Ministère de l'enseignement supérieur et de la recherche scientifique

Université de Mohamed el-Bachir el-Ibrahimi

Bordj Bou Arreridj

Faculté des mathématiques et de l'informatique

Département d'informatique

MEMOIRE DE FIN D'ETUDES

Réalisé en vue de l'obtention du diplôme de master en Informatique

Option : Ingénierie de L'informatique Décisionnelle

Thème

Clustering hiérarchique des données qui concernent l'épidémie Covid_19

Présenté par :

- Benfredj Nour Elhouda
- Guessabi Chahrazad

Soutenu publiquement le / / devant le jury composé de :

Président	***M.HadjBarakt***	Professeur à L'U.EI Bachir El Ibrahimi-BBA.
Examinateur	***Khelif Hakima***	MCB à L'U.EI Bachir El Ibrahimi-BBA.
Encadrant	***M. NailiMakhlouf***	MCB à L'U.EI Bachir El Ibrahimi- BBA

Année universitaire : 2020-2021.

Remerciements

Je remercie ALLAH tout puissant de m'avoir donné la force et la patience pour terminer ce travail de thèse.

Ce travail n'aurait pas pu voir le jour sans l'aide et l'encadrement du Dr Naili Makhlouf, je le remercie vivement pour la qualité de son encadrement, pour sa patience de ce mémoire. J'adresse mes remerciements les plus sincères aux membres du jury qui ont bien voulu examiner ce modeste travail.

Mes remerciements s'adressent également à tous nos professeurs pour leur générosité et la patience dont ils ont su faire preuve malgré leurs charges professionnelles.

Dédicace

Je dédie ce travail à mes très chers parents, que Dieu tout puissant les protège.

À mes chers frères Rabeh et sa chère épouse, Hamza et Islam

À ma tante maternelle Fatima et sa fille Kenza

À mon fiancé et sa famille

À tous mes enseignants

À ma chère amie Chahrazad

À tous ceux qui on contribué de près ou de loin pour que ce projet soit possible,

Je vous remercie.

Benfredj Nour Elhouda

Dédicace

Ce mémoire est dédié :

A mes très chers parents Saad et Houria

A mes frères et mes sœurs

A mes amiset mes copains

A ma très chère personne Benfredj Nour el Houda

Guessabi Chahrazad

Table des matières

Liste des figures

Liste des tableaux

Introduction Générale

Contexte du travail

Le COVID-19 a posé d'énormes problèmes de santé dans le monde entier en raison de son niveau élevé de contagion et de sa propagation géographique rapide. Le monde interconnecté a aidé à diffuser le virus à une telle vitesse, atteignant une couverture des pays qui a conduit l'Organisation mondiale de la santé (OMS) à déclarer COVID-19 comme pandémie au début de 2020. En seulement 12 mois, au 31 décembre 2020, il était 82,8 millions d'infections confirmées et 1,8 million de décès à travers le monde [1,2].

Les enseignements tirés des épidémies et pandémies précédentes, telles que la pandémie de grippe de 1918, le syndrome respiratoire aigu sévère (SRAS) en 2002-2003 ou le virus de la grippe H1N1 (grippe porcine) en 2008-2010, ont montré que les mesures de santé publique avaient une influence significative sur l'impact de la maladie, notamment en termes de mortalité globale. La quarantaine volontaire et obligatoire, l'interdiction des rassemblements de masse et des grands événements, la fermeture des écoles et des lieux de travail, et l'isolement des ménages/régions étaient quelques-unes des mesures appliquées par les gouvernements pour réduire la mortalité due aux maladies [1,5]. Les pandémies et les épidémies ont également montré que ces maladies peuvent avoir un impact important sur les économies [1,3]. Pour cette raison, les pays et régions doivent décider quelles mesures d'atténuation mettre en œuvre et quand les appliquer afin d'éviter d'atteindre des pics qui submergeraient les services de santé mais aussi définir des mesures agissant comme modérateurs des effets négatifs de la maladie sur l'économie, un équilibre ce n'est pas facile à atteindre [1]. L'imposition des restrictions et des blocages susmentionnés a généré une lourde charge de conséquences économiques dans de nombreux pays, déclenchant une augmentation spectaculaire des taux de chômage ainsi que des fermetures d'entreprises. Cela a été suivi de répercussions sociales, renforçant ainsi l'exigence d'une compréhension de l'évolution de cette pandémie, notamment en termes d'éventuels profils de pays différents évoluant à travers la planète [2-5].

Dans une perspective de diagnostic global, l'utilisation de méthodes et de techniques de science des données et d'apprentissage automatique constitue une opportunité pour la recherche d'atteindre cet objectif. Plus précisément, les données ouvertes partagées par des organisations réputées permettent un accès presque en temps réel à des données détaillées à travers le monde et permettent d'effectuer une collecte et une analyse de données impératives afin de progresser avec la compréhension et la prise de décision nécessaires [4, 6,7].

De plus, les techniques de science des données peuvent être des outils puissants pour comprendre les phénomènes en jeu, permettant le développement de politiques de soutien et facilitant les décisions qui peuvent optimiser les ressources, atteindre un bon équilibre entre la santé et l'économie, et finalement aussi sauver des vies.

Problématiques

La problématique principale traitée dans ce mémoire est celle d'analyser et divisées les pays européens en groupes en fonction à la fois de leurs cas et du nombre de décès. En termes de méthodologies de recherche, nous avons utilisé des techniques de science des données, telles que des visualisations de données et des tests statistiques, pour faire ce que l'on appelle souvent dans l'exploration de données. La caractérisation et la description des données, c'est-à-dire résumer les données par groupe et comparer les groupes [9]. Nous avons également utilisé des techniques d'apprentissage automatique non supervisées, notamment pour regrouper les pays en fonction de leur similitude en termes de cas de COVID-19 et de profils temporels de décès. De plus, nous avons effectué une analyse préliminaire des relations entre les cas et les décès causés par COVID-19 et les indicateurs de développement de certains pays.

Objectifs

Les résultats de l'analyse de relation entre les cas et les décès causés par COVID-19 et les mesures de développement des pays, peuvent être utilisés par les décideurs politiques pour prendre de meilleures décisions pour contrôler la pandémie. Cette analyse peut aider à mettre en évidence les politiques publiques les plus et les moins importantes pour minimiser le taux de mortalité COVID-19 d'un pays.

Et pour contenir la pandémie, les pays d'un cluster coopèrent, partagent des informations et apprennent des erreurs ou des stratégies (selon le cas) des pays d'autres clusters. Entre autres avantages, cela peut empêcher les pays du groupe de cas faiblement confirmés de progresser vers le groupe de cas hautement confirmés.

Contribution

Notre contribution se situe à plusieurs niveaux :

- Premièrement, nous donnons une présentation globale sur les Épidémies et leur propagation, notamment pour ce qui est Covid et tout ce qui s'y rapporte.
- Deuxièmement, Nous expliquons la fouille de données. En particulier, la présentation de La méthode et techniques utilisée dans notre étude.
- Enfin, Nous avons appliqué ces techniques d'apprentissage automatique non supervisées sur la base de données choisie afin de pouvoir regrouper les pays en fonction de leur similitude en termes de cas de COVID-19 et de profils temporels de décès.

Organisation du mémoire

Le reste de la thèse est structurée en trois chapitres, et elle se termine par une conclusion générale dans laquelle nous avons présenté le bilan et les perspectives de ce travail.

D'abord, le chapitre 1, intitulé « les épidémies », Dans ce chapitre, nous donnons un aperçu sur les épidémies surtout le coronavirus et les efforts et techniques déployées pour surveiller et suivre la propagation de cette épidémie. Nous commençons par présenter aperçu de l'épidémiologie et coronavirus. Ensuite, nous basculons vers veille médicale. Enfin, nous abordons les efforts pour surveiller les données nécessaires à la flambée d'épidémies.

Le chapitre 2, intitulé « Fouille de données », Dans ce chapitre nous voulons reconnaitre les différentes techniques de la fouille de données (data mining), afin d'avoir un aperçu complet sur eux, pour identifier les techniques appropriées pour l'utiliser dans la résolution des problèmes trouvé dans l'introduction général.

Le chapitre 3 consacrés à la description de l'approche proposée et la manière de l'utiliser. Dans ce chapitre, nous appliquons l'analyse de cluster, l'une des techniques d'exploration de données pour décrire la prévalence du COVID-19 dans L'Europe et identifiant les pays les plus touchés en des cas et des morts, en utilisant le langage de programmation Python.

Nous terminons cette thèse par une conclusion générale qui présente une synthèse des travaux réalisés et nous citons également quelques perspectives pour des travaux futurs.

Chapitre I : Les épidémies

I.1 Introduction

De très nombreuses maladies infectieuses, qu'elles soient d'origine virale, bactérienne ou parasitaire, ont toujours accompagné l'espèce humaine. Elles ont, en particulier dans le cadre des nombreuses épidémies, ainsi influé, et dans certains cas, façonné de façon très importante la dynamique évolutive des populations humaines des derniers millénaires [10].

Dernièrement, la complexité et le volume des connaissances médicales évoluent rapidement, telles que les données médicales, socio-économiques, démographiques et environnementales, qui obligent l'ensemble des responsables en santé publique de gérer toujours plus d'informations pour suivre et surveille la survenue d'une maladie épidémique. Ils ont besoin d'outils et de modèles et des techniques dans la veille informationnelle concernant la propagation de l'épidémie.

Dans ce chapitre, nous donnons un aperçu sur les épidémies surtout le coronavirus et les efforts et techniques déployées pour surveiller et suivre la propagation de cette épidémie. Nous commençons par présenter aperçu de l'épidémiologie et coronavirus. Ensuite, nous basculons vers veille médicale. Enfin, nous abordons les efforts pour surveiller les données nécessaires à la flambée d'épidémies.

I.2 Epidémies

I.2.1 Définition

Une épidémie est la propagation rapide d'une maladie à un grand nombre de personnes dans une population donnée sur une courte période de temps. Par exemple, dans les infections à méningocoques, un taux d'attaque supérieur à 15 cas pour 100 000 personnes pendant deux semaines consécutives est considéré comme une épidémie [11].

Selon son étymologie grecque (Demos signifiant peuple), ce mot s'applique initialement aux maladies touchant les humains si la maladie s'étend rapidement à une part importante de la planète, on parle alors de pandémie [12].

Figure I. 1.Large propagation des épidémies dans le monde [13].

I.2.2 Types d'épidémies

Les épidémies peuvent être classées selon leur mode de propagation dans une population [14] :

I.2.2.1 Épidémies de source commune

Dans certaines éclosions de source commune, les cas-patients peuvent avoir été exposés sur une période de plusieurs jours, semaines ou plus. Dans une épidémie continue de source commune, la gamme des expositions et la gamme des périodes d'incubation ont tendance à aplatir et à élargir.

Par exemple, la tragédie du gaz de Bhopal en Inde et la maladie de Minamata au Japon résultant de la consommation de poisson contenant une forte concentration de méthyl mercure.

Source commune ponctuelle

Les membres de la population à risque sont exposés à l'agent causal sur une courte période le nombre de cas augmente rapidement, atteint un sommet, puis diminue graduellement. Ceci se traduit par une courbe asymétrique dans laquelle le mode est décalé vers la gauche du centre. La courbe épidémique dans ce type de transmission suit généralement une distribution log-normale.

Les principales caractéristiques d'une épidémie « ponctuelle » sont :

1. La courbe épidémique monte et descend rapidement, sans vagues secondaires
2. L'épidémie à tendance à être explosive, il y a un regroupement de cas dans un intervalle de temps étroit, et plus important encore, tous les cas se développent au cours d'une période d'incubation de la maladie.

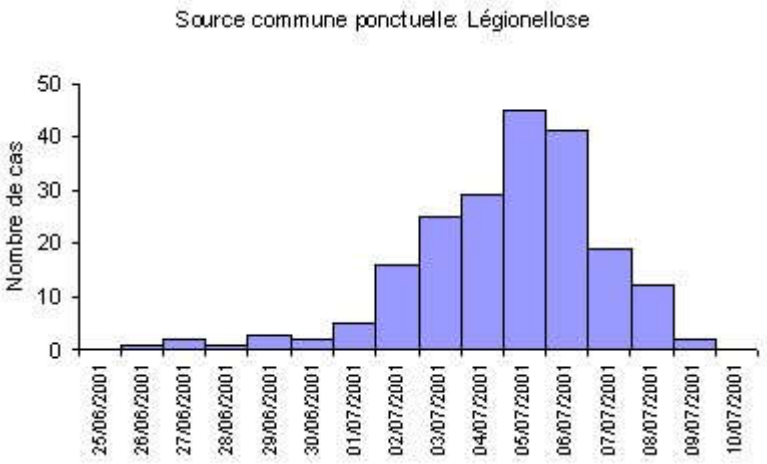

Figure I. 2.Courbe épidémique source commune ponctuelle [14].

Source commune continue

L'exposition est alors prolongée (par convention elle excède la durée d'incubation de la maladie), l'ascension rapide du nombre de cas est suivie par une phase en plateau, la diminution progressive du nombre de cas ne survient que lorsque l'exposition cesse.

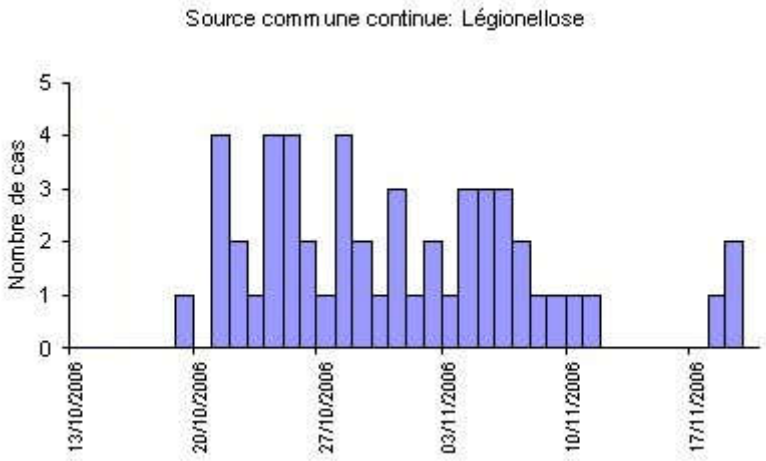

Figure I. 3.Courbe épidémique source commune continue [14].

Source commune intermittente

La courbe montre un profil irrégulier les cas, groupés ou non, surviennent à intervalles variables ce qui reflète des expositions répétées.

Il est difficile dans ce dernier type de profil de déterminer si la source est commune avec émission irrégulière ou si les sources sont multiples et variées. L'écart de temps entre les cas pourrait aussi suggérer une transmission de personne à personne séparée d'une période d'incubation, mais les pics successifs n'augmentent pas de taille et ne fusionnent pas, comme ce serait le cas lors d'une propagation infectieuse où une personne atteinte en infecte plusieurs.

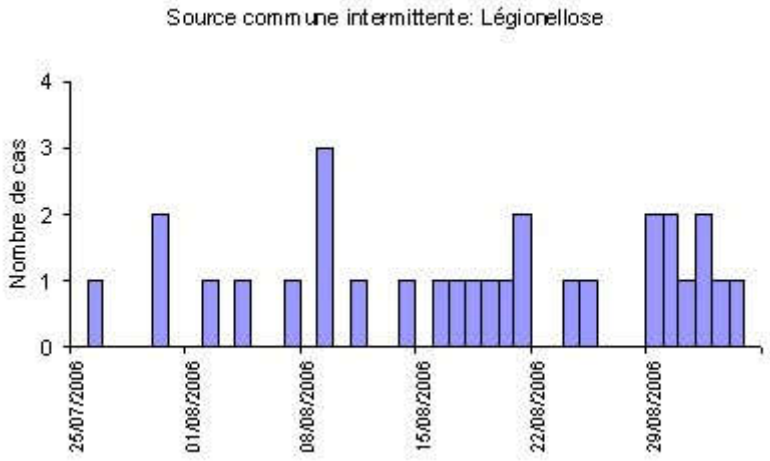

Figure I. 4.Courbe épidémique source commune intermittente [14].

I.2.2.2 Épidémies par propagation

Épidémie par propagation « propagatedoutbreak » ou transmission disséminée Elle est caractérisée par une courbe présentant des pics multiples aux sommets de plusieurs vagues d'amplitude croissante ; la décroissance est lente. Ce type de profil se retrouve dans les maladies à transmission interhumaine l'épidémie n'est propagée de personne à personne mais également dans les maladies à transmission vectorielle [15] [16].

Les modes de transmission peuvent aussi être associés ou se succéder dans une même épidémie, par exemple source commune ponctuelle associée ou suivie d'une transmission interhumaine

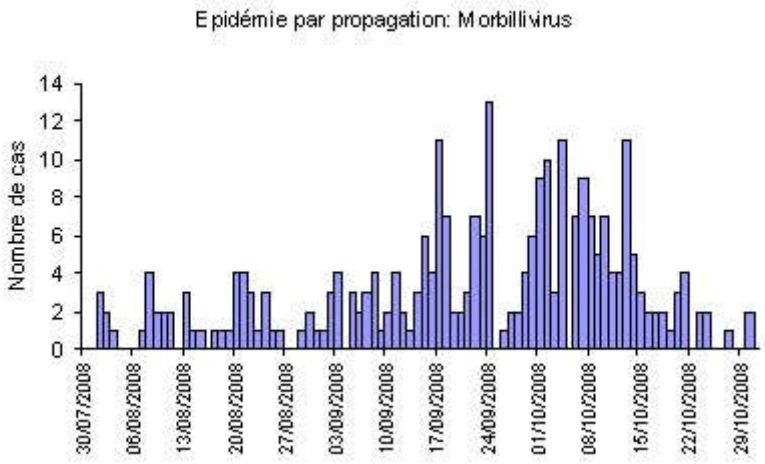

Figure I. 5. Courbe épidémique : épidémie par propagation [15].

I.2.2.3 Épidémies mixtes

Certaines épidémies ont des caractéristiques à la fois d'épidémies de source commune et d'épidémies propagées. Le schéma d'une éclosion de source commune suivie d'une propagation secondaire de personne à personne n'est pas rare. C'est ce qu'on appelle les épidémies mixtes.

I.2.3 Transmission d'épidémies

Les conditions qui régissent le déclenchement des épidémies incluent les approvisionnements alimentaires infectés tels que l'eau potable contaminée et la migration de populations de certains animaux, tels que les rats ou les moustiques, qui peuvent agir comme vecteurs de maladies [17].

- **Transmission aéroportée** : La transmission aéroportée est la propagation d'une infection par des noyaux de gouttelettes ou de la poussière dans l'air. Sans l'intervention de vents ou de courants d'air, la distance sur laquelle l'infection aéroportée à lieu est courte, disons 10 à 20 pieds.
- **Transmission arthropode** : La transmission arthropode à lieu par un insecte, soit mécaniquement par le biais d'une trompe ou d'un pied contaminé, soit biologiquement lorsqu'il y a croissance ou réplication d'un organisme dans l'arthropode.

- **Transmission biologique** : impliquant un processus biologique, par exemple, le passage d'un stade de développement de l'agent infectieux chez un hôte intermédiaire. Opposé à la transmission mécanique.
- **Transmission par contact :** L'agent pathogène se transmet directement par morsure, succion, mastication ou indirectement par inhalation de gouttelettes, consommation d'eau contaminée, déplacement dans des véhicules contaminés.
- **Transmission cyclopropagative:** L'agent subit à la fois un développement et une multiplication dans le véhicule émetteur.
- **Transmission développementale :** L'agent subit un certain développement dans le véhicule de transmission.
- **Transmission fécale-orale :** L'agent infectieux est excrété par l'hôte infecté dans les fèces et acquis par l'hôte sensible par ingestion de matériel contaminé.
- **Transmission horizontale :** Diffusion latérale à d'autres dans le même groupe et en même temps.
- **Transmission propagative:** L'agent se multiplie dans le véhicule de transmission.
- **Transmission verticale :** D'une génération à l'autre, peut-être par voie transovarienne ou par infection intra-utérine du fœtus. Certains rétrovirus sont transmis dans la lignée germinale, c'est-à-dire que leur matériel génétique est intégré dans l'ADN de l'ovule ou du sperme

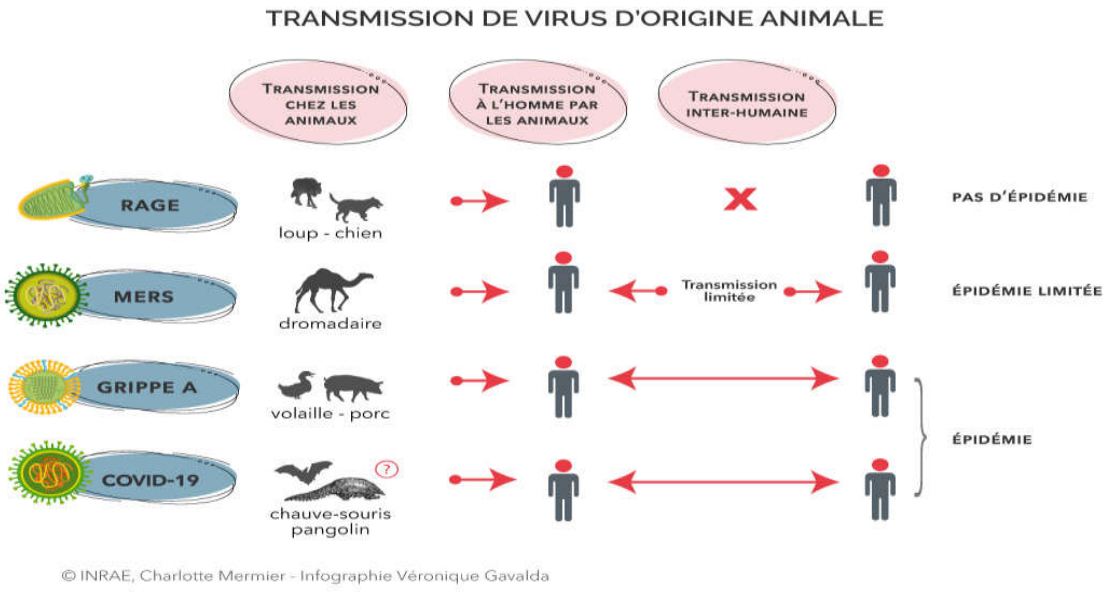

Figure I. 6. Image montrant un exemple de propagation de virus animaux [18].

I.2.4 Les pandémies les plus mortelles dans l'Histoire

La pandémie de la peste noire

La pandémie de la peste noire, causée par la bactérie Yersinia pestis a sévi en Asie, au Moyen-Orient, au Maghreb et en Europe. Elle se déclare pour la première fois en 1334 dans la province de Hubei en Chine. De 1347 à 1352, L'image montre les bactéries responsables [19].

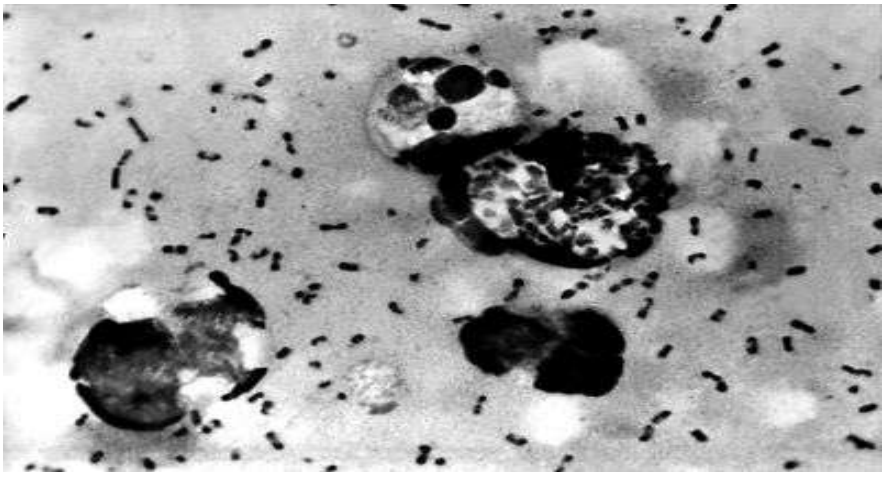

Figure I. 7. La bactérie Yersinia pestis, agent de la peste [19].

La peste noire fait 25 millions de victimes en Europe, ce qui correspond environ à la moitié de la population européenne à l'époque et 25 millions de morts dans le reste du monde, notamment en Chine, en Inde, en Egypte, en Perse et en Syrie. La peste noire est principalement transmise par les poux, les piqûres de puces et les rats.

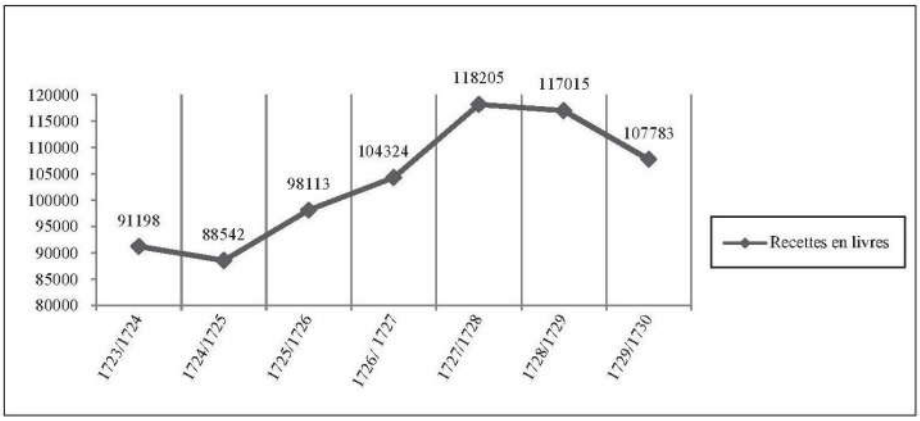

Figure I. 8. Peste de 1720-1721 : un choc financier et économique [19].

La grippe espagnole

La grippe espagnole, maladie causée par une souche de type A H1N1 particulièrement violente, est une pandémie qui a contaminé plus d'un tiers de la population mondiale entre 1918 et 1919. Elle aurait tué, selon l'Institut Pasteur, plus de 50 millions de personnes, soit 5 fois plus que lors des batailles de la Première Guerre mondiale. Très peu de régions dans le monde ont échappé à cette pandémie [20].

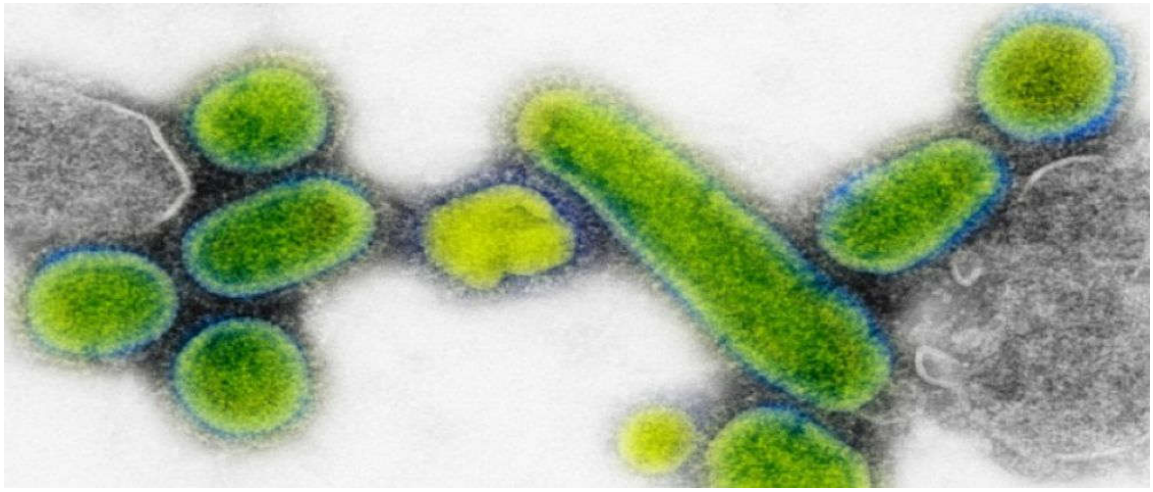

Figure I. 9. Virus de la grippe espagnole [20].

I.3 Covid19

I.3.1 Définition

Les premiers cas humains de COVID-19, la maladie causée par le nouveau coronavirus causant COVID-19, par la suite nommé SARS-CoV-2 ont été signalés pour la première fois par des responsables de la ville de Wuhan en Chine en décembre 2019.

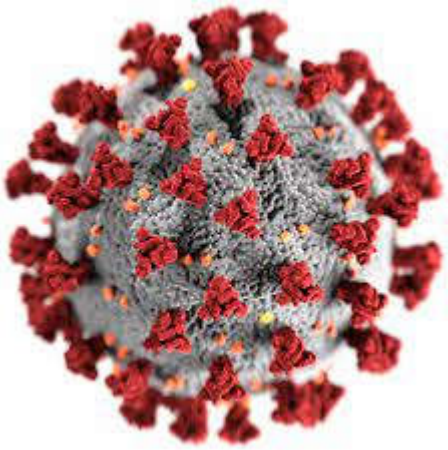

Figure I. 10. Corona virus [22].

Rétrospective des enquêtes menées par les autorités chinoises ont identifié des cas humains avec apparition de symptômes début décembre 2019. Alors que certains des premiers cas connus avaient un lien avec un marché alimentaire de gros à Wuhan, d'autres non. Beaucoup des premiers patients étaient soit des propriétaires de stands, des employés du marché ou des visiteurs réguliers de ce marché. Échantillons environnementaux prises sur ce marché en décembre 2019 ont été testées positives pour le SRAS-CoV-2, suggérant en outre que le marché en La ville de Wuhan a été à l'origine de cette épidémie ou a joué un rôle dans l'amplification initiale de l'épidémie [23].

I.3.2 Les symptômes du COVID-19

Les symptômes du COVID-19 sont variables, allant de symptômes bénins à une maladie grave. Les symptômes courants comprennent des maux de tête, une perte d'odorat et de goût, une congestion nasale et un écoulement nasal, une toux, des douleurs musculaires, des maux de gorge, de la fièvre, de la diarrhée et des difficultés respiratoires. Les personnes atteintes de la même infection peuvent présenter des symptômes différents et leurs symptômes peuvent changer avec le temps. Trois groupes communs de symptômes

ont été identifiés : un groupe de symptômes respiratoires avec toux, expectorations, essoufflement et fièvre, un groupe de symptômes musculo-squelettiques avec des douleurs musculaires et articulaires, des maux de tête et de la fatigue, un groupe de symptômes digestifs avec des douleurs abdominales, des vomissements et de la diarrhée. Chez les personnes sans troubles antérieurs des oreilles, du nez et de la gorge, la perte du goût combinée à la perte de l'odorat est associée au COVID-19 [24].

Figure I. 11. Symptômes du Corona virus [24].

Parmi les personnes qui présentent des symptômes, 81 % ne développent que des symptômes légers à modérés et peuvent devenir amnésiques après la guérison (jusqu'à une pneumonie légère), tandis que 14 % développent des symptômes graves et 5 % des patients souffrent de symptômes critiques (insuffisance respiratoire, choc ou dysfonctionnement de plusieurs organes). Au moins un tiers des personnes infectées par le virus ne développent à aucun moment des symptômes visibles. Ces porteurs asymptomatiques ont tendance à ne pas se faire tester et peuvent propager la maladie. D'autres personnes infectées développeront des symptômes plus tard, appelés "pré-symptomatiques", ou présenteront des symptômes très légers et pourront également propager le virus.

Comme c'est souvent le cas avec les infections, il y a un délai entre le moment où une personne est infectée pour la première fois et l'apparition des premiers symptômes. Le délai médian pour COVID-19 est de quatre à cinq jours. La plupart des personnes symptomatiques présentent des symptômes dans les deux à sept jours suivant l'exposition, et presque toutes présenteront au moins un symptôme dans les 12 jours [25].

I.3.3Transmission du Corona virus

La maladie est principalement transmise par voie respiratoire lorsque les personnes inhalent des gouttelettes et des particules que les personnes infectées libèrent lorsqu'elles respirent, parlent, toussent, éternuent ou chantent. Les personnes infectées sont plus susceptibles de transmettre le COVID-19 lorsqu'elles sont physiquement proches. Cependant, l'infection peut se produire sur de plus longues distances, en particulier à l'intérieur [17].

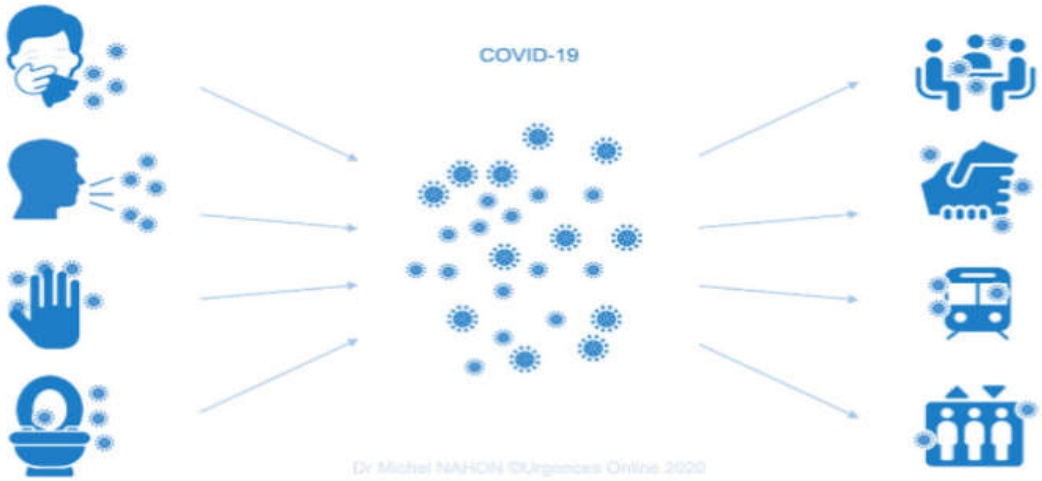

Figure I. 12. Principaux modes de transmission [26].

L'infectiosité commence dès trois jours avant l'apparition des symptômes, et les gens sont plus contagieux juste avant et pendant l'apparition des symptômes. Il diminue après la première semaine, mais les personnes infectées restent contagieuses jusqu'à 20 jours. Les gens peuvent propager la maladie même s'ils sont asymptomatiques.

Les particules infectieuses varient en taille, des aérosols qui restent en suspension dans l'air pendant de longues périodes à des gouttelettes plus grosses qui restent en suspension dans l'air ou tombent au sol. Divers groupes utilisent des termes tels que « aéroporté » et « gouttelette » à la fois de manière technique et générale, ce qui crée une confusion autour de la terminologie. De plus, la recherche COVID-19 a redéfini la compréhension traditionnelle de la façon dont les virus respiratoires se transmettent. Les plus grosses gouttelettes de liquide respiratoire ne voyagent pas loin et peuvent être inhalées ou atterrir sur les muqueuses des yeux, du nez ou de la bouche pour être infectées. Les aérosols sont les plus concentrés lorsque les personnes sont à proximité, ce qui facilite la transmission virale lorsque les personnes sont physiquement proches, mais la transmission par voie aérienne peut se produire à de plus

longues distances, principalement dans des endroits mal ventilés dans ces conditions, de petites particules peuvent rester en suspension dans l'air pendant des minutes ou des heures [27].

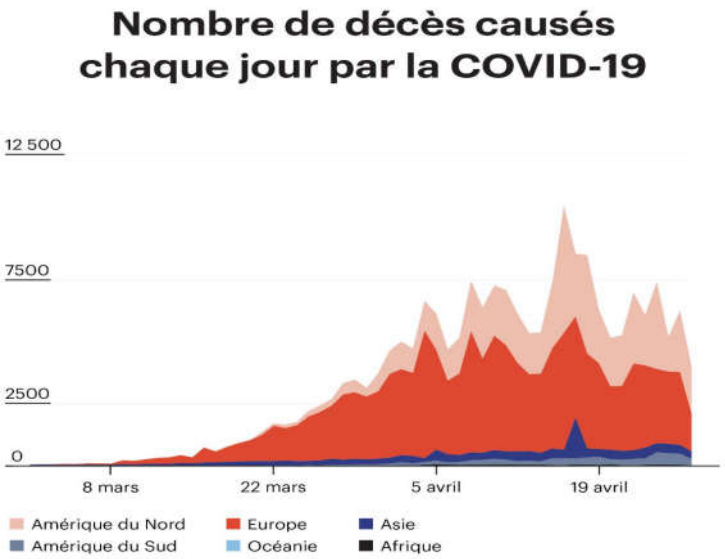

Figure I. 13. Image montrant Combien de décès sont causés par le coronavirus dans le monde [27].

Si le SRAS-CoV-2 est originaire d'Asie, ce continent s'en tire relativement mieux que l'Europe et l'Amérique du Nord. (Le bref pic dans le nombre de morts déclarés en Asie à la mi-avril est dû à la déclaration rétroactive par les autorités chinoises de 1290 morts à Wuhan.) En Amérique du Sud, le nombre de décès se maintient à un niveau stable et relativement peu élevé depuis plusieurs semaines. En Afrique et en Océanie, les courbes sont assez basses pour être invisibles à l'œil nu sur cette figure. Ainsi donc, à moins que le bilan officiel de certains pays soit sévèrement erroné, la COVID-19 est pour l'instant un problème affligeant essentiellement l'Occident.

I.3.4 Les mesures préventives

Les mesures préventives pour réduire les risques d'infection comprennent se faire vacciner, rester à la maison, porter un masque en public, éviter les endroits bondés, se tenir à distance des autres, ventiler les espaces intérieurs, gérer les durées d'exposition potentielles, se laver les mains à l'eau et au savon souvent et pendant au moins vingt secondes, en pratiquant une bonne hygiène respiratoire et en évitant de se toucher les yeux, le nez ou la bouche avec des mains non lavées [28] [29].

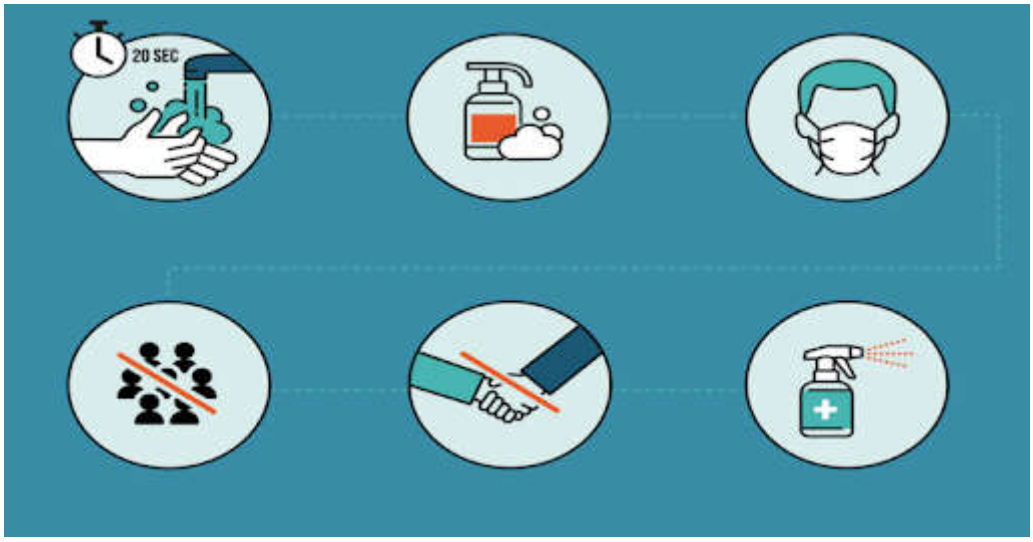

Figure I. 14.Prévention du Corona virus [28].

- **Vaccin**

Un vaccin COVID-19 est un vaccin destiné à fournir une immunité acquise contre le coronavirus 2 du syndrome respiratoire aigu sévère (SRAS-CoV-2). Les vaccins COVID-19 sont largement célébrés pour leur rôle dans la réduction de la propagation, causés par COVID-19.

De nombreux pays ont mis en œuvre des plans de distribution échelonnés qui donnent la priorité aux personnes les plus à risque de complications, comme les personnes âgées, et celles à haut risque d'exposition et de transmission, comme les professionnels de la santé. L'utilisation provisoire d'une dose unique est envisagée pour étendre la vaccination au plus grand nombre de personnes possible jusqu'à ce que la disponibilité du vaccin s'améliore [30].

- **Masques faciaux et hygiène respiratoire**

L'OMS et le CDC américain recommandent aux individus de porter des couvre-visages non médicaux dans les lieux publics où il existe un risque accru de transmission et où les mesures de distanciation sociale sont difficiles à maintenir. Cette recommandation vise à réduire la propagation de la maladie chez les personnes asymptomatiques et pré- symptomatiques et est complémentaire aux mesures préventives établies telles que la distanciation sociale.

- **Éviter les espaces intérieurs bondés et la ventilation**

Le CDC recommande d'éviter les espaces intérieurs bondés. À l'intérieur, l'augmentation du taux de renouvellement d'air, la diminution de la recirculation de l'air et l'augmentation de l'utilisation de l'air extérieur peuvent réduire la transmission. L'OMS recommande une ventilation et une filtration de l'air dans les espaces publics pour aider à éliminer les aérosols infectieux.

- **Lavage des mains et hygiène**

Une hygiène minutieuse des mains après une toux ou un éternuement est requis. L'OMS recommande également de se laver les mains souvent à l'eau et au savon pendant au moins vingt secondes, surtout après être allé aux toilettes ou lorsque les mains sont visiblement sales, avant de manger et après s'être mouché. Lorsque le savon et l'eau ne sont pas disponibles, le CDC recommande d'utiliser un désinfectant pour les mains à base d'alcool avec au moins 60% d'alcool.

- **Distanciation sociale**

La distanciation sociale (également connue sous le nom de distanciation physique) comprend des actions de contrôle des infections destinées à ralentir la propagation de la maladie en minimisant les contacts étroits entre les individus. Les méthodes incluent les quarantaines ; restrictions de voyage ; et la fermeture d'écoles, de lieux de travail, de stades, de théâtres ou de centres commerciaux. Les individus peuvent appliquer des méthodes de distanciation sociale en restant à la maison, en limitant les déplacements, en évitant les zones surpeuplées, en utilisant des salutations sans contact et en se distanciant physiquement des autres. De nombreux gouvernements imposent ou recommandent désormais une distanciation sociale dans les régions touchées par l'épidémie.

I.4 Conclusion

Le rôle des médias face aux épidémies et maladies infectieuses : l'épidémie de coronavirus comme modèle, et elle visait à démontrer l'importance de l'information sanitaire et son rôle dans la sensibilisation aux moyens de prévenir les épidémies et les maladies infectieuses, identifier les rumeurs qui ont accompagné l'émergence de l'épidémie de virus Corona et connaissent leurs tendances, et ont abordé les points de vue d'un certain nombre de spécialistes et de chercheurs. Connaissance des méthodes de prévention de l'épidémie, et l'étude a utilisé une approche exploratoire ou exploratoire pour mener l'étude.

Chapitre II : Fouille de données

II.1 Introduction

Les techniques d'exploration de données sont de plus en plus utilisées dans le domaine de la santé. Comme pour prédire certains indicateurs de santé, le Découvrez des informations ou des problèmes cachés, ou trouvez des problèmes sur la section médicale.

Dans ce chapitre nous voulons reconnaitre les différentes techniques de data mining, afin d'avoir un aperçu complet sur eux, pour identifier les techniques appropriées pour l'utiliser dans la résolution des problèmes trouvé dans l'introduction général.

II.2 Fouille de données

II.2.1 Définition

Il n'est pas surprenant que l'exploration de données, en tant que sujet véritablement interdisciplinaire, puisse être définie de différentes manières. Même le terme data mining ne présente pas vraiment tous les principaux composants dans l'image. Pour désigner l'extraction de l'or à partir de roches ou de sable, on dit or l'exploitation minière au lieu de l'extraction de roches ou de sable. De même, l'exploration de données aurait dû être plus [34].

Figure II. 1. Fouille de données - recherche de connaissances (modèles intéressants) dans les données [34]

Bien nommée « extraction de connaissances à partir de données », qui est malheureusement un peu longue. Cependant, à court terme, l'exploration des connaissances peut ne pas refléter l'accent mis sur l'extraction à partir de grandes quantités de données. Néanmoins, l'exploitation minière est un terme vivant caractérisant le processus qui trouve un petit ensemble de pépites précieuses à partir d'une grande quantité de matières premières

(Figure 2.1). Ainsi, un tel abus de langage portant à la fois « données » et « extraction » est devenu un choix populaire. En outre, de nombreux autres termes ont une signification similaire à l'exploration de données par exemple, l'exploration de connaissances à partir de données, l'extraction de connaissances, l'analyse de données/modèles, l'archéologie des données et le dragage de données. De nombreuses personnes traitent des mini-données. [34]

La fouille de données ou le Data Mining (DM) est un processus d'extraction et de découverte de modèles dans de grands ensembles de données impliquant des méthodes à l'intersection de l'apprentissage automatique, des statistiques et des systèmes de base de données.[35] L'exploration de données est un sous-domaine interdisciplinaire de l'informatique et des statistiques dont l'objectif global est d'extraire des informations (avec des méthodes intelligentes) à partir d'un ensemble de données et de transformer les informations en une structure compréhensible pour une utilisation ultérieure.[35][36][37] L'exploration de données est l'étape d'analyse du processus de « découverte des connaissances dans les bases de données », ou KDD Comme le montre la figure 2.2. Ce processus consiste en une série d'étapes de prétraitement des données au post-traitement des résultats de la fouille de données.

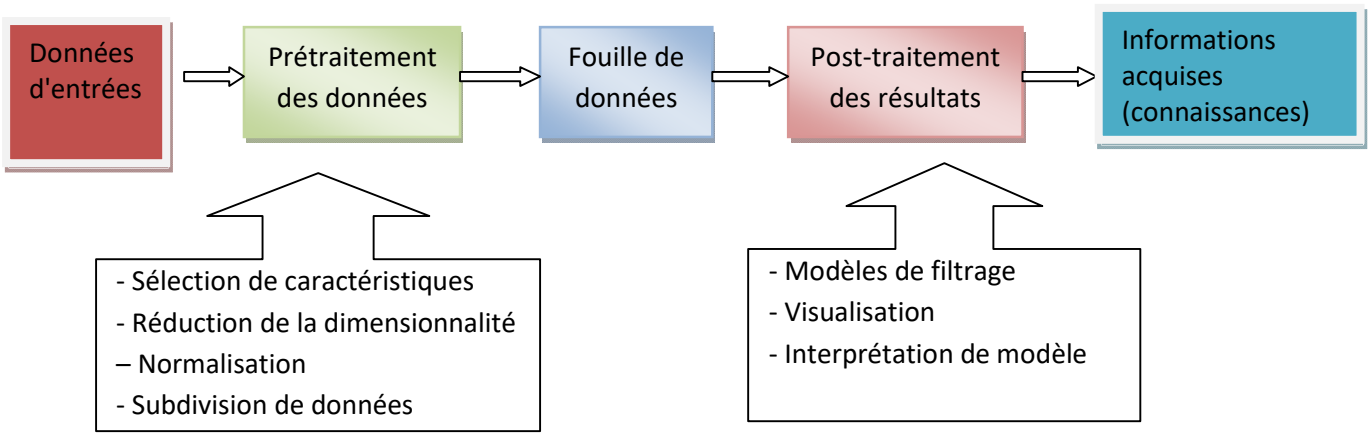

Figure II. 2. Le processus d'extraction des connaissances à partir de données [38].

II.2.2Processus du Fouille de données

Les bases de données actuelles du monde réel sont très sensibles aux données bruyantes, manquantes et incohérentes en raison de leur taille généralement énorme (souvent plusieurs giga-octets ou plus) et de leur origine probable à partir de sources multiples et hétérogènes. Des données de faible qualité conduiront à des résultats d'extraction de mauvaise qualité.

Il existe un certain nombre de techniques de prétraitement des données. Le nettoyage des données peut être appliqué pour supprimer le bruit et corriger les incohérences dans les données. L'intégration des données fusionne les données de plusieurs sources dans un magasin de données cohérent, tel qu'un entrepôt de données. Des transformations de données, telles que la normalisation, peuvent être appliquées. Par exemple, la normalisation peut améliorer la précision et l'efficacité des algorithmes d'extraction impliquant des mesures de distance. La réduction des données peut réduire la taille des données en les agrégeant, en éliminant les fonctionnalités redondantes ou en les regroupant, par exemple. Ces techniques ne sont pas mutuellement exclusives ; ils peuvent travailler ensemble. Par exemple, le nettoyage des données peut impliquer des transformations pour corriger des données erronées, telles que en transformant toutes les entrées d'un champ de date dans un format commun. Les techniques de traitement des données, lorsqu'elles sont appliquées avant l'exploitation minière, peuvent considérablement améliorer la qualité globale des modèles exploités et / ou le temps nécessaire à l'extraction proprement dite.

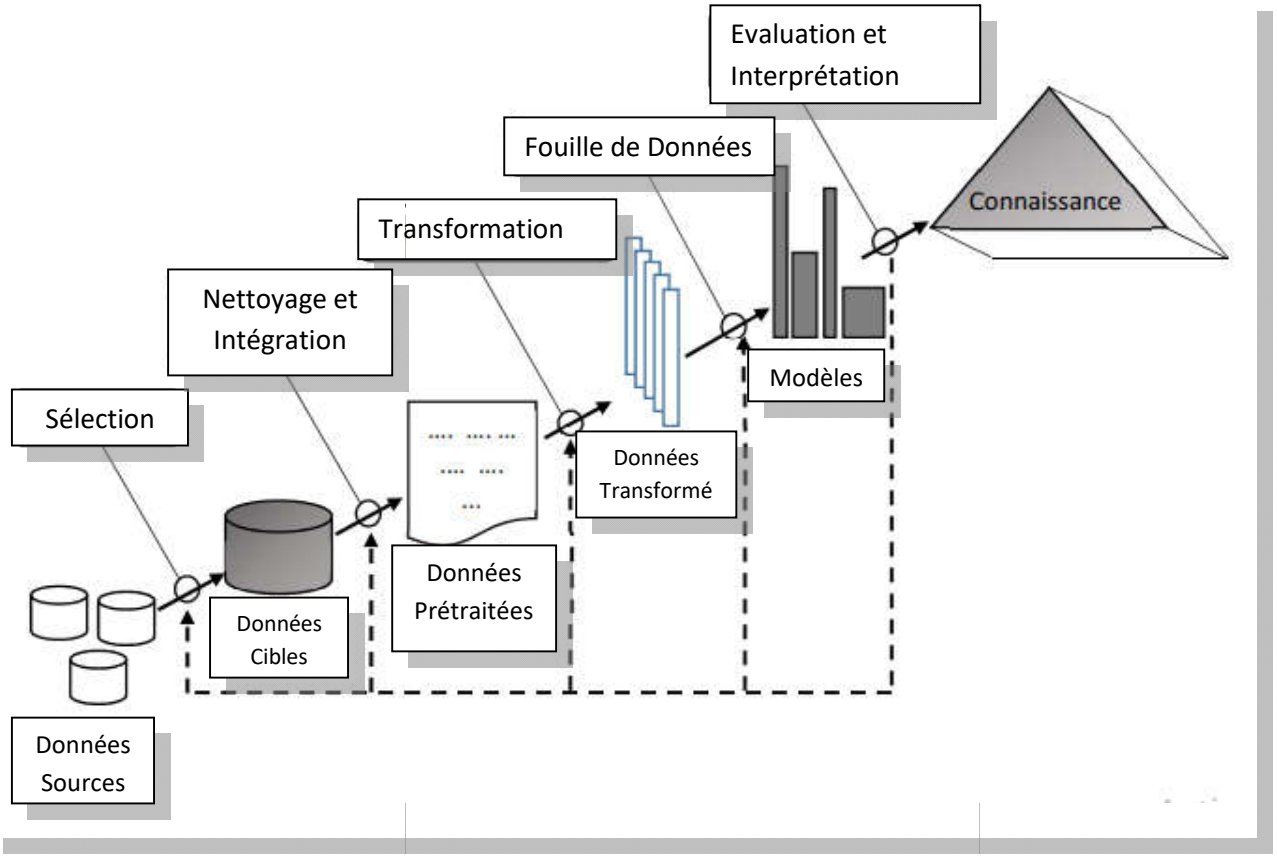

Figure II. 3. Les étapes du processus de découverte des connaissances [39].

Sélection des données : lorsque les données pertinentes pour la tâche d'analyse sont extraites de la base de données)

Nettoyage des données : Les routines de nettoyage des données permettent de « nettoyer » les données en remplissant les valeurs manquantes, en lissant les données bruyantes, en identifiant ou en supprimant les valeurs aberrantes et en résolvant les incohérences. Si les utilisateurs pensent que les données sont sales, il est peu probable qu'ils fassent confiance aux résultats de l'exploration de données qui leur a été appliquée. De plus, des données sales peuvent semer la confusion dans la procédure d'extraction, entraînant une sortie peu fiable. Bien que la plupart des routines de minage comportent des procédures pour traiter des données incomplètes ou bruyantes, elles ne sont pas toujours robustes. Au lieu de cela, ils peuvent se concentrer pour éviter de sur appliquer les données à la fonction modélisée. Par

conséquent, une étape de prétraitement utile consiste à exécuter vos données via certaines routines de nettoyage de données.

Intégration de données certains attributs représentant un concept donné peuvent avoir des noms différents dans différentes bases de données, ce qui entraîne des incohérences et des redondances.

Une grande quantité de données redondantes peut ralentir ou perturber le processus de découverte des connaissances. De toute évidence, en plus du nettoyage des données, des mesures doivent être prises pour éviter les redondances lors de l'intégration des données. En règle générale, le nettoyage et l'intégration des données sont effectués comme une étape de prétraitement lors de la préparation des données pour un entrepôt de données. Un nettoyage supplémentaire des données peut être effectué pour détecter et supprimer les redondances pouvant résulter de l'intégration des données.

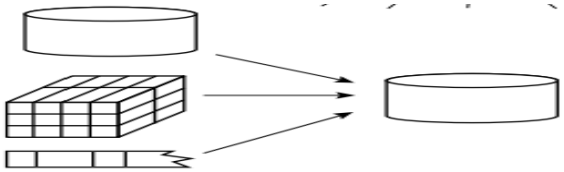

Figure II. 4. Intégration de données [41].

Les opérations de transformation de données : telles que la normalisation et l'agrégation, sont des procédures de prétraitement de données supplémentaires qui contribueraient au succès du processus d'exploration de données.

-2, 32, 100, 59, 48 ⟶ -0.02, 0.32, 1.00, 0.59, 0.48

Fouille de données : un processus essentiel où des méthodes intelligentes sont appliquées pour extraire modèles de données)

Évaluation des modèles (pour identifier les modèles vraiment intéressants représentant les connaissances sur la base de mesures de l'intérêt

Présentation des connaissances : où les techniques de visualisation et de représentation des connaissances sont utilisées pour présenter les connaissances extraites aux utilisateurs)

Les étapes 1 à 4 dans La figure 2.3, sont différentes formes de prétraitement des données, où les données sont préparées pour l'extraction. L'étape d'exploration de données

peut interagir avec l'utilisateur ou une base de connaissances. Les modèles intéressants sont présentés à l'utilisateur et peuvent être stockés en tant que nouvelles connaissances dans la base de connaissances. La vue précédente montre l'exploration de données comme une étape du processus de découverte des connaissances, bien qu'elle soit essentielle car elle découvre des modèles cachés pour l'évaluation. Cependant, dans l'industrie, dans les médias et dans le milieu de la recherche, le terme data mining est souvent utilisé pour désigner l'ensemble du processus de découverte de connaissances (peut-être parce que le terme est plus court que la découverte de connaissances à partir de données). Par conséquent, nous adoptons une vision large de la fonctionnalité d'exploration de données : l'exploration de données est le processus de découverte de modèles et de connaissances intéressants à partir de grandes quantités de données. Les sources de données peuvent inclure des bases de données, des entrepôts de données, le Web, d'autres référentiels d'informations ou des données diffusées dans le système dynamiquement [40].

II.3 Clustering (Segmentation)

II.3.1 Définition

Le clustering défini par les seules données. Le clustering est également appelé apprentissage non supervisé ou segmentation. Cela peut être considéré comme un partitionnement ou une segmentation des données en groupes qui peuvent ou non être disjoints figure 2.6. Le regroupement est généralement réalisé par déterminer la similitude entre les données sur des attributs prédéfinis. Les données les plus similaires sont regroupées en grappes. Depuis les clusters ne sont pas prédéfinis, un expert du domaine est souvent requis pour interpréter la signification des grappes créées [42].

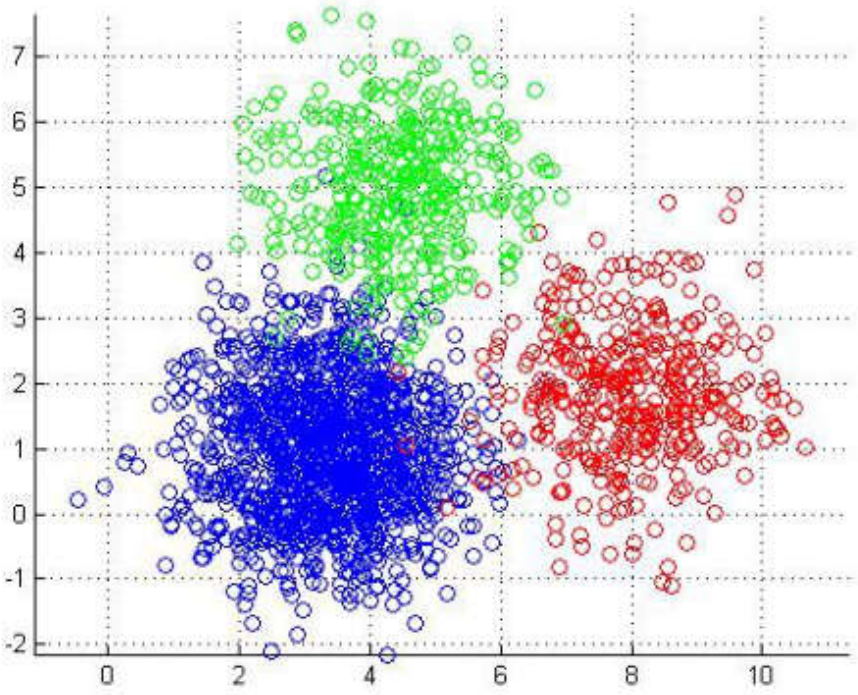

Figure II. 5. Points de données groupées [43].

II.3.2 Techniques de clustering

Les algorithmes de clustering sont classés [44] selon :

- Le type d'entrée
- Critère de clustering définissant la similarité entre les objets
- Concepts sur lesquels les techniques d'analyse de clustering

II.3.2.1 Clustering K-Means

Le clustering K-means [45,46] est l'algorithme de clustering partitionne le plus largement utilisé. Il commence par choisir K points représentatifs comme centroïdes initiaux. Chaque point est ensuite attribué au centroïde le plus proche en fonction d'une mesure de proximité particulière choisie. Une fois les clusters formés, les centroïdes de chaque cluster sont mis à jour. L'algorithme répète ensuite ces deux étapes de manière itérative jusqu'à ce que les centroïdes ne changent pas ou qu'un autre critère de convergence relâché alternatif soit satisfait.

AlgorithmeClustering K-Means

1. Sélectionnez K points comme centroïdes initiaux.
2. Répéter
   2.1.Formez K clusters en affectant chaque point à son centroïde le plus proche.
   2.2.Recalculez le centroïde de chaque cluster.
3. Jusqu'à ce que le critère de convergence soit satisfait.

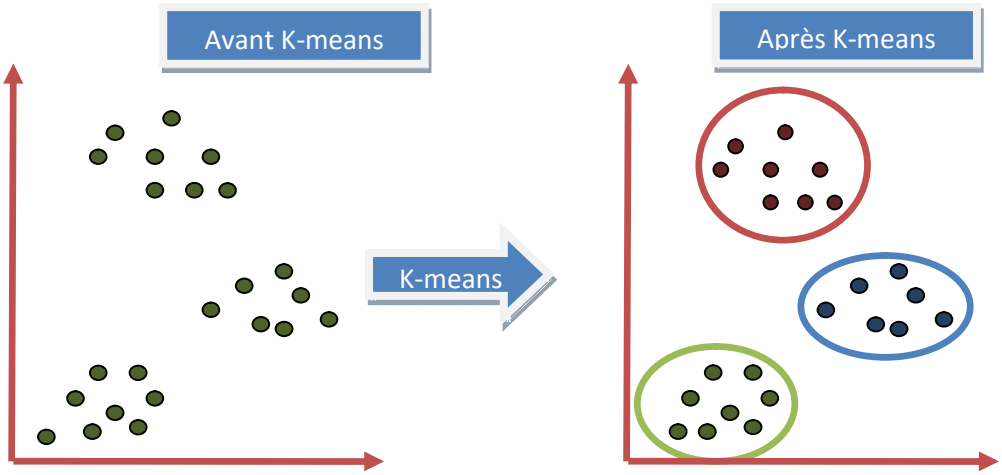

Figure II. 6. Algorithm K-means

II.3.2.2Clustering Hiérarchique

Définition

Le clustering hiérarchique est une famille générale d'algorithmes de clustering qui construisent des clusters imbriqués en les fusionnant ou en les divisant successivement. Cette hiérarchie de clusters est représentée sous la forme d'un arbre (ou dendrogramme). La racine de l'arbre est l'unique cluster qui rassemble tous les échantillons, les feuilles étant les clusters avec un seul échantillon [47].

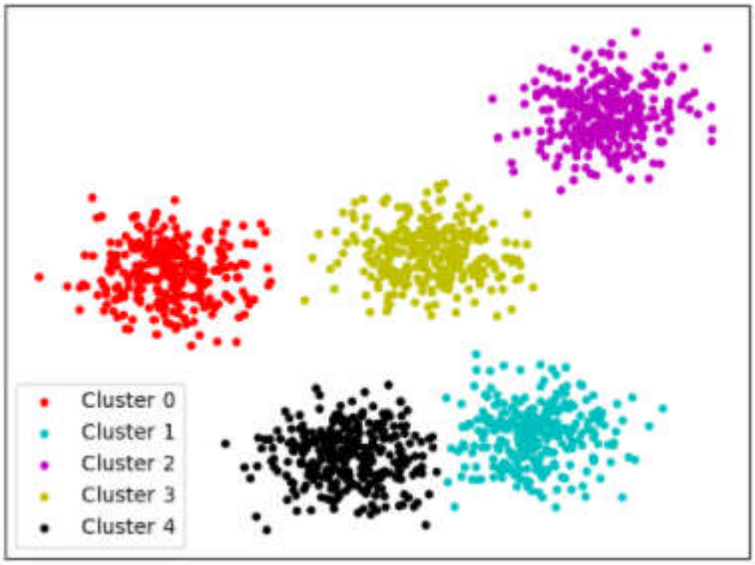

Figure II. 7. Clustering Hiérarchique [48].

Type de clustering hiérarchique

Cette technique de clustering est divisée en deux typesmontrés dans la figure 2.9 [43], à savoir, le clustering hiérarchique aggloméré et le clustering hiérarchique division.

a. Clusteringhiérarchique aggloméré

Le clustering hiérarchique aggloméré est le type le plus courant de clustering hiérarchique utilisé pour regrouper des objets en clusters en fonction de leur similarité. C'est une approche ascendante où chaque observation commence dans son propre cluster, et les paires de clusters sont fusionnées au fur et à mesure que l'on monte dans la hiérarchie [43].

L'algorithme de clustering hiérarchique aggloméré comprend les étapes suivantes :

1- Faire de chaque point de données un cluster à un seul point → forme N clusters
2- Prendre les deux points de données les plus proches et en faire un seul cluster → forme N-1 clusters.
3- Répéter.
    3.1- Prendre les deux clusters les plus proches et en faire un cluster → Formes N-2 clusters.
4- Jusqu'à ce qu'il ne vous reste qu'un seul cluster.

b. Clusteringhiérarchique division

Le clustering hiérarchiquedivisionest une méthode de clustering descendante où nous attribuons toutes les observations à un seul cluster, puis partitionnons le cluster en deux clusters les moins similaires. Enfin, nous procédons récursivement sur chaque cluster jusqu'à ce qu'il y ait un cluster pour chaque observation. Cette approche de clustering est donc exactement à l'opposé du clustering aggloméré [43].

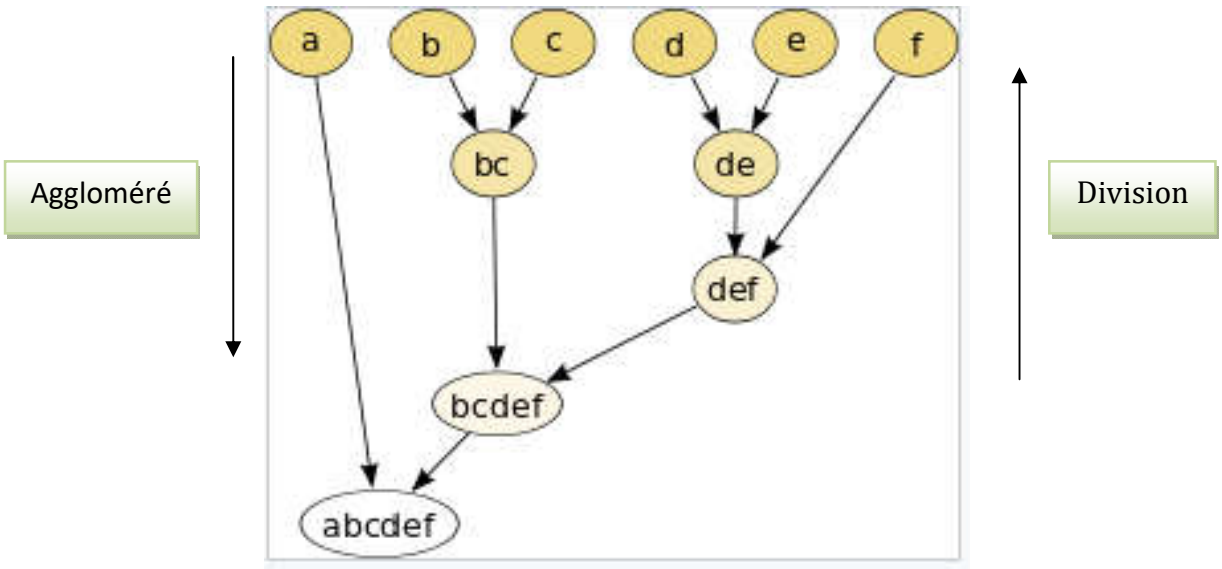

Figure II. 8. Types de clustering hiérarchique [43].

Le dendrogramme

Le dendrogramme est un diagramme en forme d'arbre qui montre la relation hiérarchique entre les observations. Il contient la mémoire des algorithmes de clustering hiérarchique [43].

Un dendrogramme peut être un graphique à colonnes (comme dans la figure ci-dessous) ou un graphique à lignes. Certains dendrogrammes sont circulaires ou ont une forme fluide, mais le logiciel produira généralement un graphique en lignes ou en colonnes. Peu importe la forme.

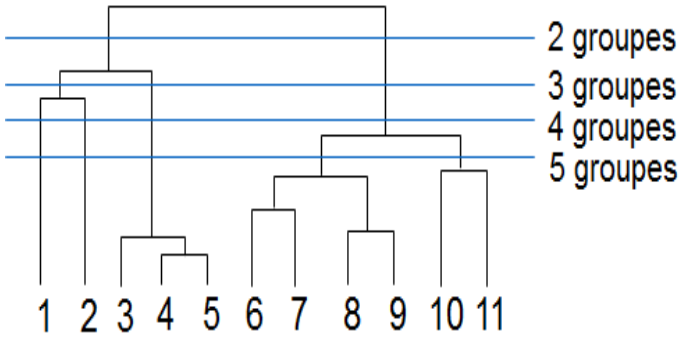

Figure II. 9. Simple exemple de dendrogramme [43].

II.3.2.3 Clustering MeanShift

Le clustering MeanShift vise à découvrir des blobs dans une densité d'échantillons homogène. Il s'agit d'un algorithme basé sur le centroïde, qui fonctionne en mettant à jour les candidats pour que les centroïdes soient la moyenne des points dans une région donnée. Ces candidats sont ensuite filtrés dans une étape de post-traitement pour éliminer les quasi-doubles pour former l'ensemble final de centroïdes.

L'algorithme n'est pas hautement évolutif, car il nécessite plusieurs recherches de voisins les plus proches pendant l'exécution de l'algorithme. La convergence de l'algorithme est garantie, mais l'algorithme cessera d'itérer lorsque le changement des centroïdes est faible [49].

Algorithme Clustering MeanShift

1. Sélectionnez K points aléatoires comme modes de distribution.
2. Répéter
    2.1. Pour chaque mode x donné, calculez le vecteur de décalage moyen mh(x).
    2.2.Mettre à jour le point x = mh(x).
3. Jusqu'à ce que les modes deviennent stationnaires et convergent.

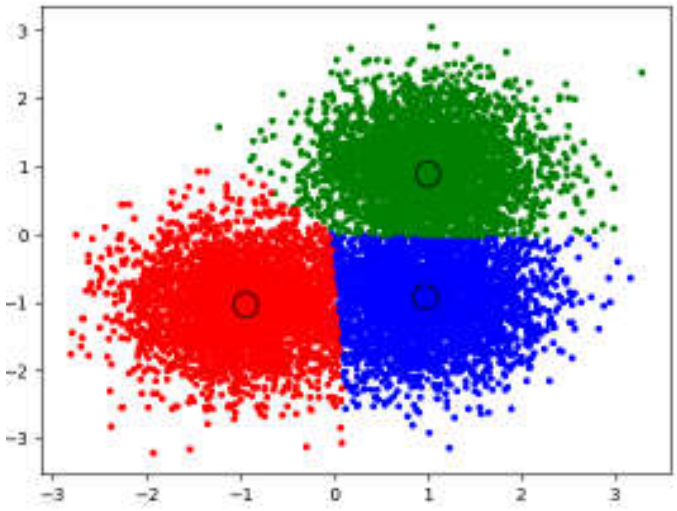

Figure II. 10. Exemple de Clustering MeanShift [49].

II.3.2.4 Clustering spatial basé sur la densité (DBSCAN)

C'est une technique de clustering qui ne nécessite pas la spécification du nombre de clusters est DBSCAN. Cependant, DBSCAN nécessite essentiellement deux paramètres : échantillons eps et min. Eps spécifie la distance maximale entre deux échantillons de données dans lesquels l'un d'eux est censé être le voisinage d'un autre qui est un point central d'un cluster. Min échantillons définit le nombre minimum d'échantillons qui doivent être dans le voisinage avec l'échantillon de base. Il suppose qu'un cluster est une région dense avec des points de données supérieurs au nombre minimum d'échantillons dans la plage d'eps du point centrale chaque groupe est séparé des autres par une densité plus faible.

II.3.2.5 Clustering Spectral

Le clustering spectral utilise l'algorithme des k-moyennes comme étape de son algorithme. Dernièrement, il est devenu populaire en raison de sa mise en œuvre simple et de ses excellentes performances avec le clustering basé sur des graphes.

Spectral Clustering effectue une intégration de faible dimension de la matrice d'affinité entre les échantillons, suivie d'un regroupement, par exemple, par K-Means, des composants des vecteurs propres dans l'espace de faible dimension. Il est particulièrement efficace en termes de calcul si la matrice d'affinité est clairsemée et que le solveur amg est utilisé pour le problème des valeurs propres (remarque, le solveur amg nécessite que le module pyamg soit installé) [47].

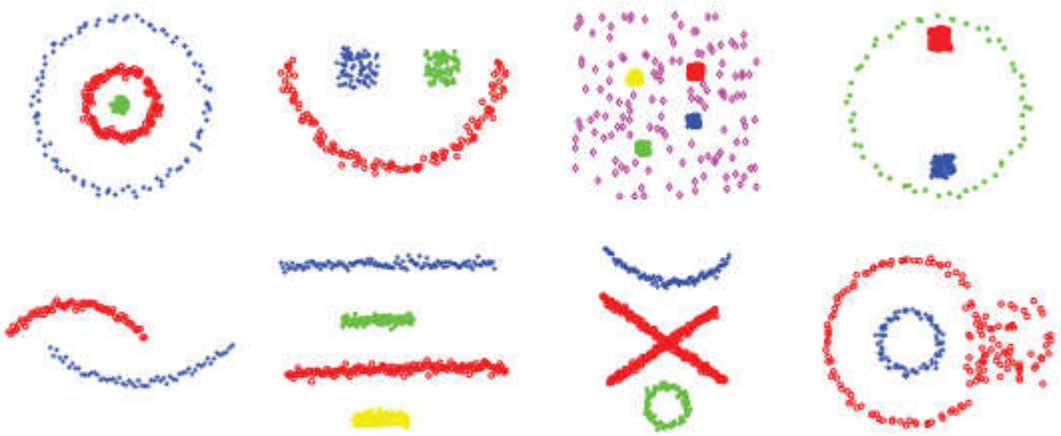

Figure II. 11. Regroupement spectral à réglage automatique [47].

II.3.3 Qualités d'un clustering

- Comme il est montré dans la figure suivante, une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité importante intra-classe
  - Similarité faible inter-classe
- La qualité d'un clustering dépend de :
  - La mesure de similarité utilisée
  - L'implémentation de la mesure de similarité
- La qualité d'une méthode de clustering est évaluée par son habilité à découvrir certains ou tous les "patterns" cachés

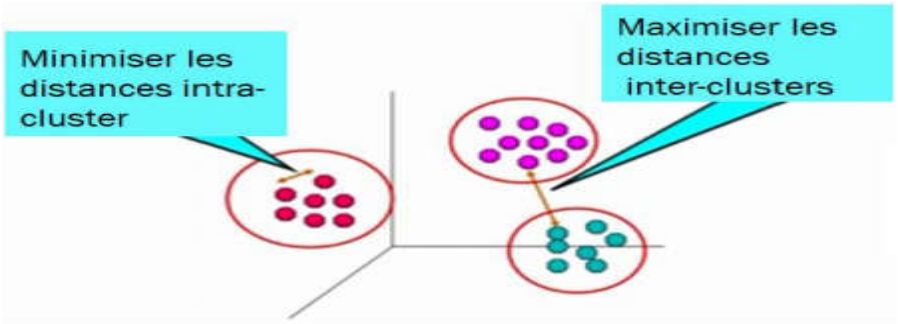

Figure II. 12. Objectifs du clustering [47].

II.3.4 Exemples d'applications

- Marketing : segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats.
- Environnement : identification des zones terrestres similaires (en termes d'utilisation) dans une base de données d'observation de la terre.
- Assurance : identification de groupes d'assurés distincts associés à un nombre important de déclarations.
- Planification de villes : identification de groupes d'habitations suivant le type d'habitation, valeur, localisation géographique,etc.
- Médecine : Localisation de tumeurs dans le cerveau
- Nuage de points du cerveau fournis par le neurologue
- Identification des points définissant une tumeur

II.4 Applications de feuille de données dans le domaine d'épidémies

L'exploration de données médicales a un grand potentiel pour explorer les modèles cachés dans les ensembles de données du domaine médical. Ces modèles peuvent être utilisés pour le diagnostic clinique. Cependant, les données médicales brutes disponibles sont largement diffusées, de nature hétérogène et volumineuse. Ces données doivent être collectées sous une forme organisée. Ces données collectées peuvent ensuite être intégrées pour former un système d'information hospitalier. La technologie d'exploration de données offre une approche orientée utilisateur des modèles nouveaux et cachés dans les données. L'exploration de données et les statistiques s'efforcent toutes deux de découvrir des modèles et des structures dans les données. Les statistiques traitent uniquement des nombres hétérogènes, tandis que l'exploration de données traite des champs hétérogènes.

Les épidémiologistes travaillant dans le domaine de la santé publique appliquée disposent d'une myriade de sources de données potentielles. Plusieurs facteurs doivent être pris en compte lors de l'identification des sources de données pertinentes pour mener une enquête sur le terrain. Il s'agit notamment des objectifs et de la portée de l'enquête, de l'existence ou de l'accès aux données requises, de la mesure dans laquelle les données provenant de différentes sources peuvent être combinées en pratique, des méthodes et de la faisabilité de la collecte de données primaires et des ressources (p. Ex. Personnel, financement) disponibles [51].

Les données épidémiologiques sont essentielles pour cibler et mettre en œuvre des mesures de contrôle fondées sur des données probantes pour protéger la santé et la sécurité du public. Nulle part les données ne sont plus importantes que lors d'une enquête épidémiologique sur le terrain pour identifier la cause d'un problème de santé publique urgent nécessitant une intervention immédiate. Bon nombre des étapes de la conduite d'une enquête sur le terrain reposent sur l'identification des données existantes pertinentes ou la collecte de nouvelles données qui répondent aux objectifs clés de l'enquête.

En fait, l'exploration de données dans le secteur de la santé reste aujourd'hui, pour l'essentiel, un exercice académique avec seulement quelques réussites pragmatiques. Les universitaires utilisent des approches d'exploration de données comme les arbres de décision, clusters, réseaux de neurones et séries chronologiques pour publier des recherches. Les soins de santé, cependant, ont toujours été lents à intégrer la dernière recherche sur la pratique quotidienne [52].

II.5 Conclusion

Nous avons fait un survol sur les techniques de data mining. Et nous avons constaté qu'ils sont décomposés en deux catégories, la première se compose des techniques supervisées et la deuxième se compose des techniques non supervisées. Dans ces deux catégories, il existe de nombreuses techniques avec des caractéristiques différentes et avec des points fortes et d'autres faibles. Nous allons utiliser ces informations pour choisir les techniques nécessaires pour résoudre les problèmes que nous avons mentionnés dans l'introduction générale.

Chapitre III : Le cas D’étude

III.1 Introduction

Il y a un effort mondial de la communauté des chercheurs pour explorer l'impact médical, économique et sociologique de la pandémie de COVID-19. De nombreuses disciplines différentes essaient de trouver des solutions et de conduire des stratégies à une grande variété de problèmes très cruciaux.

L'exploration de données, un processus de découverte des caractéristiques silencieuses des big data, est l'une de ces techniques qui sont aujourd'hui devenues plus populaires pour traiter un volume massif d'ensembles de données sur les maladies infectieuses. Dans ce chapitre, nous appliquons l'analyse de cluster, l'une des techniques d'exploration de données pour décrire la prévalence du COVID-19 dans L'Europeet identifiant les pays les plus touchés en des cas et des morts, en utilisant le langage de programmation Python.

Certaines caractéristiques extraites intéressantes sont discutées et des suggestions pour de futures recherches dans ce domaine sont également présentées.

III.3 L'environnement et les packages pythons utilisés

III.3.1Le langage de programmation python

C’est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions, il est ainsi similaire à Perl, Ruby, Scheme, Smalltalk et Tcl [53].

Le langage Python est placé sous une licence libre proche de la licence BSD4 et fonctionne sur la plupart des plates-formes informatiques, des smartphones aux ordinateurs centraux5, de Windows à Unix avec notamment GNU/Linux en passant par macOS, ou encore Android, iOS, et peut aussi être traduit en Java ou .NET. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser.

III.3.2 L'environnement python

Un environnement virtuel est simplement un répertoire quelque part dans vos fichiers de projet qui contient tous les fichiers et dépendances nécessaires pour exécuter des scripts python que vous utilisez pour exécuter le projet sur lequel vous travaillez. Il y a des choses plus techniques qui se passent en arrière-plan que le simple clonage du répertoire, mais lorsque vous inspectez visuellement le contenu de ce répertoire, vous remarquerez qu'il est très similaire au répertoire sur lequel Python est installé sur votre système. Ce que vous pouvez ensuite faire, c'est installer toutes les dépendances externes dont votre projet a besoin pour s'exécuter (comme les demandes) dans cet environnement virtuel au lieu de les installer sur votre installation globale de python. Cela donne l'avantage que si vous avez besoin d'une version spécifique d'une dépendance telle que des requêtes ou un interpréteur python spécifique, vous n'avez pas besoin de désinstaller la version que vous avez installée globalement, puis d'installer la version pour laquelle votre projet appelle chaque fois que vous changez de projet [54].

Figure III. 1. Environnements virtuels python [54].

III.3.3 Les package pythons utilisés

III.3.3.1 Package Matplotlib

Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy. Matplotlib est distribuée librement et gratuitement sous une licence de style BSD [55].

Matplotlib.pyplot est une collection de fonctions qui font fonctionner matplotlib comme MATLAB. Chaque fonction pyplot apporte des modifications à une figure : par exemple, crée une figure, crée une zone de traçage dans une figure, trace des lignes dans une zone de traçage, décore le tracé avec des étiquettes, etc.

Dans Matplotlib.pyplot, divers états sont conservés à travers les appels de fonction, de sorte qu'il garde une trace d'éléments tels que la figure actuelle et la zone de traçage, et les fonctions de traçage sont dirigées vers les axes actuels (veuillez noter que "axes" ici et dans la plupart des endroits dans la documentation fait référence à la partie des axes d'une figure et non au terme mathématique strict pour plus d'un axe)

III.3.2.2 Package pandas

Pandas (tout en minuscules) est une boîte à outils d'analyse de données basée sur Python qui peut être importée en utilisant import pandas as pd. Il présente une gamme variée d'utilitaires, allant de l'analyse de plusieurs formats de fichiers à la conversion d'une table de données entière en un tableau matriciel NumPy. Cela fait des pandas un allié de confiance dans la science des données et l'apprentissage automatique.

Semblable à NumPy, pandas traite principalement des données dans des tableaux 1D et 2D, cependant, les pandas gèrent les deux différemment [56].

III.3.2.3 Package Clustering

Les algorithmes de clustering sont utiles dans la théorie de l'information, la détection de cibles, les communications, la compression et d'autres domaines [57].

Scipy.cluster.hierarchy, Le module de hiérarchie fournit des fonctions pour le clustering hiérarchique et agglomératif. Ses fonctionnalités incluent la génération de clusters

hiérarchiques à partir de matrices de distance, le calcul de statistiques sur les clusters, la coupure de liens pour générer des clusters plats et la visualisation de clusters avec des dendrogrammes [57].

III.3.2.4 Package Scikit-Learn

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs 2 notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche [58].

Le regroupement de données non étiquetées peut être effectué avec le module sklearn.cluster.

Chaque algorithme de clustering se décline en deux variantes : une classe, qui implémente la méthode fit pour apprendre les clusters sur les données de train, et une fonction, qui, étant donné les données de train, renvoie un tableau d'étiquettes d'entiers correspondant aux différents clusters [58].

Le module sklearn.cluster rassemble des algorithmes de clustering non supervisés populaires.

III.4 Les étapes d'implémentation

III.4.1Définition de Domain

La pandémie de Covid-19 due au coronavirus SARS-CoV-2, frappe tous les continents dont l'Europe. Après que l'Asie a été le foyer initial de cette épidémie au début de l'année 2020, l'Europe devient courant mars le nouveau foyer central de l'épidémie devenue pandémie [59].

Réaction d'urgence de l'UE face à la pandémie de COVID-1, Depuis le début de la pandémie de COVID-19, l'UE coopère avec ses États membres afin de protéger la santé et le bien-être des citoyens de l'UE et de sauver des vies.

La réaction de l'UE face à la COVID-19 s'articule autour de quatre priorités [59] :

- Limiter la propagation du virus.
- Assurer la fourniture de matériel médical.

- Promouvoir la recherche sur des traitements et des vaccins.

Soutenir les emplois, les entreprises et l'économie.

Ces priorités ont été approuvées par les dirigeants de l'UE en mars 2020 afin d'orienter la réaction d'urgence de l'UE face à la pandémie de COVID-19. Au plus fort de la crise de la COVID-19, les dirigeants de l'UE se sont rencontrés régulièrement par vidéoconférence pour examiner et évaluer la situation et coordonner les actions [59].

La pandémie actuelle de COVID-19 se déroule dans un monde fortement interconnecté. Cette interconnexion explique pourquoi elle est devenue universelle en si peu de temps et pourquoi elle a stimulé la création d'une grande quantité de données ouvertes pertinentes. Dans cette étude, nous utilisons des outils de science des données pour explorer ces données ouvertes, Une méthode d'analyse de cluster non hiérarchique a été utilisée, pour identifier les similitudes des pays en termes d'ensembles complets d'indicateurs selon le nombre des cas et des morts et identifiant les pays européens les plus touchés depuis 6 mois de la pandémie à partir de mars 2021 [59].

Ces données sont utiles car divers pays sont regroupés en fonction de COVID-19 données épidémiologiques, qui peuvent être utiles pour distinguer objectivement les pays avec des Propagation et résultats du COVID-19.

Cette recherche explicative peut aider les décideurs politiques du gouvernement à prendre de meilleures décisions concernant la gestion de la crise sanitaire due à la pandémie de COVID-19.

III.4.2 Collection et description des données

Le fichier de données téléchargeable contient des informations sur les nouveaux cas et décès de COVID-19 signalés dans les pays de l'UE / EEE. Chaque ligne contient les données correspondantes pour un certain jour et par pays.

Le fichier est mis à jour quotidiennement. Vous pouvez utiliser les données conformément à la politique en matière de droits d'auteur de l'ECDC.

L'ECDC utilise plusieurs sources d'informations par pays. Les sources d'information sont les ministères de la Santé ou les instituts nationaux de santé publique (sites Web, comptes officiels Twitter ou comptes officiels Facebook). De plus amples informations sont

disponibles sur https://www.ecdc.europa.eu/en/covid-19/datacollection ou https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country.

Les données incluses dans ce fichier sont collectées par l'ECDC Epidemic Intelligence à partir de diverses sources et sont affectées par la stratégie locale de test, la capacité du laboratoire et l'efficacité de la surveillance systèmes. La comparaison de la situation épidémiologique concernant le COVID-19 entre les pays ne doit donc pas se fonder uniquement sur ces taux Cependant, au niveau de chaque pays depuis 6 mois de la pandémie à partir de mars 2021, cet indicateur peut être utile pour suivre la situation nationale au fil du temps.

Les politiques de test et le nombre de tests effectués pour 100 000 personnes varient considérablement dans l'UE / EEE.

Le nombre quotidien de cas et de décès de COVID-19 déclarés doit être utilisé en combinaison avec d'autres facteurs, notamment les politiques de dépistage, le nombre de tests effectués, le test de positivité, la surmortalité et les taux d'admissions à l'hôpital et à l'unité de soins intensifs (USI), lors de l'analyse de la situation épidémiologique dans un pays. La plupart de ces indicateurs sont présentés pour les États membres de l'UE / EEE dans le rapport Country Overview.

Toutes les informations relatives à la base de données sont présentées dans le tableau suivant.

Variable	Définition	Code
dater	Date du rapport"JJ / MM / AAAA"	String
Jour		Int
Mois		Int
Ans		Int
Cas	Nombre de nouveaux cas signalés	Int
Morts	Nombre de décès nouvellement signalés	Int
PaysEtTerritoires	Nom du pays ou territoire	String
CodeTerritoirePays	Code ISO à 3 lettres	String
DonnPop2020	Données Eurostat 2019	Int
Continent-Exp	Nom du continent faisant rapport	String

Tableau III. 1. Informations relatives à la base de données.

La base de données est représentée sur la Tableau III.2.

	A	B	C	D	E	F	G	H	I	J
1	dateRep	jour	mois	ans	cas	morts	PaysEtTerritoires	DonnPop2020	CodeTerritoirePays	continentExp
2	03/09/2021	2021	09	03	1549	3	Austria	8901064	AUT	Europe
3	02/09/2021	2021	09	02	1821	6	Austria	8901064	AUT	Europe
4	01/09/2021	2021	09	01	1265	1	Austria	8901064	AUT	Europe
5	31/08/2021	2021	08	31	1092	2	Austria	8901064	AUT	Europe
6	30/08/2021	2021	08	30	1339	5	Austria	8901064	AUT	Europe
7	29/08/2021	2021	08	29	1408	4	Austria	8901064	AUT	Europe
8	28/08/2021	2021	08	28	1490	2	Austria	8901064	AUT	Europe
9	27/08/2021	2021	08	27	1535	2	Austria	8901064	AUT	Europe
10	26/08/2021	2021	08	26	1618	2	Austria	8901064	AUT	Europe
11	25/08/2021	2021	08	25	1014	3	Austria	8901064	AUT	Europe
12	24/08/2021	2021	08	24	971	2	Austria	8901064	AUT	Europe
13	23/08/2021	2021	08	23	1174	2	Austria	8901064	AUT	Europe
14	22/08/2021	2021	08	22	1309	1	Austria	8901064	AUT	Europe
15	21/08/2021	2021	08	21	1321	2	Austria	8901064	AUT	Europe
16	20/08/2021	2021	08	20	1179	0	Austria	8901064	AUT	Europe
17	19/08/2021	2021	08	19	1293	0	Austria	8901064	AUT	Europe
18	18/08/2021	2021	08	18	864	0	Austria	8901064	AUT	Europe
19	17/08/2021	2021	08	17	715	0	Austria	8901064	AUT	Europe
20	16/08/2021	2021	08	16	871	2	Austria	8901064	AUT	Europe
21	15/08/2021	2021	08	15	997	2	Austria	8901064	AUT	Europe
22	14/08/2021	2021	08	14	967	2	Austria	8901064	AUT	Europe
23	13/08/2021	2021	08	13	879	0	Austria	8901064	AUT	Europe
24	12/08/2021	2021	08	12	894	1	Austria	8901064	AUT	Europe
25	11/08/2021	2021	08	11	584	1	Austria	8901064	AUT	Europe

Tableau III. 2. La base de données (tableau de données initiales).

III.4.3 Nettoyage de données

Dans cette partie nous avons supprimé les données non pertinentes. Dans l'ensemble de données d'origine, 10 variables collectent des informations générales sur 30 pays telles que date, jour, mois, ans, cas, morts, Pays, données Population 2020, Code Territoire Pays, continent, etc. Dans cette étude, nous considérons uniquement les variables qui sont liés aux cas et morts de chaque pays. Au total, nous utilisons 3 variables, qui sont expliquées dans la Tableau III.3.

1	PaysEtTerritoires	cas	morts
2	Austria	1549	3
3	Austria	1821	6
4	Austria	1265	1
5	Austria	1092	2
6	Austria	1339	5
7	Austria	1408	4
8	Austria	1490	2
9	Austria	1535	2
10	Austria	1618	2
11	Austria	1014	3
12	Austria	971	2
13	Austria	1174	2
14	Austria	1309	1
15	Austria	1321	2
16	Austria	1179	0
17	Austria	1293	0
18	Austria	864	0
19	Austria	715	0
20	Austria	871	2
21	Austria	997	2
22	Austria	967	2
23	Austria	879	0
24	Austria	894	1
25	Austria	584	1

Tableau III.3. Les 3 variables de base de données.

III.4.4 Préparation des données

La préparation des données consiste à rassembler, combiner, structurer et organiser les données afin de pouvoir les analyser dans le cadre de programmes d'informatique décisionnelle [60].

Donc, nous collectons le nombre de cas et de décès pour chaque pays afin de faciliter le processus de transfert et de compréhension des données.

Le tableau suivant montre les caractéristiques de chaque attribut, à savoir, la plus grande valeur et la plus petite valeur en termes de nombre de cas et de décès avec la moyenne et l'écart type.

Variable	Max	Min	Moyenne	Ecart Type
Cas	6799240	3301	1232976.6	1700664.921
Morts	12229352	33	61872.96667	2225617746

Tableau III. 4. Caractéristiques des attributs (nombre de morts et nombre de malades).

- **Normalisation de données**

La normalisation est une méthode de prétraitement des données qui permet de réduire la complexité des modèles et pour réduire les cas de double valeur. En transférant une base de données à l'un des formulaires normaux répertoriés, le schéma cible bénéficie d'une redondance moindre que le schéma source.

La mise en échelle min-max transforme chaque valeur numérique x vers une autre valeur x' ∈ [0,1] en utilisant la valeur minimale et la valeur maximale dans les données. Cette normalisation conserve la distance proportionnelle entre les valeurs d'une caractéristique [42].Pour ce faire on utilise la formule suivante :

$$x = \frac{\mathrm{x} - x_{min}}{x_{max} - x_{min}} \in [0,1]$$

Après avoir appliqué la normalisation aux caractéristiques du nombre des cas et des morts, nous obtenons le Tableau III.5.

Id	PaysEtTerritoires	Cas	Morts
0	Autriche	0,1	0,008
1	Belgique	0,174	0,02
2	Bulgarie	0,067	0,015
3	Croatie	0,054	0,006
4	Chypre	0,016	0,0003
5	Tchéquie	0,246	0,024
6	Danemark	0,05	0,002
7	Estonie	0,02	0,001
8	Finlande	0,018	0,0007
9	France	1	0,093
10	Allemagne	0,585	0,075
11	Grèce	0,086	0,011
12	Hongrie	0,119	0,024
13	Islande	0,001	0
14	Irlande	0,051	0,004
15	Italie	0,669	1
16	Lettonie	0,02	0,02
17	Liechtenstein	0	0,00002
18	Lituanie	0,043	0,003
19	Luxembourg	0,01	0,0006
20	Malte	0,004	0,0003
21	Pays-Bas	0,285	0,014
22	Norvège	0,023	0,0006
23	Pologne	0,424	0,061
24	Portugal	0,152	0,014
25	Roumanie	0,161	0,028
26	Slovaquie	0,114	0,01
27	Slovénie	0,039	0,003
28	Espagne	0,716	0,068
29	Suède	0,165	0,011

Tableau III. 5. Le tableau de données après la normalisation.

III.4.5 Traitement de données (Clustering hiérarchique)

L'analyse de cluster est un ensemble de techniques numériques qui créent des clusters ou des groupes d'éléments ayant des caractéristiques similaires [62]. Dans le clustering hiérarchique, les éléments sont regroupés en suivant certaines étapes qui peuvent commencer à partir d'un cluster pour chacun des éléments classés et aboutir à un cluster unique contenant tous les éléments.

Les méthodes hiérarchiques sont classées en techniques d'agglomération et de division. Certaines méthodes de regroupement hiérarchique agglomératif sont le lien moyen, le lien complet, le lien simple et le lien centroïdal, Les méthodes de lien permettent de garantir que les clusters sont bien séparés, car on prend en compte la distance entre les clusters. Cependant, elles ne garantissent pas que les clusters soient resserrés sur eux-mêmes (c'est la notion d'inertie intraclasse). Pour résoudre cela, il existe la méthode de Ward qui, à chaque itération, c'est-à-dire à chaque fois que 2 clusters sont regroupés en 1, cherche à minimiser l'augmentation d'inertie intraclasse due au regroupement des 2 clusters. Par défaut, on préfère en général utiliser la méthode de Ward. Fonctionnant tous de la même manière. Chacune de ces techniques à un dendrogramme [63] [64].

Toutes les analyses ont été effectuées en Python, à l'aide des packages généralement appliqués en science des données, à savoir NumPy, Pandas, Matplotlib, scipy et sklearn.

L'entraînement d'une Classification Ascendante Hiérarchique est facilité avec la librairie Scikit-Learn.

La CAH va ensuite rassembler les individus de manière itérative afin de produire un dendrogramme ou arbre de classification, Pour que l'axe des X représente la distance et l'axe des Y représente le nombre de chaque pays. La classification est ascendante car elle part des observations individuelles, elle est hiérarchique car elle produit des classes ou groupes de plus en plus vastes, incluant des sous-groupes en leur sein. En découpant cet arbre à une certaine hauteur choisie, on produira la partition désirée dans Figure III.2.

Le code Python utilisé pour dessiner ce graphique est :

```
plt.figure(figsize=(10, 7))
plt.title("Covid Dendograms")
dend = shc.dendrogram(shc.linkage(covid, method='ward'))
plt.show()
```

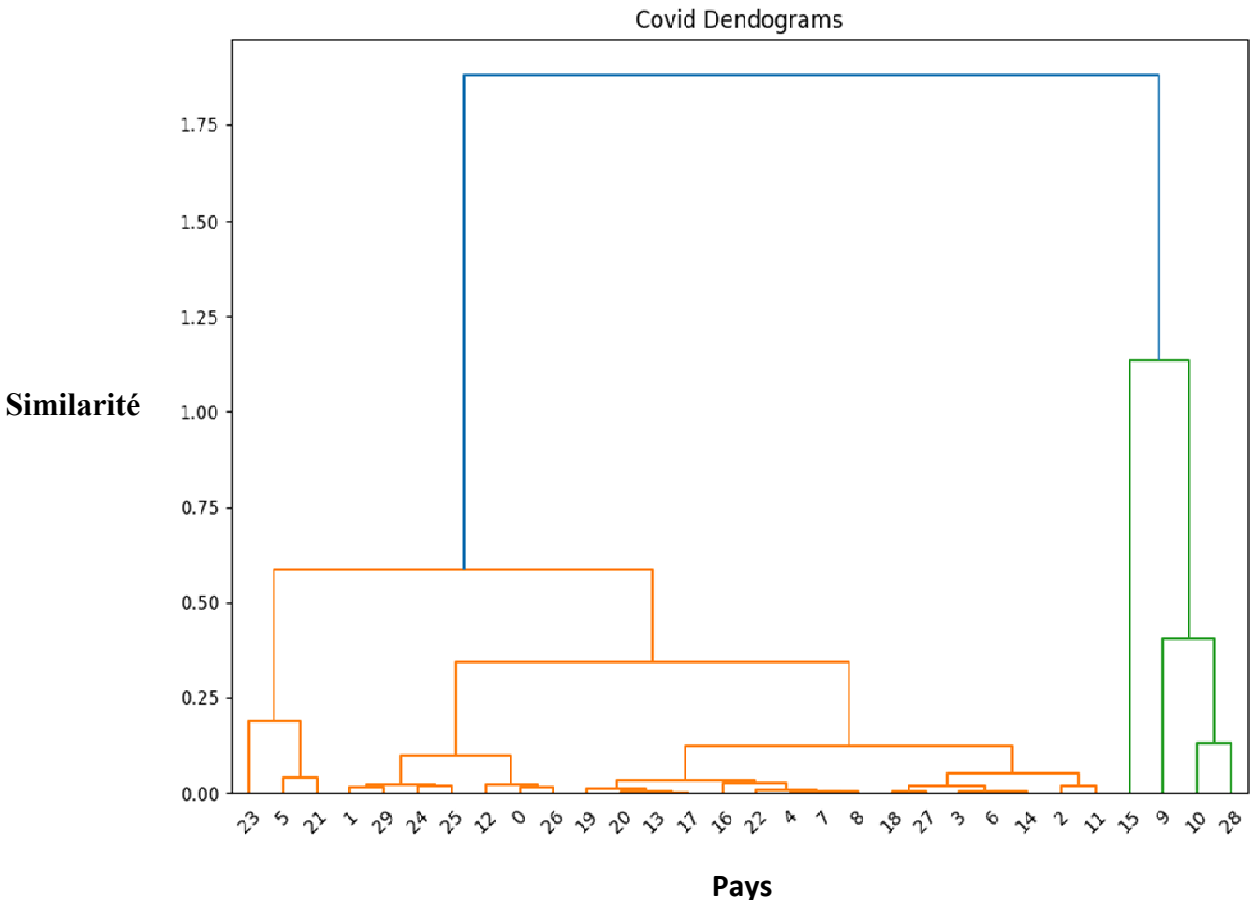

Figure III. 2. Dendrogramme covid19.

Un découpage en deux classes minimise ce critère. Cependant, si l'on souhaite réaliser une analyse un peu plus fine, un nombre de classes plus élevé serait pertinent. Nous allons donc retenir un découpage en 4 classes. Indiqués dans le tableau III.6.

Cluster	Pays
1	Italie
2	France
2	Allemagne
2	Espagne
3	Belgique
3	Bulgarie
3	Croatie
3	Chypre
3	Danemark
3	Estonie
3	Finlande
3	Grèce
3	Hongrie
3	Islande
3	Irlande
3	Lettonie
3	Liechtenstein
3	Lituanie
3	Luxembourg
3	Malte
3	Norvège
3	Portugal
3	Autriche
3	Suède
3	Slovaquie
3	Slovénie
3	Roumanie
4	Pologne
4	Pays-Bas
4	Tchéquie

Tableau III. 6. Les types de Clusters

Les groupes sont dessinés dans un plan à 2 dimensions sous forme de points colorés, pour faciliter la compréhension de leur répartition, Comme le montre la Figure III.3. Pour que l'axe des X représente les cas et l'axe des Y représente les décès.

Le code Python utilisé pour dessiner ce graphique est :

```
cluster = AgglomerativeClustering(n_clusters=4, affinity='euclidean',
linkage='ward')
cluster.fit_predict(covid)
plt.figure(figsize=(10, 7))
plt.scatter(covid[:,0], covid[:,1], c=cluster.labels_, cmap='rainbow')
```

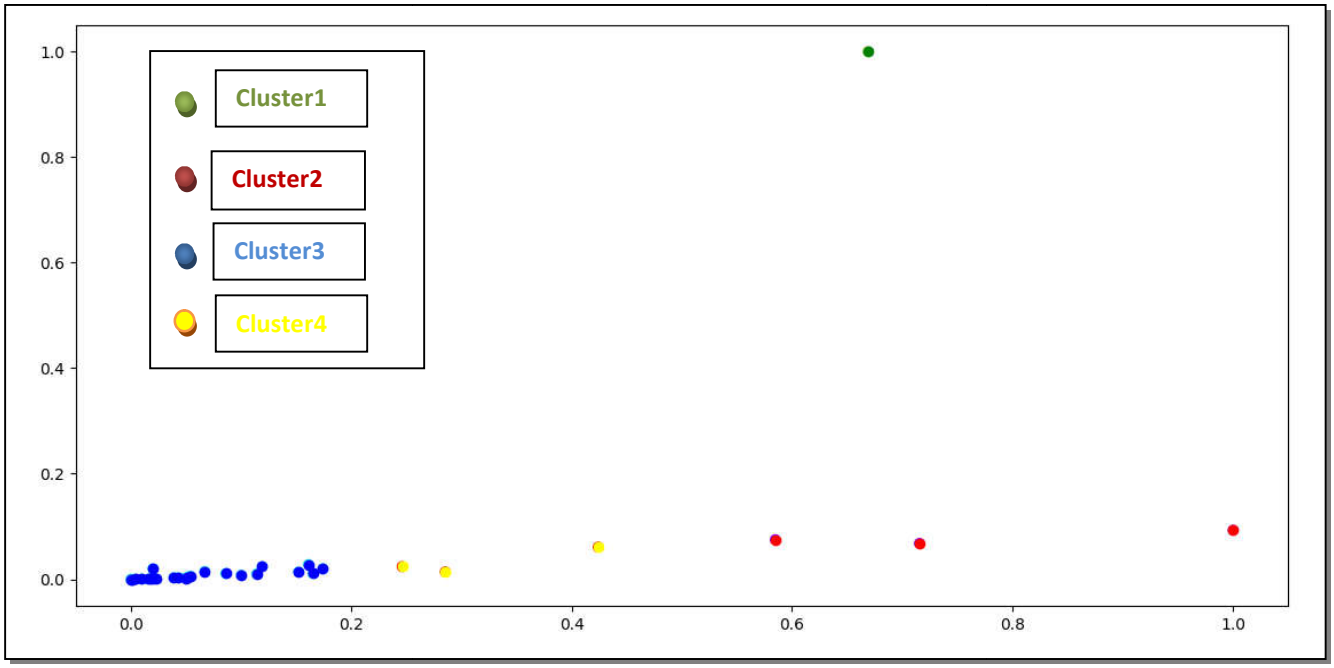

Figure III. 4. Représentation des clusters.

III.4.6 Analyse des résultats

Cette section présente les résultats des analyses visualisations de données, ainsi qu'une discussion des résultats obtenus.

- **Cluster 1 :**

C’est le cluster avec le pourcentage de cas le plus élevé dans les clusters et les décès moyens les plus élevés pour 100 000 habitants. Ce groupe comprend l’Italie la région est plus

durement touchée que partout ailleurs dans l'Union européenne (UE) par la Covid-19, La population moyenne de pays dans ce cluster est 56 923 524 personnes.

Il s'adonne que l'Italie fut le premier pays occidental à être touché par le coronavirus. Il n'y a pas vraiment de raison pour cela, mais cela implique deux choses. D'un, le cas italien semble pire en partie parce que la maladie a juste eu plus de temps pour se répandre. Alors de deux, il est aussi possible que l'éclosion précoce en Italie explique en partie pourquoi ce pays n'a pas été capable d'imposer des mesures d'isolement social à temps pour diminuer le nombre de gens atteints son gouvernement, ou sa population, ou les deux, auraient été plus ou moins pris par surprise. C'est du moins une hypothèse qui a été soulevée par divers experts dans le monde [65].

Alors que les statistiques relatives aux nombres de cas sont soumises aux aléas des politiques de test, plus ou moins restrictives selon les pays. La situation sanitaire de l'Italie, la gestion de la crise, ses effets secondaires mais également les perspectives de sortie de crise de l'Italie sont significatives à plus d'un titre, car elles représentent la première réponse de la part d'un Etat membre de l'Union européenne (UE) [65].

- **Cluster 2 :**

C'est le cluster avec les cas et décès moyens les plus élevés après l'Italie pour 100 000 habitants. Ce groupe comprend plusieurs pays ouest tels que la France, l'Allemagne et Espagne connus publiquement pour être fortement touchés par la pandémie. La population moyenne des pays dans ce cluster est 61 millions de personnes, Ce groupe contient des pays développés ainsi que des pays en développement.

L'Espagne prend des mesures de couvre-feu plus ou moins strictes sont mises en place. À compter du 15 mars, la France en fait autant le 17 mars, suivie de l'Allemagne le 22 mars.

Les mesures de strict confinement prises alors répondent en premier lieu à cette situation d'urgence, résultant du fait que ces États européens ne s'étaient pas sérieusement préparés à une crise épidémique de grande ampleur. Cette double situation de saturation des systèmes de santé et de confinement a des effets sur la santé de la population qui vont au-delà de ceux résultant directement de la Covid-19 : l'offre de soins diminue car de nombreux actes sont reportés dans le temps par les établissements hospitaliers, et dans le même temps la demande de soins diminue par crainte du coronavirus. Le confinement a aussi des impacts sur

la santé psychique d'une partie de la population, notamment les personnes âgées privées de visites [66].

- **Cluster 3 :**

C'est le cluster avec un nombre modéré de cas et de décès pour 100 000 habitants par rapport aux premiers groupes. Ce groupe comprend plusieurs pays Du nord et de l'ouest tels que la Belgique, Suède, Portugal, Roumanie, Hongrie, Autriche, Danemark, Grèce, Slovaquie, Luxembourg, malte, Islande, Liechtenstein, Lettonie, Irlande, Norvège, chypre, Estonie, Finlande, Lituanie, Slovénie, Croatie et Bulgarie. La population moyenne des pays dans ce cluster est 60 millions de personnes.

Les précautions sanitaires prises par les pays les moins touchés, des gestes barrières aux réductions de circulation, ont permis de faire baisser la mortalité ordinaire liée à la corona. « Il est important de souligner, cependant, qu'une surmortalité du Covid ou une surmortalité générale ne sont pas forcément synonymes de réponses gouvernementales moins efficaces face au virus », selon l'OCDE. Certains pays avaient un handicap de départ, du fait d'une population plus âgée, avec plus d'obésité, un secteur touristique plus développé, une population plus dense [67].

Le positionnement différent des pays en clusters en termes de cas et de décès suggère une relation possible entre la capacité à lutter contre la pandémie, à savoir la capacité de test, le vieillissement et les conditions de santé.

En fin, il est difficile de trouver des modèles sur le Vieux Continent. L'Estonie, la Finlande ou la Norvège s'en sortent un peu mieux à la fois sur le plan sanitaire et économique. Ils sont avantagés par une faible densité de population et « un niveau de confiance relativement élevé dans leur gouvernement ». Toutefois, aucun pays européen n'a aussi bien géré la pandémie [67].

- **Cluster 4 :**

C’est le cluster avec les décès moyens quotidiens les plus bas pour 100 000 habitants, ce cluster présente un profil moyen très plat sans vagues significatives. Ce groupe est principalement composé de Pologne, Tchéquie, Pays-Bas. La population moyenne des pays dans ce cluster est 20 millions de personnes.

En raison des mesures préventives suivies et du manque de densité et de mouvement de population à l'intérieur du pays par rapport à d'autres groupes, ces pays ont enregistré la moyenne la plus faible de cas et de décès.

Les institutions avaient préparé une communication de crise et le secteur de la santé publique à une épidémie. Le pays a aussi utilisé les nouvelles technologies pour surveiller les populations et retracer les cas contacts. La population a ainsi majoritairement respecté les mesures, au risque de se voir infliger des amendes particulièrement onéreuses.

III.5 Conclusion

Bien que COVID-19 soit une pandémie mondiale, ont été trouvés concernant à la fois les cas enregistrés et les profils temporels de décès pour chaque pays, lorsque des données mises à l'échelle et synchronisées ont été utilisées,

De plus, des clusters ont pu être trouvés par apprentissage non supervisé et ont été explorés, il existe des relations intéressantes mais Une grande variabilité a également été constatée dans les valeurs à l'échelle des cas par rapport aux décès observées dans les pays.

Les pays présentant un nombre plus élevé de cas pour 100 000 habitants ne sont que partiellement corrélée avec celles qui ont le plus grand nombre de décès pour 100 000 populations. Habituellement, les pays plus développés ont pu augmenter le nombre de tests par rapport aux moins développés, ces derniers souffrant également de pires conditions sanitaires ainsi que des mécanismes de réponse de santé publique plus faibles, mais ils ont également des populations plus jeunes, et donc certains effets peuvent compenser les autres.

Conclusion Générale

Conclusion générale

L'objectif principal de ce travail de thèse est celle d'analyser et divisées les pays européens en groupes en fonction à la fois de leurs cas et du nombre de décès à l'aide des techniques de science des données.

Notre étude est motivée par la grande importance du domaine de la fouille de données, Dans le domaine épidémiologique, nous avons souvent besoin de connaître certaines caractéristiques inconnues du modèle étudié. Plusieurs techniques d'estimation de ces derniers ont été proposées, tels que : la data mining, les modèles bayésiens, etc.

La méthode qu'on a utilisée pour faire face à ces problèmes est basée sur l'utilisation des techniques de science des données, telles que des visualisations de données et des tests statistiques, pour faire ce que l'on appelle souvent dans l'exploration de données, et également utilisé des techniques d'apprentissage automatique non supervisées.

Afin de réaliser les objectifs cités précédemment, nous avons présenté d'abord un aperçu sur les épidémies surtout le coronavirus et les efforts et techniques déployés pour surveiller et suivre la propagation de cette épidémie, Dans le deuxième chapitre, nous avons reconnaitre les différentes techniques de data mining afin d'avoir un aperçu complet sur eux, Dans le troisième chapitre, nous présentons, méthode basée sur les clustering hiérarchique pour analyser et divisées les pays européen en groupes en fonction à la fois de leurs cas et du nombre de décès. Le nombre total de clusters et les adhésions aux clusters de chaque pays sont déterminés par algorithme. Quatre grappes sont formées en appliquant les clustering hiérarchique sur les cas confirmés de COVID-19 et les cas de décès par COVID-19. Le groupe 1 se compose de 1 pays. Le groupe 2 contient les pays développés avec le deuxième pourcentage moyen de groupe le plus élevé pour les cas confirmés de COVID-19 et les cas de décès. Le groupe 3 comprend de 23 pays affichant le pourcentage moyen de groupe le plus élevé de cas confirmés de COVID-19 et de cas de décès de COVID-19. Le groupe 4 contient 3 pays en développement et a le pourcentage moyen le moins élevé de cas confirmés de COVID-19 et de cas de décès dus à COVID-19.

Les études futures devraient continuer à mettre à jour les informations et extraire d'autres observations et évolutions du profil temporel. De plus, des recherches ultérieures sont conseillées pour essayer de modéliser, par exemple, les décès par habitant avec un certain nombre de variables nationales et voir ce qui peut être conclu à partir de cette analyse de

modélisation. L'utilisation d'un ratio de population par zone (densité de population) peut aussi apporter une autre perspective assez intéressante et éclairante, puisque la contagion du COVID-19 est le fer de lance de la proximité entre humains.

Enfin, des études prévisibles devraient examiner quels impacts peuvent être dérivés du nombre de personnes par habitant qui ont reçu des vaccins et les profils de temps de vaccination correspondants à l'échelle et synchronisés.

Bibliographie

- [1]. Johns Hopkins University COVID-19 Map. Available online: https://coronavirus.jhu.edu/map.html (accessed on 31 December 2020).
- [2]. Nicola, M.; Alsafi, Z.; Sohrabi, C.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha, M.; Agha, R. The Socio-Economic Implications of the Coronavirus Pandemic (COVID-19): A Review. Int. J. Surg. 2020, 78, 185–193. [CrossRef] [PubMed]
- [3]. Pak, A.; Adegboye, O.A.; Adekunle, A.I.; Rahman, K.M.; McBryde, E.S.; Eisen, D.P. Economic Consequences of the COVID-19 Outbreak: The Need for Epidemic Preparedness. Front. Public Health 2020, 8. [CrossRef] [PubMed]
- [4]. Antonio, N.; Rita, P. March 2020: 31 Days That Will Reshape Tourism. Curr. Issues Tour. 2020, 1–16. [CrossRef]
- [5]. Sarkodie, S.A.; Owusu, P.A. Global Assessment of Environment, Health and Economic Impact of the Novel Coronavirus (COVID-19). Environ. Dev. Sustain. 2020. [CrossRef] [PubMed]
- [6]. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Deep Learning Applications for COVID-19. J. Big Data 2021, 8, 18. [CrossRef] [PubMed]
- [7]. Zohner, Y.E.; Morris, J.S. COVID-TRACK: World and USA SARS-COV-2 Testing and COVID-19 Tracking. BioData Min. 2021, 14. [CrossRef] [PubMed]
- [8]. Alvarez, E.; Brida, J.G.; Limas, E. Comparisons of COVID-19 Dynamics in the Different Countries of the World Using Time-Series Clustering. Health Econ. 2020. [CrossRef]
- [9] Han, J.; Kamber, M.; Pei, J. Data Mining: Concepts and Techniques, 3rd ed.; Elsevier: Waltham, MA, USA, 2012
- [10] Introduction : Antoine Gessain, Jean-Claude Manuguerra, Dans Les virus .émergents (2006), pages 3 à 10
- [11] Morens DM, Taubenberger JK, Fauci AS. The persistent legacy of the 1918 influenza virus, N Engl J Med, 2009, vol. 361 (pg. 109-13)
- [12] Cohen J. Here comes swine flu phase 6, severity 12 June 2009 Available at: http://blogs.sciencemag.org/scienceinsider/2009/06/swine-flu-who-r.html. Accessed 24 August 2009
- [13] https://www.futura-sciences.com/sante/actualites/medecine-peste-noire-serait-ancetre-toutes-pestes-actuelles-34081/
- [14] Claude Chastel (préf. François Denis), *Virus émergents : vers de nouvelles pandémies ?*, Paris, Vuibert, 2006, 316 p. (ISBN 978-2-7117-7198-1)
- [15] Desenclos J, Vaillant V, DelarocqueAstagneau E, Campèse C, Che D, Coignard B et al. Les principes de l'investigation d'une épidémie dans une finalité de santé publique. Med Mal Infect 2007.
- [16] Magnus M. Outbreak investigations. In: Essentials of infectious disease epidemiology. Boston- London: Jones & Bartlett; 2008.
- [17] Enserink M. Swine flu: WHO "really very close" to using the P word 9 June 2009 Available at: http://blogs.sciencemag.org/scienceinsider/2009/06/here-comes-phas.html. Accessed 24 August 2009
- [18] SELF-STUDY Course SS1978 : Principles of Epidemiology in Public Health Practice Third Edition An Introduction to Applied Epidemiology and Biostatistics, October 2006 Updated May 2012
- [19] Green MS, Swartz T, Mayshar E, Lev B, Leventhal A, Slater PE, Shemer J: When is an epidemic an epidemic. January 2002

- [20] Callow PP: Epidemic. The Encyclopedia of Ecology and Environmental Management. Oxford: Blackwell Science Ltd.
- [21] Martin PM, Martin-Granel E (June 2006): 2,500-year evolution of the term epidemic. Emerging Infectious Diseases.
- [22] Coronavirus disease 2019 (COVID-19) Situation Report – 94. Data as received by WHO from national authorities by 10:00 CEST, 23 April 2020
- [23] "Q&A on coronaviruses (COVID-19)". World Health Organization *(WHO). 17 April 2020.* Archived *from the original on 14 May 2020*. Retrieved 14 May 2020.
- [24] "Symptoms of Coronavirus". *U.S.* Centers for Disease Control and Prevention *(CDC). 13 May 2020.* Archived *from the original on 17 June 2020*. Retrieved 18 June 2020.
- [25]"Symptoms of Coronavirus". U.S. Centers for Disease Control and Prevention
- [34] J. Han, M. Kamber, and J. Pei. Data mining Second Edition: concepts and techniques. Morgan Kaufmann Pub.
- *[35]* ACM SIGKDD:Data Mining Curriculum. *2006-04-30*. Retrieved 2014-01-27
- [36] Clifton, Christopher (2010). Encyclopædia Britannica: Definition of Data Mining. Retrieved 2010-12-09
- [37] Hastie, Trevor*;* Tibshirani, Robert*;* Friedman, Jerome *(2009).* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Archived from* the original *on 2009-11-10*. Retrieved 2012-08-07.
- [38] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining: Pearson, 2005
- [39]U. M. Feyyad, "Data mining and knowledge discovery: making sense out of data," IEEE Expert, vol. 11, pp. 20-25, 1996
- [40] J. Han, M. Kamber, and J. Pei. Data mining Third Edition: concepts and techniques. Morgan Kaufmann Pub, 2011.
- [41] M. Kantardzic. Data mining: concepts, models, methods, and algorithms. WileyInterscience, 2003.
- [42] Margaret H.dunham.data mining: introductory and advanced topics.
- [43] https://www.kdnuggets.com/2019/09/hierarchical-clustering.html
- [44] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Clustering algorithms and validity measures, IEEE, 2001, pp.3-22
- [45] Minho Kim and R. S. Ramakrishna. New indices for cluster validity assessment. Pattern Recognition Letters, 26(15):2353–2363, 2005.
- [46] Johann M. Kraus, Christoph Mssel, Gnther Palm, and Hans A. Kestler. Multi-objective selection for collecting cluster alternatives. Computational Statistics, 26:341–353, 2011.
- [47]https://scikit-learn.org/stable/modules/clustering.html
- [48]https://datascientest.com/algorithme-des-k-means
- [49]https://scikit-learn.org/stable/modules/clustering.html#mean-shift
- [50] What is Data Mining in Healthcare? By David Crockett, Ryan Johnson, and Brian Eliason.
- [51] APPLICATION OF DATA MINING TO HEALTH CARE: Chirag Gandhi, Nakul Soni.
- [52] https://link.springer.com/article/10.1007/s12553-021-00553-7
- [53] (en) « *Python License* »
- [54]https://kyletk.com/index.php/2017/10/28/python-benefits-using-virtual-environment/

- [55] « *Matplotlib for Python Developers - Preface* » [archive], novembre 2009 (consulté le 22 janvier 2014)
- [56] https://www.educative.io/edpresso/what-is-pandas-in-python
- [57] https://docs.scipy.org/doc/scipy/reference/cluster.html
- [58] « *Release history — scikit-learn 0.19.dev0 documentation* »
- [59] https://www.consilium.europa.eu/fr/policies/coronavirus/
- [60] https://whatis.techtarget.com/fr/definition/preparation-des-donnees
- [61] https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/
- [62].Everitt BS, Hothorn T. A handbook of statistical analyses using R. 2nd ed. 2010.
- [63].Everitt BS, Landau S, Leese M, Stahl D. Cluster Analysis. 2011.
- [64]. Saraçli S, Do an N, Do an I. Comparison of hierarchical cluster analysis methods by cophenetic correlation. J Inequalities Appl [Internet]. 2013 Apr 23 [cited 2020 Sep 24]; 2013(1):1–8. Available from: https://link.springer.com/articles/10.1186/1029-242X-2013-203
- [65].https://www.lesoleil.com/actualite/vos-questions-sur-la-covid-19/covid-19-pourquoi-est-ce-pire-en-italie-2cb521bc75f5aef314cde825d11c3200
- [66] en) Giles Colclough, Penelope Dash, and Lieven Van der Veken, « *Understanding and managing the hidden health crisis of COVID-19 in Europe* » [archive], sur *McKinsey*, 2 juin 2020.
- [67] https://www.lesechos.fr/economie-france/social/quels-sont-les-pays-deurope-ou-le-covid-a-ete-le-plus-meurtrier-1266408

Résumé

Résumé - Le coronavirus a un indice de reproduction de base élevé (R0) et a provoqué la pandémie mondiale de COVID-19. Les gouvernements mettent en œuvre des mesures de confinement qui entraînent des retombées économiques dans de nombreux pays. Les décideurs politiques peuvent prendre de meilleures décisions s'ils disposent des indicateurs liés à la propagation de la maladie. La fouille de données est une technique essentielle dans le processus d'extraction de connaissances à partir de données. Cela nous permet de modéliser les connaissances extraites à l'aide d'un formalisme ou d'une technique de modélisation. Cette étude propose une méthode basée sur les clustering hiérarchique pour analyser et divisées les pays européen en groupes en fonction à la fois de leurs cas et du nombre de décès. Le nombre total de clusters et les adhésions aux clusters de chaque pays sont déterminés par algorithme. Quatre grappes sont formées en appliquant les clustering hiérarchique sur les cas confirmés de COVID-19 et les cas de décès par COVID-19. Le groupe 1 se compose de 1 pays. Le groupe 2 contient les pays développés avec le deuxième pourcentage moyen de groupe le plus élevé pour les cas confirmés de COVID-19 et les cas de décès. Le groupe 3 comprend de 23 pays affichant le pourcentage moyen de groupe le plus élevé de cas confirmés de COVID-19 et de cas de décès de COVID-19. Le groupe 4 contient 3 pays en développement et a le pourcentage moyen le moins élevé de cas confirmés de COVID-19 et de cas de décès dus à COVID-19. Les résultats produits peuvent être utilisés par les décideurs politiques pour prendre de meilleures décisions pour contrôler la pandémie. Cette analyse peut aider à mettre en évidence les politiques publiques les plus et les moins importantes pour minimiser le taux de mortalité COVID-19 d'un pays.

Mots clés – COVID-19, La fouille de données, clustering hiérarchique.

Abstract

Abstract - The corona virus has a high basal reproduction index (R0) and has caused the global COVID-19 pandemic. Governments are implementing containment measures that have economic fallout in many countries. Policymakers can make better decisions if they have indicators related to the spread of the disease. Data mining is an essential technique in the process of extracting knowledge from data. This allows us to model the extracted knowledge using formalism or a modeling technique. This study proposes a method based on hierarchical clustering to analyze and divide European countries into groups according to both their cases and the number of deaths. The total number of clusters and cluster memberships for each country are determined by algorithm. Four clusters are formed by applying hierarchical clustering on confirmed cases of COVID-19 and cases of death from COVID-19. Group 1 consists of 1 country. Group 2 contains the developed countries with the second highest group average percentage for confirmed COVID-19 cases and death cases. Group 3 comprises of 23 countries with the highest cluster average percentage of confirmed COVID-19 cases and COVID-19 death cases. Group 4 contains 3 developing countries and has the lowest average percentage of confirmed COVID-19 cases and death cases from COVID-19. The results produced can be used by policy makers to make better decisions to control the pandemic. This analysis can help highlight the most and least important public policies to minimize a country's COVID-19 death rate.

Keywords - COVID-19, Data mining, hierarchical clustering.

ملخص

ملخص - يحتوي فيروس كورونا على مؤشر تكاثر أساسي مرتفع (R0) وقد تسبب في انتشار جائحة كوفيد-19 العالمي. تقوم الحكومات بتنفيذ إجراءات احتواء لها تداعيات اقتصادية في العديد من البلدان. يمكن لواضعي السياسات اتخاذ قرارات أفضل إذا كانت لديهم مؤشرات تتعلق بانتشار المرض. يعد استخراج البيانات تقنية أساسية في عملية استخراج المعرفة من البيانات. هذا يسمح لنا بنمذجة المعرفة المستخرجة باستخدام شكليات أو أسلوب النمذجة. تقترح هذه الدراسة طريقة تعتمد على المجموعات الهرمية لتحليل الدول الأوروبية وتقسيمها إلى مجموعات وفقًا لحالاتها وعدد الوفيات. يتم تحديد العدد الإجمالي للمجموعات وعضوية المجموعات لكل بلد من خلال الخوارزمية. يتم تشكيل أربع مجموعات من خلال تطبيق المجموعات الهرمية على الحالات المؤكدة لـ كوفيد-19 وحالات الوفاة من كوفيد-19. تتكون المجموعة 1 من دولة واحدة. تحتوي المجموعة 2 على البلدان المتقدمة التي لديها ثاني أعلى متوسط نسبة لحالات كوفيد-19 المؤكدة وحالات الوفاة. تتألف المجموعة 3 من 23 دولة بها أعلى متوسط نسبة مئوية لحالات كوفيد-19 المؤكدة وحالات الوفاة كوفيد-19. تحتوي المجموعة 4 على 3 دول نامية ولديها أقل متوسط نسبة لحالات كوفيد-19 المؤكدة وحالات الوفاة من كوفيد-19. النتائج التي يتم الحصول عليها يمكن استخدامها من قبل صانعي السياسات لاتخاذ قرارات أفضل للسيطرة على الوباء. يمكن أن يساعد هذا التحليل في تسليط الضوء على أهم وأقل السياسات العامة أهمية لتقليل معدل وفيات كوفيد-19 في بلد ما.

الكلمات الرئيسية - كوفيد-19 ، التنقيب في البيانات ، التجميع الهرمي.