



Détectez des faux billets avec R ou Python

Jérémie Quéret | Data Analyst

PROJET 12 – JANVIER 2025

PROJET 12



CONTEXTE

L'ONCFM cherche à créer un algorithme capable d'identifier automatiquement les faux billets en euros pour lutter contre la contrefaçon.

01

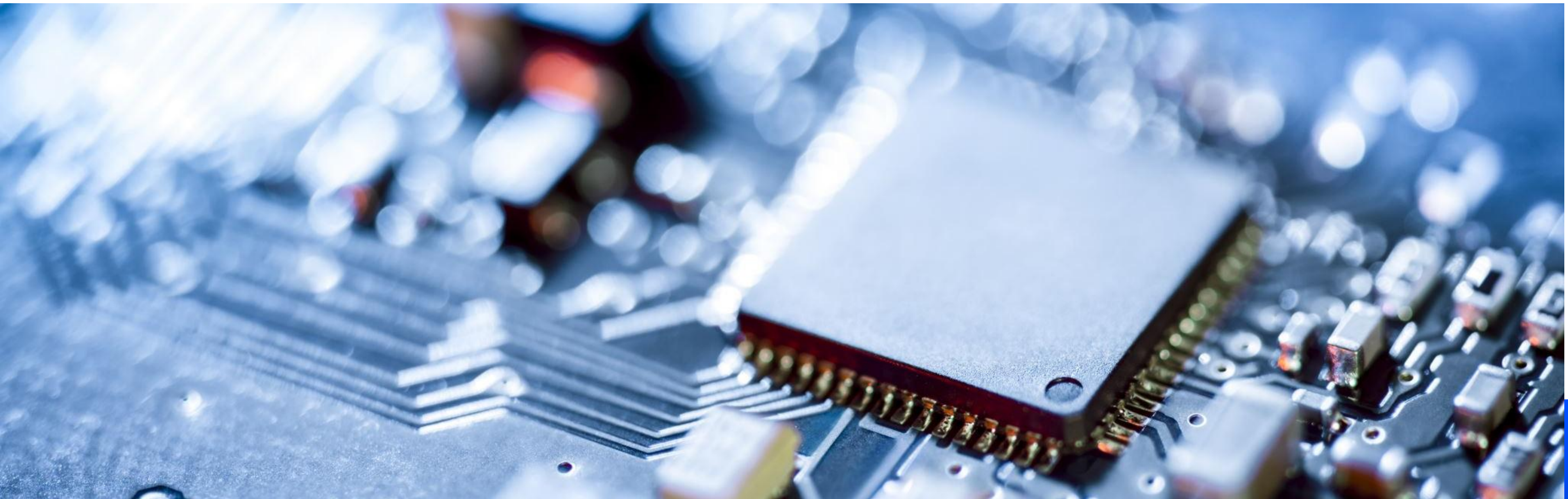
Algorithmes

- Kmeans
- Régression Logistique
- Knn
- Random Forest

02

Livrable

Notebook Python ou R.



Les variables



Length

La longueur du billet (en mm)



Height_left

La hauteur du billet (mesurée sur le côté gauche, en mm)



Height_right

La hauteur du billet (mesurée sur le côté droit, en mm)



Diagonal

La diagonale du billet (en mm).



Margin_low

La marge entre le bord inférieur du billet et l'image (en mm)



Margin_up

La marge entre le bord supérieur du billet et l'image (en mm)

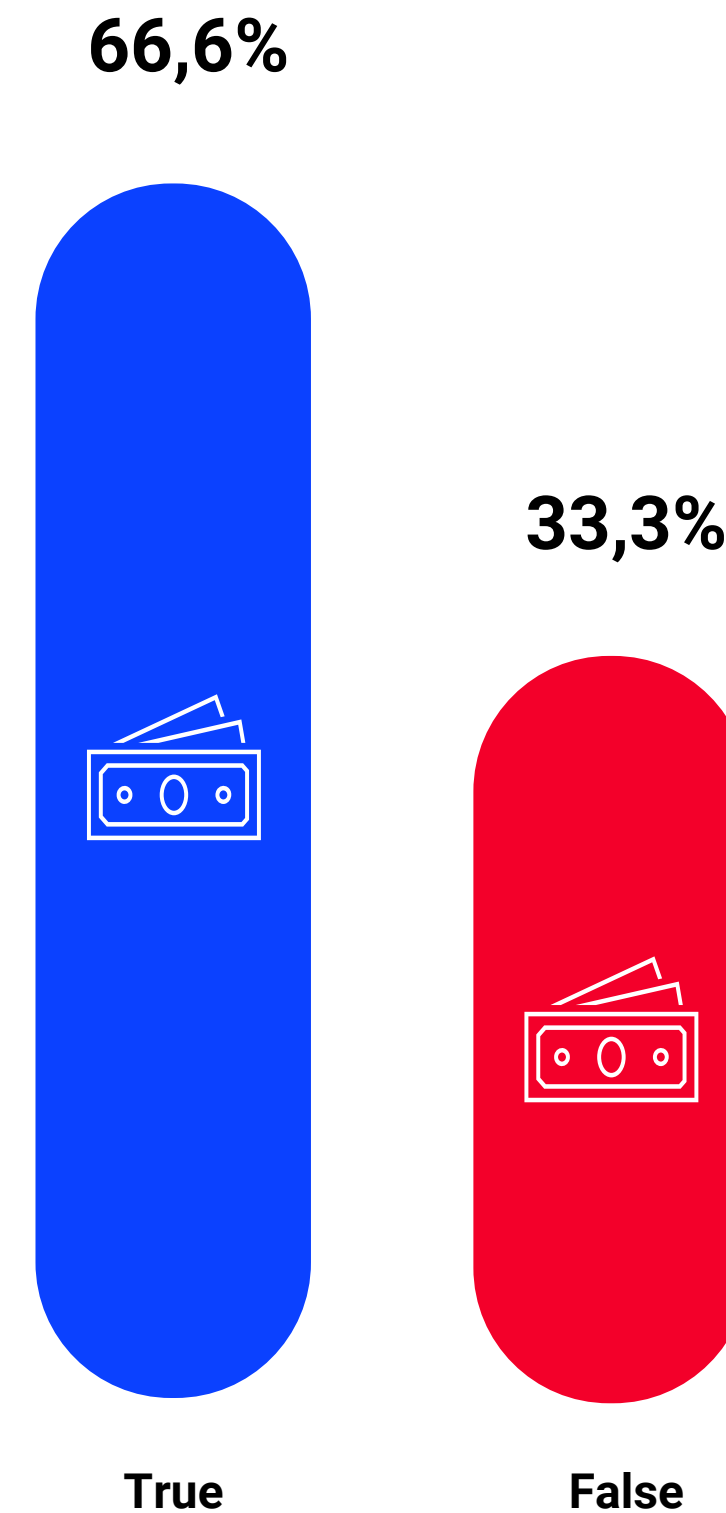


La variable cible

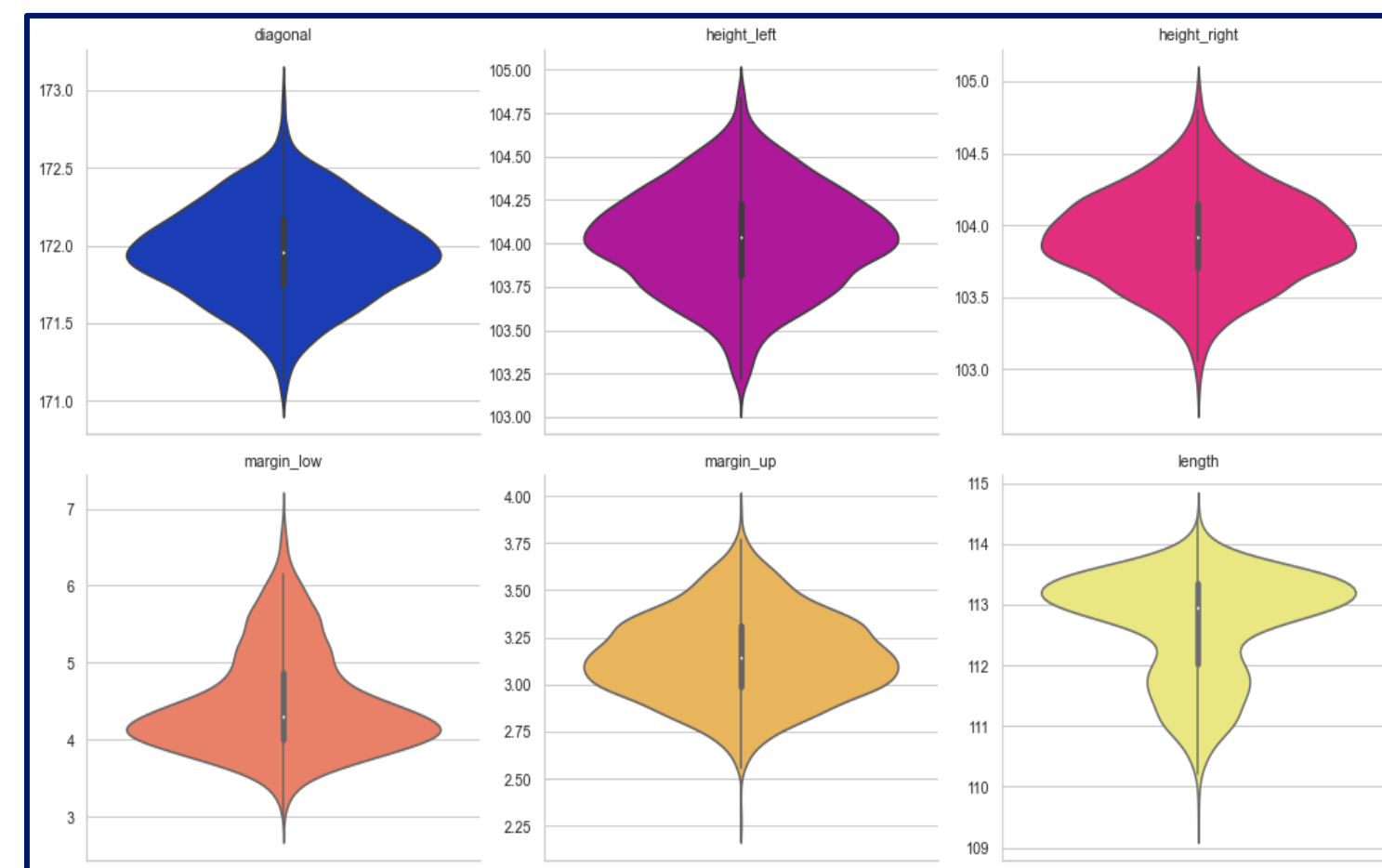
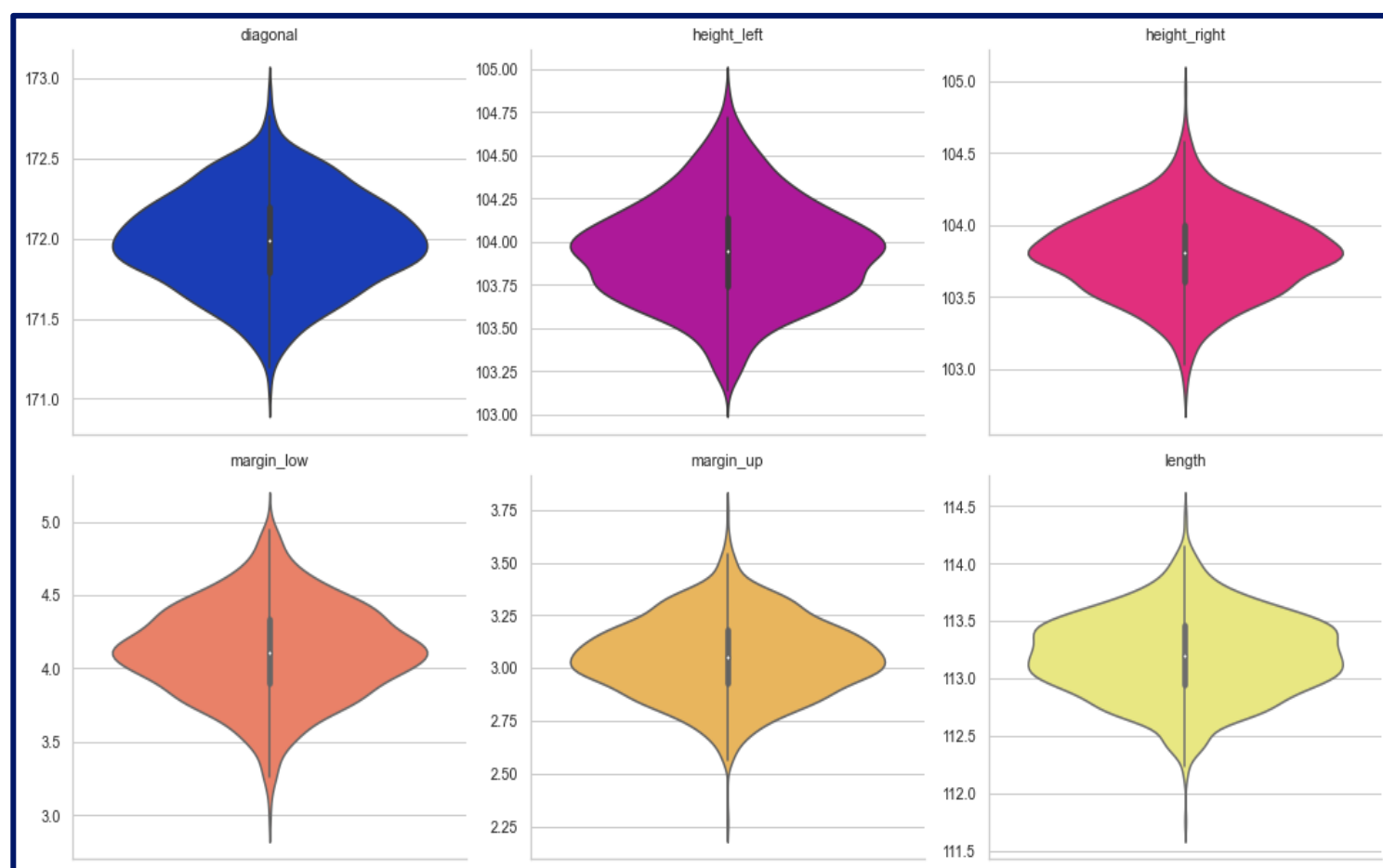
Is_genuine

1000 vrais billets

500 faux billets



Le vrai du faux, ça se voit !



Imputation des données manquantes

37 valeurs à imputer



Modèle linéaire (sel. backward)

OLS Regression Results

```
=====
Dep. Variable:          margin_low    R-squared:                0.469
Model:                  OLS           Adj. R-squared:            0.468
Method:                 Least Squares  F-statistic:              429.7
Date:                  Wed, 22 Jan 2025  Prob (F-statistic):      5.37e-200
Time:                  11:31:56        Log-Likelihood:           -1012.8
No. Observations:      1463           AIC:                     2034.
Df Residuals:          1459           BIC:                     2055.
Df Model:               3
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	23.1826	5.443	4.259	0.000	12.506	33.859
height_right	0.2745	0.043	6.378	0.000	0.190	0.359
margin_up	0.2805	0.065	4.343	0.000	0.154	0.407
length	-0.4269	0.018	-23.962	0.000	-0.462	-0.392

```
=====
Omnibus:                81.653    Durbin-Watson:            1.888
Prob(Omnibus):          0.000    Jarque-Bera (JB):        108.015
Skew:                   0.514    Prob(JB):                3.51e-24
Kurtosis:               3.847    Cond. No.                 6.59e+04
=====
```

Variables sélectionnées :

['height_right', 'margin_up', 'length']

R2 0.4691

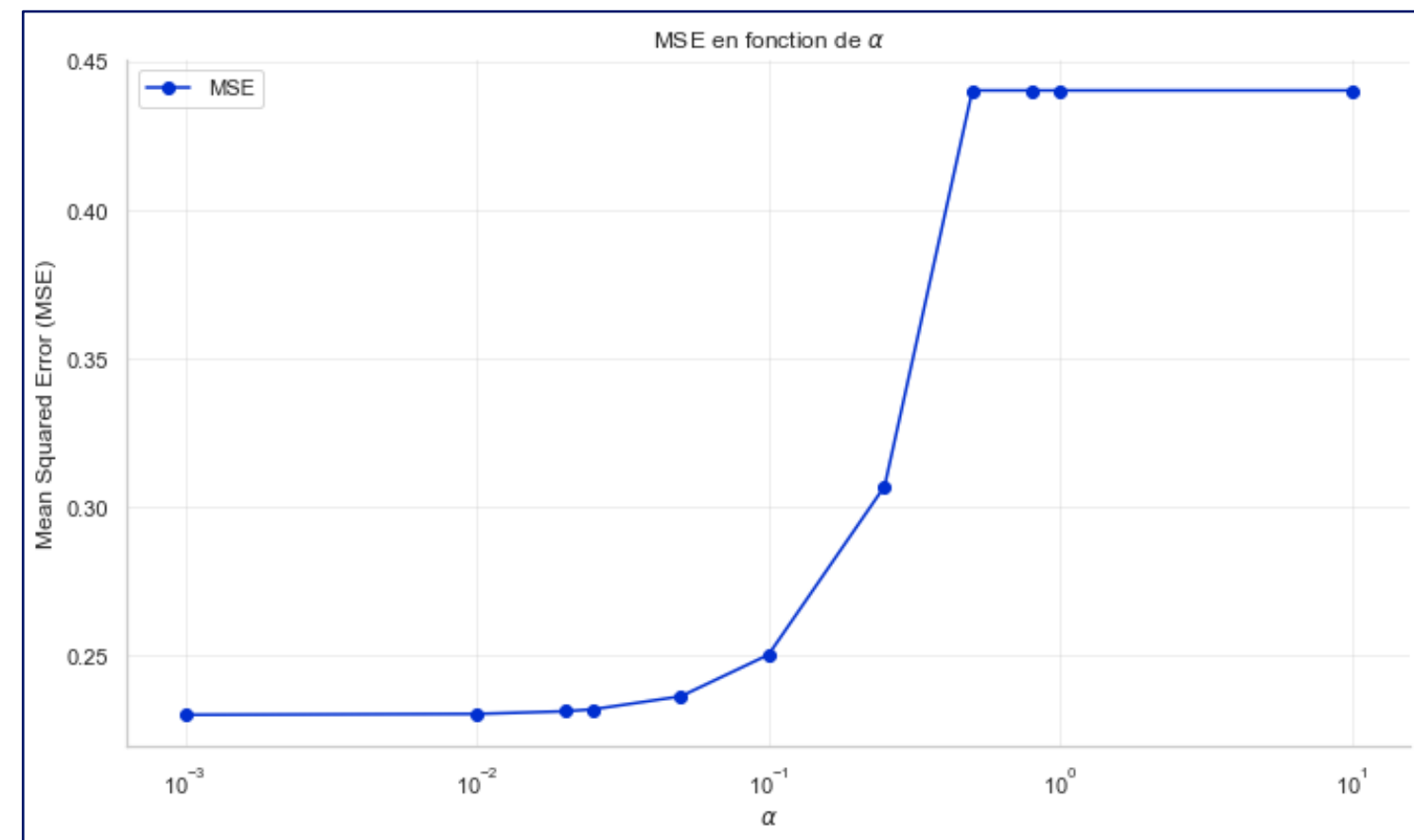
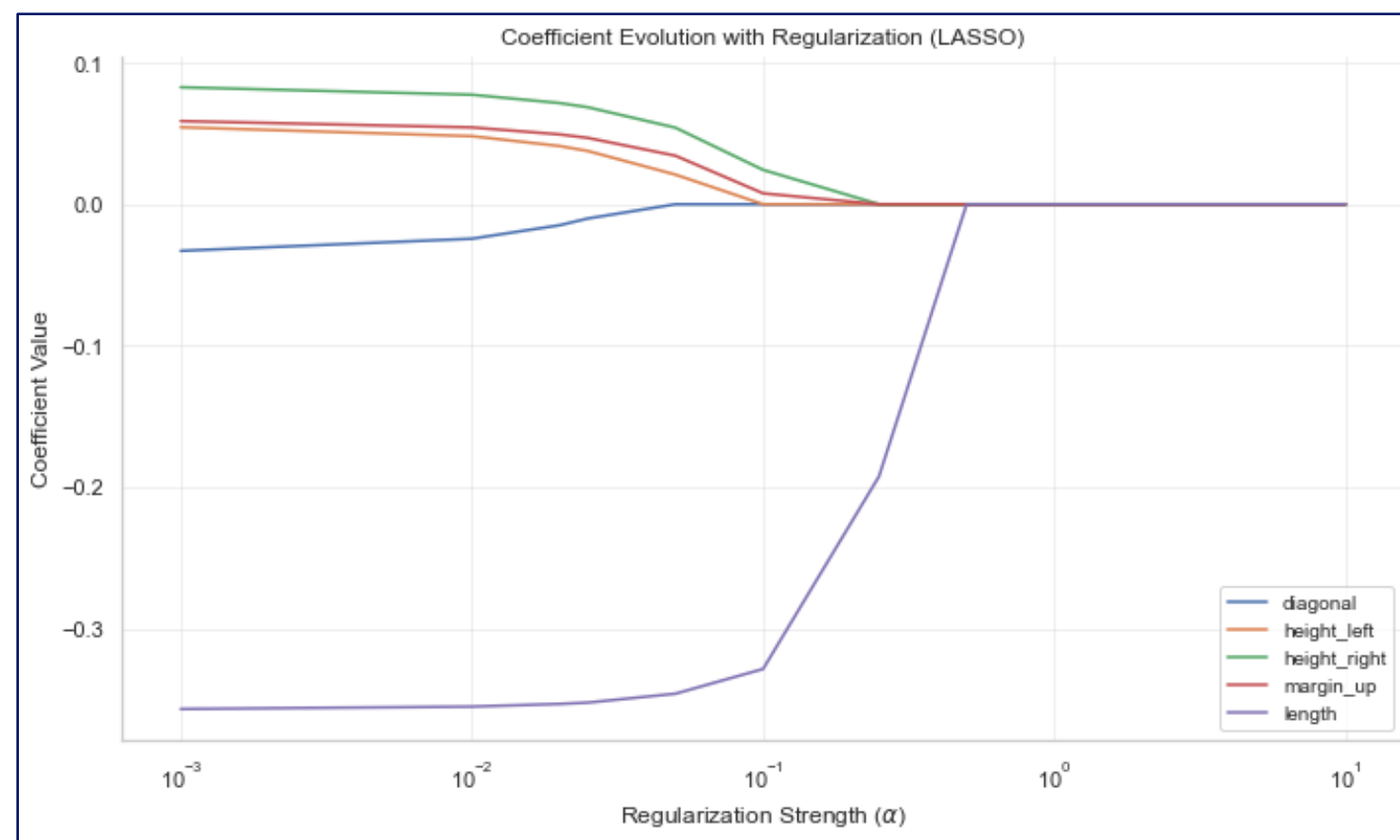
MAE 0.3748

MSE 0.2338

RMSE 0.4835

Régression Lasso

Variables sélectionnées : ['diagonal', 'height_left', 'height_right', 'margin_up', 'length']



R2 0.4773

MAE 0.3722

MSE 0.2302

RMSE 0.4797

Ajustement du modèle

Transformation inverse de la variable cible (margin_low)

R2

0.4576

MAE

0.0181

MSE

0.0005

RMSE

0.0229



Validité du modèle



Normalité des résidus

Les résidus ne suivent pas une loi normale.



homoscédasticité

Homoscédasticité confirmé (variance des résidus constante)



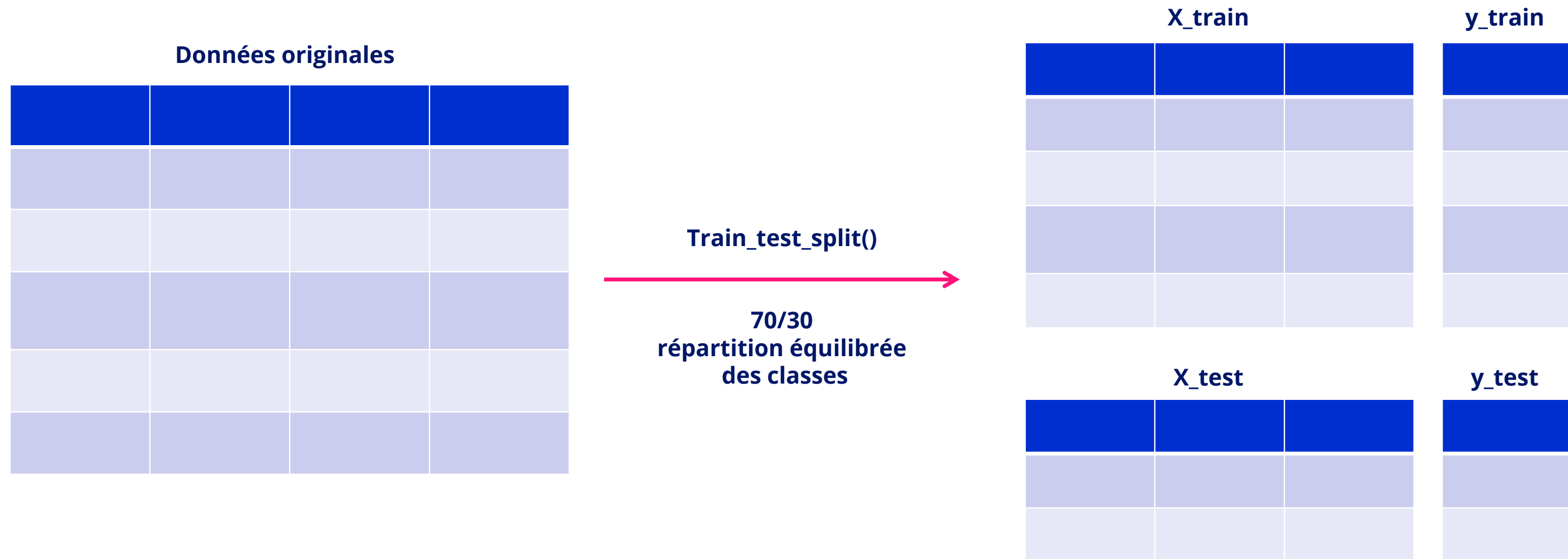
Multicolinéarité

$1 < VIF < 5$:
Niveau modéré de corrélation, acceptable.

Développement et test des algorithmes

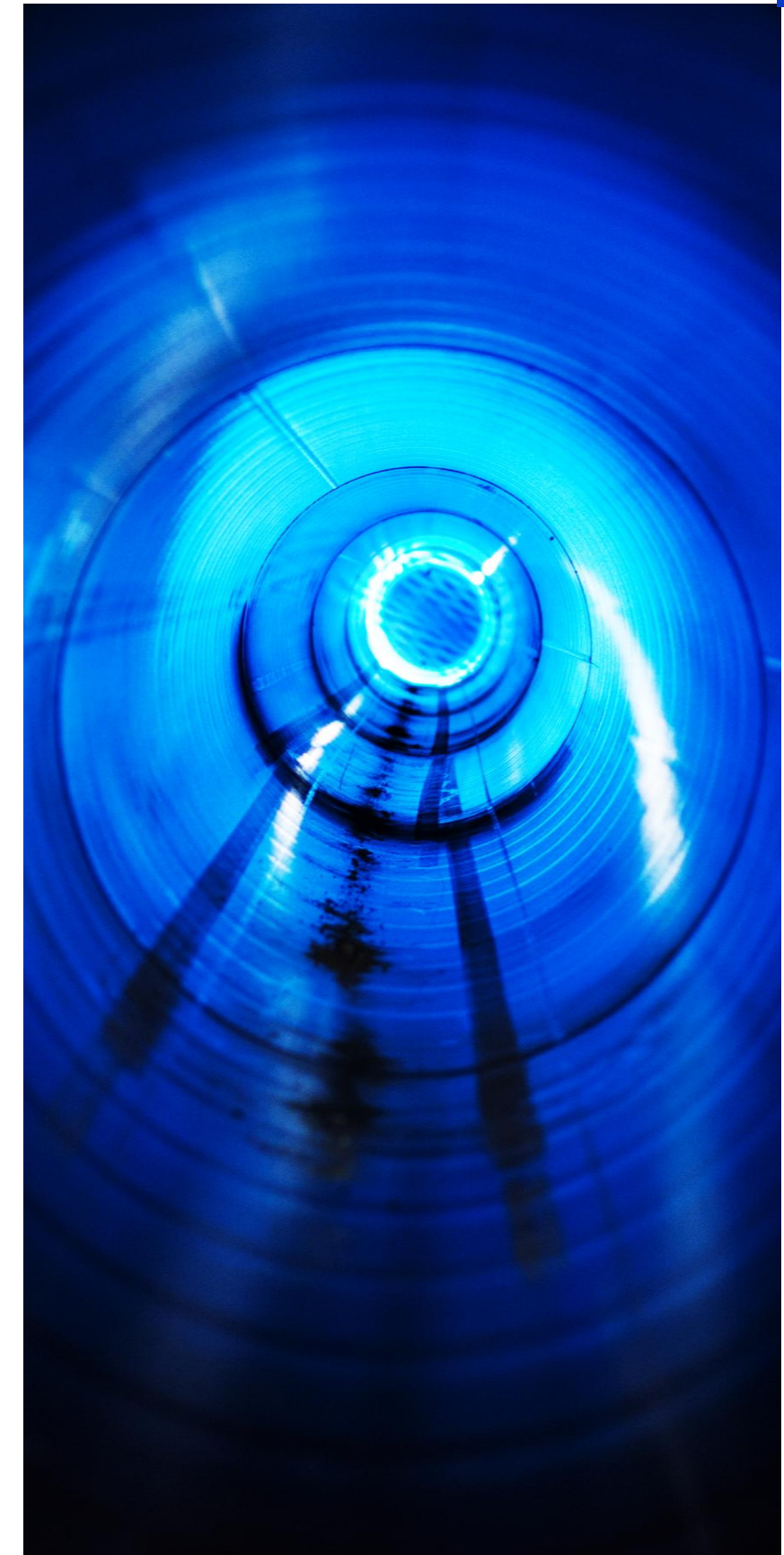
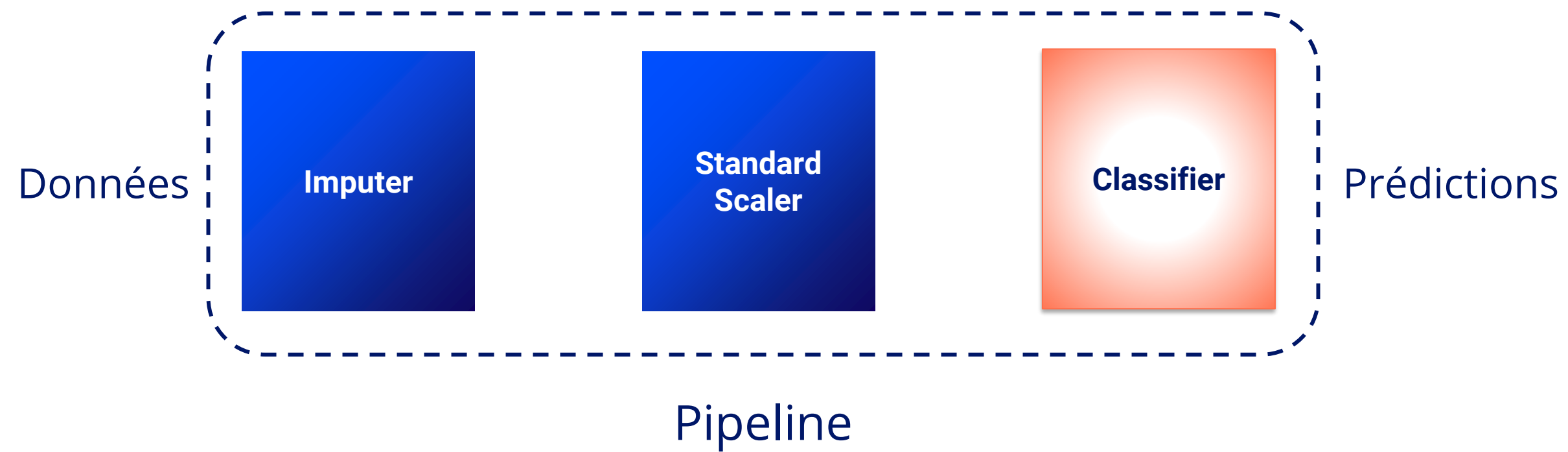


Diviser le jeu de données



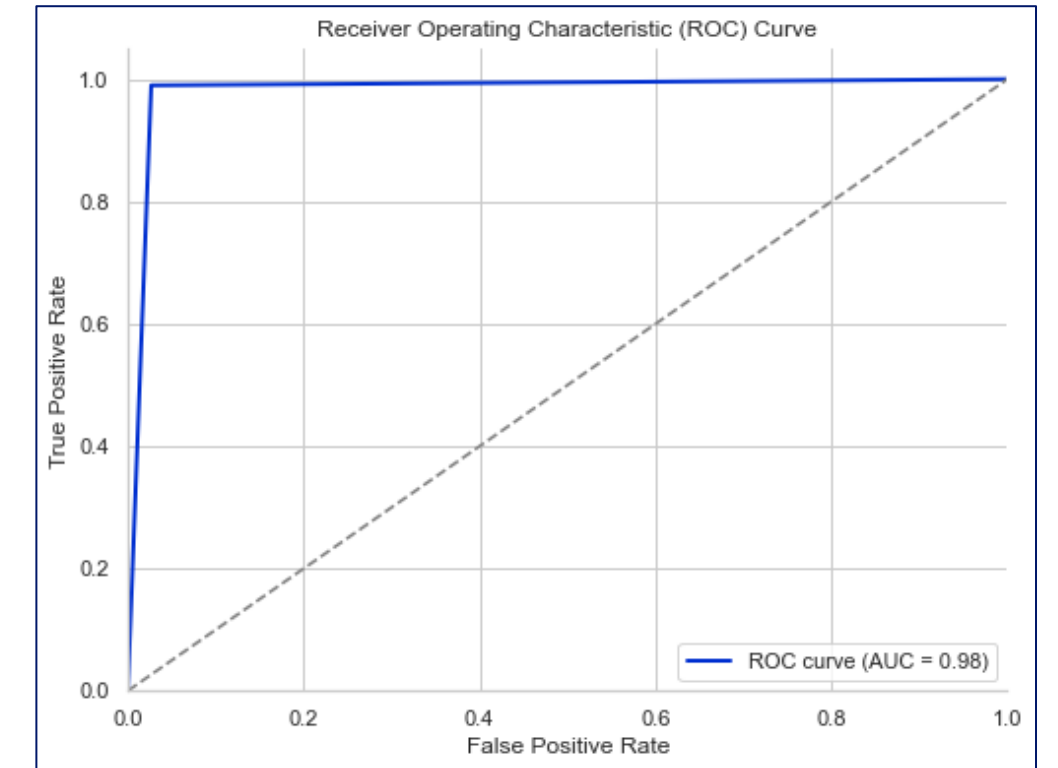
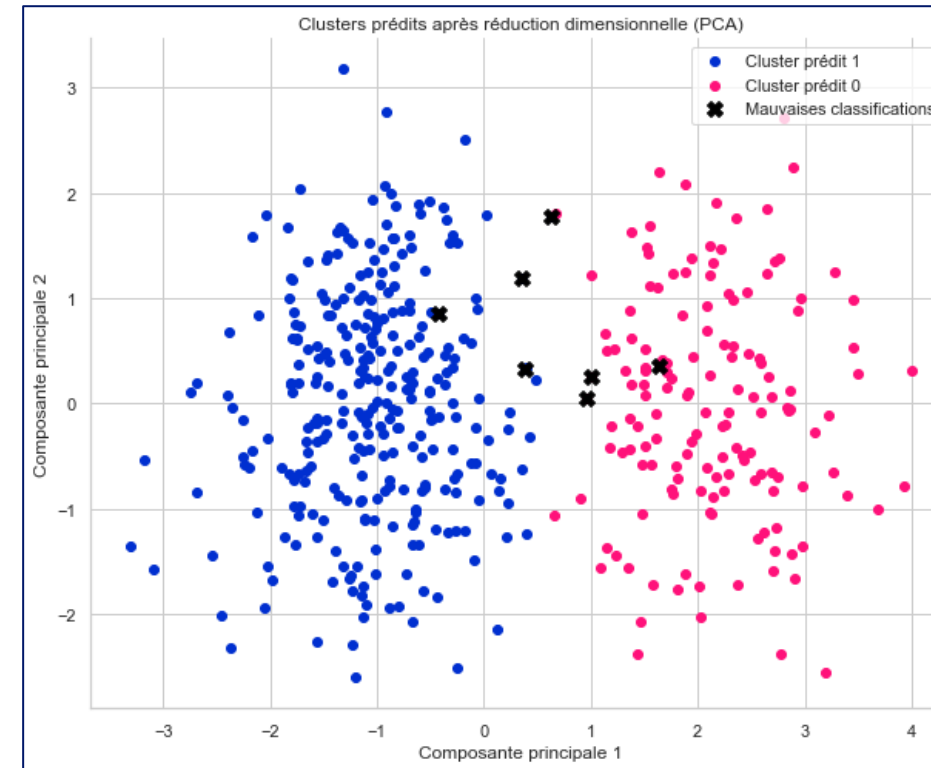
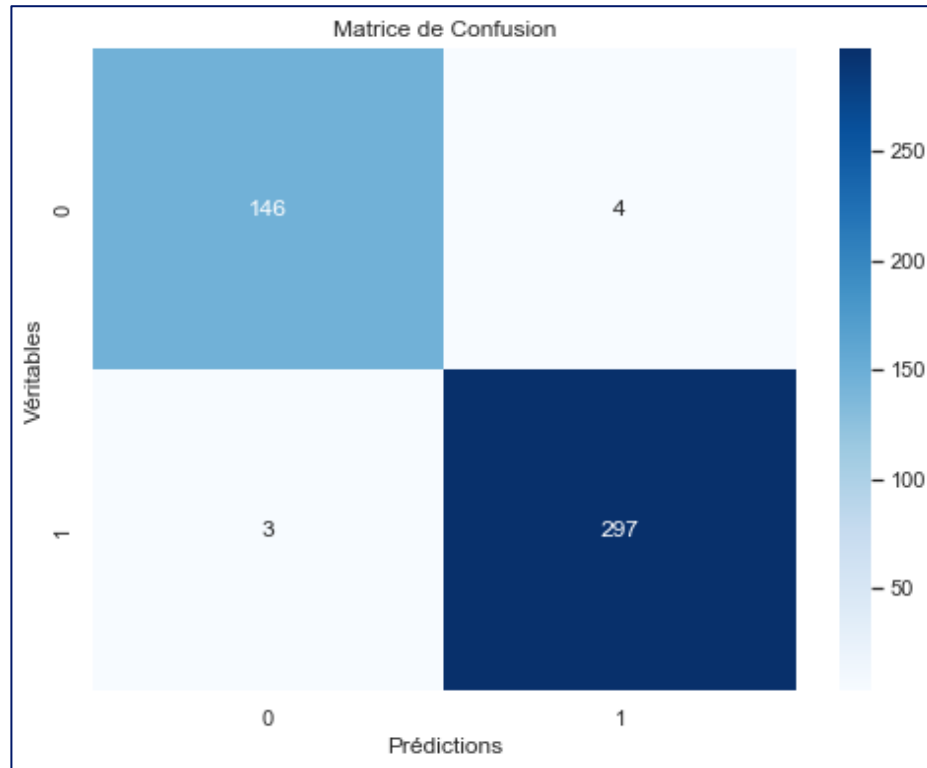


Workflow & Pipeline



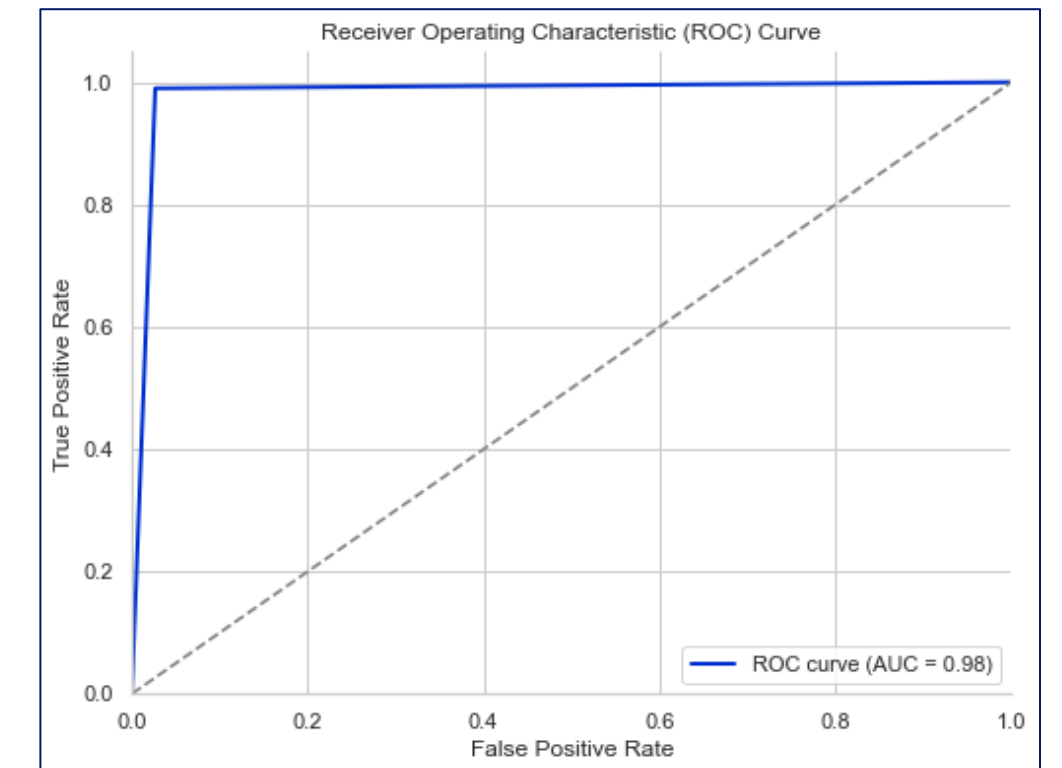
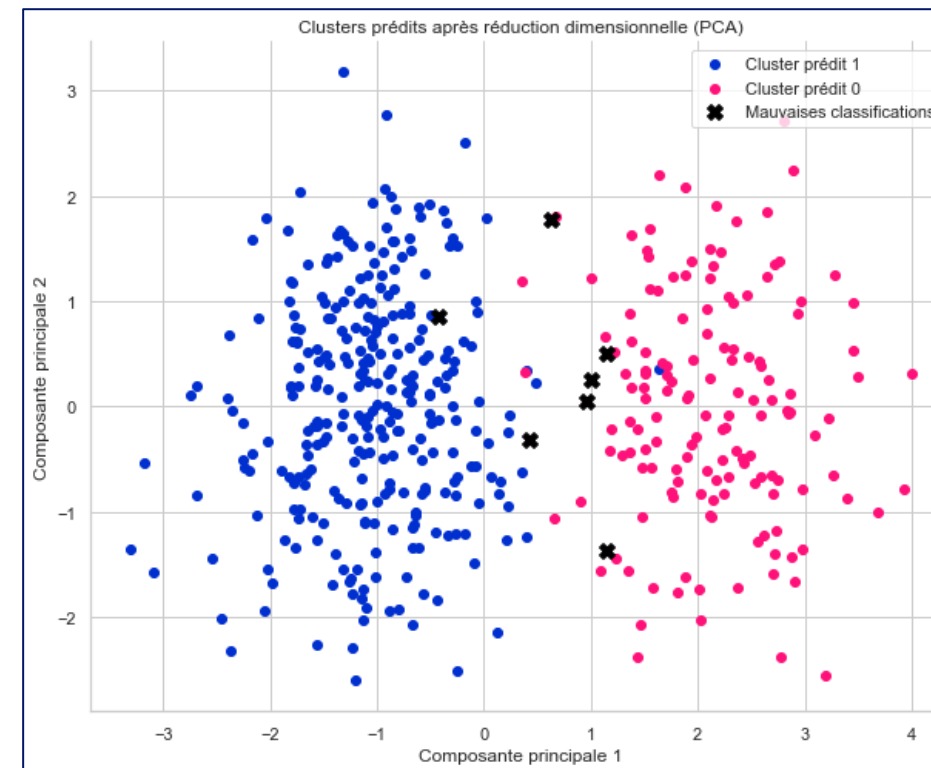
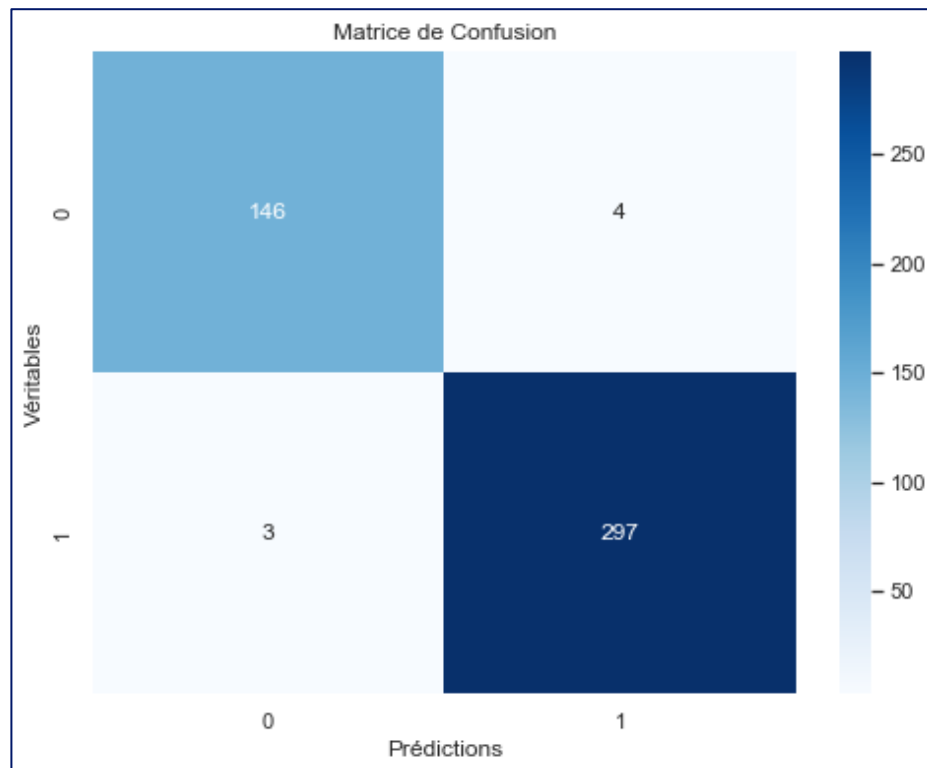


Kmeans



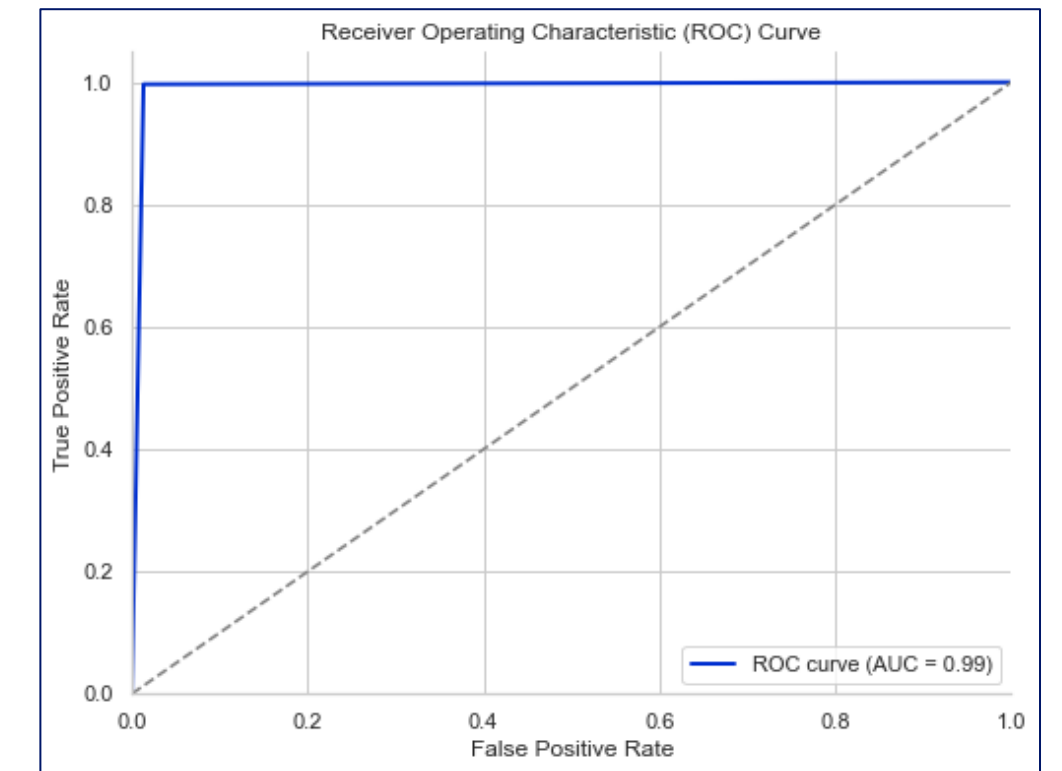
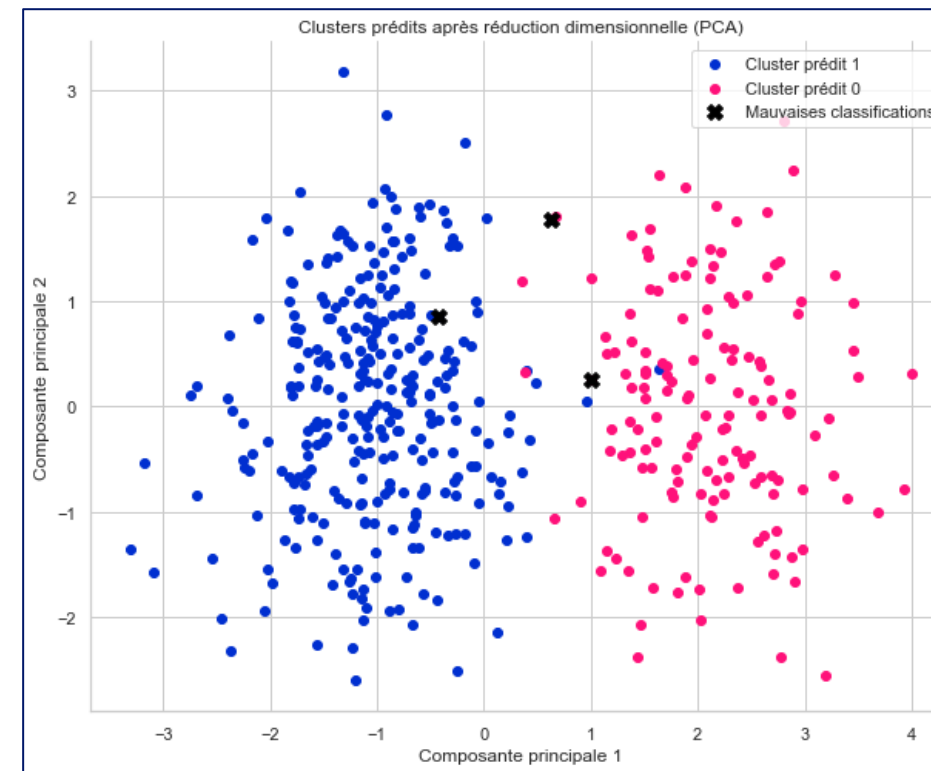
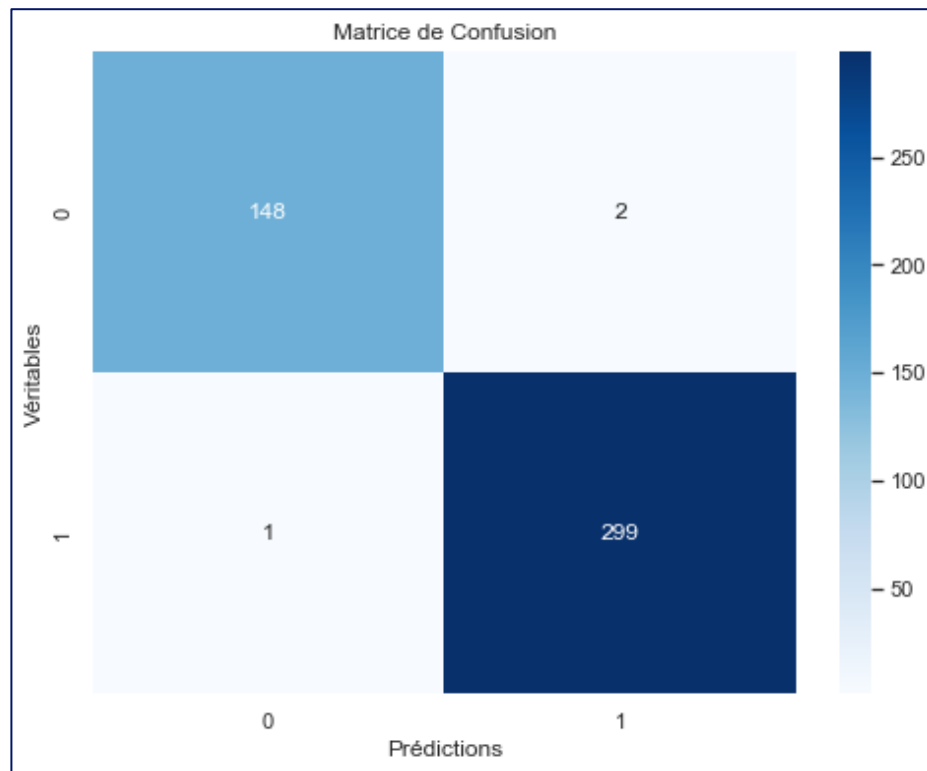


Knn



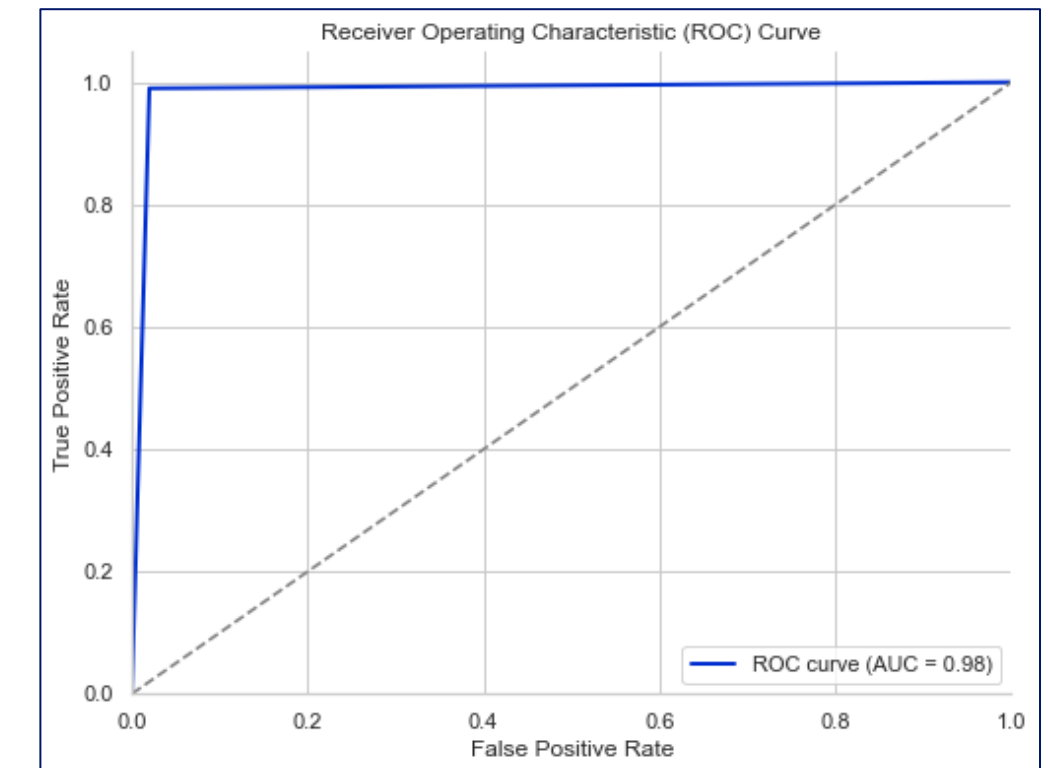
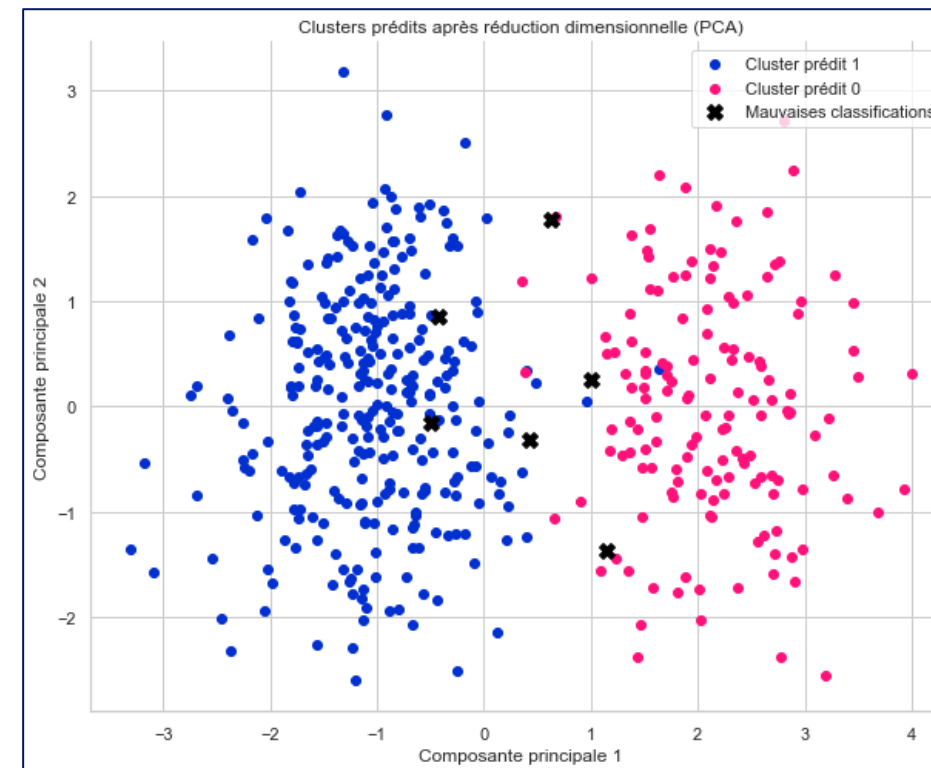
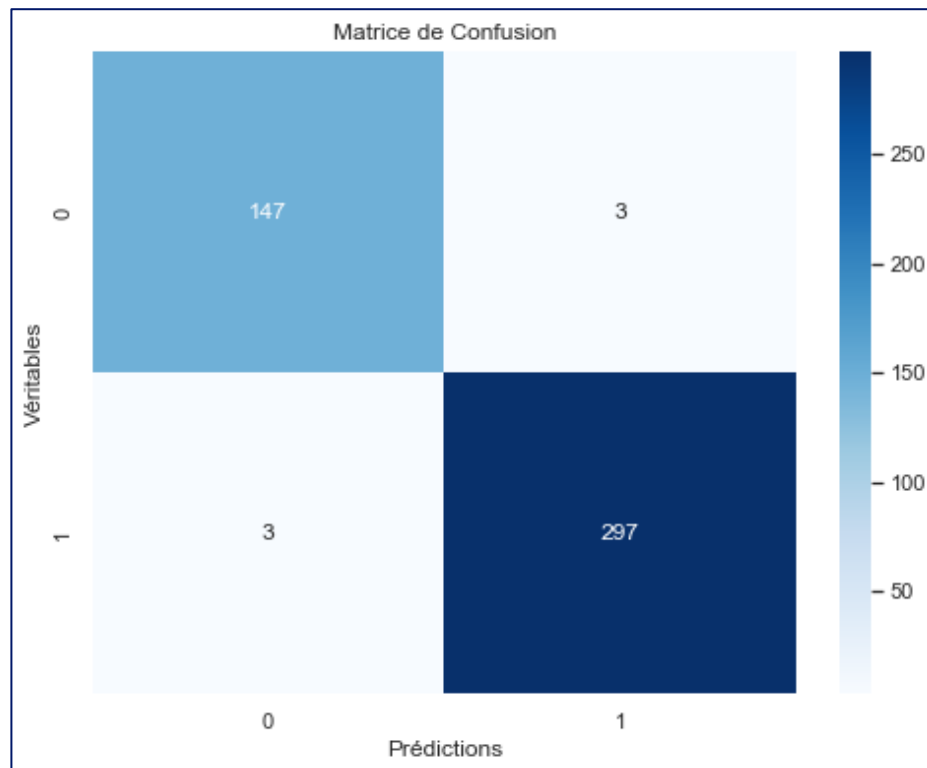


Régression Logistique





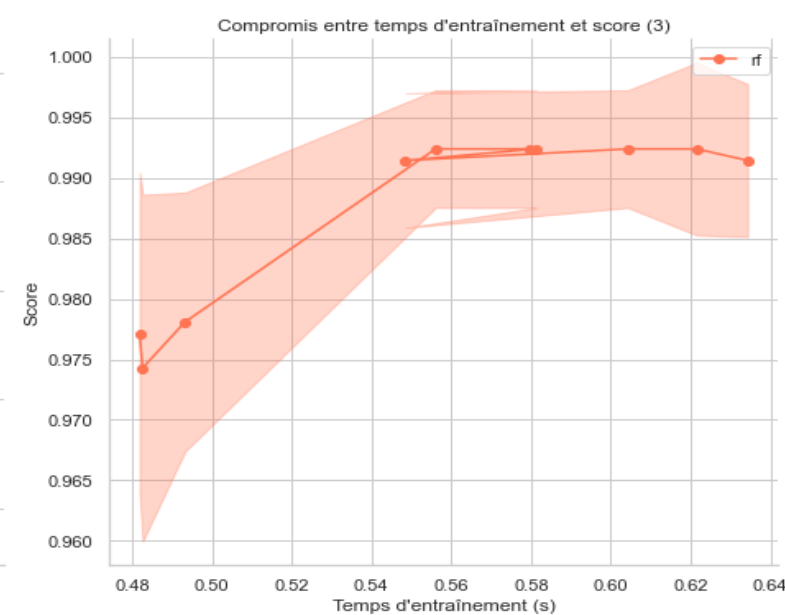
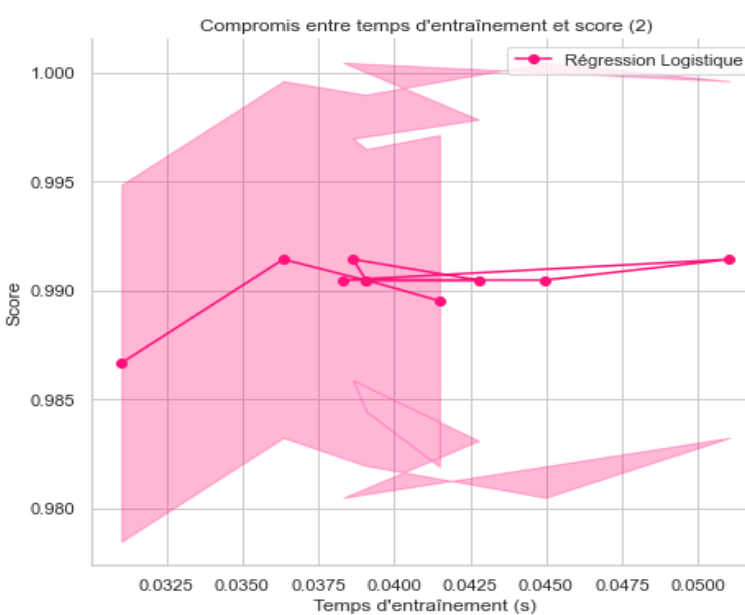
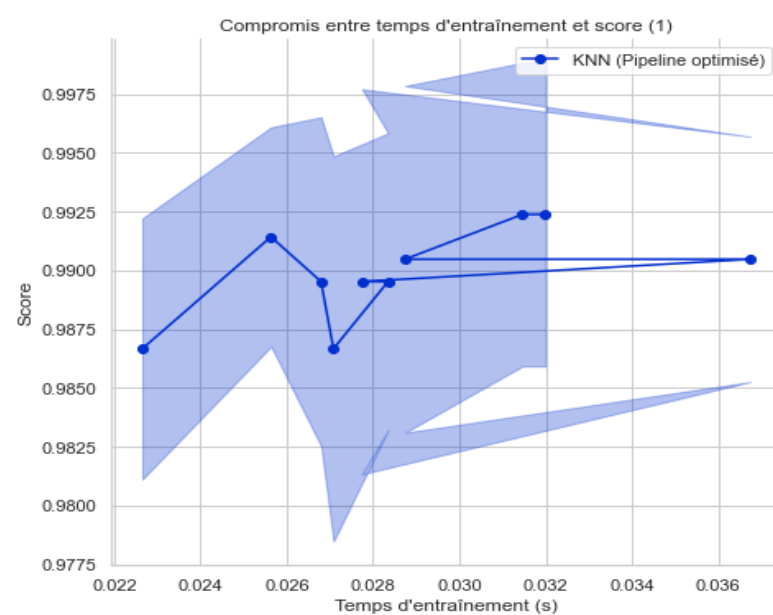
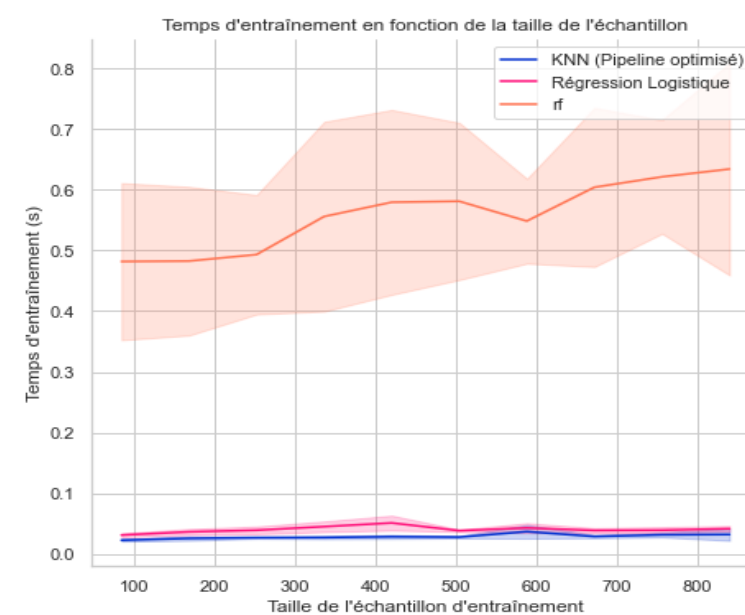
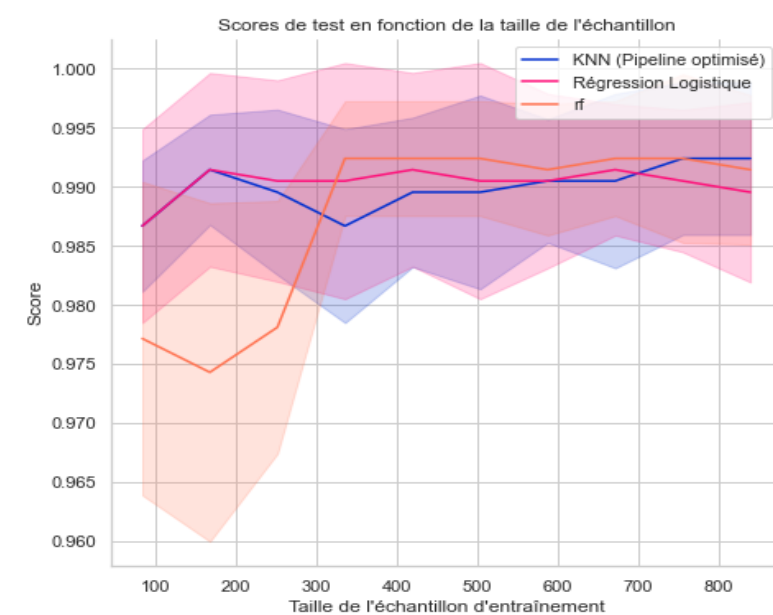
Random Forest



Synthèse

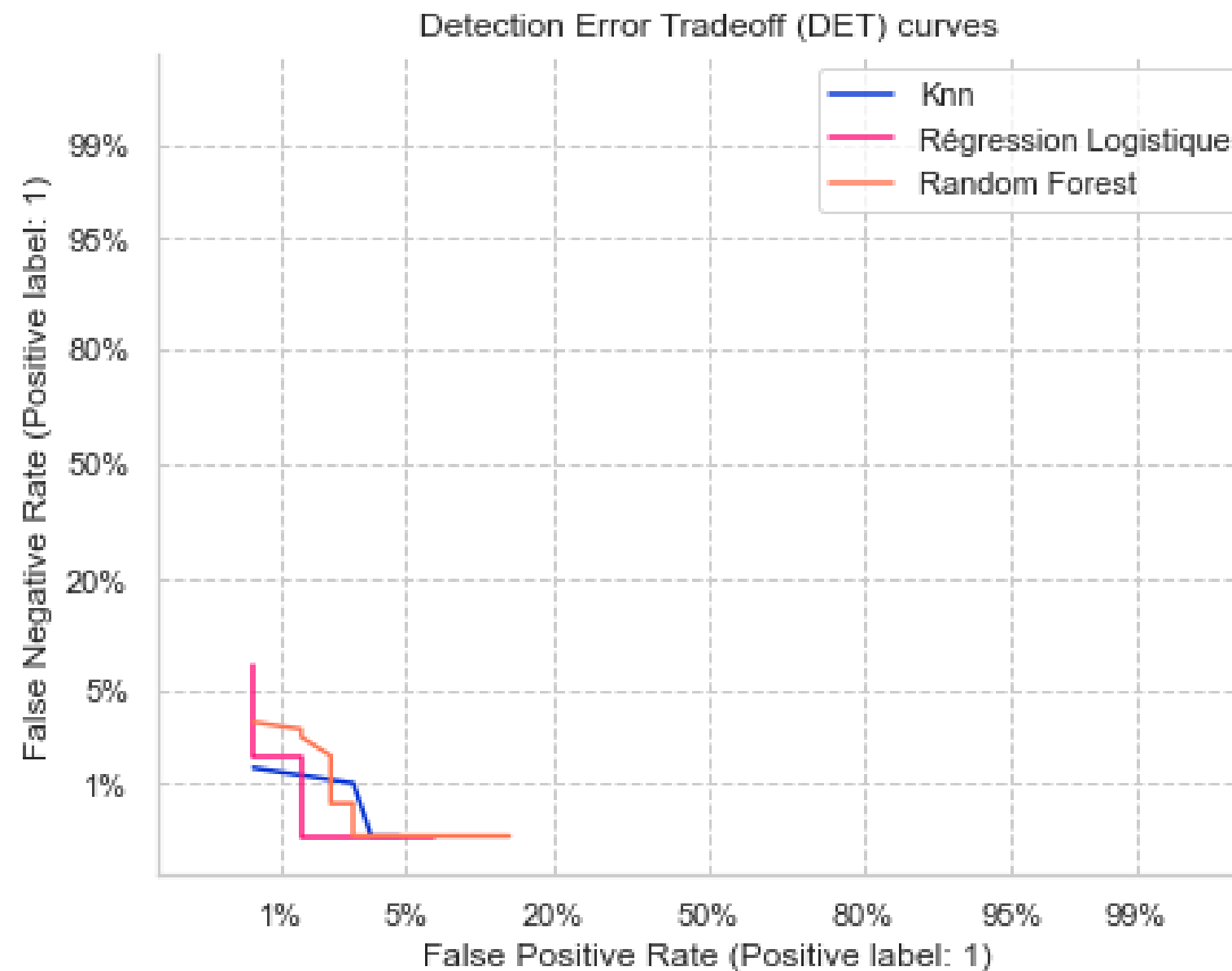
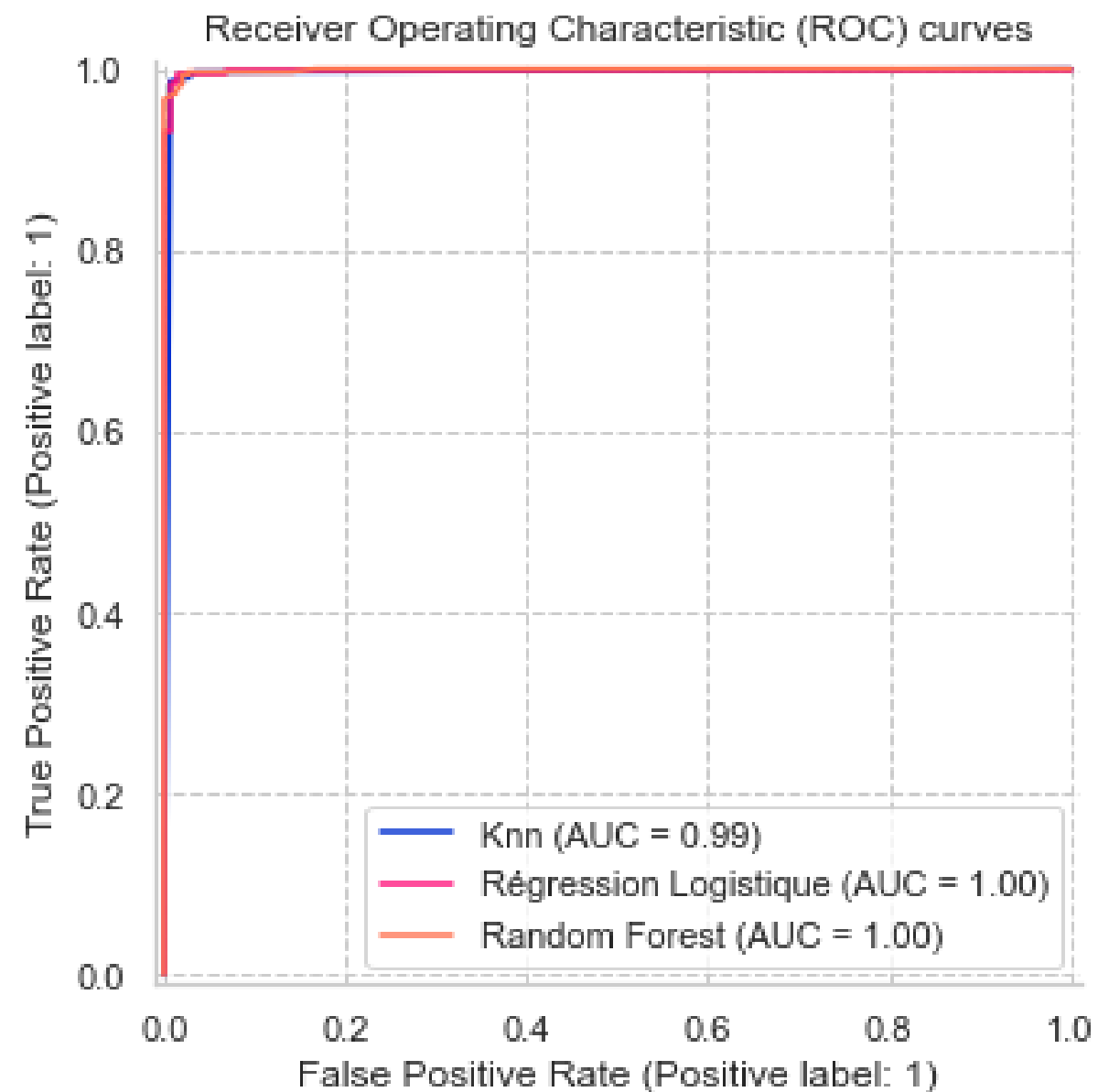
Modèle	Accuracy	F1-Score	Precision	Recall	Auc
Kmeans	0.984444	0.988353	0.986711	0.990000	0.981667
Knn	0.984444	0.988353	0.986711	0.990000	0.981667
Régression Logistique	0.993333	0.995008	0.993355	0.996667	0.991667
Random Forest	0.986667	0.990000	0.990000	0.990000	0.985000

Performances & scalabilité





Logit sélectionné



Seuil : 0,50

FNR : 0.0033

FPR : 0.0133