Projet 10 DataV2 Detection de faux billets

Barrios Mathieu

Sommaire

- Analyse du datasets
- Régression linéaire pour trouvé les Nan
- Choix du modéle de classification

Analyse du dataset

- 1500 billets
- 1463 non null dans margin_low
  Donc 37 nan

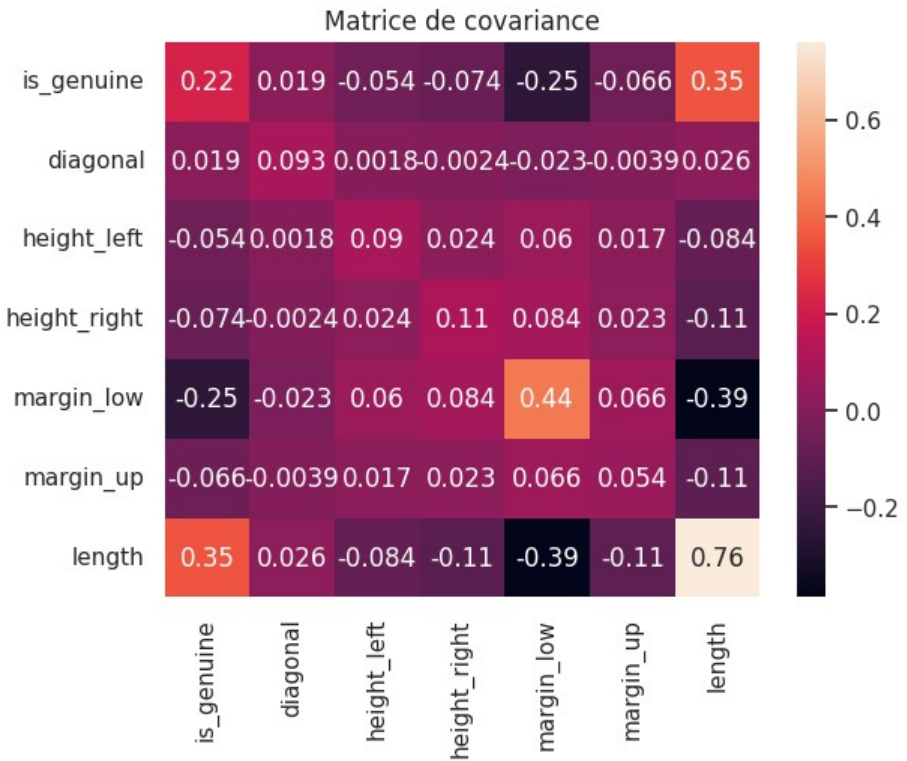
  is_genuine en booléen

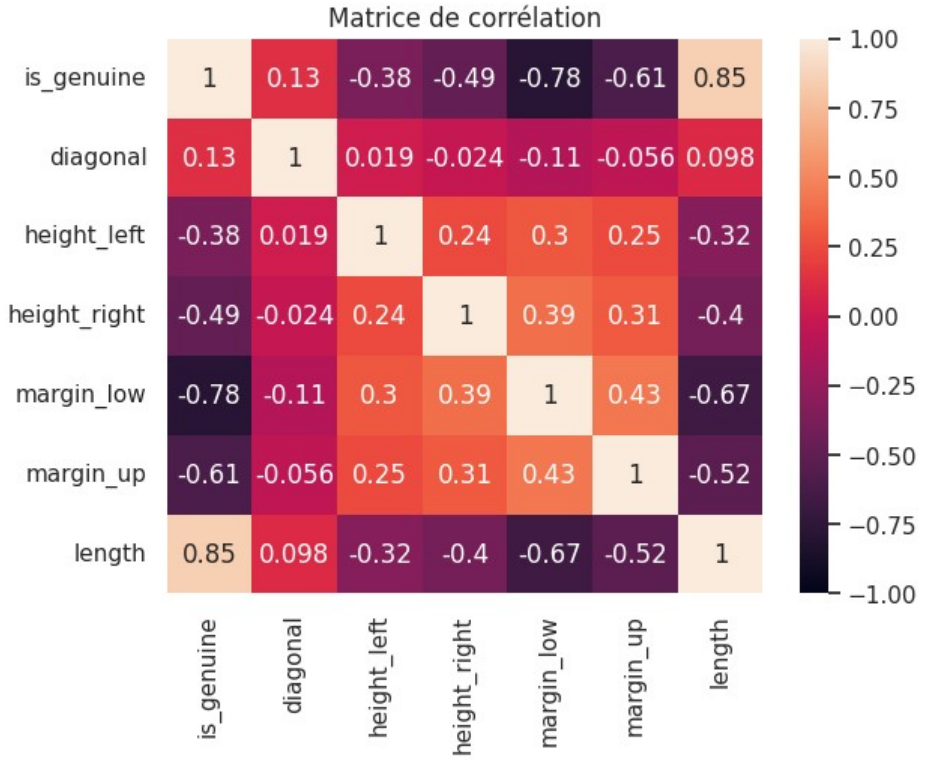
  Le reste en float.

```
################################## INFO  ###
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   is_genuine    1500 non-null   bool
 1   diagonal      1500 non-null   float64
 2   height_left   1500 non-null   float64
 3   height_right  1500 non-null   float64
 4   margin_low    1463 non-null   float64
 5   margin_up     1500 non-null   float64
 6   length        1500 non-null   float64
dtypes: bool(1), float64(6)
memory usage: 71.9 KB
```

```
Dans les données il y a  1000 vrais billets et 500 faux billets
il y a donc  33.33 % de faux billets dans le fichier exemple
il y a donc  66.67 % de vrais billets dans le fichier exemple
```

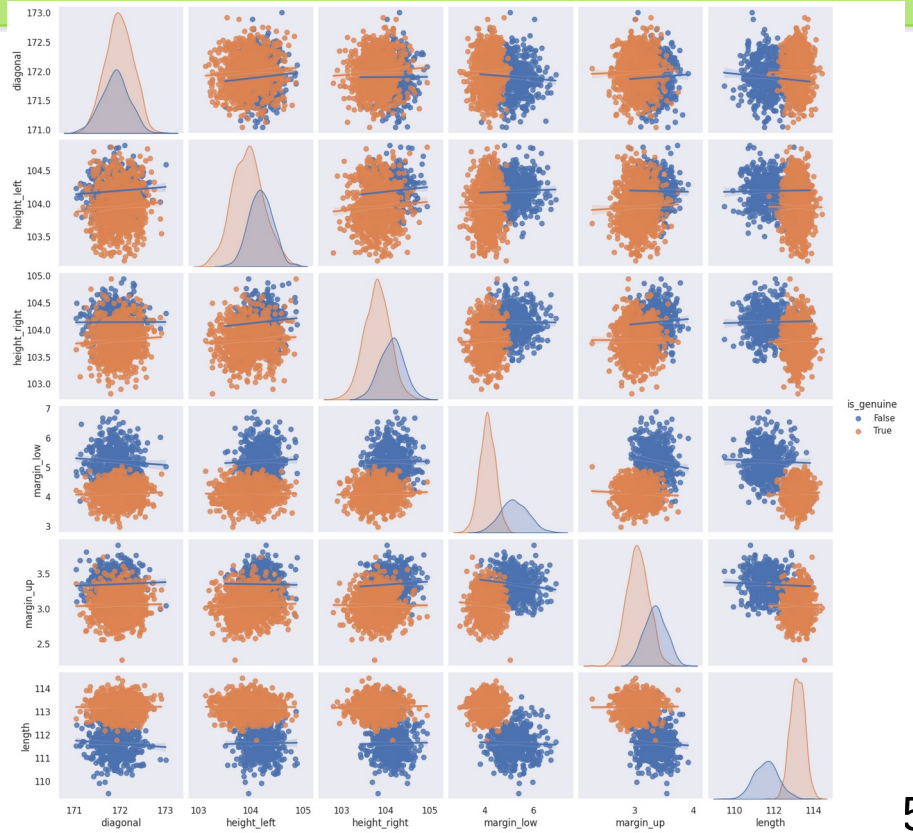
Analyse du dataset





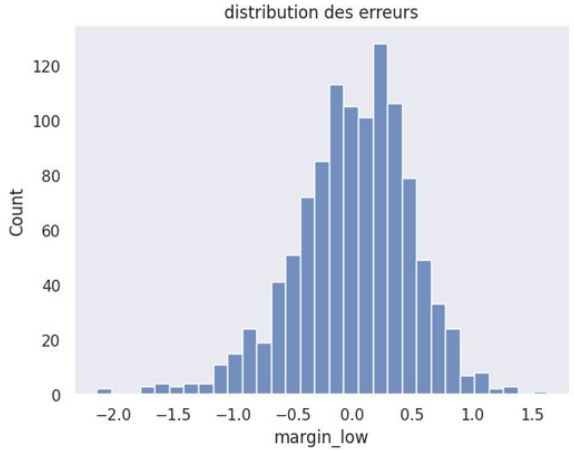
Sommaire

- Les distributions ont l’air normales
- Les distribution length, margin_low ne se supperposent pas trop.
- Margin_low: la forme des nuages de point ne forment pas une droite.



Régression Linéaire sur margin_low : regression simple

```
intercept (const dans statmodels) 60.29860050536974
Coefficients:
 [-0.49535341]
Erreur des moindres carrés train : 0.24
Coefficient de determination train : 0.43
erreur max train:  2.137264358995428
score train : 0.434071142898532 44
score test : 0.4720533533553434
```




```
test de Normalité des erreurs

Interprétation du test:
H0 : La série suit une loi Normale
H1 : La série ne suit pas une loi Normale
Étant donné que la p-values est inférieure au niveau de signification alpha = 0.05 ,
on doit rejeter l'hypothèse nulle H0 et retenir l'hypothèse H1.
Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inferieur à  0.0 %
```

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.434
Model:	OLS	Adj. R-squared:	0.434
Method:	Least Squares	F-statistic:	839.9
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	1.58e-137
Time:	15:11:25	Log-Likelihood:	-780.54
No. Observations:	1097	AIC:	1565.
Df Residuals:	1095	BIC:	1575.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	60.2986	1.926	31.307	0.000	56.519	64.078
length	-0.4954	0.017	-28.981	0.000	-0.529	-0.462

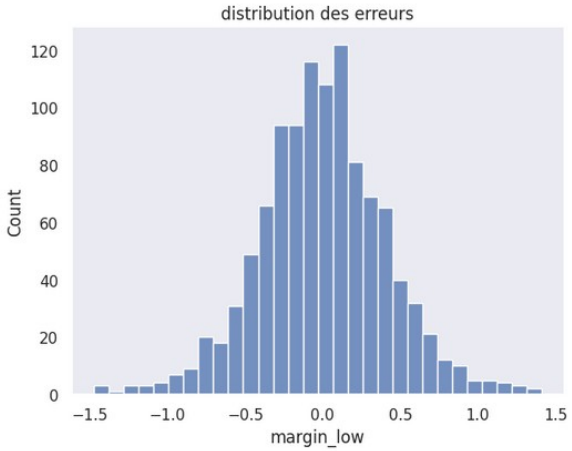
Omnibus:	67.128	Durbin-Watson:	2.026
Prob(Omnibus):	0.000	Jarque-Bera (JB):	91.326
Skew:	0.531	Prob(JB):	1.48e-20
Kurtosis:	3.932	Cond. No.	1.46e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.46e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Sans grande surprise, la régression simple ne donne pas de bon résultats.

Régression Linéaire sur margin_low : regression multiple

```
intercept (const dans statmodels) -3.245772764801453
Coefficients:
 [-1.213588    0.01031642  0.00276712  0.04723002 -0.25875826  0.02123901]
Erreur des moindres carrés train : 0.17
Coefficient de determination train : 0.62
erreur max train:  1.474356447418458
score train : 0.6236750185782756
score test : 0.5897599623896639
```




```
test de Normalité des erreurs

Interprétation du test:
H0 : La série suit une loi Normale
H1 : La série ne suit pas une loi Normale
Étant donné que la p-values est inférieure au niveau de signification alpha = 0.05 ,
on doit rejeter l'hypothèse nulle H0 et retenir l'hypothèse H1.
Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inferieur à  0.01 %
```

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.624
Model:	OLS	Adj. R-squared:	0.622
Method:	Least Squares	F-statistic:	301.1
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	2.80e-227
Time:	15:11:25	Log-Likelihood:	-580.39
No. Observations:	1097	AIC:	1175.
Df Residuals:	1090	BIC:	1210.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.2458	9.769	-0.332	0.740	-22.415	15.923
is_genuine	-1.2136	0.057	-21.200	0.000	-1.326	-1.101
diagonal	0.0103	0.041	0.249	0.804	-0.071	0.092
height_left	0.0028	0.044	0.063	0.950	-0.084	0.090
height_right	0.0472	0.045	1.057	0.291	-0.040	0.135
margin_up	-0.2588	0.068	-3.806	0.000	-0.392	-0.125
length	0.0212	0.027	0.784	0.433	-0.032	0.074

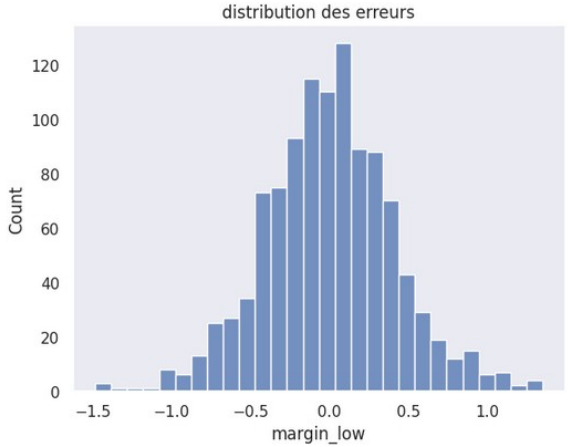
Omnibus:	15.370	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.022
Skew:	-0.007	Prob(JB):	2.24e-06
Kurtosis:	3.754	Cond. No.	1.99e+05

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.99e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

La régression multiple offre de meilleurs résultats.

Régression Linéaire sur margin_low : regression multiple

```
intercept (const dans statmodels) 5.9959161450569685
Coefficients:
 [-1.1552154  -0.23786822]
Erreur des moindres carrés train : 0.17
Coefficient de determination train : 0.61
erreur max train:  1.484291622539219
score train : 0.6070240343759 45
score test : 0.6426382405704203
```




```
test de Normalité des erreurs

Interprétation du test:
H0 : La série suit une loi Normale
H1 : La série ne suit pas une loi Normale
Étant donné que la p-values est inférieure au niveau de signification alpha = 0.05 ,
on doit rejeter l'hypothèse nulle H0 et retenir l'hypothèse H1.
Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inferieur à  0.11 %
```

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.607
Model:	OLS	Adj. R-squared:	0.606
Method:	Least Squares	F-statistic:	844.9
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	1.31e-222
Time:	15:11:25	Log-Likelihood:	-587.69
No. Observations:	1097	AIC:	1181.
Df Residuals:	1094	BIC:	1196.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9959	0.229	26.166	0.000	5.546	6.446
is_genuine	-1.1552	0.034	-34.205	0.000	-1.221	-1.089
margin_up	-0.2379	0.068	-3.500	0.000	-0.371	-0.105

Omnibus:	11.026	Durbin-Watson:	2.081
Prob(Omnibus):	0.004	Jarque-Bera (JB):	16.564
Skew:	-0.025	Prob(JB):	0.000253
Kurtosis:	3.600	Cond. No.	64.8

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Régression finale : pas de normalité des erreurs, R^2 relativement faible.

Classement : régression logistique

Données avec régression

rapport sur données d'entrainement :

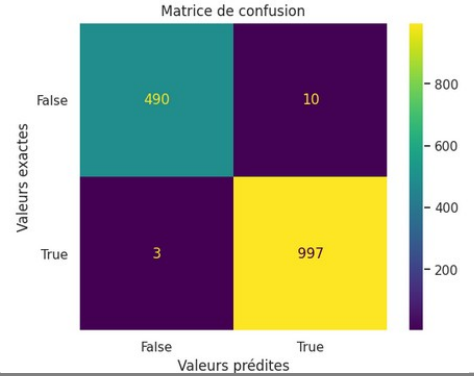
	precision	recall	f1-score	support
False	0.99	0.98	0.99	380
True	0.99	1.00	0.99	745
accuracy			0.99	1125
macro avg	0.99	0.99	0.99	1125
weighted avg	0.99	0.99	0.99	1125

rapport de classification jeu de test

	precision	recall	f1-score	support
False	0.99	0.98	0.99	500
True	0.99	1.00	0.99	1000
accuracy			0.99	1500
macro avg	0.99	0.99	0.99	1500
weighted avg	0.99	0.99	0.99	1500

rapport sur données de test:

	precision	recall	f1-score	support
False	1.00	0.98	0.99	120
True	0.99	1.00	1.00	255
accuracy			0.99	375
macro avg	1.00	0.99	0.99	375
weighted avg	0.99	0.99	0.99	375



Données sans régression

rapport sur données d'entrainement :

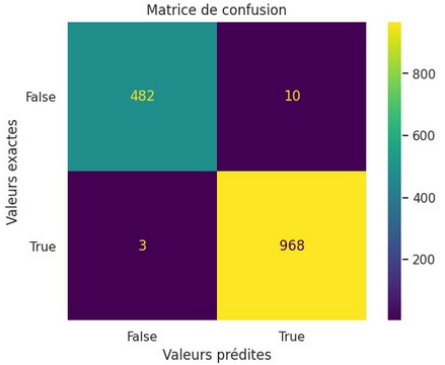
	precision	recall	f1-score	support
False	1.00	0.98	0.99	357
True	0.99	1.00	0.99	740
accuracy			0.99	1097
macro avg	0.99	0.99	0.99	1097
weighted avg	0.99	0.99	0.99	1097

rapport de classification jeu de test

	precision	recall	f1-score	support
False	0.99	0.98	0.99	492
True	0.99	1.00	0.99	971
accuracy			0.99	1463
macro avg	0.99	0.99	0.99	1463
weighted avg	0.99	0.99	0.99	1463

rapport sur données de test:

	precision	recall	f1-score	support
False	0.99	0.98	0.98	135
True	0.99	0.99	0.99	231
accuracy			0.99	366
macro avg	0.99	0.98	0.99	366
weighted avg	0.99	0.99	0.99	366



Classement : Kmeans

Données avec régression

rapport sur données d'entrainement :

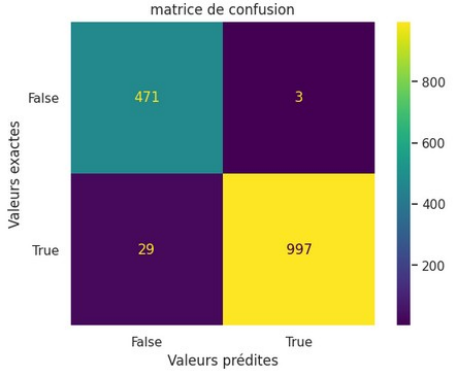
	precision	recall	f1-score	support
False	0.99	0.95	0.97	380
True	0.97	1.00	0.99	745
accuracy			0.98	1125
macro avg	0.98	0.97	0.98	1125
weighted avg	0.98	0.98	0.98	1125

rapport sur données totales

	precision	recall	f1-score	support
False	0.99	0.94	0.97	500
True	0.97	1.00	0.98	1000
accuracy			0.98	1500
macro avg	0.98	0.97	0.98	1500
weighted avg	0.98	0.98	0.98	1500

rapport sur données de test:

	precision	recall	f1-score	support
False	0.99	0.93	0.96	120
True	0.97	1.00	0.98	255
accuracy			0.97	375
macro avg	0.98	0.96	0.97	375
weighted avg	0.97	0.97	0.97	375



Données sans régression

rapport sur données d'entrainement :

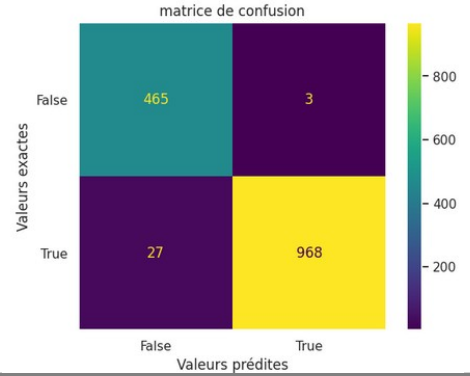
	precision	recall	f1-score	support
False	0.99	0.95	0.97	387
True	0.97	1.00	0.98	710
accuracy			0.98	1097
macro avg	0.98	0.97	0.98	1097
weighted avg	0.98	0.98	0.98	1097

rapport sur données totales

	precision	recall	f1-score	support
False	0.99	0.95	0.97	492
True	0.97	1.00	0.98	971
accuracy			0.98	1463
macro avg	0.98	0.97	0.98	1463
weighted avg	0.98	0.98	0.98	1463

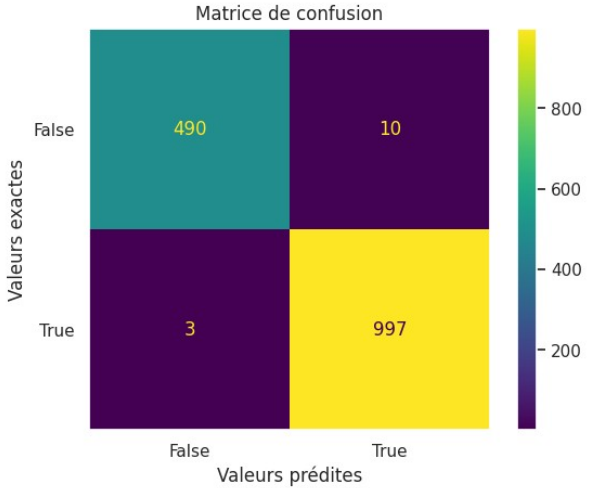
rapport sur données de test:

	precision	recall	f1-score	support
False	1.00	0.94	0.97	105
True	0.98	1.00	0.99	261
accuracy			0.98	366
macro avg	0.99	0.97	0.98	366
weighted avg	0.98	0.98	0.98	366



Choix du modéle

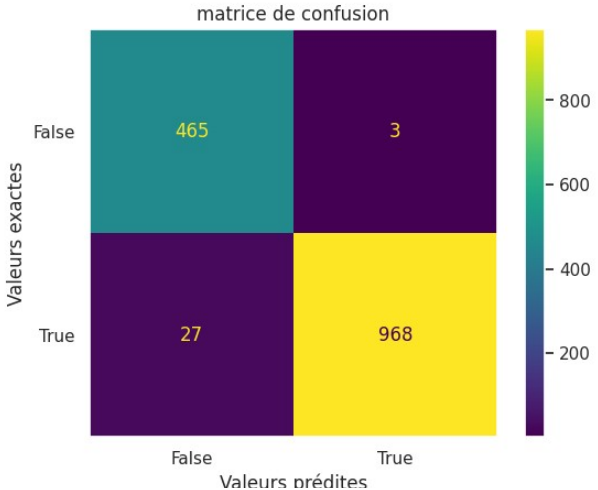
Régression logistique avec régression



rapport de classification jeu de test

	precision	recall	f1-score	support
False	0.99	0.98	0.99	500
True	0.99	1.00	0.99	1000
accuracy			0.99	1500
macro avg	0.99	0.99	0.99	1500
weighted avg	0.99	0.99	0.99	1500

Kmeans sans régression



rapport sur données totales

	precision	recall	f1-score	support
False	0.99	0.95	0.97	492
True	0.97	1.00	0.98	971
accuracy			0.98	1463
macro avg	0.98	0.97	0.98	1463
weighted avg	0.98	0.98	0.98	1463