

A scenic view of the Seattle skyline during sunrise or sunset. The Space Needle is prominent on the left, and Mount Rainier is visible in the background on the right. The sky is clear and blue.

ALEXANDRE ROGUES  
PARCOURS DATA SCIENCE  
AVRIL 2024

# PROJET 4



# ANTICIPER LES BESOINS EN CONSOMMATION DE BATIMENTS

## FEATURE ENGINEERING MACHINE LEARNING



# CONTEXTE ET PROBLÉMATIQUE

- Seattle a pour objectif d'atteindre la neutralité carbone d'ici 2050.
- Les bâtiments non résidentiels sont cruciaux pour cette ambition en raison de leur consommation énergétique importante et de leurs émissions de CO<sub>2</sub>.
- Des relevés énergétiques annuels coûteux et chronophages. Une méthode prédictive basée sur des données existantes pourrait réduire ces coûts et accélérer les évaluations.
- Comment pouvons-nous utiliser les données structurelles des bâtiments pour prédire efficacement leur consommation d'énergie et leurs émissions de CO<sub>2</sub> sans relevés annuels supplémentaires ?



# OBJECTIFS



## # 1

Développer un modèle prédictif pour estimer la consommation d'énergie et les émissions de CO<sub>2</sub> des bâtiments non résidentiels à Seattle



## # 2

Évaluer l'efficacité de l'ENERGY STAR Score comme prédicteur dans le modèle

## # 3

Proposer une approche de modélisation qui minimise les coûts et optimise l'utilisation des données existantes



# VUE D'ENSEMBLE DES DONNÉES

SOURCES, TYPES, VOLUME



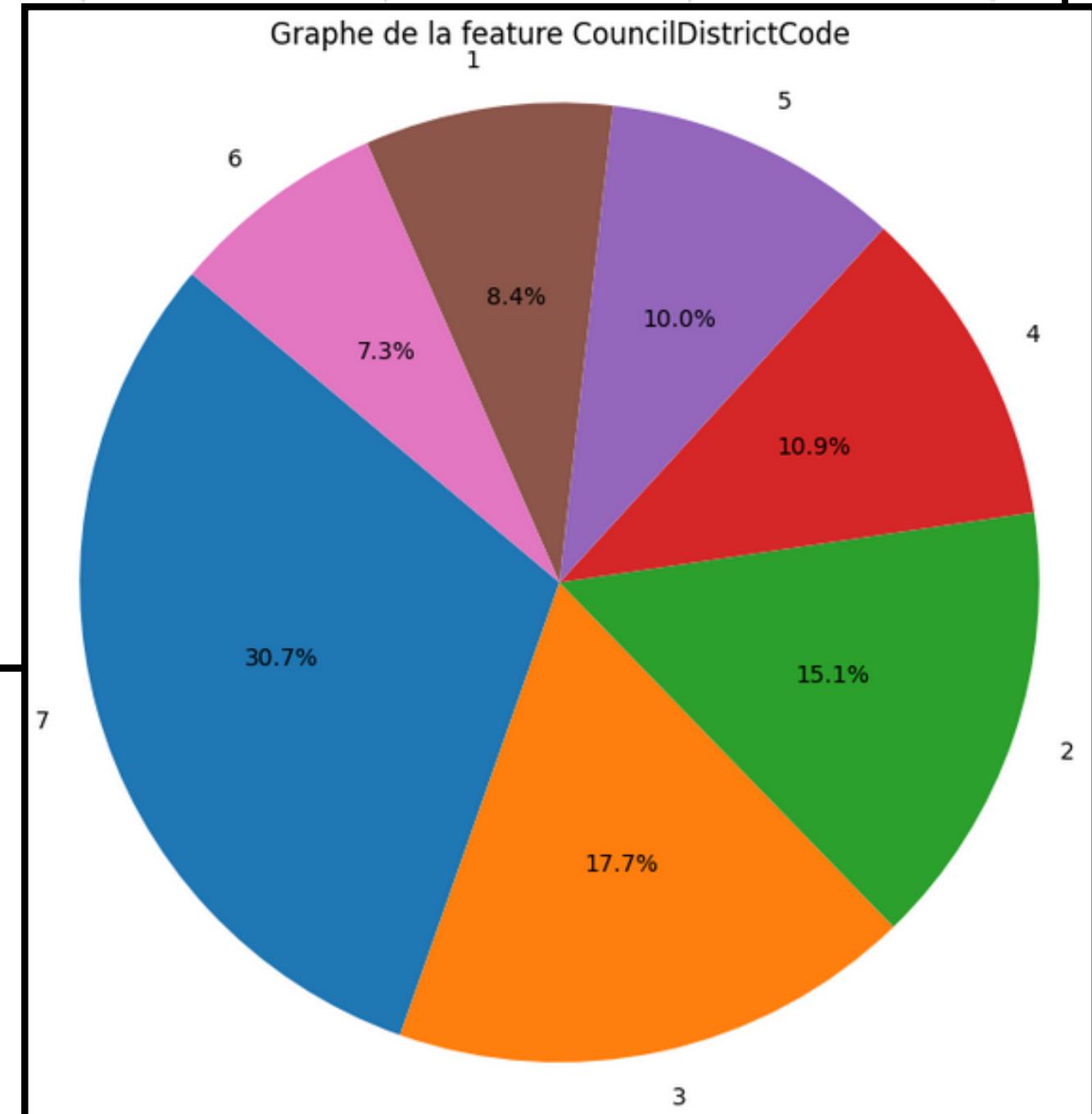
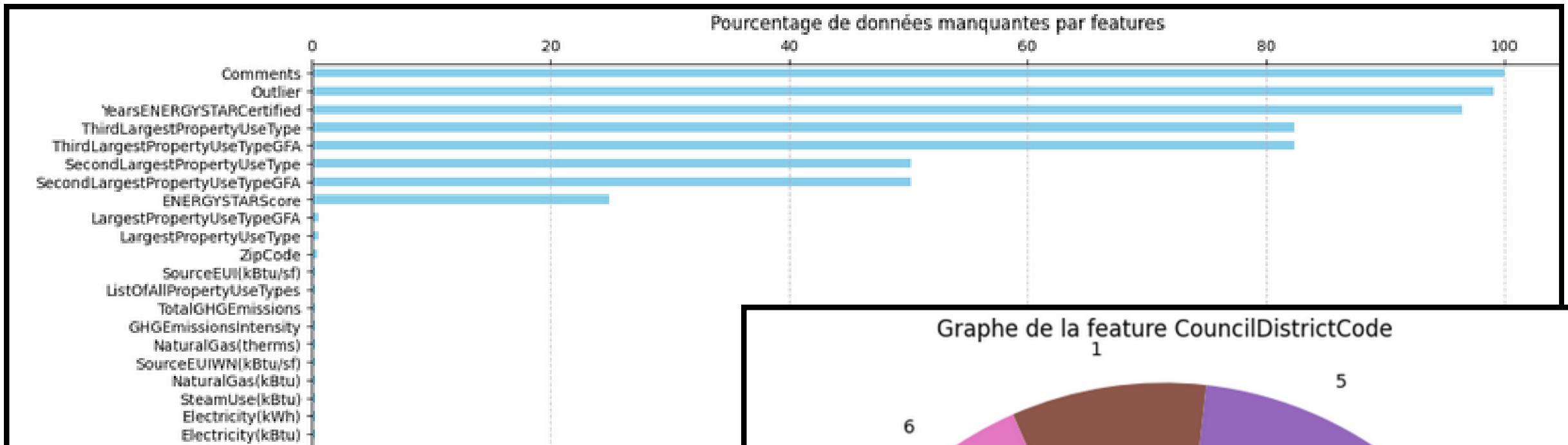
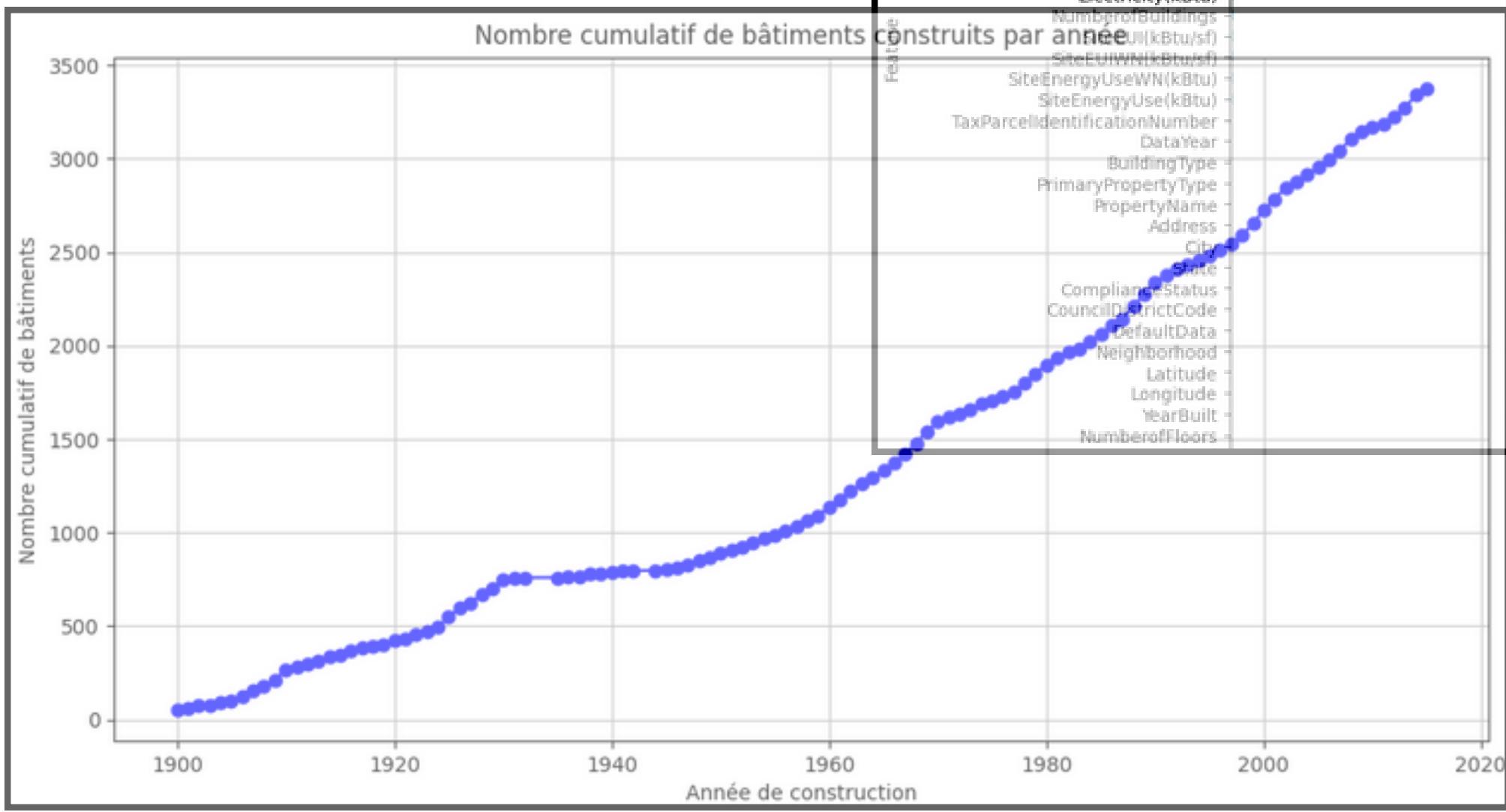
Des relevés effectués par les agents de la ville  
de Seattle en 2016 auprès des propriétaires

Différentes catégories de données :

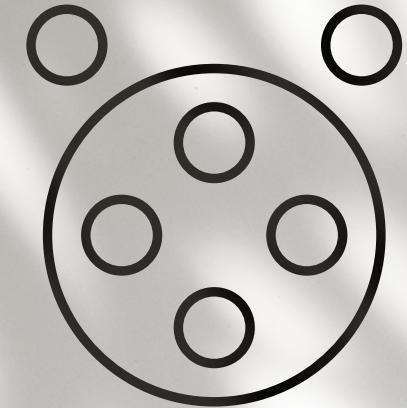
- surface, taille
- usages
- emplacement
- âge
- énergétiques
- émissions
- energystarscore
- identification
- qualification

Quasiment 3400 bâtiments  
46 caractéristiques par bâtiments

# DESCRIPTION



# NETTOYAGE



Suppression des individus outliers high and low  
default data true, et non compliant status



Sélection des individus BuildingType =/ Multifamily \*\*

3500 → 1500

# 1ST FEATURE ENGINEERING

- Proratisation type energies
- YearBuilt => age décades
- CouncilDistrictCode
- Neighboordhood
- Proratisation gfa parking/building

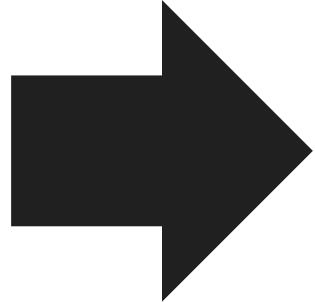


# 2ND FEATURE ENGINEERING

## Catégorisation des types d'usages

Office  
Financial Office  
Bank Branch  
Medical Office  
Non-Refrigerated Warehouse  
Refrigerated Warehouse  
Distribution Center  
Self-Storage Facility  
Retail Store  
Supermarket/Grocery Store  
Strip Mall  
Shopping Mall  
Automobile Dealership  
Wholesale Club/Supercenter  
K-12 School  
College/University  
Adult Education  
Other - Education  
Pre-school/Daycare  
Hospital (General Medical & Surgical)  
Senior Care Community  
Residential Care Facility  
Urgent Care/Clinic/Other Outpatient  
Other/Specialty Hospital

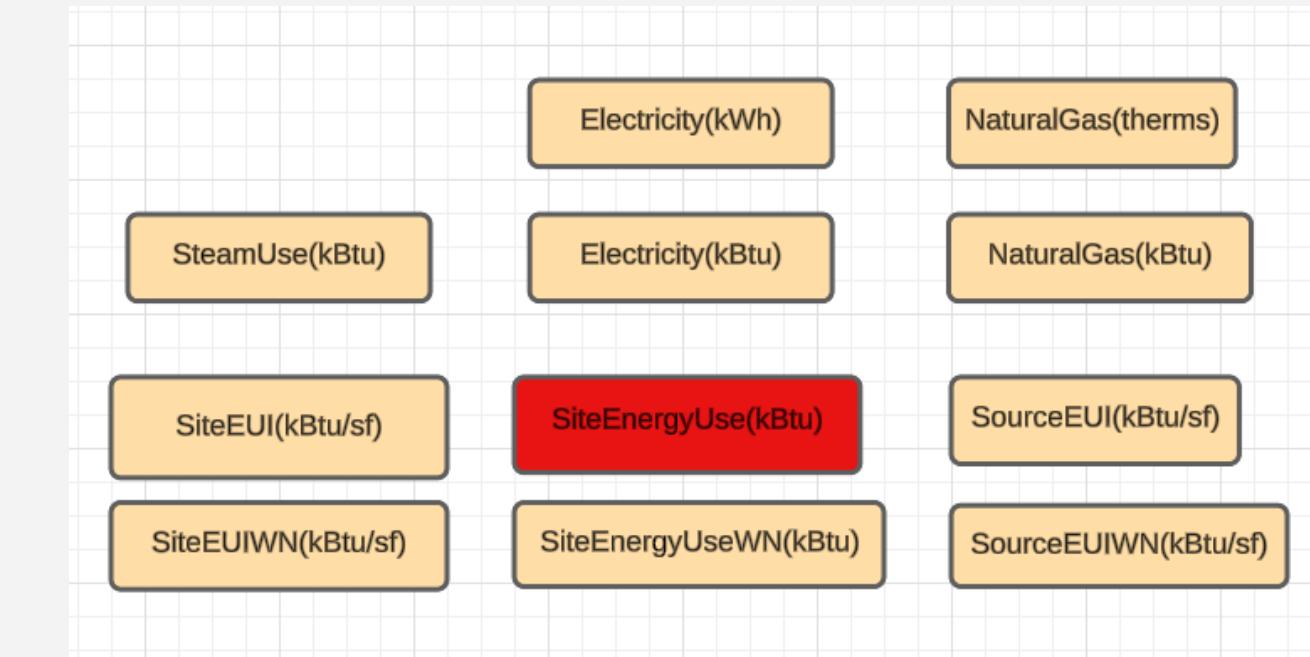
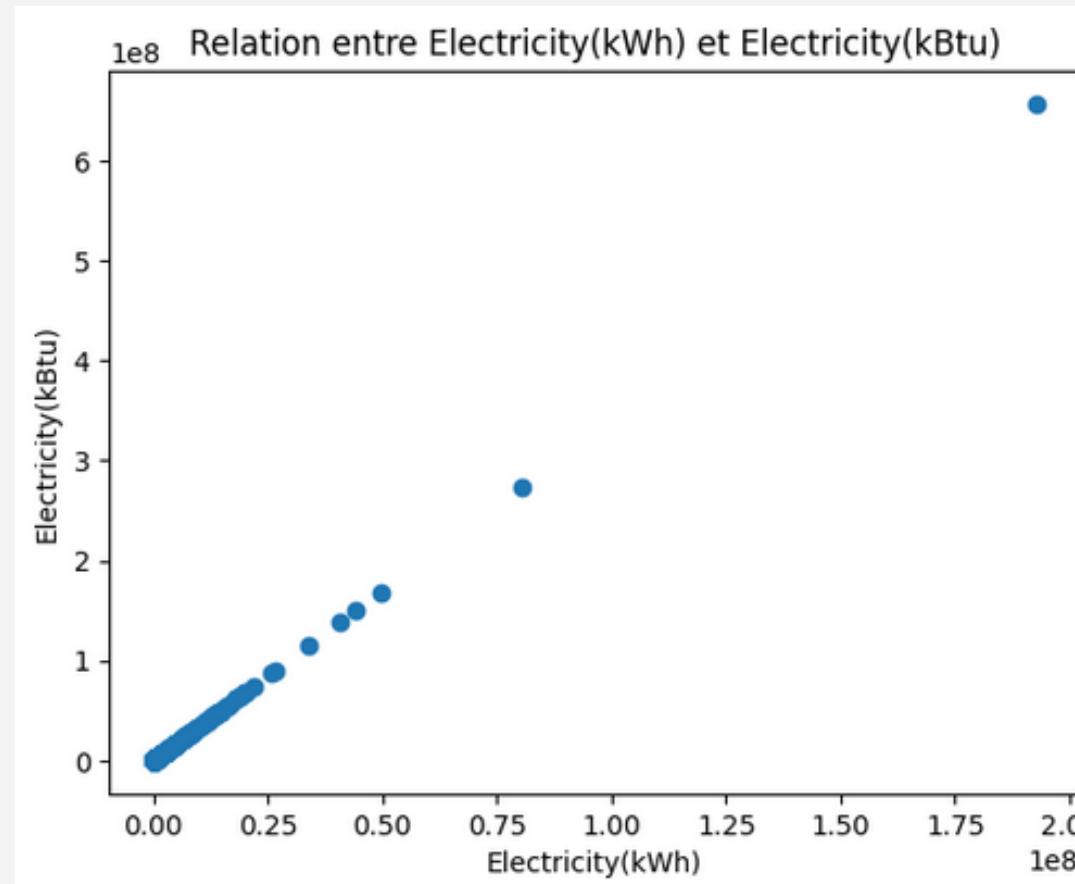
Hotel  
Residence Hall/Dormitory  
Other - Lodging/Residential  
Fitness Center/Health Club/Gym  
Museum  
Performing Arts  
Movie Theater  
Manufacturing/Industrial Plant  
Laboratory  
Data Center  
Repair Services (Vehicle, Shoe, Locksmith, etc)  
Worship Facility  
Social/Meeting Hall  
Police Station  
Fire Station  
Courthouse  
Library  
Prison/Incarceration  
Restaurant  
Other - Recreation  
Other - Entertainment/Public Assembly  
Other - Services  
Personal Services (Health/Beauty, Dry Cleaning, etc)  
Other - Public Services  
Other - Utility  
Parking



- Admin&Offices
- Warehouses
- Retail
- Education
- Health
- Leisure
- Industrial
- Public
- Other
- Parking

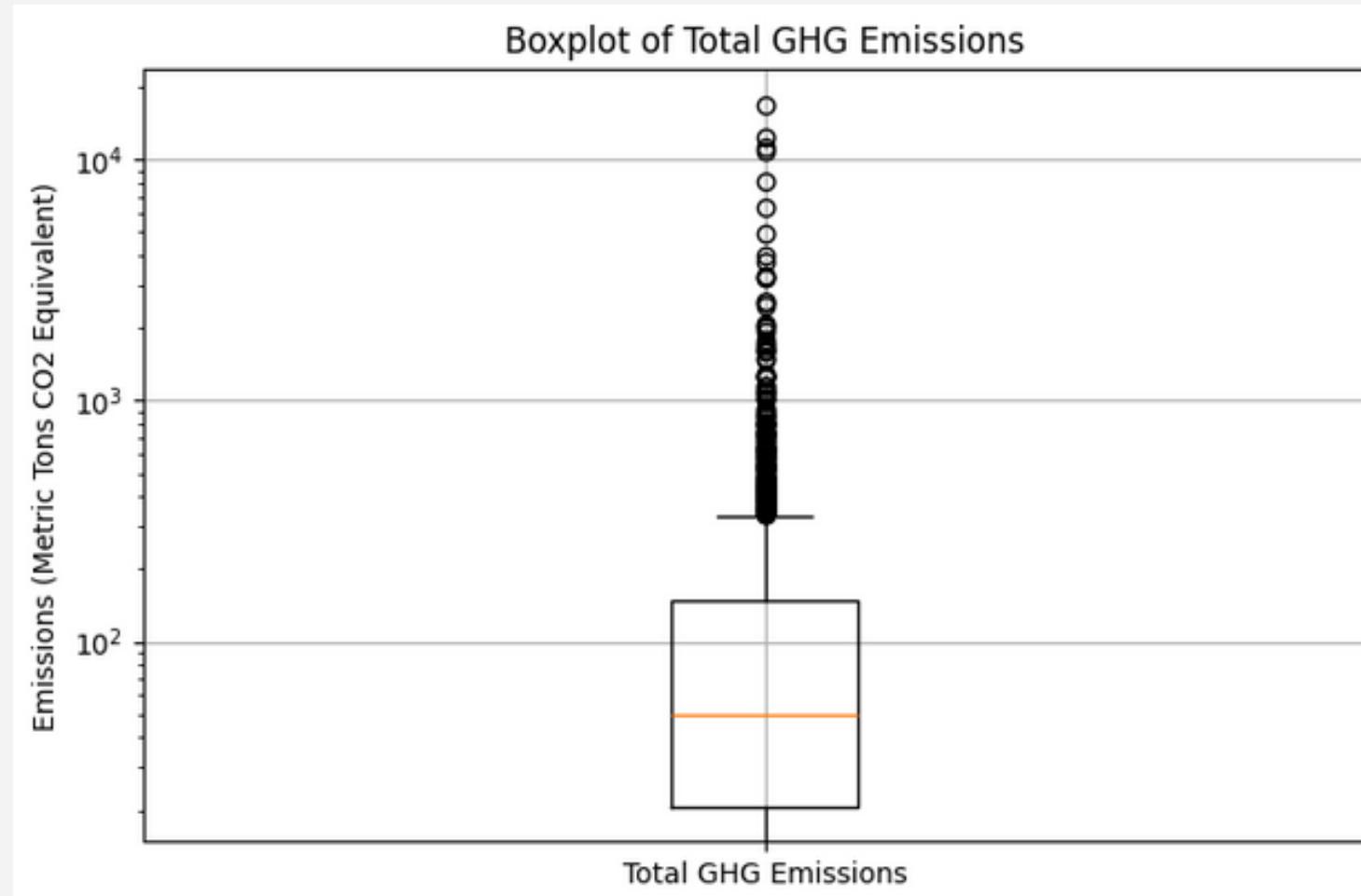
# LES CIBLES

Consommation d'énergie



# LES CIBLES

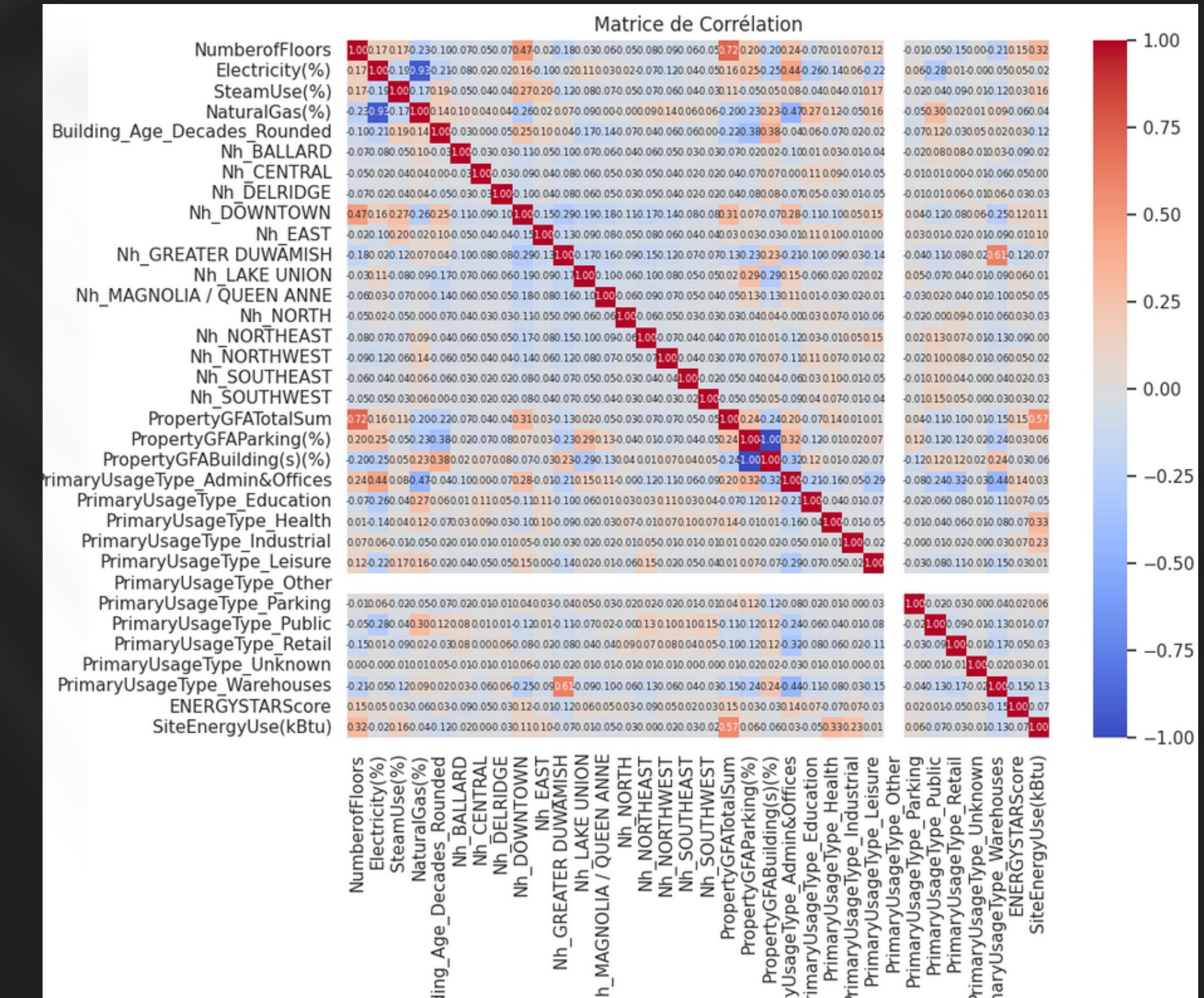
Gaz à effet de serre



passage au logarithme

# Vérification de la non corrélation des variables avec les cibles

(ci-contre SiteEnergyUse)



## **Division Train/Test**

Séparation des données en 80% pour l'entraînement, 20% pour les tests

## **Standard Scaler**

Normalise les features : moyenne à 0, écart type à 1, équilibre l'importance des variables.

## **Transformation Logarithmique des Cibles**

Applique le logarithme aux cibles pour réduire l'asymétrie et stabiliser la variance.

## **Cross-Validation**

Utilise plusieurs sous-ensembles pour entraîner et valider le modèle, assure la stabilité.

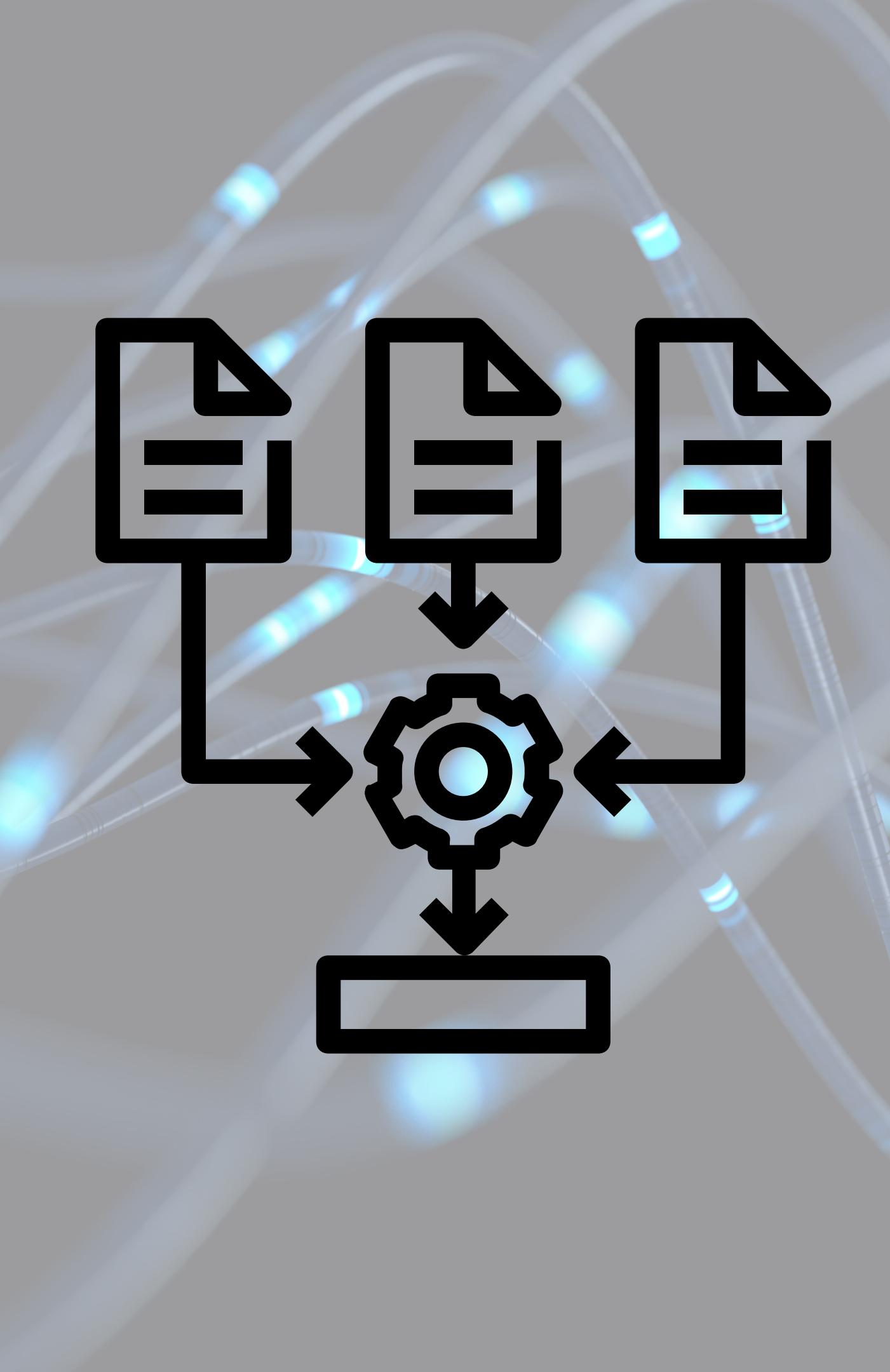
## **Grid Search**

Optimise les hyperparamètres par validation croisée, cherche la meilleure combinaison possible.

---

# **PREPARATION**

---



# MODELISATION

- ElasticNet : Combinaison de deux techniques pour mieux gérer les données et choisir les informations les plus importantes.
- SVM (Support Vector Machines) : Trouve la meilleure ligne ou surface qui sépare ou ajuste les données pour prédire les résultats.
- Gradient Boosting : Améliore les prédictions en construisant plusieurs modèles simples successivement, chacun corrigéant les erreurs du précédent.
- RandomForest : Utilise plusieurs arbres de décision pour obtenir des prédictions plus précises et stables que celles d'un seul arbre.

Les métriques évaluent la performance des modèles, guident les ajustements et permettent de comparer objectivement différents modèles.

### **R<sup>2</sup> (Coefficient de Détermination)**

*Avantages* : Mesure la proportion de la variance expliquée par le modèle, facile à interpréter.

*Inconvénients* : Peut être trompeur avec des données non-linéaires ou des échantillons de taille réduite.

### **MAE (Erreur Absolue Moyenne)**

*Avantages* : Mesure l'erreur moyenne en unités originales, robuste aux outliers.

*Inconvénients* : Moins sensible aux grands écarts que RMSE, difficile à contextualiser sans référence.

### **RMSE (Erreur Quadratique Moyenne)**

*Avantages* : Sensible aux grands écarts, couramment utilisée, donne un poids plus lourd aux grandes erreurs.

*Inconvénients* : Peut être influencée par des valeurs aberrantes.

# **METRIQUES**

# ENERGYSTARSCORE

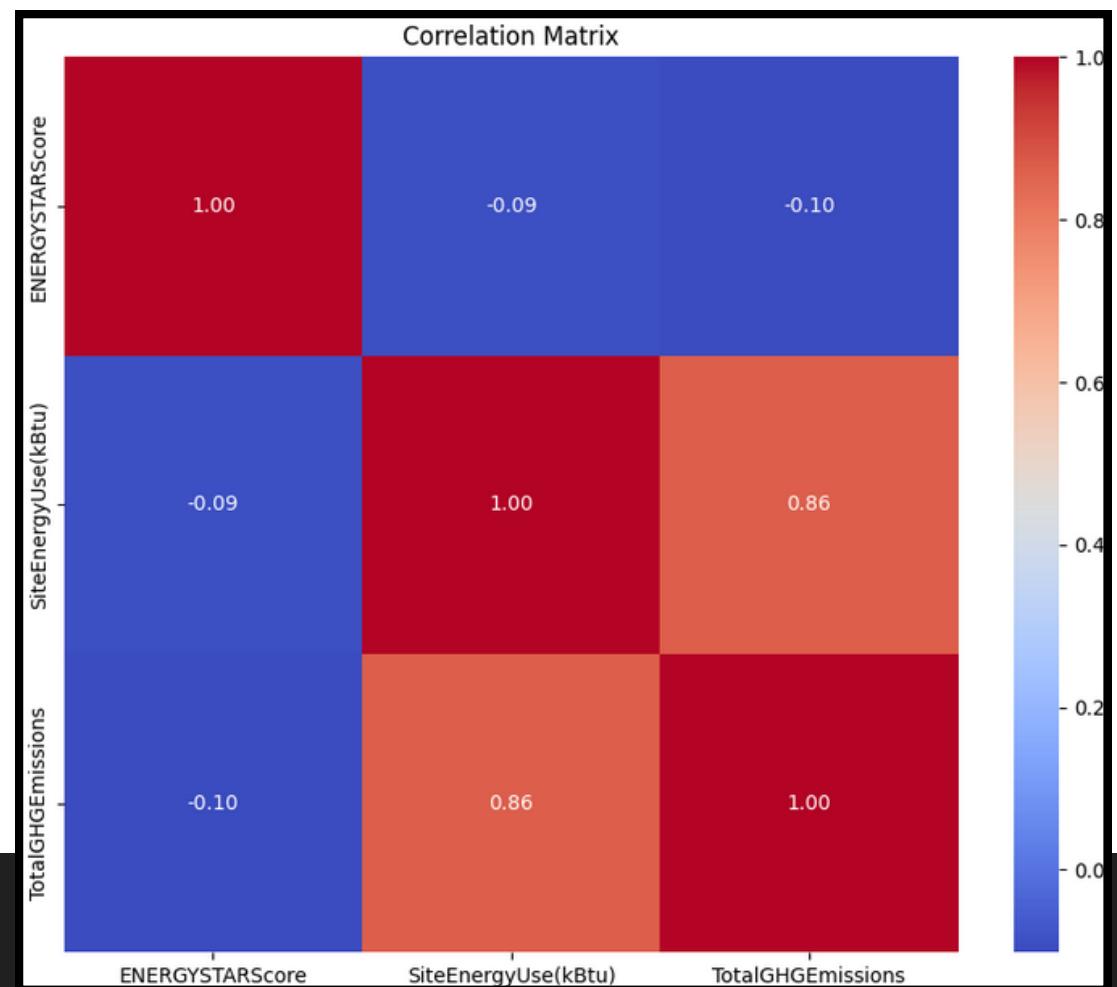
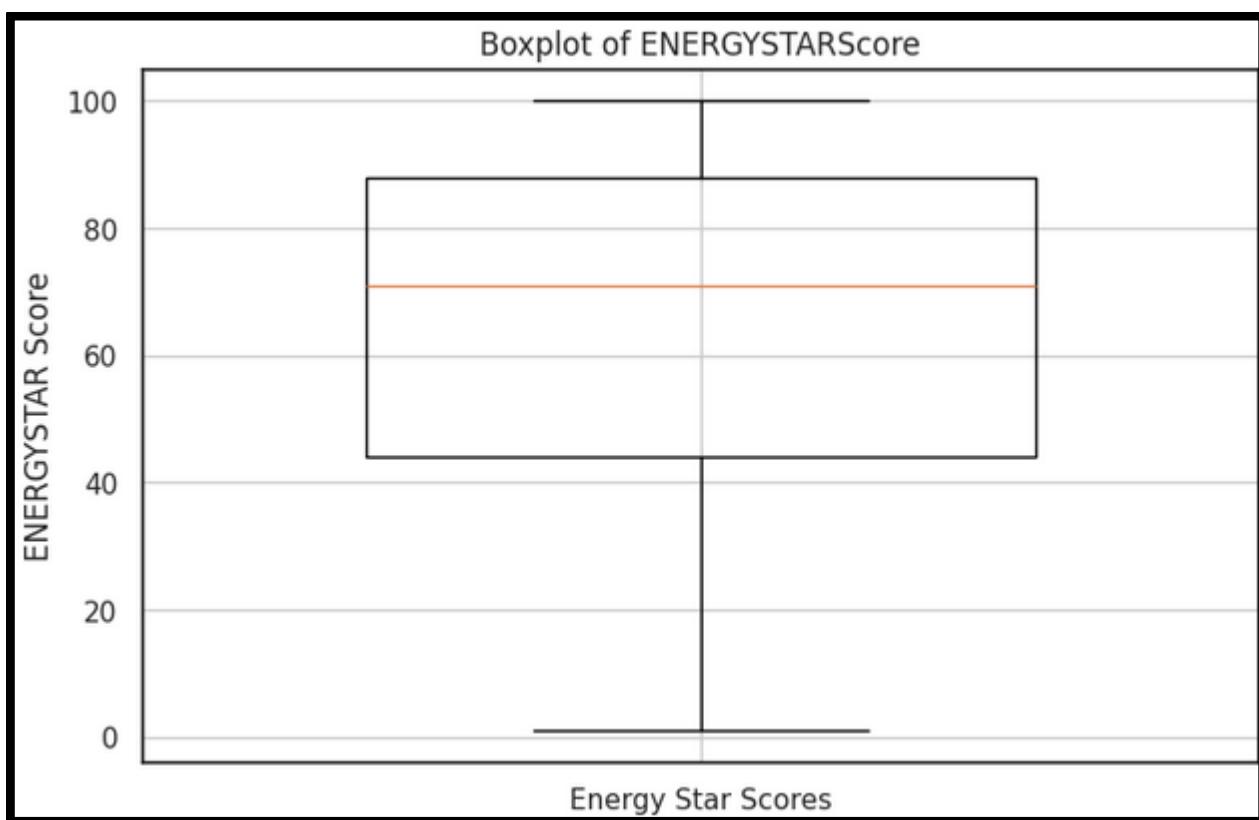
**Définition :** Indicateur de performance énergétique des bâtiments, échelle de 1 à 100.

**Benchmark :** Compare la consommation énergétique à celle de bâtiments similaires nationalement.

**Moyenne :** Score de 50 indique une efficacité moyenne.

**Certification :** Bâtiments avec score  $\geq 75$  peuvent obtenir la certification Energy Star.

**Impact :** Aide à identifier les économies potentielles et améliorer les pratiques énergétiques.



## RESULTATS : Consommation d'énergie

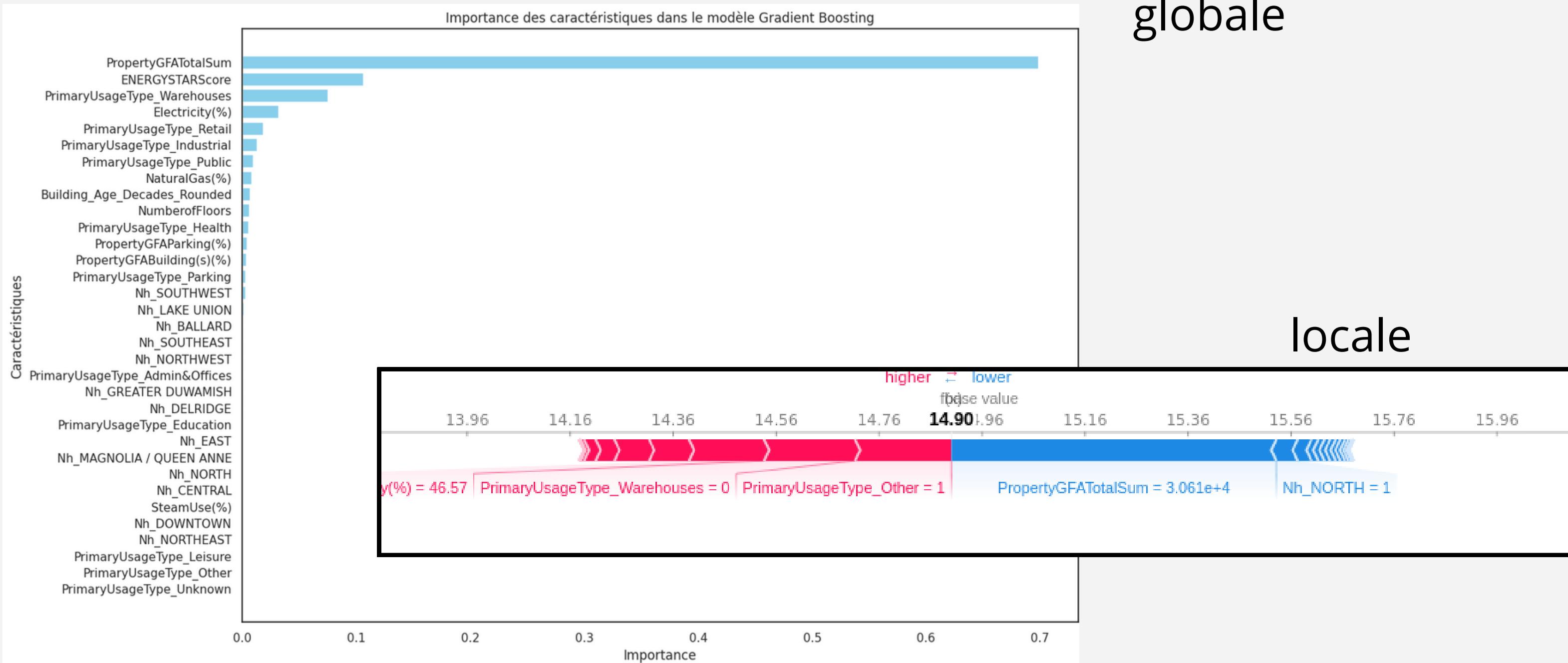
	<b>EN</b>	<b>RF</b>	<b>GB</b>	<b>SVM</b>
1er Feature Engineering	0.06	0.60	0.63	0.53
2nd Feature Engineering	0.06	0.71	0.73	0.56
ENERGYSTAR Score	0.19	0.82	0.86	0.72

## **RESULTATS : Gaz à effet de serre**

	<b>EN</b>	<b>RF</b>	<b>GB</b>	<b>SVM</b>
1er Feature Engineering	0.09	0.70	0.72	0.63
2nd Feature Engineering	0.10	0.77	0.79	0.66
ENERGYSTAR Score	0.14	0.84	0.88	0.77

# FEATURE IMPORTANCE

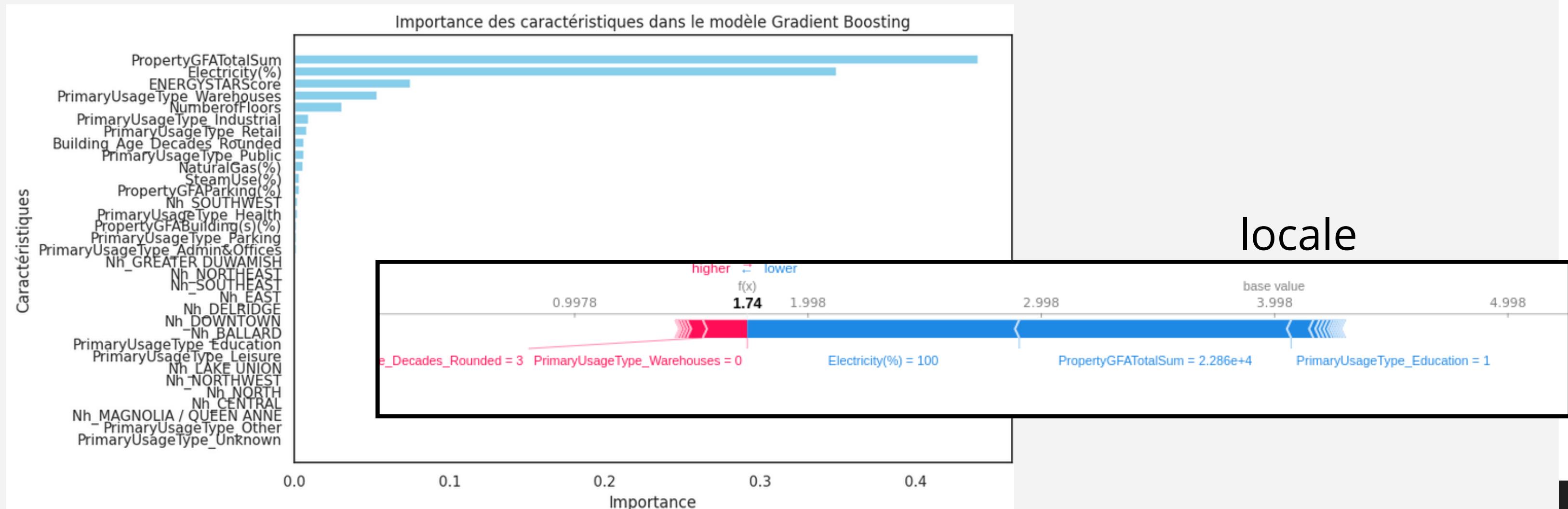
Consommation d'énergie



# FEATURE IMPORTANCE

Gaz à effet de serre

globale



# REMARQUE

Lors de l'intégration de la feature EnergyStarScore dans notre ensemble de données d'entraînement, environ 500 individus sur 1500 ont été perdus, en raison de valeurs manquantes.

Cette réduction de la taille de l'échantillon peut influencer les performances des modèles.

Il est donc important de prendre en compte cet impact lors de la comparaison des résultats entre différents modèles

---

**MERCI**

---

