FORMATION DATA SCIENTIST

PROJET 4 : ANTICIPEZ LES BESOINS EN CONSOMMATION DE BÂTIMENTS

Soutenu par:
Kourouma Sekouba Aissatou

SOMMAIRE

1.MISSION

Prédire les émissions de CO2 et la consommation totale d'énergie des bâtiments non résidentiels de Seattle en 2050

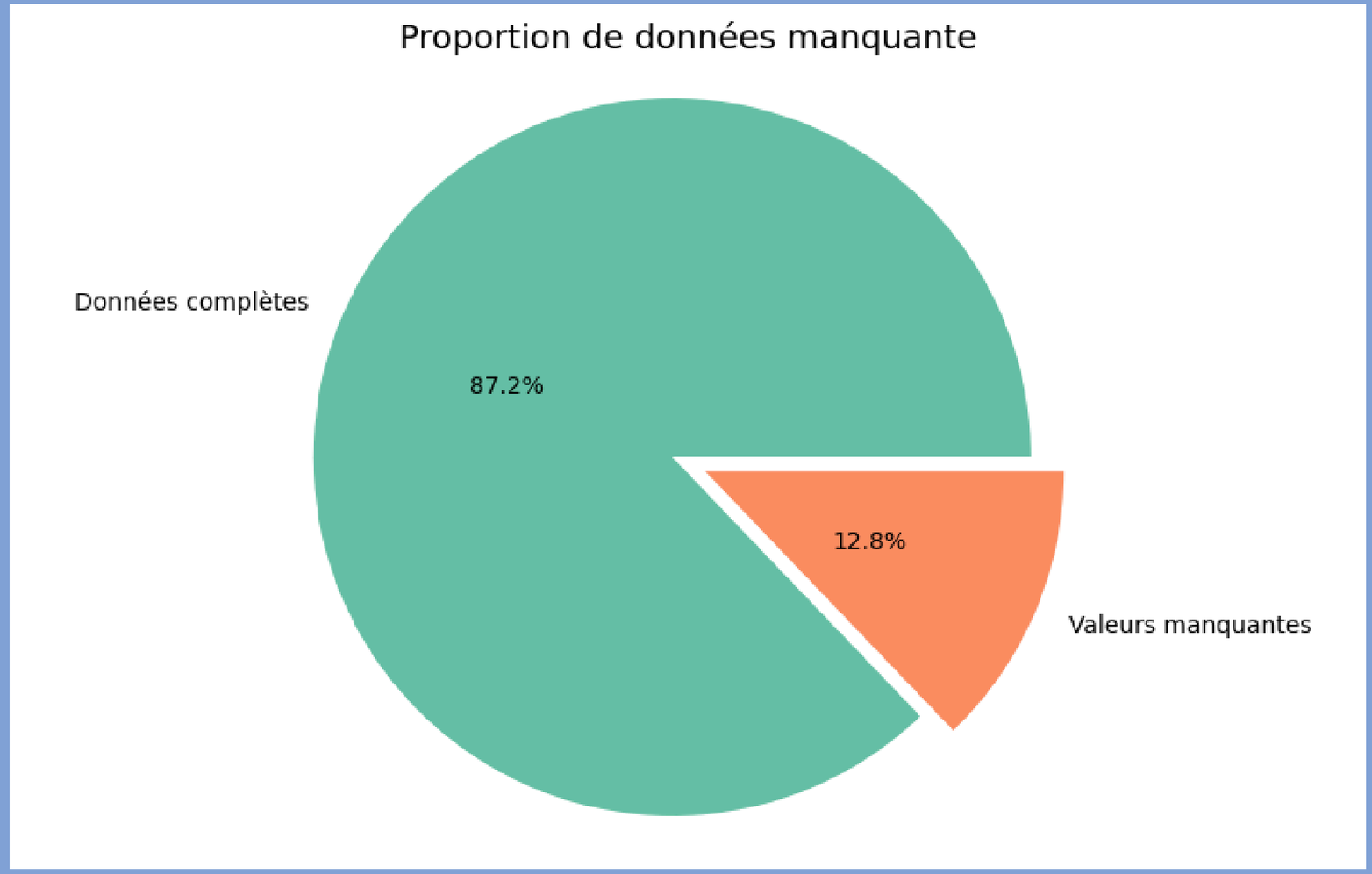
→ 2 Modèles différents

Seattle

- Peut-on prédire efficacement à partir des données collectées en 2016, afin de soutenir les objectifs de neutralité carbone de la ville pour 2050 ?
- Quelle est la précision des modèles de prédiction et quelles variables jouent un rôle crucial dans ces prédictions ?
- Comment optimiser les ressources et les efforts de la ville de Seattle en utilisant les données existantes pour prévoir les consommations et émissions futures, tout en minimisant les coûts opérationnels ?

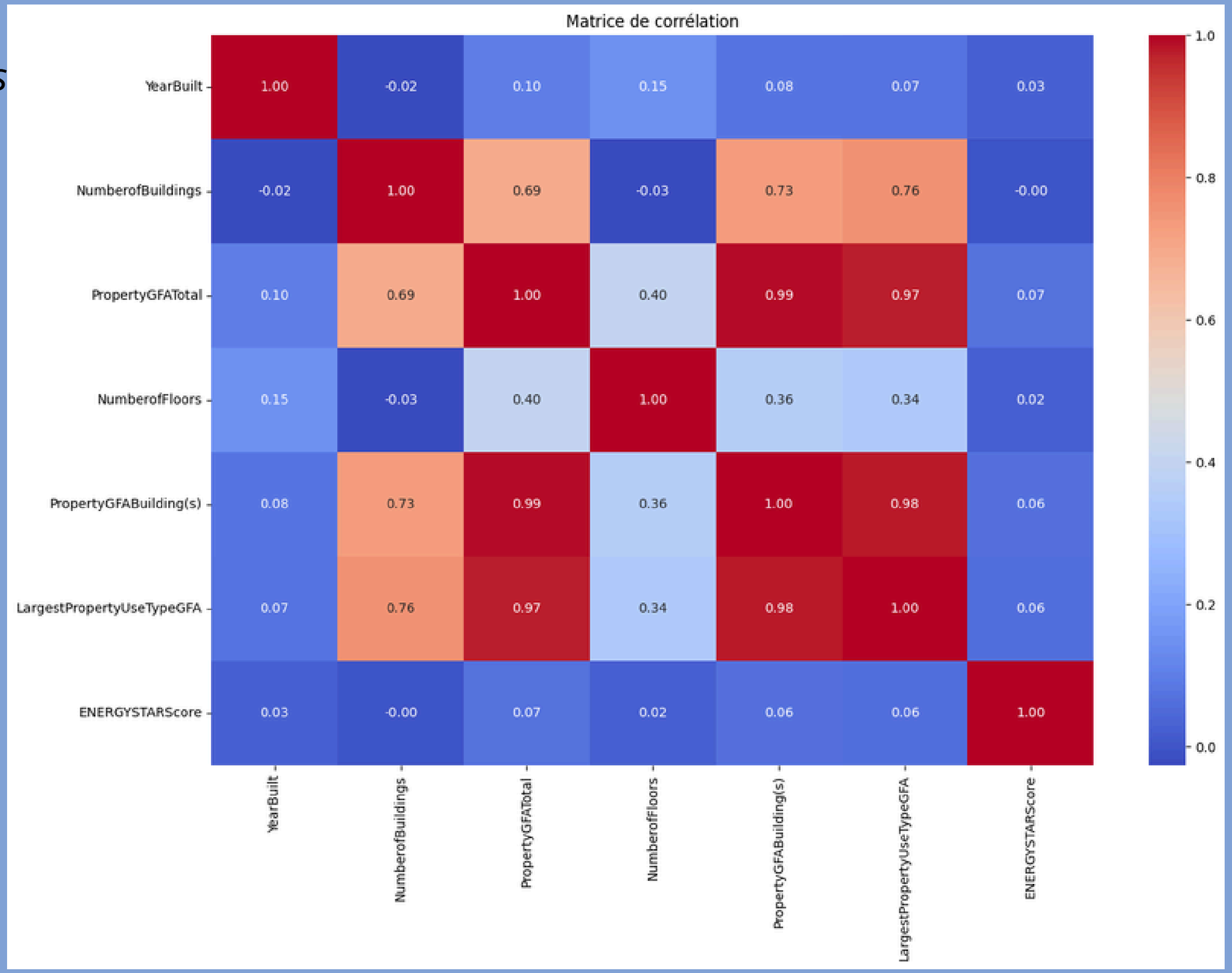
2. PRESENTATION DU JEU DE DONNÉE

Le jeux de donnée comporte une taille de (3376, 46)

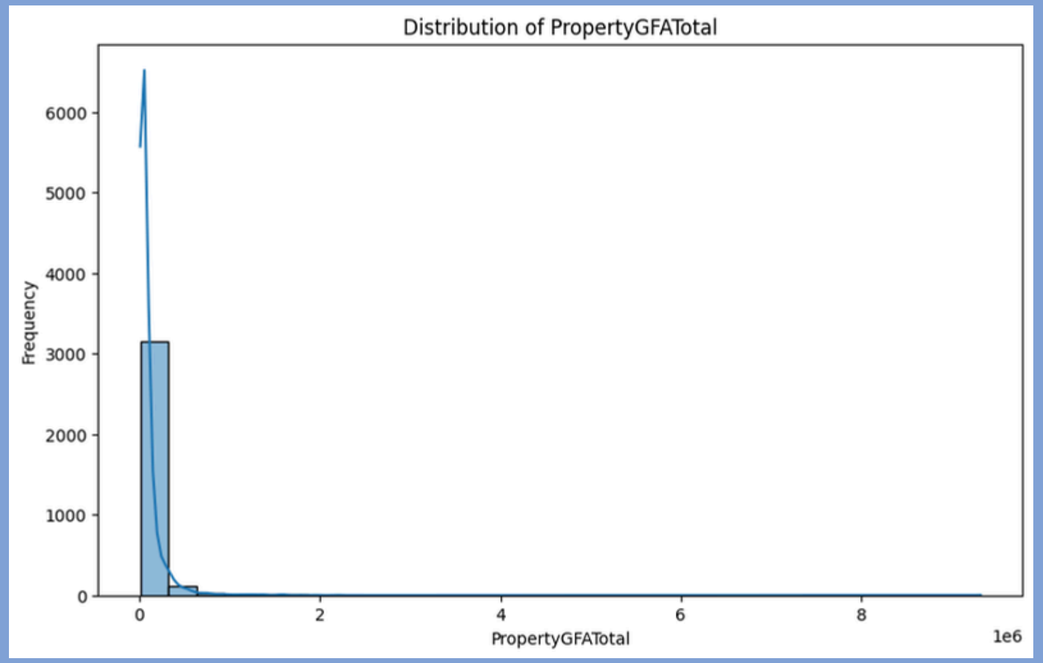


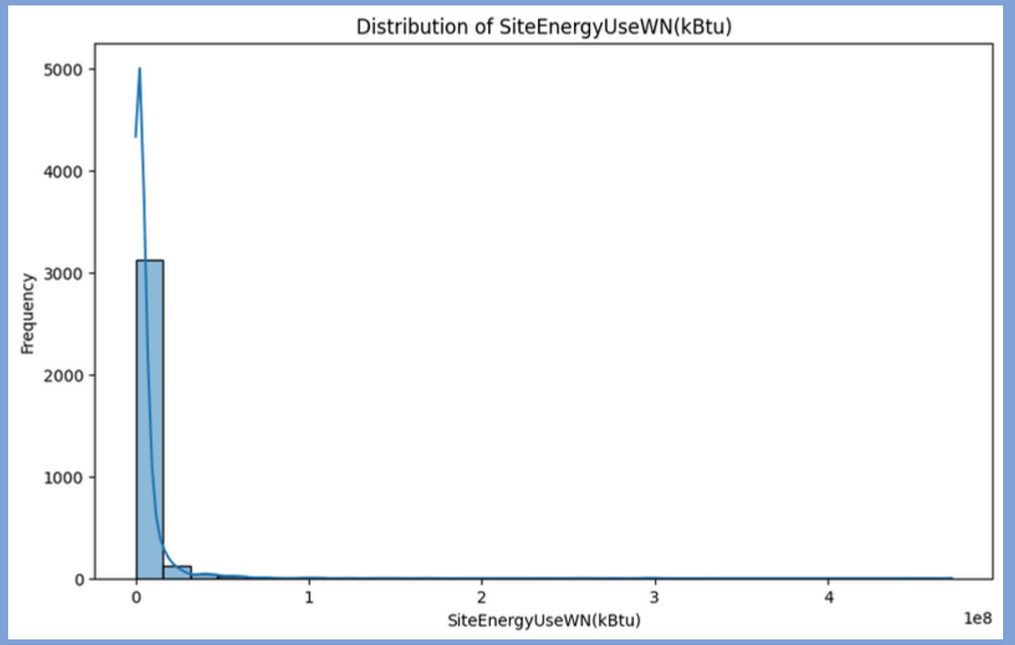
3. ANALYSE EXPLORATOIRE

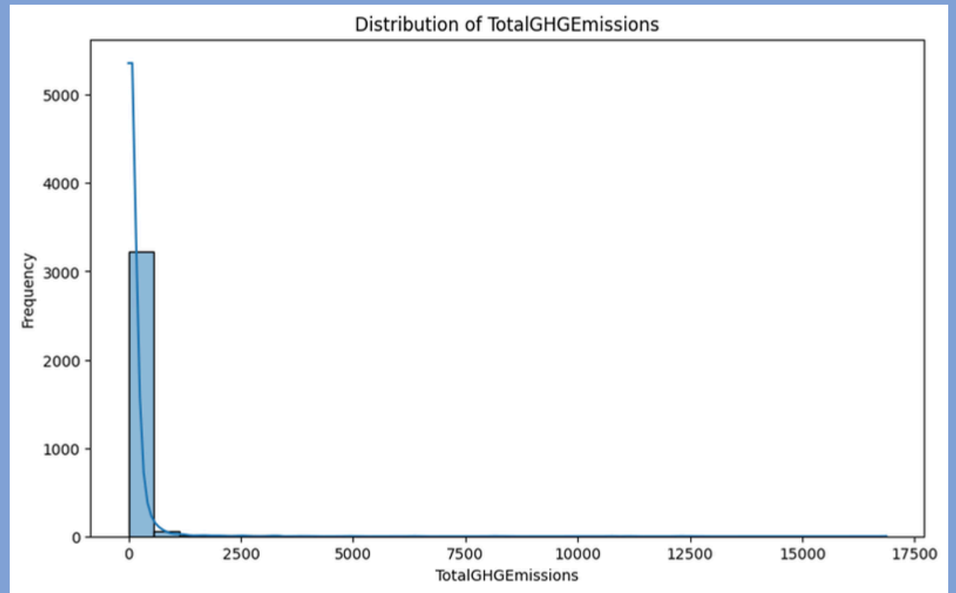
Corrélation des variables pertinentes

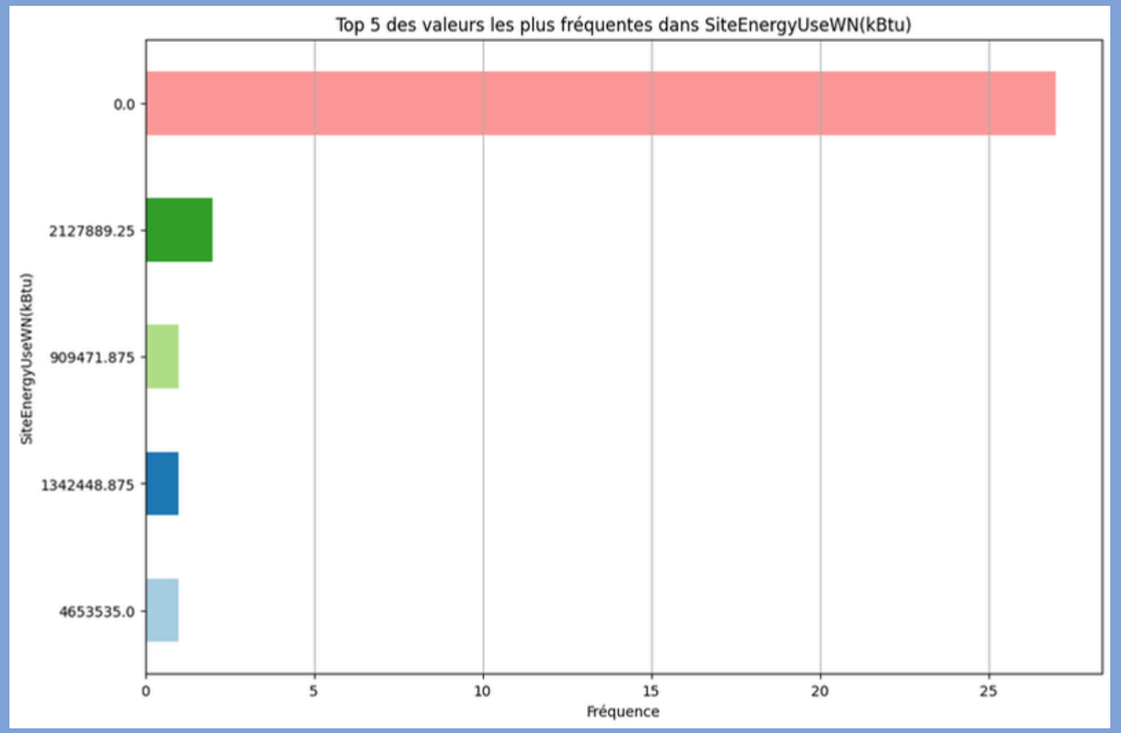


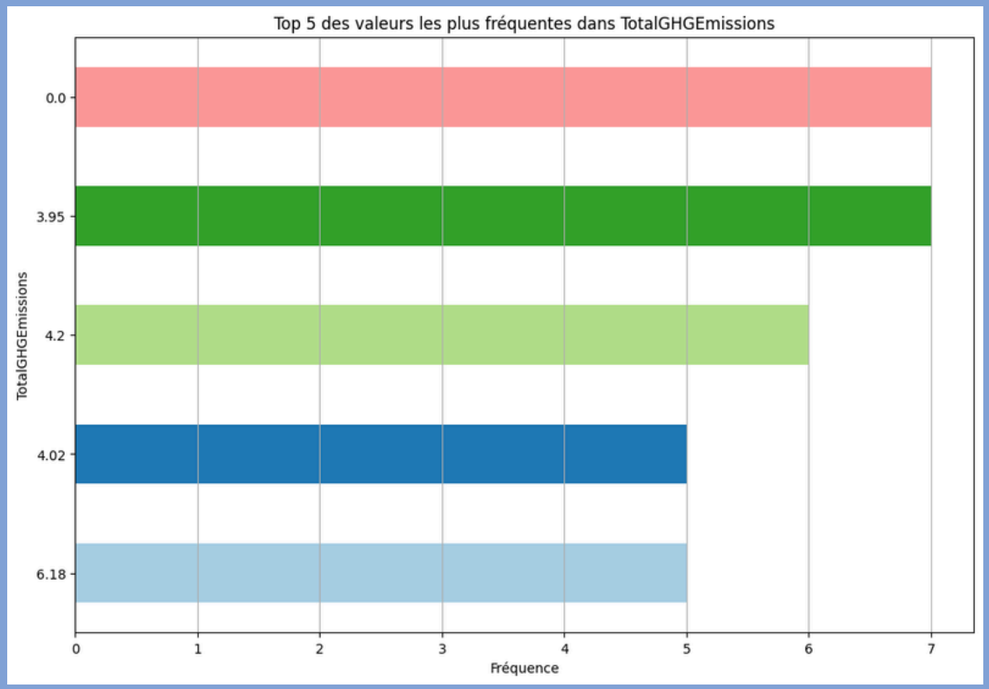
3. ANALYSE EXPLORATOIRE

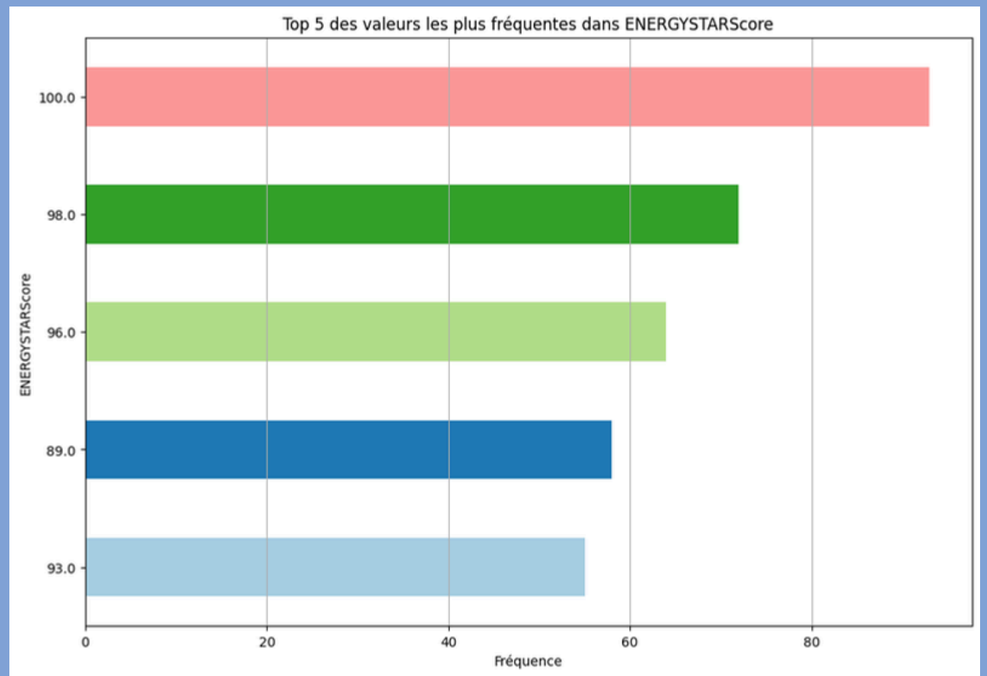


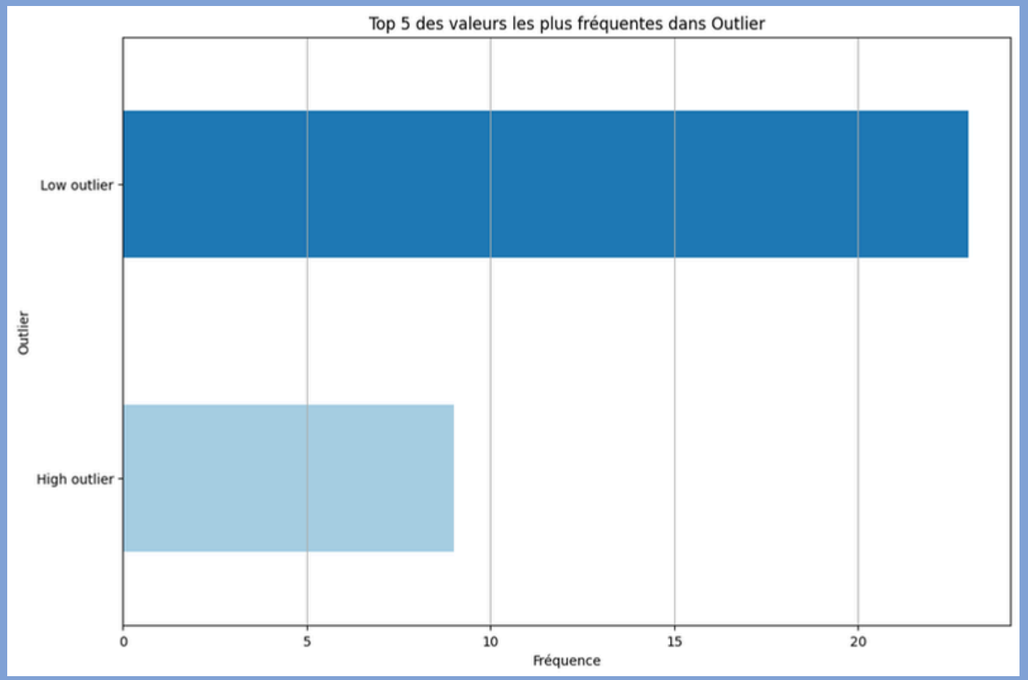




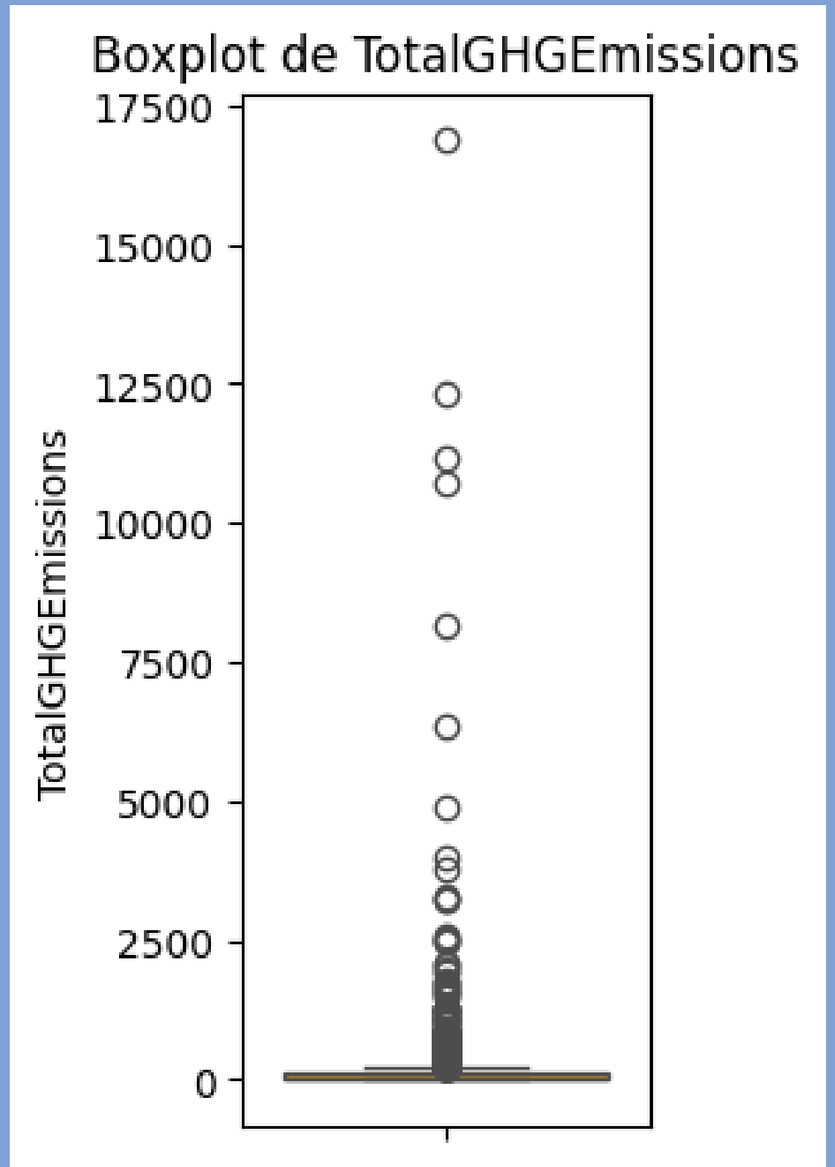

Top 5 des valeurs les plus fréquentes dans SiteEnergyUseWN(kBtu)
SiteEnergyUseWN(kBtu)
0.0
2127889.25
909471.875
1342448.875
4653535.0
0
5
10
15
20
25
Fréquence


Top 5 des valeurs les plus fréquentes dans TotalGHGEmissions
TotalGHGEmissions
0.0
3.95
4.2
4.02
6.18
0
1
2
3
4
5
6
7
Fréquence

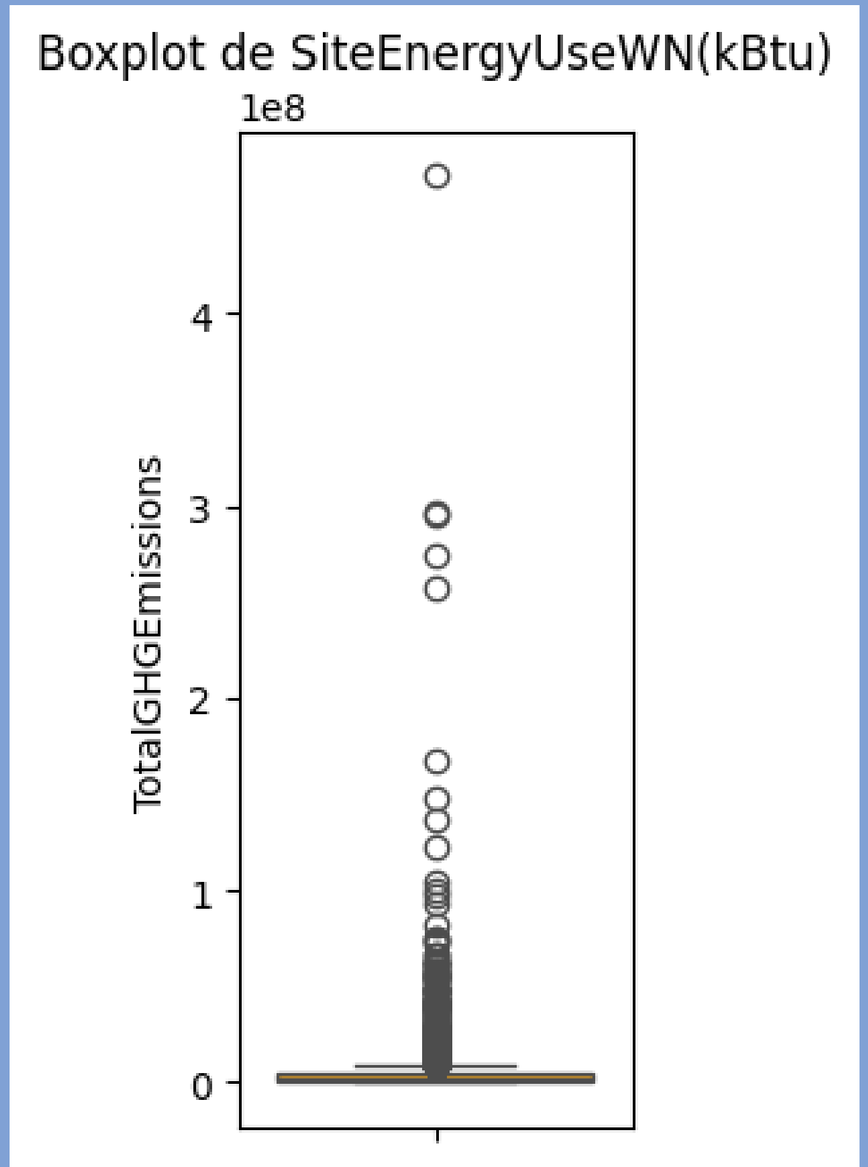

Top 5 des valeurs les plus fréquentes dans ENERGYSTARScore
ENERGYSTARScore
100.0
98.0
96.0
89.0
93.0
0
20
40
60
80
Fréquence


Top 5 des valeurs les plus fréquentes dans Outlier
Outlier
Low outlier
High outlier
0
5
10
15
20
Fréquence

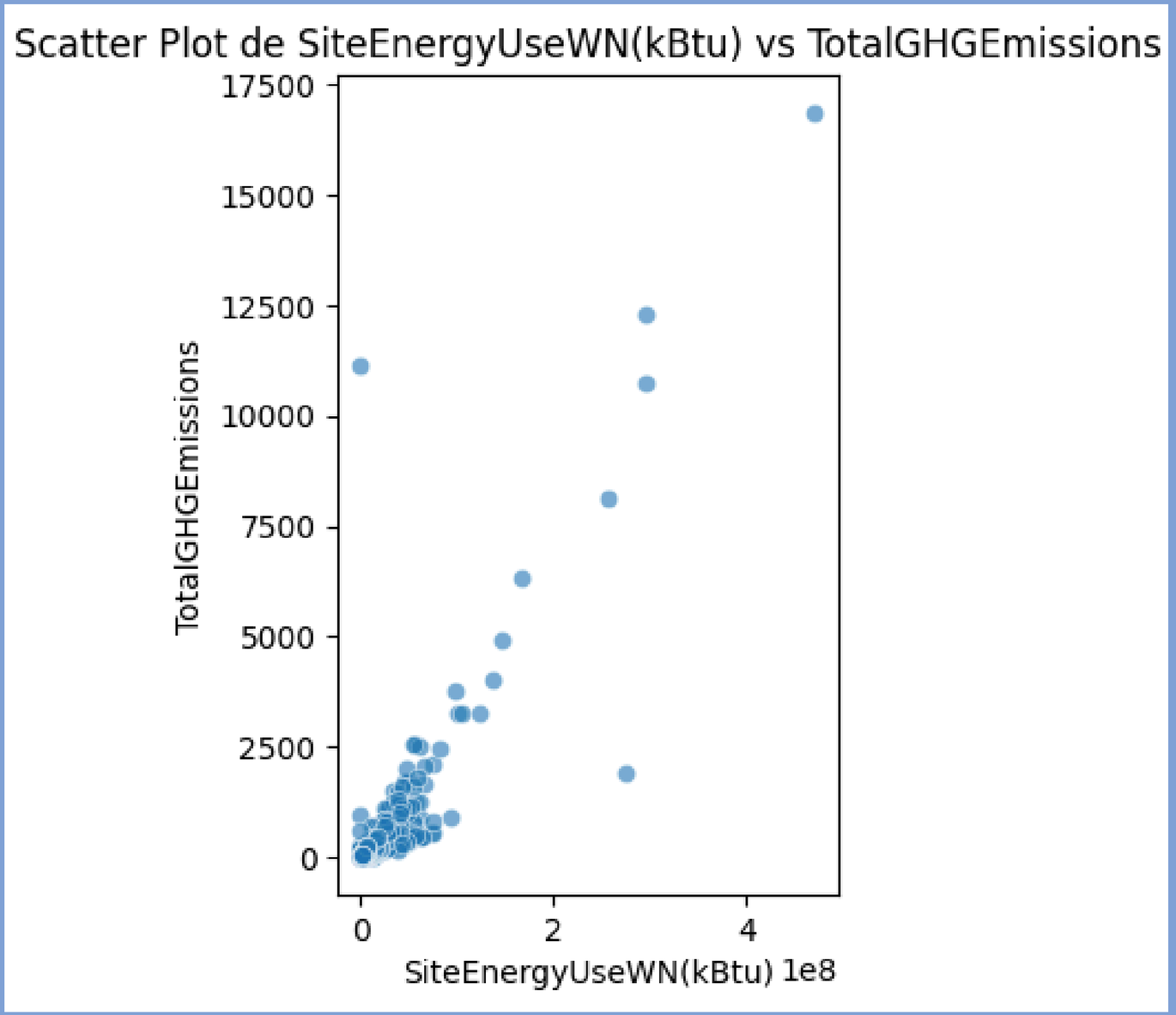
Boxplot de la target TotalGHGEmissions



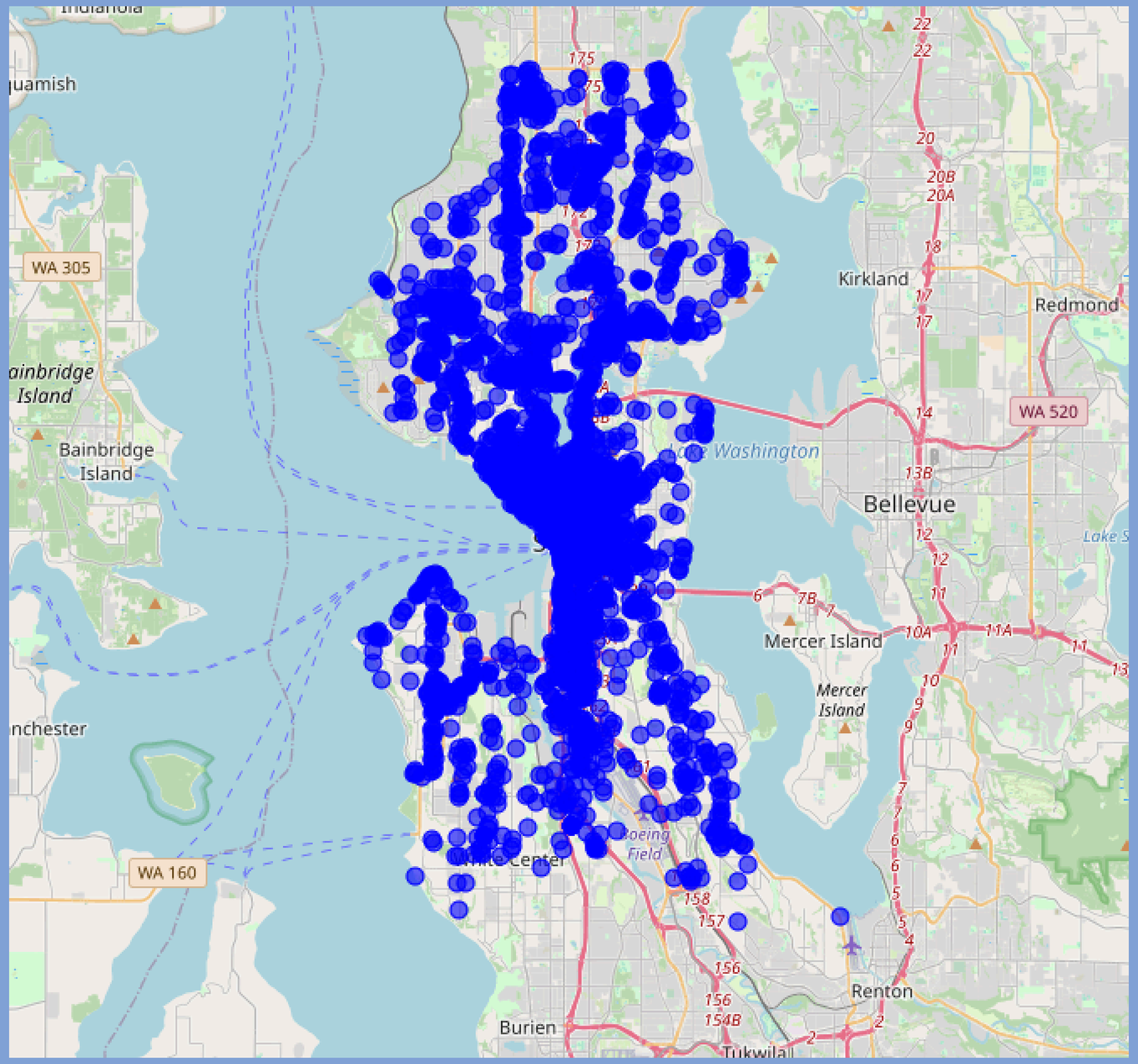
Boxplot de la target SiteEnergyUseWN(kBtu)



Analyse bivarié des 2 target



Distribution géographique des bâtiments



4. MODELE DE PREDICTION

LES TARGET: TOTALGHGEMISSIONS' ET 'SITEENERGYUSE(KBTU)'.

Pour chaque model choisi nous avons élaboré comme ci-dessous

- Séparation du jeu de donnée
  - Choix du model de prediction
    - Choix du préprocesseur
      - Choix du grille de paramètre

ElasticNet, SVM, RandomForestRegressor, Gradient Boosting

Paramètre des différents modèles

paramètres des modèles			
Regression lineair elastic net	**Suport Vector Machine**	**RandomForestRegressor**	**Gradient Boosting**
regressor__alpha':1e-3, 1e-2, ...1e14, 1e15	regressor__C': [0.1, 1, 10, 100, 1000, 10000],	regressor__n_estimators': [100, 200, 300],	regressor__n_estimators': [100, 200, 300],
regressor__l1_ratio': 0.1, 0.2,...0.8, 0.9	regressor__epsilon': [0.1, 0.2, 0.5],	regressor__max_depth': [None, 10, 20, 30],	regressor__learning_rate': [0.01, 0.1, 0.2],
	regressor__kernel': ['linear', 'rbf']	regressor__min_samples_split': [2, 5, 10],	regressor__max_depth': [3, 5, 7],
		regressor__min_samples_leaf': [1, 2, 4]	regressor__min_samples_split': [2, 5, 10]

Régression ElasticNet

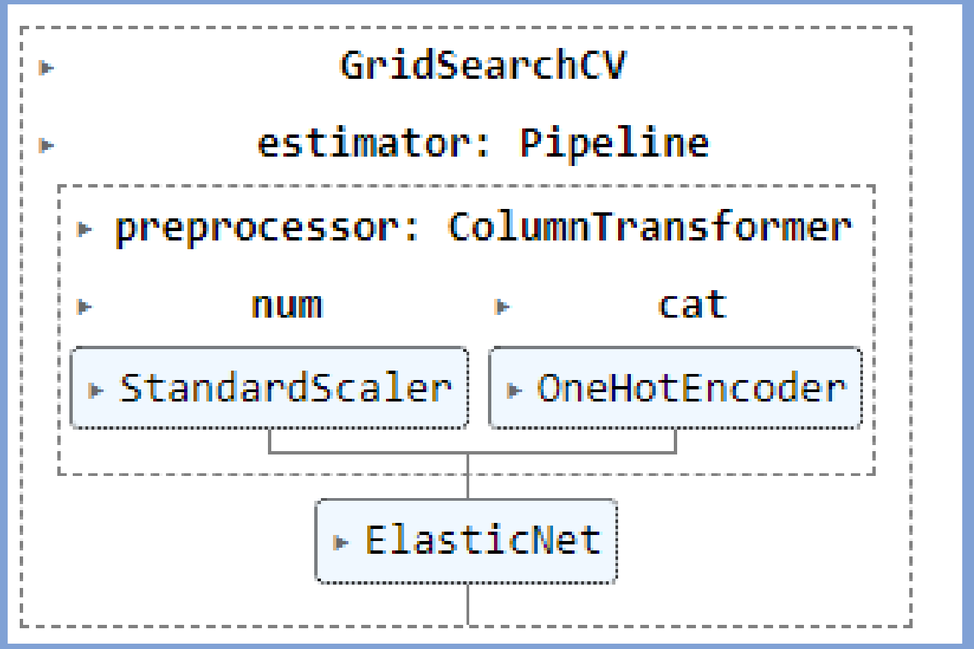
$$\text{Coût Ridge} = \text{RSS} + \alpha \sum_{j=1}^{p} \beta_j^2$$

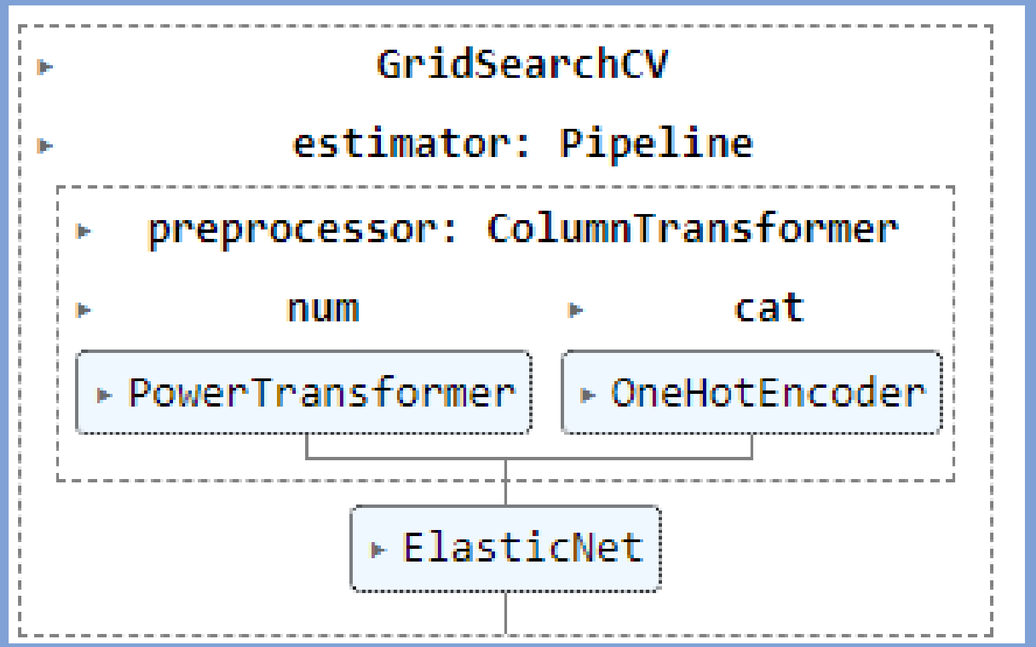
$$\text{Coût Lasso} = \text{RSS} + \alpha \sum_{j=1}^{p} |\beta_j|$$

- α1 et α2 : hyperparamètres de régularisation qui contrôlent la force des pénalités L1 et L2 respectivement.

- α2=0 purement Lasso.
- α1= 0 purement Ridge.

$$\text{Coût Elastic Net} = \text{RSS} + \alpha_1 \sum_{j=1}^{p} |\beta_j| + \alpha_2 \sum_{j=1}^{p} \beta_j^2$$





Avantage:

- Gestion de la multi colinéarité
- Sélection de variables
- Robustesse

RandomForest

- largement utilisée pour les tâches de classification et de régression en raison de sa capacité à améliorer la précision des prédictions tout en réduisant le risque de surapprentissage (overfitting)

Fonctionne la Forêt Aléatoire

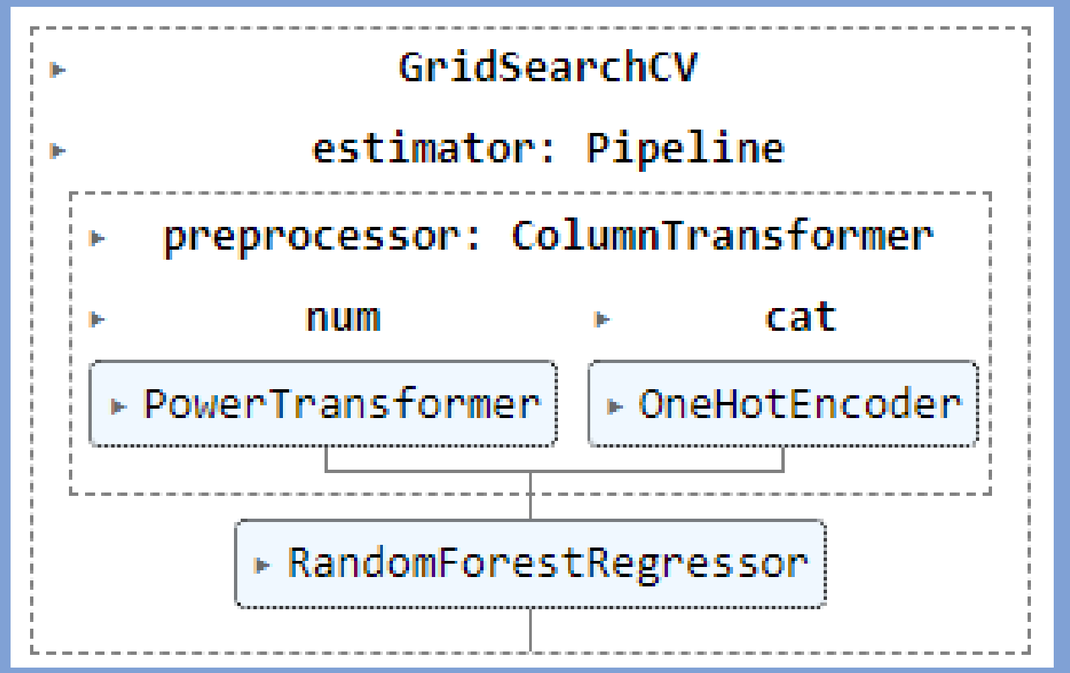
- Bagging (Bootstrap Aggregating)
- Sélection Aléatoire des Caractéristiques
- Prédiction Finale

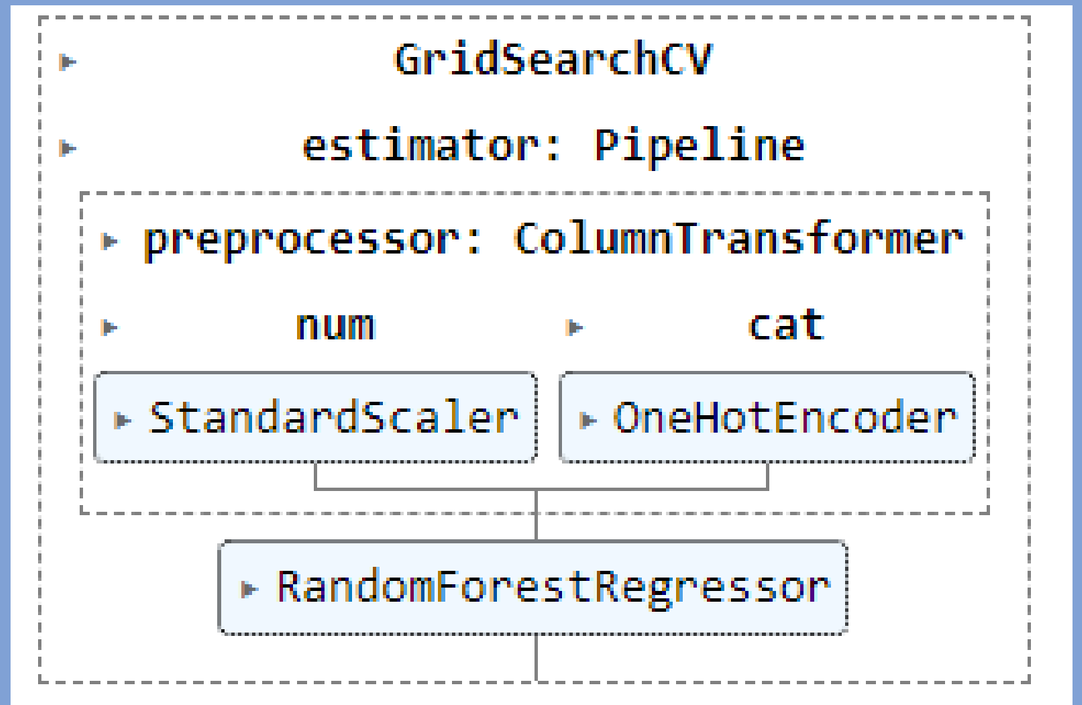
Avantages

- Robustesse et réduction du surapprentissage
- Capacité à gérer les grandes dimensions

Inconvénients

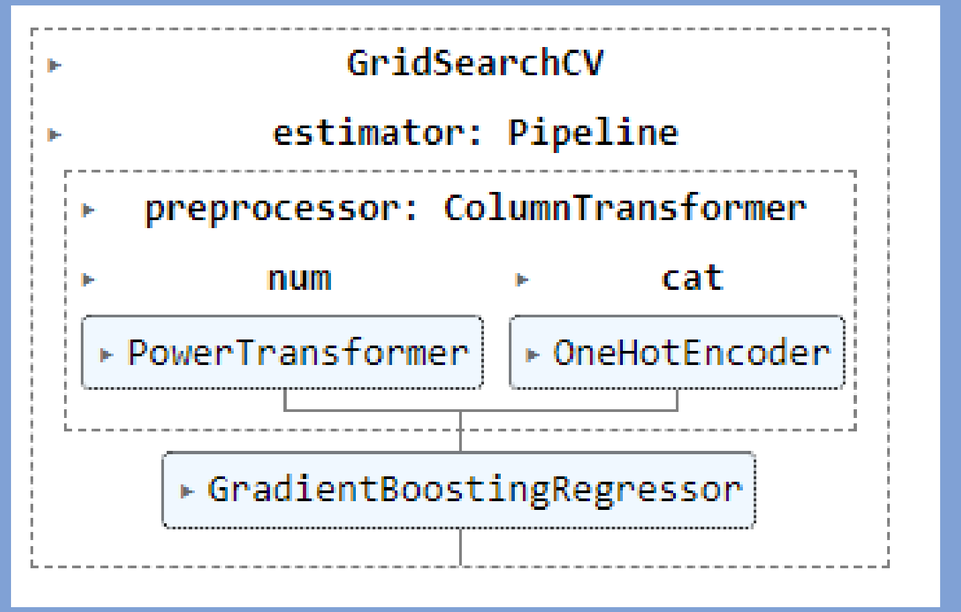
- Temps de calcul
- Sensibilité aux paramètres





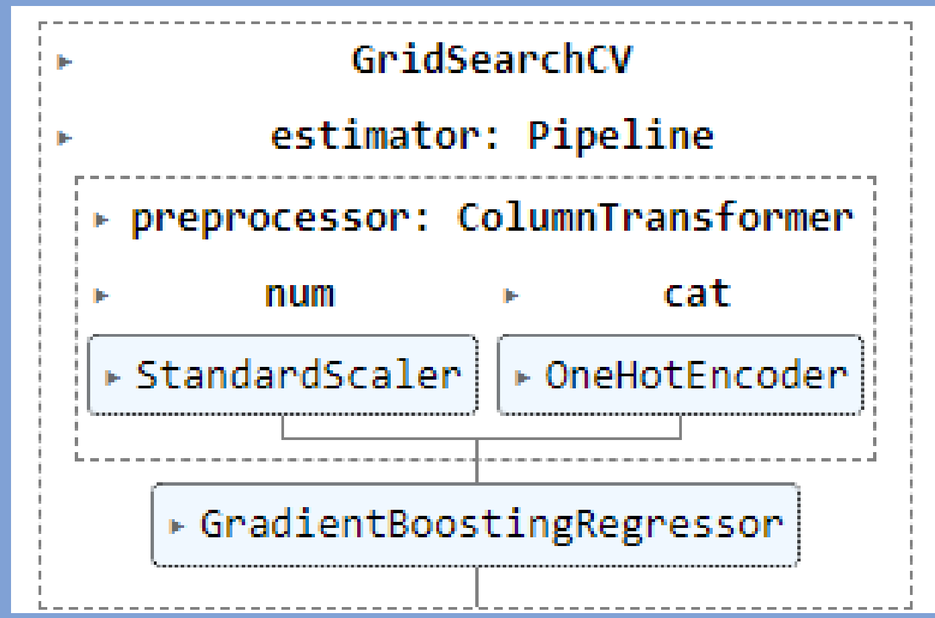
GradientBosting

- Classification et de régression
- combinent plusieurs modèles faibles, généralement des arbres de décision, pour former un modèle robuste et précis.



Fonctionnement le Gradient Boosting

- Initialisation
- Étape Itérative
- Facteur d'Apprentissage (Learning Rate): *Un petit taux d'apprentissage rend le processus plus lent mais peut conduire à une meilleure généralisation*



Avantages

- Puissance et flexibilité
- Réduction du biais

Inconvénients: Tendance au surapprentissage (overfitting)

Support Vector Machine (SVM)

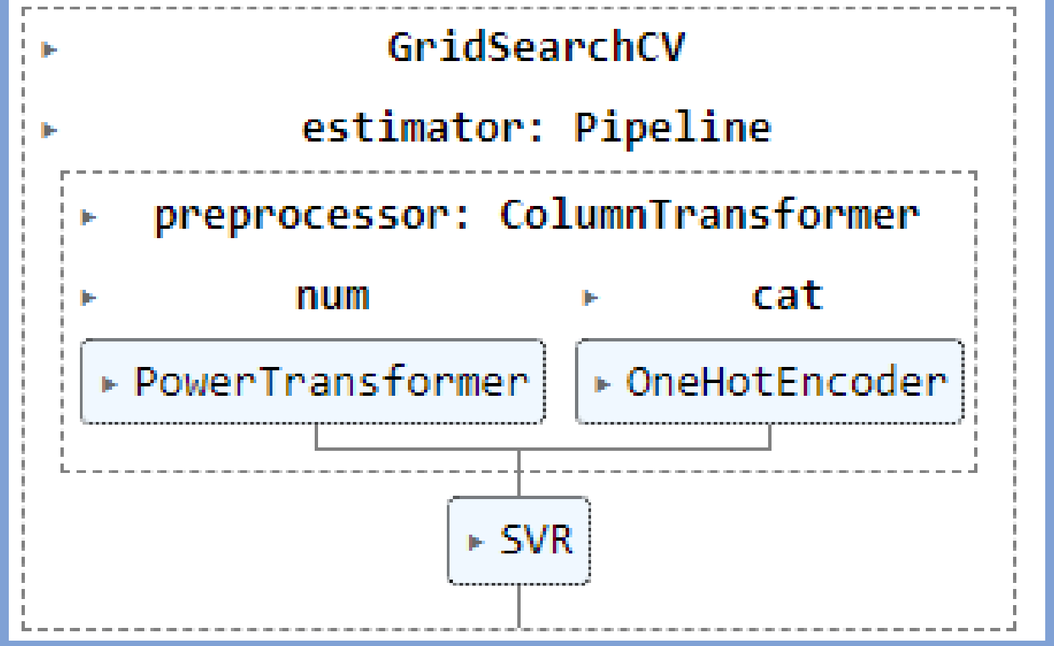
- principalement pour les classification, bien qu'elles puissent aussi être adaptées à la régression.
- L'idée centrale d'un SVM est de trouver un hyperplan qui sépare les données en classes distinctes

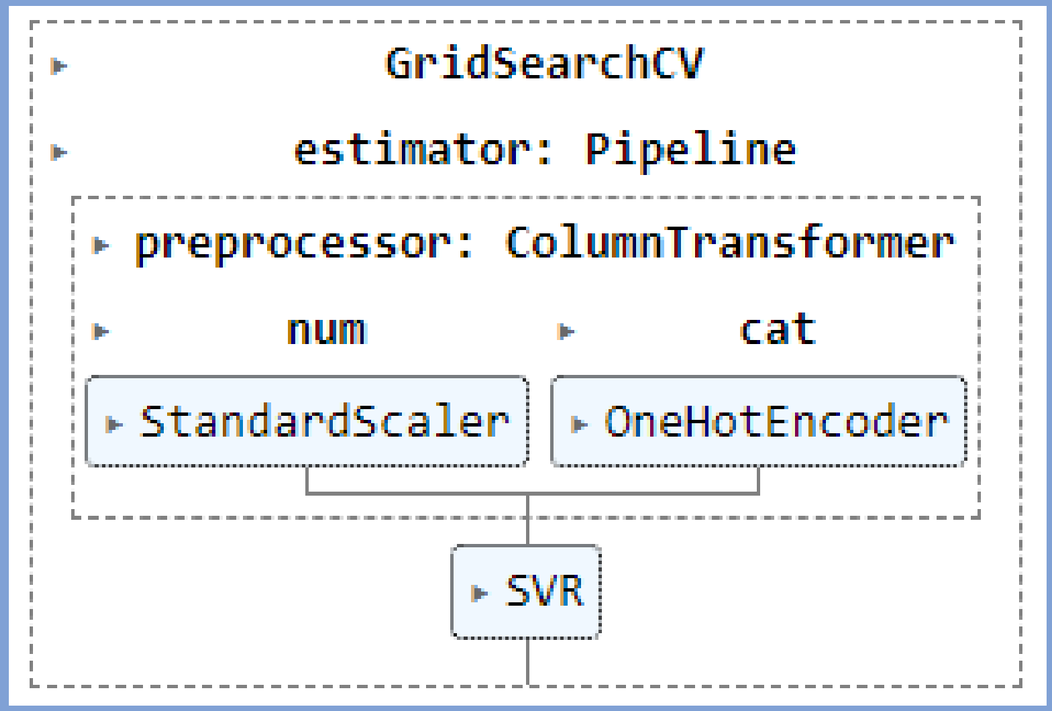
Classification Linéaire

- un SVM cherche à trouver un hyperplan (une ligne en 2D, un plan en 3D, etc.) qui divise les classes de manière à maximiser la marge.

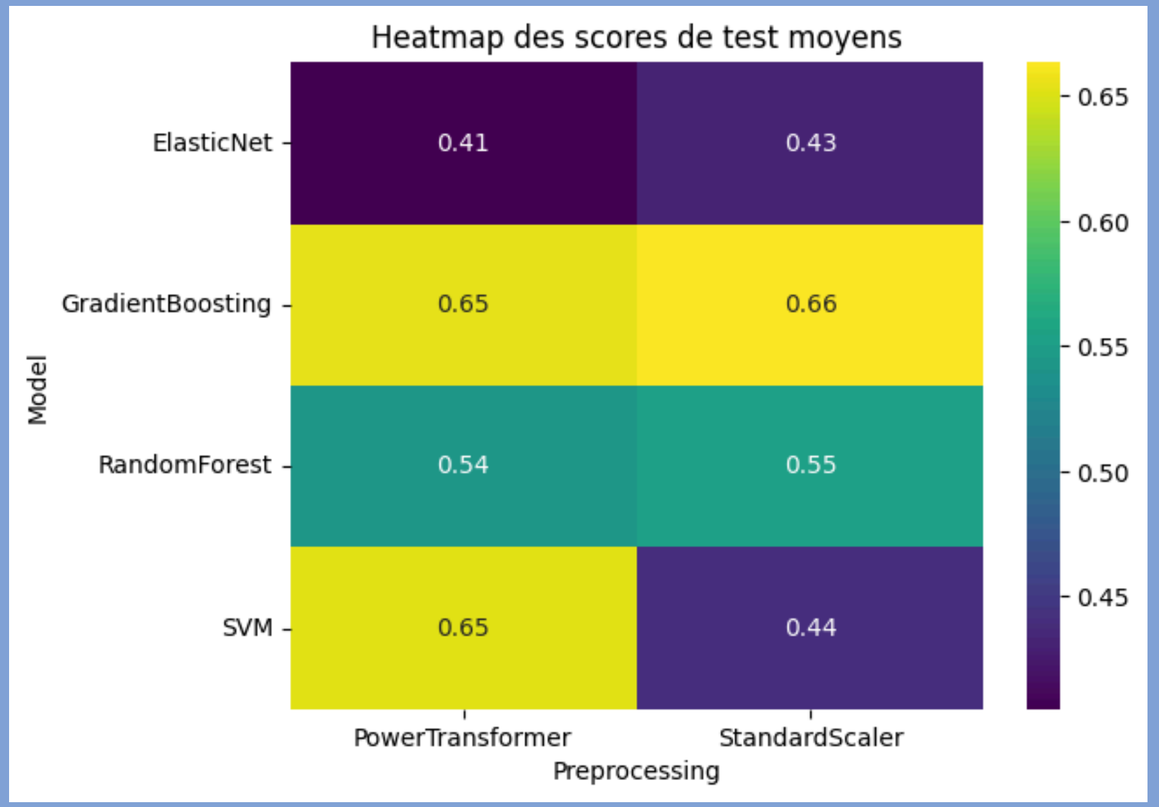
Cas Non Linéaire et le "Truc du Kernel"

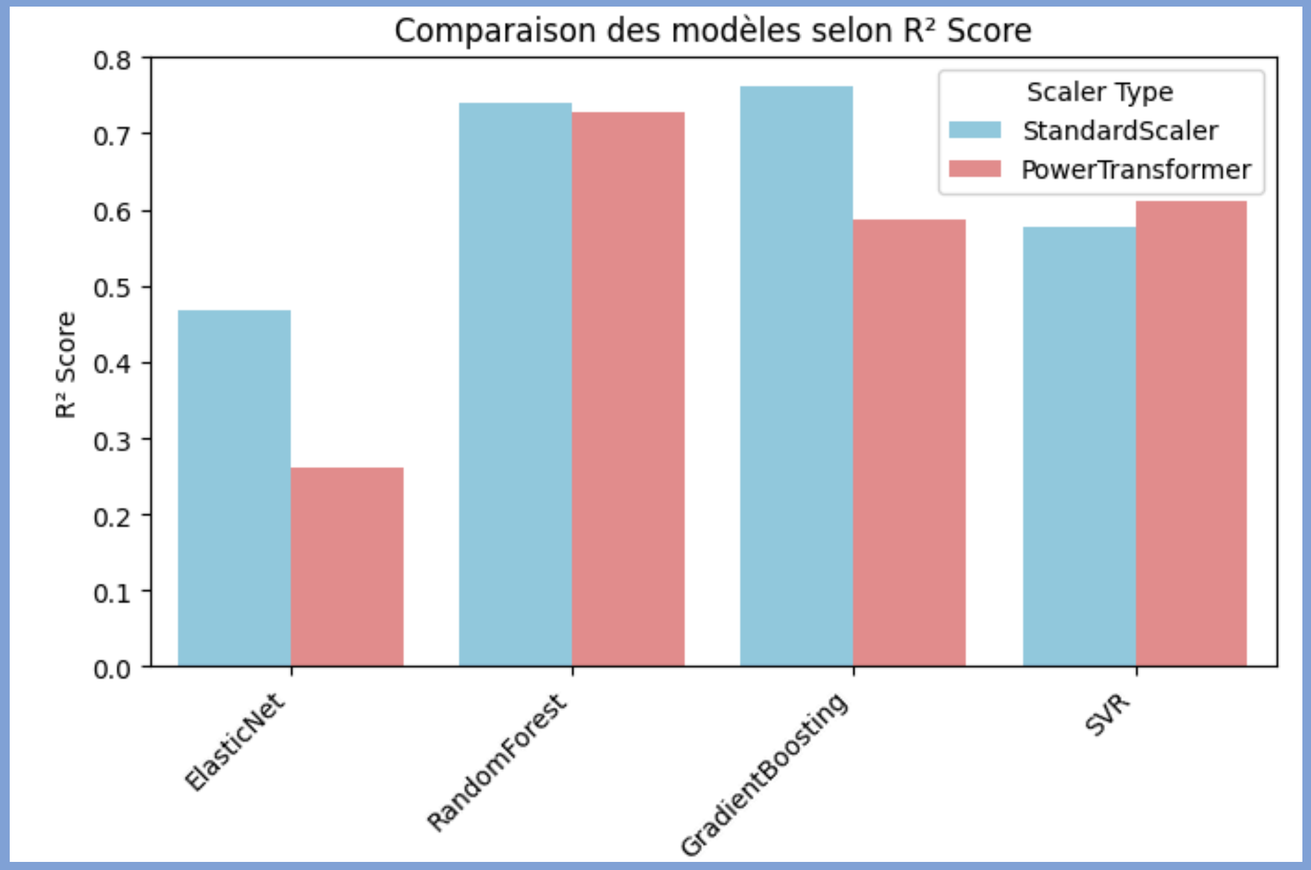
- Si les données ne sont pas séparables, un SVM peut en projetant les données dans un espace de dimension supérieure où une séparation linéaire devient possible. Cette projection est réalisée une fonction de kernel.





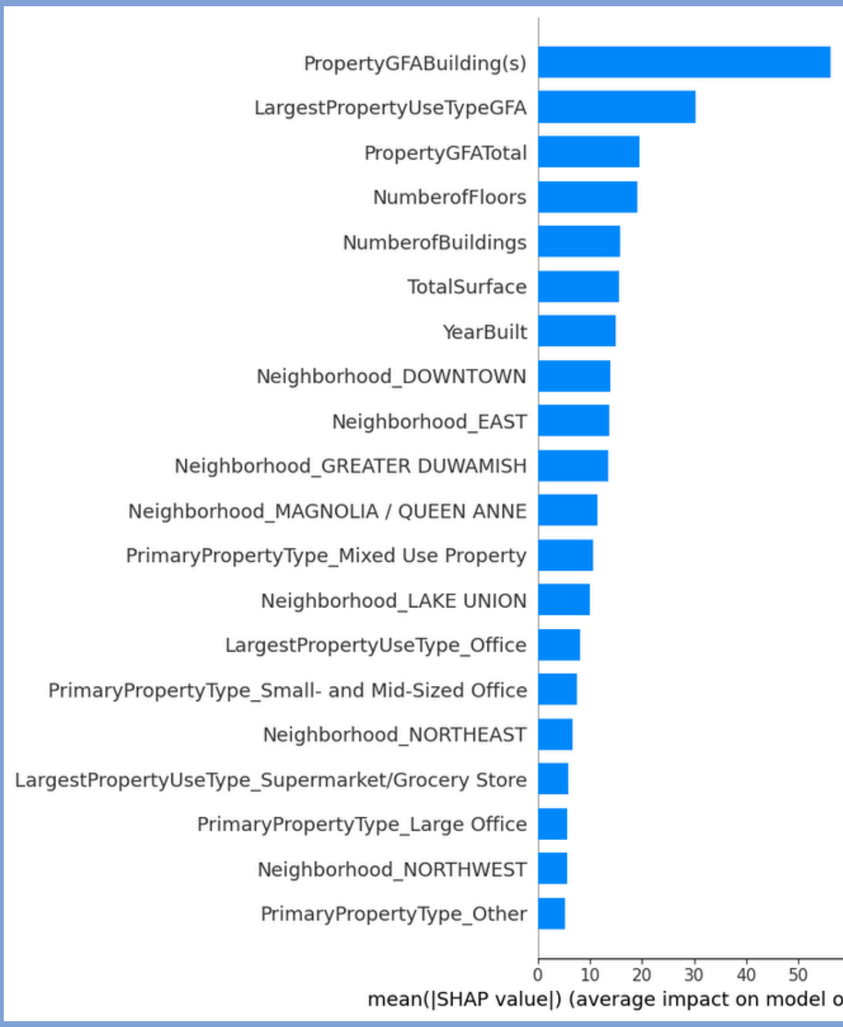
Meilleurs score des modèles de prédiction des émissions de CO2 (TotalGHGEmissions)



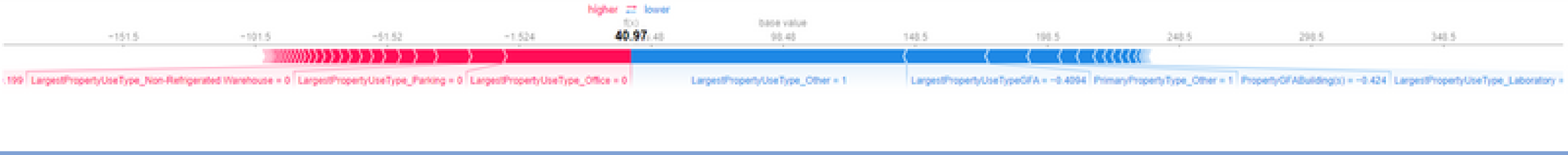


5. IMPORTANCE DES VARIABLES POUR LA PRÉDICTION D'ÉMISSIONS DE CO2

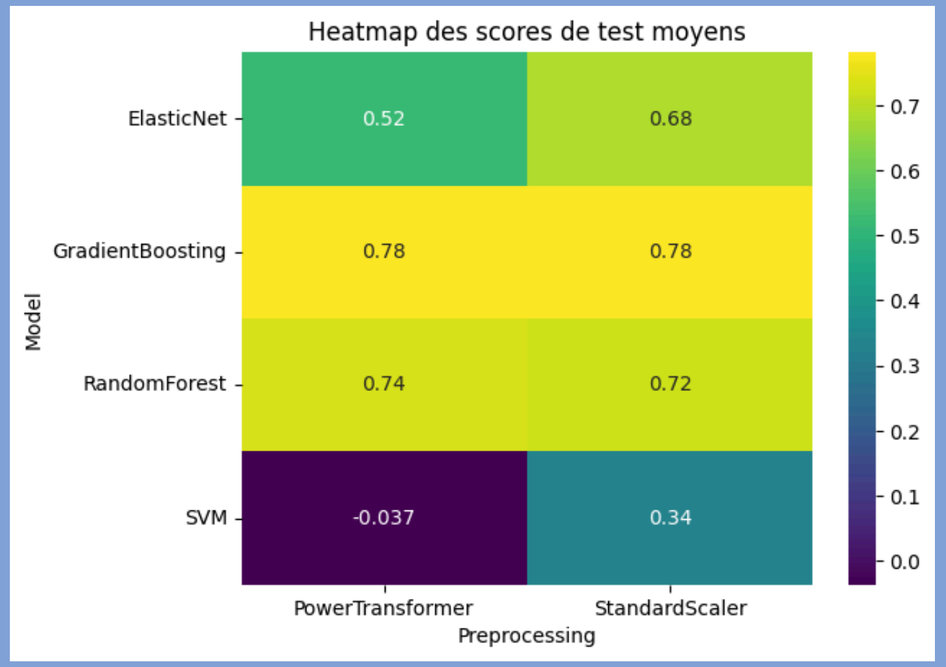
Importance globale

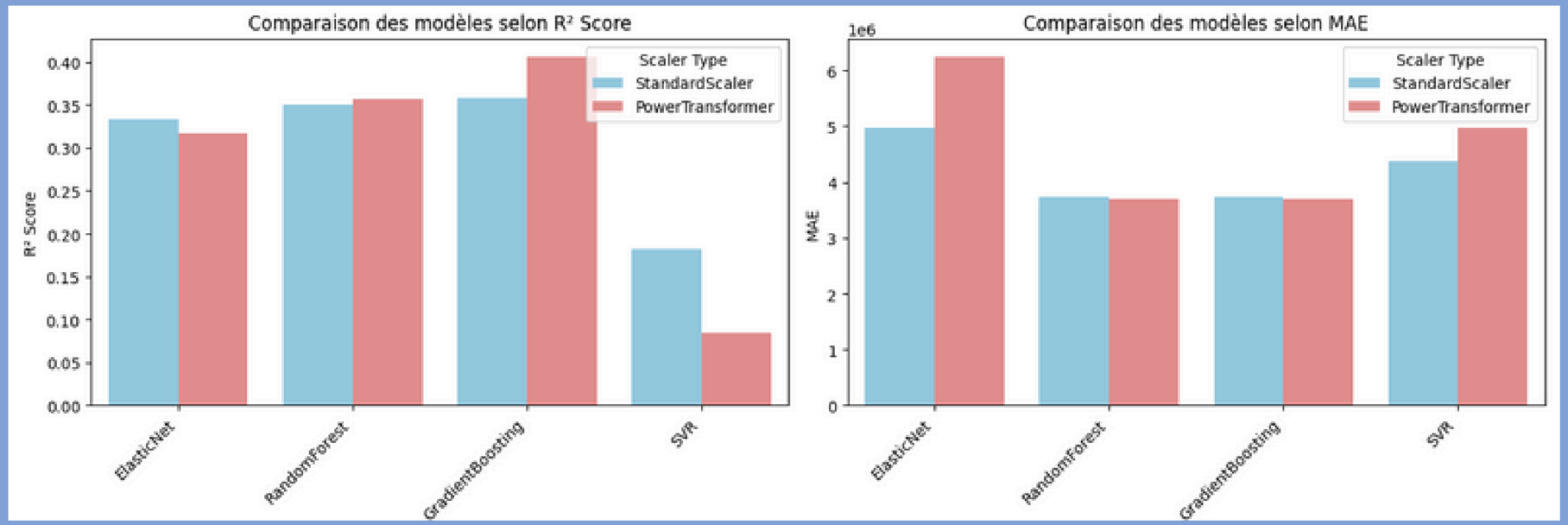


Importance locale



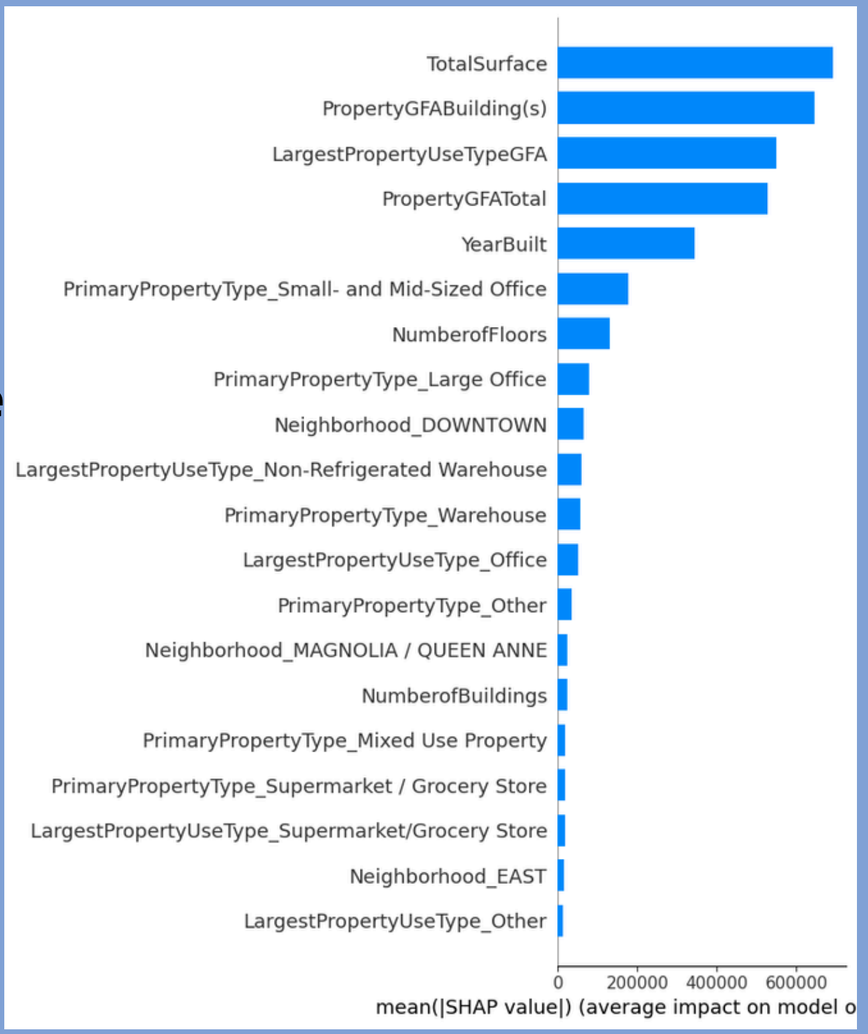
Meilleurs score des modèles de prédiction de la consommation d'Energie





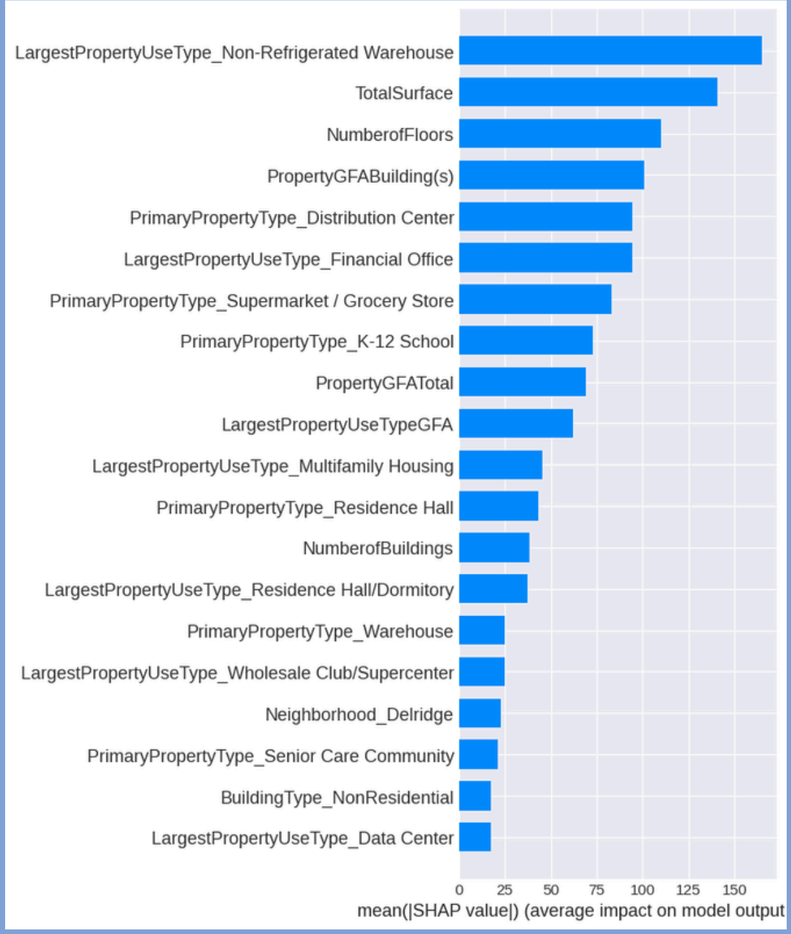
5. IMPORTANCE DES VARIABLES POUR LA PRÉDICTION DE LA CONSSOMATION D'ENERGIE

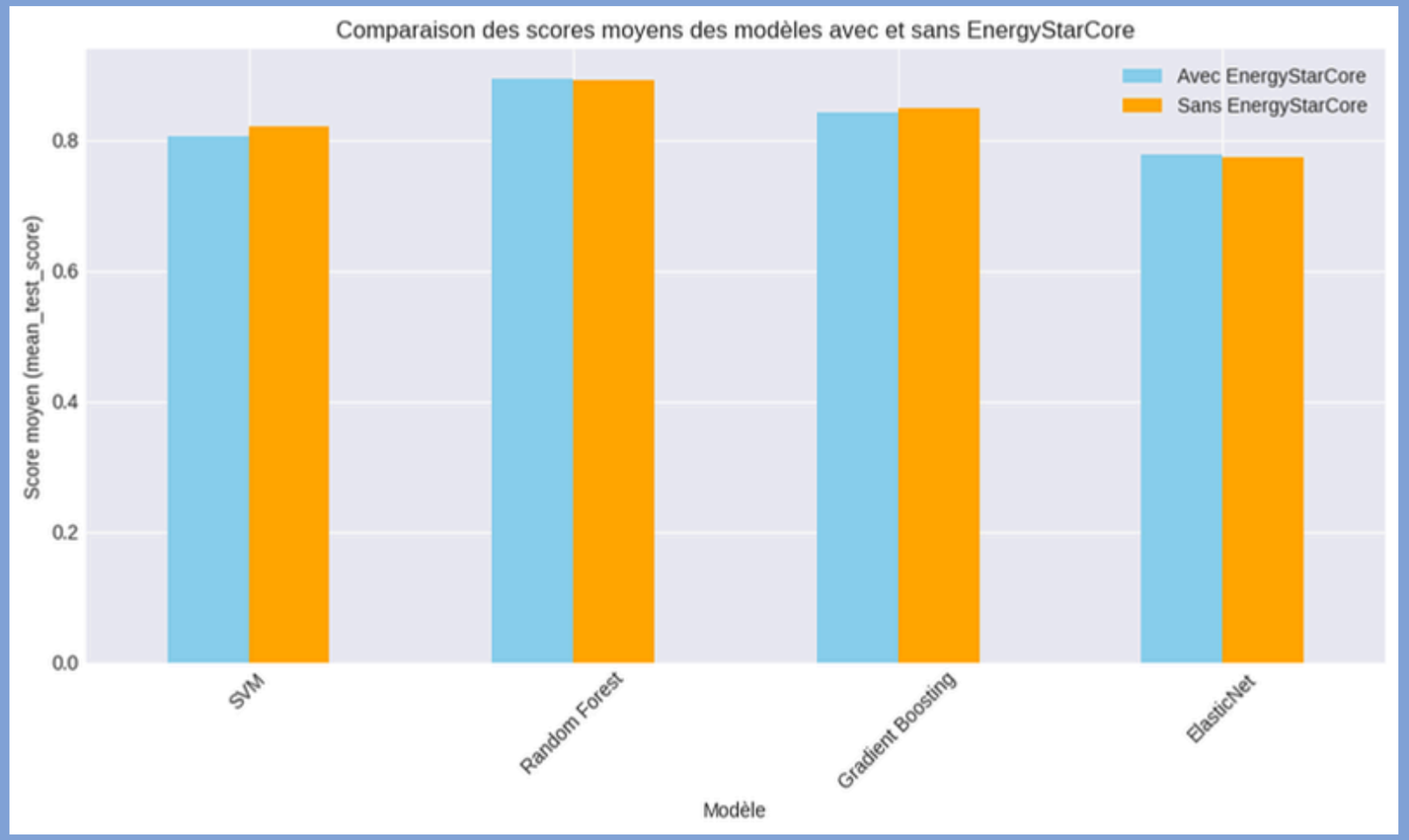
Importance globale



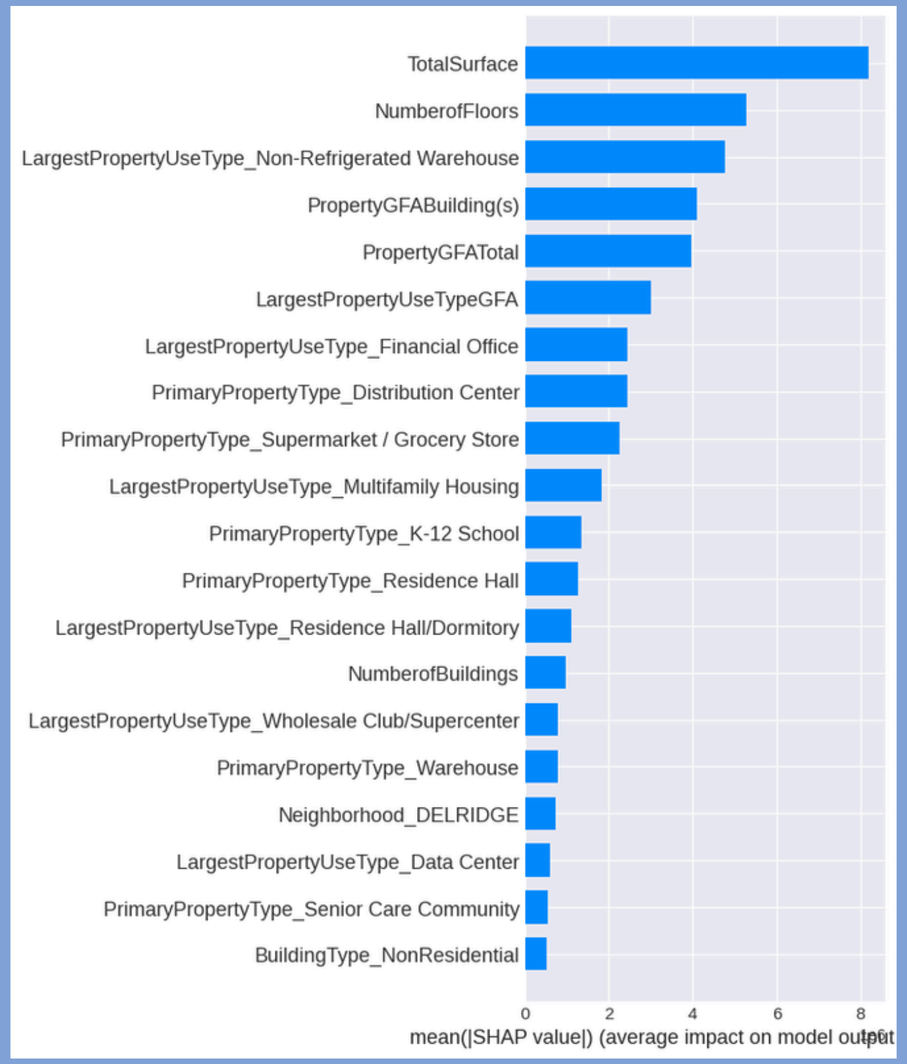
Importance locale

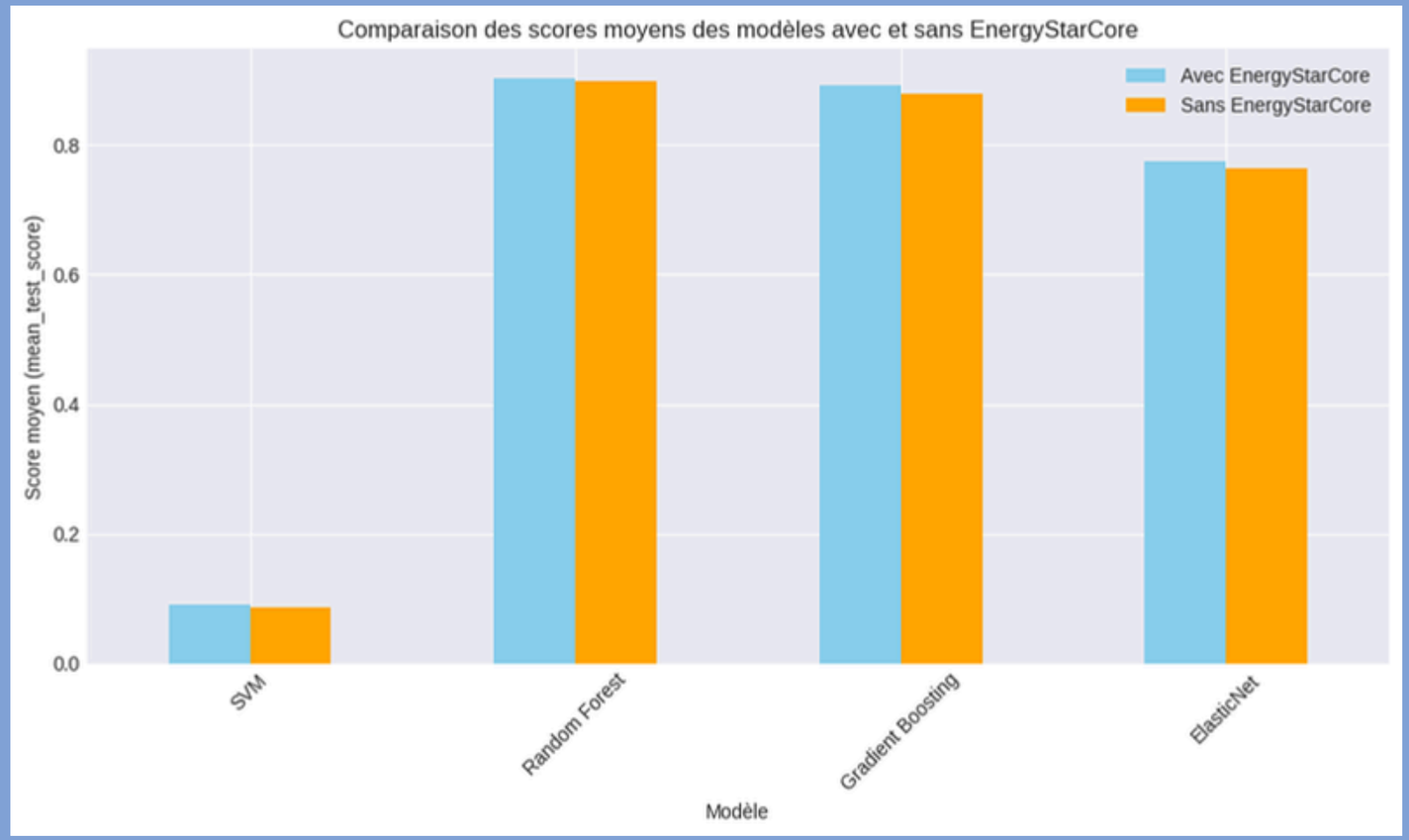
6. L'INTÉRÊT DE L'ENERGY STAR SCORE POUR LA PRÉDICTION D'ÉMISSIONS DE CO2





7. L'INTÉRÊT DE L'ENERGY STAR SCORE POUR LA PRÉDICTION DE LA CONSSOMATION D'ENERGIE





8. CONCLUSION

MERCI !