Projet 4 : Anticipez les besoins en consommation de bâtiments

Neutralité carbone 2050

Abdessalem BOUZAYANI

5/12/2022

Sommaire

- Problématique
- Jeu de données
- Analyse exploratoire
- Prédiction consommation d'énergie
- Prédiction émission de GES
- Conclusions et perspectives

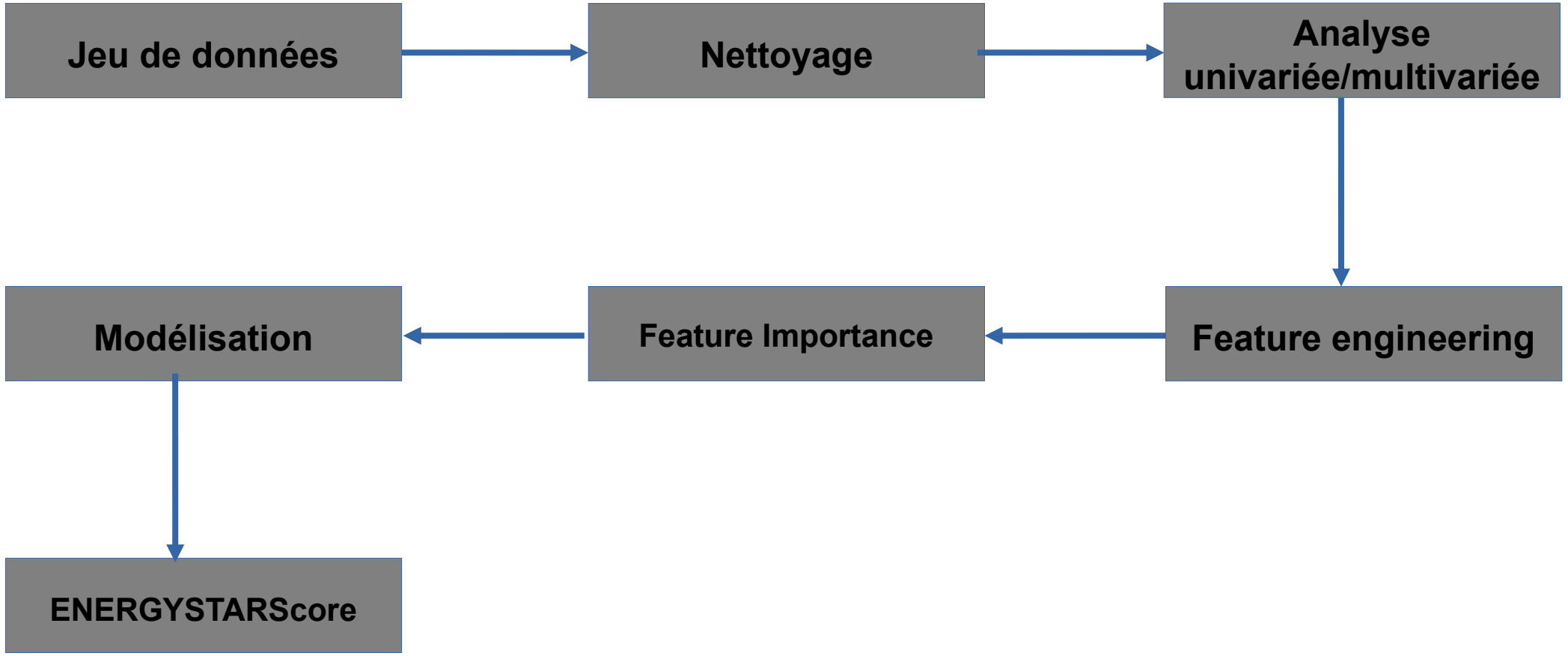
Problématique

- **Ville neutre en émissions de carbone en 2050**

- **A partir d'un jeu de données, on doit répondre aux questions :**

  - **Analyse exploratoire**
  - **Prédiction des émissions de CO2 et de la consommation totale d'énergie**
  - **Tester différents modèles de prédiction**
  - **Évaluer l'intérêt de l'ENERGY STAR Score**

Feuille de route



Jeu de données

Seattle

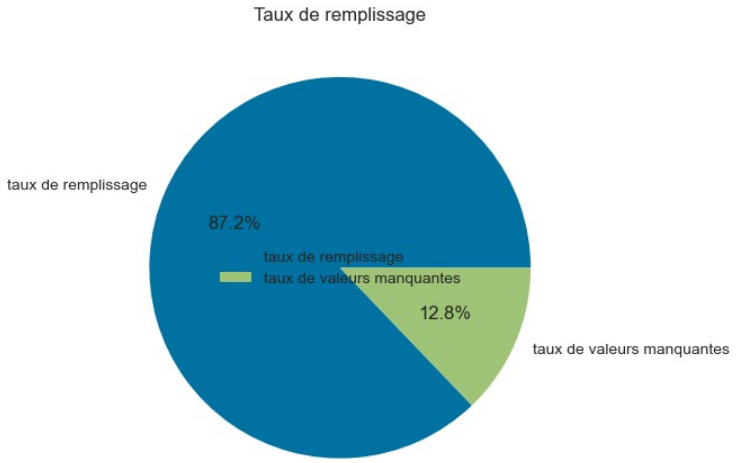
Consommation d'énergie : SiteEnergyUse_kBtu
Emission de GES : TotalGHGEmissions

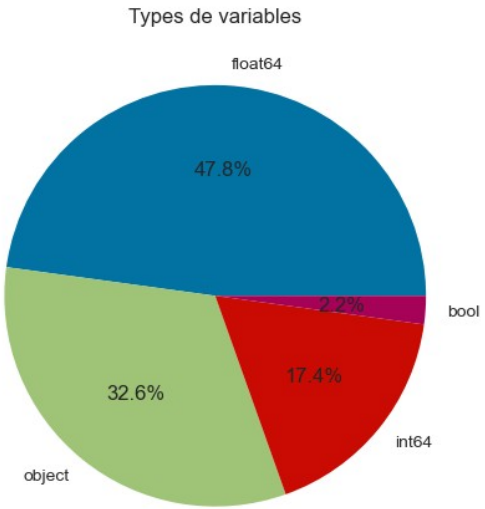
46 colonnes

(informations géographiques, informations générales, informations énergitiques

3376 lignes

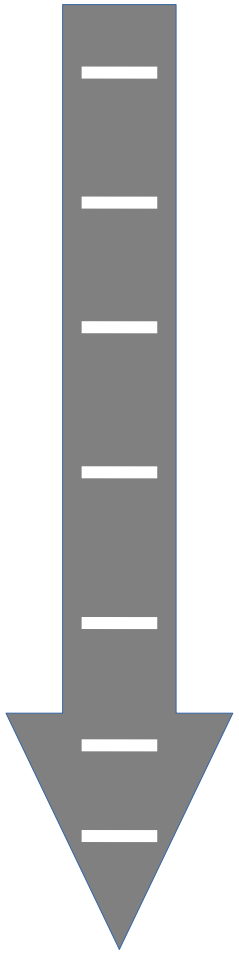
(Immeuble)





Nettoyage

- Supprimer les colonnes et lignes vides
- Supprimer les lignes dupliquées
- Supprimer les bâtiments résidentiels
- Sélectionner les colonnes utiles
- Traiter les valeurs manquantes
- Traiter les valeurs aberrantes



3376, 46

3376, 45

3376, 45

1668, 45

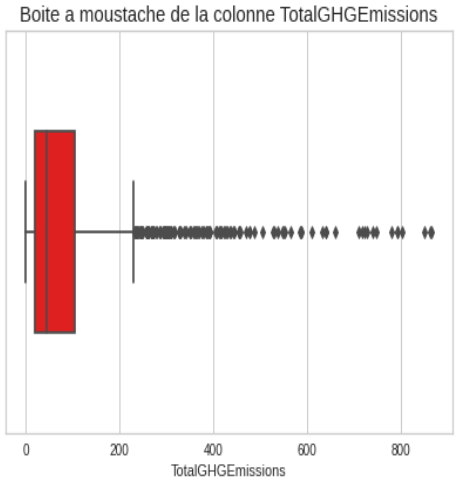
1668, 25

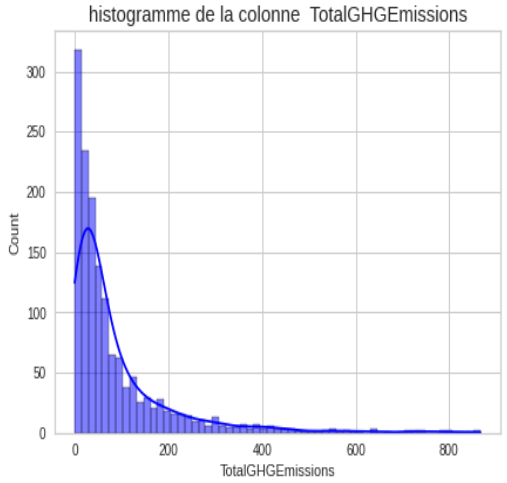
1666, 25

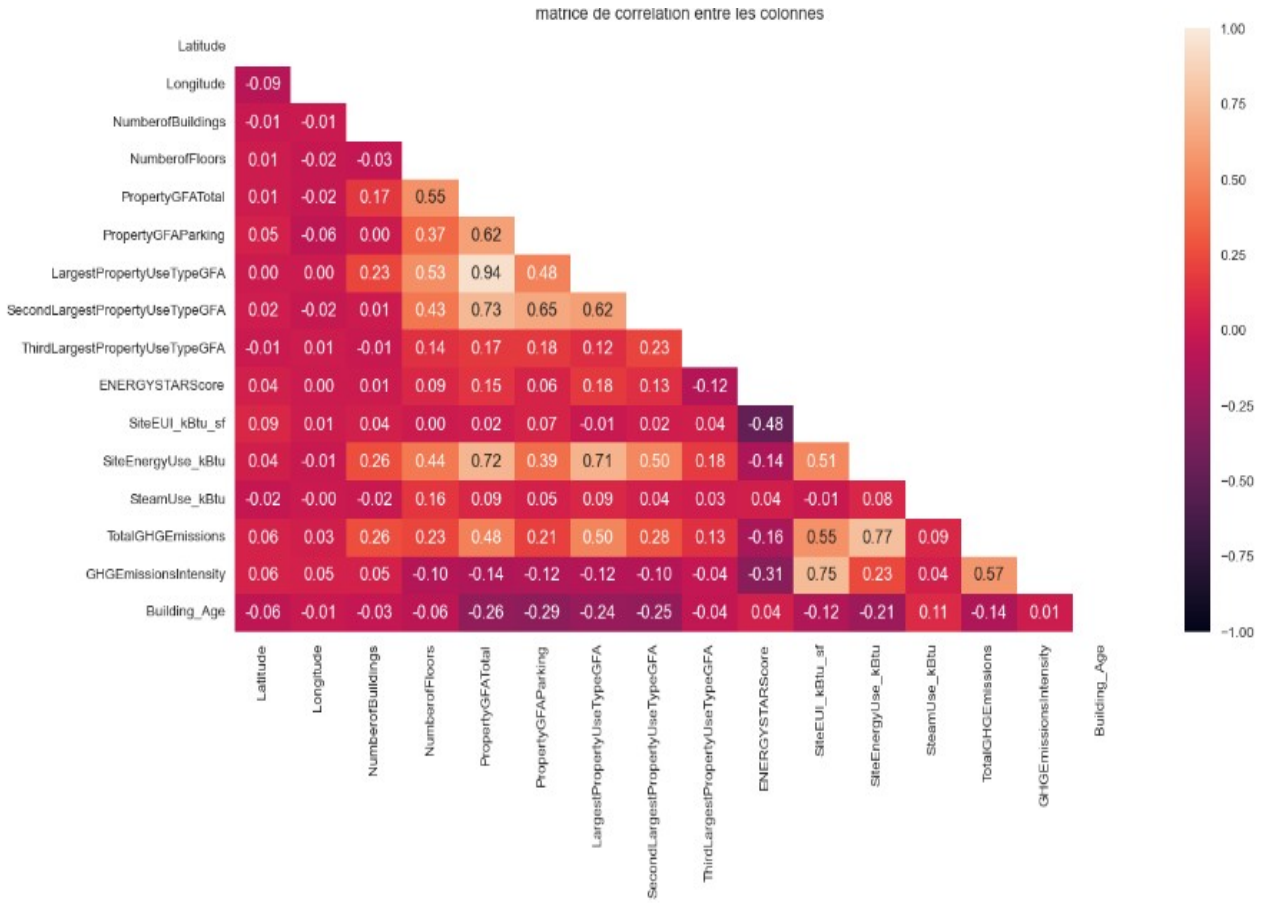
1552,25

Analyse univariée/multivariée

```
moyenne : 88.72
mediane : 44.3
mode : 0    6.3
dtype: float64
variance : 14935.61
skewness : 2.95
kurtosis : 10.87
ecart type : 122.21
min : 0.12
25% : 18.68
50% : 44.3
75% : 103.63
max : 866.23
```





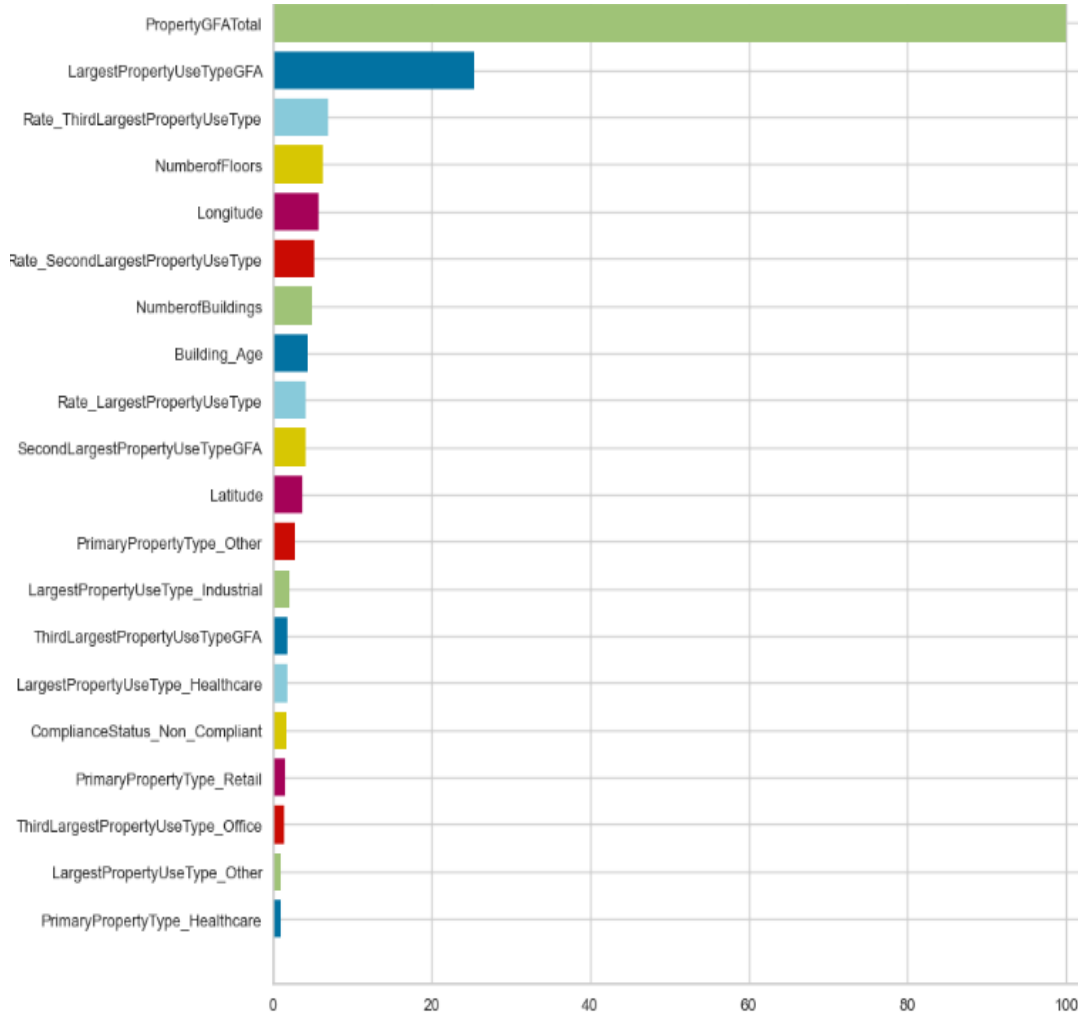


Feature engineering

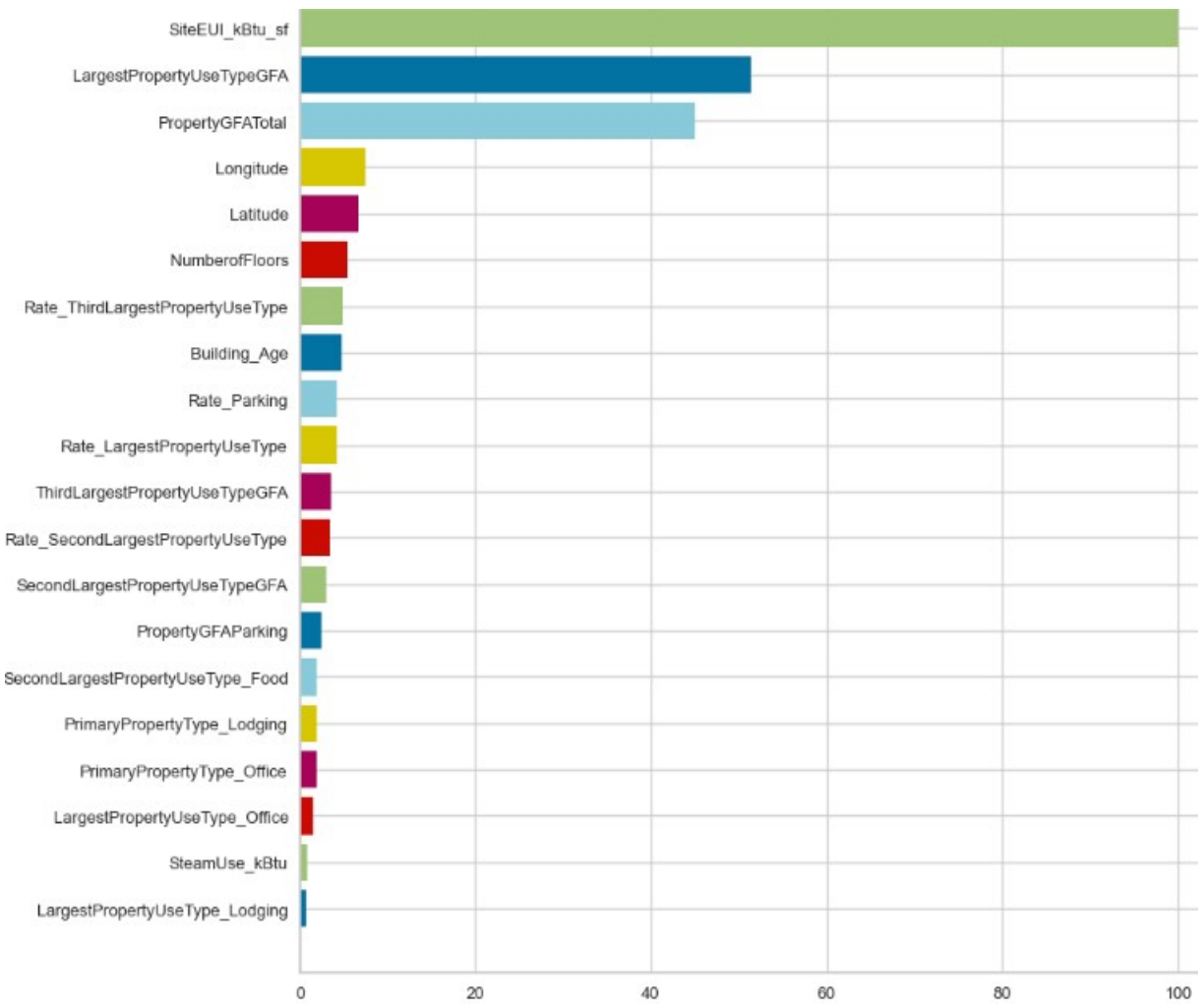
Variables	Transformations
DataYear, YearBuilt	Building_Age
PrimaryPropertyType SecondLargestPropertyUseType ThirdLargestPropertyUseType,LargestPropertyUseType	Dictionnaire de données
Address	New_Address (rue,avenue,...)
PropertyGFAParking LargestPropertyUseTypeGFA SecondLargestPropertyUseTypeGFA ThirdLargestPropertyUseTypeGFA	Rate_Parking Rate_LargestPropertyUseType Rate_SecondLargestPropertyUseType Rate_ThirdLargestPropertyUseType
Variables catégorielles	One hot encoding
Variables continues	Normalisation, log + 1

Feature importance

Consommation d'énergie



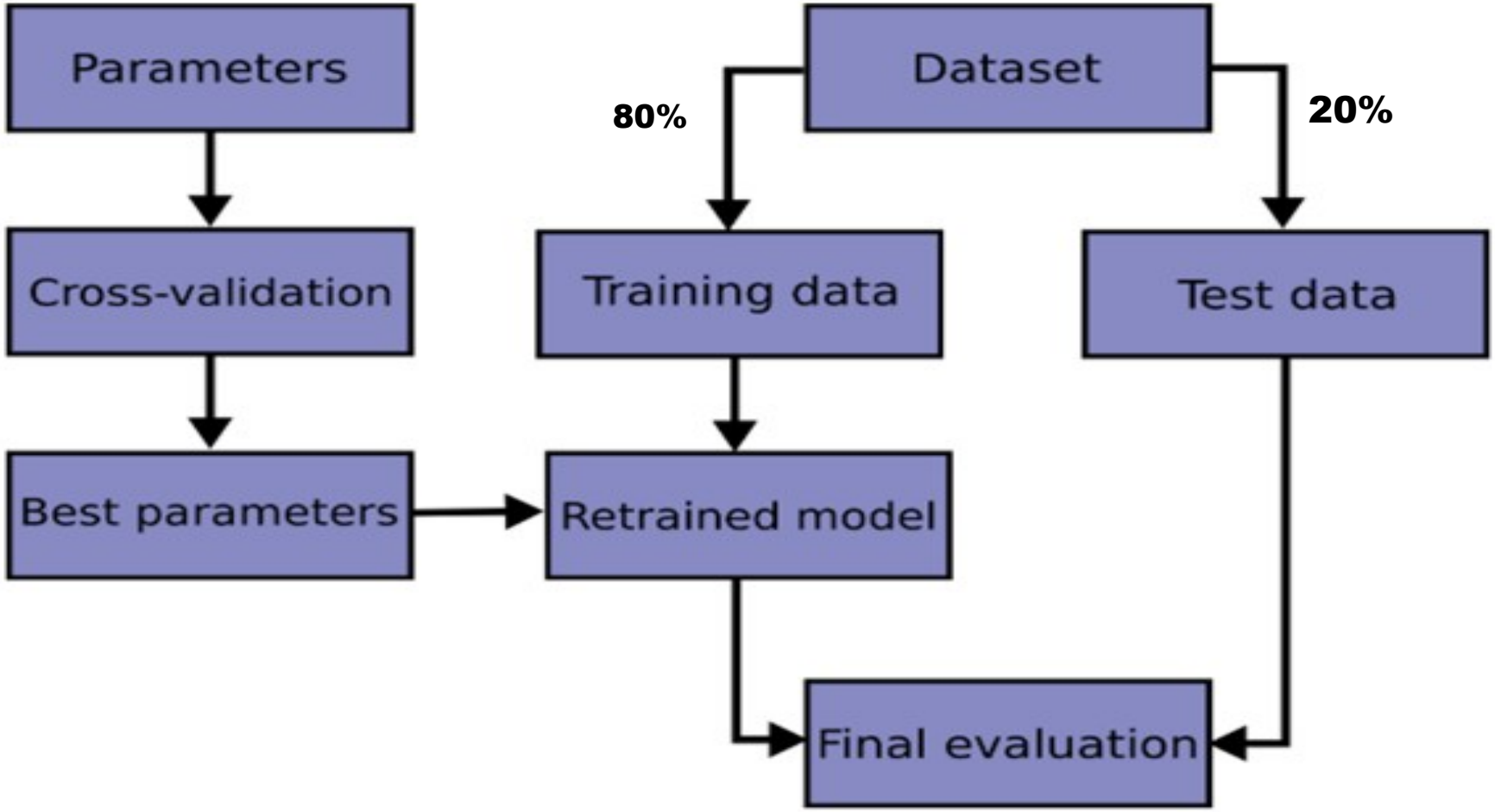
Émission de GES



Modélisation

Modèle de régression	Type
DummyRegressor	naif
LinearRegression	linéaire
Lasso	linéaire
Ridge	linéaire
ElasticNet	linéaire
KernelRidge	à noyaux
Svr	à noyaux
KNeighborsRegressor	non linéaire
DecisionTreeRegressor	non linéaire
ExtraTreesRegressor	ensembliste
RandomForestRegressor	ensembliste
AdaBoostRegressor	ensembliste
GradientBoostingRegressor	ensembliste
XGBRegressor	ensembliste
LGBMRegressor	ensembliste

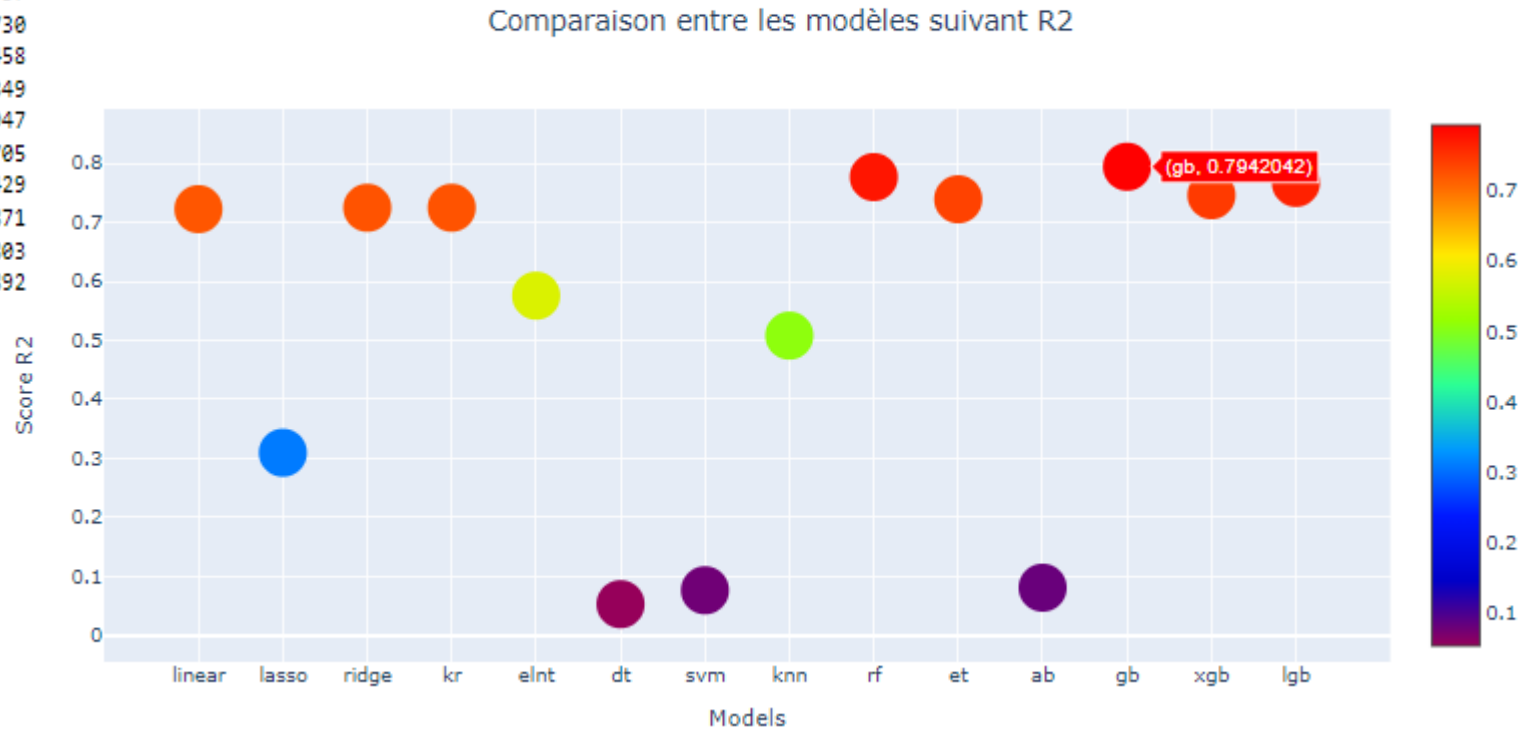
Modélisation



Modélisation - Consommation d'énergie

Initialisation : par défaut
Transformation : RobustScaler

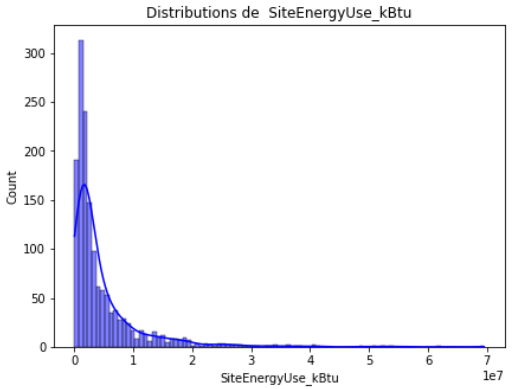
	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	gb	0.794204	0.332402	0.576544	0.284931	1.265635
1	rf	0.776681	0.360706	0.600588	0.298888	1.013917
2	lgb	0.766847	0.376589	0.613669	0.306230	0.453443
3	xgb	0.746016	0.410236	0.640497	0.323316	0.750599
4	et	0.739074	0.421448	0.649190	0.309100	0.828937
5	kr	0.724677	0.444703	0.666861	0.406983	0.088730
6	ridge	0.724668	0.444717	0.666871	0.406981	0.013458
7	linear	0.722378	0.448417	0.669639	0.410122	0.016349
8	elnt	0.575639	0.685430	0.827907	0.536316	0.007947
9	knn	0.507964	0.794739	0.891481	0.475674	0.062705
10	lasso	0.309394	1.115471	1.056158	0.708723	0.014429
11	ab	0.080692	1.484871	1.218553	1.105398	0.671871
12	svm	0.076349	1.491886	1.221428	0.641423	0.406803
13	dt	0.052743	1.530014	1.236937	0.461613	0.072892

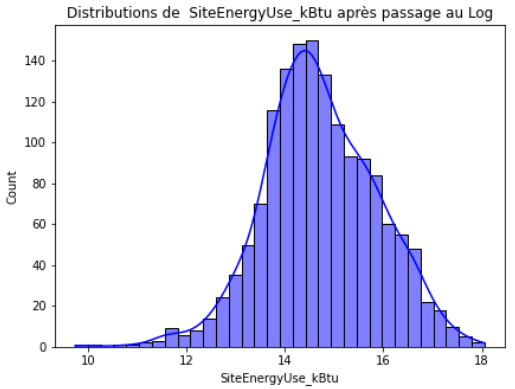


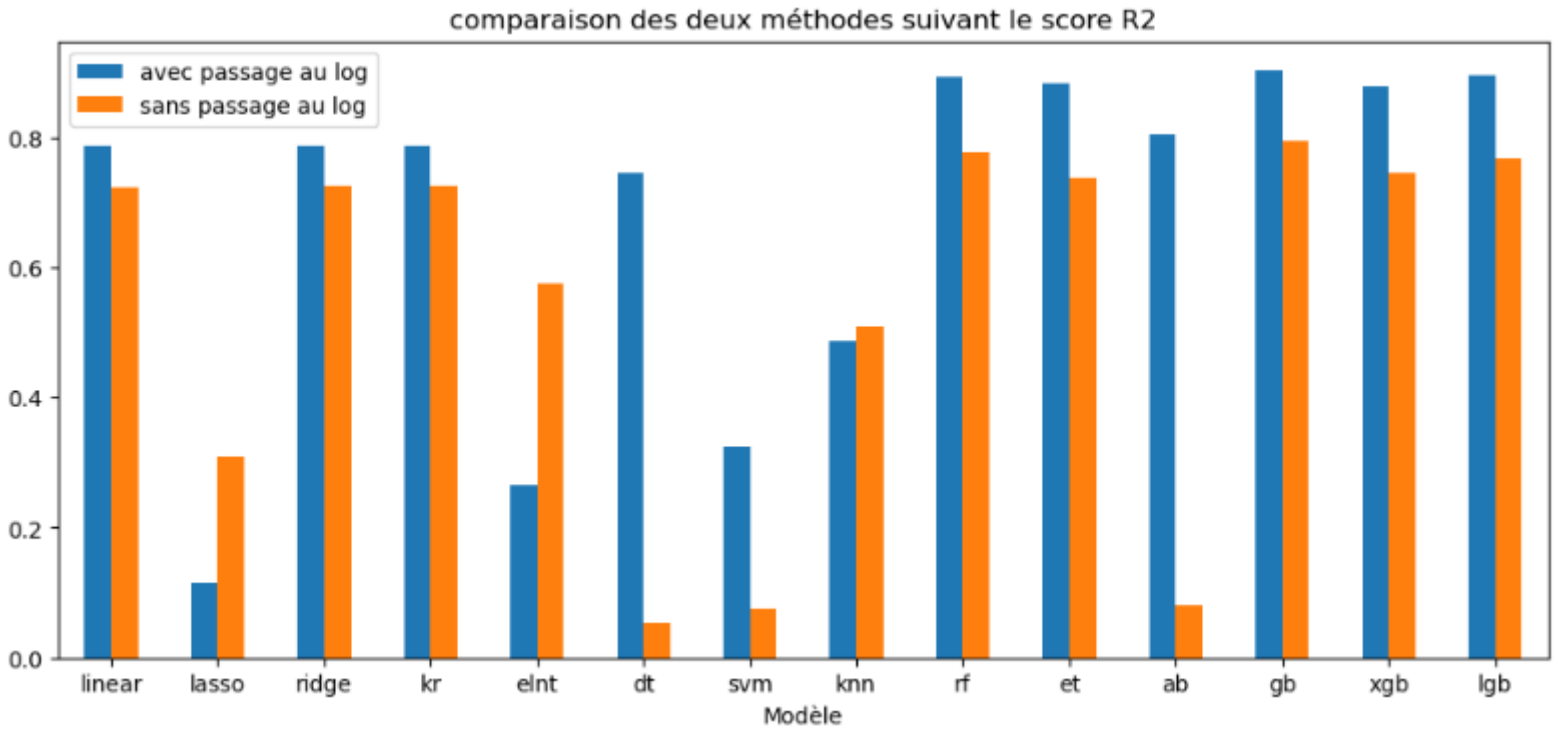
Modélisation - Consommation d'énergie

Initialisation : par défaut
Transformation : log + 1

Modèle	R2	MSE	RMSE	MAE	Durée_exéc
gb	0.902697	0.135824	0.368542	0.260889	0.506980
lgb	0.895342	0.146091	0.382219	0.270405	0.126202
rf	0.894191	0.147697	0.384314	0.278214	0.344403
et	0.884159	0.161702	0.402121	0.280648	0.364045
xgb	0.879172	0.168662	0.410685	0.291475	0.265576
ab	0.803740	0.273958	0.523410	0.420542	0.303368
kr	0.787842	0.296149	0.544196	0.359159	0.031111
ridge	0.787756	0.296269	0.544306	0.359413	0.004930
linear	0.787262	0.296959	0.544940	0.362979	0.008988
dt	0.744734	0.356323	0.596928	0.423906	0.031280
knn	0.487830	0.714932	0.845536	0.608576	0.031220
svm	0.324538	0.942870	0.971015	0.732392	0.216081
elnt	0.264291	1.026968	1.013394	0.782893	0.000000
lasso	0.114114	1.236599	1.112025	0.880037	0.006058







Modélisation - Consommation d'énergie

Optimisation

Modèle	Hyperparamètre	Valeur par défault	Grille de recherche	Meilleure
RandomForest	n_estimators	100	[10, 50, 100]	100
	min_samples_split	2	[2, 5, 10]	2
	min_samples_leaf	3	[1, 3, 4]	1
	max_depth	None	[10,50,100]	50
GradientBoosting	n_estimators	100	[100,500,1000]	500
	min_samples_split	2	[2, 5,10]	2
	min_samples_leaf	1	[1, 3,5]	1
	max_depth	3	[1,5,10]	10
	learning_rate	0,1	[0.01,0.1,0.5]	0.01
LGBM	n_estimators	100	[100,500,1000]	100
	max_depth	-1	[1,3,5,10]	10
	learning_rate	0,1	[0.01,0.05,0.1, 0.5,0.9]	0.1

Modélisation - Consommation d'énergie

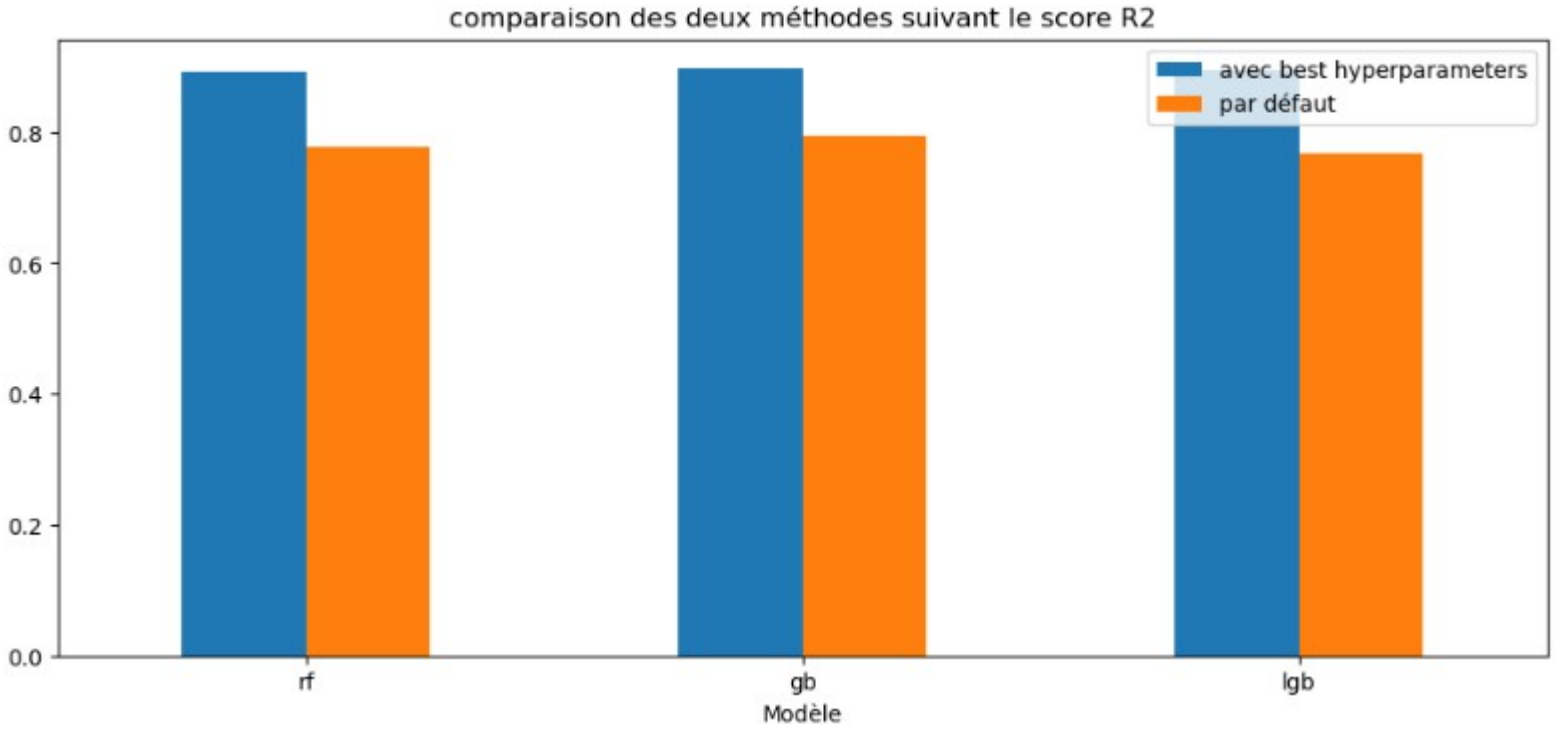
Comparaison avant et après optimisation

Par défaut

	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	rf	0.776681	0.360706	0.600588	0.298888	1.013917
1	gb	0.794204	0.332402	0.576544	0.284931	1.265635
2	lgb	0.766847	0.376589	0.613669	0.306230	0.453443

Avec meilleurs hyperparamètres

	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	rf	0.891692	0.151187	0.388827	0.281843	3.515804
1	gb	0.896740	0.144139	0.379656	0.267553	2.499980
2	lgb	0.893864	0.148155	0.384909	0.275171	0.375549



Modélisation - Consommation d'énergie

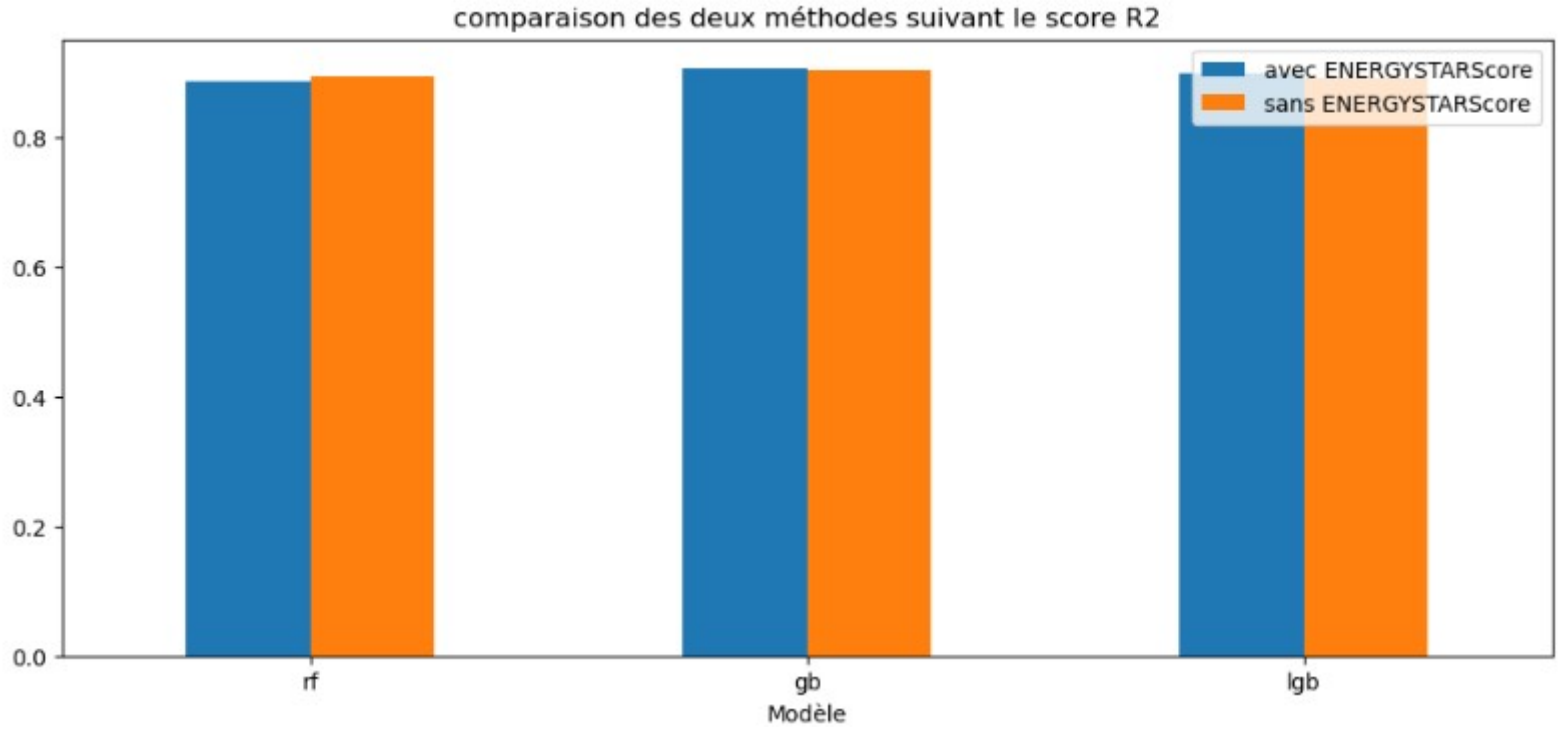
ENERGY STAR Score ?

Sans ENERGY STAR Score

	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	rf	0.891692	0.151187	0.388827	0.281843	3.515804
1	gb	0.896740	0.144139	0.379656	0.267553	2.499980
2	lgb	0.893884	0.148155	0.384909	0.275171	0.375549

Avec ENERGY STAR Score

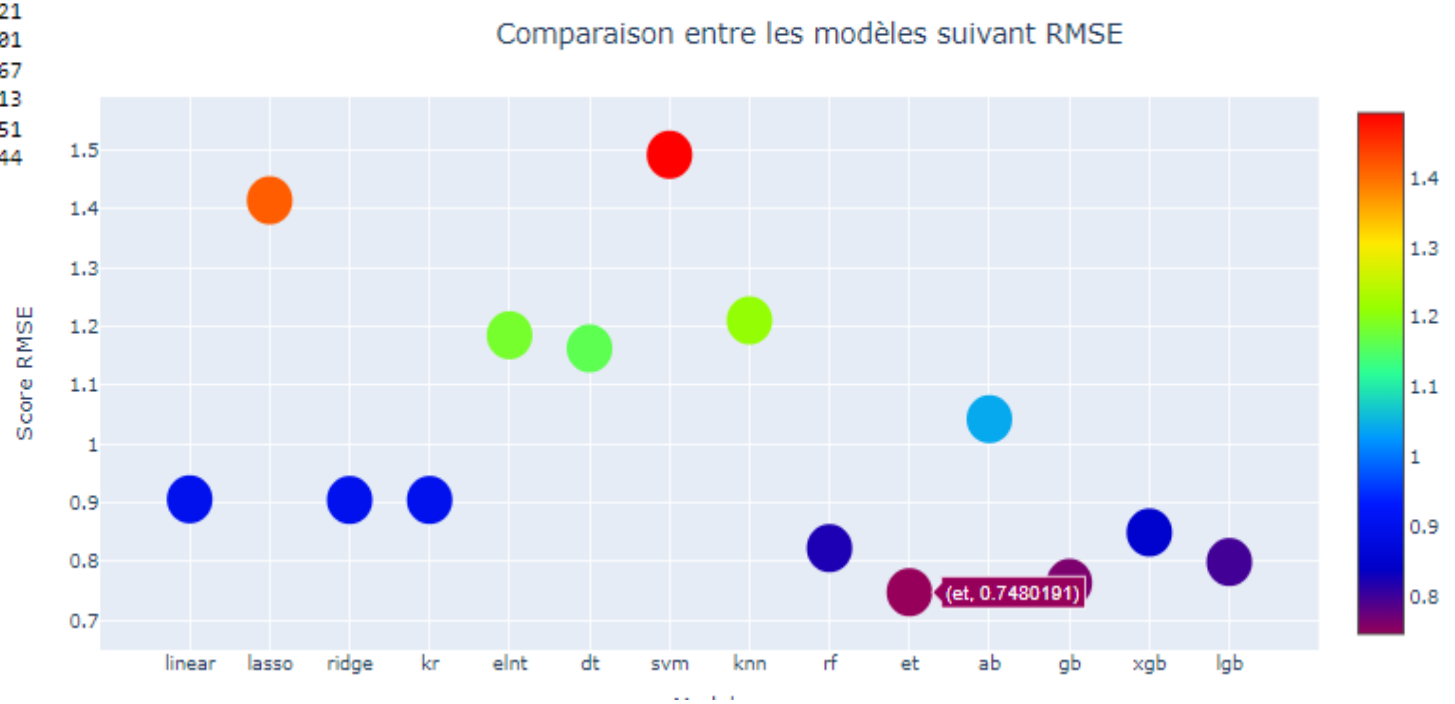
	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	rf	0.884901	0.139628	0.373669	0.261200	3.703316
1	gb	0.898456	0.123184	0.350976	0.246716	2.610486
2	lgb	0.903304	0.117303	0.342496	0.248315	0.344468



Modélisation – Emission de GES

Initialisation : par défaut
Transformation : RobustScaler

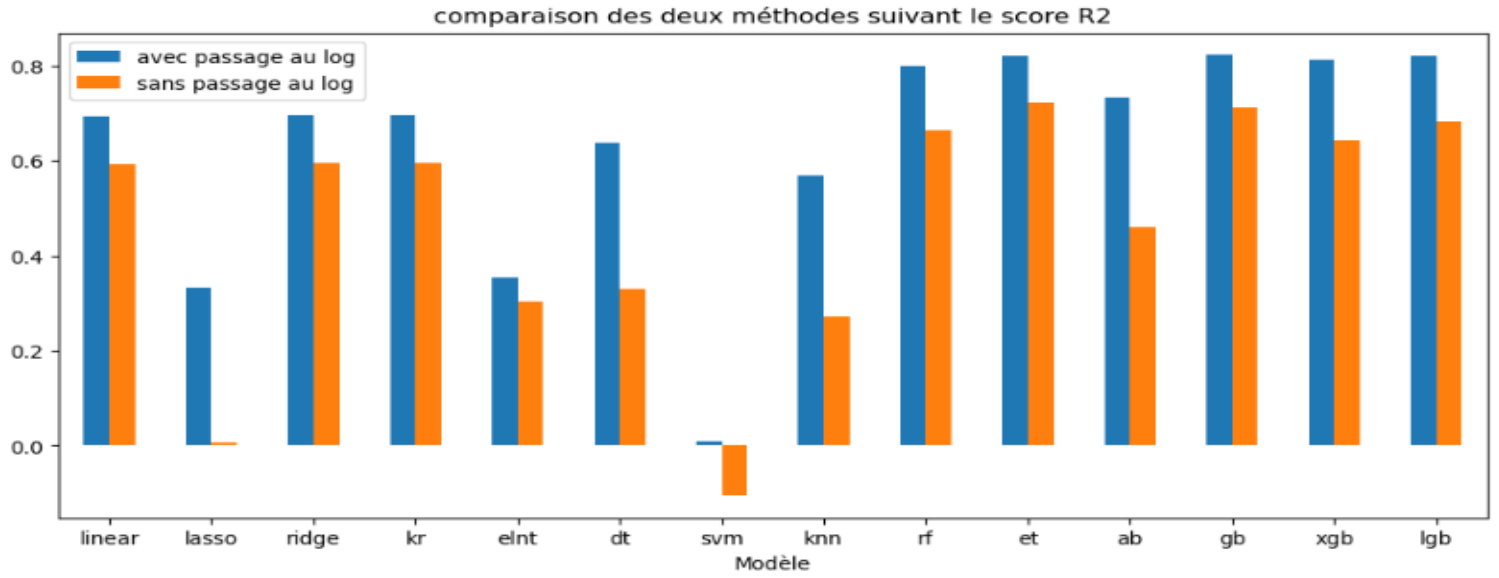
	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	et	0.722283	0.559533	0.748019	0.411645	0.814018
1	gb	0.710210	0.583857	0.764105	0.420278	1.235622
2	lgb	0.682826	0.639028	0.799392	0.424148	0.454440
3	rf	0.663838	0.677284	0.822973	0.446056	1.022741
4	xgb	0.641691	0.721905	0.849650	0.437608	0.753896
5	ridge	0.593530	0.818938	0.904952	0.559860	0.016042
6	kr	0.593508	0.818981	0.904976	0.559878	0.079844
7	linear	0.592496	0.821021	0.906102	0.565416	0.018057
8	ab	0.460544	1.086872	1.042532	0.859823	0.628121
9	dt	0.328837	1.352230	1.162854	0.616575	0.064601
10	elnt	0.302755	1.404777	1.185233	0.732474	0.014067
11	knn	0.272793	1.465144	1.210431	0.660983	0.063013
12	lasso	0.006636	2.001386	1.414703	0.935096	0.008751
13	svm	-0.105316	2.226940	1.492294	0.778327	3.456544



Modélisation – Emission de GES

Initialisation : par défaut
Transformation : log + 1

	Modèle	R2	MSE	RMSE	MAE	Durée exéc
0	gb	0.822068	0.307060	0.554130	0.463648	1.313914
1	lgb	0.820193	0.310296	0.557042	0.446686	0.410718
2	et	0.820008	0.310616	0.557329	0.455445	0.920984
3	xgb	0.811043	0.326086	0.571039	0.455285	0.785001
4	rf	0.799977	0.345183	0.587523	0.486704	1.019876
5	ab	0.733090	0.460611	0.678684	0.603347	0.857031
6	kr	0.696343	0.524026	0.723896	0.592419	0.108280
7	ridge	0.695955	0.524696	0.724359	0.592567	0.017339
8	linear	0.691827	0.531820	0.729260	0.594304	0.016001
9	dt	0.636756	0.626858	0.791743	0.604804	0.074383
10	knn	0.567826	0.745811	0.863604	0.687938	0.063218
11	elnt	0.352916	1.116685	1.056733	0.827020	0.023875
12	lasso	0.332103	1.152602	1.073593	0.836278	0.015614
13	svm	0.008130	1.711689	1.308315	1.013537	0.438401



Modélisation – Emission de GES

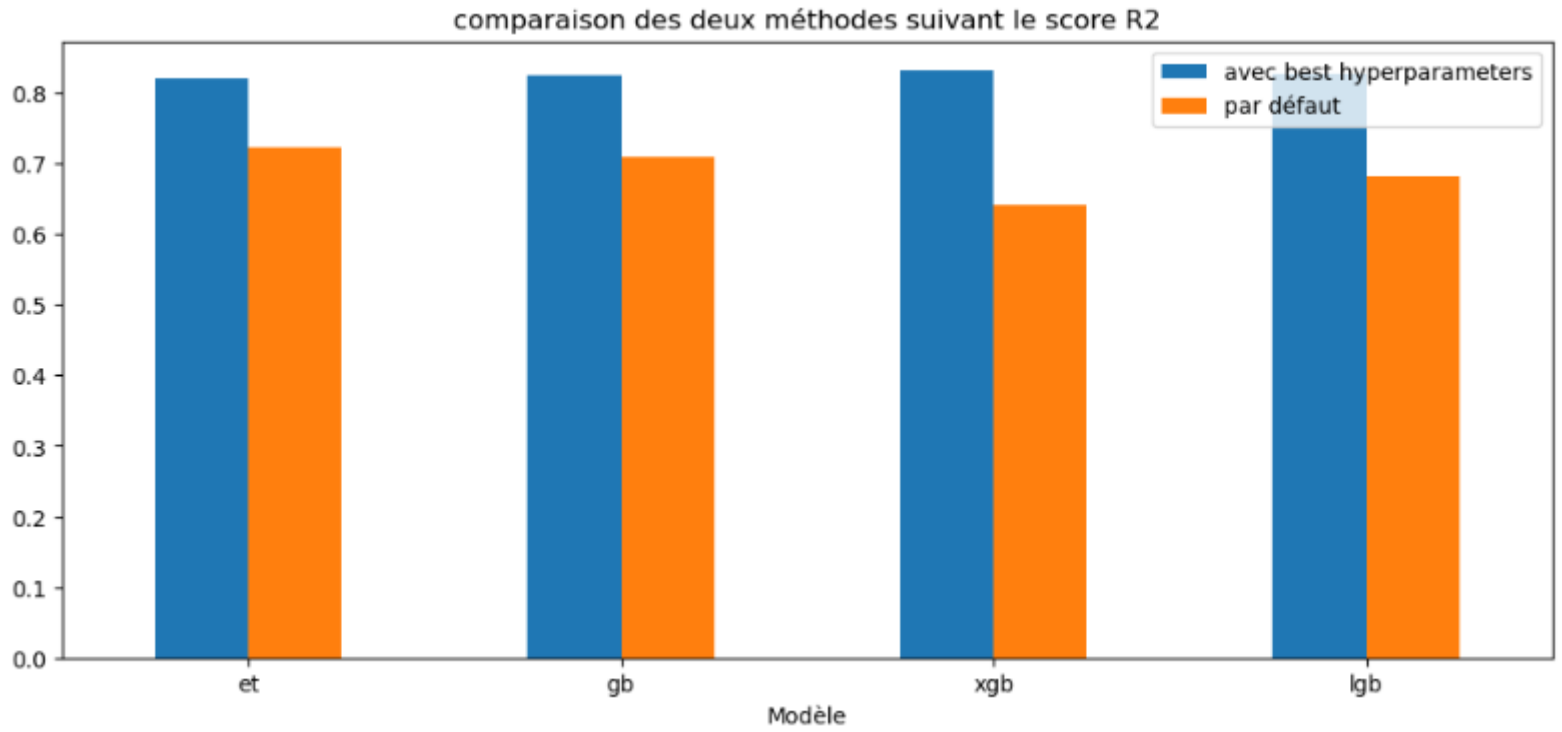
Comparaison avant et après optimisation

Par défaut

	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	et	0.722283	0.559533	0.748019	0.411645	0.814018
1	gb	0.710210	0.583857	0.764105	0.420278	1.235622
2	xgb	0.641691	0.721905	0.849650	0.437608	0.753896
3	lgb	0.682826	0.639028	0.799392	0.424148	0.454440

Avec meilleurs hyperparamètres

	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	et	0.819731	0.311093	0.557757	0.459193	2.539839
1	gb	0.825939	0.300380	0.548070	0.444665	4.100525
2	xgb	0.831618	0.290579	0.539054	0.433949	0.691635
3	lgb	0.827619	0.297482	0.545419	0.441016	0.533867



Modélisation – Emission de GES

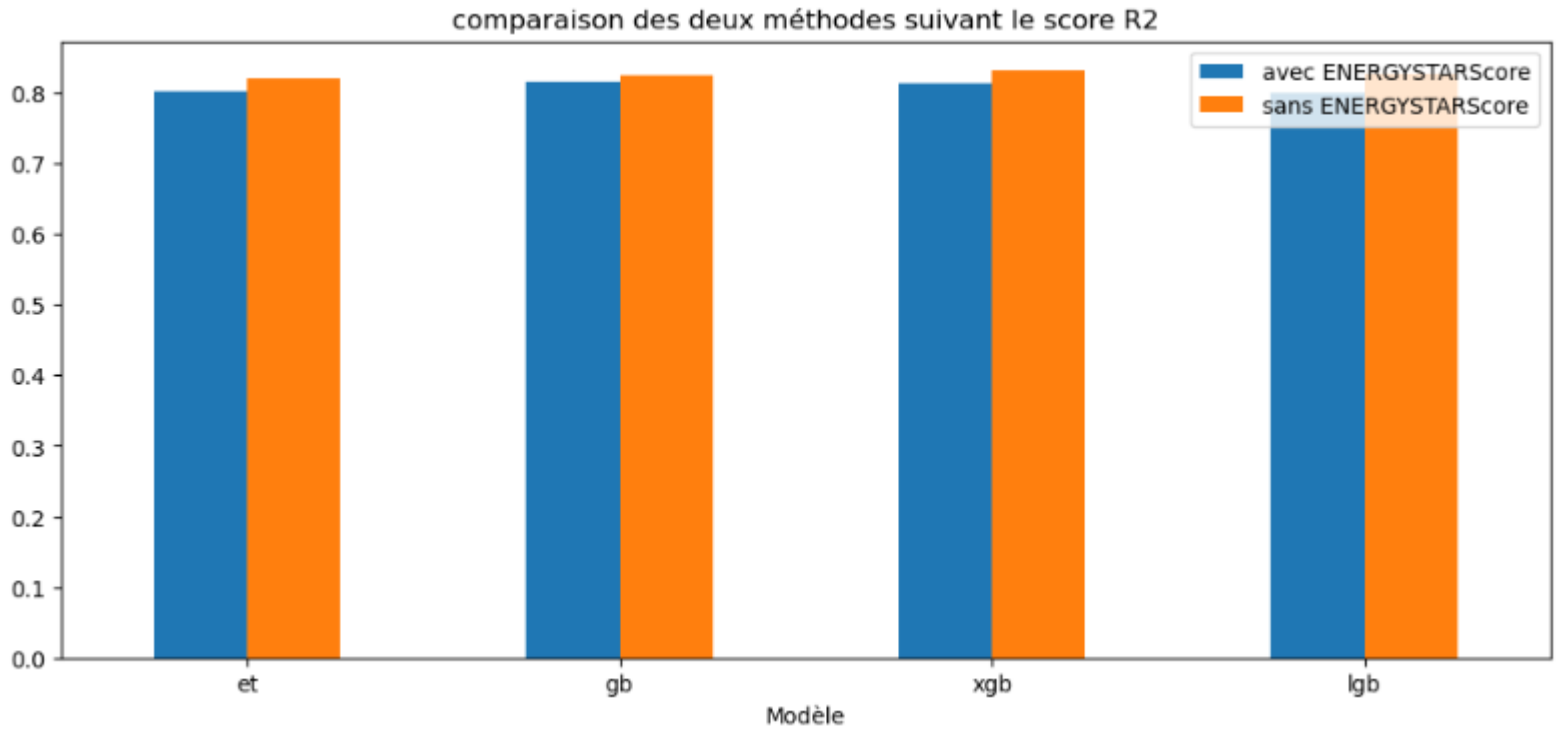
ENERGY STAR Score ?

Sans ENERGY STAR Score

	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	et	0.819731	0.311093	0.557757	0.459193	2.539839
1	gb	0.825939	0.300380	0.548070	0.444665	4.100525
2	xgb	0.831618	0.290579	0.539054	0.433949	0.691635
3	lgb	0.827619	0.297482	0.545419	0.441016	0.533867

Avec ENERGY STAR Score

	Modèle	R2	MSE	RMSE	MAE	Durée_exéc
0	et	0.802284	0.306083	0.553248	0.437300	1.867907
1	gb	0.815094	0.286253	0.535026	0.423544	1.773230
2	xgb	0.814588	0.287036	0.535757	0.431540	0.583740
3	lgb	0.800148	0.309390	0.556229	0.451852	0.203757



Conclusion

- **Consommation d'énergie :**

  - **Gradient Boosting (R2, RMSE)**
  - **Faible amélioration avec ENERGY STAR Score**

- **Émission de GES :**

  - **eXtreme Gradient Boosting (R2, RMSE)**
  - **Dégradation avec ENERGY STAR Score**

Perspectives

- **Jeu de données plus grand**
- **Recherche des hyperparametres : random search et approches bayésiennes**
- **Interpréter les prédictions des modèles : Shap ou Lime**

Merci de votre attention

Contact : bouzaieni@gmail.com