Questions :

- Chaque donnée sera automatiquement passée au log pour améliorer les résultats via la fonction np.log1p ?

Le but ici est de tendre mes distributions sur une gaussienne. Idéalement tu l’ajoutes en tant qu’hyperparamètres après tes transformations numériques.

- Comment interpréter les features importance ?

Ca serait bon que SHAP et l’importance soit un peu similaire.

- Comment interpréter l’overfitting de certains modèles ? C’est ok vu la perf sur le test ? Vu que les perfs de test c’est classe ça va. Idéalement il faudrait plus de données.

##### Sélectionner, transformer et créer les variables pertinentes d'un modèle d'apprentissage supervisé ou non supervisé (feature engineering) afin de réaliser un apprentissage optimal.

CE1 Vous avez identifié les variables catégorielles (qualitatives).

CE2 Les variables catégorielles identifiées ont été transformées en fonction du besoin (par exemple via OneHotEncoder ou TargetEncoder).

CE3 Vous avez créé de nouvelles variables à partir de variables existantes.

CE4 Vous avez réalisé des transformations mathématiques lorsque c'est requis pour transformer les distributions de variables.

CE5 Vous avez normalisé les variables lorsque c'est requis.

##### Sélectionner, entraîner et évaluer des modèles d’apprentissage supervisé, afin de choisir le modèle le plus performant pour la problématique métier.

CE1 Vous avez choisi la ou les variables cibles pertinentes. Cela signifie dans le cadre de ce projet :

* Vous avez choisi la variable cible pour la consommation totale d’énergie,
* Vous avez choisi la variable cible pour la consommation de CO2.

CE2 Vous avez vérifié qu'il n’y a pas de problème de data leakage (c'est-à-dire, des variables trop corrélées à la variable cible et inconnues a priori dans les données en entrée du modèle).

CE3 Vous avez choisi des métriques adaptées pour évaluer la performance d'un algorithme (par exemple : R2, RMSE, MAE en régression.)

CE4 Vous avez exploré d'autres indicateurs de performance que le score pour comprendre les résultats (par exemple : coefficients des variables en fonction de la pénalisation, visualisation des erreurs en fonction des variables du modèle, temps de calcul...).

CE5 Vous avez séparé les données en train/test pour les évaluer de façon pertinente et détecter l'overfitting.

CE6 Vous avez mis en place une validation croisée (via GridsearchCV, RandomizedSearchCV ou équivalent) afin de réaliser une Cross Validation sur le set de train et d’optimiser les hyperparamètres.

CE7 Vous avez testé au minimum 4 algorithmes de famille différente (par exemple : ElasticNet, SVM, GradientBoosting, RandomForest). Vous avez présenté une synthèse comparative de l'ensemble des résultats. Vous avez justifié le choix final de l'algorithme et des hyperparamètres.

CE8 Vous avez réalisé l’analyse de l’importance des variables (feature importance) globale sur l’ensemble du jeu de données et locale sur chaque individu du jeu de données.