

II/2015

AKEB ADAM

PROJET 3

Anticipez les besoins en
consommations de bâtiments



Seattle



Introduction & Contexte

Présentation de l'entreprise

J'ai travaillé pour la ville de Seattle sur un projet de prédiction de la consommation énergétique et des émissions de CO₂ des bâtiments non résidentiels, dans le cadre de la stratégie "zéro carbone 2050".

L'objectif : utiliser les données structurelles des bâtiments pour estimer leur empreinte énergétique, et identifier les facteurs les plus impactants.



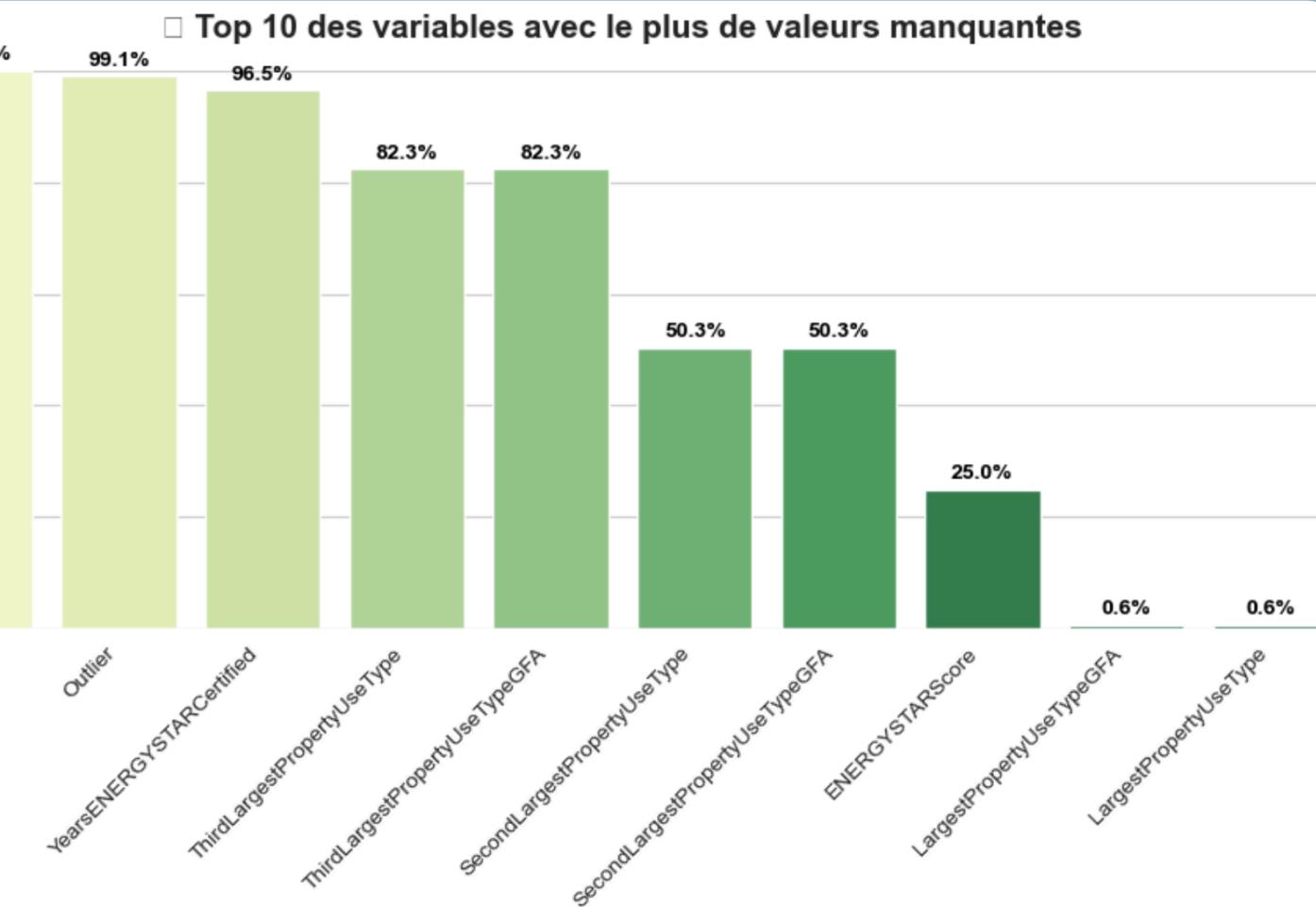
Jeu de données

Le jeu de données que j'ai utilisé provient de la ville de Seattle et regroupe environ 3 300 bâtiments mesurés en 2016.

Chaque ligne représente un bâtiment, et chaque colonne décrit une caractéristique : structurelle, géographique ou énergétique.

Les données contiennent à la fois :

- Des variables numériques (surface, âge, étages, score énergétique, etc.),
- Des variables catégorielles (type de bâtiment, quartier, statut de conformité...)



Les deux variables cibles à prédire :

- SiteEnergyUseWN(kBtu)** = consommation totale d'énergie indépendamment de la météo
- TotalGHGEmissions** = émissions totales de gaz à effet de serre (en tonnes de CO₂)

Sommaire

de la présentation

- 01 Nettoyage des données
- 02 Analyse exploratoire
- 03 Feature Engineering
- 04 Analyse des corrélations
- 05 Préparation du modèle
- 06 Modélisation
- 07 Optimisation & Validation croisée
- 08 Interprétation du modèle
- 09 Résultats finaux et comparaison

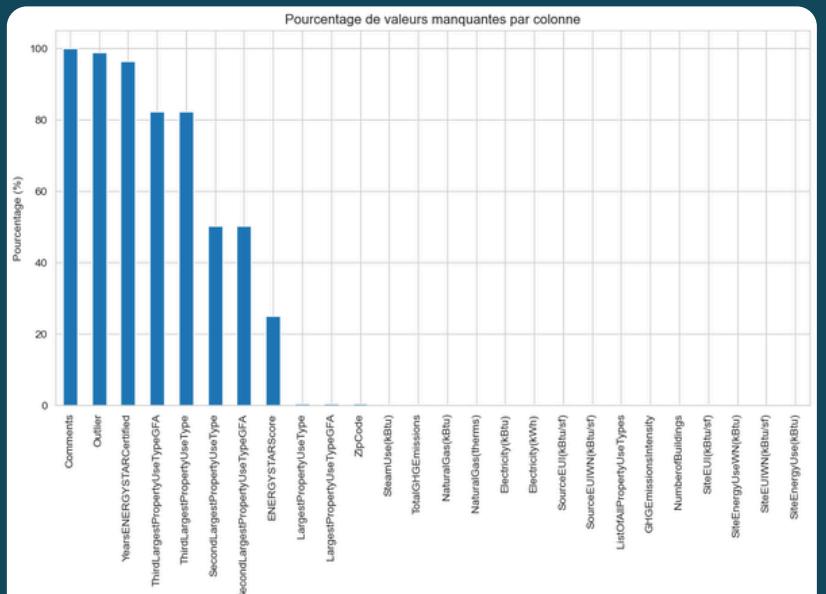
01

Nettoyage des données

Nombre de lignes avant filtrage : 3376
Nombre de colonnes avant filtrage : 46

Nombre de lignes après filtrage : 1629
Nombre de colonnes après filtrage : 39

Objectif 1



Objectif 2

- DataYear : 2016
- City : Seattle
- State : WA

Nombre de colonnes avant suppression : 46
Nombre de colonnes après suppression : 42

Nombre de colonnes avant suppression : 42
Nombre de colonnes après suppression : 39

Objectif 3

```
to_keep = [  
    'NonResidential',  
    'Nonresidential COS',  
    'Nonresidential WA',  
    'SPS-District K-12',  
    'Campus'
```

Nombre de lignes avant filtrage : 3376
Nombre de lignes après filtrage : 1668

Nombre de lignes avant suppression des résidentielles : 1668
Nombre de lignes résidentielles détectées : 39
après suppression : 1629

02

Analyse exploratoire

Les bâtiments les plus vastes sont aussi les plus consommateurs et les plus émetteurs.

Les bâtiments multi-usages (en orange) se distinguent globalement par des valeurs plus élevées sur ces trois axes. Quelques points isolés traduisent des situations extrêmes mais plausibles (par exemple, hôpitaux ou data centers).

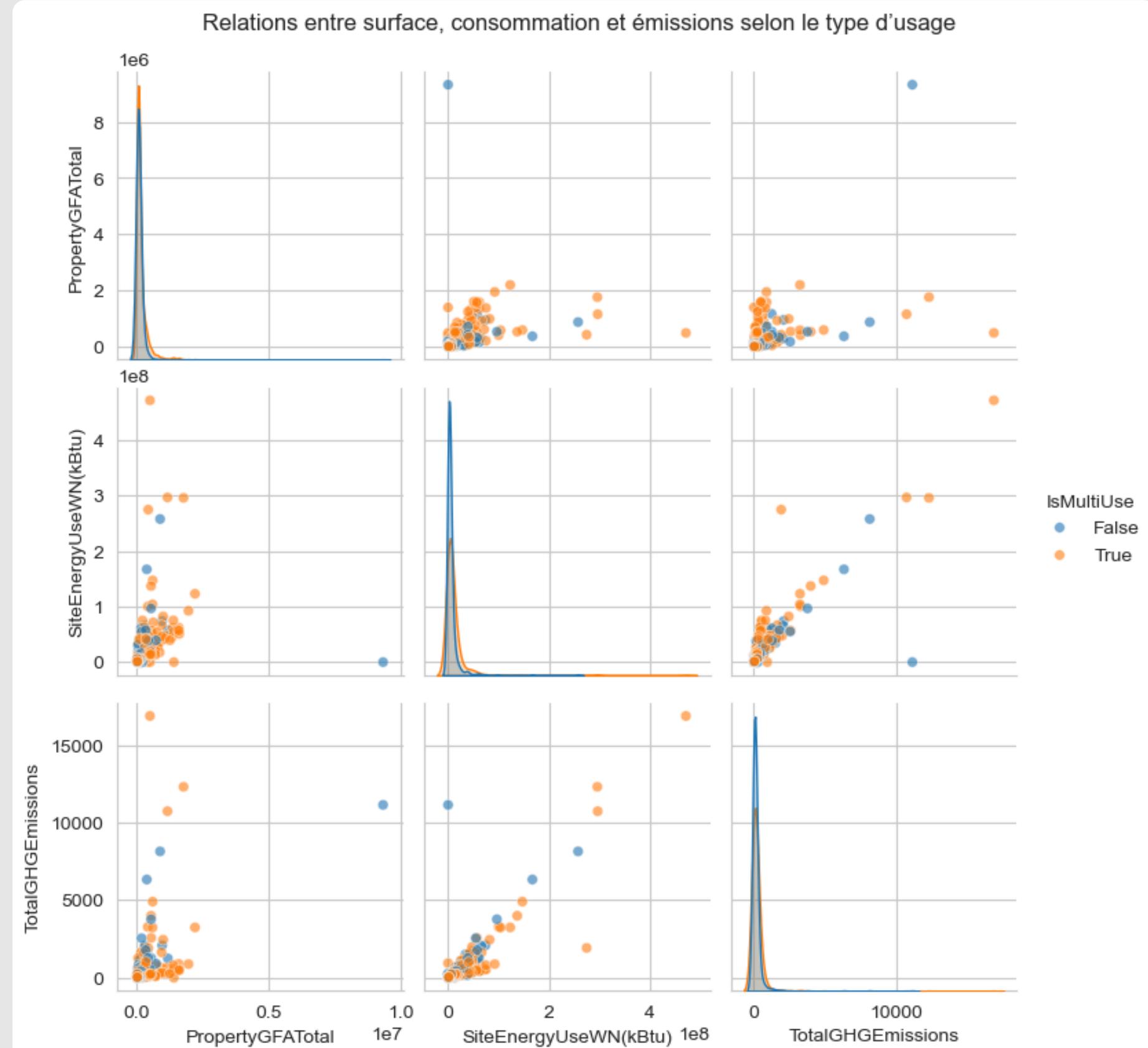
Ces observations confirment que :

La surface totale (PropertyGFATotal) est un prédicteur essentiel,

La variable IsMultiUse est pertinente pour différencier les profils énergétiques.

Les valeurs extrêmes devront être vérifiées mais probablement conservées, car elles traduisent des cas réels importants pour la politique énergétique de la ville.

VIRER TEXTE



03

Feature Engineering I

Caractéristiques
structurelles

*BuildingAge, BuildingAgeClass, IsOldBuilding, PropertyGFA_Total, IsLargeBuilding,
AvgFloorArea, ApproxBuildingVolume, HasParking, ParkingRatio*

Caractéristiques
d'usage

UseTypeCount, IsMultiUse

Indicateurs
énergétiques

EnergyIntensity, GasShare, ElectricShare, InefficiencyScore

Indicateurs
environnementaux

EmissionsIntensity, DistanceFromCenter_km

```
cols_to_drop = [
    'OSEBuildingID',                      # Identifiant unique (aucune valeur prédictive)
    'PropertyName',                        # Nom du bâtiment (texte non exploitable)
    'Address',                            # Adresse (texte, unique pour chaque ligne)
    'ZipCode',                            # Code postal : trop granulaire, peu informatif sans encodage spatial
    'TaxParcelIdentificationNumber',       # Identifiant administratif (aucun lien avec la consommation)
    'CouncilDistrictCode',                # Code politique, non pertinent pour la consommation énergétique
    'Neighborhood',                       # Nom de quartier (texte, trop de catégories)
    'DefaultData',                         # Donnée interne au dataset (booléen sans signification métier)
    'ComplianceStatus',                  # Statut administratif (non explicatif)
    'ListOfAllPropertyUseTypes',          # Déjà résumé par IsMultiUse et UseTypeCount
    'LargestPropertyUseType',             # Texte, redondant avec PrimaryPropertyType
]
```

Nombre de colonnes avant suppression : 55
Nombre de colonnes après suppression : 41

04

Analyse des corrélations

surface totale
(PropertyGFATotal)

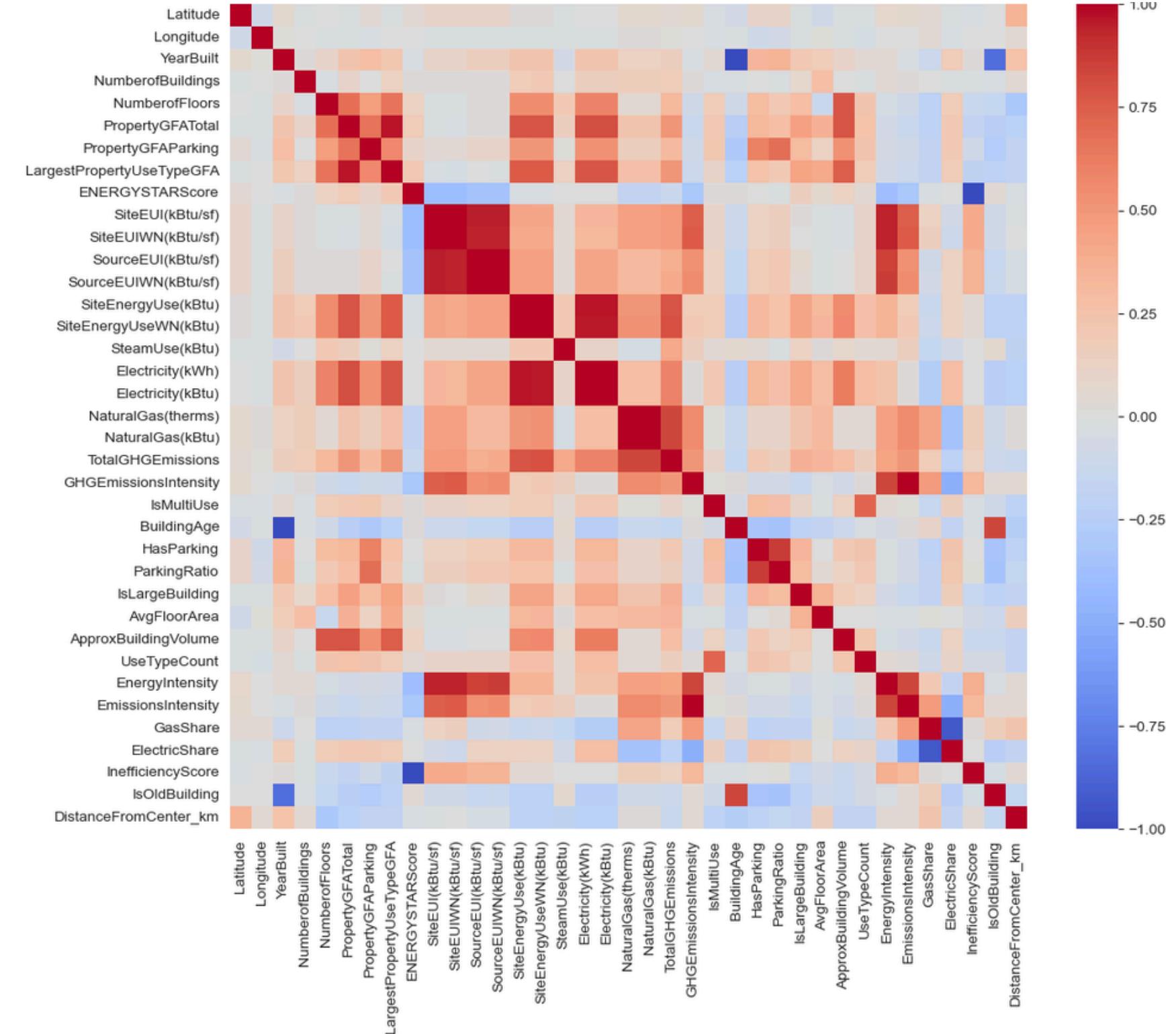
consommation énergétique
(SiteEnergyUseWN)

Émissions
(TotalGHGEmissions)

```
cols_to_drop_corr = [
    'IsLargeBuilding', # dérivé de PropertyGFATotal
    'HasParking', # dérivé de PropertyGFAParking
    'IsOldBuilding', # dérivé de BuildingAge
    'EmissionsIntensity', # redondant avec EnergyIntensity
    'SiteEUI(kBtu/sf)', # doublon non normalisé
    'SourceEUI(kBtu/sf)', # doublon non normalisé
    'SourceEUIWN(kBtu/sf)' # doublon de SiteEUIWN(kBtu/sf)
]
```

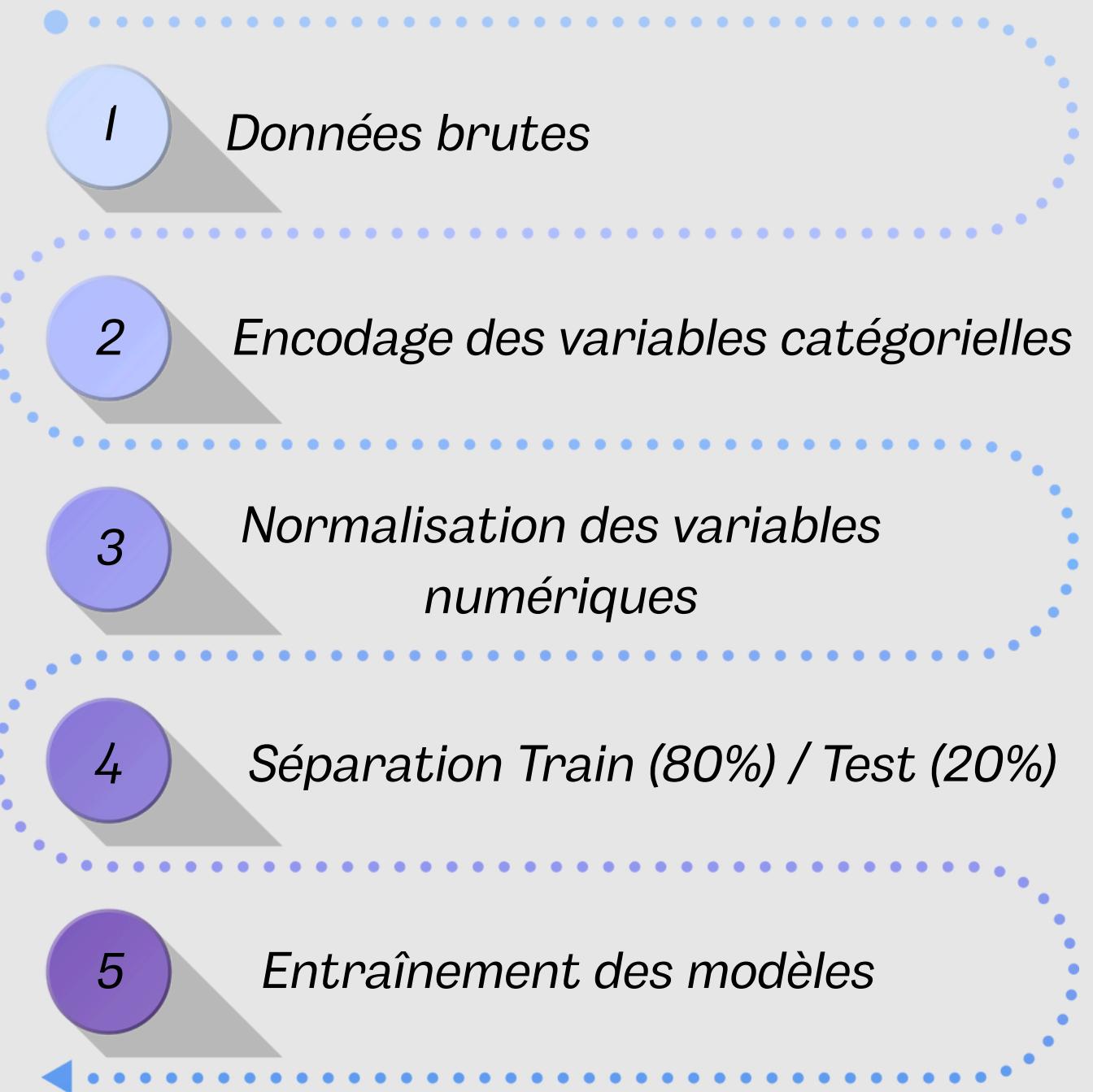
Nombre de colonnes avant suppression des corrélations : 41
Nombre de colonnes après suppression des corrélations : 34

Matrice de corrélation – Variables quantitatives



Étapes de préparation

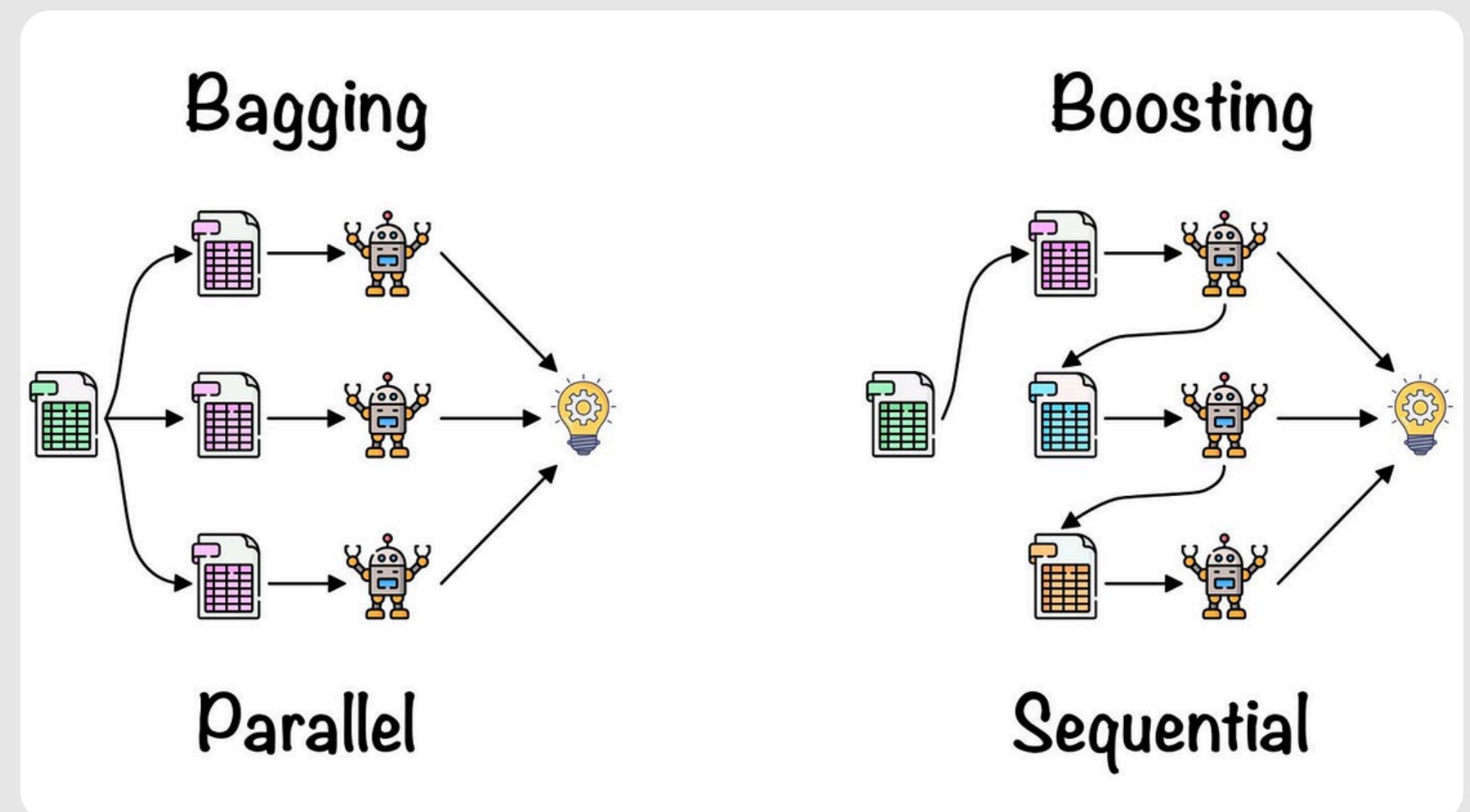
- **Encodage** des variables catégorielles → OneHotEncoder
- **Normalisation** des variables numériques → StandardScaler
- Séparation du jeu de données : 80 % entraînement / 20 % test
- Double cible :
 - SiteEnergyUseWN(kBtu) → Consommation énergétique
 - TotalGHGEmissions → Émissions de CO₂
- Objectif : garantir un apprentissage stable et des comparaisons équitables entre modèles



Données préparées pour l'entraînement:
X_train shape: (1241, 56)
X_train_scaled shape: (1241, 56)
y_energy_train shape: (1241,)
y_emissions_train shape: (1241,)

Expliquer le bagging et boosting

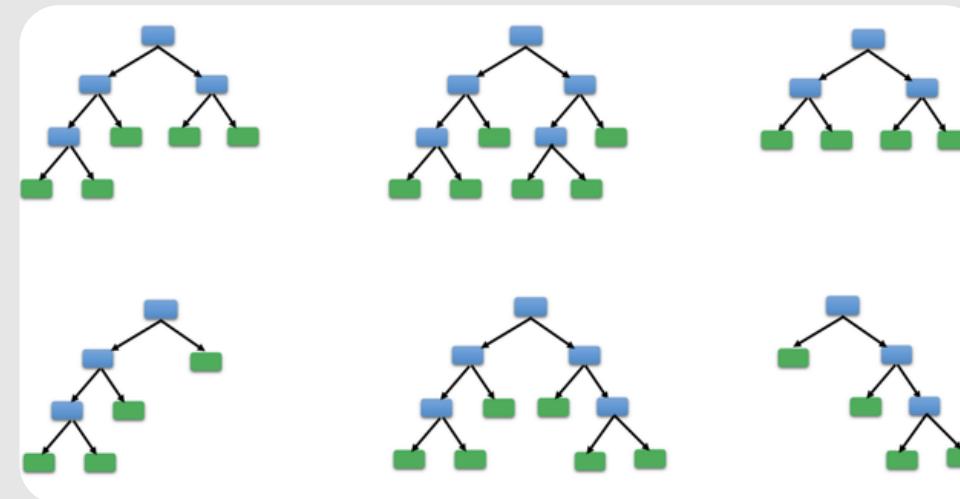
- Bagging = modèles en parallèle, indépendants → réduit variance
- Boosting = modèles en série, dépendants → réduit biais



Random Forest & LightGBM

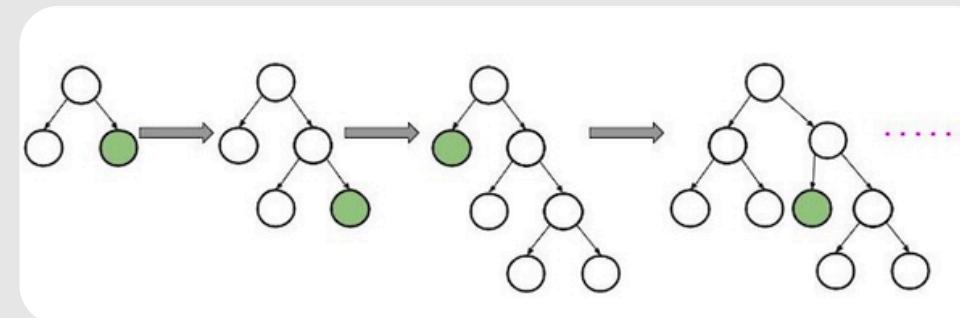
Random Forest :

- *n_estimators* → combien d'arbres
- *max_depth* → profondeur
- *max_features* → diversité des arbres
- *min_samples_split / leaf* → éviter l'overfit
- *bootstrap* → essentiel pour le bagging



LightGBM :

- *num_leaves* → complexité : le plus important
- *learning_rate* → vitesse d'apprentissage
- *n_estimators* → nombre d'arbres
- *feature_fraction / bagging_fraction* → anti-overfit
- *min_data_in_leaf* → feuilles minimum
- *lambda LI/L2* → régularisation



08

Modélisation

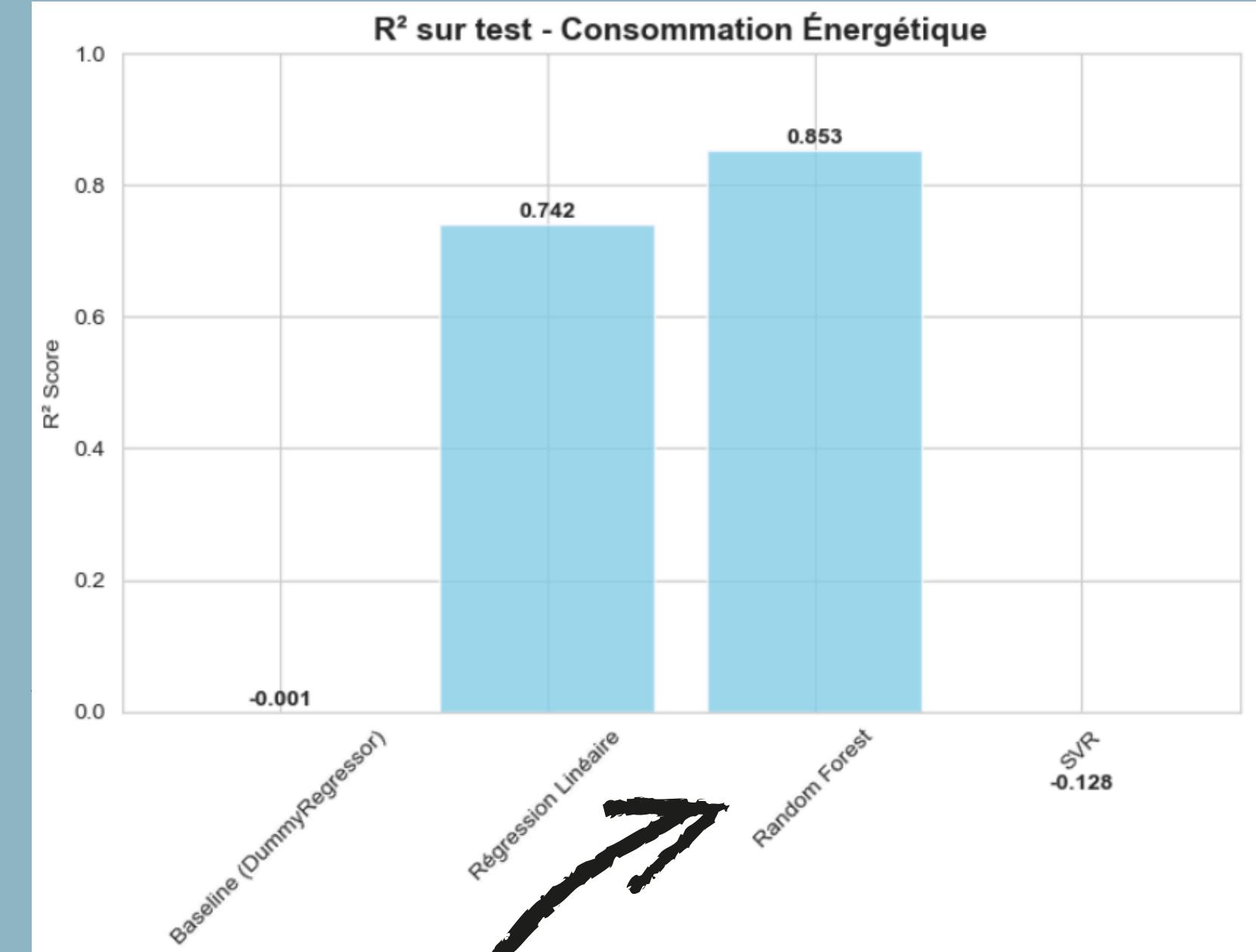
Objectif : prédire SiteEnergyUseWN(kBtu) & TotalGHGEmissions

Modèles testés :

- Baseline (DummyRegressor)
- Régression Linéaire
- SVR (Support Vector Regressor)
- Random Forest

Métriques utilisées :

- R² (qualité de fit)
- MAE (erreur moyenne)
- RMSE (grandes erreurs)
- Scaling appliqué pour les modèles sensibles à l'échelle (LinReg, SVR)



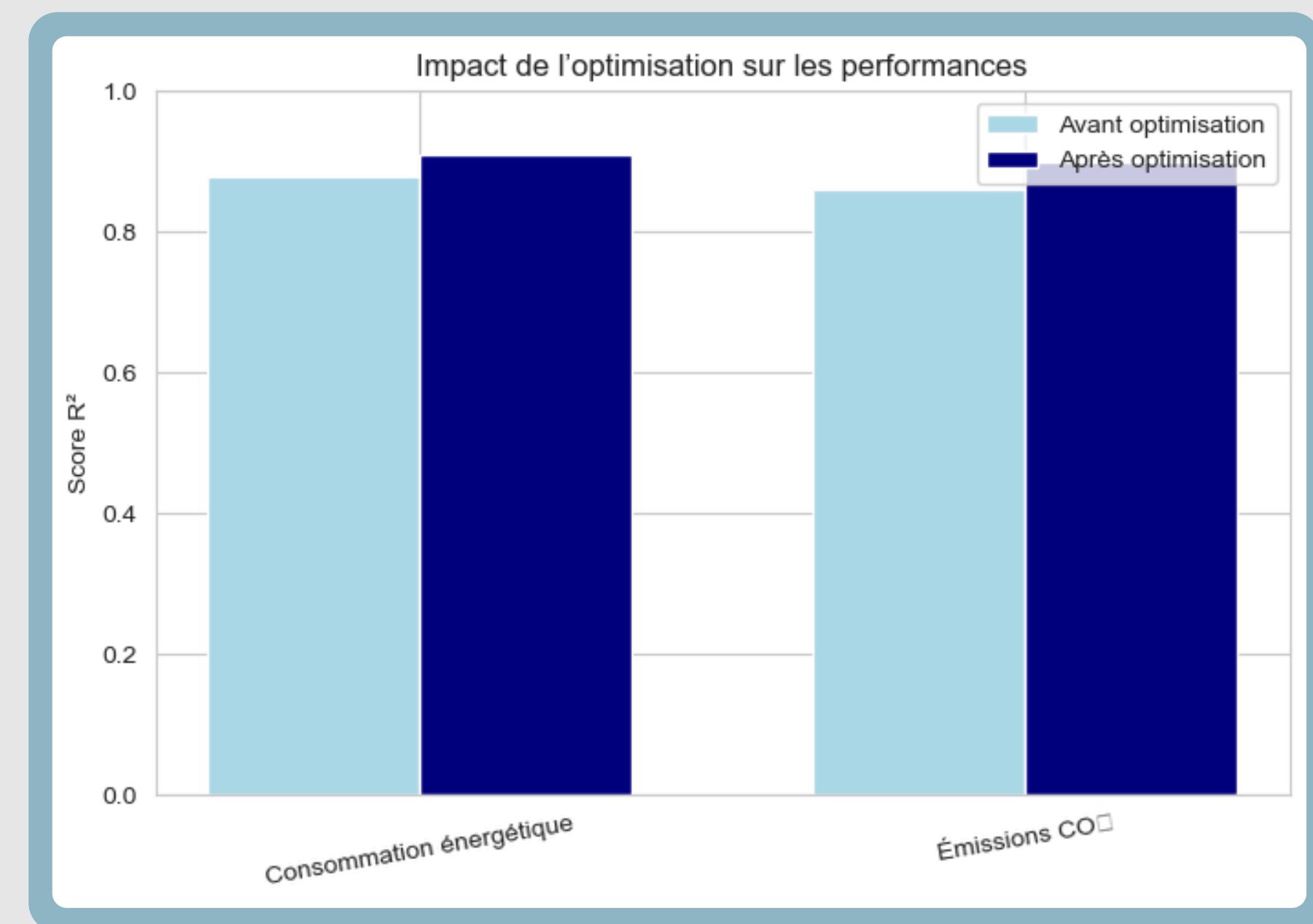
Un R²_test autour de 0.85

Optimisation & Validation croisée

Validation croisée (5-fold)

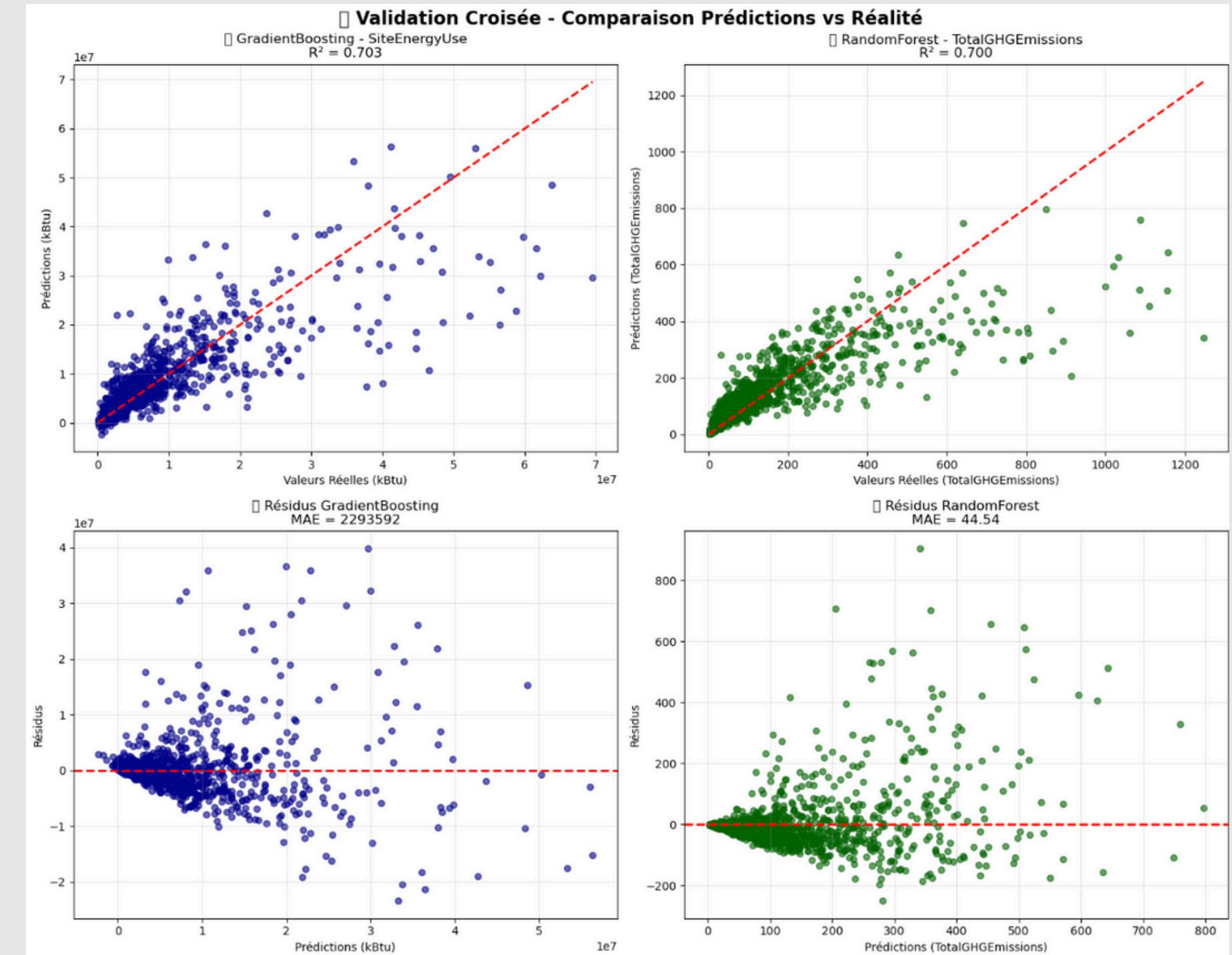
- 5-fold cross-validation sur le meilleur modèle
- Mesure : R^2 moyen sur chaque fold
- Faible variance entre folds → modèle stable
- Vérification de l'overfitting :
- écart Train vs Test < 0.1 → pas de surapprentissage

Les performances restent cohérentes entre les folds → le modèle généralise bien.



Interprétation du modèle

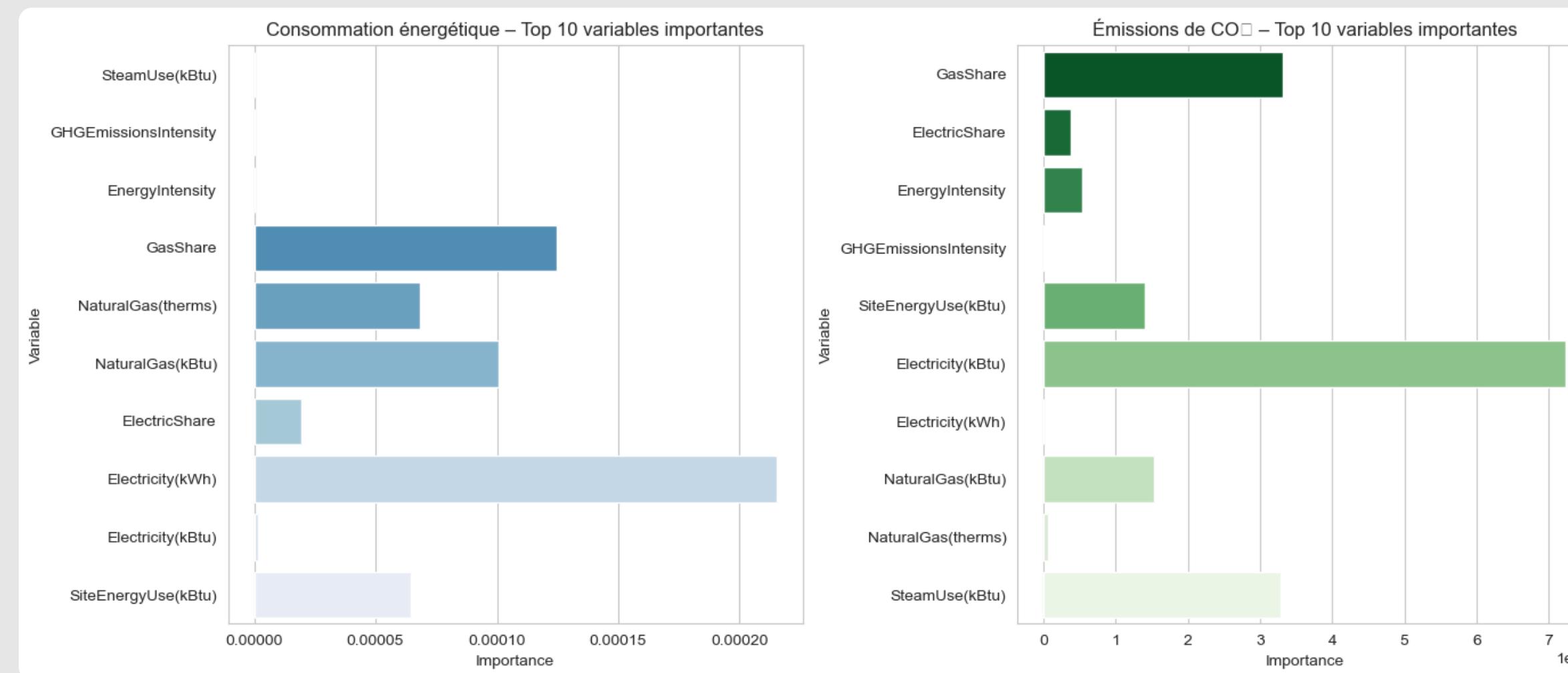
Les deux modèles montrent de bonnes performances en validation croisée, confirmant leur capacité à généraliser sur de nouvelles données. La validation croisée nous donne confiance dans la robustesse de nos prédictions.



II

Interprétation du modèle

- Analyse réalisée avec la feature *importance* de la Random Forest et la permutation *importance* pour les autres modèles.
 - Ces indicateurs mesurent la contribution de chaque variable à la qualité des prédictions.
 - On identifie les facteurs les plus influents sur la consommation d'énergie et les émissions.



Résultats finaux et comparaison

- *Les modèles finaux (optimisés) ont été évalués sur les jeux de test.*
- *L'optimisation a permis d'améliorer le score R^2 tout en réduisant les erreurs (MAE, RMSE).*
- *Les performances sont stables entre train et test, ce qui confirme la bonne généralisation du modèle.*

Échantillon de prédictions – Random Forest (Energy):
Échantillon 1: Réel=62160304.00, Prédit=46641633.60, Erreur=15518670.40 (25.0%)
Échantillon 2: Réel=8886716.00, Prédit=12964131.11, Erreur=4077415.11 (45.9%)
Échantillon 3: Réel=7437745.00, Prédit=20743772.51, Erreur=13306027.51 (178.9%)
Échantillon 4: Réel=3220635.25, Prédit=2471620.38, Erreur=749014.87 (23.3%)
Échantillon 5: Réel=3034711.50, Prédit=2844402.21, Erreur=190309.29 (6.3%)

Les modèles finaux présentent une amélioration stable après optimisation, sans surapprentissage.

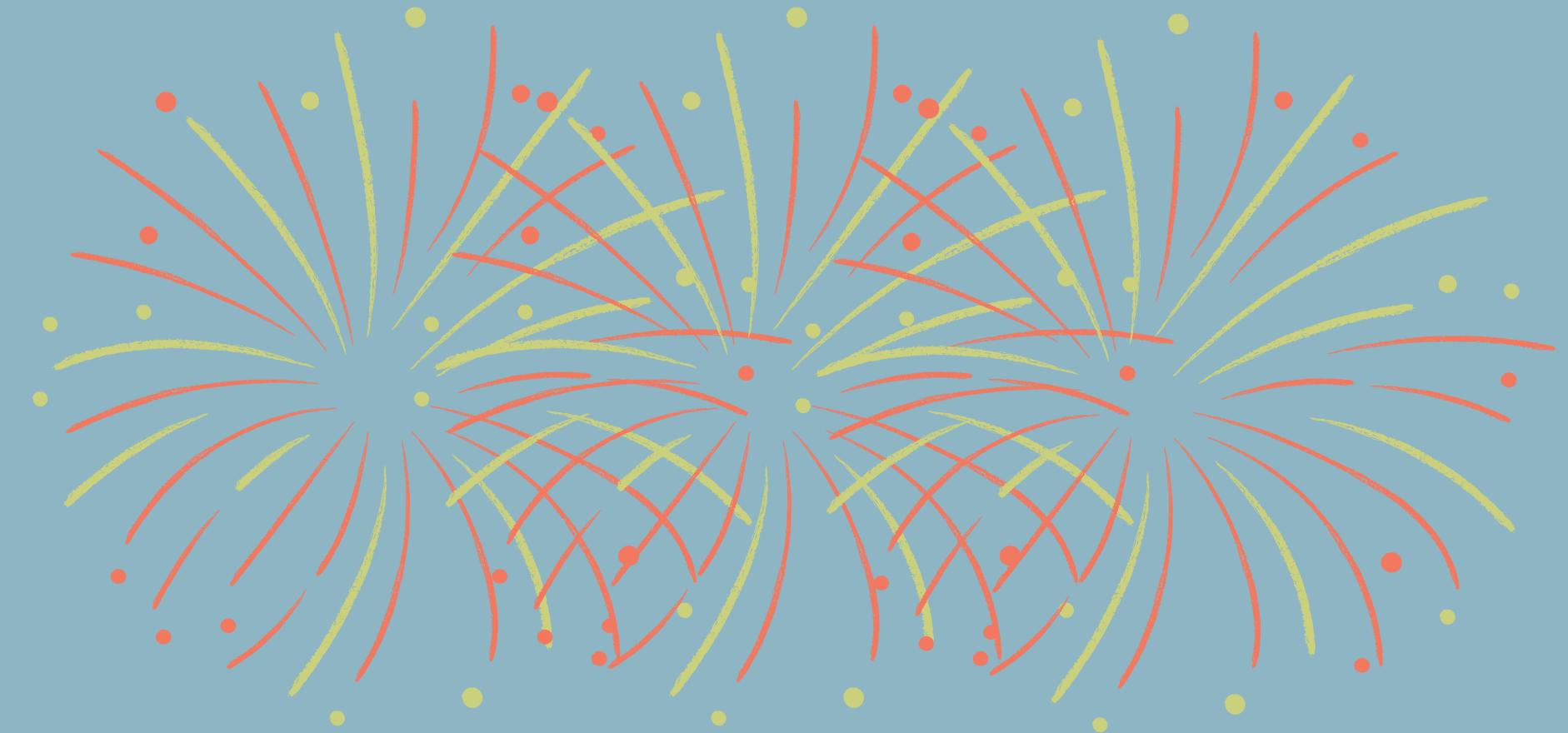
Ils capturent efficacement la relation entre les caractéristiques des bâtiments et leur impact énergétique et environnemental.

II/2015

AKEB ADAM

PROJET 3

Anticipez les besoins en
consommations de bâtiments



Merci pour
votre attention