Détecteur de billets contrefaits

Conception & tests initiaux

Sommaire

1 – Téléchargement & vérification des données
- 1.1 – Inspection
- 1.2 – Régression linéaire multiple
  - 1.2.1 – Régression & analyse des résidus
  - 1.2.2 – Valeurs atypiques & influentes
  - 1.2.3 – Régressions Ridge & Lasso
- 1.3 – Statistiques descriptives

2 – K-Means & optimisation
- 2.1 – Optimisation des paramètres
- 2.2 – Clustering & analyse des centroïdes
- 2.3 – Analyse en Composantes Principales
- 2.4 – Algorithme final : test & analyse de la performance

3 – Régression logistique & optimisation
- 3.1 – Choix des variables significatives
- 3.2 – Optimisation des hyperparamètres
- 3.3 - Algorithme final : test & analyse de la performance

ONCFM

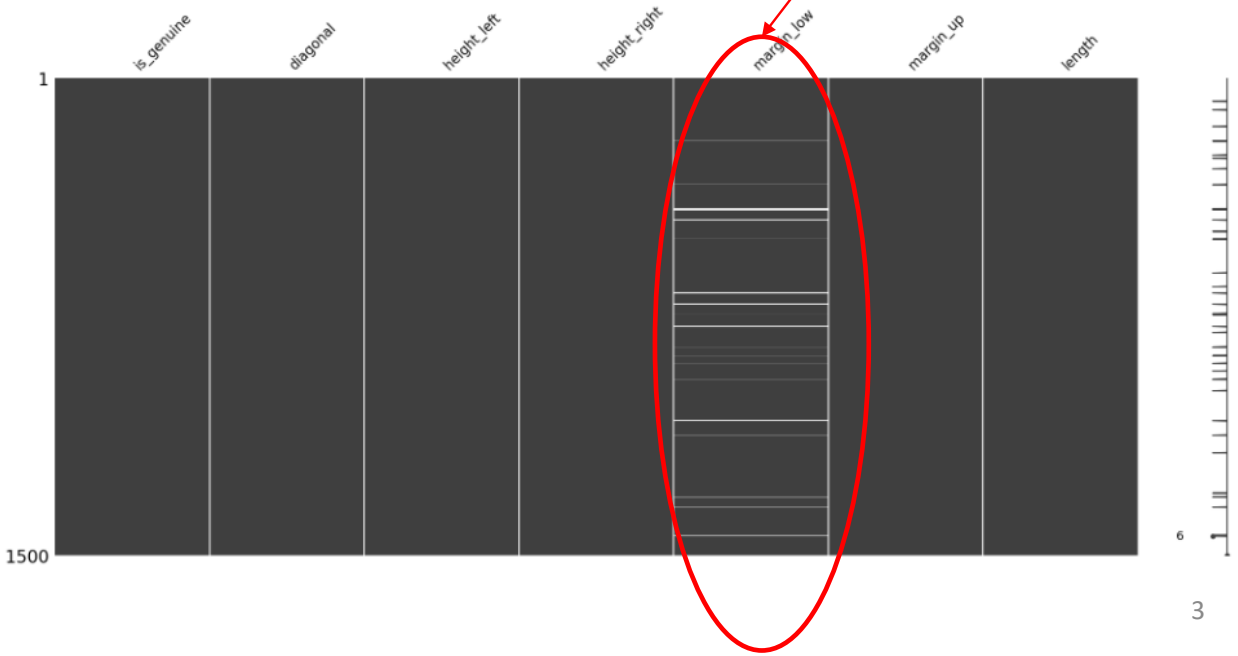
1 – Téléchargement & vérification des données

1.1 – Inspection

- Typage correct des données (1 booléenne et 6 quantitatives continues)
- Aucune valeur négative
- Pas de doublons
- Quelques outliers *a priori* atypiques et non aberrants
  - à analyser plus en detail
- 37 valeurs manquantes à compléter dans la colonne margin_low (29 vrais & 8 faux billets)
  - Régression linéaire multiple
- Pas de clé primaire
  - Index comme clé artificielle

```
#   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
```

	diagonal	height_left	height_right	margin_low	margin_up	length
count	1500.000000	1500.000000	1500.000000	1463.000000	1500.000000	1500.00000
mean	171.958440	104.029533	103.920307	4.485967	3.151473	112.67850
std	0.305195	0.299462	0.325627	0.663813	0.231813	0.87273
min	171.040000	103.140000	102.820000	2.980000	2.270000	109.49000
25%	171.750000	103.820000	103.710000	4.015000	2.990000	112.03000
50%	171.960000	104.040000	103.920000	4.310000	3.140000	112.96000
75%	172.170000	104.230000	104.150000	4.870000	3.310000	113.34000
max	173.010000	104.880000	104.950000	6.900000	3.910000	114.44000



ONCFM

1 – Téléchargement & vérification des données
1.2 – Régression linéaire multiple
1.2.1 – Régression…

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.477
Model:	OLS	Adj. R-squared:	0.476
Method:	Least Squares	F-statistic:	266.1
Date:	Tue, 20 Jun 2023	Prob (F-statistic):	2.60e-202
Time:	10:10:43	Log-Likelihood:	-1001.3
No. Observations:	1463	AIC:	2015.
Df Residuals:	1457	BIC:	2046.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	22.9948	9.656	2.382	0.017	4.055	41.935
diagonal	-0.1111	0.041	-2.680	0.007	-0.192	-0.030
height_left	0.1841	0.045	4.113	0.000	0.096	0.272
height_right	0.2571	0.043	5.978	0.000	0.173	0.342
margin_up	0.2562	0.064	3.980	0.000	0.130	0.382
length	-0.4091	0.018	-22.627	0.000	-0.445	-0.374

Omnibus:	73.627	Durbin-Watson:	1.893
Prob(Omnibus):	0.000	Jarque-Bera (JB):	95.862
Skew:	0.482	Prob(JB):	1.53e-21
Kurtosis:	3.801	Cond. No.	1.94e+05

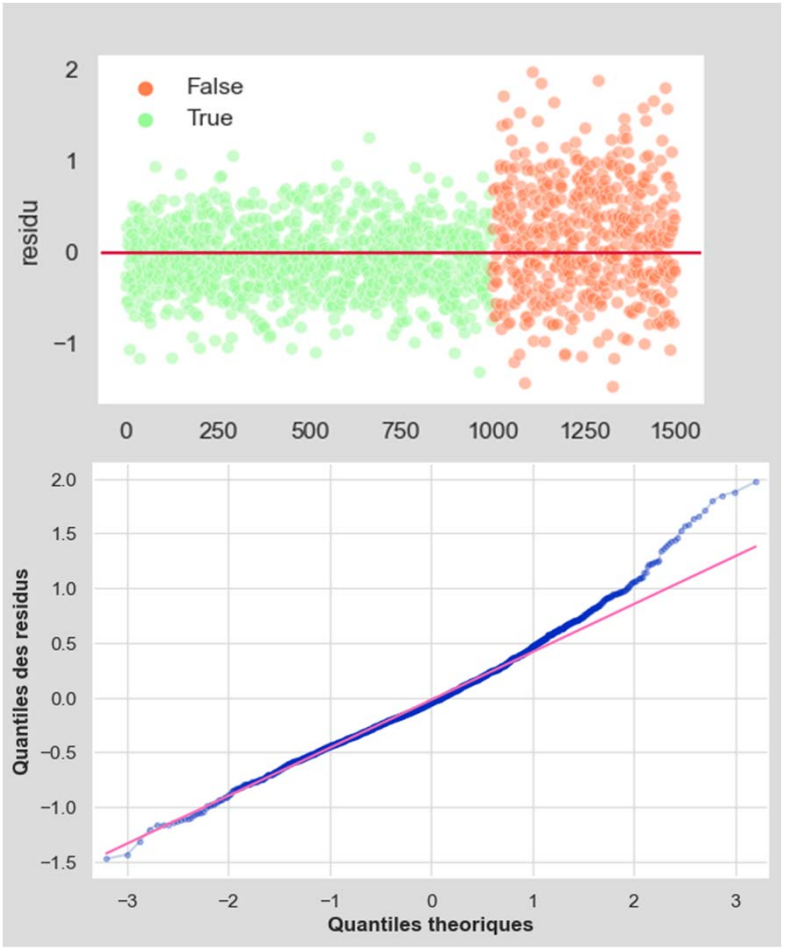
- qualité de prévision médiocre
- très proche de zéro ⇨ modèle significatif
- Toutes les variables sont significatives au seuil α=5%
- problème de colinéarité potentiel
- VIFs < 10 ⇨ ok

variable	vi_factor
diagonal	1.01
height_left	1.14
height_right	1.23
margin_up	1.40
length	1.58

ONCFM

1 – Téléchargement & vérification des données
1.2 – Régression linéaire multiple
1.2.1 – … & analyse des résidus

- Moyenne très proche de zéro mais…
- … hétéroscédastiques (tests de White et Breusch-Pagan conduisent au rejet de H_0 : les variances des résidus ne sont pas constantes) et…
- …non-Gaussiens (tests de Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov conduisent tous au rejet de H_0 : les résidus ne suivent pas une loi Normale)
  - résultats confirmés par le QQ plot et la droite de Henry
  - résidus plus dispersés sur les faux billets



ONCFM

1 – Téléchargement & vérification des données

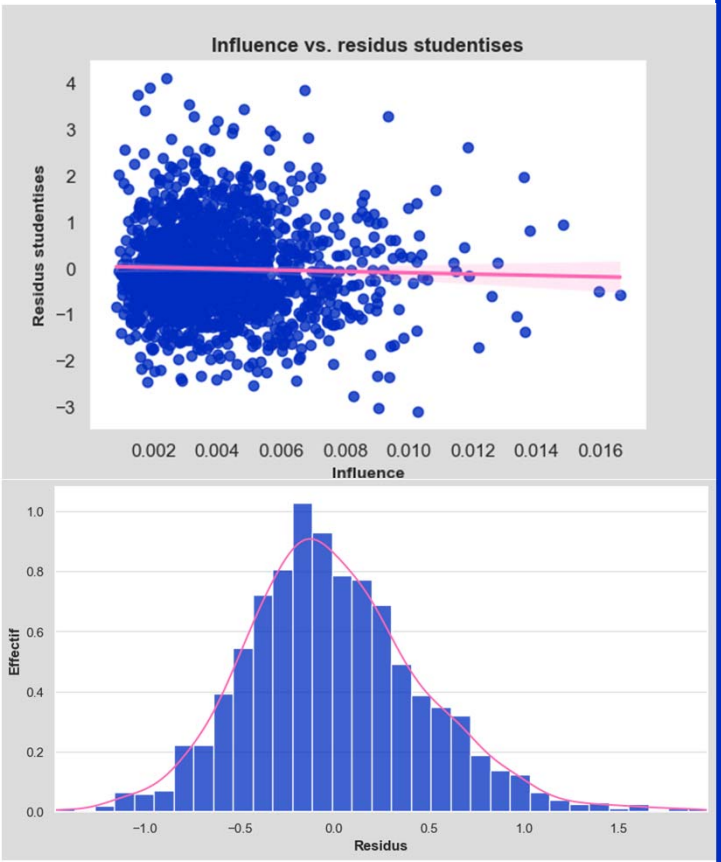
1.2 – Régression linéaire multiple

1.2.2 – Valeurs atypiques & influentes

- L'analyse des résidus studentisés fait apparaître 7 individus atypiques et influents
- Les résultats de la régression linéaire sans ces individus varient peu ⟹ on les conserve donc

1.2.3 – Régressions Ridge & Lasso

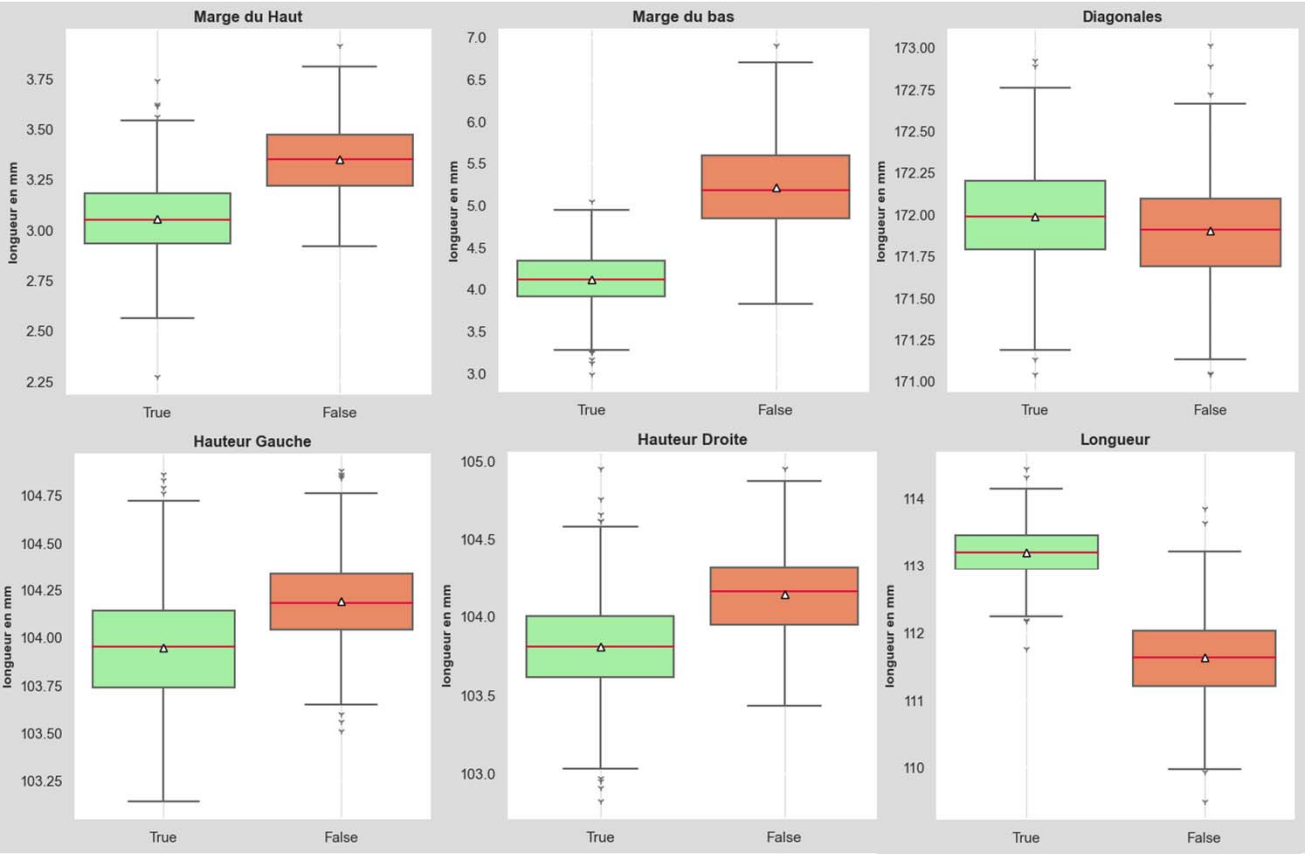
- N'ont pas permis de corriger les problèmes sur les résidus ni d'améliorer significativement le R2
  - Poursuite de l'analyse
    - moyenne suffisamment proche de zéro
    - échantillon suffisamment grand (1463)
    - courbe "relativement en cloche"
    - paramètres significatifs avec α=5%
    - modèle significatif avec α=5% (proba. F-statistic)

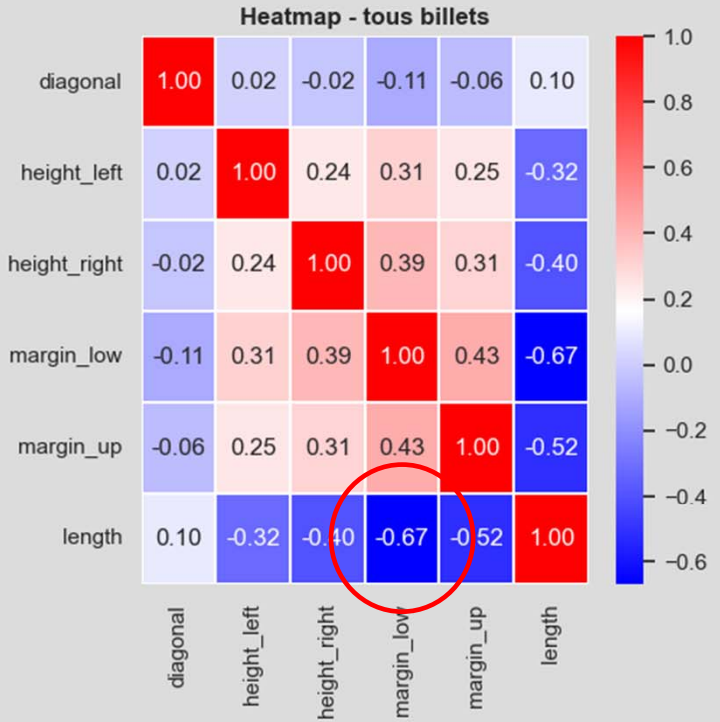


ONCFM

1 – Téléchargement & vérification des données

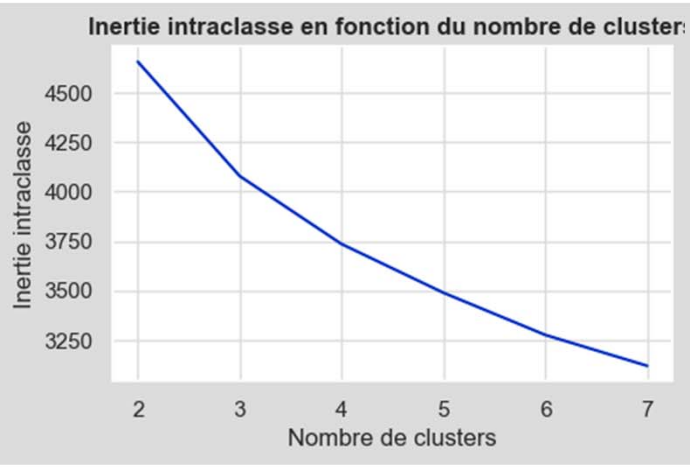
1.3 – Statistiques descriptives

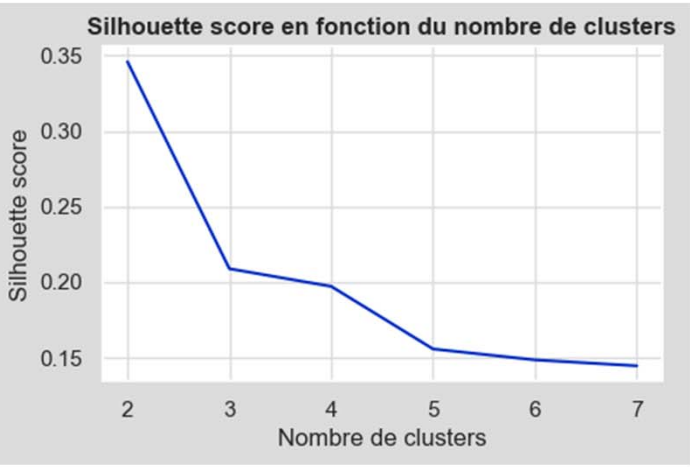


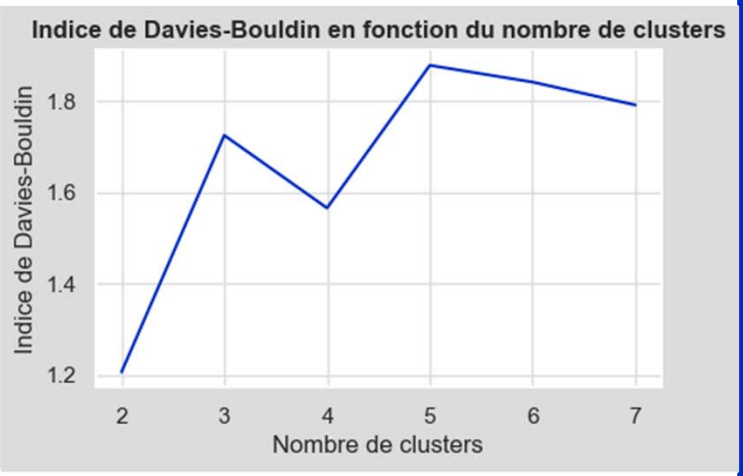


2 – K-Means & optimisation

2.1 – Optimisation des paramètres sur le training set







- Séparation training/testing set - 80/20
- K-means testé sur 2, 3 et 4 clusters
- Algorithme Kmeans ++, modèle elkan, initialisation du random state

- Indicateurs de performance:

	2 clusters	3 clusters	4 clusters
Rand score	0.9688	0.972	0.9591
Adjusted Rand Score	0.9366	0.9431	0.9168
MIB score	0.874	0.888	0.8509
Adjusted MIB score	0.8739	0.8879	0.8508
Score de Fowlkes-Mallows	0.9723	0.9752	0.964

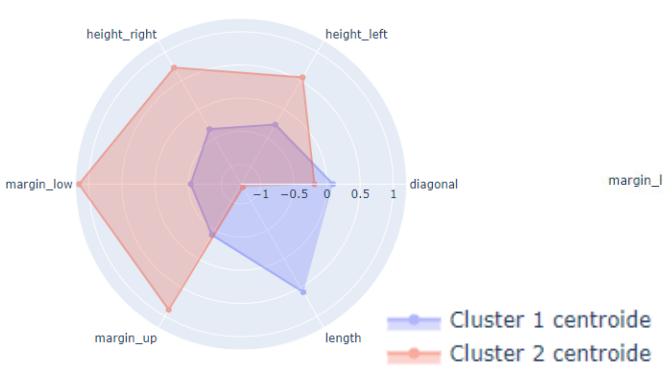
  - Modèle à 3 clusters renvoie les meilleures performances

ONCFM

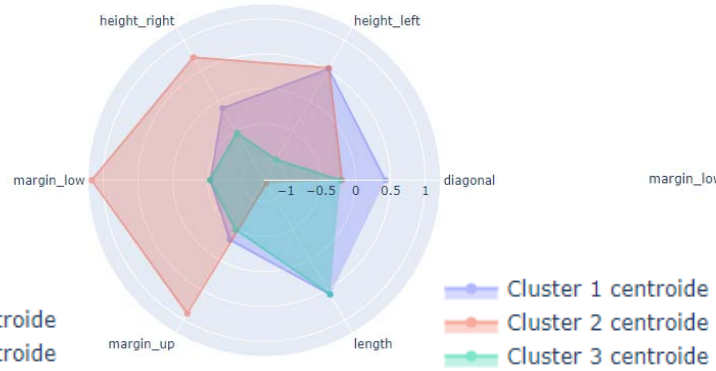
2 – K-Means & optimisation

2.2 – Clustering & analyse des centroïdes sur le training set

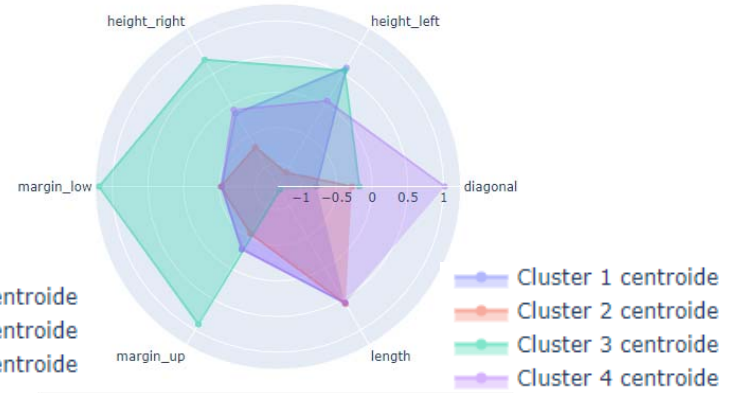
❖ 2 clusters:

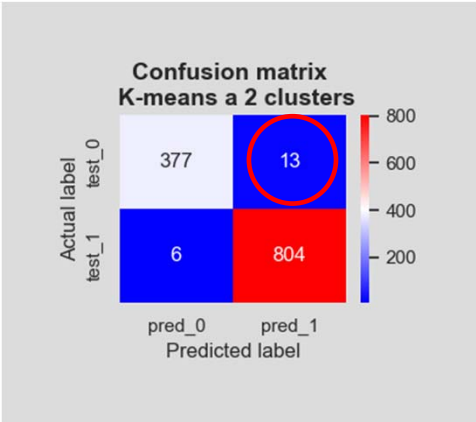


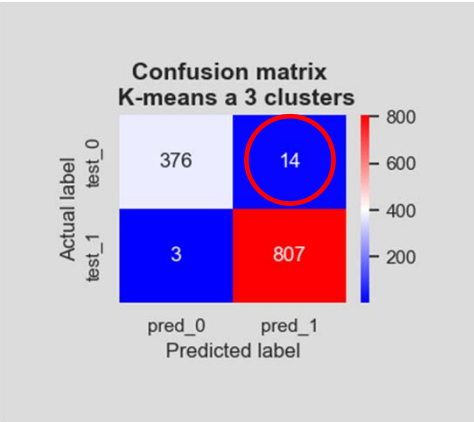
❖ 3 clusters:

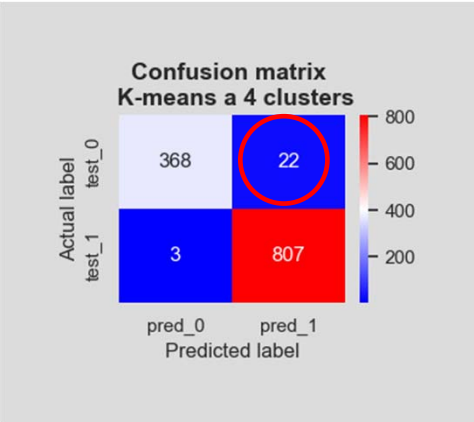


❖ 4 clusters:







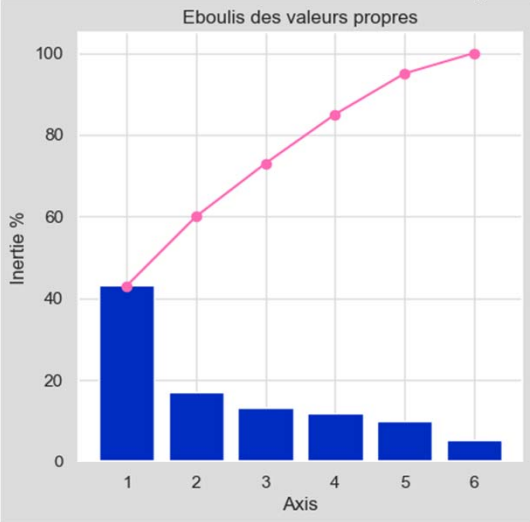


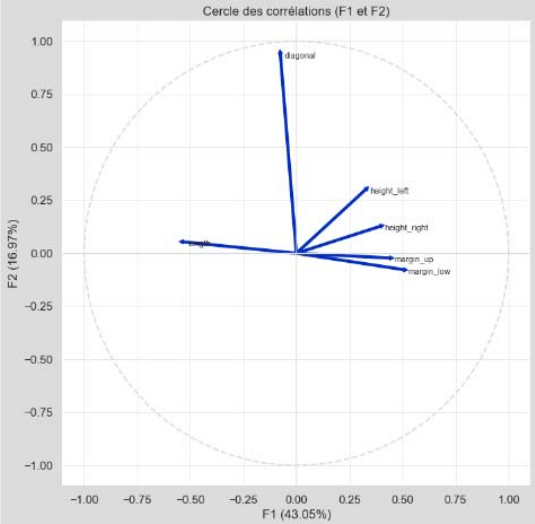
➢ Modèle à 3 clusters présente le meilleur compromis de performances

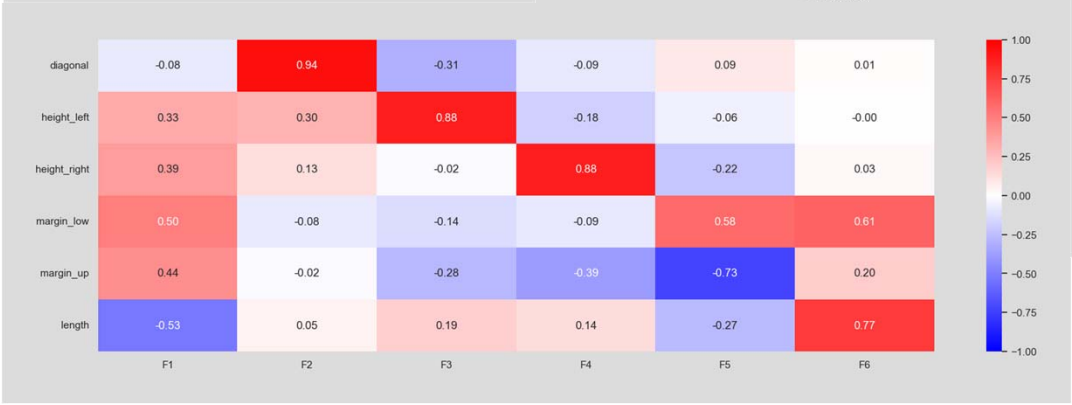
ONCFM

2 – K-Means & optimisation

2.3 – Analyse en Composantes Principales





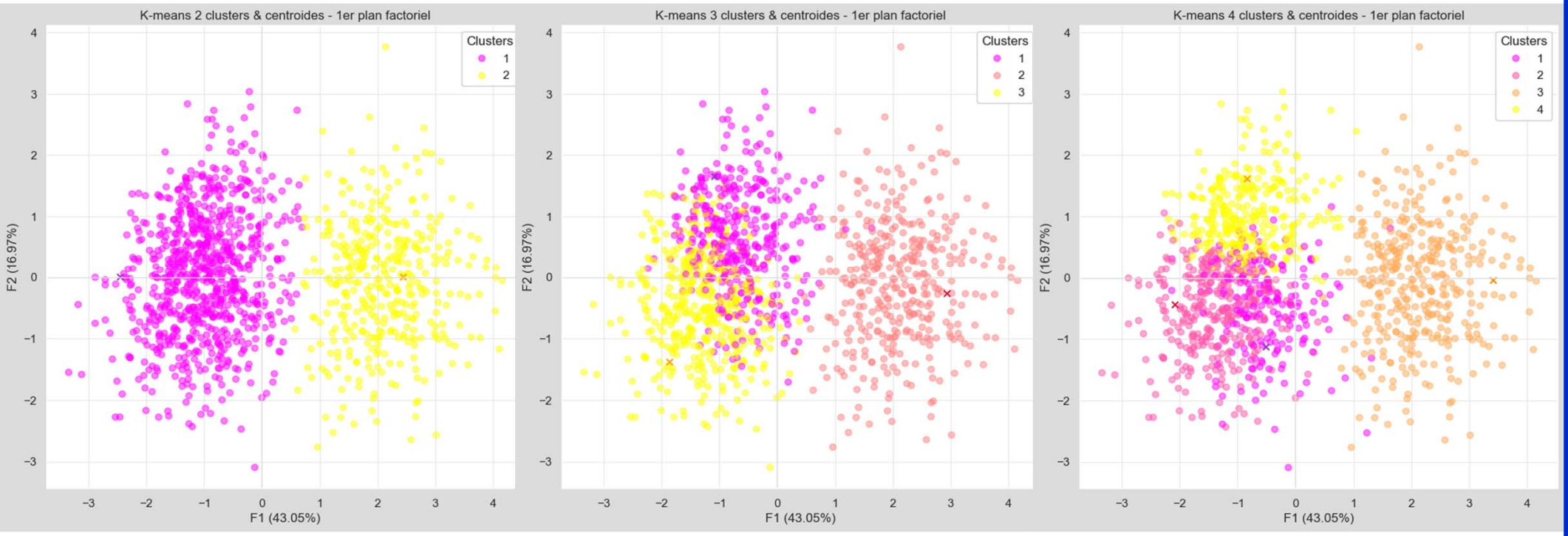


- >60% de la variance capturée avec 2 axes
- 1er axe dominé par longueur, marges & hauteur droite

ONCFM

2 – K-Means & optimisation

2.3 – Analyse en Composantes Principales



- Subdivision du cluster 1 des vrais billets à gauche bien visible ⇨ confirme l'analyse des centroïdes
- Résultat utilisé pour allouer automatiquement les étiquettes aux numéros de clusters

ONCFM

2 – K-Means & optimisation

2.4 – Algorithme final : test & analyse de la performance

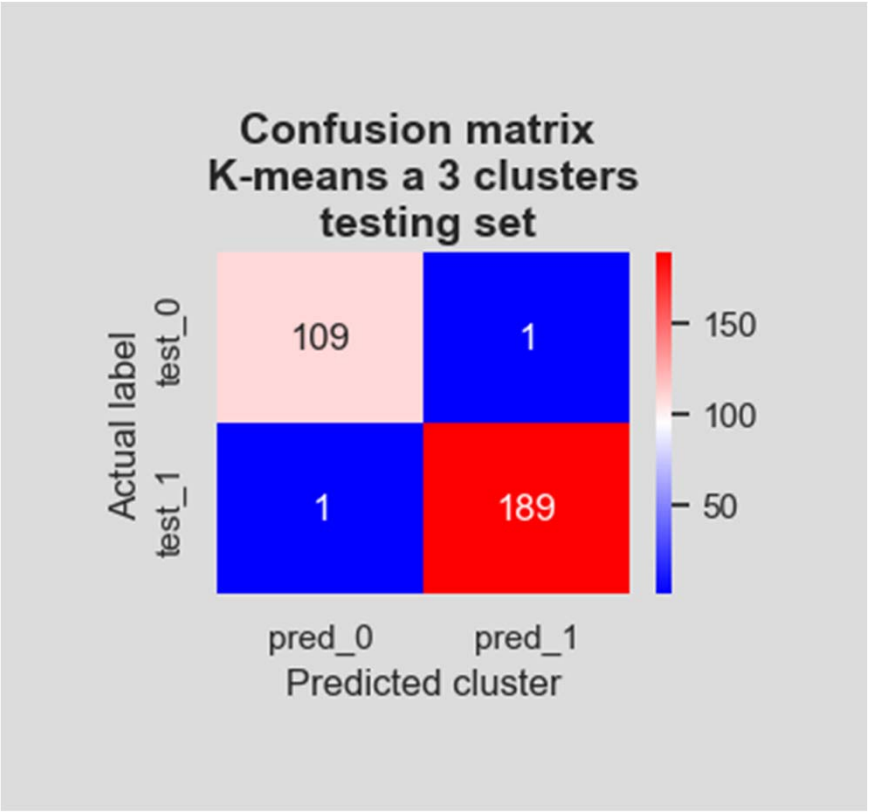
❖ Résultats sur le jeu de test :

3 clusters	Testing set	Training set
Rand score	0.9867	0.972
Adjusted Rand Score	0.9732	0.9431
MIB score	0.9394	0.888
Adjusted MIB score	0.9392	0.8879
Score de Fowlkes-Mallows	0.9875	0.9752

➢ légèrement meilleurs que sur le training set

❖ Résultats du test sur l'échantillon de production non étiqueté :

```
Le billet ref. A_1 est un faux billet.
Le billet ref. A_2 est un faux billet.
Le billet ref. A_3 est un faux billet.
Le billet ref. A_4 est un vrai billet.
Le billet ref. A_5 est un vrai billet.
```



ONCFM

3 – Régression logistique & optimisation

3.1 – Choix des variables significatives

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	1200
Model:	GLM	Df Residuals:	1193
Model Family:	Binomial	Df Model:	6
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-36.650
Date:	Tue, 20 Jun 2023	Deviance:	73.299
Time:	10:10:54	Pearson chi2:	1.96e+03
No. Iterations:	10	Pseudo R-squ. (CS):	0.6988
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	
const	-81.6224	257.021	-0.318	0.751	-
diagonal	-0.4097	1.150	-0.356	0.722	
height_left	-1.8585	1.276	-1.457	0.145	
height_right	-2.2433	1.093	-2.052	0.040	
margin_low	-5.3109	0.970	-5.473	0.000	
margin_up	-8.8103	2.100	-4.196	0.000	
length	5.6117	0.892	6.292	0.000	

- R2 indique un bon “fit” du modèle…

- … mais certaines variables ne sont pas significatives : on les élimine

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	1200
Model:	GLM	Df Residuals:	1195
Model Family:	Binomial	Df Model:	4
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-37.803
Date:	Tue, 20 Jun 2023	Deviance:	75.605
Time:	10:10:54	Pearson chi2:	2.45e+03
No. Iterations:	10	Pseudo R-squ. (CS):	0.6983
Covariance Type:	nonrobust		

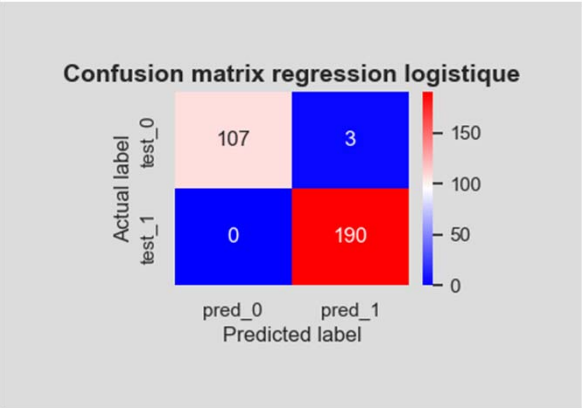
	coef	std err	z	P>|z|	[0.025	0.975]
const	-296.0525	152.148	-1.946	0.052	-594.258	2.153
height_right	-2.7652	1.123	-2.462	0.014	-4.966	-0.564
margin_low	-5.2333	0.861	-6.081	0.000	-6.920	-3.547
margin_up	-8.6477	2.027	-4.267	0.000	-12.620	-4.676
length	5.6477	0.856	6.600	0.000	3.971	7.325

ONCFM

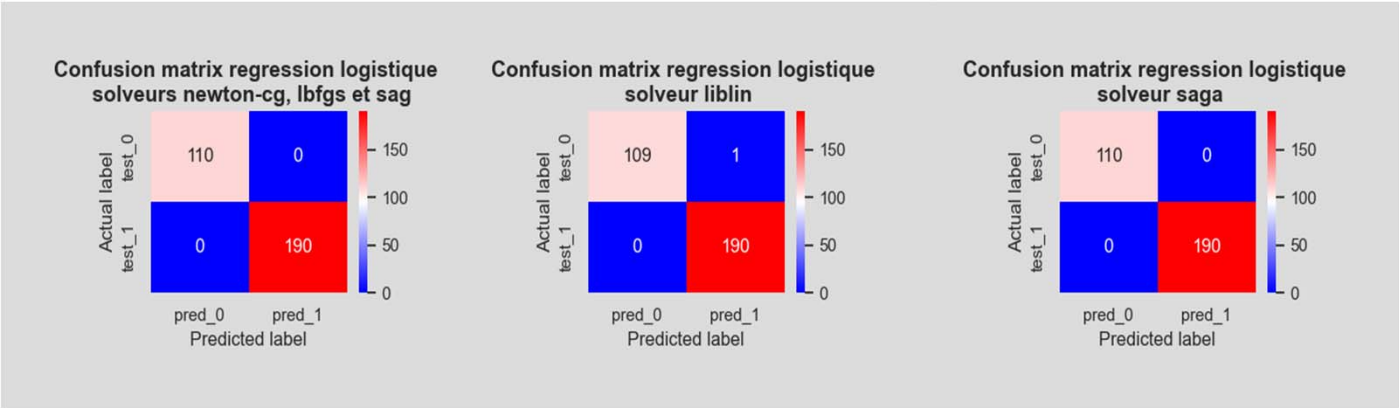
3 – Régression logistique & optimisation

3.2 – Optimisation des hyperparamètres

❖ Sans optimisation :



❖ Utilisation de gridsearchCV :



Regression logistique	Non optimisee	Optimisee nsag	Optimisee liblin	Optimisee saga
R2 training set	0.9908	0.9867	0.9908	0.9875
R2 testing set	0.99	1	0.9967	1
Precision score testing set	0.9845	1	0.9948	1
Recall score testing set	1	1	1	1

➢ Le solveur saga offre le meilleur compromis performances / efficacité computationnelle

ONCFM

3 – Régression logistique & optimisation
3.3 - Algorithme final : test & analyse de la performance

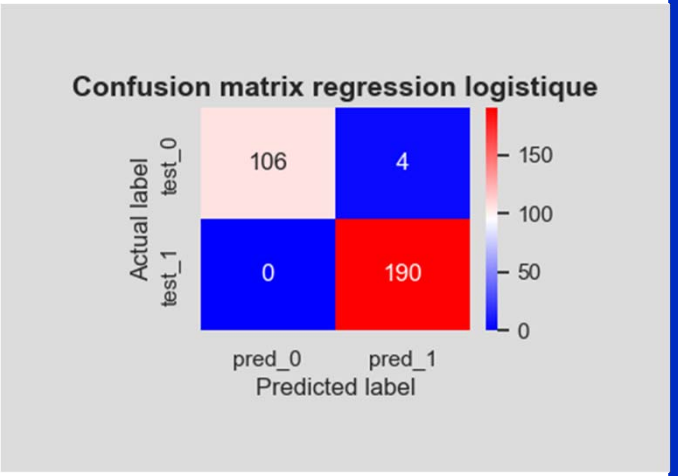
❖ Résultats sur le jeu de test :

➢ Paramètres finaux :

```
{'C': 0.001,
 'class_weight': None,
 'l1_ratio': 0,
 'max_iter': 500,
 'penalty': 'none',
 'random_state': 42,
 'solver': 'saga'}
```

➢ Seuil de probabilité = 50%

➢ R2: 0.9867
➢ Precision score: 97.94%
➢ Recall score: 100%



❖ Résultats du test sur l'échantillon de production non étiqueté :

id	proba	labels_pred_reglog
A_1	0.00013690	FALSE
A_2	0.00001730	FALSE
A_3	0.00005827	FALSE
A_4	0.99503518	TRUE
A_5	0.99997437	TRUE

```
Le billet ref. A_1 est un faux billet.
Le billet ref. A_2 est un faux billet.
Le billet ref. A_3 est un faux billet.
Le billet ref. A_4 est un vrai billet.
Le billet ref. A_5 est un vrai billet.
```

ONCFM

Conclusion

ONCFM