Rapport d'Analyse de Données COVID-19

Table des matières

1 Introduction

Le but de cette analyse est de comprendre les facteurs influençant le diagnostic du COVID-19 et les chances de survie des patients dans la population étudiée.

Nous cherchons a dégager une méthodologie pour analyser ces données et interpréter les relations significatives.

2 Caractéristiques et Défis du Dataset

Le jeu de données utilisé, fourni par le gouvernement mexicain, contient des informations anonymisées sur 1 048 576 patients, avec 21 caractéristiques, incluant notamment des antécédents médicaux et des symptômes associés au COVID-19.

2.1 Avantages et Limitations de la Taille du Dataset

L'un des atouts majeurs de ce jeu de données est sa taille considérable, avec plus d'un million d'observations (1 048 576 patients). Cette ampleur confère plusieurs avantages significatifs :

- **Robustesse statistique :** La grande taille de l'échantillon fiables même sur des observations plus subtiles.

- **Détection de patterns rares :** Le volume important de données permet d'identifier des schémas ou corrélations qui pourraient être invisibles dans un échantillon plus restreint.
- **Représentativité :** Un échantillon important offre une meilleure représentation de la population générale, réduisant les biais potentiels.

2.2 Présentation des données du dataset

Le dataset inclue des données sur les pré-conditions médicales et l'état de santé des patients. Ce dataset comporte 21 variables distinctes, détaillées ci-dessous :

- **Sexe (sex)** : Sexe du patient, une valeur binaire.
- **Âge (age)** : Âge du patient en années.
- **Classification (classification)** : Résultat du test COVID-19. Les valeurs de 1 à 3 signifient que le patient a été diagnostiqué positif au COVID-19 et les autres valeurs un diagnostique négatif.
- **Type de patient (patient type)** : Type de soin reçu par le patient. La valeur 1 indique que le patient est retourné à domicile, et la valeur 2 qu'il a été hospitalisé.
- **Pneumonie (pneumonia)** : Indique si le patient présente une inflammation des Poumons, binaire.
- **Grossesse (pregnancy)** : Indique si le patient est enceinte, binaire.
- **Diabète (diabetes)** : Indique si le patient est atteint de diabète binaire.
- **Maladie pulmonaire obstructive chronique (COPD)** : Indique si le patient est atteint de bronchopneumopathie chronique obstructive, binaire.
- **Asthme (asthma)** : Indique si le patient souffre d'asthme, binaire.
- **Immunosuppression (inmsupr)** : Indique si le patient est immunodéprimé, binaire.
- **Hypertension (hypertension)** : Indique si le patient est atteint d'hypertension artérielle, binaire.
- **Maladie cardiovasculaire (cardiovascular)** : Indique si le patient souffre d'une maladie cardiaque ou vasculaire, binaire.
- **Maladie rénale chronique (renal chronic)** : Indique si le patient est atteint d'une maladie rénale chronique, binaire.
- **Autres maladies (other disease)** : Indique si le patient souffre d'une autre maladie non listée, binaire.
- **Obésité (obesity)** : Indique si le patient est obèse, binaire.
- **Tabagisme (tobacco)** : Indique si le patient est fumeur, binaire.
- **Unité médicale (usmr)** : Spécifie le niveau de l'unité médicale ayant traité le patient (premier, deuxième ou troisième niveau).

- **Type d'institution médicale (medical unit)** : Type d'institution du système national de santé qui a fourni les soins.
- **Intubation (intubed)** : Indique si le patient a été placé sous respirateur artificiel, binaire.
- **Admission en soins intensifs (ICU)** : Indique si le patient a été admis en unité de soins intensifs, binaire.
- **Date de décès (date died)** : Indique la date de décès du patient le cas échéant, ou "9999-99-99" si le patient est vivant.

2.3 Spécificités des Données Médicales Binaires

La nature principalement binaire des variables dans ce dataset (présence/absence de symptômes ou conditions) impose des contraintes importantes sur les méthodes d'analyse applicables :

- **Restrictions sur les outils statistiques :** Plusieurs méthodes classiques d'analyse multivariée ne sont pas adaptées :
  - L'analyse en composantes principales (ACP) est peu pertinente sur des variables binaires.
  - Les techniques de régression linéaire standard sont moins appropriées.
- **Adaptations nécessaires :**
  - Utilisation du coefficient Phi plutôt que la corrélation de Pearson classique.
  - Recours à des méthodes spécifiques comme l'algorithme Apriori, particulièrement adapté aux données binaires.
  - Emploi de la distance cosinus plutôt qu'euclidienne pour le clustering.

Ces caractéristiques ont influencée nos choix méthodologiques tout au long de l'analyse.

3 Prétraitement des Données

L'analyse de ce jeu de données a été réalisée en utilisant plusieurs méthodes détaillées ci-dessous. Mais avant de pouvoir appliquer ces méthodes, nous avons dû commencer par un prétraitement des données pour traiter les valeurs manquantes et les incohérences.

Les données contenaient plusieurs valeurs manquantes, codées par les valeurs 97 et 99. Un nettoyage initial a été effectué pour :

- Remplacer les valeurs 97 et 99 par des valeurs nulles (NA).
- Supprimer ou imputer les valeurs manquantes en fonction de la distribution des variables.
- Encoder les variables binaires (où 1 = "oui" et 0 = "non") pour faciliter l'analyse.

Ce nettoyage a permis de garantir la qualité des données pour les étapes suivantes d'analyse et de modélisation.

4 Analyse de Corrélation

4.1 Méthodologie

Une analyse de corrélation a été effectuée pour évaluer les relations entre les variables binaires présentes dans le dataset. Cette analyse a permis de déterminer les associations potentielles entre différentes caractéristiques, en utilisant le coefficient de corrélation de Phi.

Contrairement à un coefficient de corrélation standard, qui pourrait être faussé par les distributions non continues de ces données, le coefficient de Phi permet d'obtenir une mesure plus fiable de l'association linéaire entre deux variables binaires.

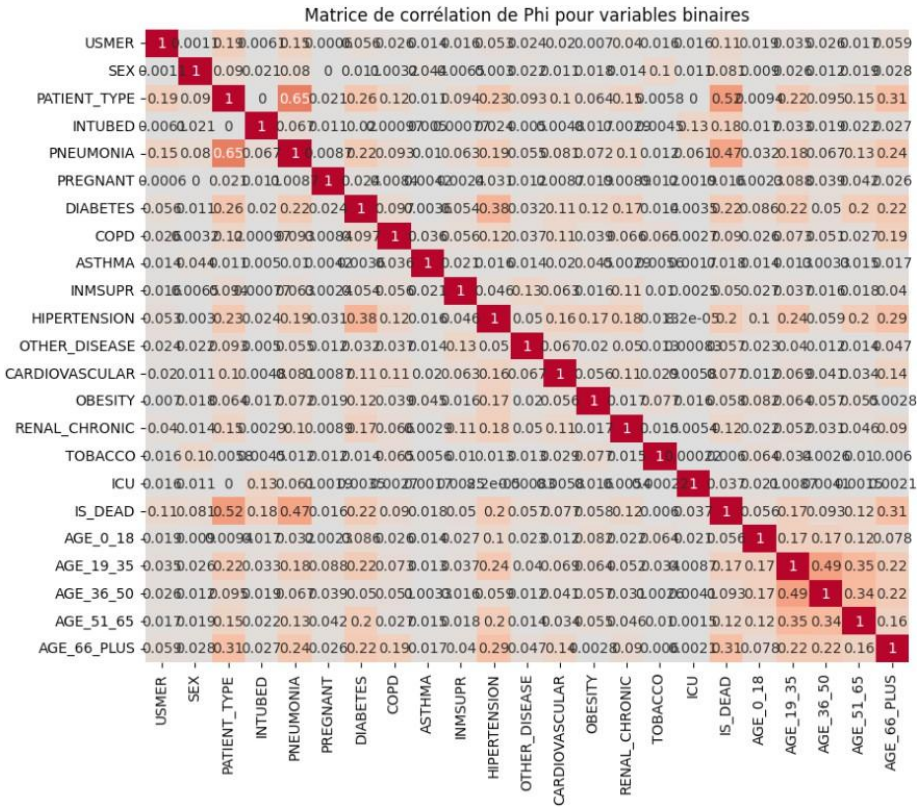

Figure 1: Matrice de corrélation Phi

4.2 Analyse

Dans cette matrice de corrélation, plusieurs facteurs se distinguent. Notamment, on observe une corrélation relativement forte entre le décès du patient et un diagnostic de COVID-19 (caractéristique PATIENT_TYPE), avec une valeur de 0,52, ainsi qu'entre le décès du patient et un diagnostic de pneumonie, avec une corrélation de 0,47. Ces relations suggèrent que ces deux facteurs pourraient avoir un impact significatif sur les chances de survie des patients.

On remarque également une forte corrélation (0,65) entre le diagnostic de pneumonie et celui de COVID-19, ce qui est logique étant donné que les formes graves de COVID-19 peuvent entraîner une pneumonie. Une autre corrélation notable existe entre un âge supérieur à 66 ans et plusieurs facteurs de risque, tels que le diabète, l'hypertension, la présence de COPD (bronchopneumopathie chronique obstructive), ainsi que les cas de décès. Bien que ces corrélations soient plus faibles dans les tranches d'âge plus jeunes,

on observe tout de même que les groupes d'âge de 51 à 65 ans et de 19 à 35 ans sont plus touchés par ces facteurs de risque que les groupes de 36 à 50 ans et de 0 à 18 ans.

Enfin, certaines corrélations notables, bien que non directement liées au COVID-19, sont également visibles, comme celle entre l'hypertension et le diabète, ou entre l'obésité et l'hypertension. Bien que pertinentes, ces corrélations ne font pas partie du sujet central de notre analyse.

figure 14,figure 15,figure 16,figure 17

5 Clustermap

La clustermap est une méthode de visualisation qui permet de classifier des groupes d'individus ou de caractéristiques similaires en fonction de leurs valeurs. Cette représentation graphique est particulièrement utile pour explorer et identifier les relations et les structures sous-jacentes des données. En appliquant des techniques de regroupement elle organise les données de manière à mettre en évidence les patterns et les similarités au sein du jeu de données. Les groupes de couleurs similaires dans la clustermap indiquent un degré élevé de similarité entre les éléments, qu'il s'agisse d'individus ou de caractéristiques. Par exemple, si deux individus partagent des caractéristiques proches, ils seront regroupés sous une même couleur, ce qui suggère qu'ils présentent des profils similaires en termes de variables mesurées.

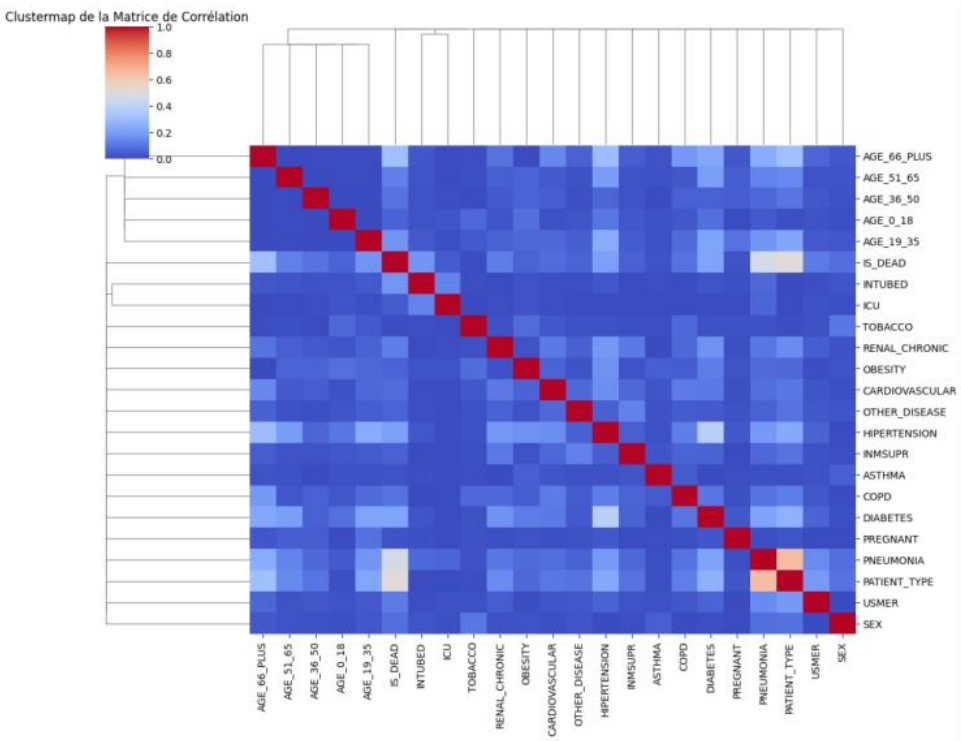

Figure 2: Clustermap

Exemples de groupes observés

- *Hypertension et diabète*
- *PATIENT_TYPE et PNEUMONIA*
- *PATIENT_TYPE, PNEUMONIA et IS_DEAD*

6 Réduction de la Dimensionnalité

Dans le cadre de l'analyse des données, une normalisation a été appliquée pour préparer les données à des analyses ultérieures. Étant donné que les données sont de nature binaire, l'utilisation de la distance cosinus s'est avérée plus appropriée que la distance euclidienne pour calculer la matrice de dissimilarité.

Pour la réduction de la dimensionnalité, la technique de Multi-Dimensional Scaling (MDS) a été choisie.Cette méthode permet de projeter les données dans un espace de dimension inférieure tout en préservant les relations de dissimilarité entre les observations. En projetant le dataset en deux dimensions, MDS facilite la visualisation des similarités et des différences entre les observations, permettant ainsi d'identifier des regroupements naturels ou des tendances qui seraient difficilement visibles dans un espace de plus grande dimension.

figure 18

7 Clustering

Dans cette section, l'objectif est de segmenter les données en groupes distincts en utilisant la méthode KMeans. Le choix du nombre optimal de clusters est essentiel pour garantir une séparation significative des données.

7.1 Détermination du Nombre Optimal de Clusters

Pour déterminer le nombre optimal de clusters dans notre analyse, la méthode du coude a été appliquée. Cette approche a permis de visualiser l'inertie en fonction du nombre de clusters, en traçant la courbe d'inertie pour différents nombres de groupes. Après évaluation de cette courbe, le choix de trois clusters a été retenu. Ce nombre représente un compromis entre une bonne séparation des données et la simplicité du modèle, facilitant une interprétation claire des résultats.

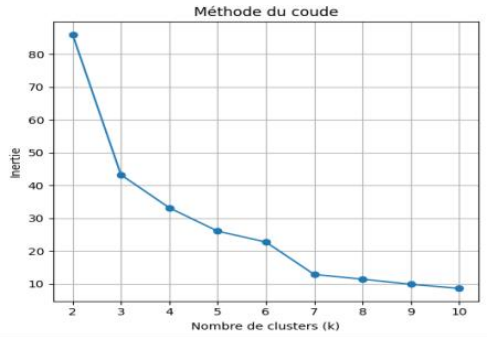

Figure 3: Méthode du coude

7.2 Implémentation des Clusters

7.3 Clusters

L'implémentation de la méthode KMeans a été choisie pour segmenter les données en trois clusters. Cette approche permet de classifier les observations en groupes distincts, facilitant ainsi l'identification de patterns et de similarités au sein de l'ensemble de données.

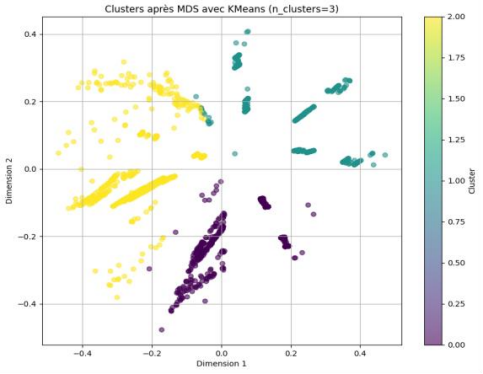

Figure 4: Répartition des données en trois clusters.

7.4 Analyse des Valeurs Influentes par Cluster

Pour chaque cluster, une analyse des valeurs des variables par rapport à leur moyenne a été réalisée. Les histogrammes ci-dessous illustrent les résultats, permettant d'identifier, pour chaque cluster, les variables dont les valeurs dépassent leur moyenne. Cette approche aide à comprendre les caractéristiques distinctives de chaque cluster en repérant les variables qui se démarquent par des valeurs plus élevées.

figure 5,figure 6,figure 7

Cluster	Variables Influentes
Cluster 0	Pneumonie, Diabète, Hypertension, Obésité, Âge 51-65
Cluster 1	Asthme, Âge 0-18, Âge 19-35
Cluster 2	Pneumonie, Diabète, Immunosuppression, Hypertension, Autres maladies, Cardiovasculaire, Obésité, Insuffisance rénale, Maladie chronique, ICU, Décédé, Âge 66 et plus

Les histogrammes ci-dessous illustrent les valeurs influentes identifiées pour chaque cluster.

7.5 Le score de silhouette obtenu

Le score de silhouette obtenu, avec une valeur moyenne de 0,5, indique une séparation modérée entre les groupes formés. Ce coefficient de silhouette évalue la cohésion au sein de chaque cluster et la séparation entre les clusters, ce qui suggère ici une structure de groupe relativement bien définie. Les valeurs de silhouette des différents clusters sont globalement similaires, ce qui signifie que les clusters sont d'une qualité relativement homogène.

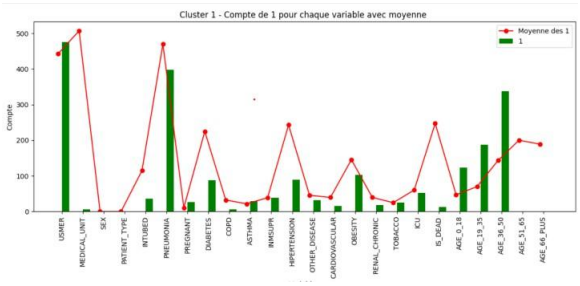

Figure 5: Variables influentes pour le Cluster 1.

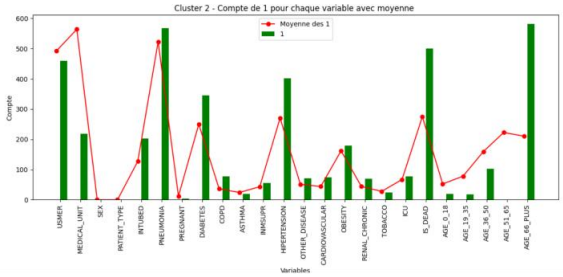

Figure 6: Variables influentes pour le Cluster 2.

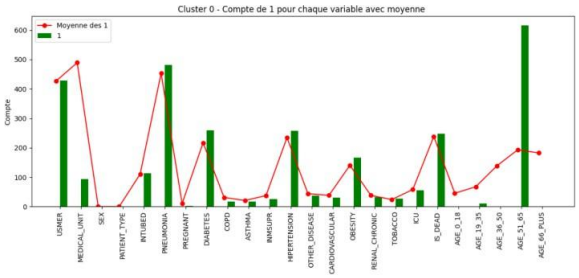

Figure 7: Variables influentes pour le Cluster 0.

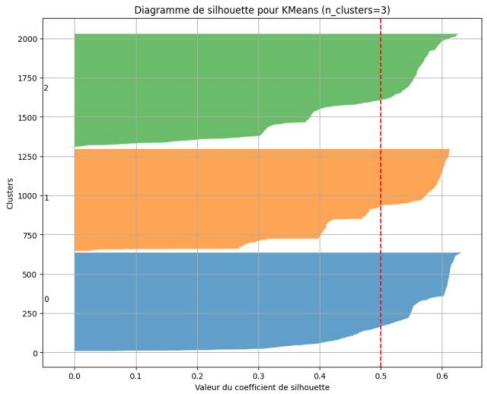

Figure 8: Score de silhouette des clusters.

7.6 Conclusion sur le Clustering

L'analyse révèle que la variable IS DEAD est particulièrement présente dans le Cluster 2, qui regroupe majoritairement des individus âgés de 66 ans et plus, souvent atteints de multiples maladies graves. Cette association met en lumière le lien entre l'âge avancé, les affections chroniques multiples, et un risque accru de décès. Le profil de ce cluster souligne la nécessité de soins intensifs et d'une vigilance accrue pour cette population, qui constitue le groupe le plus vulnérable de l'étude.

8 Comparaison d'histogrammes

8.1 Méthodologie

Cette partie de l'analyse se concentre sur la comparaison des variables corrélées, en fonction des distributions d'âge. Nous avons choisi cette méthode car l'âge est la seule variable du dataset qui peut raisonnablement être représentée sous forme d'histogramme. Il est cependant possible de créer des histogrammes pour d'autres variables en fonction de l'âge, par exemple, pour analyser plus en détail la relation entre l'âge et les décès, nous pouvons comparer l'histogramme de l'âge des patients décédés avec celui de la population générale. Ensuite, nous calculons la courbe de densité de probabilité (KDE) pour chaque histogramme et réalisons une analyse basée sur la forme des courbes ainsi que sur le coefficient de corrélation de Pearson.

Le coefficient de corrélation de Pearson mesure la force et la direction de la relation linéaire entre deux variables continues. Dans le contexte de la comparaison des KDE, il permet d'évaluer dans quelle mesure les distributions d'âge des différents groupes (par exemple, les patients décédés et les patients vivants) sont similaires ou différentes.

8.2 Analyse des résultats

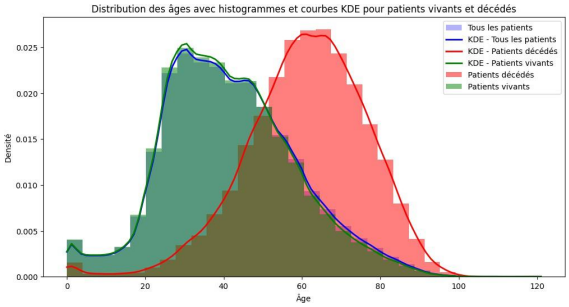

Figure 9: Histogrammes et KDE des âges des personnes décédées, vivantes et de la population générale

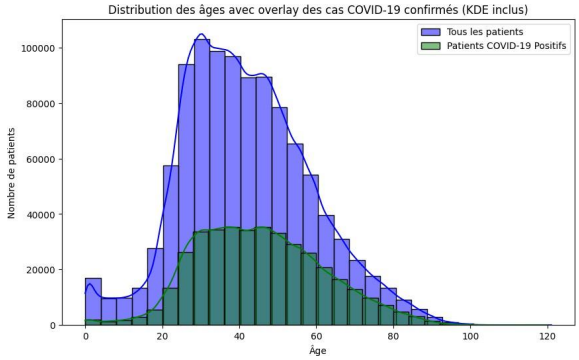

Figure 10: Histogrammes et KDE des âges des patients diagnostiqués du COVID et de la population générale

Age et Décès Dans la figure 9, on observe que la distribution des décès est nettement plus concentrée autour des âges proches de 65 ans. On peut en déduire que ce sont les personnes dans cette tranche d'âge qui ont constitué la majorité des décès. Le décalage clair des décès par rapport à l'histogramme de l'âge de la population générale indique que l'âge est un facteur déterminant dans la survie des patients.

La corrélation de Pearson entre les KDE des patients décédés et tous les patients est de 0.30, ce qui indique une faible similitude, confirmant l'interprétation visuelle. La corrélation parfaite de 1.00 entre les KDE des patients vivants et tous les patients est attendue, car les patients vivants représentent la majorité du groupe total. Enfin, la corrélation de 0.27 entre les patients décédés et vivants montre une faible similitude dans leurs distributions d'âge, ce qui souligne encore que l'âge joue un rôle important dans les chances de survie des patients.

Âge et diagnostic COVID La figure 10 montre, à première vue, une grande similarité entre la distribution des âges des personnes diagnostiquées avec la COVID-19 et celle de la population générale. Cependant, la corrélation de Pearson entre les deux KDE

est de 0.1551, ce qui indique une faible similarité. Cette différence peut s'expliquer par la présence d'un pic dans la population générale autour des 25-30 ans, tandis que la distribution des patients diagnostiqués avec la COVID-19 est relativement constante entre 30 et 50 ans. Il semble donc que la tranche d'âge des 25-30 ans ait été davantage représentée par des personnes se rendant à l'hôpital pour d'autres raisons que la COVID-19 ou pensant être infectées sans l'être. Il est également possible que cette tranche d'âge spécifique soit plus difficile à diagnostiquer pour la COVID-19.

9 Analyse frequent pattern et Apriori

L'algorithme Apriori est utilisé ici pour générer des règles d'association, permettant d'identifier les cooccurrences fréquentes entre différentes variables de notre jeu de données.

Nous avons choisi cette méthode car elle est adaptée aux datasets principalement composés de données binaires. Elle nous permet notamment de générer un graphe qui offre une vision plus claire des relations entre les différentes caractéristiques.

9.1 Traitement des données

Avant d'appliquer l'algorithme Apriori, les variables doivent être converties en format binaire. La plupart des données étant déjà dans ce format, peu de transformations ont été nécessaires. Cependant, pour la variable MEDICAL_UNIT, qui représente l'unité médicale où le patient a été traité, une transformation a été nécessaire afin de l'adapter au format requis par l'algorithme. Nous avons aussi séparer l'age en tranches d'ages de 20 ans.

9.2 Utilisation du Leverage comme métrique

Nous avons choisi la métrique de *leverage* pour évaluer les règles d'association. Le *leverage* permet de détecter des relations significatives en filtrant les associations qui pourraient être peu informatives malgré des valeurs élevées de support ou de confiance. Cette approche est particulièrement utile pour identifier les relations importantes entre les attributs de santé dans notre analyse.

9.3 Graph

Un graphe (figure 11) a été utilisé pour visualiser les règles d'association obtenues avec l'algorithme Apriori. Chaque nœud représente un attribut, et chaque arête correspond à une règle d'association entre les antécédents et les conséquents. Le poids de chaque arête est défini par la valeur de la métrique *leverage* et est représenté par la couleur de l'arête, permettant ainsi de visualiser l'importance relative de chaque règle d'association.

9.4 Analyse

9.4.1 Analyse générale

On observe dans le graphe 11 que la relation ayant le leverage le plus fort est entre Patient_type et Pneumonia. Patient type inque si le patient a été hospitalisée ou si il est retournée chez lui. On peut donc en déduire que la présance de pneumonie est le facteur le plus important important d'hositalisation.

On vois aussi que le sex du patient semble avoir un impact direct sur la probabilitée que le patient ait été diagnositquer du covid. La plupart des relation sont relativement attendu comme l'unitée médical 12 et le diagnostique qui semblerais indiquer que cette unitée était spécialisée dans le traitement des cas covid. Mais on remarque la relation

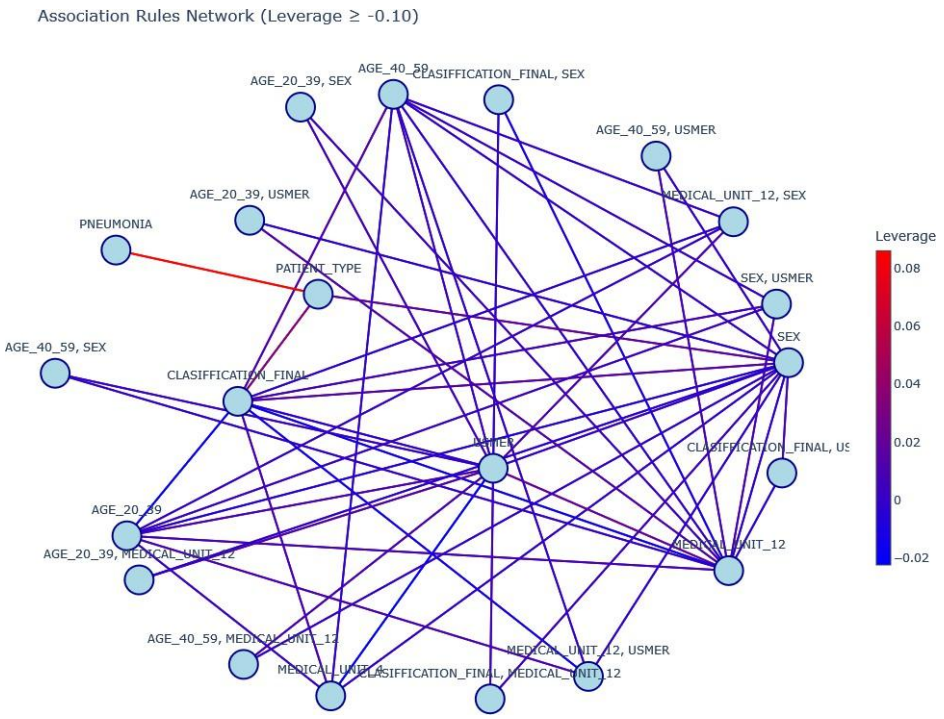

Figure 11: graphe des règles généré par Apriori

(SEX)-¿(CLASIFFICATION FINAL) qui a un lien avec le diagnostique covid. De plus le sex est présent dans plusieur règles composée en lien avec la classification final. La règle d'association entre le sexe et le diagnostique covid a une confiance de 40%, indiquant une relation modérée dans la mesure ou on a 50.07% de femmes dans le dataset. Le *lift* de 1.070 suggère une légère influence du sexe sur la classification, bien que l'effet soit faible. Le *leverage* de 0.013 indique une faible augmentation des observations par rapport à l'indépendance, rendant cette règle pertinente mais non distinctive.

9.5 Analyse de la règle (SEX) ⇒ (*CLASIFFICATION_FINAL*)

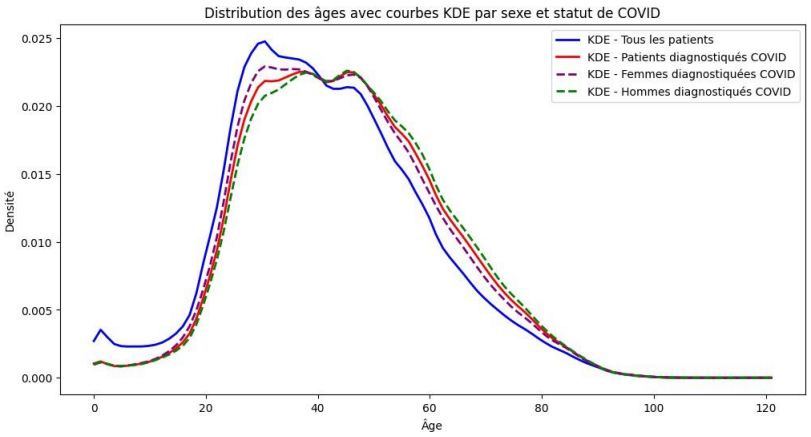

Figure 12: Distribution des âges avec courbes KDE par sexe et statut de COVID

Nous avons également calculé un ensemble de statistiques concernant la population

selon le sexe, qui sont présentées dans le tableau 1.

	Femme	Homme	Différence
Taux de Survie tout patients	97.93%	96.25%	1.68%
Taux de Survie (cas de COVID)	95.91%	93.41%	2.50%
Âge Moyen	41.32	42.27	0.95
Pourcentage de la population	51.02%	48.98%	2.04%

Table 1: Résumé des taux de survie, de l'âge moyen, et des pourcentages par sexe dans le Dataset, avec différence entre les sexes.

Pour mieux comprendre cette relation, nous pouvons revenir à notre méthode d'analyse initiale (vue dans la section 8) en créant un histogramme de la distribution des cas de COVID par âge et par sexe. Cela nous donne le graphe de la figure 12. Grâce à ce graphe, on peut observer une légère différence dans l'âge des personnes diagnostiquées avec le COVID en fonction de leur sexe. En effet, il y a une différence notable d'infections entre les hommes et les femmes autour de 30 ans, où les femmes sont largement plus victime d'infections

En examinant les taux de survie par sexe, on remarque que les femmes sont légèrement plus susceptibles de survivre à leur hospitalisation que les hommes (1.68% de plus), ce qui explique en partie la règle *(SEX) → (CLASSIFICATION FINAL)*. Cette tendance est encore plus marquée chez les patients atteints de COVID (2.5% de chances supplémentaires de survie).

En prenant en compte ces informations, nous pouvons formuler l'hypothèse suivante quant à la cause de cette différence : dans la population observée à l'hôpital, les femmes ont tendance à être plus jeunes en moyenne (41.32 ans pour les femmes contre 42.27 ans pour les hommes). Or, il est bien établi que les patients plus âgés sont plus susceptibles de mourir du COVID. Il semble donc que la différence de taux de survie entre les hommes et les femmes puisse être partiellement expliquée par cette différence d'âge moyen dans l'échantillon.

Cependant, cette différence d'âge moyen ne semble pas être explicable par notre dataset. De plus, dans la population générale du Mexique, les hommes ont tendance à être plus jeunes que les femmes, avec un âge médian de 28 ans pour les hommes contre 30 ans pour les femmes en 2020 (source : "https://www.statista.com/").

10 Conclusion

Cette analyse met en évidence plusieurs éléments clés concernant les facteurs démographiques et cliniques influençant la gravité et la mortalité liées à la COVID-19. L'âge se révèle être un facteur déterminant, les patients de plus de 66 ans présentant des taux plus élevés de pneumonie, d'admission en soins intensifs et de décès par rapport aux plus jeunes. Les comorbidités telles que le diabète, l'hypertension et l'obésité sont également associées à des issues graves de la COVID-19, en particulier chez les patients âgés. Une légère disparité de mortalité entre les sexes est observée, les hommes montrant un taux de mortalité légèrement plus élevé, possiblement en raison de différences d'âge au sein de la population étudiée.

L'analyse de clustering identifie trois groupes de patients distincts : un groupe à haut risque composé de patients avec de multiples comorbidités et une mortalité élevée, un groupe à faible risque comprenant de jeunes patients, et un groupe intermédiaire. Enfin, l'extraction de règles d'association met en évidence le lien étroit entre le diagnostic de COVID-19 et la pneumonie, celle-ci constituant un facteur clé de l'hospitalisation.

En somme, cette étude apporte des informations pertinentes pour mieux comprendre les facteurs de risque et orienter des stratégies de prévention et de traitement ciblées, particulièrement pour les populations les plus vulnérables.

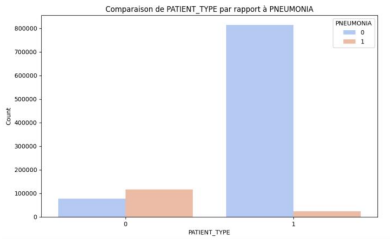

Figure 13: PATIENTTYPE et PNEUMONIA

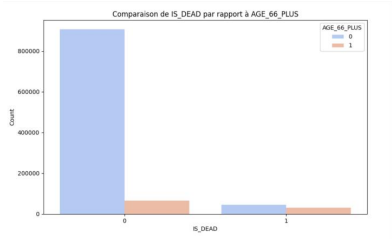

Figure 14: AGE66PLUS et ISDEAD

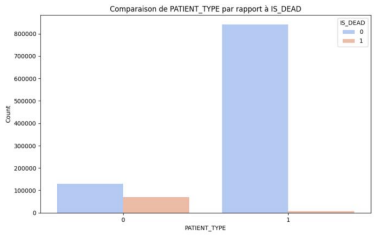

Figure 15: ISDEAD et PATIENTTYPE

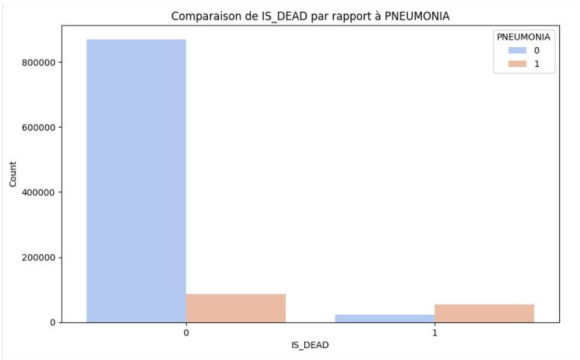

Figure 16: ISDEAD et PNEUMONIA

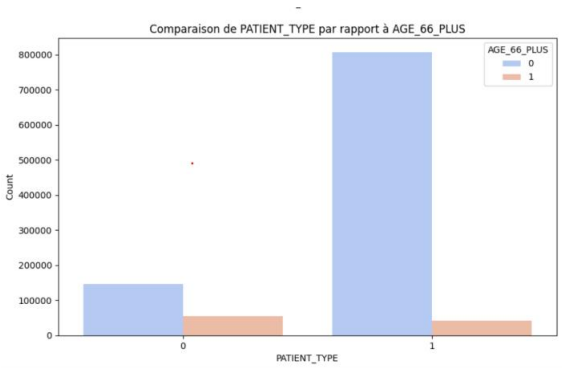

Figure 17: AGE66PLUS et PATIENTTYPE

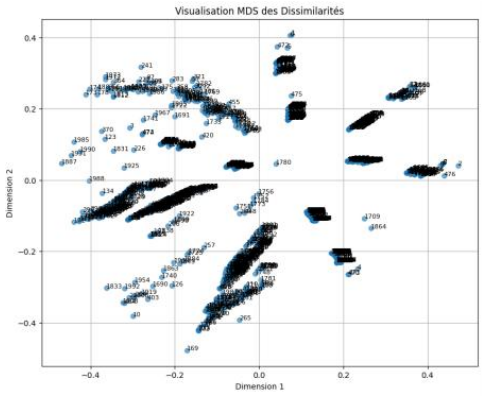

Figure 18: MDS