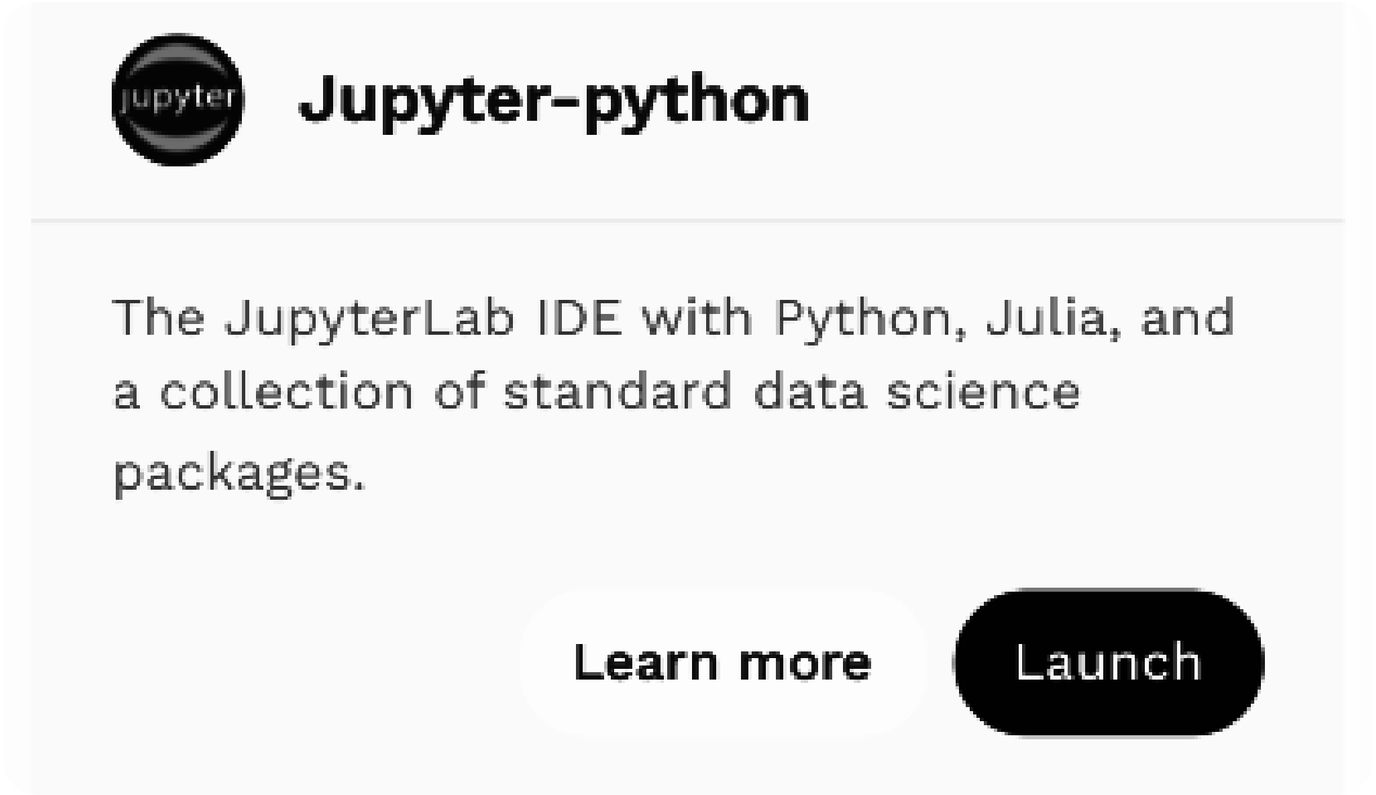
[T u x a e]

[Ateliers Data]/[3. Régression. Prédire le prix des maisons ?]

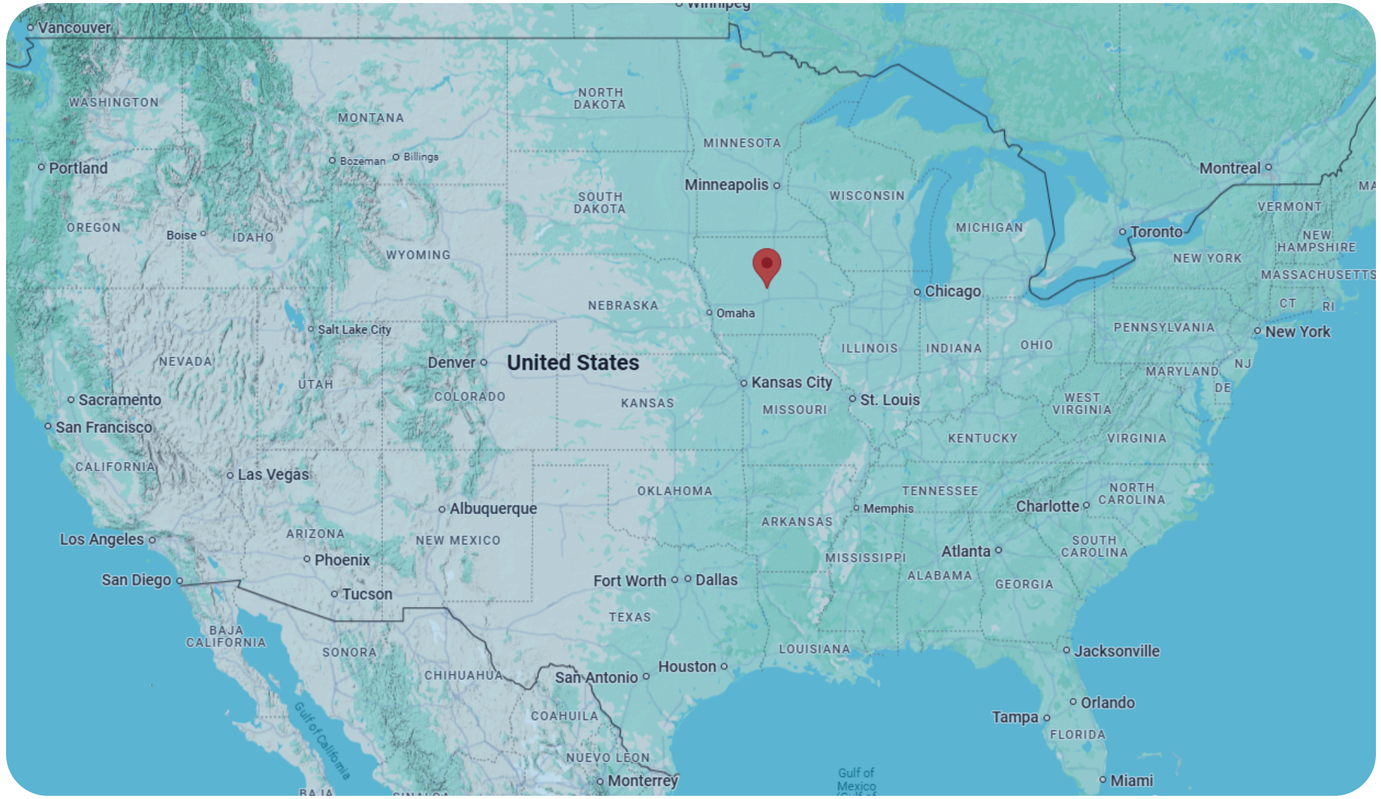
Introduction



https://shorturl.at/LbZDW

Le problème

Le problème

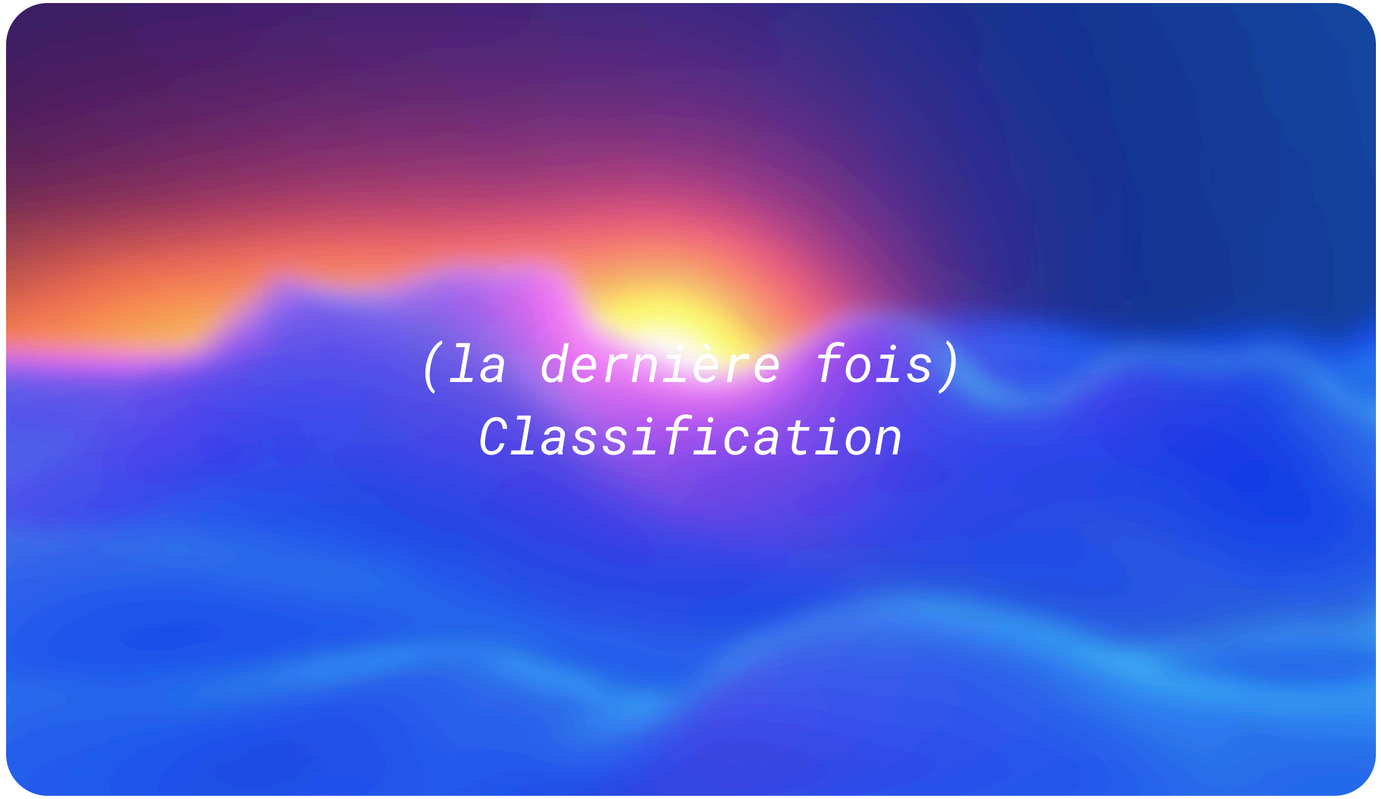


Ames, Iowa

Le problème

Ames, Iowa

Le problème



Le problème

Le dataset

Le dataset

Boston Housing Dataset
(Harrison and Rubinfeld, 1978)

506 observations, 14 variables

Le dataset

Dean De Cock, 2011

(https://jse.amstat.org/v19n3/decock.pdf)

Le dataset

Dean De Cock, 2011

(https://jse.amstat.org/v19n3/decock.pdf)

3970 observations, 113 variables

Le dataset

Ames Housing Dataset
(De Cock, 2011)

1460 observations, 80 variables

Le dataset

34 variables numériques

Le dataset

34 variables numériques

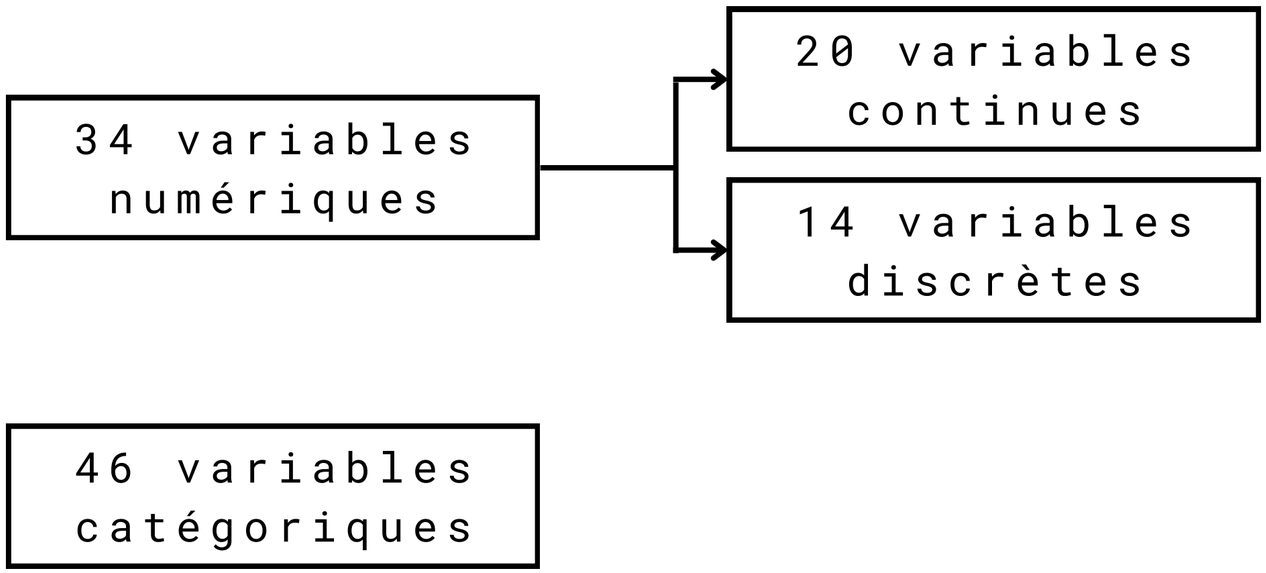
46 variables catégoriques

Le dataset

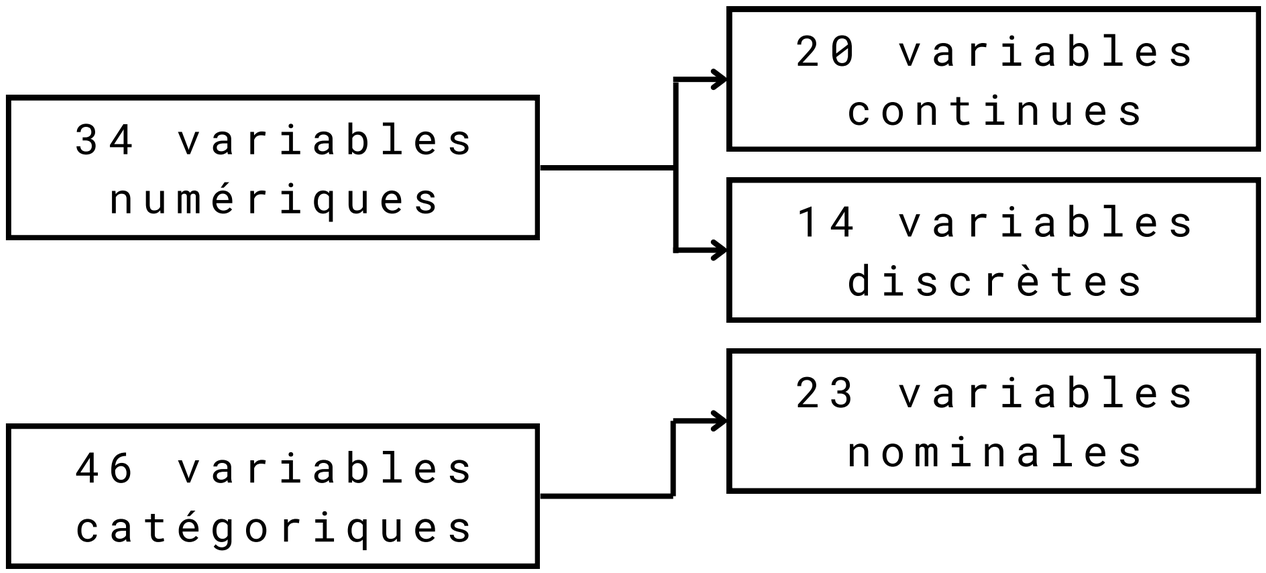
34 variables numériques → 20 variables continues

46 variables catégoriques

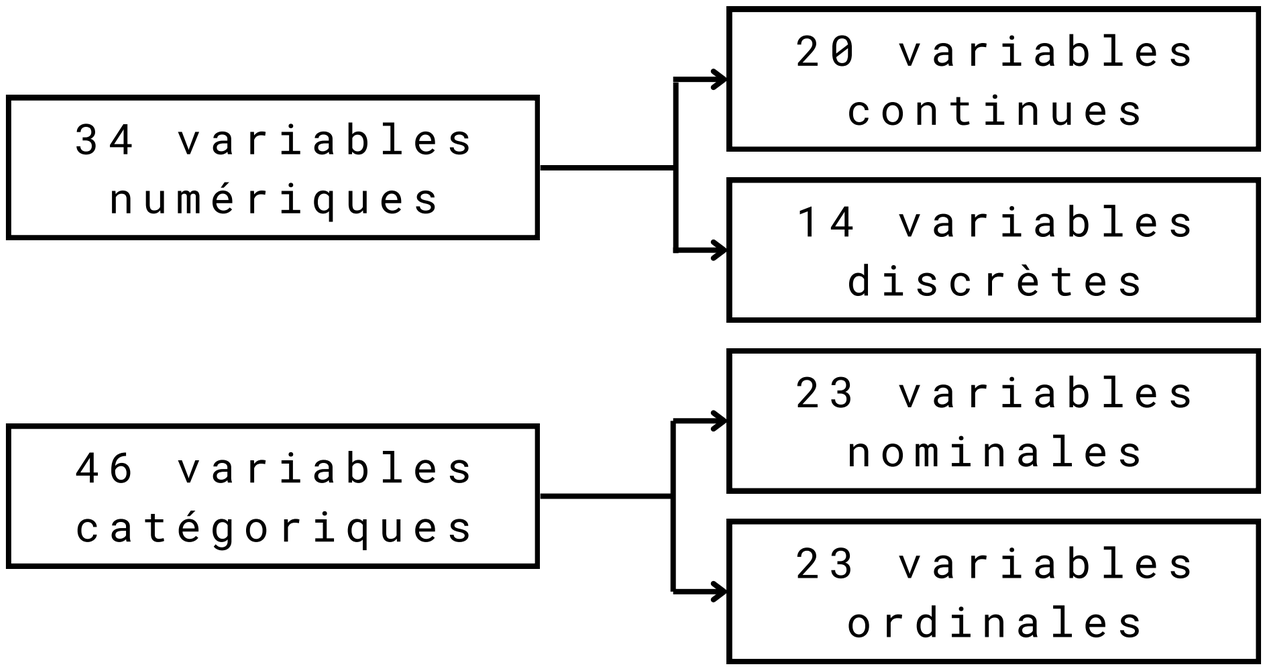
Le dataset


34 variables numériques
20 variables continues
14 variables discrètes
46 variables catégoriques

Le dataset


34 variables numériques
20 variables continues
14 variables discrètes
46 variables catégoriques
23 variables nominales

Le dataset


34 variables numériques
20 variables continues
14 variables discrètes
46 variables catégoriques
23 variables nominales
23 variables ordinales

Le dataset

- 34 variables numériques
  - 20 variables continues
  - 14 variables discrètes
- 46 variables catégoriques
  - 23 variables nominales
  - 23 variables ordinales

GarageArea

Taille du garage en pieds carrés.

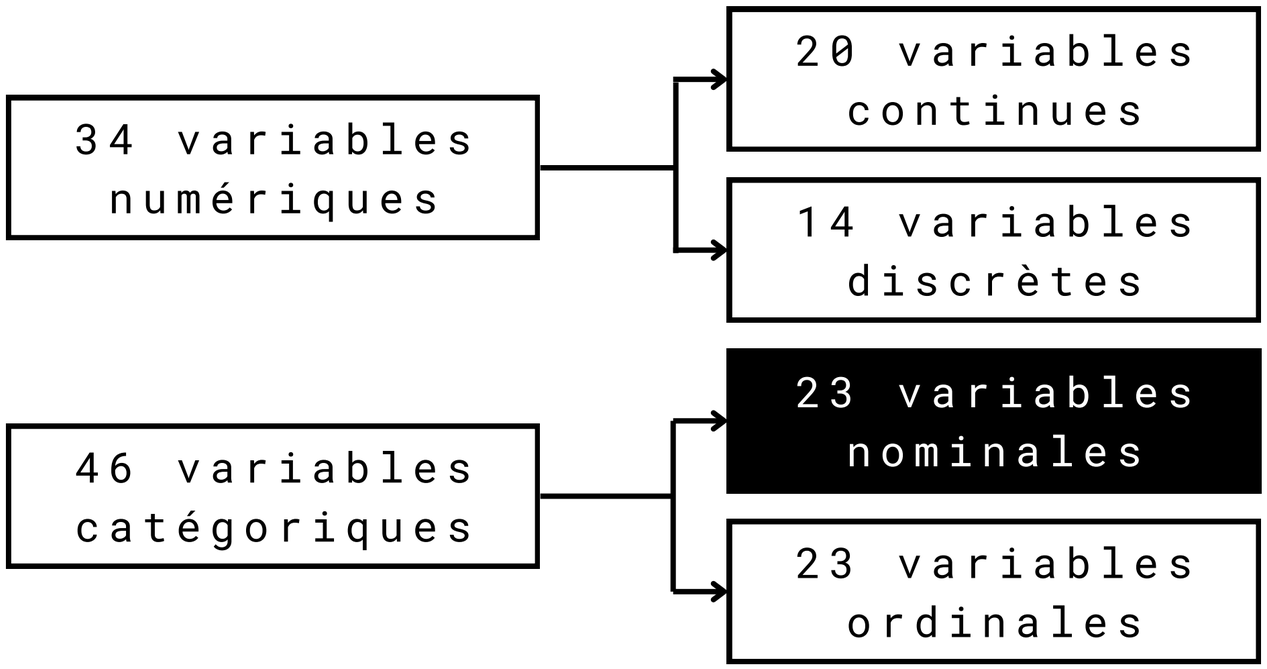
Le dataset

- 34 variables numériques
  - 20 variables continues
  - 14 variables discrètes
- 46 variables catégoriques
  - 23 variables nominales
  - 23 variables ordinales

GarageCars

Capacité du garage en nombre de voitures.

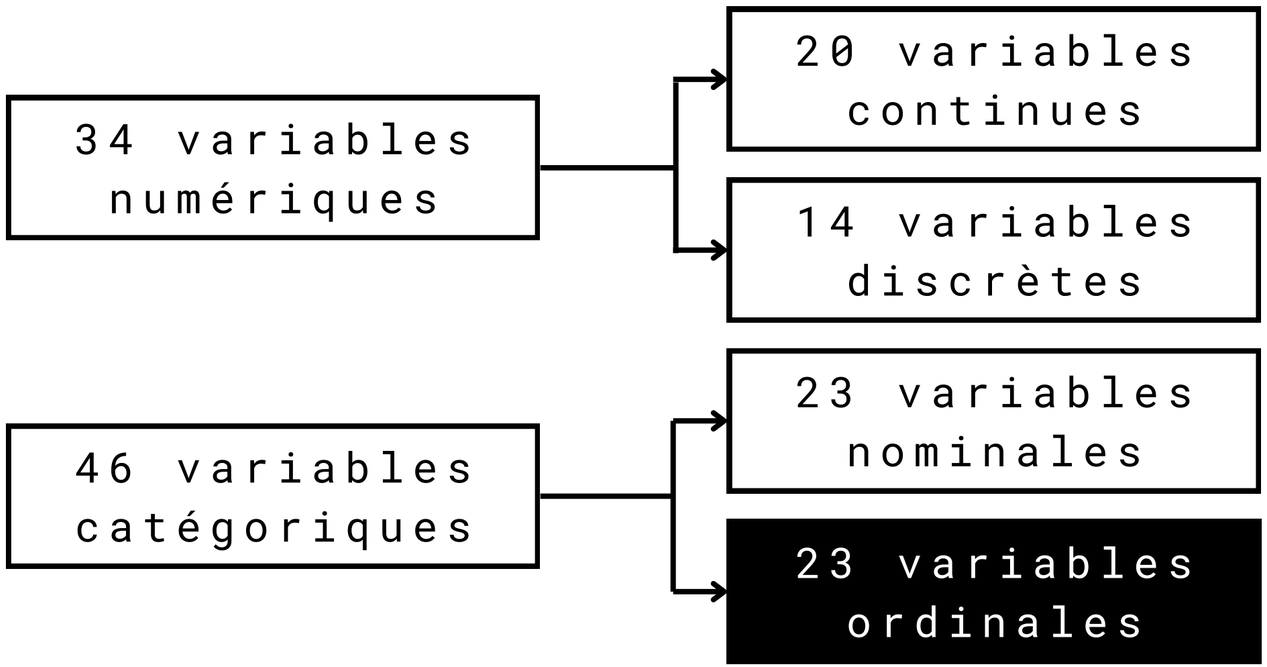
Le dataset



RoofMatl

- *ClyTile* : Argile ou tuile
- *CompShg* : Bardeaux standards (composites)
- *Membran* : Membrane
- *Metal* : Métal
- *Roll* : Rouleau
- *Tar&Grv* : Gravier et goudron
- *WdShake* : Bardeaux de bois (secoués)
- *WdShngl* : Bardeaux de bois (classiques)

Le dataset



GarageQual

- *Ex* : Excellente
- *Gd* : Bonne
- *TA* : Typique/Moyenne
- *Fa* : Passable
- *Po* : Mauvaise
- *NA* : Pas de garage (non applicable)

Le dataset

- 34 variables numériques
  - 20 variables continues
  - 14 variables discrètes
- 46 variables catégoriques
  - 23 variables nominales
  - 23 variables ordinales

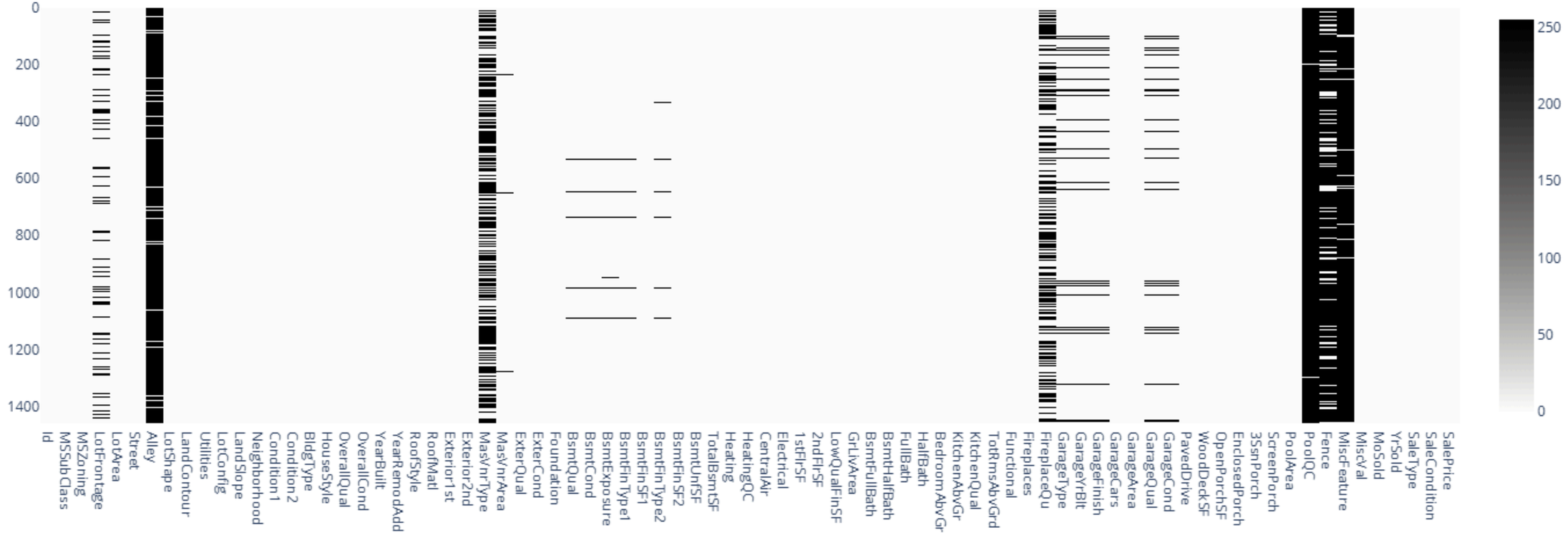
SalePrice

Prix de vente en $

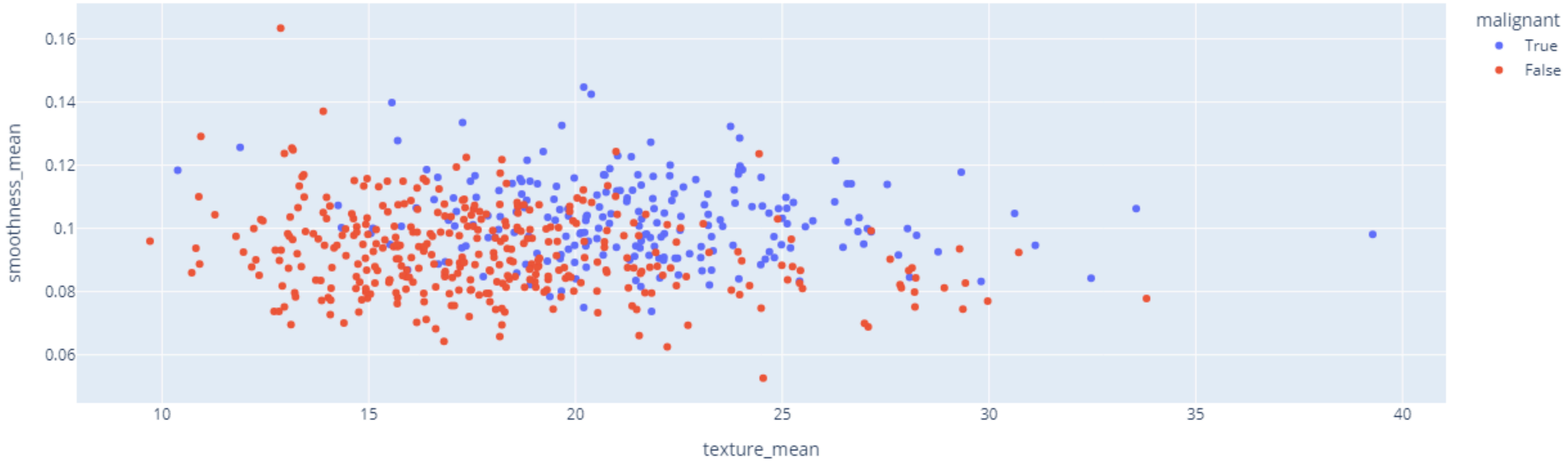
> > > > > > > > >

Au boulot !

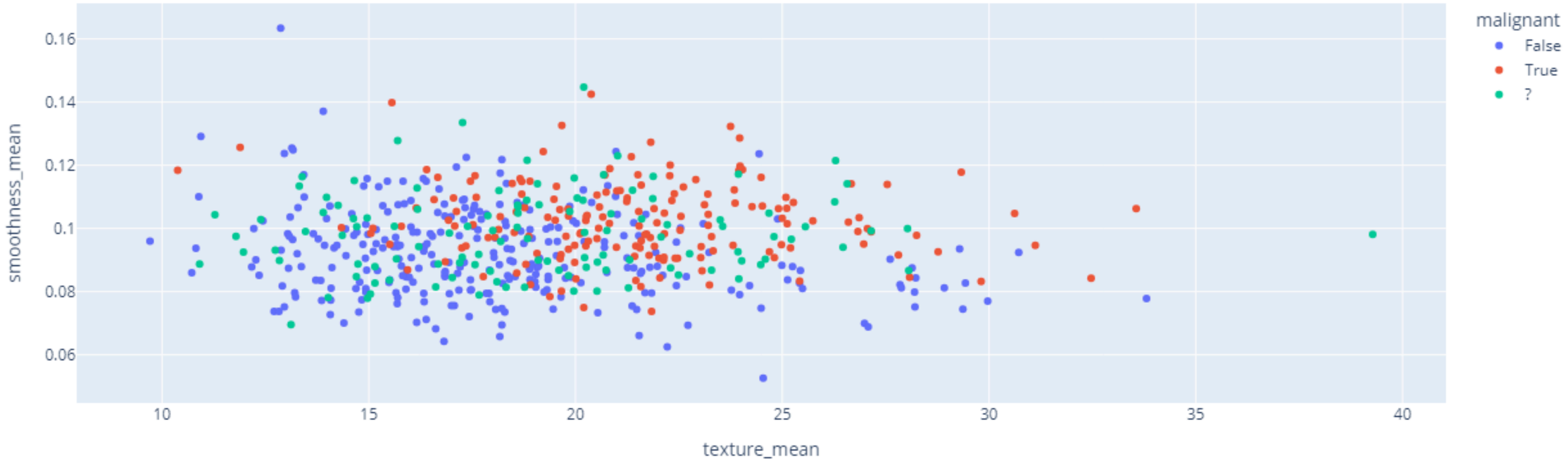
Analyse des données



Rappel du dernier atelier



Rappel du dernier atelier



Métriques

Métriques

Exactitude
(Accuracy)

Métriques

Exactitude (Accuracy)

Précision (Precision)

Métriques

Exactitude (Accuracy)

Précision (Precision)

Rappel (Recall)

Métriques

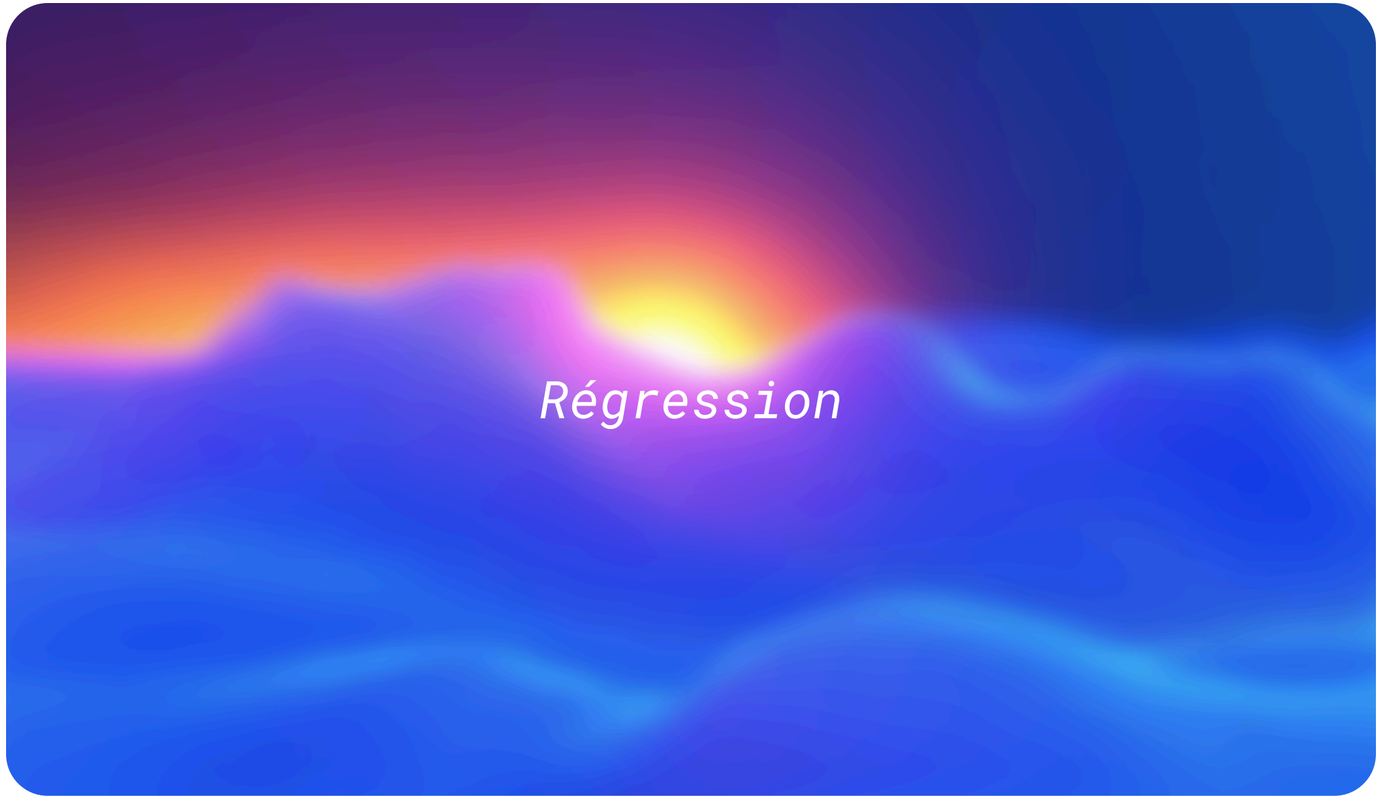
Exactitude (Accuracy)

Précision (Precision)

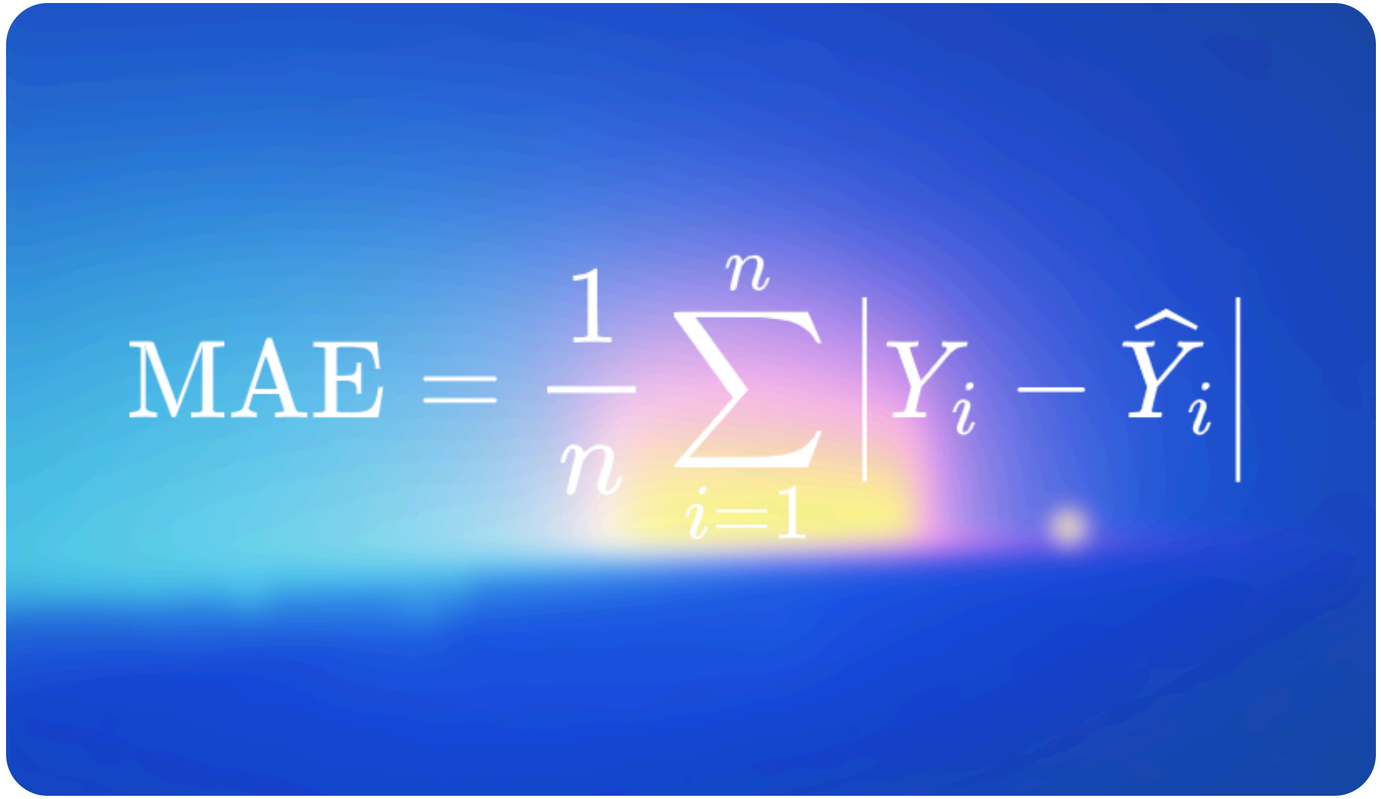
Rappel (Recall)

Score F1 (F1 Score)

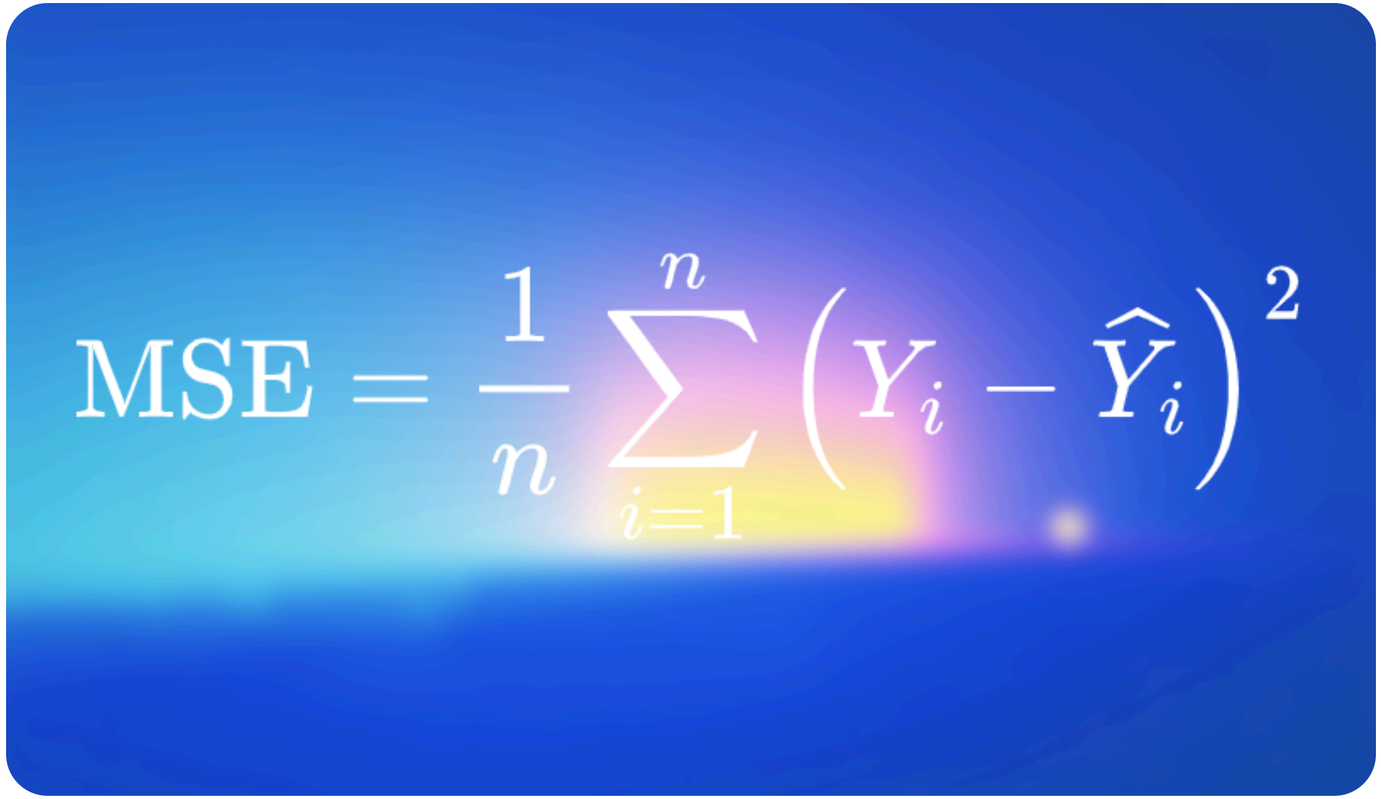
Métriques



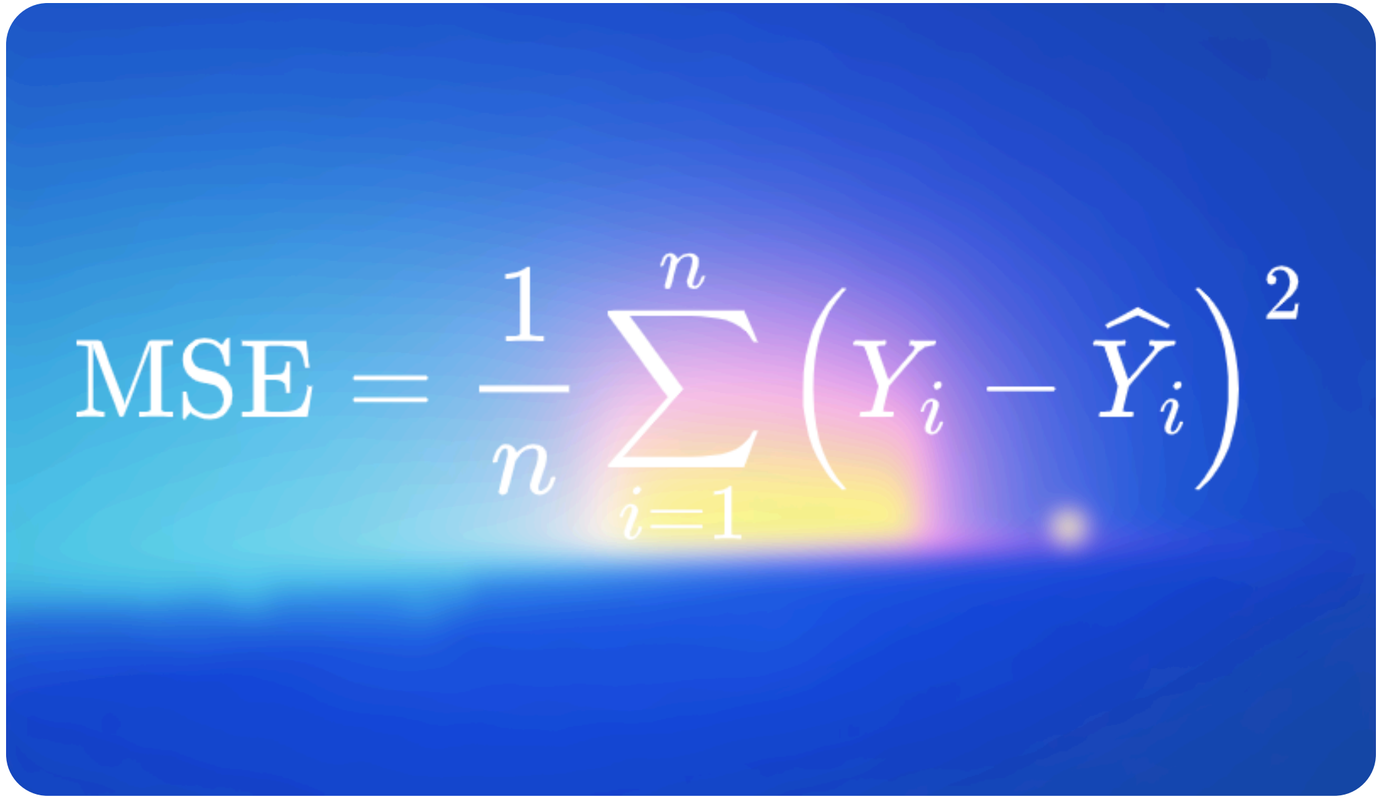
Erreur Absolue Moyenne



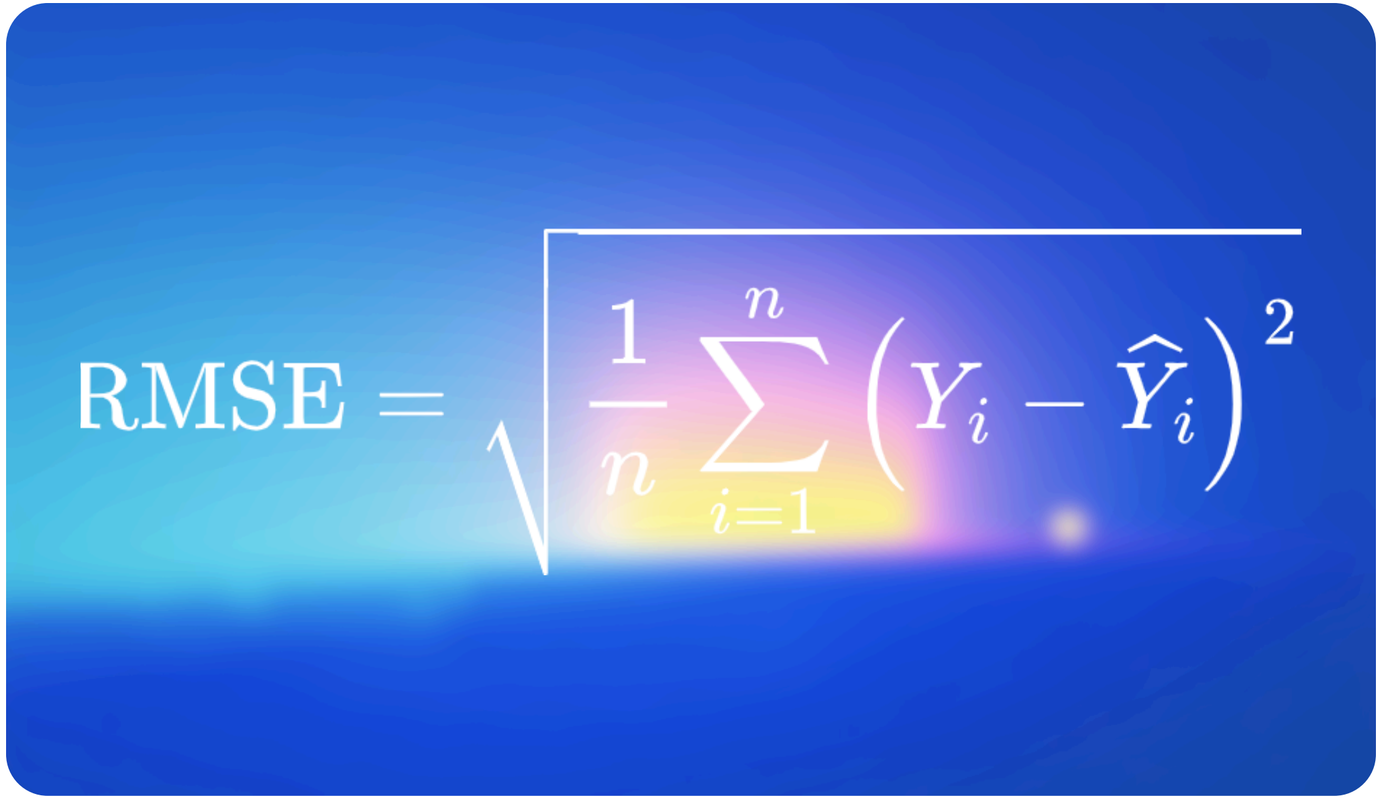
Erreur Quadratique Moyenne



Erreur Quadratique Moyenne



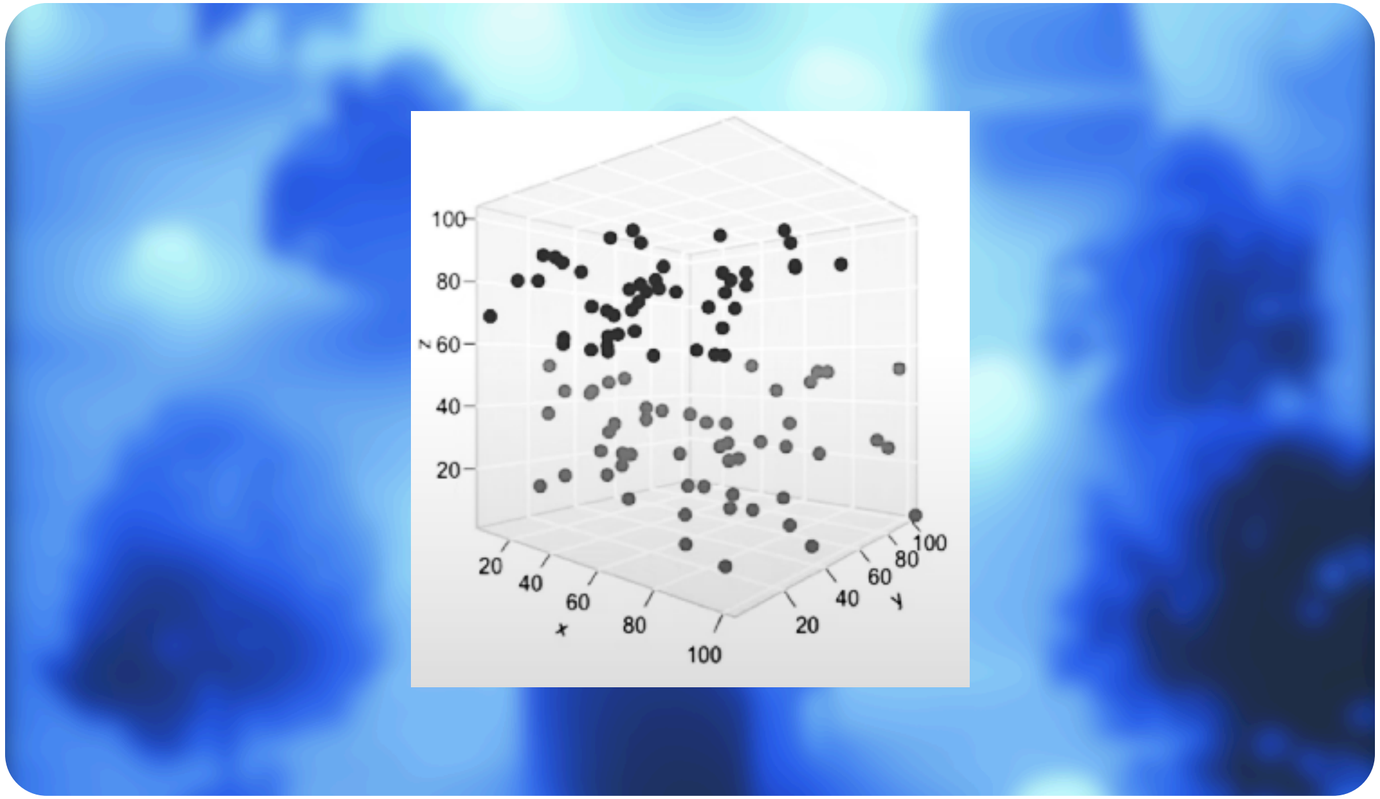
Racine de l'Erreur Quadratique Moyenne



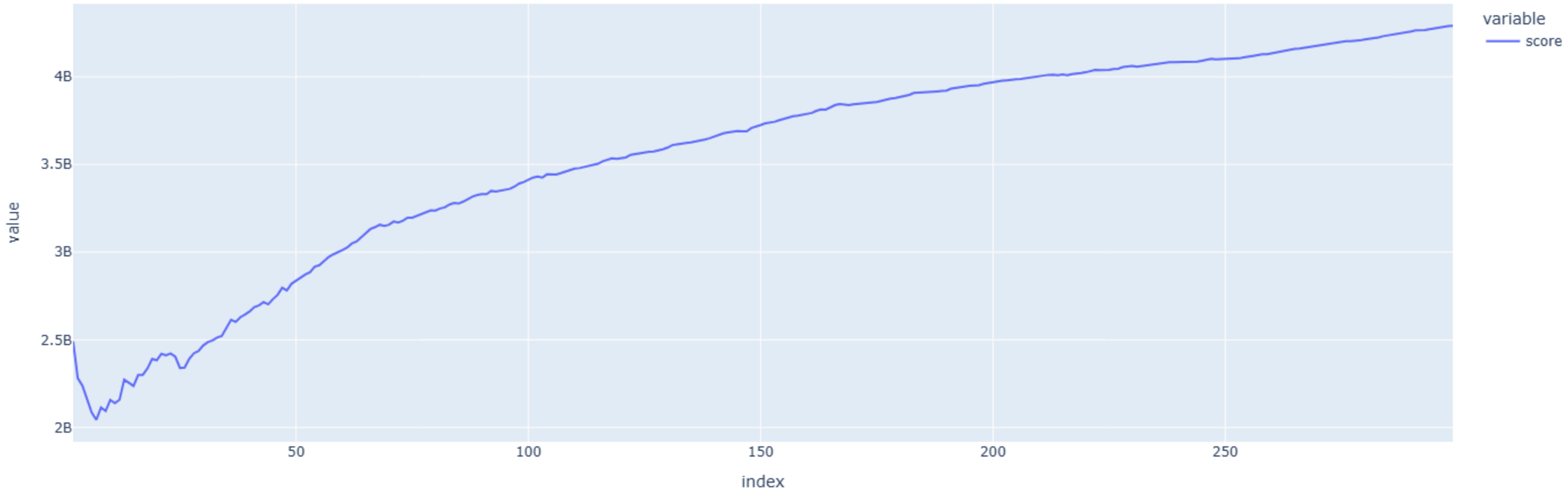
Validation croisée

kNN

kNN

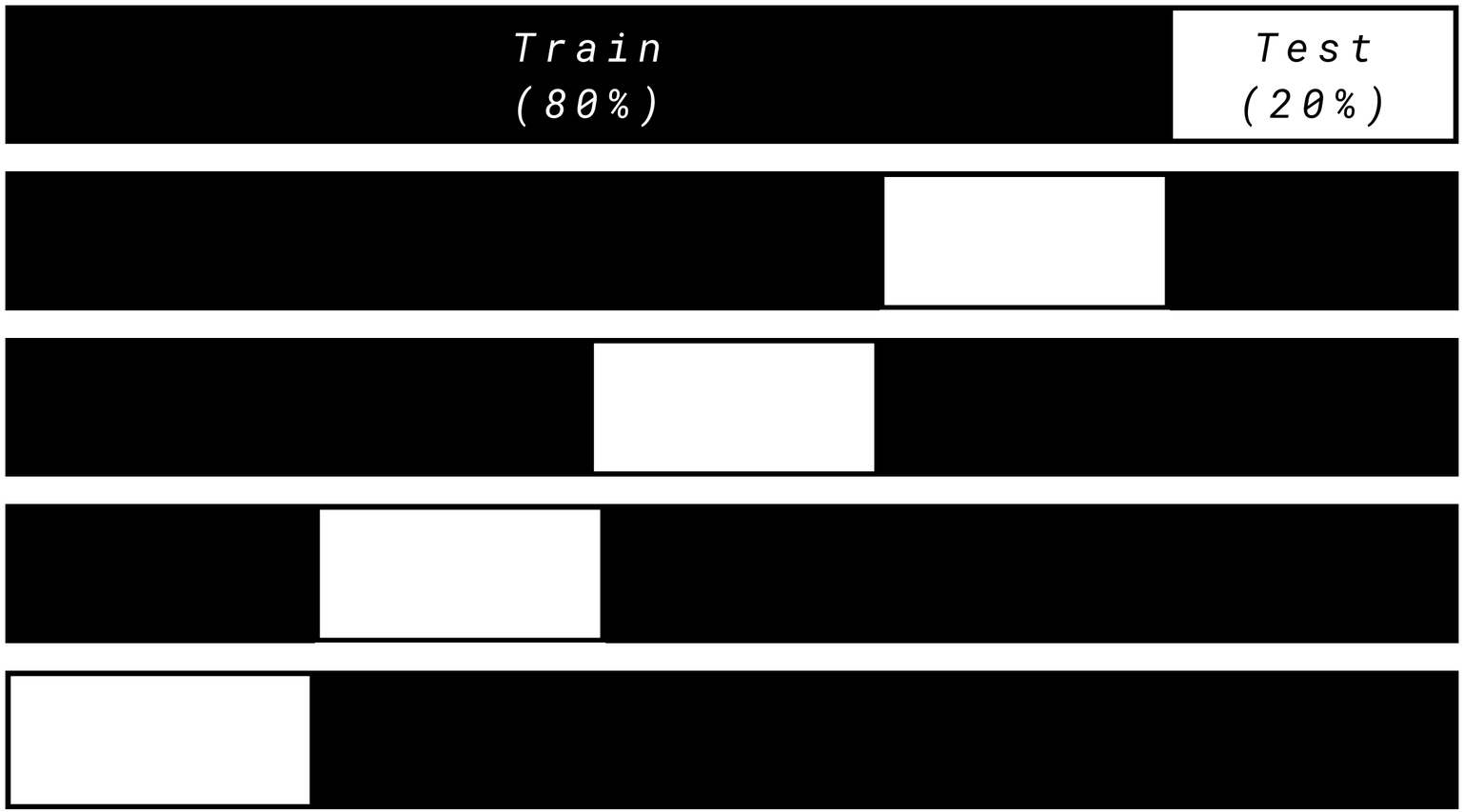


kNN > Hyper-paramètres

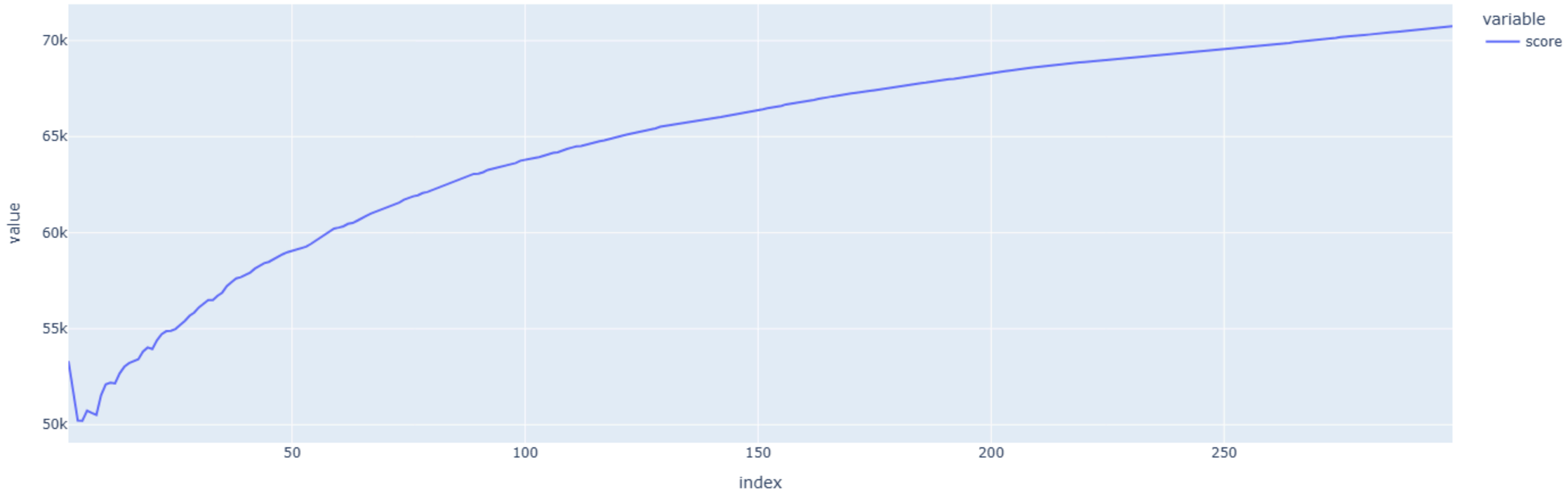


Validation croisée > KFold

Validation croisée > KFold



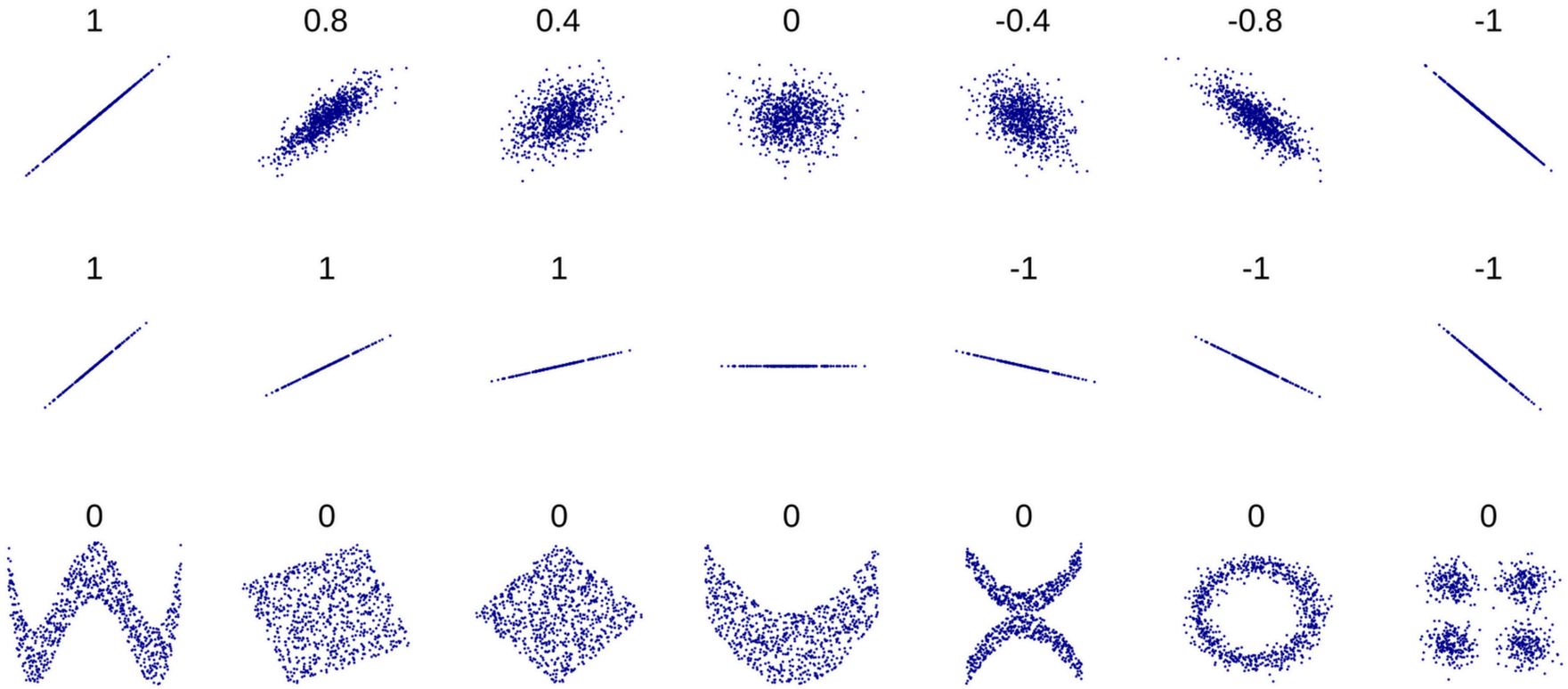
kNN > Hyper-paramètres sous KFold



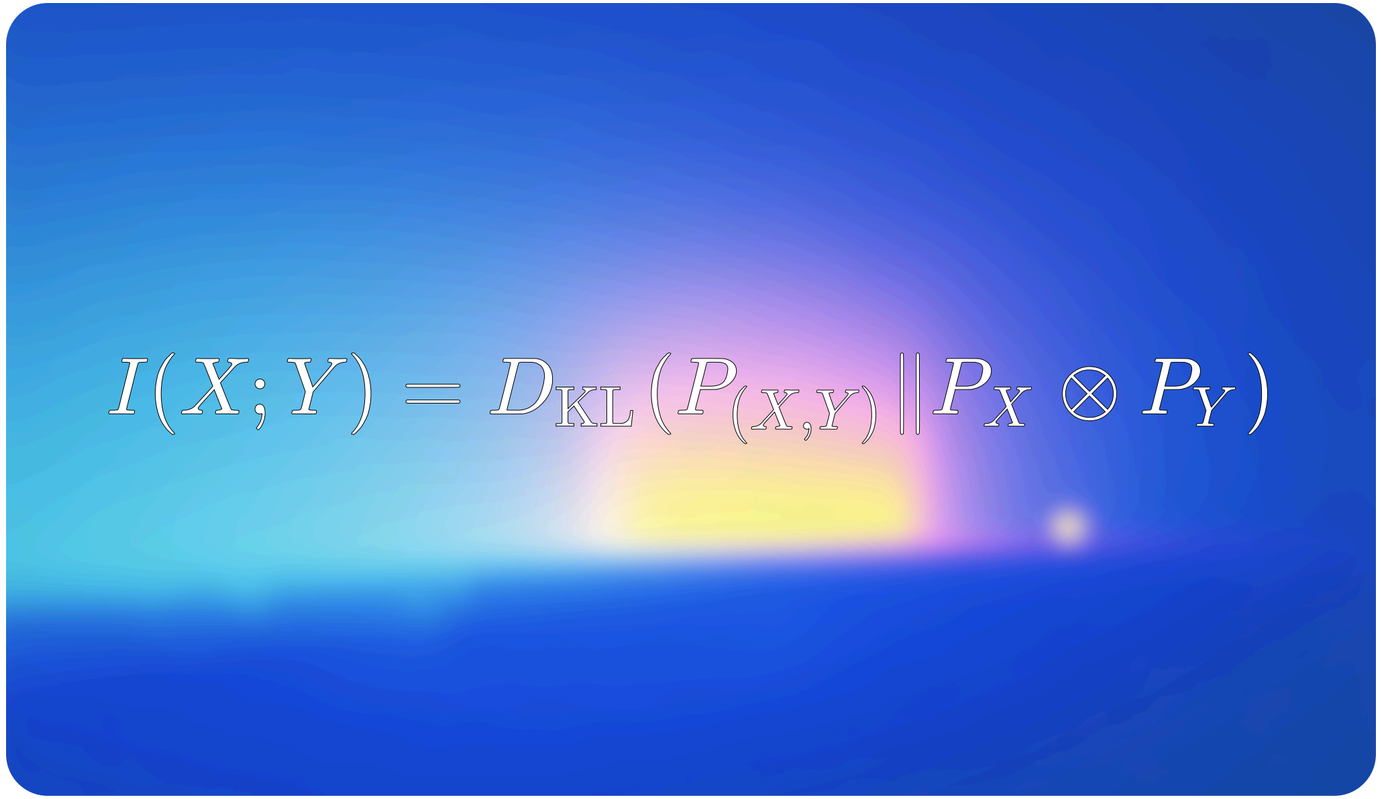
Corrélation

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Corrélation

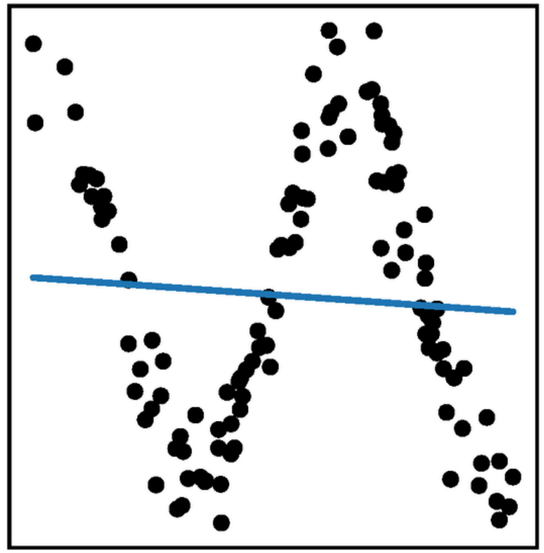


Information mutuelle



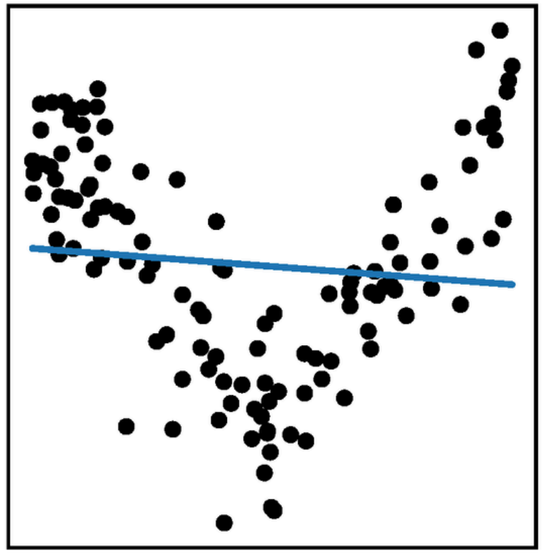
Information mutuelle

(a) $y = \sin(x) + \varepsilon$



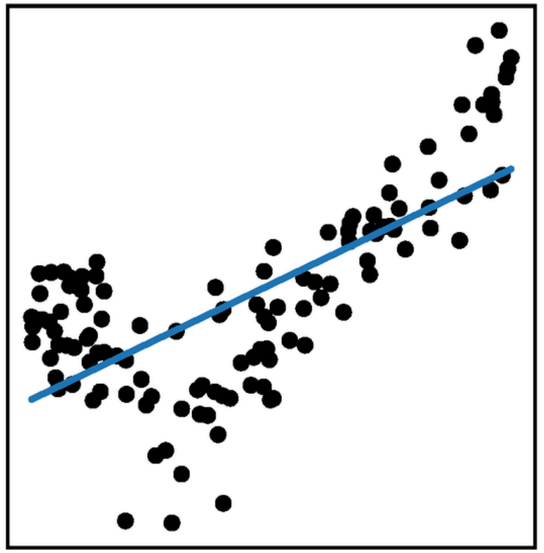
Original: −0.07
Transformed: 0.96
MI: 0.94

(b) $y = |x| + \varepsilon$



Original: −0.10
Transformed: 0.85
MI: 0.85

(c) $y = |x + 0.4| + \varepsilon$



Original: 0.70
Transformed: 0.89
MI: 0.89

[Et voilà...]

Fin de l'atelier :)

SIAHAAN--GENSOLLEN Rémy, 2025