique | 481 | 585 |
| Date de début de MAJ | 248 | 818 |
| Date de fin de MAJ | 248 | 818 |
| Total de réduction carbone - MAJ | 248 | 818 |
| Période MAJ | 248 | 818 |
| Co-Certification MAJ | 114 | 952 |
| Ecosystème | 168 | 898 |
| Poisson nom scientifique | 161 | 905 |
| Type de sol Dominant | 107 | 959 |

Fig. 4.1 : Nombre de types différents dans les bases de connaissances

Les données ont été recueillies à partir de différentes sources, notamment l'API Verra , des rapports de description du projet et de plusieurs bases de connaissance.

# 5 Ingénierie des caractéristiques

Lors de l'ingénierie des caractéristiques, plusieurs considérations doivent être prises en compte :

## 5.1 Élimination des données inutiles :

Les caractéristiques qui ne contribuent pas de manière significative à la prédiction du résultat souhaité peuvent être éliminées. Cela peut être dû à un manque de corrélation avec la cible, ou à une absence de variation dans les données. L'élimination de caractéristiques inutiles réduit la complexité du modèle et peut améliorer les performances tout en accélérant l'apprentissage. Dans notre jeu de données, certaines sont éliminées comme :

- Celles qui avaient un manque de données trop conséquent : « Type de sol Dominant », « Écosystème », les données sur les poissons et les mises à jour de projets.
- Celles inutiles : « ID », « Nom de projet », « Adhérent » et « Groupe », nom et répertoire du rapport PDF.

## 5.2 Détection des Caractéristiques Redondantes :

Dans le jeu de données, certaines sont éliminées, car les données sur les dates de début et de fin sont corrélées à la période. La suppression des homographes et des termes similaires est une étape importante dans le prétraitement des données en traitement automatique du langage naturel. Cela peut aider à améliorer la qualité des données et à éviter les ambiguïtés lors de l'analyse du texte. Les données suivantes ont un traitement spécifique pour être exploitables :

- Type de sol : suppressions des termes similaires comme « sand - sandy », « silt - silty », « loam - loamy ». Transformation des différentes terminologies en un seul terme.
- Produit forestier : suppression du terme « paper » qui peut signifier à la fois un document écrit et une feuille de papier, en fonction du contexte dans lequel il est utilisé. Le terme « feuille de papier » est une marchandise qui peut être produite dans certains projets écologiques.
- Agriculture : suppression du terme avec « dates » qui représente soit le temps soit le fruit.

## 5.3 Ajout de Caractéristiques Manquantes :

Certaines caractéristiques importantes manquent, il peut être nécessaire de les ajouter. Cela est le cas pour des informations cruciales pour la prédiction. Dans notre jeu de données, certaines données sont ajoutées dans les colonnes comme Région, Pluie annuelle, Altitude et Catégorie.

## 5.4 Étude de Corrélation :

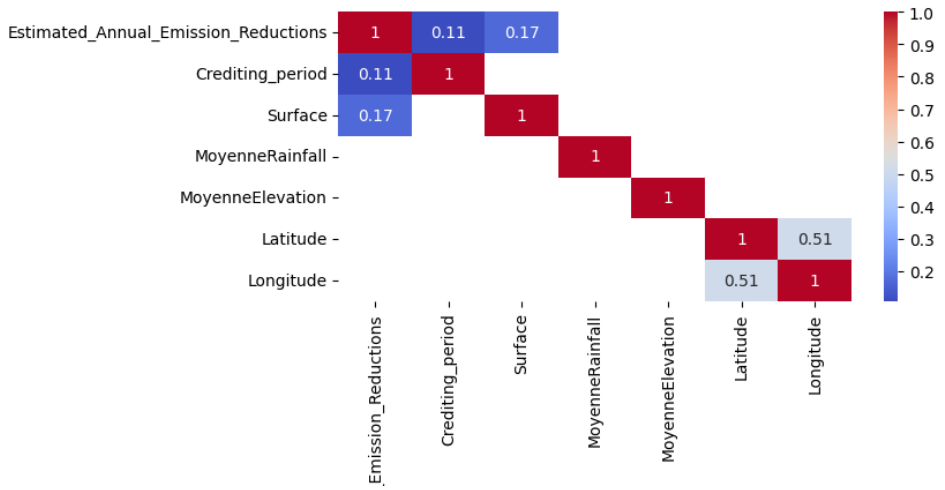


Fig. 5.1 : Corrélation

L'analyse de corrélation met en évidence la complexité des données en montrant des relations non linéaires entre les variables. Il est nécessaire de conserver toutes les données.

Pour résumer, l'ingénierie des caractéristiques vise à optimiser la qualité et la pertinence des données d'entrée pour les modèles d'apprentissage automatique. Cela implique l'élimination des caractéristiques inutiles ou redondantes, la sélection de caractéristiques pertinentes, l'étude des corrélations et l'ajout de caractéristiques manquantes. De ce fait, un processus d'ingénierie des caractéristiques bien exécuté contribue à la création de modèles plus précis, plus interprétables et plus performants.

Pour faire mon analyse de données, j'ai réduit les lignes à 1012 et les colonnes à 16. Ce sont des colonnes spécifiques et influentes pour le modèle de prédiction. Dans ce tableau les prétraitements ont déjà été fait. Vous trouverez ci-dessous le tableau montrant la proportion des données :

| Nom de colonne | Remplis | Vide |
|---|---|---|
| Méthodologie | 1012 | 0 |
| Estimation annuelle de la réduction carbone | 1012 | 0 |
| Catégorie | 1012 | 0 |
| Surface | 1012 | 0 |
| Période du projet | 1012 | 0 |
| Estimation carbone productivité par an | 1012 | 0 |
| Pluie annuelle | 1012 | 0 |
| Altitude | 1012 | 0 |
| Latitude | 1012 | 0 |
| Longitude | 1012 | 0 |
| Région | 1012 | 0 |
| Pays | 1012 | 0 |
| Type de sol | 972 | 40 |
| Agriculture | 935 | 77 |
| Écorégion | 921 | 91 |
| Produit forestier | 913 | 99 |
| Espèce | 546 | 466 |

Fig. 5.2 : Jeu de donnée d'analyse après prétraitements

# 6 Exploration et Visualisation des données après collectage

## 6.1 Exploration

La quantité de données manquantes pour certaines valeurs de ma base de données sont importantes (pour les espèces : 466 vides sur 1012 projets).

J'ai trouvé la raison de ce manque d'espèce renseigné. Dans les 466 projets il y a 233 projets, la moitié, qui sont des projets de culture de riz. Cela explique qu'il n'y a pas d'espèces d'arbres trouvées.

J'ai constaté pour l'autre moitié que :

- Ce sont des projets jeunes qui n'ont pas encore défini ce qu'ils vont planter, mais qui ont renseigné les types de produits forestiers qu'ils vont exploiter.
- Ou bien ce sont des projets liés à d'autres cultures.

## 6.2 Visualisation

Nous allons explorer les différents projets récoltés sur Verra. Les visualisations dans cette partie se font à l'aide de l'outil Tableau. Tableau est une puissante plateforme d'analyse exploratoire et de visualisation de données qui permet aux utilisateurs de transformer des données brutes en informations significatives et exploitables.

Nous pouvons observer la localisation des 1012 projets AFOLUs récoltés par des points sur la carte. La couleur indique la région de la zone de projet soit : l'Afrique, l'Asie, l'Amérique Latine, l'Europe, le Moyen-Orient, l'Amérique du nord et l'Océanie.


Fig. 6.1 : Localisation des projets AFOLUS par région du monde

On remarque que certains projets ont une latitude et longitude incorrecte. Ces projets se retrouvent parfois dans l'océan.

Dans l'exemple ci-dessous, on remarque que le pays est le Paraguay qu'il se trouve bien en Amérique latine mais que les positions longitude et latitude sont mauvaise. Pour ce projet, la longitude et latitude publié sur Verra sont mauvaise.

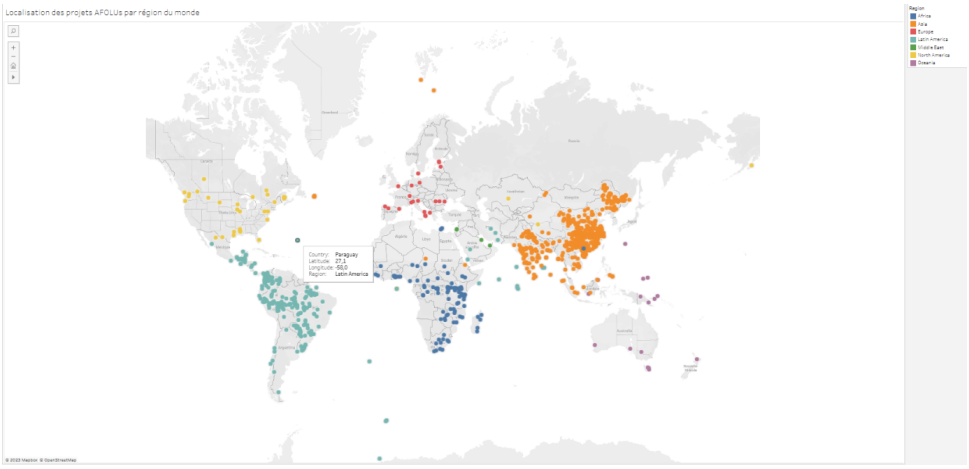


Fig. 6.2 : Localisation des projets AFOLUS par région du monde

On observe que pour les 1012 projets on a une répartition différente en fonction du statut d'avancement. On remarque que majoritairement les projets sont en attentes d'être validé par Verra.

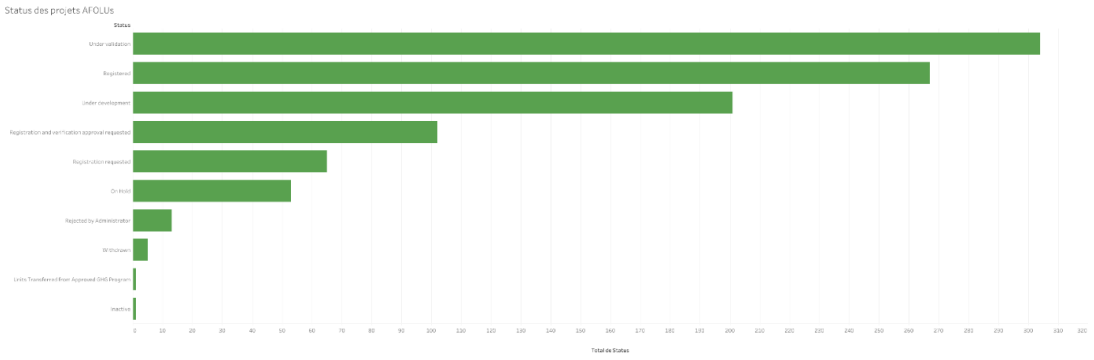


Fig. 6.3 : Status des projets AFOLUs

Pour les projets AFOLUs lancés on observe 218 Enregistrées, 27 en standby et 1 Unités transférées du programme de GES approuvé. Globalement on a 246 projets en opération et 766 en cours de développement.

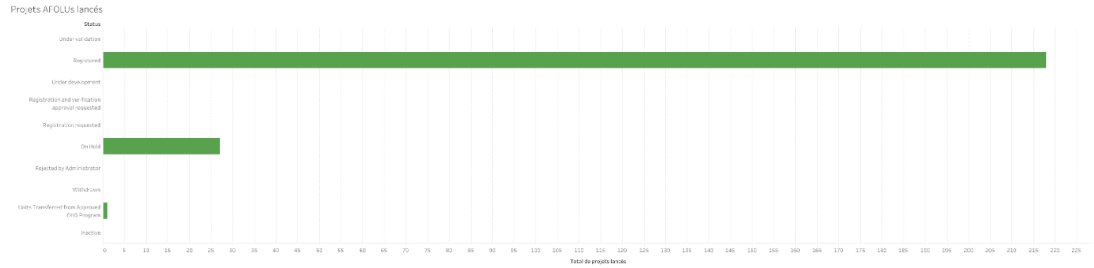


Fig. 6.4 : Projets AFOLUs lancés

La visualisation suivante montre, pour tous les projets lancés, la quantité de réduction des émissions annuelles pour chaque région. On observe que l'Amérique latine a la quantité la plus élevée de réduction des émissions annuelles.

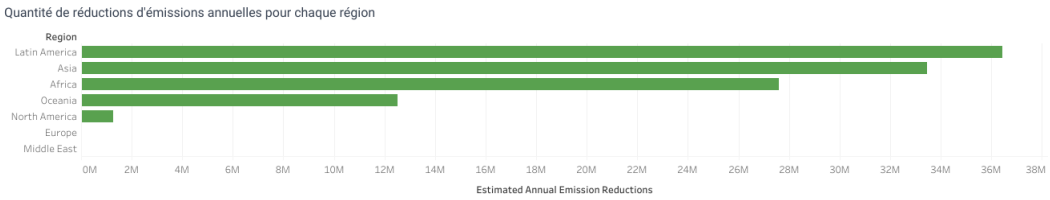


Fig. 6.5 : Quantité de réductions d'émissions annuelles pour chaque région

Un grand nombre de visualisations peuvent être faites grâce aux informations collectées. En manipulant les données on peut mieux comprendre et apprendre des différents projets AFOLUs existants.

# 7 Décomposition du problème

Pour pouvoir résoudre le problème, ce dernier a été découpé en sous-problèmes ordonnés dans une séquence logique d'étapes pour les résoudre. Cette décomposition et ordonnancement du problème est une étape essentielle dans la résolution.

Le problème Principal est :

> **La prédiction de l'estimation de la production de crédits carbones pour un projet spécifique.**

Pour cela un modèle prédictif va être produit.

Pour le produire, la séquence des étapes de résolution suivante va être faite :

1. Collecte
2. Prétraitement des Données : Nettoyage et Transformation
3. Analyse Exploratoire et Visualisation des Données
4. Modélisation et Entraînement : Sélection du Modèle et Division des Données et Entraînement
5. Évaluation et Amélioration

# 8 Conclusion

En résumé, la décomposition du problème en sous-problèmes, leur séquencement en étapes logique, l'évaluation continue et l'amélioration sont des éléments clés pour résoudre efficacement des problèmes complexes. Cette approche permet de mieux comprendre le problème, de le traiter de manière plus structurée et d'aboutir à des solutions de qualité.

# Chapitre 3

# État de l'art : Datascience pour de la prédiction

## Sommaire

# 1 Introduction

La prédiction du potentiel de production de crédits carbones d'un projet par les techniques de data science constitue un outil supplémentaire pour identifier les projets carbones les plus intéressants.

Dans cette partie, nous verrons les différentes méthodes pour résoudre le problème.

# 2 Algorithmes

Dans cette partie nous allons voir différentes méthodes pour répondre au problème :

- **Analyse de Texte et de Données Non Structurées**
- **Prétraitement et traitement des données collectées**
- **Algorithme de prédiction scoring**

## 2.1 Analyse de Texte et de Données Non Structurées

Les rapport PDF des projets AFOLUs publiés sur Verra fournissent des informations précieuses pour la prédiction des taux de réduction carbone. Les techniques de traitement du langage naturel peuvent extraire des informations pertinentes de ces sources et ainsi améliorer la fiabilisé de la capture en prenant en compte le contexte, les négations, les homonymes, ect. Voici quelques-unes des principales solutions et bibliothèques utilisées pour le traitement du langage naturel (NLP) :

**Bibliothèque en Python :**

- **NLTK (Natural Language Toolkit) :** Fournit des outils et des ressources pour le traitement du langage naturel, combinant des fonctionnalités de NLP avec la possibilité d'utiliser des expressions régulières pour le traitement de texte.
- **SpaCy :** Une bibliothèque NLP en Python qui offre des fonctionnalités avancées de traitement du langage naturel, tout en permettant l'utilisation de règles basées sur des expressions régulières pour effectuer des opérations de tokenization, d'extraction d'entités, etc.
- **Bibliothèques de Manipulation de Texte :** Outre les bibliothèques spécifiques de NLP, des bibliothèques de manipulation de texte plus générales de Python comme les regex peuvent être utilisées avec des techniques spécifiques de traitement de langage naturel.

## 2.2 Prétraitement et traitement des données collectées

### 2.2.1 Création de vecteurs de caractéristiques

Le "One-Hot Encoding" est une technique de représentation des données catégoriques sous forme de vecteurs binaires, où chaque catégorie unique est associée à une colonne avec des valeurs 0 ou 1 pour indiquer sa présence ou son absence. Cela permet aux algorithmes d'apprentissage automatique de travailler avec des données catégoriques.

Le "Label Encoding" est une technique d'encodage utilisée en science des données et en apprentissage automatique pour représenter des données catégorielles sous forme de valeurs numériques. Chaque catégorie unique de la variable catégorielle se voit attribuer une valeur numérique distincte dans un ordre séquentiel.

### 2.2.2 Normalisation

La normalisation des données numériques est une étape cruciale dans la préparation des données pour l'entraînement de modèles d'apprentissage automatique. L'objectif principal de la normalisation est de mettre toutes les variables à la même échelle, ce qui améliore les performances des algorithmes et garantit une convergence plus rapide lors de l'entraînement.

J'ai testé trois méthodes de normalisations ScikitLearn :

- **MinMAx :** La mise à l'échelle Min-Max transforme les valeurs d'une caractéristique pour qu'elles soient comprises dans une plage spécifiée, souvent [0, 1].
- **Z-score :** La normalisation z-score, également appelée standardisation, transforme les valeurs d'une caractéristique pour qu'elles aient une moyenne de 0 et un écart type de 1.
- **Robuste :** La normalisation robuste est similaire à la normalisation z-score, mais utilise la médiane au lieu de la moyenne et l'écart interquartile (IQR) au lieu de l'écart type.

## 2.3 Algorithme de prédiction scoring

Pour traiter des données complexes pour prédire le taux de réduction carbone, voici les différents algorithmes que je peux envisager d'utiliser en Python adaptés à mes données :

### 2.3.1 Arbres de Décision et Ensembles d'Arbres :

**"Decision Tree Regression"** est une technique d'apprentissage automatique utilisée pour prédire des valeurs numériques en utilisant un arbre de décision. Cet arbre divise les données en sous-groupes en fonction des caractéristiques, puis attribue une valeur de sortie à chaque feuille de l'arbre. Lorsqu'une nouvelle donnée est présentée à l'arbre, elle suit un chemin à travers les nœuds de l'arbre en fonction de ses caractéristiques pour aboutir à une valeur de prédiction numérique. Cette méthode est utile pour la modélisation de relations complexes entre les caractéristiques et les valeurs cibles dans les données numériques.

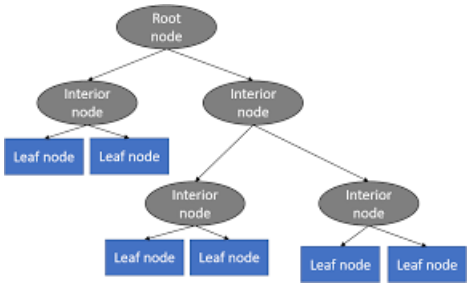


Fig. 2.1 : Decision Tree Regression

**Random Forest Regressor** est une technique d'apprentissage automatique qui est utilisée pour effectuer des prédictions en combinant les résultats de plusieurs arbres de décision. Chaque arbre de décision est construit sur une partie aléatoire des données d'entraînement et donne une prédiction. Ensuite, les prédictions de tous les arbres sont agrégées pour produire une prédiction finale. Cette approche permet généralement d'obtenir des prédictions plus robustes et moins sensibles au surapprentissage que les arbres de décision individuels. Random Forest est largement utilisé pour la classification et la régression, ainsi que pour la sélection de caractéristiques importantes dans les données.

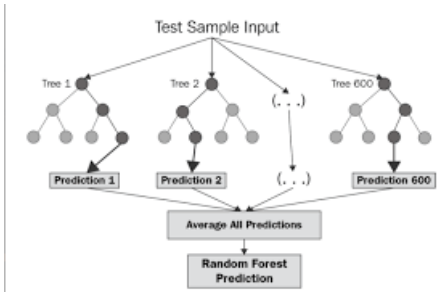


Fig. 2.2 : Random Forest Regressor

### 2.3.2 Réseaux de Neurones :

**"MLP Regressor"** (Multi-Layer Perceptron Regressor) est un modèle d'apprentissage automatique utilisé pour effectuer des tâches de régression, c'est-à-dire pour prédire des valeurs numériques continues. Il est basé sur un réseau de neurones artificiels à plusieurs couches, où chaque couche contient des neurones qui transforment les données d'entrée pour produire une valeur de sortie continue. Le MLP Regressor est capable de modéliser des relations complexes entre les caractéristiques d'entrée et la valeur de sortie souhaitée, en adaptant automatiquement ses poids et ses biais lors de l'apprentissage à partir des données d'entraînement. C'est un outil puissant pour la prédiction.

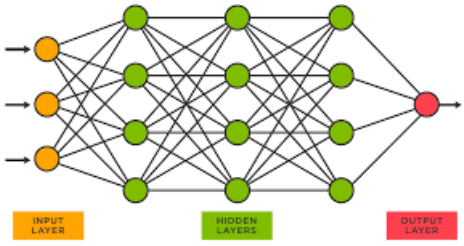


Fig. 2.3 : Multi-Layer Perceptron Regressor

### 2.3.3 Algorithmes de Régression :

La régression linéaire( **Linear Regressor**) est une technique d'apprentissage automatique qui cherche à modéliser la relation linéaire entre une variable de sortie et une ou plusieurs variables d'entrée, même lorsque les données sont complexes. Elle utilise une formule linéaire pour représenter cette relation et ajuste les coefficients de manière à minimiser l'erreur de prédiction. Cela permet de prédire une valeur numérique continue, même lorsque les relations entre les variables sont plus nuancées. Pour des données complexes, d'autres méthodes plus sophistiquées peuvent être nécessaires, mais la régression linéaire reste un outil précieux pour de nombreuses applications.

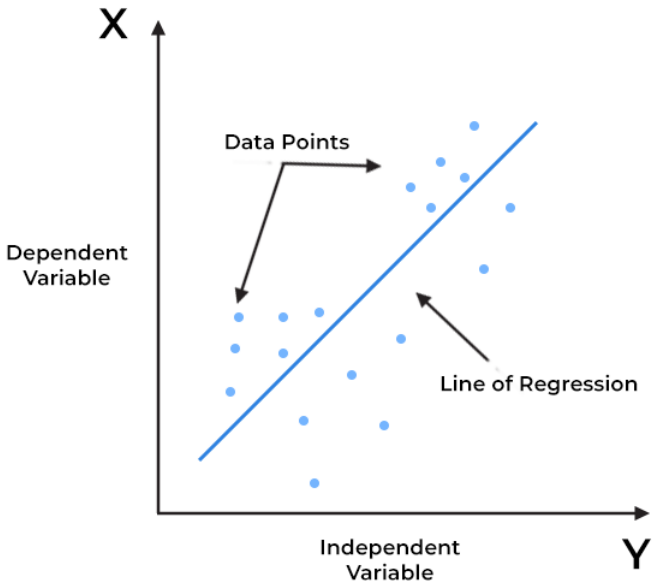


Fig. 2.4 : Régression linéaire

# 3 Conclusion

En conclusion, il est crucial de reconnaître que la résolution de problèmes en data science ne se limite pas à l'application d'un seul algorithme ou d'une seule approche. Différentes solutions peuvent répondre de manière variable à la complexité des problèmes. Les choix doivent être éclairés par une comparaison approfondie et une étude des différentes méthodes disponibles, en tenant compte à la fois de la précision, le délai imparti et du coût associé à chaque approche.

Je vais évaluer les performances de plusieurs algorithmes pour deux aspects de ma tâche :

Normalisation :

La comparaison entre les méthodes de normalisation telles que MiniMax, Robust et Z-Score est essentielle pour déterminer la meilleure approche en fonction des caractéristiques spécifiques des données et des besoins du modèle d'apprentissage automatique. Cette comparaison permet de prendre en compte des facteurs tels que la sensibilité aux valeurs aberrantes, l'impact sur la convergence du modèle, l'interprétabilité des données et leur impact sur la performance globale du modèle. En fin de compte, le choix de la méthode de normalisation dépendra de ces considérations et de la meilleure adéquation avec la tâche d'apprentissage automatisé en cours.

Algorithme de prédiction scoring :

Pour l'algorithme de prédiction, je vais explorer les modèles DecisionTree Regressor , RandomForest Regressor , MLP Regressor et Linear Regressor.

- **Linear Regressor :** La régression linéaire est un modèle simple mais puissant pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Elle suppose une relation linéaire entre les variables.
- **DecisionTree Regressor :** Les arbres de décision sont capables de capturer des relations non linéaires et des interactions entre les variables. Ils divisent récursivement les données en sous-groupes en fonction des caractéristiques.
- **Randomforest Regressor :** Les forêts aléatoires sont constituées d'ensembles d'arbres de décision. Chaque arbre est construit sur un sous-échantillon des données d'entraînement, et les prédictions finales sont agrégées à partir des prédictions de chaque arbre.
- **MLP Regressor :** Les réseaux de neurones, en particulier les réseaux de neurones profonds, sont des modèles puissants capables de capturer des relations complexes et non linéaires

J'ai opté pour l'expérimentation de ces méthodes mentionnées précédemment afin de déterminer quel modèle est le mieux adapté à mes données complexes.

# Deuxième partie

# Système réalisé

# Table des matières

# Chapitre 4

# Implémentation du système

## Sommaire

# 1 Introduction

Le système que j'ai réalisé comporte des développements pour couvrir les étapes suivantes :

- Collecte et Exploration des données,
- Prétraitement / Traitements des données,
- Séparation des données,
- Algorithme de prédiction,
- Évaluation du modèle,

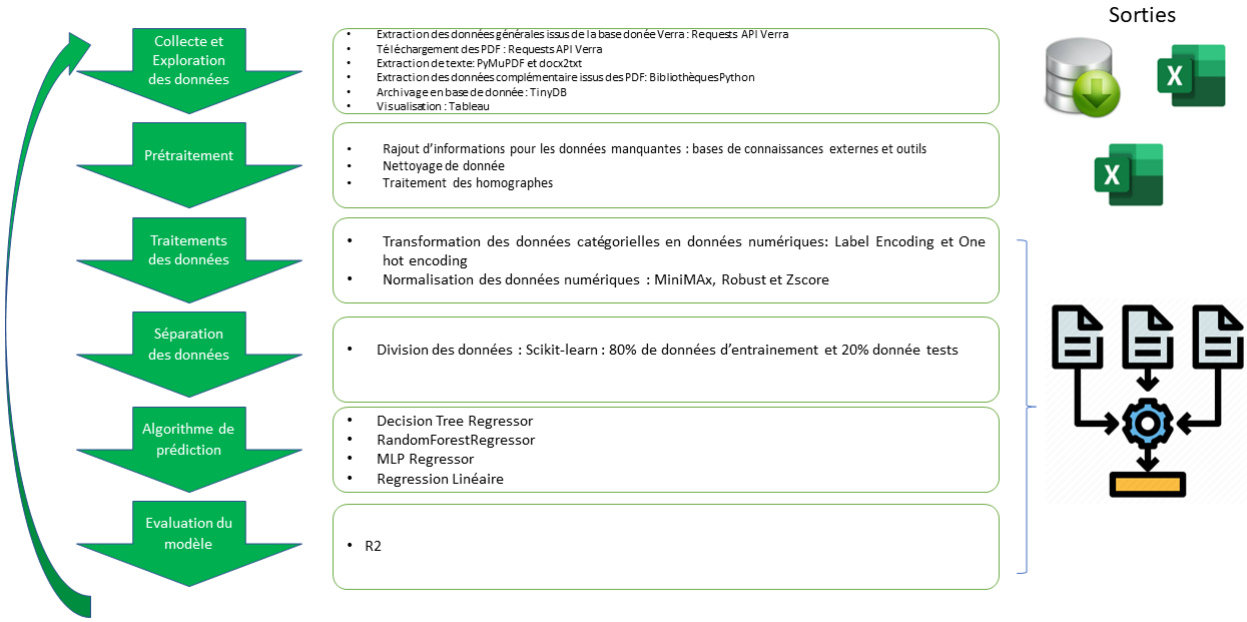


Fig. 1.1 : Le systeme

# 2 Collecte et Exploration des données

Pour extraire des données générales et télécharger les rapports de description de projet de la base de données Verra, j'ai utilisé "requests" sur l'API Verra. La

bibliothèque "requests" en Python permet de faire des requêtes vers des sites. Elle est utilisée pour obtenir des données depuis Internet, envoyer des informations à des serveurs web, et interagir avec des API. Elle simplifie la communication entre le programme Python et le monde extérieur via HTTP.

Pour extraire le texte des rapports de description, j'ai utilisé la bibliothèque PyMuPDF pour les rapports PDF et docx2txt pour les rapports DOCX.

Pour extraire les données détaillées des rapports de description j'ai cherché les entités nommées dans le texte. Pour extraire une mesure selon un contexte j'ai utilisé des regex.

Les données extraites sont ensuite archivées dans une base de données à l'aide de TinyDB, une base de données NoSQL en Python, qui nous permet de stocker et de gérer efficacement ces informations pour une utilisation ultérieure.

Enfin, pour visualiser les données, j'ai utilisé Tableau, une plateforme de visualisation de données qui me permet de créer des tableaux de bord interactifs et informatifs pour mieux comprendre et analyser les données extraites et archivées dans la base de données TinyDB.

# 3 Prétraitements et Traitements des données

Le prétraitement des données est important, car il permet d'atteindre les points suivants : qualité des données, modèles plus précis, réduction des erreurs, temps de traitement réduit, consistance et simplification.

Dans le pré-traitement des données, j'ai fait les différentes étapes suivantes :

- Rajout d'informations pour les données manquantes,
- Nettoyage de donnée et transformation des homographes.

Dans le traitement des données, j'ai fait les différentes étapes suivantes :

- Transformation des données catégorielles en données numériques,
- Normalisation des données numériques.

## 3.1 Prétraitements

### 3.1.1 Nettoyage de donnée et suppression des homographes :

Le nettoyage de données consiste à préparer des données brutes en supprimant les erreurs, les doublons et les valeurs inutiles pour les rendre utilisables. La suppres-

sion des homographes est le processus de distinction des mots qui s'écrivent de la même manière mais ont des significations différentes en fonction du contexte.

### 3.1.2 Rajout d'informations pour les données manquantes :

L'ajout de nouvelles données provenant de sources externes est une pratique courante pour enrichir et améliorer la qualité de l'ensembles de données. Cela est particulièrement bénéfique pour des analyses plus approfondies et pour entraîner des modèles d'apprentissage automatique plus précis.

Les données suivantes ont un traitement spécifique pour être complétées :

- **Traitement de la pluie annuelle :** Pour répondre au manque d'information pour cette donnée, j'ai collecté sur « Our World in Data » les pluies annuelles en moyennes de chaque pays.
- **Traitement de l'altitude :** j'ai utilisé l'API « The National Map - Elevation Point Query Service » qui renvoie l'altitude en mètres pour une latitude et longitude spécifique. J'ai pu confirmer grâce à des visualisation et cette API, que certaine donné de longitude et latitude sont mauvaises. L'API me renvoie 0 altitude lorsque le projet est dans l'océan.
- **Traitement des catégories :** Les catégories sont déterminées par rapport à une méthodologie. On peut compléter cette donnée grâce à une base de connaissance que ma tutrice m'a renseignée. Ce tableau indique les catégories possibles liées à une méthodologie spécifique.
- **Région :** Pour compléter cette donnée je m'appuie sur le pays du projet et je complète à l'aide des connaissances issues des autres projets. Pour les exceptions, j'ai constitué une base de connaissances.

Cas particulier :

Ajouter des données de manière inappropriée peut altérer la signification et la validité des informations contenues dans ces données. C'est pour cela que j'ai gardé les données telles qu'elles sont pour les données suivantes :

- Agriculture
- Espèces
- Produits forestiers
- Écorégion
- Type de sol

## 3.2 Traitements

### 3.2.1 Transformation des données catégorielles en données numériques

Transformer des données catégorielles en données numériques est un processus essentiel dans le domaine de la science des données et de l'apprentissage automatique. Cela permet de préparer les données pour l'entraînement du modèle prédictif.

Les données catégorielles ont un traitement par l'encodage « One Hot Encoding » :

- Type de sol
- Catégorie
- Agriculture
- Méthodologie
- Écorégions
- Espèces
- Produits forestiers

Chaque colonne de données contient une liste d'éléments, et chaque élément de cette liste est transformé en une représentation binaire distincte en utilisant la technique "One-Hot Encoding". Ces nouvelles colonnes binaires sont ensuite intégrées dans mon ensemble de données, ce qui entraîne une augmentation de la taille du nombre de colonnes.

Le "Label Encoding" a été appliqué à une colonne indiquant les régions. Cela signifie que chaque région unique dans cette colonne a été associée à une valeur numérique unique.

### 3.2.2 Normalisation des données numériques

Les données normalisées par MinMax, Robust et Zscore méthodes sont les suivantes :

- Période du projet,
- Surface,
- Pluie annuelle,

- Altitude,
- Longitude Latitude,
- Estimation annuelle de la réduction. carbone

Pour conclure, le prétraitement des données est essentiel pour garantir que les données sont fiables, pertinentes et de haute qualité. Le prétraitement des données joue un rôle clé dans la création de modèles précis, la prise de décisions éclairées et l'obtention d'informations significatives à partir des données brutes.

# 4 Séparation des données

Afin de vérifier l'efficacité de mes modèles de Machine Learning, le jeu de données initial est divisé en deux ensembles : un training set et un test set. Le training set est utilisé pour entraîner le modèle sur une partie des données, tandis que le test set est utilisé pour évaluer les performances de ce modèle sur l'autre partie des données. La fonction `train_test_split` de la bibliothèque Scikit-Learn (`sklearn`) de Python permet de réaliser cette séparation en deux ensembles.

J'ai effectué la séparation des données en allouant 80% au training set et 20% au test set.

J'ai choisi un nombre pour la variable "random_state" afin d'assurer la reproductibilité du code. Le fait de choisir un nombre entier pour "random_state" garantit que les données sont divisées de la même manière à chaque appel de la fonction, ce qui rend le code reproductible. Pour déterminer cette valeur, j'ai exécuté mes différents algorithmes jusqu'à obtenir l'ensemble de données d'entraînement et de test avec le meilleur score d'évaluation.

La variable cible de ma prédiction est la colonne "Estimation carbone productivité par an".

# 5 Algorithme de prédiction

Pour répondre au problème suivant : **Potentiel de production de crédits Carbone pour un projet en cours de TotalEnergies** . J'ai mis en œuvre 4 algorithmes de prédiction qui sont issus de la librairie Sklearn :

- **Régression lineaire :** sans spécifier de paramètres supplémentaires.
- **DecisionTree Regressor :** sans spécifier de paramètres supplémentaires.

- **Randomforest Regressor :** 1000 arbres, une profondeur maximale de 200 pour chaque arbre, et il exploite tous les cœurs de processeur disponibles pour accélérer le processus d'entraînement.
- **MLP Regressor :** avec deux couches cachées contenant respectivement 200 et 100 neurones et il limite l'entraînement à 1000 itérations.

# 6 Évaluation du modèle

J'ai utilisé R2 afin d'évaluer si mon modèle est correct :

- **R2 (Coefficient de Détermination) :** C'est une mesure statistique qui évalue la proportion de variance expliquée par un modèle par rapport à la variance totale des données. En d'autres termes, il indique à quel point les prédictions d'un modèle s'ajustent aux valeurs réelles. R2 varie de 0 à 1, où 1 signifie que le modèle explique parfaitement la variance des données, et 0 signifie qu'il n'explique aucune variance. Plus R2 est proche de 1, meilleure est la qualité du modèle.

# 7 Conclusion

En conclusion, le prétraitement des données est essentiel pour garantir la qualité des données et améliorer la performance des modèles de machine learning. J'ai appliqué des techniques de nettoyage, de suppression des homographes, d'enrichissement des données manquantes, de transformation des données catégorielles en numériques, et de normalisation. Les données ont été divisées en ensembles d'entraînement et de test pour évaluer les modèles, et quatre algorithmes ont été utilisés pour prédire la production de crédits carbone. L'évaluation s'est basée sur le coefficient de détermination R2 pour mesurer la précision des modèles.

# Chapitre 5

# Expérimentations et résultats

## Sommaire

# 1 Tests sur les algorithmes et prétraitements

Avec la méthode d'évaluation R2, je vais présenter les résultats obtenus des 4 algorithmes de prédiction suivants différentes méthodes de normalisation. Le tableau ci-dessous représente les résultats obtenus pour le jeu de données contenant les 1012 projets.

| | Prétraitement | | |
|---|---|---|---|
| **Algorithme de prédiction** | **Minmax** | **zscore** | **robust** |
| Régression Linéaire | ~0 | ~0 | ~0 |
| DecisionTreeRegressor | 0,99 | 0,99 | 0,99 |
| RandomForestRegressor | 0,89 | 0,89 | 0,85 |
| MLP Regressor | 0,44 | 0,69 | 0,85 |

Fig. 1.1 : Représentation des résultats obtenus pour le jeu de données contenant les 1012 projets

On observe pour ce jeu de donné que la meilleure combinaison est la méthode Robust, Minimax ou Robust avec DecisionTree Regressor avec un taux de précision à 99

Pour ce jeu de donnée les algorithmes RandomForest Regressor et MLP Regressor combiné à Robust donnent aussi des bons résultats en fonction de prétraitement.

Au contraire on voit que la Régression Linéaire donne une précision négative pour chaque pré-traitement. On peut en déduire que le modèle n'est pas adapté.

# 2 Déduction globales

Je constate que le pré-traitement Robust, MiniMax et Zscore combiné à DecisionTree Regressor donne les meilleurs résultats.

Je constate que la régression linéaire n'est pas adaptée à notre jeu de donnée. On peut en déduire qu'il n'existe pas une relation linéaire entre les variables de notre jeu de donnée.

Cependant on a globalement des bons résultats pour DecisionTree Regressor, RandomForest Regressor et MLP Regressor combiné au pré-traitement Robust. Ces 3 algorithmes de prédiction sont meilleurs car ils sont tous les 3 capables de

modéliser des relations complexes entre les caractéristiques d'entrée et la variable cible. Cela leur permet de capturer des modèles non linéaires et des interactions entre les variables.

# Conclusion

Durant ce stage complet et enrichissant au sein de l'entreprise TotalEnergies, je suis ravie de partager les enseignements et les accomplissements qui ont marqué cette expérience passionnante dans le domaine des projets AFOLUs.

Durant ces deux mois, j'ai eu l'opportunité de mettre en œuvre et d'approfondir mes compétences en programmation en travaillant sur diverses thématiques comme : la collecte d'informations sur différentes sources, le nettoyage de données, les prétraitements, les visualisations avec Tableau, les analyses ainsi que de la prédiction. J'ai pu mettre en pratique les connaissances théoriques et pratiques acquises au cours de ma formation universitaire et les adapter aux problématiques réelles de l'entreprise.

J'ai appris à travailler efficacement au sein d'une équipe multidisciplinaire, à l'écoute et à la compréhension du besoins métier, à planifier la prestation dans les 3 mois, à communiquer mes idées de manière claire et à gérer les priorités dans un environnement dynamique.

Les résultats obtenus sont également gratifiants en seulement 2 mois. Mon projet a permis à l'ensemble de l'équipe TENBS d'être sensibilisé aux étapes et au potentiel des sciences des données.

Ce projet peut évoluer. Il peut être amélioré sur les points suivants :

- Rendre dynamique la collecte et le stockage des données lorsqu'un nouveau projet est publié sur Verra,
- Augmenter la base de données en rajoutant des projets inscrits sur le registre du Gold Standard,
- Optimiser la collecte des données en utilisant une méthode NLP,
- Faire un algorithme qui détermine les colonnes les plus représentatives pour la réduction carbone,
- Faire un algorithme qui détecte les potentielles valeurs aberrantes et confirmer avec l'équipe technique si c'est bien une valeur aberrante ou bien un cas

exceptionnel,

- Faire de la validation croisée.

En conclusion, ce stage chez TotalEnergies a été une expérience formatrice. Il a renforcé ma conviction de carrière en tant que Data Scientiste et m'a donné l'assurance nécessaire pour aborder avec confiance les défis futurs. Je tiens à exprimer ma gratitude envers toute l'équipe qui m'a accueilli et guidé ainsi que mon Université Paris 8 pour l'ensemble des connaissances et projets réalisés qui m'ont permis de pouvoir produire dans ces délais. Je reste enthousiaste à l'idée de continuer à apprendre et à contribuer dans ce domaine en constante évolution

# Glossaire

| | |
|---|---|
| **ACoGS** | Avoided Conversion of Grasslands and Shrublands |
| **AFOLU** | Agriculture, Foresterie et Autres Usages des Terres |
| **ALM** | Agricultural Land Management |
| **API** | Application programming interface |
| **ARR** | Afforestation Reforestation and Revegetation |
| **COMEX** | Comité exécutif |
| **GES** | Gaz à effet de serre |
| **IEEE** | Institute of Electrical and Electronics Engineers |
| **IFM** | Improved Forest Management |
| **JSON** | JavaScript Object Notation |
| **TENBS** | TotalEnergies Nature Based Solutions NBS |
| **NLP** | Natural language processing |
| **REDD** | Reduced Emissions Deforestation and Degradation |
| **R2** | Coefficient de détermination |
| **VCS** | Verified Carbon Standard |
| **WRC** | Wetlands Restoration and Conservation |

# Table des figures

# Table des matières