CLUSTERING
- L3 MATH, ISD, TP 4 (PARTIE 1) -

Ce TP vise à initier aux différentes phases de la classification supervisée : préparation des données, mise en oeuvre des méthodes de classification non supervisée, réalisation de graphiques et de cartes, interprétation.

Partie 1 - Aide internationale

À l'issue de divers programmes de financement, l'ONG "International Humanitarian" a mobilisé une somme d'environ 10 millions de dollars. Le directeur général de l'organisation se trouve maintenant à la croisée des chemins, devant prendre des décisions stratégiques et efficientes quant à l'allocation de ces fonds. Les enjeux cruciaux au moment de cette prise de décision sont principalement liés aux nations présentant un besoin urgent d'assistance. En tant qu'analyste de données, notre tâche consiste à évaluer et classer les pays en fonction de facteurs socio-économiques et sanitaires influant sur le développement global des nations. À la suite de cette analyse, notre recommandation portera sur les pays devant retenir l'attention prioritaire du PDG, recevant ainsi la plus haute priorité pour l'affectation des ressources.

Préparation des données

Les données, disponibles via le lien suivant https://perso.univ-rennes1.fr/valerie.monbet/ISD/Country-data-ONG.csv, contiennent pour un ensemble de 167 pays les variables

- Nom du pays,
- Décès d'enfants de moins de cinq ans pour 1000 naissances vivantes,
- Exportations de biens et services en % du PIB total,
- Importations de biens et services, exprimées en % du PIB total,
- Revenu net par personne,
- La mesure du taux de croissance annuel du PIB total,
- Le nombre moyen d'années que vivrait un nouveau-né si les schémas de mortalité actuels restaient inchangés,
- Le nombre d'enfants nés de chaque femme si les taux actuels de fécondité par âge restent inchangés.

Elles datent de 2021.

Ouvrir un nouveau notebook sous google colab [1].

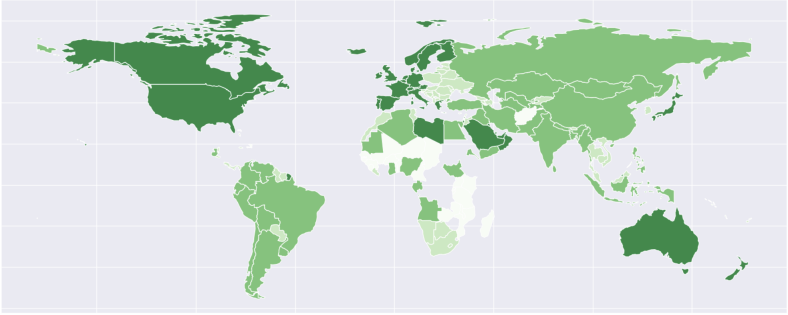
1. Charger les données avec la méthode `read_csv` de pandas. La table sera nommée `pays`.
2. Utiliser la méthode `head` pour voir les premières lignes de la table.
3. Utiliser la méthode `info` pour obtenir une description synthétique de la table : nombre de pays, nombre de variables, type des variables.

1. depuis le drive, cliquer sur le bouton "+ nouveau" et dans la liste choisir "google colaboratory"

4. La table contient-elle des données manquantes ?
5. Normaliser les données pour la suite (ie centrer et réduire avec la fonction StandardScaler de scikit learn.). La table obtenue sera nommée `pays_sc`.

Classification hiérarchique ascendante

6. Aidez vous du code vu en cours pour réaliser la classification hiérarchique ascendante et tracer le dendogramme.
   On utilisera la distance euclidienne et la méthode de Ward.
7. D'après le dendogramme, combien de classes est-il raisonnable de considérer ?
8. Dans cette application, il semble raisonnable de considérer la distance euclidienne et la méthode de ward, mais vous pouvez aussi essayer d'autres distances et d'autres méthodes.
9. Utiliser geopandas pour représenter les classes sur une carte choroplèthe. Vous pouvez vous inspirer du code vu en cours de visualisation.
   Vous ne chercherez pas à améliorer le graphique à ce stade.
   Analyser et critiquer la carte que vous obtenez.



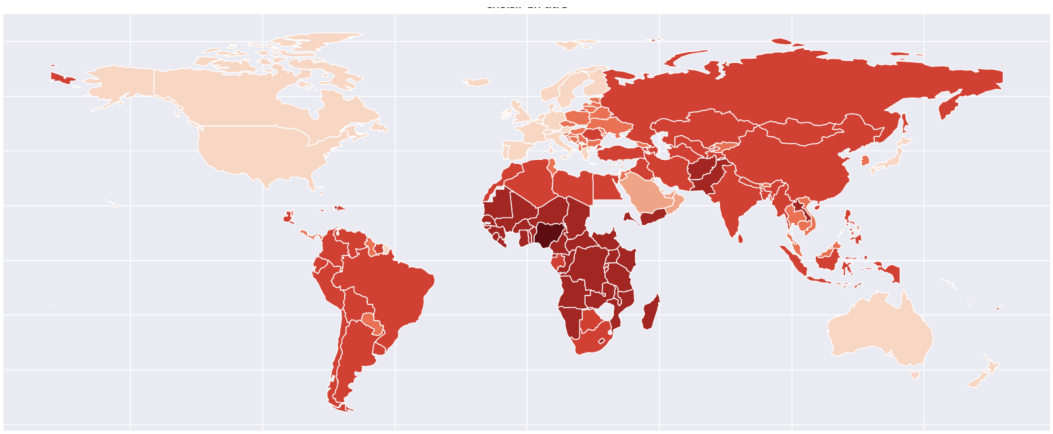
Kmeans

10. Aidez vous du code vu en cours de clustering pour réaliser la classification par l'algorithme des kmeans. Dans un premier temps, on pourra utiliser le nombre de classes déterminé à partir de la CAH.
11. Faire le kmeans pour 5 à 15 classes et tracer l'évolution de la variance intra-classe et du coefficient silhouette. Si la variance intraclasse est normalisée les deux indices peuvent être tracés sur le même graphique.
    Choisir un nombre de classes à partir de ce graphique. On gardera ce nombre de classes pour la suite.
12. Réaliser l'analyse en composante principales des données et interpréter les deux premiers axes factoriels.
13. Tracer un boxplot de la première composante principale en fonction des classes. Les classes sont-elles ordonnées de façon logiques ?
14. Utiliser les projections des centres de classe sur le premier axes factoriel de l'ACP pour ordonner les classes.
    Créer une nouvelle variable de classe nommée `niveau d'urgence`, ayant le même

nombre de modalités que la variable classe obtenue par kmeans. La variable variable de classe nommée `niveau d'urgence` sera ordonnée selon la coordonnée de son centre sur le premier axe factoriel.
Aidez-vous des commandes suivantes pour ordonner les classes.

```
cluster_pca = pca.transform(kmeans.cluster_centers_) # projection
order = cluster_pca[:,0].argsort()
ranks = order.argsort()
ranks
```

15. Représenter les classes sur le 1er plan factoriel de l'ACP. On utilisera une couleur par classe et on ajoutera les noms de pays à chaque point.
16. Tracer un boxplot du nombre de décès infantile en fonction du niveau d'urgence. Commenter le graphique obtenu.
17. Représenter le niveau d'urgence sur une carte du monde. Utiliser la carte et vos analyses pour faire des recommandations au président de l'ONG.



DBSCAN

18. Mettre enoeuvre la méthode DBSCAN avec la distance euclidienne (voir le code du cours de clustering)
19. Comparer les 3 classifications obenues. Pour réaliser la comparison, ous ordonnerez les classe selon le niveau d'urgence défini à partir du premier axes factoriel de l'ACP.