

**Projet SMA : Algorithmes de Machine Learning**

**( PCA - Gaussian Mixture - KNN )**



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

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

**Plan**

1. **La description des Algorithmes**

* Principal Component Analysis (PCA)
* Gaussian Mixture
* K-Nearest Neighbors (KNN)

1. **Les caractéristiques des Algorithmes**

* Caractéristiques de PCA, Gaussian Mixture, KNN

1. **Les modes de fonctionnement des trois algorithmes**

* KNN, GMMs, PCA

1. **Les avantages et Les Inconvénients**

* Les avantages
* Les inconvénients

1. **Champs d'Application des algorithmes et étude comparative**

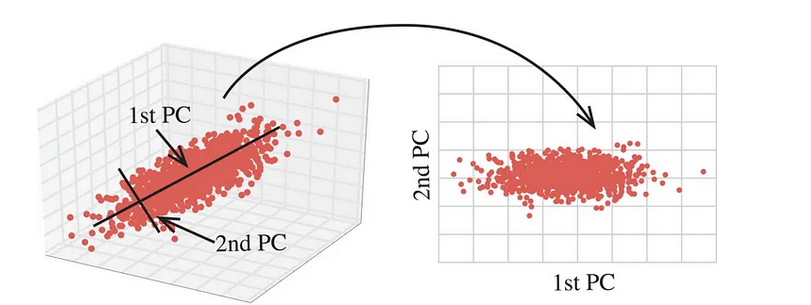
* Champs d'Application
* Etude comparative

1. **Etude d'un cas**

**I. La description des Algorithmes**

**1.Principal Component Analysis (PCA)**

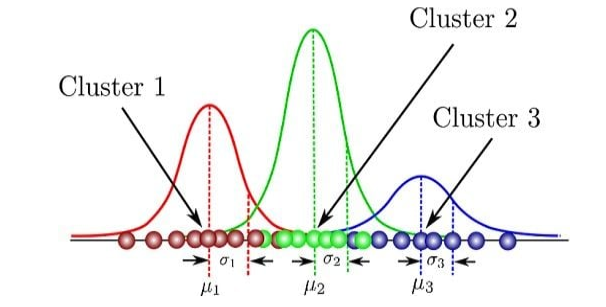
PCA est une technique de réduction de dimensionnalité qui vise à transformer un ensemble de variables corrélées en un ensemble de variables non corrélées, appelées composantes principales. Elle est largement utilisée pour explorer et visualiser des données à haute dimensionnalité, ainsi que pour simplifier la structure des données tout en préservant les informations importantes.



* La première image, des données tridimensionnelles avec les axes X, Y et Z.
* La deuxième image représente un espace bidimensionnel avec PC1 et PC2 comme axes.

**2.Gaussian Mixture**

Le modèle de mélange gaussien (Gaussian Mixture Model ou GMM en anglais) est un modèle probabiliste utilisé pour représenter la distribution de probabilité d'un ensemble de données. Contrairement aux méthodes de classification traditionnelles, où chaque observation est attribuée à une seule classe, le GMM suppose que les données sont générées à partir d'un mélange de plusieurs distributions gaussiennes (d'où le terme "mélange").



* Voilà, nous pouvons voir qu'il y a trois fonctions gaussiennes, donc **K = 3**. Chaque gaussienne explique les données contenues dans chacun des trois clusters disponibles.

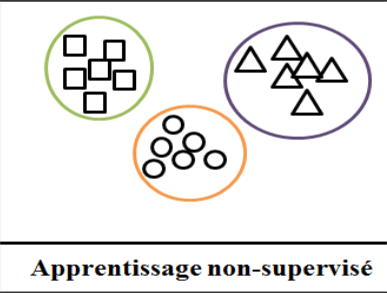
**3.Nearest Neighbors (KNN)**

L'algorithme des k plus proches voisins (KNN) est l'une des techniques les plus simples et les plus populaires en apprentissage automatique, utilisée pour la classification et la régression. KNN est un algorithme d'apprentissage supervisé non paramétrique, ce qui signifie qu'il ne fait aucune hypothèse sur la distribution des données. Au lieu de cela, il classe ou prédit de nouvelles observations en se basant sur les exemples d'entraînement les plus similaires qui lui sont les plus proches dans l'espace des caractéristiques.

**II. Les caractéristiques des Algorithmes**

**1.Caractéristiques de PCA**

* **Réduction de dimensionnalité :** PCA est principalement utilisé pour réduire la dimensionnalité des données en transformant un ensemble de variables corrélées en un nouvel ensemble de variables non corrélées, appelées composantes principales. Cela permet de représenter les données de manière plus compacte tout en conservant au maximum les informations importantes.
* **Non supervisé :** PCA est un algorithme d'apprentissage non supervisé, ce qui signifie qu'il n'a pas besoin d'étiquettes de classe pour apprendre à partir des données.



Il explore simplement la structure intrinsèque des données.

* **Capture de la variance maximale** : PCA cherche à capturer la variance maximale des données dans les premières composantes principales. Les premières composantes principales représentent donc les directions dans lesquelles les données varient le plus.
* **Matrice de Covariance :** Étant donné un ensemble de données X avec n observations et p caractéristiques, nous calculons la matrice de covariance Σ comme suit :

Insertion de l’image...

**Où Xˉ est la moyenne de X.**

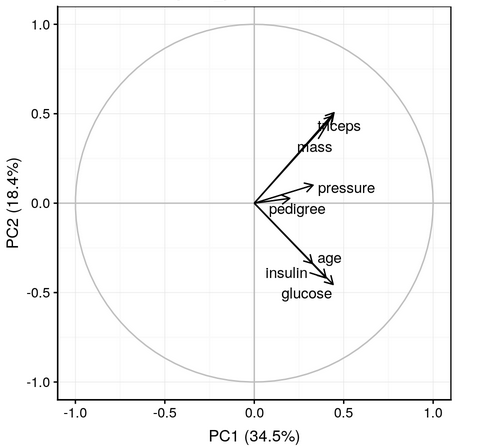
* **Décomposition en Valeurs Propres :** Nous décomposons la matrice de covariance Σ en ses vecteurs propres et ses valeurs propres :

Insertion de l’image...

L'ACP sélectionne les k vecteurs propres correspondant aux k plus grandes valeurs propres pour former le nouveau sous-espace, et ous projetons les données originales sur le nouveau sous-espace formé par les composantes principales sélectionnées :

Insertion de l’image...

* **Orthogonalité des composantes** : Les composantes principales sont orthogonales entre elles, ce qui signifie qu'elles sont indépendantes les unes des autres. Cela permet de réduire la redondance des informations dans les données transformées.



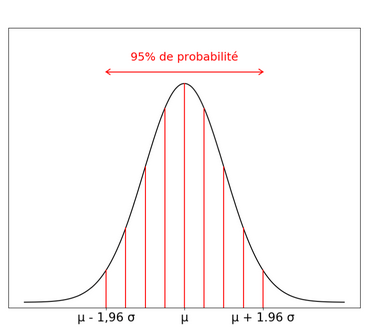
Ce graphique s’interprète comme suit :

1. Plus la norme (longueur) du vecteur qui représente une variable est grande et se rapproche d'un (matérialisé par le clé gris), plus la variable est bien représentée dans le plan choisi. On évitera d’interpréter ici les variables qui ont des normes petites, comme pedigree ou pressure.
2. Des vecteurs qui pointent dans la même direction représentent des variables directement corrélées entre elles. C’est le cas de glucose, insuline et age d’une part, et par ailleurs aussi de masse et triceps.
3. Des vecteurs qui pointent en directions opposées représentent des variables inversement proportionnelles. Il n’y en a pas ici.

* **L'interprétabilité :** Bien que les composantes principales soient des combinaisons linéaires des variables originales, elles peuvent souvent être interprétées intuitivement en termes des variables originales. Par exemple, une composante principale peut représenter une combinaison linéaire de caractéristiques spécifiques dans les données.
* **Préserve la structure des données :** PCA conserve la structure globale des données, ce qui signifie qu'elle peut être utilisée pour visualiser des données à haute dimensionnalité dans un espace de dimension réduite tout en préservant les relations entre les points.
* **Sensibilité à l'échelle des variables :** PCA est sensible à l'échelle des variables, ce qui signifie que les variables avec des échelles différentes peuvent avoir un impact disproportionné sur les composantes principales. Par conséquent, il est souvent nécessaire de standardiser les données avant d'appliquer PCA.

**2.Caractéristiques de Gaussian Mixture**

**Modèle probabiliste :** Un GMM considère que les données sont générées à partir d'un processus probabiliste où chaque composante gaussienne représente une classe ou un groupe latent.



**Mélange de gaussiennes :** Le modèle suppose que les données sont issues d'un mélange de plusieurs distributions gaussiennes, chacune avec sa propre moyenne et sa propre matrice de covariance.

**Paramètres :** Les paramètres du modèle comprennent les poids (ou proportions) de chaque composante gaussienne, ainsi que les moyennes et les matrices de covariance de chaque composante.

Les poids des composantes gaussiennes doivent satisfaire les contraintes suivantes :

Insertion de l’image...

Avec Insertion de l’image...

Chaque composante gaussienne a une matrice de covariance qui est une matrice symétrique d×d où d est le nombre de dimensions des données.

Insertion de l’image...

En combinant ces éléments, nous pouvons définir le modèle de mélange de gaussienne

Insertion de l’image...

**Flexibilité :** Les GMM peuvent modéliser des distributions de données complexes, notamment celles qui ne sont pas nécessairement bien représentées par une seule distribution gaussienne.

**Estimation des paramètres** : Les paramètres du GMM (poids, moyennes, matrices de covariance) sont souvent estimés à l'aide de l'algorithme de maximisation de l'espérance (EM), également appelé algorithme d'espérance-maximisation.

1. Les paramètres du GMM (poids, moyennes, matrices de covariance) sont initialement choisis de manière aléatoire ou par une méthode heuristique.
2. Calcul des responsabilités γ pour chaque point de données xn et chaque composante gaussienne k

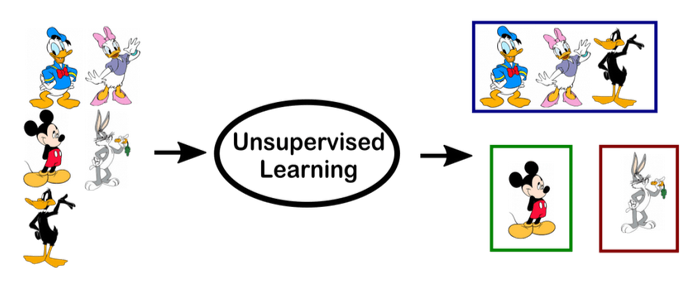
Insertion de l’image...

1. Mise à jour des paramètres en maximisant la fonction de log-vraisemblance du modèle

Insertion de l’image...

1. Répéter les étapes E et M jusqu'à ce que la convergence soit atteinte

**Utilisation en classification non supervisé** : Les GMM sont couramment utilisés pour la segmentation et la classification non supervisée des données, où l'objectif est de découvrir la structure sous-jacente des données sans étiquettes préalables.



**Sensibilité à l’initialisation :** Comme l'algorithme EM peut converger vers des minima locaux, l'initialisation des paramètres du GMM peut avoir un impact significatif sur la qualité des résultats.

Par conséquent, il est crucial de choisir soigneusement les valeurs initiales des paramètres pour obtenir des résultats de clustering optimaux avec un modèle de mélange de gaussiennes.

**3.Caractéristiques de KNN**

**Non paramétrique :** Contrairement à de nombreux autres algorithmes d'apprentissage automatique, KNN est un algorithme non paramétrique, ce qui signifie qu'il n'a pas de paramètres fixes qui doivent être ajustés avant la modélisation.

Néanmoins, il est important de prendre en compte ses limites en termes de calcul et de sensibilité au bruit lors de sa mise en œuvre.

**Basé sur l’instance :** L'algorithme des k plus proches voisins (KNN) est un exemple d'algorithme d'apprentissage basé sur l'instance. Il ne génère pas explicitement de modèle lors de la phase d'apprentissage, mais il stocke simplement les exemples d'entraînement dans leur intégralité pour les utiliser comme référence lors de la phase de prédiction. Voici comment celafonctionne, sans formules mathématiques, mais plutôt une explication conceptuelle :

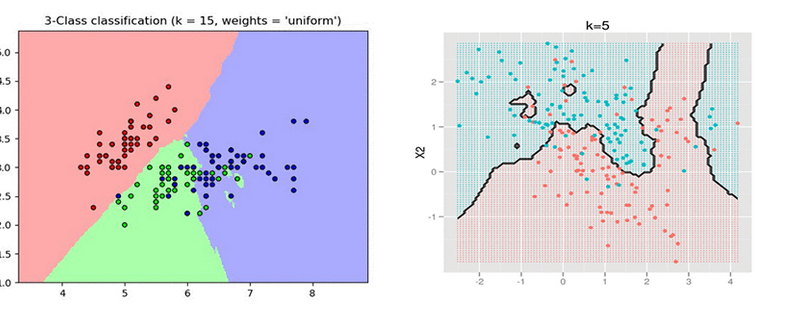
Phase d'apprentissage :

* L'algorithme KNN se contente de stocker les données d'entraînement dans leur intégralité, sans calculer de modèle paramétrique ou d'hypothèses spécifiques.

hase de prédiction :

* Lorsqu'une nouvelle donnée à prédire est présentée, l'algorithme KNN examine les k exemples les plus proches dans l'espace des caractéristiques parmi les données d'entraînement

**Classification et régression :** KNN peut être utilisé pour la classification (prédiction de classes discrètes) et la régression (prédiction de valeurs continues) en fonction de la façon dont la fonction de distance est calculée et de la manière dont les prédictions sont agrégées.



**III. Les modes de fonctionnement des trois algorithmes**

**1.KNN**

**Charger les données :** Les données d'entraînement sont chargées dans l'algorithme. Soit D l'ensemble de données chargé.

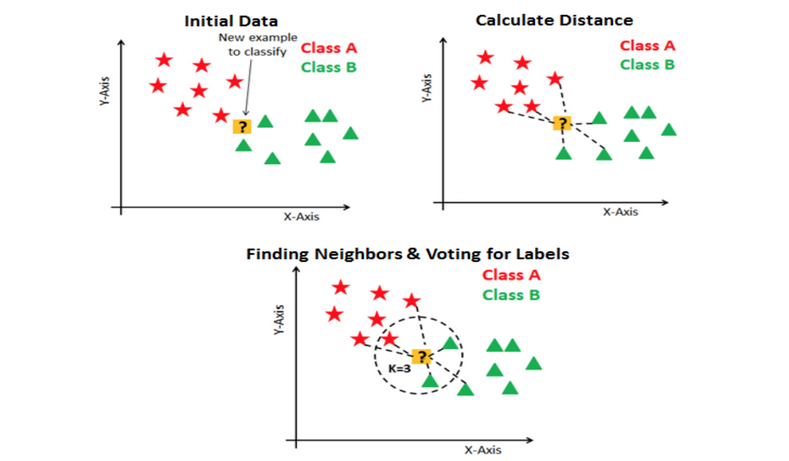
**Initialiser la valeur de k :** Vous choisissez un entier positif k qui représente le nombre de voisins à considérer lors de la prédiction.

**Itérer sur les données d'entraînement :** Pour chaque point de données dans l'ensemble de données d'entraînement D, les étapes suivantes sont répétées pour obtenir la classe prédite pour l'exemple de test xtest .

**Calculer la distance :** La distance entre l'exemple de test xtest et chaque ligne de données d'entraînement xi est calculée en utilisant la distance euclidienne

Insertion de l’image...

où xtest,j et xi,j représentent respectivement la j-ème caractéristique de l'exemple de test et de l'exemple d'entraînement.



**Trier les distances :** Les distances calculées sont triées par ordre croissant.

Obtenir les k premières lignes triées **:** Les k premières lignes du tableau trié sont sélectionnées.

**Obtenir la classe majoritaire :** La classe la plus fréquente parmi ces k voisins est identifiée. C'est-à-dire qu'on compte le nombre d'exemples de chaque classe parmi les k voisins et on choisit la classe qui apparaît le plus souvent.

**Retourner la classe prédite :** La classe prédite pour l'exemple de test xtest est la classe majoritaire identifiée dans l'étape précédente.

**2.GMMs**

**Initialisation :** Initialement, les paramètres du modèle doivent être initialisés. Cela inclut les moyennes, les covariances et les poids des composantes gaussiennes.

**Calcul de la densité de probabilité :** Pour chaque point de données xi , on calcule la densité de probabilité qu'il appartienne à chaque composante gaussienne. La densité de probabilité d'un point xi appartenant à la j-ème composante gaussienne est donnée par :

Insertion de l’image...

où μj est le vecteur moyen (moyenne) de la j ème composante, Σj est sa matrice de covariance, et d est la dimension de l'espace des caractéristiques.

**Estimation de la responsabilité :** On estime la "responsabilité" de chaque composante gaussienne pour chaque point de données. La responsabilité rij de la j ème composante gaussienne pour le point xi est donnée par la probabilité que le point xi a été généré par la composante j, normalisée par la somme des probabilités de toutes les composantes gaussiennes :

Insertion de l’image...

où πj est le poids (ou la proportion) de la j-ème composante gaussienne dans le mélange.

**Mise à jour des paramètres :** Les paramètres du modèle (moyennes, covariances, poids) sont mis à jour en maximisant la log-vraisemblance des données. Les nouvelles estimations des paramètres sont calculées en utilisant les responsabilités estimées :

* Moyenne mise à jour : **μj =∑i=1N rij ∑i=1N rij xi**

* Matrice de covariance mise à jour : **Σj =∑i=1N rij ∑i=1N rij (xi −μj )(xi −μj )T**
* Poids mis à jour : **πj =N∑i=1N rij**

où N est le nombre total de points de données.

**Répéter :** Les étapes 2 à 4 sont répétées jusqu'à ce que les paramètres convergent, c'est-à-dire jusqu'à ce que les changements dans les paramètres deviennent négligeables ou jusqu'à ce qu'un nombre maximal d'itérations soit atteint.

**Utilisation du modèle :** Une fois que les paramètres ont convergé, le modèle peut être utilisé pour estimer la probabilité qu'un nouveau point de données appartienne à chaque composante gaussienne

**3.PCA**

**Centrage des données :** Tout d'abord, les données sont centrées pour que chaque variable ait une moyenne de zéro. Cela est réalisé en soustrayant la moyenne de chaque variable à ses observations respectives.

La formule pour centrer les données est :

Insertion de l’image...

où xi est la valeur brute de la variable, et xˉ est la moyenne de la variable.

**Calcul de la matrice de covariance :** Ensuite, la matrice de covariance des données centrées est calculée. La matrice de covariance mesure la variabilité conjointe entre les différentes variables.

Soit X la matrice des données centrées, où chaque ligne représente une observation et chaque colonne représente une variable. La matrice de covariance Σ est calculée comme suit :

Insertion de l’image...

où N est le nombre d'observations.

**Décomposition en valeurs propres :** Ensuite, la matrice de covariance est décomposée en valeurs propres et vecteurs propres. Les vecteurs propres représentent les directions principales des données, tandis que les valeurs propres indiquent l'importance de ces directions principales.

Soit **λ1,λ2,...,λp** les valeurs propres de la matrice de covariance Σ, triées par ordre décroissant, et **v1,v2,...,vp** les vecteurs propres correspondants.

Les valeurs propres représentent la quantité de variance expliquée par chaque composante principale.

**Sélection des composantes principales :** Vous choisissez les k premières composantes principales qui captent la plupart de la variance dans les données. Habituellement,

**Projection des données :** Enfin, les données sont projetées sur les k premières composantes principales pour obtenir une nouvelle représentation des données dans un espace de dimension réduite.

**IV. Les avantages et Les Inconvénients**

**1.Les avantages**

**1.1 PCA**

* PCA est utile pour réduire la dimensionnalité des données, ce qui simplifie l'analyse en créant de nouvelles variables qui capturent l'essentiel de l'information.
* En réduisant la dimensionnalité, PCA facilite également la visualisation des relations entre les observations, en particulier en les représentant dans des espaces à deux ou trois dimensions.
* Une autre qualité de PCA est qu'il produit des composantes principales non corrélées, ce qui peut aider à interpréter les relations entre les variables.
* De plus, en réduisant la dimensionnalité, PCA peut améliorer les performances des modèles d'apprentissage automatique en accélérant l'entraînement.
* Il est également utile pour identifier les variables les plus importantes dans les données et pour traiter le bruit en ne conservant que les dimensions significatives.
* En résumé, PCA est un outil polyvalent pour l'analyse des données, offrant des avantages significatifs en termes de réduction de la dimensionnalité, de visualisation des données et d'amélioration des performances des modèles.

**1.2 KNN**

* KNN offre une facilité d'implémentation appréciable, car il est relativement simple à comprendre et à mettre en œuvre, sans nécessiter de modélisation complexe.
* De plus, contrairement à de nombreux autres algorithmes, il ne nécessite pas de phase d'apprentissage formelle, ce qui le rend rapide à utiliser.
* Sa flexibilité est également un avantage majeur, car il peut être utilisé pour les problèmes de classification et de régression, tout en étant compatible avec différents types de données, y compris les données numériques, catégoriques et binaires.

Enfin, les prédictions de KNN sont souvent interprétables, ce qui signifie qu'elles sont faciles à comprendre et à expliquer, car elles se basent sur les voisins les plus proches.

**1.3 GMMs**

* Large gamme de produits répondant à divers besoins de transport.
* Positionnement en tant que leader dans l'innovation technologique, notamment dans les domaines des véhicules électriques et autonomes.
* Solide réputation et longue histoire dans l'industrie automobile, ce qui renforce la confiance des consommateurs.
* Présence mondiale à travers un vaste réseau de concessionnaires et de
* Services, assurant une accessibilité globale des produits.

**2.Les inconvénients**

**2.1 PCA**

**Perte d’interprétabilité des variables** : Les relations entre les variables originales et les composantes principales générées par PCA peuvent être difficiles à interpréter, ce qui peut rendre l'analyse moins intuitive.

**Sensibilité aux données aberrantes :** Les données aberrantes peuvent perturber la structure des données et affecter les résultats de PCA, en particulier si elles sont présentes en grand nombre ou si elles ont un impact important sur la variance des données.

**Assomption de linéarité :** PCA suppose que les relations entre les variables sont linéaires. Si les relations sont non linéaires, PCA peut ne pas être efficace pour capturer la structure des données, ce qui peut conduire à une perte d'information importante.

**2.2 KNN**

**Sensibilité à la dimensionnalité :** KNN peut être sensible à la dimensionnalité des données, car il considère toutes les caractéristiques pour calculer la similarité entre les points. Dans les espaces de grande dimension, la distance entre les points peut devenir moins significative, ce qui peut affecter les performances de l'algorithme.

**Calcul intensif d la distance :** Pour chaque prédiction, KNN calcule la distance entre le point de test et tous les points d'apprentissage. Cela peut être computationnellement coûteux, surtout avec de grandes quantités de données.

**Besoin de déterminer le paramètre k :** Le choix du nombre de voisins (k) à considérer peut avoir un impact significatif sur les performances de KNN

**2.3 GMMs**

**Dépendance aux carburants fossiles :** Malgré ses progrès dans les véhicules électriques, GM reste fortement dépendant des ventes de véhicules à essence et diesel. Les fluctuations des prix du pétrole et les préoccupations croissantes concernant l'impact environnemental des émissions de gaz à effet de serre peuvent affecter la demande de ses produits.

**Problèmes de rappel et de qualité** : Comme tout grand fabricant automobile, GM a été confronté à des rappels de véhicules en raison de défauts de fabrication ou de conception. Ces rappels peuvent entraîner des coûts importants, des dommages à la réputation de la marque et une perte de confiance des consommateurs.

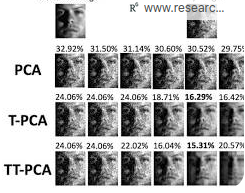
**Dépendance aux cycles économiques** : L'industrie automobile est sensible aux cycles économiques, avec des ventes de véhicules souvent corrélées à la santé économique globale.

**V. Champs application des algorithmes et étude comparative**

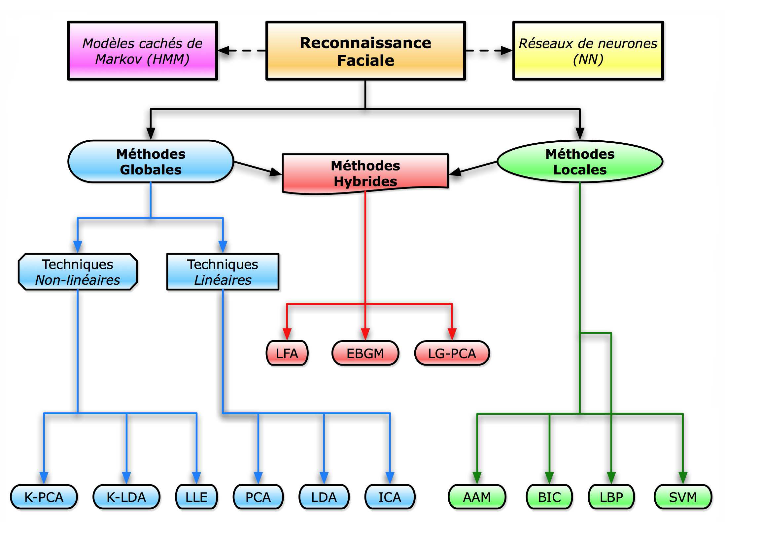
**1.champs d’application**

**1.1PCA**

**Compression d'images** : La PCA peut être utilisée pour compresser des images en réduisant le nombre de pixels nécessaires pour représenter l'image. Cela peut être utile pour le stockage et la transmission d'images.

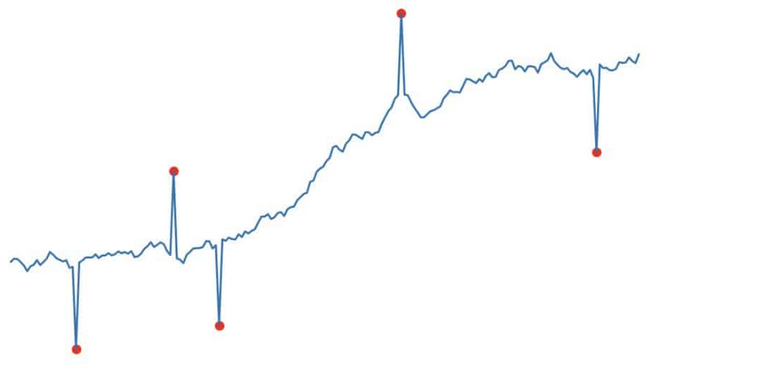


**Reconnaissance faciale** : La PCA peut être utilisée pour reconnaître des visages en extrayant des caractéristiques clés des images faciales. Cela peut être utile pour la vérification d'identité et la sécurité.



La figure 6 fournit une **classification** des algorithmes principaux de reconnaissance faciale.

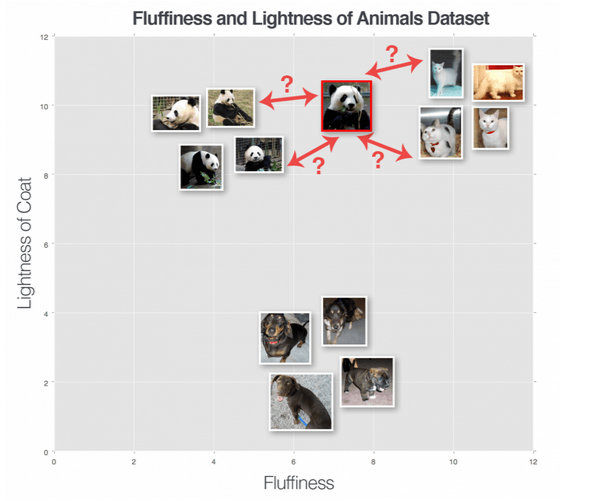
**Détection d'anomalies** : La PCA peut être utilisée pour détecter des anomalies dans les images en identifiant les images qui diffèrent des autres images du jeu de données. Cela peut être utile pour la détection de défauts dans les produits manufacturés ou pour l'identification de tumeurs dans les images médicales.



Il s’agit de suites consécutives de points qui s’écartent de manière significative de la distribution jointe, bien que chaque observation prise individuellement ne soit pas nécessairement une anomalie

**1.2 KNN**

**Classification d'images :** KNN peut être utilisé pour classifier des images dans différentes catégories, telles que des images de chats et de chiens, ou des images de fleurs et d'arbres.



**Recommandation d'images :** KNN peut être utilisé pour recommander des images à des utilisateurs en fonction des images qu'ils ont aimées dans le passé.

**Détection d'objets :** KNN peut être utilisé pour détecter des objets dans des images en identifiant les zones de l'image qui sont similaires à des exemples d'objets connus.

**1.3 GMM**

**Segmentation d'images :** Les GMM peuvent être utilisés pour segmenter des images en différentes régions, telles que le premier plan et l'arrière-plan, ou différents objets dans une scène.

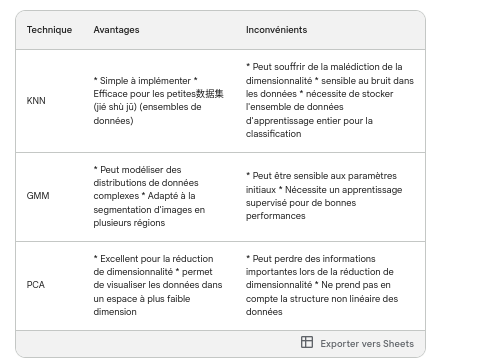


**Apprentissage par renforcement :** Les GMM peuvent être utilisés pour apprendre des politiques dans des problèmes d'apprentissage par renforcement, où un agent doit interagir avec son environnement pour maximiser une récompense.

**Traitement du langage naturel** : Les GMM peuvent être utilisés pour modéliser la distribution des mots dans un corpus de texte, ce qui peut être utile pour des tâches telles que la reconnaissance d'entités nommées et la classification de texte.

**2.Étude comparative**

Les trois techniques, KNN (K plus proches voisins), GMM (Modèles de mélange gaussiens) et PCA (Analyse en composantes principales), sont toutes utilisées pour des tâches de traitement d'images, mais elles le font de manières différentes et ont chacune leurs avantages et leurs inconvénients. Voici un tableau comparatif pour vous aider à choisir la technique la plus adaptée à votre besoin :



Pour Choisir la bonne technique dépend de la tâche spécifique que vous essayez d'accomplir

Voici quelques points supplémentaires à considérer :

* **Type de tâche :** Si vous essayez de classer des images dans différentes catégories, le KNN ou le GMM pourraient être de bons choix. Si vous essayez de réduire la dimensionnalité d'un ensemble de données d'images tout en préservant autant d'informations que possible, la PCA est la technique la plus appropriée.
* **Taille des données :** Le KNN peut devenir lent et inefficace pour les grands ensembles de données car il faut comparer le nouveau point à tous les points de l'ensemble d'apprentissage.
* **Structure des données :** Si vous savez que les données ont une structure linéaire, la PCA peut être un bon choix. Si vous soupçonnez que les données ont une structure non linéaire, le KNN ou le GMM pourraient être de meilleurs choix.

**VI. Etude d'un cas**

**Contexte :**

Selon le Fonds monétaire international (FMI), l'aide au développement est l'aide accordée par les gouvernements et d'autres organismes pour soutenir le développement économique, environnemental, social et politique des pays en développement.

**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.

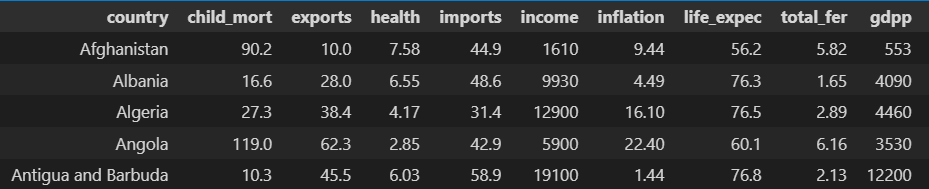
Par conséquent, l'objectif est de catégoriser les pays en utilisant certains facteurs socio-économiques et de santé qui déterminent le développement global du pays. Ensuite, vous devez suggérer les pays sur lesquels le PDG doit se concentrer le plus.

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

**Description des Données et Distribution :**

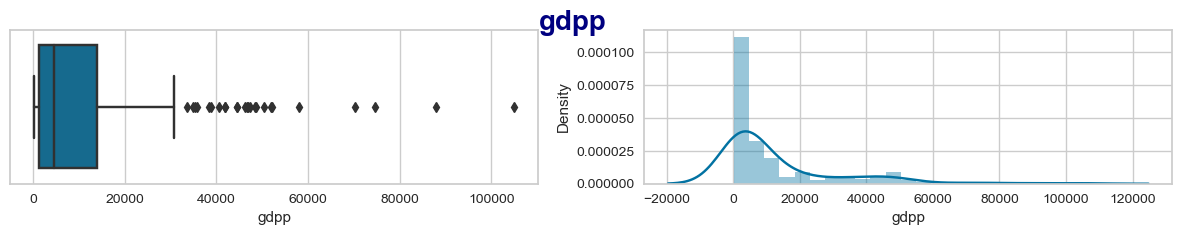
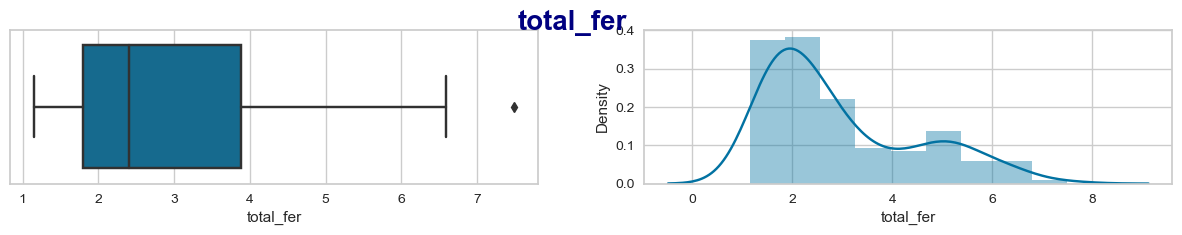
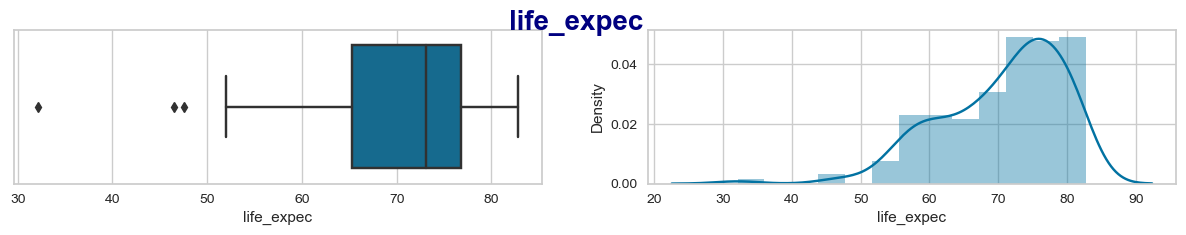
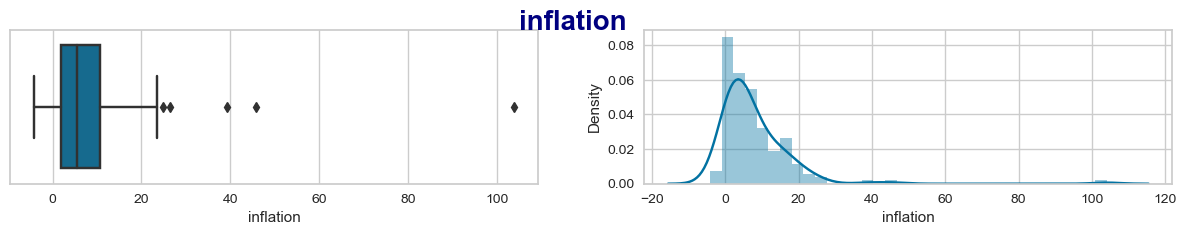
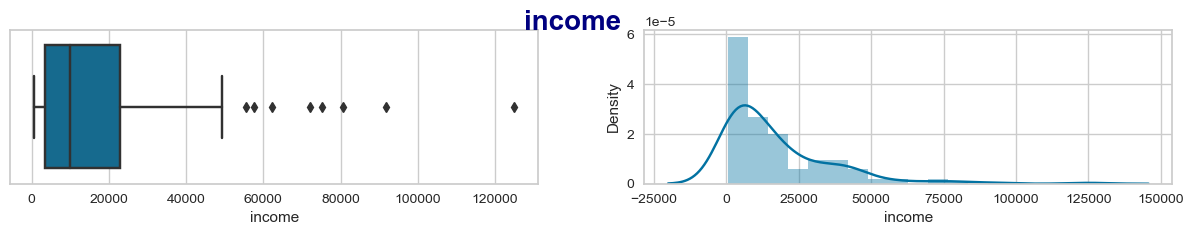
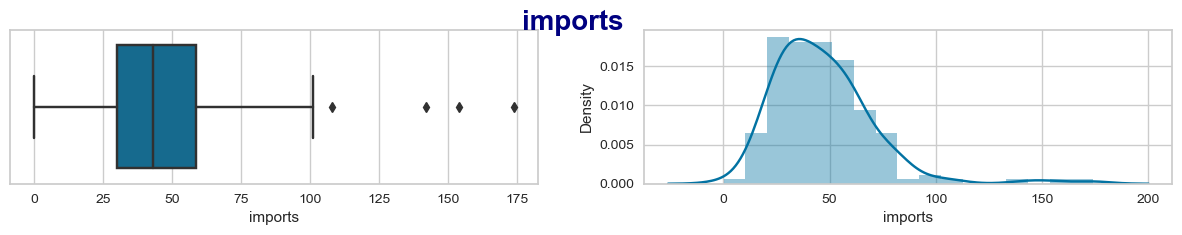
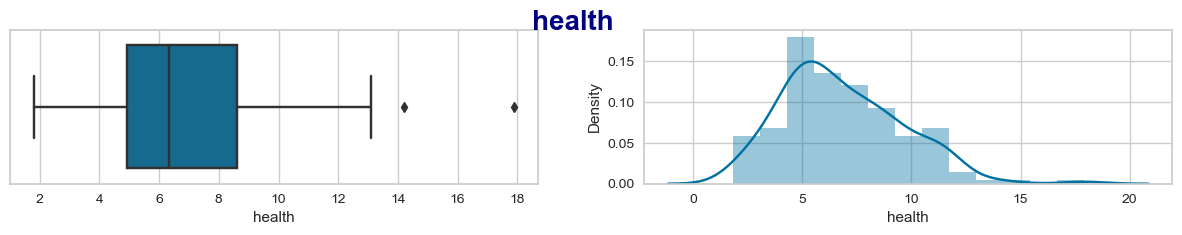
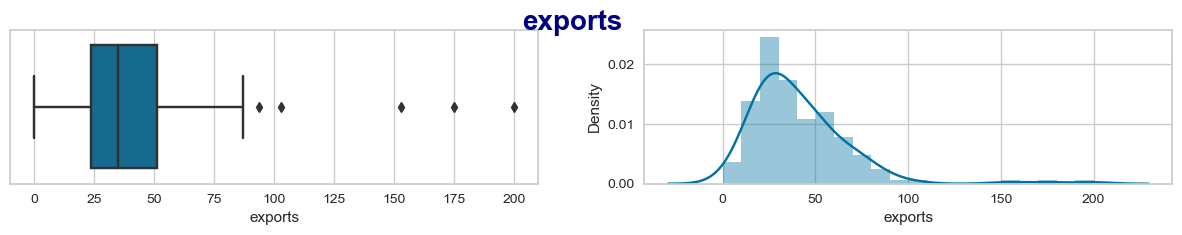
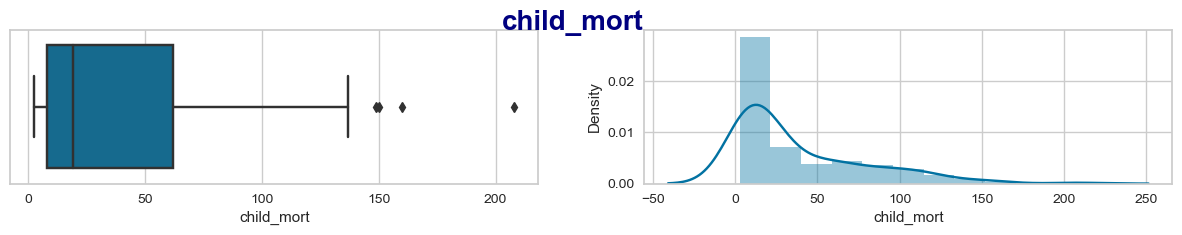
1. **Description des caractéristiques**

* **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



1. **Visualisation :**

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 :**

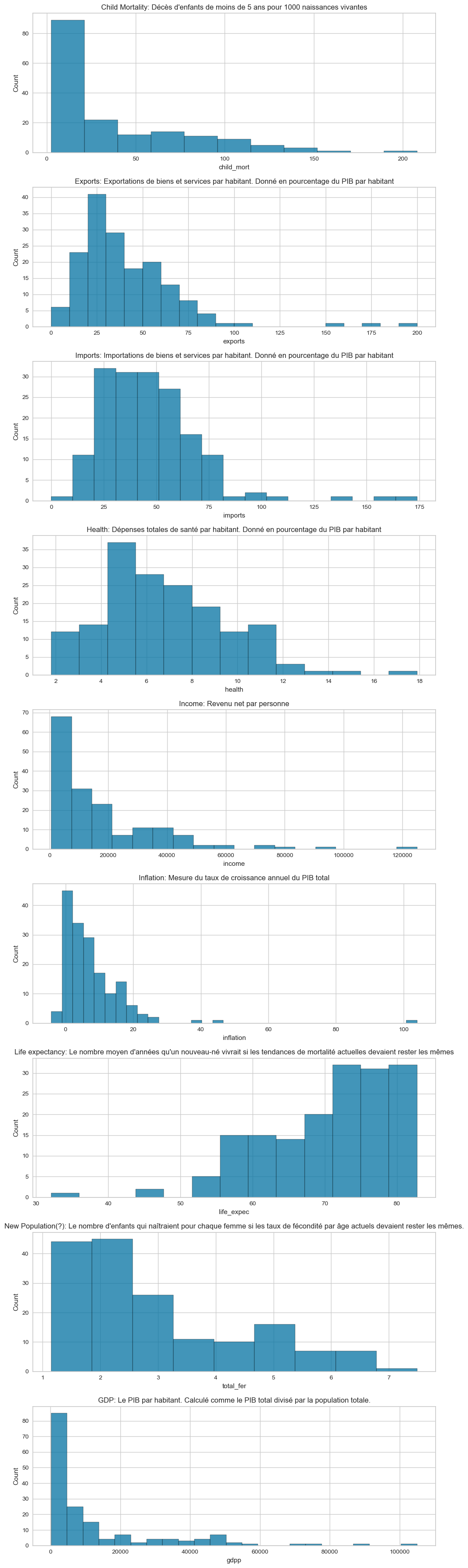
* Jeu de données de petite taille.
* Aucune valeur manquante.
* Aucune valeur en double.
* Quelques valeurs aberrantes et distribution inclinée.

**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 !

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 !

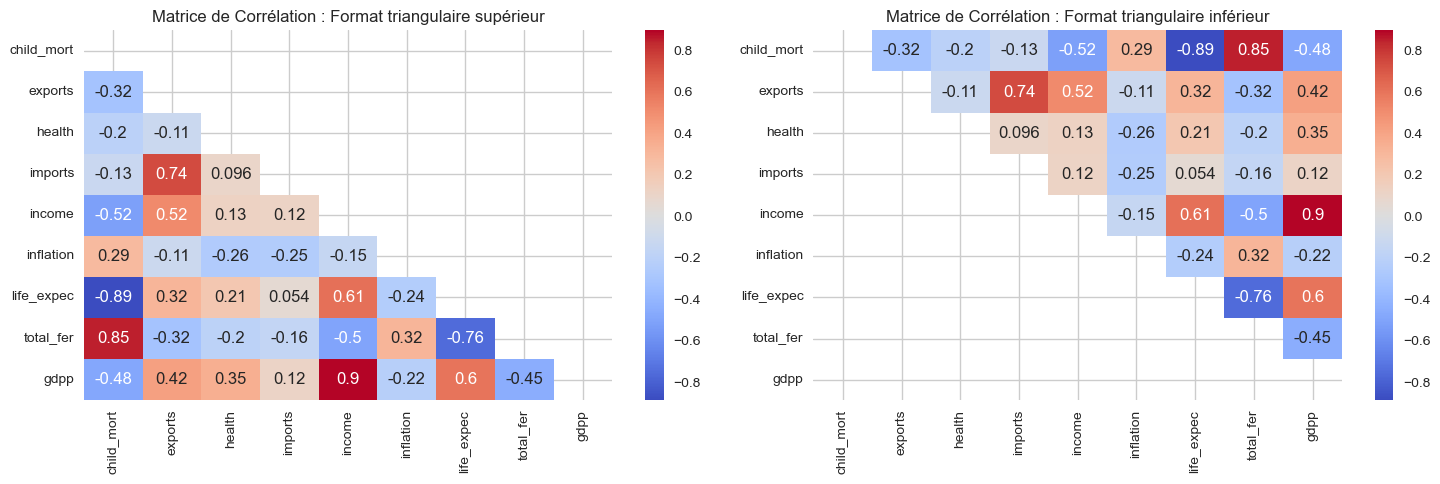
1. **Distribution des Données :**

**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:**

1. **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.

1. **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.

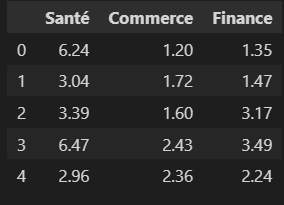
* Les 3 catégories des caractéristiques sont :

**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 !



1. **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

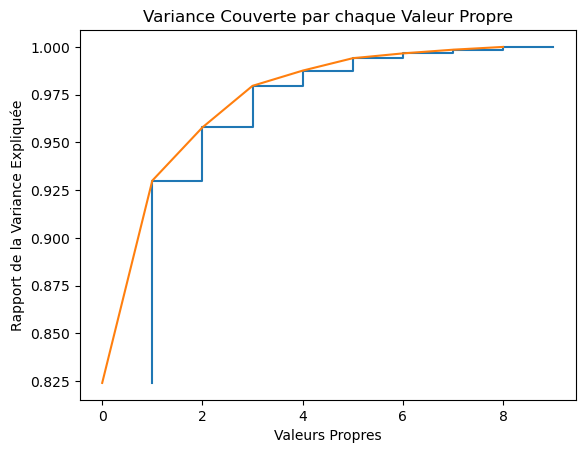
Nous avons 2 options pour la mise à l'échelle des données : 1) Normalisation 2) Standardisation. Comme la plupart des algorithmes supposent que les données sont distribuées normalement (gaussienne), la normalisation est effectuée pour les caractéristiques dont les données ne présentent pas de distribution 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 !

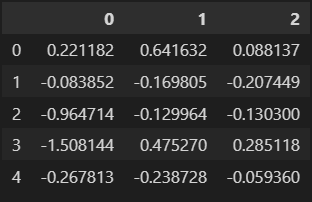
Standardisation : Aucune des caractéristiques n'est standardisée pour les données ci-dessus.

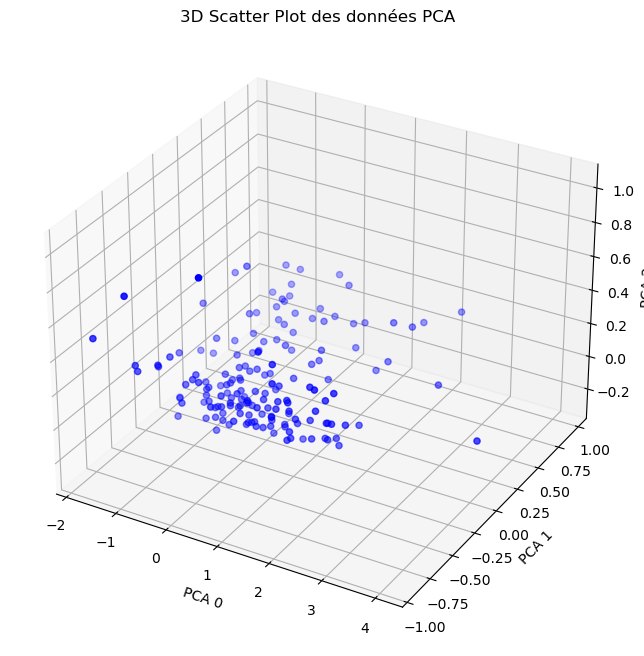
1. **Principal Component Analysis (PCA) :**

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



* 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 :

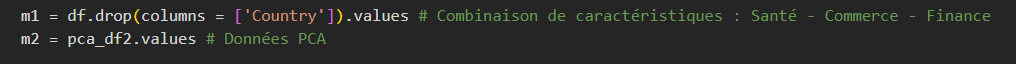
* Combinaison de caractéristiques : Santé - Commerce - Finance
* Données PCA

**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 :

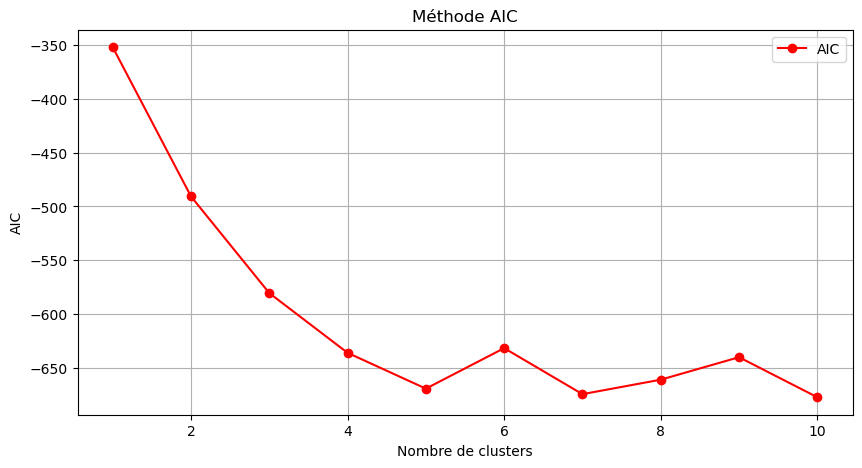
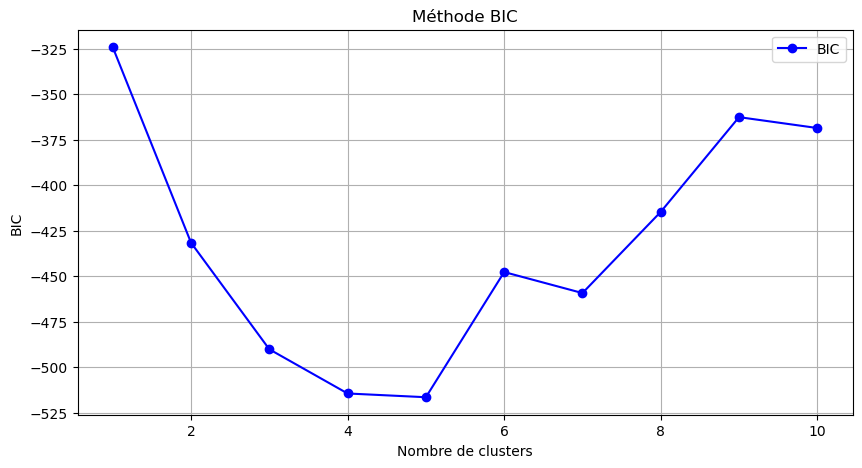
* Sélectionne aléatoirement des paramètres initiaux pour les clusters.
* Attribue chaque point de données au cluster qui maximise la vraisemblance des données observées.
* Met à jour les paramètres de chaque cluster en fonction des points de données qui lui sont attribués.
* Répète ce processus jusqu'à 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.

1. **Combinaison de caractéristiques : Santé - Commerce - Finance**

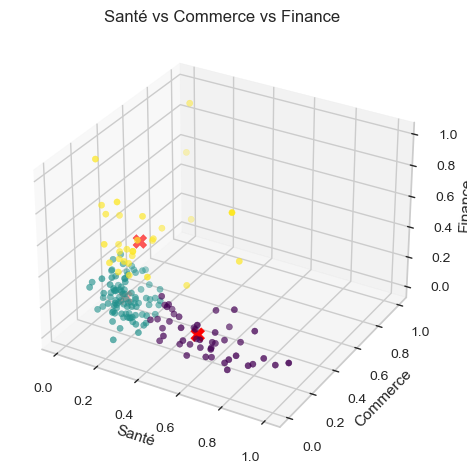
**BIC Methode et AIC Methode :**



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

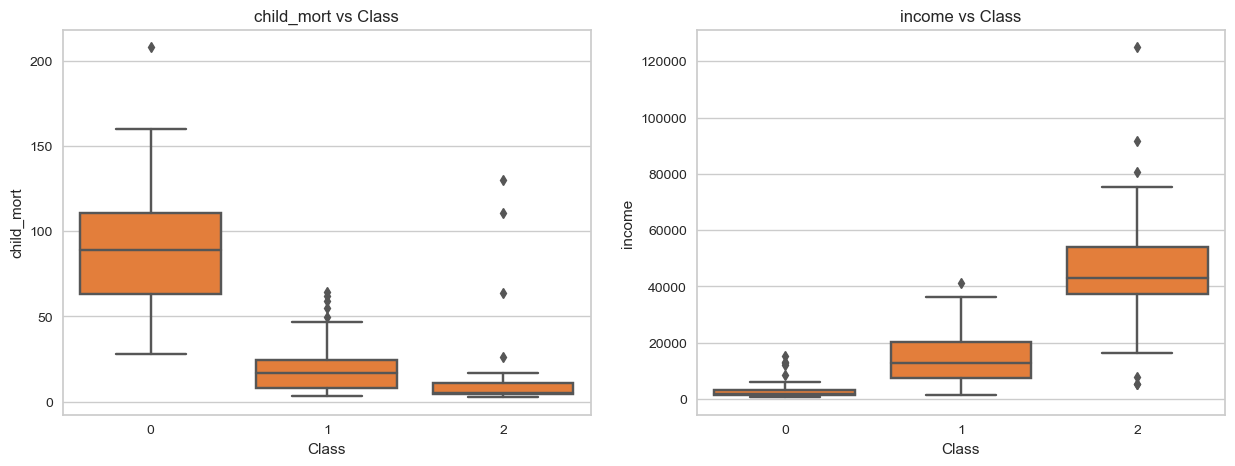
* k : Clusters = 3

**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.

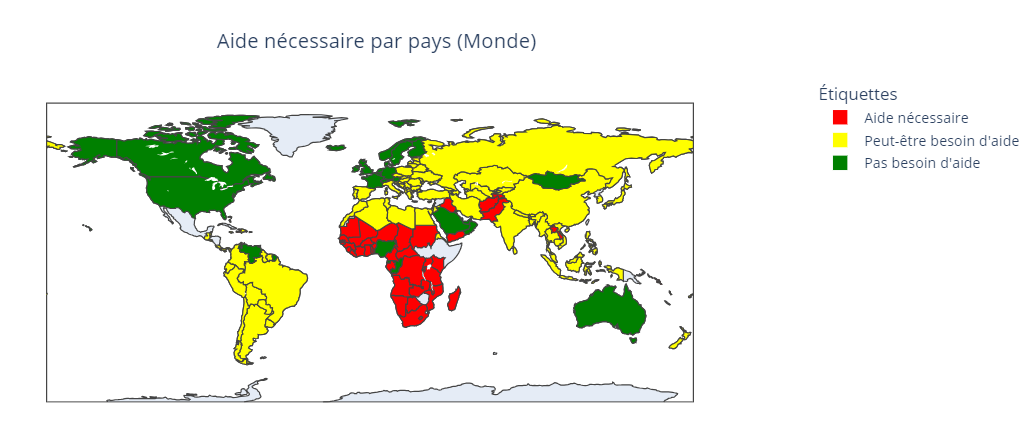


À partir du graphique ci-dessus, nous pouvons conclure :

- 0 : Aide nécessaire

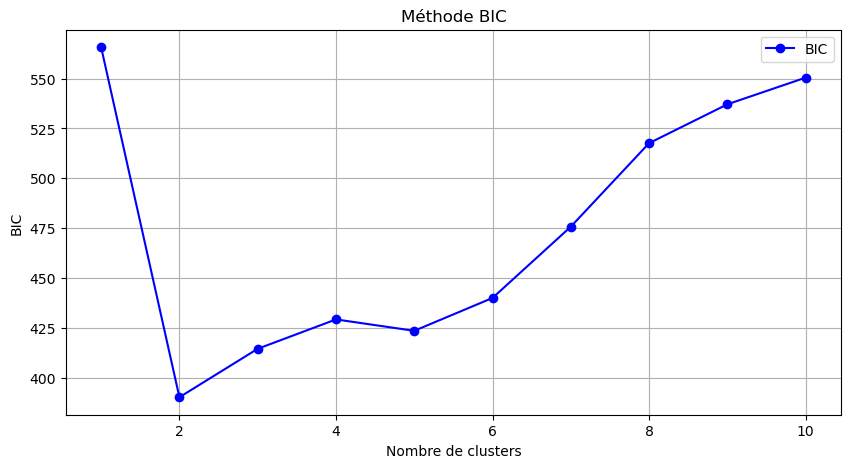
- 1 : Peut-être besoin d'aide

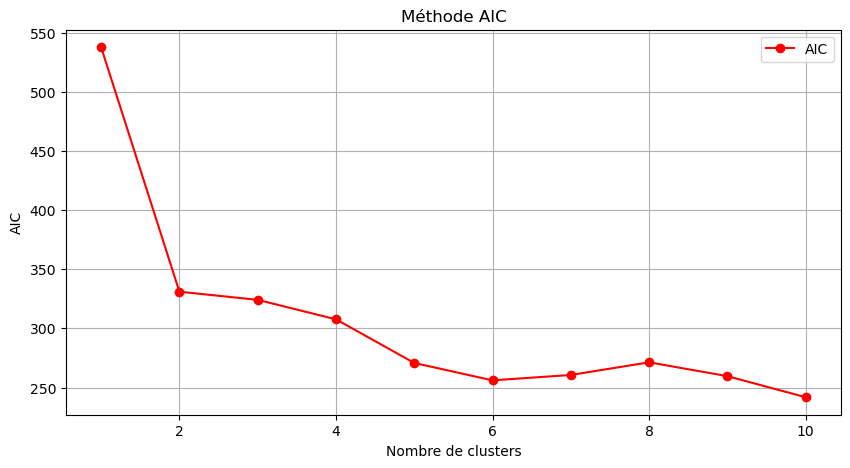
- 2 : Pas besoin d'aide



1. **PCA :**

**BIC Methode et AIC Methode :**

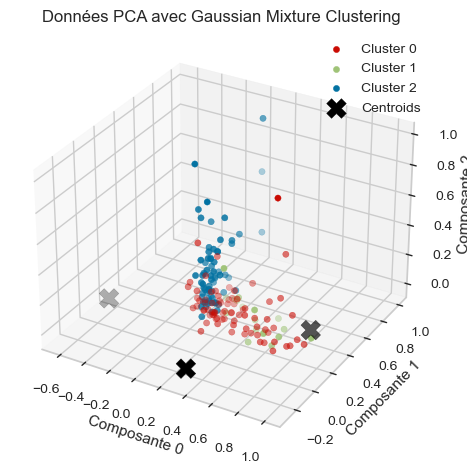




À partir des résultats des deux méthodes ci-dessus, nous sélectionnons :

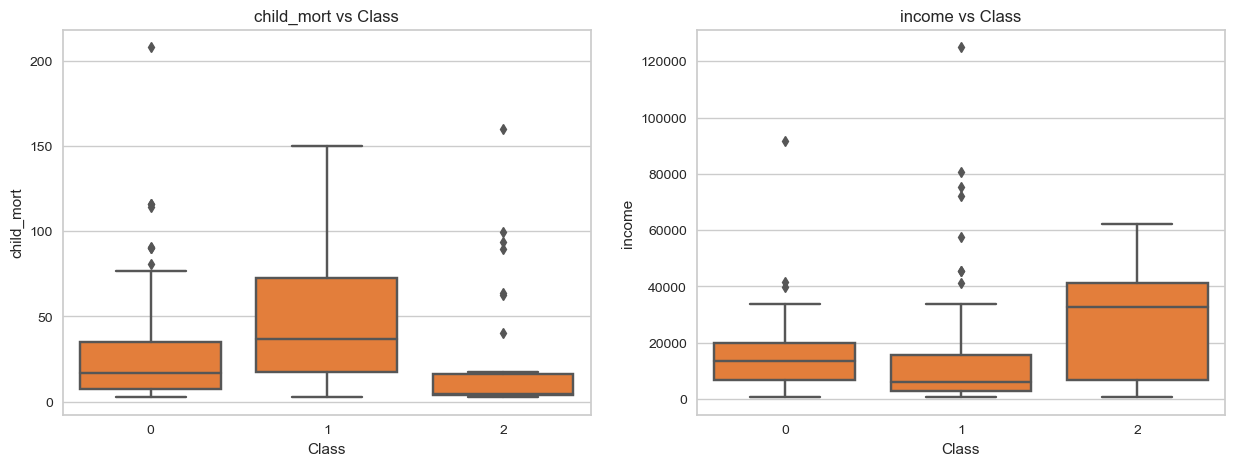
* k : Clusters = 3

**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.

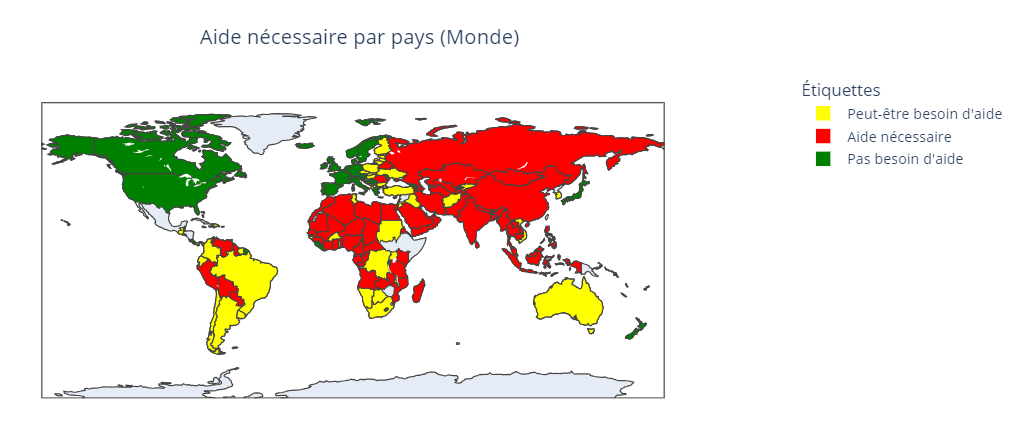


À partir du graphique ci-dessus, nous pouvons conclure :

- 0 : Peut-être besoin d'aide

- 1 : Aide nécessaire

- 2 : Pas besoin d'aide



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



