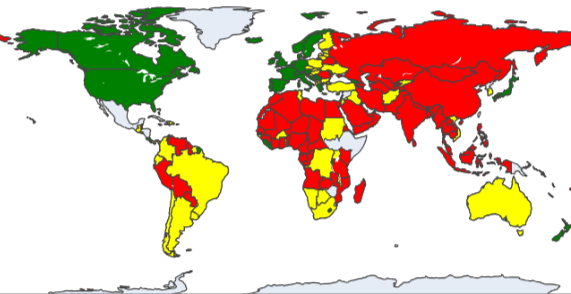


**Prédiction des Bénéficiaires de l'Aide :**

Une Approche Basée sur l'Apprentissage Automatique pour Prioriser les Pays en Besoin



**Rédigé par : ESSALHI SARA - Kaissi Houda**

**Supervisé par : Pr. El Mokhtar EN-NAIMI**

**Plan**

**I. Énoncé du problème (extrait du jeu de données)**

**II. Stratège du travail**

**III. Implémentation**

* Bibliothèques et Ensembles de Données
* Traitement des Données
* Principal Component Analysis (PCA)
* Modélisation - Apprentissage non supervisé
* Modélisation - Apprentissage Supervisé (KNN)

**IV. Mise en œuvre du Modèle Flask**

**Énoncé du problème (extrait du jeu de données) :**

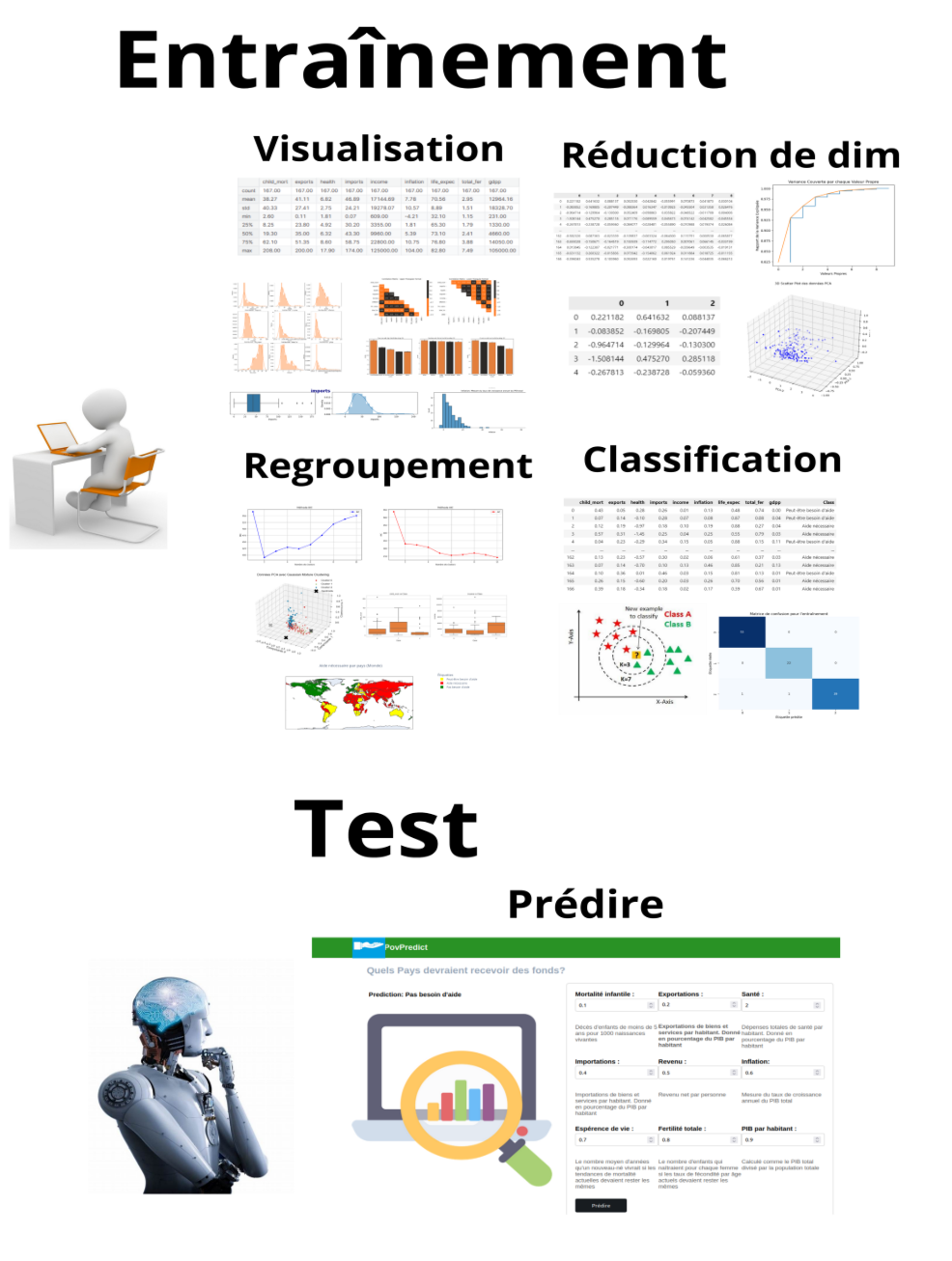
HELP International a réussi à récolter environ 10 millions de dollars. Maintenant, le PDG de l'ONG doit décider comment utiliser cet argent de manière stratégique et efficace.

Ainsi, le PDG doit prendre la décision de choisir les pays qui ont le plus besoin d'aide.

Par conséquent, l'objectif est de **catégoriser les pays en utilisant certains facteurs socio-économiques et de santé qui déterminent le développement global du pays**. Ensuite, vous devez suggérer les pays sur lesquels le PDG doit se concentrer le plus.

**Quels pays devraient recevoir des fonds et pourquoi ?**

**Stratège du travail :**



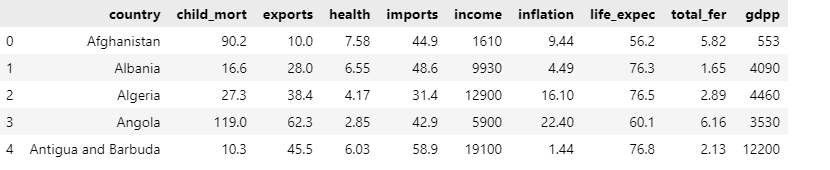
**Implémentation :**

**1. Bibliothèques et Ensembles de Données :**

**Importer les Bibliothèques et les Packages :**



**Importer les Données :**



**Description des caractéristiques :**

* **country** : Nom du pays

* **child\_mort** : Décès d'enfants de moins de 5 ans pour 1000 naissances vivantes

* **exports** : Exportations de biens et services par habitant. Donné en pourcentage du PIB par habitant

* **health** : Dépenses totales de santé par habitant. Donné en pourcentage du PIB par habitant

* **imports** : Importations de biens et services par habitant. Donné en pourcentage du PIB par habitant

* **income** : Revenu net par personne

* **inflation** : Mesure du taux de croissance annuel du PIB total

* **life\_expec** : Le nombre moyen d'années qu'un nouveau-né vivrait si les tendances de mortalité actuelles devaient rester les mêmes

* **total\_fer** : Le nombre d'enfants qui naîtraient pour chaque femme si les taux de fécondité par âge actuels devaient rester les mêmes

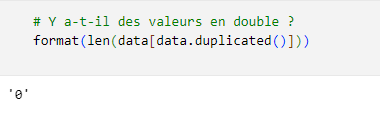
* **gdpp** : Le PIB par habitant. Calculé comme le PIB total divisé par la population totale

**2. Traitement des Données :**

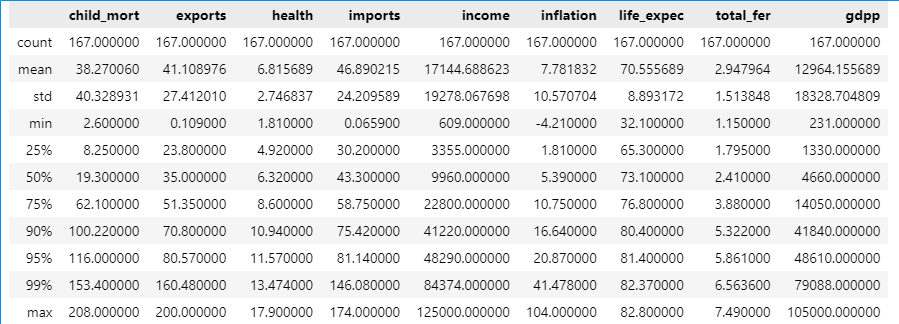
* Valeurs manquantes



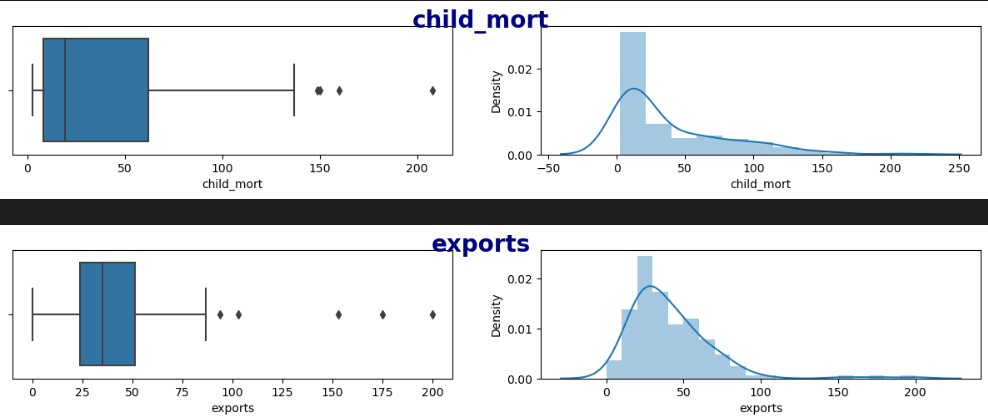
* Valeurs en double

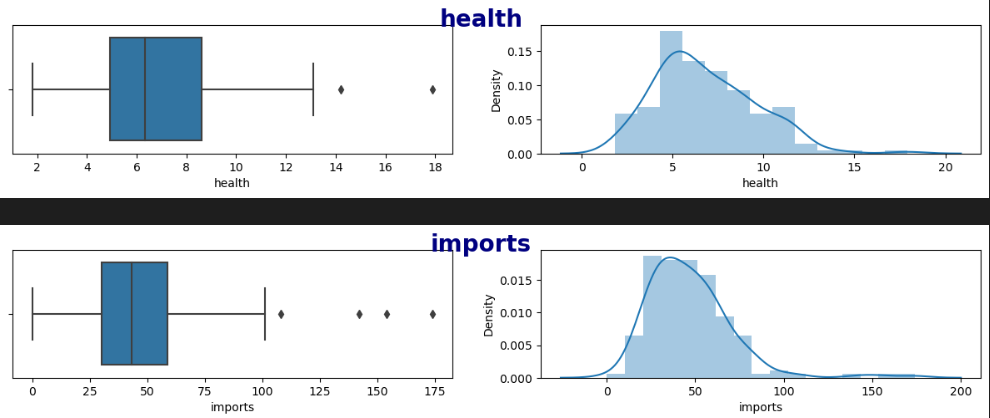


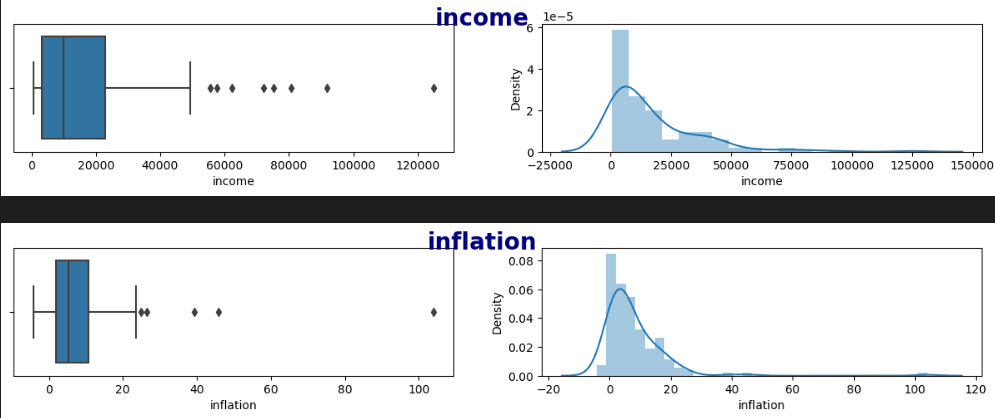
* Mesures statistiques standard

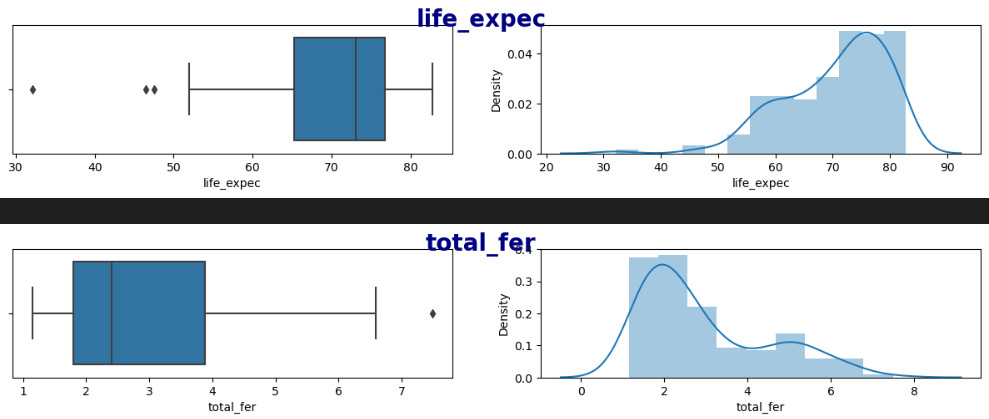


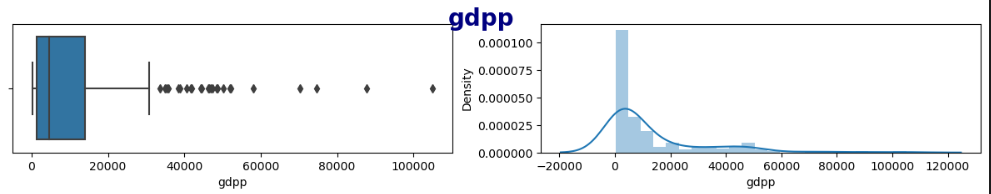
* Visualisation de toutes les colonnes avec des boîtes à moustaches (Boxplot) et des graphiques de distribution (Distplot) pour identifier la distribution des données et les valeurs aberrantes.











* **life\_expec** affiche une distribution des données inclinée à gauche ou négativement inclinée.
* **health** affiche une distribution des données normalement distribuée.
* Toutes les **autres caractéristiques** montrent une distribution des données inclinée à droite ou positivement inclinée.
* La distribution des données du **country** n'est pas affichée car il s'agit de données textuelles et elle a le même nombre de valeurs uniques que la longueur du dataframe.

**Constatations :**

1. Jeu de données de petite taille.
2. Aucune valeur manquante.
3. Aucune valeur en double.
4. Quelques valeurs aberrantes et distribution inclinée.

**Les caractéristiques des pays économiquement défavorisés sont les suivantes :**

- Le revenu par habitant du pays est très faible.

- Population élevée qui entraîne une non-disponibilité des ressources.

- Chômage en raison de ressources limitées.

- Faible richesse nationale qui entraîne un faible capital.

- Distribution inéquitable des richesses et des revenus.

- Manque d'aménagements éducatifs adéquats et donc illettrisme prévalent.

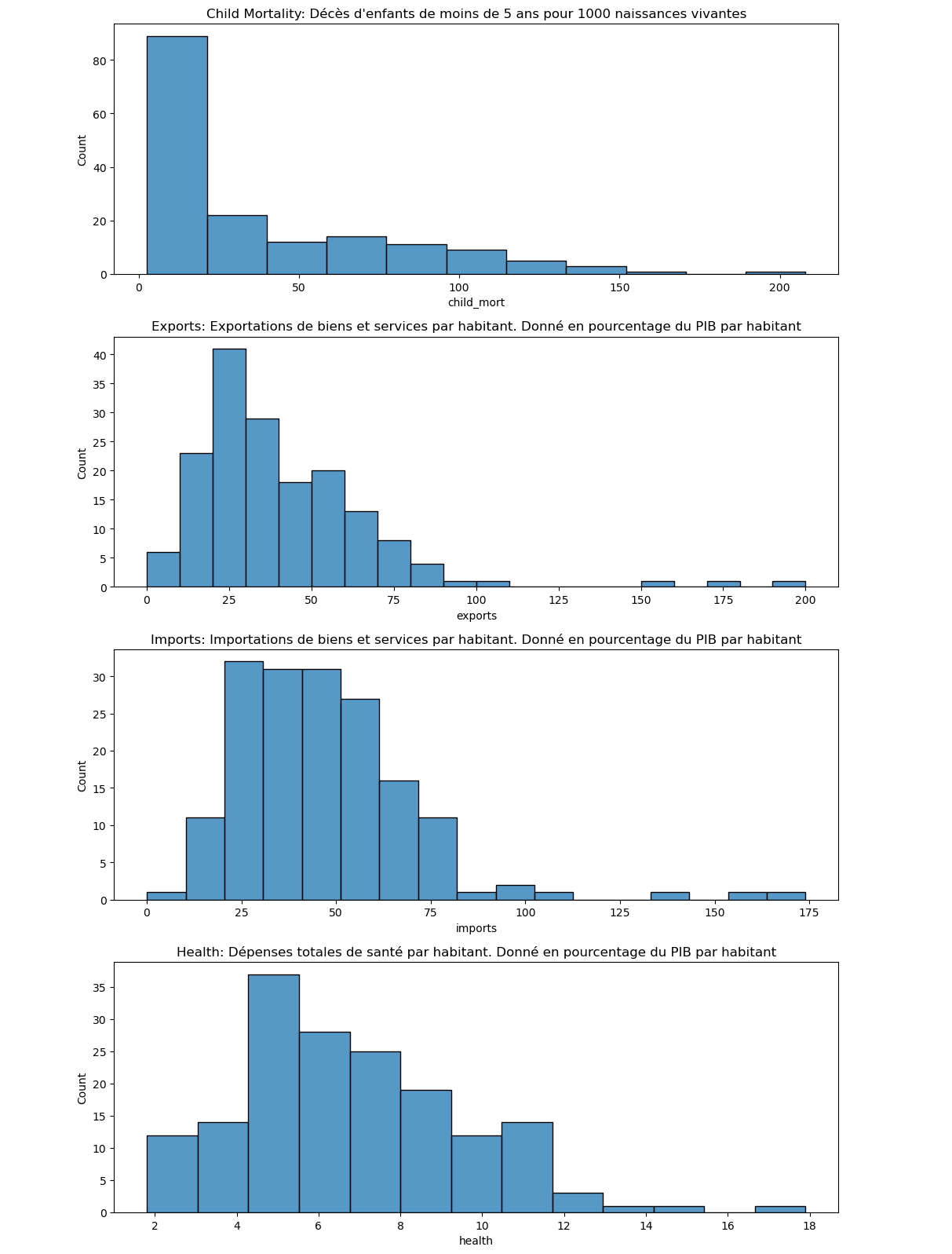
- Niveau de vie faible.

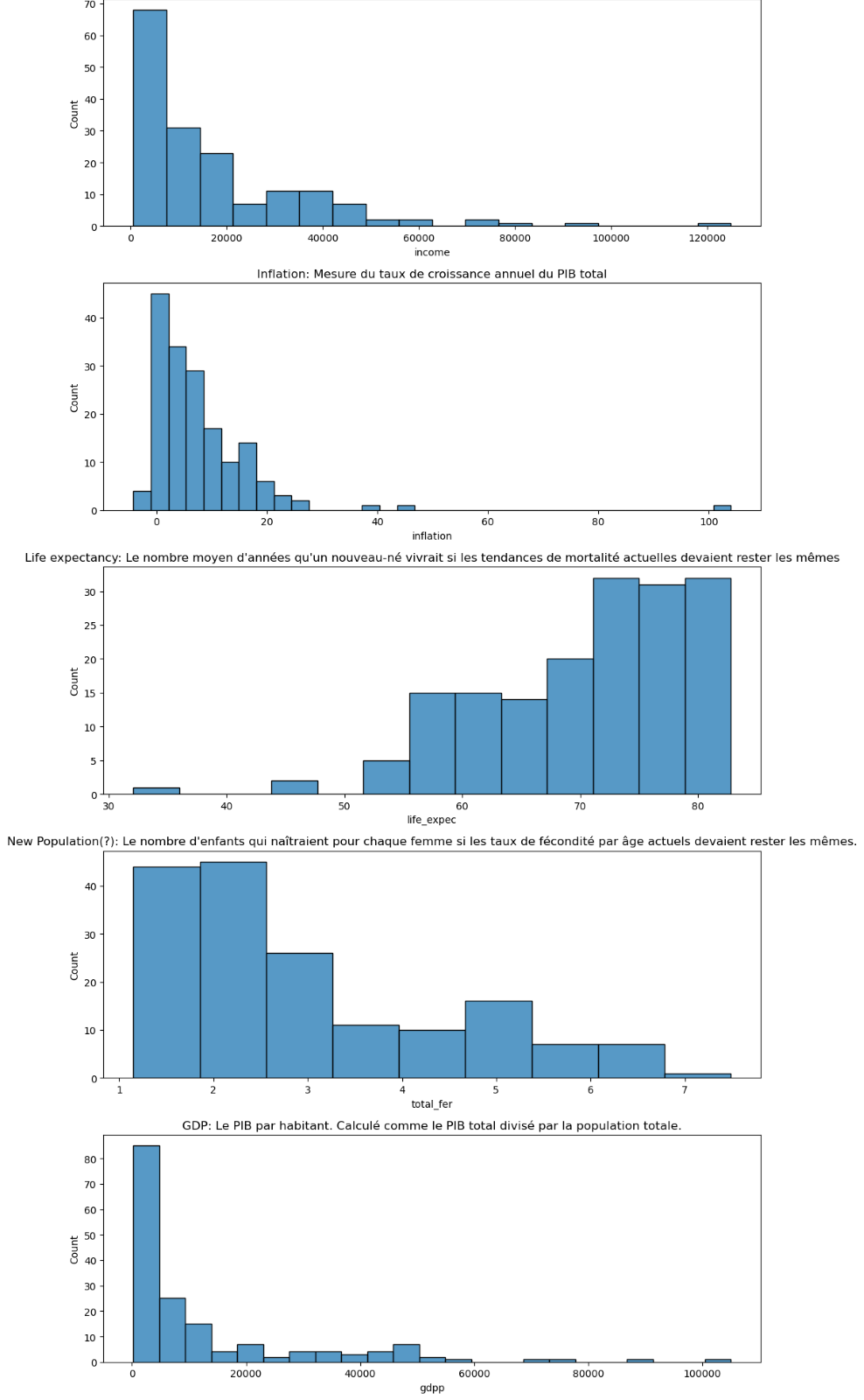
- Aucune avancée technique.

- Services de santé médiocres associés à des taux de natalité et de mortalité élevés.

La Fondation HELP doit cibler les pays qui présentent les caractéristiques ci-dessus. Nous allons maintenant visualiser les données et trouver les pays qui se trouvent aux extrêmes et au centre de chaque caractéristique pour indiquer les pays qui ont besoin d'aide !

**Distribution des Données :**





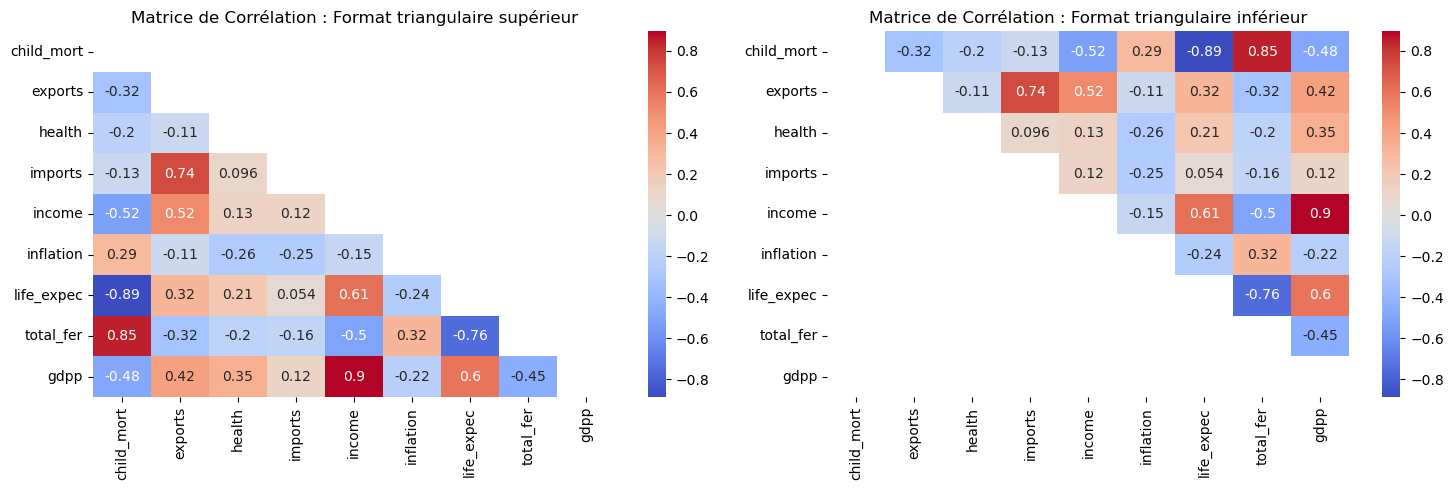
**Constatations :**

* En examinant la distribution des données, nous pouvons constater que certaines caractéristiques ont effectivement des valeurs aberrantes.
* Dans le cadre de cette analyse, les valeurs aberrantes ne seront pas supprimées car elles pourraient être considérées comme très informatives en ce sens qu'elles pourraient indiquer les pays en situation critique et ayant besoin d'aide.
* Par exemple, la mortalité infantile est un indicateur fort de la pauvreté et de la nécessité, donc les valeurs aberrantes dans cette caractéristique montrent qu'il y a des pays avec un nombre de mortalité infantile plus élevé que la normale/critique.

**Feature Engineering :**

**Matrice de Corrélation :**

Nous utilisons une Heatmap pour trouver les relations entre les caractéristiques.



De nombreuses caractéristiques présentent des relations entre elles.

* child\_mort augmente clairement lorsque income, gdpp et exports diminuent. L'augmentation de l'inflation entraîne également des cas de child\_mort élevés. Les conditions économiques agissent malheureusement comme un facteur important !
* L'augmentation des exports entraîne clairement une augmentation du gdpp, income et imports.
* Les dépenses de health ont une légère augmentation de life\_expec et diminuent également child\_mort.

* income et gdpp par habitant affichent une valeur de corrélation très élevée de 0,9. Du point de vue de health, income élevé a entraîné life\_expec plus longue mais a également diminué total\_fer de manière significative.
* Comme prévu, une forte inflation a un effet négatif sur les caractéristiques financières. Une forte inflation affiche un total\_fer et child\_mort élevés. Cela décrit les caractéristiques typiques d'une nation en retard.
* Selon les données, life\_expec plus élevée est associée à un total\_fer plus faible. gdpp par habitant plus élevé a conduit à une augmentation des dépenses de health.

**Catégoriser et Normaliser :**

Nous pouvons clairement voir que certaines caractéristiques sont essentiellement de la même catégorie et qu'elles réagissent de la même manière à d'autres caractéristiques de différentes catégories.

* Les 3 catégories des caractéristiques sont :

    - **santé** : child\_mort, health, life\_expec, total\_fer

    - **commerce** : imports, exports

    - **finances** : income, inflation, gdpp

* Par conséquent, nous allons regrouper ces caractéristiques dans ces catégories et les normaliser !



**Mise à l'échelle ou Scaling :**



* Le modèle d'apprentissage automatique ne comprend pas les unités des valeurs des caractéristiques. Il traite l'entrée simplement comme un nombre, mais ne comprend pas la vraie signification de cette valeur. Ainsi, il devient nécessaire de mettre à l'échelle les données.

* Par exemple : Age = Années ; FastingBS = mg / dl ; Charges = Devise

* Nous avons 2 options pour la mise à l'échelle des données :

1) Normalisation

2) Standardisation.

Comme la plupart des algorithmes supposent que les données sont distribuées normalement (gaussienne), la normalisation est effectuée pour les caractéristiques dont les données ne présentent pas de

distribution normale et la standardisation est effectuée pour les caractéristiques qui sont normalement distribuées et dont les valeurs sont énormes ou très petites par rapport aux autres caractéristiques.

* **Normalisation** : Les caractéristiques de santé, de commerce et de finances sont normalisées !

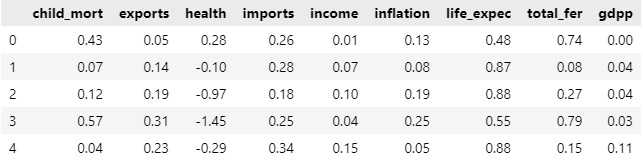
* **Standardisation** : Aucune des caractéristiques n'est standardisée pour les données ci-dessus.

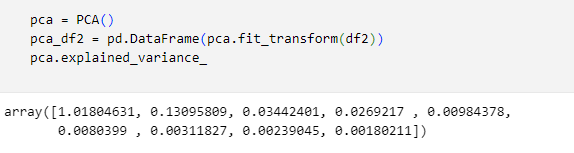
**3. Principal Component Analysis (PCA) :**

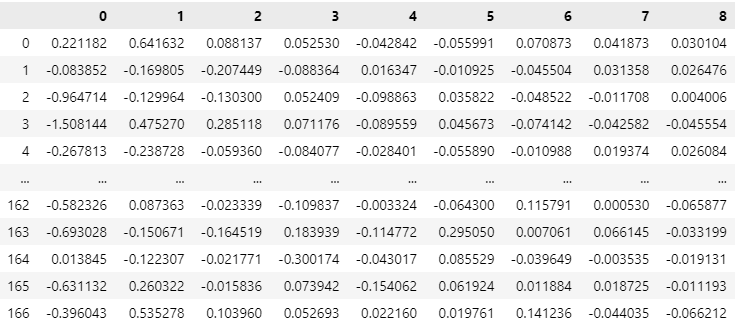
Il s'agit d'une méthode de réduction de dimensionnalité qui est de préférence utilisée pour un problème d'apprentissage non supervisé.

Avant de passer les données pour une transformation PCA, la normalisation et la standardisation sont très importantes.

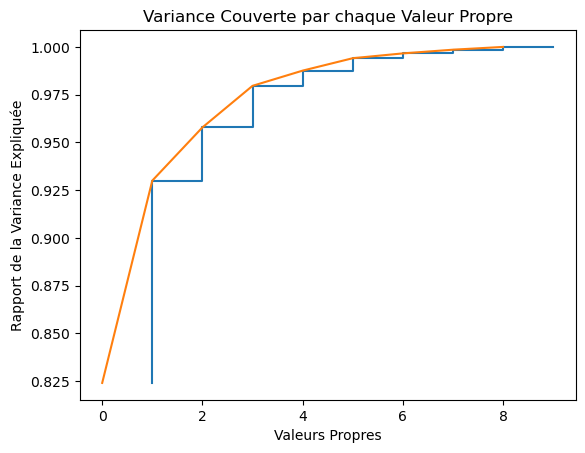




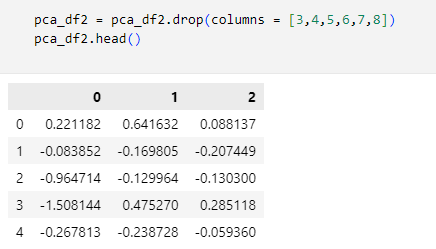


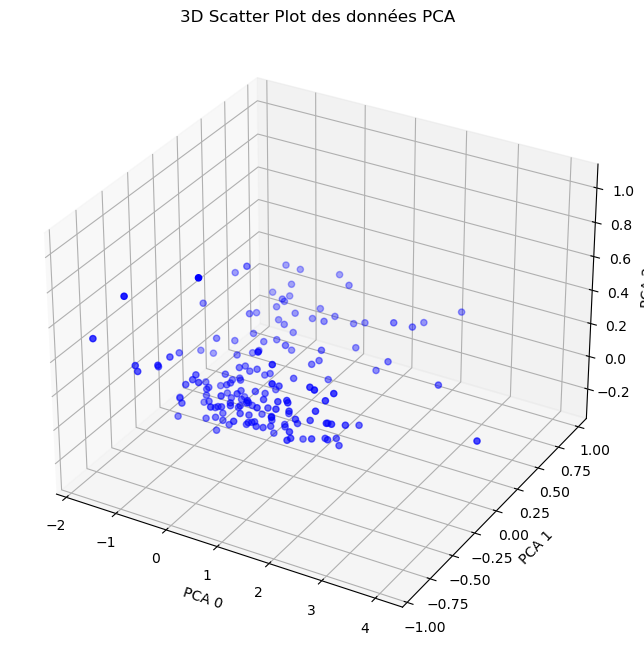


Voici les valeurs de variance de chaque caractéristique présente dans l'ensemble de données.



* C'est une méthode très efficace où nous ajoutons les variances de toutes les caractéristiques de manière cumulative.
* Généralement, les valeurs propres avec plus de 95 % de ratio de variance sont sélectionnées.
* Elles correspondent aux colonnes du dataframe PCA généré.
* Dans ce cas, nous sélectionnons la valeur propre : 2 car les étapes générées ont des variances significatives et donc les autres caractéristiques sont dominées par leurs variances.





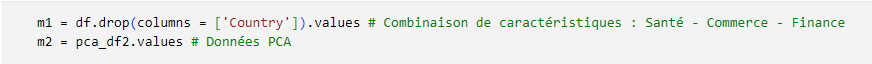
De cette manière, nous réduisons les dimensions !

Nous allons maintenant passer à la section modélisation et comparer les performances des 2 ensembles de données :

* Combinaison de caractéristiques : Santé - Commerce - Finance
* Données PCA

**4. Modélisation - Apprentissage non supervisé :**

* Apprentissage non supervisé : Il s'agit d'un problème où la variable cible / caractéristique est inconnue. Les problèmes d'apprentissage non supervisé surviennent largement dans le domaine médical où de multiples mesures sont prises et les maladies sous-jacentes sont inconnues.
* Ainsi, la recherche de motifs à l'aide de techniques de visualisation nous fournit des informations sous-jacentes qui ne peuvent pas être comprises en regardant simplement les données tabulaires.
* Cependant, lorsque la taille des ensembles de données devient énorme avec trop de caractéristiques, la visualisation des données peut être une tâche fastidieuse à exécuter. Par conséquent, nous nous tournons vers une approche algorithmique.



**Gaussian Mixture :**

Le clustering GM est un algorithme probabiliste utilisé pour l'apprentissage non supervisé, capable de gérer à la fois des caractéristiques numériques et catégorielles.

* C'est un algorithme polyvalent pour l'apprentissage non supervisé, chargé d'identifier des clusters au sein des données, en utilisant une approche probabiliste :

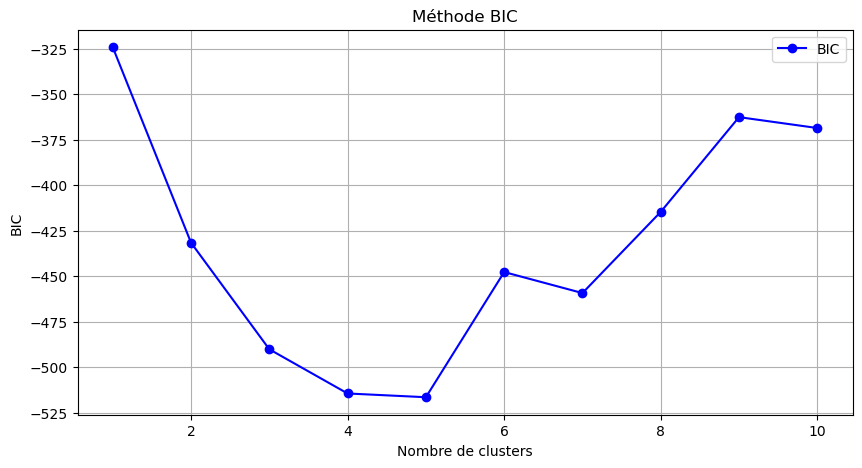
* Sélectionne aléatoirement des paramètres initiaux pour les clusters.
* Attribue chaque point de données au cluster qui maximise la vraisemblance des données observées.
* Met à jour les paramètres de chaque cluster en fonction des points de données qui lui sont attribués.
* Répète ce processus jusqu'à ce que la convergence soit atteinte, où les paramètres ne changent plus de manière significative.

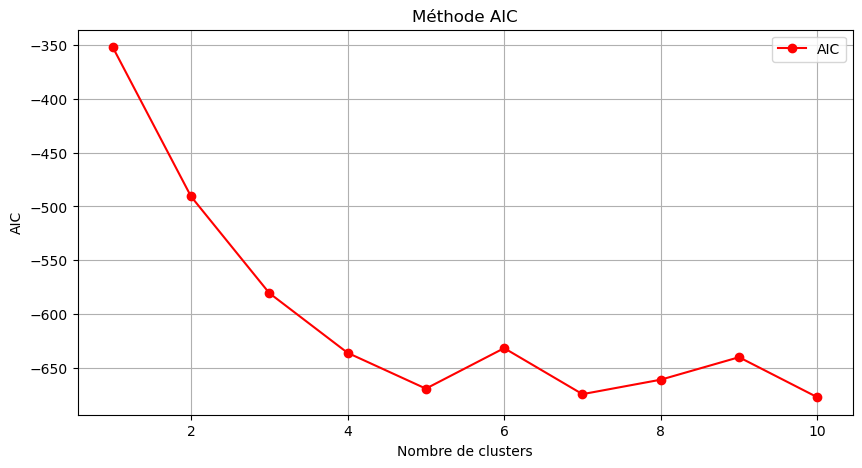
* Nombre de clusters : Ce paramètre hyperdéfini le nombre de clusters dans lesquels les données seront partitionnées. Pour sélectionner une valeur appropriée pour k, nous utilisons deux méthodes statistiques :

* **BIC** (Critère d'Information Bayésien) : Cette méthode évalue la qualité d'ajustement d'un modèle aux données, en pénalisant les modèles avec plus de paramètres. Le modèle avec la plus faible valeur de BIC est choisi comme nombre optimal de clusters.
* **AIC** (Critère d'Information d'Akaike) : Similaire au BIC, l'AIC mesure également la qualité d'ajustement d'un modèle aux données, mais avec un terme de pénalisation différent pour le nombre de paramètres. Le modèle avec la plus faible valeur d'AIC est sélectionné comme nombre optimal de clusters.

**Combinaison de caractéristiques : Santé - Commerce – Finance**

**BIC Methode et AIC Methode :**

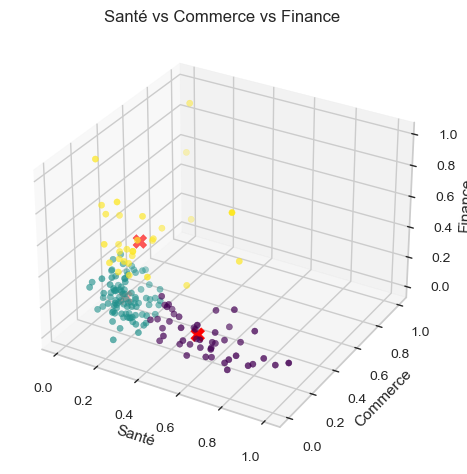




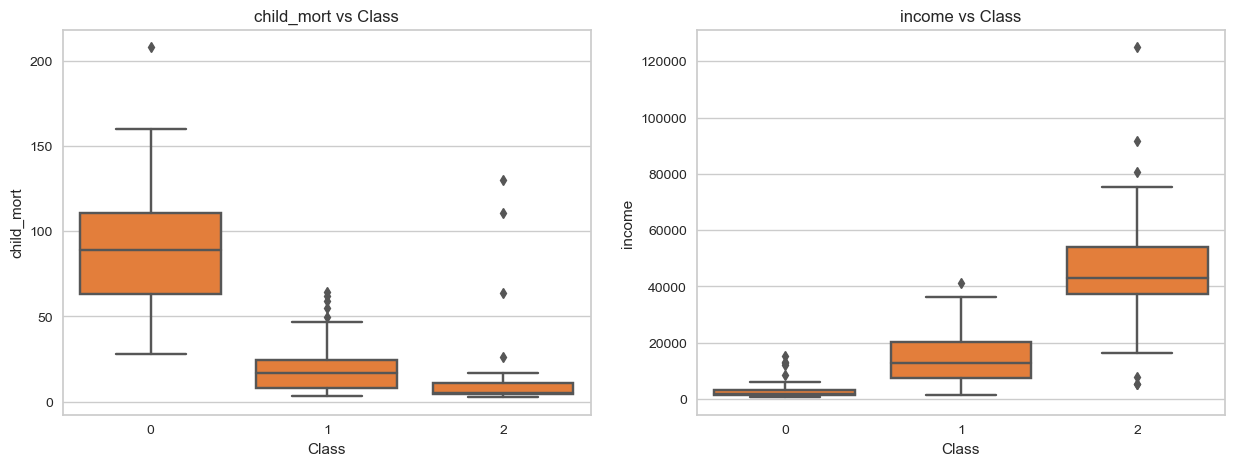
À partir des résultats des deux méthodes ci-dessus, nous sélectionnons :

  - k : Clusters = 3

**Gaussian Mixture :**



* Nous vérifions à nouveau scatterplot du income et du taux de child\_mort par rapport aux clusters étiquetés pour confirmer les valeurs des clusters !
* Nous savons qu'un faible income et une forte child\_mort sont des signes d'un pays économiquement défavorisé.

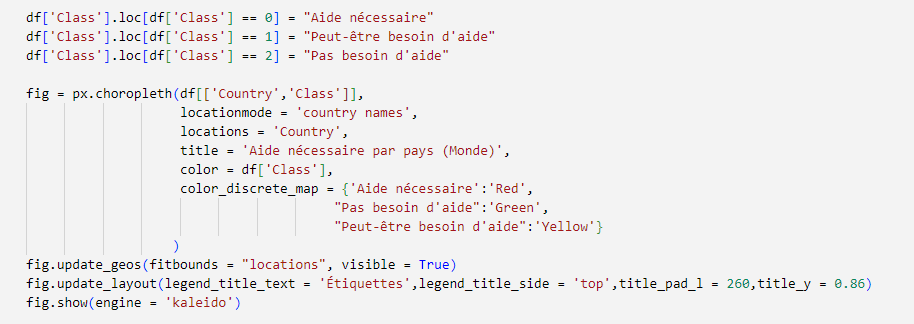


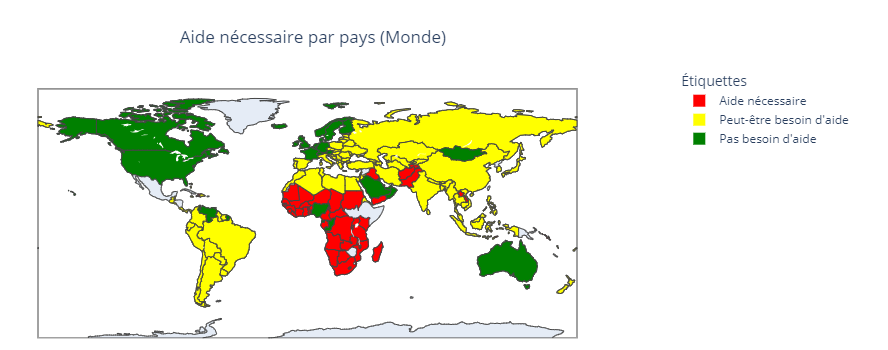
À partir du graphique ci-dessus, nous pouvons conclure :

**- 0 : Aide nécessaire**

**- 1 : Peut-être besoin d'aide**

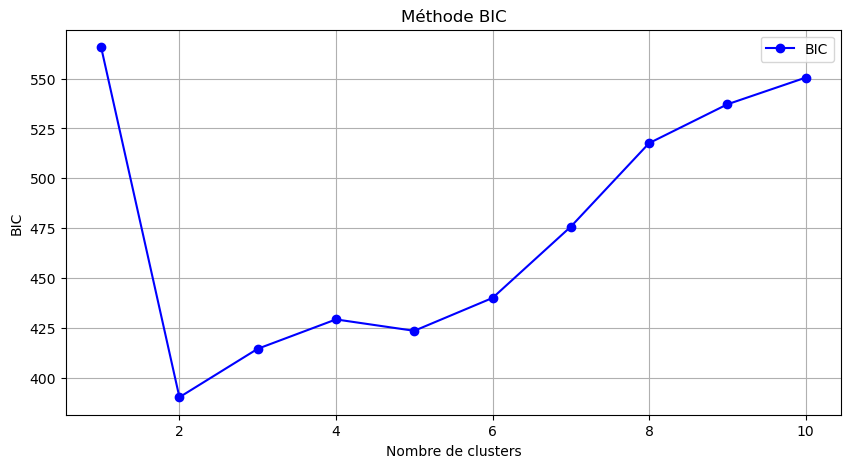
**- 2 : Pas besoin d'aide**

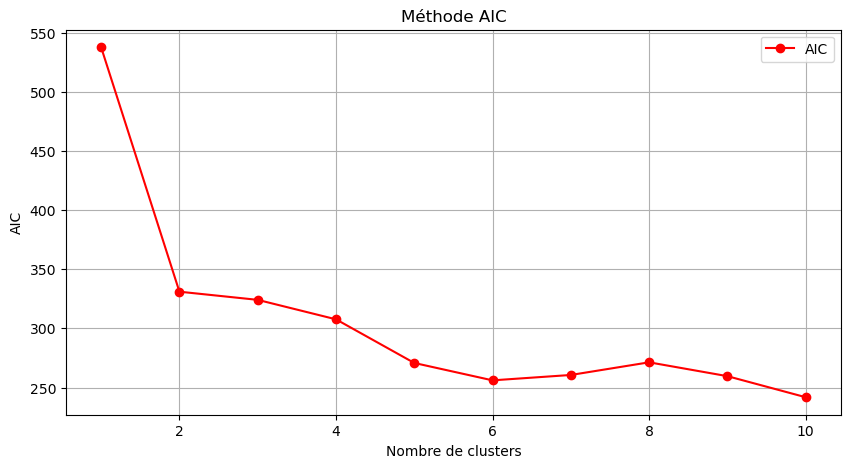




**PCA :**

**BIC Methode et AIC Methode :**

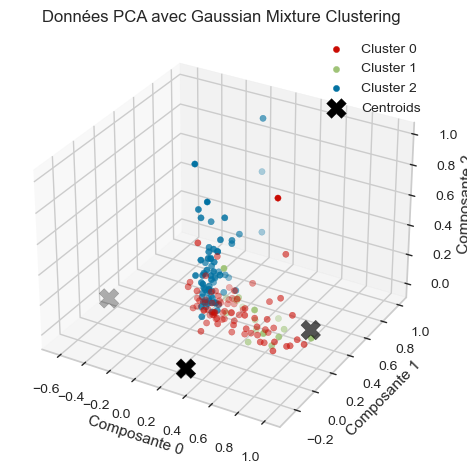




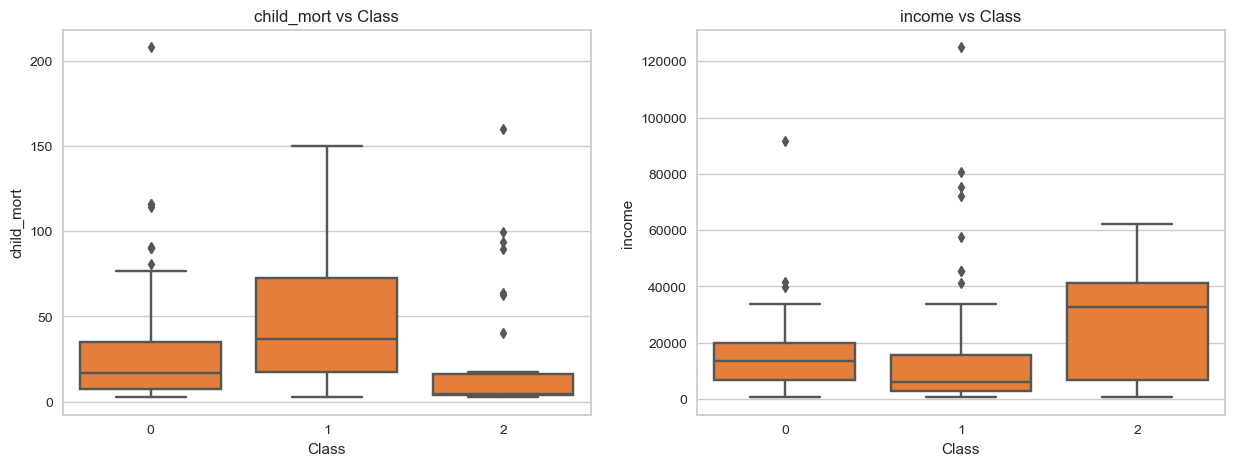
À partir des résultats des deux méthodes ci-dessus, nous sélectionnons :

  - k : Clusters = 3

**Gaussian Mixture :**



* Nous vérifions à nouveau scatterplot du income et du taux de child\_mort par rapport aux clusters étiquetés pour confirmer les valeurs des clusters !
* Nous savons qu'un faible income et une forte child\_mort sont des signes d'un pays économiquement défavorisé.

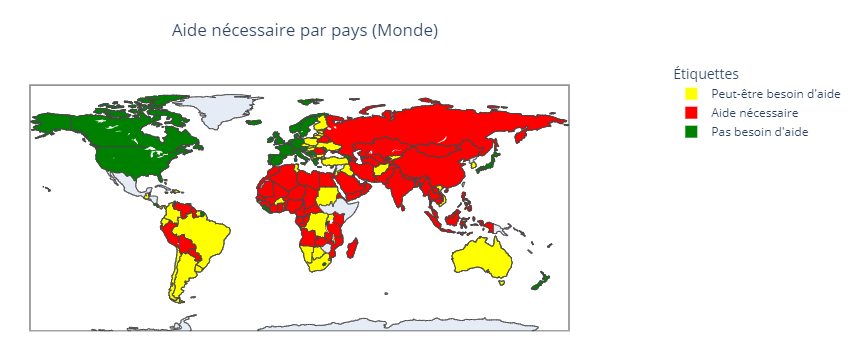


À partir du graphique ci-dessus, nous pouvons conclure :

**- 0 : Peut-être besoin d'aide**

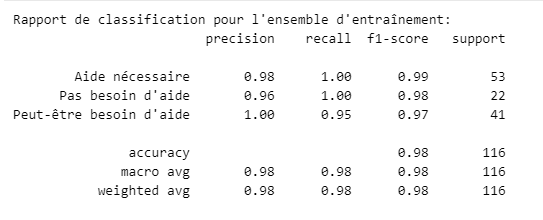
**- 1 : Aide nécessaire**

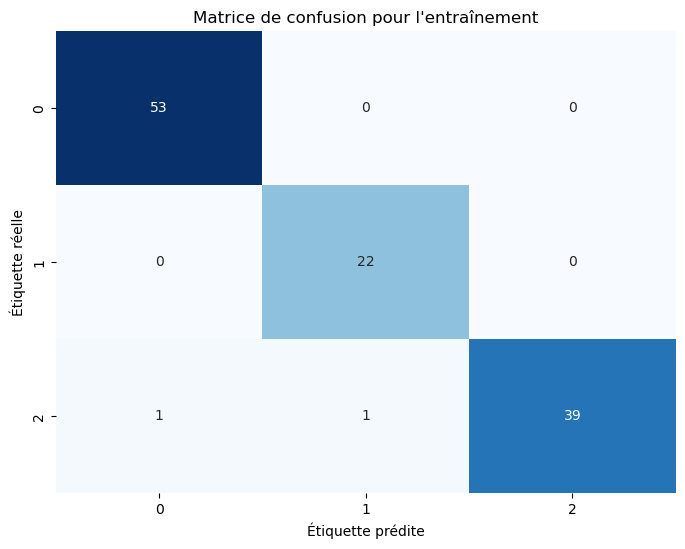
**- 2 : Pas besoin d'aide**

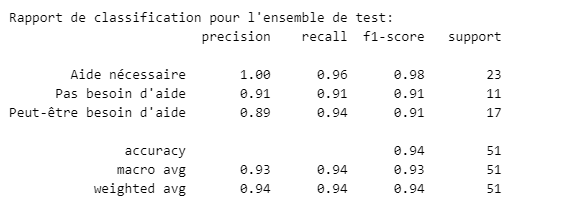


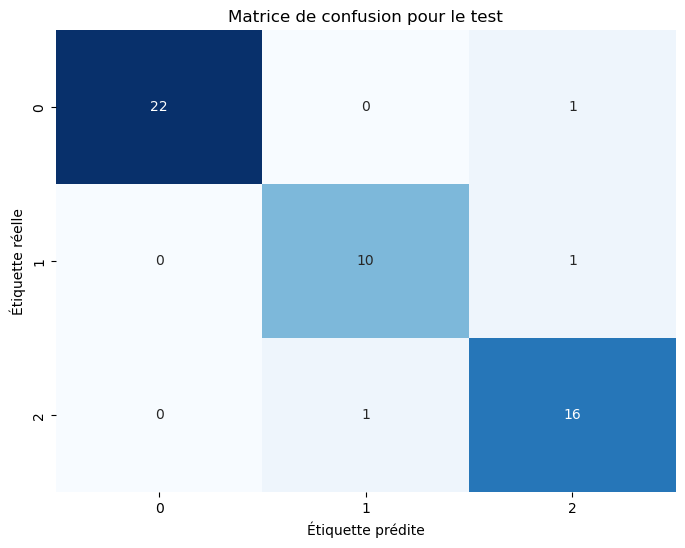
**5. Modélisation - Apprentissage Supervisé (KNN) :**











**Sauvegarder le model :**



**Mise en œuvre du Modèle Flask :**

**Structure de Projet :**

* **images** : Un répertoire contenant des images utilisées dans votre application Flask.

* **static** : Un répertoire contenant des fichiers statiques tels que des feuilles de style CSS (dans le sous-répertoire "css") et des scripts JavaScript (dans le sous-répertoire "js").

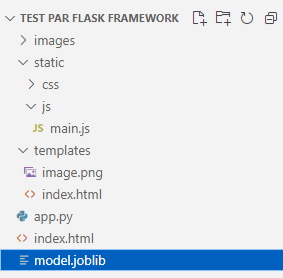
* **templates** : Un répertoire contenant des modèles HTML utilisés pour générer des pages Web dynamiques dans votre application Flask.

* **main.js** : Un fichier JavaScript utilisé dans votre application Flask.

* **Index.html** : Une page HTML principale ou d'accueil pour votre application Flask.

* **app.py :** Le fichier principal de votre application Flask qui contient le code pour l'application web, y compris le chargement du modèle "model.joblib" et son utilisation pour effectuer des prédictions.

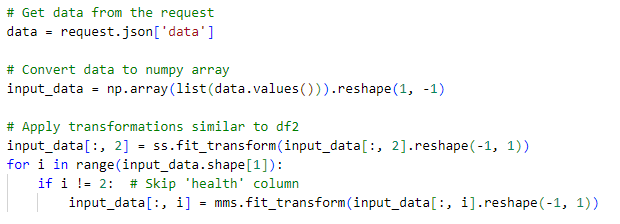
* **model.joblib :** Le modèle sauvegardé que votre application Flask utilise pour effectuer des prédictions. Ce modèle peut avoir été préalablement entraîné en utilisant des données et des techniques d'apprentissage automatique.



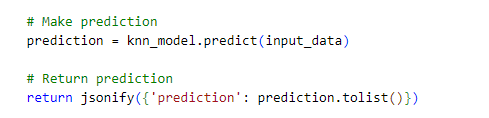
**Charger le modèle :**

10962282545ec544a4a943f99f32aa08.png

**Recevoir et appliquer la normalisation et standarisation Sur les données de l’utilisateur :**



**Appel le modèle et passer les données d’utilisateur pour lui et envoi la prédiction sous forme json :**



Cette interface utilisateur fournit une manière conviviale pour les utilisateurs de spécifier les caractéristiques des pays et d'obtenir des prédictions sur les pays qui devraient recevoir des fonds en fonction de ces caractéristiques.



