PROJET N°4

ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

Soutenance de Projet
26 aout 2021

PROGRAMME

I - Rappel de la problématique

II- Préparation du jeu de données

III-Pistes de modélisations

IV. Présentation du modèle final

RAPPEL DE LA PROBLÉMATIQUE

LA VILLE DE SEATTLE

VILLE NEUTRE EN ÉMISSIONS DE CARBONE EN 2050

Seattle

OBJECTIF

- Prédire les émissions de CO2 et la consommation totale d'énergie des bâtiments non destinés à l'habitation.
- À partir des relevés minutieux effectués en 2015 et en 2016.
- Evaluer l'intérêt de l'ENERGY STAR Score

MISSION

- Réaliser une courte analyse exploratoire.
- Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

PRÉPARATION DU JEU DE DONNÉES

HOMOGÉNÉISATION DES FICHIERS

ANALYSE DES COLONNES

- 37 colonnes communes
- 10 colonnes uniquement dans les données 2015
- 9 colonnes uniquement dans les données 2016

RENOMMER LES COLONNES AVEC LA MÊME SÉMANTIQUE: EX: 'COMMENT' ET 'COMMENTS'

DÉCOMPACTAGE DES DONNÉES DE LOCALISATION DE 2015: 'LATITUDE, LONGITUDE'

- 2015: localisation
- 2016: State, City, Address, ZipCode, Longitude, Latitude

FUSION DES FICHIERS:

- Le dataFrame a 6716 lignes et 49 colonnes.
- Le taux de remplissage global du DataFrame est de : 76.67%

NETTOYAGE DES DONNÉES

Suppression des lignes correspondant à des habitations en nous basant sur la variable BuildingType- (Shape après la suppression : 3318, 49)

Fusion des lignes correspondant à des bâtiments dupliqués

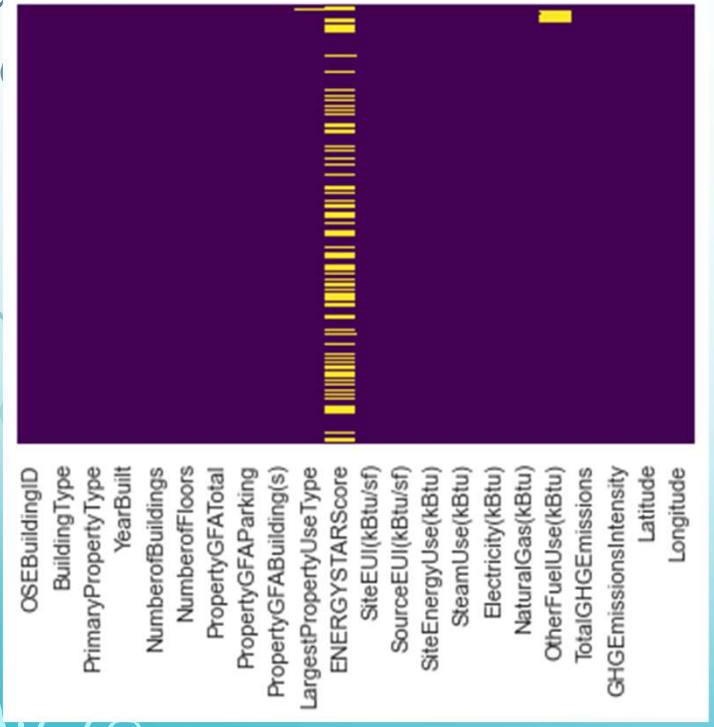
Suppression des colonnes inutiles comme 'Location', 'PropertyName'

les colonnes énergétiques :

- **Les colonnes avec les suffixes WN : "Weather Normalized“**
- **Les variables redondantes: 'Electricity(kWh)', 'NaturalGas(therms)'**

Suppression des variables avec plus 90% valeurs manquantes- (Shape :1697 ,22)

LES VALEURS MANQUANTES



LargestPropertyUseType	0.707130	Suppression des lignes avec NaN 12 ligne
OtherFuelUse(kBtu)	3.182086	Remplace les valeurs manquantes par la moyenne
ENERGYSTARScore	31.997643	Imputation par KNN

TRAITEMENT DES VALEURS ABERRANTES

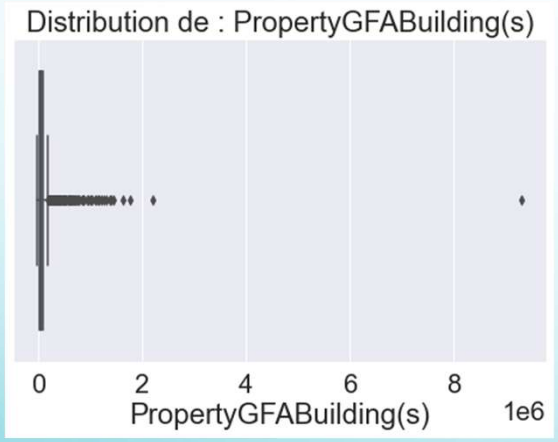
ETAPE 1

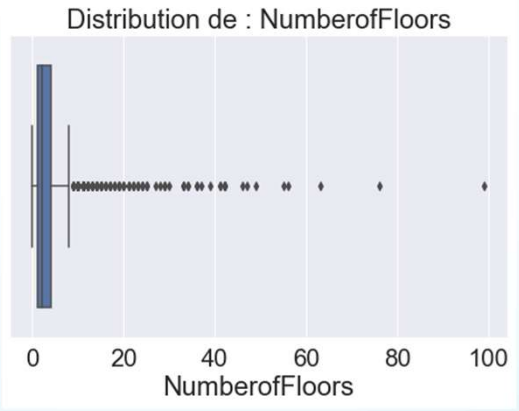
- **Suppression des valeurs négatives:**
  - ❖ **PropertyGFABuilding(s)**
  - ❖ **PropertyGFAParking**
- **Suppression de NumberofFloors:**
  - ❖ **Les lignes avec 0 étage et plus de 76 étages**

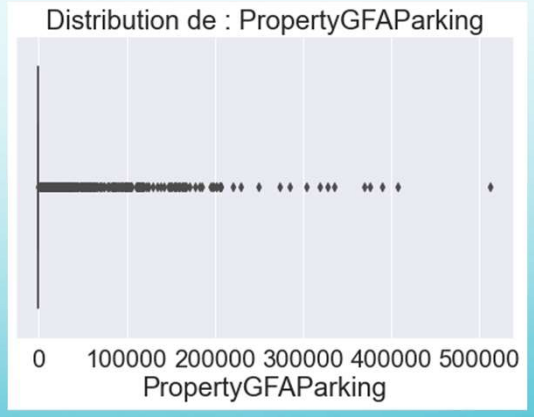
ETAPE 2: MODÈLE D'ISOLATION FOREST:

LES VARIABLES À ENLEVER AVANT DE FAIRE UNE ISOLATION FOREST

- **OSEBuildingID, YearBuilt, Latitude, Longitude,**
- **NumberofFloors**
- **ENERGYSTARScore,**
- **Les variable catégorique: BuildingType, PrimaryPropertyType, LargestPropertyUseType**

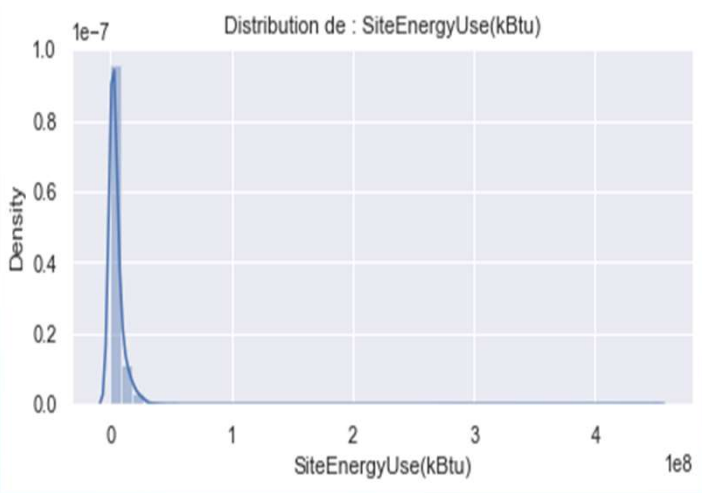


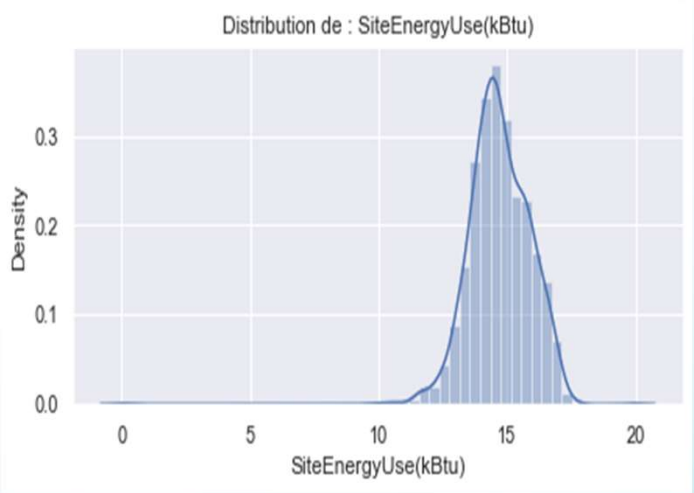




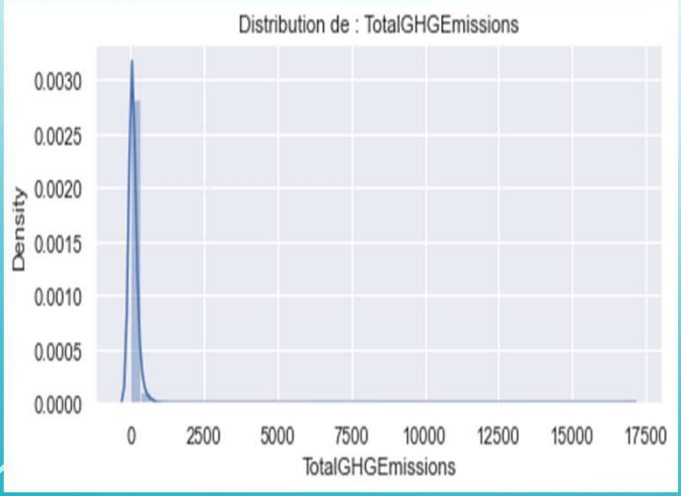
TRAITEMENT DES VALEURS ABERRANTES

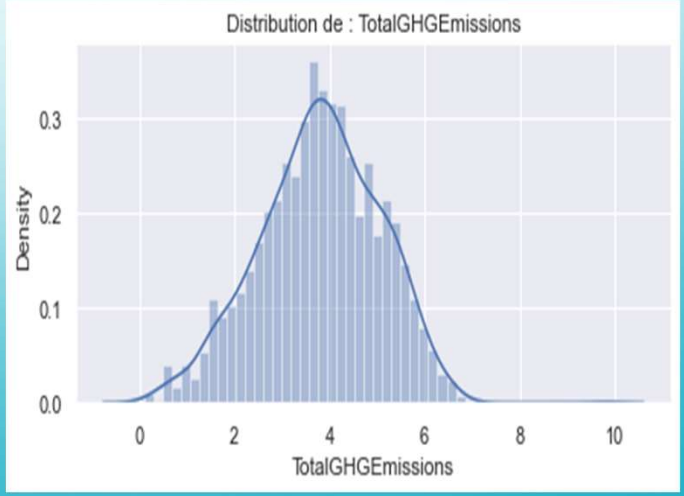
CONSOMMATION D'ENERGY





EMISSION CO2

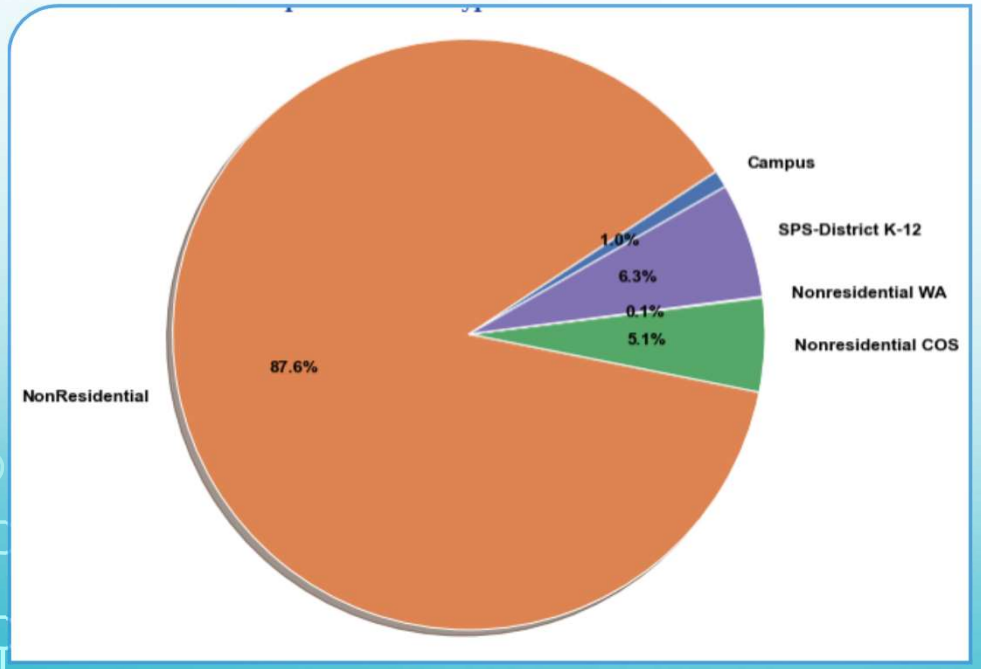


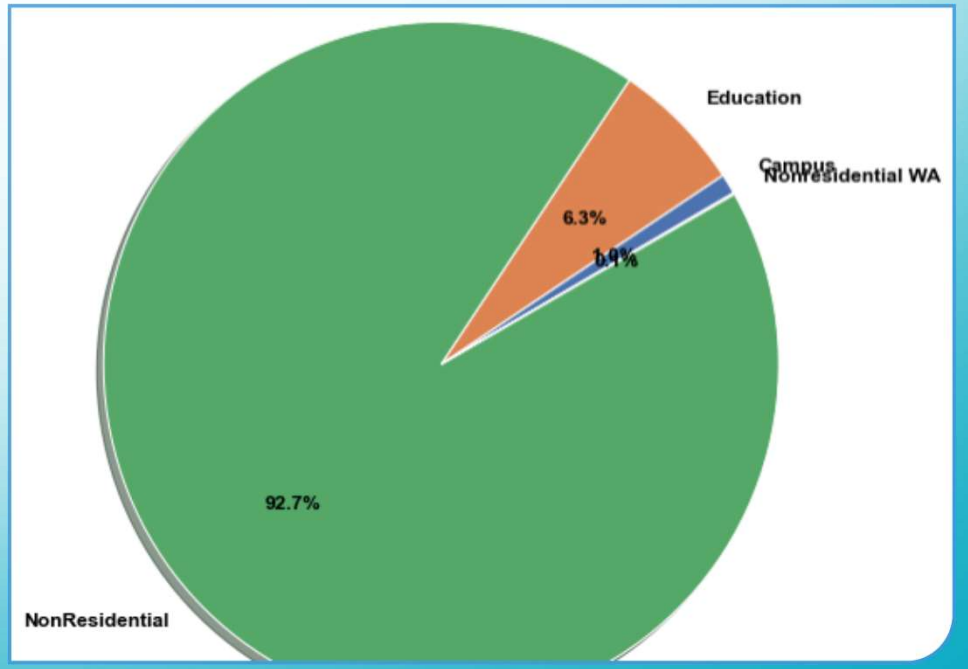


TRANSFORMATION LOGARITHMIQUE

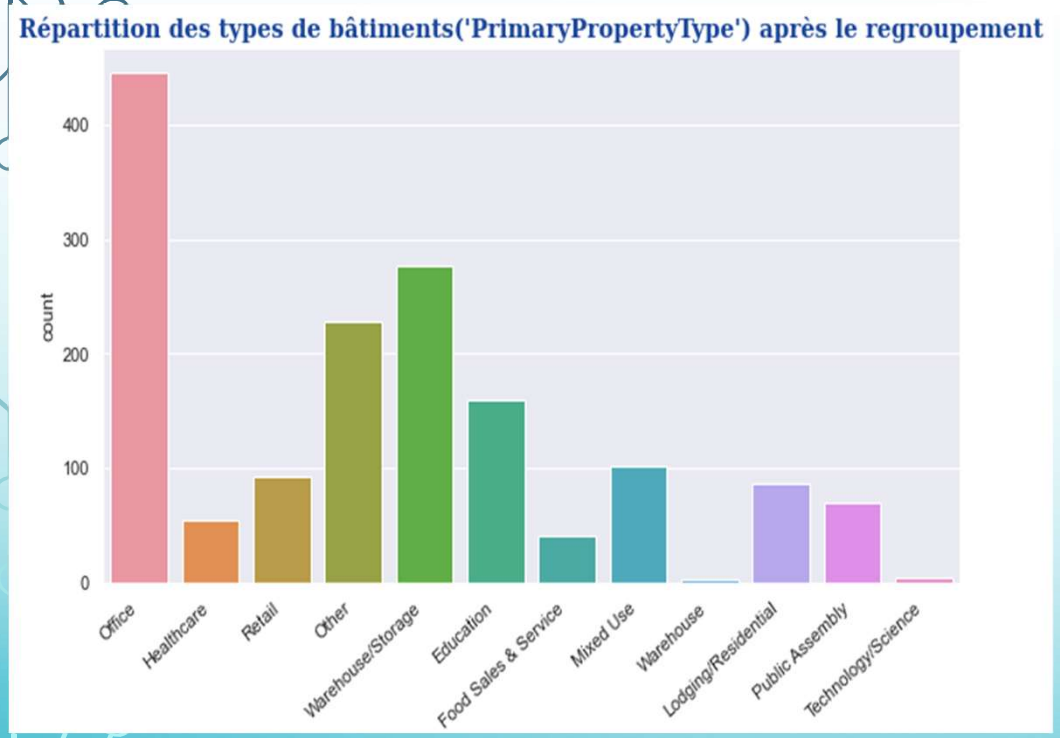
LES VARIABLES CATÉGORIQUES

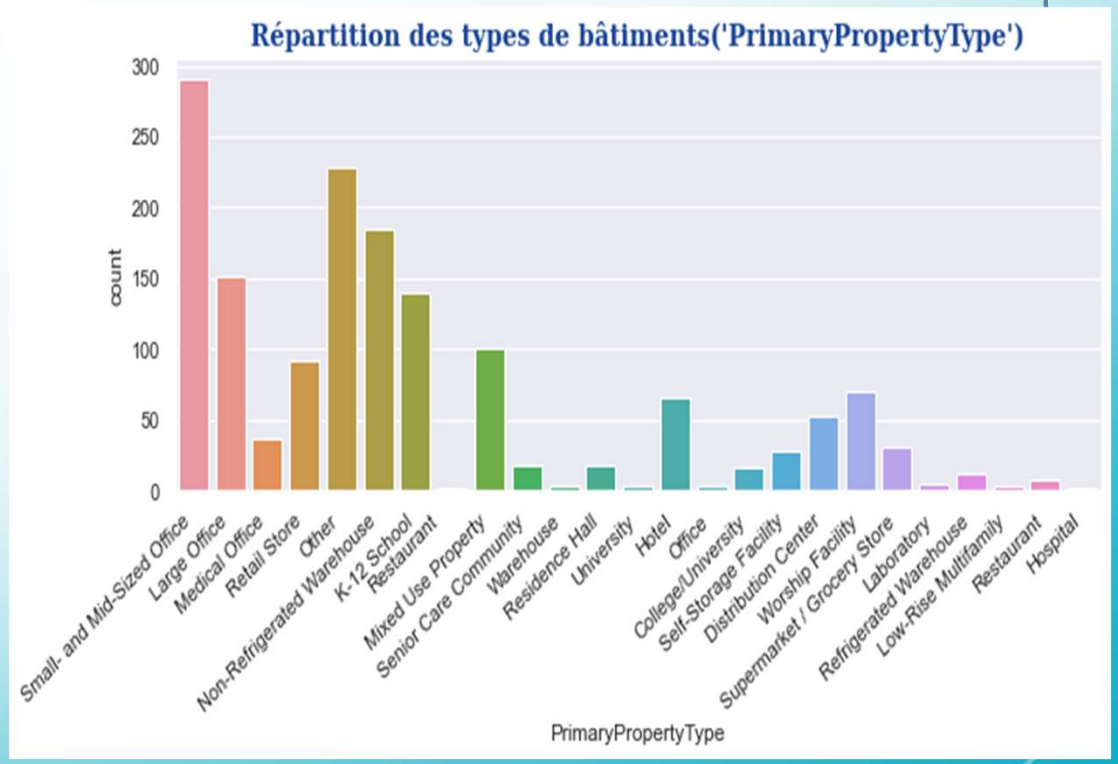
- **BUILDINGTYPE**





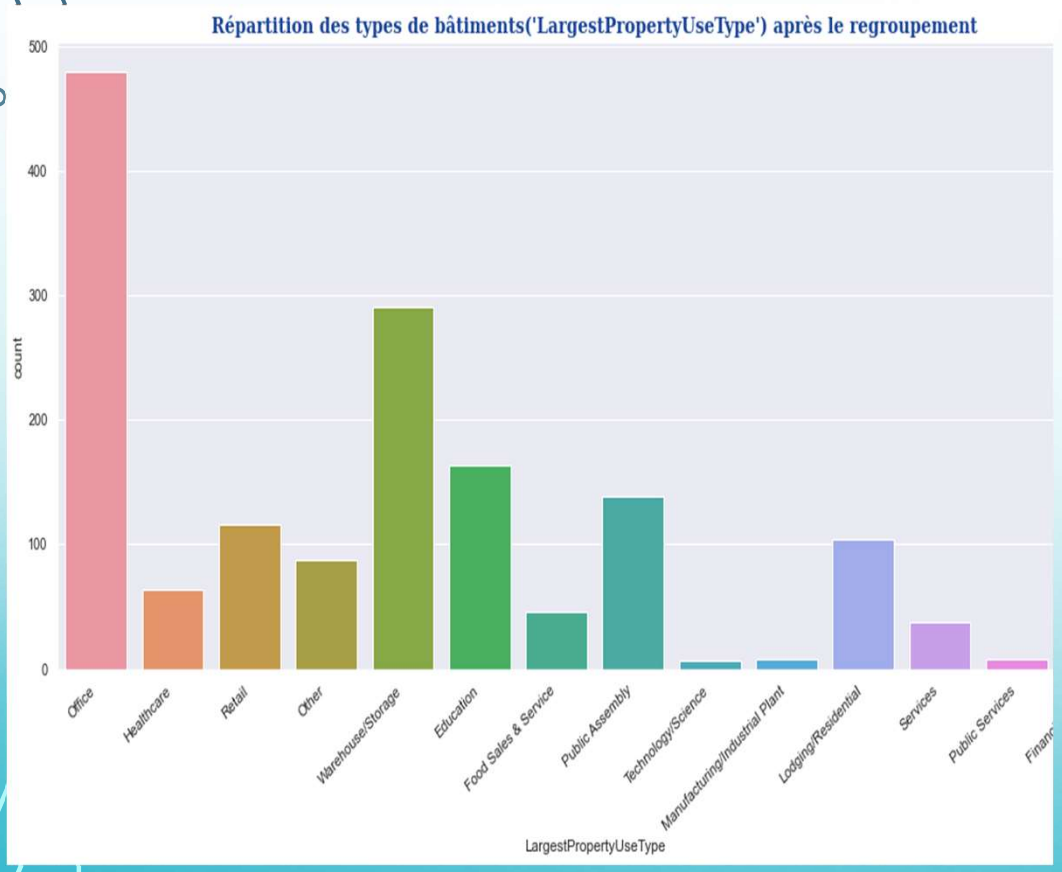
PRIMARYPROPERTYTYPE

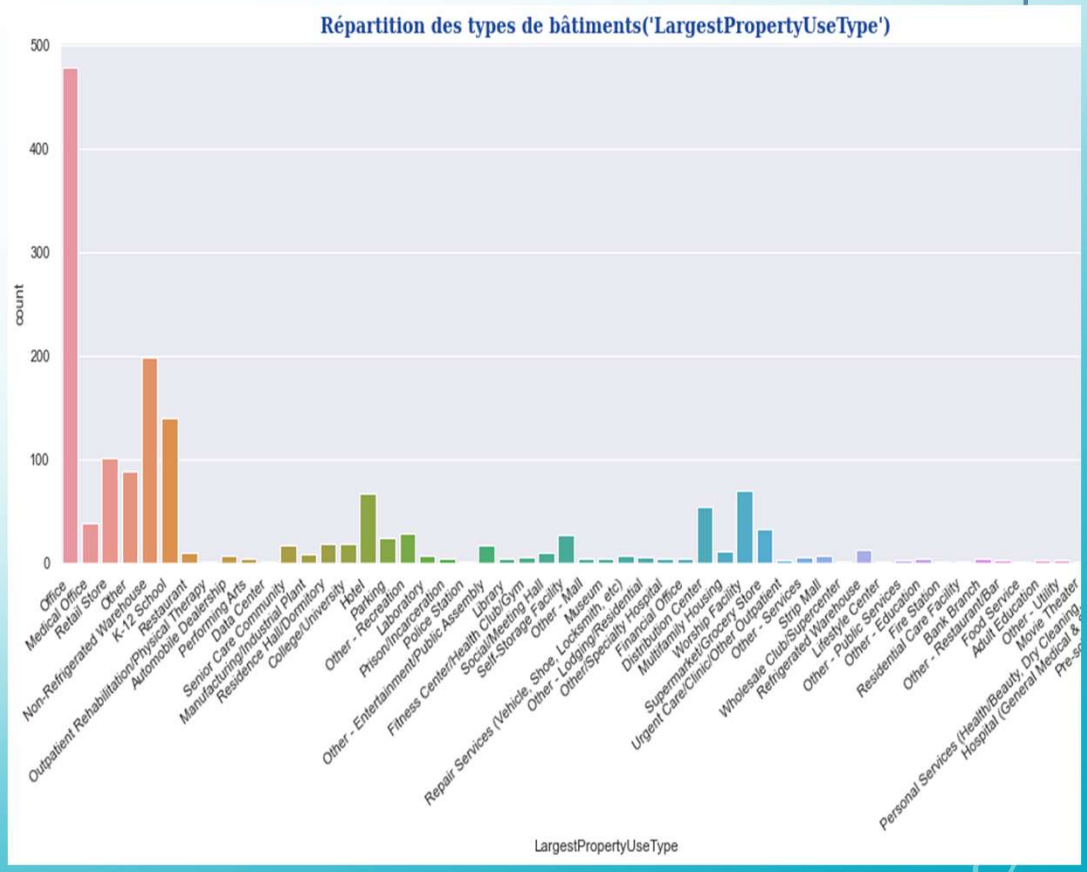




*** Regroupement selon la classification de l'Agence de Protection de l'Environnement**

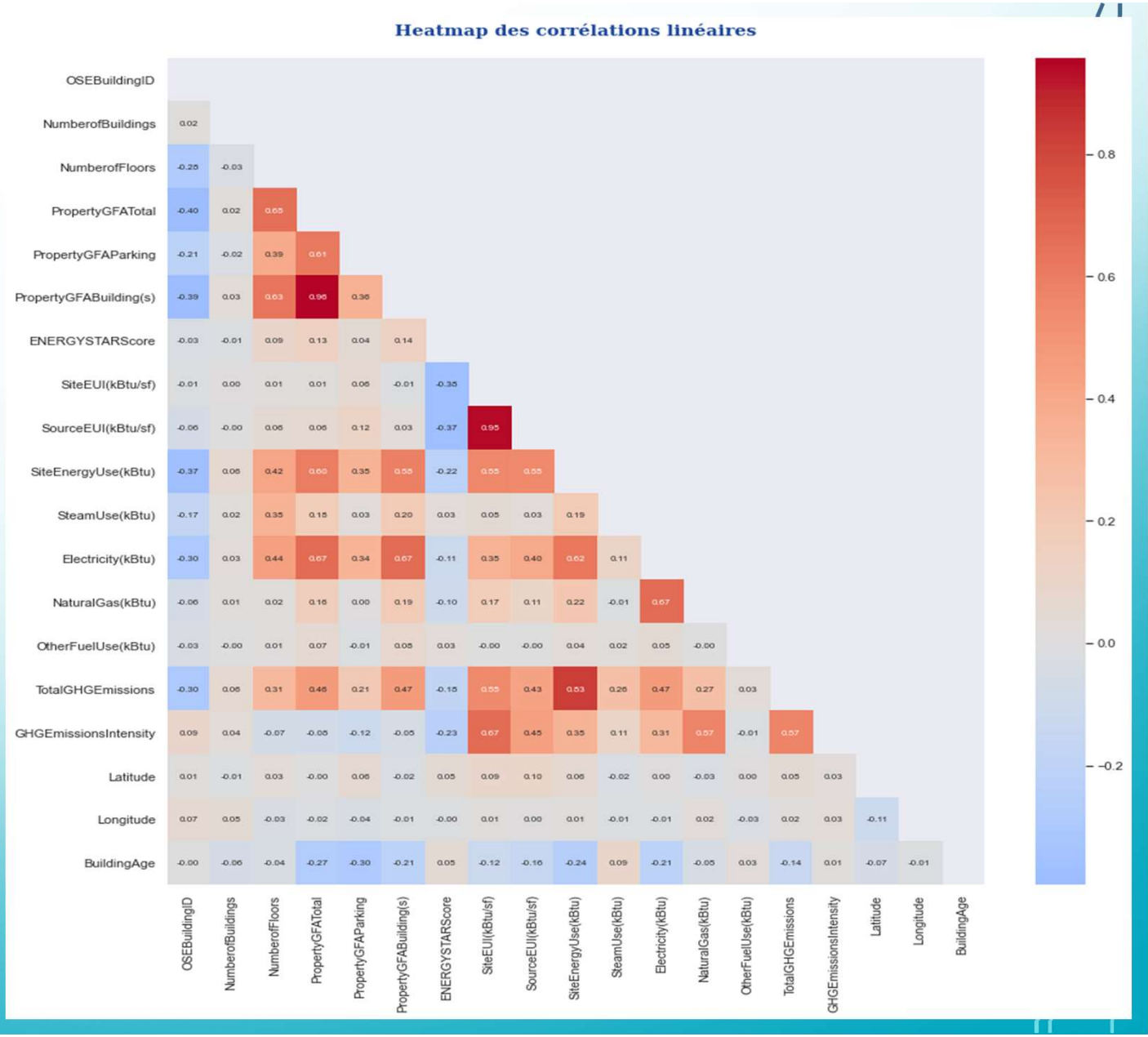
LARGESTPROPERTYUSETYPE





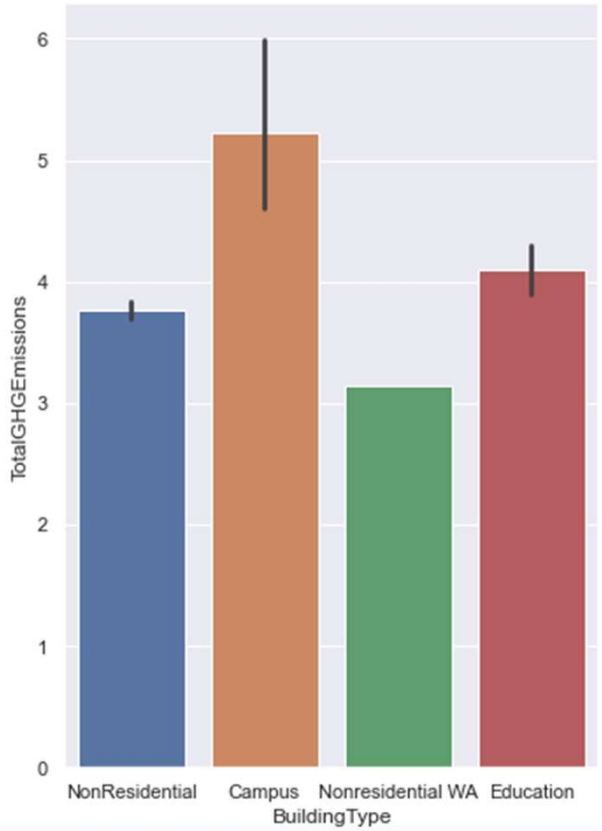
LES CORRÉLATIONS ENTRE LES VARIABLES

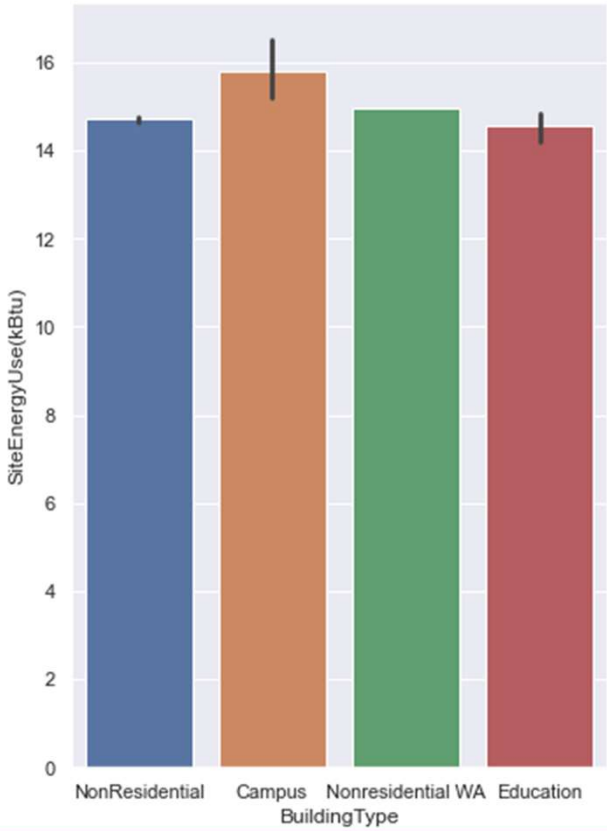
- **Energy Star Score : pas de corrélation notable**
- **PropertyGFABuilding**
- **Les deux variables cibles très corrélées**



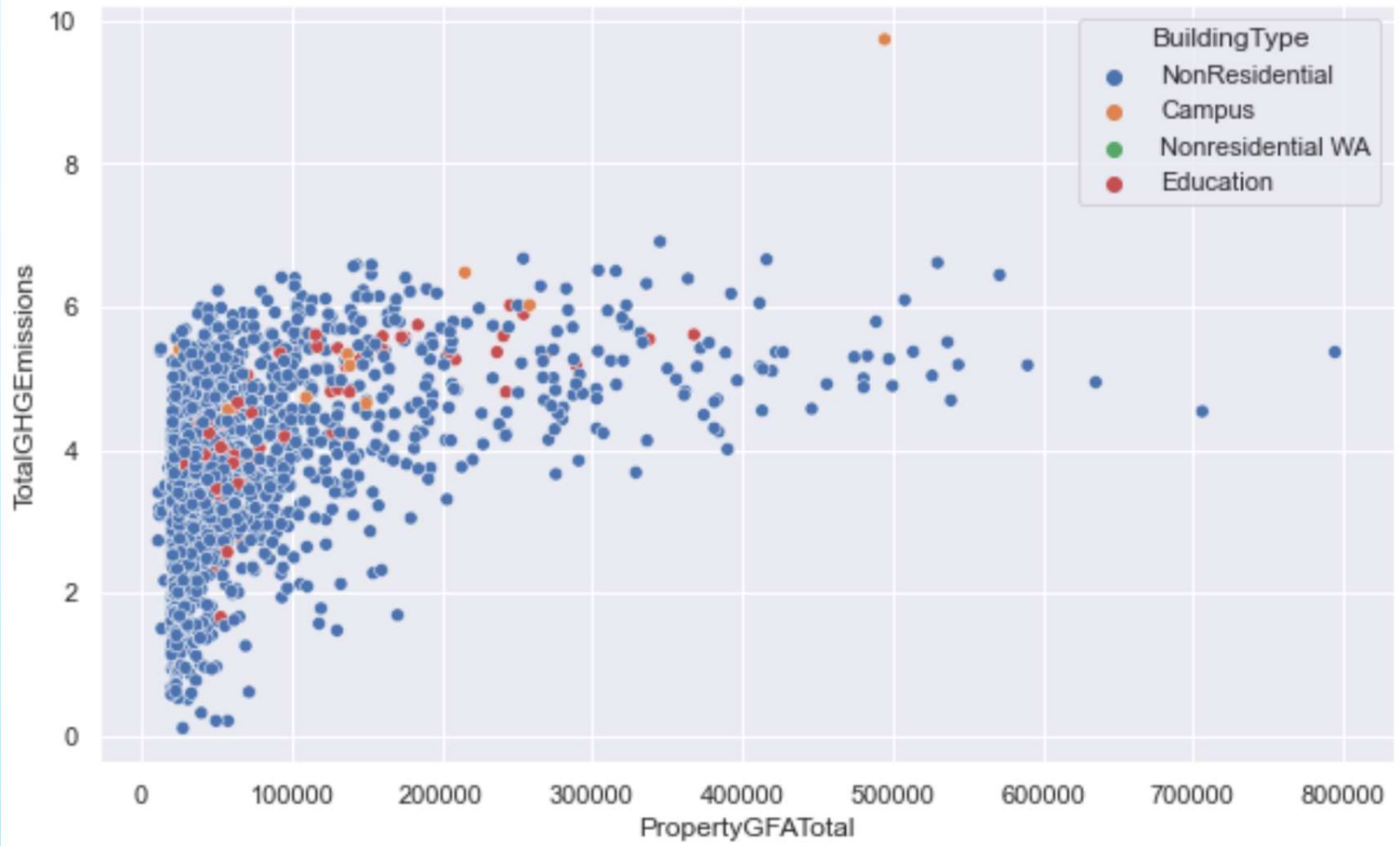
gfdgfs

Répartition de la consommation d'énergie et emissions de CO2 en fonction du type de bâtiment





Emission de CO2 par surface totale au sol et par type de bâtiment



PISTES DE MODÉLISATION

PRÉTRAITEMENT

- SAUVEGARDE SÉPARÉE DU ENERGY STAR SCORE
- SÉPARATION DES VARIABLES EXPLICATIVES (FEATURES) ET DES ÉTIQUETTES (TARGETS)

PRÉPARATION DES JEUX D'ENTRAINEMENT ET DE TEST

X = TOUTES LES VARIABLES SAUF 'SITEENERGYUSE(KBTU)','TOTALGHGEMISSIONS'

Y = UNIQUEMENT 'SITEENERGYUSE(KBTU)','TOTALGHGEMISSIONS'

- PRÉPARATION DES DONNÉES NUMÉRIQUES ET CATÉGORIQUES
  - Les variables numériques:

    Standardisation
  - Les variables catégoriques : PrimaryPropertyType
    One-hot-encoding (transformation en numériques)
  - Fusionner les variables numériques et catégoriques

MODÈLE BASELINE: DUMMYREGRESSOR

METTRE EN PLACE UN MODÈLE BASELINE

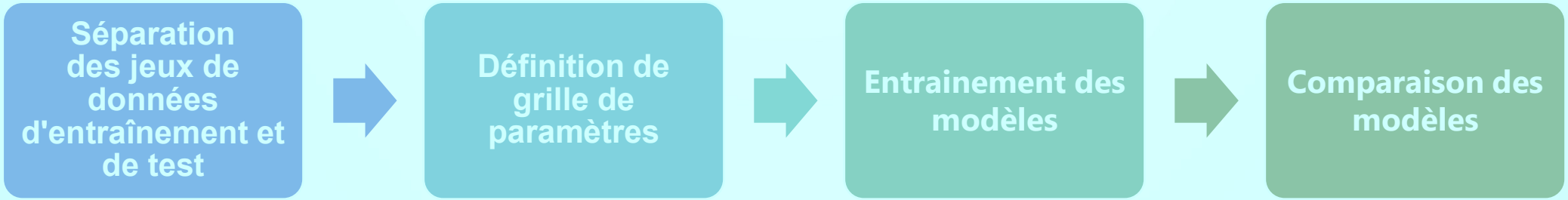
POUR ÉVALUER LES PERFORMANCES DE NOS FUTURS MODÈLES

MAE	Modèle	R2	RMSE	time
0.9636	DummyRegressor	-0.000019	1.185507	13.136921

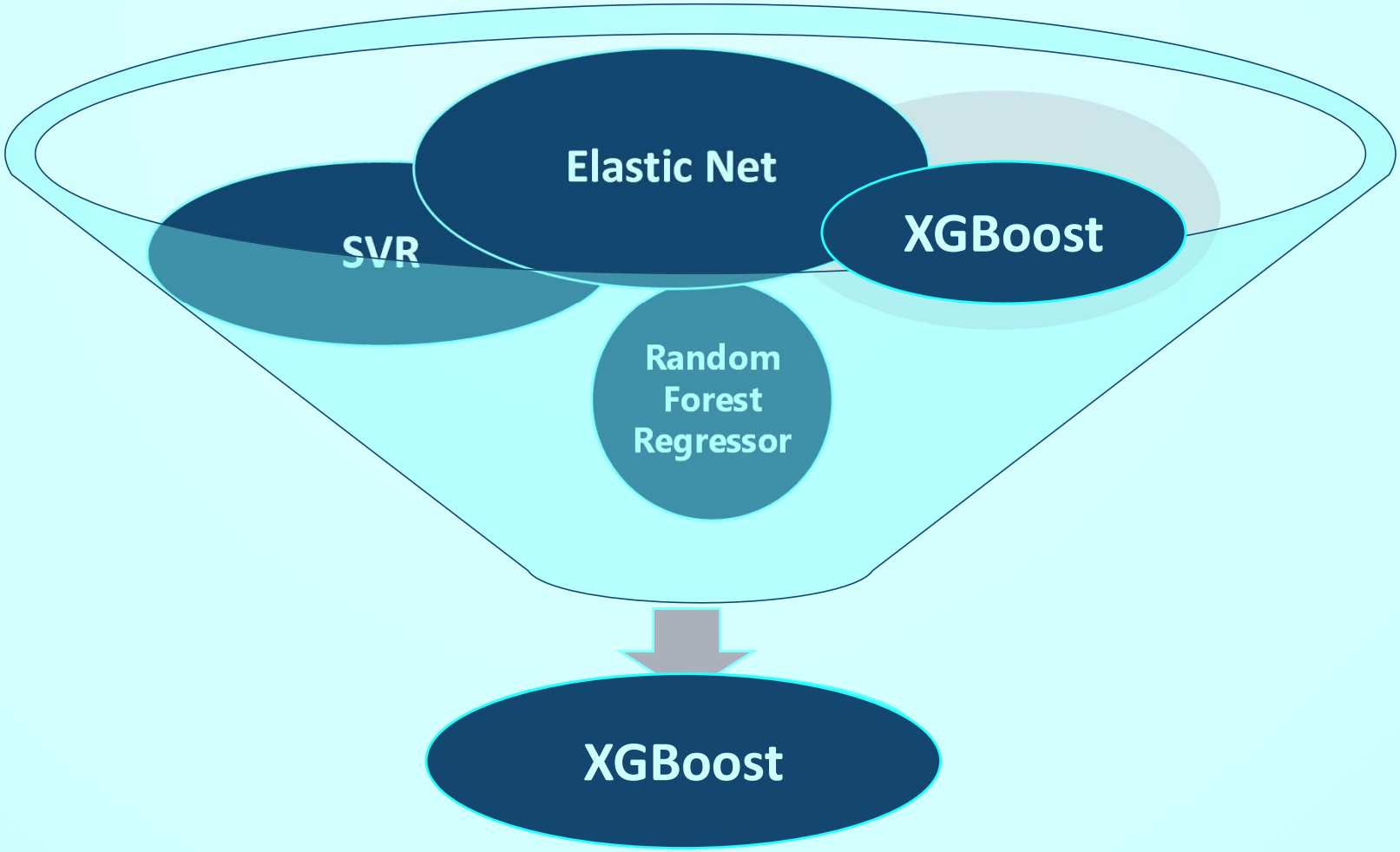
PRÉDICTION DE LA CONSOMMATION TOTALE D'ÉNERGIE

(siteenergyuse)

PROCESSUS



MODÈLES ENTRAÎNÉS



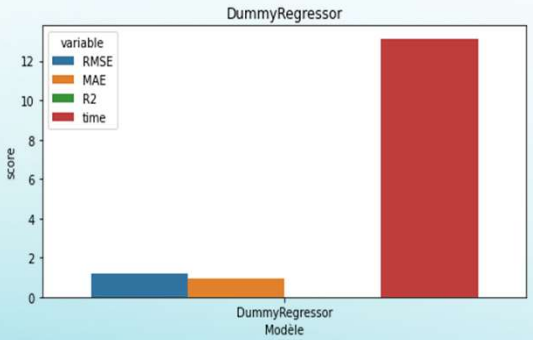
Les Métriques de l'évaluation

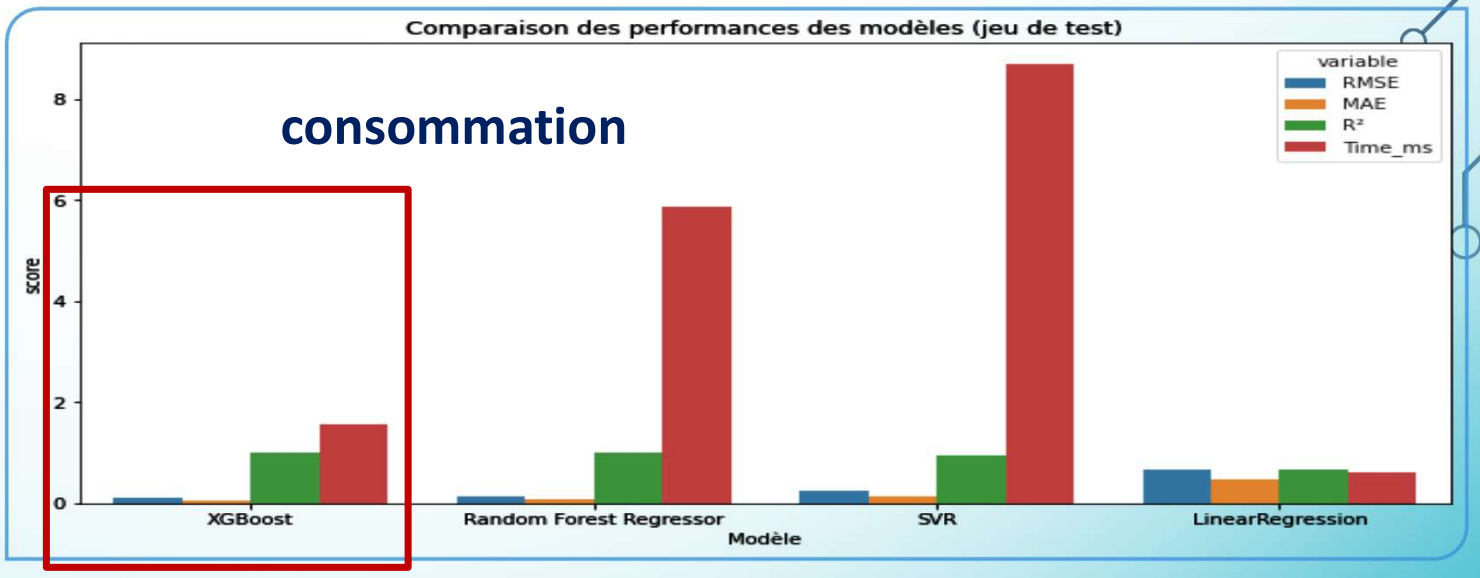
CONSOMMATION

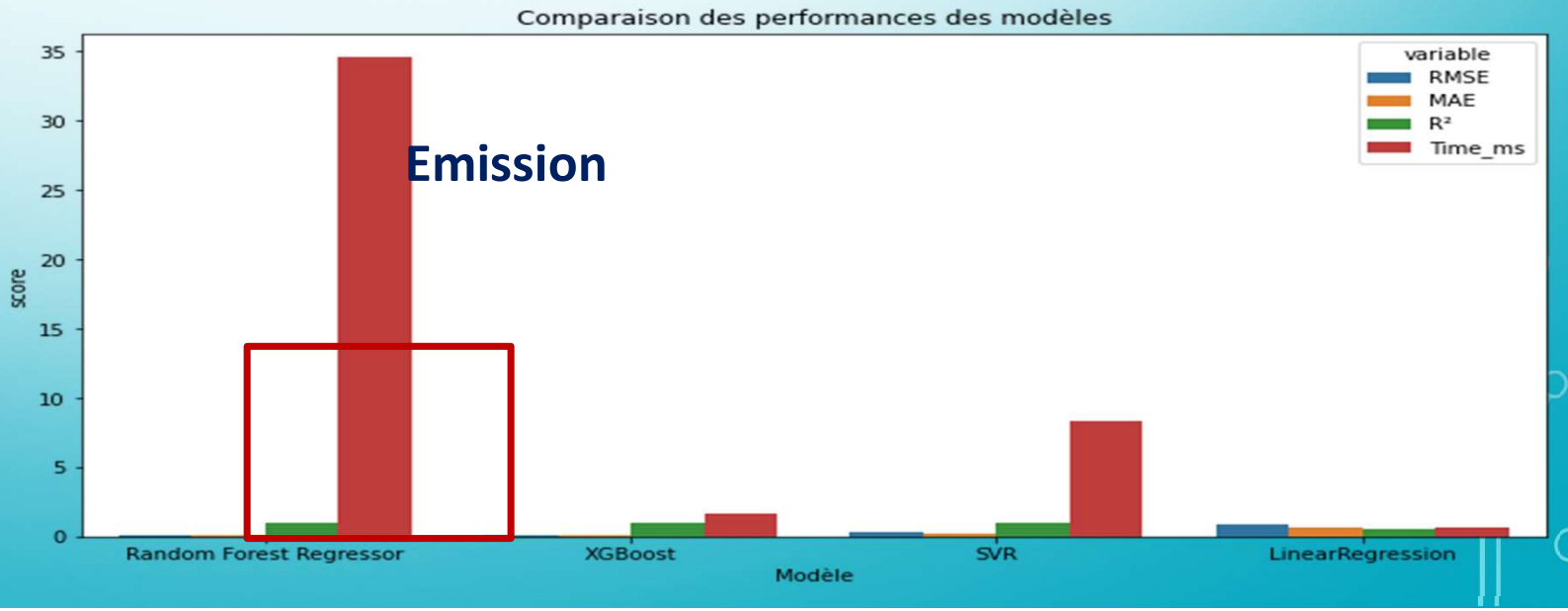
MODÈLE	RMSE	MAE	R^2	TIME_MS	CV_SCORE(RMSE)
XGBoost	0.119115	0.063015	0.989904	1.561197	0.274948
Random Forest Regressor	0.126818	0.072198	0.988556	5.874823	0.310197
SVR	0.259227	0.130223	0.952186	8.691777	0.395331
LinearRegression	0.672611	0.473024	0.678095	0.609981	0.691647

Emission

MODÈLE	RMSE	MAE	R^2	TIME_MS	CV_SCORE(RMSE)
RANDOM FOREST REGRESSOR	0.093586	0.054499	0.994876	34.579260	0.127989
XGBOOST	0.095544	0.066745	0.994659	1.638885	0.116742
SVR	0.309781	0.189442	0.943856	8.377229	0.342345
LINEARREGRESSION	0.883711	0.649457	0.543104	0.619562	0.777396

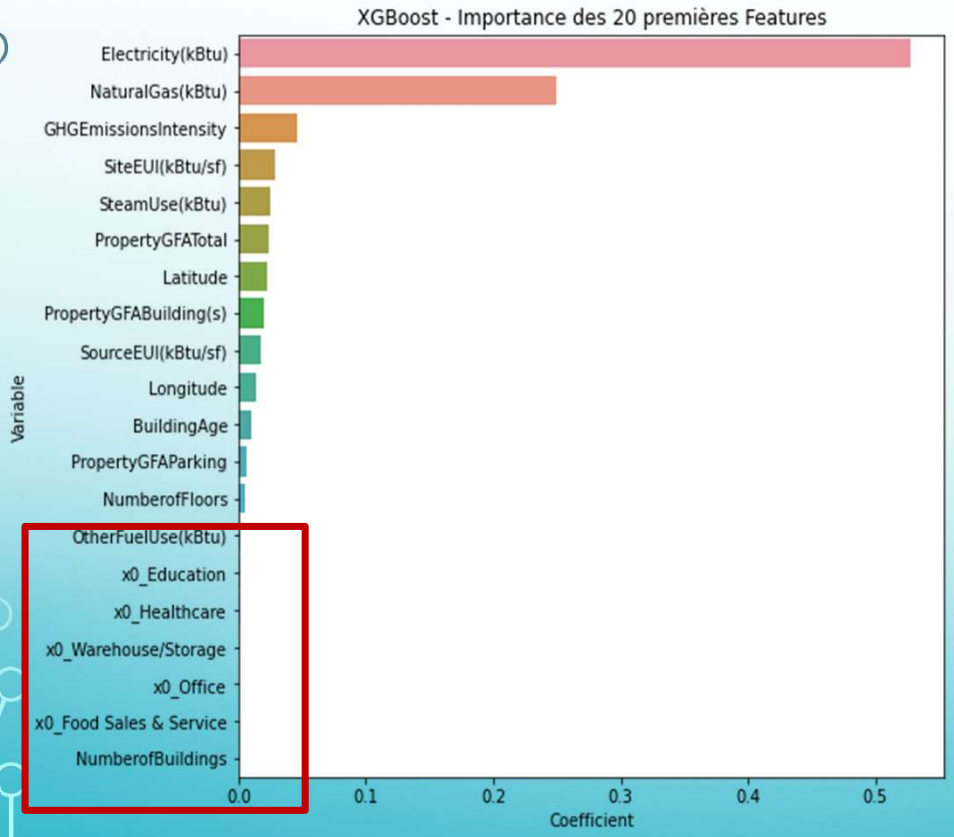

DummyRegressor
variable
RMSE
MAE
R2
time
score
DummyRegressor
Modèle


Comparaison des performances des modèles (jeu de test)
consommation
variable
RMSE
MAE
R²
Time_ms
score
XGBoost
Random Forest Regressor
SVR
LinearRegression
Modèle

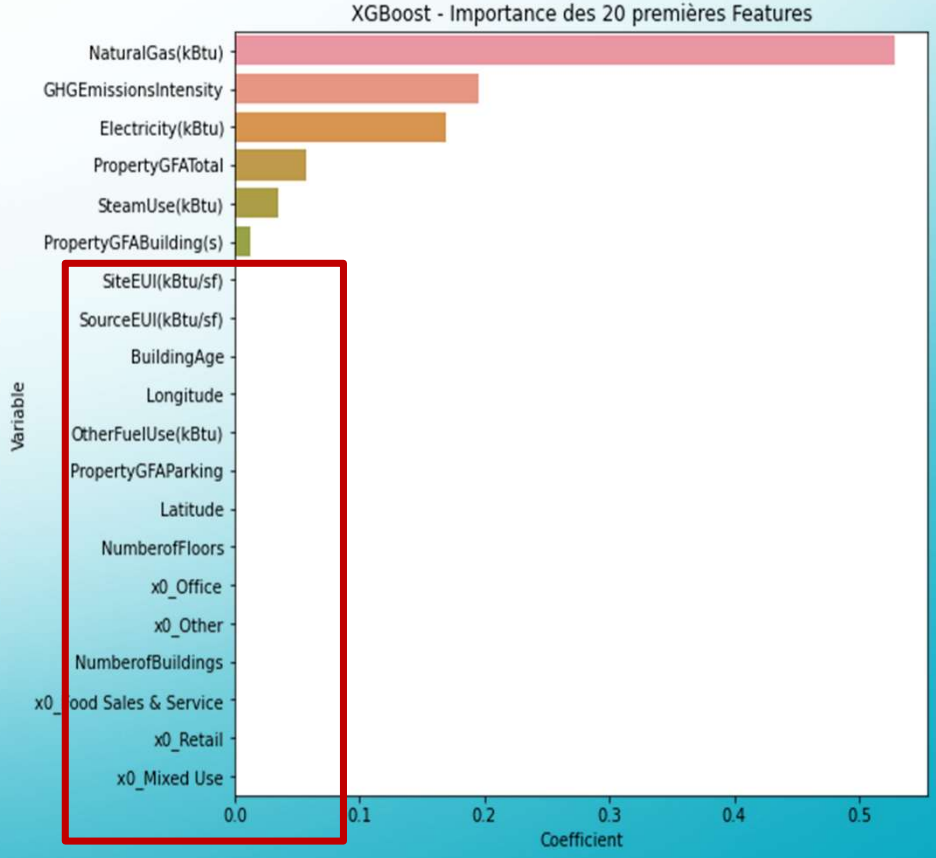

Comparaison des performances des modèles
Emission
variable
RMSE
MAE
R²
Time_ms
score
Random Forest Regressor
XGBoost
SVR
LinearRegression
Modèle

IMPORTANCE DES VARIABLES

Consommation



Emission CO2



l'intérêt de ENERGY STAR Score

Modèle	RMSE	MAE	R^2	Time	Cv_score(RMSE)
XGBoost	0.096436	0.067150	0.994559	1.710076	0.117061
XGBoost_Energy_Star	0.094976	0.065890	0.994723	1.703106	0.114633

MERCI DE
VOTRE ATTENTION