

[Tu x a e]



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

Introduction



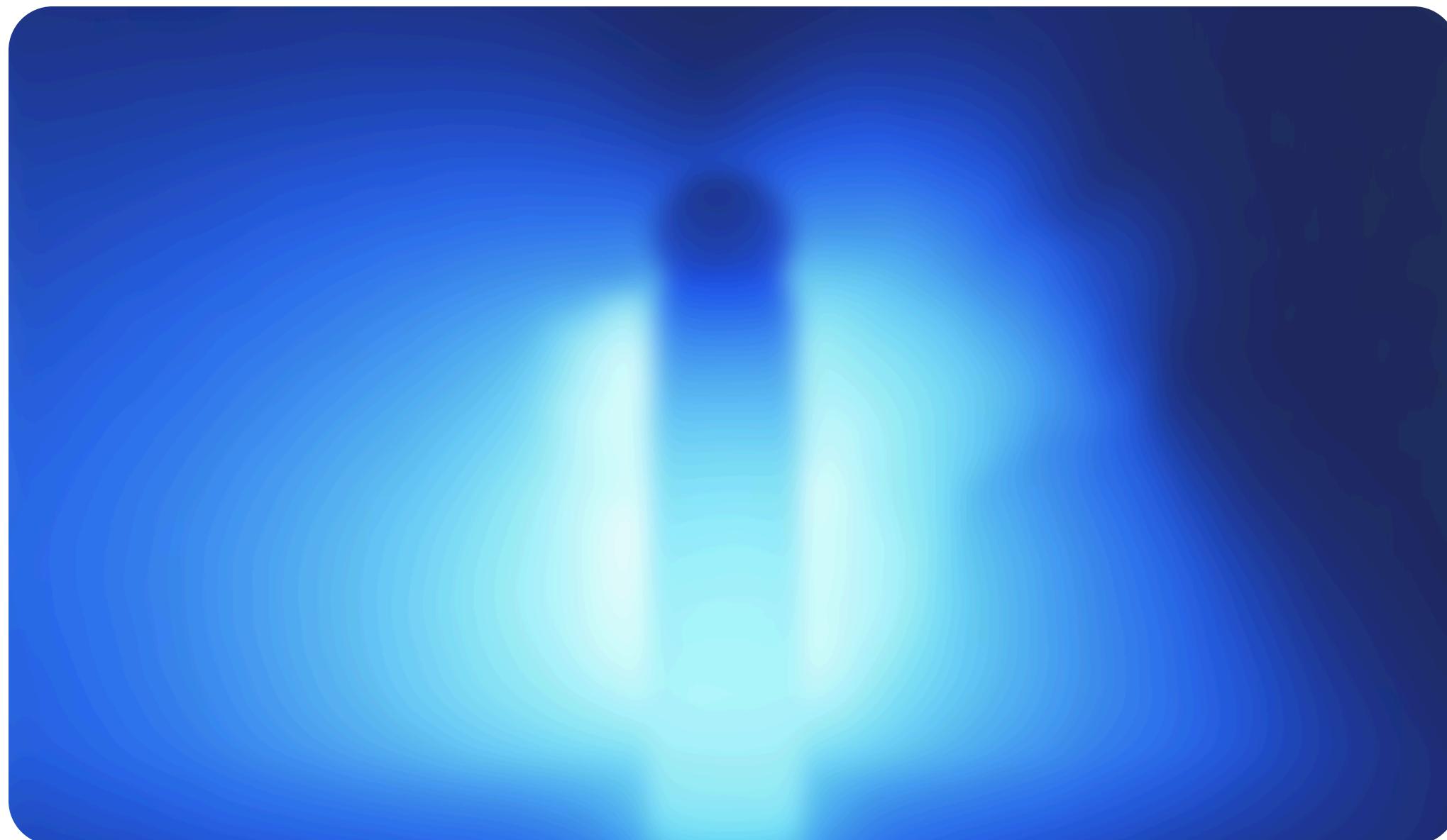
Jupyter-python

The JupyterLab IDE with Python, Julia, and a collection of standard data science packages.

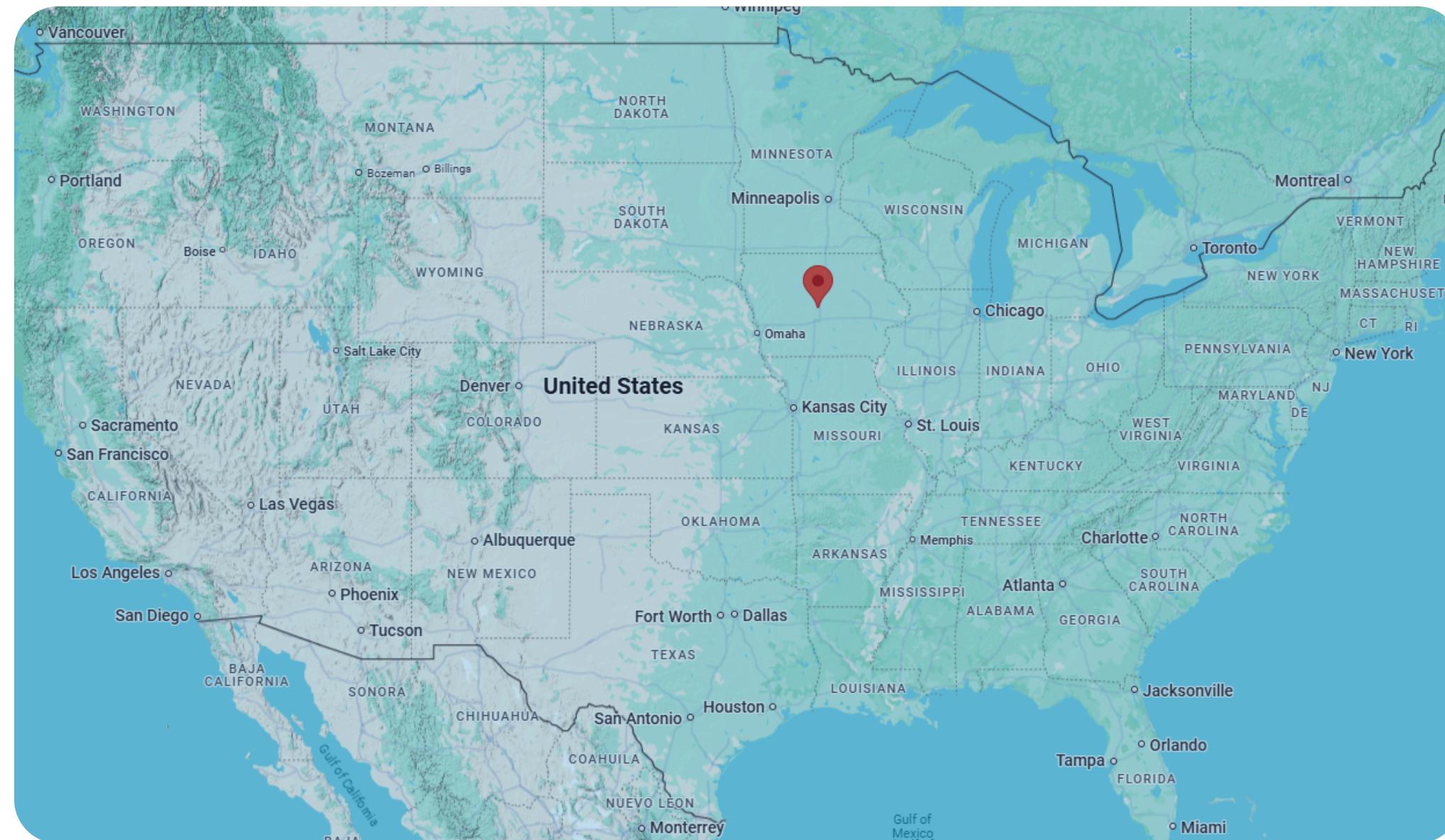
[Learn more](#) [Launch](#)

<https://shorturl.at/LbZDW>

Le problème



Le problème



Ames, Iowa

Le problème



Ames, Iowa

Le problème



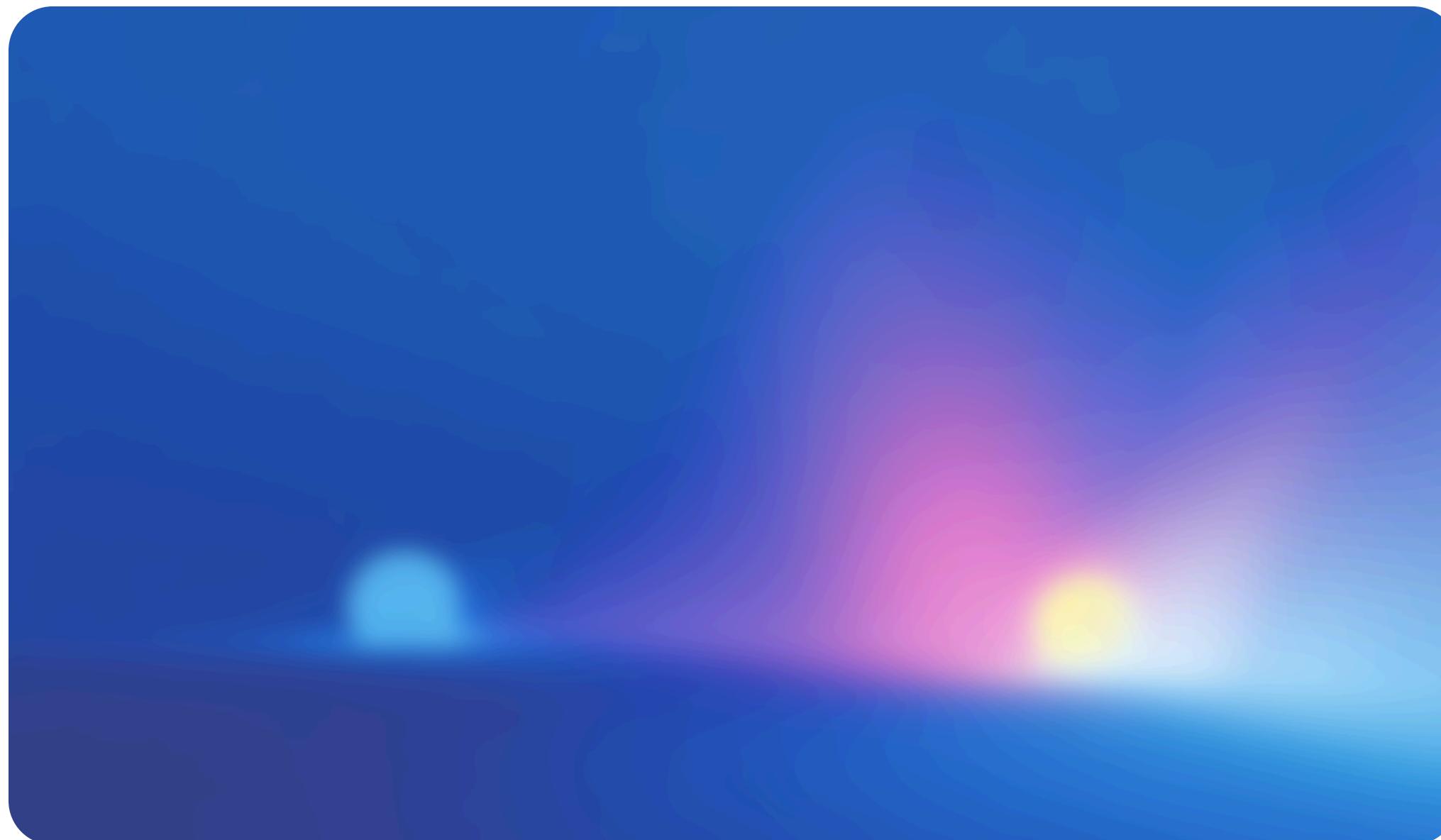
(la dernière fois)
Classification

Le problème



Régression

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>)

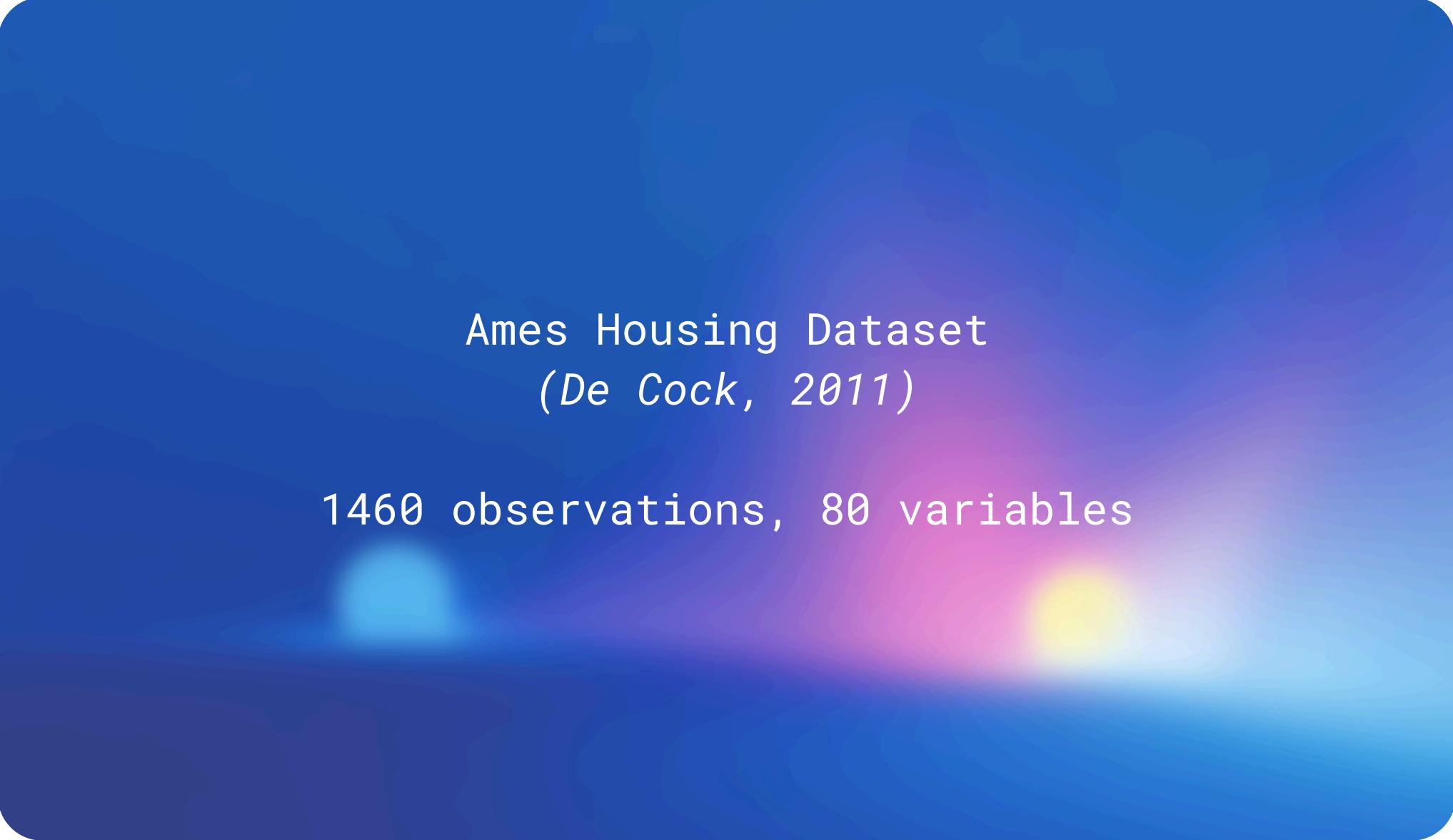
Le dataset

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

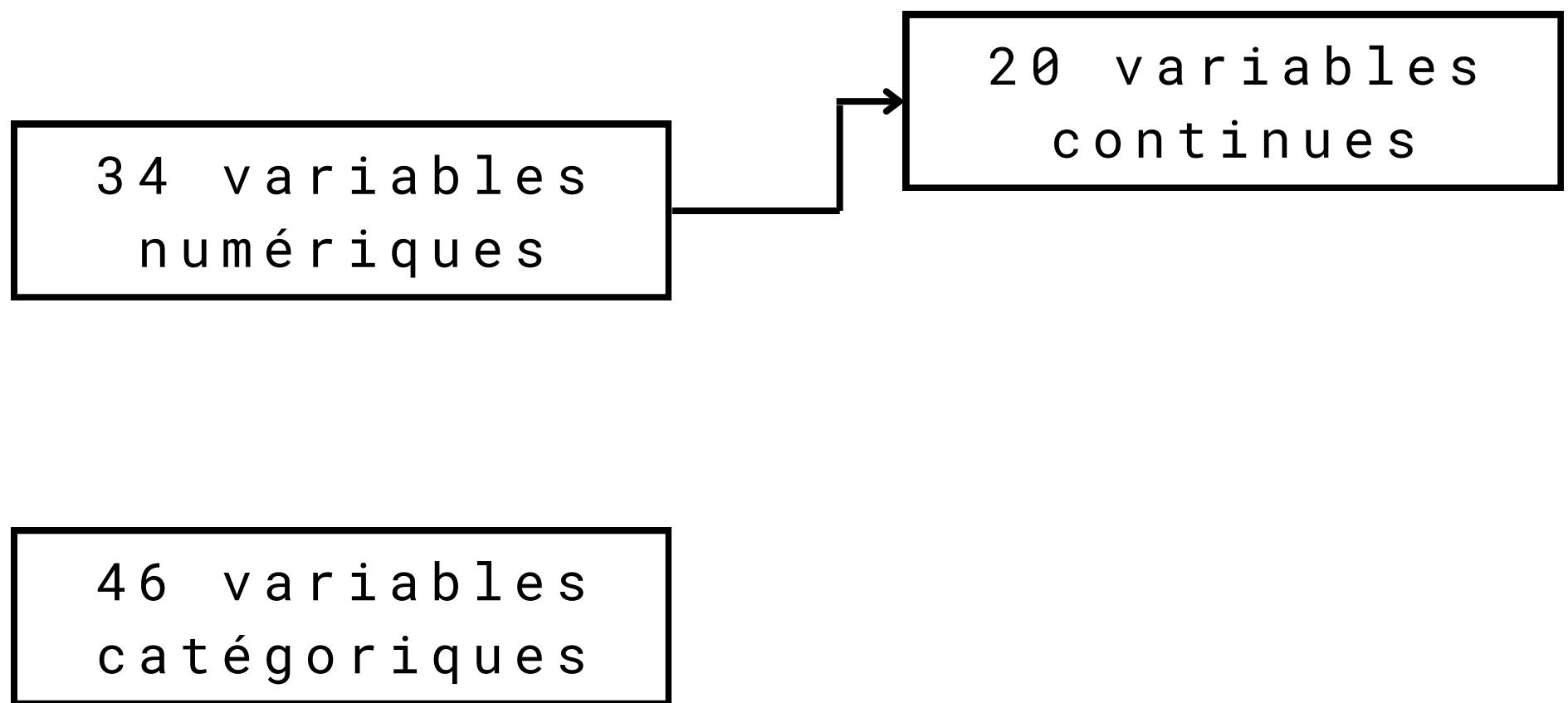
34 variables
numériques

Le dataset

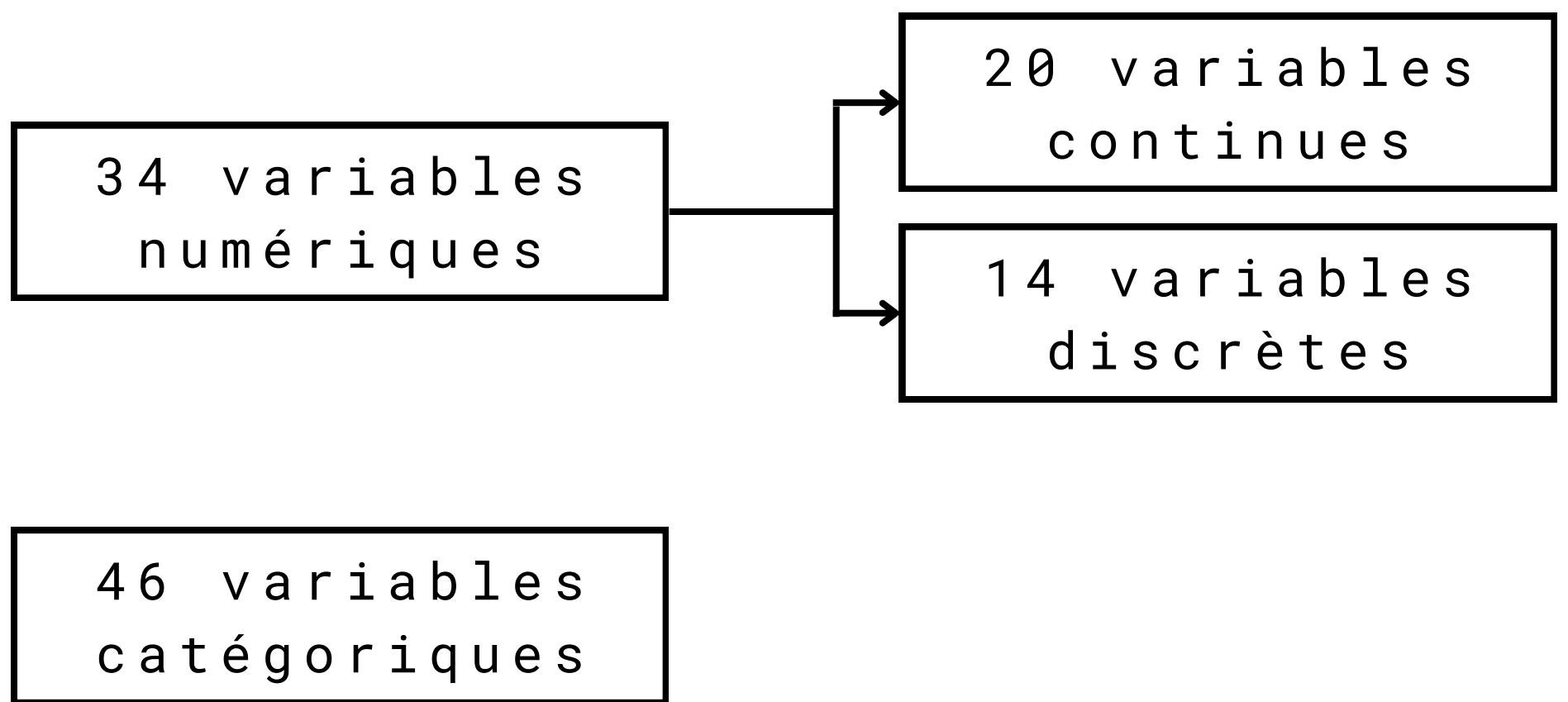
34 variables
numériques

46 variables
catégoriques

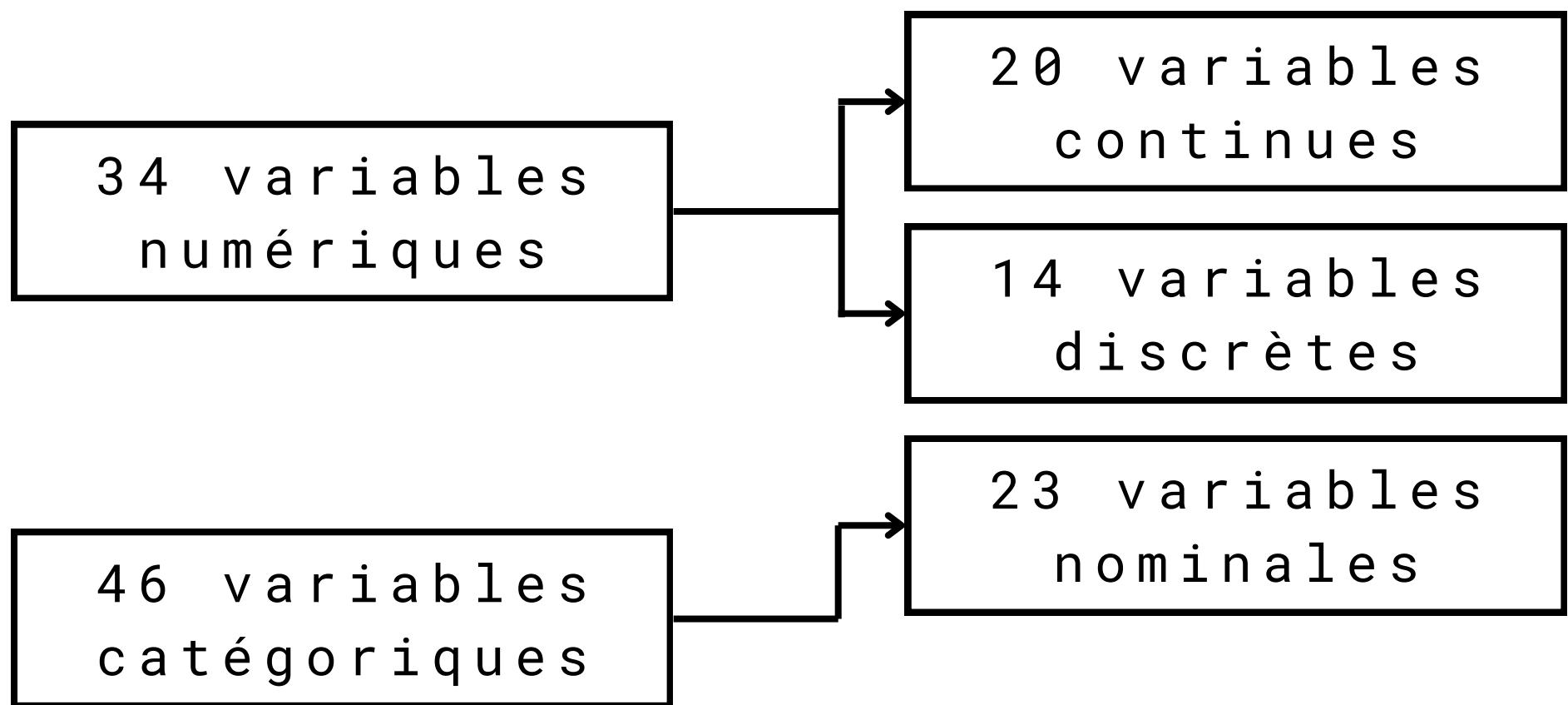
Le dataset



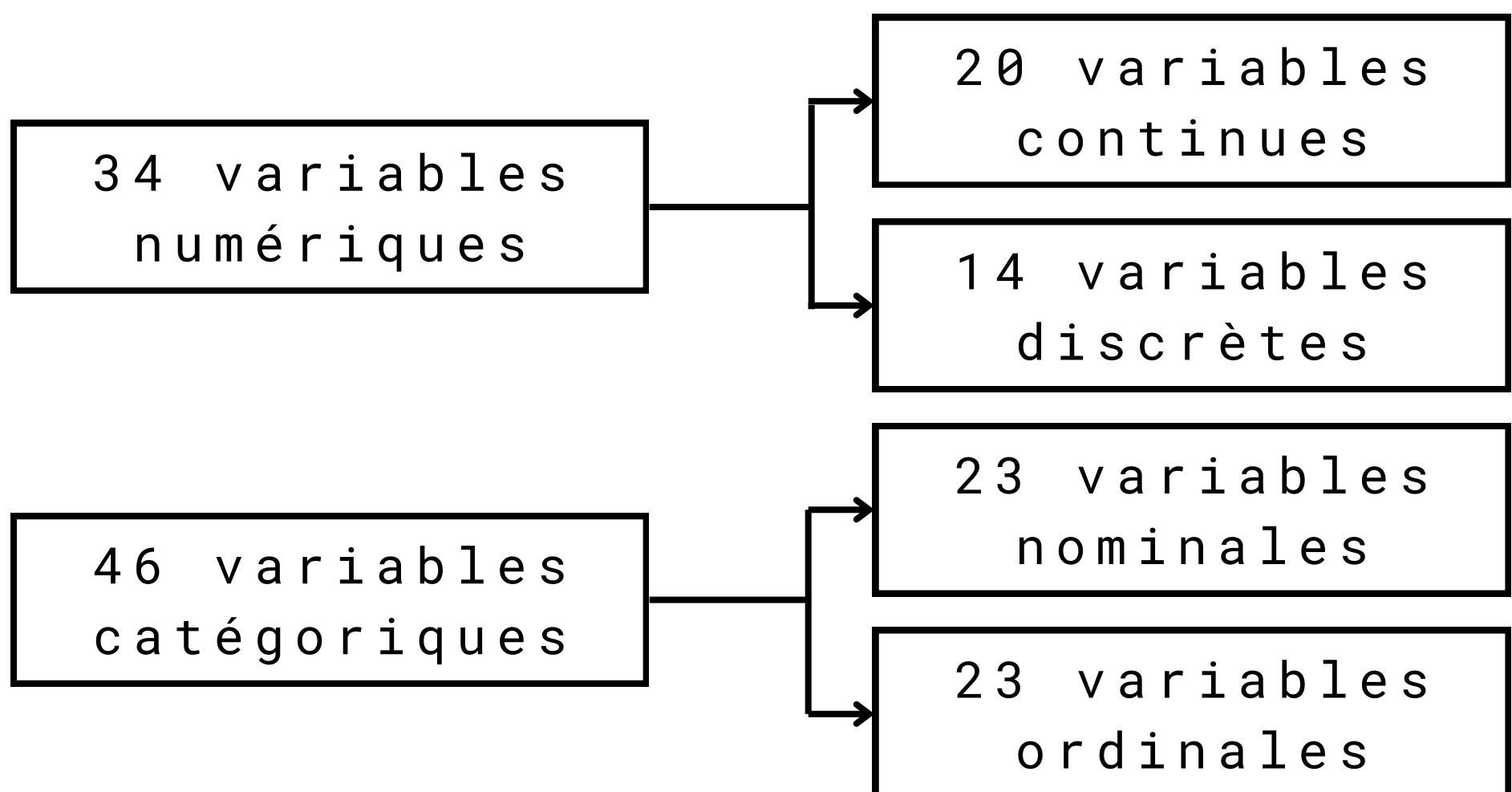
Le dataset



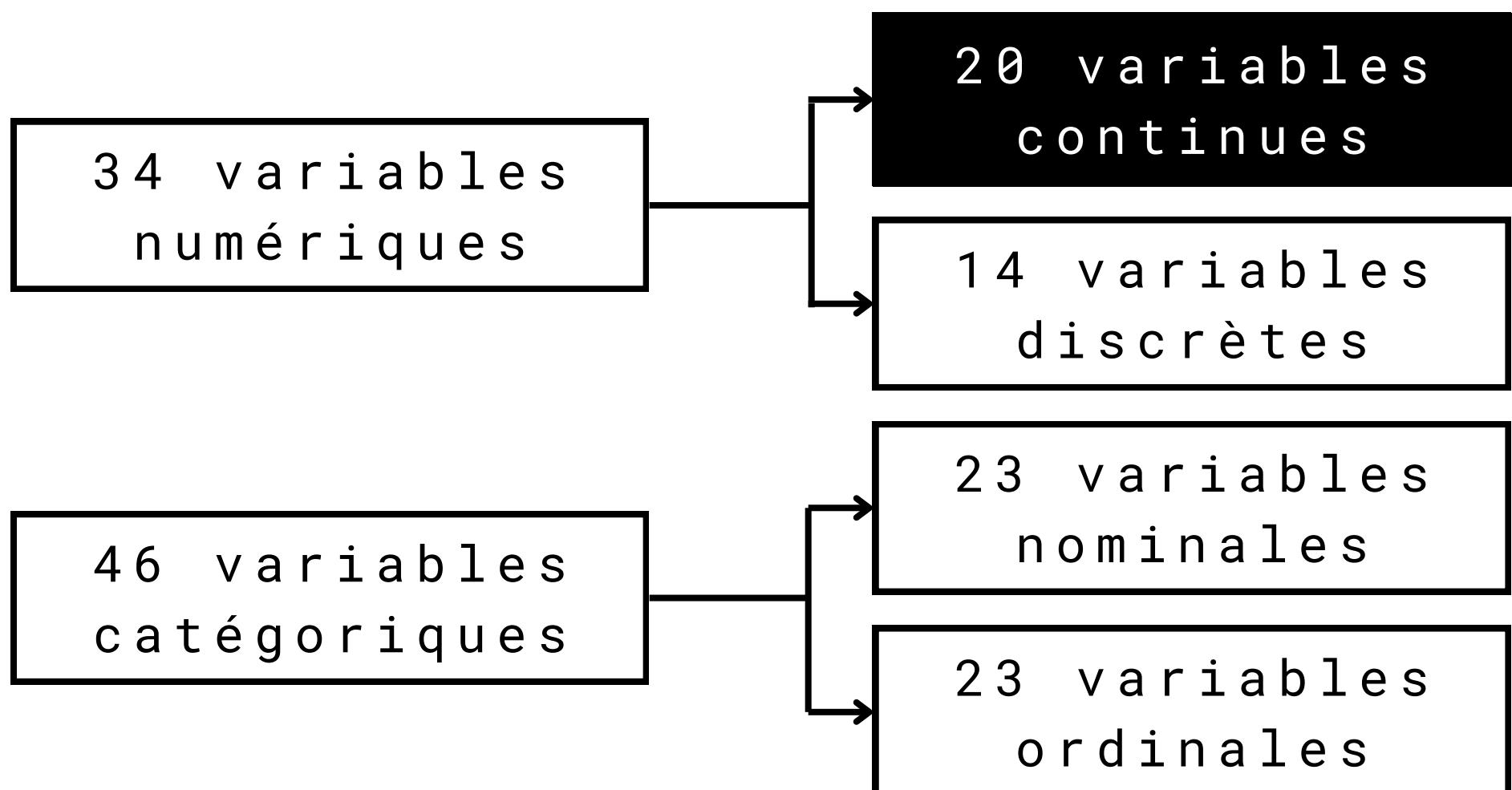
Le dataset



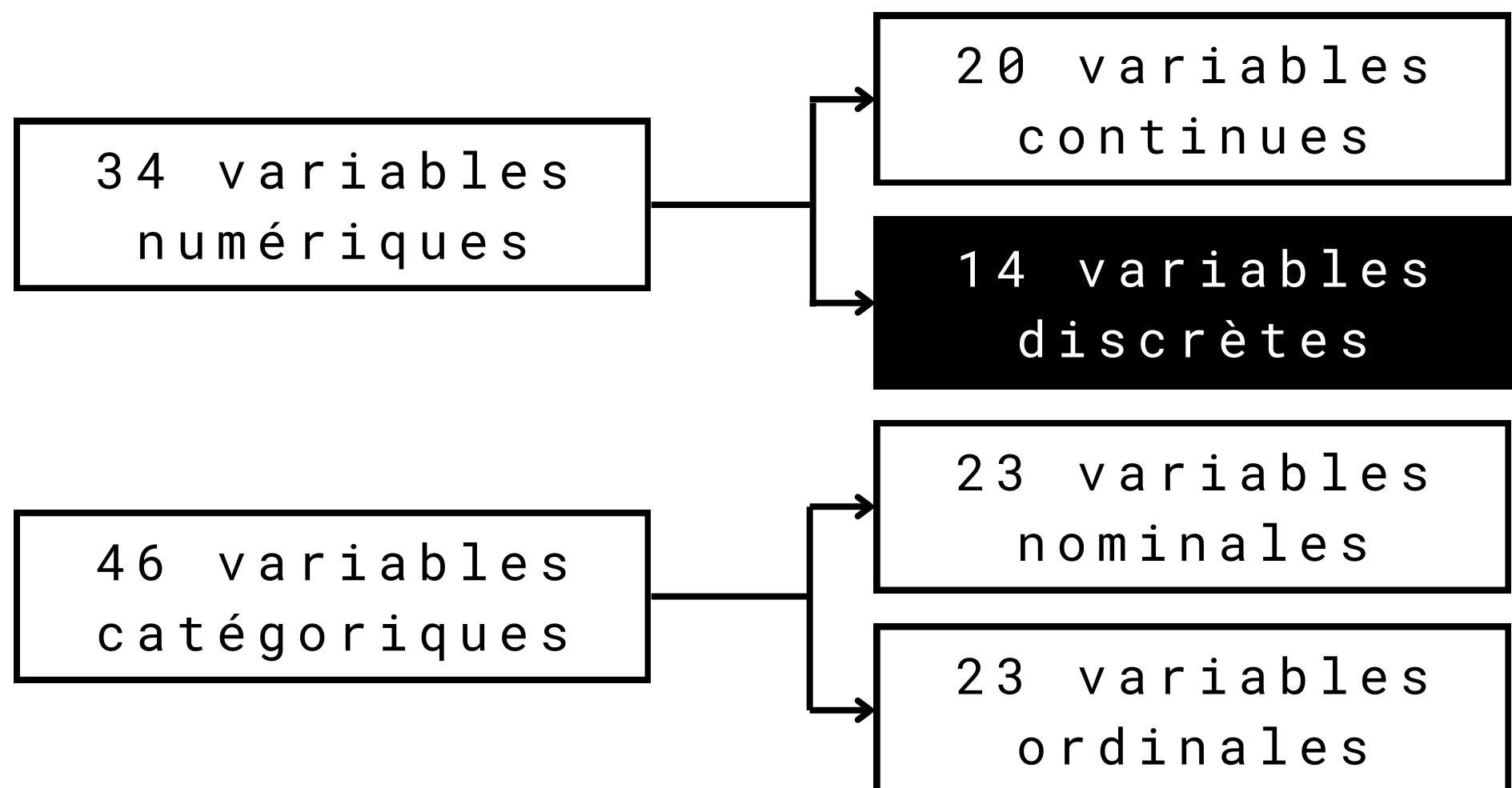
Le dataset



Le dataset



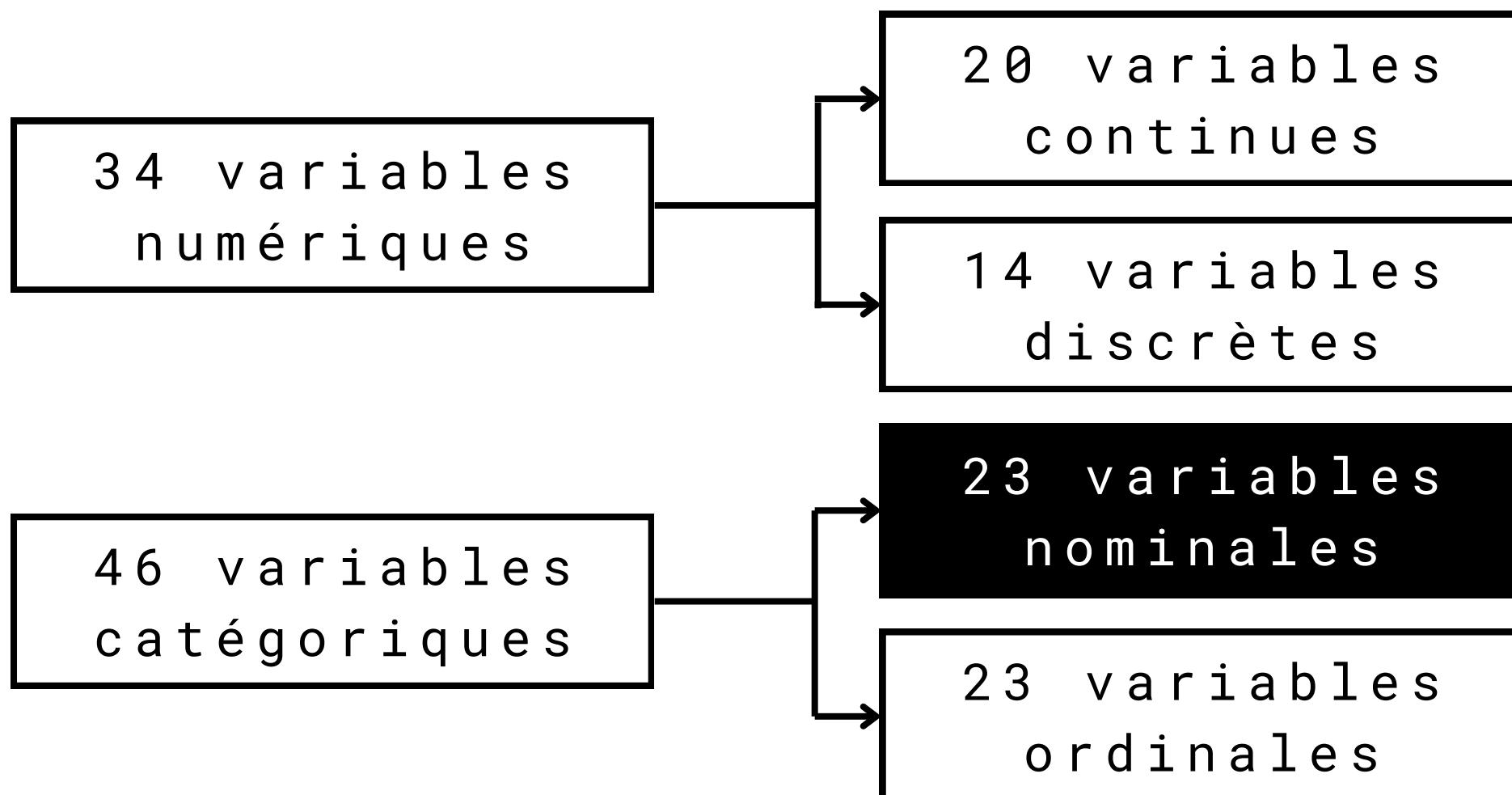
Le dataset



GarageCars

Capacité du garage en nombre de voitures.

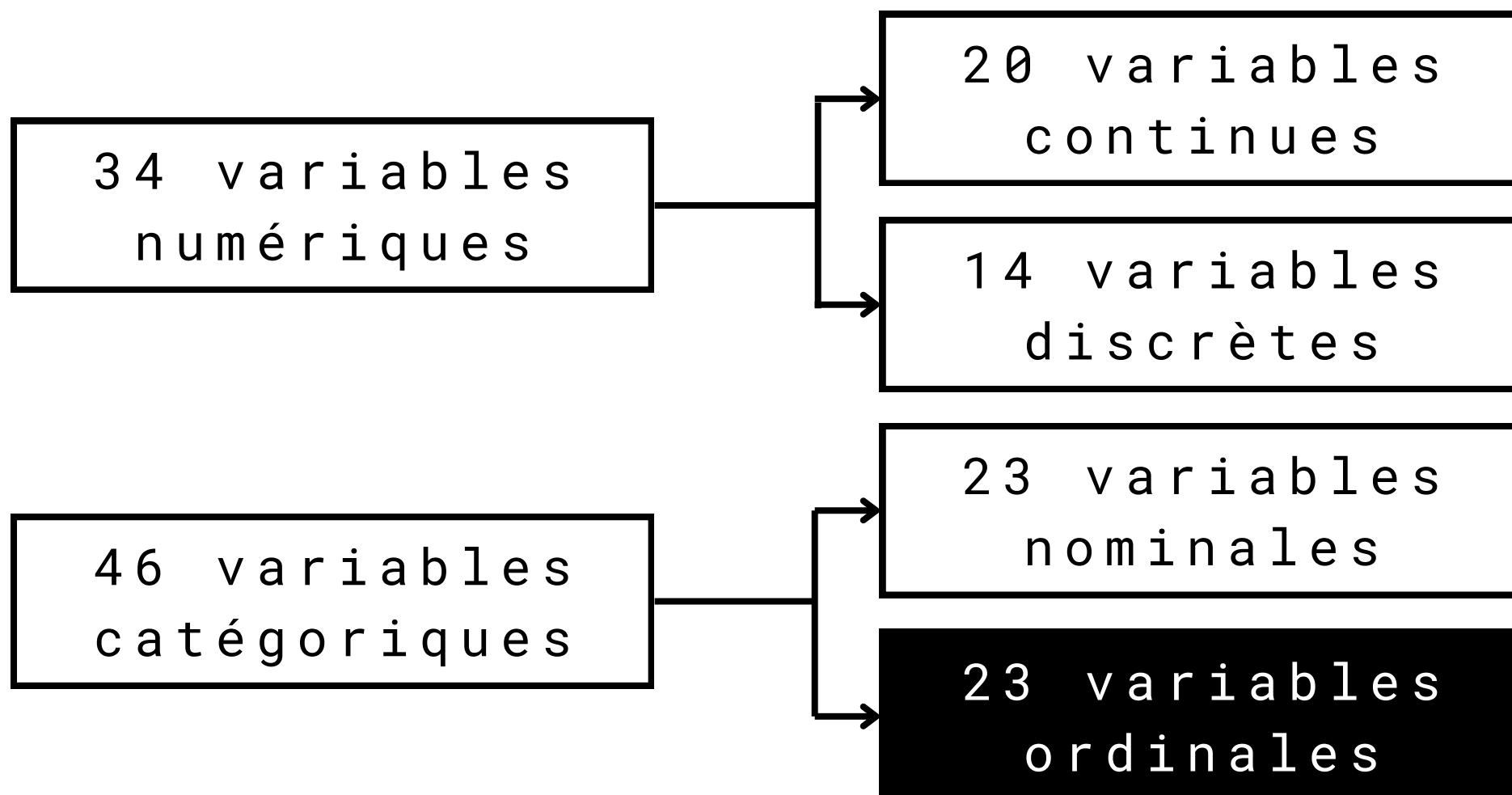
Le dataset



RoofMat1

- *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)

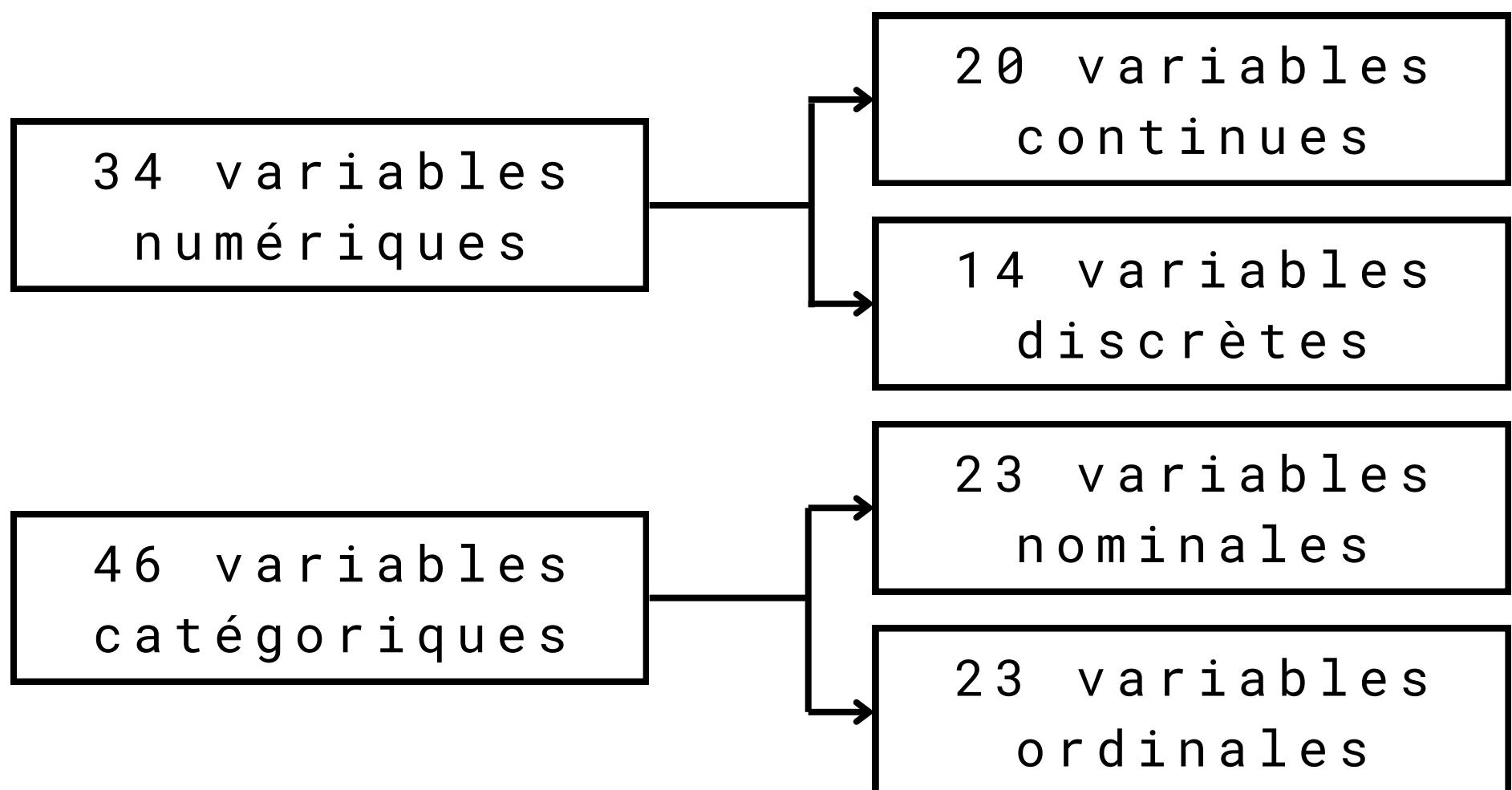
Le dataset



GarageQual

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

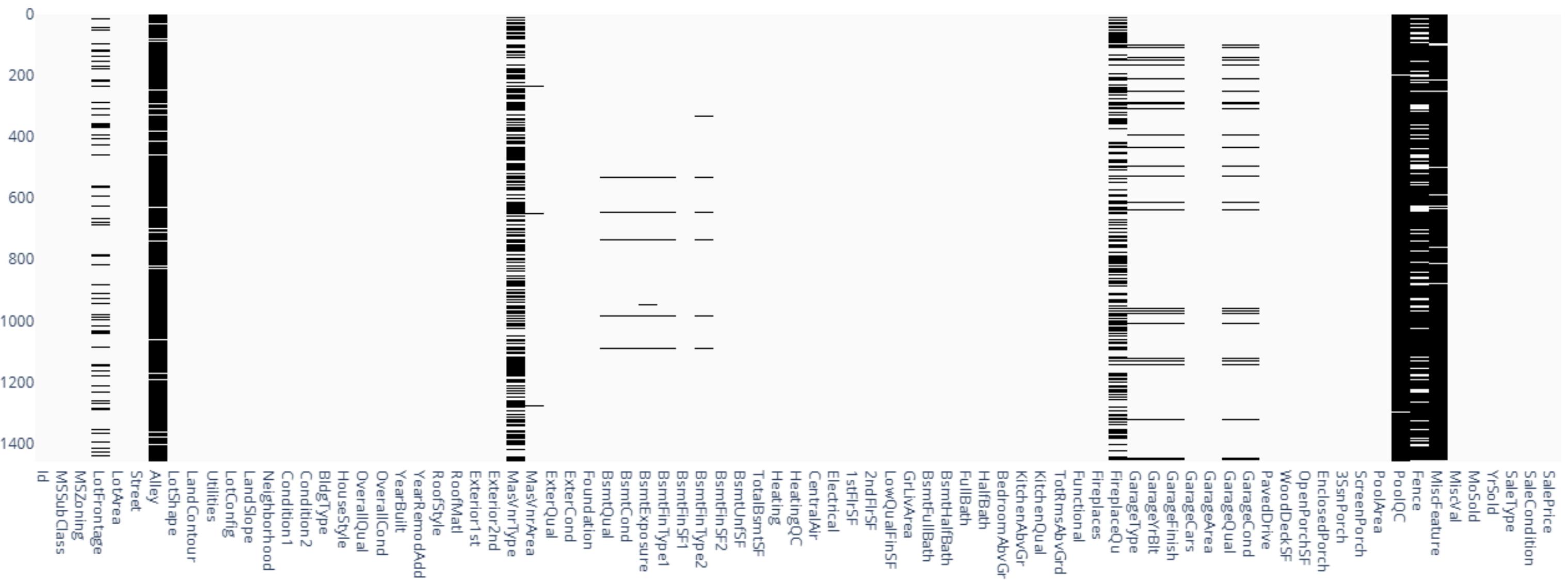
Le dataset



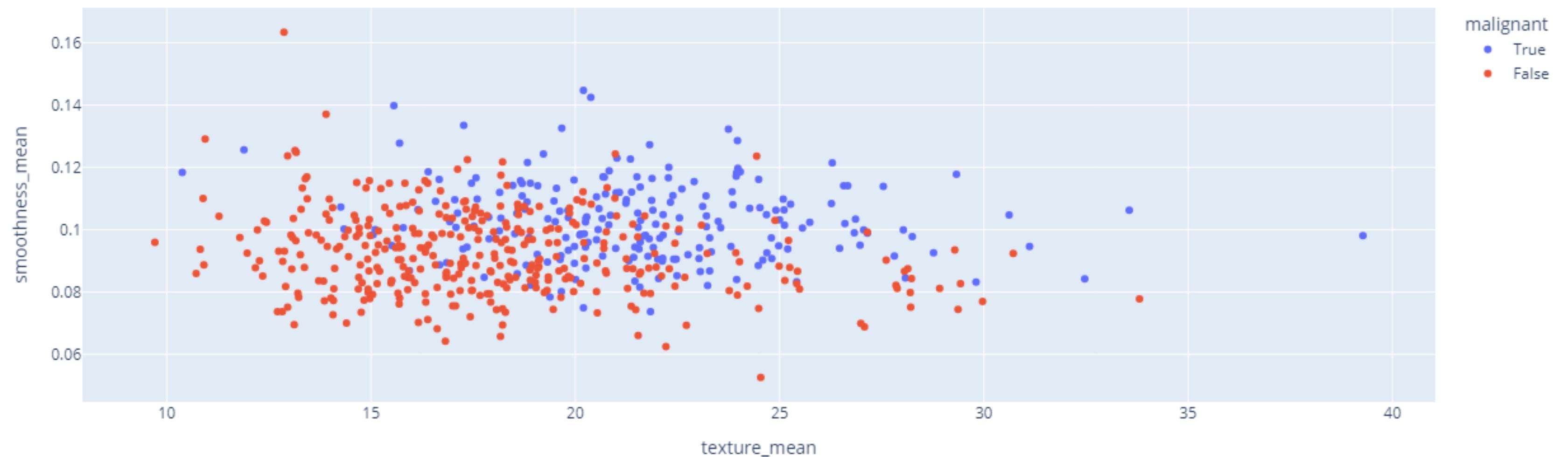
> > > > > > >

Au travail !

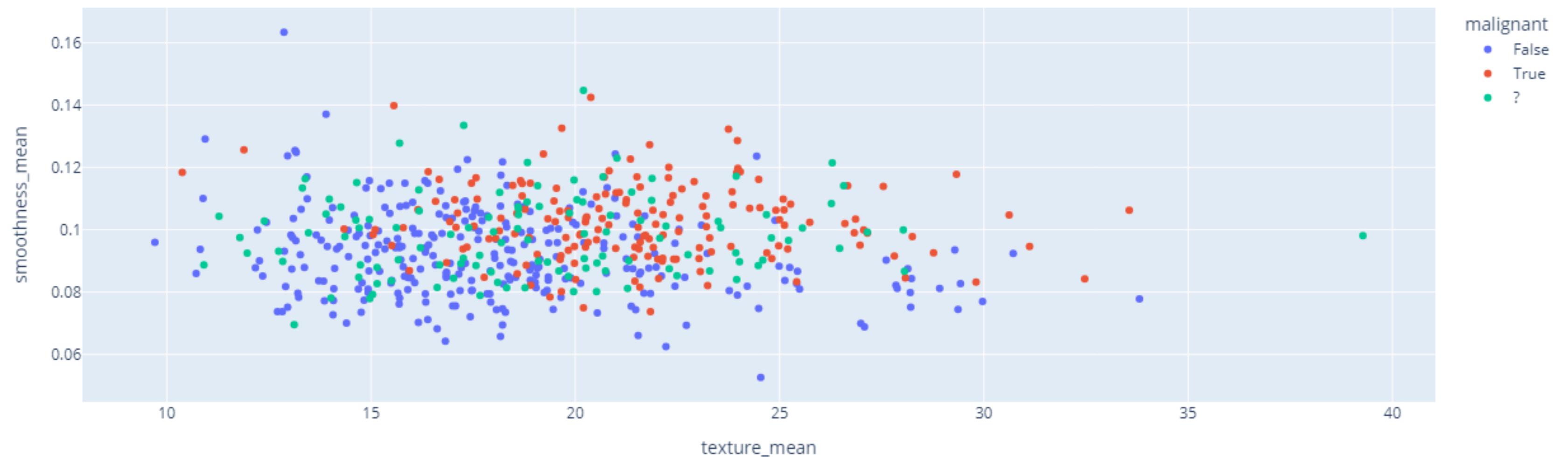
Analyse des données



Rappel du dernier atelier



Rappel du dernier atelier

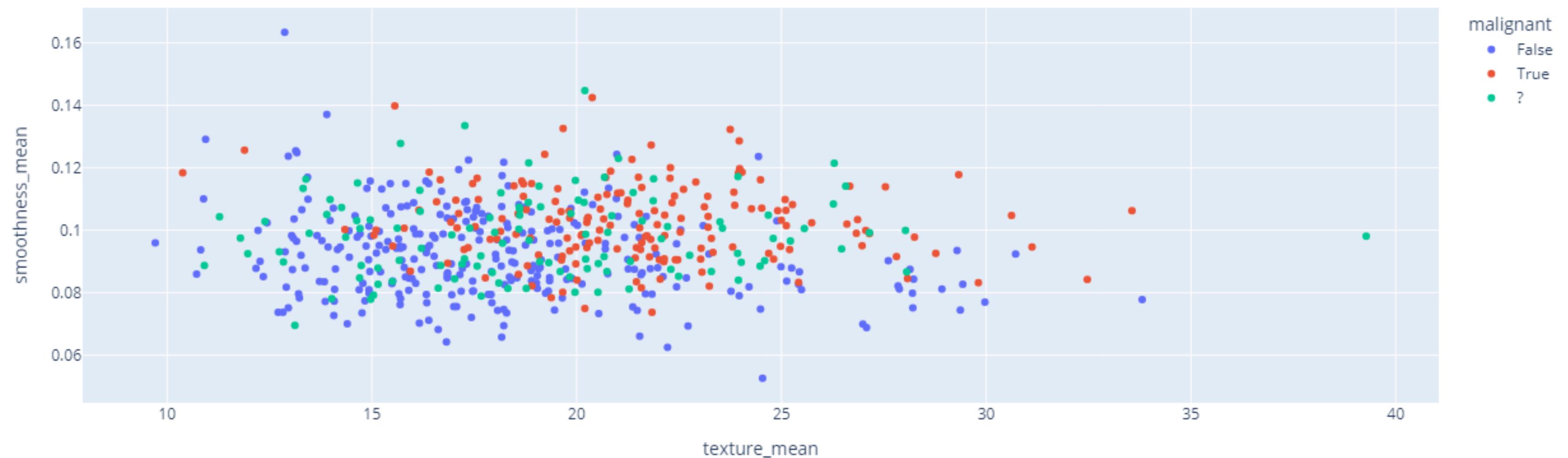


Information mutuelle

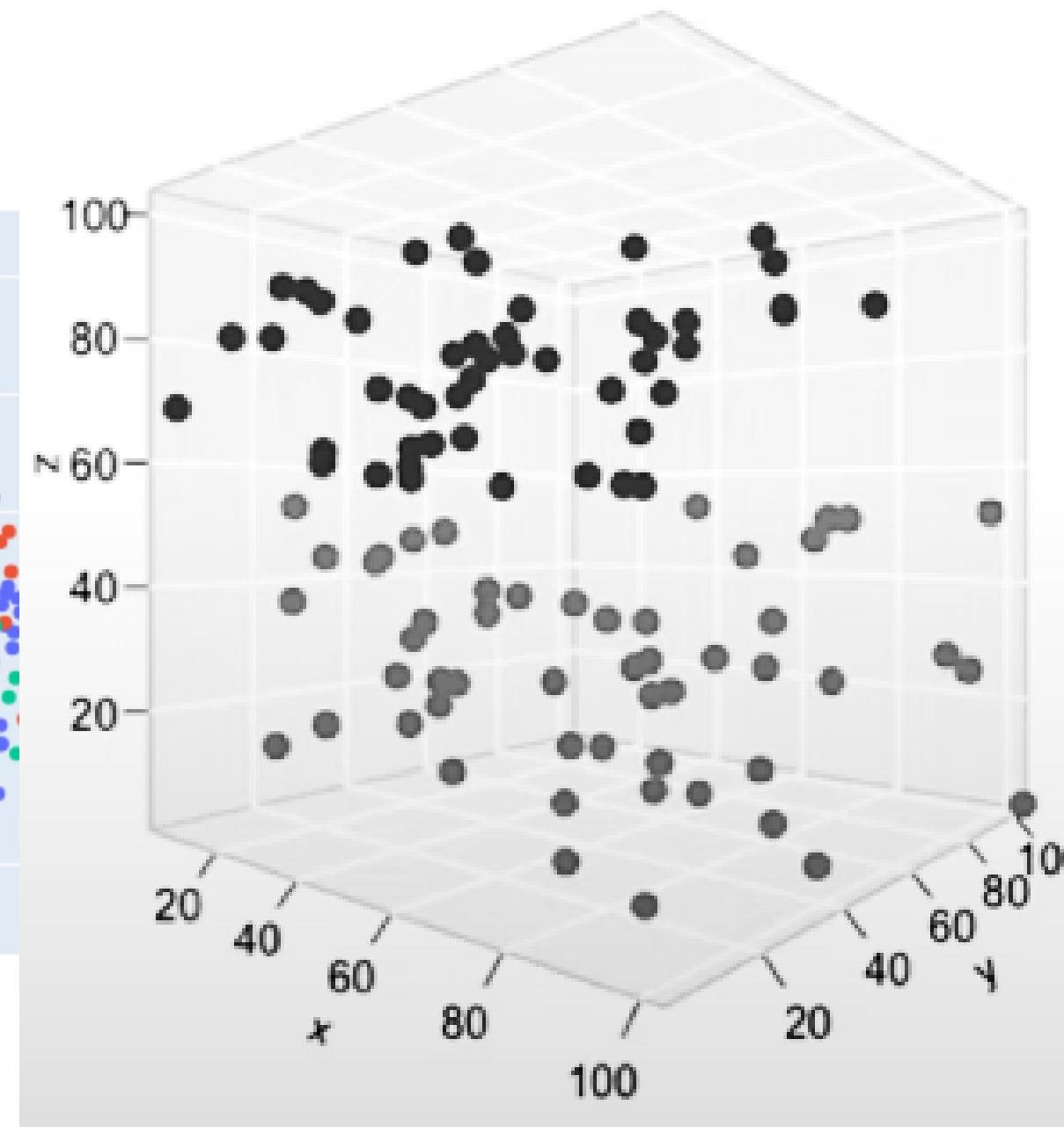
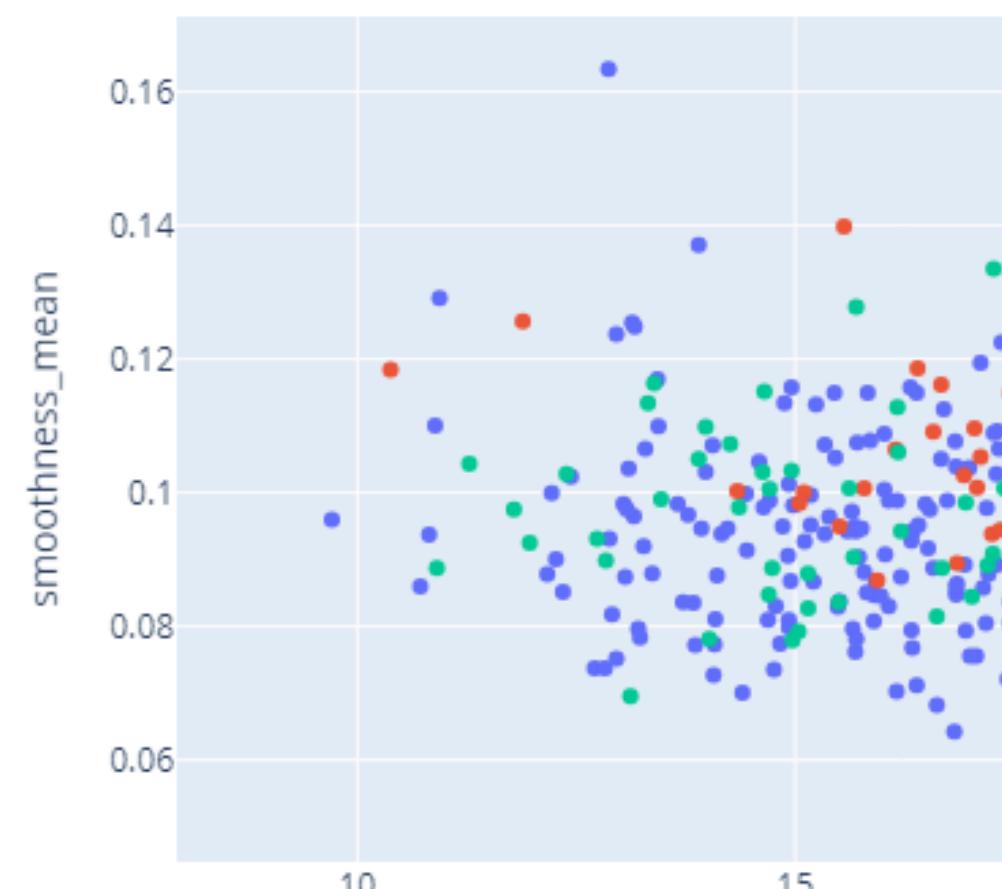
$$I(X;Y) = D_{\text{KL}}(P_{(X,Y)} \| P_X \otimes P_Y)$$

kNN

kNN

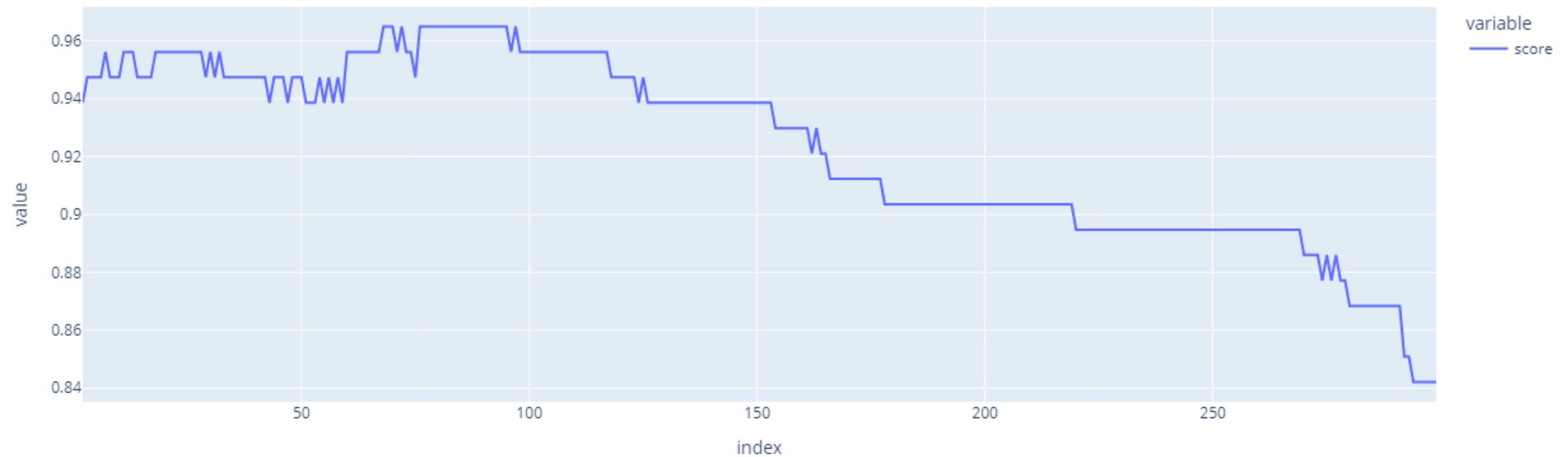


kNN

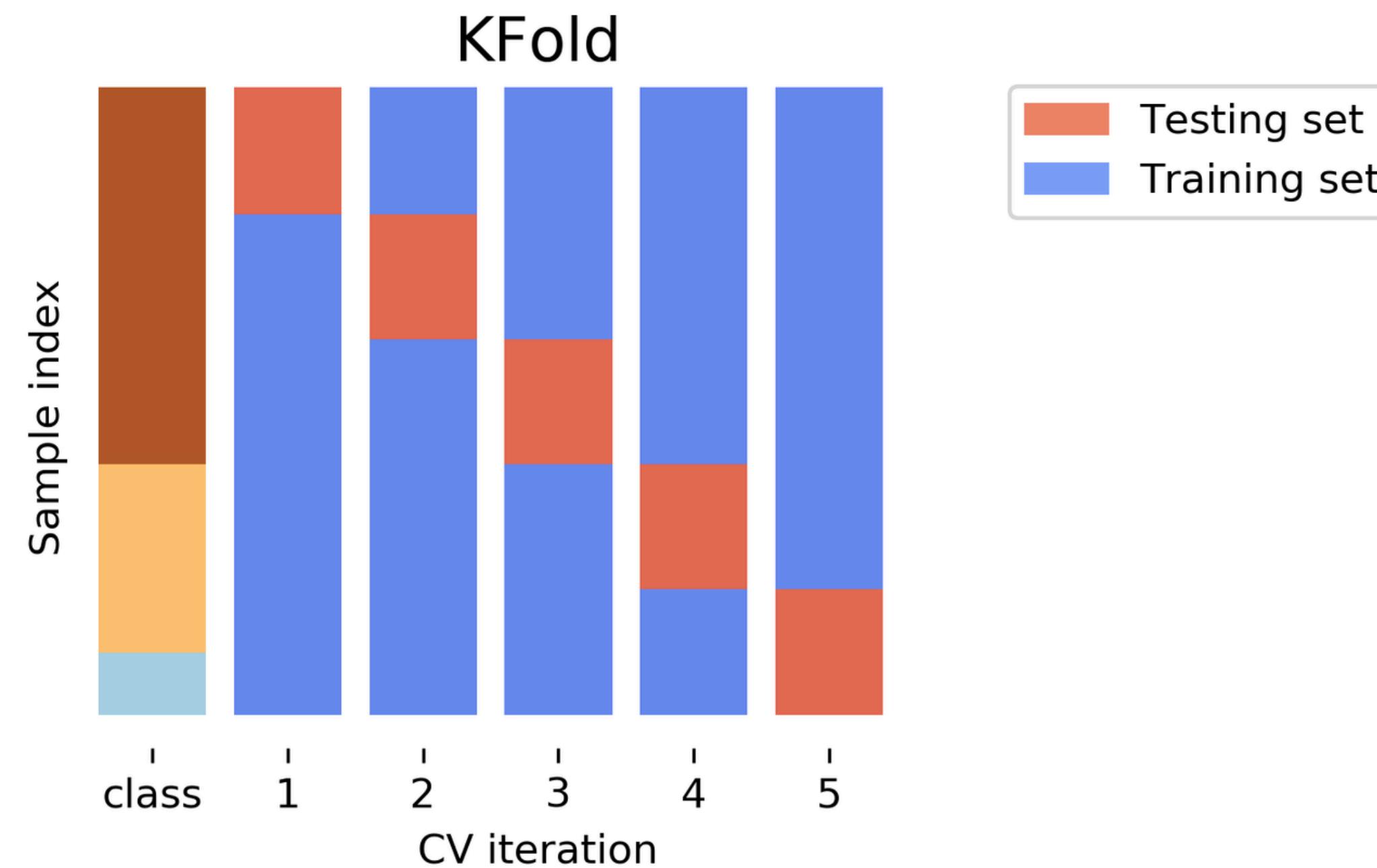


malignant
• False
• True
• ?

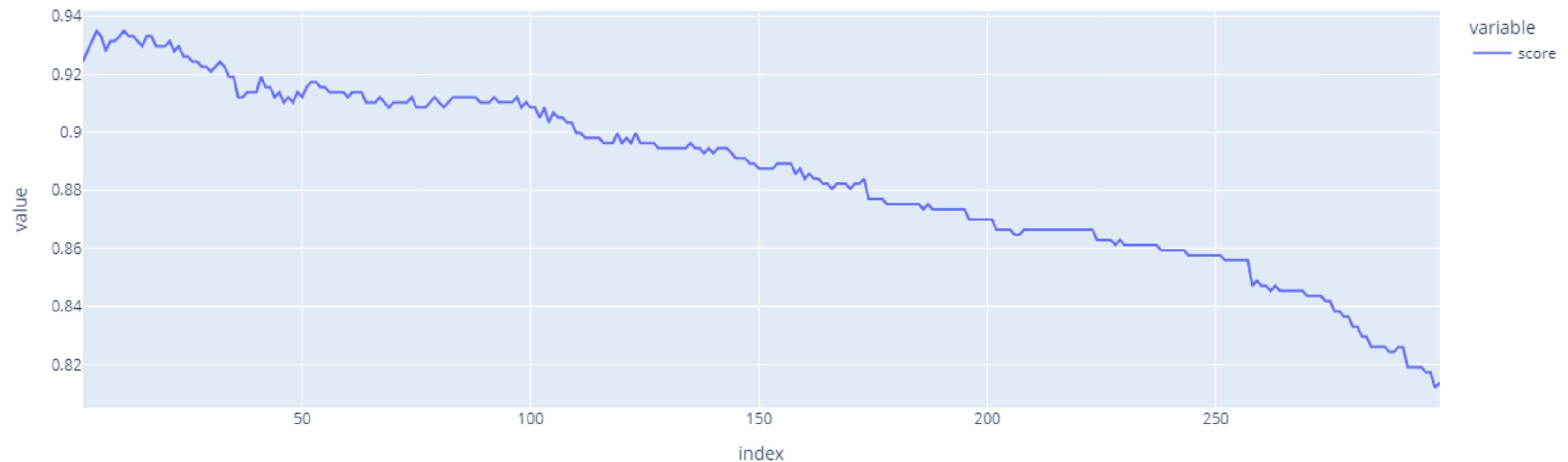
k NN > Hyper Paramètres



Validation croisée



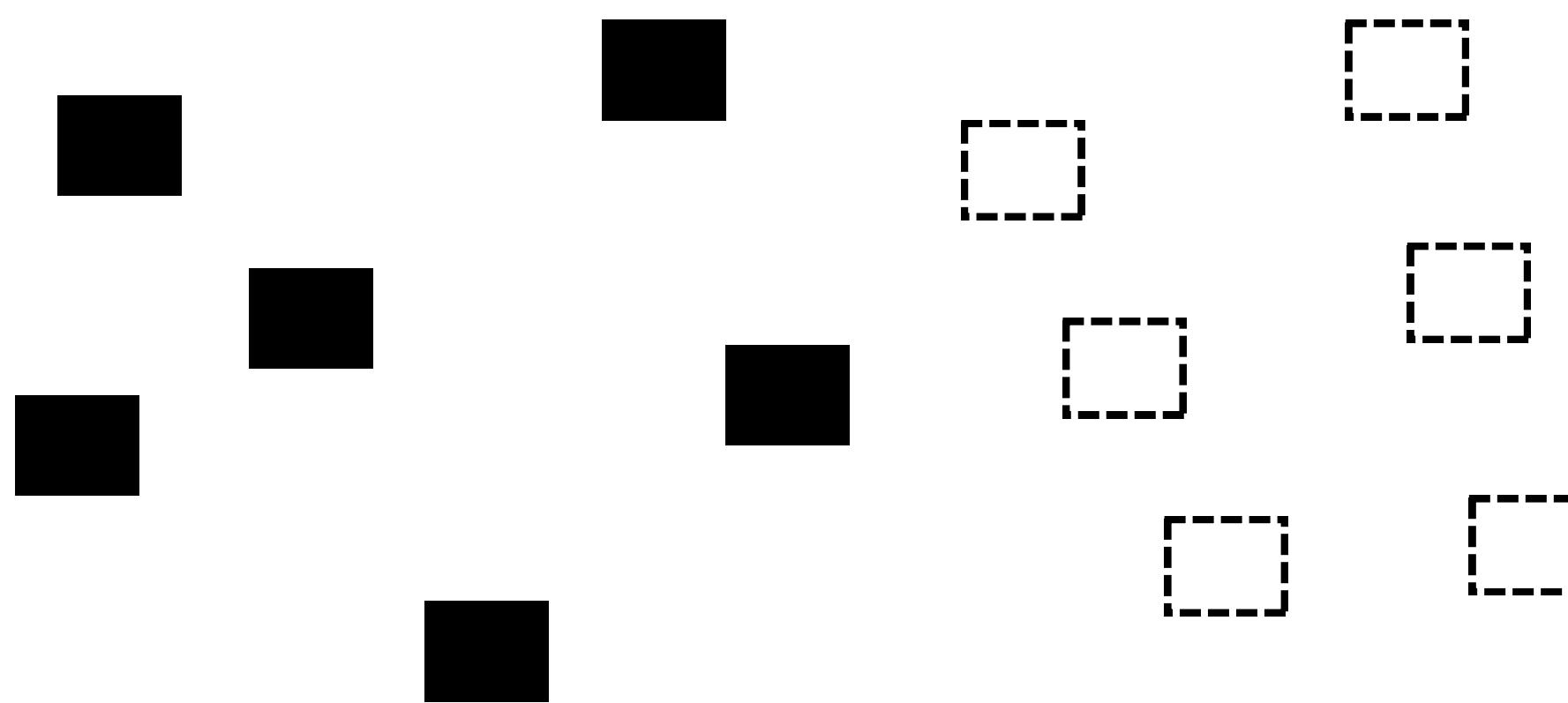
kNN > Validation croisée



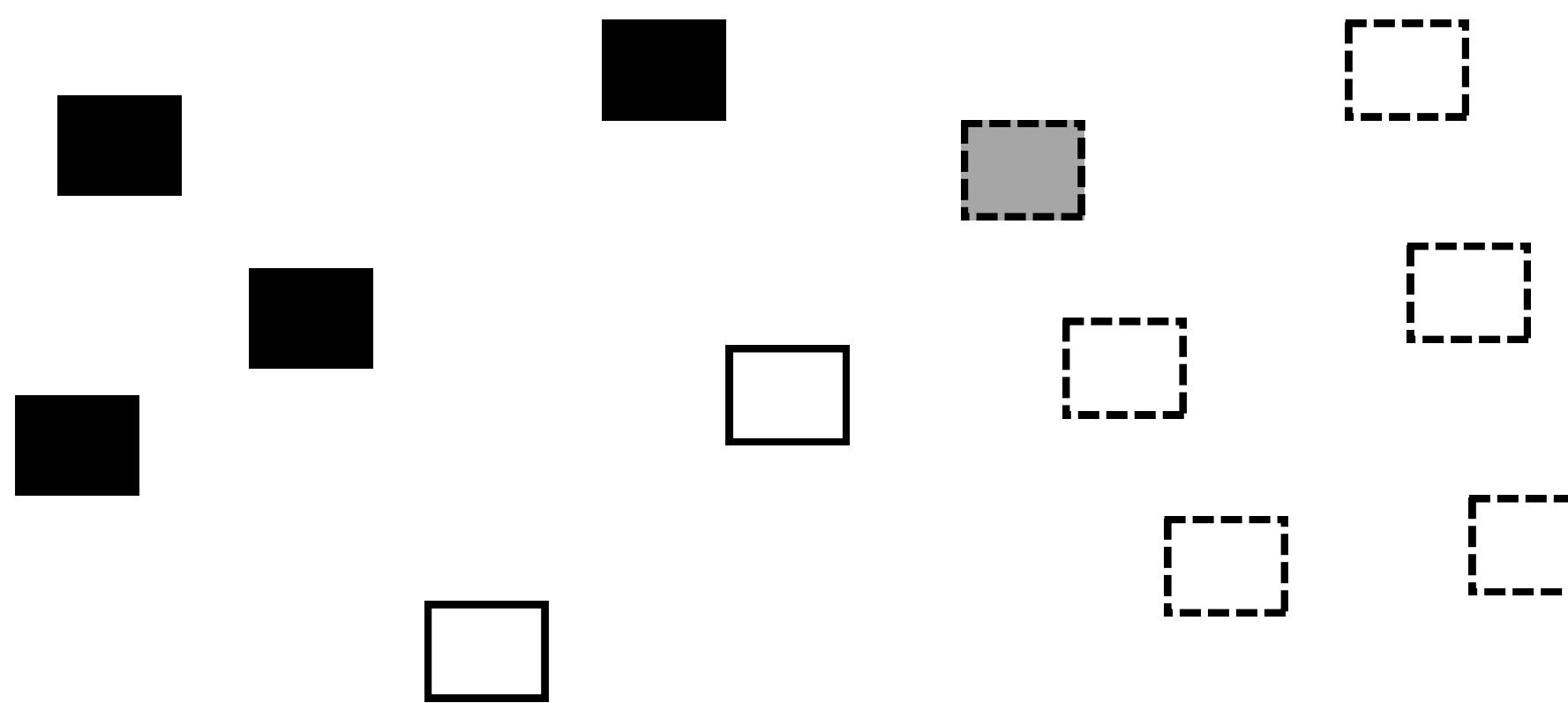
Précision, Rappel, F1

Exactitude / Accuracy

Précision, Rappel, F1



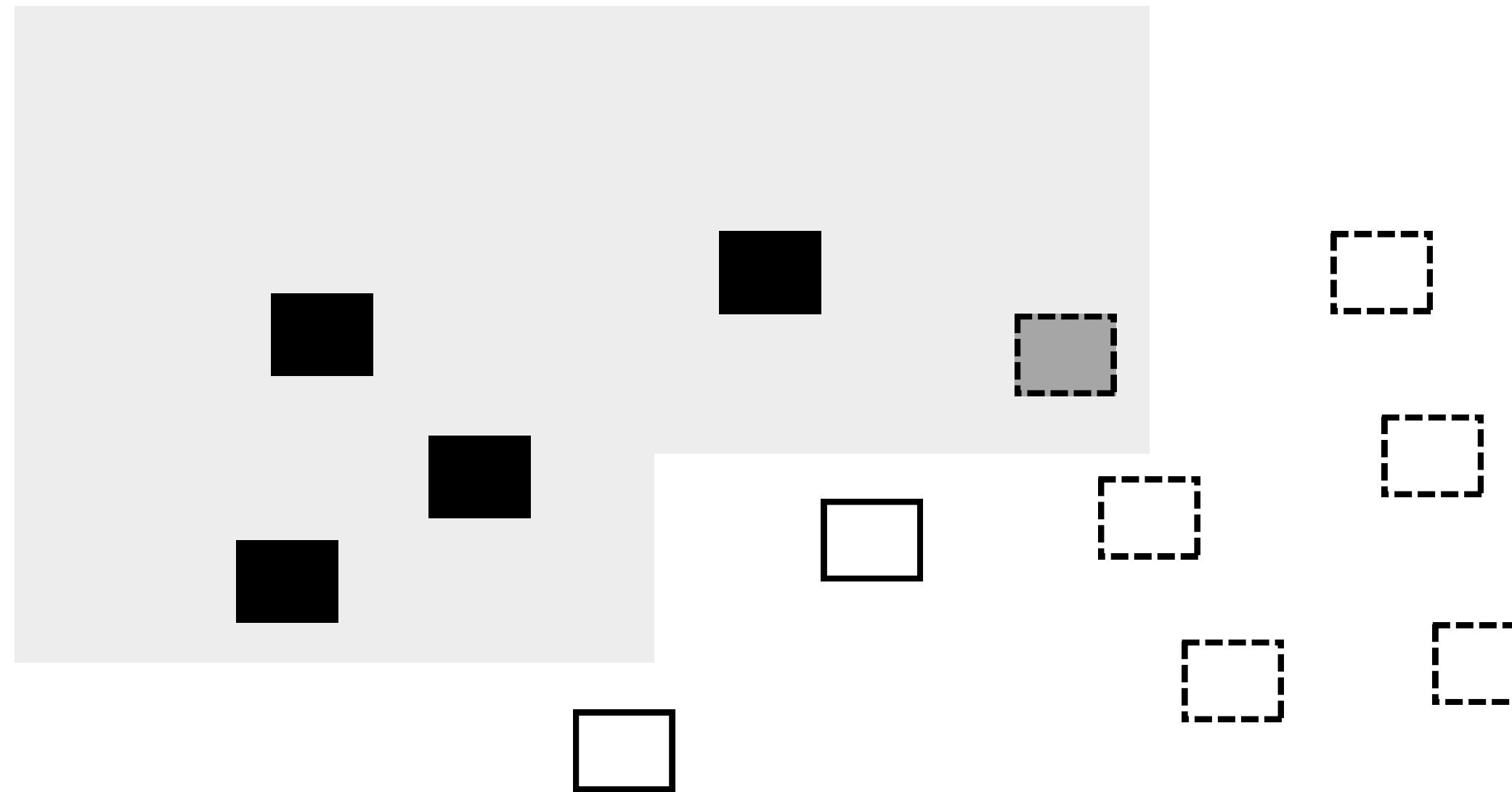
Précision, Rappel, F1



Précision, Rappel, F1

Précision :

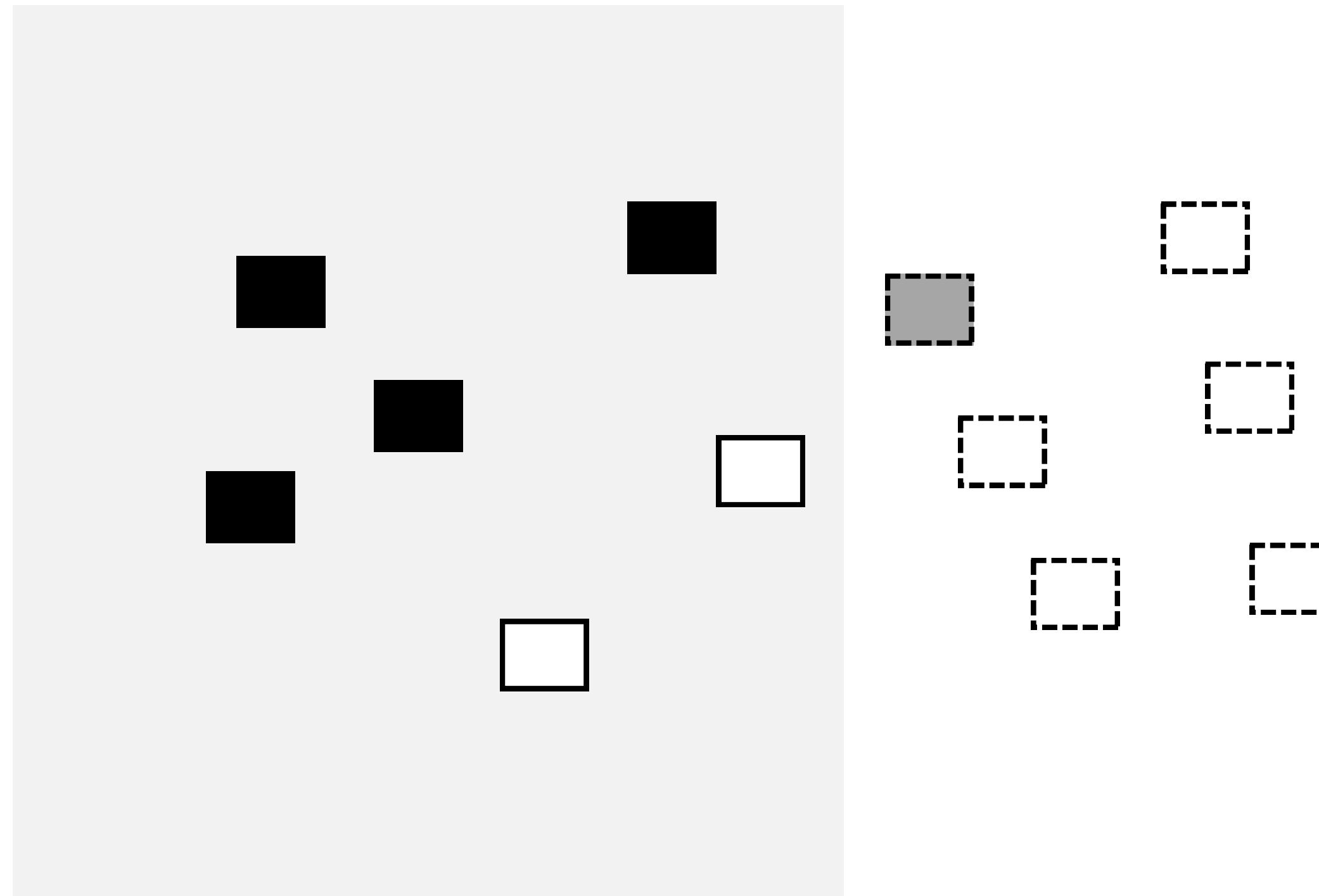
$$VP / (VP + FP)$$



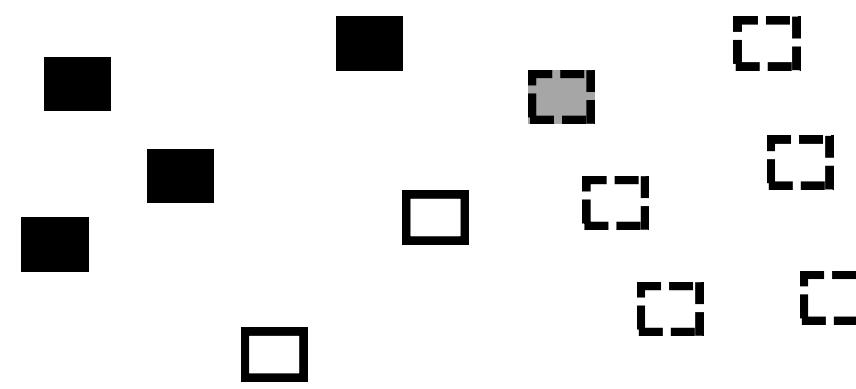
Précision, Rappel, F1

Rappel :

$$VP / (VP + FN)$$

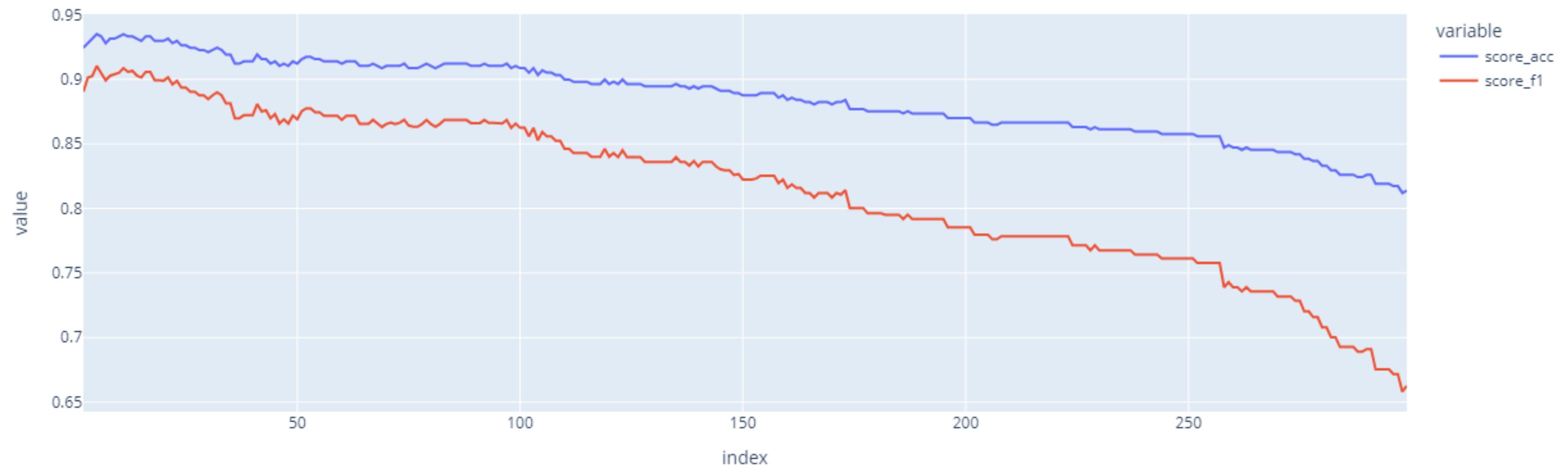


Précision, Rappel, F1

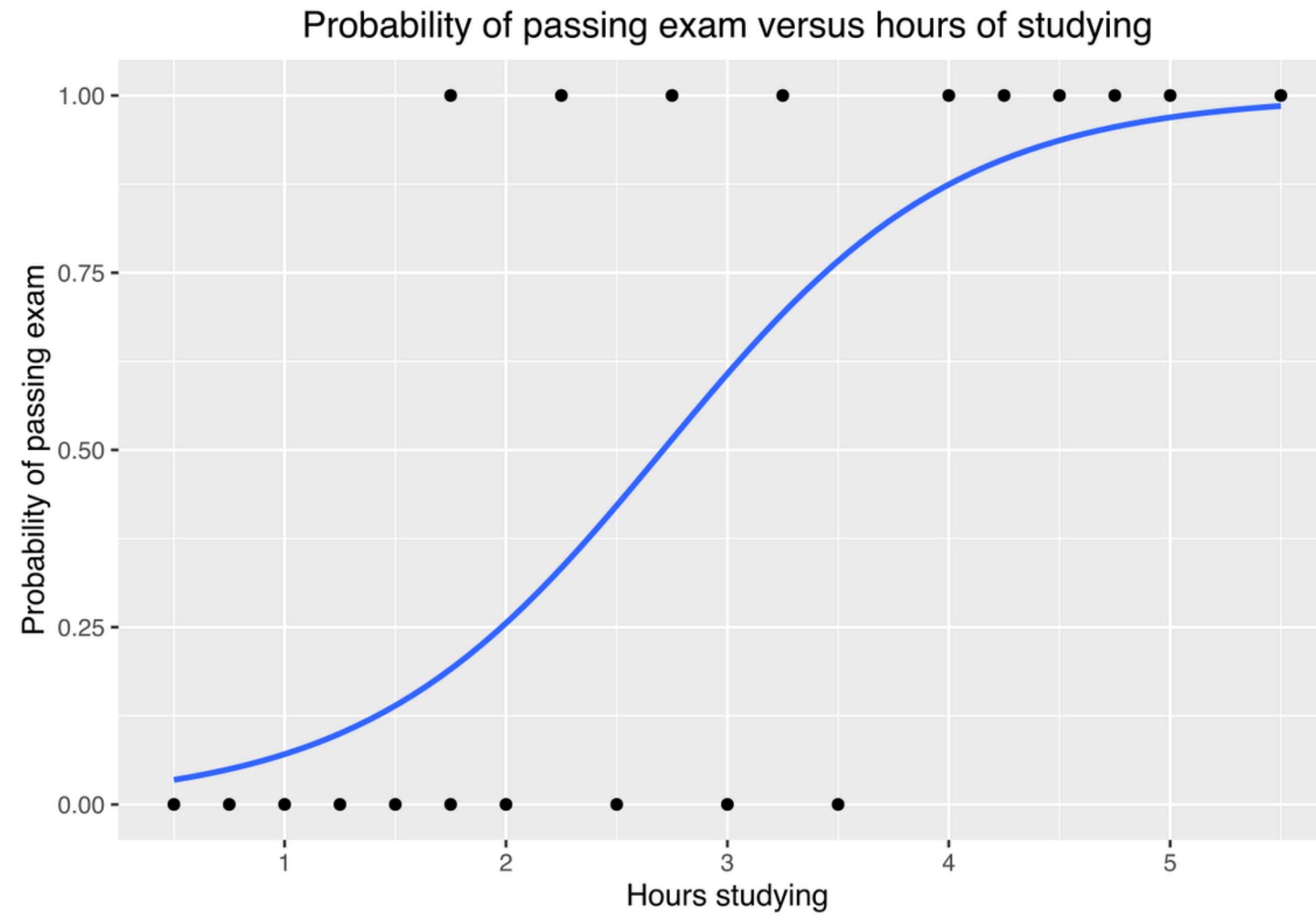


$$\begin{aligned} \text{F1 Score} &= \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \\ &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \end{aligned}$$

kNN > F1 cross-validation



Regression logistique



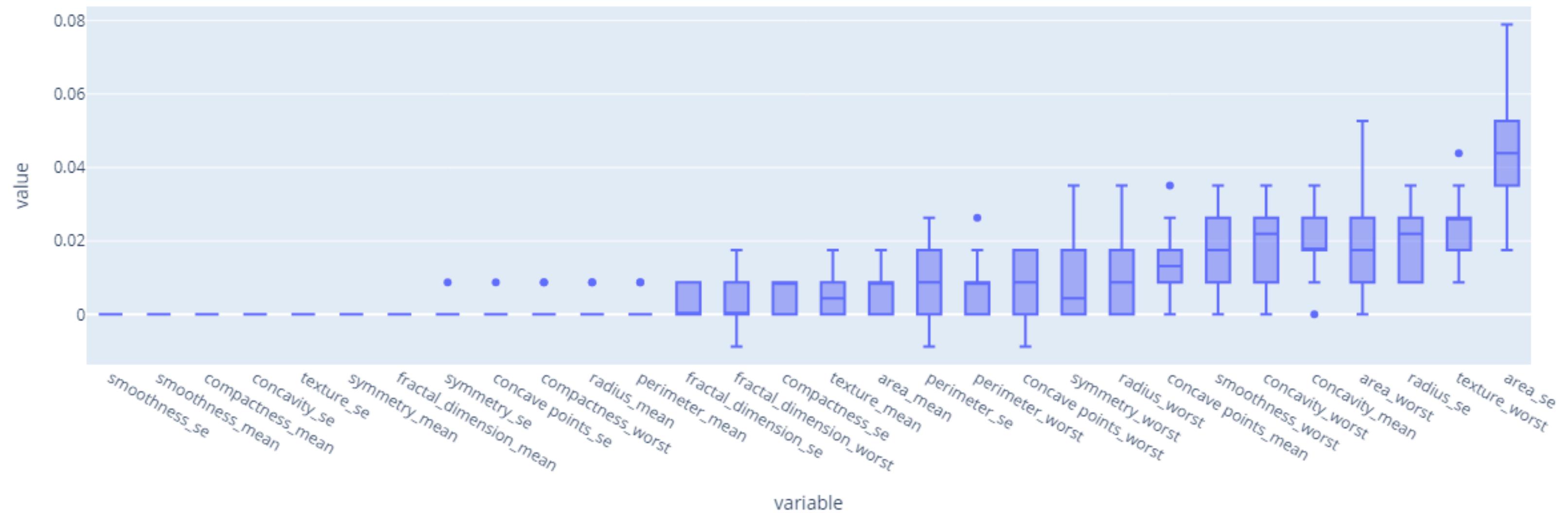
Standardisation, normalisation

$$z = \frac{x - \mu}{\sigma}$$

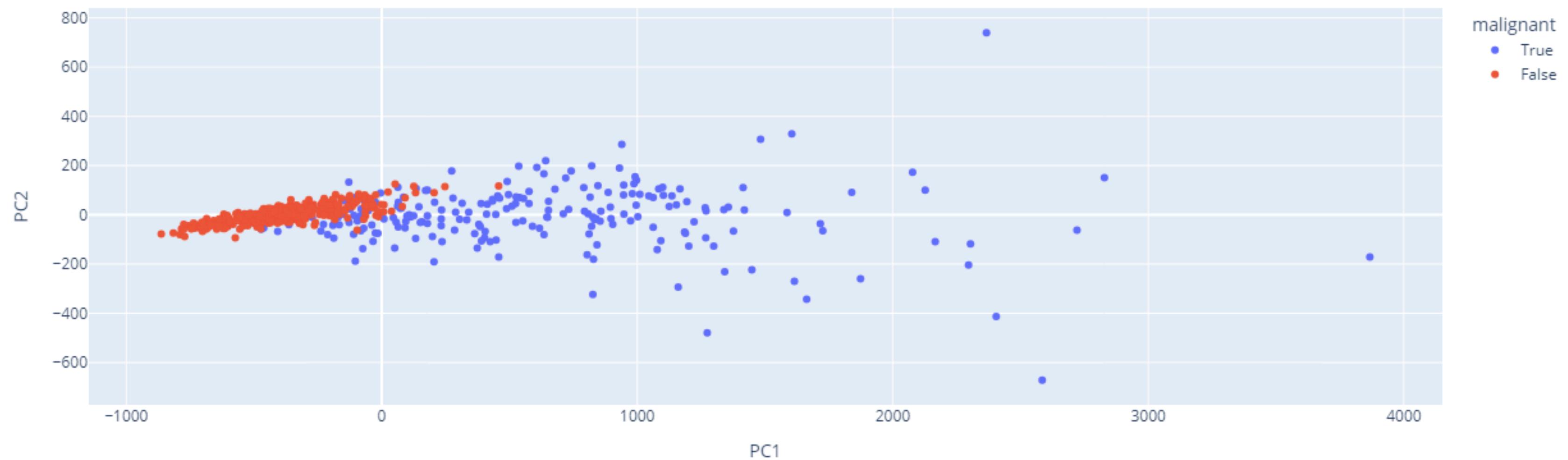
Standardisation, normalisation

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

Interprétabilité



A C P



[E t v o i l à . . .]

Fin de l'atelier :)

SIAHAAN - GENSOLLEN Rémy , 2025