Projet de Machine Learning

Deadline: 16 Février 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** : The name of the country or region for which the data is reported.
- **Year** : The year for which the data is reported, ranging from 2000 to 2020.
- **Access to electricity (% of population)** : The percentage of population with access to electricity.
- **Access to clean fuels for cooking (% of population)** : The percentage of the population with primary reliance on clean fuels.
- **Renewable-electricity-generating-capacity-per-capita** : Installed Renewable energy capacity per person
- **Financial flows to developing countries (US Dollars)** : Aid and assistance from developed countries for clean energy projects.
- **Renewable energy share in total final energy consumption (%)** : Percentage of renewable energy in final energy consumption.
- **Electricity from fossil fuels (TWh)** : Electricity generated from fossil fuels (coal, oil, gas) in terawatt-hours.
- **Electricity from nuclear (TWh)** : Electricity generated from nuclear power in terawatt-hours.
- **Electricity from renewables (TWh)** : Electricity generated from renewable sources (hydro, solar, wind, etc.) in terawatt-hours.
- **Low-carbon electricity (% electricity)** : Percentage of electricity from low-carbon sources (nuclear and renewables).
- **Primary energy consumption per capita (kWh/person)** : Energy consumption per person in kilowatt-hours.
- **Energy intensity level of primary energy (MJ/2011 PPP GDP)** : Energy use per unit of GDP at purchasing power parity.
- **Value-co2-emissions (metric tons per capita)** : Carbon dioxide emissions per person in metric tons.
- **Renewables (% equivalent primary energy)** : Equivalent primary energy that is derived from renewable sources.
- **GDP growth (annual %)** : Annual GDP growth rate based on constant local currency.
- **GDP per capita** : Gross domestic product per person.
- **Density (P/Km2)** : Population density in persons per square kilometer.
- **Land Area (Km2)** : Total land area in square kilometers.
- **Latitude** : Latitude of the country's centroid in decimal degrees.
- **Longitude** : Longitude of the country's centroid in decimal degrees.

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 une analyse descriptive unidimensionnelle des données.
2. Des transformations des variables quantitatives vous semblent-t-elles pertinentes ?
   N.B. Curieusement, la variable **Density (P/Km2)** n'est pas considérée comme une variable numérique, convertissez-là en une variable numérique. Convertissez la variable **Year** en une variable qualitative.
3. Visualisez la grande hétérogénéité des émissions de CO_2 entre les pays. Quels sont les 5 pays les plus émetteurs de CO_2 ?
4. Déterminer 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)**.
5. Poursuivez avec une analyse descriptive multidimensionnelle. Utilisez des techniques de visualisation : par exemple scatterplot, correlation plot ... Analysez les dépendances entre les variables quantitatives.
6. Réalisez une analyse en composantes principales des variables quantitatives et interprétez les résultats.
7. 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. Dans un premier temps, vous allez créer un jeu de données comportant seulement les observations qui n'ont pas de valeur manquante. Il reste 2768 observations.
2. Divisez ce jeu de données 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 ?
3. 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.
4. Comparez les différents modèles optimisés sur votre échantillon test. Quels sont les modèles les plus performants ? Quel est le niveau de précision obtenu ?
5. Interprétation et retour sur l'analyse des données : Vos résultats sont-ils cohérents avec l'analyse préliminaire des données, par exemple en ce qui concerne l'importance des variables ?
6. 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.

Organisation et rapport à rendre

Vous réaliserez le projet par groupe de 3 ou 4 étudiant.e.s. **Deadline: 16 février 2024**. Comme livrable, vous rendrez un rapport au format pdf ne dépassant pas 30 pages. Il doit comprendre une introduction, une description succincte des algorithmes utilisés, une interprétation des résultats, une conclusion, etc. De plus, vous rendrez deux notebooks Jupyter, l'un en R, l'autre en Python. N'oubliez pas de commenter votre code. Le dépôt se fera sur Moodle : chaque groupe téléchargera un fichier zip contenant le rapport (format pdf) et les notebooks.

L'évaluation tiendra compte de la présentation du rapport et de la rédaction (clarté, argumentation, etc.), de la cohérence de l'étude, de la qualité de présentation des notebooks, des interprétations des résultats (graphiques et autres).