

**Département d’Informatique**

**Master :**

**Intelligence artificielle et analyse des données**

**Projet 2**

**Intitulé** :

**Acidification des océans**

**Réalisé par :**: **Encadré par** :

* **HADDOUCHE Jamal Pr. BEKRI Ali**

**MR. BA-ICHOU Ayoub**

Soutenu le xx devant le jury :

[par ordre alphabétique des noms de famille]

**xxx, Professeur à la Faculté des Sciences- Meknès**

**xxx, Professeur à la Faculté des Sciences- Meknès**

**Année Universitaire: 2022-2023**

Table des matiéres

Introduction.............................................................................................................................................5

[I. Matériels et méthodes 4](#_Toc125750937)

[A. Ensemble des données 4](#_Toc125750938)

[1. Pacifica dataset 4](#_Toc125750939)

[B. Modéle d’apprentissage automatique et profond utilisé 4](#_Toc125750940)

[1. Support vector regressor 4](#_Toc125750941)

[2. Random Forest Regressor 4](#_Toc125750942)

[3. Xgboost: Extreme gradient boosting 4](#_Toc125750943)

[4. Artificial Neural Networks 4](#_Toc125750944)

[5. Les métriques 4](#_Toc125750945)

[6. Grid Search CV 5](#_Toc125750946)

[II. Mise en pratique et résultats 5](#_Toc125750947)

[A. Nettoyage des données et extraction des caractéristiques 5](#_Toc125750948)

[B. Résultats obtenus par les modéles d’apprentissages automatiques 6](#_Toc125750949)

[C. Résultats obtenus par les modeles d’apprentissages profond 8](#_Toc125750950)

[1. Experience pour choisir le nombre des couches et des neurones de chaque couche 8](#_Toc125750951)

[III. Discussion 10](#_Toc125750952)

[A. Discussion des résultats obtenus par les modéle d’apprentissage automatique 10](#_Toc125750953)

[B. Discussion des résultat obtenus par le réseaux de neurones artificiels 11](#_Toc125750954)

[IV. Simulation 11](#_Toc125750955)

[A. Outils 11](#_Toc125750956)

[B. Interfaces 12](#_Toc125750957)

Conclusion et perspectives....................................................................................................................16

**Introduction**

En 2015, 32.1 milliards de tonnes de CO2 ont été libérés dans l'atmosphère suite aux activités humaine, principalement la combustion des carburants fossiles : charbon, gaz naturel, pétrole. Les oceans constituent des puits pour l'absorption du carbone atmosphérique, et participent ainsi à contrebalancer le réchauffement climatique : ils capturent 800 kilos de CO2 de l'atmosphère par seconde. Sans ces puits de carbone maritime, l'effet de serre serait bien plus massif sur terre. L'absorption du C02 atmosphérique par les océans conduit à leurs acidifications (diminution du pH). Le pH des eaux de mers est passé de 8,16 au milieu du 19ème siècle à 8,05 aujourd'hui, ce qui correspond à une acidité accrue de 10%.

L'acidification des océans, est devenue alors l'un des problèmes majeurs qui influence négativement sur l'environnement, sur l'Homme, ainsi que sur plusieurs secteurs économiques et socio- économiques. Vue l'importance de ce problème, plusieurs projets de recherche expérimentaux ont été réalisés pendant plusieurs années depuis 1936, sur différentes régions de différent océans. Ces études ont permet de récolter, une grande quantité d'information (des mesures), sur les paramètres qui jouent un rôle direct, ou indirect sur le phénomène d'acidification des océans, à savoir la température, la salinité, l'alcalinité, l'oxygène dissous, la pression, la profondeur, le CO2 totale et le CO2 anthropique. Ces informations ont été organisées et stockées dans des bases de données. L'objectif de notre projet est de développer une application de calculer le pH de l'eau de mer (acidification) sur base de certains paramètres physico-chimique comme la pression la température totale carbonique , phosphate et la salinité.

# Matériels et méthodes

## Ensemble des données

### Pacifica dataset

PACIFICA était un projet collaboratif international pour la synthèse de données sur le carbone intérieur de l’océan et ses paramètres associés dans l’océan Pacifique. L’Organisation des sciences marines du Pacifique Nord (PICES), Section du carbone et du climat (S-CC) a soutenu le projet. Ils ont fusionné des ensembles de données hydrographiques 14/ hydrochimiques d’un total de 213 croisières, y compris celles des croisières menées entre la fin des années 1980 et 2000 mais non stockées dans GLODAP, ainsi que les ensembles de données CLIVAR / CO2 Repeat Hydrography des années 2000. Les valeurs d’ajustement ont été suggérées pour tenir compte des compensations analytiques dans les données de carbone inorganique dissous, d’alcalinité totale, de salinité, d’oxygène et de nutriments (nitrate et nitrite, phosphate et acide silicique) pour chaque campagne à la suite du contrôle de qualité secondaire.

La base de données PACIFICA se compose d’ensembles de données provenant d’un total de 306 croisières qui comprennent également 34 ensembles de données des croisières du programme hydrographique WOCE dans l’océan Pacifique menées dans les années 1990.

## Modéle d’apprentissage automatique et profond utilisé

### Support vector regressor

La régression par vecteur de support, comme son nom l'indique, est une régression qui prend en charge les régressions linéaires et non linéaires. Cette méthode fonctionne sur le principe de la machine à vecteur de support. Dans la régression simple, l'idée est de minimiser le taux d'erreur tandis que dans la SVR, l'idée est d'ajuster l'erreur à l'intérieur d'un certain seuil, ce qui signifie que le travail de la SVR est d'approcher la meilleure valeur dans une marge donnée appelée ℇ-tube.

### Random Forest Regressor

Une forêt aléatoire est un méta-estimateur qui ajuste un certain nombre d'arbres de décision de classification sur divers sous-échantillons de l'ensemble de données et utilise la moyenne pour améliorer la précision prédictive et contrôler l'ajustement excessif .

### Xgboost: Extreme gradient boosting

Le boosting de gradient est une technique d'apprentissage automatique pour les problèmes de régression et de classification, qui produit un modèle de prédiction sous la forme d'un ensemble de modèles de prédiction faibles, généralement des arbres de décision .

### Artificial Neural Networks

Les réseaux de neurones artificiels sont des systèmes informatiques conçus pour simuler la manière dont le cerveau humain analyse et traite les informations. Les réseaux neuronaux artificiels ont des capacités d'auto-apprentissage qui leur permettent de produire un meilleur résultat au fur et à mesure que des données sont disponibles. Ainsi, si le réseau est formé sur davantage de données, il sera plus précis car ces réseaux neuronaux apprennent à partir d'exemples.

### Les métriques

Pour évaluer la performances d’un modèle d’apprentissage , on a utilisé les métriques : le coefficient () (Eq.8), L'erreur quadratique moyenne (MSE) (Eq.9), the moyenne quadratique (RMSE) (Eq.10), et l‘Erreur absolue moyenne (MAE) (Eq. 11). Les equations sont comme suit:

(8)

(9)

(10)

(11)

Si la valeur du score MSE est plus petite, cela signifie que vous êtes très proche de déterminer la ligne de meilleur ajustement, ce qui dépend également des données sur lesquelles vous travaillez, donc parfois il n'est pas possible d'obtenir une petite valeur de score MSE.

### Grid Search CV

Le grid search cv essaie toutes les combinaisons exhaustives de valeurs de paramètres et choisit la meilleure d'entre elles.

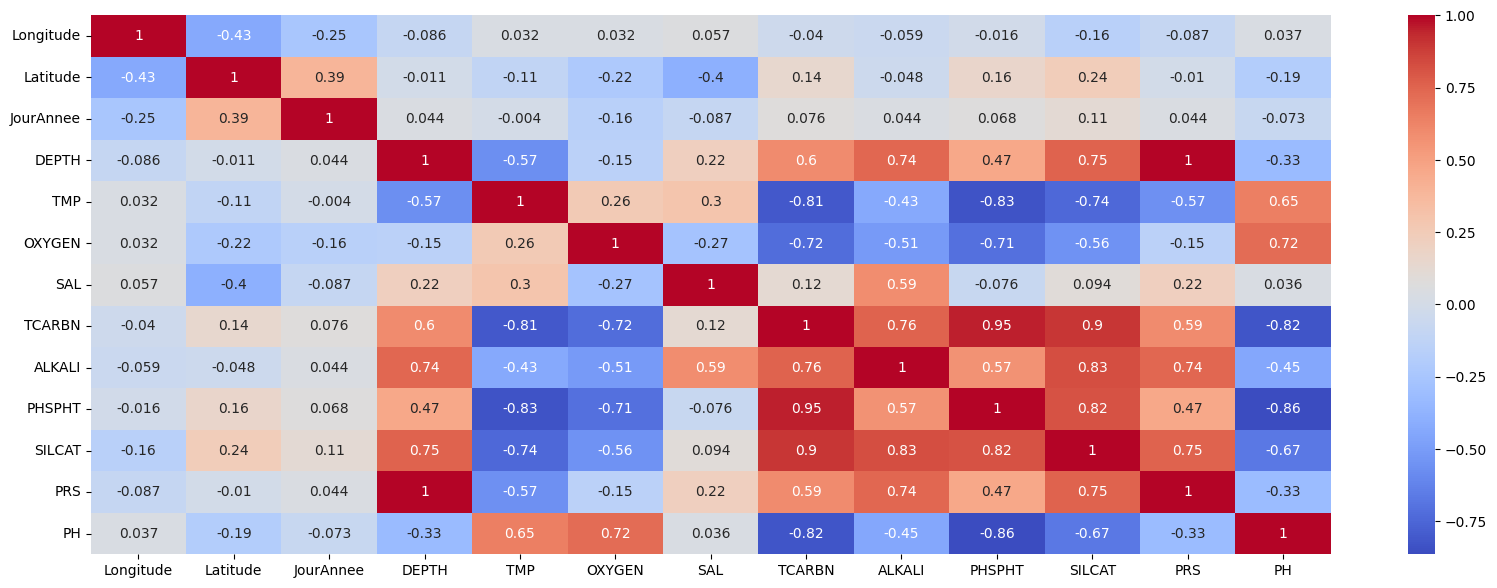
Nous avons choisi ces algorithmes car ils sont devenus plus populaire dans le domaine d'apprentissage automatique en raison de leurs évolutivité.

# Mise en pratique et résultats

## Nettoyage des données et extraction des caractéristiques

Pour le nettoyage des données on a supprimer tous les lignes contenant des valeurs null.

Pour l extraction des caractéristique à utiliser dans l’entrainement de mes modéles, on a utiliser la matrice de correlation (Une **matrice** utilisée pour évaluer la **dépendence** entre plusieurs variables en même temps. Le résultat est une table contenant les **coefficients de corrélation** entre chaque variable et les autres.)



**Fig. 1.** Matrice de correlation

D’après la matrice de corrélation on a choisi les cinq variables qui influencent plus dans la variation de PH , ces variables sont :

**Température Phosphate**

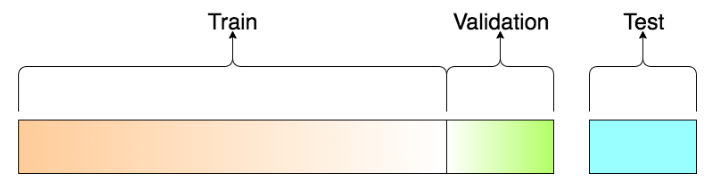
**Oxygéne Totale carbonique**

**Silicate**

## Résultats obtenus par les modéles d’apprentissages automatiques

Dans cette section, les trois modèles (SVR, RF, XGBoost) d'apprentissage automatique ont été utilisés pour prédire la valeur PH, chaque modèle étant entraîné avec chaque combinaison possible de variable.

**Scénario suivi :** Une crosse validation de 0,8 train et 0,2 validation de données de d’entrainement puis un test avec 20% de données .



60% Données

20%

Données

20%

Données

**Résultats obtenus:**

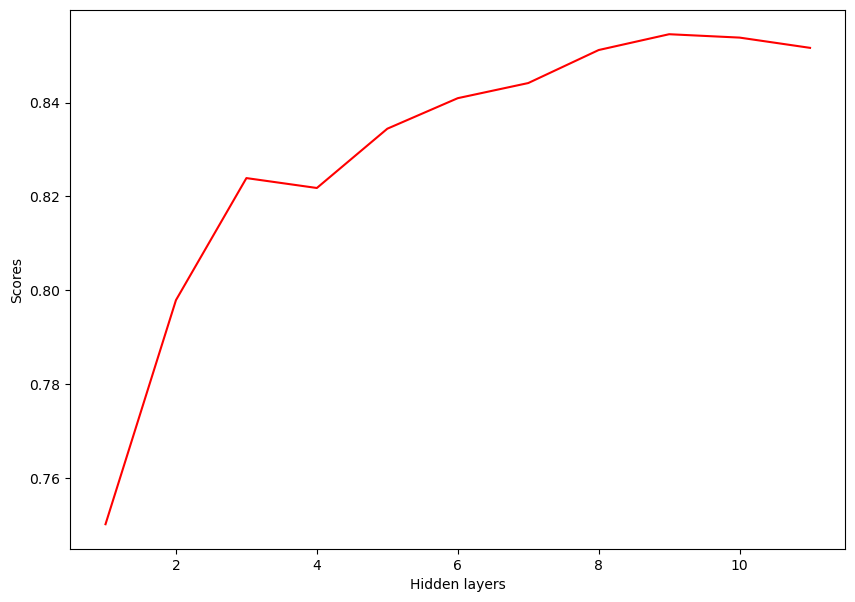
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nombre de variable | Combinaison | SVR | RF | XGB | Best  Model |
| 5 variables | 5variables | R2= 0.8850  Mae=0.0558 | R2 =0.8021  Mae=0.0867 | R2 =0.8335  Mae=0.0717 | Svr |
| 4 VARIABLE | SILCAT\_TCARBN\_OXYGEN\_TMP\_ | R2 0.8849  Mae=0.0558 | R2 0.7971  Mae=0.0874 | R2 0.8239  Mae=0.0748 | Svr |
| PHSPHT\_TCARBN\_OXYGEN\_TMP\_ | R2 0.8666  Mae=0.0621 | R2 0.8008  Mae=0.0868 | R2 0.8207  Mae=0.0774 | Svr |
| PHSPHT\_SILCAT\_OXYGEN\_TMP\_ | R2 0.8648  Mae=0.0622 | R2 0.7996  Mae=0.0866 | R2 0.8262  Mae=0.0741 | Svr |
| PHSPHT\_SILCAT\_TCARBN\_TMP | R2 0.8591  Mae=0.0623 | R2 0.7996  Mae=0.0866 | R2 0.8253  Mae=0.0741 | Svr |
| PHSPHT\_SILCAT\_TCARBN\_OXYGEN\_ | R2 0.8798  Mae=0.0572 | R2 0.7987  Mae=0.0868 | R2 0.82169  Mae=0.0757 | Svr |
| 3 Variables | TCARBN\_OXYGEN\_TMP\_ | R2 0.8634  Mae=0.0629 | R2 0.7866  Mae=0.0900 | R2 0.8041  Mae=0.0819 | Svr |
| SILCAT\_OXYGEN\_TMP\_ | R2 0.8637  Mae=0.0625 | R2 0.7805  Mae=0.0913 | R2 0.8108  Mae=0.0787 | Svr |
| PHSPHT\_OXYGEN\_TMP\_ | R2 0.8406  Mae=0.0683 | R2 0.7909  Mae=0.0893 | R2 0.8057  Mae=0.0828 | Svr |
| SILCAT\_TCARBN\_TMP\_ | R2 0.8520  Mae=0.0638 | R2 0.7797  Mae=0.0900 | R2 0.8051  Mae=0.0792 | Svr |
| PHSPHT\_TCARBN\_TMP\_ | R2 0.8357  Mae=0.0676 | R2 0.7917  Mae=0.0884 | R2 0.8056  Mae=0.0820 | Svr |
| PHSPHT\_SILCAT\_TMP\_ | R2 0.8547  Mae=0.0631 | R2 0.7874  Mae=0.0893 | R2 0.8100  Mae=0.0796 | Svr |
| SILCAT\_TCARBN\_OXYGEN | R2 0.8786  Mae=0.0575 | R2 0.7867  Mae=0.0908 | R2 0.8046  Mae=0.0811 | Svr |
| PHSPHT\_TCARBN\_OXYGEN | R2 0.8542  Mae=0.0655 | R2 0.7928  Mae=0.0893 | R2 0.8067  Mae=0.0826 | Svr |
| PHSPHT\_SILCAT\_OXYGEN | R2 0.8589  Mae=0.0645 | R2 0.7894  Mae=0.0903 | R2 0.8062  Mae=0.0802 | Svr |
| PHSPHT\_SILCAT\_TCARBN | R2 0.8444  Mae=0.0673 | R2 0.7853  Mae=0.0901 | R2 0.7986  Mae=0.0834 | Svr |
| 2 Variables | OXYGEN\_TMP | R2 0.8183  Mae=0.0726 | R2 0.7754  Mae=0.0925 | R2 0.7834  Mae=0.0889 | Svr |
| TCARBN\_TMP | R2 0.8115  Mae=0.0722 | R2 0.7746  Mae=0.0906 | R2 0.7771  Mae=0.0887 | Svr |
| SILCAT\_TMP | R2 0.8110  Mae=0.0708 | R2 0.7235  Mae=0.0973 | R2 0.7319  Mae=0.0943 | Svr |
| PHSPHT\_TMP | R2 0.8409  Mae=0.0681 | R2 0.7860  Mae=0.0891 | R2 0.7891  Mae=0.0872 | Svr |
| TCARBN\_OXYGEN | R2 0.8390  Mae=0.0683 | R2 0.7784  Mae=0.0918 | R2 0.7826  Mae=0.0888 | Svr |
| SILCAT\_OXYGEN | R2 0.8415  Mae=0.0691 | R2 0.7613  Mae=0.0952 | R2 0.7662  Mae=0.0918 | Svr |
| PHSPHT\_OXYGEN | R2 0.8329  Mae=0.0711 | R2 0.7848  Mae=0.0909 | R2 0.7851  Mae=0.0892 | Svr |
| SILCAT\_TCARBN | R2 0.8217  Mae=0.0719 | R2 0.7638  Mae=0.0928 | R2 0.7704  Mae=0.0903 | Svr |
| PHSPHT\_TCARBN | R2 0.8285  Mae=0.0707 | R2 0.7771  Mae=0.0908 | R2 0.7788  Mae=0.0891 | Svr |
| PHSPHT\_SILCAT | R2 0.8299  Mae=0.0713 | R2 0.7771  Mae=0.0914 | R2 0.7812  Mae=0.0887 | Svr |
| 1 variable | TMP\_ | R2 0.6451  Mae=0.0988 | R2 0.5502  Mae=0.1164 | R2 0.5470  Mae=0.1167 | Svr |
| OXYGEN\_ | R2 0.6283  Mae=0.1067 | R2 0.5873  Mae=0.0984 | R2 0.5859  Mae=0.1259 | Svr |
| TCARBN\_ | R2 0.7259  Mae=0.0896 | R2 0.6940  Mae=0.1259 | R2 0.6924  Mae=0.0986 | Svr |
| SILCAT\_ | R2 0.6791  Mae=0.0988 | R2 0.6086  Mae=0.1089 | R2 0.6078  Mae=0.1093 | Svr |
| PHSPHT\_ | R2 0.8160  Mae=0.0742 | R2 0.7618  Mae=0.0928 | R2 0.7611  Mae=0.0928 | Svr |

**Table 2.**Résultats d’apprentissage automatique

## Résultats obtenus par les modeles d’apprentissages profond

### Experience pour choisir le nombre des couches et des neurones de chaque couche

Pour trouver le meilleur modèle ANN, on a tenté une expérience où on commence avec une seule couche et une boucle qui ajoute des neurones, puis on choisi le nombre de neurones qui nous a donné le meilleur score, puis on fixe la couche avec le nombre de neurones et on passe à la couche suivante et choisi le nombre de neurones qui a donné le meilleur score et ainsi de suite jusqu'à ce que le score commence à diminuer.



**Fig. 1.** Score obtenu avec chaque nombre de couche

Comme vous pouvez le voir dans le graphique, le meilleur score est 0,8545. Le réseau neuronal qui a donné ce score avait neuf couches :

1ère couche : 12 neurones. 2ème couche : 10 neurones.

3ème couche : 28 neurones. 4ème couche : 21 neurones.

5ème couche : 21 neurones. 6ème couche : 18 neurones.

7ème couche : 18 neurones. 8ème couche : 19 neurones.

9ème couche : 25 neurones..

Dans cette partie on a ajouté d’autres caractéristique à nos données , ces caractérestiques sont : Longitude, Latitude et la profondeur ou les mesures sont pris.

Pour bien comprendre on a entrainé les modèles avec les trois précidente variables:( Longitude, Latitude et la profendeur) plus toutes les variables ou avec une combinaison de quatre, trois,deux ou une variable;

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre de variable** | **Combinaison a Ajouter** | **R2** | **Mae** |
| **5 variables** | 5 variables | 0.8179 | 0.0756 |
| **4 variables** | SILCAT\_TCARBN\_OXYGEN\_TMP\_ | 0.7929 | 0.0900 |
| PHSPHT\_TCARBN\_OXYGEN\_TMP\_ | 0.7763 | 0.0954 |
| PHSPHT\_SILCAT\_OXYGEN\_TMP\_ | 0.7944 | 0.0741 |
| PHSPHT\_SILCAT\_TCARBN\_TMP\_ | 0.8500 | 0.0724 |
| PHSPHT\_SILCAT\_TCARBN\_OXYGEN\_ | 0.8055 | 0.0760 |
| **3 variables** | TCARBN\_OXYGEN\_TMP\_ | 0.6385 | 0.1119 |
| SILCAT\_OXYGEN\_TMP\_ | 0.8205 | 0.0823 |
| PHSPHT\_OXYGEN\_TMP\_ | 0.7626 | 0.0999 |
| SILCAT\_TCARBN\_TMP\_ | 0.8436 | 0.0761 |
| PHSPHT\_TCARBN\_TMP\_ | 0.6135 | 0.1086 |
| PHSPHT\_SILCAT\_TMP | 0.8172 | 0.0807 |
| SILCAT\_TCARBN\_OXYGEN\_ | 0.8117 | 0.0702 |
| PHSPHT\_TCARBN\_OXYGEN\_ | 0.6937 | 0.0968 |
| PHSPHT\_SILCAT\_OXYGEN\_ | 0.8374 | 0.0808 |
| PHSPHT\_SILCAT\_TCARBN\_ | 0.8353 | 0.0800 |
| **2 variables** | OXYGEN\_TMP\_ | 0.8271 | 0.0758 |
| TCARBN\_TMP\_ | 0.8330 | 0.0768 |
| SILCAT\_TMP\_ | 0.6972 | 0.0956 |
| PHSPHT\_TMP\_ | 0.8302 | 0.0811 |
| TCARBN\_OXYGEN\_ | 0.7605 | 0.0841 |
| SILCAT\_OXYGEN\_ | 0.8435 | 0.0770 |
| PHSPHT\_OXYGEN\_ | 0.8424 | 0.0774 |
| SILCAT\_TCARBN\_ | 0.7785 | 0.0825 |
| PHSPHT\_TCARBN\_ | 0.7921 | 0.0770 |
| PHSPHT\_SILCAT\_ | 0.7996 | 0.0930 |
| **1 variables** | TMP\_ | 0.7030 | 0.1138 |
| OXYGEN\_ | 0.8194 | 0.0811 |
| TCARBN\_ | 0.7935 | 0.0840 |
| SILCAT\_ | 0.8247 | 0.0736 |
| PHSPHT\_ | 0.8090 | 0.0746 |

**Table 4.** Résultats de réseaux de neurones

# Discussion

## Discussion des résultats obtenus par les modéle d’apprentissage automatique

Lorsque on utilize des modéles d’apprentissage automatique, on observe que le SVR donne un bon score lorsqu’il est entrainé avec la combinaison de quatre variables suivants(Silicate, Totale Carbonique, Oxygène, Température) .

**Fig. 3.**Compariason des meilleurs score obtenus.

## Discussion des résultat obtenus par le réseaux de neurones artificiels

Lorsque on utilise les résaux de neurones, on observe que la meilleure combinaison se compose des variables suivants: Longitude, Latitude, Depth, Silicat, Tcarbon, Phosphate, Température.

**Fig. 3.**Compariason des meilleurs score obtenus.

# Simulation

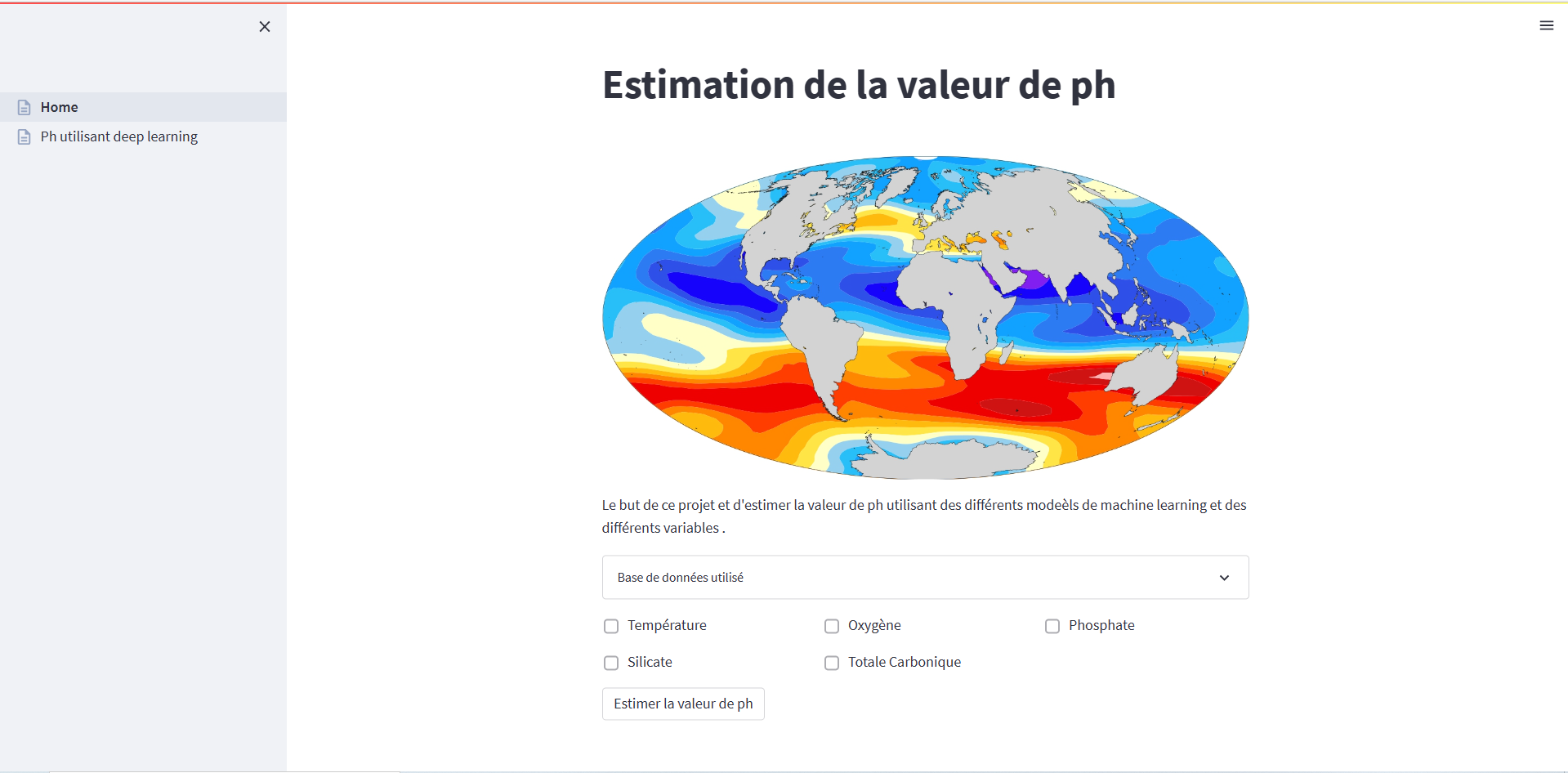
## Outils

Jupyter est une**évolution du projet IPython**. Jupyter permet de réaliser des calepins ou notebooks, c'est-à-dire des programmes contenant à la fois du texte en markdown et du code. Ces calepins sont utilisés en science des données pour explorer et analyser des données.

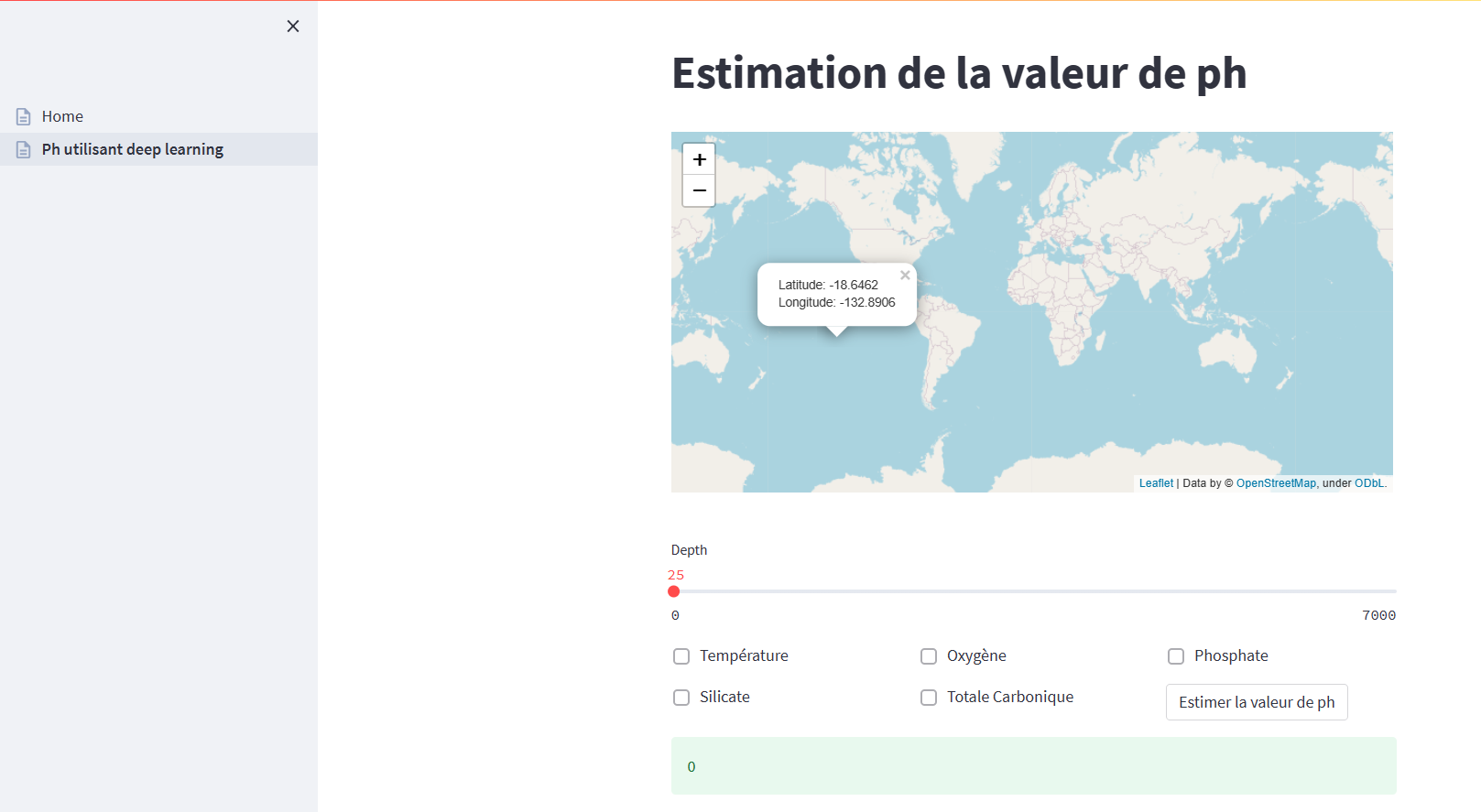


Streamlit est un outil open-source pour créer des applications web interactives en utilisant simplement Python. Il permet aux développeurs de créer rapidement et facilement des applications de visualisation de données, de démonstration de modèles d'IA et d'autres types d'applications web interactives en utilisant uniquement leur code Python existant. Il vise à rendre l'expérience de développement d'applications web en Python aussi simple que possible.

## Interfaces



Cette interface permet à l’utilisateur de faire des predictions de ph avec n’importe quelle combinaison, cette interface estime la valeur Ph utilisant les modèles d’apprentissage automatique .



Cette interface permet à l’utilisateur de faire des predictions de ph avec n’importe qelle combinaison, cette interface estime la valeur Ph utilisant les modèles de d‘apprentissage profond .



Cette Interface permet à l’utilisateur de calculer les valeurs de ph de plusieurs mesures et affiche ces valeurs sur une chart , il suffit d’importer le ficher de format csv , et choisir la combinaison avec lesquel il veut faire les predictions. Il lui permet aussi de telecharger le fichier le fichier csv labilisé.

**Conclusion et perspectives**

L’objectif de ce projet et de développer des modéles d’apprentissages automatiques (XGBoost, SVR, RF) et d’apprentissages profond qui sont capable d’estimer la valeur de PH avec moins de variables d’entré .

Les résultats ont montré que des bons performances ont été obtenues lorsque on a utilisé l’apprentissage automatiqe,Aussi des meilleures performances ont obtenus lors de l’utilisation d’apprentissage profond.

Ce projet était une occasion pour mettre en évidence et déployer sur le plan pratique mes connaissances en apprentissages automatiques et apprentissages profond et d’avoir en réelle l’un de leurs pratique dans le domaine de la biologie. Néanmoins, ce projet pourra être amélioré par l’utilisation d’autre modèles comme le CNN et LSTM ainsi que l’algorithme génétique pour prédire la valeur d’acidification océanique.