Questions :

* Is Pytorch good enough for the formation and actual work or should I check TensorFlow and/or Keras as well ?
* Project 3 :
  + Can I use multiple years of data as I only have a few thousands lines ?
  + Can I simply clean the lines with na in important columns and use multiple years as dataset or should I do some inputation for this project ?
  + Should I extract the number from the address ?
  + Only keep ComplianceStatus == ‘Compliant’ ?
  + I need to filter !BuildingType.contains(Multifamily) right ?
  + What is column DefaultData ?

1. Réaliser une courte analyse exploratoire.

- De l’univariante sur les colonnes considérées + une ACP suffirait ? Je peux aussi faire une bivariante avec un boxplot, fais gaffe aux colonnes String, mapping first ? Addresse ?

2. Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

- Quels types de modèles ?

Lis bien le bouquin du debout à environ p.100.

Pour 60/20/20, les 20 derniers sont pour une évaluation à la fin qui sera la réelle.

Donc n’utilise pas une année prochaine pour l’évaluation mais juste prends un gros dataset. Seulement 2016 en fait.

Un modèle comprend la prépa des données aussi, pas seulement l’algorithme. Tout ça se fait avec des pipelines. Voici comment faire un modèle :

1. D’abord je commence avec seulement des colonnes numérique et j’utilise le RandomForest, et finit par une cross-validation. Ca va être ta première version.

2. Tu vas ensuite rajouter des features comme la proportion de l’énergie. Quelques versions en plus.

3a. Tu vas ajouter les String mais avec chaque valeur est un flag bool, 1 ou 0.

3b. Tu vas ensuite ajouter les String convertis en chiffre avec un mapping. Tu as du Onehot encoding, Target encoding, je crois que ces 2 s’appellent.

4. Normaliser les données ou les standardiser.

5a. Linear regression, random forest, …

5b. Pour chaque algo tu as ses paramètres à optimiser, ils s’appellent les hyperparamètres. Pour trouver les valeurs optimisées : Grid-SearchCV.