

RAPPORT D'ANALYSE



Catégorisation de Pays pour HELP International

RÉALISÉE PAR

Ezzagrani Habiba

Table des matières :

I - Introduction :.....	2
1. Contexte du projet & Problématique :.....	2
2. Objectifs de l'analyse :.....	2
II - Choix de l'Ensemble de Données :.....	2
1. Raisons du choix de l'ensemble de données de HELP International :.....	2
2. Description des attributs et de leur signification :.....	3
III - Objectifs de l'analyse :.....	4
1. Définition des objectifs principaux :.....	4
2. Précision sur l'utilisation de techniques d'apprentissage non supervisé :.....	4
3. Regroupement :.....	5
IV - Exploration des Données :.....	5
1. Résumé des mesures prises pour nettoyer et préparer les données :.....	5
2. Visualisations préliminaires pour comprendre les distributions et les tendances :	6
V - Normalisation des données :.....	8
VI - Modèles d'Apprentissage Non Supervisé :.....	11
1. Entraînement de différentes variantes de modèles pour le Clustering :.....	11
2. Description des différentes approches de modélisation utilisées :.....	11
VII - Benchmarking & Résultats:.....	12
VIII - Sélection du Modèle Final :.....	19
IX - Environnement du travail :.....	19
X - Conclusions :.....	20
XI - Suggestions pour les Prochaines Étapes :.....	21
Annexe :.....	21

I - Introduction :

1. Contexte du projet & Problématique :

Dans ce projet de catégorisation des pays, HELP International est une organisation humanitaire d'envergure internationale, se trouve confrontée à une tâche complexe et importante. Ayant réussi à mobiliser une somme considérable, estimée à environ 10 millions de dollars, la direction de cette organisation cherche à élaborer une stratégie d'allocation de ces fonds, visant à apporter une aide efficace aux nations les plus défavorisées. Cette initiative s'inscrit dans un souci d'optimisation et d'impact maximal de l'assistance humanitaire offerte.

2. Objectifs de l'analyse :

Le principal objectif de cette analyse consiste à opérer une catégorisation pertinente des pays en utilisant une série d'indicateurs socio-économiques et de santé. Cette démarche vise à évaluer de manière exhaustive le niveau de développement global de chaque nation. L'objectif sous-jacent à cette catégorisation est de pouvoir identifier avec précision les pays qui nécessitent de façon pressante une assistance spécifique et personnalisée. L'analyse de ces indicateurs permettra de cibler efficacement les régions les plus vulnérables et de mettre en œuvre des solutions adaptées pour une aide humanitaire adéquate et prompte.

II - Choix de l'Ensemble de Données :

1. Raisons du choix de l'ensemble de données de HELP International :

Le jeu de données de HELP International contient des attributs cruciaux tels que le taux de mortalité infantile, les dépenses de santé, les revenus, etc., offrant une vue globale des conditions socio-économiques et de santé de chaque pays. Ces données permettent une évaluation approfondie des besoins de chaque pays en termes de développement.

The screenshot shows the Microsoft Power BI desktop interface. The ribbon menu is visible at the top with tabs like Accueil, Insertion, Mise en page, Formules, Données (highlighted), Révision, Affichage, and Partager. Below the ribbon are several icons for data management: Obtenir des données (Power Query), Actualiser tout, Types de données, Trier et filtrer, Outils de données, Analyse de scénarios, Plan, and Outils d'ana. A message bar indicates "Afficher uniquement Votre compte hezzagran@insea.ac.ma n'autorise pas la..." and "Utiliser un autre compte". The main area displays a table titled "Country-data" with columns: O1, A, B, C, D, E, F, G, H, I, J, K. The table contains data for 32 countries, including country names, child mortality rates, exports, imports, income levels, inflation rates, life expectancy, total fertility rates, and GDP per capita. The table is sorted by child mortality rate. The bottom of the screen shows the Power BI ribbon, a search bar, and a zoom slider set to 100%.

O1	A	B	C	D	E	F	G	H	I	J	K
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	
1	Afghanistan	90.2		10 7.58	44.9	1610 9.44	56.2	5.82		553	
2	Albania	16.6		28 6.55	48.6	9930 4.49	76.3	1.65		4090	
3	Algeria	27.3	38.4	4.17	31.4	12900 16.1	76.5	2.89		4460	
4	Angola		119 62.3	2.85	42.9	5900 22.4	60.1	6.16		3530	
5	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100 1.44	76.8	2.13		12200	
6	Argentina	14.5	18.9	8.1		18700 20.9	75.8	2.37		10300	
7	Armenia	18.1	20.8	4.4	45.3	6700 7.77	73.3	1.69		3220	
8	Australia	4.8	19.8	8.73	20.9	41400 1.16		82 1.93		51900	
9	Austria	4.3	51.3		11 47.8	43200 0.873	80.5	1.44		46900	
10	Azerbaijan	39.2	54.3	5.88	20.7	16000 13.8	69.1	1.92		5840	
11	Bahamas	13.8		35 7.89	43.7	22900 -0.393	73.8	1.86		28000	
12	Bahrain	8.6	69.5	4.97	50.9	41100 7.44		76 2.16		20700	
13	Bangladesh	49.4		16 3.52	21.8	2440 7.14	70.4	2.33		758	
14	Barbados	14.2	39.5	7.97	48.7	15300 0.321	76.7	1.78		16000	
15	Belarus	5.5	51.4	5.61	64.5	16200 15.1	70.4	1.49		6030	
16	Belgium	4.5	76.4	10.7	74.7	41100 1.88		80 1.86		44400	
17	Belize	18.8	58.2	5.2	57.5	7880 1.14	71.4	2.71		4340	
18	Benin		111 23.8	4.1	37.2	1820 0.885	61.8	5.36		758	
19	Bhutan	42.7	42.5	5.2	70.7	6420 5.99	72.1	2.38		2180	
20	Bolivia	46.6	41.2	4.84	34.3	5410 8.78	71.6	3.2		1980	
21	Bosnia and Herzegovina	6.9	29.7	11.1	51.3	9720 1.4	76.8	1.31		4610	
22	Botswana	52.5	43.6	8.3	51.3	13300 8.92	57.1	2.88		6350	
23	Brazil	19.8	10.7	9.01	11.8	14500 8.41	74.2	1.8		11200	
24	Brunei	10.5	67.4	2.84		80600 16.7	77.1	1.84		35300	
25	Bulgaria	10.8	50.2	6.87		15300 1.11	73.9	1.57		6840	
26	Burkina Faso		116 19.2	6.74	29.6	1430 6.81	57.9	5.87		575	
27	Burundi	93.6	8.92	11.6	39.2	764 12.3	57.7	6.26		231	
28	Cambodia	44.4	54.1	5.68	59.5	2520 3.12	66.1	2.88		786	
29	Cameroon		108 22.2	5.13		2660 1.91	57.3	5.11		1310	
30	Canada	5.6	29.1	11.3		40700 2.87	81.3	1.63		47400	
31	Cape Verde	26.5	32.7	4.09	61.8	5830 0.505	72.5	2.67		3310	

2. Description des attributs et de leur signification :

Les attributs comprennent :

- *child_mort*: Taux de mortalité des enfants de moins de 5 ans pour 1000 naissances vivantes.
- *exports*, *imports*, *income*, *inflation*, *life_expec*, *total_fer*, *gdpp*: Indicateurs socio-économiques tels que les exportations, les importations, le revenu par habitant, l'inflation, l'espérance de vie, etc.

Ces indicateurs sont cruciaux pour évaluer le niveau de développement de chaque pays.

The screenshot shows the Microsoft Power BI Data Dictionary application window. The menu bar includes Accueil, Insertion, Mise en page, Formules, Données (selected), Révision, Affichage, Partager, Outils d'analyse, and Solveur. A message at the top says "Afficher uniquement Votre compte hezzagragani@insea.ac.ma n'autorise pas la..." and "Utiliser un autre compte". The main area displays a table with columns A and B:

	A	B
1	Column Name	Description
2	country	Name of the country
3	child_mort	Death of children under 5 years of age per 1000 live births
4	exports	Exports of goods and services per capita. Given as %age of the GDP per capita
5	health	Total health spending per capita. Given as %age of GDP per capita
6	imports	Imports of goods and services per capita. Given as %age of the GDP per capita
7	Income	Net income per person
8	Inflation	The measurement of the annual growth rate of the Total GDP
9	life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
10	total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
11	gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.
12		
13		

At the bottom, there are buttons for data dictionary, plus, and minus, along with accessibility and zoom controls.

III - Objectifs de l'analyse :

1. Définition des objectifs principaux :

L'objectif central de cette analyse réside dans le regroupement rigoureux des nations, usant de méthodes d'apprentissage non supervisé, particulièrement du clustering. Cette démarche ambitieuse vise à élaborer une segmentation précise des pays selon leurs caractéristiques socio-économiques et de santé. L'objectif ultime est de permettre à HELP International d'identifier avec acuité les régions nécessitant une aide humanitaire urgente et significative.

2. Précision sur l'utilisation de techniques d'apprentissage non supervisé :

Cette étude approfondie se concentrera sur l'application rigoureuse de méthodes de clustering pour regrouper les nations en fonction de leurs similarités dans les indicateurs socio-économiques et de santé. L'objectif sous-jacent à cette méthodologie est de simplifier l'identification et la

compréhension des groupes de pays partageant des profils similaires, permettant ainsi à HELP International de cibler efficacement les régions les plus vulnérables et de déployer des ressources appropriées pour répondre à leurs besoins spécifiques.

3. Regroupement :

Le choix du regroupement (clustering) parmi d'autres techniques d'apprentissage non supervisé découle de sa capacité à identifier les similarités et les disparités intrinsèques entre les pays en se basant sur un ensemble complexe d'indicateurs socio-économiques et de santé. Contrairement à d'autres approches telles que la réduction de dimension ou l'analyse des anomalies, le clustering offre la possibilité de créer des groupes homogènes de pays partageant des caractéristiques communes. Cette méthodologie permettra à HELP International de dégager des clusters distincts de nations ayant des besoins similaires, facilitant ainsi l'identification précise des régions les plus nécessiteuses et orientant efficacement l'allocation des ressources pour une action humanitaire ciblée.

IV - Exploration des Données :

1. Résumé des mesures prises pour nettoyer et préparer les données :

Une série de mesures ont été prises pour garantir la qualité et la cohérence des données. Cela inclut la gestion des valeurs manquantes, la vérification de la cohérence des valeurs, la normalisation des échelles des indicateurs pour éviter les biais, et le traitement des éventuelles valeurs aberrantes. Ces actions ont permis d'assurer que les données utilisées pour l'analyse sont fiables et aptes à être exploitées pour les besoins du clustering.

➤ **Visualisation des données après le traitement et le nettoyage :**

```

    ↗ child_mort    exports     health    imports      income \
count  167.000000  167.000000  167.000000  167.000000  167.000000
mean   38.270060  41.108976  6.815689  46.890215  17144.688623
std    40.328931  27.412010  2.746837  24.209589  19278.067698
min    2.600000  0.109000  1.810000  0.065900  609.000000
25%   8.250000  23.800000  4.920000  30.200000  3355.000000
50%   19.300000  35.000000  6.320000  43.300000  9960.000000
75%   62.100000  51.350000  8.600000  58.750000  22800.000000
max   208.000000 200.000000 17.900000 174.000000 125000.000000

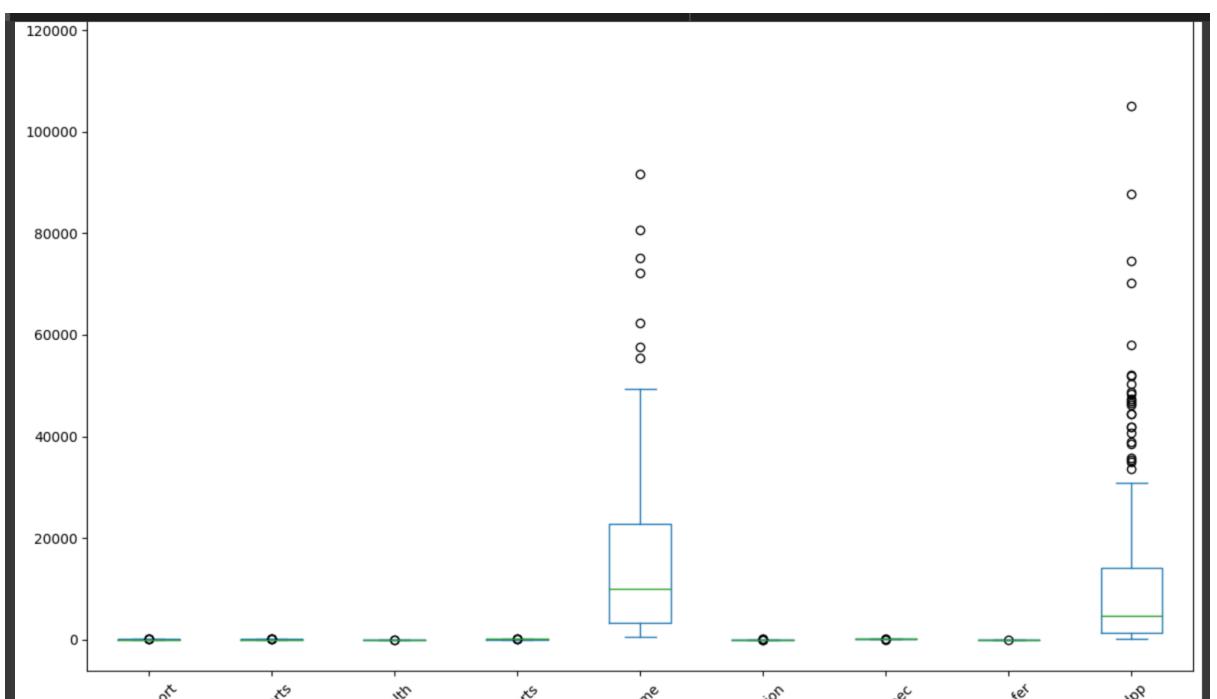
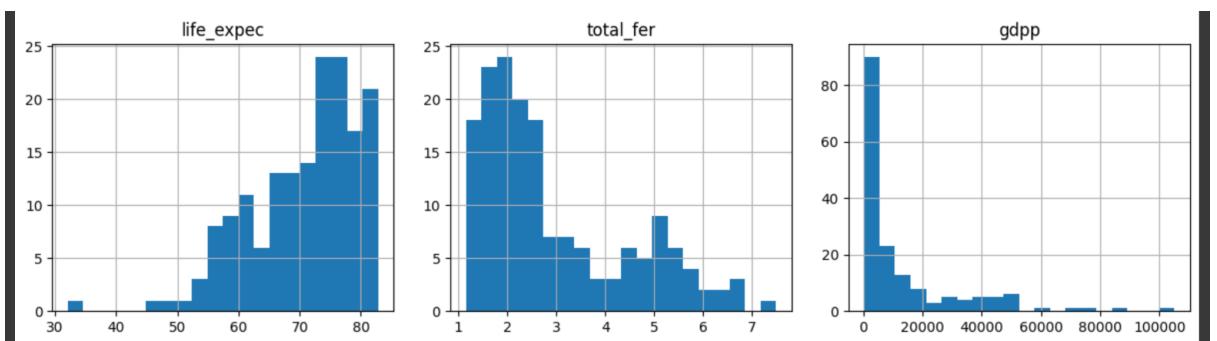
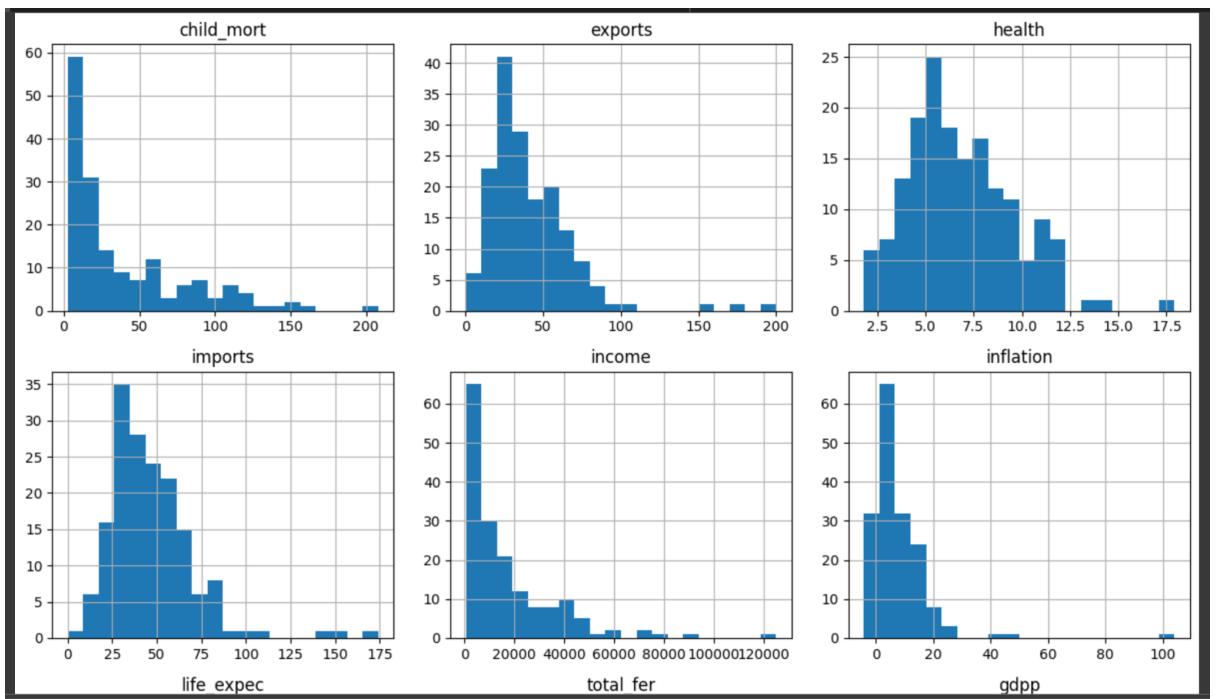
            inflation  life_expec  total_fer      gdpp
count  167.000000  167.000000  167.000000  167.000000
mean   7.781832  70.555689  2.947964  12964.155689
std    10.570704  8.893172  1.513848  18328.704809
min   -4.210000  32.100000  1.150000  231.000000
25%   1.810000  65.300000  1.795000  1330.000000
50%   5.390000  73.100000  2.410000  4660.000000
75%   10.750000 76.800000  3.880000  14050.000000
max   104.000000 82.800000  7.490000  105000.000000
<ipython-input-3-0a9169e15e2d>:8: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated.
data.fillna(data.mean(), inplace=True) # Remplacer les valeurs manquantes par la moyenne des colonnes

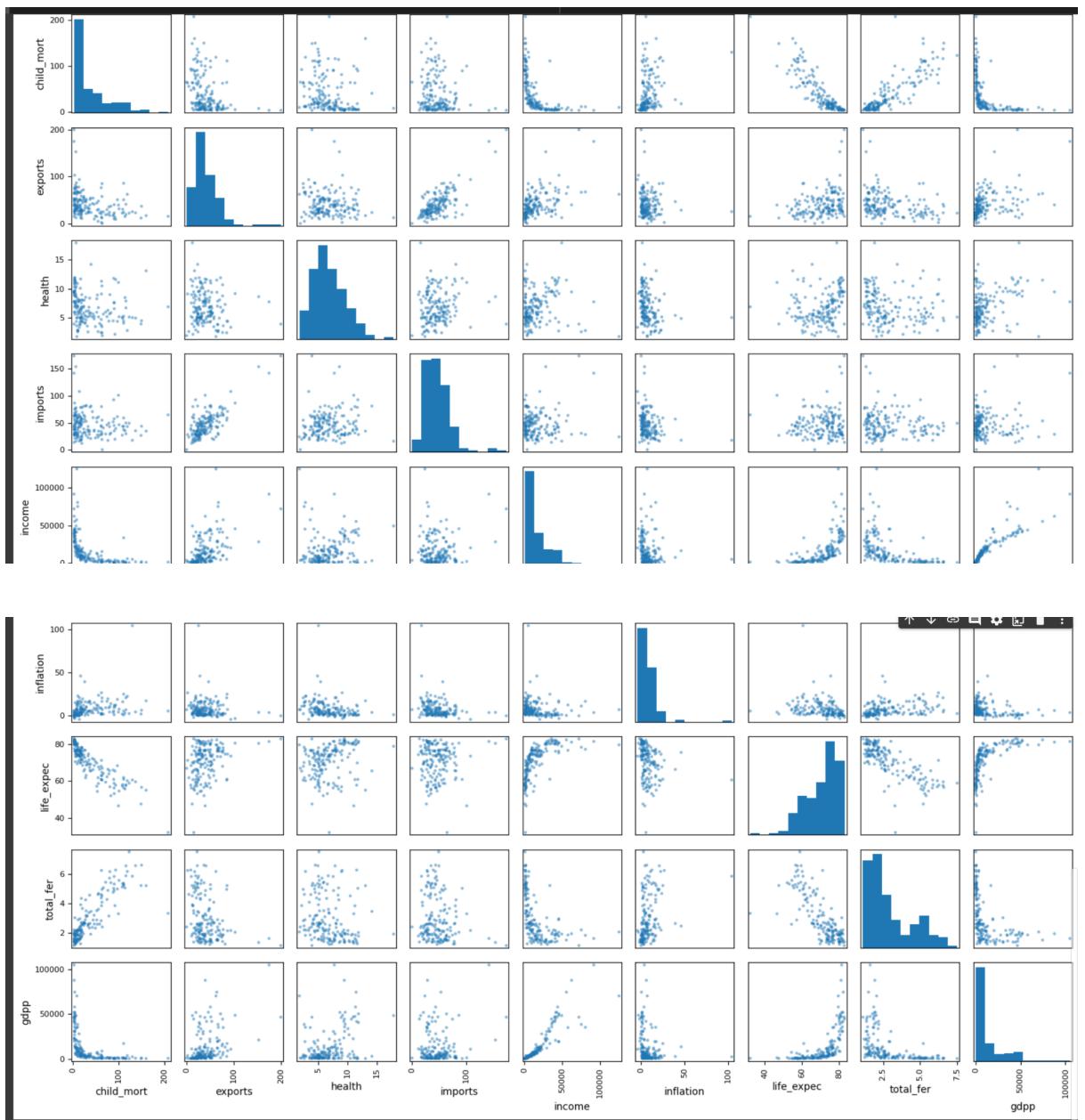
```

2. Visualisations préliminaires pour comprendre les distributions et les tendances :

Des visualisations exploratoires ont été réalisées pour comprendre la distribution et les tendances présentes dans les indicateurs socio-économiques et de santé des différents pays. Des graphiques, tels que des histogrammes, des graphiques en boîte, des diagrammes de dispersion et des cartes géographiques, ont été utilisés pour analyser les variations, les corrélations et les schémas qui émergent des données. Ces visualisations préliminaires ont fourni un aperçu des structures potentielles à explorer davantage lors du processus de clustering.

➤ **Visualisation graphique :**



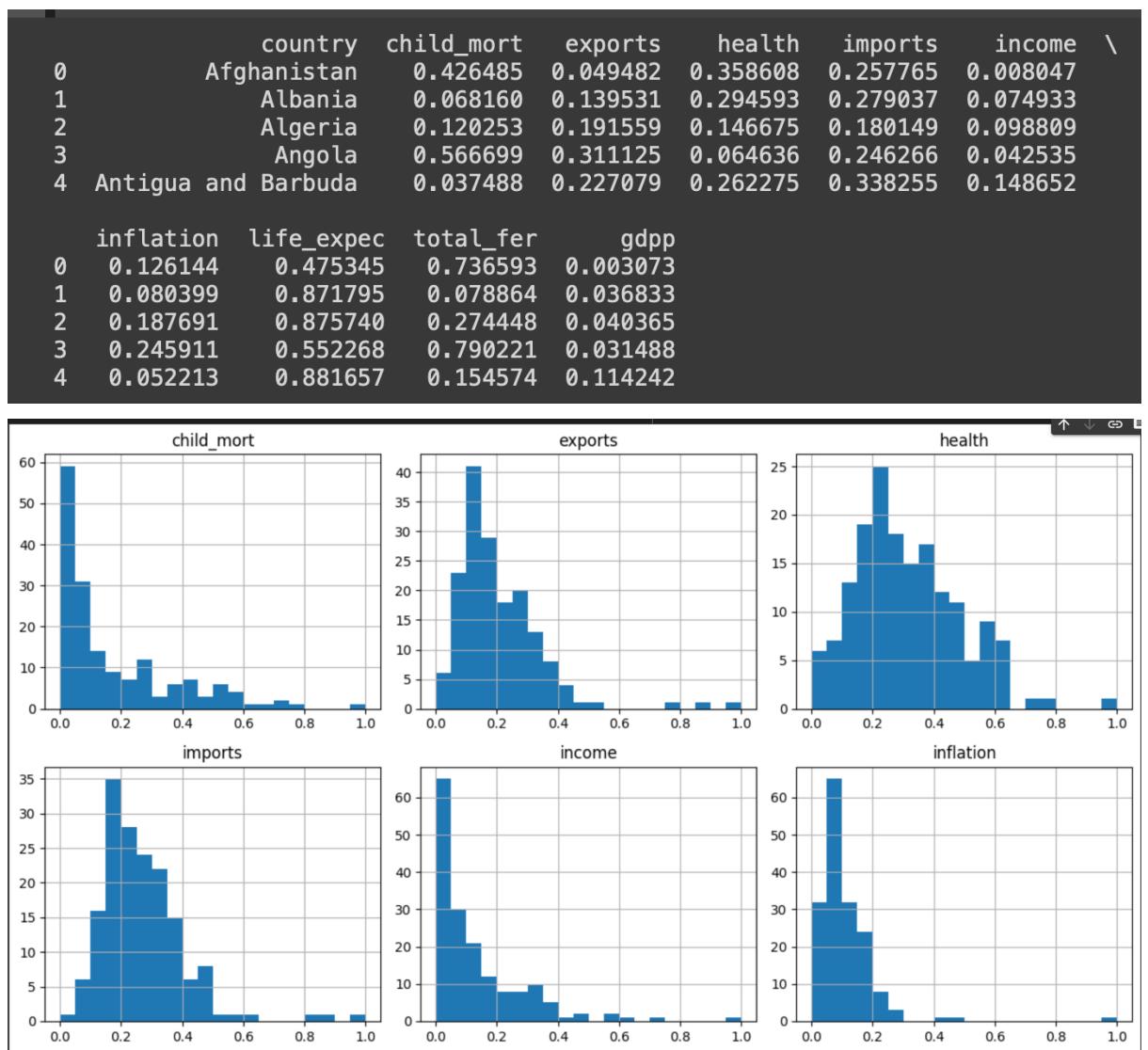


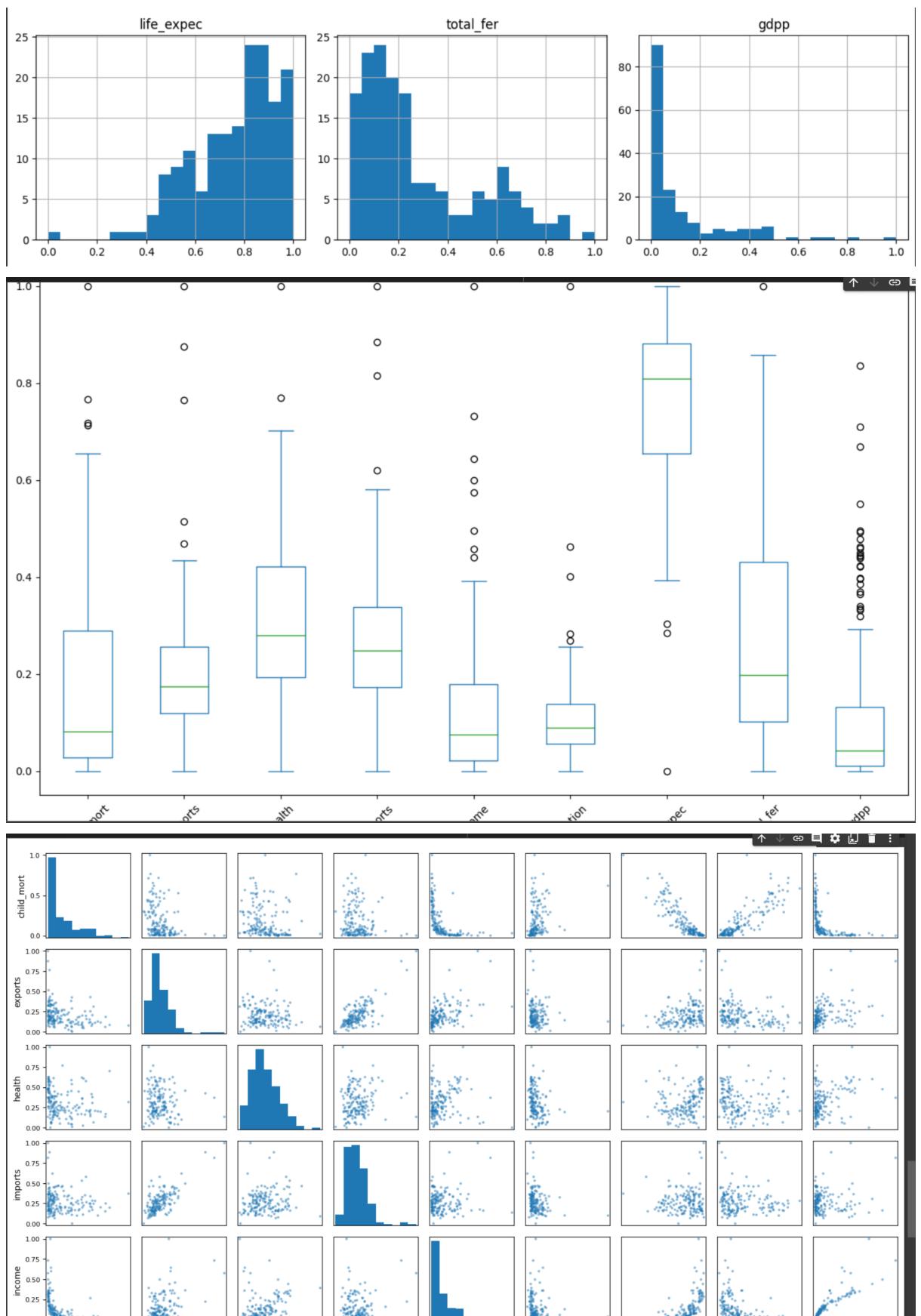
V - Normalisation des données :

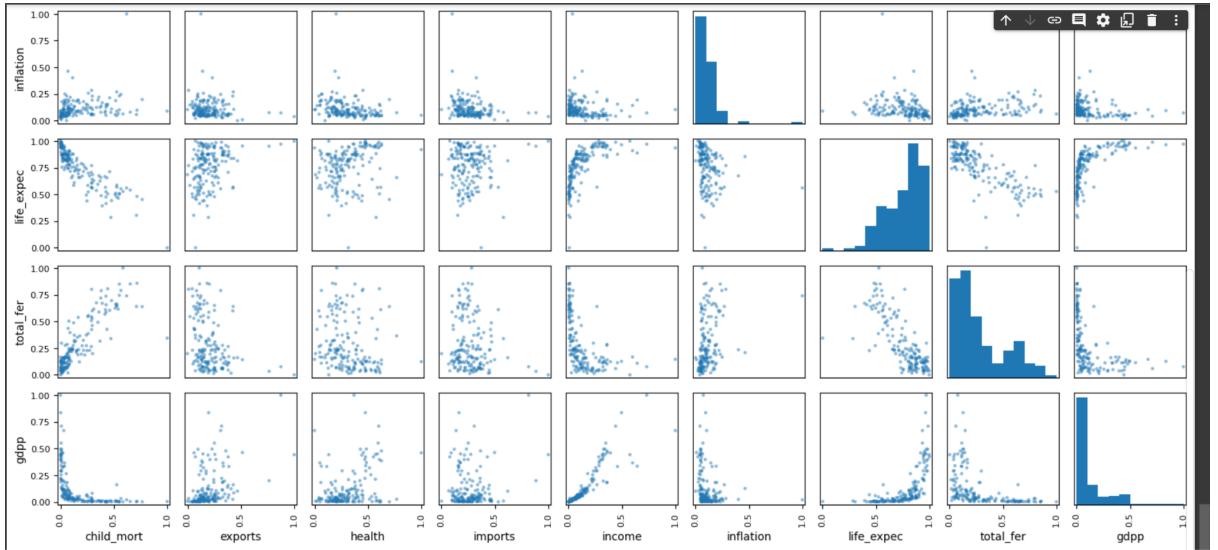
La normalisation des données est un processus utilisé pour mettre à l'échelle les valeurs des différentes variables d'un ensemble de données vers une plage commune. Elle est souvent utilisée dans les méthodes d'apprentissage automatique pour garantir que les variables avec des échelles différentes ont un impact équilibré sur l'analyse. Si on revient à nos statistiques descriptives , elles montrent des échelles différentes pour chaque colonne. Certaines variables ont des plages de valeurs très différentes, par exemple, 'income' a une valeur maximale bien plus

élevée que 'exports' ou 'health'. Ces différences dans les plages de valeurs peuvent influencer négativement les algorithmes qui s'appuient sur des calculs de distances. Donc, on doit appliquer la normalisation, elle est souvent recommandée dans le cadre de l'apprentissage automatique, surtout lorsque les variables ont des échelles très différentes. Elle permet de ramener toutes les variables à une échelle comparable, généralement entre 0 et 1, en préservant les relations entre les données. Cela facilite la convergence des algorithmes d'apprentissage et améliore souvent les performances.

➤ Visualisation graphique normalisée:







VI - Modèles d'Apprentissage Non Supervisé :

1. Entraînement de différentes variantes de modèles pour le Clustering :

Dans le cadre de cette étude, plusieurs approches d'apprentissage non supervisé ont été explorées pour regrouper les pays en fonction de leurs indicateurs socio-économiques et de santé. Trois variantes de modèles de clustering ont été entraînées et évaluées pour leur capacité à identifier des structures significatives au sein des données. Parmi les techniques examinées figuraient **le K-Means Clustering**, une méthode de partitionnement des données en clusters, **le clustering hiérarchique**, permettant de créer une hiérarchie de clusters, et enfin, **le DBSCAN**, privilégiant une détection de densité pour identifier des clusters de formes arbitraires. Chacune de ces approches a été étudiée avec des paramètres divers pour déterminer la configuration optimale offrant une meilleure compréhension des différents groupes de pays en situation de besoin. Les résultats obtenus ont éclairé la décision finale sur la méthode de regroupement la plus appropriée pour cette analyse spécifique.

2. Description des différentes approches de modélisation utilisées :

- **K-Means Clustering** :Le K-Means est un algorithme de clustering largement utilisé pour partitionner les données en K clusters. Son fonctionnement repose sur la minimisation de la variance intra-cluster. Il itère pour assigner les observations à des clusters de manière à minimiser la somme des carrés

des distances entre les points et le centroïde du cluster auquel ils sont assignés.

- **Clustering hiérarchique** :Le clustering hiérarchique est une méthode qui crée une hiérarchie de clusters. Il existe deux types : le clustering agglomératif (bottom-up) et le clustering diviseur (top-down). L'agglomératif commence avec chaque observation comme un cluster distinct et fusionne progressivement les clusters les plus similaires. L'indice de similarité est souvent mesuré par la distance euclidienne ou d'autres métriques comme la distance de Manhattan.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** :DBSCAN est un algorithme basé sur la densité qui identifie des clusters en se basant sur la densité des points dans un voisinage spécifié. Il définit deux paramètres clés : le rayon $\text{eps}(\text{Epsilon})$ pour définir le voisinage d'un point, et minPts pour spécifier le nombre minimal de points dans ce voisinage pour qu'un point soit considéré comme central. Les points se trouvant à une distance eps les uns des autres sont considérés comme faisant partie du même cluster. La technique permet également de détecter les valeurs aberrantes en les considérant comme des points isolés.

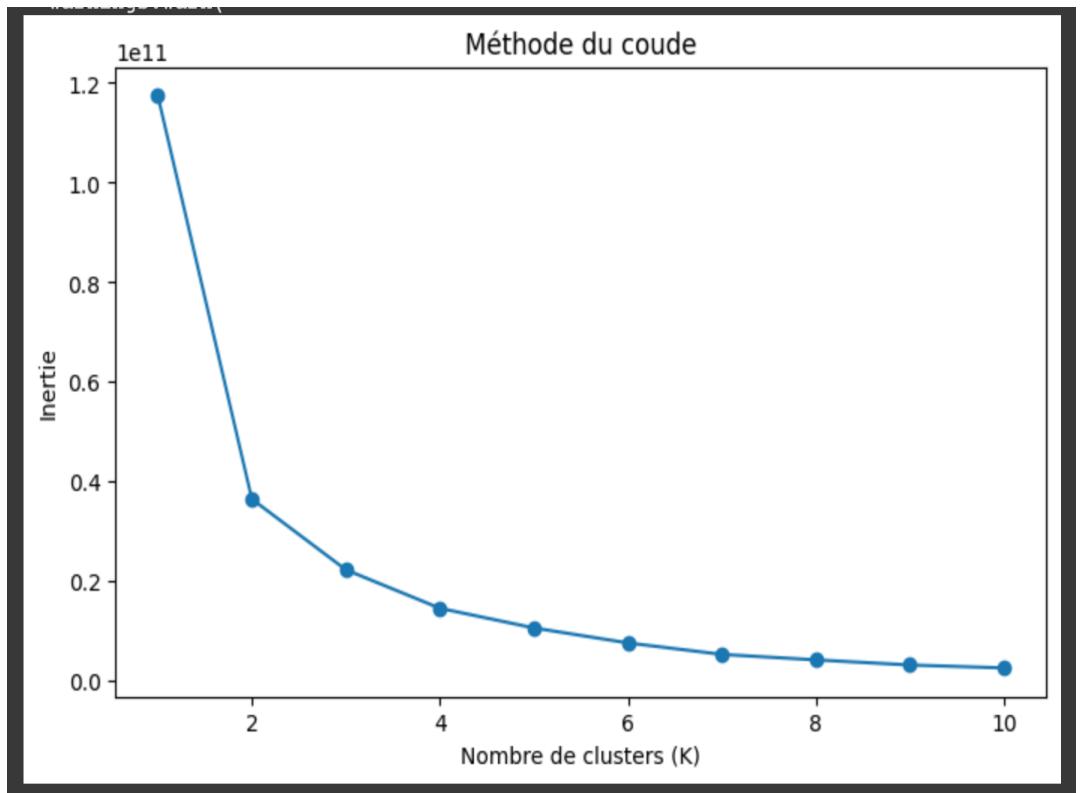
Ces trois méthodes diffèrent dans leur approche de la formation de clusters et dans leur sensibilité aux formes, densités et tailles des clusters dans les données. Chacune offre des avantages et des limitations distincts en fonction des caractéristiques des données à analyser.

VII - Benchmarking & Résultats:

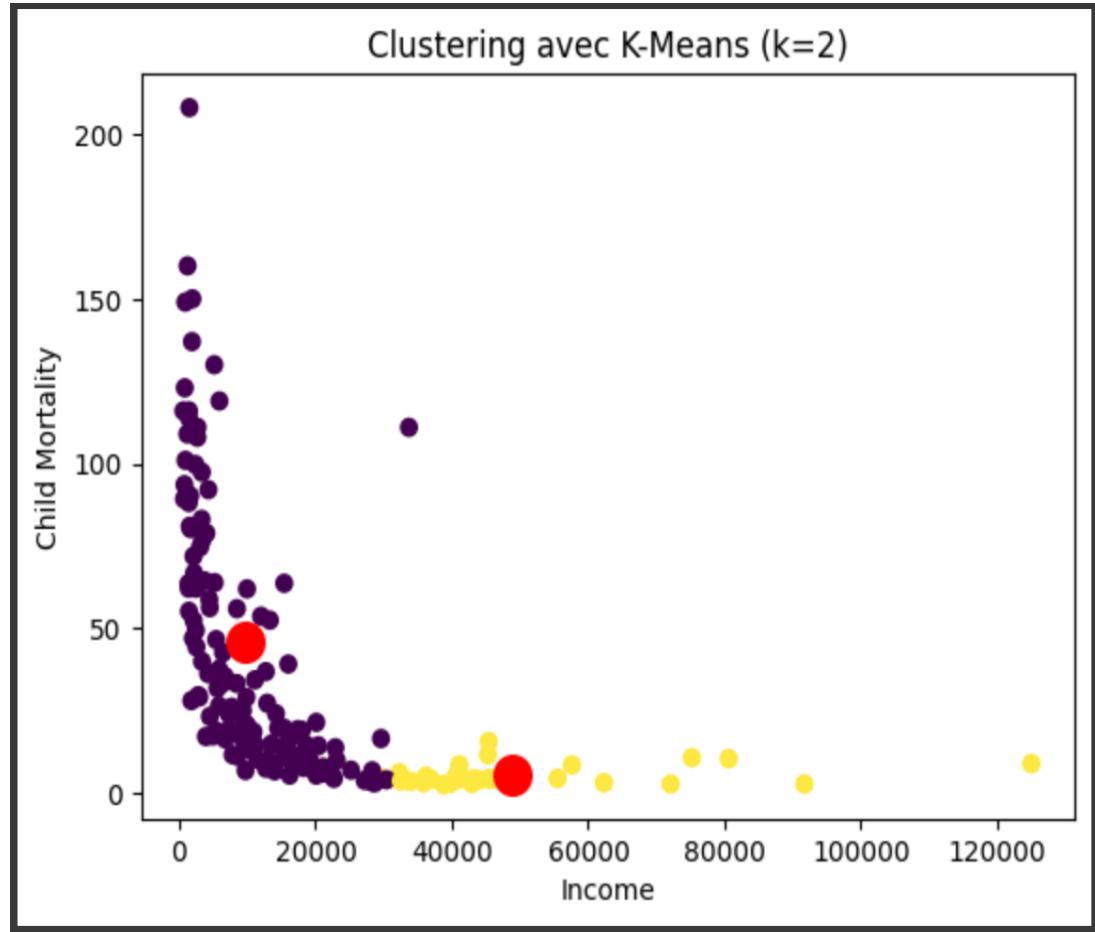
Le benchmarking va être appliqué sur les données non normalisées pour bien visualiser la différence entre les modèles .Dans ces visualisations, bien que le clustering soit effectué sur toutes les caractéristiques, nous ne pouvons pas représenter graphiquement plus de deux dimensions à la fois. Pour pouvoir visualiser le résultat du clustering sur plusieurs caractéristiques (qui peuvent être multidimensionnelles), nous utilisons les techniques de la réduction de la dimension ou projection des données dans un espace à deux dimensions, en revanche , l'objectif de ce projet est le regroupement et non pas la réduction des dimensions .

➤ **K-Means Clustering** :Le choix du nombre de clusters (K) dans l'algorithme KMeans est souvent basé sur différentes techniques. Dans ce cas , j'ai choisie la méthode du coude (Elbow Method), elle est implémentée en utilisant la somme des carrés des distances intra-cluster (inertia) pour différents nombres de clusters (K) dans l'algorithme KMeans .

- *Méthode de coude* : En vérifiant le graphe ci-dessus , on peut voir clairement le point d'inflexion , donc , on peut déduire que $k=2$.



- *Clustering avec K-means ou K=2* : Lorsque on effectue un clustering avec K-Means, les étiquettes (labels) attribuées à chaque point correspondent aux clusters auxquels ces points sont assignés. Dans le cas d'un clustering binaire avec k=2, les labels seront soit 0 soit 1, représentant les deux clusters distincts créés par l'algorithme K-Means.

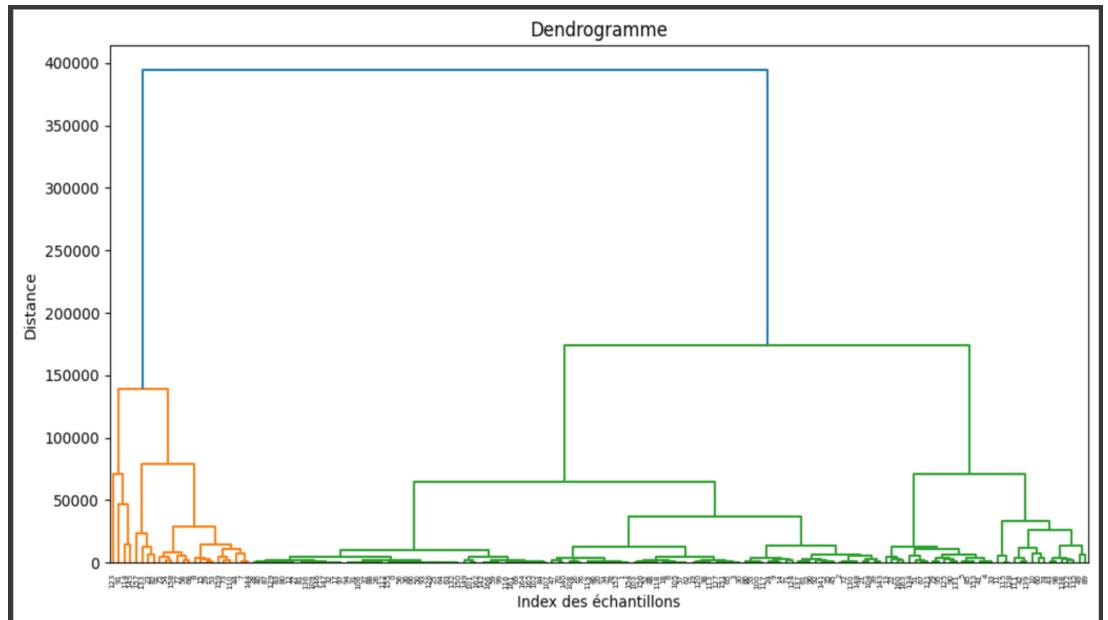


Les deux grands points rouges représentent les centres des clusters obtenus après l'application de l'algorithme K-Means avec k=2. Lorsque l'on applique, il identifie deux centres de cluster dans l'espace des attributs. Ces centres sont les moyennes des attributs de chaque cluster et sont représentés comme des points rouges dans la visualisation. Ils indiquent la position centrale de chaque cluster calculée par l'algorithme K-Means pour minimiser la distance des points de données au centre de leur cluster respectif.

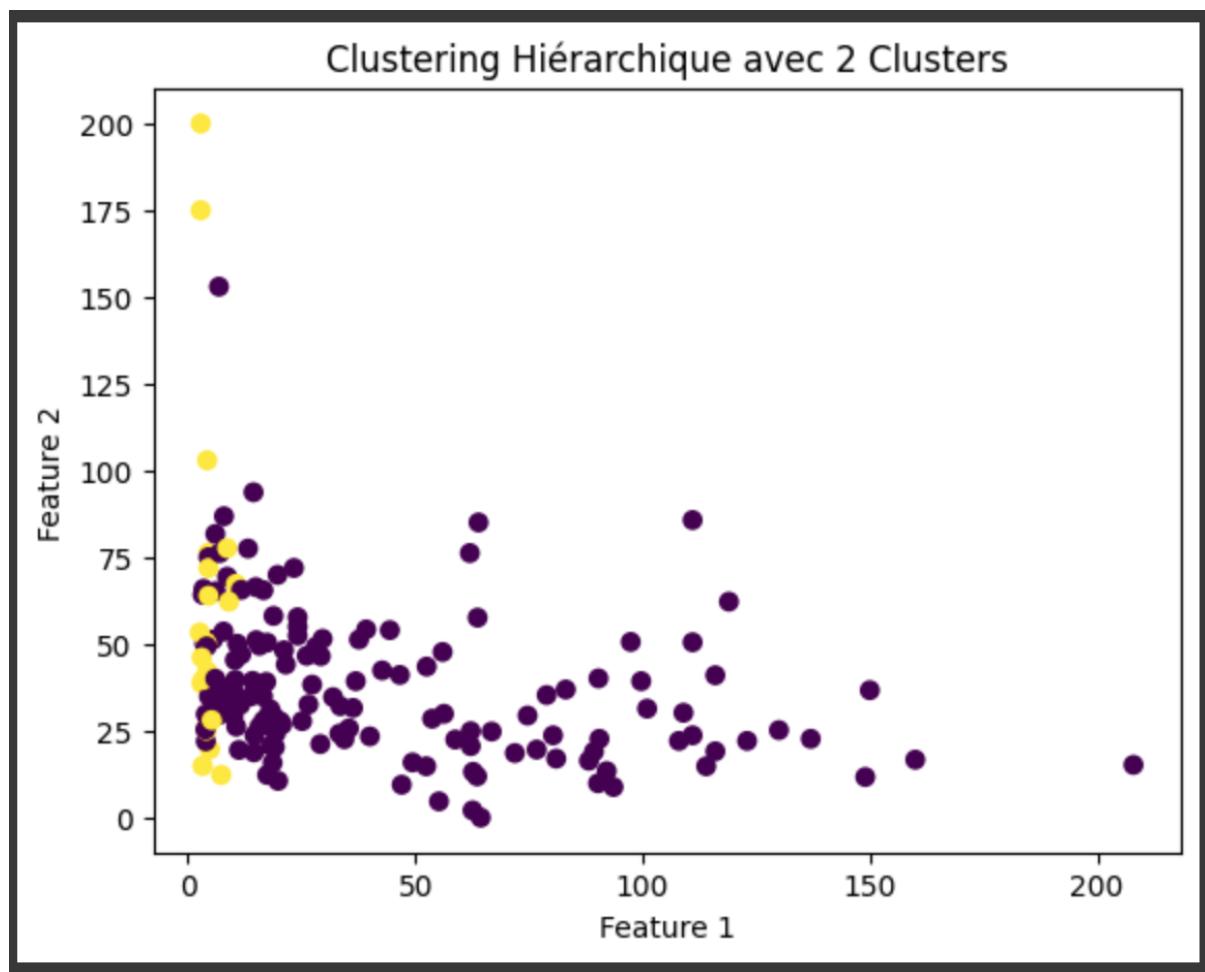
- **Clustering hiérarchique :** Le nombre optimal de clusters peut varier d'un algorithme de clustering à un autre. La méthode du coude peut donner des résultats différents selon l'algorithme utilisé, car chaque algorithme utilise ses propres critères pour déterminer le nombre optimal de clusters. Les mécanismes internes des algorithmes de clustering diffèrent, ce qui peut conduire à des recommandations différentes en termes de nombre de clusters optimaux. Lorsqu'on utilise la méthode du coude ou d'autres méthodes pour déterminer le nombre optimal de clusters, on peut observer des différences entre les résultats obtenus avec KMeans et ceux obtenus avec le clustering hiérarchique ([Agglomerative Clustering](#)). Chaque

algorithme a ses propres métriques internes pour évaluer la qualité des clusters, ce qui peut influencer le choix du nombre optimal de clusters.

- *Le dendrogramme* : Le dendrogramme est généralement créé avant de décider du nombre de clusters à choisir. Il permet de visualiser la structure hiérarchique des données avant de les regrouper. En observant le dendrogramme, on peut repérer visuellement les zones où les branches fusionnent, ce qui nous donne des indications sur le nombre optimal de clusters à sélectionner. Il existe beaucoup de méthodes utilisées pour calculer la distance entre les clusters à chaque étape de fusion lors de la construction du dendrogramme . Dans ce cas, j'ai utilisé **Linkage de Ward** qui vise à minimiser la variance lors de la fusion des clusters, favorisant ainsi la formation de clusters de taille égale. Notre $k=2$.



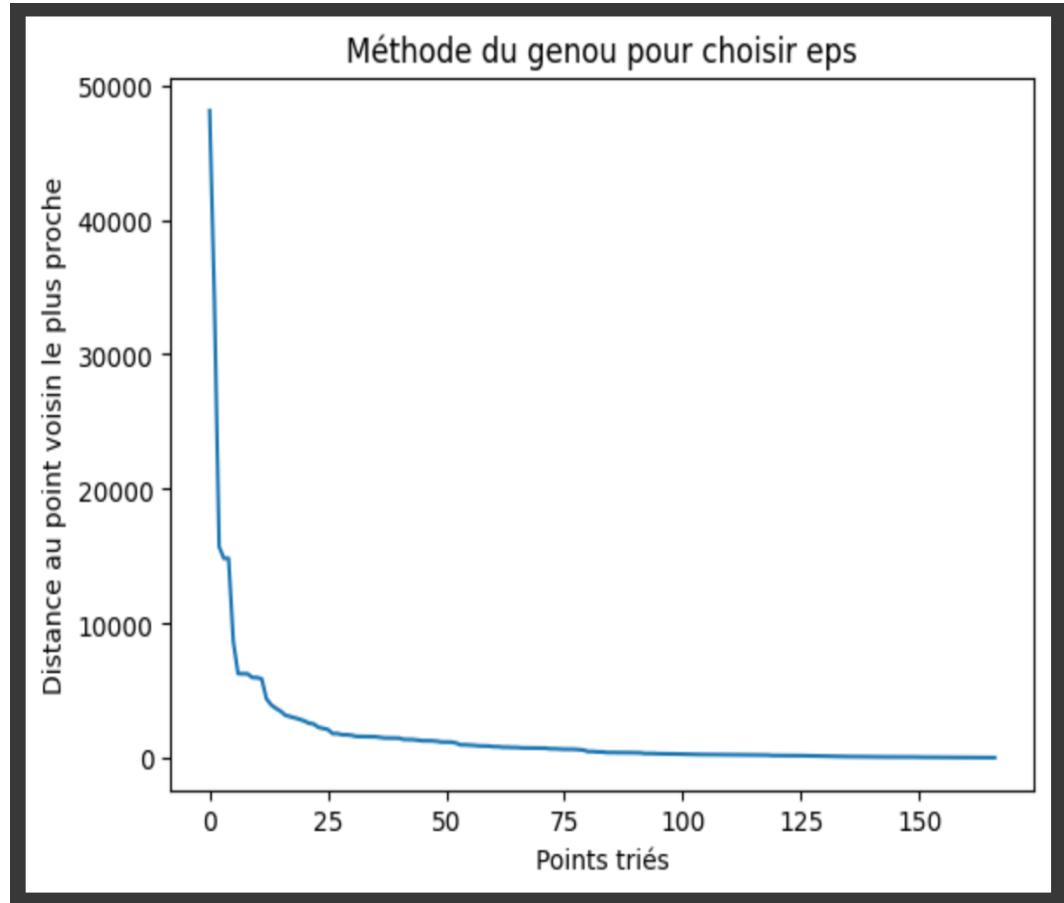
- *Clustering avec clustering Hiérarchique :*



➤ **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

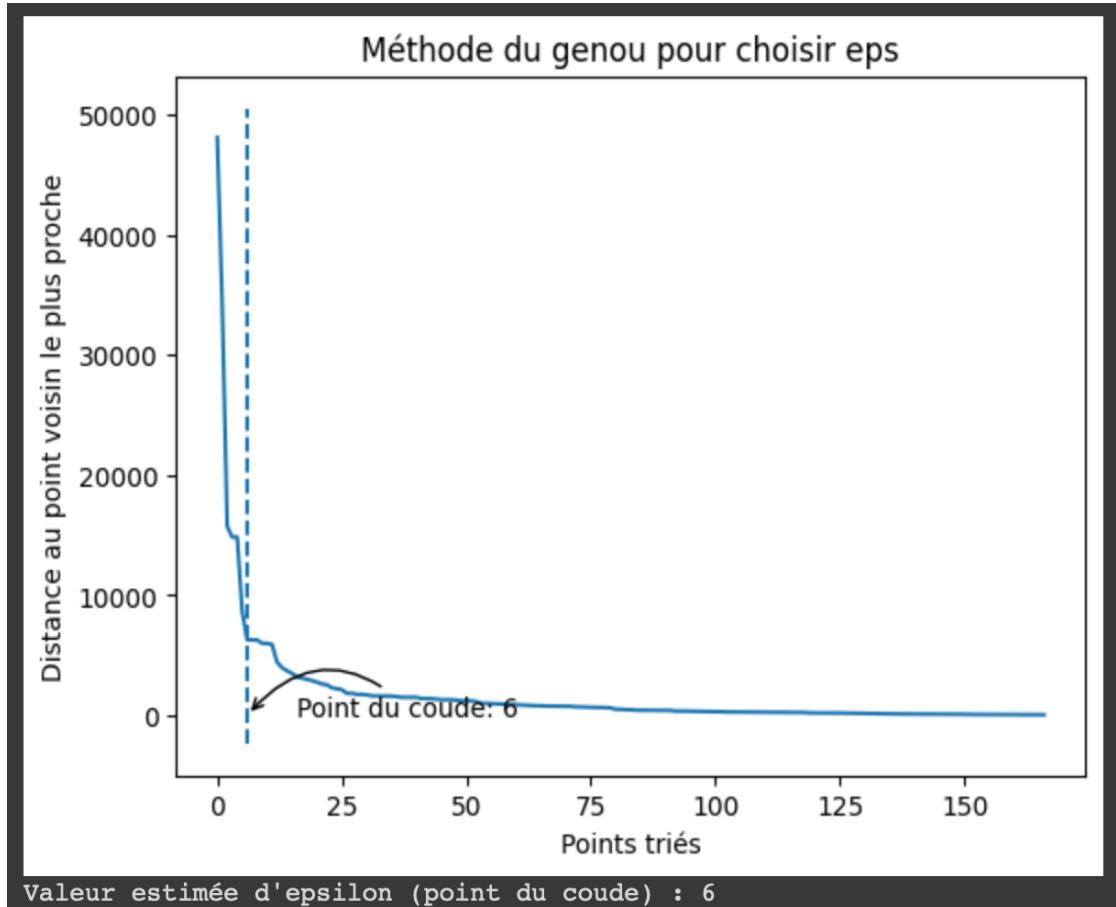
:Le choix des valeurs pour eps et min_samples dans DBSCAN est important pour obtenir des résultats pertinents.

- *Méthode du genou (Elbow Method)* : On trace le graphique de la distance entre chaque point et ses k points les plus proches en ordre décroissant. Après , on identifie le point où la courbe commence à devenir moins raide (le "coude") comme valeur d'epsilon.



On peut voir que ce graphique n'est pas bien lisible pour l'estimation de la valeur de epsilon , donc je vais utiliser la méthode knee .

- *La méthode knee* : La méthode Knee est une technique qui identifie le point où la courbe prend un angle aigu ou forme un "coude". Ce point est souvent considéré comme un choix judicieux pour déterminer un seuil, une valeur ou un paramètre dans diverses analyses. Dans le contexte de DBSCAN, cette méthode vise à sélectionner la valeur optimale pour le paramètre `eps`, qui représente la distance maximale entre deux échantillons pour qu'ils soient considérés dans le même voisinage. En identifiant le point du coude sur le graphique des distances, on peut estimer un bon `eps` pour le clustering DBSCAN.



- *Clustering avec DBSCAN* : en utilisant DBSCAN, les valeurs de la matrice peuvent inclure des étiquettes de cluster, mais aussi des valeurs de -1. Les valeurs de -1 représentent les points considérés comme des valeurs aberrantes ou du bruit par l'algorithme DBSCAN, n'appartenant à aucun cluster spécifique. Les valeurs autres que -1 représentent les différents clusters numérotés à partir de 0, 1, 2, etc. Les points affectés à un cluster auront un numéro de cluster attribué, tandis que ceux considérés comme du bruit seront marqués -1 dans la sortie.
 - *Résultat avec epsilon=6* : on remarque qu'il est possible que l'algorithme DBSCAN ait considéré la majorité des points comme du bruit si les paramètres, en particulier epsilon, ne sont pas bien ajustés. Cela peut arriver si l'epsilon est trop petit ou si les données ne sont pas bien adaptées à DBSCAN.

```

[-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1]

```

VIII - Sélection du Modèle Final :

Notre prochaine étape consiste à sélectionner le modèle le plus adapté à nos données. Étant donné que DBSCAN a produit principalement des valeurs de cluster -1, suggérant une sélection inadéquate de la valeur d'epsilon, nous allons nous concentrer sur les deux modèles restants : K-means et le Clustering Hiérarchique. Nous allons évaluer ces modèles en utilisant le score de silhouette, une mesure qui évalue à la fois la cohésion et la séparation entre les clusters. Les scores de silhouette proches de 1 indiquent une meilleure séparation et une cohésion plus forte des clusters. Il est possible d'ajuster les paramètres de ces algorithmes pour mieux s'adapter à la structure inhérente de nos données.

```

Score de silhouette pour K-Means : 0.7256314906273207
Score de silhouette pour Clustering hiérarchique : 0.7228222144066916

```

On peut voir que les deux sont convenables pour nos données particulièrement K-means .

IX - Environnement du travail :



Kaggle est une plateforme web interactive qui propose des compétitions d'apprentissage automatique en science des données. La plateforme fournit des jeux de données, des notebooks et des didacticiels gratuits dont les scientifiques de données ont besoin pour réaliser leurs projets d'apprentissage automatique.



Python est un langage de programmation interprété, multi paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet.



Colaboratory, souvent raccourci en "Colab", est un produit de Google Research. Colab permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté au machine learning, à l'analyse de données et à l'éducation.

X - Conclusions :

Les clusters identifiés à travers les méthodes de clustering, notamment K-means et Clustering Hiérarchique, ont fourni des insights cruciaux pour HELP International. Ces regroupements de pays, basés sur des indicateurs socio-économiques et de santé, ont permis de catégoriser les nations selon leurs besoins en aide humanitaire. Cependant, malgré l'efficacité de ces méthodes, il est essentiel de souligner la complexité inhérente à la représentation multidimensionnelle des données. Bien que le clustering ait été effectué sur plusieurs caractéristiques, la visualisation graphique se limite à deux dimensions. Ce défi de visualisation multidimensionnelle pourrait être surmonté par des techniques de réduction de dimensionnalité, mais il est important de noter que notre objectif principal réside dans l'identification des clusters significatifs pour orienter les actions d'HELP International.

En considérant les scores de silhouette, indicateurs de la qualité des clusters, ces méthodes ont démontré une certaine efficacité dans la création de clusters distincts. Cependant, chaque approche de clustering présente des avantages et des limites propres. Par exemple, K-means nécessite la spécification préalable du nombre de clusters (K), tandis que le Clustering hiérarchique offre une vue hiérarchique des clusters sans nécessiter un nombre prédéterminé. En outre, le DBSCAN a montré des difficultés avec le choix des paramètres epsilon et min_samples, conduisant à des résultats moins satisfaisants.

Dans ces visualisations, bien que le clustering soit effectué sur toutes les caractéristiques, nous ne pouvons pas représenter graphiquement plus de deux dimensions à la fois. Pour pouvoir visualiser le résultat du clustering sur plusieurs caractéristiques (qui peuvent être multidimensionnelles), nous utilisons les techniques de réduction de la dimension ou projection des données dans un espace à deux dimensions. Cependant, il est important de souligner que l'objectif fondamental de ce projet reste l'identification de regroupements significatifs pour aider à l'allocation stratégique des ressources d'aide d'urgence, et non la réduction des dimensions.

XI - Suggestions pour les Prochaines Étapes :

- **Exploration de données supplémentaires** : Intégrez des données supplémentaires telles que des indicateurs culturels, politiques ou environnementaux pour obtenir une image plus holistique des besoins des pays. Cela pourrait améliorer la précision des clusters.
- **Analyse de sensibilité des modèles** : Explorez la sensibilité des modèles à différents paramètres. Effectuez des analyses plus détaillées en variant les paramètres pour chaque algorithme de clustering afin d'observer l'impact sur les clusters identifiés.
- **Enrichissement des données** : Envisager d'enrichir les données en utilisant des techniques telles que l'imputation de données manquantes ou la normalisation différente pour voir l'effet sur les clusters.
- **Validation externe** : Validez les clusters identifiés en les comparant à des informations externes, comme des rapports d'organisations internationales ou des études académiques, pour confirmer si les clusters correspondent à des profils de pays réels.
- **Implémentation de la solution** : Développez des recommandations spécifiques pour HELP International basées sur les clusters identifiés. Créez des stratégies d'intervention ciblées pour répondre aux besoins de chaque groupe de pays.

Annexe :

- Lien vers code source :
[https://colab.research.google.com/drive/1aDCWX7jss9n1zSHmMO2rWy4I4k_0Gp30
?usp=sharing](https://colab.research.google.com/drive/1aDCWX7jss9n1zSHmMO2rWy4I4k_0Gp30?usp=sharing)