OpenClassRooms

Parcours Ingénieur Machine Learning

P3 – Seattle

Aoufi Nizar

Déroulement

- Présentation de la mission
- Nettoyage & Analyse exploratoire
- Modélisation
- Exploitation des résultats

Présentation du projet

Contexte

- City of Seattle
- Ville neutre en émissions de carbone

Nettoyage

Gestion des doublons

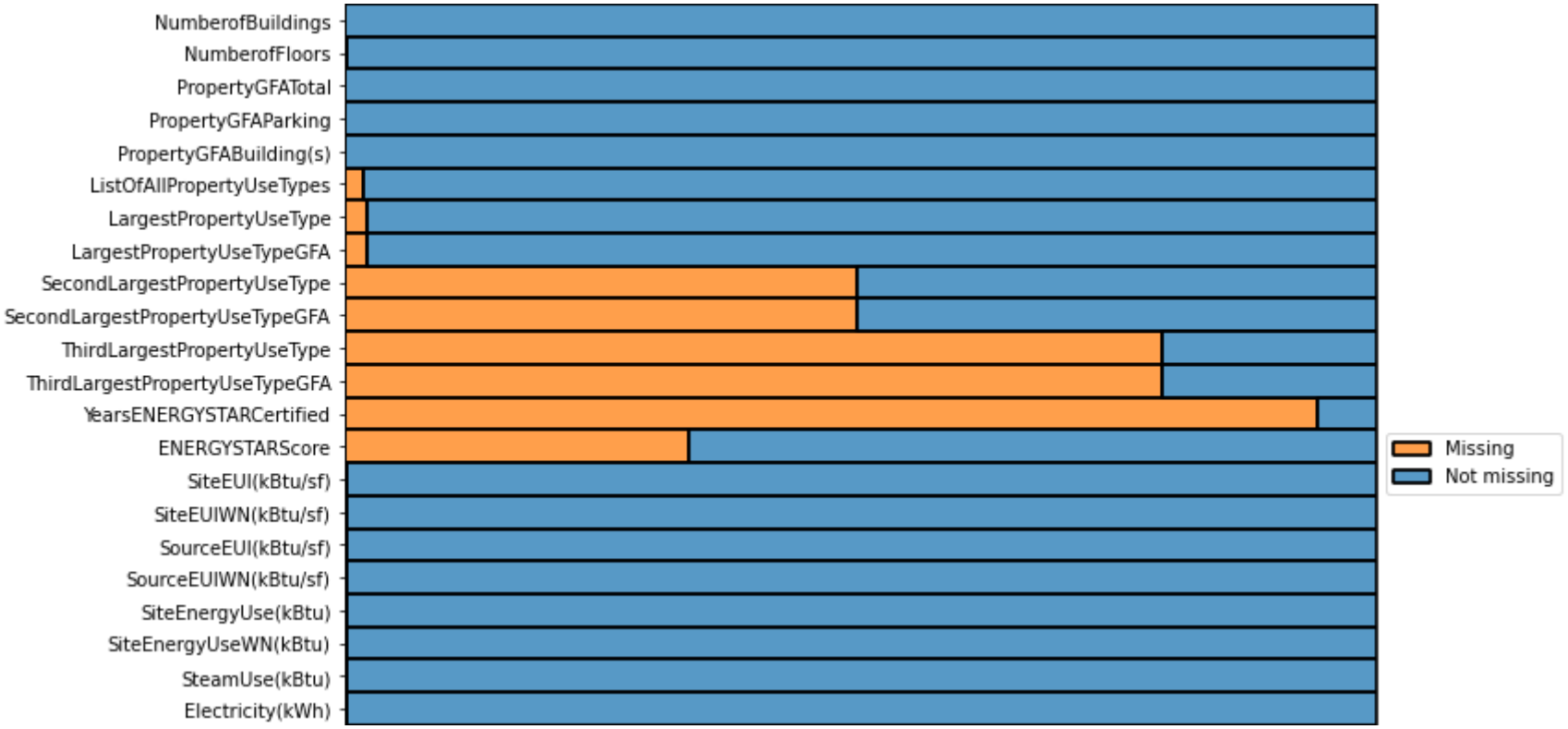
- Deux années différentes, colonnes doubles
- Décompactage de la localisation
- Fusion, conservation données récentes
- Harmonisation du format des données

Suppression bâtiments résidentiels

Nettoyage

Gestion des valeurs manquantes

- Seuil à 35 %



Nettoyage

Dataset composé de 1595 lignes et 11 colonnes :

- 7 attributs numériques

```
['BuildingAge', 'NumberofBuildings', 'PropertyGFATotal', 'PropertyGFABuilding(s)', 'PropertyGFAParking', 'LargestPropertyUseTypeGFA', 'ENERGYSTARScore']
```

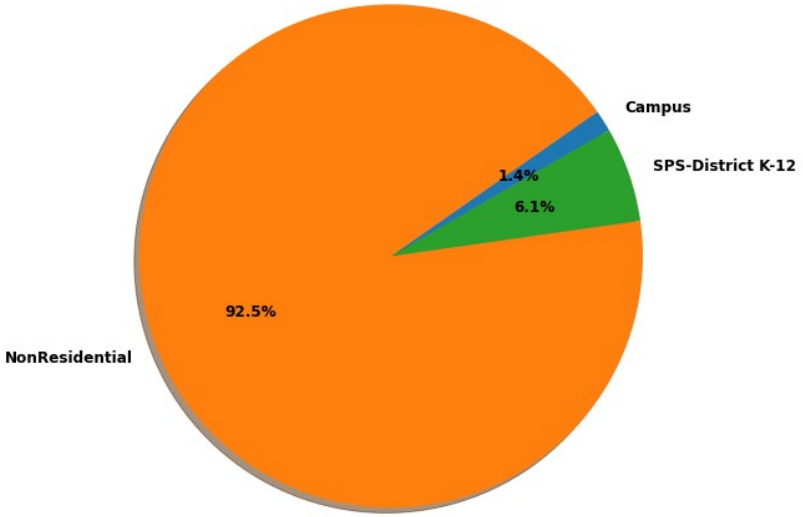
- 4 attributs catégoriques

```
['BuildingType', 'LargestPropertyUseType', 'Neighborhood', 'PrimaryPropertyType']
```

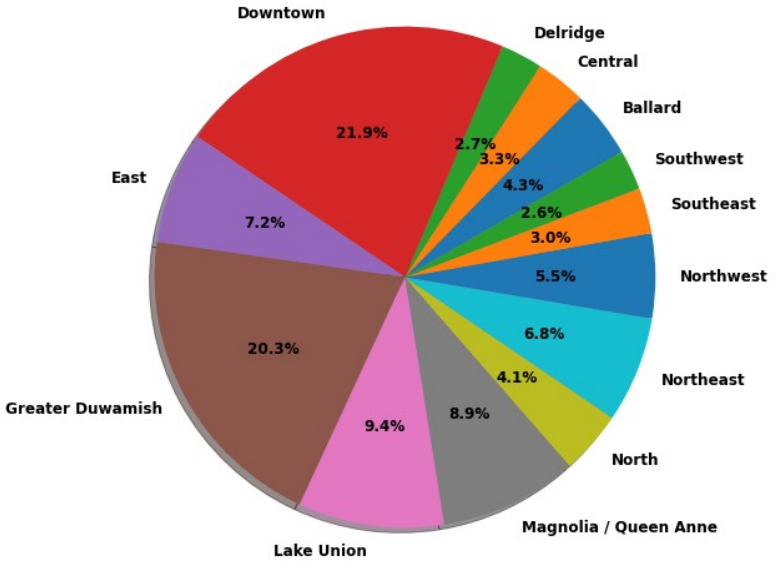
Après transformation one-hot-encoding 102 colonnes.

Analyse exploratoire

Types de bâtiments

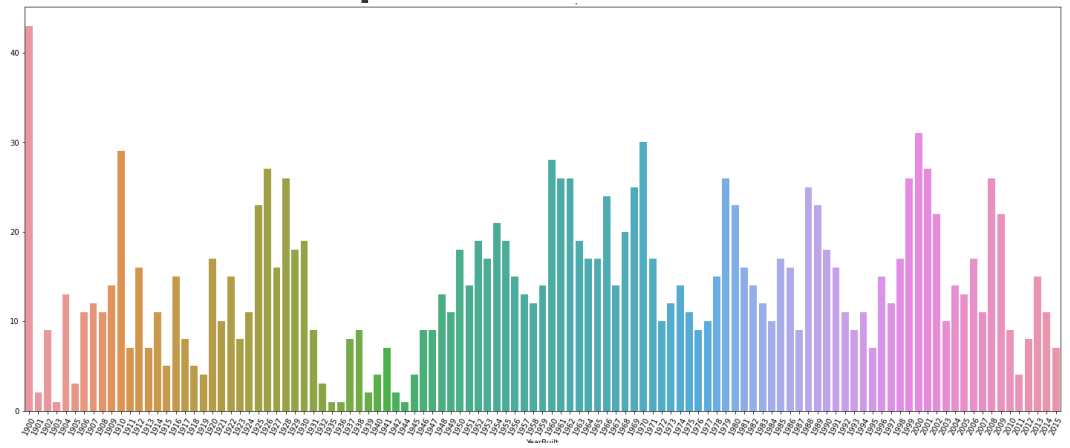


Répartition par quartier

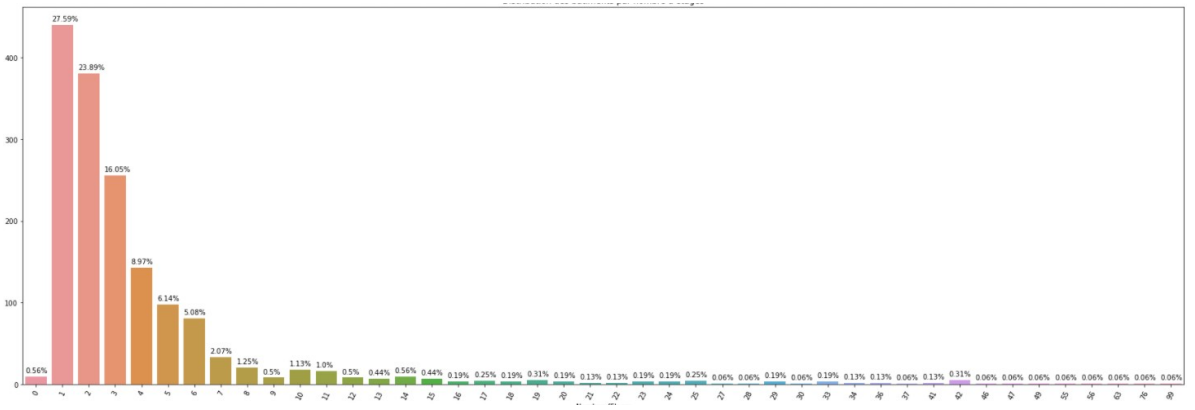


Analyse exploratoire

Distribution par année de construction

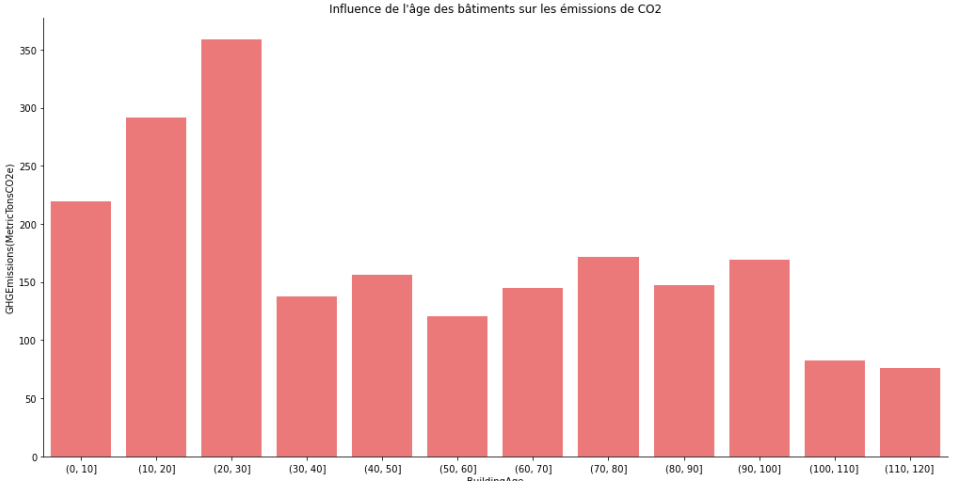


Distribution du nombre d’étages

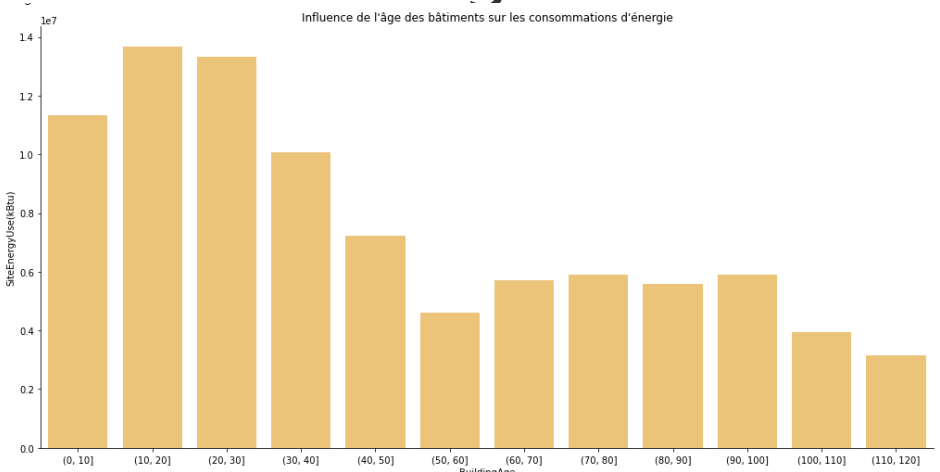


Analyse exploratoire

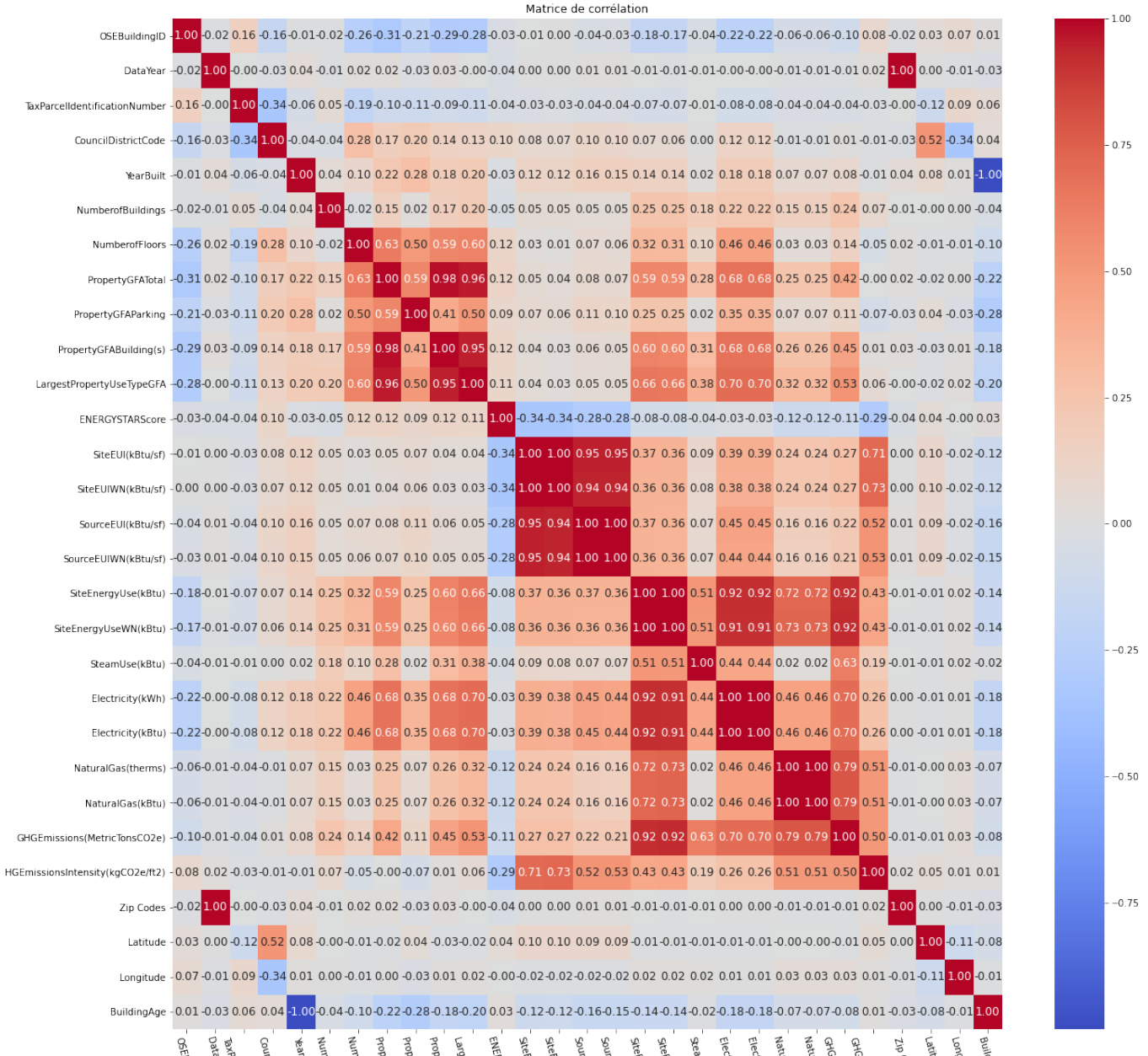
Émission de CO^2 en fonction de l'âge



Consommation d'énergie en fonction de l'âge

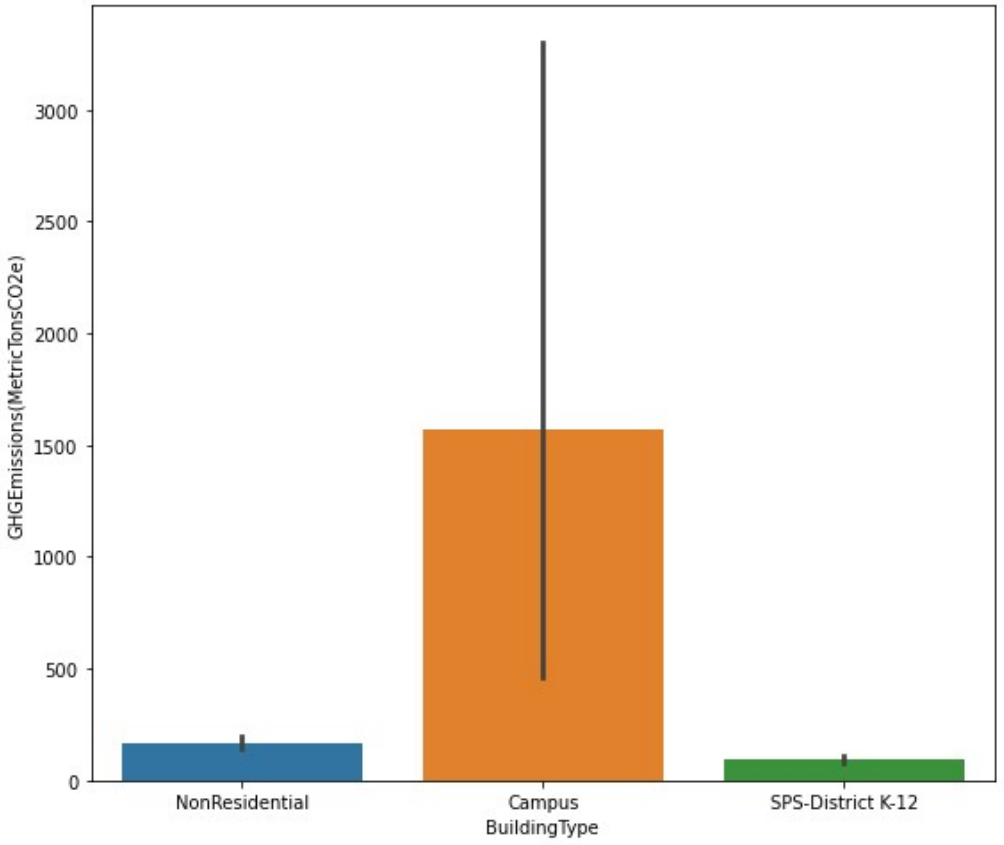


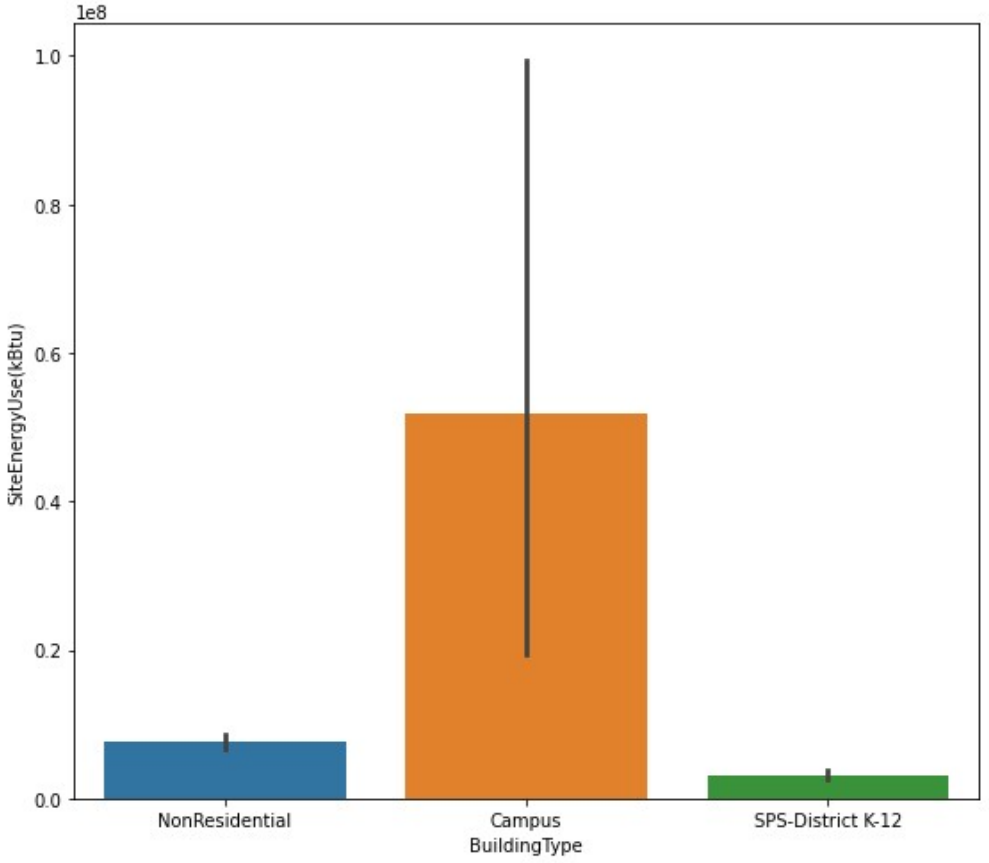
Analyse exploratoire



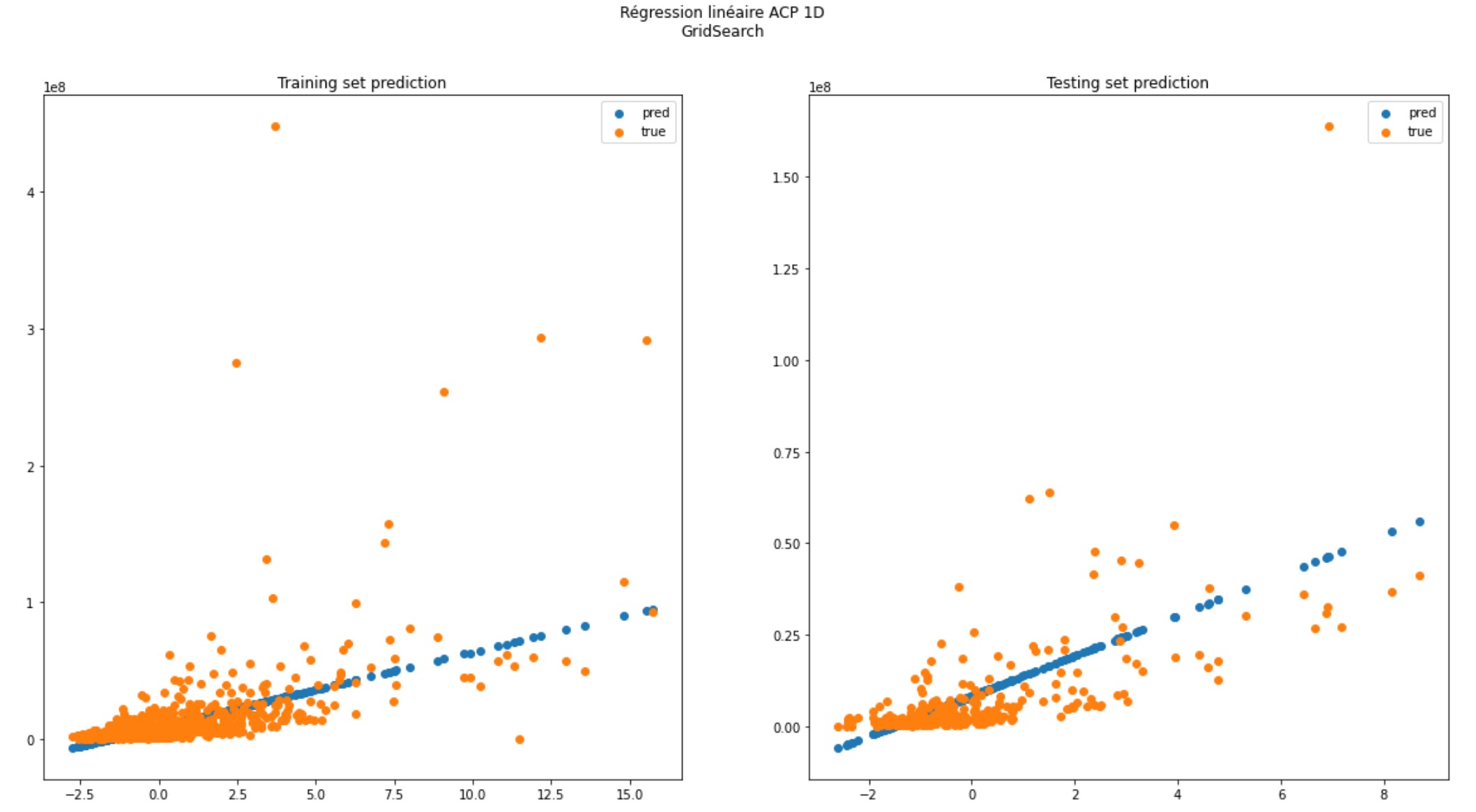
Analyse exploratoire

Consommation et émission en fonction du type de bâtiment



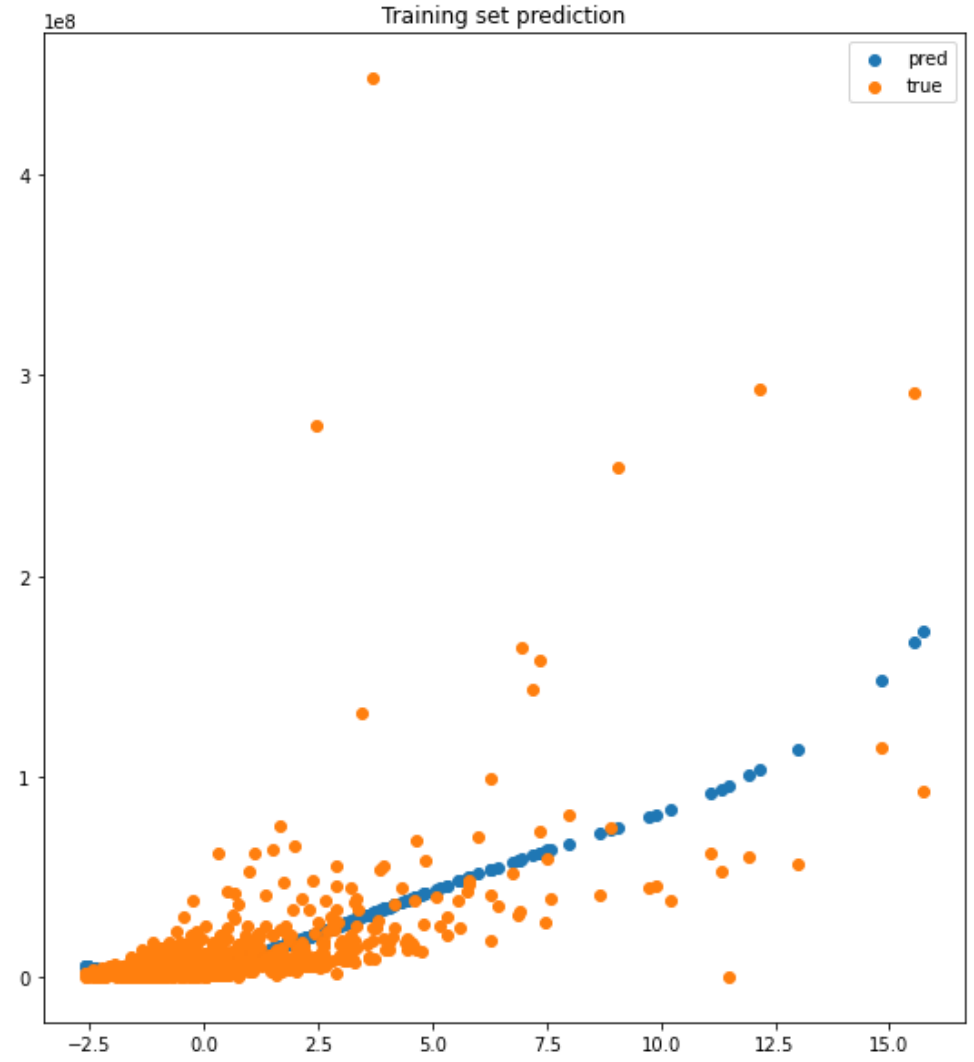


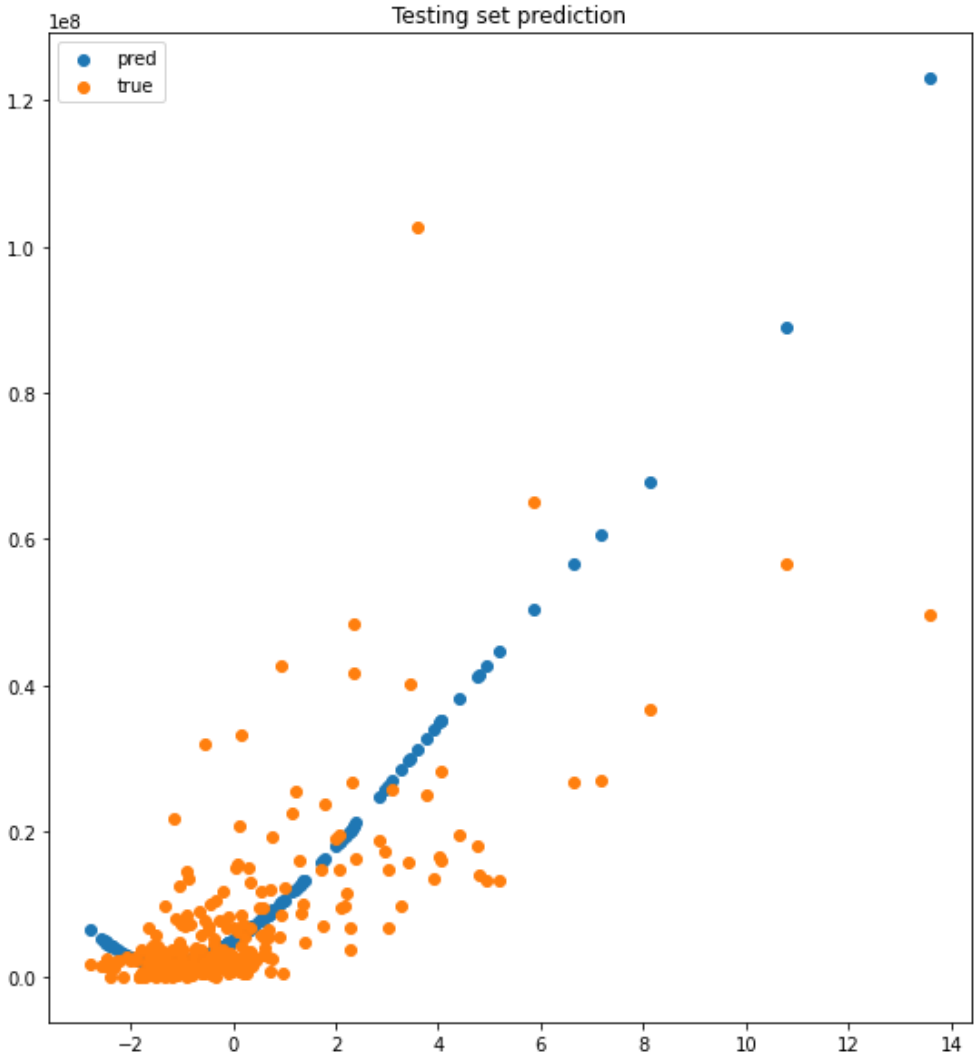
Modèles de régression linéaire



Modèles de régression linéaire

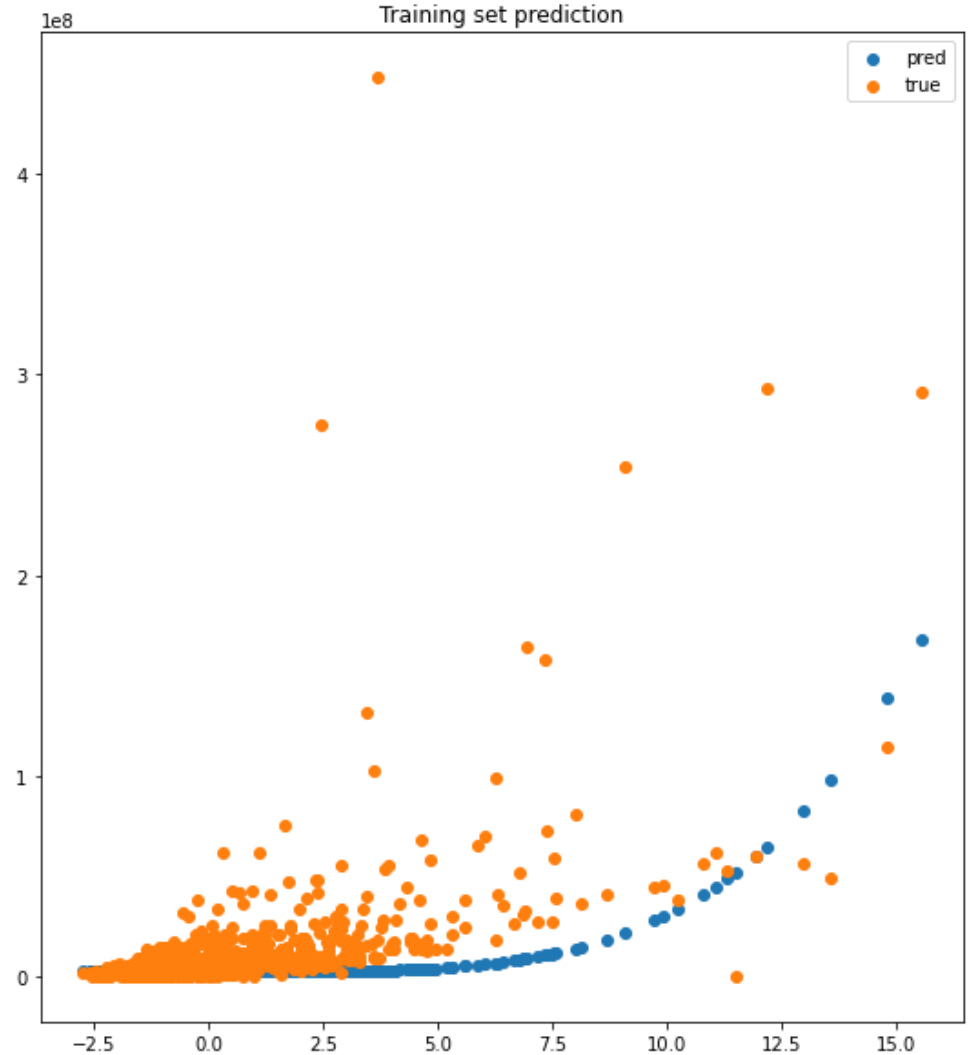
Régression KernelRidge ACP 1D
GridSearch

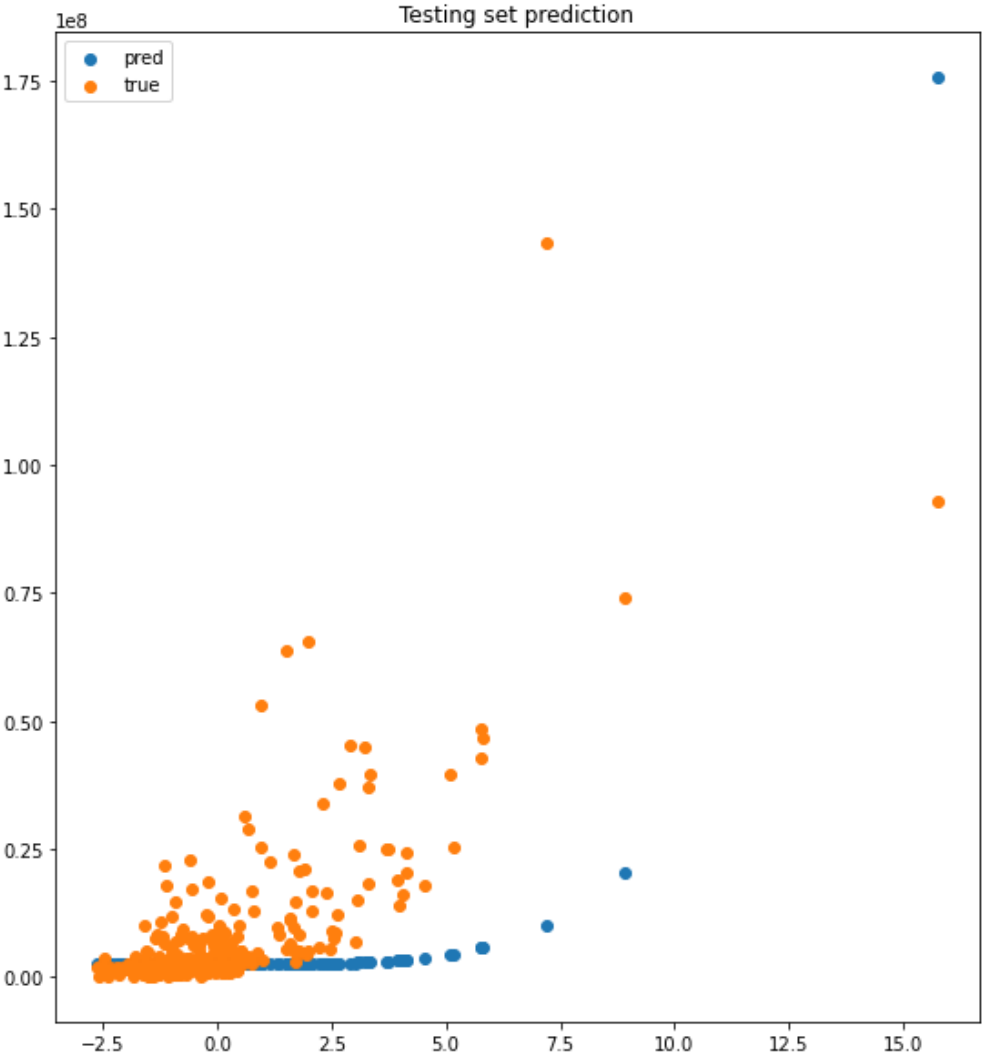




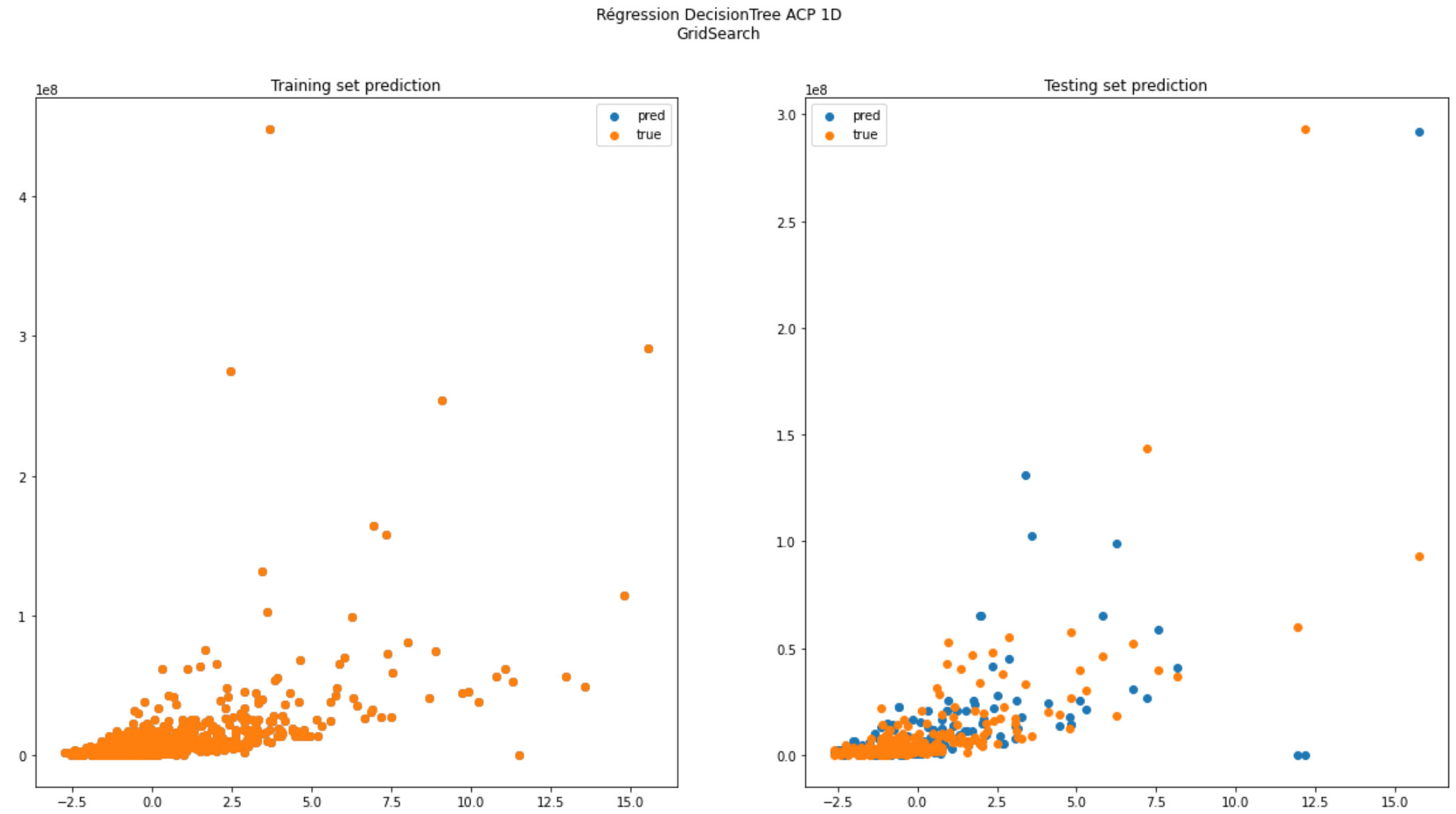
Modèle à vecteur support

SVR ACP 1D
GridSearch



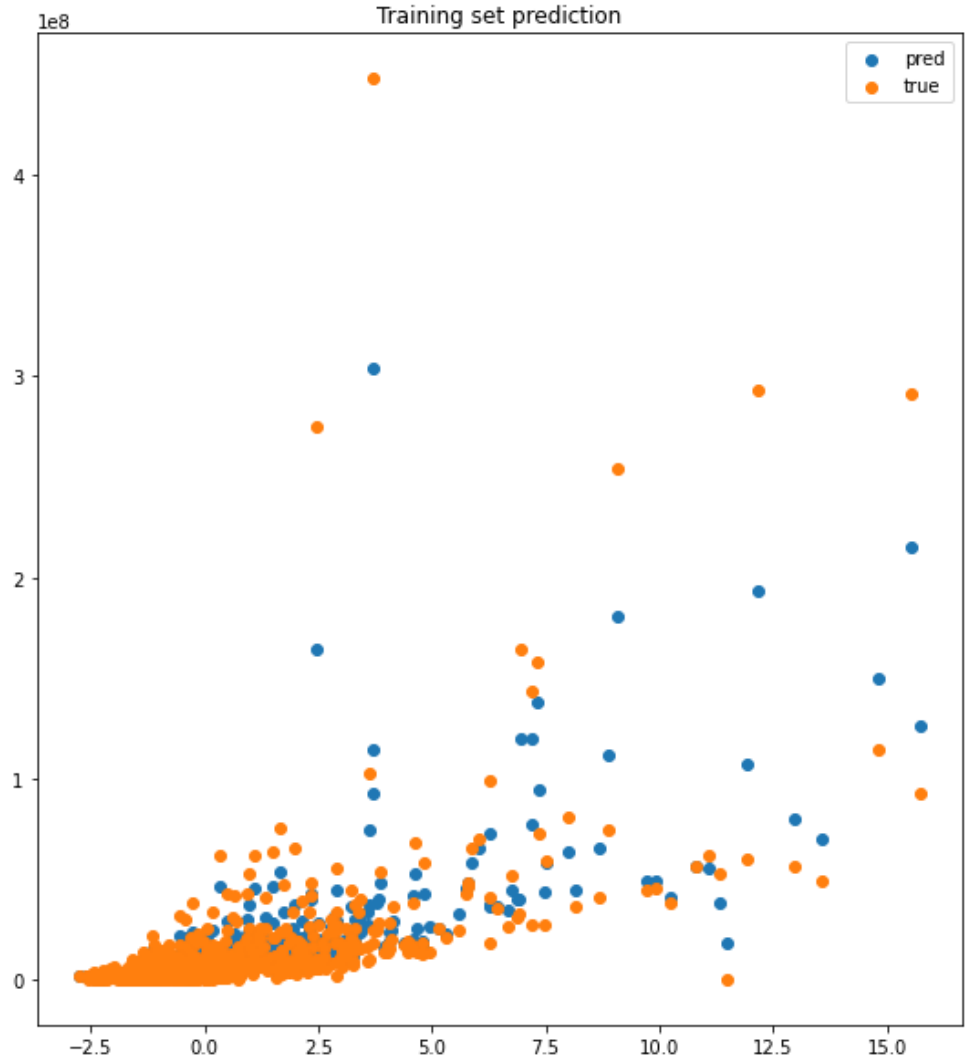


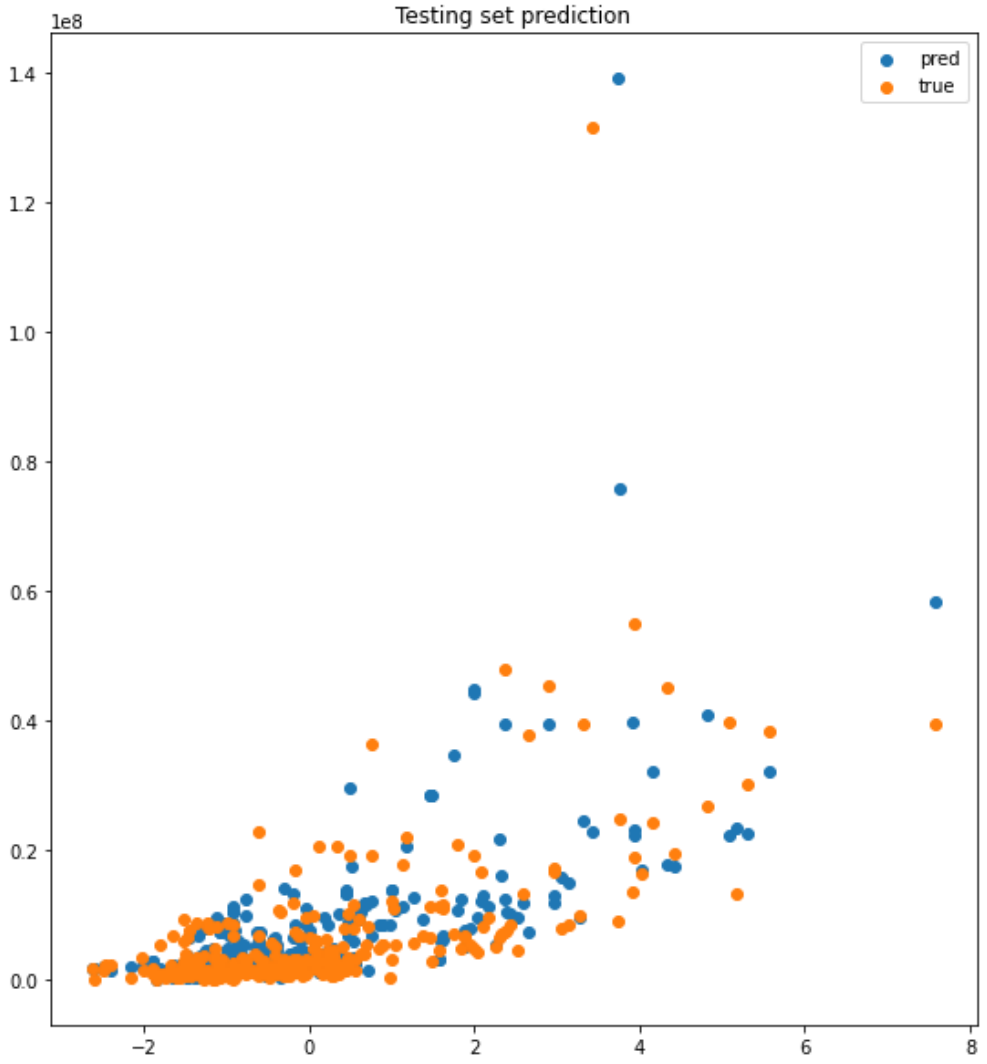
Modèle arbre de décision



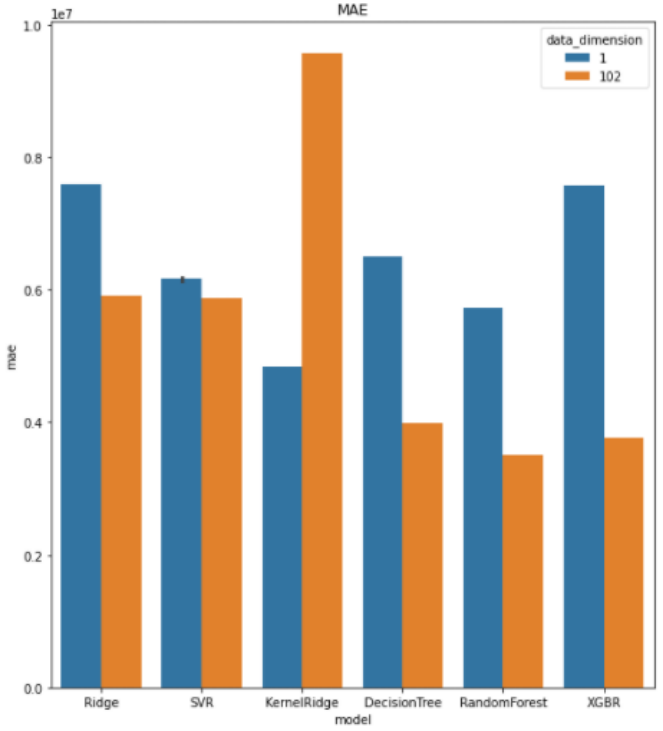
Modèle de forêt aléatoire

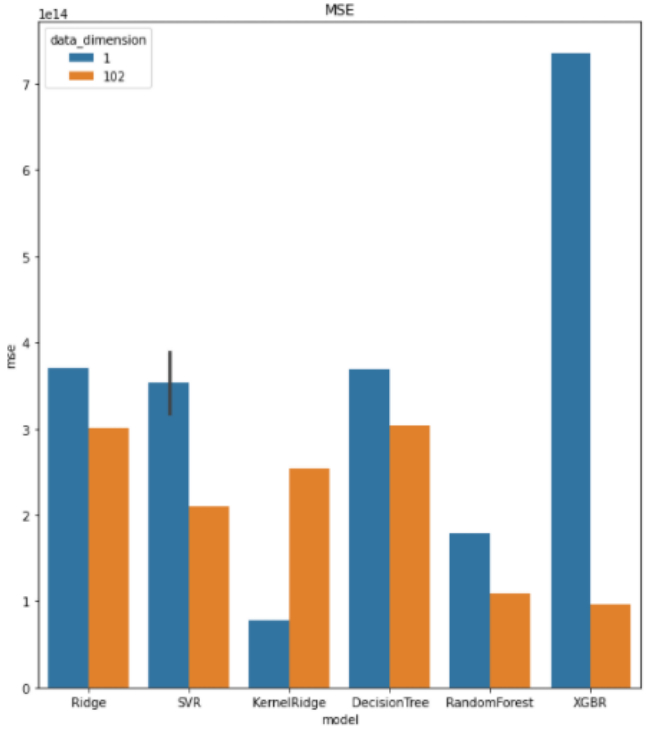
Régression RandomForest ACP 1D
GridSearch

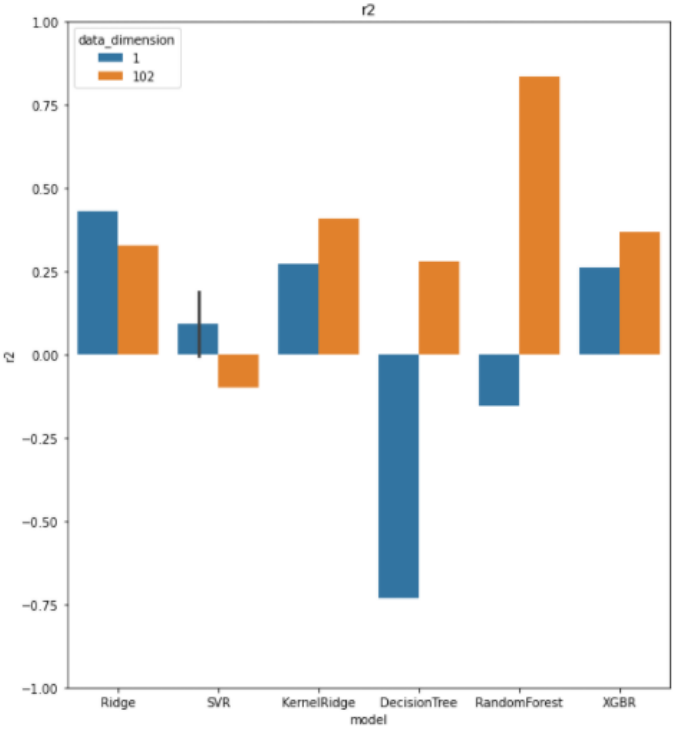




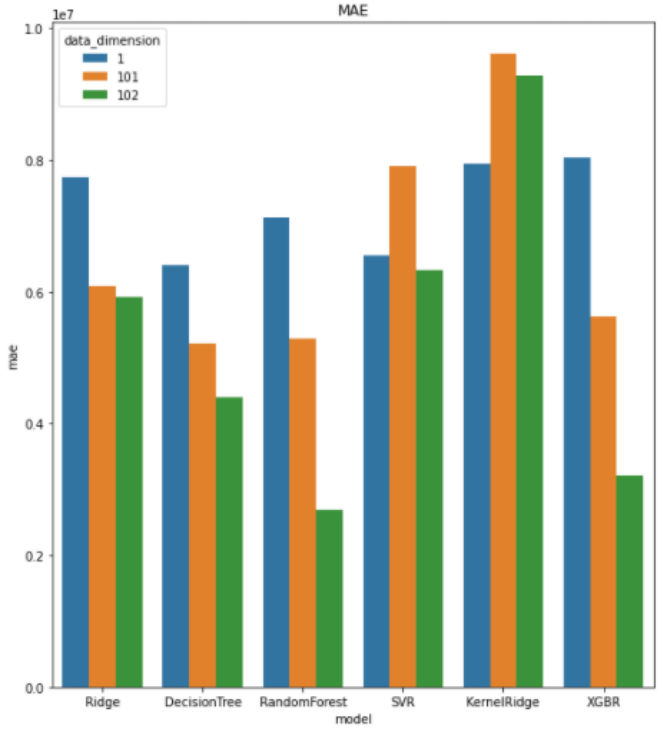
Résultats prédiction énergie

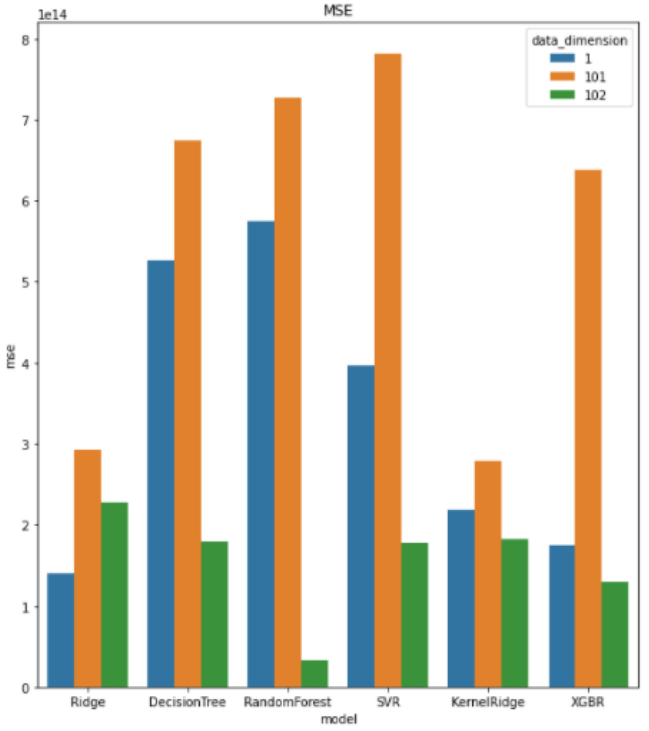


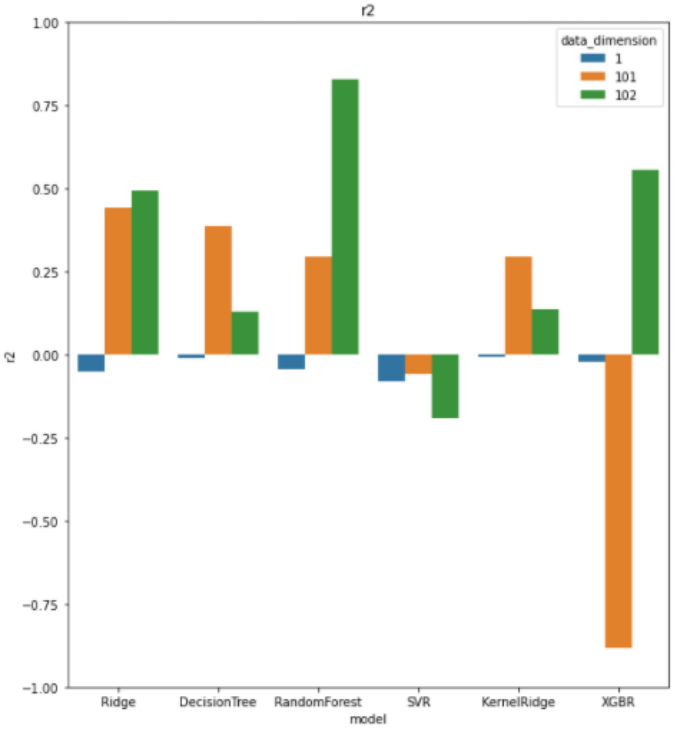




Résultats prédiction CO^2







Conclusion

- L’EnergyStar Score améliore les prédictions
- Modèle utilisé : Random Forest