Projet Analyse de donnée

Membres :
Quentin POTIRON
Coumba Bocar KANE

1. Description de notre jeu de données

Nous avons choisi le dataset “owid-covid-data.csv” (Our World in Data) car il présente des informations assez récentes, entre 2020 et 2024, concernant la pandémie du COVID 19. Ces données sont récoltées et mises en ligne par une association anglaise à but non lucratif qui a pour but la collecte et l’analyse de données. La plupart des données et projets sont open source et sont d’origine assez vérifiées.

Le dataset est plutôt complet avec plus de 400 000 lignes et une soixantaine de variables par pays et par date. Il y a tous les pays reconnus par l’ONU en plus d’avoir directement fait les calculs par continent, pour la terre entière et par catégorie de pays (développés, en voie de développement, …).

De plus, le sujet étudié nous intéresse particulièrement car relève du domaine médical. L’étude de ce type de données peut aider à la gestion de futures catastrophes sanitaires mondiales ou plus locales dont il est, d’après plusieurs études, assez probable de voir apparaître de nouveau à cause du dérèglement climatique.

2. Etude des données

Comme il y a un grand nombre de variables différentes, toutes n’ont pas été traitées dans cette étude. De plus, certaines variables sont redondantes, c’est-à-dire qu’elles peuvent être calculées à partir d’autres variables (Ex : nombre de cas, nombre de cas par million et population).

Nous nous sommes surtout concentrés sur l’étude des variables “nombre de cas” et “nombre de décès”. Des graphes similaires peuvent être entrepris pour d’autres variables : ‘nombre de tests”, “nombre d’hospitalisations”, etc…

3. Différents Graphes

Nous avons commencé par faire des graphes simples afin d’appréhender les données comme celui montrant l'évolution d’une certaine variable pour un pays dans le temps. Ensuite, pour une comparaison par pays, à une date donnée, sur une carte géographique, est affiché le nombre de cas et le nombre de décès suivant la couleur et la taille du point.

Ensuite, afin de mieux étudier les différences entre pays, il y a la représentation graphique de la répartition des pays suivant un type de variables ainsi que l’affiche de boxplot sur différentes périodes. Afin d’éviter de rendre le graphique illisible avec une

surabondance de boxplots, il a été décidé d'étudier directement l'évolution des différents paramètres (moyenne, médiane, percentile).

4. Clustering

Pour cette tâche de clustering, il a été question de regrouper les pays qui se ressemblent du point de vue de l'épidémie et de leur situation économique. Cela permet de trouver des groupes de pays avec des caractéristiques communes (exemple des pays avec beaucoup de cas mais peu de décès, des pays avec beaucoup de décès et peu d'hôpitaux).

Nous avons implémenté et comparé trois méthodes de clustering :

- K-means : Cette méthode crée des groupes de pays qui se ressemblent le plus possible à l'intérieur de chaque groupe. On a illustré deux versions : une avec une valeur de k automatique et une autre avec une valeur de k fixe.
- Gaussian Mixture Models (GMM) : Cette méthode suppose que chaque groupe de pays suit une distribution en forme de cloche (gaussienne). Elle donne une probabilité qu'un pays appartienne à chaque groupe et peut identifier des groupes de forme ovale.
- DBSCAN : Cette méthode regroupe les pays selon leur proximité. Elle fonctionne bien pour trouver des groupes de formes irrégulières et pour identifier les pays "différents" qui n'appartiennent à aucun groupe (les valeurs aberrantes).

Nous utilisons la PCA pour la visualisation 2D des clusters et des métriques d'évaluation : silhouette score et inertie / elbow.

5. Indice de développement humain

L'indice de développement humain est une représentation simplifiée du niveau de vie des habitants d'un pays. Il a donc été réfléchi de s'il y avait un lien entre l'indice de développement et le nombre de cas du Covid-19. Cette représentation a donc été faite avec l'aide des résultats du clustering.