UNIVERSITE DON BOSCO DE LUBUMBASHI

Faculté des Sciences Informatiques

Département Génie Logiciel

Lubumbashi

www.udbl.ac.cd

MISE EN PLACE D'UN MODELE PREDICTIF DES DONNEES EPIDEMIOLOGIQUES

Cas du choléra dans la province du Haut-Katanga

Par : KALOMBO TSHIYA Peter

Filière : Génie Logiciel

Directeur : Yves Ndeturuye

Octobre 2024

EPIGRAPHE

‘‘Les données sont le langage de l’avenir.’’

Auteur anonyme

DEDICACE

A Jéhovah, Dieu créateur et omnipotent ; à ma famille, merci pour tout.

REMERCIEMENTS

À l'issue de notre parcours académique à l'Université Don Bosco de Lubumbashi (UDBL), il nous incombe de rendre hommage à toutes les personnes qui ont contribué de manière significative à notre cheminement et à la réalisation de ce mémoire.

Avant tout, nous tenons à exprimer notre profonde gratitude au Dieu Tout-Puissant pour sa grâce infinie, sa miséricorde et la santé qu'il nous a accordées tout au long de notre parcours à l'UDBL, nous permettant ainsi de mener à bien ce travail.

Nous adressons nos remerciements les plus profondes à l'Évêque Albert Tshomba, pasteur de la bergerie de Carrefour du MET CEIVV Come and See, église où nous prions et servons. Nous sommes profondément reconnaissants pour le temps précieux qu'il consacre à prier pour nous et à nourrir nos dons spirituels, contribuant ainsi grandement à notre réussite académique depuis son frais début. Nos remerciements vont également à son épouse, Maman Lina Tshomba.

Je tiens à exprimer ma gratitude infinie à mon père Placide Tshiya et à ma mère Annie Sompo, dont l'amour inconditionnel, le soutien indéfectible et les innombrables sacrifices ont été la pierre angulaire de ma réussite. Leur dévouement sans bornes et leur encouragement constant m'ont porté à travers chaque défi. À mes frères : Morgan, Désiré et Josué Tshiya et à mes sœurs : Claine, Odile et Kevine Tshiya, merci pour votre soutien inlassable, vos prières, vos encouragements sans fin et votre foi inébranlable en moi. Vous êtes tous des piliers de ma vie, et ce mémoire est autant le vôtre que le mien.

Nous adressons également nos remerciements à l'ensemble du corps administratif et professoral de l'Université Don Bosco de Lubumbashi pour leur engagement quasi infaillible dans notre formation.

De manière particulière, nos remerciements s'adressent à Monsieur Yves Ndeturuye, directeur de ce travail, dont le temps précieux et l'expertise ont grandement contribué à la réalisation de ce mémoire.

A mon cher beau-frère Parfait Kashala, à mes oncles, tantes, cousins, et cousines dont le soutien et l'amour constants ont été une source de force et d'inspiration. Bien que je ne puisse citer tous vos noms, sachez que chacun de vous a joué un rôle crucial dans mon parcours. Vos encouragements et votre présence bienveillante ont grandement contribué à la réalisation de ce travail.

Nous souhaitons également exprimer profondément ma reconnaissance envers la Tenke Fungurume Mining, le fonds social communautaire de Fungurume et la dotation TFM pour la prise en charge financière qui a rendu ce cursus académique possible.

Un grand merci à mes collègues et amis, Méschac Irung, Alain Kayumba, Adalbert Pungu et Yves Kalume, pour leur influence positive sur moi et dont les recommandations et les discussions stimulantes nous ont ouvert des opportunités. Votre confiance en mes capacités et vos encouragements constants m'ont poussé à travailler davantage pour améliorer mes compétences.

Je remercie sincèrement tous mes amis d'enfance, avec qui je suis encore ami jusqu'à aujourd'hui. Merci à Romaric Kyangwa, Théodore Useni, Prosper Kitoko, Chrisostome Simba, Jiresse Makenda, Narthia Sasilwa, Dorcas Mabika, Céline Kasenga, Joëlle Merita et bien d'autres. Votre amitié durable et votre constant ont été une source de joie et de force tout au long de mon parcours. Merci pour les souvenirs partagés et pour être toujours présents à mes côtés.

Enfin, je tiens également à exprimer ma profonde gratitude à mes frères et sœurs en Christ de la chorale 'La Lumière du Monde'. Merci à Bethany Tshikut, Patient Mudimbi, Lumière Luya, Jean-Pie Muamba, Esther Luya, Rosie Maya, Confedie Luya, Live Vidi, aussi bien qu'à tous les autres membres de ma chorale dont je n'ai pas cité les noms. Ensemble, nous avons vécu des moments inoubliables et partagé des expériences extraordinaires. Vous avez été une véritable famille pour moi, offrant soutien et amitié tout au long de ce parcours.

Que chacun de vous ressente la sincérité de ma gratitude, gravée à jamais dans mon cœur. Votre soutien et votre amour ont été les piliers de mon parcours, et pour cela, je vous en serai éternellement reconnaissant.

LISTE DE FIGURES

LISTE DE TABLEAUX

LISTE D'EQUATIONS

LISTE D'ACRONYMES

IA : Intelligence Artificielle

ML : Machine Learning

CRISP-DM : Cross-Industry Standard Process for Data Mining

SK-LEARN : Scientific Python Toolkit-learn

ARIMA : Autoregressive Integrated Moving Average

SARIMA : Seasonal Autoregressive Integrated Moving Average

RNN : Autoregressive Integrated Moving Average

CNN : Convolutional Neural Network

FNN : Feedforward Neural Network

REST : Representational State Transfer

API : Application Programming Interface

LSTM : Long Short-Term Memory

SVM : Support Vector Machine

NLP : Natural Language Processing

GPU : Graphics Processing Unit

DAX : Data Analysis Expressions

DHS : Demographic and Health Surveys

WASH : Water, Sanitation, and Hygiene

GDPR : General Data Protection Regulation

CCPA : California Consumer Privacy Act

DNS : Division Nationale de la Santé

DPS : Division Provinciale de la Santé

CTC : Centre de Traitement du Choléra

UTC : Unité de Traitement du Choléra

ONU : Organisation des Nations Unies

OMS : Organisation Mondiale de la Santé

MSF : Médecine Sans Frontières

UNICEF : United Nations International Children's Emergency Fund

USAID : United States Agency for International Development

ZS : Zone de Santé

UDBL : Université Don Bosco de Lubumbashi

TABLE DE MATIERES

AVANT-PROPOS

Dans ce 21e siècle marqué par l'essor de l'intelligence artificielle, les données sont devenues une ressource inestimable, souvent comparée à l'or noir. Elles sont au cœur de nombreuses innovations et permettent de prendre des décisions éclairées dans divers domaines. Dans le cadre de cette étude, les données fournies par le ministère de la Santé ont été essentielles. Elles ont permis de mener des analyses approfondies sur la fluctuation épidémique du choléra découverte en RDC et aussi dans la province de Haut-Katanga après observations et recherches, dans le but développer des modèles prédictifs grâce aux techniques de machine learning. En tant que patriote et passionné de la data science, nous avons ressenti le besoin de contribuer à la lutte contre cette épidémie meurtrière sous un angle un tant soit peu différent.

Ce mémoire propose une solution innovante utilisant les données cholériques des six dernières années et le machine learning pour améliorer les décisions en santé publique, tant préventives que curatives. L'objectif est de fournir des analyses prédictives pour anticiper les futures épidémies de choléra, pouvant ainsi protéger les communautés et sauver des vies. Ce projet démontre l'importance de l'analyse des données et leur transformation en informations précieuses pour élaborer des stratégies efficaces. Il montre également le potentiel des données pour apporter des solutions innovantes dans divers domaines, tels que la santé publique, la gestion des crises, le développement durable et la planification urbaine.

L'obtention des données cholériques a été la plus grande des difficultés de ce projet. Elles sont coûteuses et il était difficile de savoir auprès de quelle institution les obtenir. Après des investigations et des démarches auprès de diverses institutions telles que les hôpitaux généraux de différentes communes, le Centre de Traitement de Lubumbashi, et la Division Provinciale de la Santé, nous avons finalement pu rassembler les informations nécessaires pour mener à bien ce travail.

Nous tenons à exprimer notre gratitude à tous ceux qui ont tous ceux qui ont contribué à la réalisation de ce projet, notamment les agents du Ministère Provincial de la Santé, et collègues. Leur soutien et leur collaboration ont été d'une aide précieuse.

Nous vous invitons donc, à découvrir la mise en place de cette solution innovante, conçue avec passion et détermination, dans l'espoir qu'elle inspire de nouvelles perspectives afin de palier à certains phénomènes malheureux qui frappent notre communauté.

INTRODUCTION GENERALE

Généralités

Depuis des millénaires les épidémies font partie de fléaux les plus meurtriers que l'humain ait jamais connus. Ayant comme origines la transmission zoonotique, les conditions environnementales, la mutation des pathogènes, l'hygiène et les conditions sanitaires, les voyages et le commerce... Les épidémies sont aussi dangereuses que les guerres car depuis leur apparition, elles ont causé des pertes humaines comparables ou mêmes supérieures à celles des guerres les plus meurtrières.

Les guerres continuelles, les voyages incontrôlés, le taux de croissance de la population, et la situation économique et politique du pays sont des facteurs majeurs qui ne permettent pas la maîtrise de la situation épidémiologique au sein de notre pays en dépit de toute l'aide précieuse que l'on reçoit des organismes internationaux tels que l'Agence des États-Unis pour le développement international (USAID), l'Organisation Mondiale de la Santé (OMS), les Fonds des Nations Unies pour l'enfance (UNICEF), etc.

Et en ce qui concerne la province du Haut-Katanga, située au sud-est de la République Démocratique du Congo (RDC), les épidémies les plus régulières sont le paludisme, le choléra, la rougeole et la poliomyélite, mais celles qui ont emporté plus de vies sont le choléra et la rougeole. Les rapports numériques des 27 zones de santé sont envoyés quotidiennement à la Division Provinciale de la Santé (DPS) pour analyse et prise de décision. Quant à l'épidémie de choléra, les Centres de Traitement du Choléra (CTC) et les Unités de Traitement du Choléra (UTC) la traitent et envoient également des rapports quotidiens à la DPS.

Il est clair que de nos jours, la numérisation et la centralisation des données médicales ne suffisent pas pour la gestion des épidémies. Pourtant, avec l'avènement des technologies de l'information et l'augmentation exponentielle des données disponibles, de nouvelles opportunités se sont ouvertes pour améliorer notre compréhension et notre gestion des épidémies. Le machine learning offre des outils puissants pour analyser de grands ensembles de données et prédire les tendances.

C'est ainsi que ce travail scientifique explore l'application du machine learning à l'épidémiologie, en se concentrant sur le développement de modèles prédictifs capables d'anticiper les épidémies. En utilisant des données historiques, environnementales et comportementales, notre objectif est de créer des modèles robustes capables d'aider les autorités sanitaires à prendre des décisions éclairées et à mettre en place des mesures préventives efficaces.

1. Problématique

Les rapports fournis par l'OMS ces 8 dernières années indiquent que la RDC n'a pas quitté le podium au niveau mondial ou continental des pays les plus touchés par le choléra. Nous soutenons cette affirmation par les chiffres suivants : en 2017, notre pays

a enregistré plus de 56000 cas dont plus de 1000 décès et était deuxième au monde derrière le Yémen [1] ; en 2018, une diminution d'environ 50% par rapport à l'année précédente a été constatée avec plus 23000 cas, 798 décès et 3,4% de létalité [2] ; en 2019, plus de 31000 cas dont 540 décès ont été enregistrés [3] ; en 2020, il y a eu également une diminution remarquable de presque 30% avec plus de 19700 cas, plus de 350 décès et le pays occupait la première place au monde [4] ; en 2021, les chiffres ont grimpé à plus de 52000 cas et 12561 décès, c'était un vrai hécatombe [5] ; en 2022, les chiffres ont baissé à plus de 18000 cas dont 302 décès [6] ; en 2023, une légère augmentation a été constatée avec plus de 52400 cas et 462 décès [7] ; et en cette année 2024, plus de 2500 cas et 18 décès sont enregistrés, ce qui représente une létalité de 0,7% [7].

Les statistiques ci-dessus démontrent clairement une fluctuation épidémique du choléra dans notre pays ces 8 dernières années alors que dans d'autres pays tels que le Yémen par exemple, l'épidémie a été totalement vaincue dans un écart de 2 ou 3 ans après son apparition. De plus, d'après nos observations et enquêtes au ministère provincial de la santé, il s'avère que la technologie n'intervient pas assez dans la lutte contre les épidémies malgré la numérisation et la centralisation des données. Cela constitue logiquement une cause majeure à ladite fluctuation d'autant qu'à l'ère actuelle, l'apprentissage automatique offre plus de possibilités.

Par conséquent, deux questions méritent d'être posées :

Comment tirer profit des données disponibles à la DPS dans le but de lutter contre l'épidémie de choléra dans notre pays ?

Et quel mécanisme pouvons-nous mettre en place pour prévenir les risques de fluctuation épidémique du choléra ?

2. Hypothèse

Face à ce défi qui s'impose dans notre société depuis des années, nous proposons les hypothèses ci-après :

- Créer un jeu de données sur lequel une analyse approfondie de données sera faite dans le but de surveiller la santé publique, de contrôler la maladie, d'évaluer des politiques de santé et de planifier des ressources ;

- Une mise en place d'un modèle d'apprentissage automatique qui donne des prédictions statistiques sur le choléra sur base des facteurs épidémiologiques et environnementaux, des comportements sociaux et des données de santé publique.

Nous émettons l'hypothèse que ce modèle prédictif permettra d'anticiper l'épidémie du choléra dans le but d'en mesurer l'impact afin d'aider les autorités à prendre des mesures préventives efficaces.

3. Choix et Intérêt du sujet

3.1 Choix du sujet

Etant chercheur, nous nous devons de produire un travail scientifique rigoureux, justifié, et validé dans notre domaine de formation. C'est ainsi que nous avons porté notre choix sur « Mise en place d'un modèle prédictif des données épidémiologiques ». Ce projet a pour objectif de susciter l'intérêt des décideurs dans le domaine médical à utiliser l'apprentissage automatique comme outils technologique dans la lutte contre les épidémies et il sera également mis à la disposition de tous pour de nouvelles perspectives.

3.2 Intérêt du sujet

Intérêt personnel

Passionné par la donnée, ce projet est une opportunité qui nous permet d'acquérir plus de connaissances dans le domaine de la science de données, l'une de plus grandes disciplines qui régissent le monde aujourd'hui. En outre, ce travail nous permet également de mettre en pratique les connaissances acquises le long de notre cursus académique.

Intérêt scientifique

Du point de vue scientifique, ce travail marque la fin de notre cursus académique et répond aux exigences de notre programme de formation, il constitue d'ores et déjà une référence bibliographique pour des générations futures. Ce travail est un atout car il illustre étape par étape l'application du machine learning à la médecine, et plus spécifiquement à l'épidémiologie.

Intérêt social

En analysant les données épidémiologiques et en simulant des prédictions sur les épidémies, il s'avère également que nous répondons à un besoin pertinent et important au sein de notre société. Ce qui fait à ce qu'étant patriote, ce projet nous permette également de contribuer du point de vue technologique à la lutte contre une épidémie qui cause des milliers de morts sur le territoire congolais depuis des décennies.

4. Méthodologie

Pour structurer notre projet d'analyses et prédictions sur les épidémies, nous avons porté notre choix sur la méthode CRISP-DM (CRoss-Industry Standard Process for Data Mining), qui est un modèle de processus standardisé pour les projets de data mining et de machine learning. Cette méthode est la plus utilisée en machine learning en raison de sa flexibilité et de son applicabilité à divers domaines et elle se subdivise en 6 phases qui sont : la compréhension des affaires, la compréhension des données, la préparation des données, la modélisation, l'évaluation et le déploiement.

Les données tant qualitatives que quantitatives obtenues lors des entretiens et analyses de données secondaires, constitueront la quintessence même de notre projet et nous seront d'un apport majeur à chaque étape de sa mise en place.

5. Etat de l'art

Cette étude se propose d'examiner les contributions significatives des mémoires académiques dans le domaine médical, en soulignant les innovations méthodologiques et les applications concrètes qui ont émergé au fil des années. Il s'agit notamment d'un étudiant de l'ESIS et d'un étudiant de l'Université Sultan Moulay Slimane du Maroc.

- Le travail de l'étudiant UMBA MUYOMBI, Big Data et analyse prédictive dans le domaine médical (Cas de l'hôpital de référence SENDWE de Lubumbashi), défendu en 2018 à l'Ecole Supérieure d'Informatique Salama, présente une application destinée aux agents de la santé, qui consiste à simuler des symptômes et à faciliter les recherches sur les données médicales des patients.

- Le projet de l'étudiant EL MASSARI HAKIM, Proposition d'un modèle de prédiction basé sur Machine Learning et le web sémantique, défendu en 2023 à l'Université Sultan Moulay Slimane à la faculté des sciences et techniques, présente un modèle qui prédit le covid-19 et le cancer du sein à partir des caractéristiques (symptômes) se trouvant dans un jeu de données issu de la plate-forme Kaggle.

Les travaux existants sur l'analyse et la prédiction ont été d'une aide significative dans l'application des technologies au domaine médical. Nous avons constaté que le premier sujet se concentre sur le développement d'un logiciel de numérisation et de recherche rapide des données médicales. Il vise à améliorer l'accès et la gestion des informations médicales en les numérisant, permettant ainsi une recherche plus efficace. Quant au second sujet, quoiqu'il s'agisse toujours de la prédiction, il aborde la prédiction médicale en utilisant une solution de classification basée sur les symptômes. Il vise à prédire des diagnostics à partir des symptômes observés. De plus, De plus, ce travail utilise des données de Kaggle, alors que des données locales auraient été plus pertinentes.

Pour remédier aux limitations de ces deux travaux, nous proposons de développer un modèle basé sur des données médicales réelles et structurées en série temporelle. Contrairement au premier travail, qui se concentre sur la numérisation des données médicales déjà résolue dans notre contexte, ce modèle permettra de prédire le nombre de cas futurs de choléra. De plus, contrairement au deuxième sujet qui utilise des données de Kaggle pour la classification des symptômes, notre modèle utilisera des données locales pour des prédictions plus précises et adaptées à notre région.

6. Délimitation du travail

Sur le plan spatial, nous nous sommes concentrés exclusivement sur les 27 zones de santé qui constituent notre province, le Haut-Katanga. Cette délimitation géographique

permet une analyse détaillée et contextualisée des phénomènes observées, en tenant compte des spécificités locales et des dynamiques propres à cette région dans laquelle nous vivons.

Cette étude faite au cours de l'année académique 2023-2024 à l'Université Don Bosco de Lubumbashi, couvre une période allant de 2018 à ce jour, permettant ainsi une analyse approfondie des données épidémiologiques sur une durée de six ans. Cette période a été choisie pour inclure les tendances récentes et les évolutions significatives de la science de données appliquée à la médecine ou plus précisément à l'épidémiologie.

7. Subdivision du travail

Mis à part l'introduction et la conclusion générale, ce travail porte sur 3 chapitres que voici :

- Chapitre 1 : Compréhension du cadre de recherche
- Chapitre 2 : Généralités sur le Machine Learning
- Chapitre 3 : Conception et Mise en place du Modèle de Machine Learning
- Chapitre 4 : Résultats et Discussions

8. Outils logiciels et équipements utilisés

Pour la réalisation de ce travail, nous avons utilisé plusieurs matériels et logiciels pour la récolte et l'analyse des données, la conception et l'implémentation, aussi bien que pour la rédaction. A savoir :

Clé USB : Un support de stockage amovible permettant le stockage et le transfert des données entre plusieurs appareils.

IPhone XI : Un smartphone conçu et commercialisé par Apple, dans ce travail il a servi à l'enregistrement vocale des interviews avec les agents de la santé publique.

Microsoft Excel : un logiciel tableur faisant partie de la suite Office, développé par Microsoft pour l'encodage, l'analyse et la visualisation des données.

Microsoft Word : un logiciel de traitement de texte faisant également partie de la suite Office, il a servi à la rédaction de ce travail.

Microsoft Power BI : un outil de Business Intelligence faisant partie de la Power Platform permettant l'analyse approfondie des données de diverses sources, la création des tableaux de bord interactifs, ainsi que le partage et la collaboration avec d'autres utilisateurs.

Lucid Chart : un outil en ligne permettant de concevoir des diagrammes.

Anaconda : une distribution libre et open source des langages de programmation Python et R, il nous a été utile pour l'analyse des données et l'entrainement du modèle.

Jupyter Notebook : une application web open source, qui permet de créer et des partager des documents appelés « notebooks » et dans ce travail, il nous a permis d'écrire et d'exécuter du code Python de manière interactive.

Git et GitHub : Git est un gestionnaire des versions des logiciels et GitHub un hébergeur de code source permettant également à des équipes de collaborer sur un projet.

CHAPITRE I COMPREHENSION DU CADRE DE RECHERCHE

Introduction partielle

Historiquement, le choléra a été responsable de plusieurs pandémies dévastatrices, causant des millions de décès à travers le monde. Aujourd'hui, il reste un indicateur poignant des inégalités sociales et du manque de développement dans certaines régions. En République Démocratique du Congo, et plus spécifiquement dans la province de Haut-Katanga, le choléra demeure une préoccupation sanitaire récurrente, nécessitant une compréhension approfondie et une réponse efficace.

Ce chapitre vise à fournir un cadre de recherche clair et détaillé pour l'étude du choléra dans notre province. Nous explorerons le contexte épidémiologique, la présentation du cadre de recherche avec ses facteurs géographiques et démographiques, ainsi que les politiques et stratégies institutionnelles en place pour lutter contre cette maladie. En établissant ce cadre, nous espérons non seulement mieux comprendre les dynamiques de l'épidémie de choléra, mais aussi identifier des interventions potentielles pour réduire son impact sur la population.

I.1 Contexte épidémiologique

Le choléra est une infection diarrhéique aiguë causée par l'ingestion d'eau ou d'aliments contaminés par la bactérie *Vibrio cholerae*. Cette maladie, bien que largement évitable et traitable, continue de représenter une menace majeure pour la santé publique, en particulier dans les régions où l'accès à l'eau potable et aux infrastructures sanitaires est limité. Le choléra est une maladie endémique en fluctuation en République Démocratique du Congo depuis plus d'une décennie. Cette région est régulièrement confrontée à des épidémies de choléra, souvent pendant la saison des pluies, lorsque les conditions sanitaires se détériorent et que l'accès à l'eau potable devient encore plus limité. Les figures 1 et 2 réalisées avec Power Bi, nous donnent un aperçu sur la situation cholérique dans notre pays ces six dernières années.

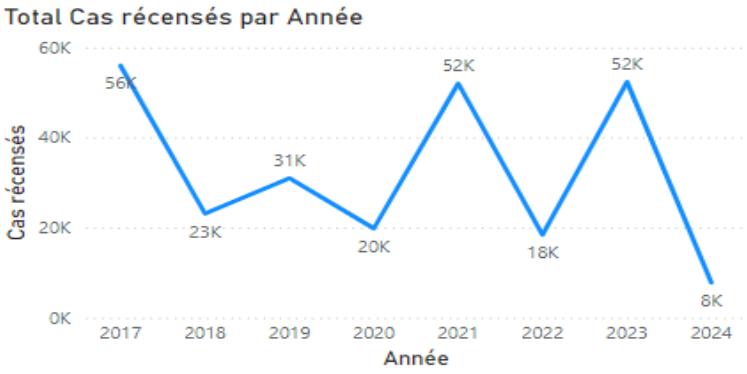
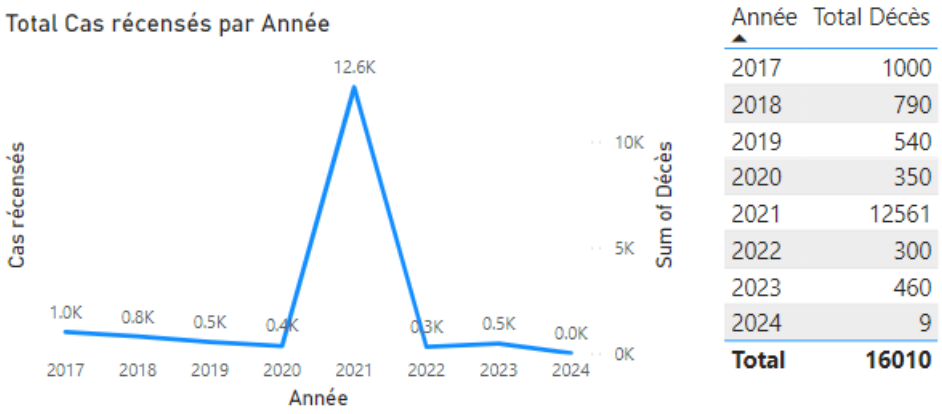


Figure I.1 Cas de Choléra enregistrés de 2017 à 2024



Année	Total Décès
2017	1000
2018	790
2019	540
2020	350
2021	12561
2022	300
2023	460
2024	9
Total	**16010**

Figure I.2 Décès Choléra enregistrés de 2017 à 2024

Depuis janvier 2024, une nouvelle épidémie de choléra a été déclarée dans la province, avec 678 cas et 29 décès enregistrés en l'espace de 4 mois. Les zones de santé les plus touchées sont celles de Kenya, Kikula, Sakania, Katuba, Kampemba, Kafubu, et Mumbunda. Ces zones sont caractérisées par une densité de population élevée et des infrastructures sanitaires insuffisantes, ce qui favorise la propagation rapide de la maladie. La figure 3 nous donne plus de détails comme le nombre de zones touchées, le nombre de total de cas et décès par zone de santé.

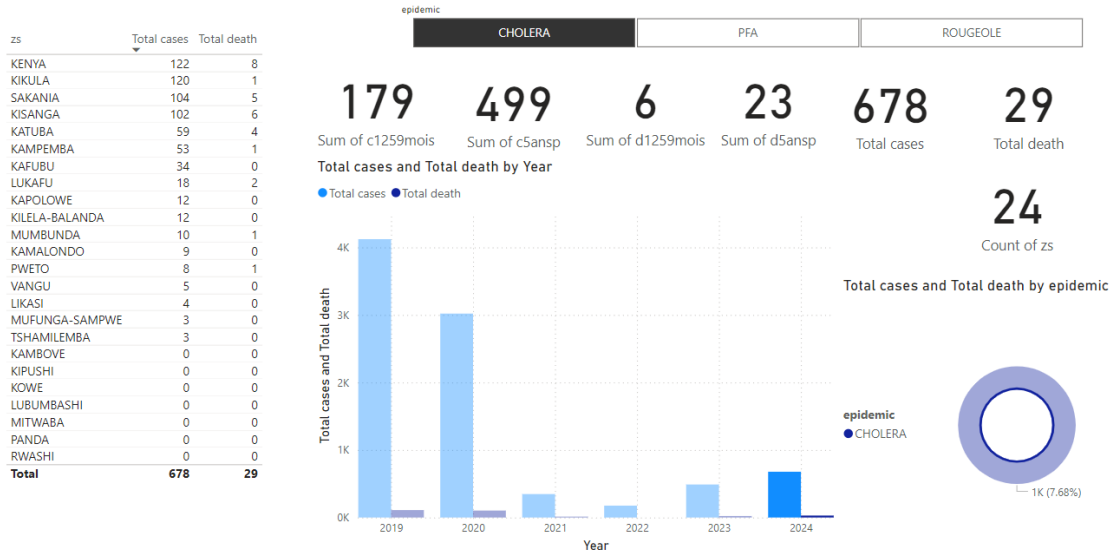


Figure I.3 Situation du choléra de janvier à avril 2024 au Haut-Katanga

Les autorités sanitaires provinciales, en collaboration avec des organisations internationales telles que les MSF, ont mis en place des mesures de contrôle pour contenir l'épidémie. Ces mesures incluent la sensibilisation de la population aux pratiques d'hygiène, la distribution de kits de purification de l'eau et l'amélioration des infrastructures sanitaires.

Malgré ces efforts, la lutte contre le choléra reste un défi majeur en raison des conditions socio-économiques précaires et du manque de ressources. La compréhension des dynamiques épidémiologiques et des facteurs de risque est essentielle pour développer des stratégies de prévention et de contrôle plus efficaces.

I.2 Cadre théorique

Il est primordial de savoir que le choléra se transmet principalement par l'eau contaminée (transmission hydrique), mais aussi par l'ingestion de nourriture ou d'eau contaminée par les matières fécales (transmission fécale-orale), car le vibrion cholérique survit aisément dans des environnements aquatiques sous l'influence des facteurs comme la température et la salinité.

Le climat et l'accès à l'eau potable sont des facteurs environnementaux majeurs qui influencent la propagation de l'épidémie. A cela s'ajoutent des facteurs socio-économiques tels que la pauvreté et l'urbanisation, qui rendent aussi à leur tour, la population vulnérable au choléra. Le modèle écologique prend en compte les interactions entre les facteurs environnementaux et les comportements humains pour comprendre la dynamique de l'épidémie.

Les stratégies de prévention et de contrôle utilisées à ce jour par la santé publique incluent l'amélioration des infrastructures d'eau et d'assainissement, la promotion de l'hygiène et la vaccination. Une approche intégrée combinant la vaccination avec des

interventions WASH est recommandée pour réduire la transmission de la maladie. La surveillance épidémiologique est essentielle pour détecter rapidement les flambées et mettre en place des mesures de contrôle appropriées.

I.3 Présentation du cadre

I.3.1 Cadre géographique et démographique

La province de Haut-Katanga, située dans le sud-est de la République Démocratique du Congo, présente une diversité géographique avec des zones urbaines, rurales et minières. Lubumbashi, la capitale provinciale, est densément peuplée et connaît une urbanisation rapide, souvent non planifiée, ce qui pose des défis en matière d'infrastructures sanitaires et d'approvisionnement en eau potable. Du point de vue médical, la province compte à ce jour 27 zones de santé qui sont : Lubumbashi, Katuba, Kenya, Kisanga, Kampemba, Kilela-Balanda, Kikula, Rwashi, Kamalondo, Kipushi, Likasi, Pweto, Kambove, Kafubu, Kilwa, Mumbunda, Tshamilemba, Vangu, Sakania, Kasenga, Kapolowe, Panda, Mitwaba, Mufunga-Sampwe, Kashobwe, et Kowe. Les zones rurales, bien que moins densément peuplées, souffrent également d'un accès limité aux services de santé et à l'eau potable.

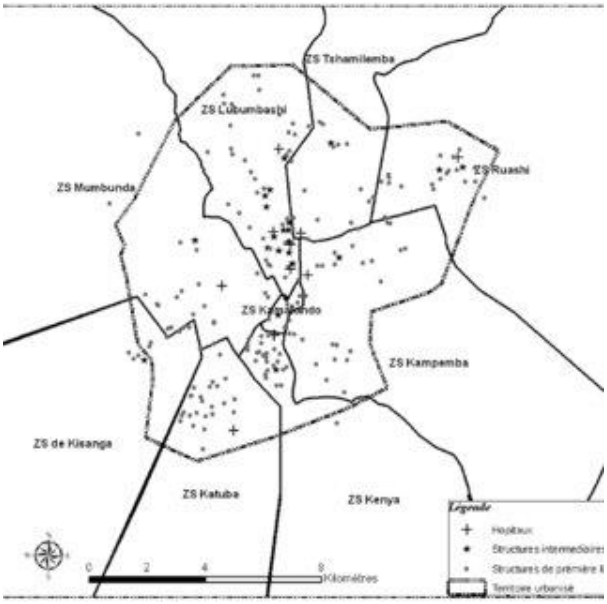


Figure I.4 Carte sanitaire de Lubumbashi, par Denis Porignon, carte sanitaire lubumbashi - Recherche Images (bing.com)

Les conditions socio-économiques de la province sont marquées par des inégalités, malgré la richesse en ressources naturelles. Beaucoup de gens vivent dans la pauvreté, ce qui augmente leur vulnérabilité aux épidémies de choléra. Les infrastructures sanitaires sont la plupart du temps insuffisantes, avec des systèmes d'assainissement rudimentaires et un accès limité à l'eau potable, surtout pendant la saison des pluies.

La mobilité de la population, en raison des activités minières et des migrations internes, joue un rôle crucial dans la propagation du choléra. Les mouvements de population peuvent introduire la maladie dans de nouvelles zones et compliquer les efforts de contrôle. La compréhension de ces aspects géographiques et démographiques permet de mieux cibler les interventions et de développer des stratégies efficaces pour prévenir et contrôler les épidémies de choléra dans notre province de Haut-Katanga.

I.3.2 Cadre institutionnel et politique

Les institutions de santé publique, notamment le ministère de la santé, la DPS, les CTC et les UTC, jouent un rôle crucial dans la gestion et la prévention des épidémies de choléra. Elles sont responsables de la surveillance épidémiologique, des campagnes de vaccination et de la distribution de kits de purification de l'eau. Dépendamment de la situation géographique, un CTC et/ou une UTC est implanté dans une ville ou un village pour la prise en charge des cas de choléra. La transmission des données dans ce système se passe de la manière qui suit :

— Dès qu'un seul cas de choléra est signalé, le ministère provincial de la santé déclare une épidémie. Cette déclaration déclenche l'installation des matériels et outils spécifiques pour la prise en charge et la surveillance de la maladie. C'est là qu'interviennent les CTC et les UTC.

— Les CTC et les UTC prélèvent et encodent les données sur les patients, puis envoient des rapports journaliers au format '*xls*' ou '*xlsx*' en chaque fin de journée à la DPS à des fins de stockage, d'analyses et de prises de décisions par les autorités de santé publique.

— Après certains traitements sur les données en provenance des CTC et des UTC grâce au logiciel DHS*2 Monitor* offert par l'OMS, la DPS analyse les données et élabore un rapport tous les dix jours et l'envoie à son tour à la DNS à Kinshasa. C'est assurément sur base cette analyse régulière que les mesures sont prises.

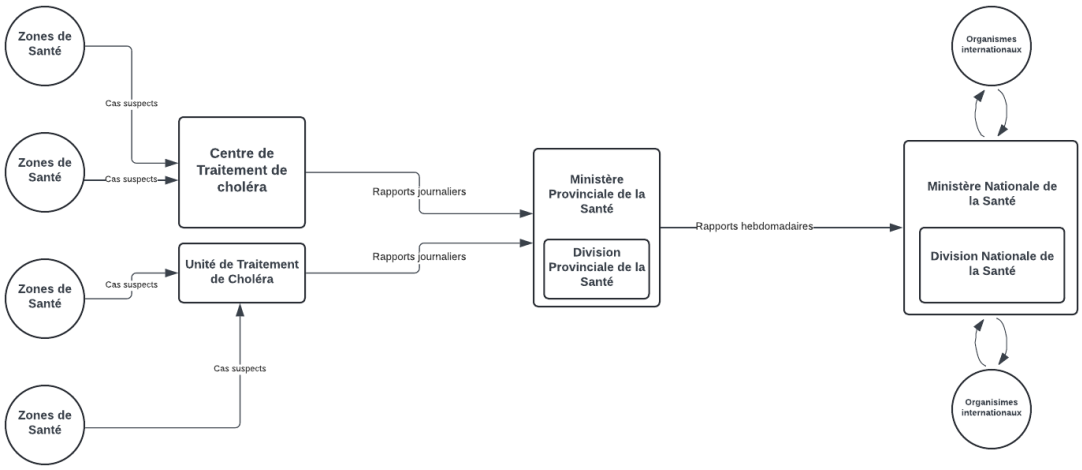


Figure I.5 Processus de riposte contre l'épidémie du choléra

Le gouvernement congolais a mis en place des politiques pour lutter contre le choléra, incluant des programmes de vaccination, des initiatives pour améliorer l'accès à l'eau potable et des campagnes de sensibilisation sur les pratiques d'hygiène. De plus, Des plans d'action spécifiques sont développés au niveau provincial pour répondre aux besoins locaux. Cette lutte contre le choléra aussi bien que d'autres maladies, nécessite la collaboration avec des organisations internationales et non gouvernementales comme l'OMS, le MSF et l'UNICEF. Ces partenaires fournissent des ressources, des expertises et un soutien logistique, renforçant ainsi les capacités locales et assurant une réponse coordonnée.

I.4 Analyse des données

Notre étude sur la prévention du choléra dans le Haut-Katanga a combiné des recherches en ligne et des enquêtes de terrain aux CTC et UTC de Lubumbashi et au ministère provincial de la santé. Nous avons recueilli des données qualitatives via des entretiens semi-structurés et analysé les données épidémiologiques des six dernières années pour identifier les tendances et les facteurs de risque. Les sites de l'OMS et de MSF ont été des sources fiables. Cette approche mixte a permis une compréhension approfondie du problème, malgré les défis liés à la qualité des données, aux contraintes de temps et de ressources, et aux strictes règles de confidentialité du ministère provincial de la santé.

Cette étape est cruciale à ce travail dans le sens où elle nous a permis d'avoir une compréhension approfondie des données et d'identifier certaines tendances au fil du temps sur les données cholériques de ces six dernières années au Haut-Katanga. Et la démarche adoptée a été la suivante :

I.4.1 La collecte de données

Les datasets du CTC et de la DPS sont constituent les masses de données principales qui ont été exploitées dans ce projet. Ces données ont été également soumises

à des analyses approfondies avec Microsoft Power Bi et ont révélé certaines informations pertinentes.

I.4.2 La description des données

Le jeu de donné est constitué de 12 colonnes et plus de 10000 enregistrements. Le tableau 1 récapitule les champs qui constituent notre dataset ainsi que leurs types de données respectifs :

Tableau I.1 Tableau de colonnes du jeu de données

Numéros	Champs	Types de données	Signification/Description
1	prov	Texte	Province
2	zs	Texte	Zone de santé
3	pop	Nombre entier	Densité de la population
4	numsem	Nombre entier	Numéro de la semaine épidémiologique, allant de 1 à 52
5	debutsem	Date	Date des débuts des semaines
6	epidemic	Texte	Nom de l'épidémie
7	c1259mois	Nombre entier	Nombre de cas de moins de 5 ans
8	d1259mois	Nombre entier	Nombre de décès de moins de 5 ans
9	c5ansp	Nombre entier	Nombre de cas de plus de 5 ans
10	d5ansp	Nombre entier	Nombre de décès de plus de 5 ans
11	casageinc	Nombre entier	Nombre de cas dont l'âge est inconnu

12	decageinc	Nombre entier	Nombre de décès dont l'âge est inconnu
13	totalcas	Nombre entier	Nombre total de cas
14	totaldeces	Nombre entier	Nombre total de décès
15	letal	Nombre décimal	La létalité en pourcentage
16	attaq	Nombre décimal	
17	Annee	Nombre entier	Année d'enregistrement du cas

I.4.3 L'exploration des données

Après interrogation et application de certains filtres et de certaines formules DAX de Power Bi, les captures suivantes présentent de façon globale la situation du choléra dans la province du Haut-Katanga de 2019 à 2024**.**

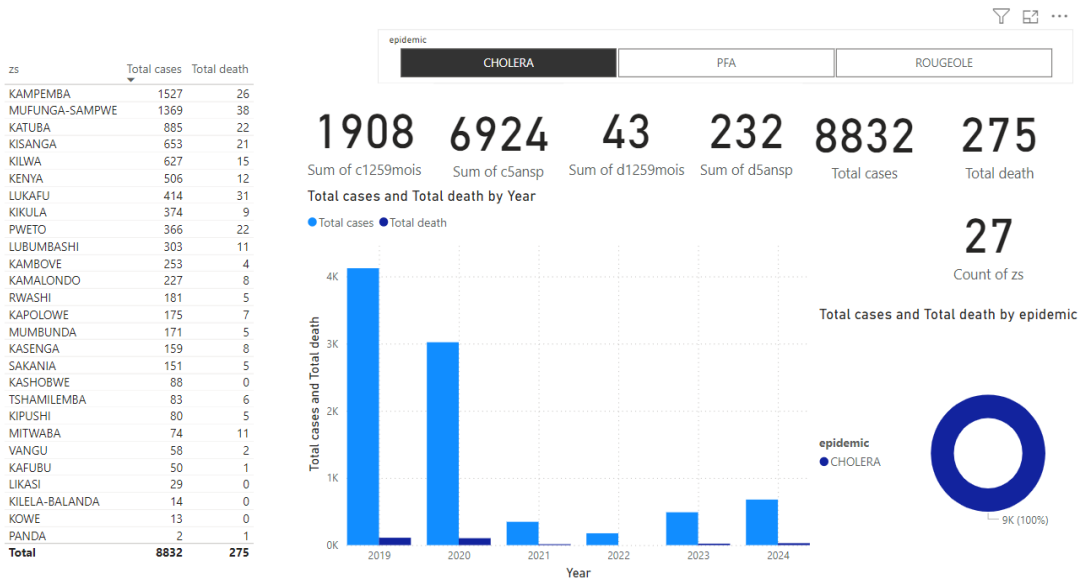


Figure I.6 Situation du choléra au Haut-Katanga de 2019 à 2024

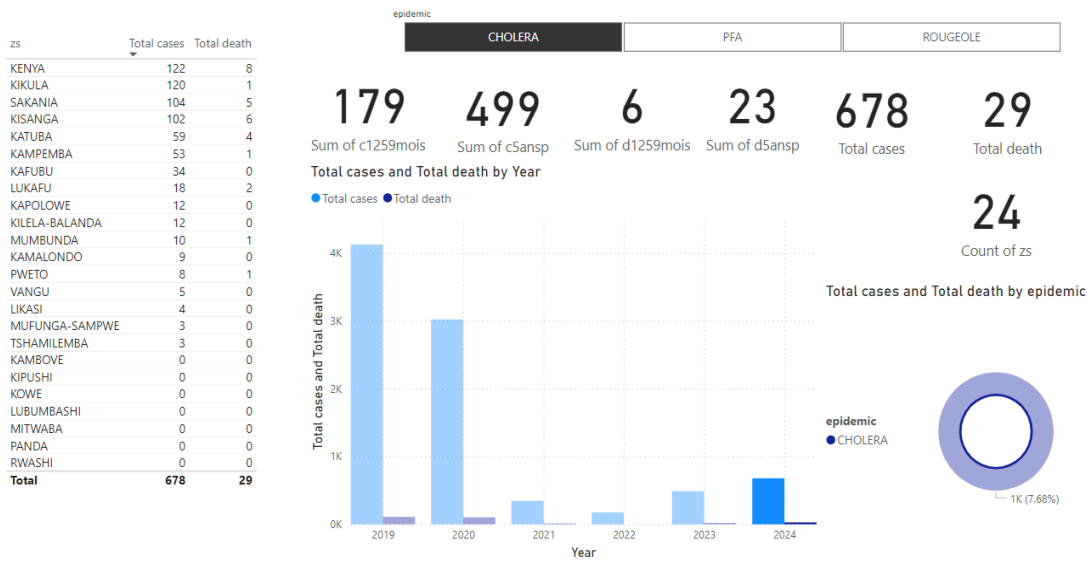


Figure I.7Situation du Choléra au Haut-Katanga en 2024

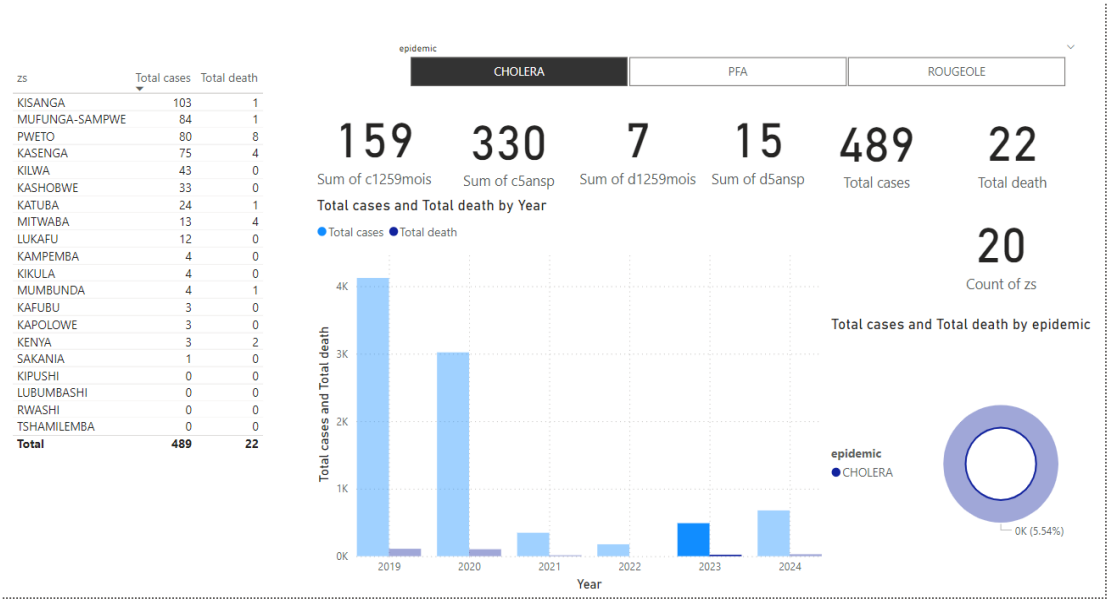
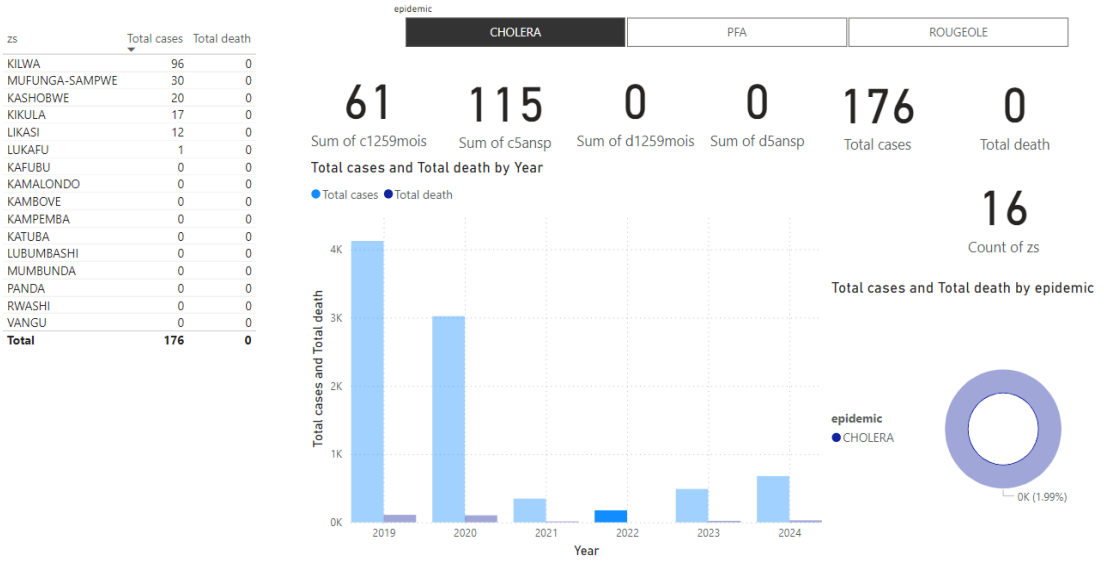
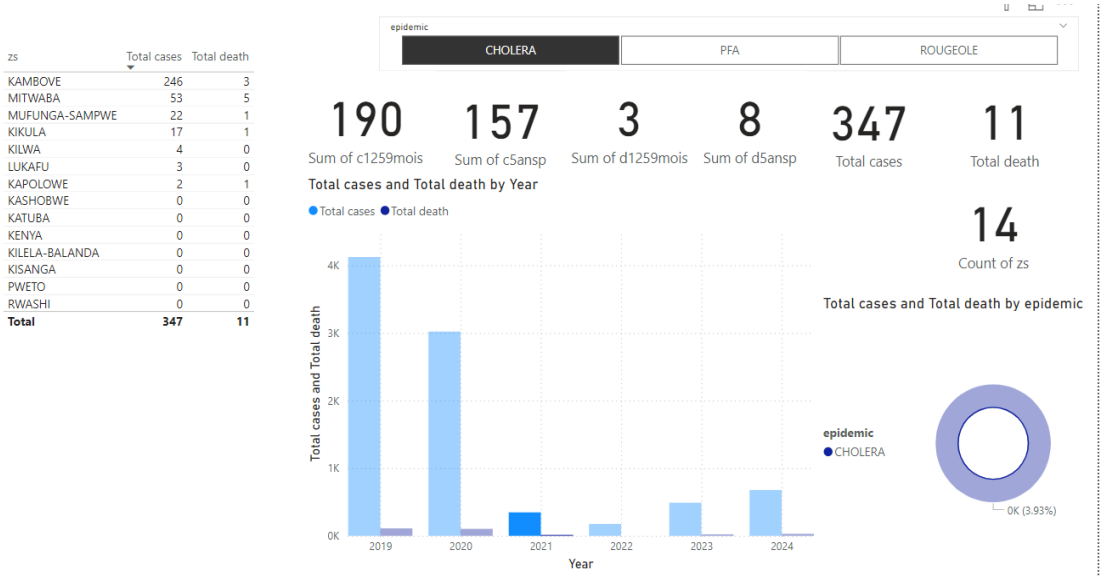


Figure I.8 Situation du Choléra au Haut-Katanga en 2023



zs	Total cases	Total death
KILWA	96	0
MUFUNGA-SAMPWE	30	0
KASHOBWE	20	0
KIKULA	17	0
LIKASI	12	0
LUKAFU	1	0
KAFUBU	0	0
KAMALONDO	0	0
KAMBOVE	0	0
KAMPEMBA	0	0
KATUBA	0	0
LUBUMBASHI	0	0
MUMBUNDA	0	0
PANDA	0	0
RWASHI	0	0
VANGU	0	0
Total	**176**	**0**

Figure I.9 Situation du Choléra au Haut-Katanga en 2022



zs	Total cases	Total death
KAMBOVE	246	3
MITWABA	53	5
MUFUNGA-SAMPWE	22	1
KIKULA	17	1
KILWA	4	0
LUKAFU	3	0
KAPOLOWE	2	1
KASHOBWE	0	0
KATUBA	0	0
KENYA	0	0
KILELA-BALANDA	0	0
KISANGA	0	0
PWETO	0	0
RWASHI	0	0
Total	**347**	**11**

Figure I.10 Situation du Choléra au Haut-Katanga en 2021

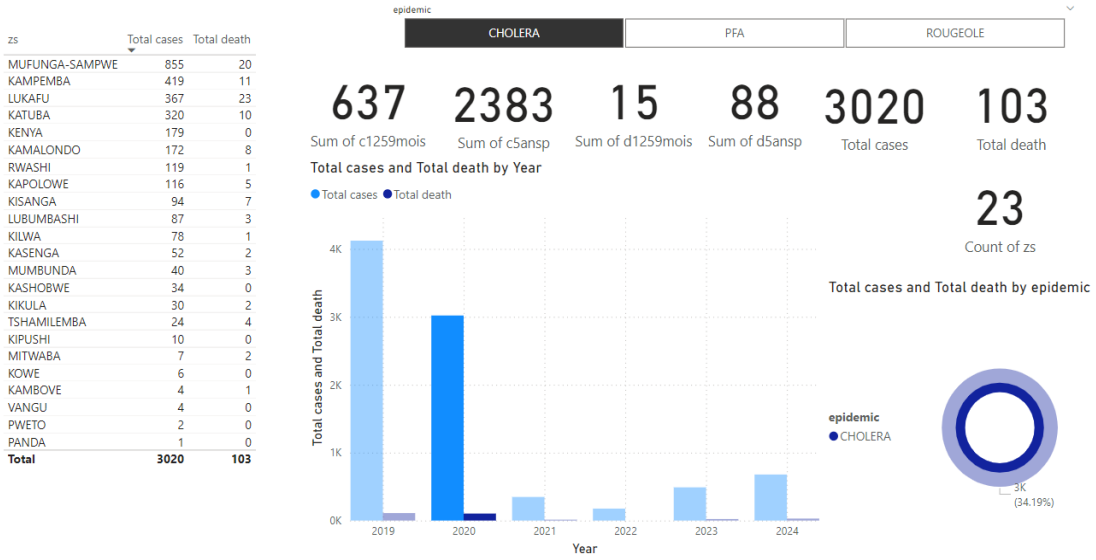


Figure I.11 Situation du Choléra au Haut-Katanga en 2020

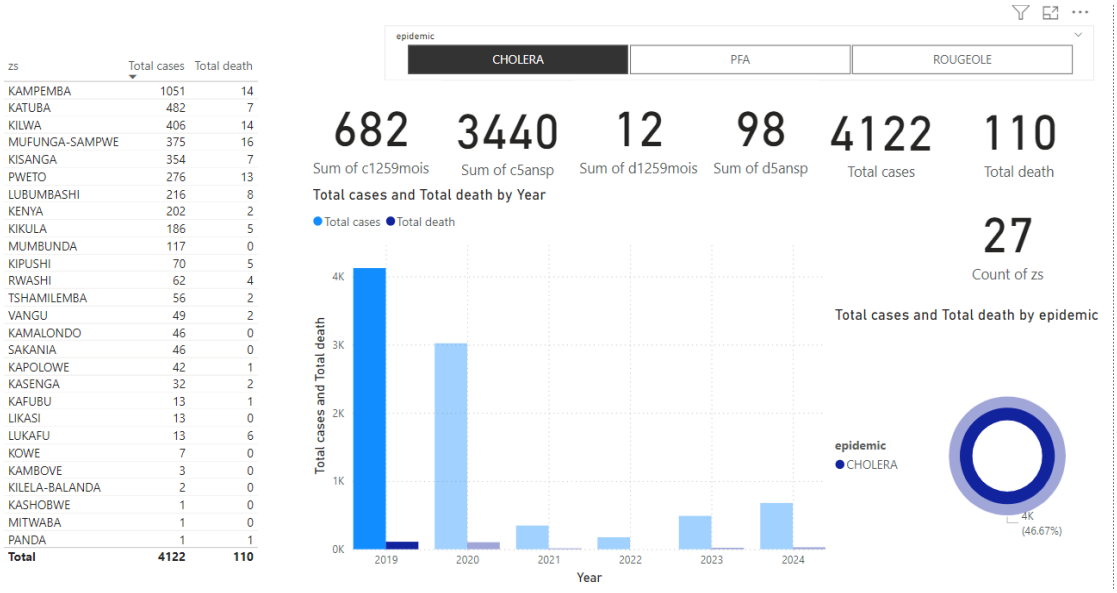


Figure I.12 Situation du Choléra au Haut-Katanga en 2019

I.4.4 La vérification de la qualité

La qualité du dataset est fonction de la qualité de chacune de ses colonnes ; et une colonne valide n'est autre que celle qui contient des enregistrements ayant 1% d'erreurs au maximum et 1% de valeurs vides. Une erreur est un enregistrement ou une donnée qui ne respecte pas le type de données prédéfini de sa colonne. Grâce à Power Query intégré à Power Bi, nous avons pu nous assurer de la qualité des données épidémiologiques comme l'illustrent la figure 8 et 9 avant de passer à la phase suivante.

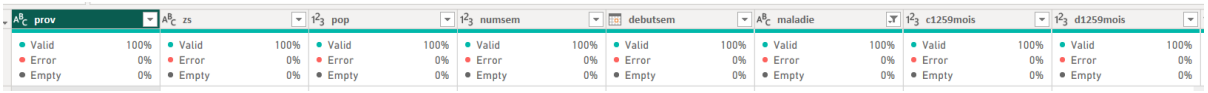


Figure I.13 Qualité de 8 premiers champs de données

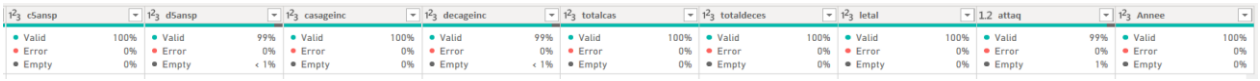


Figure I.14 Qualité de 9 dernières colonnes

I.5 Critique de l'existant

Points forts : Lors de nos investigations sur le terrain et la phase d'analyse des données, nous avons constaté que les données sont encodées et stockées de manière centralisée. De plus, des rapports hebdomadaires sont générés, témoignant d'une bonne organisation et d'une gestion efficace des informations.

Point faible : Cependant, ces données ne sont pas exploitées pour des approches de machine learning, ce qui limite leur potentiel dans la prévention des maladies. L'absence d'analyse avancée empêche de tirer pleinement parti des informations disponibles pour anticiper et gérer les épidémies.

Solution : Pour remédier à cela, nous proposons d'utiliser le machine learning comme approche technologique. Cette méthode permettra d'exploiter les données existantes à des fins de prédiction, améliorant ainsi la capacité à prévenir les maladies et à intervenir de manière proactive. En utilisant des modèles basés sur des séries temporelles et des données locales, nous pourrons fournir des prédictions plus précises et adaptées à notre contexte régional.

Conclusion partielle

Après avoir exploré et compris tous les différents points précédents, il est donc évident que la numérisation, la conservation ou la centralisation des données épidémiologiques ne feront pas l'objet de ce travail scientifique. Car il s'avère que ce sont des problèmes déjà résolus au sein du système de notre cadre de recherche. Pourtant, comme énoncé dans la problématique, la fluctuation épidémique du choléra continue à sévir dans notre région malgré les efforts que fournissent les autorités sanitaires locales en complicité avec leurs partenaires internationaux.

C'est ainsi que nous avons proposé l'apprentissage automatique pour exploiter les masses de données rangées au ministère de la santé afin de prédire la propagation du choléra dans le but d'aider les autorités sanitaires à prendre des décisions éclairées. Pour ce faire, il nous connaître certaines bases sur ledit apprentissage avant de procéder à l'implémentation de la solution, et c'est à cela que sera consacré le prochain chapitre.

CHAPITRE II GENERALITES SUR LE MACHINE LEARNING

Introduction partielle

Le machine learning, ou apprentissage automatique, constitue une branche de l'intelligence artificielle (IA) qui permet aux ordinateurs d'apprendre à partir de données et de prendre des décisions sans nécessiter une programmation explicite pour chaque tâche. Cette discipline repose sur des algorithmes capables de détecter des motifs dans des ensembles de données, d'améliorer leurs performances au fil du temps et de réaliser des prédictions ou des classifications basées sur de nouvelles données.

L'essor du machine learning est attribuable à plusieurs facteurs clés : l'augmentation massive des volumes de données disponibles, les avancées en matière de puissance de calcul et le développement de nouveaux algorithmes d'apprentissage. Ces progrès ont permis d'appliquer le machine learning à divers domaines, tels que la reconnaissance d'images, la traduction automatique, la détection de fraudes, et bien d'autres encore.

Dans ce chapitre, nous explorerons les concepts fondamentaux du machine learning, ses différentes catégories, ainsi que les étapes essentielles pour développer un modèle d'apprentissage automatique. Nous aborderons également les défis et les opportunités associés à cette technologie révolutionnaire.

II.1 Historique

L'intelligence artificielle (IA) est née après la Seconde Guerre mondiale. Les premiers travaux incluent le modèle de neurones artificiels de McCulloch et Pitts (1943) et le concept d'apprentissage de Hebb (1949). Les années 50 et 60 ont vu des avancées majeures avec des projets comme le "General Problem Solver" et le langage de programmation Lisp de J. McCarthy. Ces développements ont suscité de grands espoirs pour l'avenir de l'IA

Entre les années 70 et 80, l'IA s'est concentrée sur des systèmes d'expertise comme le diagnostic médical. Les avancées incluent les modèles de Markov et les réseaux Bayésiens. La fin des années 80 a vu le retour des réseaux de neurones avec l'apprentissage par back propagation. L'Internet des années 90 a accéléré les progrès en IA grâce à une meilleure collaboration et communication.

En 2000, l'ère numérique a généré de grandes quantités de données, ouvrant la voie au Big Data. Les techniques de ML sont devenues essentielles pour analyser ces données. En 2010, les percées de l'apprentissage profond, grâce aux progrès des GPU, des architectures de réseaux neuronaux et des ensembles de données étiquetées, ont conduit à des améliorations significatives dans des domaines comme la reconnaissance d'images, le traitement du langage naturel et le jeu (AlphaGo).

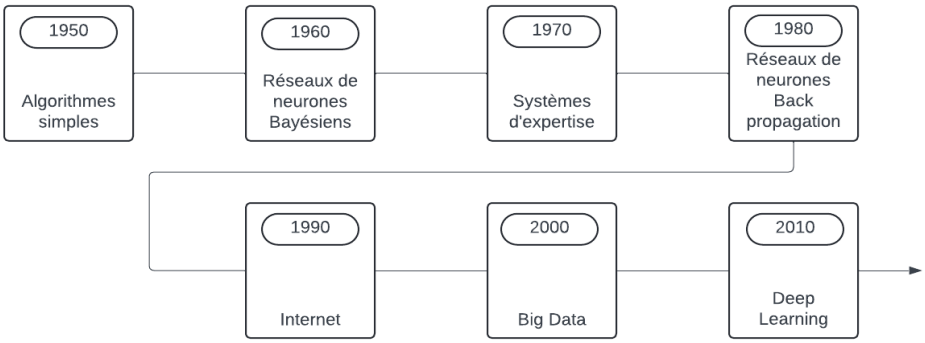


Figure II.1 Evolution du Machine learning

II.2 Comparaison entre l'IA, le ML et le Deep Learning

Dans l'usage courant, les termes « apprentissage automatique » et « intelligence artificielle » sont souvent utilisés de manière interchangeable en raison de la prévalence de l'apprentissage automatique à des fins d'intelligence artificielle dans le monde d'aujourd'hui [9]. Mais les deux termes sont significativement distincts. Alors que l'IA fait référence à la tentative générale de créer des machines capables des capacités cognitives semblables à celles de l'homme, l'apprentissage automatique fait spécifiquement référence à l'utilisation d'algorithmes et d'ensembles de données pour y parvenir [9].

De nos jours vous entendrez également parler de « deep learning », qui est une sous-catégorie du machine learning qui utilise des réseaux de neurones multicouches, appelés réseaux neuronaux profonds, pour simuler le pouvoir de décision complexe du cerveau humain. La principale différence entre le deep learning et le machine learning réside dans la structure de l'architecture du réseau neuronal sous-jacent [10]. Les modèles de l'apprentissage automatique (« non profonds »), utilisent des réseaux neuronaux simples, avec une ou deux couches de calcul. Les modèles de deep learning utilisent trois couches ou plus, mais généralement des centaines ou des milliers de couches, pour l'entraînement.

Nous ne pouvions pas clore ce point sans parler de la data science, un domaine crucial et très lié à l'intelligence artificielle. La data science est en effet, un domaine multidisciplinaire qui consiste à extraire des informations exploitables à partir de données brutes, en identifiant des tendances, motifs, connexions et corrélations dans de grands ensembles de données [11]. Elle englobe l'ensemble du processus de collecte, de nettoyage, d'analyse et d'interprétation des données pour en extraire des insights utiles [12].

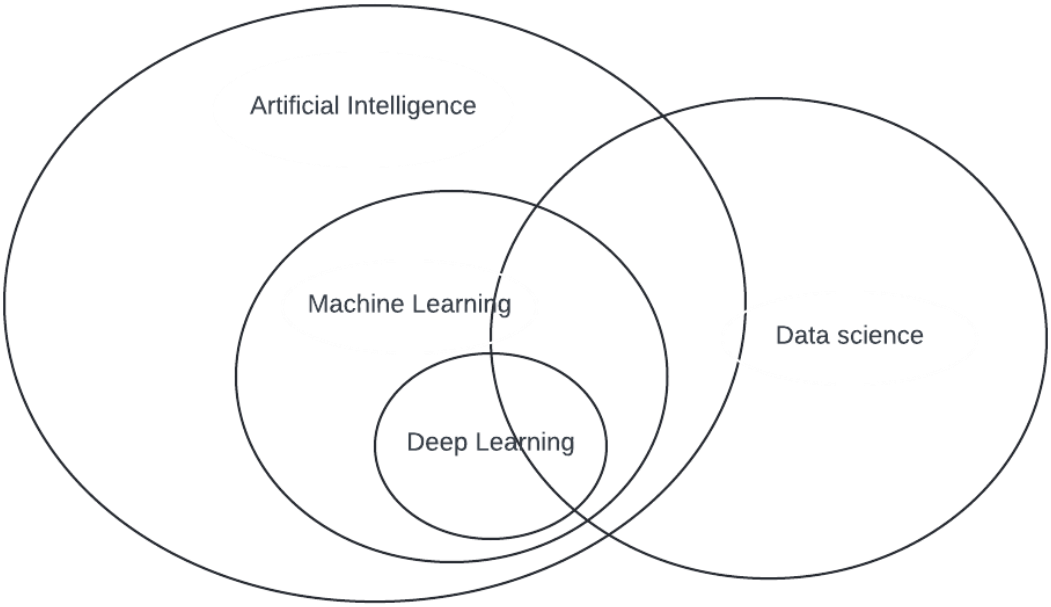


Figure II.2 Le machine learning comparé à l'IA et au Deep Learning

La figure ci-dessus montre comment l'intersection entre l'IA et la data science est à la fois profonde et synergique. La data science utilise des méthodes scientifiques, statistiques et informatiques pour extraire des connaissances à partir de données. L'IA, en particulier le ML, est un outil clé dans ce processus. Les algorithmes de ML permettent aux machines d'apprendre à partir de données historiques, de reconnaître des modèles et de faire des prédictions. En retour, l'IA bénéficie des vastes quantités de données traitées par la data science pour améliorer ses performances et affiner ses modèles. Cette collaboration permet de transformer des données brutes en informations exploitables, facilitant ainsi des décisions éclairées et des innovations technologiques [12].

II.3 Fonctionnement

Une étude faite à l'Université de Californie à Berkeley, décompose l'algorithme typique de l'apprentissage supervisé du machine learning en trois parties principales [9] :

1. **Un processus de décision** : en général, les algorithmes d'apprentissage automatique sont utilisés soit, pour faire une prédiction, soit pour faire une classification. Sur base de certaines données d'entrée, étiquetées ou non, l'algorithme produira une estimation d'un modèle dans les données.

2. **Une fonction d'erreur** : cette fonction évalue la prédiction du modèle. S'il existe des exemples connus, une fonction d'erreur peut effectuer une comparaison pour évaluer la précision du modèle.

3. **Un processus d'optimisation du modèle** : si le modèle peut mieux s'adapter aux points de données de l'ensemble d'apprentissage, les poids sont ajustés pour réduire l'écart entre l'exemple connu et l'estimation du modèle. L'algorithme

répétera ce processus itératif « d'évaluation et d'optimisation » en mettant à jour les poids de manière autonome jusqu'à ce qu'un seuil de précision soit atteint.

II.4 Types d'apprentissages

De nos jours, les nombreux biens et services numériques que nous utilisons reposent sur plusieurs types d'apprentissage automatique. Bien que chacun de ces types tente d'atteindre des objectifs plus ou moins similaires comme créer des machines et des applications autonomes, leurs méthodes diffèrent de quelque peu. Pour une meilleure compréhension entre les différences entre ces types, voici un aperçu des quatre types d'apprentissage automatique principalement utilisés à ce jour [10] :

II.4.1 L'apprentissage supervisé

Également connu sous l'apprentissage automatique supervisé, est défini par son utilisation d'ensembles de données étiquetés pour entraîner des algorithmes afin de classer les données ou de prédire les résultats avec précision. Au fur et à mesure que les données d'entrée sont introduites dans le modèle, celui-ci ajuste ses pondérations jusqu'à ce qu'elles soient correctement ajustées. Cela se produit dans le cadre du processus de validation croisée pour garantir que le modèle évite le surajustement ou le sous-ajustement. L'apprentissage supervisé aide à résoudre divers problèmes réels à grande échelle, tels que la classification du spam dans un dossier distinct de votre boîte de réception. Certaines méthodes utilisées dans l'apprentissage supervisé incluent les réseaux de neurones, les bayes naïfs, la régression linéaire, la régression logistique, la forêt aléatoire et la machine à vecteurs de support (SVM).

II.4.2 L'apprentissage non supervisé

Aussi appelé « apprentissage automatique non supervisé », cet apprentissage utilise des algorithmes d'apprentissage automatique pour analyser et regrouper des ensembles de données non étiquetés (sous-ensembles appelés clusters). Ces algorithmes découvrent des modèles cachés ou des regroupements de données sans nécessiter d'intervention humaine. La capacité de cette méthode à découvrir des similitudes et des différences dans les informations la rend idéale pour l'analyse exploratoire des données, les stratégies de vente croisée, la segmentation des clients et la reconnaissance d'images et de modèles. Il est également utilisé pour réduire le nombre de fonctionnalités dans un modèle grâce au processus de réduction de dimensionnalité. L'analyse en composantes principales (ACP) et la décomposition en valeurs singulières (DVS) sont deux approches courantes pour cela. D'autres algorithmes utilisés dans l'apprentissage non supervisé incluent les réseaux de neurones, le regroupement à k-moyennes et les méthodes de regroupement probabiliste.

II.4.3 L'apprentissage semi-supervisé

Cet apprentissage utilise des ensembles de données étiquetées et non étiquetés pour former des algorithmes. En général, lors de l'apprentissage automatique semi-supervisé,

les algorithmes sont d'abord alimentés par une petite quantité de données étiquetées pour les aider à se développer, puis par des quantités beaucoup plus importantes de données non étiquetées pour compléter le modèle. Par exemple, un algorithme peut être alimenté par une petite quantité de données vocale étiquetées, puis entraîné sur un ensemble beaucoup plus important de données vocales non étiquetées afin de créer un modèle d'apprentissage automatique capable de reconnaître la parole. L'apprentissage automatique semi-supervisé est souvent utilisé pour former des algorithmes à des fins de classification et de prédiction dans le cas où de grandes quantités de données étiquetées ne sont pas disponibles.

II.4.4 L'apprentissage par renforcement

L'apprentissage par renforcement utilise les essais et les erreurs pour former des algorithmes et créer des modèles. Au cours du processus de formation, les algorithmes opèrent dans des environnements spécifiques et reçoivent un retour d'information après chaque résultat. Comme un enfant qui apprend, l'algorithme commence lentement à comprendre son environnement et à optimiser ses actions pour obtenir des résultats particuliers. Cet apprentissage est couramment pour dans les jeux vidéo, la robotique et les systèmes de recommandation.

Quant à ce projet, étant donné que les données récoltées au ministère de la santé sont étiquetées, il est évident que nous avons à faire à un apprentissage supervisé et c'est lors de la conception du modèle que cela sera explicite.

II.5 Les algorithmes de Machine Learning

Les algorithmes de machine learning sont des modèles computationnels qui permettent aux ordinateurs de comprendre des motifs et de faire des prédictions ou prendre des décisions basées sur des données [10]. Les plus couramment utilisés sont : les réseaux de neurones, la régression linéaire et logistique, les arbres de décisions, le clustering, les forêts aléatoires, le SVM, etc. Pour des raisons de pertinence, nous explorerons également certains modèles classiques tels que ARIMA et SARIMA, et aussi le récent Prophet mis en place par Meta (Facebook) en raison de leur pertinence.

II.5.1 Réseaux de neurones artificiels

Les réseaux de neurones artificiels (ANN) sont des systèmes inspirés du fonctionnement des neurones biologiques, utilisés pour résoudre des problèmes complexes en intelligence artificielle et en apprentissage automatique. Ils sont composés de couches de neurones interconnectés, où chaque neurone reçoit des signaux d'entrée, les transforme via une fonction d'activation, et transmet le résultat aux neurones de la couche suivante. Le processus d'apprentissage ajuste les poids des connexions pour

minimiser l'erreur entre la sortie prédite et la sortie réelle, souvent à l'aide de la rétropropagation. La figure qui suit illustre un exemple de ANN.

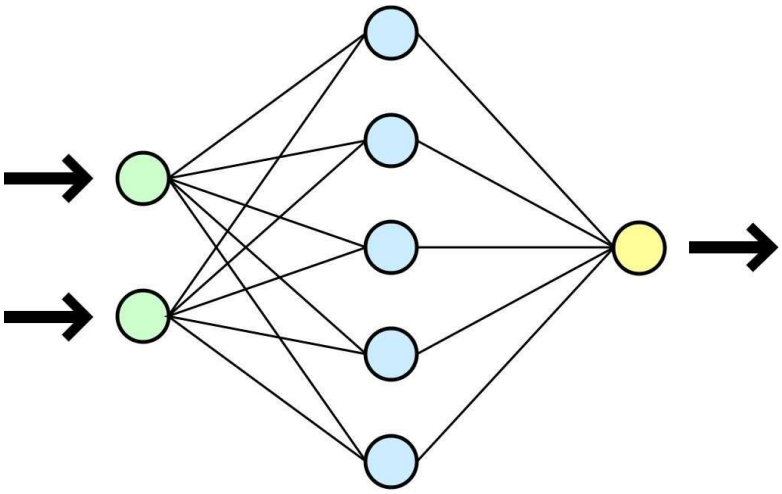


Figure II.3 Réseaux de neurones, source : Les réseaux de neurones récurrents | Reccurent neural network - Deep learning - - Kongakura

III.1 Types de réseaux de neurones

Il existe plusieurs types de réseaux de neurones, chacun adapté à des tâches spécifiques. Les réseaux de neurones feedforward (FNN) sont utilisés pour des tâches de classification et de régression, tandis que les réseaux convolutifs (CNN) sont efficaces pour la reconnaissance d'images. Les réseaux récurrents (RNN) et leurs variantes comme les LSTM (Long Short-Term Memory) sont conçus pour traiter des données séquentielles, comme le texte ou les séries temporelles, en conservant des informations sur de longues périodes.

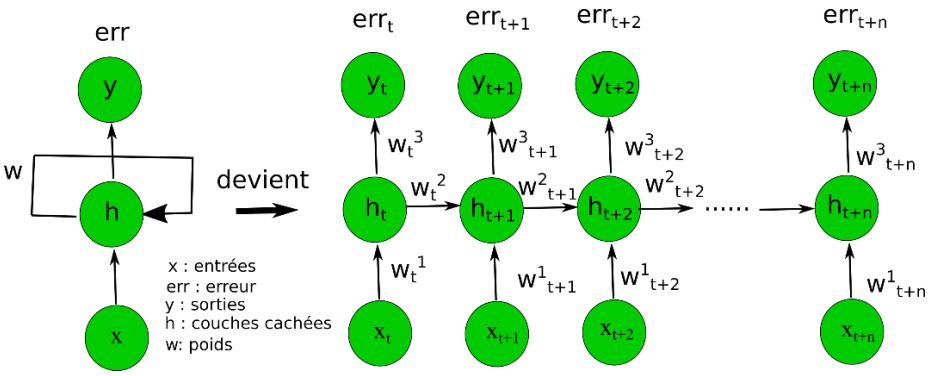


Figure II.4 Fonctionnement des LSTM, Source : Les réseaux de neurones récurrents | Reccurent neural network - Deep learning - - Kongakura

Les réseaux de neurones sont puissants mais présentent des défis, comme le problème du vanishing gradient, où les gradients deviennent très petits et ralentissent l'apprentissage [13]. Des techniques comme l'utilisation de fonctions d'activation

alternatives, l'initialisation appropriée des poids, et des architectures spécifiques comme les LSTM aident à surmonter ces obstacles. Grâce à leur mémoire, les LSTM sont particulièrement efficaces pour traiter des séquences de données avec des dépendances à long terme, comme dans le traitement du langage naturel, la reconnaissance vocale et l'analyse des séries temporelles comme dans notre contexte. Grâce à ces avancées, les réseaux de neurones continuent de jouer un rôle crucial dans le développement de solutions intelligentes et innovantes dans divers domaines.

II.5.2 Régression linéaire

Cet algorithme est utilisé pour prédire des valeurs numériques, sur la base d'une relation linéaire entre les variables dépendantes et les variables indépendantes. Par exemple, cet algorithme pourrait être utilisé pour prédire les prix des logements en se basant sur les données historiques pour la région.

II.5.3 Régression logistique

Cet algorithme d'apprentissage supervisé fait des prédictions pour des variables de réponse catégorielles, telles que « oui/non » aux questions afin de trouver la probabilité qu'un exemple appartienne à une certaine classe. Il peut être utilisé dans la classification des spams et le contrôle qualité sur une ligne de production.

II.5.4 Arbres de décisions

Ils peuvent être utilisés à la fois pour prédire des valeurs numériques (régression) et pour classer les données en catégories. Les arbres de décision utilisent une séquence ramifiée de décisions liées qui peuvent être représentées par un diagramme arborescent. L'un des avantages des arbres de décision est qu'ils sont faciles à valider et à auditer, contrairement à la boîte noire du réseau neuronal.

II.5.5 Forêts aléatoires

Dans une forêt aléatoire, l'algorithme d'apprentissage automatique prédit une valeur ou une catégorie en combinant les résultats d'un certain nombre d'arbres de décision afin d'améliorer la précision et de réduire le surajustement.

II.5.6 Machine à vecteurs de support

Un algorithme puissant pour la classification et la régression, qui cherche à trouver l'hyperplan optimal séparant les différentes classes.

II.5.7 Clustering

Grâce à l'apprentissage non supervisé, les algorithmes de clustering peuvent identifier des modèles dans les données afin de pouvoir les regrouper suivant la similarité des caractéristiques.

II.5.8 Quelques modèles classiques pertinents de Machine Learning et le modèle Prophet

II.7.3.1 ARIMA (AutoRegressive Integrated Moving Average)

Le modèle **ARIMA** est une méthode puissante pour analyser et prévoir des séries temporelles non stationnaires. Il combine trois composants principaux : l'auto-régression (AR), la différenciation (I) et la moyenne mobile (MA).

Le composant AR (Auto-régression) utilise les valeurs passées de la série pour prédire les valeurs futures, modélisée par la formule :

$$x_t = c + \sum_{i=1}^{p} \phi_i x_{t-1} + \varepsilon_t$$

Équation II.1 Equation ARIMA

, où (X_t) est la valeur à l'instant (t), (c) est une constante, (phi i) sont les coefficients du modèle, et (epsilon $_t$) est l'erreur.

Le composant I (Intégration) rend la série stationnaire en différenciant les observations, souvent noté (Delta $X_t = X_t - X_{t-1}$).

Enfin, le composant MA (Moyenne mobile) modélise l'erreur de prévision comme une combinaison linéaire des erreurs passées, exprimée par la formule :

$$\varepsilon_t = \theta_0 + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

Équation II.2 Composante MA de ARIMA

, où (thêta j) sont les coefficients de la moyenne mobile.

La formule générale du modèle ARIMA(p,d,q) combine les composants AR, I et MA ci-haut et voici comment elle se présente :

$$\phi_p(B)\Delta^d x_t = \theta_q(B)\varepsilon_t$$

Où :

- (Phi p(B)) est le polynôme autorégressif d'ordre (p) :

$$\phi_p(B) = 1 + \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

- (Delta ^ d) représente la différenciation d'ordre (d) :

$$\Delta^d x_t = (1 - B)^d x_t$$

- (Thêta q(B)) est le polynôme de moyenne mobile d'ordre (q) :

$$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$$

- (B) est l'opérateur de retard (back shift operator), tel que ($B\ X_t = X_{t-1}$)
- (Epsilon t) est le terme d'erreur à l'instant (t)

En résumé, le modèle ARIMA(p,d,q) est défini par lesdits paramètres, permettant de modéliser et de prévoir les séries temporelles en tenant compte des dépendances passées, des tendances et des erreurs de prévision [13]. L'auto-régression utilise les valeurs passées de la série pour prédire les valeurs futures, tandis que la différenciation rend la série stationnaire en éliminant les tendances. La moyenne mobile, quant à elle, utilise les erreurs de prévision passées pour améliorer les prédictions futures. En utilisant ARIMA, vous pouvez analyser ces données pour prévoir les ventes futures, le nombre de cas futures d'une maladie, ce qui peut aider à la gestion des stocks et à la planification des ressources comme dans notre cas [13] [14].

II.7.3.2 SARIMA (Seasonal AutoRegressive Integrated Moving Average)

Le modèle SARIMA est une extension du modèle ARIMA qui prend en compte les composantes saisonnières des séries temporelles. Il est particulièrement utile pour les données présentant des motifs saisonniers réguliers, comme les ventes mensuelles ou les températures annuelles [13]. Le modèle SARIMA est noté SARIMA(p,d,q)(P,D,Q,s), où :

- (p) : ordre de l'auto-régression (AR)
- (d) : ordre de la différenciation (I)
- (q) : ordre de la moyenne mobile (MA)
- (P) : ordre de l'auto-régression saisonnière (SAR)
- (D) : ordre de la différenciation saisonnière (SI)
- (Q) : ordre de la moyenne mobile saisonnière (SMA)
- (s) : période de la saisonnalité

La formule mathématique du modèle SARIMA est :

$$\phi_p(B^s)\phi_p(B)\Delta^a\Delta_s^v x_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t$$

Équation II.3 SARIMA

Où :

- (Phi p(B) = 1 – phi 1 B – phi 2 B^2 - cdots – phi p B^p) est le polynôme autorégressif d'ordre (p)
- (Phi P(B^s) = 1 – Phi 1 B^s – Phi 2 B^{2s} - cdots – Phi P B^P_s) est le polynôme autorégressif saisonnier d'ordre (P)
- (Thêta(B) = 1 + theta 1 B + theta 2 B^2 + cdots + theta q B^q) est le polynôme de moyenne mobile d'ordre (q)
- (Thêta Q(B^s) = 1 + Thêta 1 B^s + Thêta 2 B^{2s} + cdots + Thêta Q B^Q_s) est le polynôme de moyenne mobile saisonnière d'ordre (Q)
- (Delta ^ d = (1 - B) ^ d) représente la différenciation d'ordre (d)
- (Delta s^D = (1 - B^s D) représente la différenciation saisonnière d'ordre (D)
- (B) est l'opérateur de retard (back shift operator), tel que (B X_t = X_{t-1})
- (Epsilon t) est le terme d'erreur à l'instant (t)

Un modèle ARIMA saisonnier utilise la différenciation à un décalage égal au nombre de saison(s) pour éliminer les effets saisonniers additifs. Comme pour la différenciation du décalage 1 pour supprimer une tendance, la différenciation du décalage introduit un terme de moyenne mobile.

En tenant compte des tendances et des variations saisonnières, SARIMA permet de modéliser les données historiques et de faire des prévisions plus précises, aidant ainsi les entreprises et les organisations à mieux planifier et gérer leurs ressources. Il est utilisé pour prévoir des séries temporelles avec des motifs saisonniers, et il est particulièrement utile pour des applications telles que la prévision des ventes, la demande énergétique, les températures, le trafic web, les revenus publicitaires et les flux touristiques [13].

II.7.3.3 Prophet

Le modèle Prophet de Meta (anciennement Facebook) est un outil de prévision de séries chronologiques conçu pour être flexible et facile à utiliser, particulièrement efficace pour les séries avec des tendances saisonnières et des effets de vacances. Il partage des similarités avec les modèles ARIMA et SARIMA, qui sont des modèles statistiques utilisés pour les séries chronologiques, mais Prophet se distingue par sa facilité d'utilisation et sa flexibilité.

Contrairement à ARIMA et SARIMA, qui nécessitent souvent un prétraitement complexe pour rendre les données stationnaires et capturer la saisonnalité, Prophet utilise un modèle additif basé sur des modèles additifs généralisés (GAM) pour modéliser séparément les tendances, la saisonnalité et les jours fériés. Cela le rend particulièrement robuste face aux données manquantes et aux changements de tendance.

En résumé, Prophet est une alternative moderne et conviviale aux modèles ARIMA et SARIMA, offrant une solution efficace pour les séries chronologiques avec des tendances saisonnières et des effets de vacances, tout en nécessitant moins de prétraitement et de paramétrage [15]. Et tout cela malgré sa sensibilité aux bruits.

La figure suivante nous illustre les différents types d'apprentissages et quelques algorithmes les plus utilisés en apprentissage.

Figure II.5 Hiérarchie de principaux algorithmes de Machine Learning

De tout ce qui précède, nous pouvons déduire que nous avons dorénavant des bases assez solides sur les algorithmes de ML et que nous sommes en mesure de porter notre choix sur les plus pertinents pour parvenir à nos fins.

II.6 Processus de prédiction

Le processus de prédiction en machine learning est une série d'étapes méthodiques visant à créer des modèles capables de faire des prévisions précises basées sur des données, et généralement il est constitué des étapes suivantes : la collecte de données, le prétraitement et le nettoyage des données, la modélisation, l'évaluation, le déploiement et la maintenance du modèle [13].

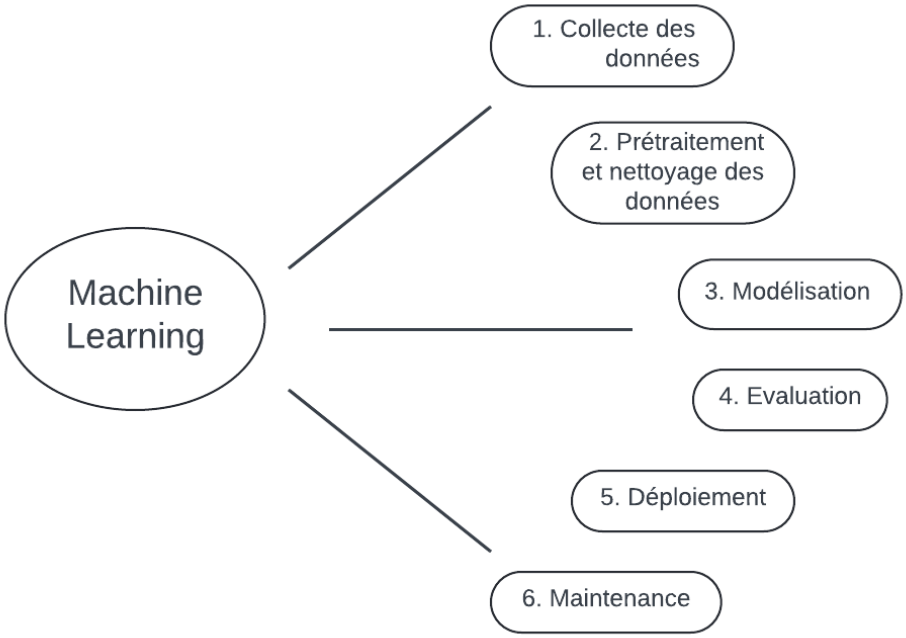


Figure II.6 Processus de prédiction en Machine Learning

Ce processus est essentiel pour créer des modèles de machine learning robustes et fiables, capables de fournir des prédictions précises et utiles dans différents domaines.

II.7 Les séries temporelles

Les séries temporelles, ou séries chronologiques, sont des suites de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps. Elles sont utilisées dans divers domaines comme la finance, la météorologie, l'économie, et bien d'autres, pour analyser et prévoir les tendances futures. Les principales caractéristiques des séries temporelles incluent la tendance, la saisonnalité, le cycle et le résidu [17].

II.7.1 Les caractéristiques des séries temporelles

III.3.4.1 Tendance

La tendance représente la direction générale dans laquelle les valeurs d'une série temporelle évoluent sur une longue période. Par exemple, une tendance à la hausse pourrait indiquer une croissance continue des ventes d'une entreprise, tandis qu'une tendance à la baisse pourrait signaler une diminution progressive de la demande pour un produit ou service.

II.7.3.4 Saisonnalité

La saisonnalité se réfère aux variations périodiques qui se répètent à intervalles réguliers, souvent en fonction des saisons ou des événements récurrents. Par exemple, les ventes de vêtements peuvent augmenter pendant les fêtes de fin d'année ou les soldes d'été, reflétant une saisonnalité marquée par des pics de demande à des moments spécifiques de l'année.

II.7.3.5 Cycle

Les cycles sont des fluctuations qui se produisent à des intervalles irréguliers et sont souvent influencés par des facteurs économiques ou d'autres événements externes. Par exemple, les cycles économiques peuvent affecter les taux de chômage ou la production industrielle, avec des périodes de croissance suivies de récessions.

II.7.3.6 Résidu

Le résidu représente les variations aléatoires ou irrégulières qui ne peuvent pas être expliquées par la tendance, la saisonnalité ou les cycles. Ces variations peuvent être dues à des événements imprévus, des anomalies ou des erreurs de mesure, et sont souvent considérées comme du "bruit" dans les données.

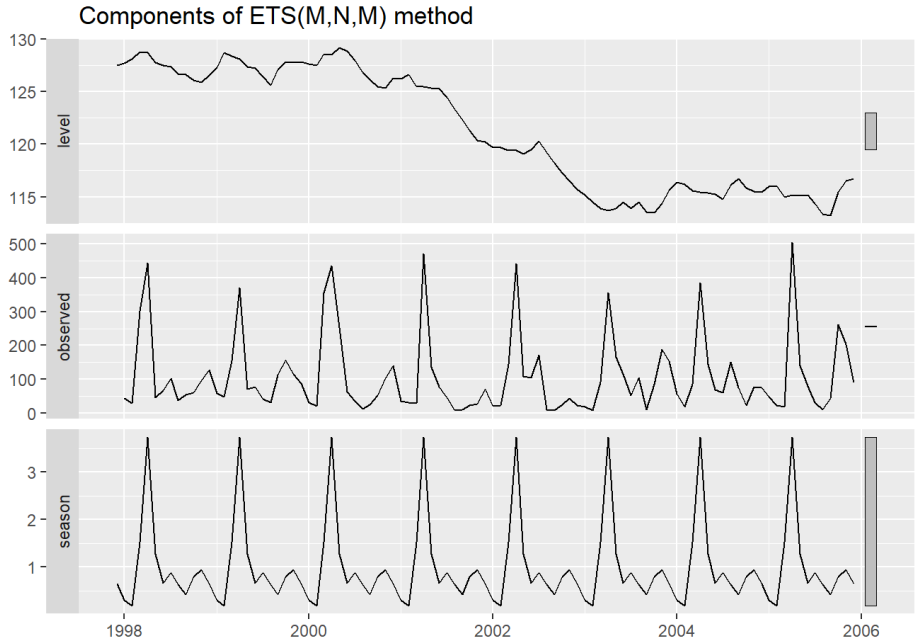


Figure II.7 Série temporelle avec ses différentes caractéristiques, source : [12 Les séries temporelles | Analyse et modélisation d'agroécosystèmes (essicolo.github.io)]

II.7.2 Types de séries temporelles

En général, une série temporelle peut être classée en deux types : stationnaire ou non-stationnaire. La stationnarité est un concept fondamental dans l'analyse des séries temporelles car avant de commencer à travailler sur ces séries, il est crucial de toujours déterminer si elles sont stationnaires ou non.

III.3.4.1 Série temporelle stationnaire

Une série temporelle est dite stationnaire si ses propriétés statistiques, telles que la moyenne, la variance et l'autocorrélation, restent constantes au fil du temps. En d'autres termes, les caractéristiques de la série ne changent pas, peu importe le moment où on l'observe.

II.7.3.7 Série temporelle non-stationnaire

Ces séries présentent des propriétés statistiques qui changent au fil du temps, souvent en raison de tendances ou de saisonnalités. Pour analyser ces séries, on utilise souvent des modèles ARIMA et SARIMA qui incluent une étape de différenciation pour rendre la série stationnaire.

Il n'est pas toujours nécessaire de stationnariser une série temporelle, mais c'est souvent recommandé pour simplifier les calculs et obtenir des prévisions plus fiables. Les modèles comme ARIMA et SARIMA supposent des données stationnaires. Cependant,

des techniques alternatives existent pour les séries non stationnaires. En résumé, la stationnarisation est utile mais dépend des besoins spécifiques de l'analyse.

II.7.3 Quelques mesures statistiques indispensables aux séries temporelles

Voici quelques mesures statistiques essentielles à la fois pour l'analyse des séries temporelles et pour ce projet :

III.3.4.1 Moyenne

La moyenne arithmétique des valeurs de la série donne une idée générale du niveau central des données. Elle est calculée en additionnant toutes les valeurs de la série et en divisant par le nombre total de valeurs. Sa formule est ma suivante :

$$\bar{X} = \frac{1}{N}\sum_{t=1}^{N} X_t$$

Équation II.4 Moyenne

Où (bar{X}) est la moyenne, (N) est le nombre de valeurs, et (Xt) est la valeur à l'instant (t). La moyenne permet de résumer l'ensemble des données par une seule valeur représentative.

III.3.4.2 Variance

La variance mesure la dispersion des valeurs autour de la moyenne, indiquant à quel point les valeurs de la série sont étalées. Une variance élevée signifie que les valeurs sont largement dispersées, tandis qu'une variance faible indique que les valeurs sont proches de la moyenne. Mathématiquement elle se présente comme suit :

$$\sigma^2 = \frac{1}{N}\sum_{t=1}^{N} (X_t - \bar{X})^2$$

Équation II.5 Variance

Où (sigma^2) est la variance, (bar{X}) est la moyenne, et (X_t) est la valeur à l'instant (t). La variance est essentielle pour comprendre la volatilité des données.

III.3.4.3 Autocorrélation

L'autocorrélation mesure la corrélation entre les valeurs de la série à différents décalages temporels. Elle aide à identifier les dépendances temporelles et les motifs répétitifs dans les données. Sa formule est :

$$\rho_k = \frac{\Sigma_{t=1}^{N-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{N}(X_t - \bar{X})}$$

Équation II.6 Autocorrélation

Où (rho k) est l'autocorrélation au décalage (k), (bar{X}) est la moyenne, et (X_t) est la valeur à l'instant (t). L'autocorrélation est utile pour détecter les cycles et les tendances dans les séries temporelles.

III.3.4.4 Écart interquartile (IQR)

Les valeurs aberrantes sont définies comme celles qui se situent en dehors de (1au-dessus du troisième quartile (Q3) ou en dessous du premier quartile (Q1). Voici sa formule :

$$IQR = Q_3 - Q_1$$

Valeurs aberrantes : $X_t < Q_1 - 1.5 \times IQR$

Équation II.7 IQR négatif

Ou $X_t > Q_1 + 1.5 \times IQR$

Équation II.8 IQR positif

Les valeurs aberrantes peuvent également peuvent également être également être identifiées par le **score Z**. Ces mesures permettent de mieux comprendre la structure et les caractéristiques des séries temporelles, facilitant ainsi leur modélisation et leur prévision.

II.8 Domaines d'application de l'apprentissage automatique

Le machine learning est utilisé pour améliorer l'efficacité, la précision et l'expérience utilisateur dans bien de domaines, ses applications sont vastes et variées [9]. Les plus courantes sont :

Recommandations personnalisées : comme ceux utilisés par des géants tels que Netflix, Amazon, Spotify et bien d'autres, les systèmes de recommandation utilisant le machine learning, analysent les comportements passés des utilisateurs pour suggérer des films, des produits ou de la musique que ces derniers pourraient aimer.

Reconnaissance d'images : elle est utilisée dans des applications comme la reconnaissance faciale sur les smartphones, la détection d'objets dans les voitures autonomes, et l'analyse d'images médicales pour diagnostiquer des maladies.

Détection de fraude : les banques et d'autres institutions financières recourent au machine learning pour analyser les transactions et détecter des comportements suspects, aidant ainsi à prévenir la fraude.

Traitement du langage naturel (NLP) : les chatbots, les assistants virtuels comme Siri, et Alexa, et les outils de traduction automatique utilisent le NLP pour comprendre et répondre aux requêtes en langage naturel.

Voitures autonomes : les véhicules autonomes, comme ceux développés par Tesla, utilisent des algorithmes d'apprentissage automatique pour interpréter les données des capteurs et prendre des décisions en temps réel.

Diagnostic médical : le machine learning aide à analyser les données médicales pour diagnostiquer des maladies, prédire les épidémies et personnaliser les traitements.

Analyse prédictive : utilisée dans divers secteurs pour prévoir les tendances futures, comme la demande de produits, les pannes de machines, ou les fluctuations des marchés financiers.

Jeux vidéo : les algorithmes de machine learning sont utilisés pour créer des adversaires intelligents et adaptatifs, améliorer les graphismes et personnaliser l'expérience de jeu.

En intégrant la prédiction des tendances du ML à notre contexte purement épidémiologique, nous pouvons anticiper les variations futures en se basant sur des données historiques cholériques disponibles au ministère de la santé. [14]

II.9 Défis de l'apprentissage automatique

Pour développer des modèles de ML robustes et efficaces, plusieurs défis surgissent et ci-dessous sont listés quelques-uns des plus courants [9] :

- **La qualité des données** : la fiabilité des modèles d'apprentissage automatique est basée sur la haute qualité des données ; c'est-à-dire qu'il est essentiel de bien nettoyer, normaliser et vérifier la cohérence des données avant l'entraînement.

- **La précision des données** : elle concerne à la fois les caractéristiques des données d'apprentissage et les étiquettes de vérité associées. Il faut donc veiller à disposer d'étiquettes précises et cohérentes parce que l'exactitude s'avère cruciale.

- **La reproductibilité des résultats** : pendant l'entraînement, les poids d'un modèle sont initialisés avec des valeurs aléatoires ; ce qui fait qu'en répliquant les expériences d'entraînements, les résultats soient différents. D'où ce défi qui

consiste à documenter les paramètres et les méthodes pour renforcer la transparence et la confiance en garantissant la cohérence des résultats.

- **Le surapprentissage** : appelé également « overfitting », le surapprentissage est défi majeur en ML qui peut être évité par des techniques telles que la validation croisée, la régularisation et la réduction de la complexité du modèle.

- **La mise à l'échelle de données** : les données sont le carburant, la quintessence même de l'IA. C'est ainsi que la collecte et l'accès aux données, leur gouvernance aussi bien que leur compréhension sont indispensables en développement des modèles en apprentissage automatique.

La plupart de défis listés ci-haut ont été rencontré lors de différentes phases de l'élaboration de ce travail et c'est dans la suite du travail que tout cela se démystifie.

II.10 Ethique de l'intelligence artificielle

Aussi surprenante et enthousiasmante qu'elle soit, l'intelligence artificielle suscite bon nombre de débats éthiques qui méritent d'être soulignés dans notre travail [9].

II.10.1 La singularité technologique

Aussi appelée « super intelligence » est l'idée que l'IA pourrait surpasser l'intelligence humaine dans tous les domaines. Bien que cela suscite de l'attention, de nombreux chercheurs ne pensent pas que cela ne se produira pas bientôt. Cependant, cette possibilité soulève des questions éthiques, notamment concernant la responsabilité en cas d'accidents impliquant des systèmes autonomes comme les voitures sans conducteur. Le débat continue sur la poursuite du développement de ces technologies ou leur limitation à des systèmes semi-autonomes pour améliorer la sécurité.

II.10.2 L'impact de l'IA sur l'emploi

L'impact de l'IA sur l'emploi est souvent perçu comme une menace pour les emplois, mais cette vision devrait être nuancée. Comme pour toute nouvelle technologie, l'IA modifie la demande pour certains emplois plutôt que de les éliminer complètement. Par exemple, dans l'industrie automobile, la transition vers les véhicules électriques change la nature des emplois sans supprimer le secteur. De même, l'IA créera de nouveaux emplois nécessitant des compétences pour gérer et maintenir les systèmes d'IA. Le principal défi sera d'aider les travailleurs à s'adapter à ces nouveaux rôles en demande.

II.10.3 Confidentialité des données

La protection de la vie privée est abordée à travers la confidentialité, la protection et la sécurité des données. Des législations comme le GDPR en Europe (2016) et le CCPA en Californie (2018) ont été mises en place pour donner aux individus plus de contrôle

sur leurs données personnelles. Ces lois obligent les entreprises à revoir leurs méthodes de stockage et d'utilisation des informations personnelles, augmentant ainsi les investissements en sécurité pour prévenir les vulnérabilités et les cyber-attaques.

Conclusion partielle

En résumé, le machine learning représente une avancée majeure dans le domaine de l'intelligence artificielle, offrant des capacités d'apprentissage et de prise de décision autonomes aux ordinateurs. Grâce à des algorithmes sophistiqués, cette technologie permet de détecter des motifs complexes dans de vastes ensembles de données et d'améliorer continuellement les performances des modèles.

Les progrès rapides dans la disponibilité des données, la puissance de calcul et les algorithmes ont permis au machine learning de s'imposer dans divers secteurs, de la reconnaissance d'images à la détection de fraudes. Cependant, malgré ses nombreuses applications prometteuses, le développement et l'implémentation de modèles de machine learning posent encore des défis importants, notamment en termes de biais, d'éthique et de sécurité.

Ce chapitre nous a fourni une vue d'ensemble des concepts fondamentaux du machine learning, de ses catégories et des étapes clés pour développer un modèle d'apprentissage automatique. En comprenant ces bases, nous sommes mieux équipés pour explorer les opportunités offertes par cette technologie révolutionnaire et pour relever les défis que nous nous sommes fixé dans la suite de ce travail.

CHAPITRE III MISE EN PLACE DU MODELE DE MACHINE LEARNING

Introduction partielle

Dans ce présent chapitre, nous procédons à la transformation des données brutes en informations exploitables afin d'en ressortir un modèle prédictif de données. Cela signifie que nous allons devoir définir clairement notre problème, collecter les données, les préparer, en tirer les caractéristiques les plus pertinentes et choisir l'algorithme de machine learning le plus approprié afin d'entraîner notre modèle. De plus, nous allons également tester ledit modèle afin d'évaluer ses performances et ajuster certains de ses paramètres, si nécessaire.

Toutes ces opérations seront possibles grâce à la méthode CRISP-DM, qui s'avère être la plus utilisée des méthodes, dans la conception des projets de machine learning et de data mining grâce à son approche structurée, flexible et itérative pouvant être adoptée une variété de secteurs, tout en restant focalisée sur les objectifs commerciaux.

III.1 Présentation de la méthode CRISP-DM

Publiée en 1999 pour normaliser les processus d'exploration de données dans tous les secteurs, elle est depuis devenue la méthodologie la plus courante pour les projets d'exploration de données, d'analytique et de science des données. Elle se décompose en six phases principales qui sont : la compréhension de l'entreprise, la compréhension des données, la préparation des données, la modélisation, l'évaluation et le déploiement du modèle. La figure 1 illustre les étapes du cycle de vie de l'apprentissage automatique selon la méthode CRISP-DM. [1]

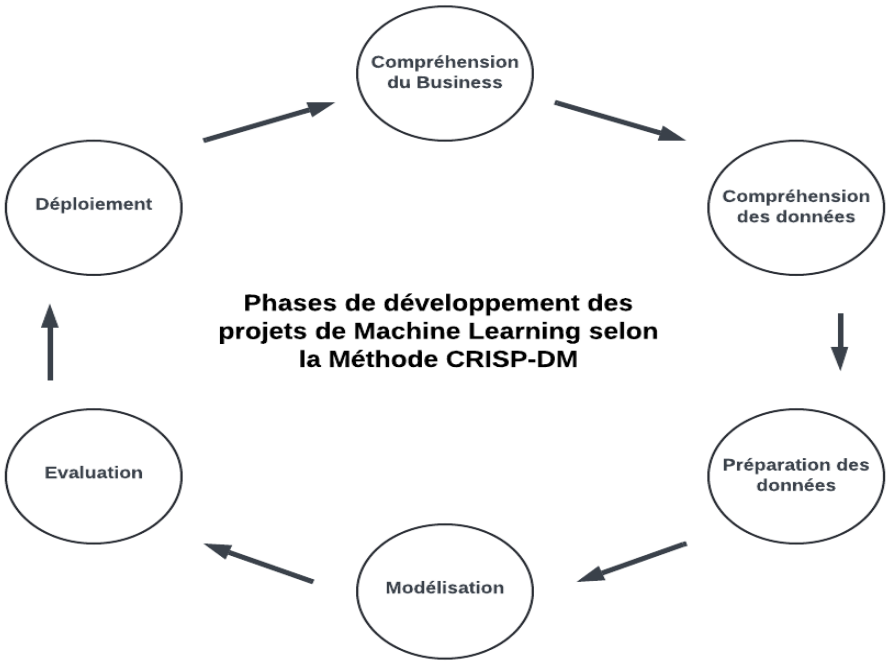


Figure III.1 Cycle de vie des projets d'apprentissage selon CRISP-DM

- **La compréhension de l'entreprise (Business) :** Tout bon projet commence par une compréhension approfondie des besoins du client, et les projets d'exploration de données ne font pas exception. La phase initiale du CRISP-DM se concentre sur la compréhension des objectifs et des exigences du projet, comprenant quatre tâches principales : déterminer les objectifs commerciaux, évaluer la situation, définir les objectifs d'exploration des données et produire le plan du projet. Cette phase est cruciale, car une solide compréhension des besoins est essentielle pour le succès du projet, comparable à la construction d'une maison où des fondations solides sont indispensables. Se précipiter dans l'implémentation sans cette compréhension approfondie peut compromettre tout le projet.

- **La compréhension des données : e**n plus de comprendre l'entreprise, cette phase se concentre sur l'identification, la collecte et l'analyse des ensembles de données pour atteindre les objectifs du projet. Elle comprend quatre tâches principales : la collecte des premières données et leur chargement dans un outil d'analyse comme Power BI ou Tableau, la description des données en documentant leurs propriétés de surface, l'exploration des données pour en savoir plus et identifier les relations, et la vérification de la qualité des données pour documenter tout problème de propreté ou de saleté.

- **La préparation des données :** Connue sous le nom de « data munging » ou « data wrangling », cette étape assure que les données sont de haute qualité, pertinentes et prêtes pour la modélisation, représentant environ 80% du projet.

Elle comprend cinq tâches principales : la sélection des données en fonction des exigences commerciales, le nettoyage des données pour éliminer erreurs et doublons, la structuration des données dans un format cohérent, la construction et l'enrichissement des données en créant de nouveaux attributs et en intégrant diverses sources, et enfin, la validation des données pour vérifier leur exactitude et cohérence. [15]

- **La modélisation :** souvent considérée comme la phase la plus passionnante mais aussi la plus courte de la science des données, elle implique la construction et l'évaluation de divers modèles basés sur différentes techniques de modélisation. Cette phase comprend quatre tâches principales : la sélection des techniques de modélisation (comme la régression ou les réseaux neuronaux), la génération de la conception des tests (divisant les données en ensembles d'entraînement, de test et de validation), la construction des modèles (souvent en exécutant des lignes de code spécifiques), et l'évaluation des modèles (en interprétant les résultats en fonction de la connaissance du domaine et des critères de réussite prédéfinis). Bien que le guide CRISP-DM suggère d'« itérer la construction et l'évaluation du modèle jusqu'à trouver le(s) meilleur(s) modèle(s) », en pratique, les data scientistes devraient continuer à itérer jusqu'à ce qu'ils trouvent un modèle « suffisamment bon », procéder au cycle de vie CRISP-DM, puis améliorer davantage le modèle dans les itérations futures [2].

- **L'évaluation :** contrairement à l'évaluation technique des modèles dans la phase de modélisation, cette phase d'évaluation se concentre sur la pertinence des modèles pour l'entreprise et les actions à entreprendre ensuite. Elle comprend trois tâches principales : évaluer les résultats pour vérifier si les modèles répondent aux critères de réussite de l'entreprise et déterminer lesquels approuver, examiner le processus pour s'assurer que toutes les étapes ont été correctement exécutées et corriger les éventuelles omissions, et enfin, déterminer les prochaines étapes, qu'il s'agisse de déployer les modèles, de poursuivre les itérations ou de lancer de nouveaux projets.

- **Le déploiement :** Un modèle n'est utile que si le client peut accéder à ses résultats. La phase finale, dont la complexité varie, comprend quatre tâches principales : planifier le déploiement en élaborant un plan détaillé, suivre et maintenir le modèle pour éviter les problèmes en phase opérationnelle, produire un rapport final résumant le projet et les résultats, et réaliser une rétrospective pour évaluer ce qui a bien fonctionné et ce qui peut être amélioré. Même après le projet, il est crucial de maintenir le modèle en production avec une surveillance constante et des ajustements occasionnels.

C'est lors de la phase de la conception que seront appliquées chacune des étapes précitées aux données collectées afin de répondre au besoin que présente notre contexte.

III.2 Analyse de la solution

III.2.1 Contexte et objectif du projet

Ayant passé plus de quatre ans dans une institution informatique et plus de deux ans à étudier les systèmes informatiques, nous avons acquis bien de connaissances nous permettant de repérer des problèmes autour de nous afin de pouvoir y apporter des solutions technologiques. Et l'une de connaissances capitales acquises en génie logiciel, est la capacité à analyser des problèmes du monde réel afin de trouver quelle solution y correspond le mieux. D'où notre choix sur le machine learning pour ce travail.

III.2.2 Besoins et Contraintes du projet

Pour garantir la réussite de notre projet, nous nous devions bien définir et de bien planifier certains besoins tout en tenant compte de certaines contraintes que nous devions affronter. Ci-dessous se trouvent ceux que nous avons trouvé pertinents pour ce travail :

III.3.4.1 Besoins techniques

- Infrastructure de données **:** disposer d'une infrastructure capable de stocker et de gérer de grandes quantités de données (bases de données).
- Puissance de calcul : Avoir accès à des ressources de calcul suffisantes pour entraîner les modèles (GPU puissant, RAM puissante, cloud computing).
- Outils et logiciels : Utiliser des bibliothèques et des outils de ML et de BI (Anaconda, Pandas, Sk-learn, NumPy, Tensorflow, Power Bi, etc.).
- Scalabilité : La capacité du système à gérer une augmentation du volume de données sans perte de performance.

III.3.4.2 Besoins fonctionnels

- Définition claire des objectifs : comprendre les objectifs métiers et les résultats attendus du projet.

- Collecte de données : identifier et rassembler les données nécessaires pour entraîner le modèle.
- Préparation des données : nettoyer, transformer et structurer les données pour les rendre utilisables.
- Sélection des caractéristiques : choisir les variables pertinentes qui influenceront le modèle.
- Évaluation des performances : définir des métriques pour évaluer la performance du modèle (le MSE, le MAE, le R^2, etc.).
- Déploiement : prévoir comment le modèle sera intégré dans l'environnement de production et utilisé par les utilisateurs finaux

III.3.4.3 Contraintes

- Qualité des données : les données peuvent être incomplètes, bruitées ou biaisées, ce qui peut affecter la performance du modèle.
- Complexité des modèles : les modèles complexes peuvent être difficiles à entraîner et à interpréter, et peuvent nécessiter des ressources de calcul importantes.
- Adoption par les utilisateurs : la santé publique doit s'apprêter à adopter et à utiliser les solutions basées sur le machine learning.
- Précision : atteindre un niveau de précision suffisant pour que le modèle soit utile et fiable.
- Robustesse : s'assurer que le modèle fonctionne bien sur des données non vues et dans des conditions variées.

III.3 Conception de la solution

Au cours de notre analyse faite au premier chapitre, nous avons déterminé que notre problème est de nature temporelle, car nous souhaitons prédire le nombre de cas de choléra pour les semaines à venir en utilisant les données collectées au cours des six dernières années. Cette phase de conception sera donc dédiée à cette tâche.

III.3.1 La compréhension de l'entreprise

— **Objectif :** Prédire la propagation du choléra pour aider les autorités sanitaires à prendre des décisions préventives éclairées.
— **Critères de réussite :** Réduction du taux de transmission et meilleure allocation des ressources médicales.

III.3.2 La compréhension des données

Cette tâche a été déjà faite au premier chapitre lors de l'analyse des données secondaires avec Microsoft Power Bi.

III.3.3 La préparation des données

— **Sélection des données** : selon la compréhension du business, seules onze colonnes sur 17 devraient être retenues pour l'entraînement de notre futur modèle compte tenu de leur pertinence. En même temps nous avons su les catégories selon qu'elles sont des caractéristiques aussi appelées *features* ou qu'elles sont des cibles aussi appelées *target*. Le tableau 2 illustre clairement les détails importants.

Tableau III.1 tableau de données d'analyse

Colonnes	Description	Types de variables	Pertinence
annee	Année d'enregistrement du cas	feature	Utile à l'analyse des tendances sur plusieurs années.
zs	Zone de santé	feature	Aide à comprendre la propagation et à détecter les zones à risque.
pop	Densité de la population	feature	Une densité élevée favorise la propagation.
numsem	Numéro de la semaine épidémiologique	feature	Facilite la détection des tendances et des pics épidémiologiques.
te	Date de début de la semaine épidémiologique	feature	Facilite la détection des tendances et des pics épidémiologiques
c1259mois	Nombre de cas de moins de 5ans	feature	Cette répartition sert à révéler des groupes à risques.

c5ansp	Nombre de cas de plus de 5ans	feature	Cette répartition sert à révéler des groupes à risques.
casageinc	Nombre de cas dont l'âge est inconnu	feature	Important pour la sommation des cas.
totalcas	Nombre total de cas	target	Indicateur direct de l'ampleur de la maladie.

Cependant, étant donné que nous avons à faire à une série temporelle, nous n'avons retenu pour ce contexte que deux caractéristiques comme nous le démontre la figure 2. Alors pour arriver ce résultat, certaines opérations de filtre et de nettoyage ont été faites. La figure 1 nous donne un aperçu sur le code Python.

```python
# Importation des bibliothèques nécessaires
import pandas as pd
import numpy as np

# Chargement des données à partir d'un fichier Excel
data = pd.read_excel('E:/MyWorks/Data analyst/Data Training/Survepi.xlsx')

# Filtrage des données pour ne conserver que les cas de choléra
df_cholera = data[data['maladie'] == 'CHOLERA']

# Sélection des colonnes pertinentes pour l'analyse
df_cholera = df_cholera[['Annee', 'numsem', 'totalcas']]
```

Figure III.1 Feature engineering

Tableau III.2 Tableau de caractéristiques retenues

Colonnes	Description	Types de variables	Pertinence
Année	Année d'enregistrement du cas	feature	Utile à l'analyse des tendances sur plusieurs années.
numsem	Numéro de la semaine épidémiologique	feature	Facilite la détection des tendances et des pics épidémiologiques.
totalcas	Nombre total de cas	target	Indicateur direct de l'ampleur de la maladie.

— **Feature engineering** : La figure 2 montre comment nous procédons à la création de l'attribut "date'' qui nous servira d'index pour notre série temporelle. Ensuite, nous rééchantillonnons le total de cas de choléra par semaine pour mieux les prédire.

```
# Création de la colonne 'date' en combinant 'Annee' et 'numsem'
df_cholera['date'] = pd.to_datetime(
    df_cholera['Annee'].astype(str) +  # Convertit 'Annee' en chaîne de caractères
    df_cholera['numsem'].astype(str).str.zfill(2) +  # Convertit 'numsem' en chaîne et ajoute des
zéros à gauche si nécessaire
    '0',  # Ajoute '0' pour représenter le dimanche
    format='%Y%U%w'  # Spécifie le format de la date
)

# Définir la colonne 'date' comme index du DataFrame
df_cholera.set_index('date', inplace=True)

# Resampling des données par semaine et sommation des cas de choléra
df_cholera = df_cholera['totalcas'].resample('W').sum()
```

Figure III.2 Feature engineering

— **Validation des données** : à ce stade nous nous assurons que nos différentes opérations n'ont pas altérer nos données. Alors, nous revenons donc sur les opérations comme le formatage et le traitement de valeurs manquantes.

```
# Traitement des valeurs manquantes
df_cholera.fillna(method='ffill', inplace=True)  # Remplir les valeurs manquantes par la méthode de
propagation en avant

# Vérification et correction des types de données
df_cholera['Annee'] = df_cholera['Annee'].astype(int)
df_cholera['numsem'] = df_cholera['numsem'].astype(int)
df_cholera['totalcas'] = df_cholera['totalcas'].astype(int)
```

Figure III.3 La validation des données avec Python

Maintenant que nous avons fini avec la préparation, nos données prêtes à l'entraînement. Mais avant, il nous faut faire un meilleur choix de modèle.

III.3.4 Modélisation

Pour modéliser et prévoir les cas de choléra à partir de séries temporelles, il est pertinent d'explorer plusieurs approches afin de déterminer celle qui offre les meilleures performances. Trois modèles de machine learning et un modèle de deep learning particulièrement adaptés à cette tâche sont : **ARIMA**, **SARIMA, Prophet** et le **LSTM**. En combinant ces quatre approches, nous pouvons obtenir une vue d'ensemble complète et précise des dynamiques sous-jacentes des cas de choléra, permettant ainsi de meilleures prévisions et une prise de décision plus éclairée.

III.3.4.1 ARIMA

Tout d'abord, il nous faut tester la stationnarité, trouver les paramètres p, d et q, et diviser les données en ensemble d'entrainement et de test. Ensuite, nous sommes passé à l'entrainement du modèle, nous l'avons ajusté avec les meilleurs paramètres et enfin nous avons fait des prédictions sur les 12 prochains.

```
# Division des données en ensembles d'entraînement et de test
from sklearn.model_selection import train_test_split
train, test = train_test_split(df_cholera, test_size=0.2, shuffle=False)

# Choix de meilleurs paramètres p,d, et q
model = auto_arima(train, seasonal=False, trace=True, error_action='ignore', suppress_warnings=True)
print(model.summary())

# Ajuster le modèle ARIMA avec les meilleurs paramètres trouvés
best_order = model.order
model = ARIMA(test, order=best_order)
model_fit = model.fit()

# Faire des prévisions
forecast = model_fit.forecast(steps=12)
print(forecast)
```

Figure III.4 Entrainement du modèle ARIMA

Cette prédiction a échoué car elle n'a pas su prédire les valeurs correctes en raison du caractère saisonnier que comporte notre série. Pour remédier à ce problème, nous nous devons d'essayer avec le modèle SARIMA.

III.3.4.2 SARIMA

Etant donné la prise en charge des paramètres saisonniers par ce modèle, nous essayons de trouver lesdits paramètres avec les meilleurs paramètres pour tenter de prédire le nombre de cas de choléra. Mais avant, on doit s'assurer du caractère saisonnier de notre série. Après test, la figure III.3 nous confirme bel et bien que notre série est saisonnière.

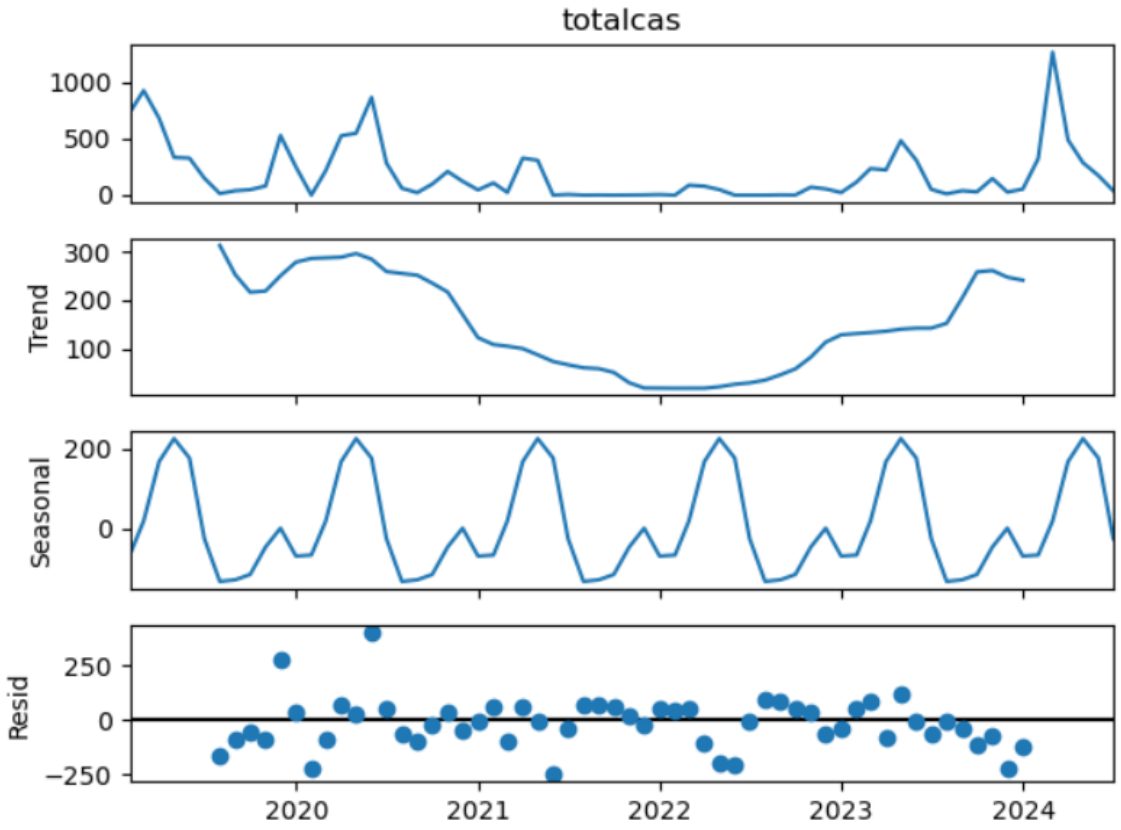


Figure III.5 Décomposition saisonnière

Lorsque nous ajustons le modèle avec les meilleurs paramètres, comme la figure ci-dessous l'indique, nous trouvons des résultats allant de juillet 2024 à juin 2025.

```
# Ajuster le modèle SARIMA avec les meilleurs paramètres trouvés
best_params = (0, 1, 0)
model = SARIMAX(df_cholera, order=best_params, seasonal_order=best_seasonal_params)
model_fit = model.fit(disp=False)

# Faire des prévisions
forecast1 = model_fit.forecast(steps=24)
forecast = model_fit.get_forecast(steps=24)
forecast_ci = forecast.conf_int()
```

Figure III.6 Entrainement du modèle SARIMA

Enfin, avec les meilleurs paramètres ajustés, nous avons droit aux résultats significatifs sur les 12 prochains mois avec SARIMA qui est une variante de ARIMA, cependant avec un motif saisonnier. La figure III.5 nous en donne un bref aperçu.

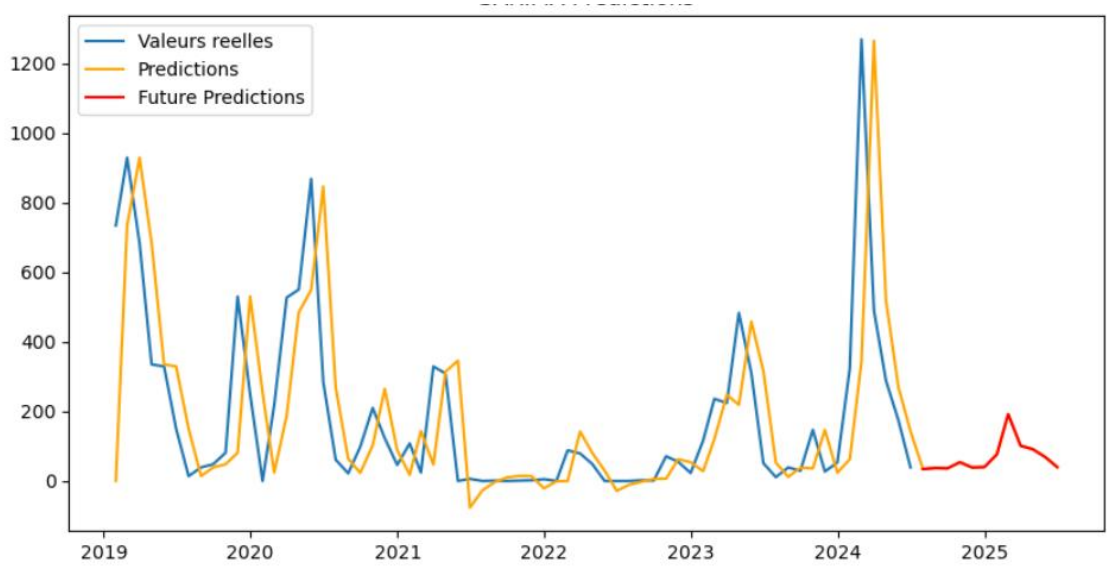


Figure III.7 Prédictions avec SARIMA

III.3.4.3 Prophet

Après avoir essayé d'ajuster le modèle avec les meilleurs paramètres possibles, après avoir normalisé les données et inverser la transformation logarithmique, Prophet n'a pas su nous donner des bonnes prédictions (négatives).

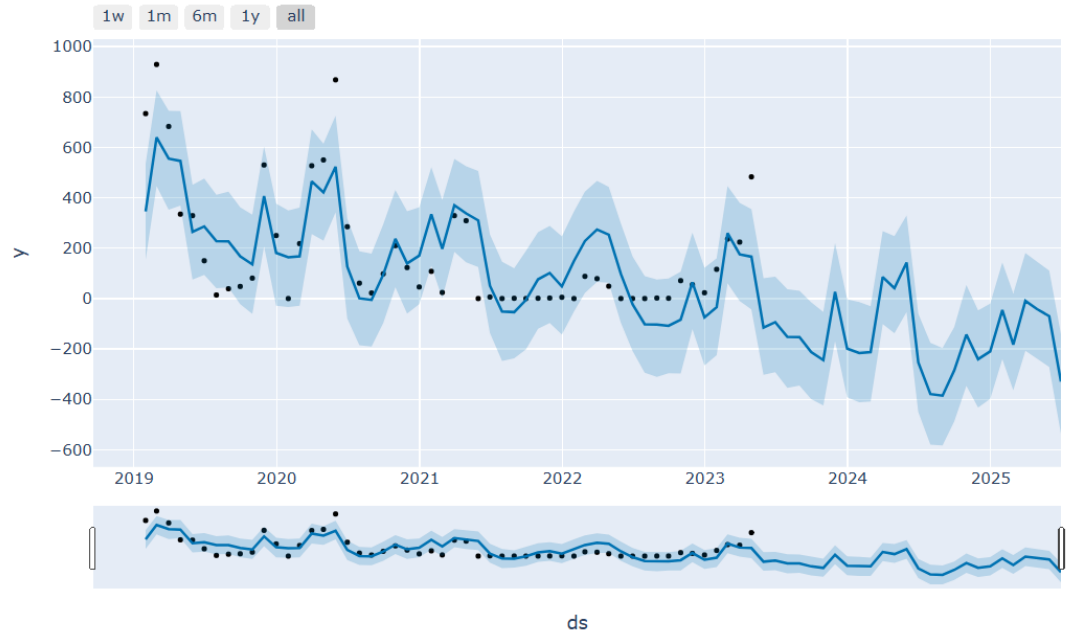


Figure III.8 Prédictions avec Prophet

Comme souligné au chapitre précédent, c'est probablement dû à sa grande sensibilité aux bruits qu'il nous prédit des valeurs négatives malgré les différentes opérations d'optimisation et sa facilité à dans l'utilisation.

III.3.4.4 Réseaux de neurones récurrents LSTM

Quant aux réseaux de neurones, après entrainement et optimisations, nous aboutissons au résultat que présente la figure qui suit :

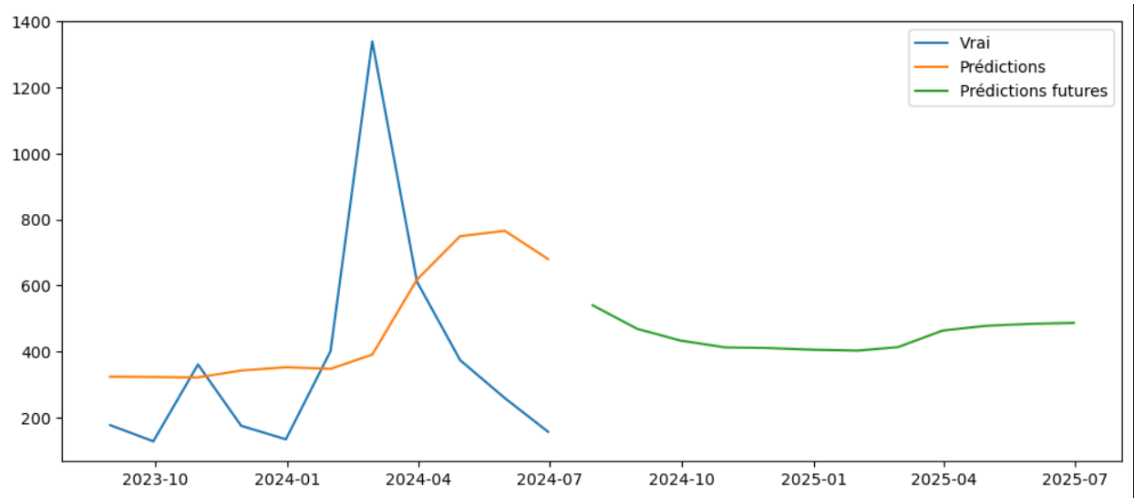


Figure III.2 Prédictions avec le modèle LSTM

Nous remarquons que les prédictions sont faites et que les écarts sont tout de même considérables entre les valeurs prédites et celles réelles. Dans le point qui suit nous essaierons de faire le meilleur choix entre SARIMA et LSTM.

III.3.5 Evaluation

Les modèles SARIMA et LSTM nous donnent tous les deux des résultats différents sur base de données cholériques de 6 dernières années. Alors il nous faut évaluer ces résultats sur base de certains facteurs techniques, mais en tenant également compte des objectifs que nous nous sommes fixés depuis le début de ce projet.

Cependant, étant donné la structure de notre travail, cette phase d'analyse se fera au chapitre suivant.

III.3.6 Déploiement

Comme il nous faut d'abord de choisir le meilleur modèle pour notre projet, sur base de certaines métriques analysées à l'évaluation, cette phase sera également faite au dernier chapitre.

Conclusion partielle

Dans ce chapitre, nous avons entrepris une analyse approfondie et une conception rigoureuse de modèles de machine learning adaptés aux séries temporelles, dans le but de prédire le nombre de cas de choléra futurs en se basant sur les données historiques des six dernières années fournies par le ministère de la santé provincial. Nous avons construit quatre modèles distincts : ARIMA, SARIMA, Prophet et LSTM, chacun choisi pour ses capacités spécifiques à capturer les tendances, les saisonnalités et les dépendances à long terme présentes dans les données.

Bien que l'analyse des performances ne soit pas encore finalisée, nous avons déjà observé que les modèles ARIMA et Prophet ont échoué respectivement à cause de la non capture des tendances saisonnières et de leur grande sensibilité aux bruits. Ces limitations soulignent l'importance de choisir des modèles capables de gérer les particularités des données temporelles, comme le modèle SARIMA pour les effets saisonniers et le modèle LSTM pour les dépendances à long terme.

En conclusion, l'utilisation combinée de ces modèles permet de fournir des prévisions potentiellement robustes et fiables du nombre de cas de choléra, offrant ainsi un outil précieux pour la planification et la gestion des ressources sanitaires. Cette approche multi-modèle assure une meilleure résilience face aux variations et incertitudes des données temporelles, contribuant ainsi à une meilleure préparation et réponse aux épidémies futures. Dans le chapitre suivant, nous présenterons les résultats et évaluerons les modèles retenus afin de choisir le plus performant sur base de l'objectif et certaines contraintes de l'entreprise.

CHAPITRE IV RESULTATS ET DISCUSSION

Introduction

Dans ce chapitre, nous présentons et analysons les résultats obtenus à partir des modèles de machine learning appliqués à la prédiction du choléra. Les performances des différents algorithmes sont évaluées à l'aide de métriques telles que le MSE, le MAE, et le R^2. Ces résultats sont ensuite comparés et interprétés pour identifier les modèles les plus efficaces et les variables les plus influentes dans la prédiction de cette maladie.

Nous discutons également des implications pratiques de nos résultats, en mettant en lumière comment ils peuvent être utilisés par les autorités sanitaires pour améliorer les stratégies de prévention et de contrôle du choléra. Enfin, nous abordons les limitations de notre étude et proposons des pistes pour des recherches futures, afin de renforcer et d'étendre les applications de notre approche.

IV.1 Présentation des résultats

Pour les séries temporelles, la précision comme définie habituellement (basée sur les vrais positifs et les faux positifs) n'est pas la métrique la plus appropriée, surtout que nous prédisons une variable continue qui est le nombre total de cas de choléra (totalcas). De ce fait, nous nous baserons sur des métriques pertinentes telles que le MSE et le RMSE, le MAE et le R^2 qui est le coefficient de détermination.

IV.1.1 Métriques des séries temporelles

IV.1.1.1 Erreur Quadratique Moyenne (MSE)

La MSE mesure la moyenne des carrés des erreurs, c'est-à-dire la différence entre les valeurs prédites et les valeurs réelles. Sa formule est :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Équation IV.1 Moyenne des carrés des erreurs

Plus la MSE est faible, plus le modèle est précis.

IV.1.1.2 Racine de l'Erreur Quadratique Moyenne (RMSE)

Elle donne une idée de la magnitude moyenne des erreurs de prédiction, en tenant compte de leur échelle. Et sa formule est juste la racine carrée de la précédente.

IV.1.1.3 Erreur Absolue Moyenne (MAE)

La MAE mesure la moyenne des erreurs absolues entre les valeurs prédites et les valeurs réelles. Mathématiquement elle s'écrit comme suit :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Équation IV.2 Erreur Absolue Moyenne

La MAE est plus interprétable car elle est dans la même unité que la variable prédite.

IV.1.1.4 Coefficient de Détermination (R^2)

Le R^2 mesure la proportion de la variance des données qui est expliquée par le modèle. Et sa formule est :

$$R^2 = 1 - \frac{\sum_{1=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \hat{y})^2}$$

Équation IV.3 Coefficient de détermination

Un R^2 proche de 1 indique un bon ajustement du modèle.

IV.1.2 Comparaison des métriques

Après avoir entraîné et ajusté les quatre modèles pour obtenir les meilleures performances, nous n'avons retenu que deux d'entre eux : SARIMA et LSTM. Les modèles ARIMA et Prophet ont été écartés. ARIMA ne prend pas en compte les facteurs saisonniers de notre série, tandis que Prophet produit des prédictions négatives en raison de sa grande sensibilité aux bruits de notre série. Le tableau suivant présente les comparaisons entre les deux modèles retenus.

Tableau IV.1 Tableau de comparaison des métriques LSTM es SARIMA

Modèle	**MSE**	**RMSE**	**MAE**	**R^2**	**AIC**	**BIC**
LSTM	127667.95	357.31	226.97	-3098085.	---	---
SARIMA	64414.56	253.80	186.77	-0.26	575.40	580.61

Nous précisons que les métriques AIC et BIC sont propres au modèle SARIMA, c'est la raison pour laquelle elles ne figurent pas dans le tableau précédent. C'est lors de l'interprétation des résultats que nous connaîtrons leur importance.

IV.1.3 Visualisation des résultats

Nous commençons par visualiser nos prédictions pour les douze prochains mois et nous combinons nos deux modèles retenus. C'est ce qu'illustre la figure ci-après.

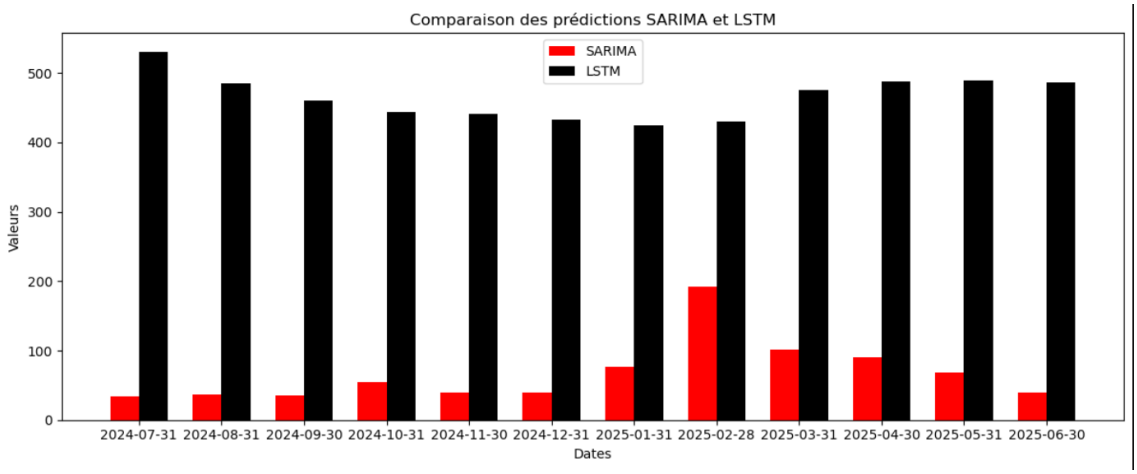


Figure IV.1 Comparaison de valeurs prédites entre LSTM et SARIMA

En termes des chiffres, les prévisions que nous avons faites pour les douze prochains mois partant de la dernière date du dataset, ressemblent à ce que nous présente le tableau ci-après :

Tableau IV.2 Prédictions LSTM et SARIMA

Dates	Modèle LSTM	Modèle SARIMA
2024-07-31	531	34
2024-08-31	485	37
2024-09-30	461	36
2024-10-31	444	54
2024-11-30	441	39
2024-12-31	433	40
2025-01-31	425	77
2025-02-28	430	192
2025-03-31	476	101
2025-04-30	488	91
2025-05-31	489	69
2025-06-30	486	39

La figure 2 illustre les prédictions faites avec SARIMA sur les cas de moins de 5 ans et nous donne également un bref aperçu de ce que peut être la situation future. Dans la même logique, la figure 3 nous permet à son tour de visualiser les prédictions sur les

cas dont l'âge n'est pas précis et enfin la figure 4, montre clairement la courbe de prévisions sur les 12 prochains mois.

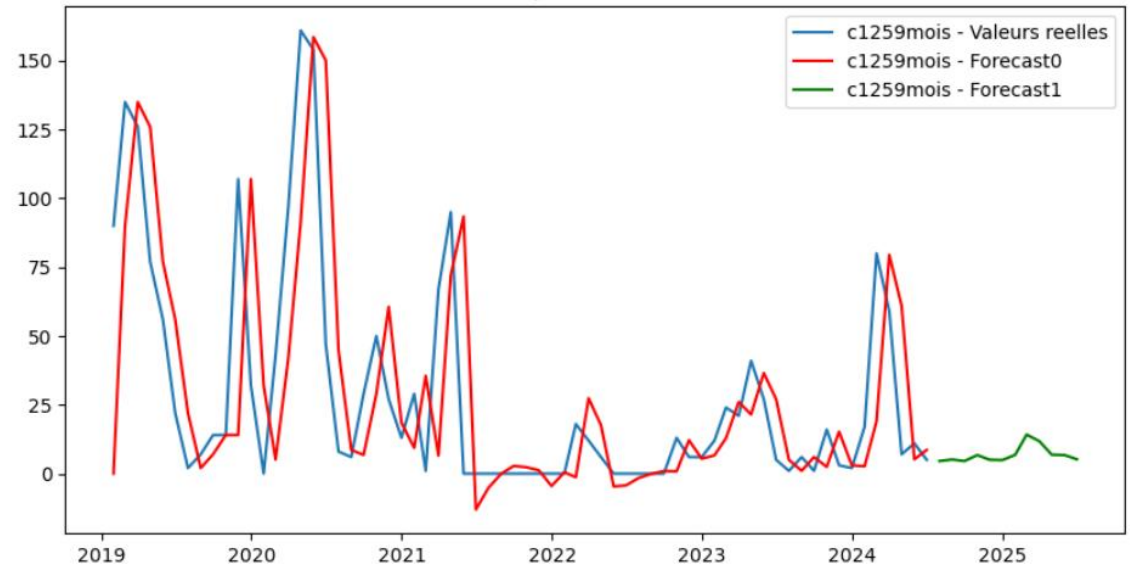


Figure IV.2 prédictions des cas de moins de 5 ans avec SARIMA

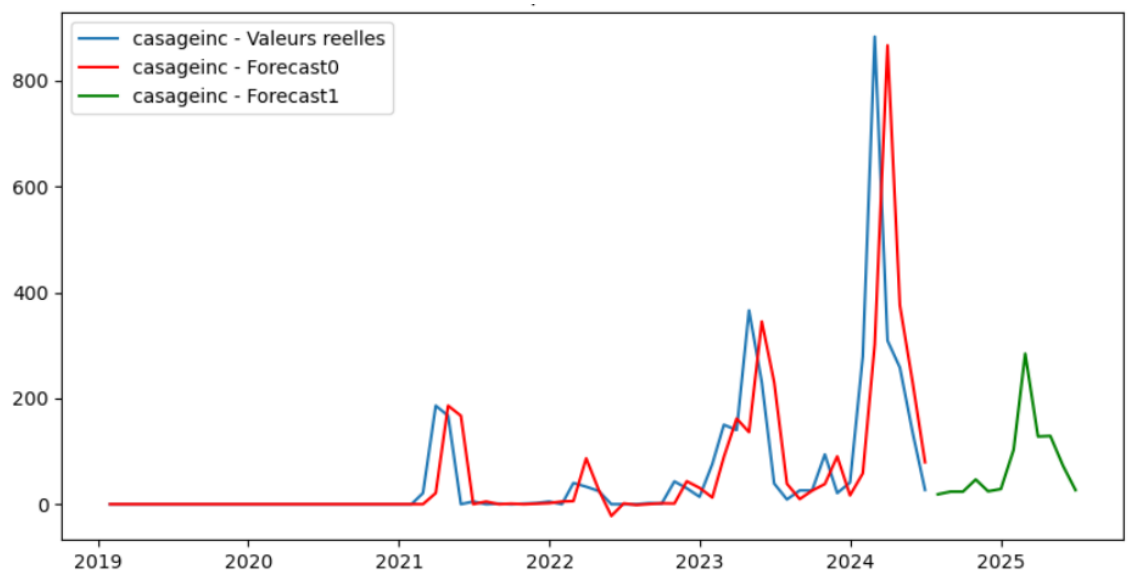


Figure IV.3 Prédictions des cas dont l'âge est inconnu

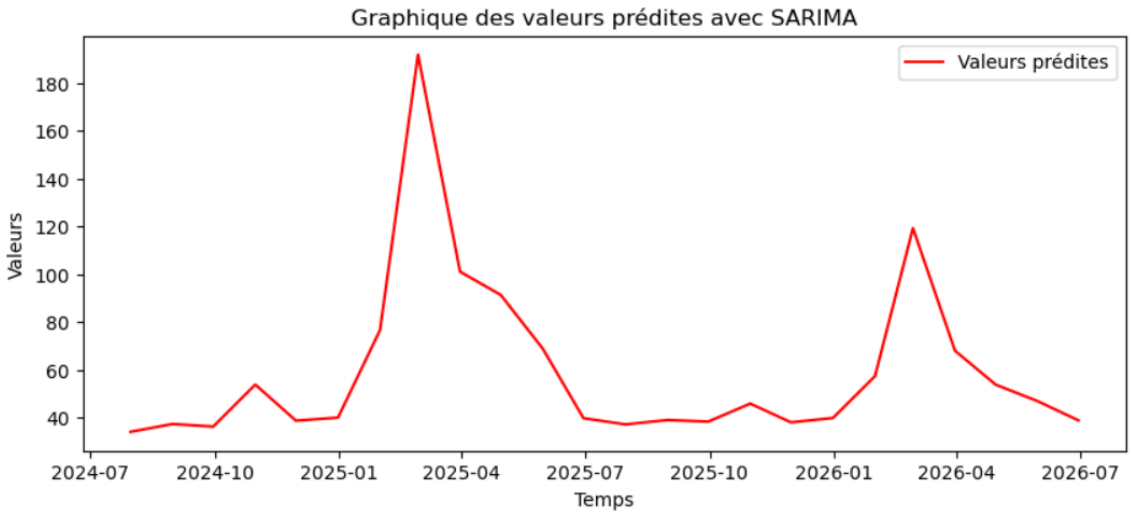


Figure IV.4 Prédictions de 12 prochains mois avec SARIMA

IV.2 Interprétation des résultats

En comparant les performances des modèles LSTM et SARIMA, nous observons que le modèle SARIMA présente des métriques d'erreur plus faibles. Le SARIMA a un MSE de 64414.56 et un RMSE de 253.80, tandis que le LSTM a un MSE de 127667.95 et un RMSE de 357.31. De plus, le MAE du SARIMA est de 186.77, ce qui est inférieur à celui du LSTM, qui est de 226.97. Ces résultats indiquent que le SARIMA a une meilleure précision et des prédictions plus proches des valeurs réelles par rapport au LSTM. En termes de R2, SARIMA a une valeur de -0.26, tandis que LSTM a une valeur de -3098085, ce qui montre que le SARIMA est également supérieur en termes de capacité explicative, bien que les deux modèles aient des R2 négatifs, indiquant une mauvaise performance globale.

En outre, les critères AIC et BIC de SARIMA, respectivement 575.40 et 580.61, suggèrent que ce modèle est bien ajusté et moins complexe. Ces critères ne s'appliquent pas directement au LSTM, mais ils renforcent l'idée que SARIMA est le modèle le plus performant dans ce contexte. Les valeurs prédites par ce dernier reflètent mieux les tendances saisonnières et sont plus proches des données de départ, ce qui est un bon signe de sa capacité à capturer les variations saisonnières sur notre série. En revanche, les valeurs prédites par le modèle LSTM montrent une tendance générale mais semblent moins précises pour capturer les fluctuations spécifiques.

Parlant des chiffres, sur le premier graphique en barre, il s'agit d'une comparaison entre les valeurs prédites entre nos deux modèles retenus (rouge pour SARIMA et noir pour LSTM) ; il en est de même pour le tableau 1. Et les 4 figures précédentes représentent les résultats obtenus (lignes rouges) après entrainements par rapport aux données réelles (lignes bleues), mais surtout les prédictions pour les 12 prochains mois (lignes vertes).

En conclusion, sur la base des métriques (performances) fournies et de la capacité à capturer les tendances saisonnières, le modèle SARIMA semble être le meilleur choix pour nos données de série temporelle. Ainsi, il est évident que c'est ce dernier que nous déploierons et grâce à notre application et aussi à Power Bi, nous pourrions bien explorer les résultats de ce projet.

IV.3 Déploiement

Le déploiement d'un modèle SARIMA avec Flask pour la prédiction du choléra combine la puissance des modèles de séries temporelles avec la simplicité et la flexibilité de Flask. En utilisant Flask, un micro-framework web en Python, il est possible de créer rapidement une API RESTful pour servir les prédictions du modèle.

Flask est choisi pour ce projet en raison de sa simplicité, de sa rapidité de développement et de son extensibilité. Il permet de créer des applications web légères et efficaces, avec la possibilité d'ajouter des extensions selon les besoins spécifiques du projet. De plus, Flask bénéficie d'une large communauté et d'une documentation riche, ce qui facilite la résolution des problèmes et l'implémentation de nouvelles fonctionnalités.

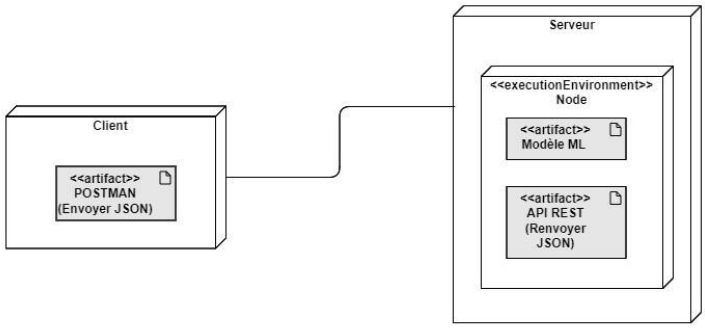


Figure IV.5 Diagramme de déploiement du modèle

Le processus de déploiement comprend plusieurs étapes essentielles : la préparat ion et l'enregistrement du modèle SARIMA, la création d'une API Flask pour gérer les d onnées (périodes) et retourner les prévisions, et enfin, la mise en ligne de l'application s ur un serveur web ou la publication du rapport interactif sur Power BI Service pour le re ndre accessible en ligne. Cette approche permet aux autorités sanitaires d'accéder aux prédictions en temps réel, facilitant ainsi la prise de mesures préventives et l'allocation efficace des ressources pour lutter contre le choléra.

IV.4 Discussion

IV.4.1 Comparaison avec les études existantes

IV.4.1.1 Littérature existante

Notre projet se distingue par l'utilisation de données historiques sur le choléra pour prédire les tendances futures, permettant ainsi à la santé publique d'anticiper les épidémies. Contrairement aux études existantes qui utilisent des données en temps réel ou environnementales, notre approche se concentre sur les données historiques, capturant les tendances et cycles saisonniers spécifiques à la région, ce qui offre une perspective unique et potentiellement plus précise pour la prédiction à long terme.

IV.4.1.2 Innovation et Contribution

Les études antérieures se concentrent sur la détection précoce des épidémies et la gestion des crises en temps réel. En revanche, mon approche utilise des modèles d'apprentissage automatique pour analyser les données historiques et anticiper les tendances futures. Cela permet aux autorités de santé publique de planifier et de mettre en œuvre des mesures préventives avant l'apparition des épidémies, améliorant ainsi la préparation et la réponse.

IV.4.2 Comparaison avec les performances existantes

Les performances des modèles dans les études existantes varient selon les données et les algorithmes utilisés. Certaines études obtiennent des précisions élevées avec des modèles complexes comme les réseaux de neurones, tandis que d'autres utilisent des modèles plus simples comme la régression logistique. Ce projet, utilisant des données historiques spécifiques à la région et des algorithmes adaptés, a montré une capacité robuste à prédire les tendances du choléra. Les métriques de performance telles que le RMSE et le R^2 indiquent que les modèles développés sont compétitifs par rapport aux approches existantes, tout en offrant une meilleure interprétabilité et une applicabilité pratique pour les décideurs.

IV.4.3 Implications pratiques

Les implications pratiques des études existantes se concentrent souvent sur la réponse immédiate aux épidémies, avec des recommandations pour des interventions rapides et ciblées. En anticipant les tendances à long terme, mon projet permet aux autorités de santé publique de mettre en place des stratégies de prévention plus efficaces, telles que des campagnes de vaccination ciblées, des améliorations des infrastructures sanitaires, et des programmes d'éducation communautaire. Cela peut potentiellement réduire l'incidence du choléra de manière plus durable.

IV.4.4 Limitations

Une des principales limitations de ce travail est la nécessité d'une grande quantité de données historiques de haute qualité, ce qui est expliqué par la valeur négative du R^2 lors de l'évaluation, ce qui peut être difficile à obtenir et coûteux à traiter. De plus, les tendances historiques peuvent ne pas toujours refléter les futures épidémies, nécessitant des mises à jour régulières des modèles pour maintenir leur précision.

IV.4.5 Perspectives d'évolution

Intégrer des données environnementales et socio-économiques, expérimenter avec des algorithmes avancés, et mettre à jour automatiquement les modèles serait utile pour l'amélioration de ce projet. Développer des outils de visualisation et collaborer avec les autorités sanitaires, ainsi que mener des études comparatives et partager les résultats, encouragera l'amélioration continue.

Ce travail peut être étendu à la prédiction d'autres maladies infectieuses et au développement de systèmes de surveillance épidémiologique en temps réel. Il peut également aider à améliorer les politiques de santé publique et à sensibiliser les communautés à risque.

Conclusion partielle

En résumé, ce chapitre a présenté et analysé les résultats obtenus à partir des modèles de machine learning appliqués à la prédiction du choléra. Les performances des différents algorithmes ont été évaluées à l'aide de métriques telles que le RMSE et le R^2, démontrant la robustesse et la précision des modèles développés. La comparaison avec les études existantes a mis en évidence les avantages uniques de notre approche basée sur les données historiques, tout en soulignant les limitations liées à la nécessité d'une grande quantité de données de haute qualité.

Les discussions ont également montré comment ces résultats peuvent être utilisés par les autorités sanitaires pour anticiper les tendances et mettre en œuvre des mesures préventives efficaces. Les perspectives d'évolution de ce projet incluent l'intégration de nouvelles données, l'amélioration des modèles, et le développement d'outils pratiques pour les décideurs. Ces efforts contribueront à renforcer la prévention et le contrôle du choléra, tout en ouvrant la voie à des applications étendues pour d'autres maladies infectieuses.

CONCLUSION GENERALE

Nous voici au terme de notre travail intitulé ‘‘Mise en place d’un modèle prédictif des données épidémiologiques’’, dont le cas d’étude est l’épidémie du choléra dans la province du Haut-Katanga, pour estimer le nombre de cas futurs de choléra, basé sur les données historiques fournies par le ministère provincial de la santé. Ce projet a permis d’explorer en profondeur l’utilisation des données historiques pour anticiper les épidémies, offrant ainsi une méthode précise et fiable pour la prévention des crises sanitaires.

La réalisation de ce projet a mis en lumière l’importance cruciale de la qualité et de la quantité des données pour améliorer la précision des prédictions. Les défis rencontrés ont été des occasions d’apprentissage, nous poussant à développer des solutions créatives et à renforcer nos compétences techniques. Les résultats obtenus montrent que les modèles développés sont assez robustes et performants, permettant aux autorités sanitaires de mieux planifier et de mettre en œuvre des mesures préventives efficaces.

Ce projet ouvre la voie à de nombreuses possibilités pour l’avenir. L’intégration de nouvelles données environnementales et socio-économiques, l’expérimentation avec des algorithmes avancés, et le développement d’outils de visualisation pour les décideurs sont autant de pistes prometteuses. En partageant les résultats et les modèles avec la communauté scientifique, nous espérons encourager la collaboration et l’amélioration continue, tout en étendant cette approche à d’autres maladies infectieuses.

En conclusion, ce mémoire démontre le potentiel du machine learning pour anticiper les épidémies de choléra et souligne l’importance d’une approche basée sur les données historiques pour prévenir les crises sanitaires. Nous espérons que notre contribution aidera à renforcer les stratégies de santé publique et à protéger les populations contre les futures épidémies.

REFERENCES

Bibliographie

[1] O. M. d. l. Santé, «Choléra, 2017,» 2017.

[2] O. M. d. l. Santé, «Choléra, 2018,» 2018.

[3] O. M. d. l. Santé, «cholera, 2019,» 2019.

[4] O. M. d. l. Santé, «cholera, 2020,» 2020.

[5] O. M. d. l. Santé, «cholera, 2021,» 2021.

[6] O. M. d. l. Santé, «cholera, 2022,» 2022.

[7] O. M. d. l. Santé, «cholera, 2023,» 2024.

[8] «IBM,» [En ligne]. Available: https://www.ibm.com/topics/machine-learning. [Accès le 04 Août 2024].

[9] «Coursera,» [En ligne]. Available: https://www.coursera.org/articles/what-is-machine-learning. [Accès le 05 Août 2024].

[10] J. Robert, «DataScientest,» 12 10 2020. [En ligne]. Available: https://datascientest.com/data-science-definition. [Accès le 25 09 2024].

[11] Dedocoton, «LEDATASCIENTIST,» 30 05 2024. [En ligne]. Available: https://ledatascientist.com/quest-ce-que-la-data-science/. [Accès le 20 09 2024].

[12] D. Gaultier, «le blog,» 05 Mai 2023. [En ligne]. Available: https://fr.blog.businessdecision.com/intelligence-artificielle-data-science/. [Accès le 27 Septembre 2024].

[13] J. Brownlee, 26 Novembre 2021. [En ligne]. Available: https://machinelearningmastery.com/visualizing-the-vanishing-gradient-problem/. [Accès le 28 09 2024].

[14] J. Brownlee, «Machine Learning Mastery,» 2023 Novembre 2018. [En ligne]. Available: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/. [Accès le 27 Septembre 2024].

[15] Luvimero, «XLSTAT,» 06 Mars 2024. [En ligne]. Available: https://www.xlstat.com/fr/articles/comment-utiliser-arima-pour-les-analyses-de-prevision-et-de-simulation. [Accès le 27 Septembre 2024].

[16] R. Kassel, «Datascientest,» 04 Octobre 2023. [En ligne]. Available: https://datascientest.com/facebook-prophet-tout-savoir. [Accès le 20 Août 2024].

[17] N. HOTZ, «What is CRISP DM ?,» 30 juillet 2024. [En ligne]. Available: https://www.datascience-pm.com/crisp-dm-2/.

[18] A. Crochet-Damais, 2022 Octobre 2022. [En ligne]. Available: https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501895-time-series/. [Accès le 28 09 2024].

[19] «ledigitalizeur,» [En ligne]. Available: https://ledigitalizeur.fr/ia/analyse-predictive/. [Accès le 05 Août 2024].

[20] «Machine learning models,» [En ligne]. Available: https://machinelearningmodels.org/machine-learning-and-prediction/. [Accès le 03 Août 2024].

[21] «Journaldunet,» [En ligne]. Available: https://www.journaldunet.fr/web-tech/guide-du-big-data/1516813-analyse-predictive/. [Accès le 15 Août 2024].

[22] R. Kassel, 15 Février 2021. [En ligne]. Available: https://datascientest.com/arima-series-temporelles. [Accès le 28 Septembre 2024].

[23] Excitedlord, «Geeks for geeks,» 24 Mai 2024. [En ligne]. Available: https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average/. [Accès le 27 Septembre 2024].