# Projet de Machine Learning

Soutenances orales: 28 et 31 Mai 2024

## Jeu de données

Les données sont issues du site du concours KAGGLE; il s'agit du jeu de données " Global Data on Sustainable Energy" (2000-2020) disponible ici:
https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy.
Le jeu de données comprend 3649 observations et 21 variables, qui représentent diverses caractéristiques liées à la consommation énergétique et à la géographie de 176 pays du monde au cours des années 2000 à 2020.

Les variables sont les suivantes:

- **Entity**: Nom du pays ou de la région pour lequel les données sont rapportées.
- **Year**: Année pour laquelle les données sont communiquées, entre 2000 et 2020.
- **Access to electricity (% of population)**: Pourcentage de la population ayant accès à l'électricité.
- **Access to clean fuels for cooking (% of population)**: Pourcentage de la population qui utilise principalement des combustibles propres.
- **Renewable-electricity-generating-capacity-per-capita**: Capacité installée d'énergie renouvelable par personne.
- **Financial flows to developing countries (US Dollars)**: Aide et assistance des pays développés pour les projets d'énergie propre.
- **Renewable energy share in total final energy consumption (%)**: Pourcentage d'énergie renouvelable dans la consommation d'énergie finale.
- **Electricity from fossil fuels (TWh)**: Électricité produite à partir de combustibles fossiles (charbon, pétrole, gaz) en térawattheures.
- **Electricity from nuclear (TWh)**: Électricité produite à partir de l'énergie nucléaire en térawattheures.
- **Electricity from renewables (TWh)**: Électricité produite à partir de sources renouvelables (hydroélectricité, énergie solaire, énergie éolienne, etc.) en térawattheures.
- **Low-carbon electricity (% electricity)**: Pourcentage d'électricité provenant de sources à faible teneur en carbone (nucléaire et énergies renouvelables).
- **Primary energy consumption per capita (kWh/person)**: Consommation d'énergie par personne en kilowattheures.
- **Energy intensity level of primary energy (MJ/2011 PPP GDP)**: Consommation d'énergie par unité de PIB à parité de pouvoir d'achat.
- **Value-co2-emissions (metric tons per capita)**: Émissions de dioxyde de carbone par personne en tonnes métriques.
- **Renewables (% equivalent primary energy)**: Équivalent énergie primaire provenant de sources renouvelables.
- **GDP growth (annual %)**: Taux de croissance annuel du PIB en monnaie locale constante.
- **GDP per capita**: Produit intérieur brut (PIB) par personne.
- **Density (P/Km2)**: Densité de population en personnes par kilomètre carré.
- **Land Area (Km2)**: Surface terrestre totale en kilomètres carrés.
- **Latitude**: Latitude du centroïde du pays en degrés décimaux.
- **Longitude**: Longitude du centroïde du pays en degrés décimaux.

L'objectif est de prédire la variable **Value-co2-emissions** à partir des autres variables.
Attention: Le jeu de données comporte beaucoup de valeurs manquantes, une étude exploratoire préalable est plus que jamais nécessaire pour se familiariser avec les données et les préparer à la phase de modélisation.

# Questions posées

## Analyse exploratoire des données

L'objectif dans un premier temps est d'explorer les différentes variables, étape préliminaire indispensable à l'analyse. Ci-dessous sont précisées quelques questions basiques. Vous pouvez compléter l'analyse selon vos propres idées.

1. Commencez par vérifier la nature des différentes variables et leur encodage. Convertissez la variable **Year** en une variable qualitative. N.B. Curieusement, la variable **Density (P/Km2)** n'est pas considérée comme une variable numérique. Convertissez-là en une variable numérique en prenant soin de ne pas transformer les nombre décimaux en `NA`. Par exemple, en R, vous pourrez utiliser la formule: `as.numeric(gsub(",",".",data$Density.n.P.Km2.))`, où `data` représente le jeu de données utilisé.

2. Déterminez le taux de valeurs manquantes pour chaque variable.
   On propose de supprimer pour ce projet les variables comportant un taux de données manquantes très important: **Renewable-electricity-generating-capacity-per-capita**, **Financial flows to developing countries (US Dollars)** et **Renewables (% equivalent primary energy)**.

3. Pour la suite de l'étude, vous allez créer un jeu de données comportant seulement les individus qui n'ont pas de valeur manquante. Il reste alors 2868 observations.

4. Commencez l'exploration par une analyse descriptive unidimensionnelle des données. Des transformations des variables quantitatives vous semblent-t-elles pertinentes ?

5. Visualisez la grande hétérogénéité des émissions de $CO_2$ entre les pays. Quels sont les 5 pays les plus émetteurs de $CO_2$ ?

6. Poursuivez avec une analyse descriptive multidimensionnelle. Utilisez des techniques de visualisation: par exemple les nuages de points (*scatterplot*), des graphes des correlations... Analysez les dépendances entre les variables quantitatives.

7. Réalisez une analyse en composantes principales des variables quantitatives et interprétez les résultats.

8. Visualisez la possible dépendance entre la variable **Year** et la variable à prédire.

### Modélisation

Nous considérons maintenant le problème de la prédiction la variable **Value-co2-emissions** à partir des autres variables du point de vue de l'apprentissage automatique, c'est-à-dire en nous concentrant sur les performances du modèle. L'objectif est de déterminer les meilleures performances que nous pouvons attendre, et les modèles qui les atteignent. Voici quelques questions pour vous guider.

1. Divisez le jeu de données sans données manquantes en un échantillon d'apprentissage et un échantillon test. Vous prendrez un pourcentage de 20% pour l'échantillon test. Pourquoi cette étape est-elle nécessaire lorsque nous nous concentrons sur les performances des algorithmes ?

2. Comparez les performances d'un modèle de régression linéaire avec/sans sélection de variables, avec/sans pénalisation, d'un SVM, d'un arbre optimal, d'une forêt aléatoire, du boosting, et de réseaux de neurones. Justifiez vos choix (par exemple le noyau pour le SVM), et ajustez soigneusement les paramètres (par validation croisée). Interprétez les résultats et quantifiez l'amélioration éventuelle apportée par les modèles non linéaires.

3. Comparez les différents modèles optimisés sur votre échantillon test. Quels sont les modèles les plus performants ? Quel est le niveau de précision obtenu ?

4. Interprétation et retour sur l'analyse des données: vos résultats sont-ils cohérents avec l'analyse exploratoire des données, par exemple en ce qui concerne l'importance des variables ?
5. Dans un second temps, vous pourrez utiliser un algorithme de complétion des valeurs manquantes et reprendre la modélisation (pour les algorithmes qui se sont montrés les plus performants) avec le jeu de données complété.

# Modalités et évaluation

Vous réaliserez le projet par groupe de 4 étudiant.e.s. L'évaluation portera sur une soutenance orale et deux notebooks Jupyter (un en R et un en Python).

**Travail à rendre:** Comme livrable, chaque groupe déposera sur Moodle :

- **au plus tard le 27 Mai à 18H**, un fichier zip contenant les deux notebooks Jupyter (R et Python),
- **au plus tard la veille de la soutenance à 18 H**, les slides de l'exposé **au format pdf**.

**Soutenances orales les 28 et 31 Mai 2024:** 20 minutes de présentation, puis 5 à 10 minutes de questions. L'exposé doit comprendre une introduction présentant les données ainsi que toutes les transformations que vous avez effectuées, une description succincte des algorithmes utilisés (en précisant bien quels hyperparamètres vous avez optimisés et comment), une interprétation des résultats, et une conclusion. Les questions pourront porter sur votre code (donc pensez à ouvrir vos notebooks et si possible les compiler juste avant la soutenance).

**Critères d'évaluation:** L'évaluation tiendra compte de la qualité de présentation orale (clarté, argumentation, interprétation des résultats etc.), de la cohérence de l'étude, de la qualité de présentation des notebooks (n'oubliez pas de commenter votre code), des interprétations des résultats (graphiques et autres).