Rapport CMI

Pierre Lague

March 2023

Le but du projet intégrateur CMI en sciences des données était d'utiliser le *machine learning* pour l'étude de l'évolution du COVID-19 à l'échelle mondiale. En Intelligence Artificielle, l'apprentissage supervisé permet de faire d'établir des modèles permettant de prédire les probabilités d'appartenance d'une observation à une classe (classification), ou bien une valeur quantitative liée (régression). L'apprentissage non-supervisé quant à lui pour objectif de trouver une structure pertinente pour un ensemble de données.
Ces technologies sont utilisées dans divers domaines, grâce aux réponses qu'elles apportent à leurs problématiques types associés.

En vue de créer des modèles afin de prédire l'évolution de l'épidémie de COVID-19. L'état de l'art porte ici sur les algorithmes d'apprentissage supervisé les plus utilisés à ces fins.
Les résultats obtenus par nos modèles à la prédiction du nombre total de cas d'infectés et du taux de mortalité par pays à une date donnée, ont été dans l'ensemble non satisfaisants. Cela est due principalement à la difficulté élevée de généraliser des modèles pour de telles variables cibles et le manque de données à disposition pour l'entraînement des modèles.

L'application web R Shiny qui a été implémentée, est l'ébauche d'un service très pratique pour analyser et comparer les données de l'épidémie de COVID-19 pour des pays choisis à une plage temporelle considérée.
De par le gain de temps non négligeable du simple lancement de l'application, et la flexibilité qu'elle offre dans la sélection des valeurs d'entrées. Ce service saura bien répondre à des exigences en termes de confort, dans l'aspect pratique de son utilisation.

0.1 Perspectives

Toujours dans l'intention d'améliorer les performances des modèles de *machine learning*. Une perspective serait de trouver de nouvelles sources de données afin de combler les valeurs manquantes fortement présentes, pour des dates sur la base d'une fréquence journalière.

En plus de l'augmentation du nombre d'observations, dont les algorithmes de *deep learning* bénéficient grandement. Avec des données provenant de nouvelles sources, on pourra améliorer les observations déjà présentent à des dates

données.
En effet, rien ne garantit la fiabilité d'un montant donné par exemple du nombre de patients dans un hôpital. Afin d'améliorer la précision des observations, il serait totalement naturel d'utiliser un agrégat comme la moyenne.

1 Introduction

Suite à l'apparition en Chine de la maladie infectieuse à coronavirus fin 2019, les vies des habitants d'une grande partie des pays du monde ont été impactées. Cette crise humanitaire s'est ensuite peu à peu inscrite sur la durée, c'est d'ailleurs le 11 mars 2020 que l'Organisation Mondiale de la Santé déclare l'épidémie de coronavirus (COVID-19) comme pandémie. En plus des moyens physiques comme le déploiement d'espaces permettant d'accueillir des malades du COVID-19, les tests et la vaccination pour lutter contre cette maladie. La collecte quotidienne et l'analyse des nouvelles données produites liées à cette thématique sont aussi d'une importance primordiale.

Ces données peuvent par exemple permettre d'accompagner les dirigeants des Etats dans leurs prises de décisions quant à de nouvelles mesures pour limiter le risque de propagation du virus. Le thème de ce travail porte sur l'utilisation du Machine Learning ou apprentissage automa- tique, pour la modélisation et la prédiction de l'évolution de l'épidémie COVID-19 au niveau mondial.

L'apprentissage automatique est une application de l'IA, désignant un ensemble d'algorithmes capables de s'améliorer eux-mêmes grâce à des données fournies. L'intérêt majeur de l'utilisation du Machine Learning dans notre cas de figure, est de pouvoir extraire de l'information aux données liées au COVID-19 suite à l'établissement de modèles prédictifs. Afin de connaître l'évolution future du COVID-19 et de piloter de façon encore plus sûre les choix à faire dans la création ou modification de mesures sanitaires.

2 Rédaction de l'état de l'art

Avant toute étude il est important de se fixer un terrain de travail balisé par des normes et des technologies d'actualités. C'est pourquoi nous avons rédigé un Etat de l'Art en Machine Learning appliqué à l'analyse statistique pour nous aider dans notre projet. L'état de l'art est l'état des connaissances dans tout domaine donné (scientifique, technique, artistique, médical, etc.) à un instant donné.

Notre état de l'art reprend alors une grande partie des modèles de machine learning utilisés à ce jour dans le milieu professionnel allant de la simple régression linéaire aux réseaux de neurones complexes.

La rédaction de l'état de l'art s'est axée sur 2 points :

- l'expérience avec le modèle / algorithme.
- leur pertinence théorique sur un problème de régression.

Les élèves de L3 se sont occupés d'analyser et d'expliquer les modèles de machine learning les plus complexes (ANN, SVM etc.) tandis que les élèves de L1 ont pu étudier des modèles plus abordables à leur niveau (la régression linéaire, KNN). Le but étant de proposer une synthèse du fonctionnement du modèle en visant un public large (professeur de statistique comme professeur d'informatique) pour que le livrable soit utilisable par le plus de profils possible. Pour chaque modèle, nous expliquons son fonctionnement en décrivant les mathématiques sous-jacentes et nous présentons un exemple d'utilisation dans la vie professionnelle.

Ainsi, en établissant un état de l'art, nous avons non seulement une base de travail d'actualité, mais cela à permis au groupe d'en savoir plus sur des modèles jusque là non étudiés.

3 Préparation des données - Description

Les données utilisées ont été récupérées sur le github mis à disposition par le corps enseignant en début d'année. Ce sont des données relativement complexes car elles expliquent un grand nombre de choses. Beaucoup de colonnes n'ont pas été utilisées et nous avons cherché un moyen de pallier au problème des valeurs manquantes.

Après une analyse préliminaire nous avons décidé de garder deux variables cibles pour l'étude : total_cases et total_deaths.

4 Construction des modèles sous Python

Pour des raisons plus techniques que pratiques, nous avons décidé de mener une analyse des modèles à étudier sous Python. Bien que cela ne réponde pas aux souhaits du jury, nous avons tenté de mener une analyse très complète en entraînant beaucoup de modèles et ceci pour nos 2 variables cibles (donc 2 analyses).

L'implémentation Après avoir arrangé les données et avoir fait en sorte qu'elles puissent être utilisées dans un modèle, nous avons implémenté plusieurs modèles de machine learning en python à l'aide de la librairie Scikit-Learn. Scikit-Learn est une librairie Python qui propose une variété de modèles ML simples à utiliser. Les modèles implémentés sont les suivants :

- Random Forest
- Artificial Neural Network
- K Nearest Neighbors
- Decision Tree
- Support Vector Machine
- Ordinary Least Squares

- Lasso Regression
- Ridge Regression

L'évaluation en utilisant la méthode de cross-validation (séparation des données en 2 datasets, test et train), nous avons entraîné chacun de ces modèles sur le dataset train, puis nous les avons évalués sur le dataset test. Chaque entraînement était plus ou moins long en fonction du modèle. Les variables explicatives sont :

- reproduction_rate
- strigency_index
- population
- median_age
- cardiovasc_death_rate
- diabetes_prevalence
- hospital_beds_per_thousand
- date (parsed)

On constate que 2 modèles ont des performances moins bonnes qu'un classificateur aléatoire (ANN et SVM). Ce type de modèle n'est donc pas adapté à notre problème. On remarque aussi que 3 modèles ont des scores R^2 supérieurs à 0.99. On suspecte ici un sur-apprentissage. Cependant, les 3 modèles de régression (OLS, Lasso et Ridge) proposent un R-squared assez bon ainsi qu'une bonne corrélation. Pour les modèles ANN et SVM, on voit que leur MSE est beaucoup plus élevée que les autres. Ceci peut être dû à la mauvaise paramétrisation des hyperparamètres. Pour la variable "total_cases", les résultats ne sont pas bien différents, on trouve que les modèles de régression (OLS, Lasso et Ridge) sont beaucoup plus adaptés que les autres.

En somme, les modèles les plus fiables sont les suivants : OLS, Lasso Regression, Ridge Regression.

5 Application web

Afin de rendre accessible le fruit de notre travail à toutes personnes ayant des connaissances ou non dans ce domaine. Une application web a été développée à cet effet. L'utilisateur obtiendra facilement les résultats d'analyses descriptives, de modélisation et de prédictions.

Dans les étapes de développement d'un projet de Machine Learning, l'étape finale de déploiement du modèle (ou mise en production) est sûrement la plus délaissée. Étapes de développement d'une projet de machine Learning. Mais elle est essentielle par exemple en entreprise, où en plus de restituer les résultats

d'une recherche, on voudra fournir un produit à un client ou pour l'entreprise. De cette manière, un modèle déployé et accessible via une application ou API, est utilisable de façon très pratique par toute personne disposant de l'accès.

5.1 Shiny

Pour développer notre application web, nous avons utiliser la librairie Shiny, du langage de programmation R. La librairie Shiny permet de créer des interface graphique composée de wigdets, permettant à l'utilisateur de fournir des valeurs d'entrées.

Le rendu des résultats fournis par l'application est réactif. Les codes R de l'application web fait développée avec Shiny dépendent des valeurs courantes d'entrées sélectionnées par l'utilisateur via les wigdets. En cas de modification des valeurs d'entrées, les codes R sont réexecutés et leurs sorties affichées sont mises à jour. Quand on parle de codes R liés à une application web Shiny, on fait référence à des codes liés à la structure de base d'une application construite grâce à la librairie Shiny. Cette structure est composée de 2 parties qui sont liées, la partie UI l'interface utilisateur et la partie serveur. La partie UI regroupe la partie visuelle de l'interface web avec lequel l'utilisateur interagit. La mise en forme des aspects visuels et les widgets grâce auxquels l'utilisateur sélectionne les entrées et l'affichage des sorties.

La partie serveur vise à produire les résultats affichés en sortie sur l'interface utilisateur. Et à les mettre à jour dynamiquement quand les valeurs d'entrées sont modifiées.

5.2 Composants de l'application web

L'application web a été conçue pour recueillir des connaissances sur les données mondiale de COVID-19 sur 2 principaux axes. L'analyse descriptive et la modélisation/prédiction d'une variable d'intérêt. Afin d'obtenir des informations simples et pertinentes, l'analyse descriptive est employée sur des données que l'on groupe par pays et suivant une plage temporelle définie par l'utilisateur. Ce choix se fait de façon naturelle, au vu du caractère géo spatial et temporel des données.

Dans la partie modélisation et prédiction, l'utilisateur aura le choix du modèle de Machine Learning pour modéliser et prédire l'évolution à l'échelle mondiale du COVID-19. Là aussi la prédiction d'une variable d'intérêt comme le nombre de cas de COVID-19 se fait pour un pays donné à une date choisie. En addition à cela, l'utilisateur pourra choisir les autres valeurs d'entrées à fournir au modèle afin de calculer la prédiction. Afin de savoir comment le modèle utilisé par l'utilisateur performe sur les données, un rendu graphique sur l'ajustement du modèle pour un pays sélectionné est fourni.

La métrique choisie pour évaluer tous les modèles est la racine carrée de l'erreur quadratique moyenne (RMSE) : Le RMSE est calculé sur l'ensemble des observations.

Une idée qui s'inspire des prévisions météo, qui sont aussi obtenues grâce à des données géo spatiales et temporelles. Est de montrer une prévision étendue de l'évolution de COVID-19. On se sert des données d'entrées qui sont directement associées aux observations pour faire de l'extrapolation. Dans le but d'avoir des valeurs d'observations inconnues à des dates futures. La prévision étendue de l'évolution de COVID-19 est donc possible grâce à ce procédé qui permet d'obtenir de nouvelles valeurs d'entrées.