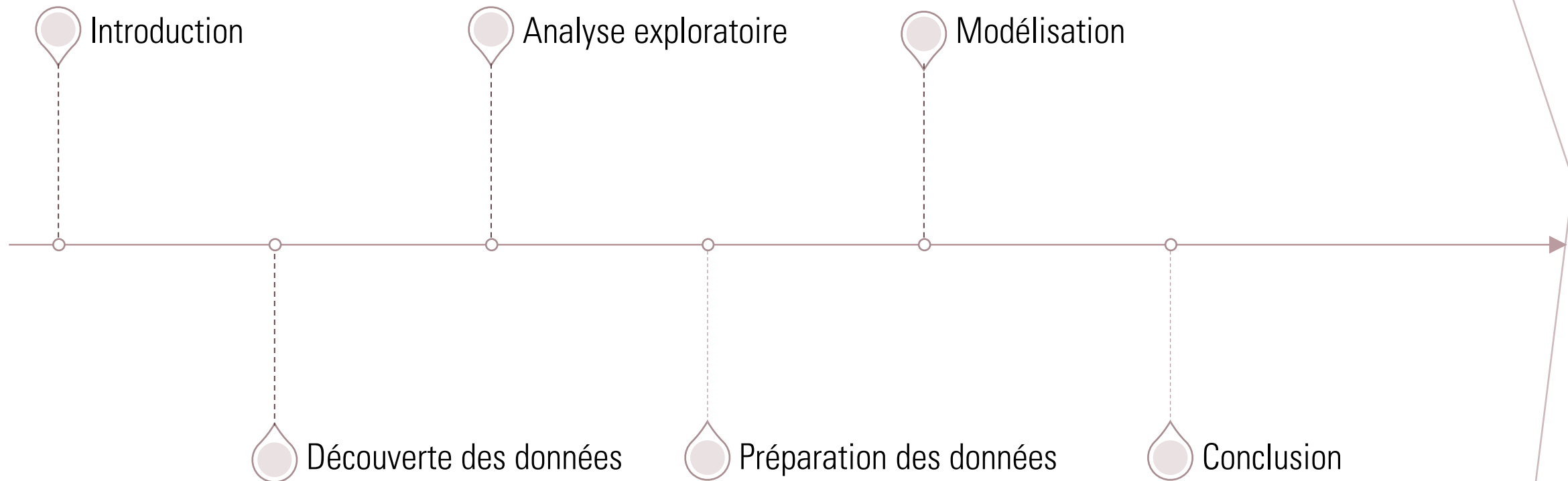


# PROJET N°3

ANTICIPEZ LES  
BESOINS EN  
CONSOMMATION DE  
BÂTIMENTS DE LA  
VILLE DE SEATTLE

FAYZ EL RAZAZ

# *PLAN DE LA PRÉSENTATION*







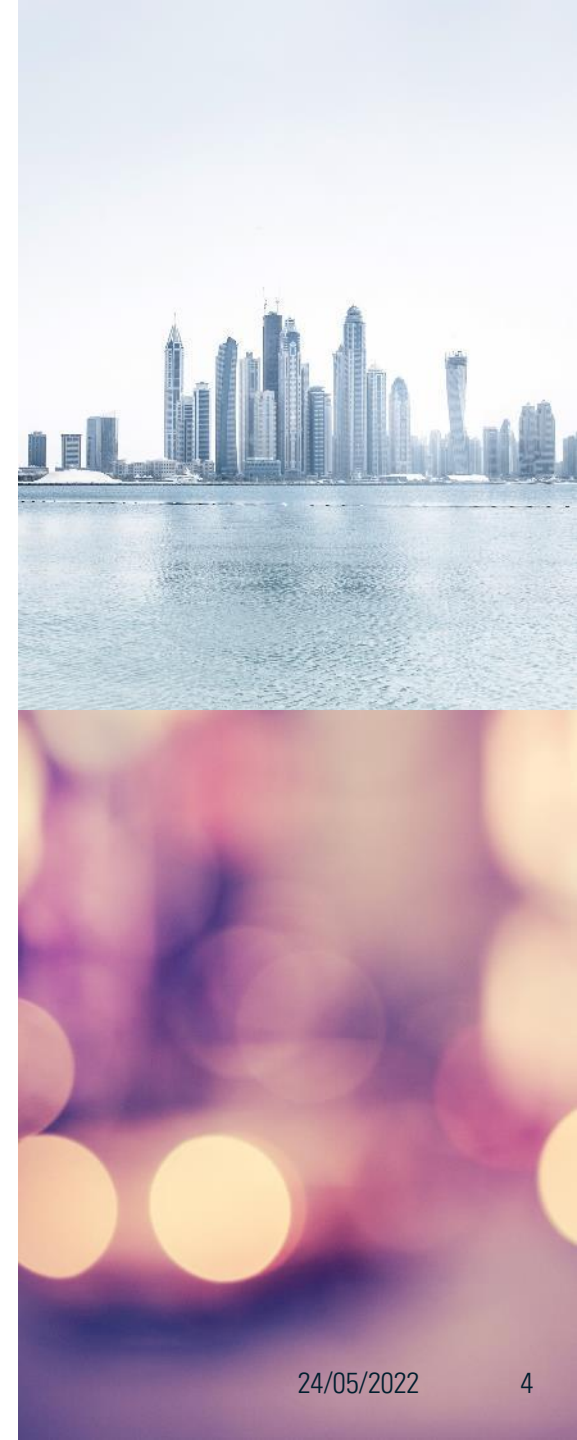
# *INTRODUCTION*

## Contexte

- Travail pour la ville de Seattle
- Objectif : ville neutre en émissions de carbone en 2050
- Etude sur la consommation et les émissions des bâtiments non destinés à l'habitation
- Données disponibles : relevés effectués par les agents de la ville
- Evaluation de l'ENERGY STAR Score

# *DÉCOUVERTE DES DONNÉES*

- Présentation des datasets
- Data cleaning
- Préparation à l'analyse exploratoire



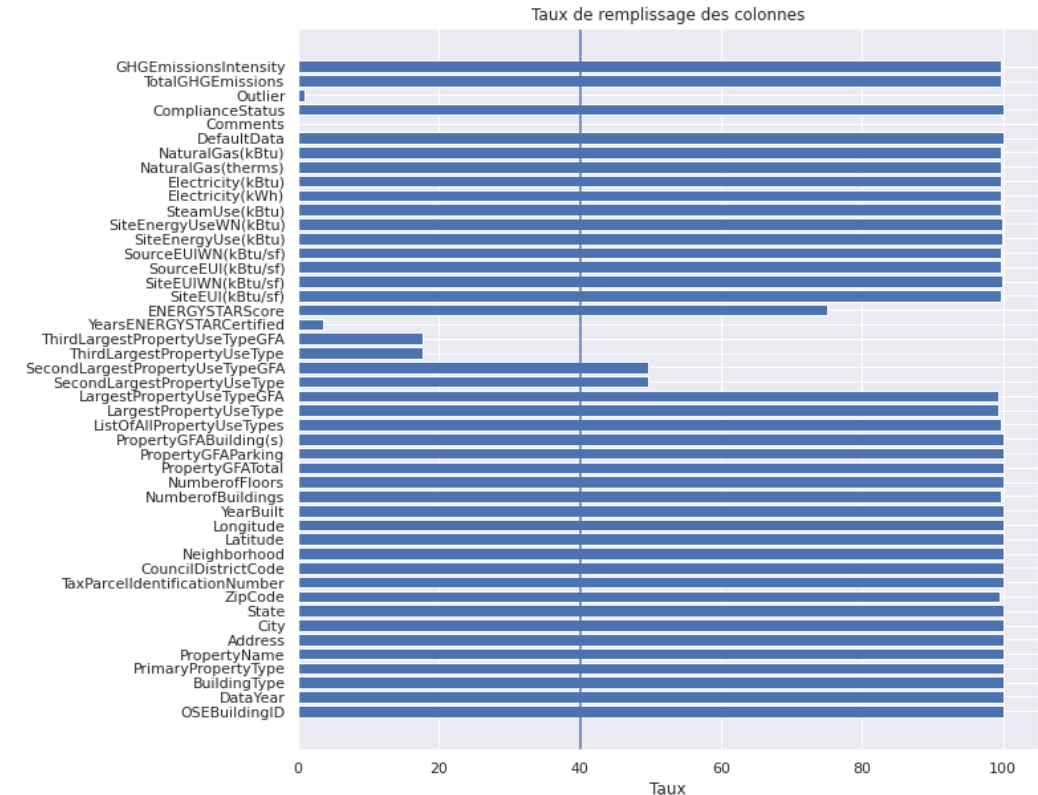
# PRÉSENTATION DES DATASETS

Deux datasets initiaux : données de 2015 et 2016

- pour 2015 : 3341 lignes 47 colonnes
- pour 2016 : 3377 lignes 46 colonnes

OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	TaxParcelIdentificationNumber	Location	CouncilDistrictCode	Neighborhood
0	1	2015	NonResidential	Hotel	MAYFLOWER PARK HOTEL	659000030	'[latitude': '47.61219025', 'longitude': '-122...	7 DOWNTOW
1	2	2015	NonResidential	Hotel	PARAMOUNT HOTEL	659000220	'[latitude': '47.61310583', 'longitude': '-122...	7 DOWNTOW
2	3	2015	NonResidential	Hotel	WESTIN HOTEL	659000475	'[latitude': '47.61334897', 'longitude': '-122...	7 DOWNTOW
3	5	2015	NonResidential	Hotel	HOTEL MAX	659000640	'[latitude': '47.61421585', 'longitude': '-122...	7 DOWNTOW
4	8	2015	NonResidential	Hotel	WARWICK SEATTLE HOTEL	659000970	'[latitude': '47.6137544', 'longitude': '-122....	7 DOWNTOW

Premier cleaning



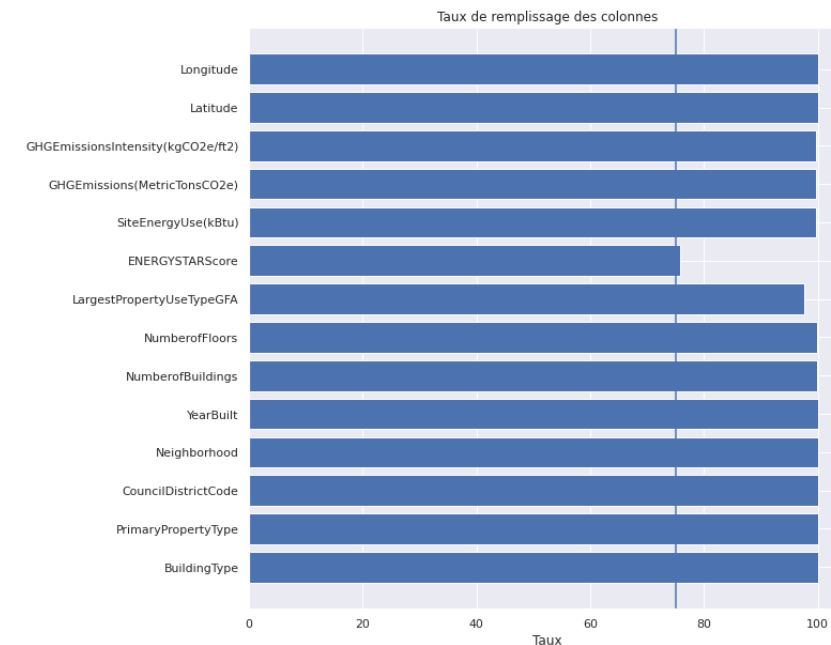
# DATA CLEANING

Suppression des features non suffisamment remplies

Suppression des features pour éviter le data leakage

Uniformisation des noms de features

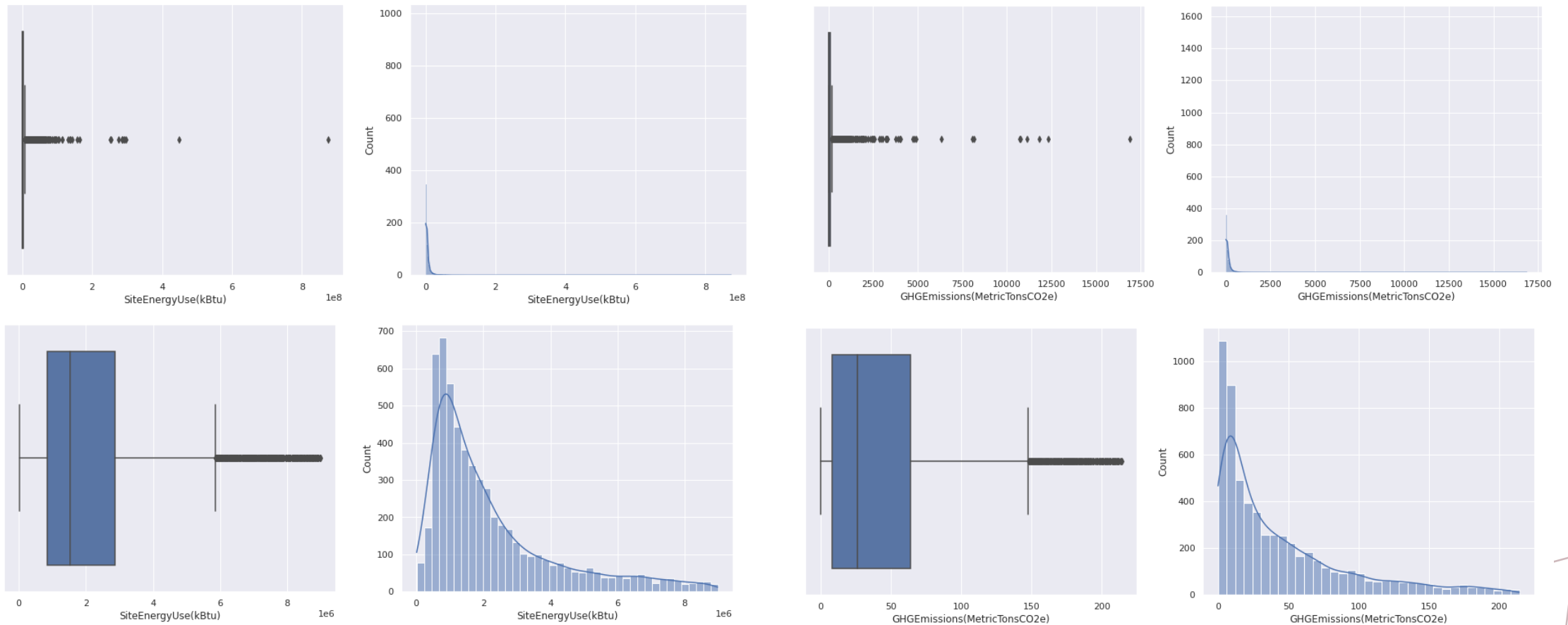
Concaténation de tables sur un maximum de features





# DATA CLEANING

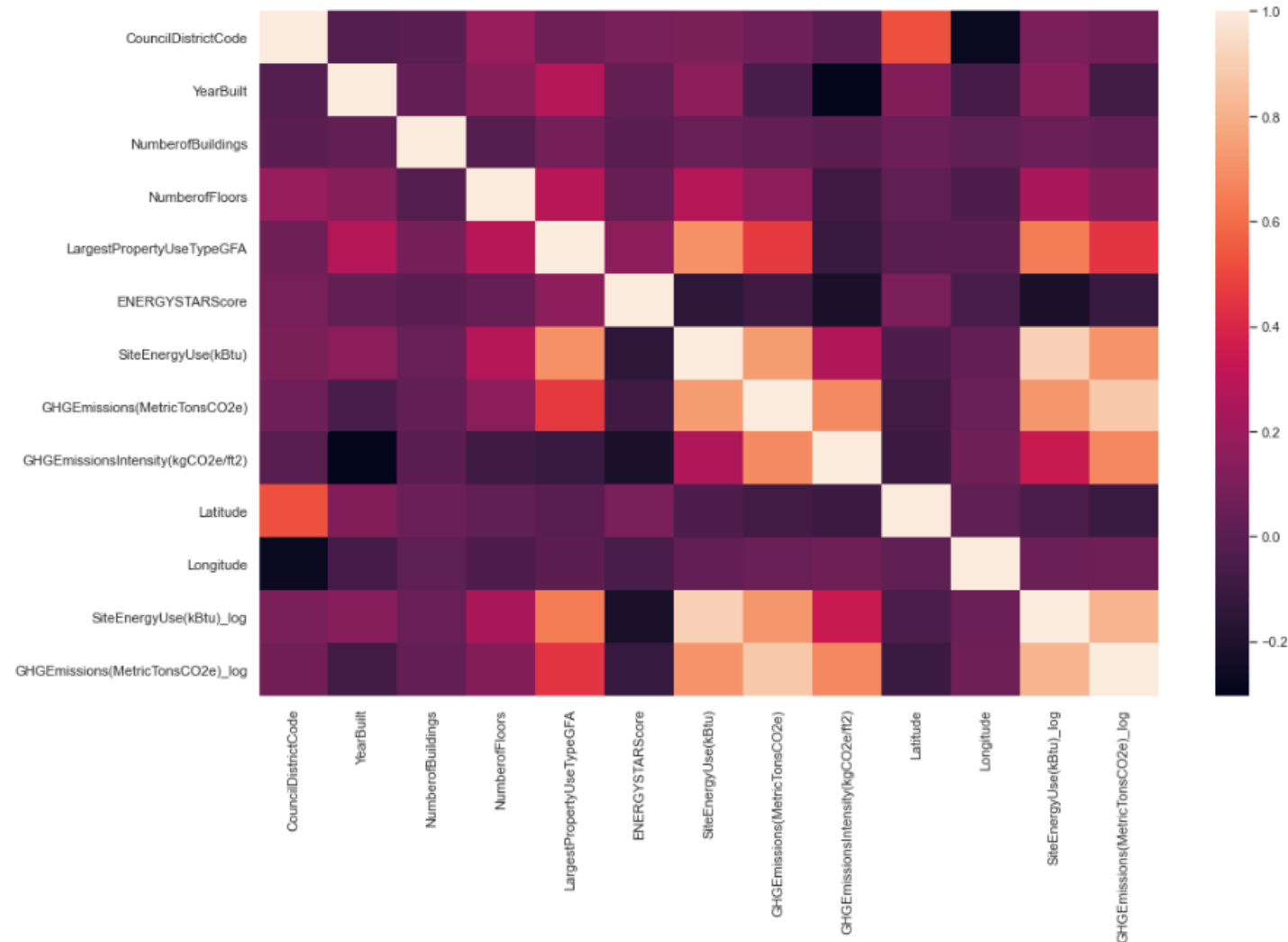
## Traitement des outliers



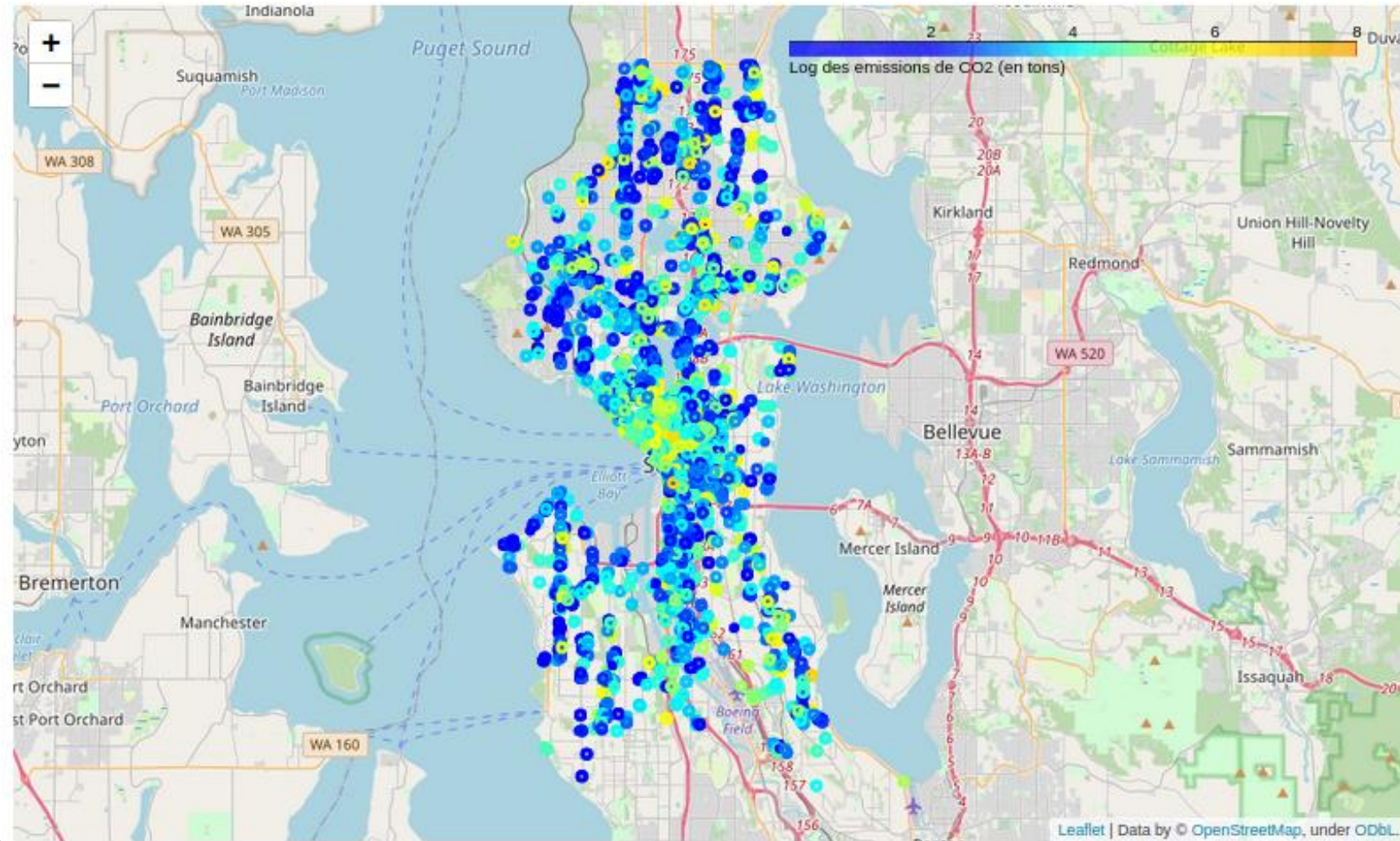
# *ANALYSE EXPLORATOIRE*



# ANALYSE EXPLORATOIRE

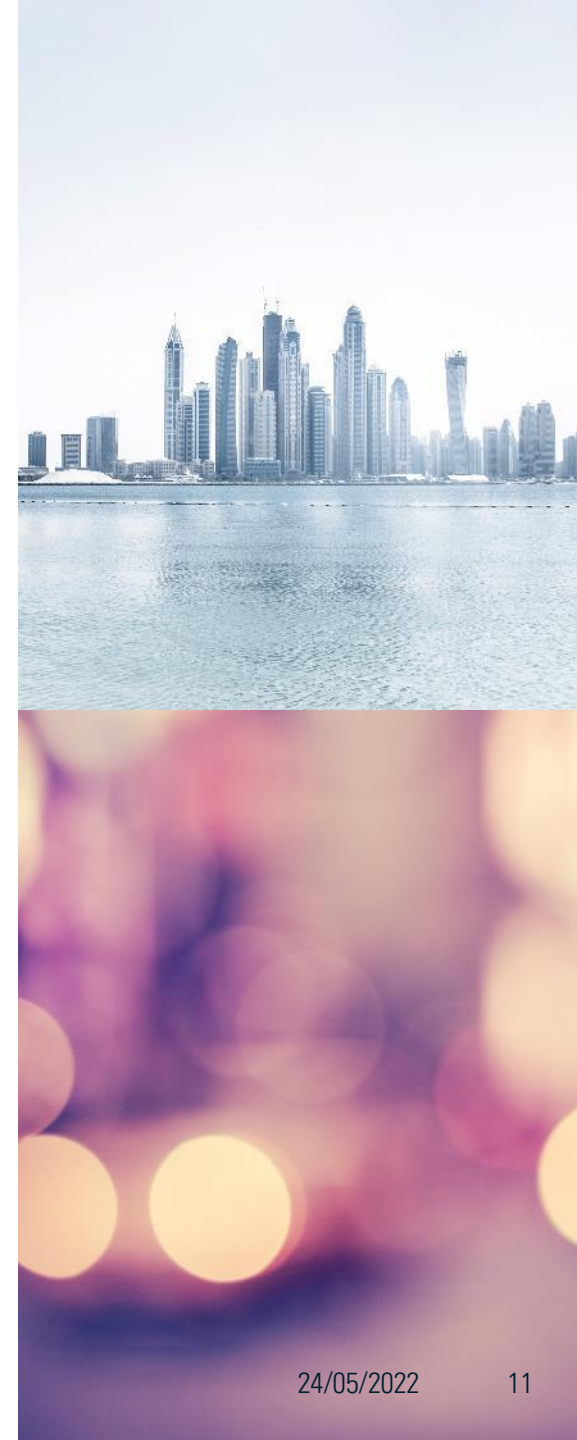


# ANALYSE EXPLORATOIRE



# *PRÉPARATION DES DONNÉES*

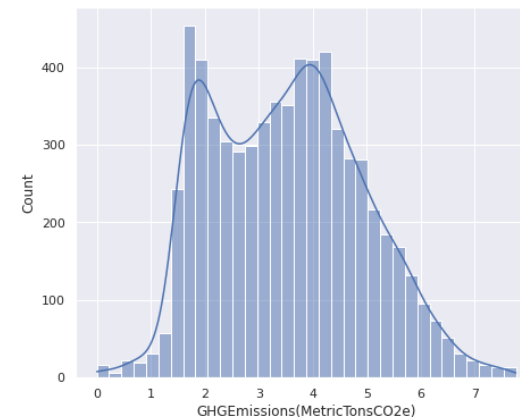
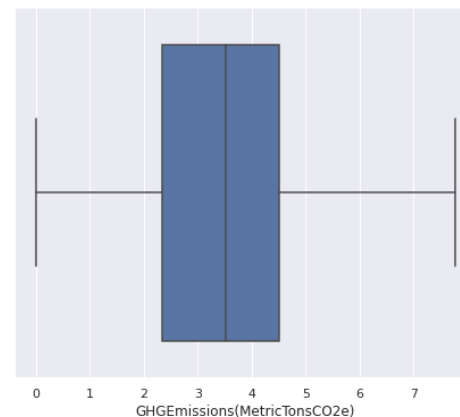
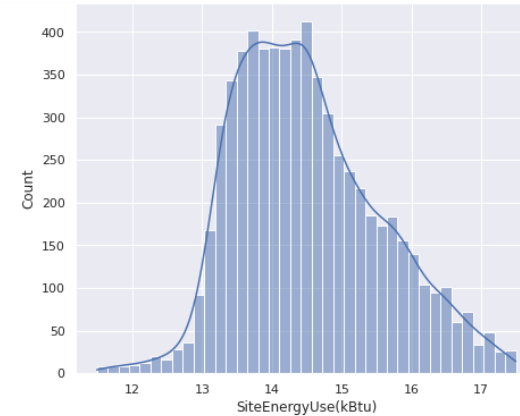
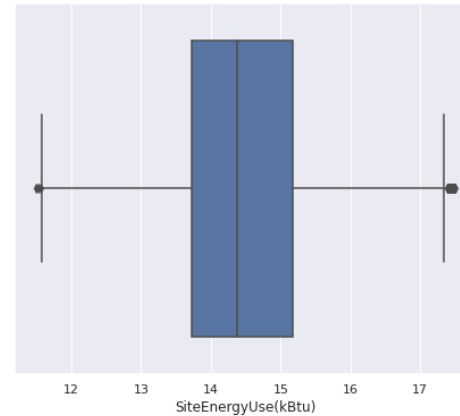
- Feature engineering
- Encodage des variables





# *PRÉPARATION DES DONNÉES*

Passage au logarithme sur les variables d'intérêt

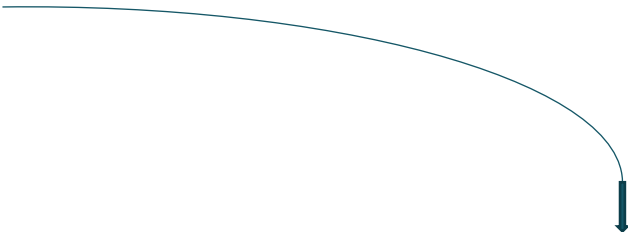


# ENCODAGE DES VARIABLES

Utilisation de OneHotEncoder

OSEBuildingID DataYear BuildingType

0	1	2015	NonResidential
1	2	2015	NonResidential
2	3	2015	NonResidential
3	5	2015	NonResidential
4	8	2015	NonResidential

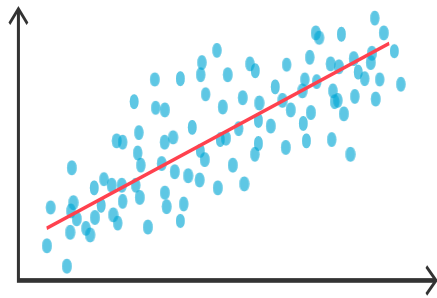


BuildingType_NonResidential	BuildingType_Nonresidential COS	BuildingType_SPS- District K-12	PrimaryPropertyType_Distribution Center
1.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0

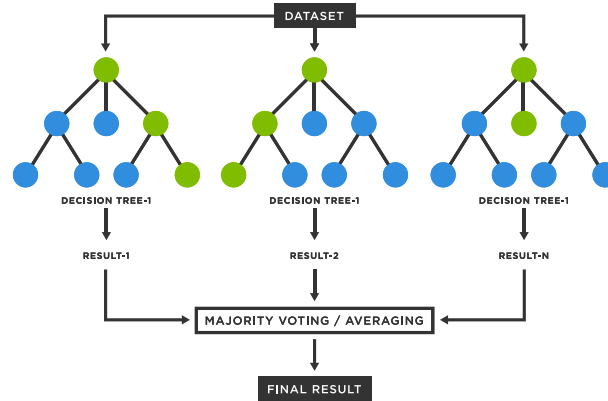
# *MODÉLISATION*



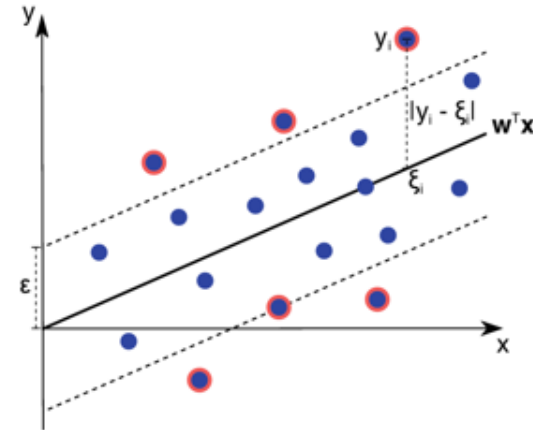
# MODÈLES UTILISÉS



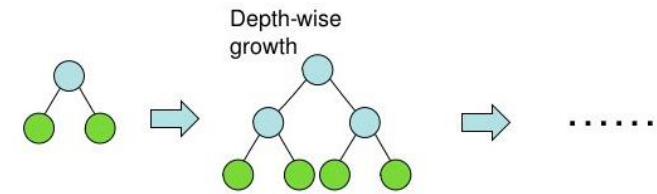
Régression



Random Forest



Support vector  
Regression



XGBoost

# *MÉTHODOLOGIE GÉNÉRALE*

Recherche des hyperparamètres optimaux avec  
GridSearch

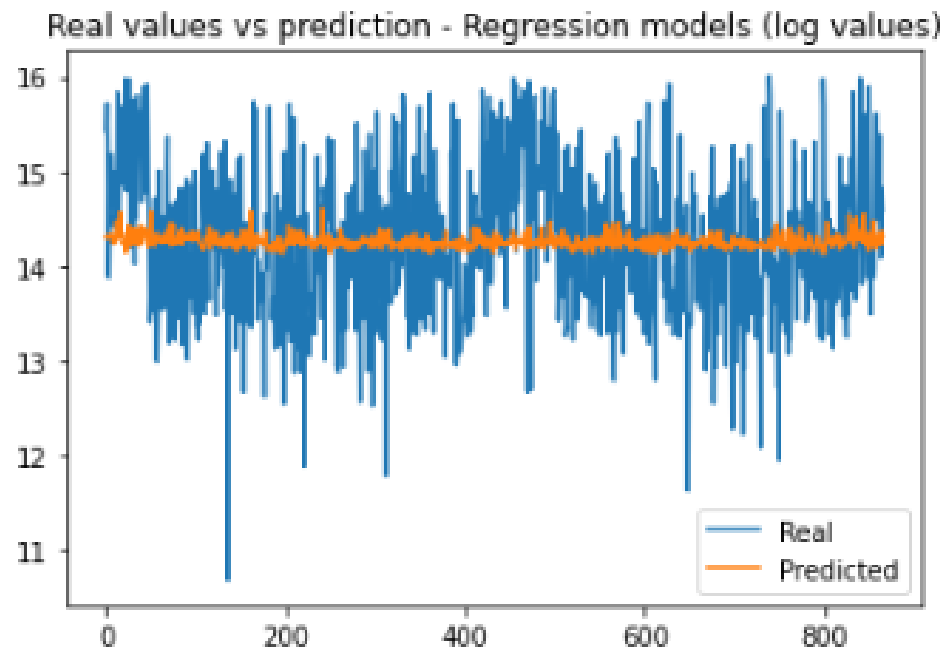
CrossValidation

Calcul du RMSE pour évaluer la performance du  
modèle sur les données de test pour contrôle de  
l'overfitting

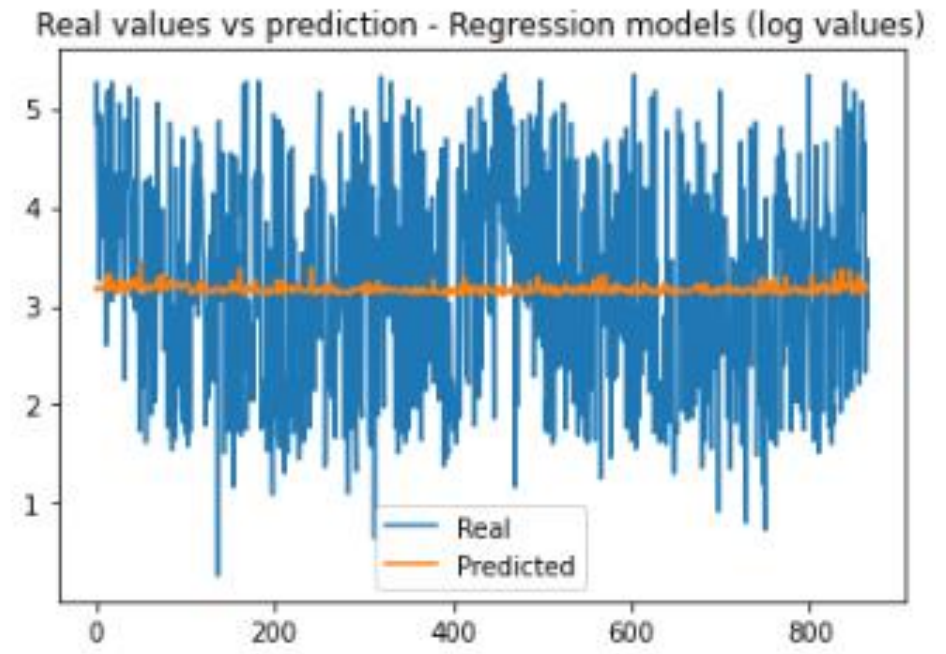
Comparaison des RMSE pour évaluer l'impact de  
l'ENERGYSTARScore

# *RÉGRESSION LINÉAIRES - RÉSULTATS*

Consommation d'énergie



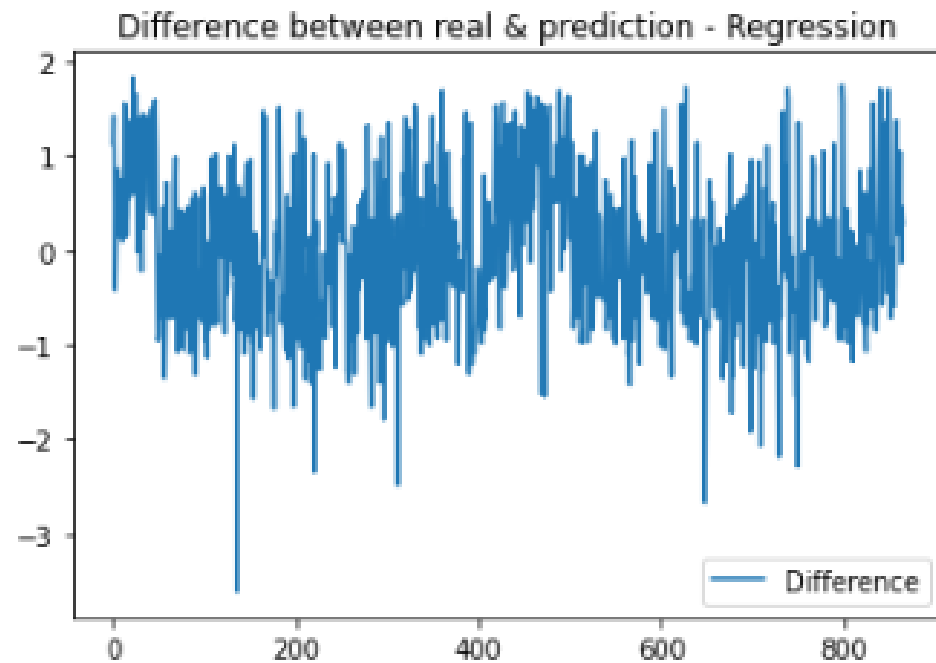
Emissions de CO2



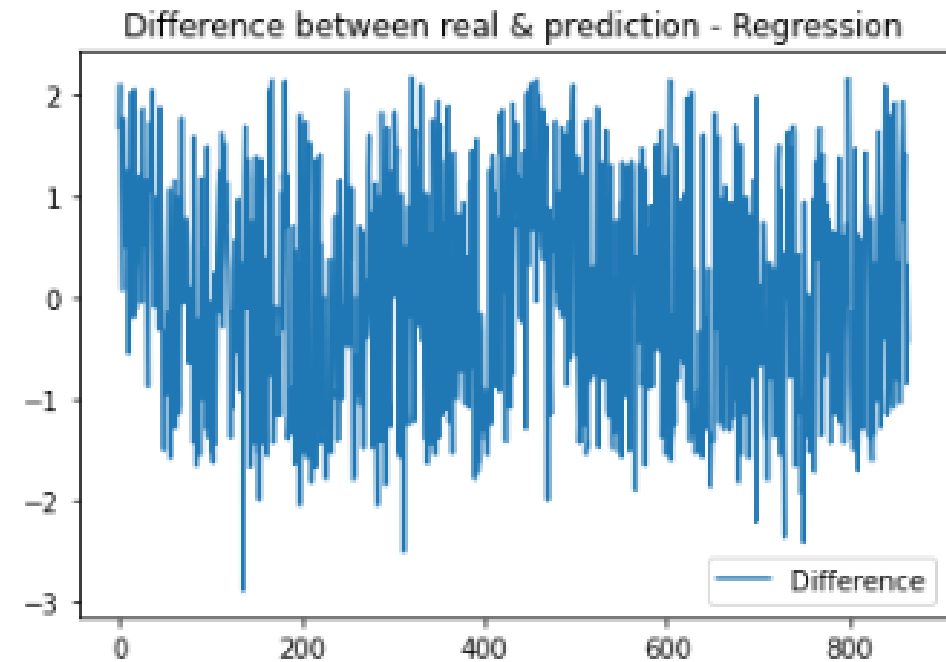


# *RÉGRESSION - DIFFÉRENCE*

Consommation d'énergie



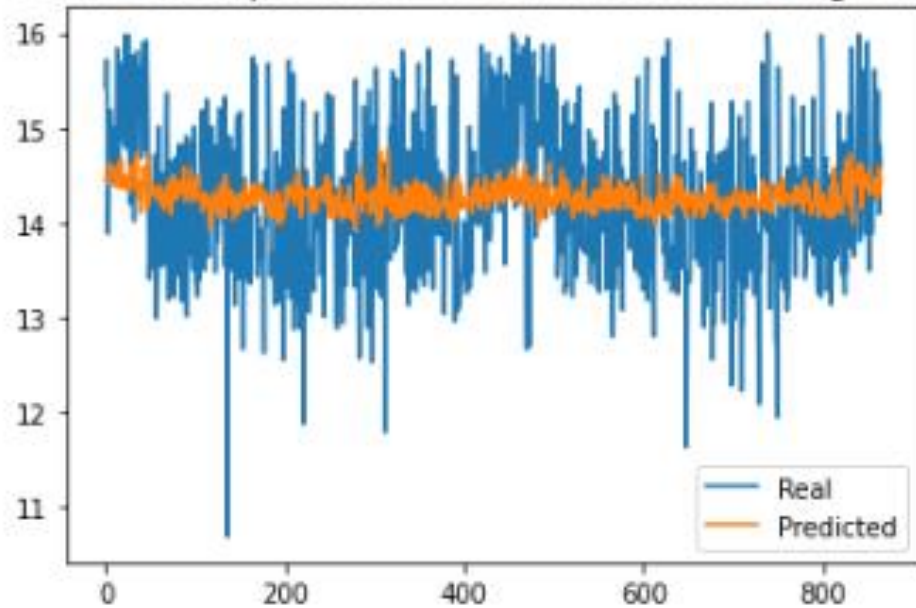
Emissions de CO2



# *RANDOM FOREST - RÉSULTATS*

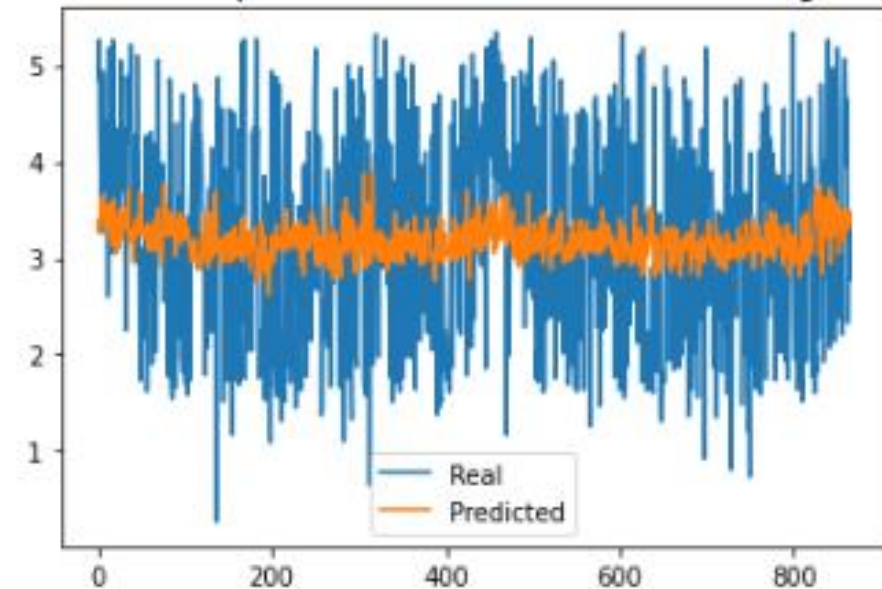
## Consommation d'énergie

Real values vs prediction - Random forest models (log values)



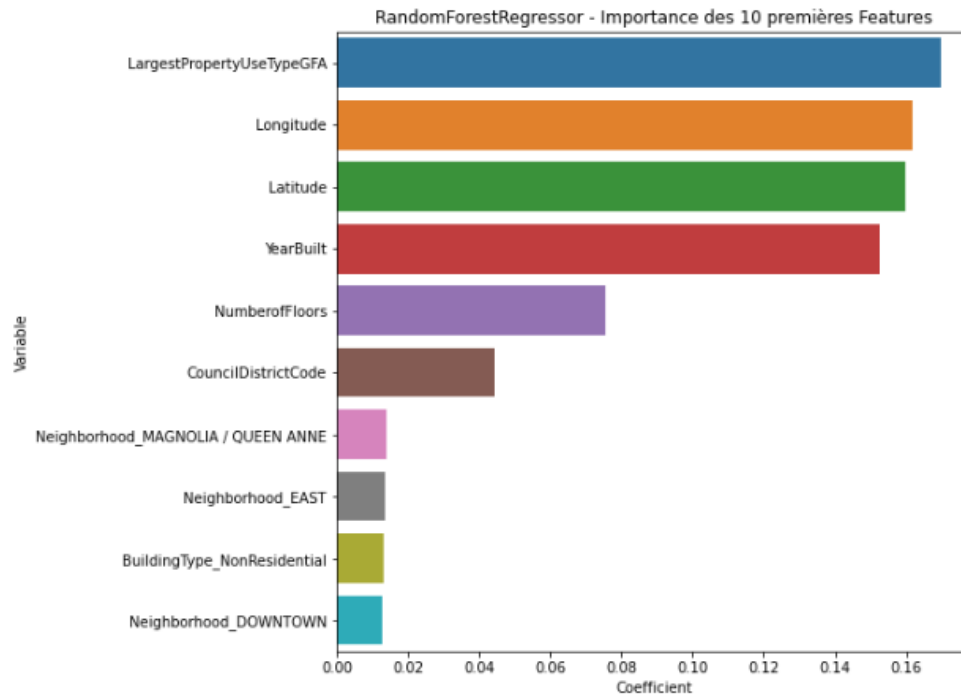
## Emissions de CO2

Real values vs prediction - Random forest models (log values)

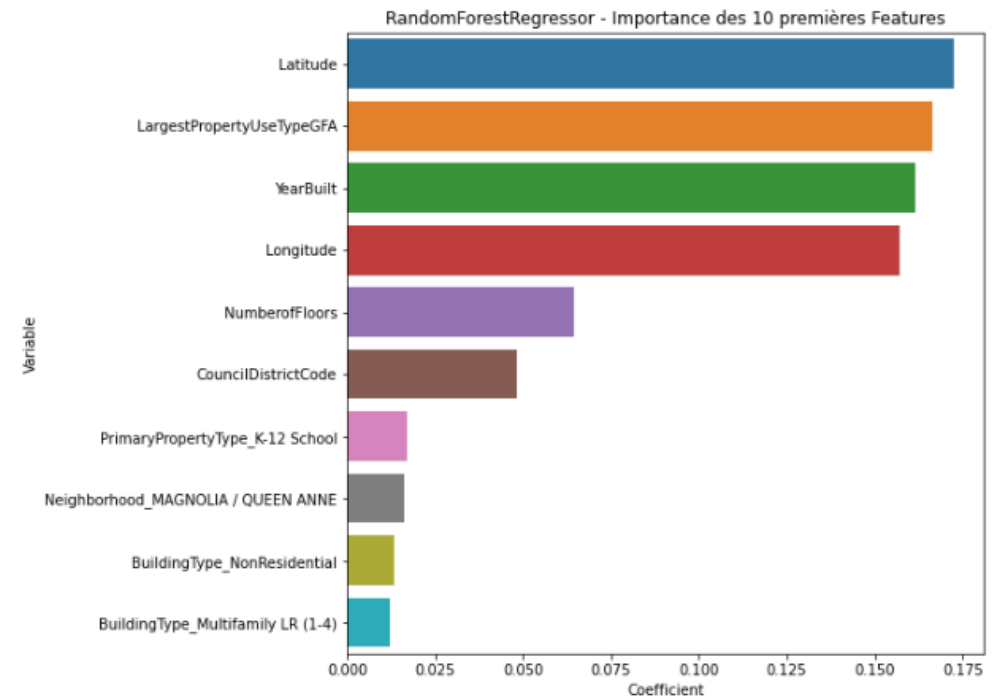


# *RANDOM FOREST - RÉSULTATS*

## Consommation d'énergie



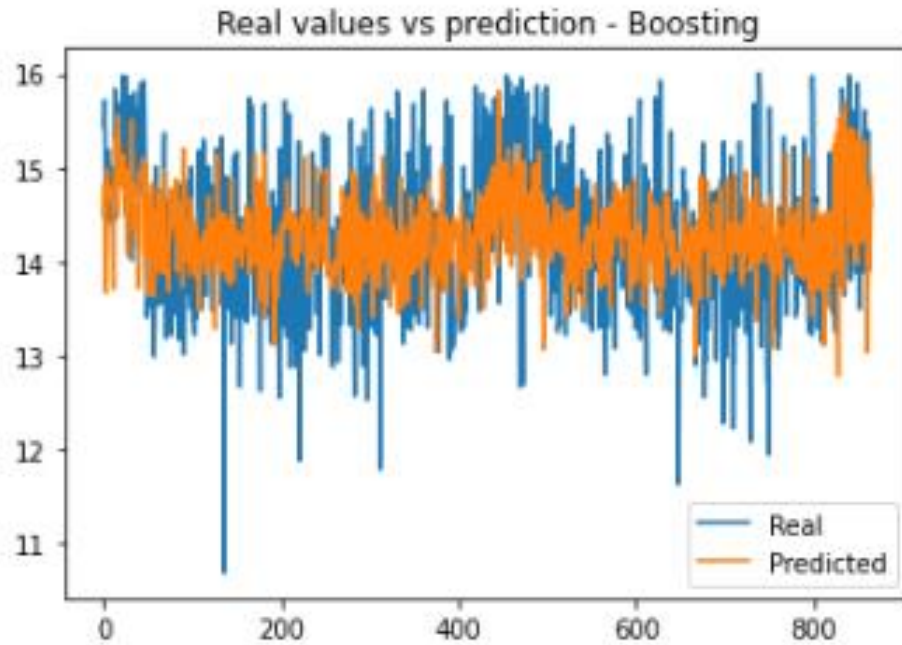
## Emissions de CO2



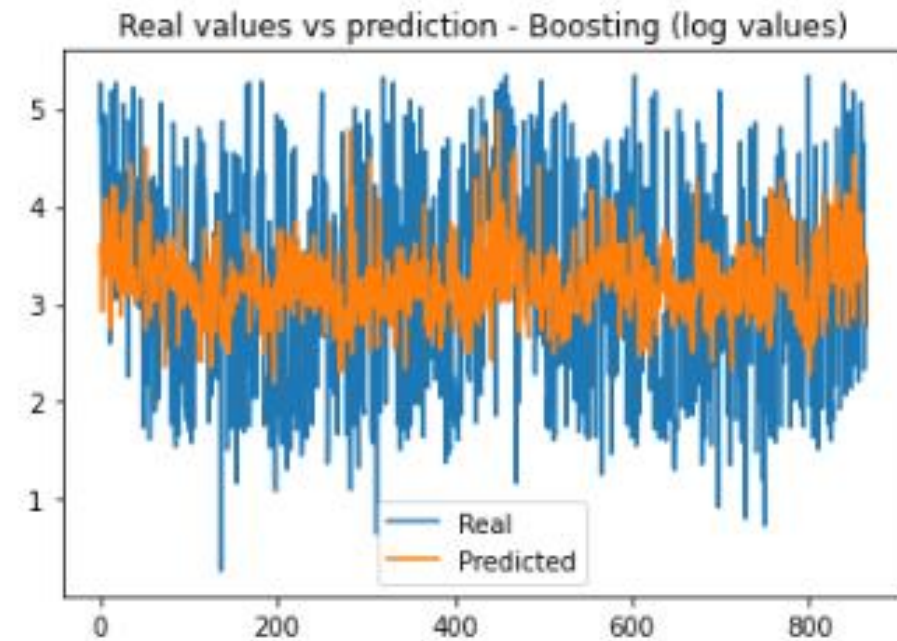


# *XGBOOST- RÉSULTATS*

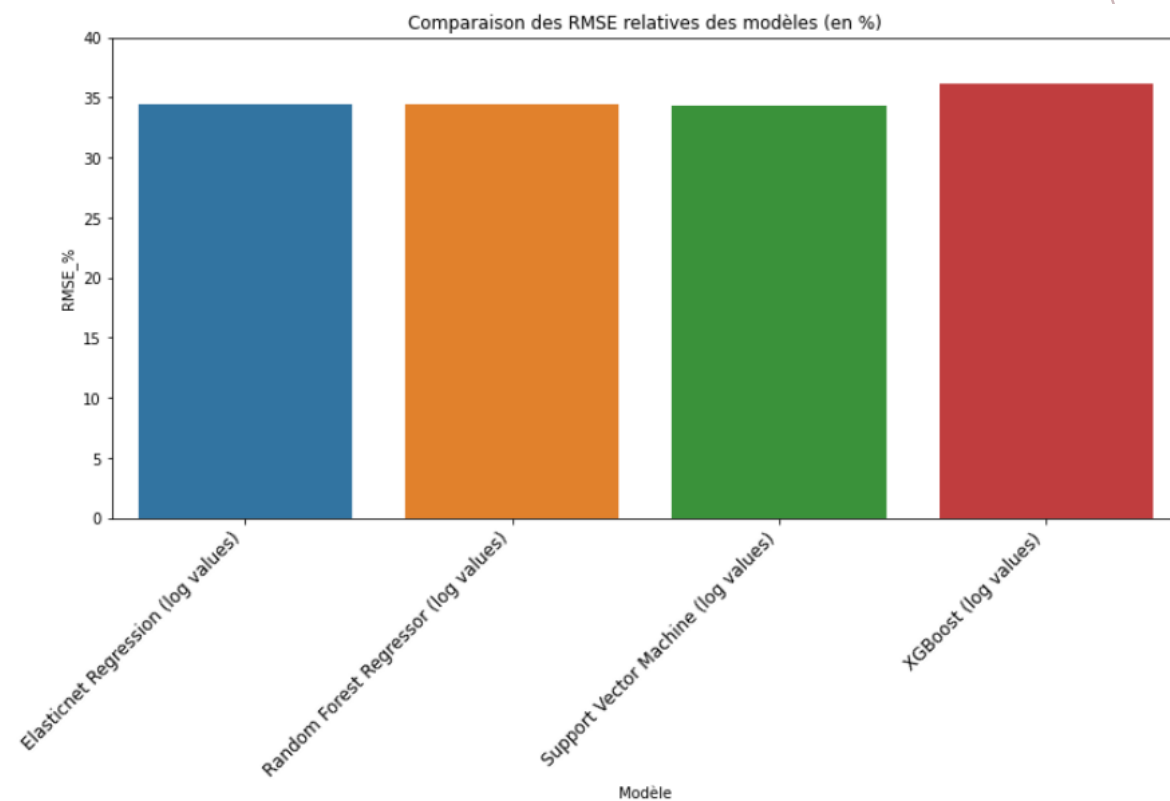
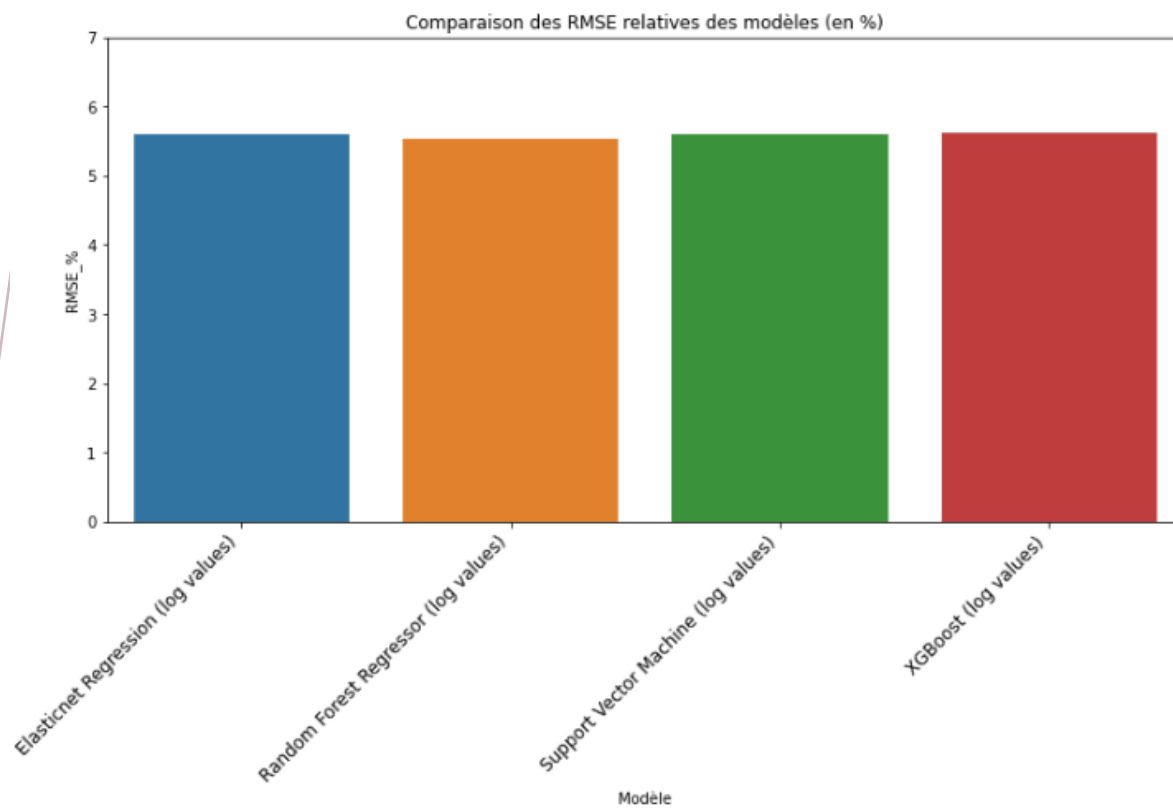
Consommation d'énergie



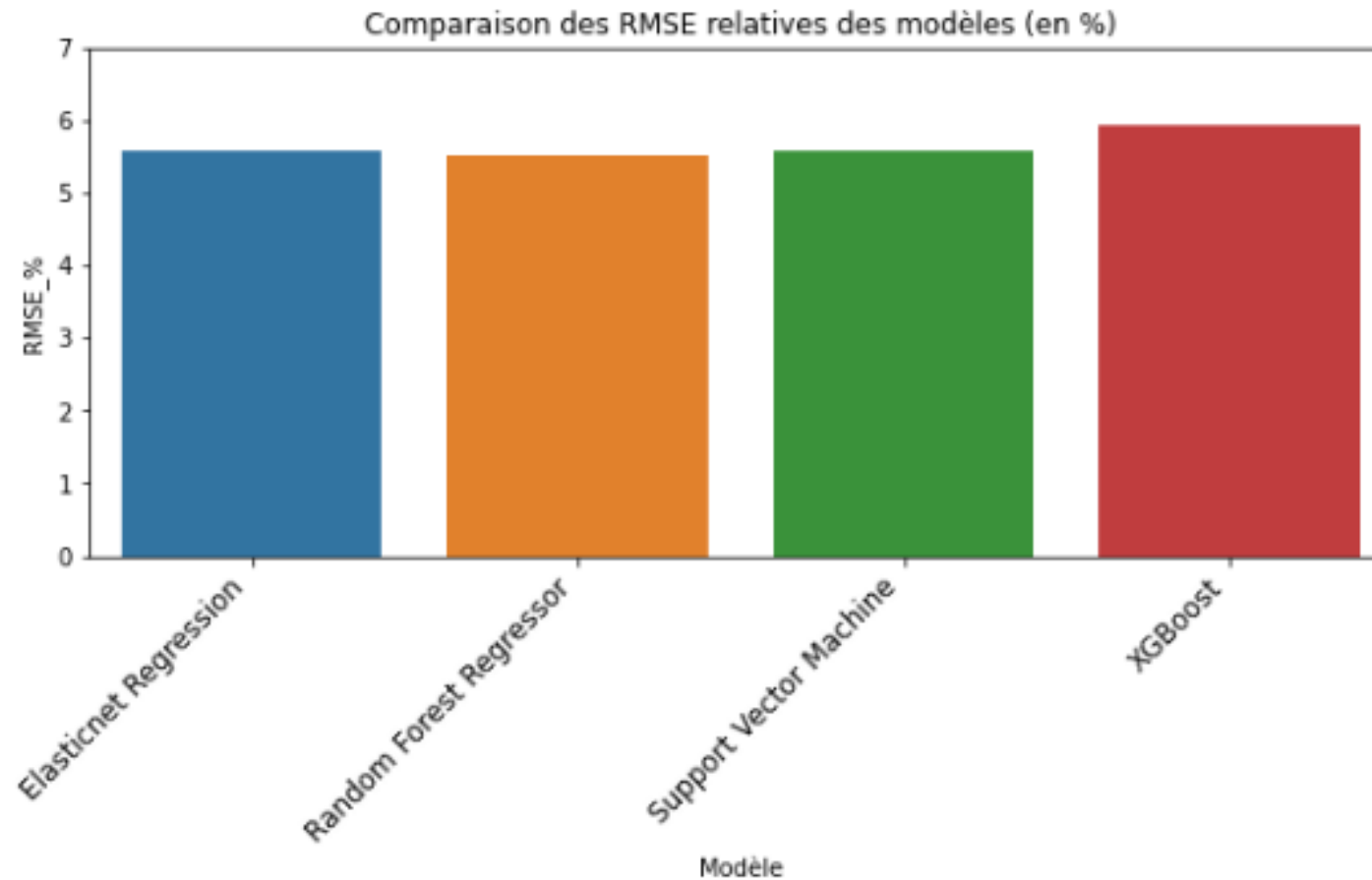
Emissions de CO2



# TABLEAU RÉCAP DES ERREURS RELATIVES



# *TABLEAU RÉCAP AVEC L'ENERGYSTARSCORE*



# CONCLUSION

Travail mené sur des données structurées

Mise en place de modèles d'apprentissage supervisé

Le meilleur modèle semble être celui des forêts aléatoires

L'ENERGY STAR Score semble inutile

Les prévisions sont imprécises, et les erreurs restent importantes

Il serait fortuit d'avoir d'autres informations sur les bâtiments (présence de panneau solaire, immeuble intelligent, matériau de construction, évolution de la consommation annuelle...) pour améliorer les prévisions