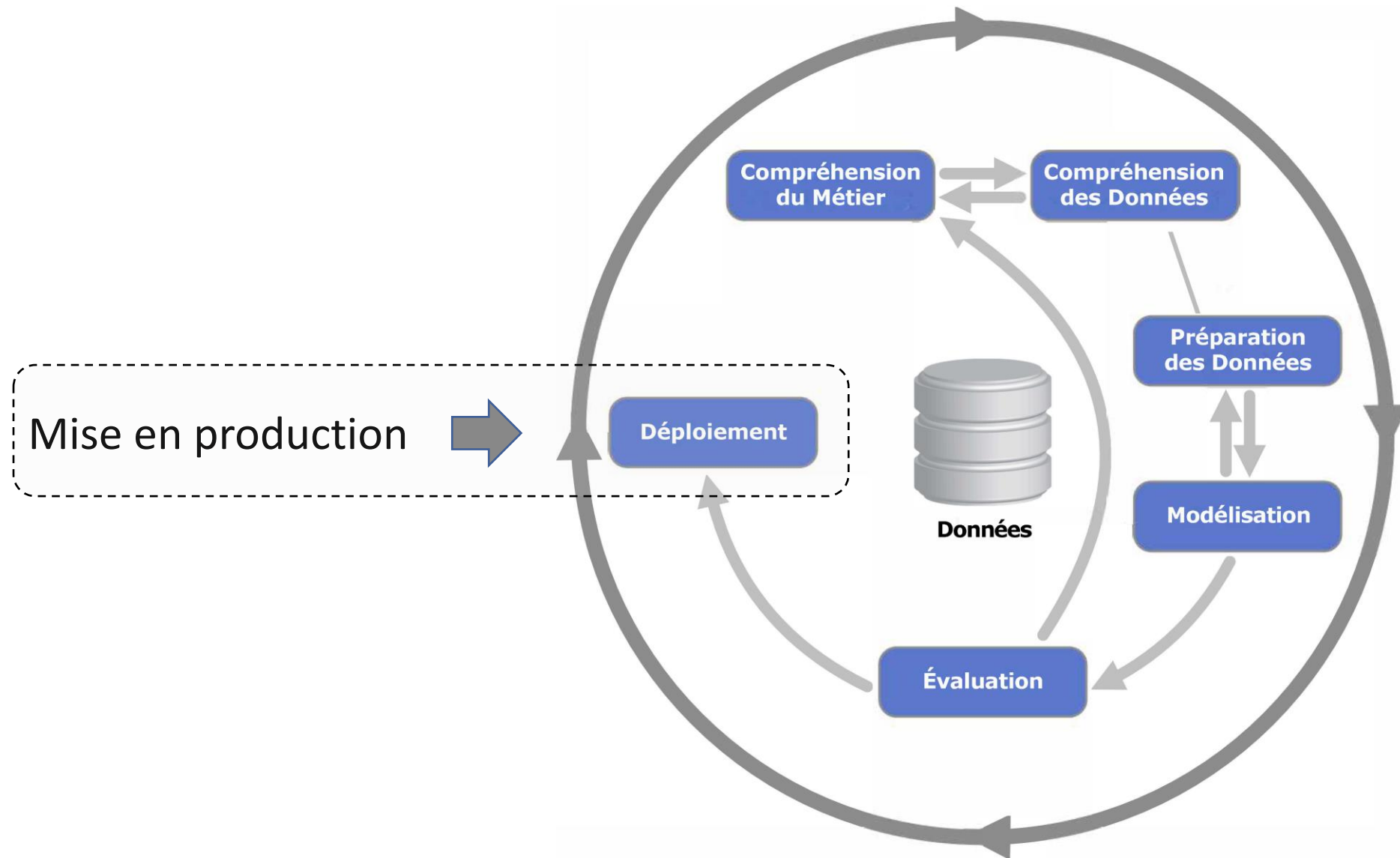
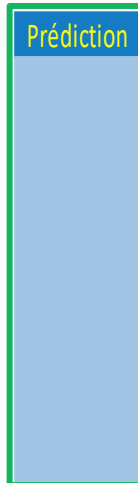
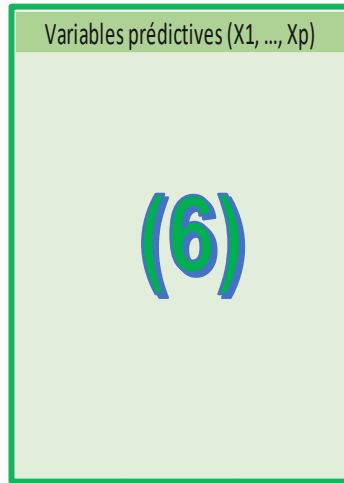


CRISP-DM



[https://fr.wikipedia.org/wiki/Cross Industry Standard Process for Data Mining](https://fr.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)

Base à classer, sans les étiquettes
(classes d'appartenance)



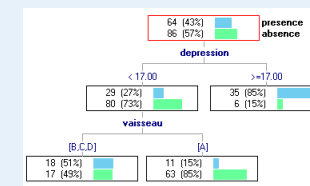
Estimation de la probabilité de mal classer

Déploiement sur la base à classer (6)

Modélisation (2)

(2)

Modélisation : Base de règles
ou Classifieur linéaire

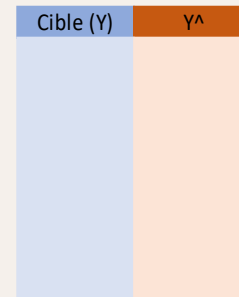


$$d(y_k, X) = a_{0,k} + a_{1,k}X_1 + a_{2,k}X_2 + \dots$$

Prédiction

Application des règles prédictives sur l'échantillon test. Construction de la prédiction \hat{Y}

(4)



Mesure des performances

Confrontation entre valeurs observées de Y et prédictions

(5)

Matrice de confusion

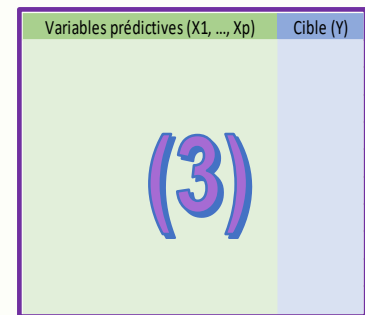
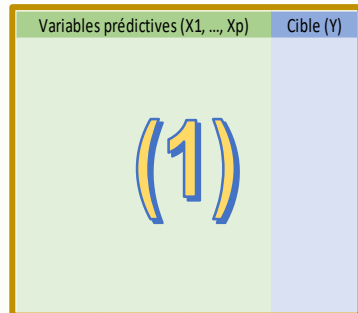
| | | \hat{Y} | |
|---|---|-----------|-----------|
| | | $+\hat{}$ | $-\hat{}$ |
| Y | + | a | b |
| | - | c | d |

Indicateurs

- Taux d'erreur
- Rappel (sensibilité)
- Précision

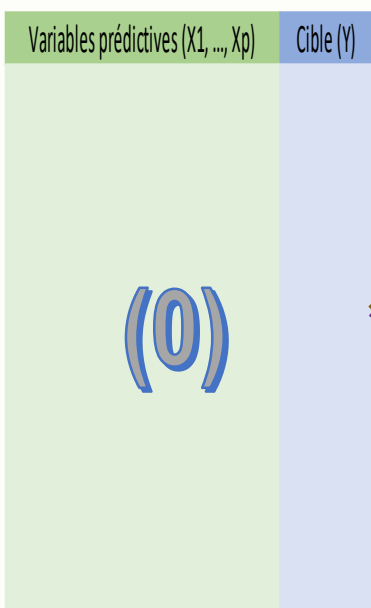
Evaluation du modèle prédictif (4 et 5)

Base d'apprentissage
(TRAIN SET)



Base de test (TEST SET)

Base initiale (DATASET)



≈ 70%

≈ 30%

Subdivision des données (0) en échantillons d'apprentissage (1) et de test (3).