Université de Ngaoundéré
Faculté des Sciences
Département de Mathématiques et Informatique

The University of Ngaoundere
Faculty of Science
Department of Mathematics and Computer Science

Mémoire présenté en vue d'obtention du diplôme de Master en Ingénierie Informatique

Parcours : **Systèmes et Logiciels en Environnements Distribués**

DEEP ANALYSIS AND PREDICTION OF CHIKUNGUNYA USING ENSEMBLE REGRESSION APPROACH

MOHAMED EL BACHIR BOUBA NGANADAKOUA
19A666FS
(Licence en Informatique)

Sous la direction de :

Dr. ABBOUBAKAR Hamadjam
Chargé de Cours
Université de Ngaoundéré

Dr. ZONGO MEYO Epse NDO
Chargé de Cours
Université de Ngaoundéré

Année académique : 2023-2024

Plan du travail

1. Introduction
2. Vue d’ensemble sur le chikungunya et l’apprentissage d’ensemble
3. l’état de l’art
4. Matériel et méthodes
5. Résultats et discussions
6. Conclusion et perspectives

Introduction

Introduction

Contexte

Le **2 août 2024** , Le Centre européen de contrôle et de prévention des maladies (**ECDC**) a indiqué qu'environ **350 000 cas** de maladie à virus chikungunya (CHIKVD) et plus de **140 décès** ont été signalés dans le monde en 2024, Ces cas proviennent de 21 pays **d'Amérique**, **d'Asie**, **d'Afrique** et **d'Europe** [ECDC, 2024]. Au total, plus de **2 millions** de cas ont été signalés depuis **2005** [WHO, 2024].

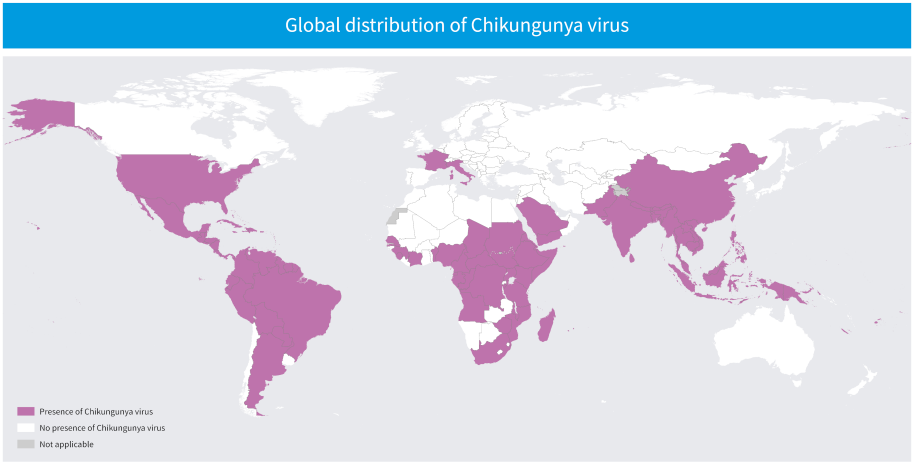

Figure: la répartion dans le monde du virus du Chikungunya [WHO, 2024]

Introduction

Les approches traditionnelles de surveillance épidémiologique se révèlent souvent insuffisantes pour anticiper les épidémies de manière proactive (préventive). Ces limites rendent nécessaire le développement de nouvelles méthodes de prédiction et d'analyse, comme l'utilisation de la régression ensembliste , afin d'améliorer la capacité à anticiper aux nouvelles épidémies.

Introduction

Les approches traditionnelles de surveillance épidémiologique se révèlent souvent insuffisantes pour anticiper les épidémies de manière proactive (préventive). Ces limites rendent nécessaire le développement de nouvelles méthodes de prédiction et d'analyse, comme l'utilisation de la régression ensembliste , afin d'améliorer la capacité à anticiper aux nouvelles épidémies.

Problématique

Comment prédire et analyser les épidémies du chikungunya en utilisant la régression ensembliste ?

Introduction

Objectif général

Développer un modèle prédictif du chikungunya, en utilisant des approches de **régression d'ensemble** dérivées de l'intelligence artificielle, en utilisant les variables climatiques.

Introduction

Objectif général

Développer un modèle prédictif du chikungunya, en utilisant des approches de **régression d'ensemble** dérivées de l'intelligence artificielle, en utilisant les variables climatiques.

Les objectifs spécifiques de ce travail sont les suivants :

1. Comprendre les concepts liés au chikungunya et la régression d'ensemble;
2. Prédire les épidémies de chikungunya en se basant sur les données climatiques et les cas rapportés au Tchad, au Brésil, et au Paraguay ;
3. Evaluer les performances obtenues dans ces 3 pays.

Vue d’ensemble sur le chikungunya et l’apprentissage d’ensemble

Vue d’ensemble sur le chikungunya

Définitions et origines

Le **chikungunya** est une maladie transmise principalement par les moustiques **femelles** *Aedes aegypti* et **Aedes albopictus**. elle est décrite pour la première fois en **1952** lors d’une épidémie dans le sud de la **Tanzanie**. Le nom vient d’un mot de la langue *Makonde*, parlée dans le sud-est de la Tanzanie et le nord du Mozambique, qui signifie ”*ce qui se plie*”.

Vue d’ensemble sur le chikungunya

Figure: Le moustique *Aedes aegypti* plein de sang [CDC]

Vue d’ensemble sur le chikungunya

Le **CHIKV** (Chikungunya Virus) se transmet selon deux cycles différents :

- **Cycle urbain**: transmission de l’homme au moustique.
- **Cycle sylvatique**: transmission de l’animal au moustique, puis à l’homme.

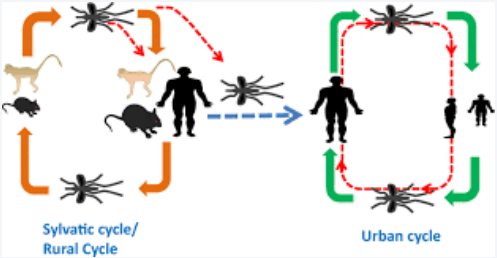

Figure: mode de transmission du Chikungunya [ResearchGate,2024]

Vue d’ensemble sur le chikungunya

Symptômes

Chez les patients symptomatiques, la maladie se déclare généralement **4 à 8 jours** après la **piqûre** d’un **moustique infecté**. Elle se caractérise par une brusque **poussée de fièvre**, souvent accompagnée de **fortes douleurs articulaires**. Les douleurs articulaires durent généralement quelques jours, mais peuvent être prolongées et durer des semaines, des mois, voire des années.

Impact du chikungunya

Au cameroun

En **2006**, plus de **400** épidémie de type dengue ont été signalés à **Kumbo** (région du Nord-Ouest du Cameroun). Bien que les investigations aient été menées un an après cette dernière épidémie, les résultats suggèrent une circulation récente du **CHIKV** dans trois villages de **Kumbo** (Cameroun occidental).

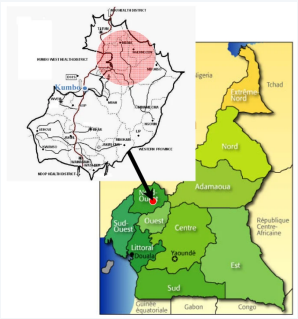

Figure: zone d'impact du CHIKV dans la region du Nord Ouest [Demanou,2010]

Impact du chikungunya

Au Tchad

Le 3 Septembre 2020, dans le rapport de **OMS** (Organisation Mondiale de la Santé) **927** cas ont été notifiés, tous pris en charge en ambulatoire, sans aucun décès. À cette date, le cumul atteignait **13 488** cas, toujours sans décès, avec de nouveaux cas suspects signalés dans les régions **Ouadaï**,**Sila** et **Wadi Fira**.

Figure: Region infecté au Tchad [Dr Jean,2020]

Impact du chikungunya

Au Brésil

le Brésil, notamment dans l'État de Minas Gerais avec **395** cas pour 100 000 habitants. Depuis son introduction au Brésil en 2014, avec **3,6 millions** de cas signalés à la **PAHO** (Pan American Health Organization), la maladie s'est déplacée du Nord-Est vers le Sud-Est.

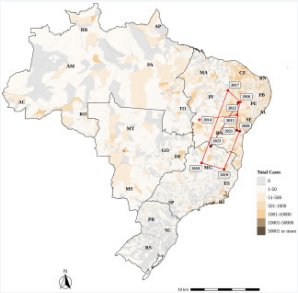

Figure: Carte des épicentres des cas de chikungunya au brésil [Research Gate,2024]

Impact du chikungunya

Au Paraguay

Des infections autochtones ont été détectées au **Paraguay** en 2013 et le CHIKV a été détecté dans le pays chaque année depuis cette date toutes associées aux mois d'été. Du 2 octobre 2022 au 10 avril 2023, un total de **118 179** infections suspectes et confirmées ont été signalées et 46 décès.

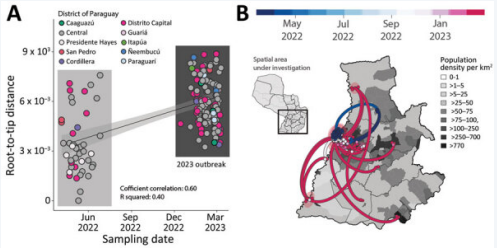

Figure: Cas de chikungunya déclarés chaque semaine au paraguay [Research Gate,2024]

Concepts sur l'apprentissage d'ensemble

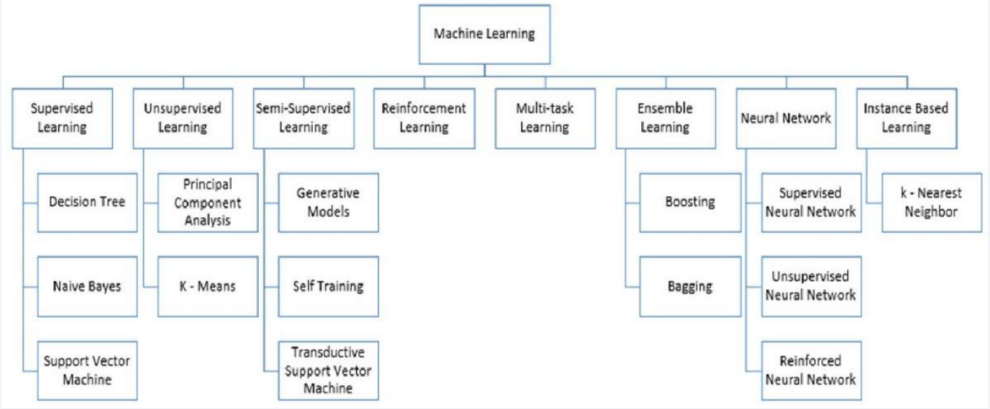

Figure: Structure arborescente du machine learning [Springer,2023]

Concepts sur l’apprentissage d’ensemble

Définition

L’**apprentissage d’ensemble** est une technique d’apprentissage automatique qui regroupe deux ou plusieurs apprenants (par exemple, des **modèles de régression** , des **réseaux neuronaux**) afin de produire de **meilleures prédictions**.

Concepts sur l’apprentissage d’ensemble

Définition

L’**apprentissage d’ensemble** est une technique d’apprentissage automatique qui regroupe deux ou plusieurs apprenants (par exemple, des **modèles de régression** , des **réseaux neuronaux**) afin de produire de **meilleures prédictions**.

Elle repose sur le principe selon lequel une collectivité d’apprenants produit une plus grande précision globale qu’un apprenant individuel, de même utilisé pour réduire la **variance** et diminuer le **biais**.

Concepts sur l'apprentissage d'ensemble

Type de modèle d'ensemble

- Les méthodes **parallèle** : entraînent chaque apprenant de base indépendamment des autres.

Concepts sur l'apprentissage d'ensemble

Type de modèle d'ensemble

- Les méthodes **parallèle** : entraînent chaque apprenant de base indépendamment des autres.
- Les méthodes **séquentielles** : forment un nouvel apprenant de base de manière à minimiser les erreurs commises par le modèle précédent formé à l'étape précédente.

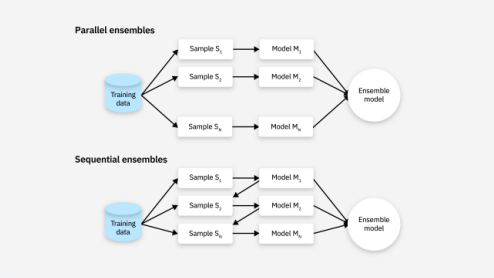

Figure: Type de modèle d'ensemble [IBM,2024]

Concepts sur l’apprentissage d’ensemble

Comment les méthodes d’ensemble combinent-elles les apprenants de base pour former un apprenant final ?

Concepts sur l'apprentissage d'ensemble

Comment les méthodes d'ensemble combinent-elles les apprenants de base pour former un apprenant final ?

Bagging

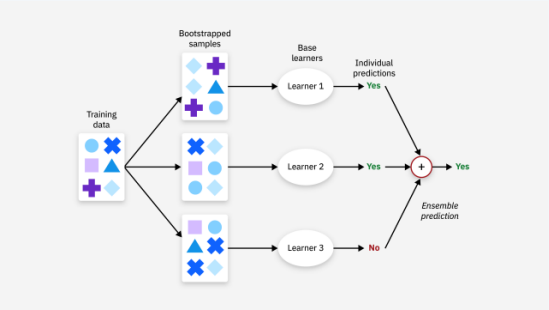

Figure: Bagging [IBM,2024]

Concepts sur l'apprentissage d'ensemble

Comment les méthodes d'ensemble combinent-elles les apprenants de base pour former un apprenant final ?

Bagging

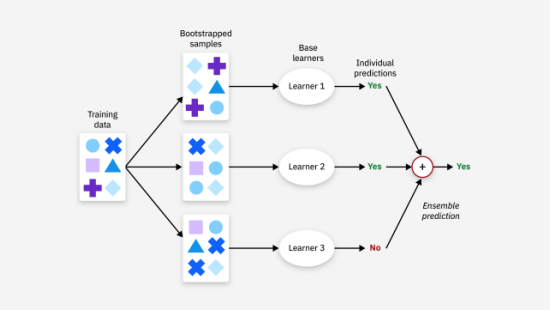

Figure: Bagging [IBM,2024]

Boosting

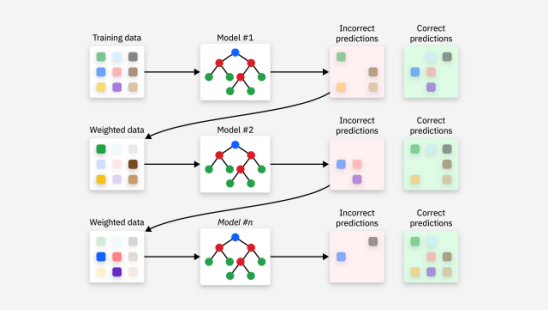

Figure: Boosting [IBM,2024]

l’état de l’art

l'état de l'art

Travaux connexes

Dans [Roster K et Al,2022], prévision de la dengue dans les villes brésiliennes basée sur l'apprentissage automatique à l'aide de variables épidémiologiques et météorologiques

- **Méthode** :
  - algorithme de machine learning (random forests, naive model, gradient boosting regression, a feed-forward neural network ou support vector regression)
  - méthode du feature selection
- **Résultat** : le modèle **random forests** était le plus performant
- **Limites** : non utilisation des modèles d'ensemble et aussi elle est basée sur les cas de Dengue

l'état de l'art

Travaux connexes

Dans [Rekha Gangula et Al,2023], prédiction de la maladie de la dengue basée sur l'apprentissage automatique d'ensemble avec des modèles d'élévation de la performance et de la précision

- **Méthode** : technique d'apprentissage automatique d'ensemble dans des intégrations hybrides
- **Résultat** : identifier les caractéristiques associées à la propagation de la maladie de la dengue et obtenir de meilleures performances
- **Limites** : utilisation des méthodes hybrides et aussi prediction des caractéristique uniquement

Matériel et méthodes

Méthodes de collecte des données

Données épidémiologiques

Les données utilisées dans cette étude proviennent de diverses sources fiables.

Méthodes de collecte des données

Données épidémiologiques

Les données utilisées dans cette étude proviennent de diverses sources fiables.

- Pour le **Tchad** : les données épidémiologiques ont été extraites d'un rapport OMS lors de l'émergence du Chikungunya en **2020**. Ce dataset couvre la période allant du *12 août 2020* au *10 novembre 2020*

Méthodes de collecte des données

Données épidémiologiques

Les données utilisées dans cette étude proviennent de diverses sources fiables.

- Pour le **Tchad** : les données épidémiologiques ont été extraites d'un rapport OMS lors de l'émergence du Chikungunya en **2020**. Ce dataset couvre la période allant du *12 août 2020* au *10 novembre 2020*
- Concernant le **Brésil** : les données ont été recueillies à partir du site de mendeley. Cet ensemble de données présente des informations cliniques, **sociodémographiques** et de laboratoire relatives aux patients confirmés atteints de *dengue* et de *chikungunya*. Il couvre la période de **2013** à **2020**, mais pour cette thèse, nous avons restreint notre analyse à l'intervalle de **2013** à **2017**.

Méthodes de collecte des données

Données épidémiologiques

Les données utilisées dans cette étude proviennent de diverses sources fiables.

- Pour le **Tchad** : les données épidémiologiques ont été extraites d'un rapport OMS lors de l'émergence du Chikungunya en **2020**. Ce dataset couvre la période allant du *12 août 2020* au *10 novembre 2020*
- Concernant le **Brésil** : les données ont été recueillies à partir du site de mendeley. Cet ensemble de données présente des informations cliniques, **sociodémographiques** et de laboratoire relatives aux patients confirmés atteints de *dengue* et de *chikungunya*. Il couvre la période de **2013** à **2020**, mais pour cette thèse, nous avons restreint notre analyse à l'intervalle de **2013** à **2017**.
- Pour le **Paraguay** : les données ont été collectées via le site de la PAHO , qui rapporte les cas de Chikungunya en temps réel, avec des enregistrements hebdomadaires variant entre 2013 et 2017.

Méthodes de collecte des données

Données climatiques

Les données climatiques pour ces trois pays ont été obtenues à partir du site **weatherandclimate**, correspondant aux mêmes intervalles temporels que les cas de Chikungunya dans chaque pays à savoir :

- Au **Tchad** : dans les villes de biltine,Abeche et Abdi oú il y'à eu l'epidemie.
- Au **Brésil** : dans les villes Amapá , Bahia ,Ceará , Espírito Santo ,Distrito Federal ,Goiás ,Maranhão , Minas Gerais, Mato Grosso do Sul, Mato Grosso ,Pará, Paraíba,Pernambuco, Piauí,Paraná,Rio de Janeiro,Rio Grande do Norte, Rondônia, Roraima, Rio Grande do Sul, Santa Catarina,Sergipe,São Paulo et Tocantins oú il y'à eu l'epidemie.
- Au **Paraguay** : asuncion,central oú il y'à eu l'epidemie.

.

impact du climat sur les cas du chikunguya

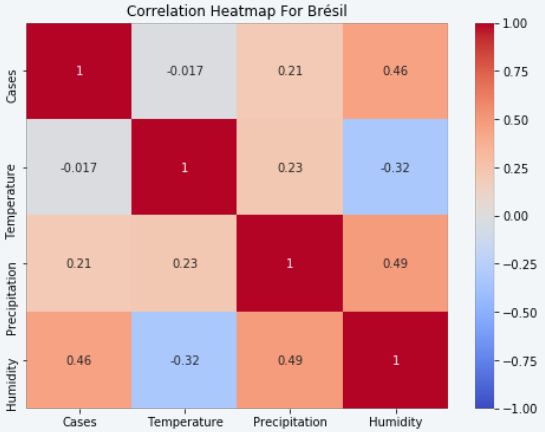

Figure: Corrélation entre les variables climatiques et le nombre de cas

Exploration et Préparation des Données

La préparation des données pour le **Tchad** et le **Paraguay** a présenté des défis importants en raison du manque de données **suffisantes**.

Exploration et Préparation des Données

La préparation des données pour le **Tchad** et le **Paraguay** a présenté des défis importants en raison du manque de données **suffisantes**.
Pour pallier ces lacunes, nous avons envisagé plusieurs méthodes de traitement des données manquantes. Les méthodes sélectionnées sont les suivantes :

Exploration et Préparation des Données

La préparation des données pour le **Tchad** et le **Paraguay** a présenté des défis importants en raison du manque de données **suffisantes**.
Pour pallier ces lacunes, nous avons envisagé plusieurs méthodes de traitement des données manquantes. Les méthodes sélectionnées sont les suivantes :

- **KNN Imputer** : Cette méthode utilise les k-plus proches voisins pour estimer les valeurs manquantes en se basant sur les données les plus proches.

Exploration et Préparation des Données

La préparation des données pour le **Tchad** et le **Paraguay** a présenté des défis importants en raison du manque de données **suffisantes**.
Pour pallier ces lacunes, nous avons envisagé plusieurs méthodes de traitement des données manquantes. Les méthodes sélectionnées sont les suivantes :

- **KNN Imputer** : Cette méthode utilise les k-plus proches voisins pour estimer les valeurs manquantes en se basant sur les données les plus proches.
- **Data augmentation** ou **Augmentation de données** : est une technique utilisée pour augmenter artificiellement la taille d'un jeu de données en générant de nouvelles données à partir des données existantes.

Feature engineering

Le **feature engineering** est une technique qui consiste à créer de nouvelles variables (features) à partir des données brutes.

Feature engineering

Le **feature engineering** est une technique qui consiste à créer de nouvelles variables (features) à partir des données brutes.

Dans le cadre de notre étude, nous avons utilisé cette approche pour améliorer les données épidémiologiques du Brésil. En effet, lors de la collecte des données sur les cas de Chikungunya, nous avons rencontré un ensemble de données comprenant à la fois des patients testés positifs pour la dengue et pour le Chikungunya. Nous avons alors croisé les données des patients testés positifs par le Chikungunya avec les dates de détection de leur maladie. Cela nous a permis de créer une nouvelle variable (**feature**) associant chaque **cas** à une date spécifique, ce qui a enrichi notre base de données épidémiologiques.

Feature engineering

Nous avons aussi utilisé de certaines technique :

Feature engineering

Nous avons aussi utilisé de certaines technique :

- **Shifting** : Cette technique est particulièrement utile pour capturer les dépendances temporelles, car elle permet au modèle d'intégrer l'influence des valeurs passées sur les valeurs futures.

Feature engineering

Nous avons aussi utilisé de certaines technique :

- **Shifting** : Cette technique est particulièrement utile pour capturer les dépendances temporelles, car elle permet au modèle d'intégrer l'influence des valeurs passées sur les valeurs futures.
- **Rolling** : Cette technique est particulièrement utile pour réduire le bruit dans les séries temporelles, Cette méthode permet de lisser les fluctuations à court terme et de mettre en évidence les tendances ou cycles à plus long terme dans les données

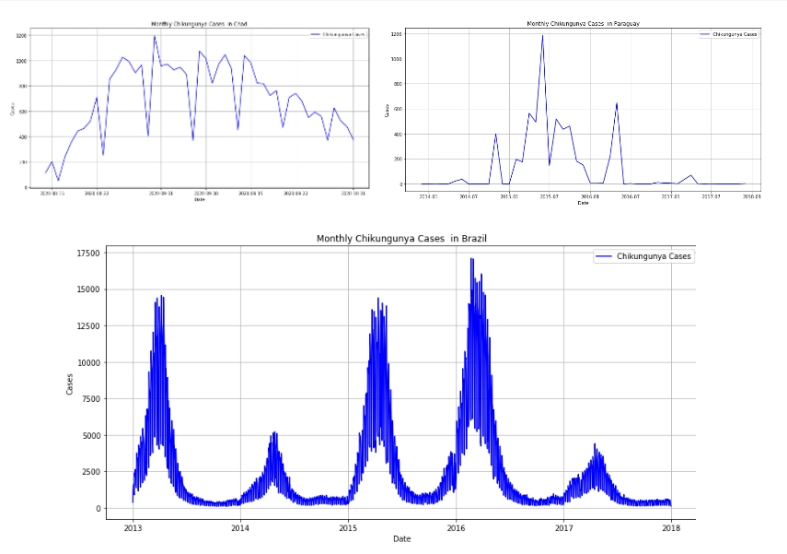

Figure: Visualisation numériques des cas du CHIKV au cours du temps dans les trois pays

Les métriques d'évaluation des modèles

Métrique	Formule mathématique
MAE(Erreur Absolue Moyenne)	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
RMSE(Racine de l'Erreur Quadratique Moyenne)	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
R^2 (Coefficient de détermination)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

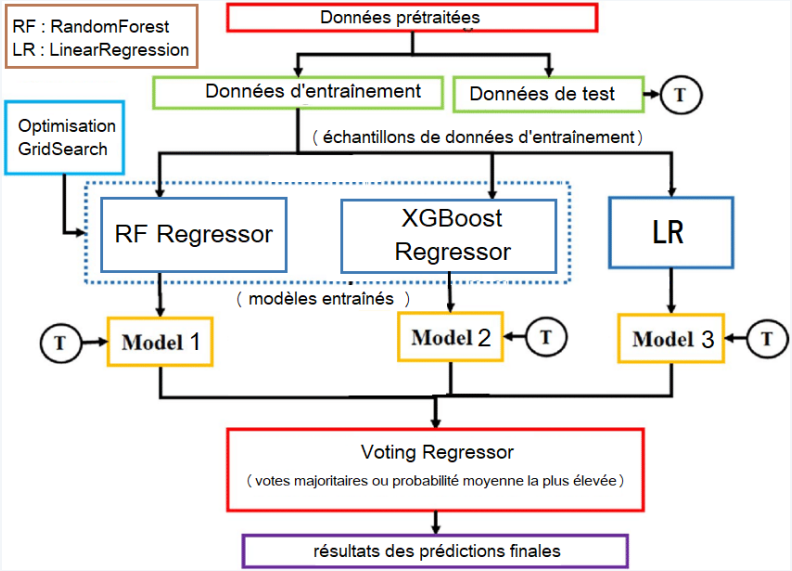

Figure: Architecture de notre méthode

Résultats et discussions

Résultats et discussions

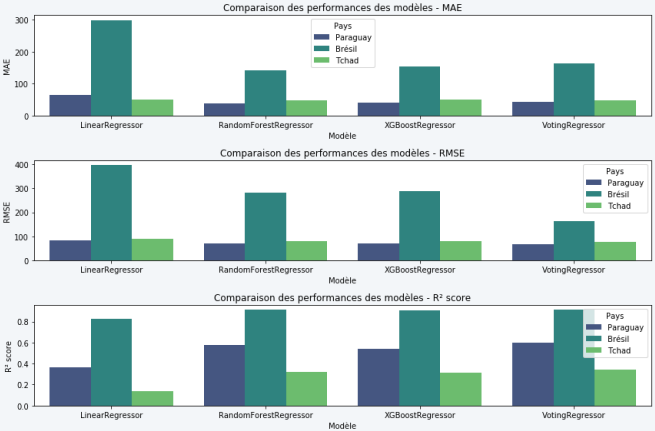

Figure: Comparaison performance

Table: Tableau de comparaison des différents modèles

Pays	Modèle	MAE	RMSE	R^2 score
Paraguay	LinearRegressor	65.4165	85.5322	0.3623
	RandomForestRegressor	38.4501	69.9400	0.5736
	XGBoostRegressor	40.4713	72.3494	0.5437
	VotingRegressor	**43.0001**	**68.1025**	**0.5957**
Brésil	LinearRegressor	299.0416	396.7418	0.8229
	RandomForestRegressor	141.0863	281.1369	0.9111
	XGBoostRegressor	154.3493	288.9261	0.9061
	VotingRegressor	**163.0350**	**163.0350**	**0.9108**
Tchad	LinearRegressor	51.5692	90.1886	0.1344
	RandomForestRegressor	49.2575	79.7674	0.3229
	XGBoostRegressor	51.1603	80.1259	0.3168
	VotingRegressor	**47.6738**	**78.5506**	**0.3434**

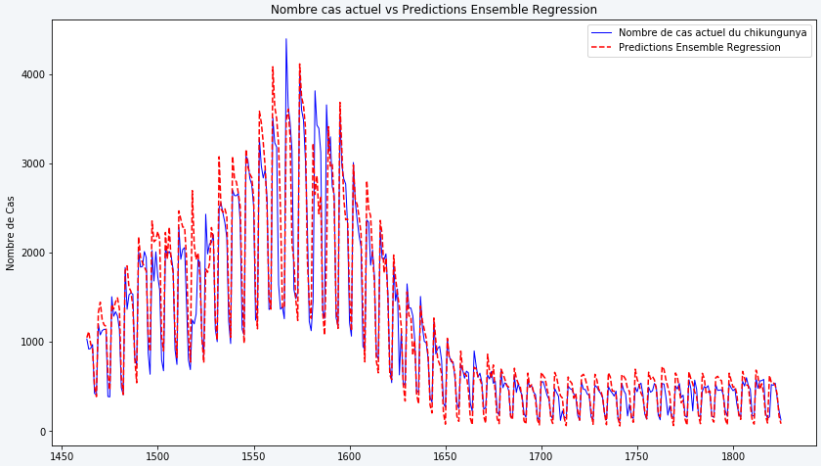

Figure: Prédiction **Brésil** sur les ensembles de donnée de test avec notre modèle d'ensemble (**VotingRegressor**)

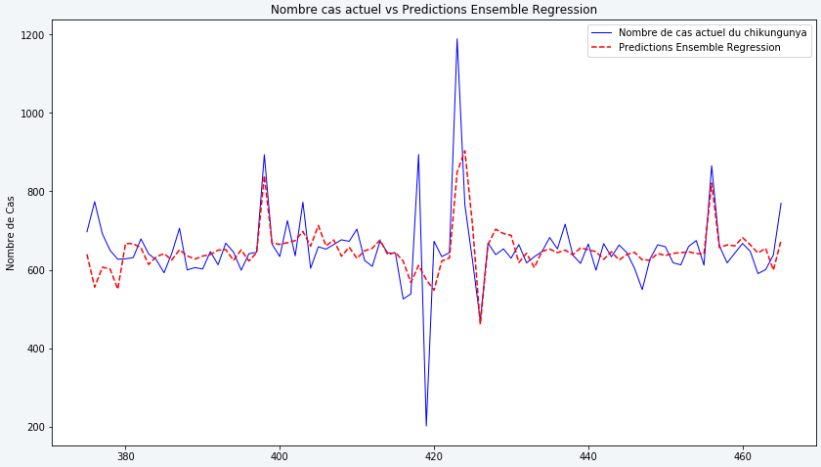

Figure: Prédiction **Tchad** sur les ensembles de donnée de test avec notre le modèle d'ensemble (**VotingRegressor**)

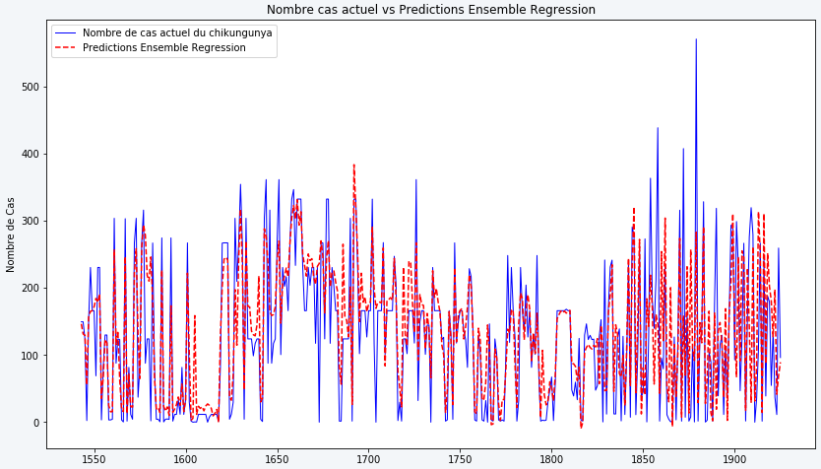

Figure: Prédiction **Tchad** sur les ensembles de donnée de test avec notre modèle d'ensemble (**VotingRegressor**)

Conclusion et perspectives

Conclusion et perspectives

L'objectif principal de cette étude était de Développer un modèle prédictif du chikungunya, en utilisant des approches de **régression d'ensemble**, en utilisant les variables climatiques. Les données épidémiologiques ont été obtenues sur le site du monitoring des cas d'épidomologie en temps réel **PAHO** (pour le Paraguay) ,dans le rapport d'**OMS**(pour le cas du Tchad) et dans le site **mendeley** pour celui du brésil tandis que les données climatiques provenaient de sources fiables telles que le site **WeatherAndClimate**. Les modèles choisis pour cette étude incluaient le **Random Forest Regressor** et le **XGBoost Regressor** optimisé via **Grid Search**, ainsi qu'un modèle d'ensemble (**Voting Regressor**) combinant **Linear Regression**, **Random Forest Regressor** et le **XGBoost Regressor** optimisé.

Conclusion et perspectives

Parmi ces modèles, notre modèle d'ensemble **Voting Regressor**, qui combine les prédictions des modèles **Linear Regression**, **Random Forest Regressor** et **XGBoost Regressor**, a affiché des performances globalement supérieures aux autres modèles individuels, avec un **MAE** minimal, un **RMSE** relativement bas et une très bonne précision (**91,08%** pour le Brésil, **34,34%** pour le Tchad et **59,57%** pour le Paraguay).
Les limites de l'étude ont montré que les variables climatiques ne suffisent pas à elles seules d'expliquer les variations des cas de chikungunya dans ces pays.

Perspectives

Intégrer d'autres facteurs environnementaux et de développer un modèle hybride pour améliorer les prévisions.

Bibliographie

- European Centre for Disease Prevention and Control (2024)
  Chikungunya worldwide overview
  https://www.ecdc.europa.eu/en/chikungunya-monthly.
- World Health Organization (2024)
  Chikungunya worldwide overview
  https://www.who.int/health-topics/chikungunya#tab=tab_1.
- CDC: Transmission of chikungunya virus.
  https://www.cdc.gov/chikungunya/php/transmission/index.html
- Infographic showing transmission cycles of chikungunya virus(chikv)
  https://www.researchgate.net/figure/Infographic-showing-transmission-cycles-of-chikungunya-virus-CHIKV-The-virus-is_fig3_339364587

Bibliographie

- Study of the optimization control of agricultural greenhouse climatic parameters by the integration of machine learning figure on springer (2023). https://link.springer.com/chapter/10.1007/978-3-031-43520-1_28
- A retrospective serological and entomological survey (2010) Demanou, M., Antonio-Nkondjio, C., Ngapana, E., Rousset, D., Paupy, C., Manuguerra, J.C., Zeller, H.: Research article chikungunya outbreak in a rural area of western cameroon in 2006
- Dr Jean Bosco NDIHOKUBWAYO, D.B.H.e.a.: Rapport de la situation Épidémiologique chikungunya
- The expansion of chikungunya in brazil- scientific figure on research-gate(2024). Available from :https://www.researchgate.net/figure/Map-of-the-epicenters-of-chikungunya-cases-each-year-and-the-accumulated-cases-per_fig1_373336451

Bibliographie

- Rapid epidemic expansion of chikungunya virus east/central/south african lineage , paraguay-scientific figure on research gate(2024).Available from: https://www.researchgate.net/figure/Expansion-of-the-chikungunya-East-Central-South-African-lineage-epidemic-in-Paraguay-A_fig2_372620121
- What is ensemble learning?.Available from: https://www.ibm.com/topics/ensemble-learning
- Machine-Learning-Based Forecasting of Dengue Fever in Brazilian Cities Using Epidemiologic and Meteorological Variables.
  Connaughton C, Rodrigues FA. Am J Epidemiol. 2022 Sep 28;191(10):1803-1812. PMID: 35584963. doi: 10.1093/aje/kwac090.
- Ensemble machine learning based prediction of dengue disease with performance and accuracy elevation patterns. doi: https://doi.org/10.1016/j.matpr.2021.07.270.

MERCI POUR VOTRE AIMABLE ATTENTION ! ! ! !