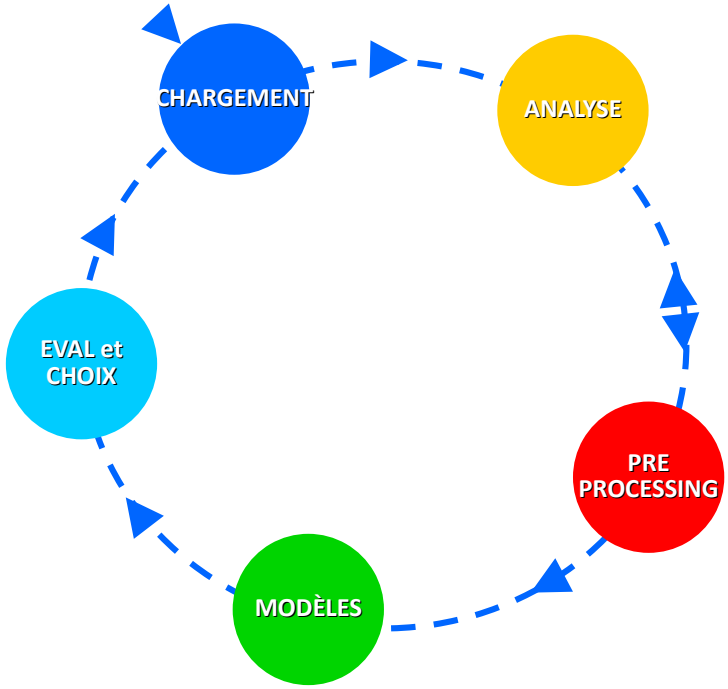
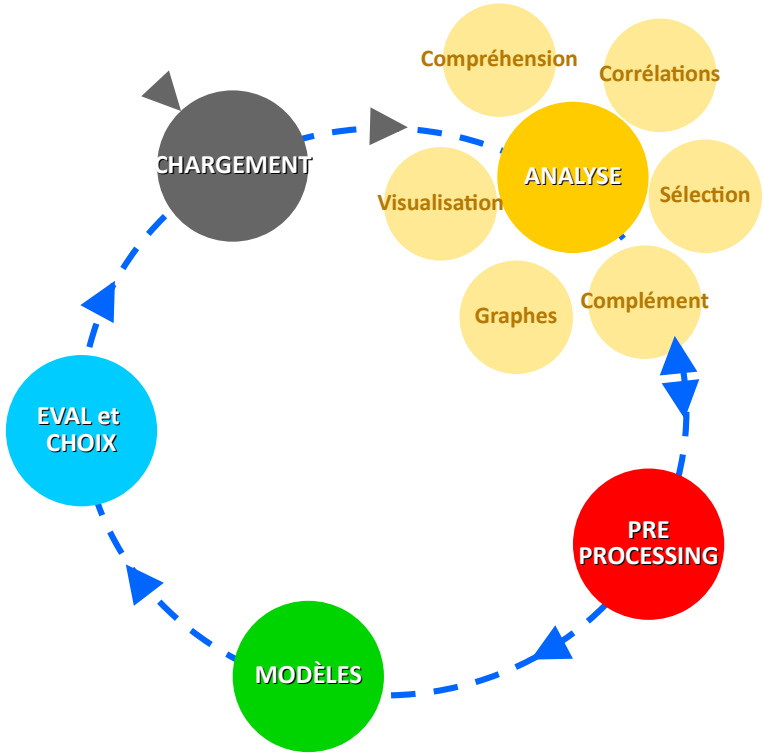
Par Aurélie RAOUL

Process général



Sommaire


CHARGEMENT
ANALYSE
Compréhension
Corrélations
Visualisation
Sélection
Graphes
Complément
PRE PROCESSING
MODÈLES
EVAL et CHOIX

ANALYSE

Analyse de la forme

Shape : 20640 x 10

Type

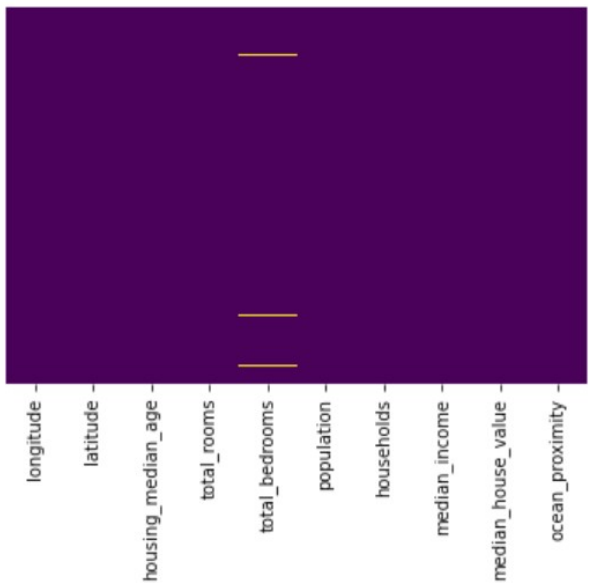
```
longitude             float64
latitude              float64
housing_median_age    float64
total_rooms           float64
total_bedrooms        float64
population            float64
households            float64
median_income         float64
median_house_value    float64
ocean_proximity        object
dtype: object
```

NaN

```
data.isna().sum()
✓ 0.4s
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms      207
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
```

Describe

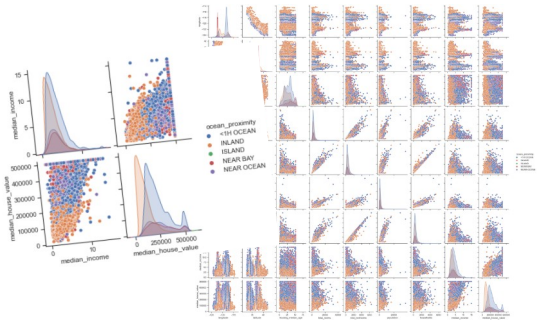
Target



ANALYSE

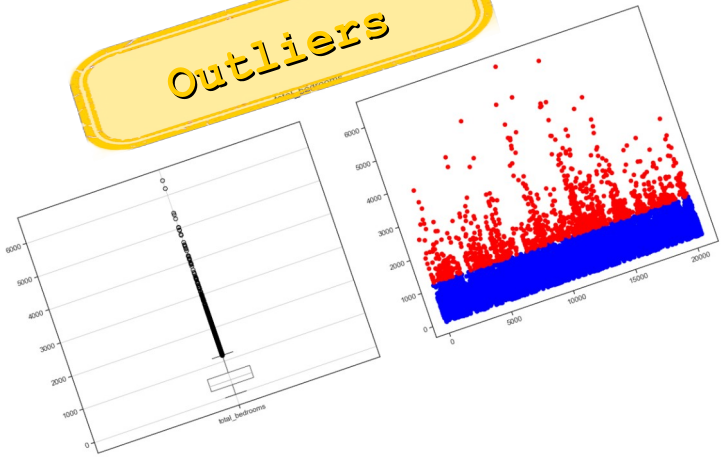
Analyse du fond

Pairplot

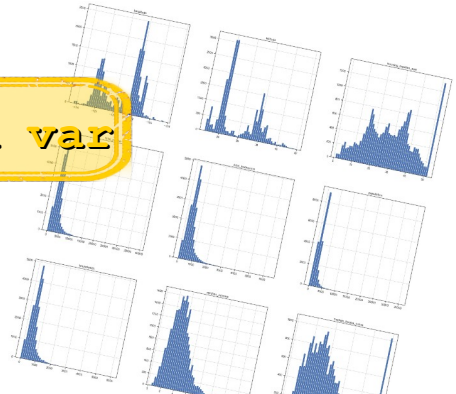


Corrélations

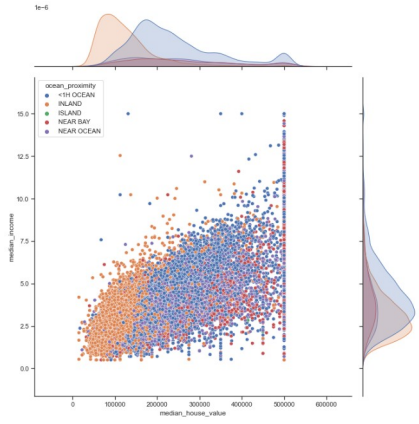
Outliers



Hist. Num var



Chi² :
- longitude
- latitude
- age
- chambres
- population

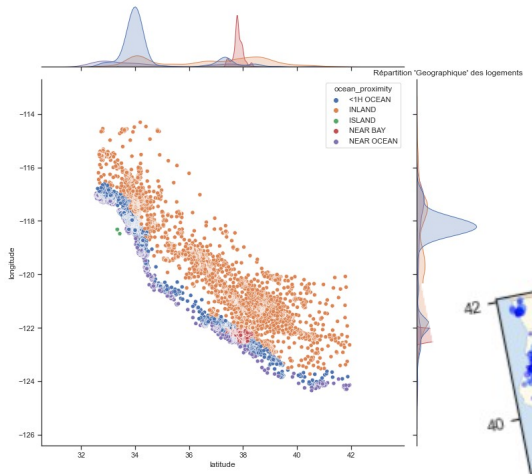


Recherches

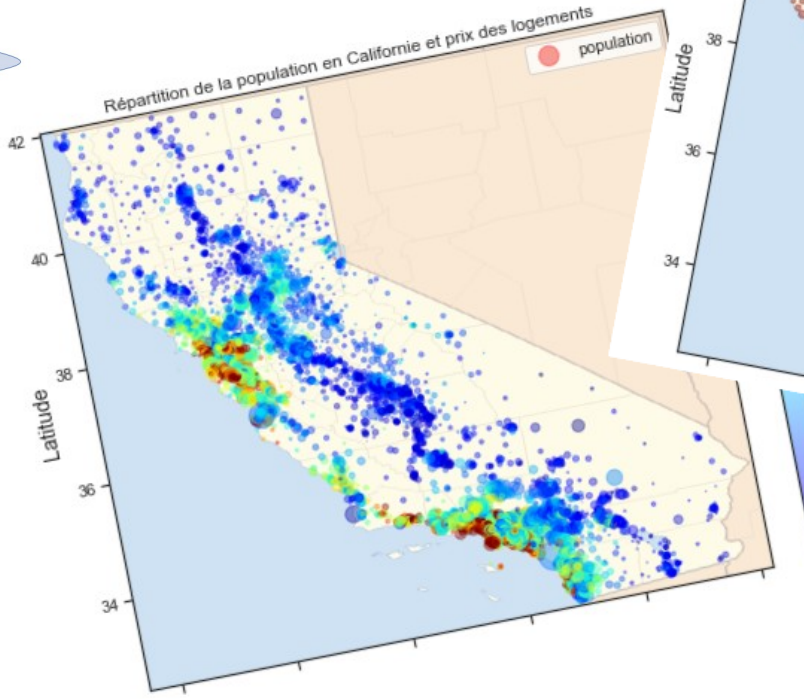
Logements californiens by Aurélie RAOUL

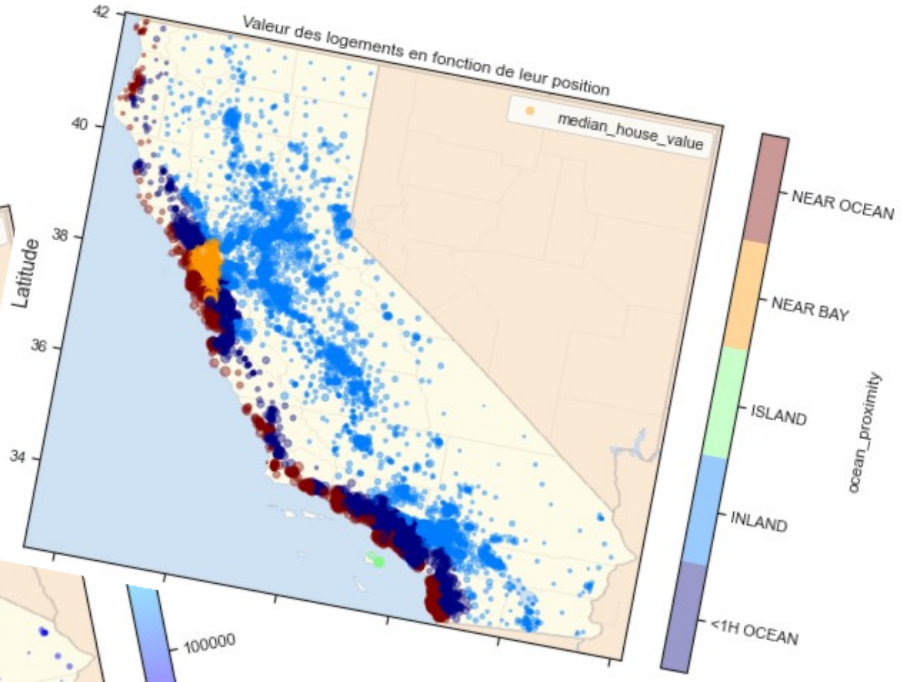
ANALYSE

2nd niveau



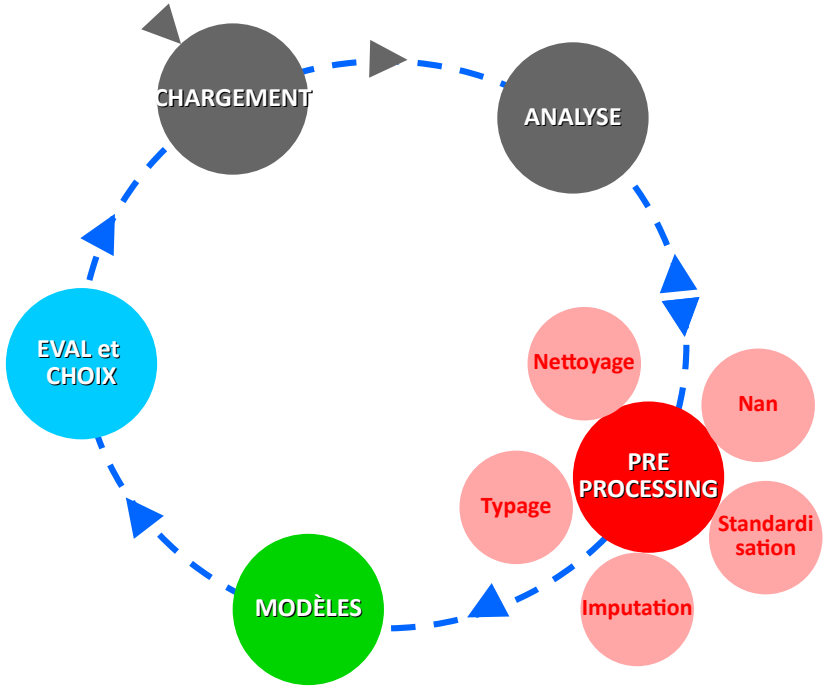
MAP





Logements californiens by Aurélie RAOUL

Sommaire


CHARGEMENT
ANALYSE
EVAL et CHOIX
MODÈLES
PRE PROCESSING
Nettoyage
Nan
Typage
Standardisation
Imputation

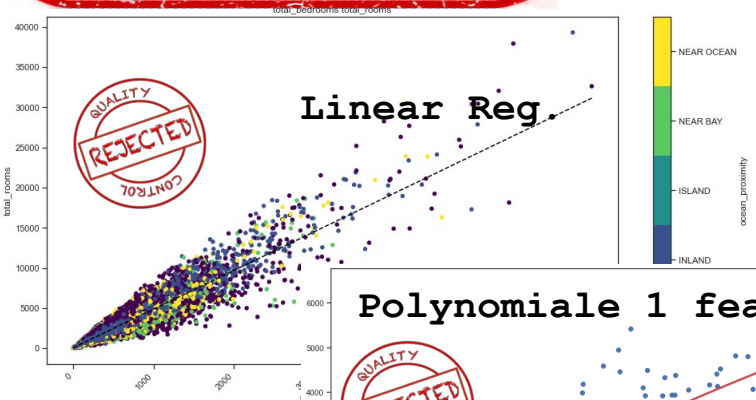
PRE-PROCESSING

Nettoyage

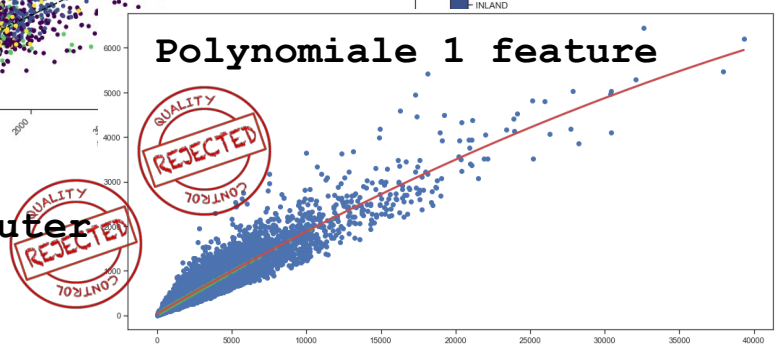
Type

ocean_proximity

Imputation / NaN
total_bedrooms



SimpleImputer



Polynomiale :
total_rooms + households + population

Encodage

OneHotEncoder

Valeurs aberrantes

Standardisation

PRE-PROCESSING

Nettoyage

APPROVED

Type

ocean_proximity

Imputation / NaN

total_bedrooms

Polynomiale :
total_rooms + households + population

Valeurs aberrantes

3 lignes DELETE

Encodage

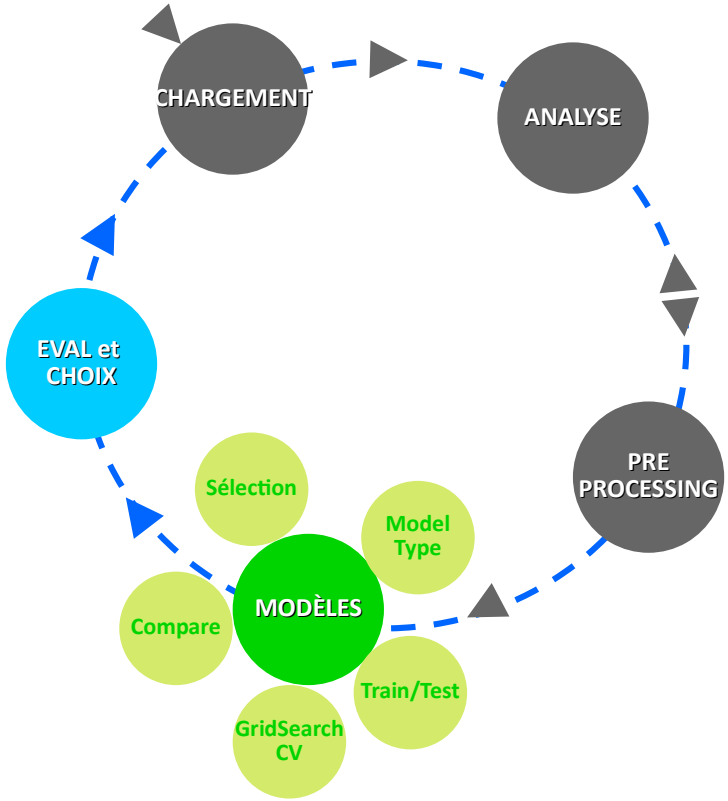
OneHotEncoder

Outliers

KEEP ✓

Logements californiens by Aurélie RAOUL

Sommaire


CHARGEMENT
ANALYSE
PRE PROCESSING
MODÈLES
Sélection
Model Type
Train/Test
GridSearch CV
Compare
EVAL et CHOIX

MODÈLES

Split dataset

Train | Test

Features ?

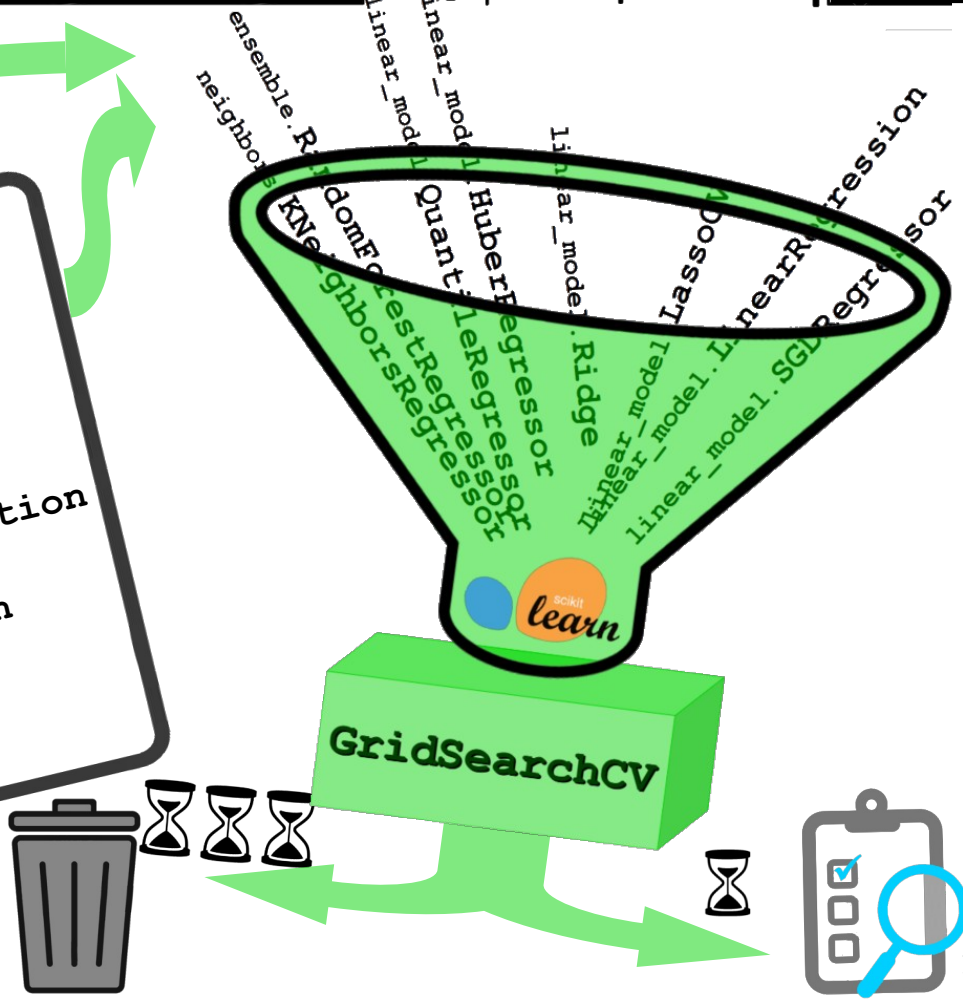
Manuelle

Variance Threshold

SelectKBest

SelectFromModel

RFECV

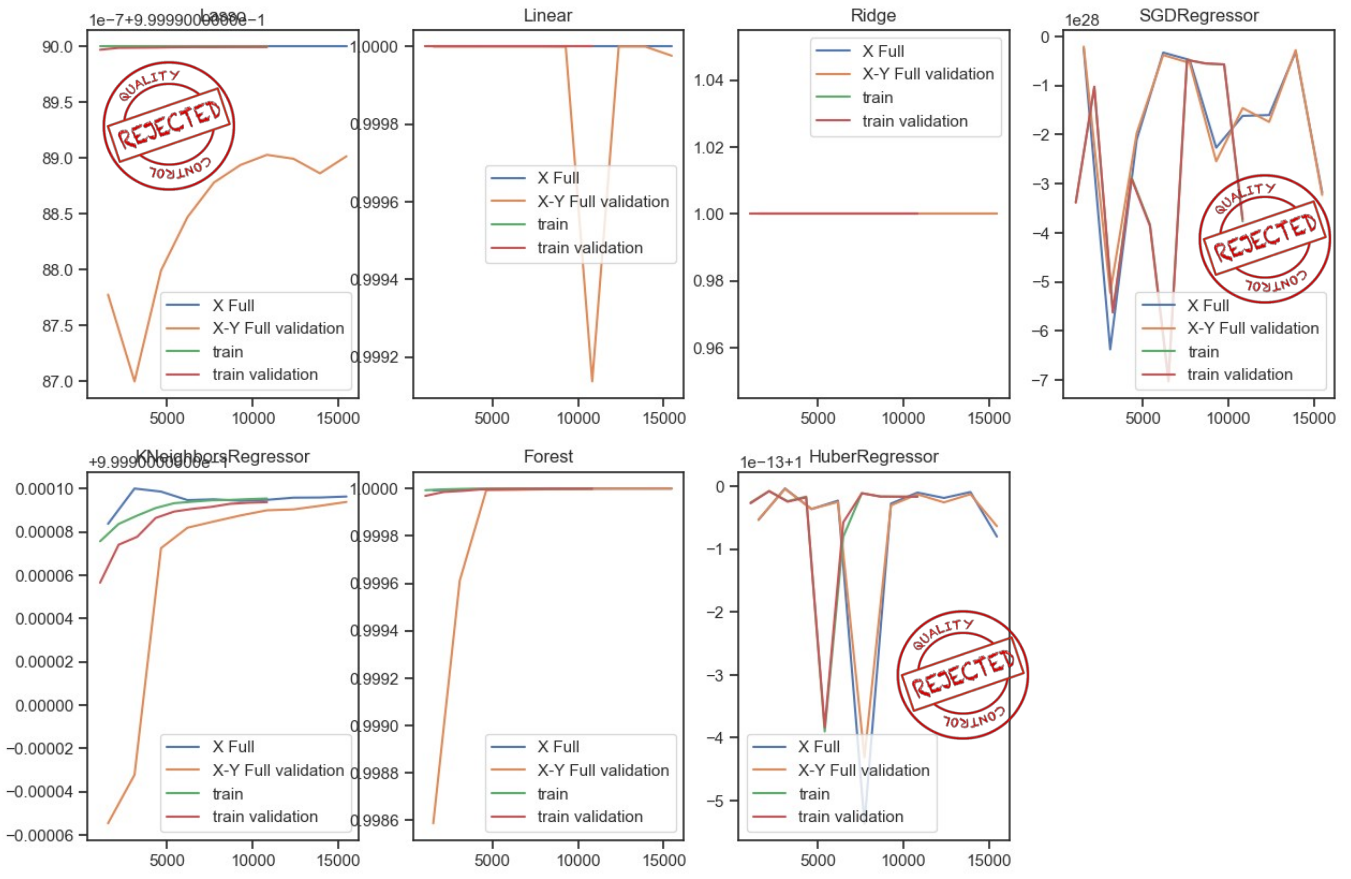


Logements californiens by Aurélie RAOUL

MODÈLES

Courbe d'apprentissage

learning_curve > x=train_sizes



MODÈLES

Train/Test → Model Type → GridSearch CV → Compare → Sélection

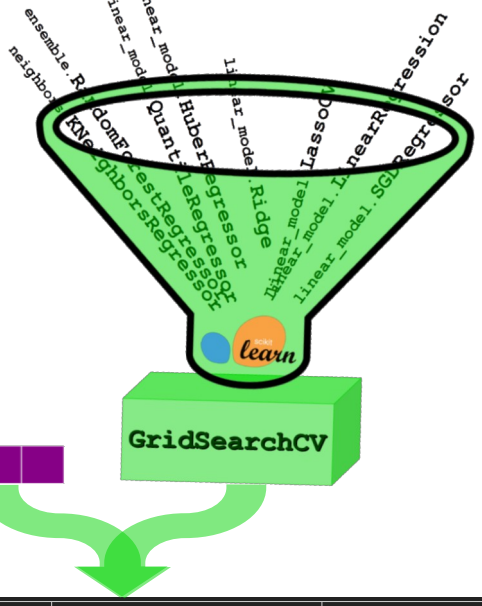
Split dataset

Train | Test

70 / 30

Features ?

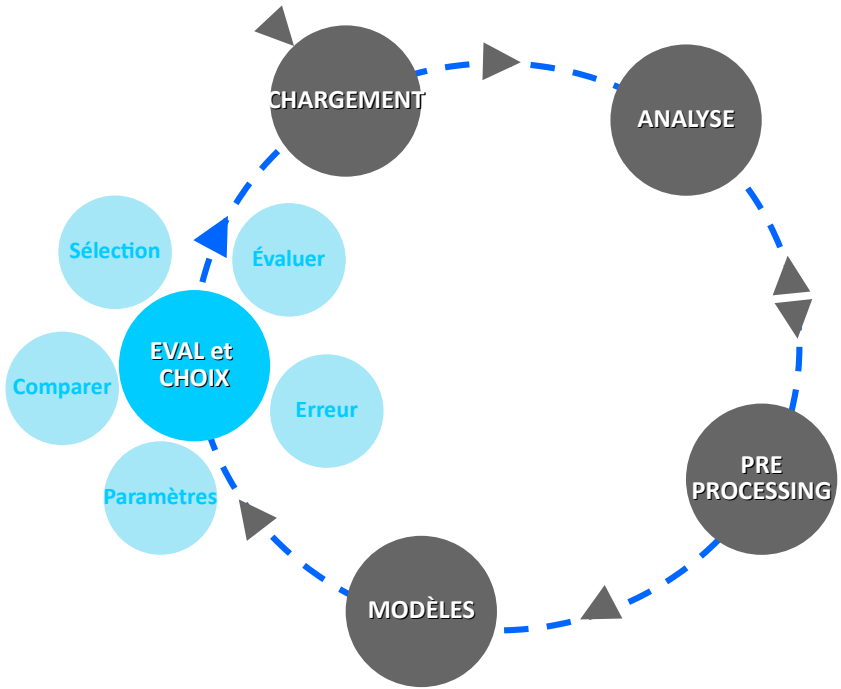
Manuelle



APPROVED

Model	Score (Toutes colonnes)	Paramètres
RandomForestRegressor	0.9999996421039528	{'randomforestregressor__max_features': 'auto', 'randomforestregressor__n_estimators': 81}
Linear	1.0	{'linearregression__fit_intercept': True, 'linearregression__normalize': False, 'linearregression__positive': True}
Ridge	1.0	{'ridge__alpha': 1, 'ridge__fit_intercept': True, 'ridge__solver': 'auto'}
Lasso	0.9999989999649406	{'lassocv__alphas': None, 'lassocv__fit_intercept': True}
KNeighborsRegressor	0.9999946882516152	{'kneighborsregressor__n_neighbors': 7}
SGDRegressor	-2.4498384279180864e+26	{'sgdregressor__fit_intercept': True, 'sgdregressor__loss': 'squared_error', 'sgdregressor__penalty': 'l1'}
HuberRegressor	0.9999999999999741	{'huberregressor__fit_intercept': False}
QuantileRegressor	Trop long à exécuter	

Sommaire


CHARGEMENT
ANALYSE
PRE PROCESSING
MODÈLES
EVAL et CHOIX
Sélection
Évaluer
Comparer
Erreur
Paramètres

ÉVALUATION ET CHOIX

Erreur | Évaluer | Comparer | Paramètres | Sélection

Métriques

Modèle	R2	MAE	MSE	RMSE	Media AE
Lasso	1.0	91.296	13289.368	115.28	76.549
Linear	1.0,	0.0	0.0	0.0	0.0
Ridge	1.0,	0.0	0.0	0.0	0.0
SGDRegressor	-2.44..e+26	1.59..e+18	3.2555..e+36	1.8031..e+18	1.405...e+18
KNeighborsRegressor	1.0	125.487	70587.303	265.683	71.429
Forest	1.0	21.317	4756.045	68.964	6.173
HuberRegressor	1.0	0.015	0.0	0.019	0.014

R2 = Coefficient de Détermination

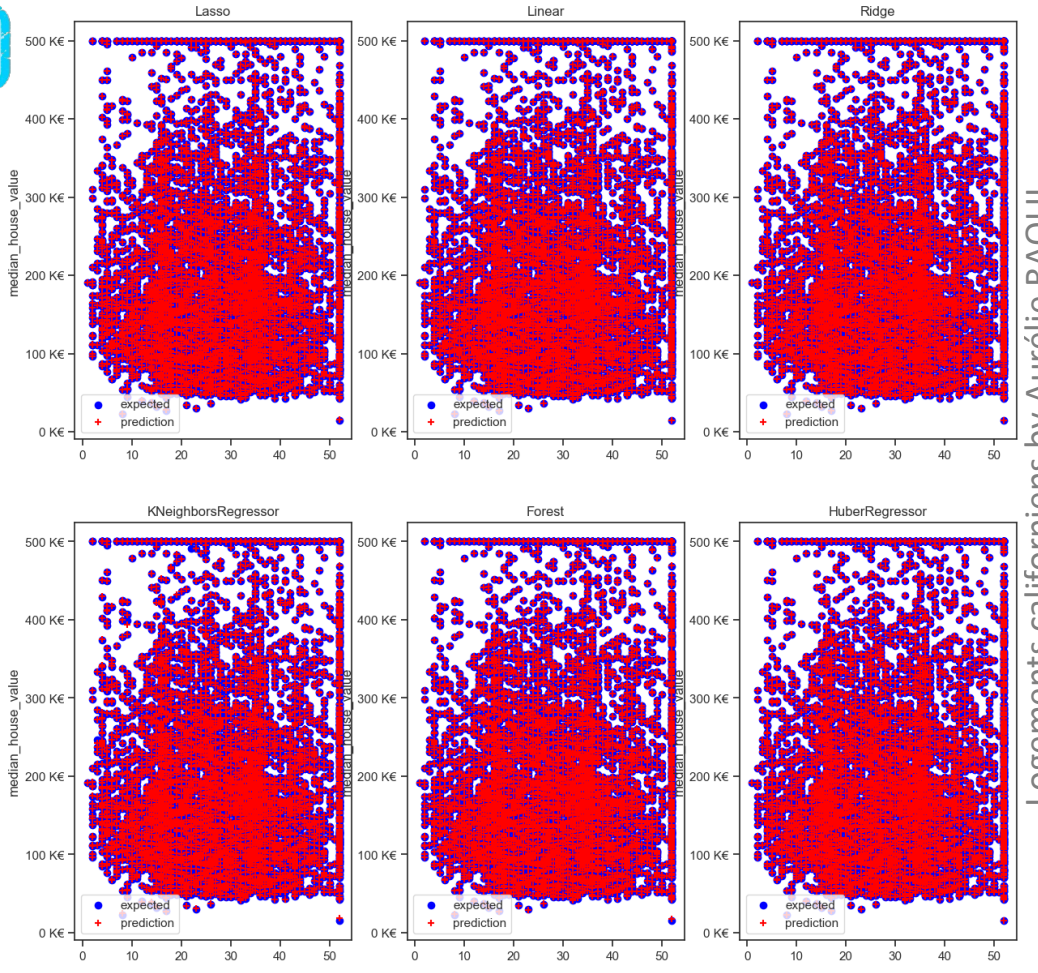
MAE = Mean Absolute Error : l'importance d'une erreur est linéaire avec son amplitude. Si le dataset contient des outliers

MSE = Mean Squared Error : vous accordez une grande importance aux grandes erreurs

RMSE = Root Mean Squared Error : Remise à l'échelle, donc racine carré de MSE

Media AE = Median Absolute Error : très peu sensible aux outliers

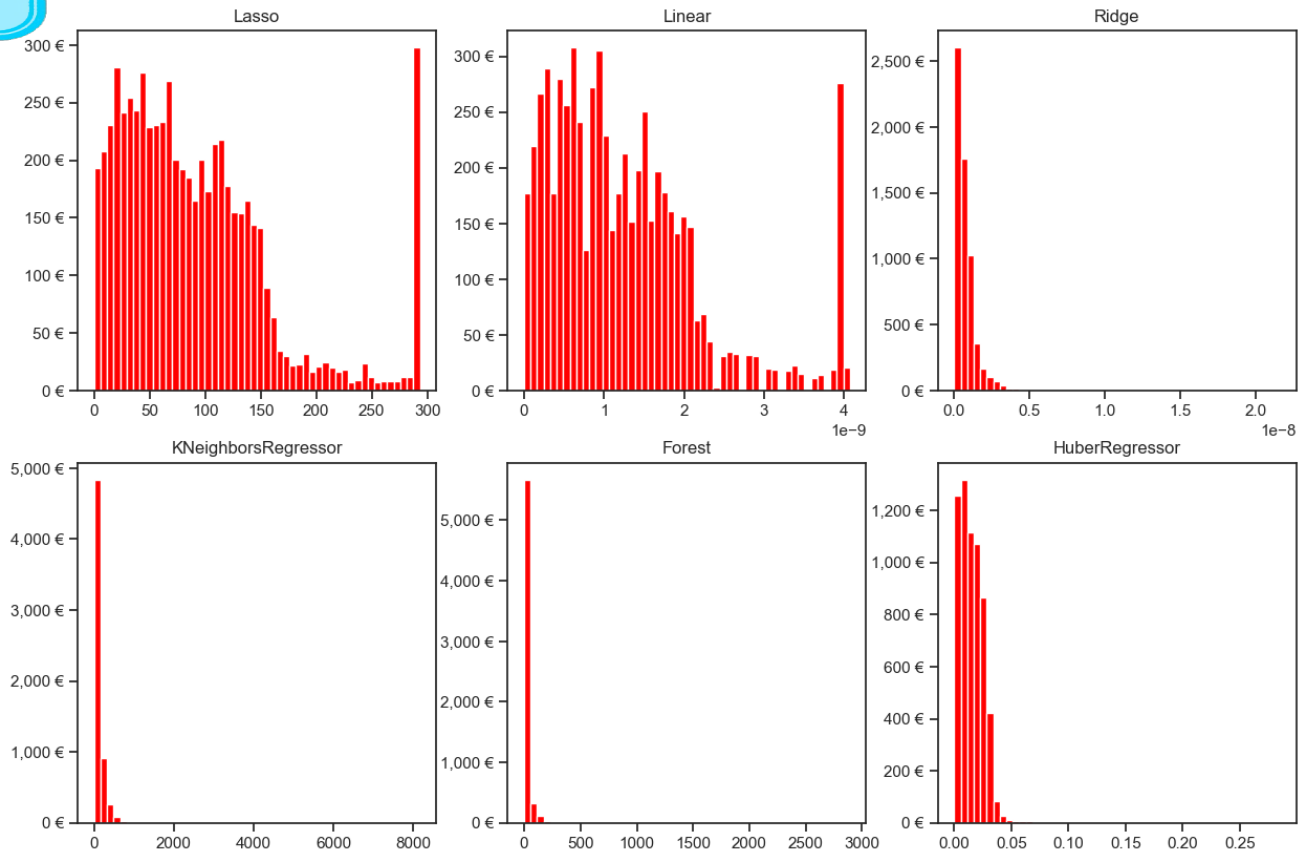
Visualisation de la cible



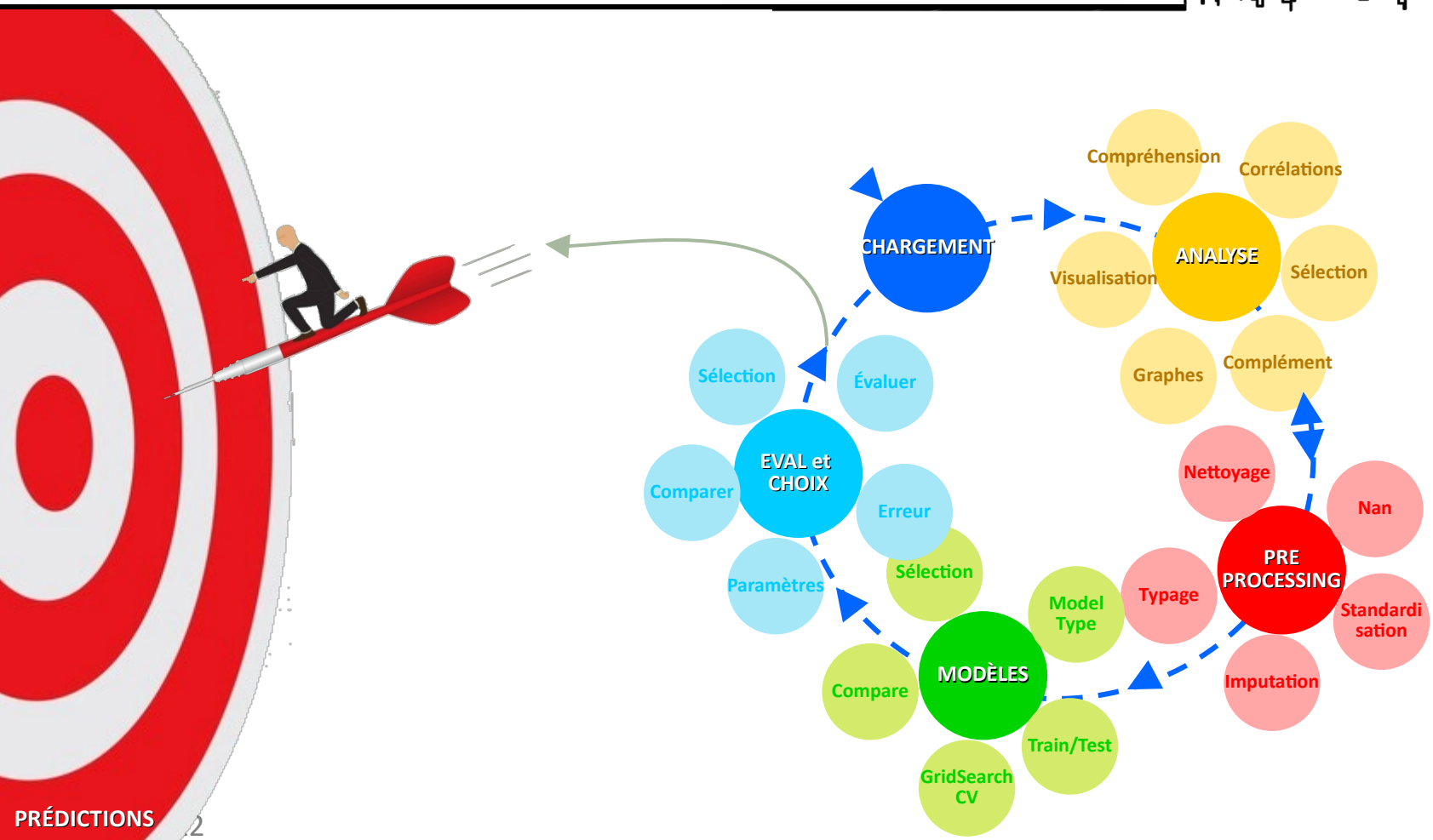
ÉVALUATION ET CHOIX

Erreurs

Représentation du montant des erreurs par nombre (montant en moins par rapport à l'attendu).



Synthèse



Logements californiens by Aurélie RAOUL

Discussion