كلية العلوم و التقنيات فاس
ⵜⴰⵖⵚⵓⵍⵜ ⵏ ⵜⵇⴰⵙⵙⵉⵏ ⴷ ⵜⵙⵙⵜⵉⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵃⵎⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰ
Université Sidi Mohamed Ben Abdellah

Projet de Fin d'Etudes

Licence Sciences et Techniques Génie Informatique

Application d'Analyse Du Big Data Pour Décrire, Prescrire & Prédire

Cas d'Etude De Sars-Cov-2

Lieu de stage : **Centre Hospitalier Universitaire Hassan II à Fès - CHU Hassan II**

Réalisé par :

Assia HAYATI

Encadré par :

Pr. Loubna LAMRINI

M. Mounir MAKHLOUK

Soutenu le 10/07/2021 devant le jury composé de :

Pr. Jamal KHARROUBI

Pr. Fatiha MRABTI

Pr. Loubna LAMRINI

Année Universitaire 2020-2021

DEDICACES

À

Mes précieux parents et mes chères sœurs, que leurs mots d'encouragement et leurs pousser vers la ténacité résonnent à mes oreilles.

Toute personne passionnée par l'analyse des Big Data.

REMERCIEMENTS

Grand merci

À

« Curiosité », qui m'a donné l'envie d'enquêter sur la boîte noire de cette science

Mes superviseurs : Mme. Loubna LAMRINI et M. Mounir MAKHLOUK, pour être mon guide, ma boussole, ma référence.

Mes professeurs, qui au fil des ans, m'ont aidé à développer une compréhension de ce domaine magique.

Les collecteurs de données, les fournisseurs open source, les partageurs de recherche, pour leur générosité, pour donner à ce travail la chance d'exister.

Tout

Je vous dis merci beaucoup

Je vous apprécie profondément

RESUME

Le 30 janvier 2020, l'organisation mondiale de la santé a officiellement déclaré la COVID-19 comme une urgence de santé publique de portée internationale. En réponse, les autorités ont imposé des réglementations, les agences ont annoncé des recommandations, les médecins ont défini des protocoles et les chercheurs ont mené des études. Ce document est une tentative d'assiéger cette pandémie, en facilitant les mesures d'atténuation et de contrôle, en raison d'une meilleure prise de décision.

Pour atteindre l'objectif décrit ci-dessus, nous allons, d'abord, mener une approche de résolution, à travers une analyse descriptive (c'est quoi le problème ?) et une autre prescriptive (Comment résoudre ce problème ?). Son implémentation consiste à identifier les facteurs primaires potentiellement contribuant aux chiffres de l'infection et de la mortalité, ce qui nous permettra de juger les politiques, éradiquer les mythes et détecter les hypothèses trompeuses. Et puis nous plongerons dans une approche préventive en utilisant l'analyse prédictive (que se passera au futur ?) et une autre prescriptive (comment l'éviter s'il est négatif ?), en créant des modèles (Régression Linéaire Multiple et les Arbres de Décisions) qui imitent le comportement des crises, pour avoir un aperçu de l'avenir.

Nos résultats montrent que Covid-19, a dépassé la loupe d'une épidémie pour être qualifier de pandémie, elle indique, également, que les pays qui ont le plus souffert sont des pays européen, ce qui classifie : la démographie, l'environnement et l'exposition internationale comme des facteurs de risque potentiels. Après, elle déclare les éléments interférant avec la crédibilité des résultats des analyses, comme : la non-normalisation et le manque des tests. Ensuite, cette analyse montre le rôle positive qu'a joué le confinement. Finalement, elle cite les pays suivant les même patterns.

Mots-clés

Big data, analyse exploratrice, analyse descriptive, analyse prescriptive, modélisation prédictive, régression linéaire multiple, k-means clustering, prise de décision, covid 19.

ABSTRACT

The 30 th of january 2020, the World Health Organisation officially declared COVID-19 as a Public Health Emergency of International Concern. In responce, politicians imposed regulations, agencies announced recommendations, doctors defined protocoles and researches conducted studies. These papers are an attempt to besiege this pandamic by facilitating mitigation and control measures, due to better decision-making.

To fulfill the objective drawn above, we shall first conduct a solving approach, through a descriptive and a prescriptive analysis, in order to combine the reported facts with the primary factors contributing towards the infection and mortality numbers, which enables us to judge the policies, diminish the myths and detecte the misleading hypothesis. And then we'll dive into a preventive approach using predictive and prescriptive analysis, by creating models (multi linear regression model as well as the decision tree) that mimics the crises' behavior, to get an insight into the future.

Our results show that Covid-19, exceeds, in deed, the dimention of an epidemic and qualifies as a pandamic, it also indicates that the countries who suffered the most were generally european ones, which classifies : the demography, the environment and the international exposure as potential risk factors. It, then, shades some light on the elements that interfere with the credibility of the analysis such as : the non-normalisation and the lack of tests. Moreover, this analysis displays the role of the lockdown. Finally, it declars countries having the same patterns.

Key-words

Big data, exploratory analysis, descriptive analysis, prescriptive analysis, predictive modelling, multiple linear regression, k-means clustering, decision-making, covid19.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABBREVIATIONS

Abbréviation	Description
BDA	Analyse des mégadonnées
CHU	Centre hospitalier universitaire
EDA	Analyse exploratoire des données
MC	Modèle de clustering
MRL	Modèle de régression linéaire
OMS	Organisation mondiale de la santé
USPPI	Urgence de santé publique de portée internationale

INTRODUCTION

Ces papiers sont le fruit d'une étude scientifique, menée au sein du CHU Hassan II. Ils visent à augmenter nos chances de gagner la bataille contre la crise sanitaire mondiale, grâce à une meilleure prise de décision. Ce, en faisant recours à deux types d'analyse : une prescriptive et une autre prédictive. L'analyse prescriptive va nous permettre d'évaluer les hypothèses proposées, et de juger les politiques mises en place pour lutter contre la pandémie actuelle du covid 19. Quant à l'analyse prédictive, on cherchera à obtenir un aperçu futuriste du post-corona virus.

Pour atteindre les objectifs mentionnés précédemment, nous discuterons le sujet comme suit :

Chapitre 1 : Contexte général du projet.
C'est le chapitre " préambule ", où nous décrivons le lieu de stage, présentons le travail connexe et discutons la problématique, la solution et le plan.

Chapitre 2 : Analyse et conception.
C'est le chapitre " décollage", dans lequel nous présentons une vue théorique en effectuant une analyse fonctionnelle et technique ainsi que la conception adoptée.

Chapitre 3 : Mise en œuvre du projet.
C'est le chapitre " atelier ", où nous représentons une interface graphique / des visualisations chacune avec sa description respective, son interprétation et ces conclusions.

CHAPITRE I

CONTEXTE GENERAL DU PROJET

1. Description Du Lieu De Stage

Sa Majesté le Roi Mohammed VI, a inauguré, le 14 janvier 2009, le Centre Hospitalier Universitaire de Fès, Maroc – CHU Hassan II.

Cet hôpital du 3ème niveau, est un établissement public qui relève de la compétence du ministère de la Santé, avec une personnalité morale légale et une autonomie financière.

Ce centre a 3 missions principales :

- ✓ Fournir les soins de base ou intensifs aux patients résidentiels / régionaux.
- ✓ Préparez le futur personnel médical / paramédical, en leur garantissant de la connaissance pratique sur le terrain.
- ✓ Ravitailler les chercheurs par les ressources nécessaires, pour mener des études liées au domaine de la santé.

Voici les dernières statistiques du CHU Hassan II :

- ✓ Superficie couverte : 12 hectares.
- ✓ Nb des Hôpitaux : 5.
- ✓ Nb des Services : 42.
- ✓ Nb des Agents : 2460.
- ✓ Nb des Salles d'opération : 28.
- ✓ Nb des Lits : 1050.
- ✓ Coût global : 1200 milliards de DH.

Ci-après, l'organigramme du CHU :

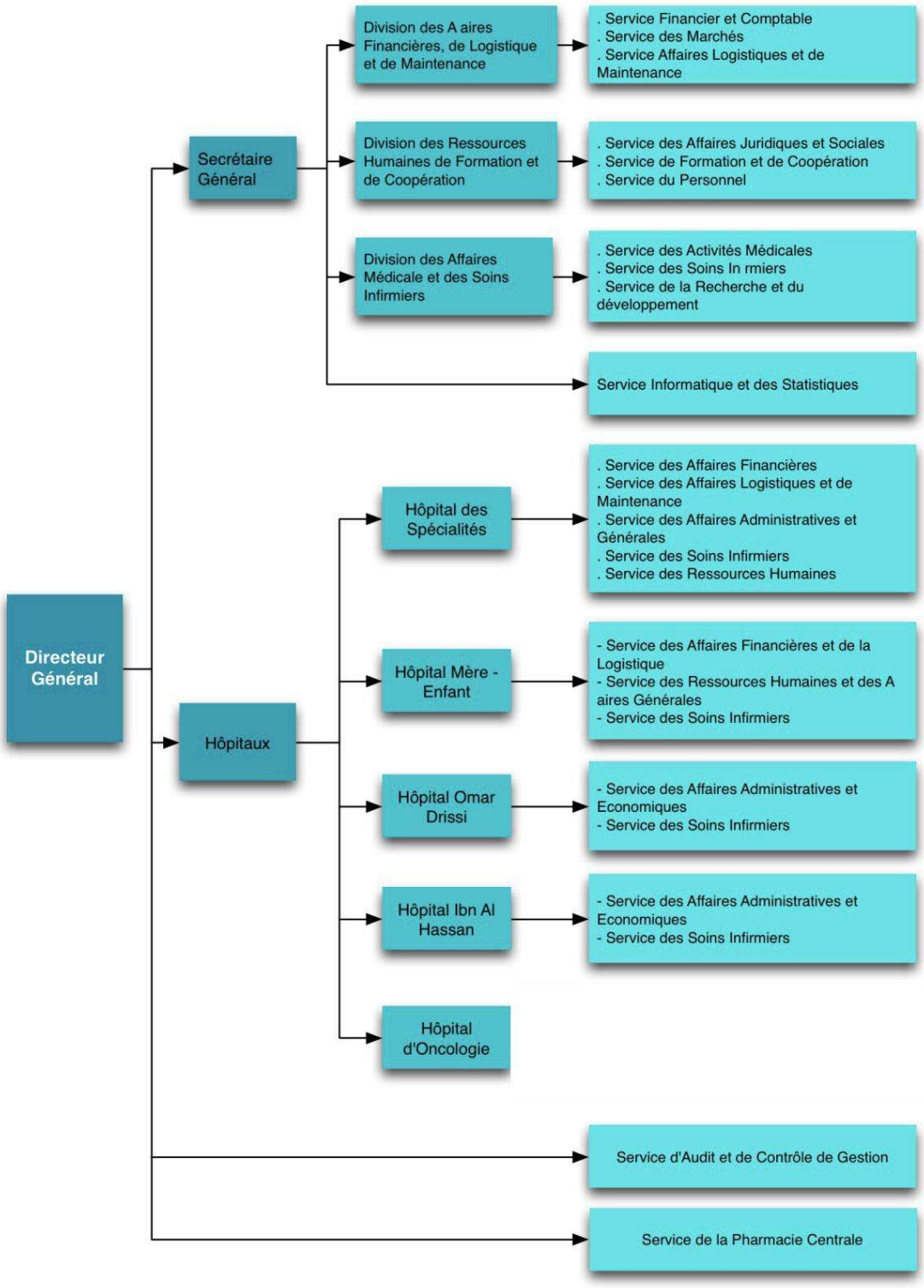


Figure1. Organigramme du CHU Hassan II, Fès (Source : www.chu-fes.ma)

2. Travaux Connexes

S'il n'y a quelque chose qui s'était propagé dans le monde entier cette dernière année plus rapidement que covid 19, ça sera les recherches. Des recherches qui aspirent obtenir des informations utiles de ce chaos de données brutes dont nous avons été bombardés. Ce travail est une tentative de rejoindre la communauté des analystes de données et de contribuer à la stratégie qui finira cette tragédie.

Cependant, le bonus présent dans cette étude par rapport à ceux existantes, réside dans son aspect de discussion multidimensionnelle, puisque l'analyse comprend une grande quantité des facteurs suspects critiques dans les domaines les plus cruciaux (médical, social, politique, environnemental, démographique et ethniques), à toutes les phases(vague) de la crise à partir du premier jour du tout premier cas signalé, dans tous les pays du monde.

3. Énoncé Du Problème

Est-ce que le pire est encore à venir ? Combien de décès supplémentaires à compter ? Quelles politiques adopter ? Quels protocoles ont désavoué ? Quels facteurs sont nos ennemis ? Quelles décisions sont nos alliées ? qu'est-ce que ces données tentent de nous dire ?

Avec tant de données brutes, se posent tant de questions.
Il est temps d'obtenir des réponses, il est temps de récupérer quelques informations.

4. Description De La Solution

Investiguer cette problématique, implique d'investir dans l'analyse big data, ainsi :

- Réalisation d'un BDA descriptif – Vue statistique :
  - ✓ Mise en place de l'environnement technique.
  - ✓ Traitement des données.
  - ✓ Génération de visualisation des données, à observer puis interpréter.
  - ✓ Évaluer la situation actuelle, comprendre le comportement de crise et obtenir quelques faits.

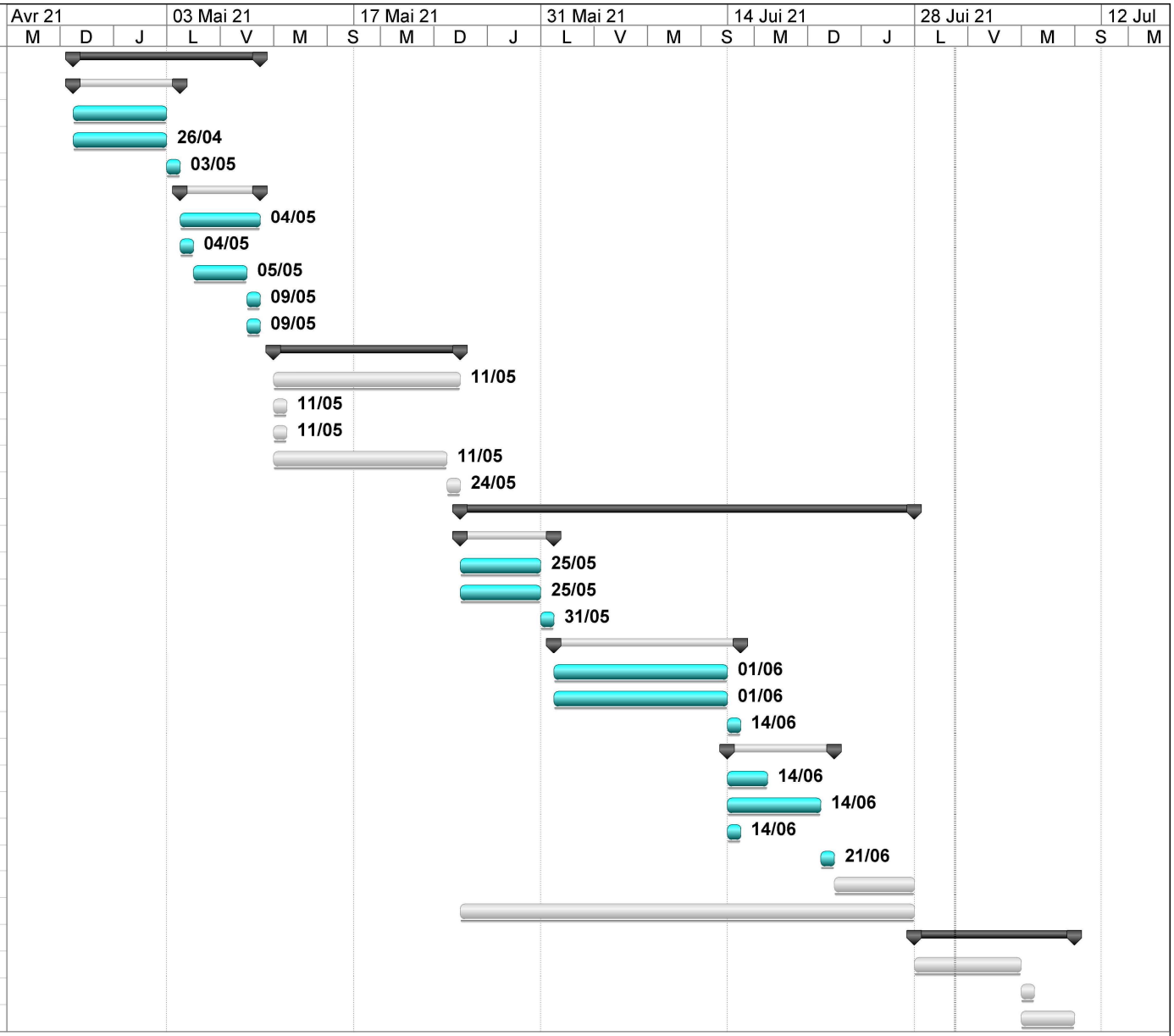
- Réalisation d'un BDA prescriptif – Vue statistique :
  - ✓ Extraire les tendances qui ont mené à des résultats positifs/négatifs spécifiques (pyramide de la population).
  - ✓ Évaluer les hypothèses en fonction du ratio (attente, réalité).
  - ✓ Jugez les politiques et les protocoles en fonction du ratio (cause = décision, conséquence = visualisation/ensemble de données).
  - ✓ Élaborer un plan pour une future crise sanitaire mondiale (Liste des tâches à faire/Liste des tâches à ne pas faire).

- Réalisation d'un BDA prédictif – Vue probabilité :
  - ✓ Création d'un modèle (MLR et MC) qui imite le comportement de la pandémie à l'aide d'une base de données d'apprentissage.
  - ✓ Évaluer le taux d'erreur / la corrélation, du modèle en effectuant des tests à l'aide d'un ensemble de données de test.
  - ✓ Déduire des prédictions (à vérifier dans le futur).

5. Plan – Diagramme De GANTT

Le diagramme suivant montre les différentes tâches qu'il fallait accomplir, dans l'ordre, avec la date de début, la date de fin et donc la durée proposée.

N°	Nom de la tâche
1	**PHASE 1 : START**
2	**Task 1 : Pick A Topic**
3	Brainstorming / Research
4	Select Criteria
5	Contact Supervisors
6	**Task 2 : Answer The 6W's Questions**
7	What ? Details
8	Why ? Objectives
9	How ? Concepts
10	With ? Hardware / Software
11	When ? Plan
12	**PHASE 2 : PREPARE WORK ENVIRONMENT**
13	Collecte Data
14	Install IDE / Packages
15	Dowload Documentation
16	Learn Language
17	Contact Supervisors
18	**PHASE 3 : IMPLEMENTATION**
19	**Task 1 : Descriptive BDA**
20	Generate Visualizations
21	Coclude : Evaluate Situaion / Describe Behavior
22	Contact Supervisors
23	**Task 2 : Prescriptive BDA**
24	Read Research / News
25	Conclude : Identify Factors
26	Contact Supervisors
27	**Task 3 : Predictive BDA**
28	Create LR Model & CM
29	Evaluate Corrolation / Error rate
30	Conclude : Forcast
31	Contact Supervisors
32	**Task 4 : Dashboard**
33	**Task 5 : Document Process / Results**
34	**PHASE 4 : END**
35	**Task 1 : Write Report**
36	**Task 2 : Contact Supervisors / Jury**
37	**Task 3 : Prepare Presentation**



Projet : pro1
Date : Jeu 01/07/21

Tâche		Jalon		Tâches externes	
Fractionnement		Récapitulative		Jalons externes	
Avancement		Récapitulatif du projet		Échéance	

CHAPITRE II

ANALYSE ET CONCEPTION

---

1. Analyse Fonctionnelle

Le Big Data Analysis requiert une “Big Analyse fonctionnelle” afin de pouvoir identifier les différentes modalités de son écosystème, la recognition de la fonction principale (fonction de haut-niveau) et sa décomposition à des sous-fonctions (fonctions de bas-niveau), ainsi que les mécanismes utilisés et les critères pris en compte. Pour ce faire, l’approche TACS connue aussi par IDEF0, ainsi que l’approche FAST seront étudiées. Le choix de ces 2 méthodes vient du fait que chacune à un champ de vision différent, et leur association donnera une idée complète de l’étude.

1.1. L’approche TACS / IDEF0

1.1.1. Acronyme

L’approche TACS- Technique d’analyse et de conception structurée.

L’approche IDEF0- Définition d’intégration pour la modélisation de fonction.

1.1.2. Tactique

La méthode s’appuie sur une technique interrogative en posant deux questions :

- ✓ Quel est l’objectif final de l’étude ?
- ✓ Quels mécanismes utilisés ?
- ✓ Quelle sont les critères à valider ?
- ✓ Quelles sont les entrées ?
- ✓ Quelles seront les sorties ?

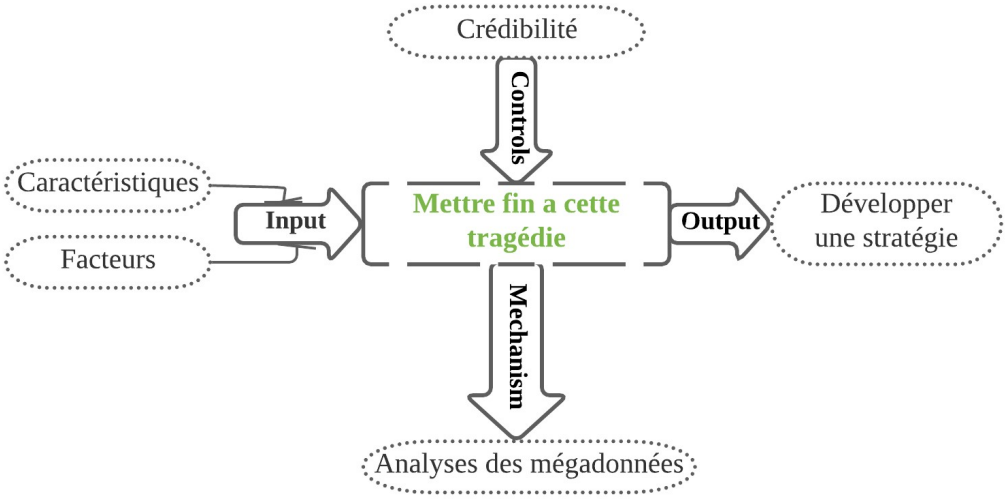


Figure3. Schéma de l'approche TACS/IEDF0.

1.2. L'approche de FAST

1.2.1. Acronyme

L'approche de FAST- Technique des systèmes d'analyse fonctionnelle.

1.2.2. Tactique

La méthode s'appuie sur une technique interrogative en posant deux questions :
1. COMMENT accomplir (fonction) ?
En lisant de haut en bas.
2. POURQUOI est-il nécessaire de (fonctionner) ?
En lisant de bas en haut.

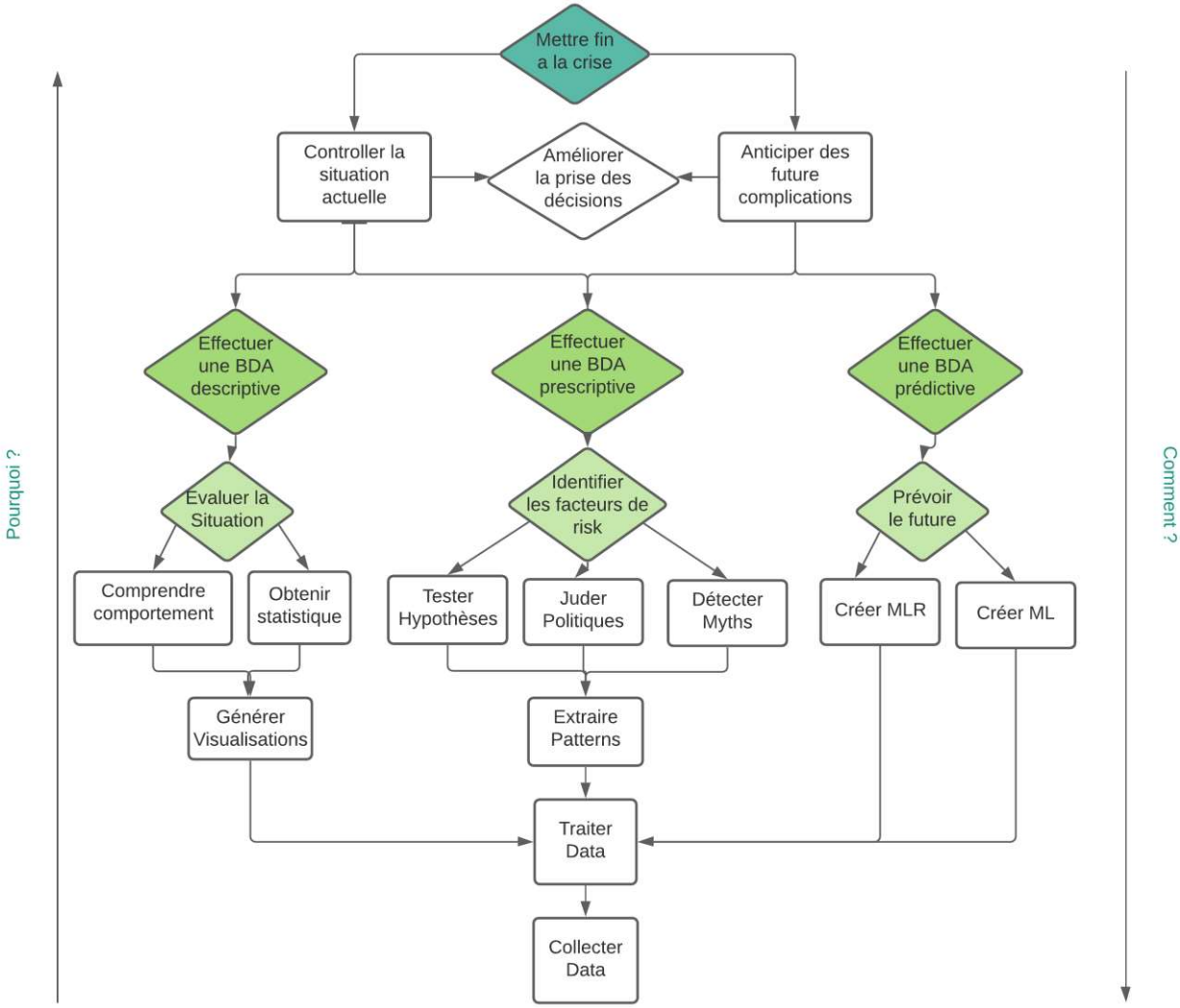


Figure4. Schéma de l'approche FAST.

2. Analyse Technique

Le Big Data requiert une “Big Analyse Technique” afin de pouvoir identifier les différents composants en termes de hardware, de software et de concept de son écosystème.

2.1. Hardware

Stocker, traiter et puis analyser un volume colossal des données, induit des besoins en termes de matériel. D’où la nécessité de se disposer d’un appareil numérique (ex : ordinateur) avec :

- ✓ Mémoire d’accès aléatoire - RAM puissante (fortement consommée dans la visualisation / génération des modèles).
- ✓ Disque dur puissant (utilisée pour le stockage des bases de données).

2.2. Software

Générer des visualisations et créer des modélisations, induit des besoins en termes de logiciel. D’où la nécessité de se disposer d’un environnement de développement intégré :

- Programmation locale :
- ✓ Installez “Anaconda”.

Qu’est-ce que c’est ?

Anaconda est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique, qui vise à simplifier la gestion des paquets et de déploiement. A noté que : “Anaconda = Collab notebook+ Python + majorité des Packages ".

- ✓ Installez PyCharm...

- Cloud Coding / Programmation en ligne :
- ✓ Utilisez “Colaboratory” .

2.3. Langage de Programmation

L’outil élu pour notre étude est le fameux langage de programmation Python.

Pourquoi Python ?

- ✓ Langue en demande.
- ✓ Librairies puissantes (pas seulement pour l'analyse de données).
- ✓ Langage de programmation à usage général.
- ✓ Communauté incroyable, documentation et références.
- ✓ Gratuit et open source.
- ✓ Intuitif et simple à apprendre.

2.4. Paquets / Modules /Bibliothèques

Devoir rédiger un code explicite pour chaque tache minuscule, peut s'avère fastidieux, d'où l'intérêt des "bibliothèques". Les bibliothèques sont des morceaux de code (fonctions, classes...etc) regroupées et mises à disposition afin de pouvoir être implémentées sans avoir à les réécrire.

On utilisera plusieurs packages pour mener cette étude, comme :

- ✓ NumPy

Bibliothèque destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.

- ✓ Pandas

Bibliothèque permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

- ✓ Matplotlib

Bibliothèque destinée à tracer et visualiser des données sous formes de graphiques

- ✓ Scikit-learn

Bibliothèque dédiée à l'apprentissage statistique (machine learning) et peut être utilisée comme middleware, notamment pour des tâches de prédiction

2.5. Data

2.5.1. La vie des données

Le type de traitement réalisé sur un élément, implique la valeur de ce dénier.

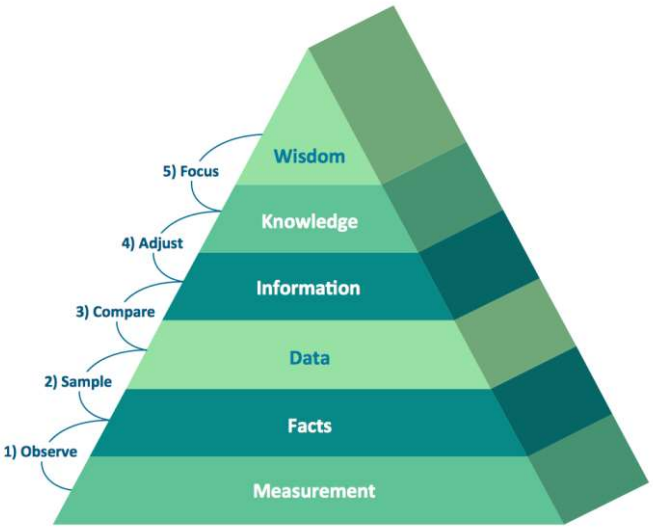


Figure5. Cycle de vie des données.

2.5.2. Base de données

Pour créer/choisir une base de données, plusieurs critères sont pris en compte :

- ✓ La crédibilité de la source.
- ✓ La régularité des mises à jour.
- ✓ Le nombre de lignes (taille des données) et le nombre de colonnes (nombre des caractéristiques).
- ✓ La validation des 5 Vs.

2.5.3. Meta data

Les données utilisées dans cette étude sont collectées de plusieurs sources (Kaggle, GitHub, site d'une organisation ou d'une université..etc), pour les utiliser, soit on les a stockées sous forme de fichier csv, ou bien on a extrait leur code depuis une page html. On a travaillé sur des données quotidiennes commençant par début janvier 2020 jusqu'à fin mai 2020 (lignes) et par pays. Ces données concernent les chiffres des cas confirmés, des décès, des cas critiques, des guérisons, des vaccinations reportés, ainsi que des caractéristiques potentiellement contribuant à ces chiffres, comme la densité de la population, la démographie, un score qu'on a calculé indiquant les politiques et les protocoles prises.

2.6. Concepts

2.6.1. Qu'est-ce que le Big Data ?

« Big Data c'est discerner des formes et les cohérences dans le désordre informationnel ambiant. »

Source : www.piloter.org

"C'est un processus d'inspection, de nettoyage, de transformation et de modélisation des données dans le but de découvrir des informations utiles, d'éclairer les conclusions et de soutenir la prise de décision."

Wikipédia

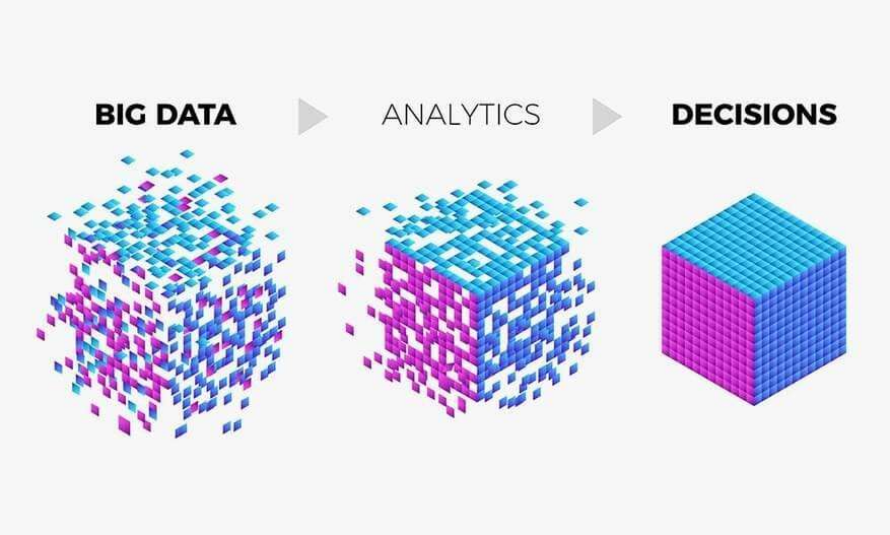


Figure6. Des big data aux décisions.

2.6.2. Historique/Evolution

Ces dernières années on a vécu un big bang du big data, vu que tout usage d'un objet connecté alimente la nébuleuse de données.

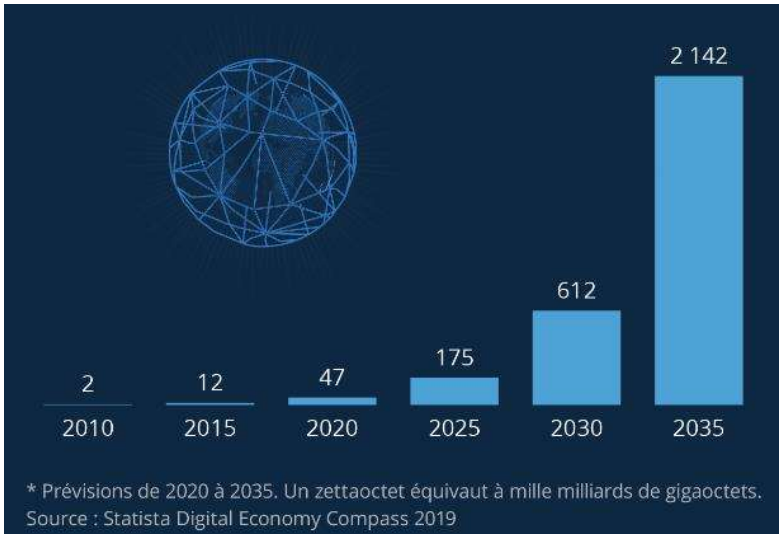


Figure7. Le big bang des big data (Source : statista)

2.6.3. Caractéristiques

Il existe cinq propriétés de définition qui peuvent aider à décomposer le terme “Big Data”. Surnommé les 5 Vs.

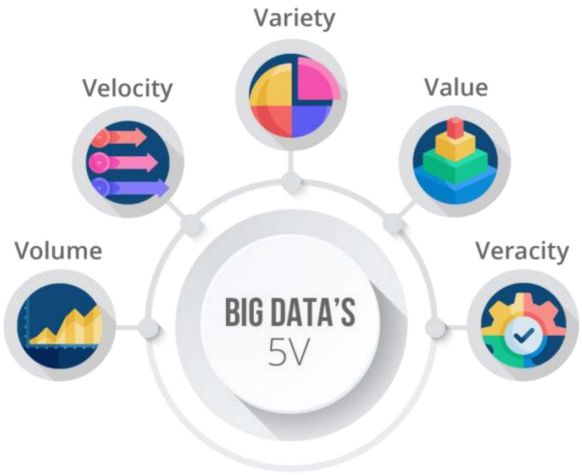


Figure8. Les 5Vs du big data

- ✓ Volume :

Cela fait référence au volume pur des données générées chaque instant. Ces volumes sont devenus tellement massifs que nous ne parlons plus en Téraoctets mais en Zettaoctets pour les quantifier.

- ✓ Vitesse :

Également appelée Vélocité, correspond à la rapidité à laquelle les data sont générées et circulent. Le Big Data permet l’analyse d’informations en temps réel et leur transmission à un rythme effréné. Ainsi, les entreprises peuvent faire preuve d’une réactivité et d’une agilité incomparables quasiment instantanément.

- ✓ Variété :

Elle désigne la multiplicité des types de données disponibles. Auparavant, les data étaient majoritairement des données structurées, faciles à classer et organiser. Aujourd’hui, de nombreuses data non-structurée comme les données textuelles sont générées à chaque seconde.

- ✓ Valeur :

Avoir accès au big data, c’est bien beau, mais ce n’est utile que si nous pouvons en faire une valeur précieuse ajoutée à l’entreprise. Il est donc crucial, avant de lancer son projet Big Data, de savoir pourquoi et comment on va le mener afin d’évaluer la future rentabilité.

- ✓ Véracité :

Elle désigne à la fiabilité de la data qui est essentielle pour pouvoir en tirer profit et la transformer en information utilisable dans l'entreprise.

Ces dernières années avec l'explosion qu'a connu le monde des big data, on parle du 6Vs, 8Vs même 100Vs, on cite par exemple :

- ✓ Vertu :

La vertu fait référence aux réglementations en matière de confidentialité et de conformité des data. L'aspect éthique et le respect des normes en vigueur concernant les données sont cruciaux pour traiter les informations.

2.6.4. Que nous disent ces données ?

Afin d'obtenir une certaine utilisation de cette énorme quantité de données, nous pouvons effectuer des analyses.

Il y a 4 principaux types d'analyse de données :

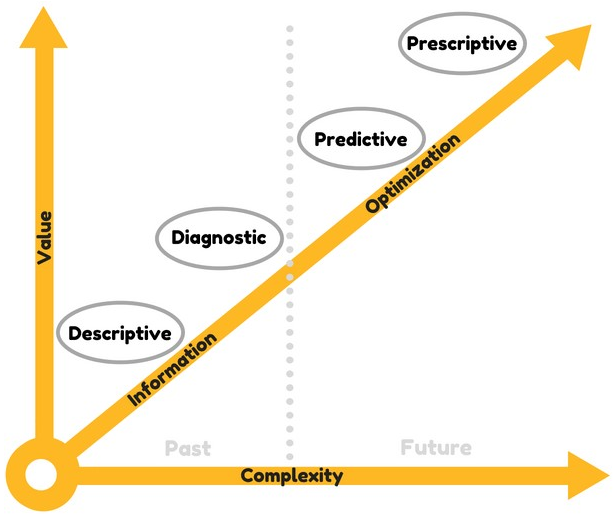


Figure9. Types d'analyses des big data

- ✓ Descriptif : Que se passe-t-il dans mon entreprise ?

-Données complètes, précises et en direct.

-Visualisations efficaces.

- ✓ Diagnostic : Pourquoi cela se produit-il ?

-Capacité d'explorer jusqu'à la racine - cause.

-Capacité d'isoler toutes les informations confusionnelles.

- ✓ Prédictif : Que risque-t-il de se passer ?

-Les stratégies d'affaires sont restées assez stables au fil du temps.

-Modèles historiques utilisés pour prédire des résultats spécifiques à l'aide d'algorithmes.

-Les décisions sont automatisées à l'aide d'algorithmes et de la technologie.

- ✓ Prescriptif : Que dois-je faire

-Actions et stratégies recommandées basées sur les tests de champion / chalenger et les résultats de la stratégie.

-Appliquer des techniques d'analyse avancées pour faire des recommandations spécifiques.

2.6.5. Phases

Il y a 4 phase principale qui il faut franchir avant de pouvoir passer de l'objectif à l'action, des données aux informations. Ce sont :

Figure10. Les phases d'analyses des big data

Note :

Le processus est circulaire plutôt que linéaire, puisque nous faisons, généralement, des allers-retours entre les phases mentionnées ci-dessus.

3. Conception - Diagrammes UML

Nous utiliserons se langage pour modéliser notre étude, afin d'avoir une vision claire sur ces différents aspects.

3.1. Diagramme Des Cas d'Utilisation

Ce modèle capture les exigences comportementales de mon étude. C'est un moyen de communiquer avec les utilisateurs et d'autres parties prenantes ce que cette analyse est destinée à faire.

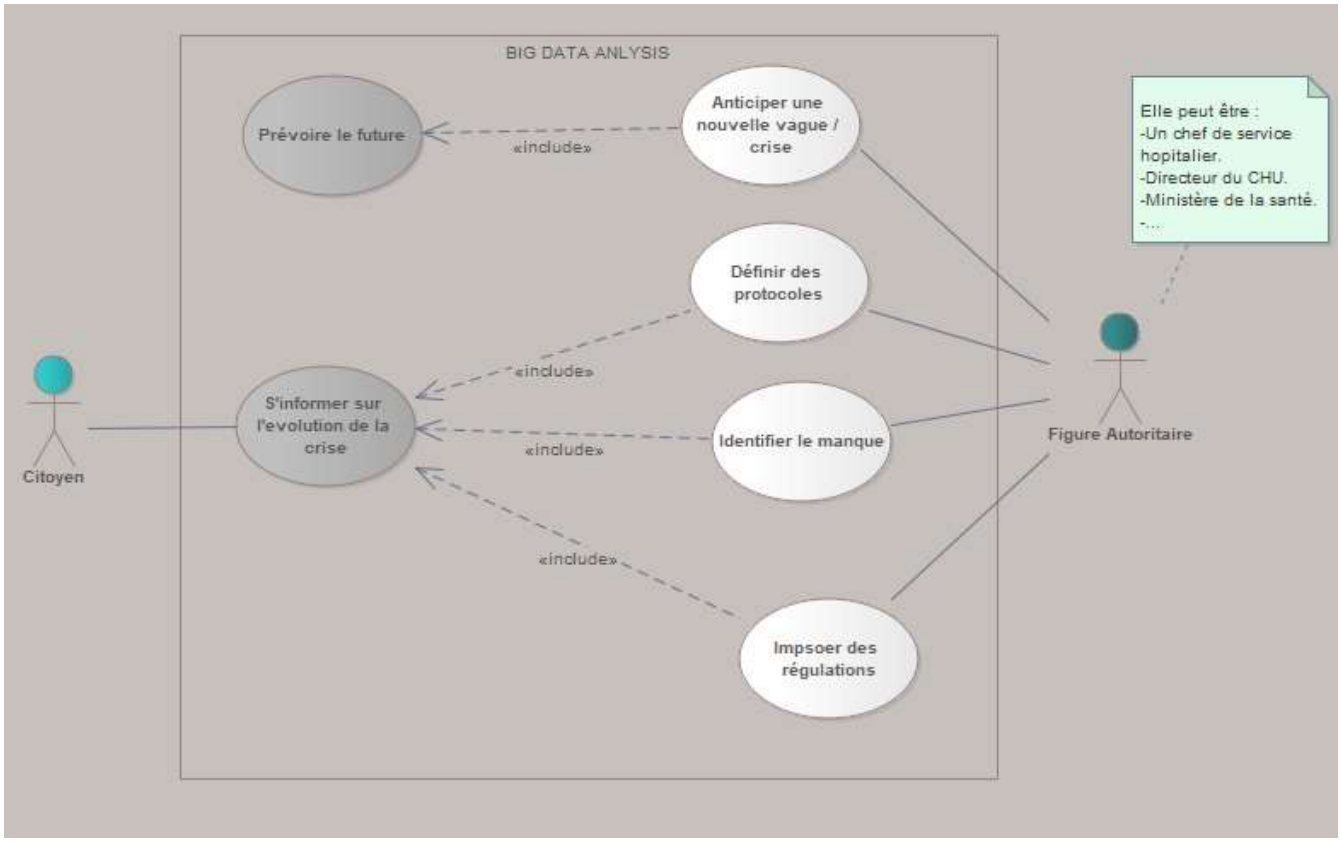


Figure11. Diagramme des cas d'utilisation.

3.2. Diagramme De Séquence

Ce diagramme met l'accent sur le comportement des messages échangés entre les acteurs (les utilisateurs) et le système, présentés dans un ordre chronologique.

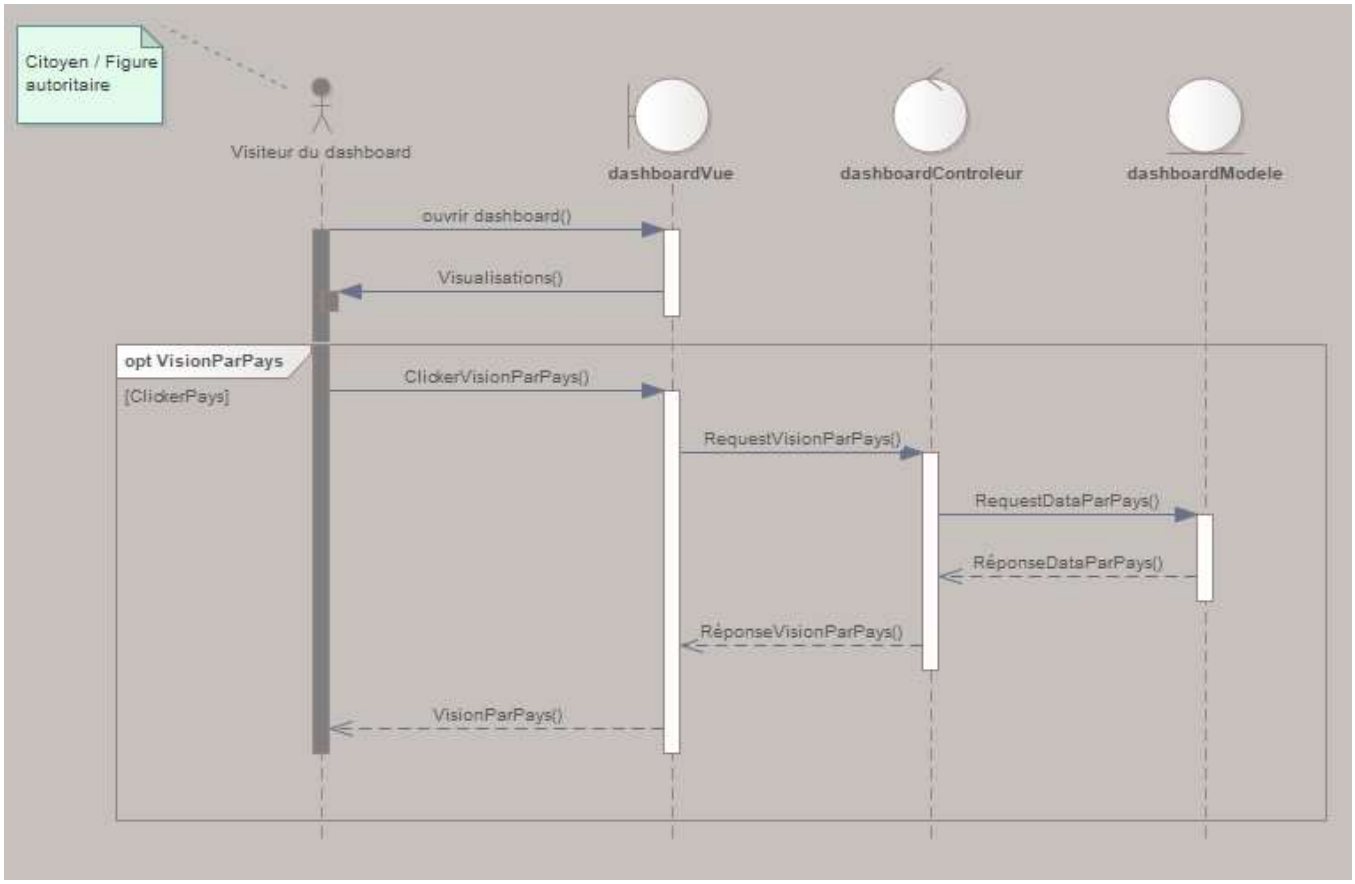


Figure12. Diagramme de séquence

3.3. Diagramme Des Composants

Ce diagramme illustre l'architecture physique d'un système en définissant sa structure, càd, en décrivant les composant dites les “blocks” et leur connectivité.

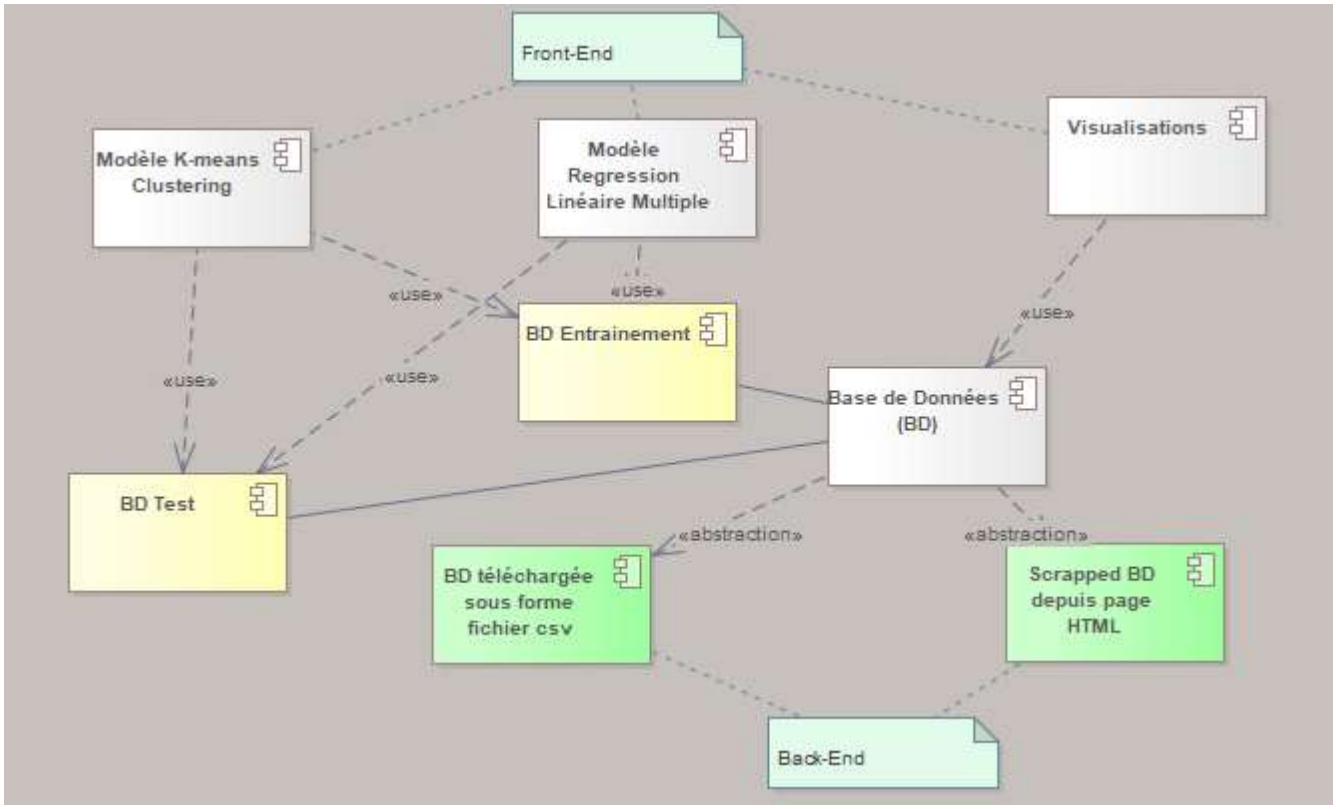


Figure13. Diagramme des composants.

CHAPITRE III

MISE EN OEUVRE DU PROJET

MISE EN SITUATION

Covid-19 est-elle une pandémie ou une épidémie ? Quel est son comportement ? Quelle est la situation actuelle de la crise ?

Répondre à ces questions nécessite une analyse descriptive.

1. Analyse descriptive

1.1. Tache 1

But

Observez le comportement d'une pandémie.

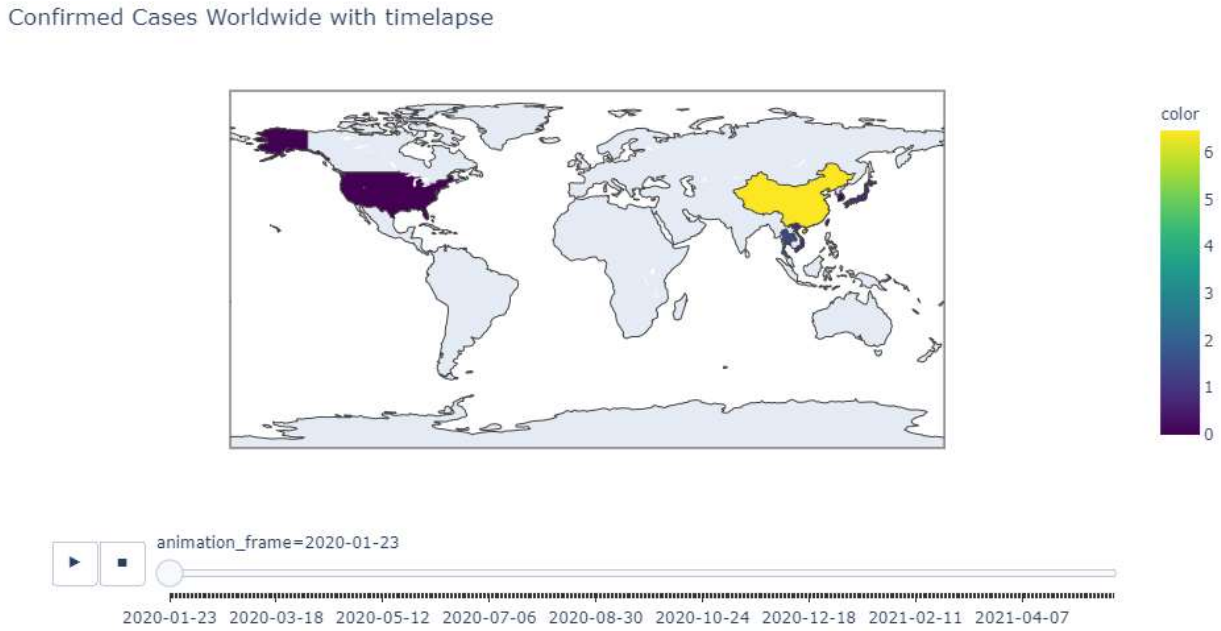


Figure14. Carte choroplèthe des cas confirmés au monde le 23/01/2020 (Source data : JHU site).

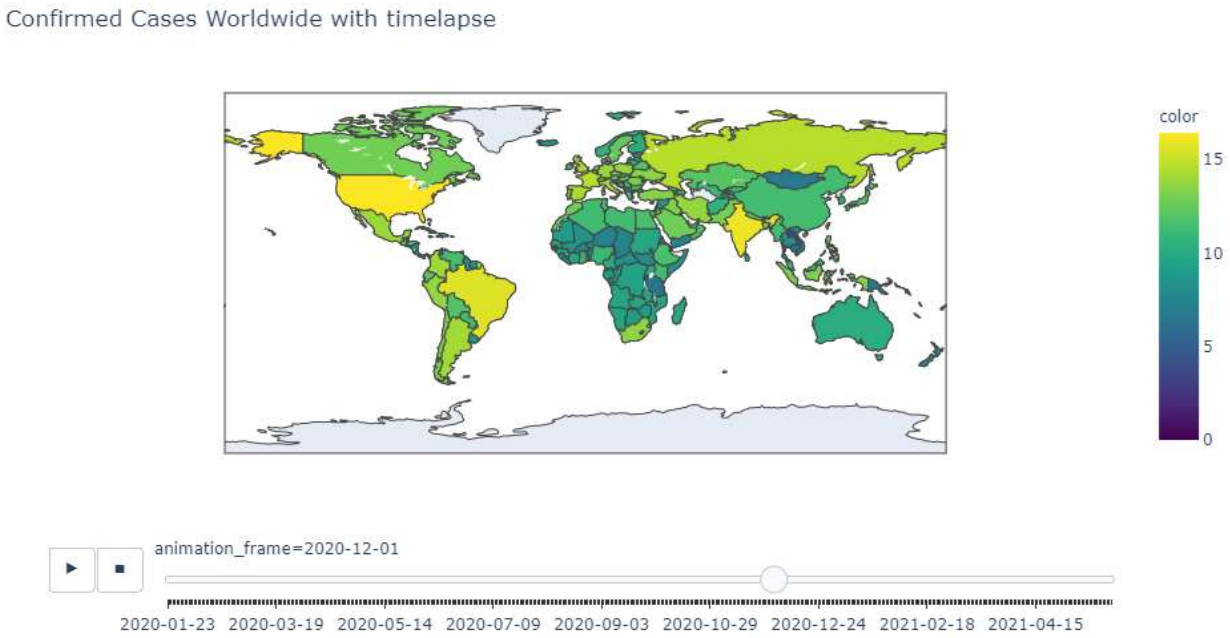


Figure15. Carte choroplèthe des cas confirmés au monde le 01/12/2020 (Source data : JHU site)

Description

Carte choroplèthe, une carte dynamique qui montre l'intensité des cas confirmés dans le monde en fonction d'un laps du temps quotidien (23/01/2020 - 31/05/2021).

Interprétations

-Cela a commencé en Chine, en Asie, avec les cas confirmés les plus élevés à signaler à janvier le 23 de 2020.

-Il a couvert le monde entier le 1er décembre 2020, avec l'apparition de nouveaux points chauds autres que l'Asie, comme l'Amérique du Nord et l'Amérique du Sud.

Conclusions

- Sars.Cov.2 a transgressé la boucle d'une épidémie, c'est une pandémie.

1.2. Tache 2

But

Obtenir un aperçu de la situation actuelle.

Figure16. Aperçu des nombres actuelle (Source Data : kaggle JHU Challenge).

Description

Parcelle de zone, un diagramme linéaire interactif qui représente le nombre total dans le monde entier (jusqu'au 31 mai 2020) des patients rétablis en orange, les actifs en vert et les cas de décès en rose.

Interprétation

-Le total des recouvrements est de 104 774 370.

-Le total des décès est de 3 484 466.

-Le nombre total d'actifs est de 59 529 145.

Conclusions

Le nombre des guérisons dépasse à la fois le nombre de décès et d'actifs, ce qui donne un sentiment de soulagement.

1.3. Tache 3

But

Comparez le Sars-Cov-2 à d'autres épidémies similaires.

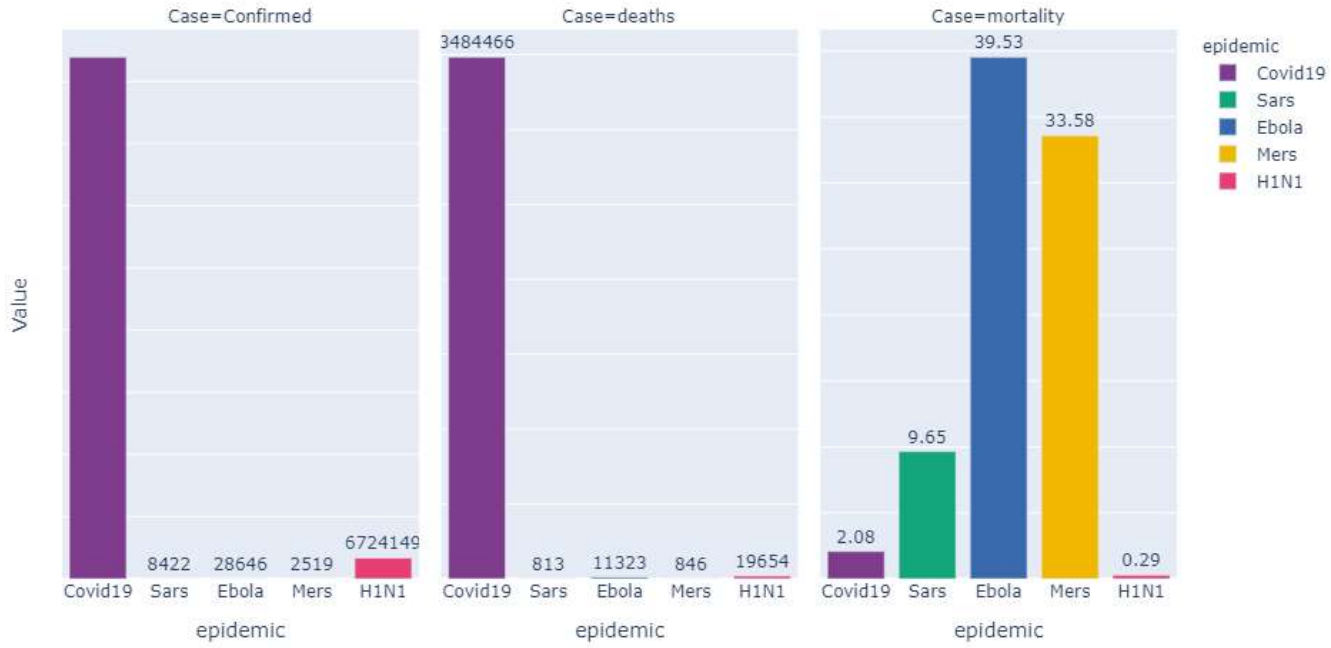


Figure17. Covid-19 Vs autres épidémies (Source Data : GitHub Covid)

Description

Un graphique à barres, associer une couleur à différentes épidémies dont le covid 19 pour comparer la sévérité du Sars-Cov-2.

Interprétation

Alors que le 19-nCov est la principale infection en termes des cas confirmés et des décès, ce n'est pas le cas en ce qui concerne le taux de mortalité.

Conclusions

L'approche des « chiffres » peut être trompeuse, vaut mieux adopter une approche de « rapport », ce qu'on appelle la normalisation des données.

MISE EN SITUATION

Qu'est-ce qui se cache derrière toutes ces données ? Quelles sont les politiques déguisés derrière la hausse/baisse des chiffres de mortalité ? Quelles sont les facteurs susceptibles à influencer l'évolution de la crise ?

Répondre à ces questions nécessite une analyse prescriptive.

2. Analyse prescriptive

2.1. Tache 1

But

Identifier les facteurs contribuant à des taux élevés / faibles, de mortalité et d'infection.

Figure18. Clusters des cas confirmés et des décès (Source data : www.worldometres.info)

Description :

Nuage de points, un graphique qui regroupe les pays qui ont un ratio similaire (décès, confirmés) sous forme de "clusters".

Interprétation

On voit bien la formation de deux amas solides :

-Cluster 1 : Pologne, Iran, Colombie, Allemagne, Espagne, Argentine.

-Cluster 2 : Royaume-Uni, Russie, France, Italie.

Discussion

-Selon les statistiques de 2017, les 4 pays du 2ème cluster figurent parmi les 10 pays les plus visités. Avec :

Classement	Pays	Nombre de visiteurs(million)
1	France	86.9
3	Italie	58.3
4	Royaume-Uni	37.7
9	Russie	24.4

Tableau1. Nb des visiteurs par pays du cluster 2(Source data : www.banquemondiale.org)

-Tous les 4 pays du 2ème cluster tombent dans l'hémisphère Nord de la terre.

-Les 4 pays du 2ème cluster sont des pays européens.

-Les 4 pays du 2ème cluster ont la même démographie, vu que la population est constituée, principalement, des personnes âgées (39 ans < âge médian < 45 ans).

-Ils ont tous imposé un confinement en mars 2020.

-3 pays parmi les 4 pays du 2ème cluster presque la même densité de population mondiale.

Pays	Population (% mondial)
Russie	1.87
Royaume-Uni	0.87
France	0.84
Italie	0.78

Tableau2. La population par pays du cluster 2(Source data : www.wikipedia.org)

Conclusions

D'après cette analyse, on peut indiquer quelques facteurs de risques susceptibles à influencer le comportement de la pandémie. Parmi eux on cite :

- ✓ L'exposition internationale.
- ✓ La démographie.
- ✓ La race.
- ✓ L'âge de la population.
- ✓ La densité urbaine.

2.2. Tache 2

But

Identifier les gouvernements qui ont pris les bonnes mesures pour arrêter la propagation du virus.

Description

Barre horizontale : un graph permettant de visualiser, aisément, une tendance (top 15).

Discussion

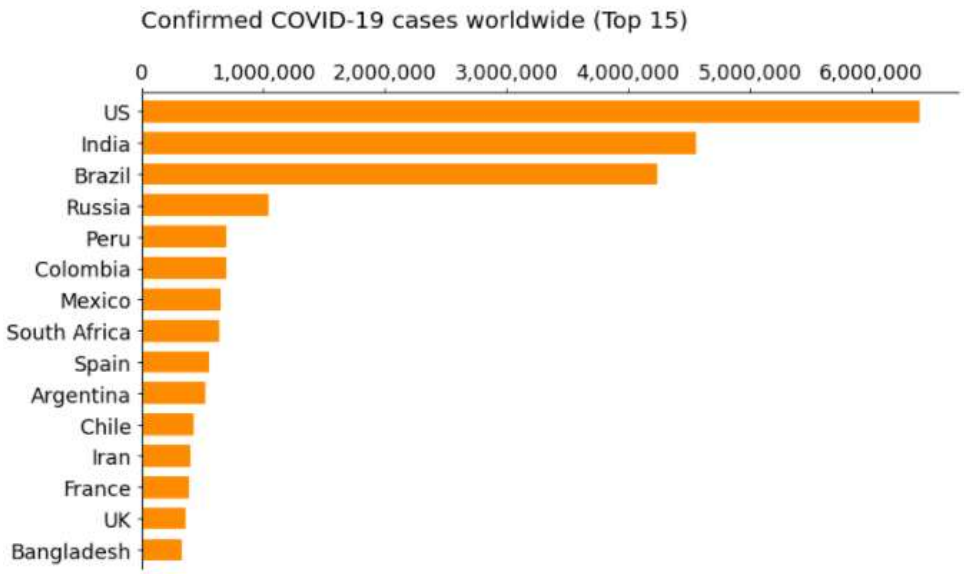


Figure19. Top 15 pays en termes des cas confirmés-sans normalisation (Source data : JHU site)

Ce classement nous dit-il quelque chose de significatif ?

Il est clair que les États-Unis ont beaucoup plus d'infections qu'au Mexique. Toutefois, cela ne signifie pas que les États-Unis sont plus touchés que le Mexique.

Pourquoi ?

Ça revient au fait que les Etats-Unis est un pays qui est beaucoup plus grand que le Mexique en termes de population.

Alors ?

Par conséquent, le nombre d'infections doit être normalisé à la population de chaque pays. Cela permettra une comparaison plus juste indépendamment de la taille du pays.

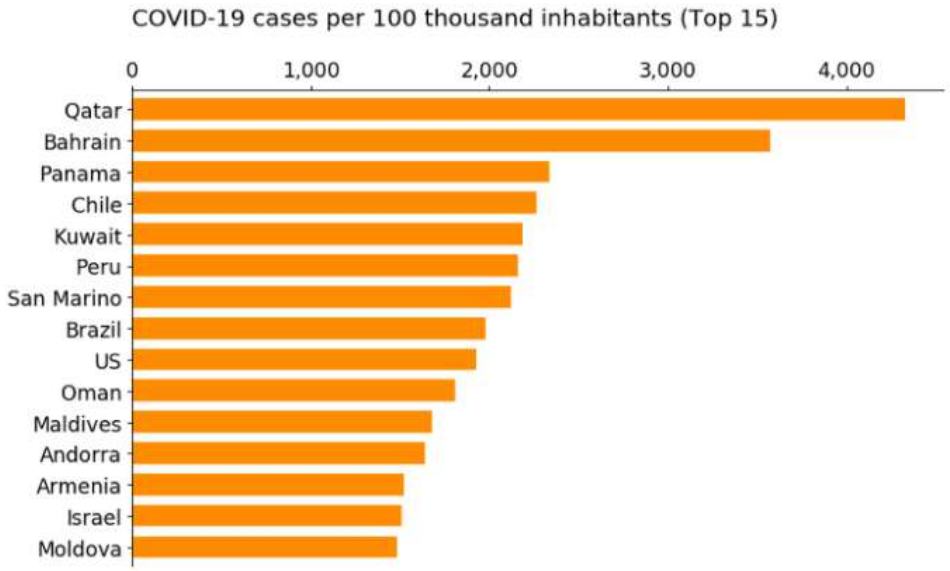


Figure20. Top 15 pays en termes des cas confirmés-avec normalisation (Source data : JHU site)

Après normalisation, on constate l'apparition d'autres pays au sommet, ce n'est plus les Etats-Unis qui le sont. Le Qatar, Bahreïn, le Panama, le Chili et Kuweit ont la plus forte densité d'infections.

On sait que la normalisation est nécessaire, la question qui se pose maintenant est :

Est-ce que la normalisation est suffisante pour vérifier la validité des résultats ?

La réponse c'est Non, car même si on ignore la différence de la transparence gouvernementale ainsi que la liberté de la presse, on ne peut pas négliger le fait que les pays ont eu des politiques de dépistage différente : plus de tests COVID-19 donnent plus de cas confirmés - et aucun test du tout impliquerait zéro cas. Nous avons donc besoin d'éliminer le facteur : la quantité de tests, de notre équation.

Et donc ?

On se basera sur le nombre normalisé de décès dus à la COVID-19, a la place des nombres des cas confirmés, vu que le nombre des décès n'est pas biaisé par le taux de dépistage.

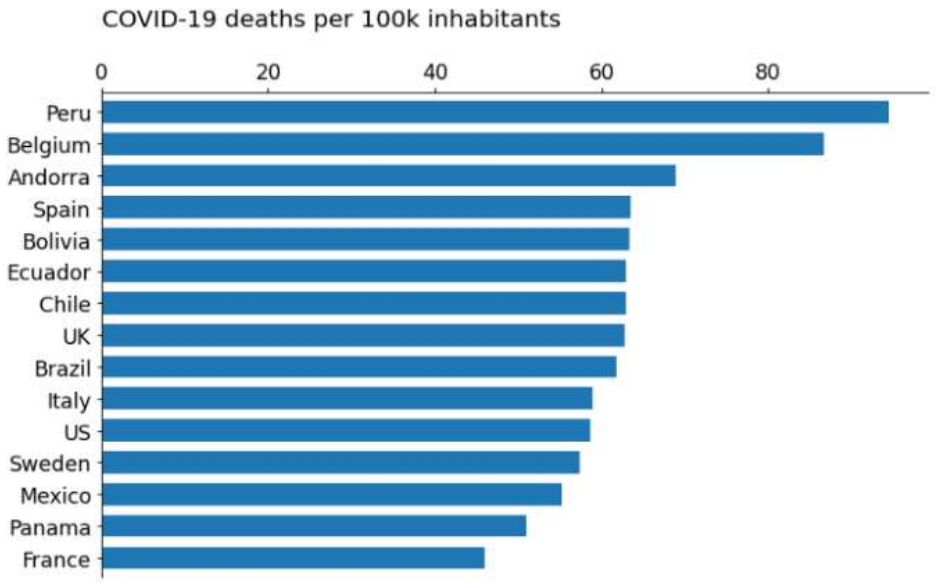


Figure21. Top 15 pays en termes des décès-avec normalisation (Source data : JHU site)

Le Pérou, la Belgique et Andorre sont les pays qui ont connu le plus grand nombre de décès (normalisé à sa population). En outre, 7 des 15 premiers pays ci-dessus appartiennent au continent européen.

L'évolution de cette quantité au fil du temps et voir ce que nous pouvons apprendre d'autre.

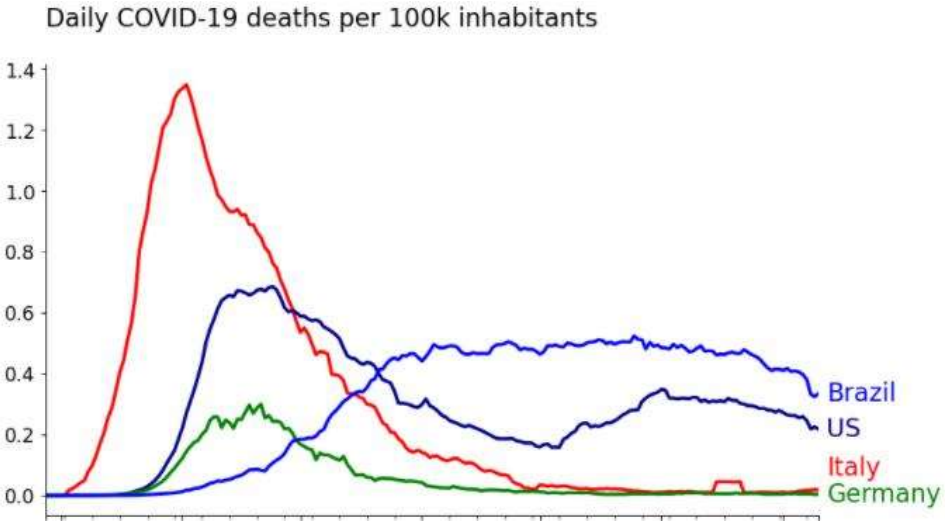


Figure22. Les décès quotidiens normaliser (Source data : JHU site).

En examinant les courbes de décès quotidiens ci-dessus, on peut extraire différents comportements épidémiologiques. Par exemple, la courbe rouge montre que l'Italie a subi de nombreux décès en avril. Pourtant, ils ont réussi à arrêter la propagation du virus. L'Italie a atteint cet objectif en fermant complètement la vie sociale pendant des mois.

L’Allemagne, la courbe verte, n’a pas été durement touchée par la première vague de COVID-19. L’État a réagi rapidement et a ralenti la vie sociale dès le début. En conséquence, le nombre de décès est revenu à presque zéro.

De l’autre côté, le Brésil et les États-Unis souffrent constamment de morts depuis avril jusqu’à aujourd’hui. Ils n’ont pas encore réussi à maîtriser le virus.

Conclusions

-Le fait de se concentrer sur le nombre normalisé de décès compense le biais du taux de dépistage.

2.3. Tache 3

But

Evaluer la politique du confinement.

Description

Chartographie Folium : une carte mondiale montrant la densité/la répartition d'une tendance.

Discussion

Appliquons la méthode d'agrégation "k-means" provenant de la technique d'apprentissage "clustering" aux courbes de décès quotidiens normalisées.

L'algorithme regroupe des pays ayant un comportement comparable. Le résultat est les trois clusters suivants :

-Le cluster vert : Pays qui sont légèrement touchés par la COVID-19.

-Le cluster rouge : Pays qui ont été très touchés lors de la première vague(avril/mai), mais qui ont contrôlé la situation

-Le cluster bleu : Pays en souffrance actuellement.

Figure23. Cluster des situation/confinement (Source data : Kaggle JHU Challenge)

Le cluster rouge est situé sur le continent européen.

Comme : l'Espagne, la France, le Royaume-Uni, l'Italie, la Belgique, la Suède, les Pays-Bas, l'Irlande et Andorre.

Le cluster bleu est concentré en Amérique du Nord et du Sud.

Comme : Etats-Unis, le Brésil, le Pérou, la Colombie, le Mexique, l'Argentine, le Chili, la Bolivie, l'Equateur et le Panama.

Conclusions

-Le fait de se concentrer sur le nombre normalisé de décès compense le biais du taux de dépistage.

-La vigilance dans la prise des décisions et surtout la rapidité d'introduction des restrictions sociales ont aidé la plupart des pays à maîtriser le virus.

MISE EN SITUATION

24% de la population mondiale a reçu, au moins, une dose du vaccin COVID-19.

3.19 billions de doses ont été administrer globalement, et 37.27 millions ont enregistrer chaque jour.

Alors, c'était quoi l'effet de la compagne de vaccination sur l'évolution de la crise ?

3. Analyse prédictive

3.1. Régression Linéaire Multiple

Objective

On souhaite savoir si, de façon générale, la combinaison (nombre de vaccinations, population, score des politique) a une influence sur le nombre des cas confirmés. Si oui, sous quelle forme cette influence peut être exprimée.

Le but est d'expliquer au mieux comment l'évolution de la crise varie en fonction de multiple facteurs, et éventuellement de prédire le nombre des cas confirmés à partir de ces facteurs.

Principe

On définit deux catégories des variables :

- ✓ La variable Y : variable cas confirmés ; c'est la variable à expliquer, appelée encore variable à régresser, variable réponse, variable dépendante (VD).
- ✓ Les variables X : variable (X1=nb de vaccinations, X2=population, X3=score de politique) ; c'est la variable explicative, appelée également régresseur, variable indépendante (VI).

Avec :

$Y = B\_1 * x\_1 + B\_2 * x\_2 + \ldots + B\_n * x\_n + A$

Et

PointErreur = (Réalité — Prédiction) ²

Dataset

Pour cette partie, on va créer notre propre base de données, en combinant ce qu'on appelera par la suite "score des caractéristiques", avec le nb des cas confirmés normalisé et le nombre des vaccinations normalisé, et ce pour chaque pays .

Le "score des caractéristiques " est désigner en calculant la somme des "indices des caractéristiques".

Un "indice des caractéristiques" représente un nombre attribué au chaque facteur de risque.

Les facteurs de risques pris en compte sont : population et le nombre des cas critiques.

Implémentation

La corrélation du graph obtenue est de 34.2%, ce qui n'est pas suffisant pour valider le model.

Conclusions

=>Ce model n'imitent pas le comportement de la pandémie.

=>Il n'a pas de relation linéaire entre les variables indépendantes (les entrées) et celle dépendante (la variable à prévoir).

Note

Différence entre régression polynomiale et régression linéaire multiple

La régression linéaire implémente son modèle par différentes caractéristiques déjà existant, or la régression polynomiale crée des caractéristiques en se basant sur un même facteur.

3.2. Arbre de Décision

Les arbres de décision sont un modèle très utilisé en apprentissage automatique supervisé. Dans ce travail, nous appliquons des idées de classification non supervisée (clustering) à la construction d'arbres semblables à des arbres de décision.

3.2.1. K-means Clustering

Principe :

Trouver des patterns dans les données, en regroupant les éléments ayant les mêmes caractéristiques dans des clusters.

Tactique :

Etape1 : Choisir un nombre arbitraire K d'éléments de la base de données, et leur associer des labels.

Etape 2 : Associer chaque élément de la BD au cluster qui le représente le mieux, en faisant appel à l'algorithme de la distance Euclidien.

Etape 3 : Recalculer les nouveaux représentant des clusters.

Etape 4 : Répéter l'étape 2 et l'étape 3 jusqu'à ce que les clusters soient inchangés.

Astuce :

Afin d'optimiser les résultats, il faut cibler la valeur de K. On peut avoir cette valeur, en faisant appelle à la méthode "Within-Cluster-Sum-of-Squares" -WCSS.

L'idée est de minimaliser la distance entre chaque point des data et le centre du cluster. Ce processus devra être itéré jusqu'à obtenir la valeur minimale des carrés de ces distances.

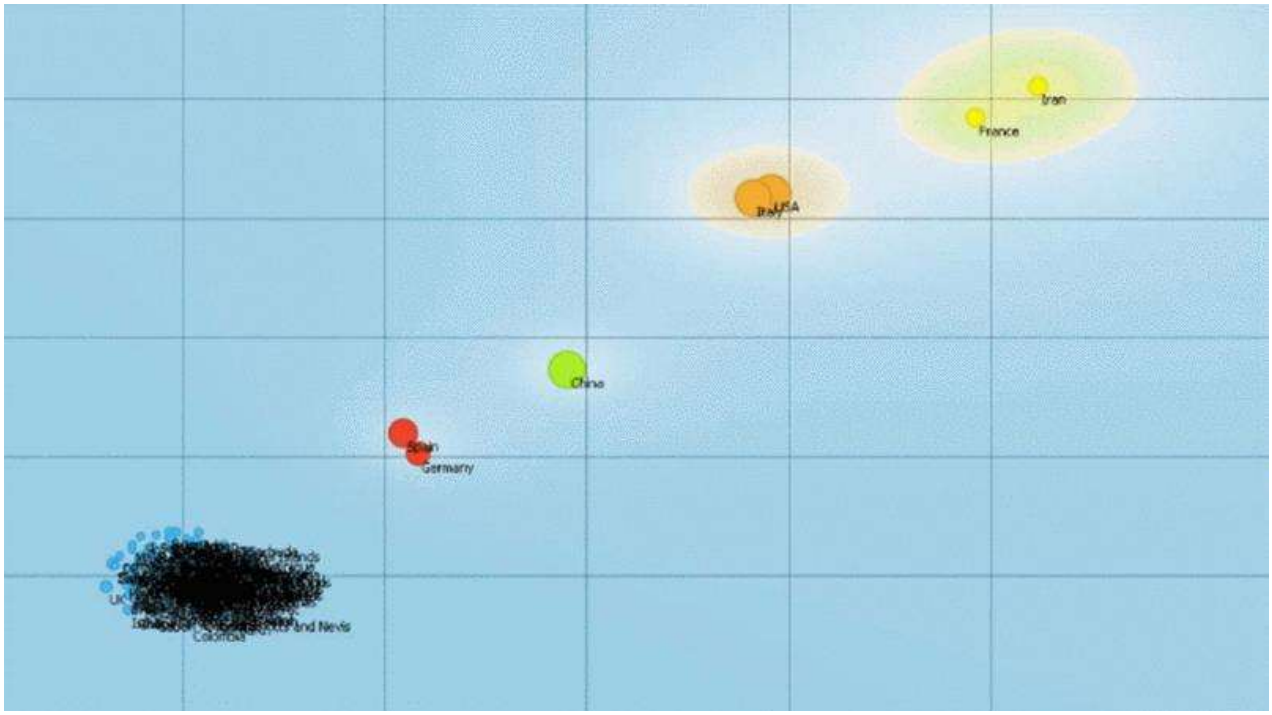


Figure24. Le clustering générale des cas confirmés de Covid-19 dans le monde.

Interprétation

Les bulles de même couleur indiquent les pays qui ont des caractéristiques similaires (nombre des cas confirmés, nombre des cas critiques, nombre des décès, nombre des vaccinations).D'où,

- ✓ La France et l'Iran forment le cluster jaune.
- ✓ L'Espagne et l'Allemagne forment le cluster rouge.
- ✓ L'Asie du sud forme le cluster bleu.

4. Tableau de bord

Il présente synthétiquement, des informations relatives à Covid-19, sous formes des visualisations. Son intérêt, est de faciliter le suivie de certains indicateurs jugés importante pour les prises de décisions.

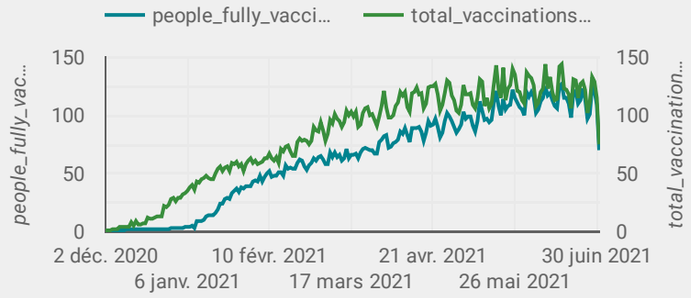
TABLEAU DE BORD POUR SARS-COV-2

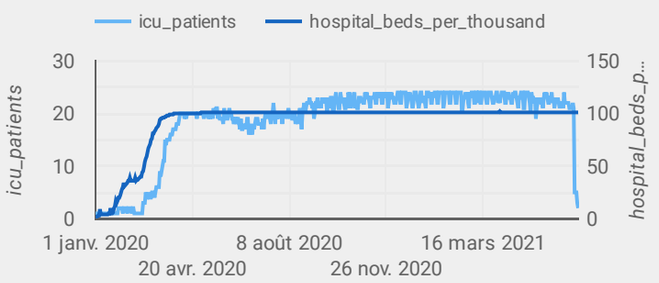
Dérnière mise a jour, le 06/07/2021

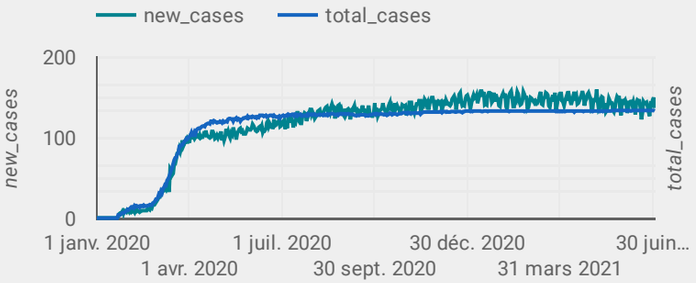
184,940,047 Cas Confirmés

4,000,847 Décès

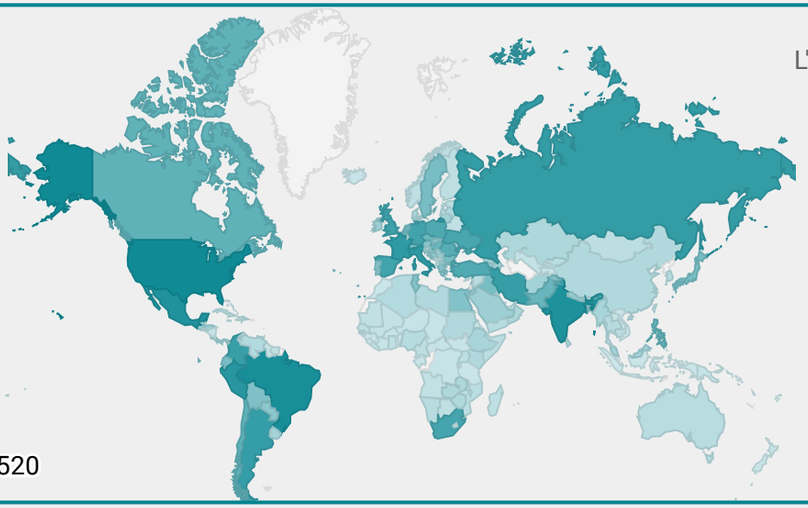
169,315,028 Guérison





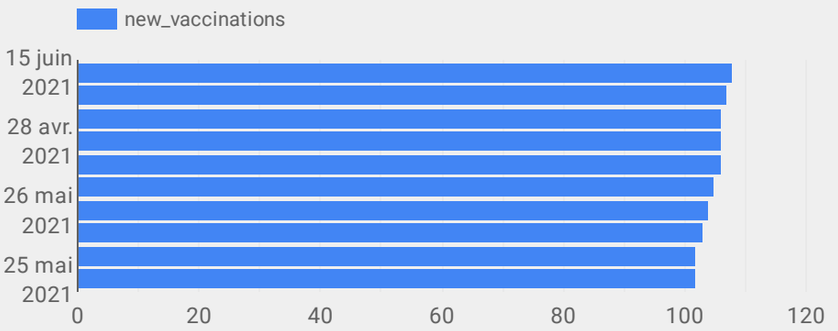


location



L'evolution des cas confirmés par jour, par pays

1 520



CONCLUSIONS ET PERSPECTIVES

En guise de conclusion, ce travail a mis en lumières l'application des big data analysis sur les données émergentes du virus Sars-Cov-2. Le choix de ce sujet dérivait de ces points forts, mais c'est lors du développement qu'on constatait ces points faibles.

Atouts

L'importance de ce travail, réside dans la convergence de plusieurs facteurs, principalement:

- ✓ Le besoin que connait le monde de ce type d'intervention harmonise avec ces objectives. (tendance +sujet critique + besoin).
- ✓ Le déploiement réel reflété par les scénarios étudiés.
- ✓ La flexibilité de l'application et l'inclusion des données structurées et non structurées.
- ✓ La documentation explicite pour les informaticiens et non informaticiens.

Limitations / Challenges

Lors de la mise en œuvre, quelques entraves se révèlent, et restreignent l'avancement de ce type de travail. On cite :

- ✓ L'accessibilité des ressources de données crédibles (informations inexactes).
- ✓ Le manque des bases de données pré-collectées, pré-nettoyées, prétraitées (ce qui prend du temps).
- ✓ La considération de tous les facteurs impliqués (interprétations trompeuses).
- ✓ Le choix des modèles qui imitant le mieux, le sujet d'étude (taux d'erreur élevé, faible corrélation).

Travaux futurs

Lors d'une future étude, il serait intéressant de Vérifier :

- ✓ La crédibilité des données (en utilisant l'analyse diagnostique) avant de les exploiter.
- ✓ D'évaluer la différence des résultats provenant de ce type d'analyse en jouant, uniquement, sur la quantité des data utilisées.

Annexe 1

Analyse Des Big Data VS Science Des Big Data

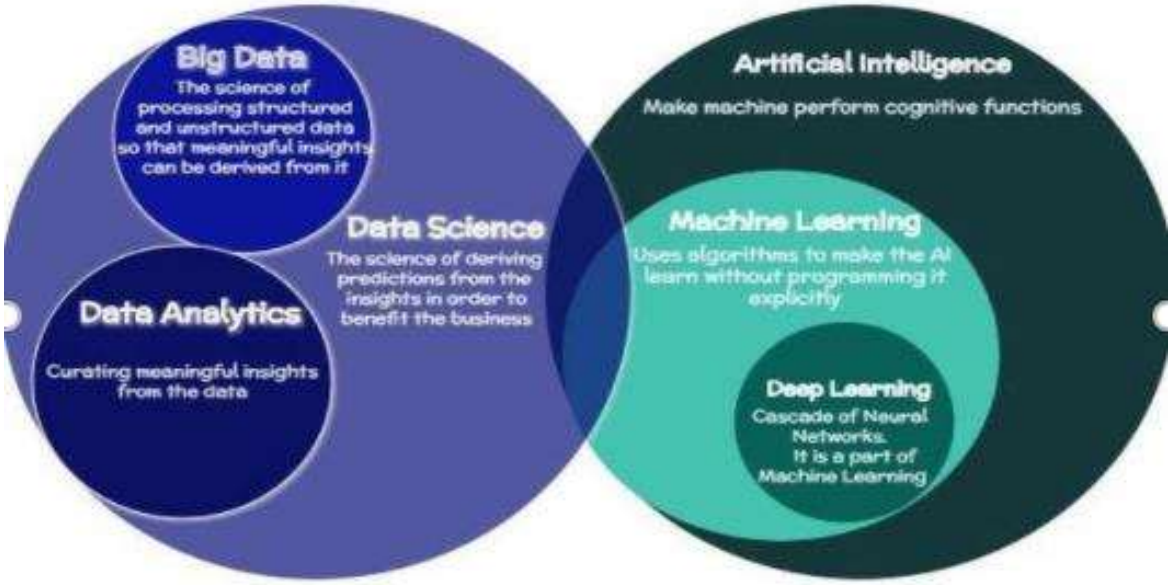


Figure26. Analyse des Big Data Vs Science des Big Data

Alors que l'analyse des données se concentre, principalement, sur la compréhension des ensembles de données et le glanage d'informations qui peuvent être transformées en actions, la science des données est centrée sur la création, le nettoyage et l'organisation des ces données. Les scientifiques des données créent et exploitent des algorithmes, des modèles statistiques et leurs propres analyses personnalisées pour collecter et façonner des données brutes en quelque chose qui peut être plus facilement compris.

« Depuis les premières étapes de la détermination de la qualité d'une source de données jusqu'à la détermination du succès d'un algorithme, la pensée critique est au cœur de chaque décision prise par les scientifiques des données et ceux qui travaillent avec eux », explique Dustin Tingley, professeur à Harvard, dans le cours en ligne Data Science Principles. « La science des données est une discipline qui repose sur une base de pensée critique. »

Les scientifiques des données jettent les bases de toutes les analyses effectuées par une organisation. Pour ce faire, ils exécutent des fonctions clés, notamment :

- ✓ Data wrangling : Le processus de nettoyage et d'organisation des données à utiliser plus facilement.

- ✓ Modélisation statistique : processus d'exécution de données à travers différents modèles, tels que les modèles de régression, de classification et d'agrégation, entre autres, pour identifier les relations entre les variables et obtenir des informations à partir des nombres.
- ✓ Programmation : processus d'écriture de programmes informatiques et d'algorithmes dans divers langages, tels que R, Python et SQL, qui peuvent être utilisés pour analyser des jeux de données volumineux plus efficacement que par le biais d'une analyse manuelle.

Le développement de vos compétences en science des données peut vous permettre de :

- ✓ Identifier et éviter les erreurs qui surviennent couramment lors de l'interprétation des ensembles de données, des métriques et des visualisations.
- ✓ Adoptez la prise de décision basée sur les données et assurez-vous que vos décisions d'entreprise sont étayées par des chiffres.
- ✓ Formulez des hypothèses, exécutez des expériences et rassemblez des preuves qui vous permettent de reconnaître les défis et les solutions de l'entreprise
- ✓ Comprendre la taille du marché, les tendances des acheteurs, la concurrence, les opportunités et les risques auxquels votre entreprise est confrontée

L'analyse d'entreprise peut être exploitée de différentes manières. Voici quelques exemples à prendre en compte :

- ✓ Budgétisation et prévision : en évaluant les données historiques sur les revenus, les ventes et les coûts d'une entreprise ainsi que ses objectifs de croissance future, un analyste peut identifier le budget et les investissements nécessaires pour faire de ces objectifs une réalité.
- ✓ Gestion des risques : En comprenant la probabilité que certains risques opérationnels se produisent et les coûts qui y sont associés, un analyste peut faire des recommandations rentables pour aider à les atténuer.
- ✓ Marketing et ventes : en comprenant les mesures clés, telles que le taux de conversion du prospect au client, un analyste marketing peut identifier le nombre de prospects que ses efforts doivent générer pour remplir le pipeline de ventes.
- ✓ Développement de produits (ou recherche et développement) : en comprenant comment les clients ont réagi aux fonctionnalités du produit dans le passé, un analyste peut aider à guider le développement, la conception et l'expérience utilisateur du produit à l'avenir.

Annexe 2

Code Pour L'Analyse Prédictive

1. Régression Linéaire Multiple

```
# Etape1 :  importer la dataset (en utilisant pandas)
import pandas as pddf = pd.read_csv ('Le chemin vers le fichier.csv')
print (df)

# Etape 2 : Implémenter le modèle MLR (en utilisant sklearn)
from sklearn import linear_model
X = df[['nbVaccinations','scoreCaractéristiques']]
 # On a ici 2 variables indépendante
Y = df['nbCasConfirmes']regr = linear_model.LinearRegression()
regr.fit(X, Y)
print('Intercept: \n', regr.intercept_)
print('Coefficients: \n', regr.coef_)
# Les output sont : "interset" et les "coefficients".Il vérifient l'équation suivante
# nbCasConfirmes = (Intercept) + (nbVaccinationCoef)*X1 + (scoreCarateristiquesCoef)*X2

# Etape 3 : Faire des Prediction (en utilisant sklearn)
NouveauVaccinations = valeurVaccinations
NouveauCaracteristiques = valeurCaracteristiques
print ('Le nombre des cas confirmés prédit est : \n', regr.predict([[NouveauVaccinations , NouveauCaracteristiques]]))
```

2. K-means Clustering

```
# ETAPE 1 : Importer les bibliothèques nécessaires
import pandas as pd

import numpy as np

import seaborn

import matplotlib.pyplot as plt

%matplotlib inline


# ETAPE 2 : Télecharger les Data
data = pd.read_csv('casConfirmesclusters.csv')Data


## Etape 3 :Déviser la base des données à une BD test et une BD train

setsample_train, sample_test = train_test_split(x, test_size=0.25)


# ETAPE 4 : Entrainer notre modèle

from sklearn.cluster import KMeans

model = KMeans(n_clusters=6)

model.fit(train_data[0])


# ETAPE 5 : Faire des prédictions

model.labels_# Prévoir à quel cluster appartient chaque point de data

model.cluster_centers_ # Le centre de chaque cluster


# ETAPE 6 : Evaluer la précision du modèle
```

REFERENCES

Webographie

https://www.sciencedirect.com (03/05/2021)
Saudi Journal of Biological Sciences (06/06/2021)
update39-covid-and-schools.pdf (who.int) (17/06/2021)
https://www.who.int (09/05/2021)
https://africacenter.org (30/05/2021)
https://www.worldometers.info (28/04/2021)

Documentation

les packages python.

Méthodes d'analyse fonctionnelle systématique pour la récupération et la documentation de la conception
L. Zehtaban et D. Roller

Articles/ Jornaux

Intitulé : Sustainability|Special Issue : Big Data Analytics amid COVID-19: Toward Sustainable Society
Auteurs: Dr. Ohbyung Kwon, Dr. Kyoung-yun "Joseph" Kim, Dr. Namgyu Kim, Dr. Namyeon Lee
Date de publication : 31 Mars 2021

Intitulé :COVID-19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions
Auteurs :Jie Sheng, Joseph Amankwah-Amoah, Zaheer Khan, Xiaojun Wang
Date de publication : 02 November 2020

Intitulé :Application of Big Data Technology for COVID-19 Prevention and Control in China: Lessons and Recommendations
Auteurs : Jun Wu, Jian Wang, Prof Dr,2 Stephen Nicholas, Prof Dr, Elizabeth Maitland, Prof Dr and Qiuyan Fan
Date de publication : 09 octobre 2020