



# PROJET N° 3. ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

---

AUTEUR: IRINA MASLOWSKI

# PLAN

- Ville neutre en émissions de carbone
  - Objectifs de la ville de Seattle
  - Objectifs de notre équipe
  - Objectifs personnels
- Données :
  - Nettoyage
  - Feature engineering
  - Exploration
  - Transformation des variables asymétriques
- Pistes de modélisation
- Modèle final
- Conclusion



# VILLE NEUTRE EN ÉMISSIONS DE CARBONE.

## OBJECTIFS DE LA VILLE DE SEATTLE

- à long terme :
  - ville neutre en émissions de carbone en 2050
- à court terme :
  - Connaitre les émissions de CO<sub>2</sub> des bâtiments non destinés à l'habitation
  - Économiser le budget sur les relevés des émissions de CO<sub>2</sub> et la consommation totale d'énergie



# VILLE NEUTRE EN ÉMISSIONS DE CARBONE.

## OBJECTIFS DE NOTRE ÉQUIPE

- Etudier les émissions de CO<sub>2</sub> et la consommation totale d'énergie des bâtiments non destinés à l'habitation
- Prédire ces valeurs pour les bâtiments pour lesquels elles n'ont pas encore été mesurées





# VILLE NEUTRE EN ÉMISSIONS DE CARBONE. OBJECTIFS DE NOTRE ÉQUIPE

- Augmenter l'efficacité de l'équipe:
  - évaluer l'intérêt de l'"ENERGY STAR Score" pour la prédiction d'émissions



# VILLE NEUTRE EN ÉMISSIONS DE CARBONE. OBJECTIFS PERSONNELS

- Réaliser une courte analyse exploratoire
- Tester différents modèles de prédiction
- Faire attention à:
  - la fuite de données
  - le traitement des différentes variables
  - effectuer une évaluation rigoureuse des performances des modèles



# PLAN

- Ville neutre en émissions de carbone
  - Objectifs de la ville de Seattle
  - Objectifs de notre équipe
  - Objectifs personnels
- Données :
  - Nettoyage
  - Feature engineering
  - Exploration
  - Transformation des variables asymétriques
- Pistes de modélisation
- Modèle final
- Conclusion





# DONNÉES

- SEA Building Energy Benchmarking : Open Data from the City of Seattle

année	2015	2016
Nombre des observations	3340	3376
Nombre de variables	47	46






# DONNÉES

- Fusion de deux dataframes:
  - *combine\_first*
- Completion de données de l'années 2016 par les données du 2015 :
  - Ajout de 56 batiments manquants



# DONNÉES.

## NETTOYAGE

- Homogénéisation des noms des colonnes
    - 2015 : GHGEmissions(MetricTonsCO2e)
    - 2016: TotalGHGEmissions
- 
- TotalGHGEmissions(MetricTonsCO2e)
  - Homogénéisation de la casse
  - Correction du type des variables
    - ZipCode : float64 → category
  - Suppressions:
    - Outlier: 'Bullitt Center' (consommation négative de l'énergie)
    - Bâtiments résidentiels et mixtes
  - Correction des valeurs erronées:
    - Seattle Chinese Baptist Church, nombre d'étage: 99 → 2

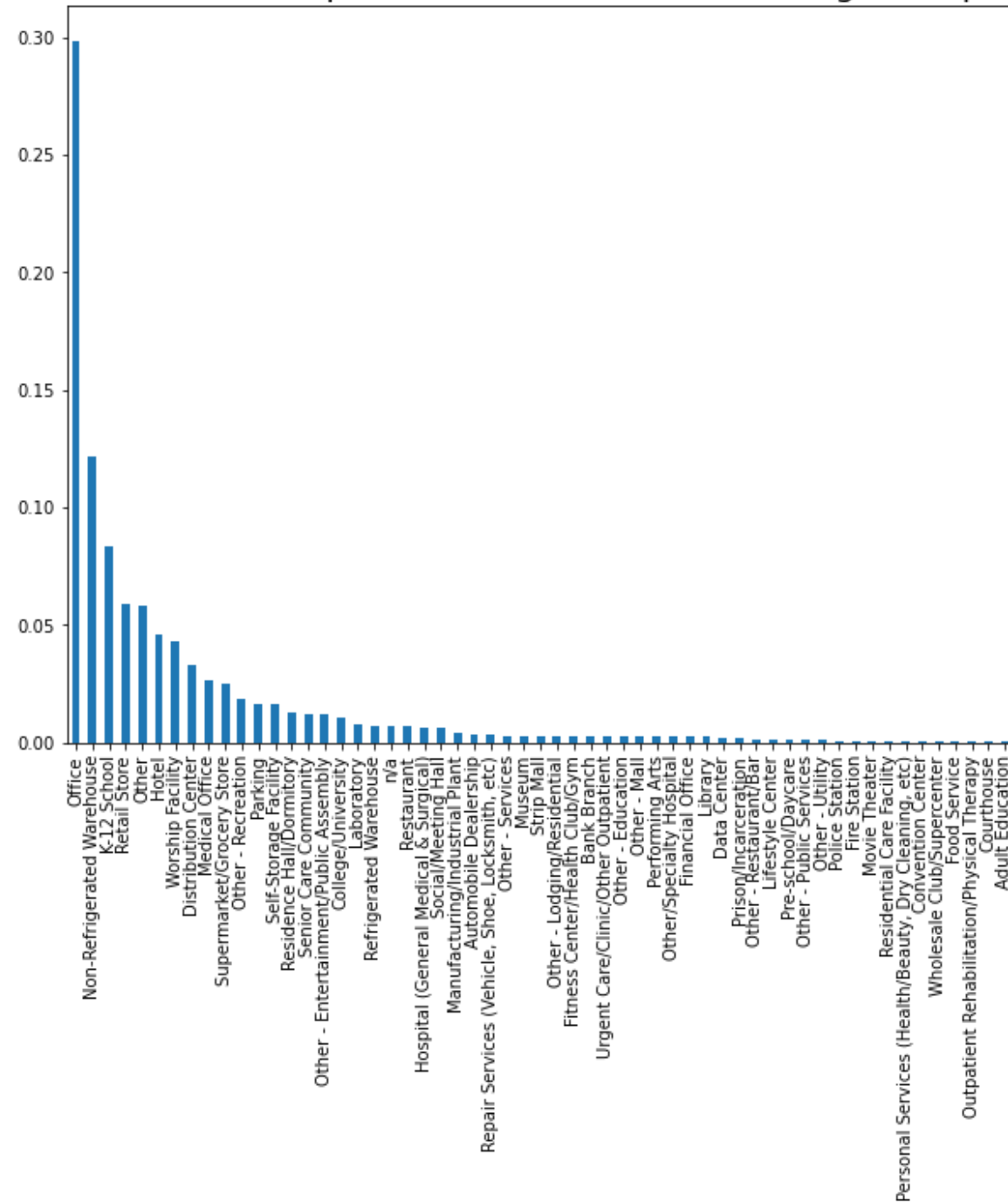
# DONNÉES.

## FEATURE ENGINEERING

- Introduction de nouvelles variables:
  - distance par rapport au centre-ville
  - la surface du sol par étage de bâtiment :
    - $\text{GFA (gross floor area) d'un étage} = \text{GFA de propriété} / \text{nombre de bâtiments} / \text{nombre d'étages}$
  - type d'énergie utilisée (valeurs binaires)
  - proportion des sources d'énergie utilisées → n'est pas utilisé pour éviter la fuite des données
  - nombre de différents types d'énergie utilisés

# DONNÉES. EXPLORATION

- Analyse univarié :
  - Les bureaux et les entrepôts prévalent

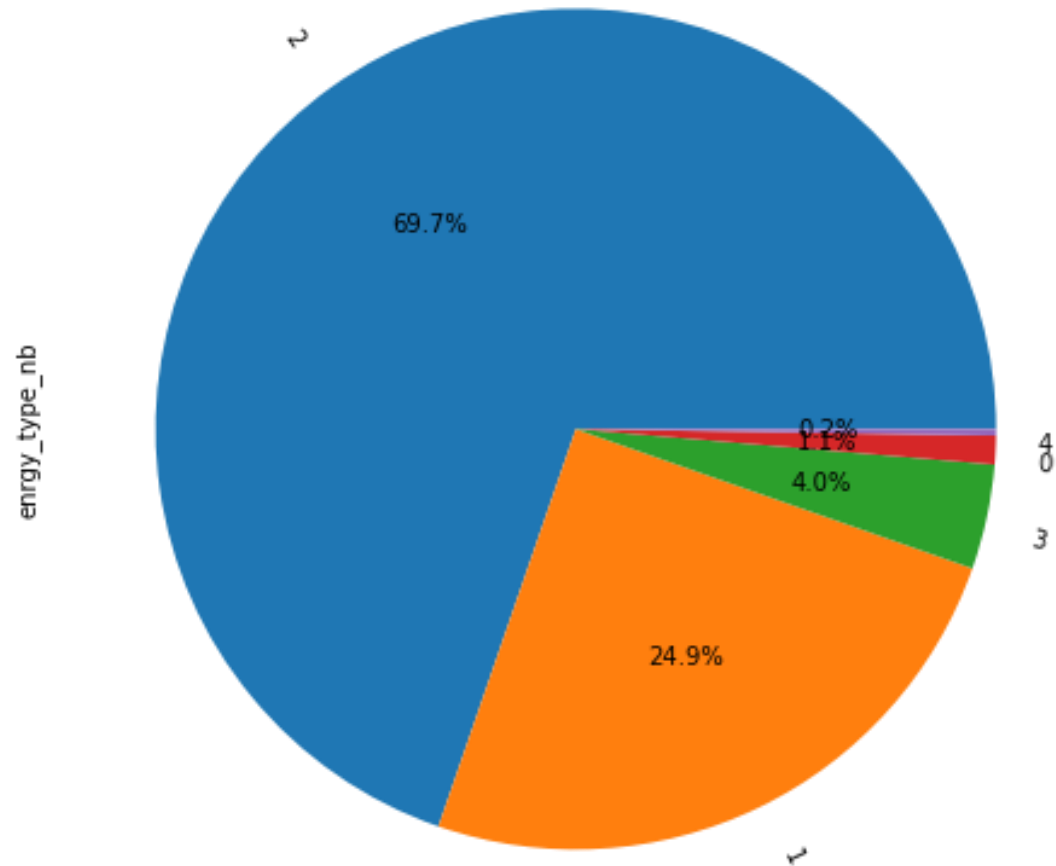




# DONNÉES. EXPLORATION

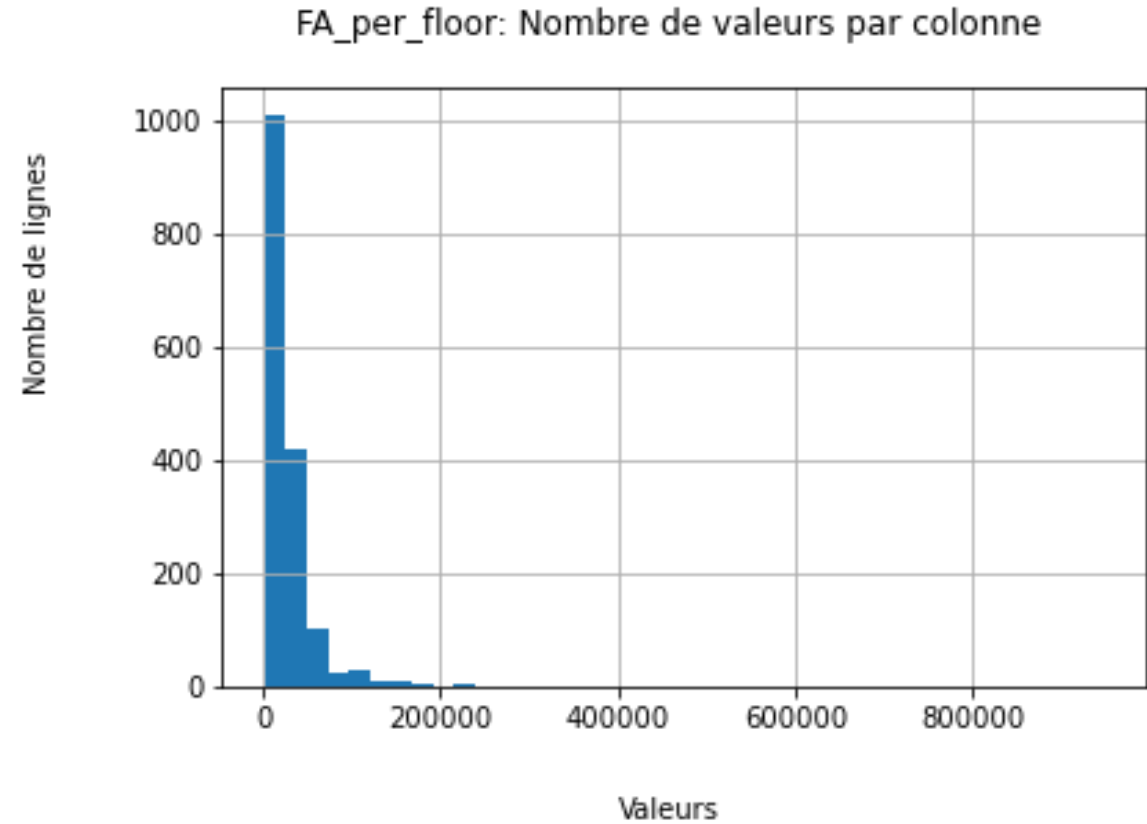
- Analyse univarié:
  - Un bâtiment utilisant 4 types de l'énergies (électricité, gaz, géothermique et autre carburant) est rare

Distribution des données qualitatifs dans la colonne 'enrgy\_type\_nb'



# DONNÉES. EXPLORATION

- Analyse univarié :
  - Les données sont très dissymétriques



# DONNÉES. EXPLORATION

- Analyse bivariée:
  - Les émissions de gaz à effet de serre sont corrélées avec la consommation de l'énergie
  - La quantité annuelle d'énergie consommée par le bien à partir de toutes les sources d'énergie est corrélée avec la superficie totale de la surface entre l'extérieur des murs d'un bâtiment
  - ENERGY STAR Score est anti-corrélé avec l'intensité des émissions de CO<sub>2</sub>
  - L'année de construction n'a pas de corrélation avec l'intensité des émissions de CO<sub>2</sub>

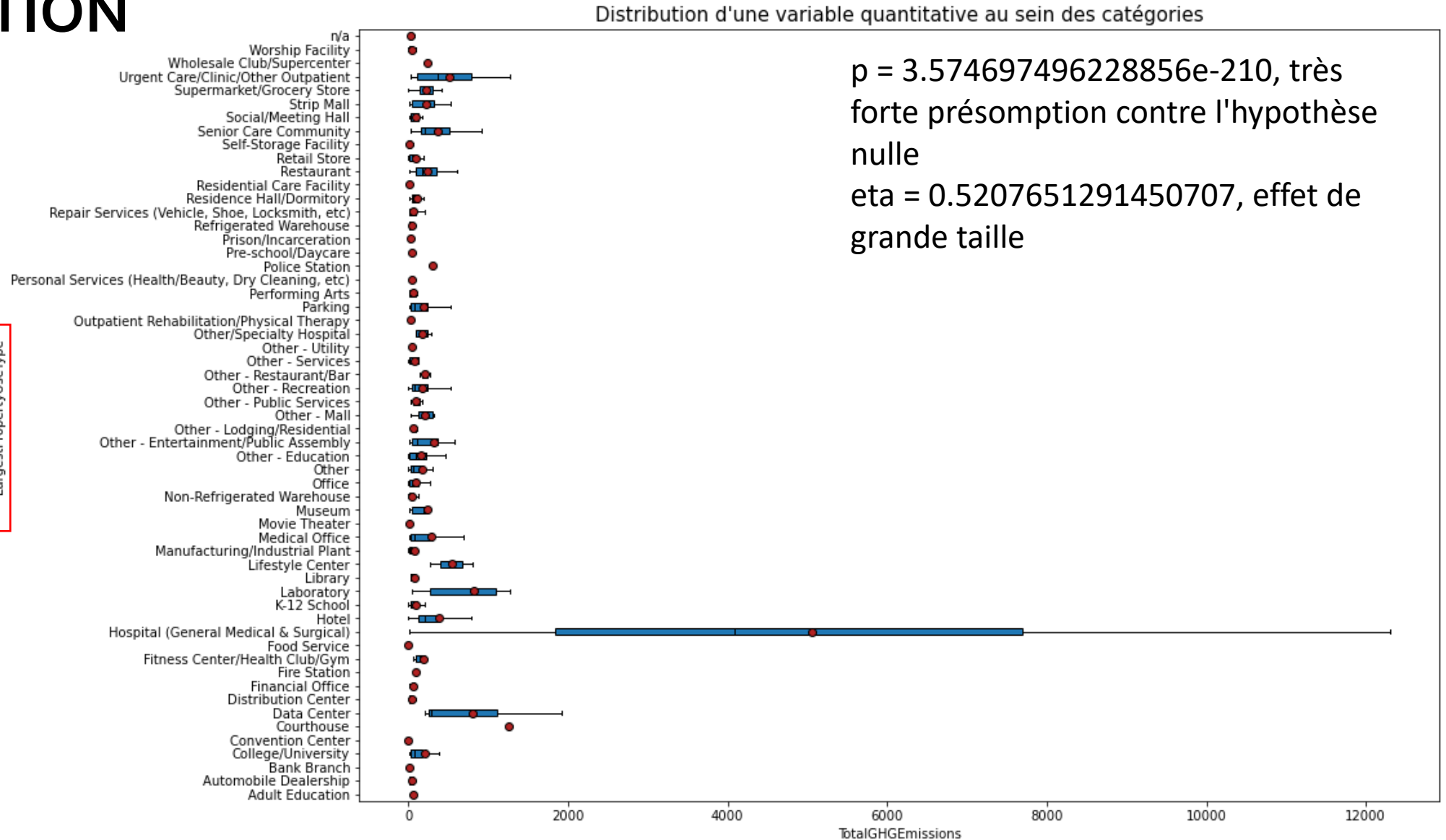
# DONNÉES. EXPLORATION

Variable cible: « émission de gaz à effet de serre »

Le type majoritaire de propriété a un effet similaire sur la **variable cible** « consommation de l'énergie » :

$p = 3.588102015410557e-164$ ,  
 $\eta^2 = 0.44918550013085207$

LargestPropertyUseType

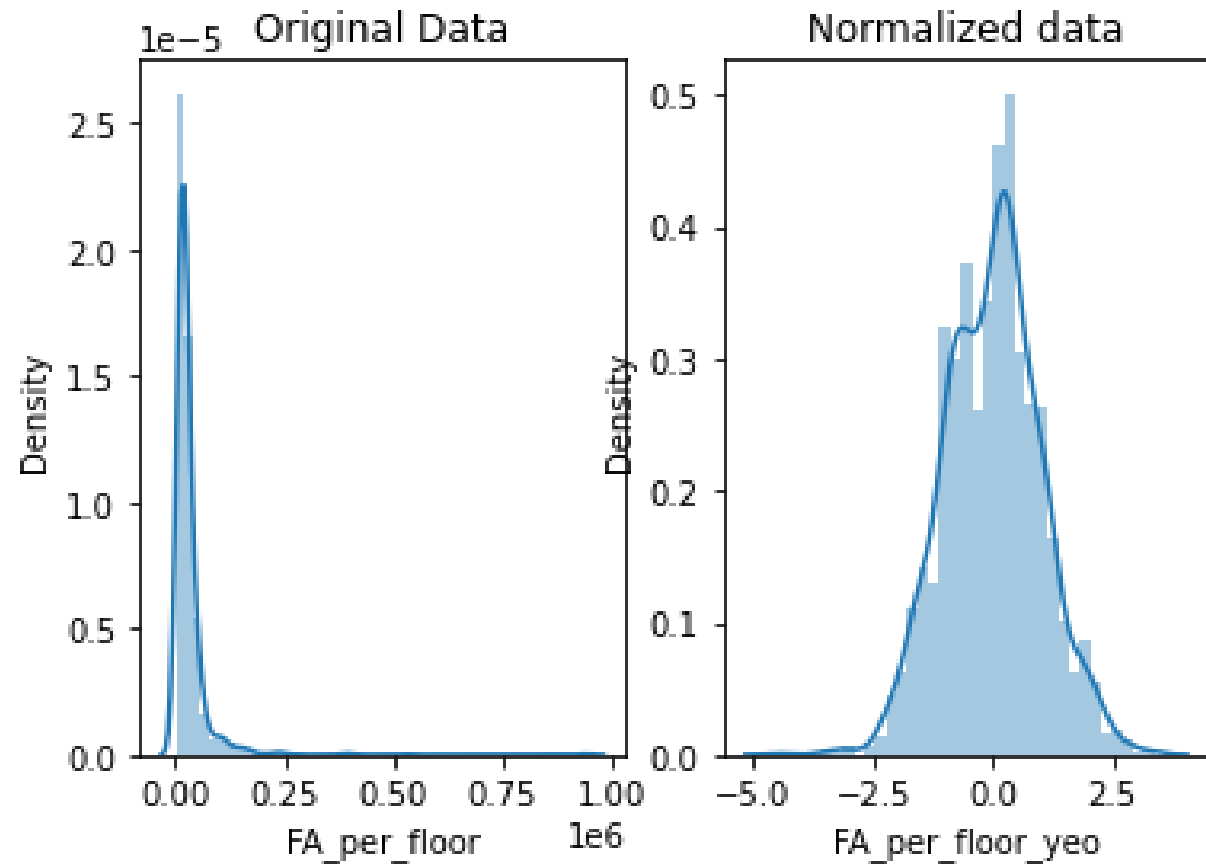




# DONNÉES.

## TRANSFORMATION DES VARIABLES ASYMÉTRIQUES

- Les transformations testées:
  - Log, racine carrée,  $1/x$ , box-cox, yeo-johnson
- Transformation finale choisie: yeo-johnson



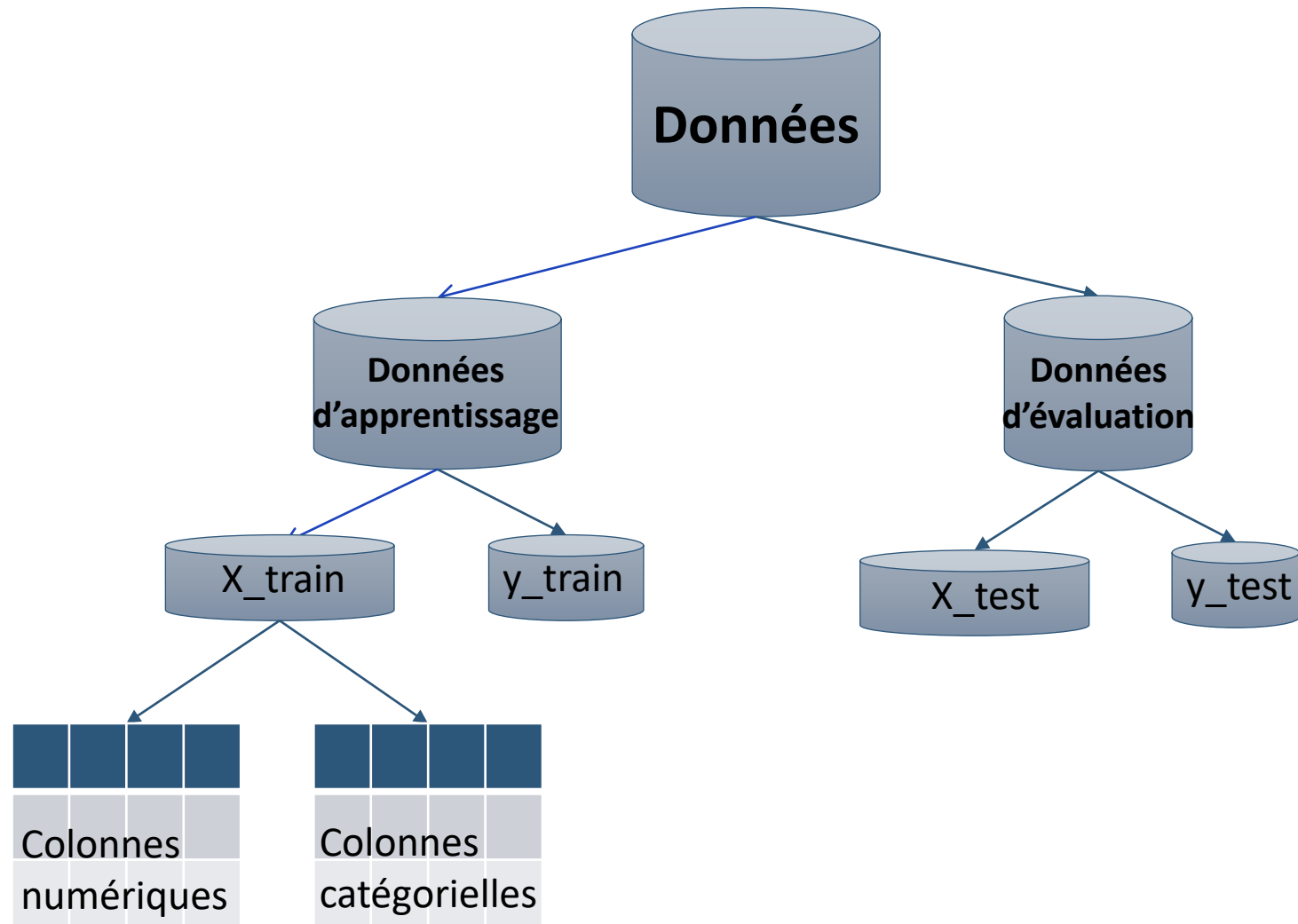
# PLAN

- Ville neutre en émissions de carbone
- Données
- Pistes de modélisation
  - Division des données
  - Création de pipeline
  - Configuration des tests
  - Processus d'apprentissage des modèles
  - Modèles testés et optimisation des hyperparamètres
  - Résultats de la validation croisée
  - Learning curves
- Modèle final
- Conclusion



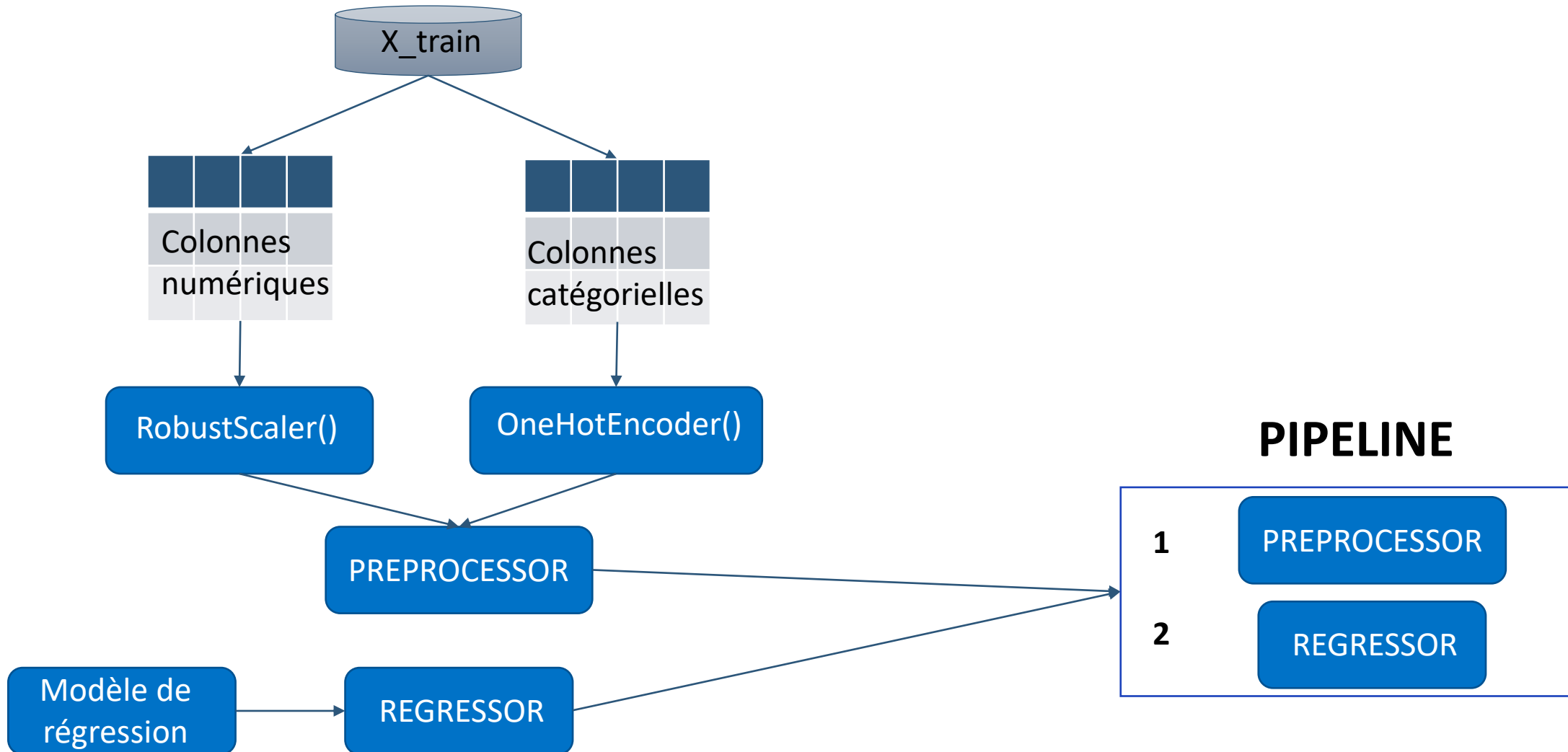
# PISTES DE MODÉLISATION

## DIVISION DE DONNÉES



# PISTES DE MODÉLISATION

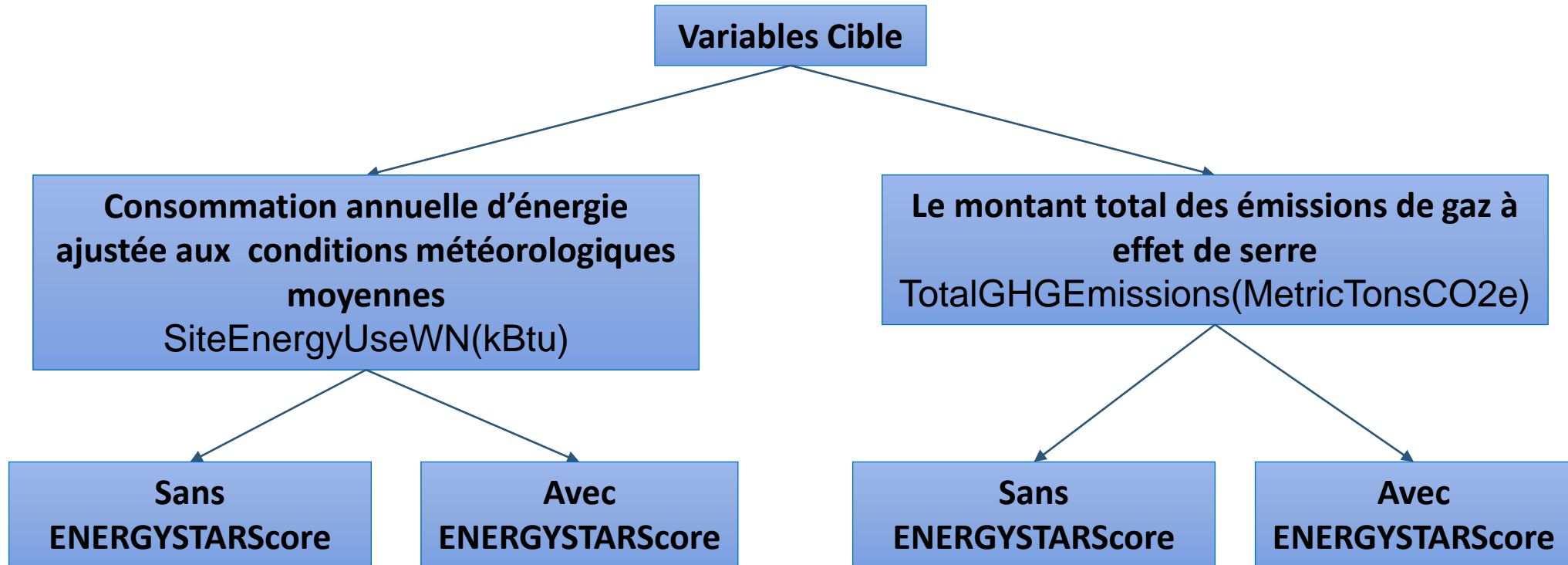
## CRÉATION DE PIPELINE





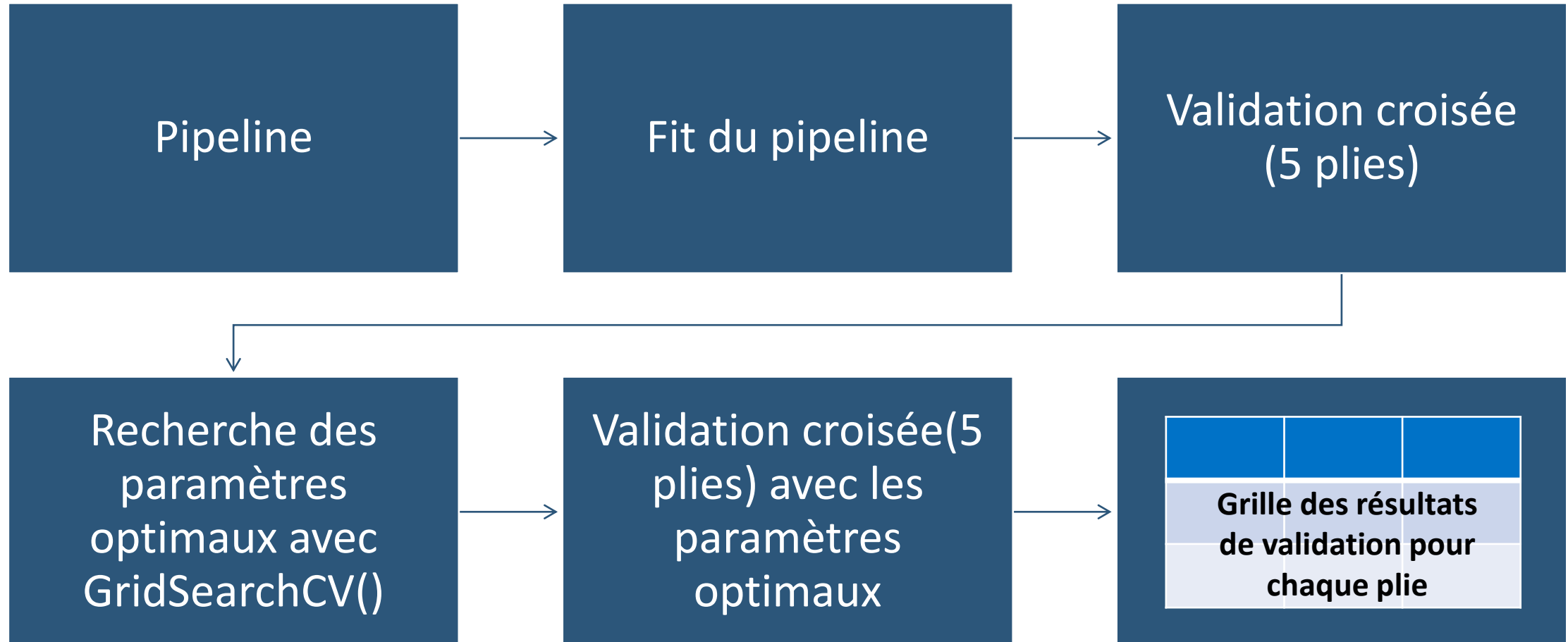
# PISTES DE MODÉLISATION

## CONFIGURATION DES TESTS



# PISTES DE MODÉLISATION

## PROCESSUS D'APPRENTISSAGE DES MODÈLES



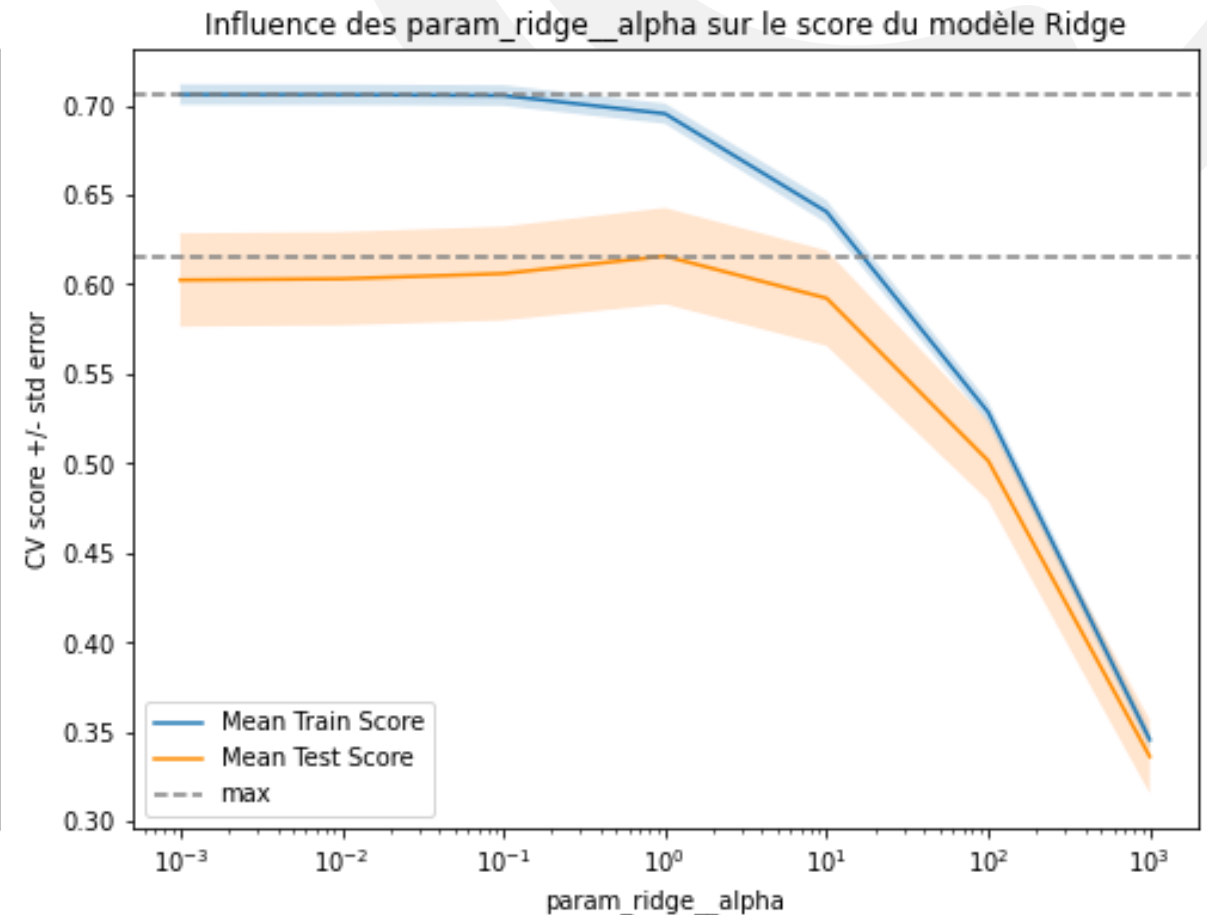
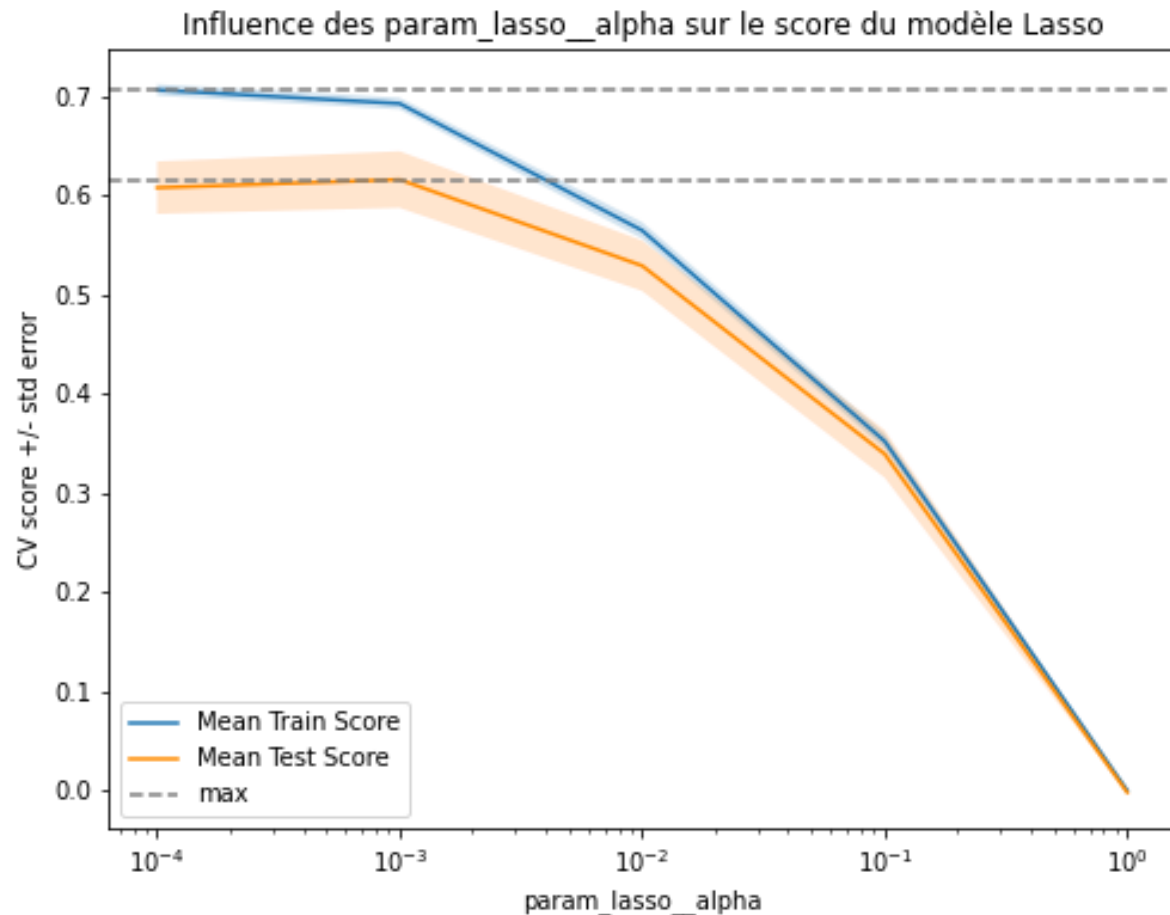
# PISTES DE MODÉLISATION

## MODÈLES TESTÉS ET OPTIMISATION DES HYPERPARAMÈTRES

Type de modèles	Nom du modèle	Paramètres du modèle à optimiser				
		Avant optimisation (par défaut) Le paramètre à optimiser est en gras	Meilleurs paramètres			
			Variable: Consommation d'énergie		Variable: Emission de gaz à l'effet de serre	
			Sans EnergyStarScore	Avec EnergyStarScore	Sans EnergyStarScore	Avec EnergyStarScore
Baseline	Dummy Regressor	mean				
Modèles linéaires	Regression linéaire					
	Régression Ridge	$\alpha = 0,5$	$\alpha = 1$			
	Régression Lasso	$\alpha = 0,1$	$\alpha = 0,001$			

# PISTES DE MODÉLISATION

## MODÈLES TESTÉS ET OPTIMISATION DES HYPERPARAMÈTRES





# PISTES DE MODÉLISATION

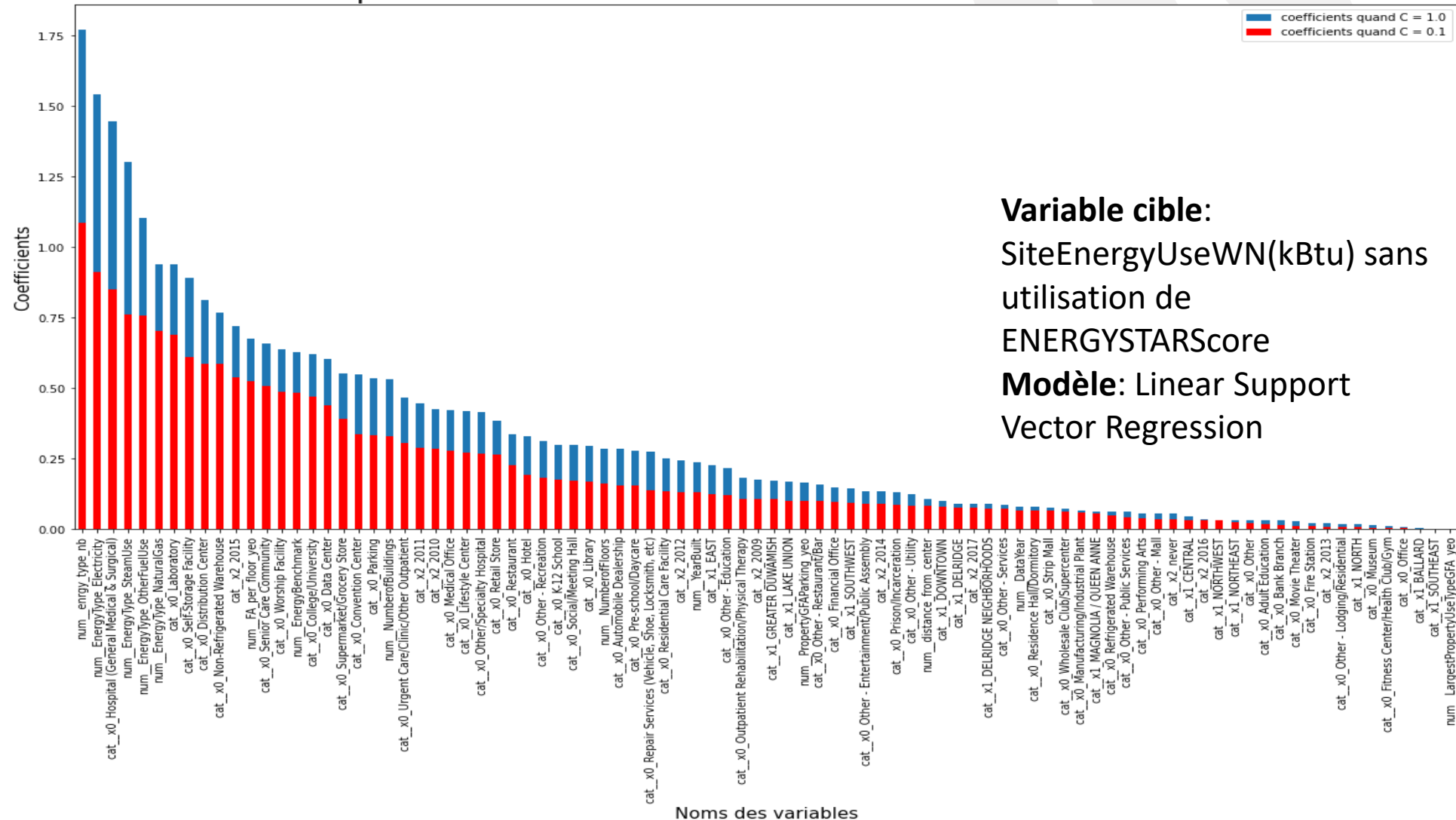
## MODÈLES TESTÉS ET OPTIMISATION DES HYPERPARAMÈTRES

Type de modèles	Nom du modèle	Paramètres du modèle à optimiser				
		Avant optimisation (par défaut) Le paramètre à optimiser est en gras	Meilleurs paramètres			
			Variable: Consommation d'énergie		Variable: Emission de gaz à l'effet de serre	
			Sans EnergyStarScore	Avec EnergyStarScore	Sans EnergyStarScore	Avec EnergyStarScore
Arbres de décision	Decision Tree Regressor	<b>max_depth=None,</b> <b>min_samples_split=2,</b> <b>min_samples_leaf=1</b>	max_depth=20, min_samples_split=0,1, min_samples_leaf=4		max_depth=20, min_samples_split=0,1, min_samples_leaf=2	max_depth=20, min_samples_split=0,1, min_samples_leaf=1
Forêts aléatoires	Random Forest Regressor	n_estimators=500, <b>max_depth=None</b>	max_depth=35	max_depth=30	max_depth=40	max_depth=20
Support Vector Machine	SVR (Linear Support Vector Regression)	<b>C=1.0,</b> dual=False, loss='squared_epsilon_insensitive'	C=0,1	C=1.0	C=0,1	

# PISTES DE MODÉLISATION

## MODÈLES TESTÉS ET OPTIMISATION DES HYPERPARAMÈTRES

Comparaison de l'effet de C sur les coefficients des variables.



# PISTES DE MODÉLISATION

## MODÈLES TESTÉS ET OPTIMISATION DES HYPERPARAMÈTRES

Type de modèles	Nom du modèle	Paramètres du modèle à optimiser			
		Avant optimisation (par défaut) Le paramètre à optimiser est en gras	Meilleurs paramètres		
			Variable cible : Consommation d'énergie		Variable cible : Emission de gaz à l'effet de serre
			Sans EnergyStarScore	Avec EnergyStarScore	Sans EnergyStarScore
Support Vector Machine	Epsilon-Support Vector Regression (SVM SVR)	<b>C=1.0</b> , epsilon=0.2, <b>gamma=0.01</b>	C=100, gamma=0.01		
Méthodes ensemblistes	Gradient Boosting Regressor	<b>n_estimators=100</b> , <b>max_depth=3</b> , <b>min_samples_split=2</b>	max_depth=1, min_samples_split=0,1, n_estimators = 700	max_depth=2, min_samples_split=0,2, n_estimators = 500	

# PISTES DE MODÉLISATION

## RÉSULTATS DE LA VALIDATION CROISÉE

**Modèle  
choisi**

	Dummy Regressor	Decision Tree Regressor	GBR	Linear SVR	Random Forest Regressor	Regression Lasso	Regres sion Linéaire	Regres sion Ridge	SVM SVR
--	--------------------	-------------------------------	-----	---------------	-------------------------------	---------------------	----------------------------	----------------------	---------

**Variable cible : Consommation d'énergie, sans EnergyStarScore**

Score de validation	-0.000083	0.591806	0.755285	0.698051	0.706606	0.704974	0.687599	0.702385	0.750089
fit_time	0.010970	42.413607	540.156380	0.378985	113.933814	0.455782	0.030916	0.490688	9.695037

**Variable cible : Consommation d'énergie, avec EnergyStarScore**

Score de validation	-0.000083	0.585452	0.795356	0.727563	0.743545	0.730595	0.707665	0.734799	0.789572
fit_time	0.021942	39.086434	622.949432	0.695175	114.887013	0.887662	0.050862	0.888658	10.462023



# PISTES DE MODÉLISATION

## RÉSULTATS DE LA VALIDATION CROISÉE

**Modèle  
choisi**

	Dummy Regressor	Decision Tree Regressor	GBR	Linear SVR	Random Forest Regressor	Regression Lasso	Regres sion Linéaire	Regres sion Ridge	SVM SVR
--	--------------------	-------------------------------	-----	---------------	-------------------------------	---------------------	----------------------------	----------------------	---------

**Variable cible: Emission de gaz à effet de serre, sans EnergyStarScore**

Score de validation	-0.000044	0.596497	0.716335	0.675687	0.681986	0.676511	0.667387	0.674224	0.745679
fit_time	0.022939	45.906229	522.355260	0.659201	136.715743	0.709100	0.052861	0.916552	9.859664

**Variable cible: Emission de gaz à effet de serre, avec EnergyStarScore**

Score de validation	-0.000044	0.613540	0.760884	0.696672	0.714335	0.700187	0.688839	0.696888	0.766443
fit_time	0.022939	40.880631	565.837129	0.750997	173.070456	0.682141	0.053855	0.867644	13.688656

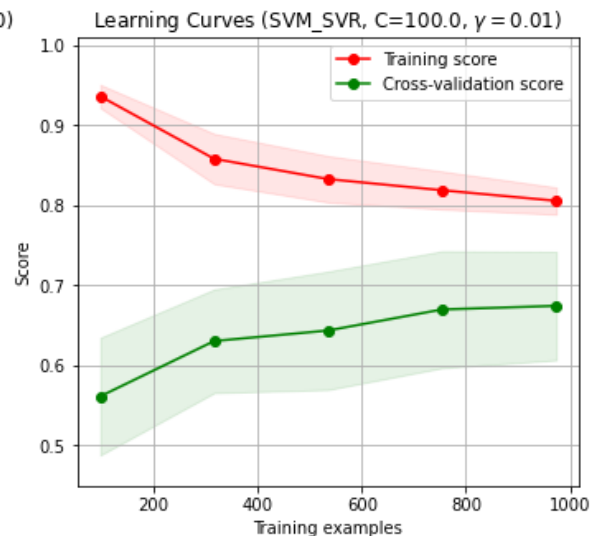
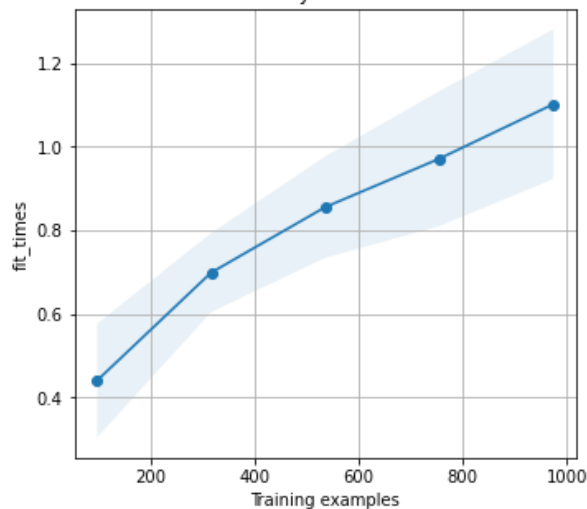
# PISTES DE MODÉLISATION

## LEARNING CURVES

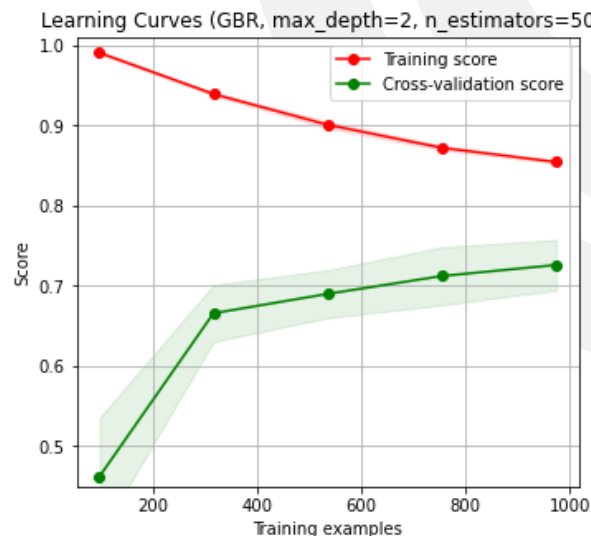
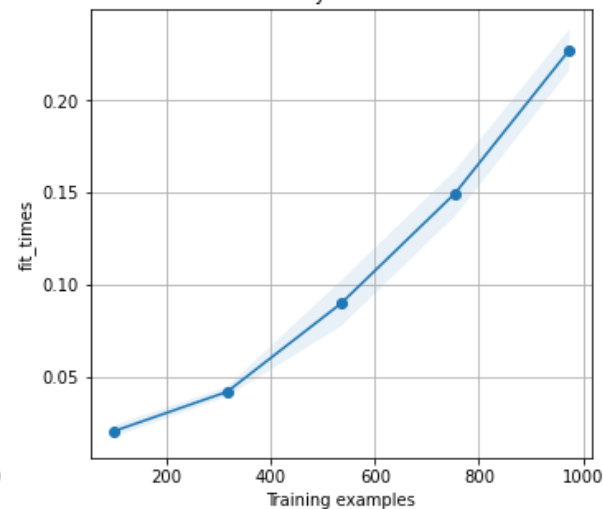
Variable cible : Consommation d'énergie



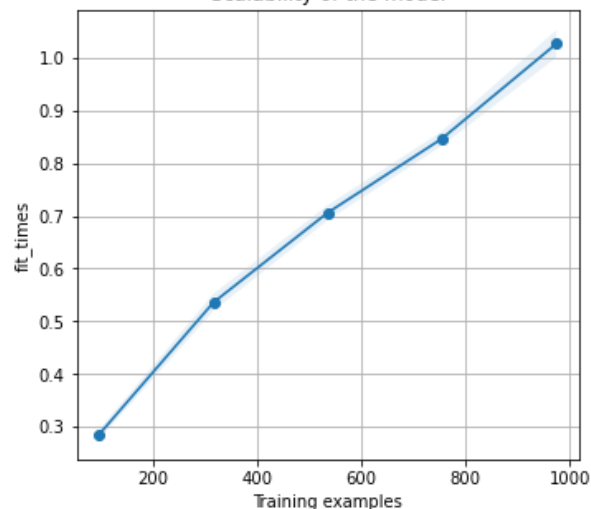
Scalability of the model



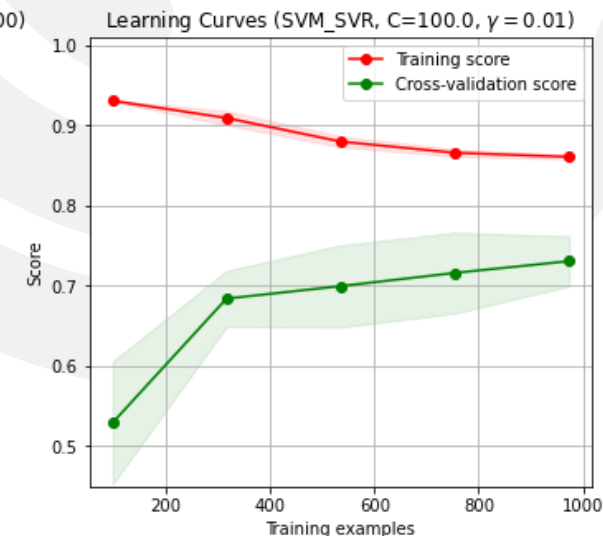
Scalability of the model



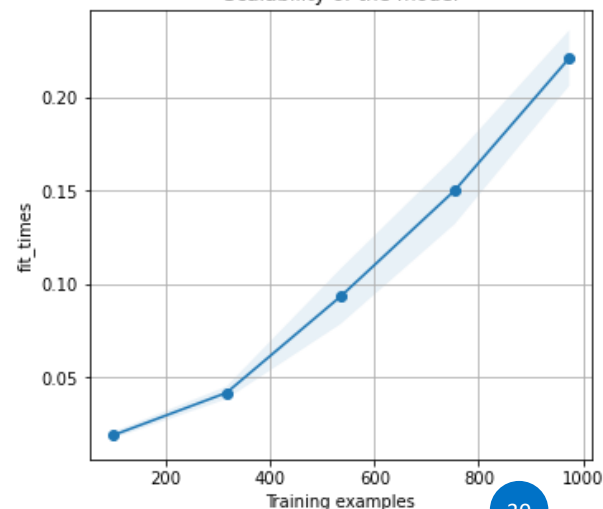
Scalability of the model



Variable cible : Emission de gaz CO<sub>2</sub>



Scalability of the model

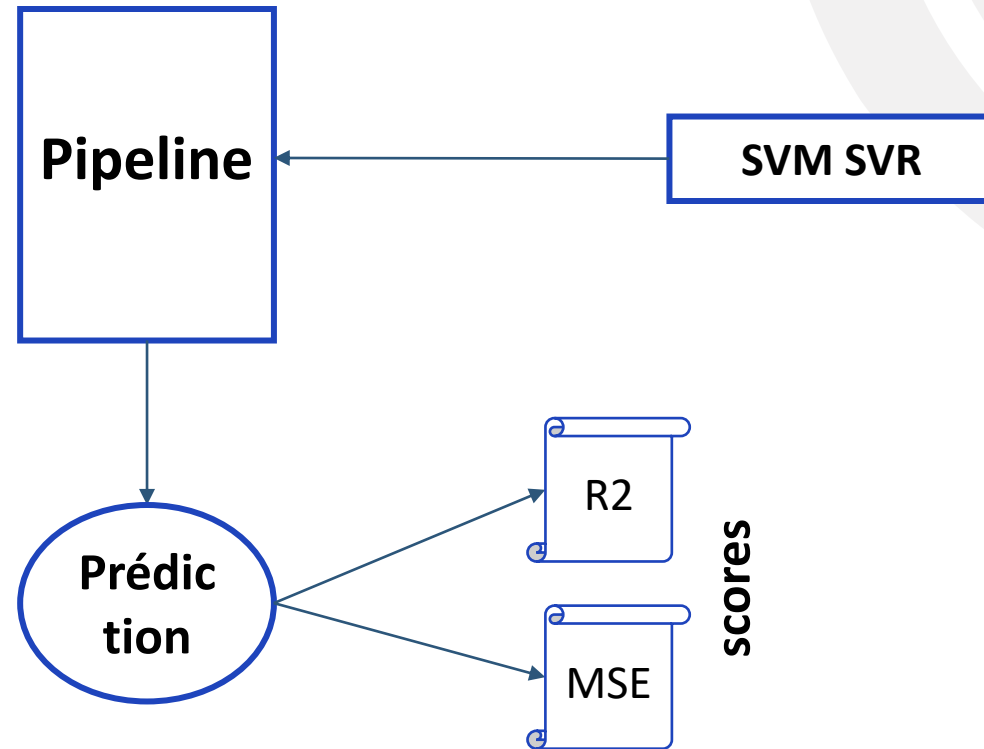


# PLAN

- Ville neutre en émissions de carbone
- Données
- Pistes de modélisation
- Modèle final
  - Prédiction
  - Résultats de prédiction
- Conclusion



# MODÈLE FINAL PRÉDICTION

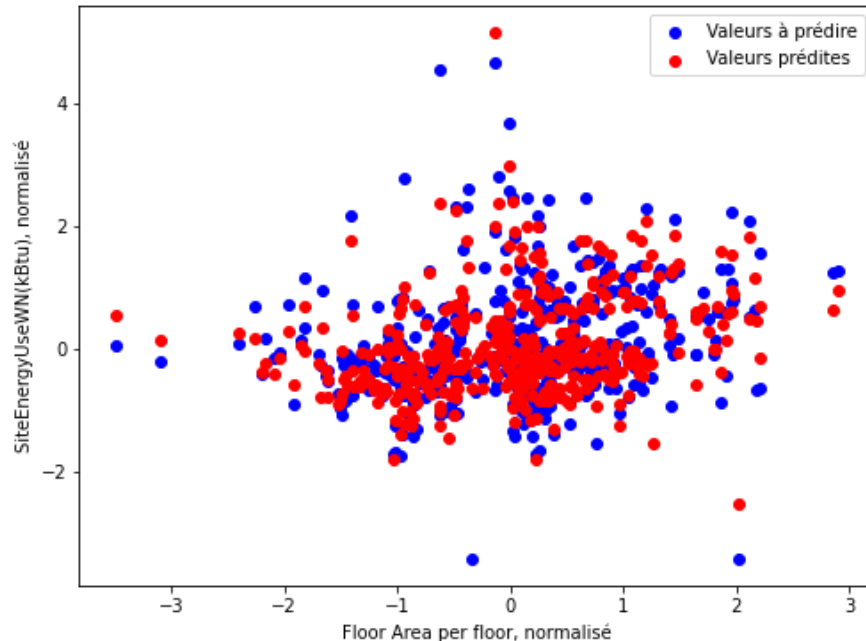


# MODÈLE FINAL

## RÉSULTAT DE PRÉDICTION

### Variable cible : Consommation d'énergie

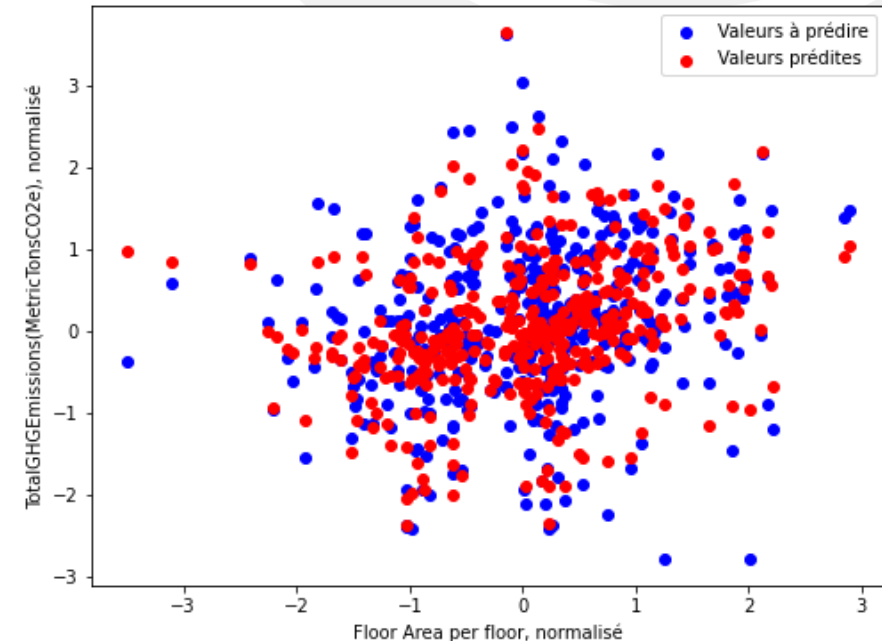
Résultats de prédiction du modèle SVM SVR avec un noyau gaussien et sans utilisation d'ENERGYSTARScore



SVM SVR:  $R^2 = 0,73$ ;  $MSE = 0,26$   
Dummy Regressor:  $R^2 = -0,02$ ;  $MSE = 0,97$

### Variable cible : Emission de gaz CO<sub>2</sub>

Résultats de prédiction du modèle SVM SVR avec un noyau gaussien et sans l'utilisation de l'ENERGYSTARScore



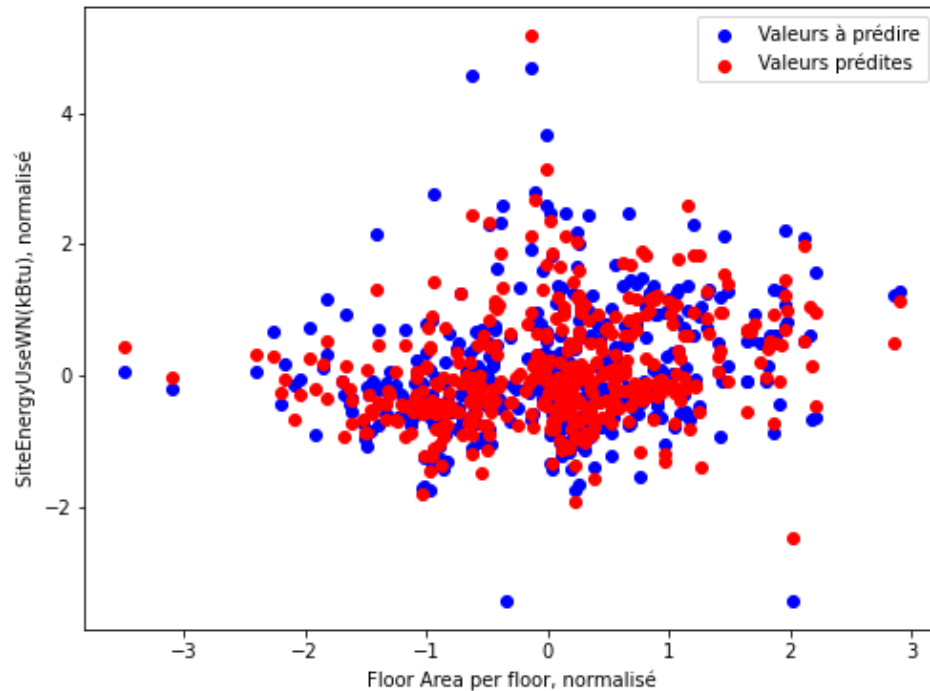
SVM SVR:  $R^2 = 0,69$ ;  $MSE = 0,30$   
Dummy Regressor:  $R^2 = -0,02$ ;  $MSE = 0,99$

# MODÈLE FINAL

## RÉSULTAT DE PRÉDICTION

### Variable cible : Consommation d'énergie

Résultats de prédiction du modèle SVM SVR avec un noyau gaussien  
et avec l'utilisation d'ENERGYSTARScore

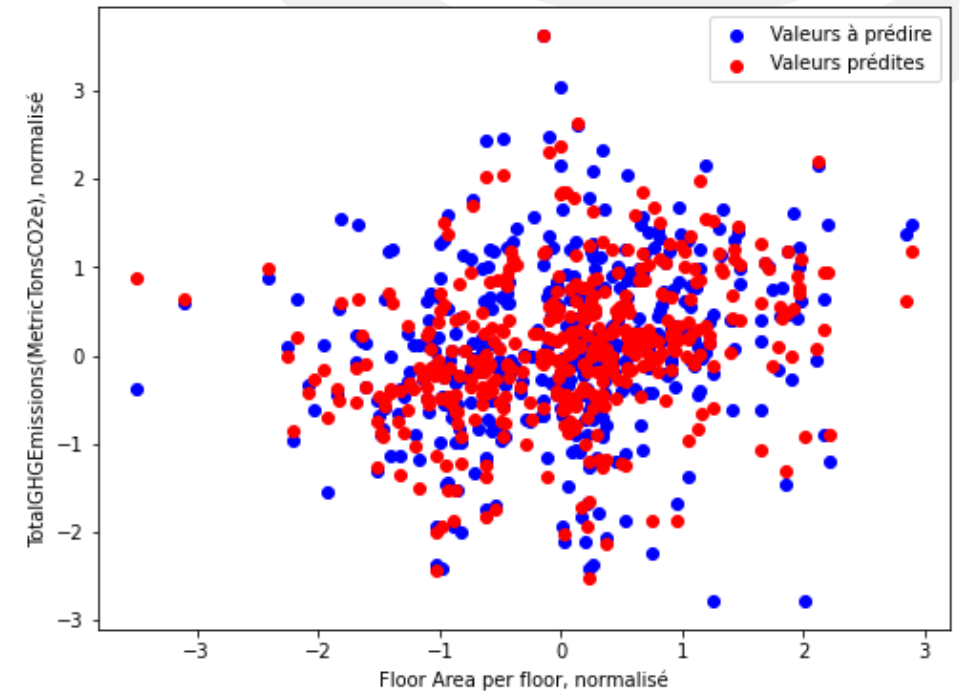


SVM SVR:  $R^2 = 0,75$ ;  $MSE = 0,24$

Dummy Regressor:  $R^2 = -0,02$ ;  $MSE = 0,97$

### Variable cible : Emission de gaz CO<sub>2</sub>

Résultats de prédiction du modèle SVM SVR avec un noyau gaussien  
et avec l'utilisation de l'ENERGYSTARScore



SVM SVR:  $R^2 = 0,71$ ;  $MSE = 0,28$

Dummy Regressor:  $R^2 = -0,02$ ;  $MSE = 0,99$



# CONCLUSION

- L'analyse exploratoire a montré que les données sont très dissymétriques et demande une transformation avant d'être fournies aux modèles.
- L'utilisation de pipeline permet d'éviter la fuite de donnée du corpus test vers le corpus d'apprentissage.
- La création de nouvelles variables permet d'éviter l'utilisation des variables corrélées avec la variable cible, ainsi que la fuite de donnée.
- 8 modèles ont été testés et évalués grâce à la validation croisée.
- Le modèle choisi est SVM avec un noyau gaussien (SVM SVR).

# CONCLUSION

Modèle	Score							
	R2				MSE			
	Variable cible : Consommation d'énergie		Variable cible : Emission de gaz CO2		Variable cible : Consommation d'énergie		Variable cible : Emission de gaz CO2	
	Sans Energy Star Score	Avec Energy Star Score	Sans Energy Star Score	Avec Energy Star Score	Sans Energy Star Score	Avec Energy Star Score	Sans Energy Star Score	Avec Energy Star Score
Dummy Regressor	-0,02	-0,02	-0,02	-0,02	0,97	0,97	0,99	0,99
SVM SVR	0,73	<b>0,75</b>	0,69	<b>0,71</b>	0,26	<b>0,24</b>	0,30	<b>0,28</b>

L'utilisation de la variable Energy Star Score améliore les résultats pour les deux variables cible.



# MERCI

---



IRINA.MASLOWSKI@GMAIL.COM

Comparaison de l'effet de recherche de meilleurs paramètres sur les coefficients des variables.

