DÉTECTION DE FAUX BILLETS

ONCFM

BRIEF

Dans le cadre de lutte contre du faux-monnayage, un algorithme de detection de faux billets est mis en place.

L'algorithme créé est basé uniquement sur l'information géométrique des 1.500 billets parmi lesquels 500 faux billets.

Les méthodes k-means et régression logistiques sont testées et la plus performante parmi elles est sélectionnées.

DONNÉES

Aperçu et nettoyage

APERÇU

variable cible

1.500 observations, 7 variables

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
0	True	171.81	104.86	104.95	4.52	2.89	112.83
1	True	171.46	103.36	103.66	3.77	2.99	113.09
2	True	172.69	104.48	103.50	4.40	2.94	113.16
3	True	171.36	103.91	103.94	3.62	3.01	113.51
4	True	171.73	104.28	103.46	4.04	3.48	112.54

variables explicatives

```
#   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
```

37 valeurs manquantes

RÉGRESSION LINÉAIRE

variable cible

Données d'entraînement

Données prédites

is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
True	171.81	104.86	104.95	4.52	2.89	112.83
True	171.46	103.36	103.66	3.77	2.99	113.09
True	172.69	104.48	103.50	4.40	2.94	113.16
False	171.75	104.38	104.17	4.42	3.09	111.28
False	172.19	104.63	104.44	5.27	3.37	110.97
False	171.80	104.01	104.12	5.51	3.36	111.95
True	171.94	103.89	103.45	NaN	3.25	112.79
True	171.93	104.07	104.18	NaN	3.14	113.08
True	172.07	103.80	104.38	NaN	3.02	112.93
False	171.57	104.27	104.44	NaN	3.21	111.87
False	172.17	104.49	103.76	NaN	2.93	111.21
False	172.08	104.15	104.17	NaN	3.40	112.29

Prédiction des valeurs manquantes

NORMALISATION

Faux billets

	margin_low
1000	4.78
1001	4.96
1002	4.97
1003	5.19
1004	5.60

y_train

Vrais billets

	margin_low
0	4.52
1	3.77
2	4.40
3	3.62
4	4.04

x_train

	diagonal	height_left	height_right	margin_up	length
0	1.240642	-1.063992	-0.859492	-0.231658	-0.377594
1	0.060943	-1.465436	0.577463	-1.236343	-0.556539
2	-1.020447	-0.216497	0.872223	0.661395	-0.670412
3	0.388637	0.630997	0.687998	-0.789817	0.582196
4	2.125416	0.274157	0.319548	-1.236343	0.142970

	diagonal	height_left	height_right	margin_up	length
0	-0.592458	3.014043	3.951689	-0.877055	-1.048097
1	-1.754297	-1.963341	-0.516407	-0.337474	-0.317635
2	2.328738	1.753105	-1.070590	-0.607264	-0.120972
3	-2.086251	-0.138301	0.453412	-0.229558	0.862342
4	-0.858021	1.089454	-1.209135	2.306471	-1.862843

x_test

	diagonal	height_left	height_right	margin_up	length
0	-1.389147	1.056272	2.185233	0.849603	-3.745187
1	-1.953469	1.421280	1.319323	0.633771	-2.284263
2	-1.322756	0.326255	0.453412	-0.175600	-5.374679
3	0.602577	1.786288	-0.170043	-0.661222	-5.599436
4	0.303818	0.658081	1.250050	1.874807	-2.565210

	diagonal	height_left	height_right	margin_up	length
0	-0.160917	-0.204666	-1.243772	1.065436	-1.160475
1	-0.194113	0.392620	1.284686	0.471897	-0.345730
2	0.270623	-0.503309	1.977414	-0.175600	-0.767150
3	-1.787492	-0.967865	-0.031497	3.061884	0.188069
4	-0.526067	0.624898	0.869049	-0.175600	-2.368547

DONNÉES FINALES

#	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	1500 non-null	float64
5	margin_up	1500 non-null	float64
6	length	1500 non-null	float64

- Prédiction ‘margin_low’ des faux billet

[5.3171461 , 5.34582653, 5.39540866, 5.41349514, 5.14499516, 5.20235225, 5.05640493, 5.27981473]

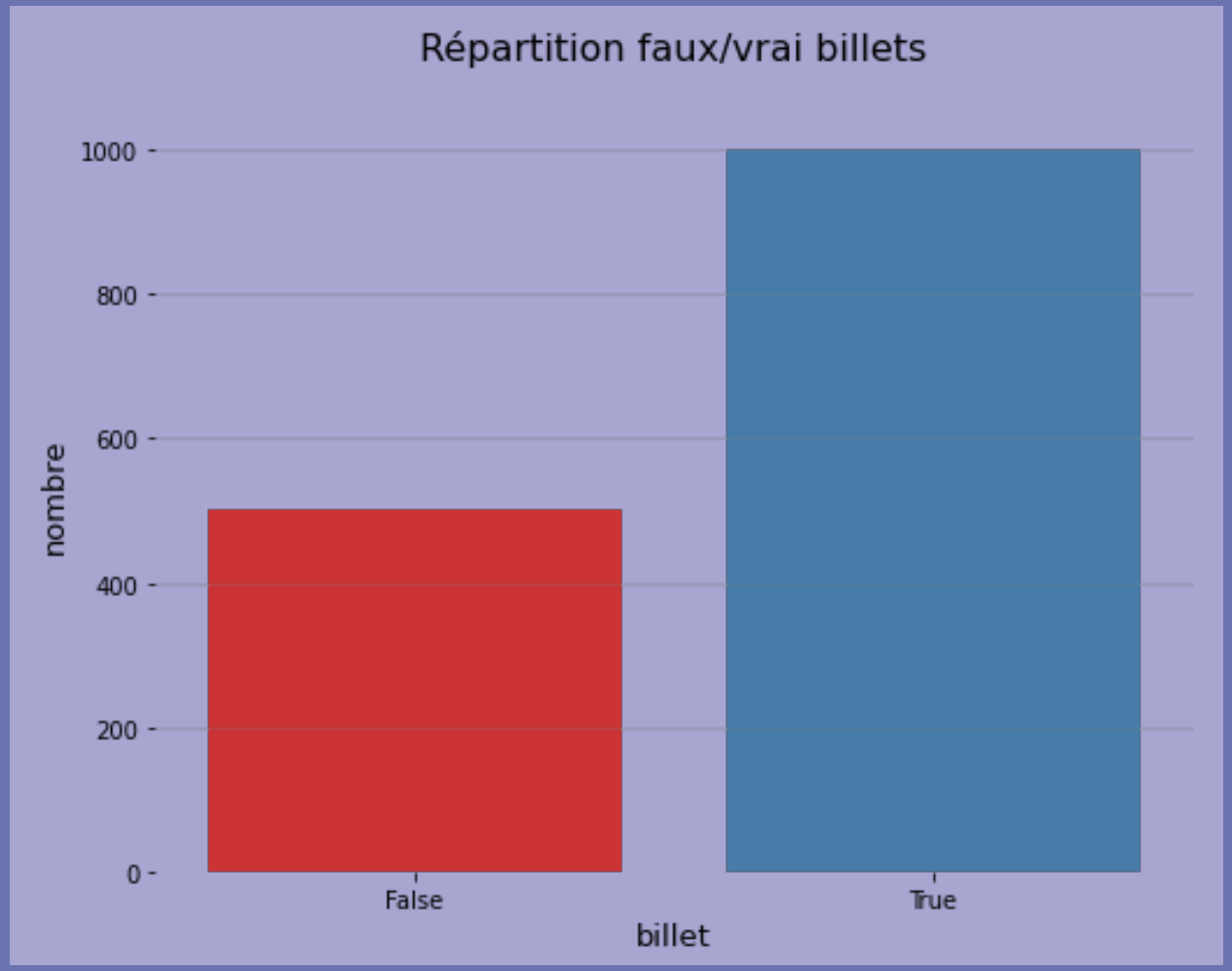
- Prédition ‘margin_low’ des billets authentiques

[4.06542749, 4.11404761, 4.13311635, 4.03501851, 4.09387695, 4.08159287, 4.09513499, 4.12611241, 4.11049199, 4.09544665, 4.10710194, 4.17684255, 4.14906315, 4.05685798, 4.12786563, 4.15924515, 4.09846281, 4.08486851, 4.1031968 , 4.13074695, 4.15539216, 4.13894295, 4.15805537, 4.09086845, 4.13689381, 4.16654165, 4.09741216, 4.09086659, 4.12789199]

EXPLORATION

Liens & Variabilité

RÉPARTITION DE CLASSES



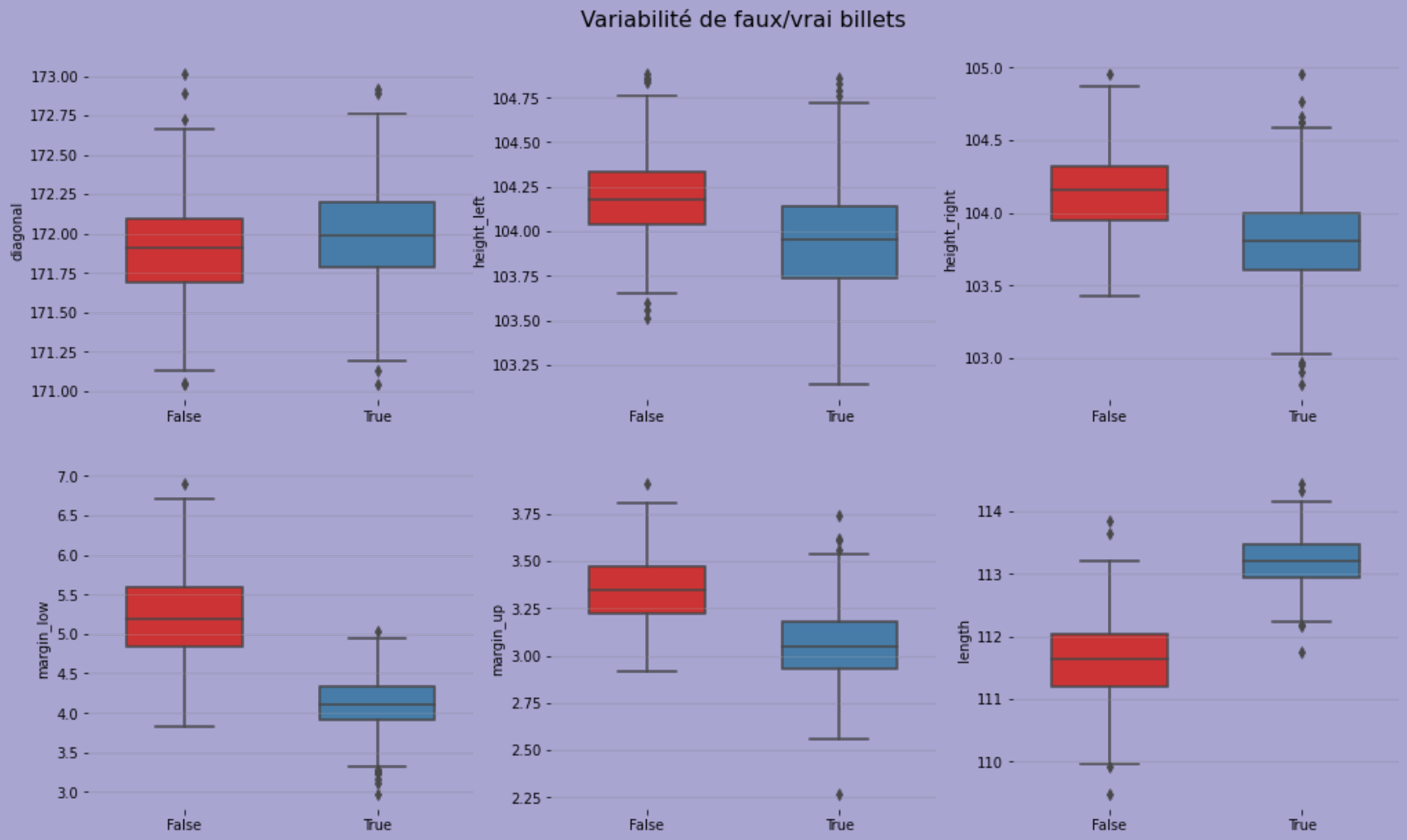
500 faux billets
1.000 billets authentiques

La quantité des billets authentiques est double des faux billets et ne nécessite pas de Oversampling.

VARIABILITÉ

La différence des deux billets est visible sur tous les variables. Tous les variables sont importants.

Surtout, les faux billets sont plus courts avec plus grandes marges que les vrais billets.

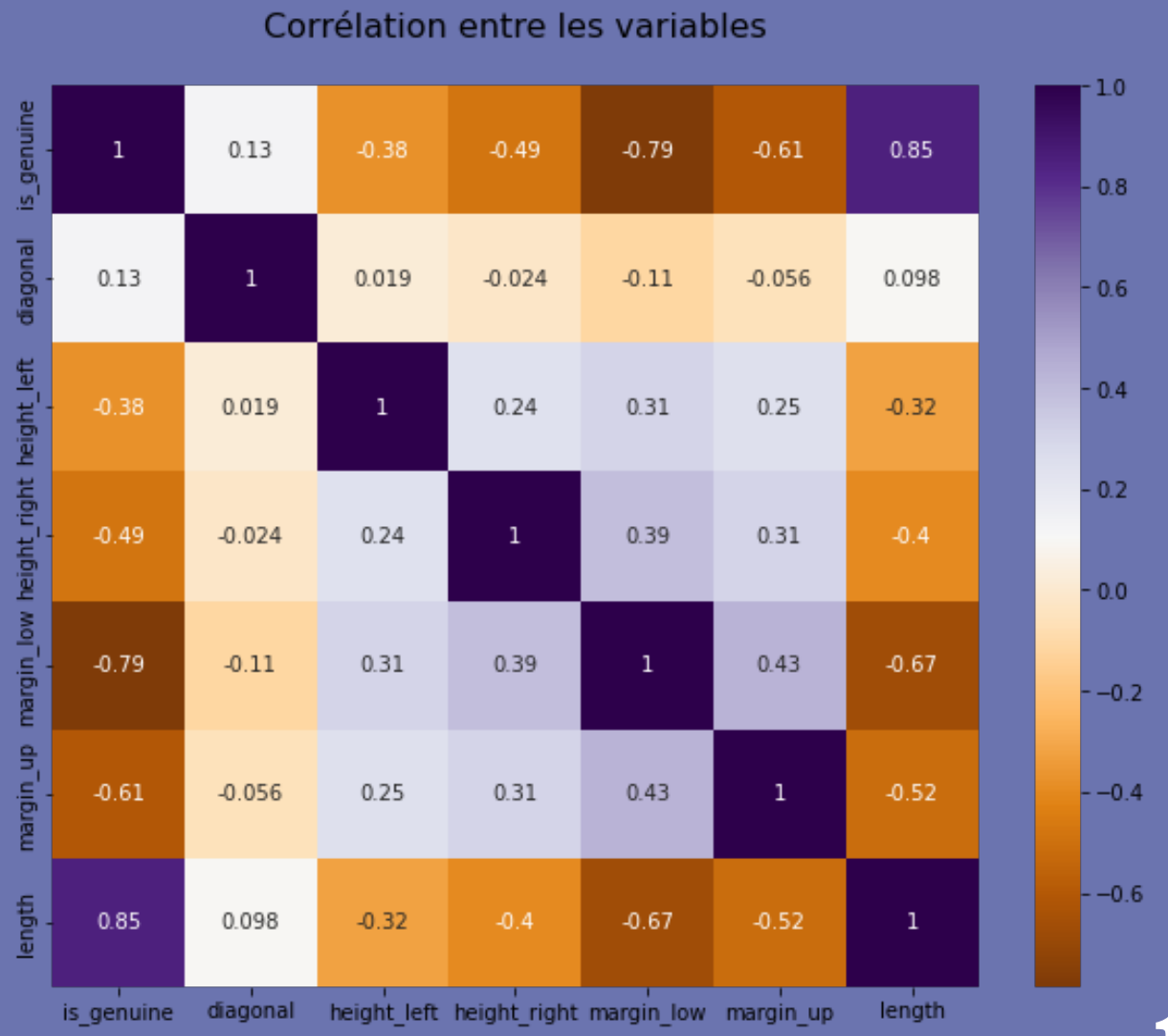


CORRÉLATION

Entre les variables explicatives, il existe des liens entre les marges et la longeur.

La variable cible semble avoir un lien avec la longeur, les marges et la taille.

ACP ne sera pas nécessarie car l'intensité de correlations et le nombre de variables ne sont pas importants.



CLASSIFICATION

K-means & Régression Logistique

NORMALISATION

	diagonal	height_left	height_right	margin_low	margin_up	length
count	1.500000e+03	1.500000e+03	1.500000e+03	1.463000e+03	1.500000e+03	1.500000e+03
mean	-7.849706e-14	4.815781e-14	-2.759974e-14	1.264842e-16	-4.056015e-16	1.707597e-15
std	1.000334e+00	1.000334e+00	1.000334e+00	1.000342e+00	1.000334e+00	1.000334e+00
min	-3.010357e+00	-2.971432e+00	-3.380166e+00	-2.269439e+00	-3.803785e+00	-3.654697e+00
25%	-6.832007e-01	-6.999333e-01	-6.460667e-01	-7.097308e-01	-6.967992e-01	-7.433186e-01
50%	5.113189e-03	3.496326e-02	-9.420867e-04	-2.651763e-01	-4.951040e-02	3.226587e-01
75%	6.934271e-01	6.696467e-01	7.056229e-01	5.787237e-01	6.840835e-01	7.582193e-01
max	3.446683e+00	2.840932e+00	3.163240e+00	3.637861e+00	3.273239e+00	2.019053e+00

Données centrées et réduites.

K-MEANS

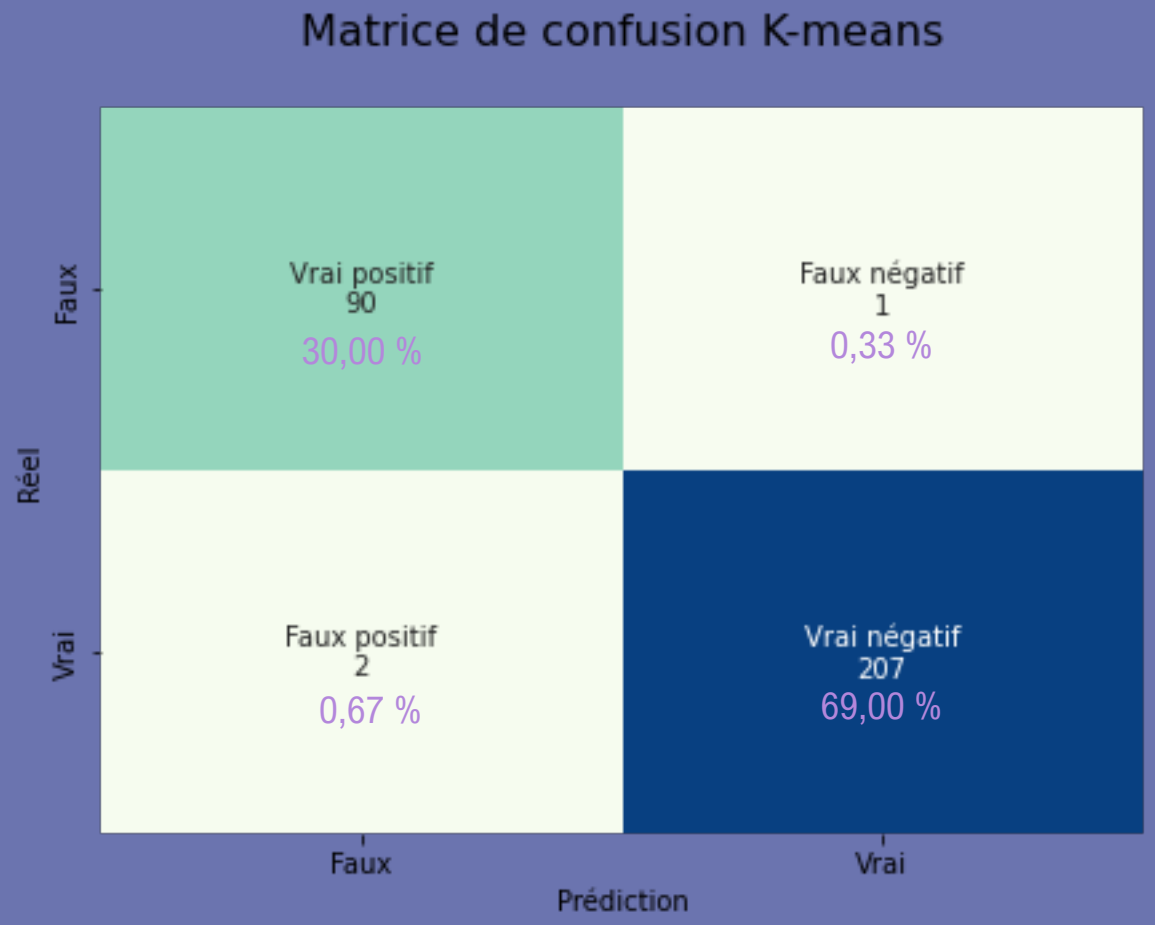
Classification non-supervisée

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length	cluster
0	True	171.81	104.86	104.95	4.52	2.89	112.83	0
1	True	171.46	103.36	103.66	3.77	2.99	113.09	1
2	True	172.69	104.48	103.50	4.40	2.94	113.16	1
3	True	171.36	103.91	103.94	3.62	3.01	113.51	1
4	True	171.73	104.28	103.46	4.04	3.48	112.54	1

Données entraînées

Prédiction

K-MEANS



99% des faux billets sont correctement détectés.
98 % de prediction positive sont correcte.

	precision	recall	f1-score	support
False	0.98	0.99	0.98	91
True	1.00	0.99	0.99	209
accuracy			0.99	300
macro avg	0.99	0.99	0.99	300
weighted avg	0.99	0.99	0.99	300

RÉGRESSION LOGISTIQUE

Classification supervisée
- 1.200 données entraînées
- 300 données testées

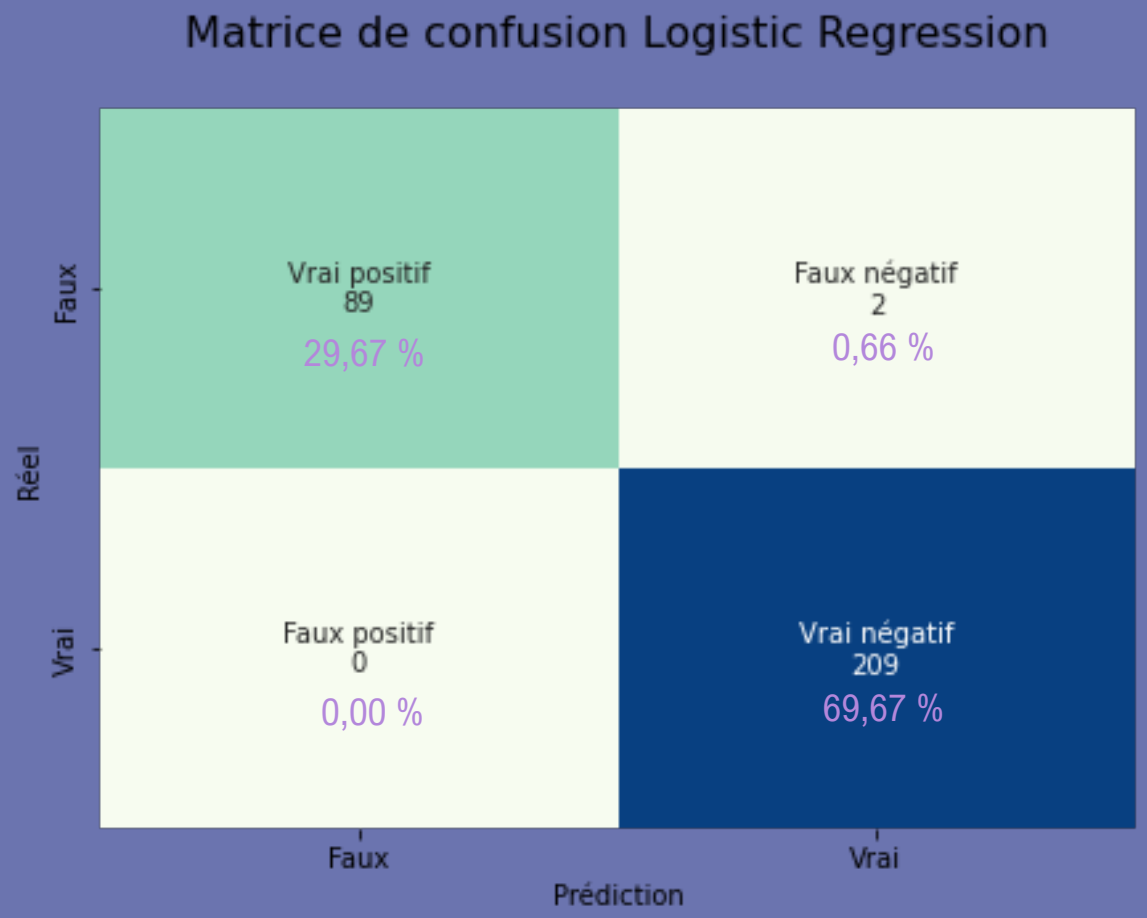
	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length	classe
0	True	171.81	104.86	104.95	4.52	2.89	112.83	True
1	True	171.46	103.36	103.66	3.77	2.99	113.09	True
2	True	172.69	104.48	103.50	4.40	2.94	113.16	True
3	True	171.36	103.91	103.94	3.62	3.01	113.51	True
4	True	171.73	104.28	103.46	4.04	3.48	112.54	True

Donnée cible

Données explicatives

Prédiction

RÉGRESSION LOGISTIQUE

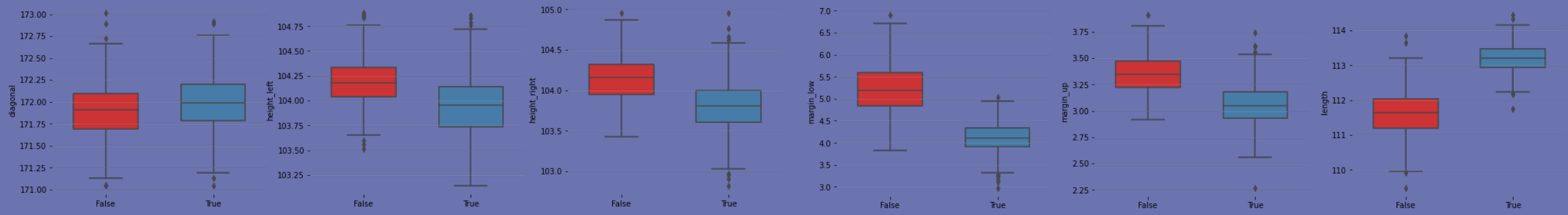


98% des faux billets sont correctement détectés
100 % de prediction positive sont correcte.

	precision	recall	f1-score	support
False	1.00	0.98	0.99	91
True	0.99	1.00	1.00	209
accuracy			0.99	300
macro avg	1.00	0.99	0.99	300
weighted avg	0.99	0.99	0.99	300

RÉGRESSION LOGISTIQUE

Variabilité des variables



La longeur et les marges ont plus de poids dans la classification

Coefficients des variables

diagonal	height_left	height_right	margin_low	margin_up	length
0.19	-0.32	-0.83	-2.73	-1.5	3.57

TEST DE L'ALGORITHME

PRÉDICTION

	diagonal	height_left	height_right	margin_low	margin_up	length	id	prediction
0	171.76	104.01	103.54	5.21	3.30	111.42	A_1	False
1	171.87	104.17	104.13	6.00	3.31	112.09	A_2	False
2	172.00	104.58	104.29	4.99	3.39	111.57	A_3	False
3	172.49	104.55	104.34	4.44	3.03	113.20	A_4	True
4	171.65	103.63	103.56	3.77	3.16	113.33	A_5	True

Parmi les 5 billets à tester, nous avons 2 billets authentiques et 3 faux billets.