ONCFM
Organisation nationale de lutte contre le faux-monnayage

Détectez des faux billets avec R ou Python

Hébert Thomas / Data Analyst

Contexte

L'ONCFM souhaite développer une application de machine learning pour détecter automatiquement les faux billets en euros, à partir de caractéristiques physiques (longueur, hauteur, etc.).

Un jeu de **1500 billets scannés** (1000 vrais / 500 faux) est fourni pour entraîner les modèles.
Sur recommandation de l'agence européenne **EMV**, quatre algorithmes seront testés en priorité : **Régression Logistique, Random Forest, KNN** et **K-means.**

L'objectif final : **mettre à disposition une application simple et efficace** pour prédire la nature d'un billet à partir de ses caractéristiques.

Sommaire

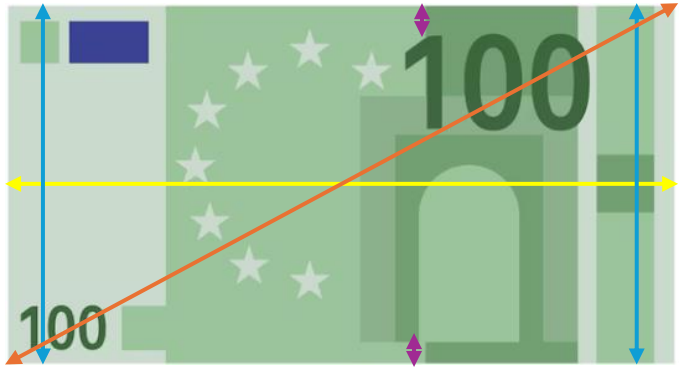
1. Traitement et analyses

a. Exploration du fichier

Le fichier comporte **1500** échantillons réparties en *7* variables :

- 1 variable booléenne (Vrai/Faux billet)
  - **1000** vrais billets
  - **500** faux billets

- 6 variables géométriques (dimensions du billet)
  - **37** valeurs manquantes dans la variable ***margin_low***

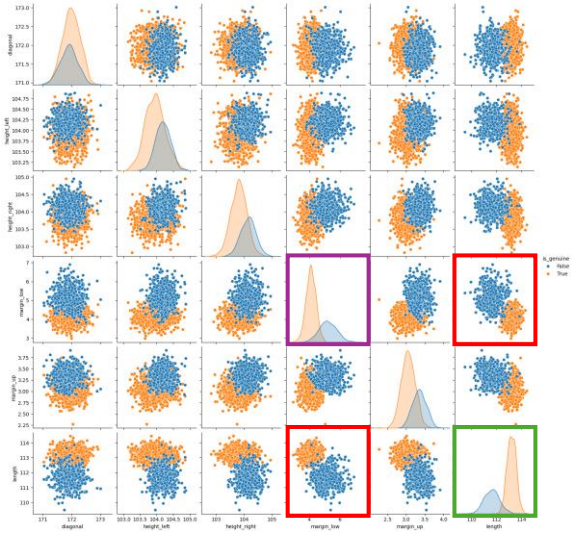


is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
True	171.81	104.86	104.95	4.52	2.89	112.83
True	171.46	103.36	103.66	3.77	2.99	113.09
True	172.69	104.48	103.5	4.4	2.94	113.16
True	171.36	103.91	103.94	3.62	3.01	113.51
True	171.73	104.28	103.46	4.04	3.48	112.54

Extrait du fichier

1. Traitement et analyses

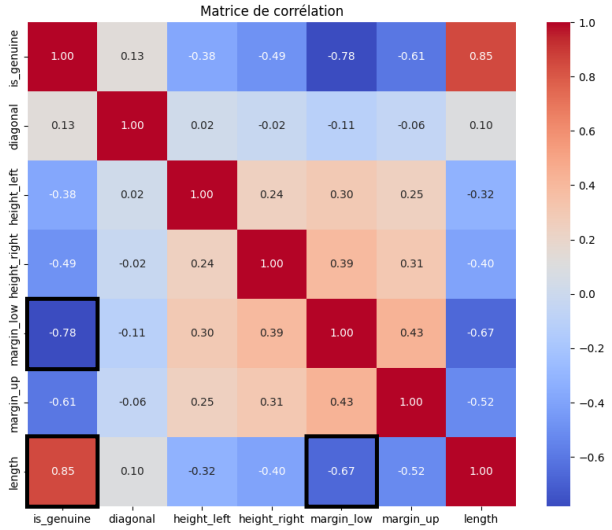
a. Exploration du fichier



- Les variables *length* et *margin_low* montrent une bonne séparation entre vrais et faux billets (pairplot et distribution)

- Corrélation avec *is_genuine* confirmée par la heatmap

- Potentiels bons candidats pour la classification

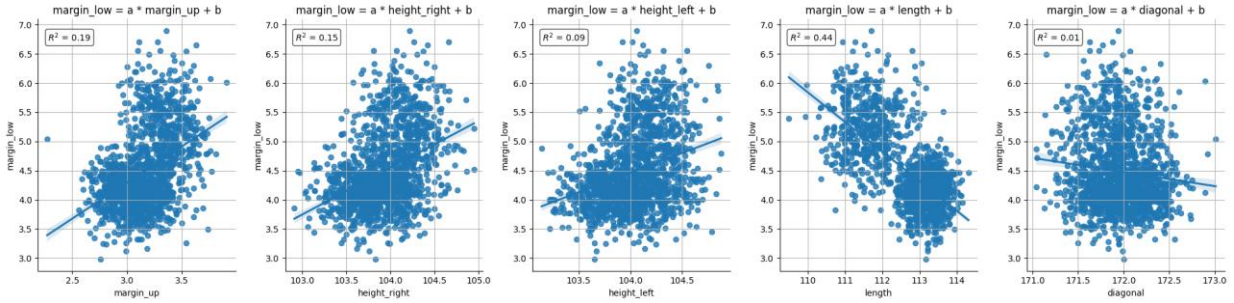


1. Traitement et analyses

b. Régression linéaire

Imputation des valeurs manquantes à la variable *margin_low* :

- Régression linéaire multiple
- Relation linéaire **confirmée** par les graphiques et coefficients
- *is_genuine* transformée en variable numérique
- VIF < 5 ➔ pas de multicolinéarité



Corrélations

margin_low	1.000000
margin_up	0.431606
height_right	0.391085
height_left	0.302643
diagonal	-0.111534
length	-0.666753
is_genuine_num	-0.783032

VIF

	Feature	VIF
1	margin_up	1.596090
2	height_right	1.320836
3	height_left	1.173711
4	diagonal	1.028199
5	length	3.613255
6	is_genuine_num	4.725732

1. Traitement et analyses

a. Régression linéaire

Modèle **cohérent** mais **peu performant** :

- **5** variables explicatives
- P-values < **5%** ➔ variables significatives
- Coefficients interprétables
- R^2 moyen : 0,54
- MSE, RMSE, MAE, MAPE : très faible

```
                            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:                Sat, 17 May 2025   Prob (F-statistic):          2.60e-202
Time:                        08:48:12   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
margin_up      0.2562      0.064      3.980      0.000       0.130       0.382
height_right   0.2571      0.043      5.978      0.000       0.173       0.342
height_left    0.1841      0.045      4.113      0.000       0.096       0.272
length        -0.4091      0.018    -22.627      0.000      -0.445      -0.374
diagonal      -0.1111      0.041     -2.680      0.007      -0.192      -0.030
==============================================================================
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
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.94e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
MSE: 0.1914495494749333
RMSE: 0.4375494823159243
MAE: 0.33537069327648383
MAPE: 0.07467677345733466
R²: 0.5457956683532408
```

1. Traitement et analyses

a. Régression linéaire

Modèle **décousu** mais **performant** :

- **6** variables explicatives (ajout de *is_genuine_num*)
- P-values proche ou > **5%** ➔ variables peu significatives
- Coefficients désordonnés
- R^2 élevé : 0,67
- MSE, RMSE, MAE, MAPE : très faible

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.617
Model:	OLS	Adj. R-squared:	0.615
Method:	Least Squares	F-statistic:	390.7
Date:	Sat, 17 May 2025	Prob (F-statistic):	4.75e-299
Time:	08:48:12	Log-Likelihood:	-774.14
No. Observations:	1463	AIC:	1562.
Df Residuals:	1456	BIC:	1599.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.8668	8.316	0.345	0.730	-13.445	19.179
margin_up	-0.2128	0.059	-3.621	0.000	-0.328	-0.098
height_right	0.0267	0.038	0.701	0.484	-0.048	0.102
height_left	0.0283	0.039	0.727	0.468	-0.048	0.105
length	-0.0039	0.023	-0.166	0.868	-0.050	0.042
diagonal	-0.0130	0.036	-0.364	0.716	-0.083	0.057
is_genuine_num	-1.1406	0.050	-23.028	0.000	-1.238	-1.043

Omnibus:	21.975	Durbin-Watson:	2.038
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37.993
Skew:	0.061	Prob(JB):	5.62e-09
Kurtosis:	3.780	Cond. No.	1.95e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.95e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

MSE: 0.13739452871780827
RMSE: 0.3706676796239568
MAE: 0.28949122166076857
MAPE: 0.06496074172211527
R²: 0.6740384594304603

1. Traitement et analyses

a. Régression linéaire

Modèle **simple** et **performant** :

- 2 variables explicatives
- P-values < 5% ➔ variables significatives
- Coefficients interprétables et cohérents
- R^2 élevé : 0,67
- MSE, RMSE, MAE, MAPE : très faible

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.617
Model:	OLS	Adj. R-squared:	0.616
Method:	Least Squares	F-statistic:	1174.
Date:	Mon, 19 May 2025	Prob (F-statistic):	1.24e-304
Time:	15:01:38	Log-Likelihood:	-774.73
No. Observations:	1463	AIC:	1555.
Df Residuals:	1460	BIC:	1571.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9263	0.198	30.003	0.000	5.539	6.314
margin_up	-0.2119	0.059	-3.612	0.000	-0.327	-0.097
is_genuine_num	-1.1632	0.029	-40.477	0.000	-1.220	-1.107

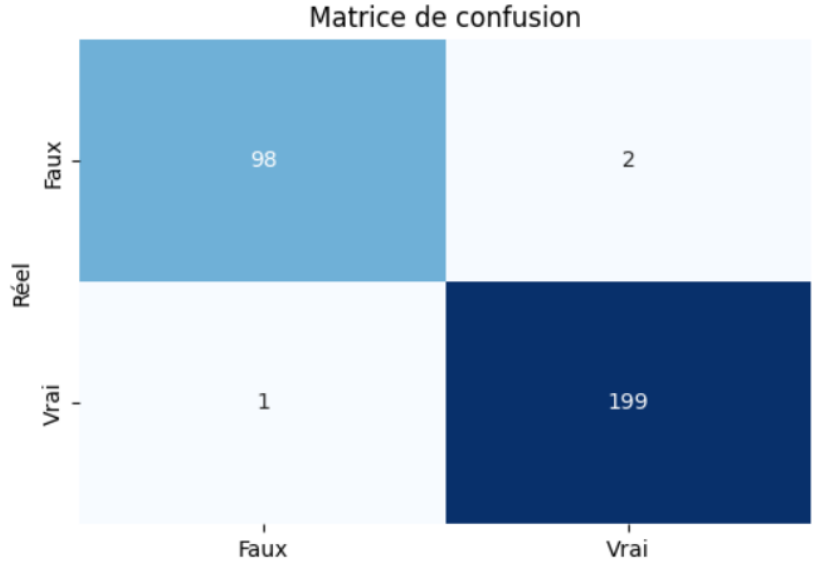
Omnibus:	22.365	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.106
Skew:	0.057	Prob(JB):	3.22e-09
Kurtosis:	3.793	Cond. No.	65.0

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
MSE: 0.13757029265899082
RMSE: 0.3709046948462513
MAE: 0.2897307135185014
MAPE: 0.06504068673145773
R^2: 0.6736214684077528

2. Algorithmes et résultats

a. Régression logistique

- Algorithme de **classification supervisée**
- Adapté aux problèmes **binaires**
- Prédit la **probabilité d'appartenance à une classe**

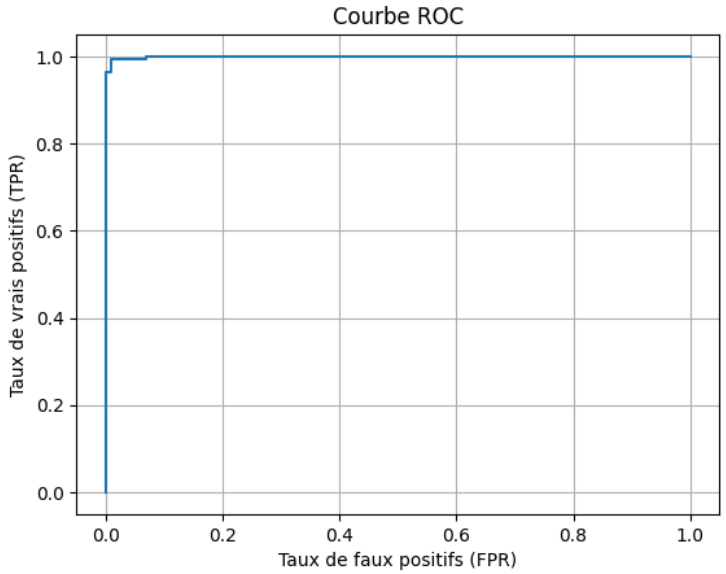


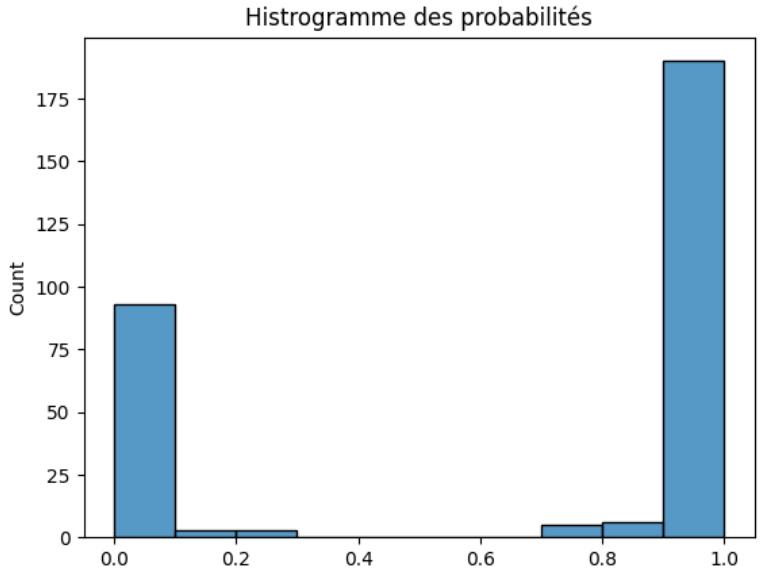
	precision	recall	f1-score	support
False	0.99	0.98	0.98	100
True	0.99	0.99	0.99	200
accuracy			0.99	300
macro avg	0.99	0.99	0.99	300
weighted avg	0.99	0.99	0.99	300

- **Précision élevée** : seulement 3 erreurs sur 300
- Rejette **1** vrai billet (faux négatif)
- Prend **2** faux billets pour des vrais (faux positifs)

2. Algorithmes et résultats

a. Régression logistique





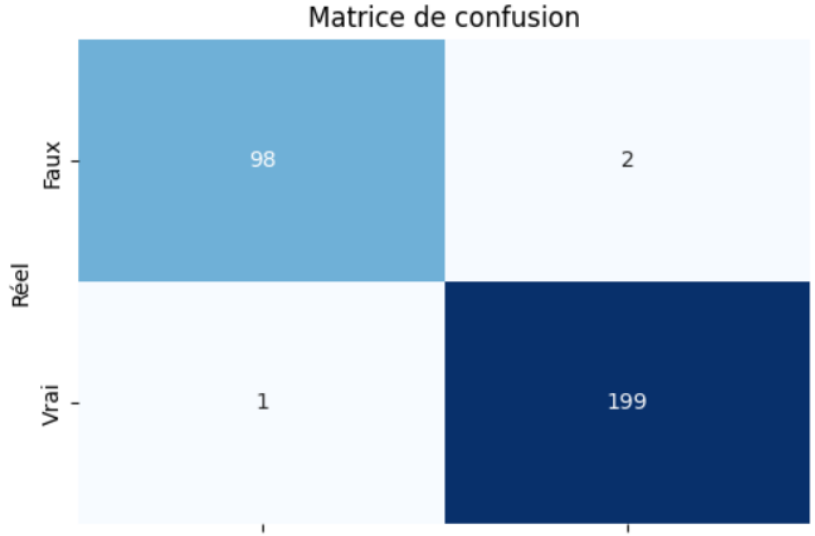
- Courbe **quasi parfaite**
- AUC = **0,99**
- Confirme une **excellente performance du modèle**

- Majorité des prédictions proches de **0 ou 1**
- **Forte confiance du modèle** dans ses prédictions
- **Bonne séparation** entre les deux classes

2. Algorithmes et résultats

b. Random forest

- Algorithme de **classification supervisée**
- Basé sur un **ensemble d'arbres de décision**
- Chaque arbre s'entraîne sur un **échantillon différent**
- La prédiction finale se fait par **vote majoritaire**



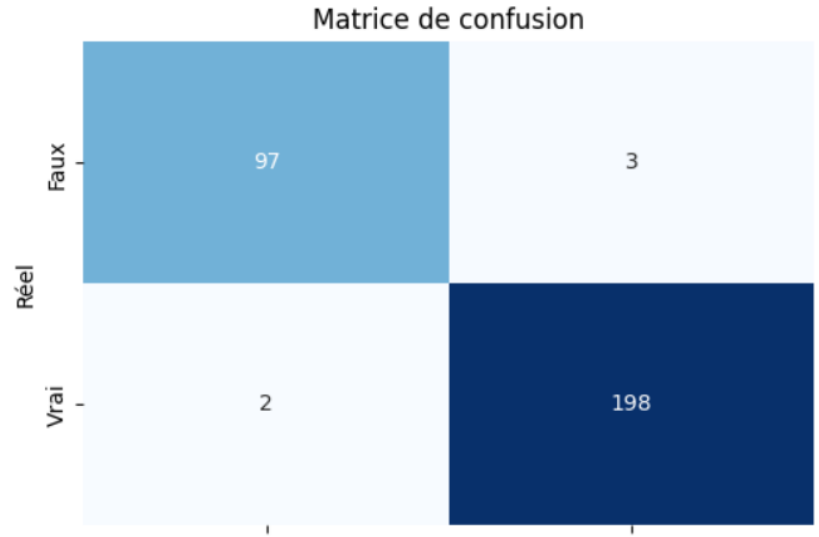
	precision	recall	f1-score	support
False	0.99	0.98	0.98	100
True	0.99	0.99	0.99	200
accuracy			0.99	300
macro avg	0.99	0.99	0.99	300
weighted avg	0.99	0.99	0.99	300

- Score identique à la régression logistique

2. Algorithmes et résultats

c. KNN

- Algorithme de **classification supervisée**
- Basé sur la notion de **proximité dans l'espace des variables**
- Prédit la classe d'un billet en regardant les **k voisins les plus proches**
- La **classe majoritaire** parmi ces voisins est choisie comme prédiction



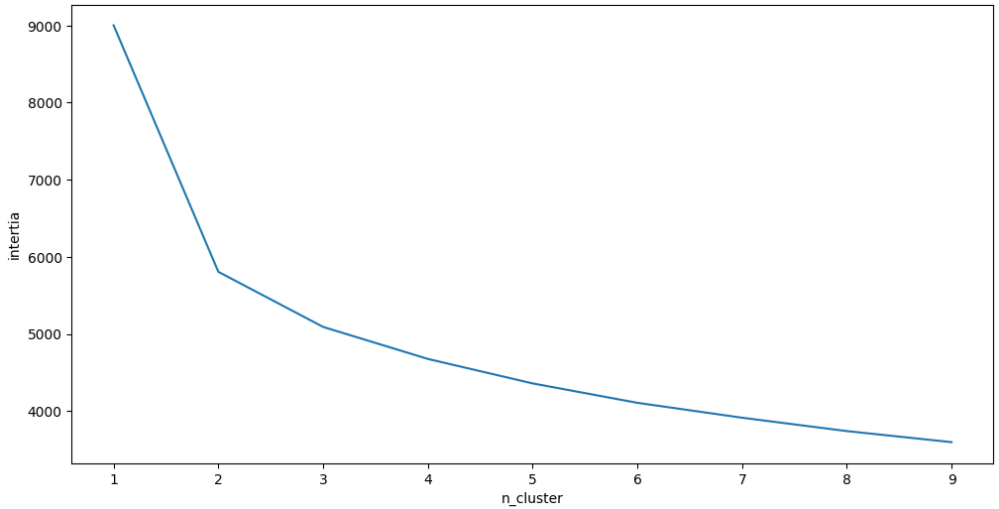
	precision	recall	f1-score	support
False	0.98	0.97	0.97	100
True	0.99	0.99	0.99	200
accuracy			0.98	300
macro avg	0.98	0.98	0.98	300
weighted avg	0.98	0.98	0.98	300

- Le modèle commet **une erreur de plus** dans chaque catégorie
- **Moins performant** que les deux algorithmes précédents

2. Algorithmes et résultats

d. K-means

- Algorithme **non supervisée**
- Regroupe les billets en **clusters** selon leurs caractéristiques
- Ne tient pas compte de **l'étiquette** (vrai/faux)
- Permet de détecter des **structures naturelles** ou des regroupements dans les données

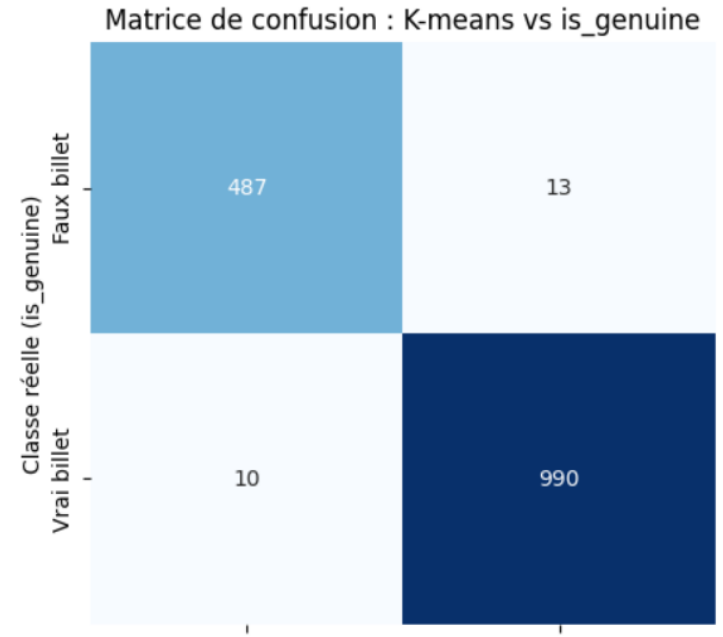


- **Méthode du coude** pour déterminer le nombre optimal de clusters
- Point d'inflexion à **k=2**
- Confirme la présence de **2 groupes distincts**
- Cohérent avec la **classification vrai/faux**

2. Algorithmes et résultats

d. K-means

- Matrice de confusion et scores globalement satisfaisants
- Bonne performance **malgré l'absence d'étiquettes**
- K-means se **défend bien face aux modèles supervisés**

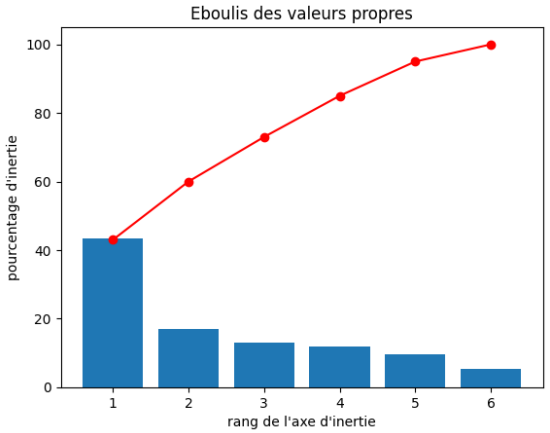


	precision	recall	f1-score	support
False	0.98	0.97	0.98	500
True	0.99	0.99	0.99	1000
accuracy			0.98	1500
macro avg	0.98	0.98	0.98	1500
weighted avg	0.98	0.98	0.98	1500

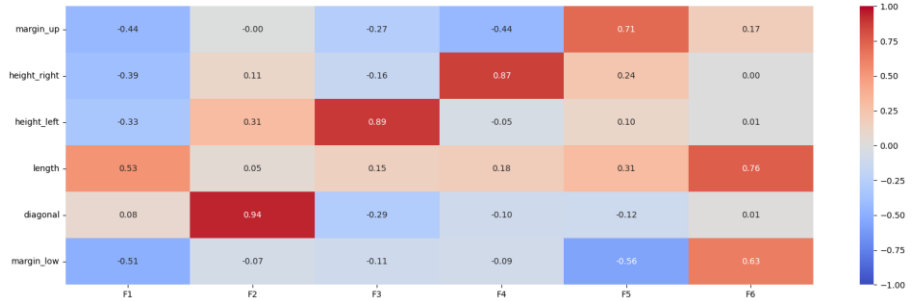
2. Algorithmes et résultats

e. ACP

- Méthode de réduction de dimension
- Résume l'information des variables en quelques axes principaux
- Facilite la visualisation et l'interprétation des données



- F1 ≈ 40% de la variance



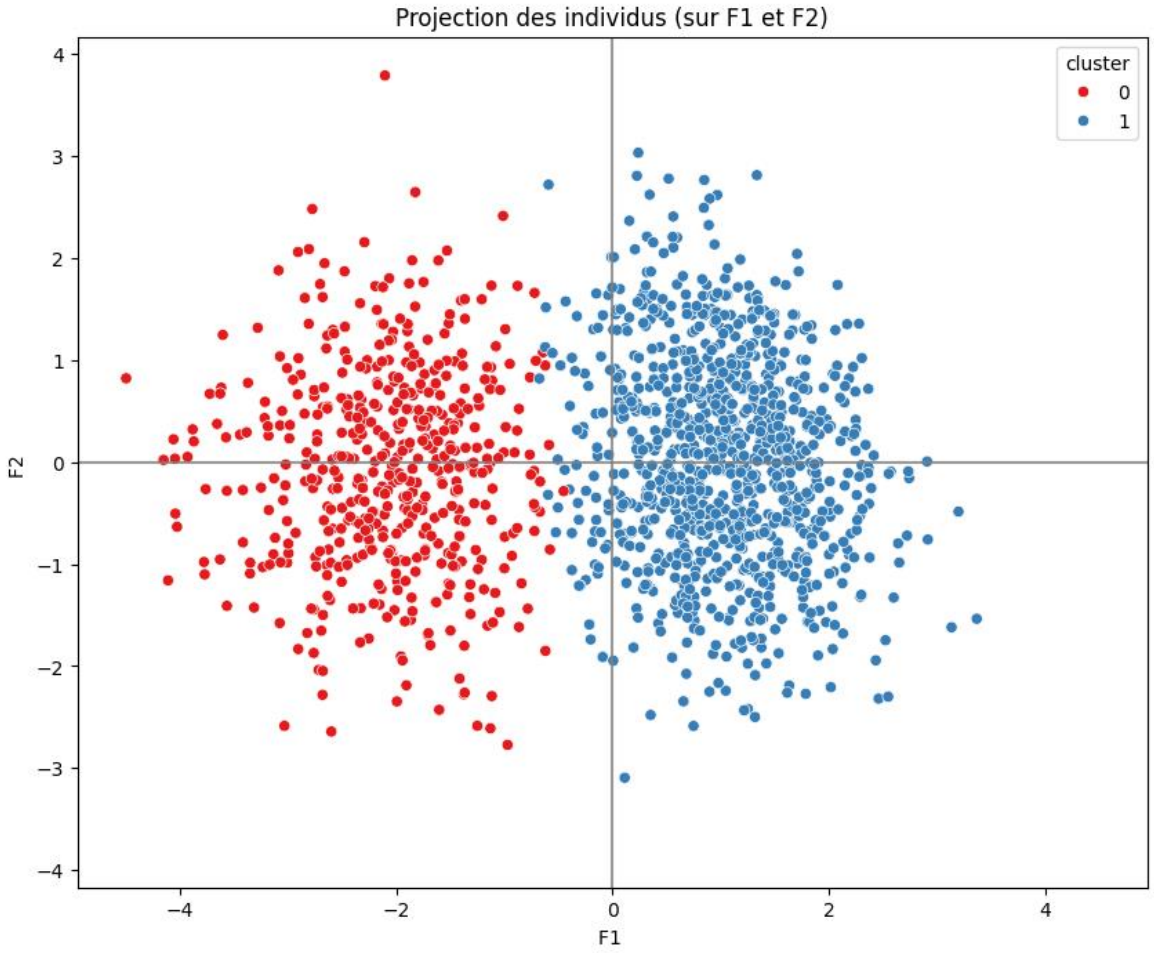
- F1 corrélé à *length* et *margin_low* ➔ 2 variables influançant le plus la véracité d'un billet.

2. Algorithmes et résultats

e. ACP

- Projection sur **F1** et **F2** : séparation nette entre vrais et faux billets.

- Les clusters du **K-means** renforce la distinction.

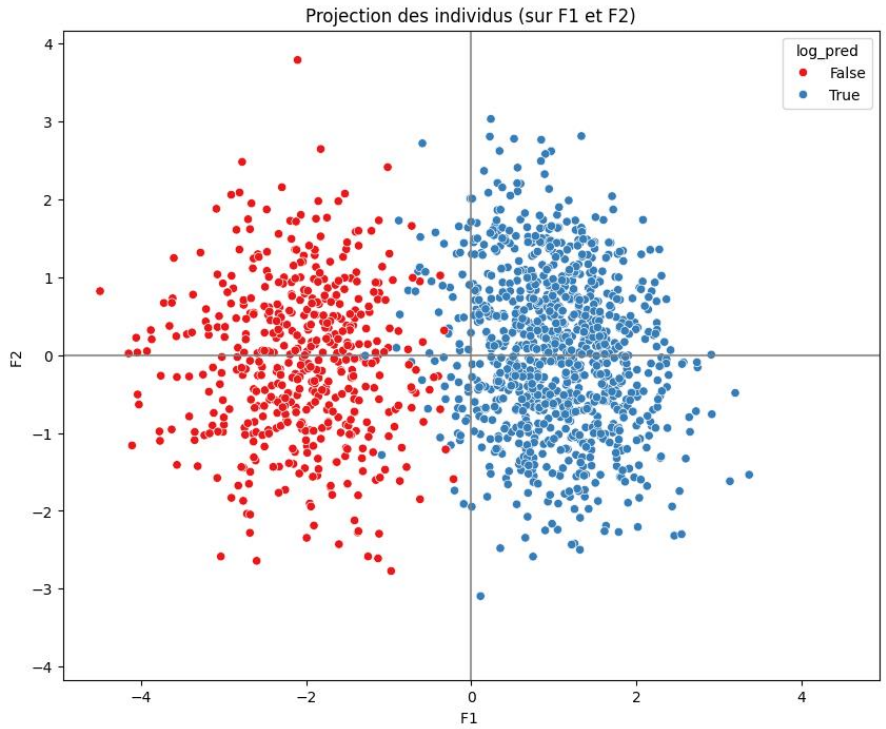


3. Modèle final et application

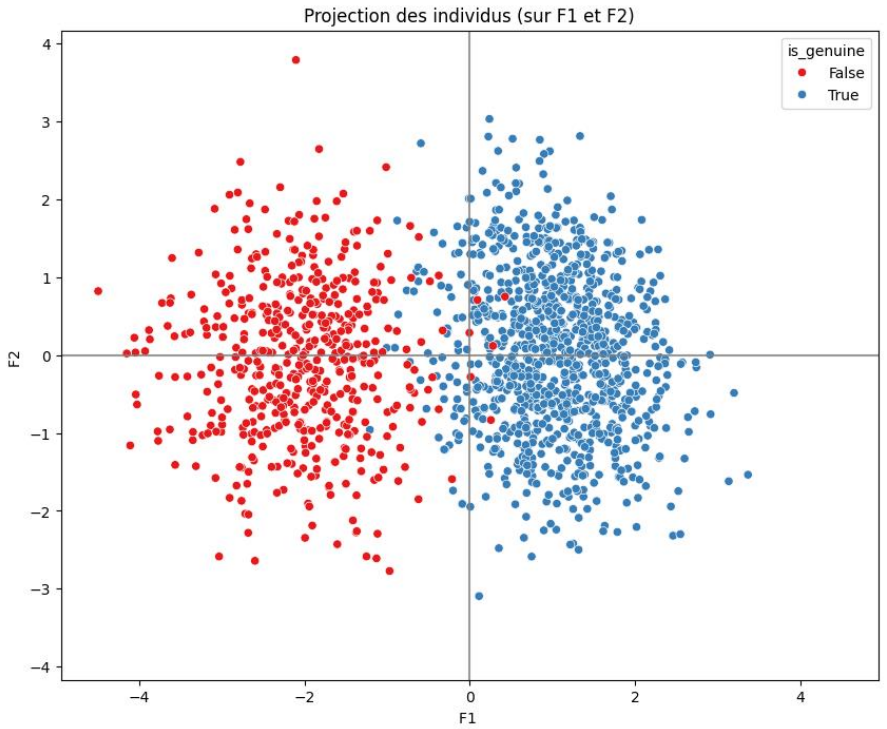
a. Choix du modèle

Régression logistique :

- meilleur compromis entre la **performance**, la **simplicité** et l'**interprétabilité**



Prédiction



Réel

3. Modèle final et application

b. Application

Application déployée sous deux formats :

- **Notebook Jupyter :** permet d'explorer les données si besoin
- **interface Streamlit** : utilisation simple par l'utilisateur

Permet de **visualiser les données**, **prédire automatiquement** et **télécharger les résultats**