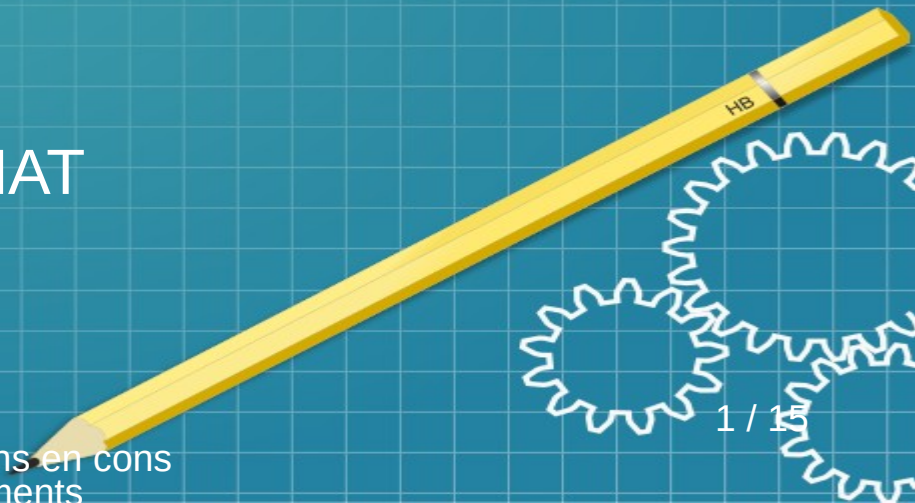
Anticipez les besoins en consommation électrique de bâtiments

Formation Data scientist
Projet n°4

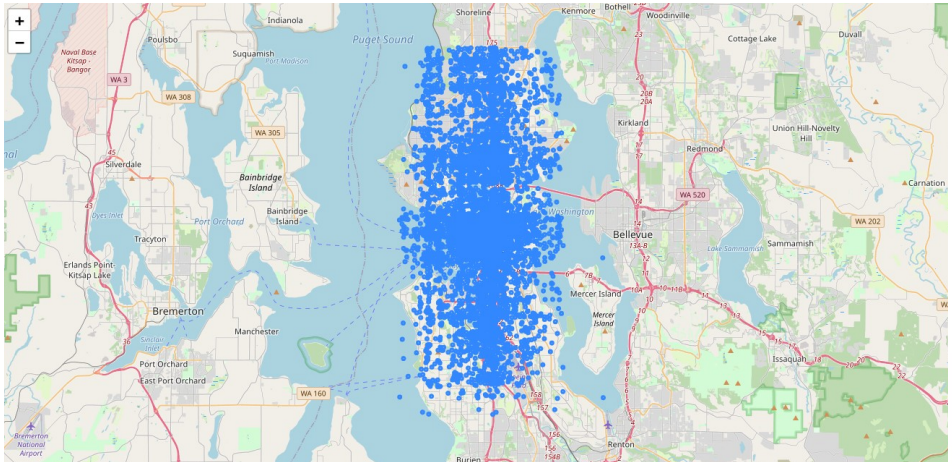
Bayram DONAT



Seattle

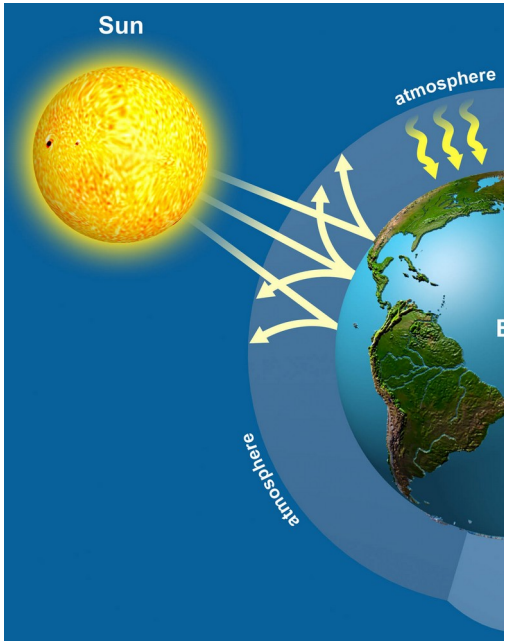
Sommaire

- Problématique, interprétation
- Solutions envisagées
- Travaux d’exploration
- Travaux de modélisation
- Modélisation finale optimisée



Problématique

- Contexte
  - Objectif de la ville de Seattle pour 2050:
    - Devenir une ville neutre en émissions CO2
  - Environnement de travail : Service concerné
    - Emissions des bâtiments non destinés à l'habitation.

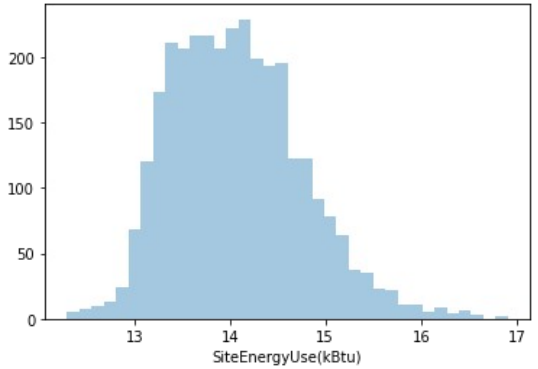


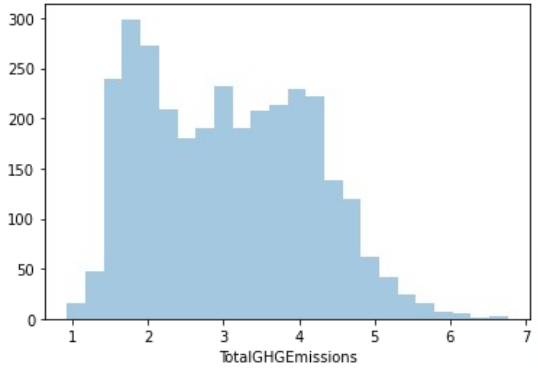
Seattle Solution envisagée

- Travail demandé :
  - A partir des relevés de 2015 et 2016, **prédire :**
    - les émissions de CO2
    - la consommation totale d'énergie
  - Evaluer l'intérêt de l'energy star score pour ces prédictions
- Mission
  - Réaliser une courte analyse exploratoire
  - Tester différents modèles de prédiction

Exploration

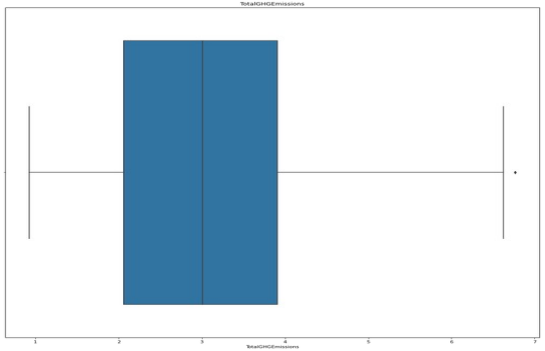
- Nettoyage : 2015 : 3340 lignes 47 colonnes + 2016 : 3376 lignes 46 colonnes
  - Fusion sur colonnes communes et colonnes similaires
  - Suppression des colonnes non communes
  - Correction des remplissages différents 2015/2016
  - Suppression des colonnes pas assez remplies (+90 % NaN)
  - Suppression Outliers (méthode interquartile)
  - Suppression des valeurs négatives
  - Passage au logarithme base 10
  - Remplacement des NaN par KNN imputer
  - 3008 lignes, 31 colonnes

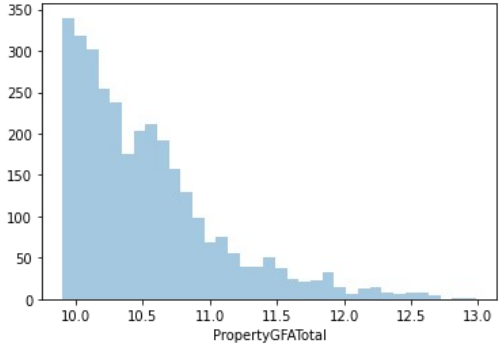


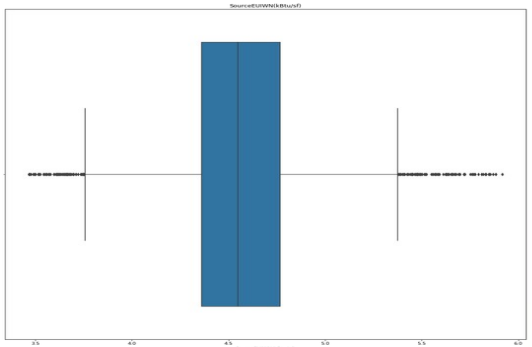


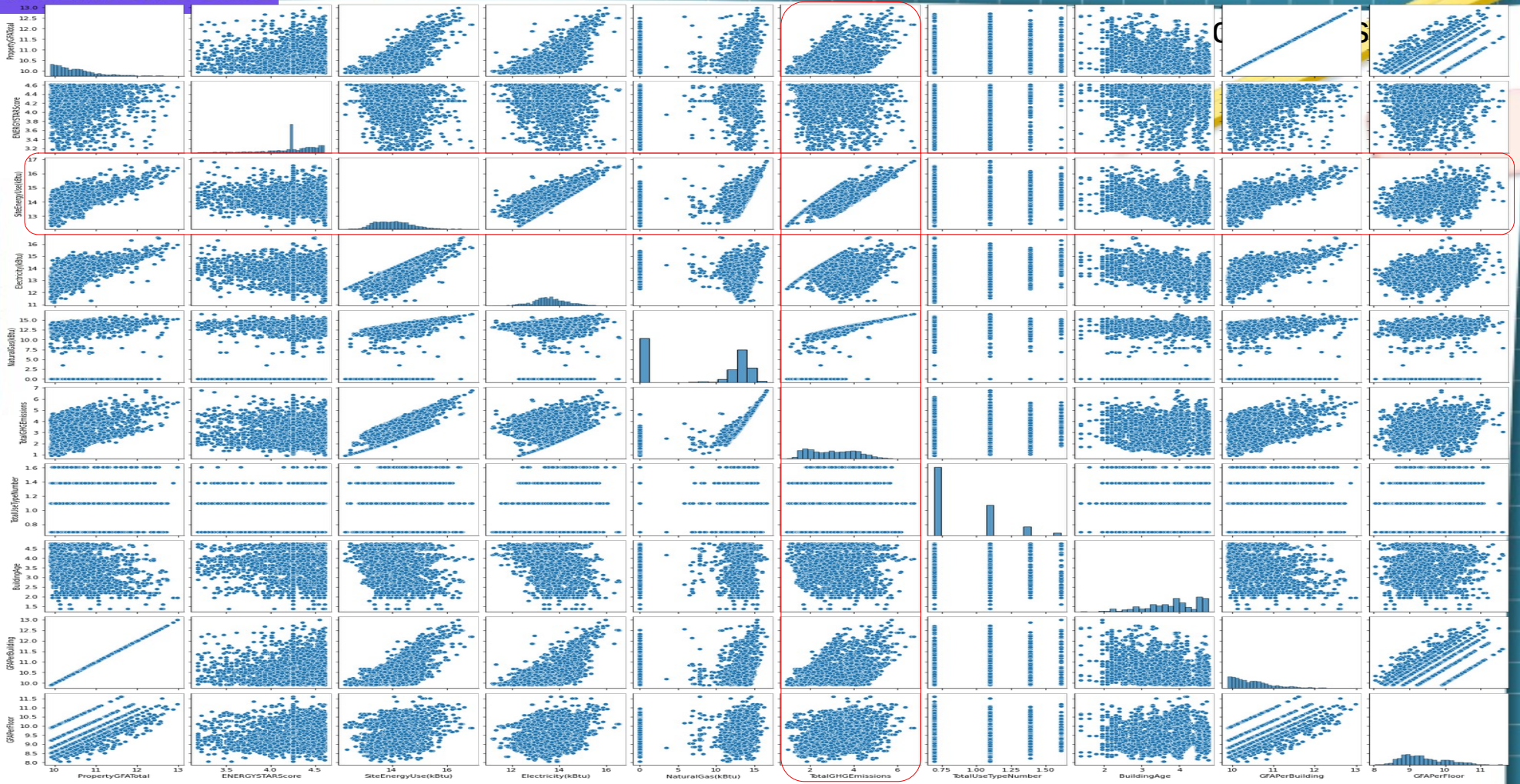
Features engineering

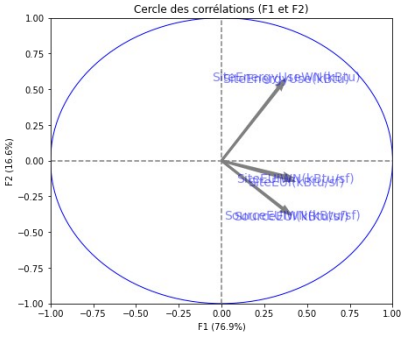
- Suppression des variables non corrélées
- Suppression une des variables dans chaque couple de variables très fortement corrélées
- Suppression des variables catégorielles sauf une

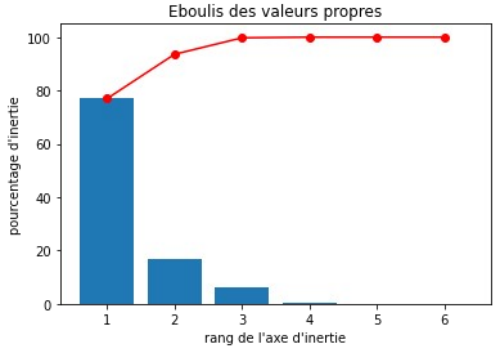


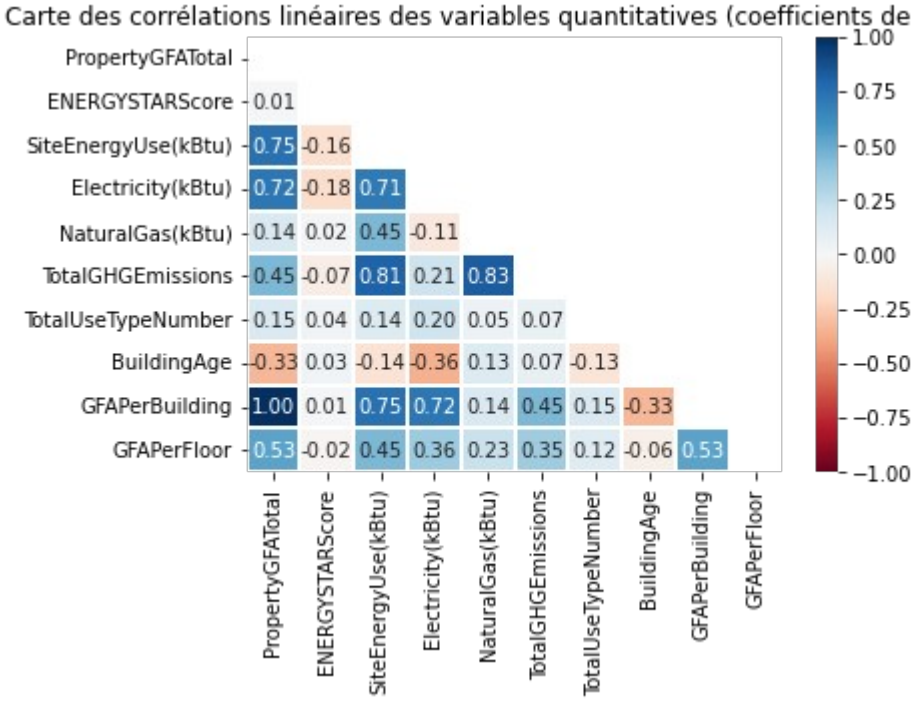





PropertyGFATotal
ENERGYSTARScore
SiteEnergyUse(kBtu)
Electricity(kBtu)
NaturalGas(kBtu)
TotalGHGEmissions
TotalUseTypeNumber
BuildingAge
GFAPerBuilding
GFAPerFloor





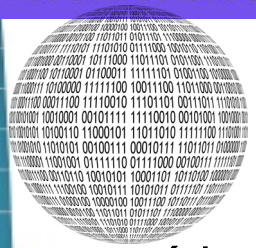


Qualitative	Quantitative	Eta_squared	Corrélation
BuildingType	GFAPerFloor	0.232881	forte
PrimaryPropertyType	GFAPerFloor	0.309613	forte
PropertyName	GFAPerFloor	0.994786	très forte
TaxParcelIdentificationNumber	GFAPerFloor	0.989107	très forte
Neighborhood	GFAPerFloor	0.201109	forte
ComplianceStatus	GFAPerFloor	0.015269	faible
Address	GFAPerFloor	0.993454	très forte

Qualitative_1	Qualitative_2	P-value
BuildingType	PrimaryPropertyType	0.000000e+00
BuildingType	TaxParcelIdentificationNumber	5.972883e-173
BuildingType	Neighborhood	2.637213e-150
BuildingType	ComplianceStatus	3.753193e-168
PrimaryPropertyType	PropertyName	1.699246e-27
PrimaryPropertyType	TaxParcelIdentificationNumber	0.000000e+00
PrimaryPropertyType	Neighborhood	0.000000e+00
PrimaryPropertyType	ComplianceStatus	5.809822e-66
PropertyName	TaxParcelIdentificationNumber	0.000000e+00
PropertyName	Neighborhood	2.034060e-08
PropertyName	Address	6.702664e-111
TaxParcelIdentificationNumber	Neighborhood	0.000000e+00
TaxParcelIdentificationNumber	Address	0.000000e+00

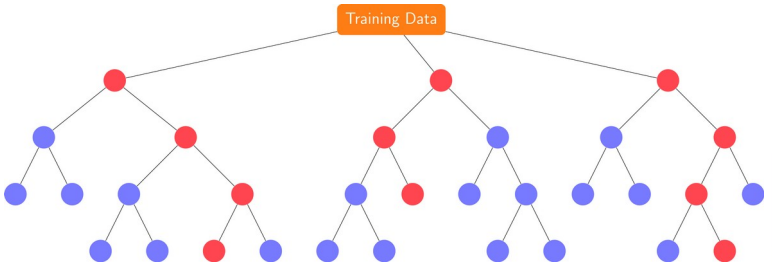
Features engineering

- Variables à étudier (targets):
  - **'TotalGHGEmissions',**
  - **'SiteEnergyUse(kBtu)'**
- Variables numériques :
  - **‘PropertyGFATotal'** : surface de plancher brut total
  - **'Electricity(kBtu)'**, : consommation annuelle d'énergie électrique
  - **'NaturalGas(kBtu)'**, : consommation annuelle de gaz naturel
  - **'TotalUseTypeNumber**', (variable créée) : Nombre utilisation
  - **'BuildingAge'**, (variable créée) : âge du batiment (à la place de l’année de construction)
  - **'GFAPerBuilding'**, (variable créée) : surface de plancher brute par bâtiment
  - **'GFAPerFloor'** (variable créée) : surface de plancher brute par bâtiment
  - **EnergyStarScore**
- Variables catégorielles :
  - **PrimaryPropertyType** : type de propriété



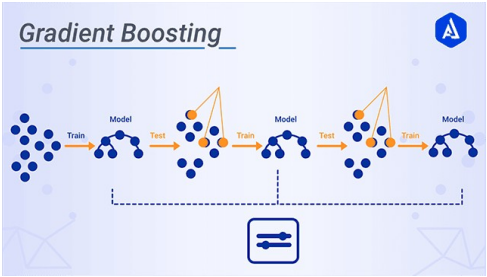
Modélisation

- Méthode
  - Création de X(Features) et y (Target)
  - Passage des variables catégorielles en numérique (OneHotEncoder)
  - Normalisation des variables numériques (StandardScaler)
  - Division BDD en entrainement et test
  - Choix des modèles d'entrainement
  - Choix des Hyperparamètres avec GridSearchCV
  - Entrainement des modèles
  - Evaluation des modèles



Modélisation

- Modèle linéaire
  - ElasticNet : tol=1e-4, alpha=1e-4, L1_ratio=0.9
- Modèle non linéaire Support Vector Machine
  - SVR : 'degree': 2, 'gamma': 'auto', 'kernel': 'poly'
- Modèles ensemblistes
  - Random Forest :
    - 'max_features': 'auto', 'min_samples_leaf': 1, 'n_estimators': 300
  - XGBoost : les mêmes + 'criterion': 'mse', 'loss': 'ls',



Modélisation : évaluation

- SiteEnergyUse(kBtu)

Modèle	RMSE	MAE	R²	Time_ms	Cv score (RMSE)
ElasticNet	0.301373	0.223752	0.816534	2.442177	0.280948
SVM SVR	0.301373	0.223752	0.816534	2.460181	0.280948
Random Forest	0.044507	0.016996	0.995999	24.558900	0.039260
XGBoost	0.042505	0.024471	0.996351	4.775324	0.040210

- TotalGHGEmissions

Modèle	RMSE	MAE	R²	Time_ms	Cv score (RMSE)
ElasticNet	0.467245	0.346584	0.820252	2.447647	0.447659
SVM SVR	0.467245	0.346584	0.820252	2.445876	0.447659
Random Forest	0.042613	0.015209	0.998505	71.589479	0.036778
XGBoost	0.034725	0.018566	0.999007	4.913400	0.033907

Modélisation : ajout energy starscore

- TotalGHGEmissions

Modèle	RMSE	MAE	R²	Time_ms	Cv score (RMSE)
ElasticNet	**0.458924**	**0.335618**	**0.826597**	2.443828	**0.439601**
SVM SVR	**0.458924**	**0.335618**	**0.826597**	**2.474696**	**0.439601**
Random Forest	**0.042130**	0.015228	**0.998539**	**68.779979**	0.037028
XGBoost	0.035866	0.018737	0.998941	**4.552764**	0.034030

Amélioration du modèle

- Choix du modèle : XGBoost
  - Meilleur performance (score)
    - Modèles ensemblistes RandomForest et XGBoost
  - Temps le plus rapide
    - XGBoost
- EnergyStarScore
  - Meilleur performance sur modèles linéaire et SVM
  - Temps plus rapide avec modèles ensemblistes

Conclusion

- Les points importants avant la prédiction
  - La compréhension des variables
  - Le nettoyage des données
  - La standardisation et la normalisation des variables
  - Le choix des variables
- Les points importants dans la prédiction
  - Les hyperparamètres
  - Accepter les erreurs du modèle