Université Abdelhamid MEHRI - Constantine 2 -

Faculté des Nouvelles Technologies de l'Informatique et de la Communication

Département des Technologies des Logiciels et Systèmes d'information

Master 1 - Génie Logiciel

Module : SAD

Mini Projet

Projet de statistiques et d'analyse de données , COVID19 Global Forecasting (Week 4)

Réalisé par :

- Ounissi Fadoua
- Bouzellifa Nora
- Alioua Chaima

2022/2023

Introduction

Covid a touché de nombreux domaines mais il était si difficile de prévoir ou d'obtenir le nombre de tous les cas, les statistique était juste au niveau des hopitaux d'une manière assez simple et primitif, mais c'est devenu compliqué et ca prend beaucoup de temps à gérer et à analyser tout ces cas. Avec le machine learning on vas essayer de faciliter la méthode dont on vas l'utiliser pour avoir une Une étude complète et approfondieà propos tout les cas malgré le grand nombre. on vas essayer aussi de predecter le nombre des cas dans les jours et les mois à venir. On va utiliser dans cette étude une regression lineaire en appliquant plusieurs modèles et par la suite choisir le meilleur modèle pour prédire les valeurs voulu pour notre dataset.

I- L'exploration des données :

L'exploration des données est une méthode d'analyse des données qui vise à découvrir des informations cachées, des modèles et des relations entre les variables.
Objectif: C'est de comprendre du mieux possible nos données et de développer une première stratégie de modélisation.

1)Analyse de la forme des données :

Nombres de lignes et de colonnes : 6 colonnes et 35995 lignes.
Présentation des variables : :
object : ProvinceState, CountryRegion, Date.
Float : ConfirmedCasses, Fatalities.
Int : Id.
2)Analyse des valeurs manquantes :
a)Visualisatin initiale :

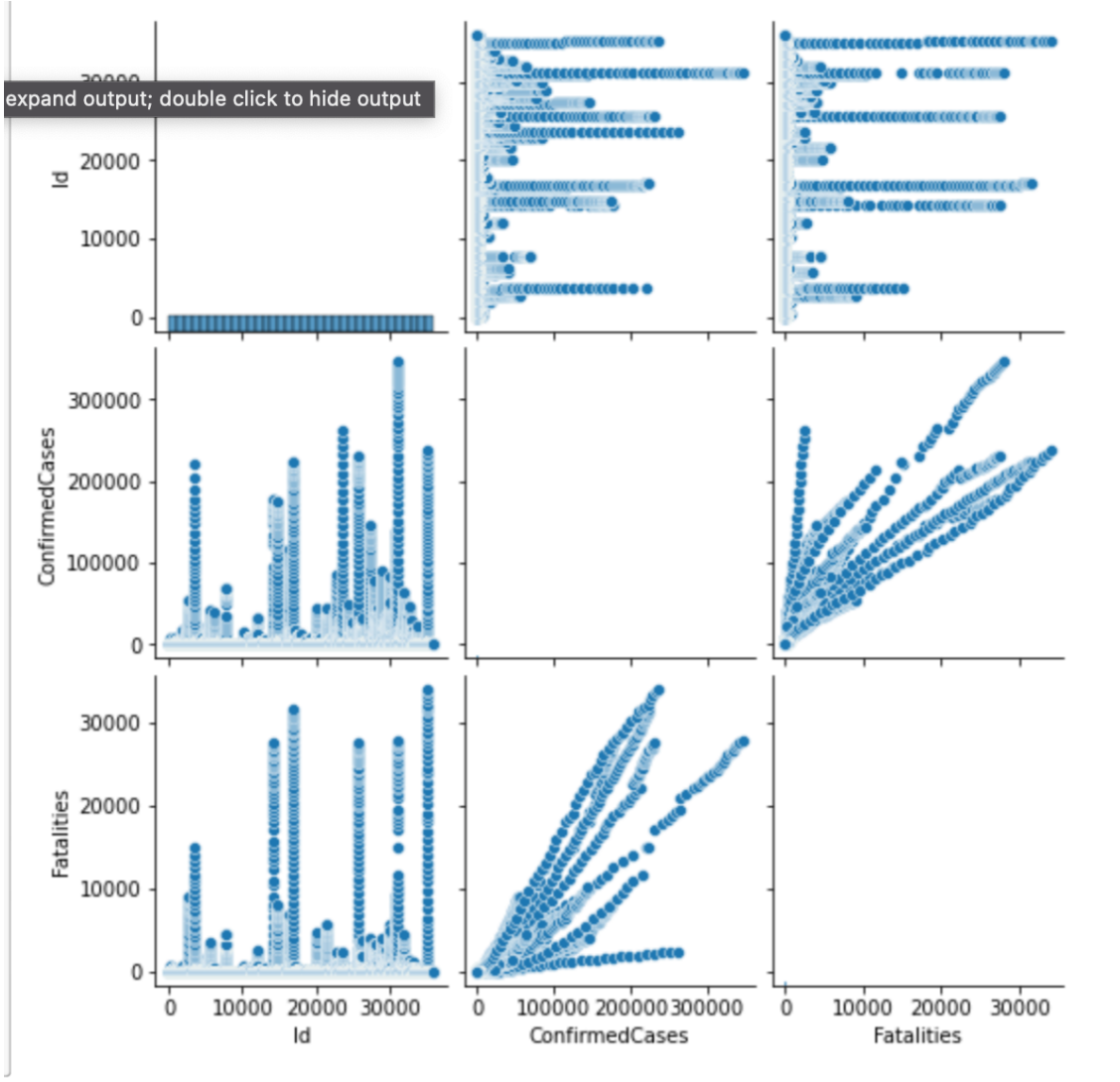
Elimination des colonnes inutiles :on a supprimé la colonne « Id ».
Visualisation de target :on a affiché le nombre d'apparence de chaque valeure de la target dans la dataset.
Visualisatin de la distribution des variables continues (« ConfirmedCases », « fatalities »):
D'après l'affichage des deux plots on remarque que : la distribution des vriables « ConfirmedCases » et « fatalities » est une distributin normale ce qui signifie que la plus part des valeurs se situent autour de la moyenne et que y'a pas des valeurs extremes et abérentes.
Le nombre total des fatalities 1879, et le nombre de « ConfirmedCases » est : 6000.
Présentation des variables qualitatives (« ProvinceState », « CountryRegion », « date ») :
Il y'a 133 catégories de « ProvinceState » et 184 catégories de « CountryRegion » et 115 catégories de la date.
-Le nombre d'apparition de chaque catégorie de la variable « date » est : 313 fois.
Ça veut dire que tous les dates sont apparus dans notre dataset avec le meme nombre donc y'a pas de problème dans ce feature.

Date	2020-01-22	2020-01-23	2020-01-24	2020-01-25	2020-01-26	2020-01-27	2020-01-28	2020-01-29	2020-01-30	2020-01-31	...	2020-05-06	202 05-(
Province_State													
Alabama	1	1	1	1	1	1	1	1	1	1	...	1	
Alaska	1	1	1	1	1	1	1	1	1	1	...	1	
Alberta	1	1	1	1	1	1	1	1	1	1	...	1	
Anguilla	1	1	1	1	1	1	1	1	1	1	...	1	
Anhui	1	1	1	1	1	1	1	1	1	1	...	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
Wyoming	1	1	1	1	1	1	1	1	1	1	...	1	
Xinjiang	1	1	1	1	1	1	1	1	1	1	...	1	
Yukon	1	1	1	1	1	1	1	1	1	1	...	1	
Yunnan	1	1	1	1	1	1	1	1	1	1	...	1	
Zhejiang	1	1	1	1	1	1	1	1	1	1	...	1	

Le nombre d'apparition de chaque catégorie de la variable « ProvinceState » est : 115 fois, qui représente le nombre des dates. Donc c'est très bien ça pose pas de problème.
-Le nombre d'appartion de toutes les catégories de la variable 'CountryRegion' est 115 fois appart (US , China, canada, France , United Kingdom, Australia, Netherlands, Denmark . ca veut dire que dans le dataset ces pays ont des lignes avec la même date (ca peut poser un problème).
-D'après la visualisation de la distribution des variable quantitatives on a remarqué que

la distribution des variables(ProvinceState et date) est une distribution normale, la distribution de la variables Countryregion n'est pas normale donc on doit netoyer ce feature dans la phase data cleaning.

Relation entre les features :



D'après ce Histogram on a remarqué que les deux features qui ont une relation forte entre eux sont : Fatalities et ConfirmedCases. On a affiché le nombre total de Fatalities et ConfirmedCases dans chaque CountryRegion et les 5 premiers CountryRegion qui ont un nombre supérieure de Fatalities et ConfirmedCases sont : US, Italy, Spain, United Kingdom, France.

Il y'a 24 CountryRegion qui ont un nombre totale null de fatalities.

On a remarqué que les valeurs de toutes les features ne sont pas ordonnées dans la dataset.

Définition : La table de la crosstab() est une table de contingence qui permet de comparer deux variables catégorielles. Elle affiche le nombre d'occurrences de chaque combinaison

de valeurs des deux variables. Elle peut être utilisée pour déterminer si les deux variables sont corrélées ou non.
Le résultat de cette fonction dans notre dataset : en prennant l'exemple du combinaison entre ProvinceState et CountryRegion. On remarque que chaque ProvinceState est apparue que dans le pays qu'il l'appartient 115 fois, et ce résultat prouve que notre dataset est correcte.

II- Pre-processing :

-Shape de notre dataset : 35995 lignes et 6 colonnes.
- Le résultat de dropDuplicates est : 35995 lignes et 6 colonnes. Ça prouve que il y'a pas des lignes dupliquées.

Selétion de nos variables dans la dataset de training :
On a éliminé la variable 'ProvinceState' dans le train dataset parceque 57 pour cent des valeurs sont manquantes et c'est une variable catégorielle que l'on peut pas remplire avec des lignes vides (20700).

Train-nettoyage-encodage :
Nettoyage de la dataset :
Il ya des pays qui ont un nombre de Fatalities différent dans une même date mais dans des 'ProvinceState' différent . vu que on a éliminer la colonne 'ProvinceState' il reste des lignes avec la même 'CountryRegion' et la même 'Date' mais un nombre de fatalities différent et ça c'est pas logique . Donc en à créer une fonction qui faire la somme des Fatalities dans chaque pays qui apparait dans la dataset dans la même date .
On a proposer comme solution une fonction qui a regler ce problème(getFatalitieByCountryDate(df)).
Après ce cleaning le nouveau shape de notre dataset est : 21160 lignes, 3 colonnes.
On a vérifié que les pays (les catégories de la variable CountryRegion) apparaitre dans la dataset 115 fois qui représente de nombre des dates qui existe (le nombre de catégories de la variable Date).
Encodage de dataset :
Encodage de la variable " CountryRegion " : On a utilisé label incoding parce que le nombre des catégories est large , le résultat est que chaque CountryRegion est converti a un nombre (de 0 à 183). 6 colonnes et 35995 lignes. le résultat de la méthode de seaborn qu'on a utilisé affiche que la plus part de nos variable sont complétement non vide et juste une seule variables « ProvinceState » contient 57 pour cent de ses valeurs sont vide.
Encodage de la variable « date » : On a remarqué que le type de la variable date est Object

, d'abord on a transformé le type de cet variable à Datetime. Après ont a fait l'encodage.

III- Modèle:

La phase de modeling est la phase finale du processus de Machine Learning, où un modèle est construit à partir des données. Cette phase comprend l'entraînement du modèle, l'évaluation et le test du modèle.

L'entrainement du modèle :

D'abord, on a diviser notre dataset à : traindataset(Xtrain, Ytrain), testdataset(Xtest,Ytest), ensuite on a initialiser un nombre 300 de nestimator qui signifie le nombre d'arbres de décision qui seront utilisés pour construire le modèle.

Après on a entrainé le modèle, ensuite on a fait des prédictions sur le jeu de test et on les a comparé avec les jeux de test réel et on a afficher R2 score :

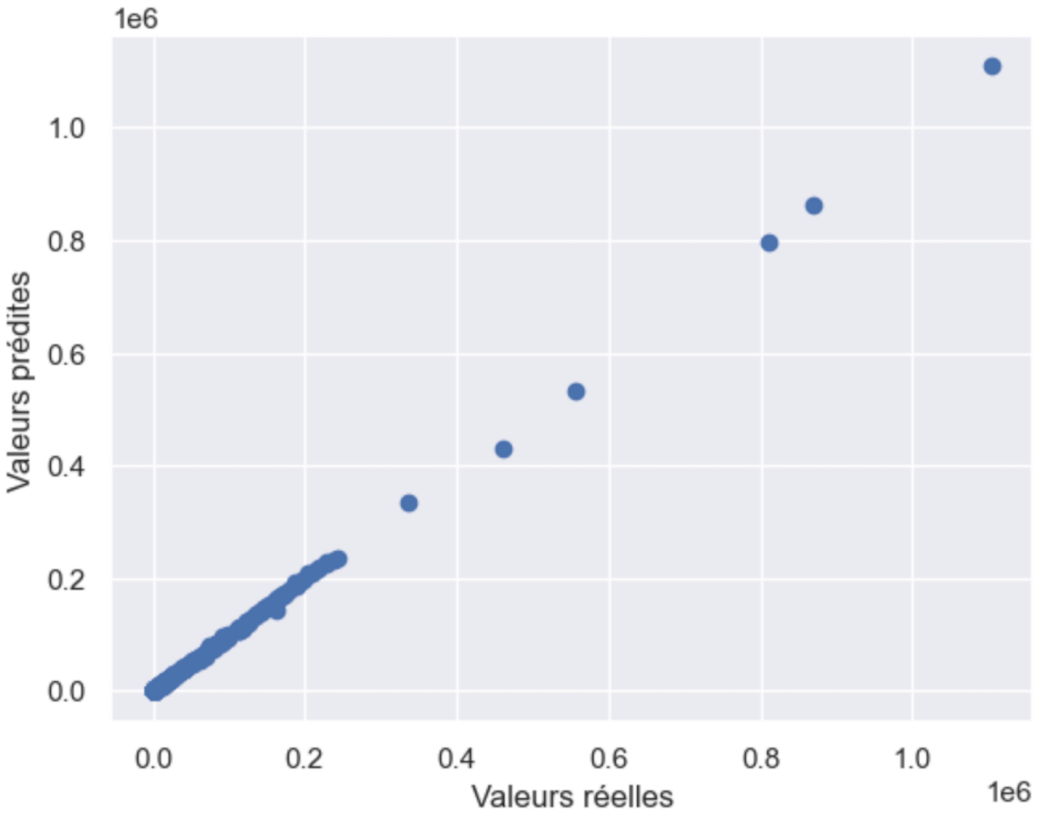
R2 score : 0.9993775166834664

Mean squared error : 0.16409664473030716

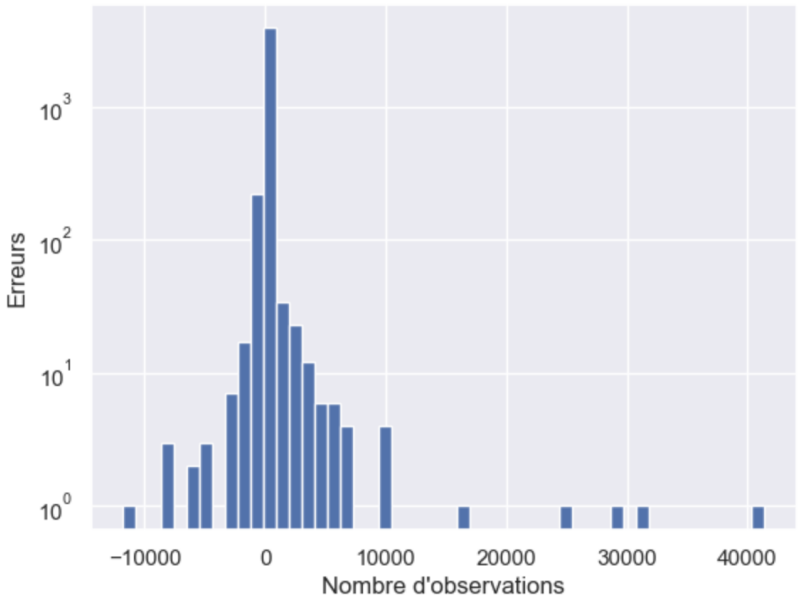
Mean absolute error : 0.02391776937614054

Visualisation des résultats :

On a essayer de visualiser les valeurs prédictées avec les valeurs réelles afin de faire une comparaison entre ces deux dernier.



on a résolut que ces deux valeurs sont très proches, donc on a réussi de faire le training de ce modèle.



IV - Conclusion :

En conclusion, le projet de machine learning de prédiction supervisée a été un succès. Les résultats obtenus montrent que la méthode de machine learning est efficace pour prédire les résultats avec une précision élevée. Les résultats obtenus sont prometteurs et peuvent être utilisés pour améliorer les processus de prédiction.

Références:

1)Vidéo 01
2)Vidéo 02
3)Vidéo 03
4)Vidéo 04
5)Documentation Numpy
6)Documentation Pandas
7)Documentation Matplotlib
8)Documentation Scikit-learn