Projet Intégrateur CMI - 2021/2022

Étude de modèles et construction d'une interface R-shiny sur des données COVID-19.

Pierre LAGUE, Diamondra RAKOTONDRAZAKA, Jean ROBIN, Damien TANNEAU, Titouan QUINTIN, Victor HAGUET, Ewen DENIAU

8 juin 2022

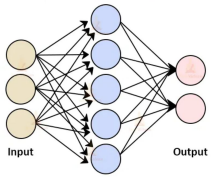
Sommaire

- Rédaction de l’état de l’art
- Préparation des données
- Développement des modèles
- Conception et implémentation de l’application web R Shiny
- Conclusion sur le projet

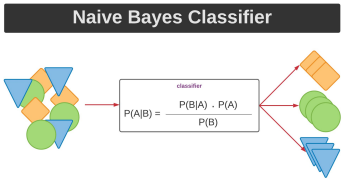
Rédaction de l'état de l'art

- Constitue une base de travail
- Permet une connaissance plus précise du modèle étudié/implémenté
- Mise à jour des pratiques et des standards du monde professionnel

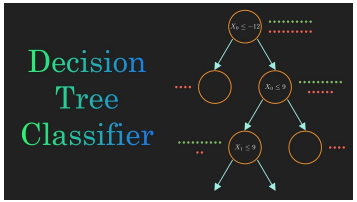
BUT -> Travailler sur des modèles/algorithmes standards et se les approprier.



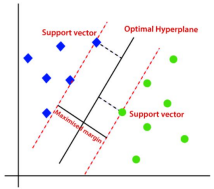
Réseau de neurones



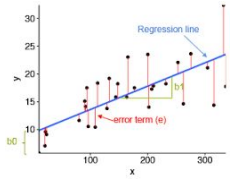
Classifieur Naïf Bayes



Arbre de décision



Support Vecteur Machine



Régression Linéaire et ses dérivées

Régression pénalisée

Pénalités

Ridge $\lambda \times (variable_1^2 + ... + variable_x^2)$

Lasso $\lambda \times (|variable_1| + ... + |variable_x|)$

Elastic-Net $\lambda_1 \times (variable_1^2 + ... + variable_x^2) + \lambda_2 \times (|variable_1| + ... + |variable_x|)$

Préparation des données - Description

Données Covid-19

COLUMN COUNT	FILE COUNT	SIZE
67	1	45.28 MB
2022-06-06 09:48	2022-06-06 09:48	2022-06-06 09:48

RECORD COUNT
166 554
2022-06-06 09:48

Description du dataset

- Données manquantes
- Colonnes inutiles

iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths
string	string	string	string	string	string	string	string
Country	Text	Country	Date (unparsed)	Decimal	Decimal	Decimal	Decimal
AFG	Asia	Afghanistan	2020-02-24	5.0	5.0		
AFG	Asia	Afghanistan	2020-02-25	5.0	0.0		
AFG	Asia	Afghanistan	2020-02-26	5.0	0.0		
AFG	Asia	Afghanistan	2020-02-27	5.0	0.0		
AFG	Asia	Afghanistan	2020-02-28	5.0	0.0		
AFG	Asia	Afghanistan	2020-02-29	5.0	0.0		
AFG	Asia	Afghanistan	2020-03-01	5.0	0.0	0.714	
AFG	Asia	Afghanistan	2020-03-02	5.0	0.0	0.0	
AFG	Asia	Afghanistan	2020-03-03	5.0	0.0	0.0	
AFG	Asia	Afghanistan	2020-03-04	5.0	0.0	0.0	
AFG	Asia	Afghanistan	2020-03-05	5.0	0.0	0.0	
AFG	Asia	Afghanistan	2020-03-06	5.0	0.0	0.0	
AFG	Asia	Afghanistan	2020-03-07	8.0	3.0	0.429	
AFG	Asia	Afghanistan	2020-03-08	8.0	0.0	0.429	
AFG	Asia	Afghanistan	2020-03-09	8.0	0.0	0.429	
AFG	Asia	Afghanistan	2020-03-10	8.0	0.0	0.429	
AFG	Asia	Afghanistan	2020-03-11	11.0	3.0	0.857	
AFG	Asia	Afghanistan	2020-03-12	11.0	0.0	0.857	
AFG	Asia	Afghanistan	2020-03-13	11.0	0.0	0.857	
AFG	Asia	Afghanistan	2020-03-14	14.0	3.0	0.857	
AFG	Asia	Afghanistan	2020-03-15	20.0	6.0	1.714	
AFG	Asia	Afghanistan	2020-03-16	25.0	5.0	2.429	
AFG	Asia	Afghanistan	2020-03-17	26.0	1.0	2.571	
AFG	Asia	Afghanistan	2020-03-18	26.0	0.0	2.143	
AFG	Asia	Afghanistan	2020-03-19	26.0	0.0	2.143	

Échantillon du dataset

Préparation des données - Engineering

Données Covid-19

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Conversion des données en types numériques
- Remplissage des valeurs manquantes (si NA < 5% insérer la moyenne des données).
- Parsing de la date

COLUMN COUNT	FILE COUNT	SIZE	RECORD COUNT
28	1	8.79 MB	134 996

Détails du dataset final

Développement des modèles - Implémentation

Langage de programmation adapté à la science des données disposant d'un grand nombre de librairies

Libraire de Python alliant puissance de calcul, vectorisation et opérations mathématiques.

Librairie de Python d'analyse et de manipulation de données. Très utile pour un volume important de données

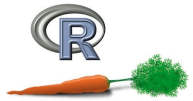
Librairie Python proposant une variété de modèles de ML (régression Lasso jusqu'à ANN). Tous les modèles seront issus de cette librairie.

Tous les modèles + comparaison

R est un langage de programmation pour le calcul statistique et les graphiques.

Dplyr est principalement un ensemble de fonctions conçues pour la manipulation de dataframes de manière intuitive.

Shiny est un paquet R qui permet de créer facilement des applications web interactives directement à partir de R.



Le paquet caret est un ensemble de fonctions qui tentent de rationaliser le processus de création de modèles prédictifs.

Application R-Shiny et quelques modèles

Développement des modèles - Entraînements

```
k-Nearest Neighbors

101247 samples
     7 predictor

Pre-processing: centered (7), scaled (7)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 101247, 101247, 101247, 101247, 101247, 101247, ...
Resampling results across tuning parameters:

  k  RMSE        Rsquared   MAE
  5  0.01461566  0.9436960  0.00229603
  7  0.01530359  0.9383102  0.00261205
  9  0.01596484  0.9328988  0.00288967

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 5.
```

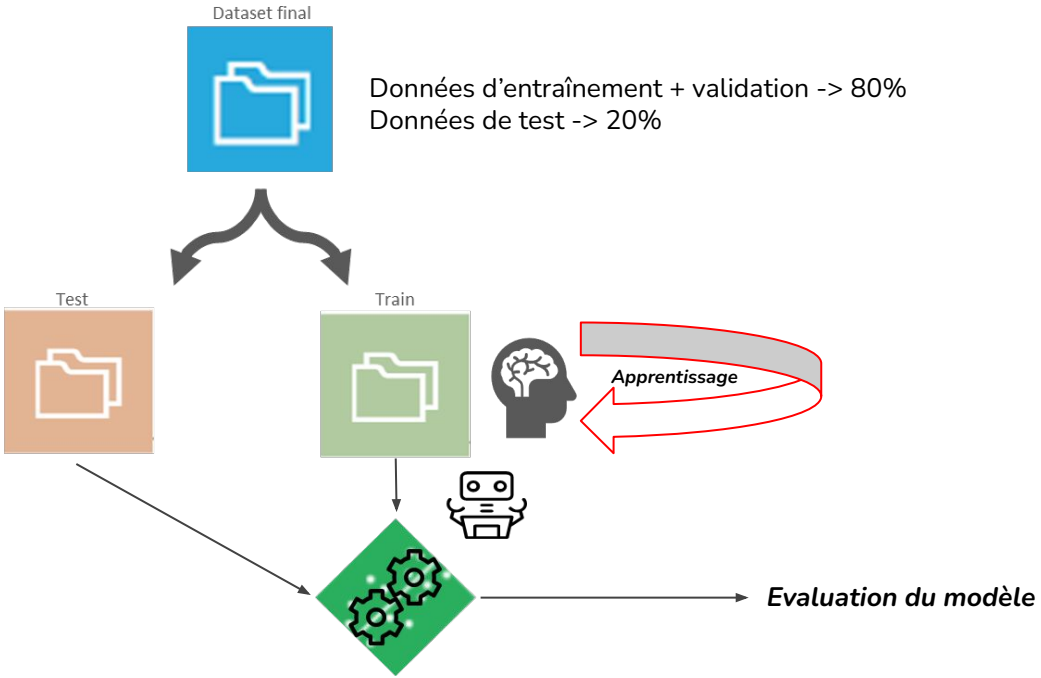
Résultats de l'optimisation de l'algorithme KNN sur les données du Covid-19

Développement des modèles - Entraînements

- Chaque modèle est entraîné avec ses hyperparamètres par défaut
- Chaque modèle est entraîné sur la même base de données train (même seed)
- Chaque modèle est entraîné sur “*total_deaths*” puis sur “*total_cases*”

Exemple de file d’entraînement sous Datalku - R^2

Développement des modèles - Evaluation

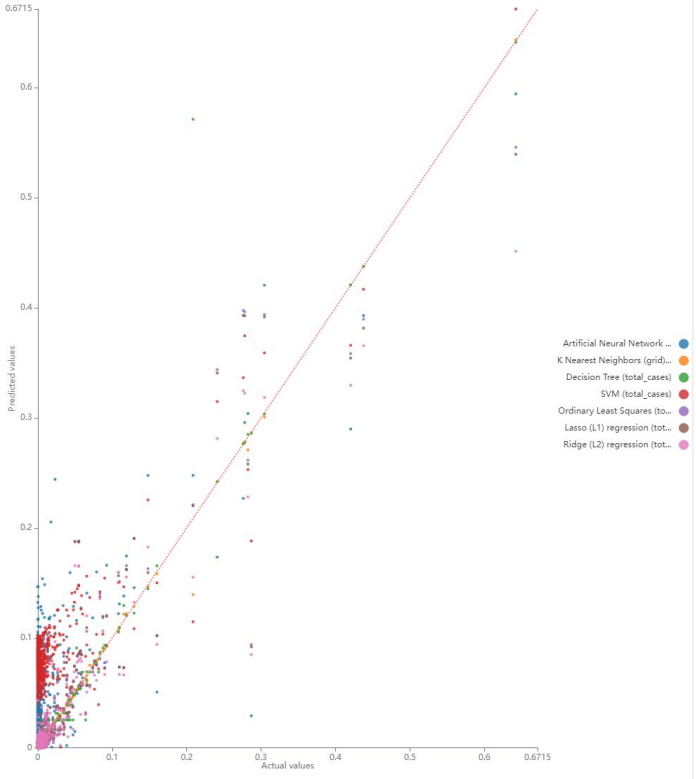


Evaluation du modèle :

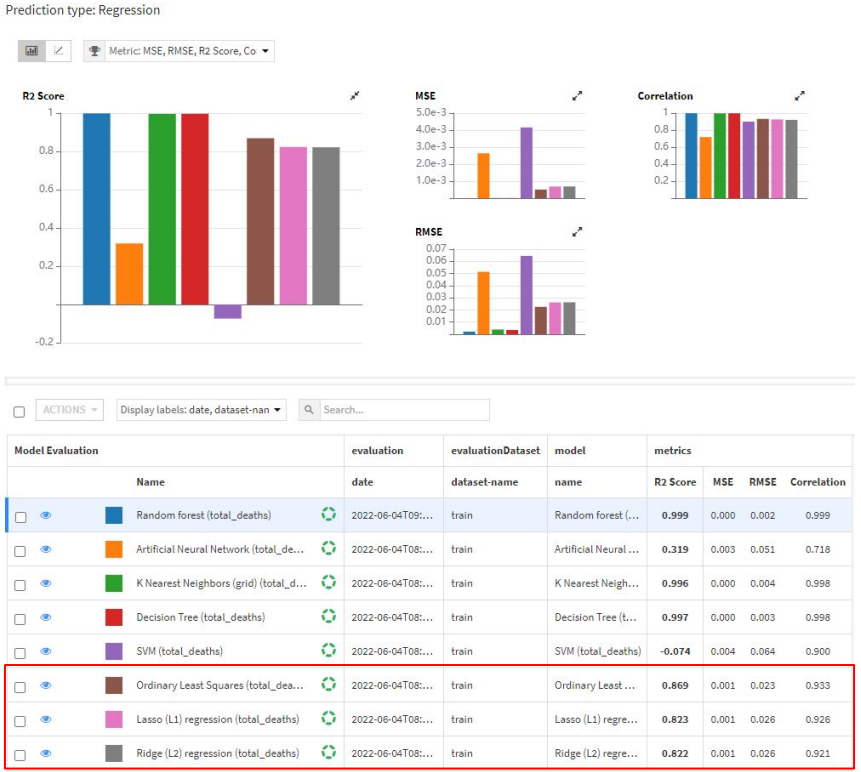
- R^2
- MSE
- RMSE
- Correlation

\+ Allure de la courbe "*actual vs predicted*"

Développement des modèles - Evaluation

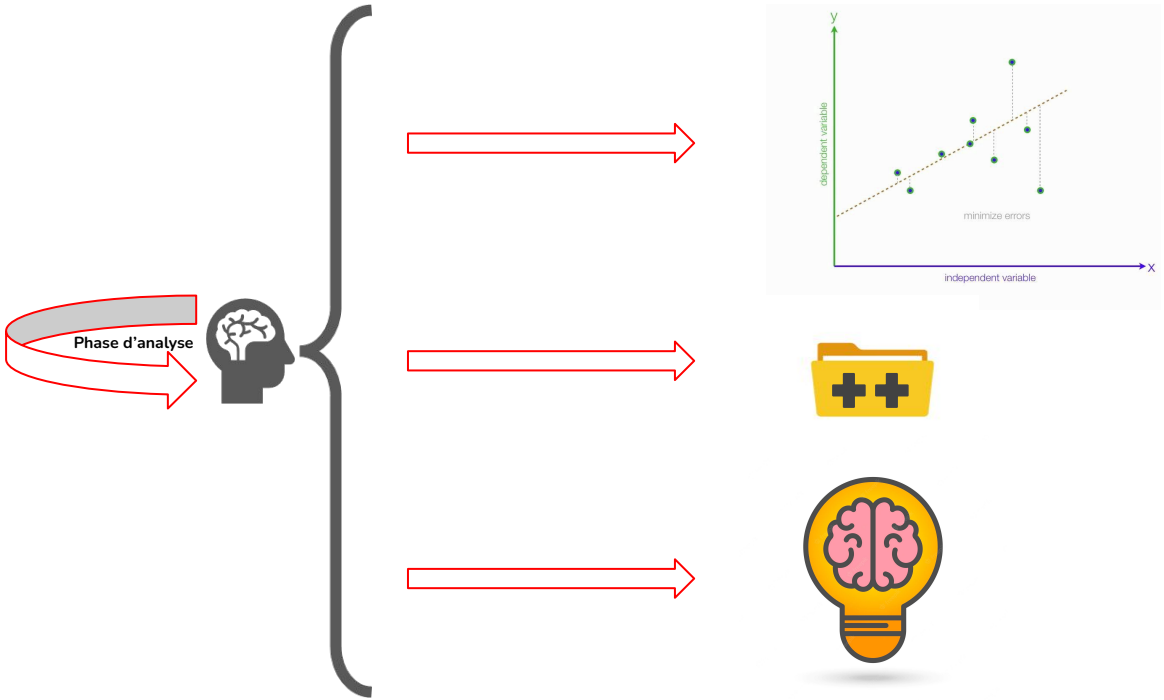


Scatter Plot des test des modèles



Mesures MSE, RMSE, R² et Correlation des chaque modèles

Développement des modèles - Conclusion

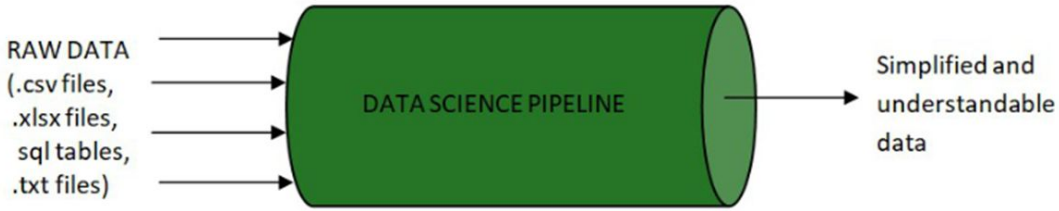


Les modèles linéaires sont plus adaptés à notre problème de régression

Il nous faut beaucoup de données préparées et sans valeurs manquantes

Les variables explicatives doivent être choisies de manière à expliquer la variable cible sur plusieurs dimensions (temps, espace)

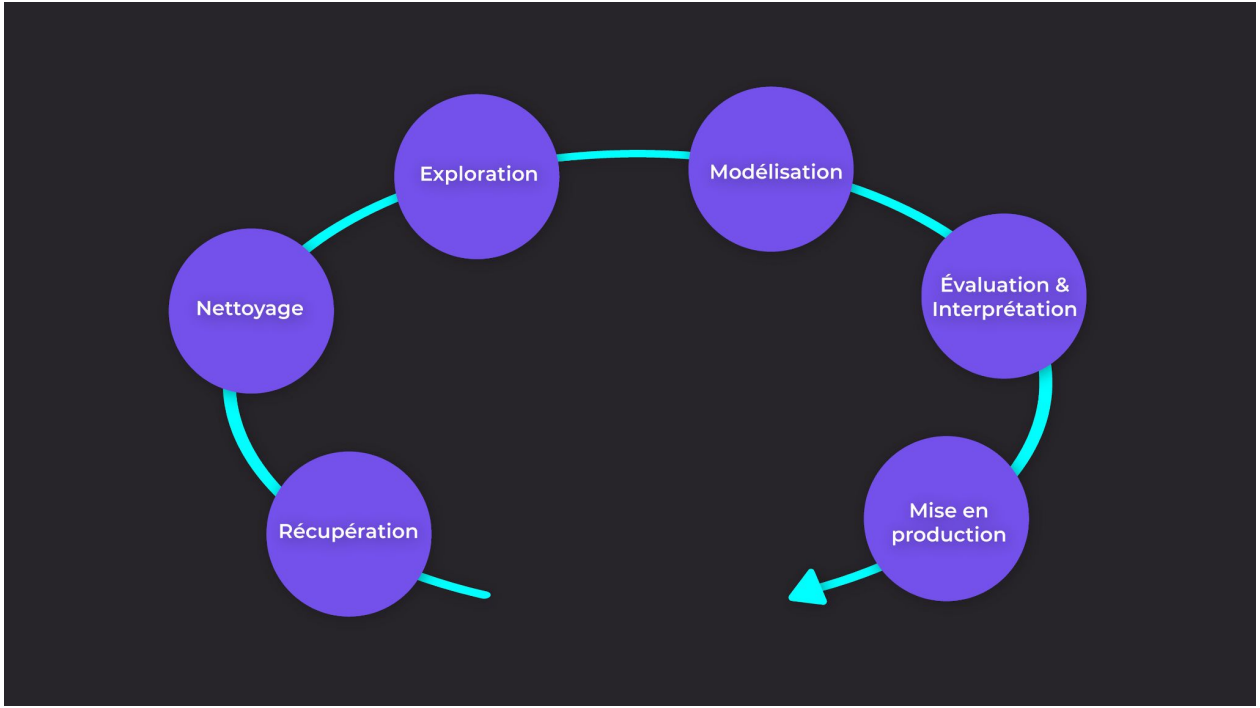
Conception et implémentation de l’application web R Shiny



Représentation simplifiée d'un data pipeline

Source : https://www.geeksforgeeks.org/whats-data-science-pipeline/

- Obtention des données
- Nettoyage des données
- Analyse exploratoire des données
- Modélisation
- Interprétation



Etapes d’un projet de machine learning

Source : https://openclassrooms.com/fr

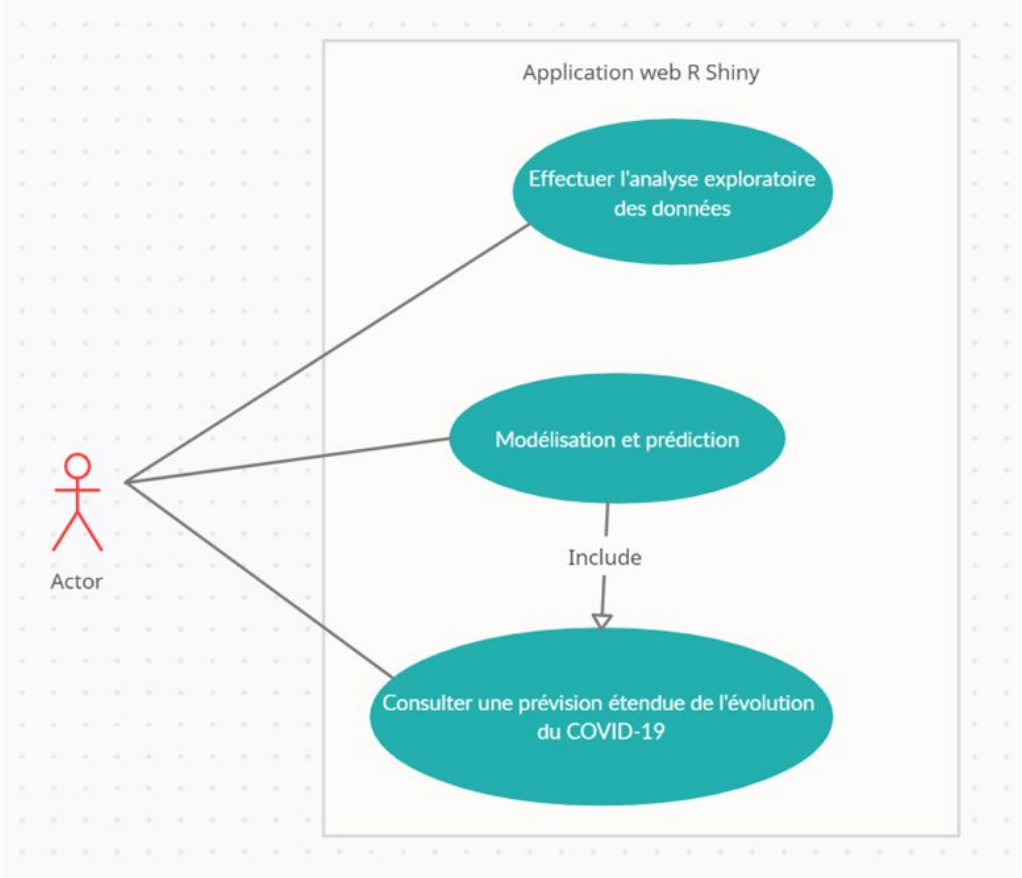
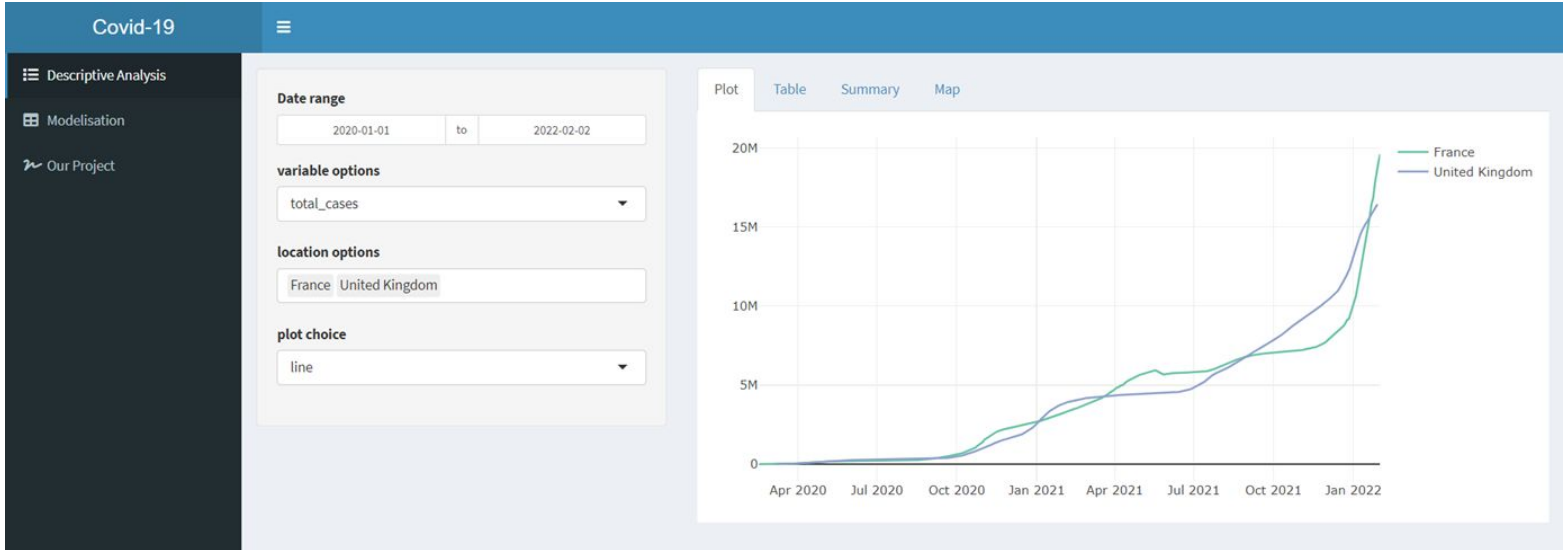
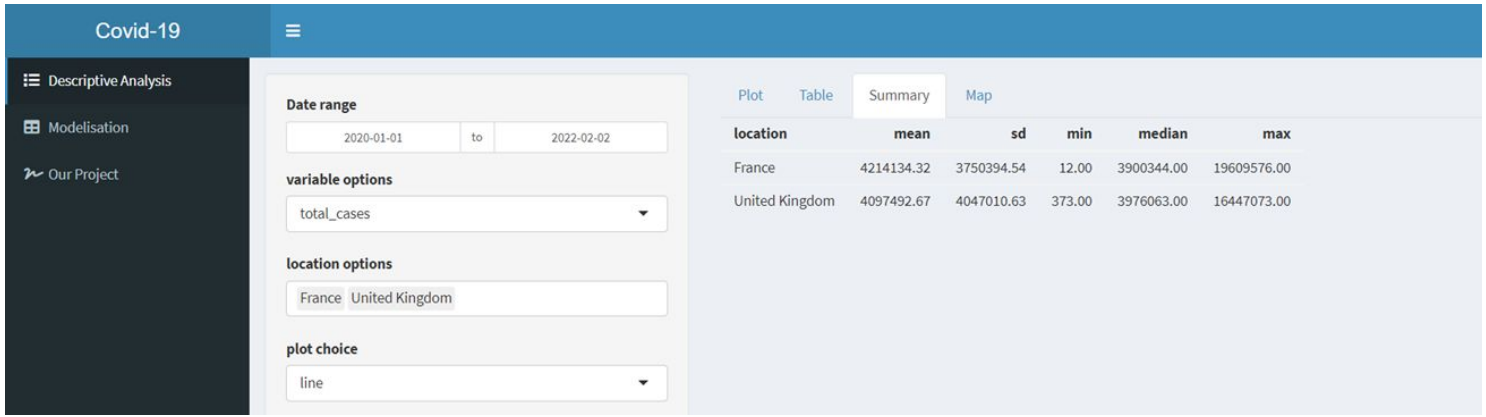


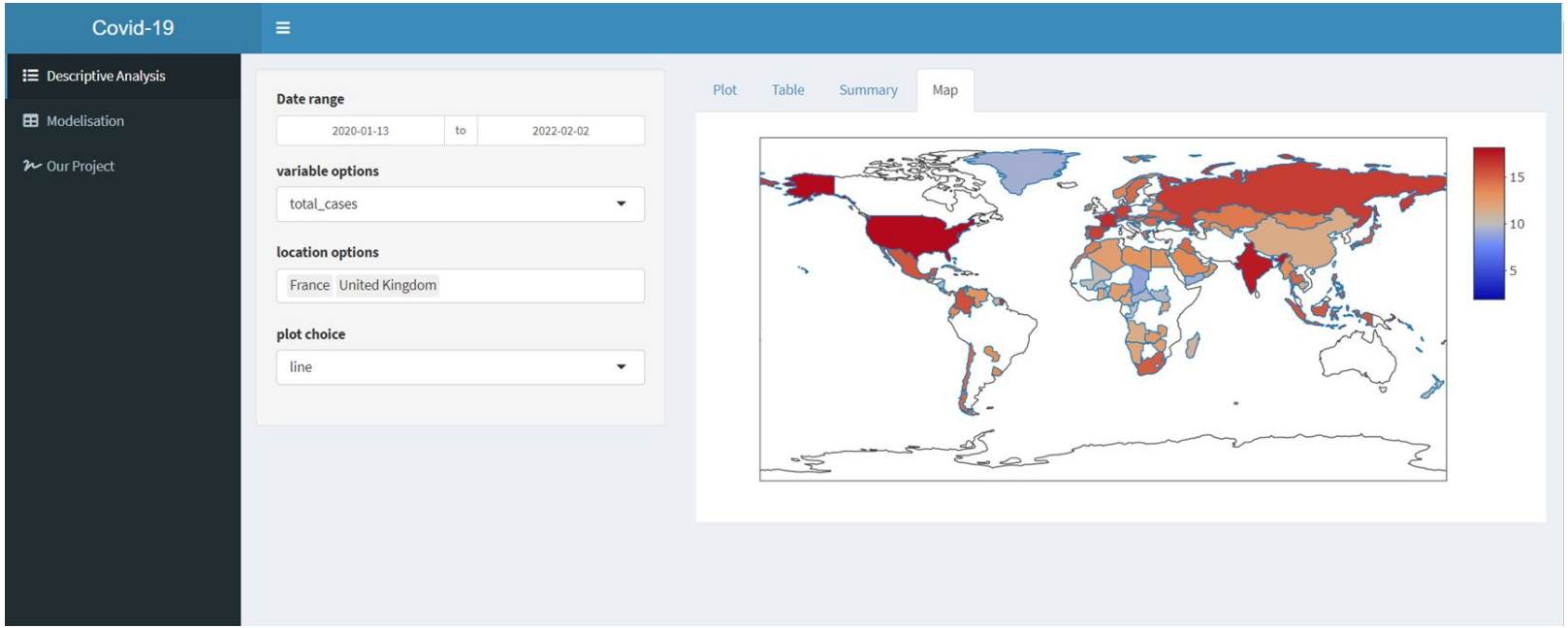
Diagramme de cas d’utilisation de l’application

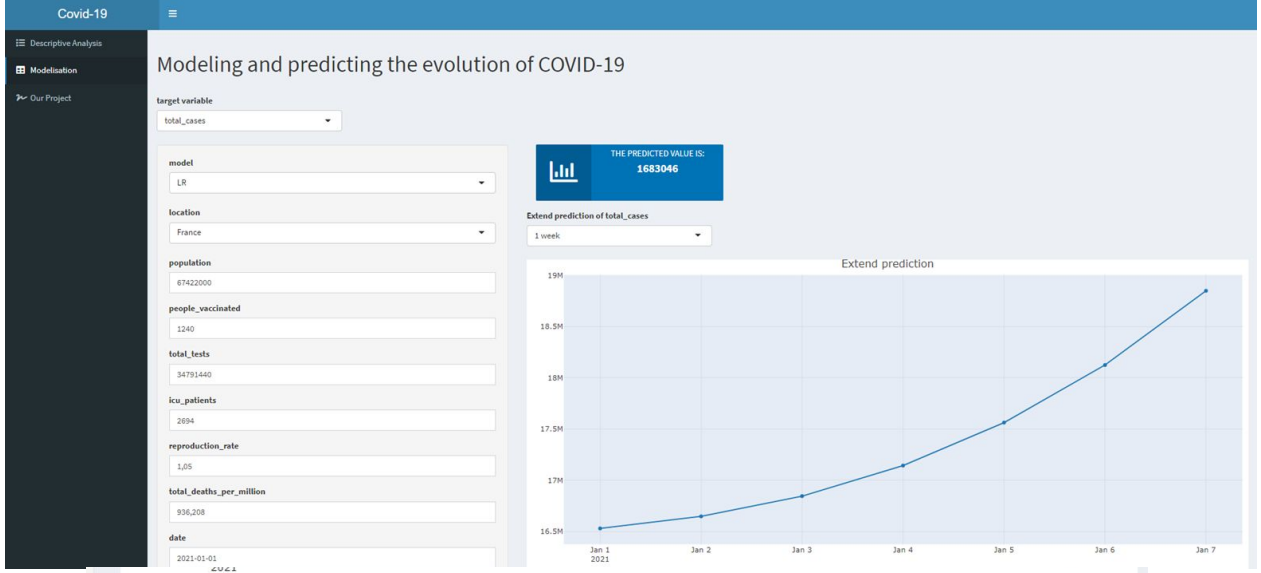
Application Shiny

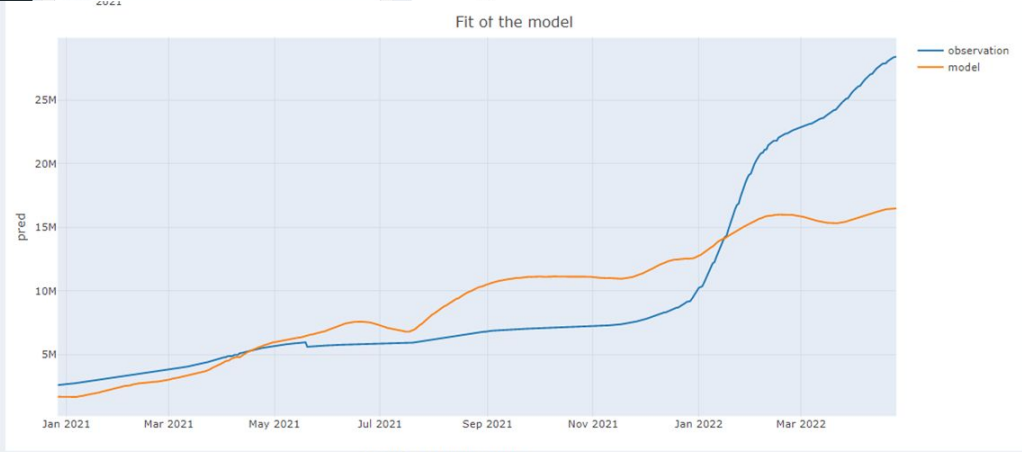
https://q1ov6nn1.shinyapps.io/app-covid19/


Covid-19
Descriptive Analysis
Modelisation
Our Project
Date range
2020-01-01
to
2022-02-02
variable options
total_cases
location options
France
United Kingdom
plot choice
line
Plot
Table
Summary
Map
20M
15M
10M
5M
0
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
France
United Kingdom


Covid-19
Descriptive Analysis
Modelisation
Our Project
Date range
2020-01-01
to
2022-02-02
variable options
total_cases
location options
France
United Kingdom
plot choice
line
Plot
Table
Summary
Map
location
mean
sd
min
median
max
France
4214134.32
3750394.54
12.00
3900344.00
19609576.00
United Kingdom
4097492.67
4047010.63
373.00
3976063.00
16447073.00


Covid-19
Descriptive Analysis
Modelisation
Our Project
Date range
2020-01-13
to
2022-02-02
variable options
total_cases
location options
France
United Kingdom
plot choice
line
Plot
Table
Summary
Map
15
10
5





LR RMSE : 2867498.94512285

Merci de votre attention