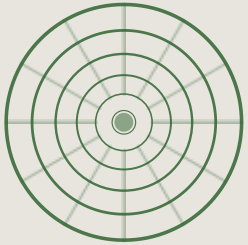
DÉTECTION

DES FAUX *BILLETS*

AVEC R OU PYTHON

1 500 billets | 6 variables | 4 modèles



Rondeau Cécile – 05/2026

HB 72090893 B

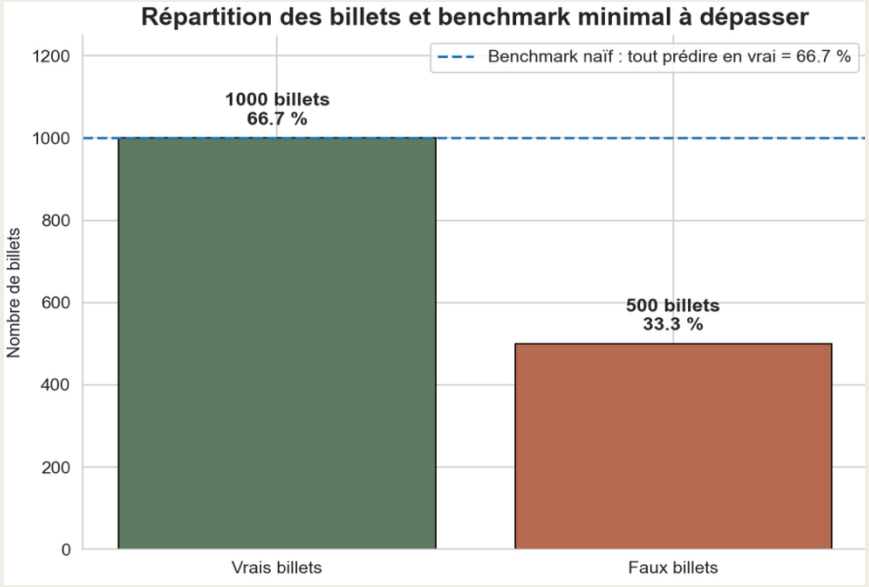
BANQUE CENTRALE

Le dataset

1 500
billets étiquetés

6
variables géométriques (mm)

2 / 1
ratio vrais / faux billets



Benchmark minimal

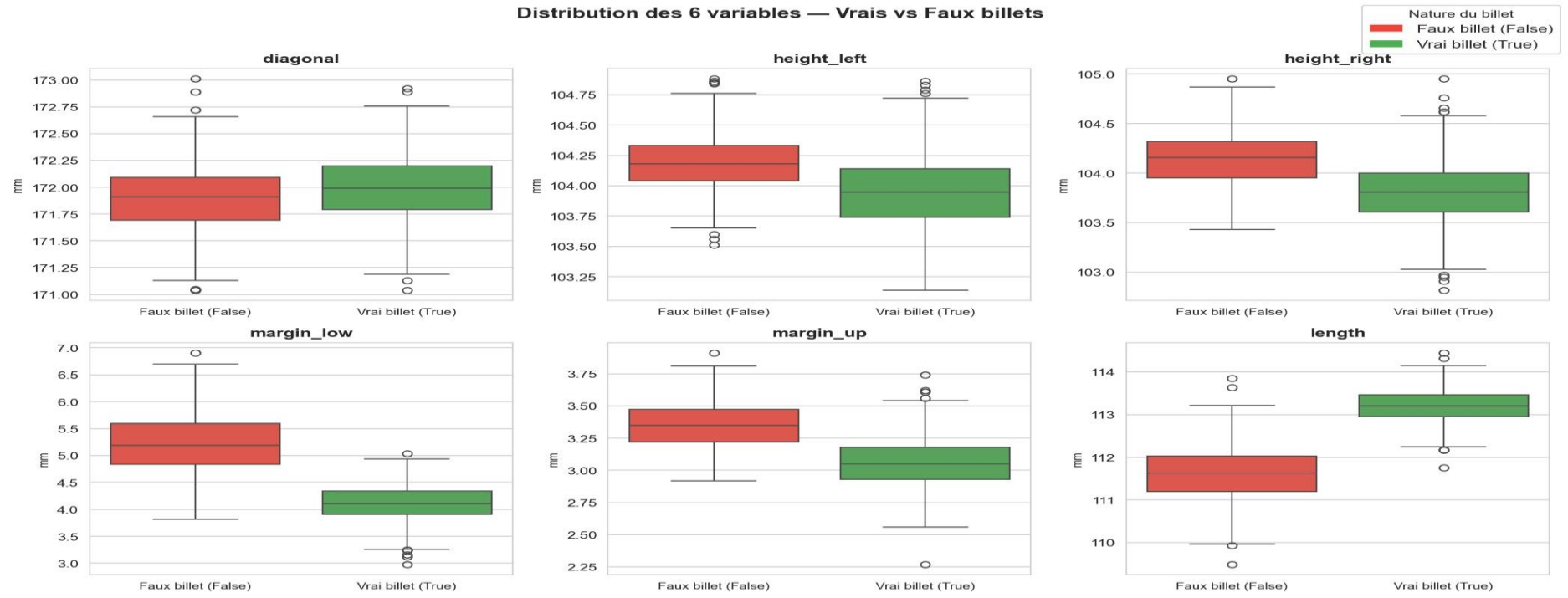
66,7 %

Le déséquilibre 2/1 impose de dépasser ce seuil avant de conclure quoi que ce soit.

Variables prédictives

`diagonal`	`height_left`	`height_right`	`margin_up`	`margin_low`	`length`

Distribution des 6 variables — Vrais vs Faux billets

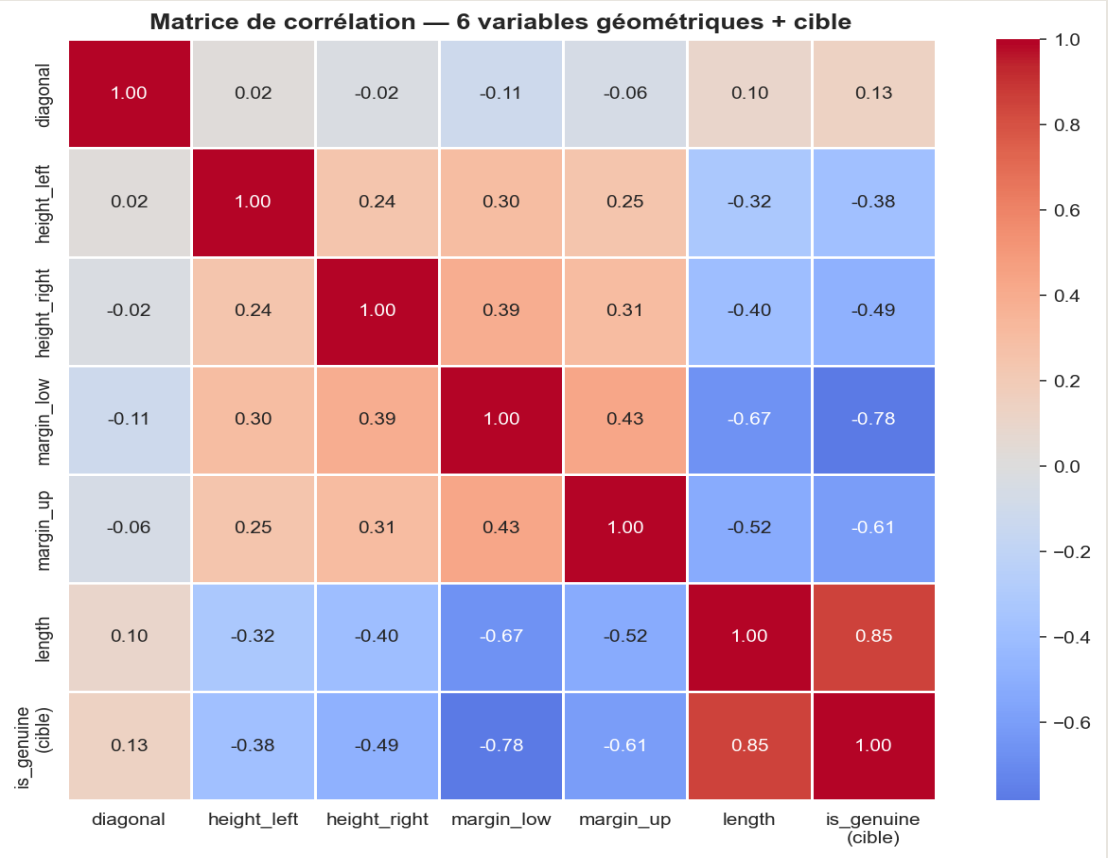


length — séparation la + nette

margin_low — 2ème séparation la + nette

length et margin_low présentent les séparations les plus nettes entre les deux classes

Matrice de corrélation — 6 variables géométriques



−0.67

margin_low ↔ length

+ 0.85

is_genuine ↔ length

− 0.78

is_genuine ↔ margin_low

Pas de redondance parfaite — toutes les variables apportent de l'information à l'ACP

Imputation de margin_low

LE PROBLÈME

37

valeurs manquantes sur margin_low

Supprimer ces lignes ferait perdre 37 billets du jeu d'entraînement. On choisit d'imputer — prédire les valeurs manquantes à partir des 5 autres variables.

Approche

Tester 3 Methodes ·

3 MÉTHODES TESTÉES

1 **Régression simple** — 1 variable

Prédit margin_low uniquement à partir de length.

2 **Régression multiple OLS** — 5 variables

Prédit margin_low à partir des 5 autres variables géométriques simultanément.

3 **Régression multiple** — 2 variables

Prédit margin_low ~ length + is_genuine .

Imputation de margin_low — Résultats

Comparaison des 3 modèles OLS — Choix du modèle d'imputation de margin_low

Modèle	RMSE	MAE	R^2	Type	Retenu
A — Simple (length seul)	0.6872	0.5518	0.9770	1 variable	Non retenu
B — Multiple (5 variables)	0.4807	0.3726	0.9888	5 variables	Non retenu
C — length + is_genuine	0.4134	0.3168	0.9917	2 variables	**Retenu**

Modèle retenu : C — length + is_genuine | Meilleurs RMSE et MAE — pas de multicolinéarité

a **Impact :** 37 valeurs imputées — aucune ligne supprimée
Jeu d'entraînement complet conservé (1 500 billets)

VALIDATION DU MODELE C

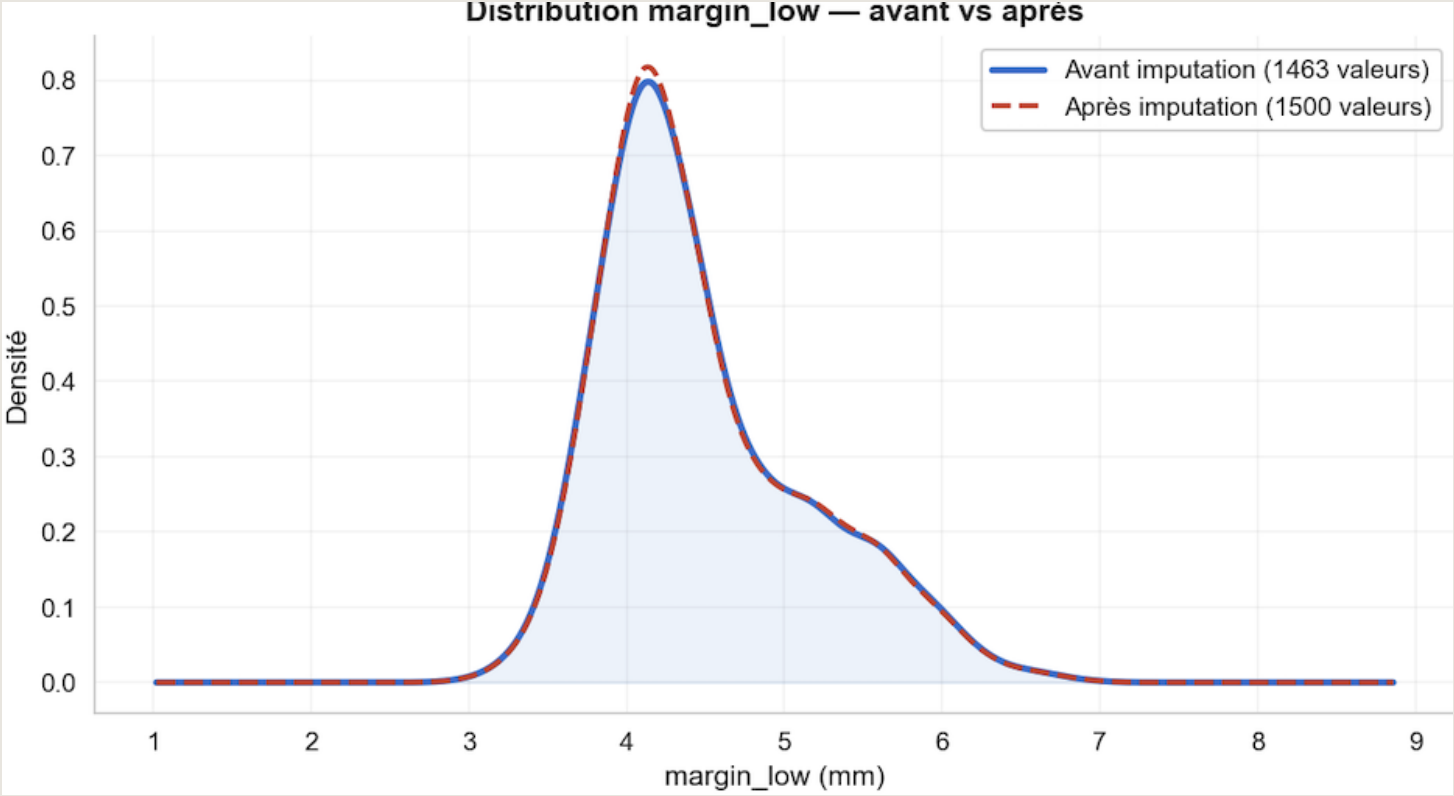
5 CONDITIONS OLS VÉRIFIÉES

1. **Linéarité**
   Points suivent les droites
2. **Homoscédasticité**
   Dispersion résidus constante
3. **Normalité**
   QQ-plot aligné
4. **Multicolinéarité**
   VIF = 3.03 < 5
5. **Autocorrélation**
   Durbin-Watson = 2.03 ≈ 2

Modèle C retenu — RMSE le plus bas · 5 conditions validées · Modèle valide pour imputer les 37 valeurs manquantes

Vérification post-imputation — margin_low

Distribution statistique préservée avant et après imputation



1 500

lignes après imputation

4.427

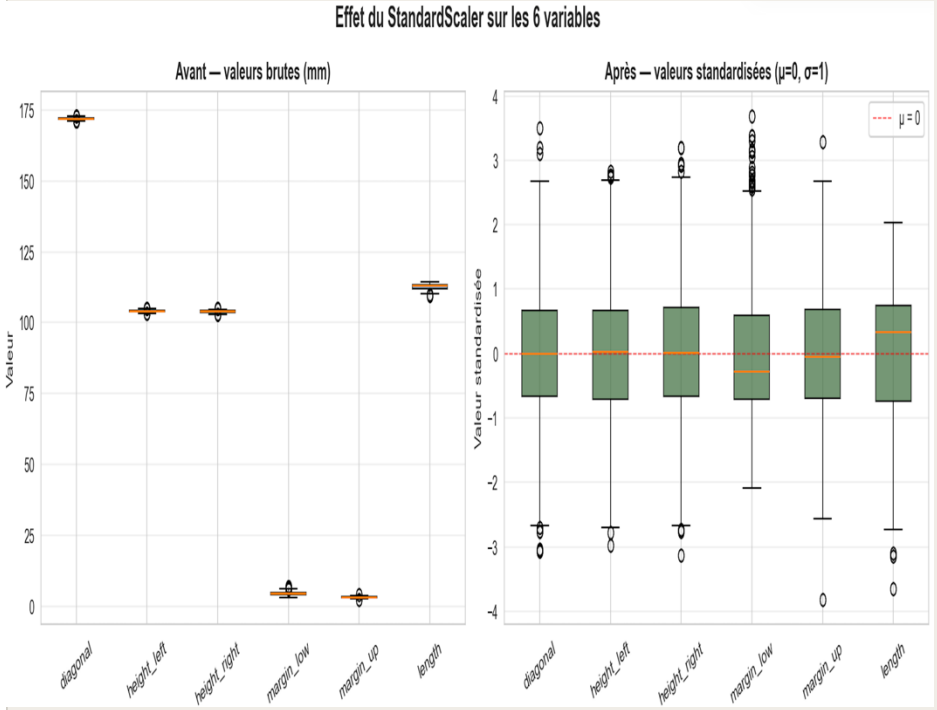
moyenne préservée

0.659

écart-type préservé

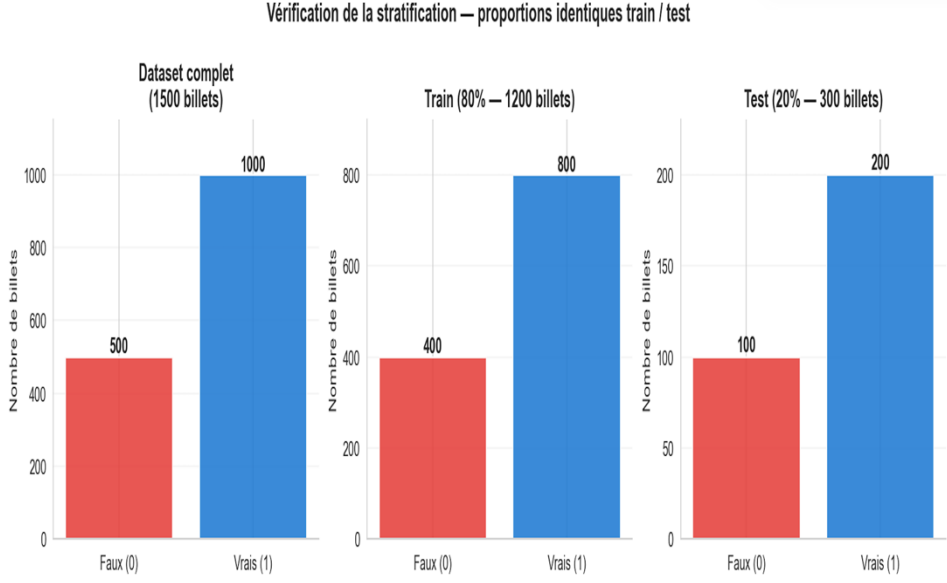
Préparation — split 80/20 & standardisation

EFFET DU STANDARDSCALER SUR LES 6 VARIABLES



Avant : amplitudes très différentes (3mm à 172mm) · Après : toutes centrées sur 0

SPLIT STRATIFIÉ 80/20



1 200
billets train (80%)

300
billets test (20%)

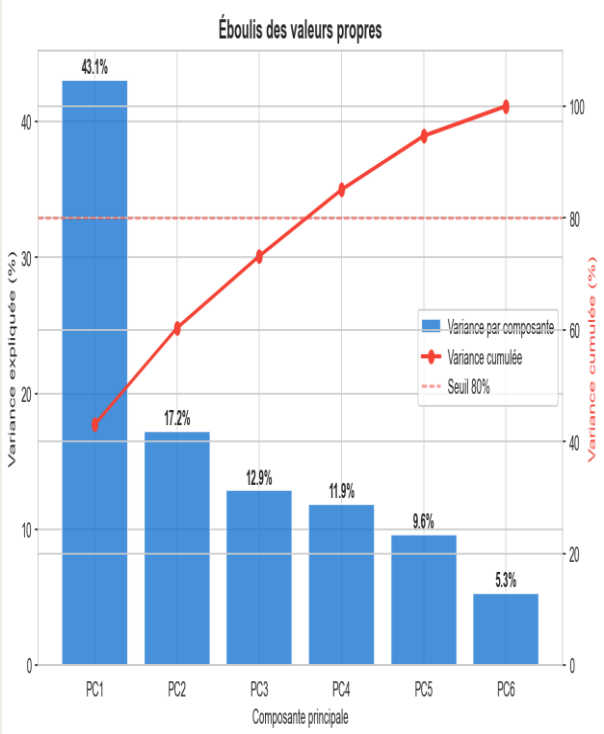
RÈGLE ANTI-LEAKAGE

- ✓ fit(X_train) — scaler apprend sur le train
- ✓ transform(X_train) + transform(X_test)
- ✗ jamais fit(X_test) — évite la fuite

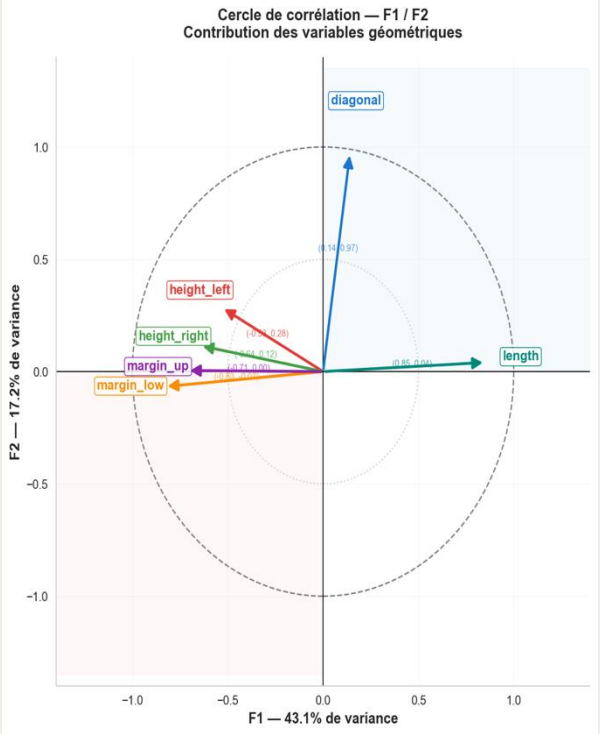
Le scaler ne voit jamais le test · les proportions vrais/faux sont préservées dans les deux ensembles

ACP — réduction de dimension

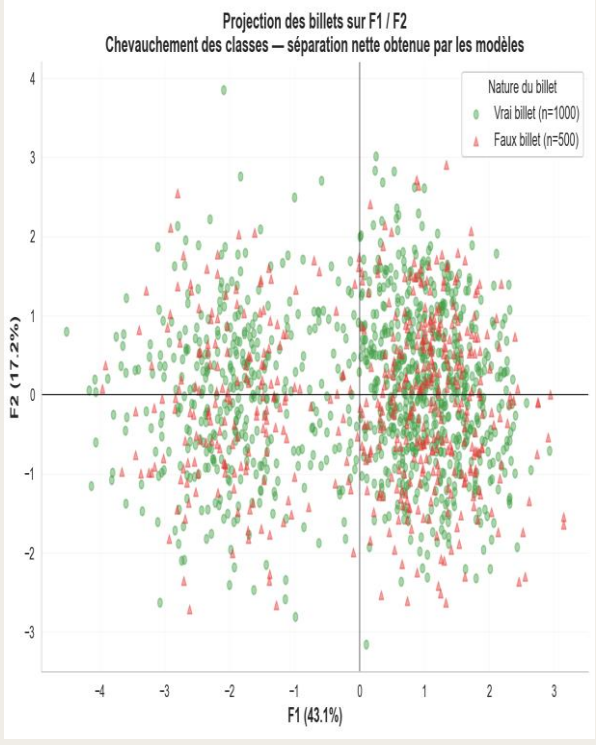
ÉBOULIS DES VALEURS PROPRES



CERCLE DE CORRÉLATION F1/F2

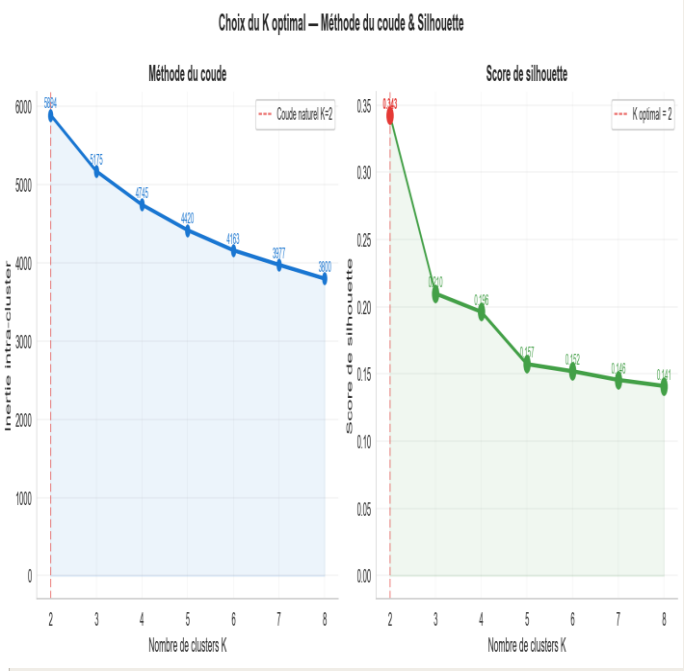


PROJECTION DES INDIVIDUS



K-Means — exploration non supervisée · 2 groupes naturels confirmés

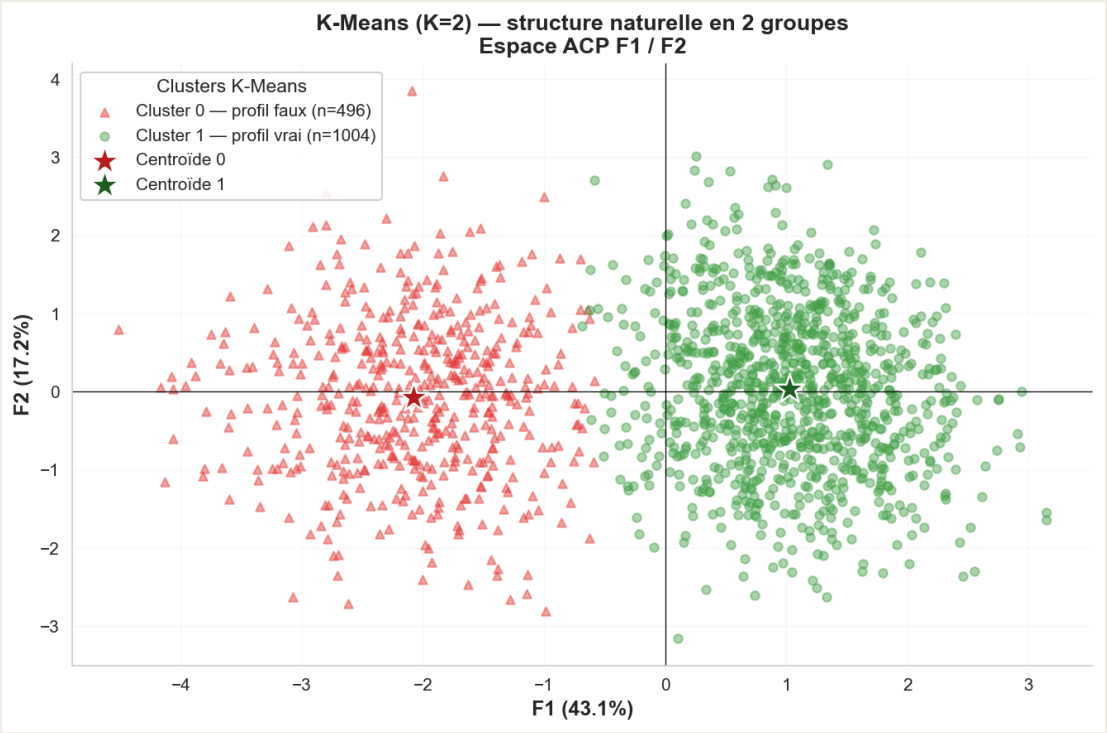
CHOIX DU K OPTIMAL



K=2
clusters retenus

0.343
score silhouette

K-MEANS VS CLASSES RÉELLES



3 modèles supervisés testés

1

Régression logistique

Linéaire · probabiliste

Calcule une probabilité d'appartenir à la classe « faux billet ». Si probabilité ≥ seuil → faux billet détecté.

ATOUT

Interprétable — les coefficients montrent l'influence de chaque variable.

2

KNN

Non-paramétrique

Classe un billet selon les K billets les plus proches dans l'espace des 6 variables. Vote majoritaire.

ATOUT

Simple, aucune hypothèse sur la distribution des données.

3

Random Forest

Ensembliste

Combine plusieurs centaines d'arbres de décision entraînés sur des sous-ensembles aléatoires.

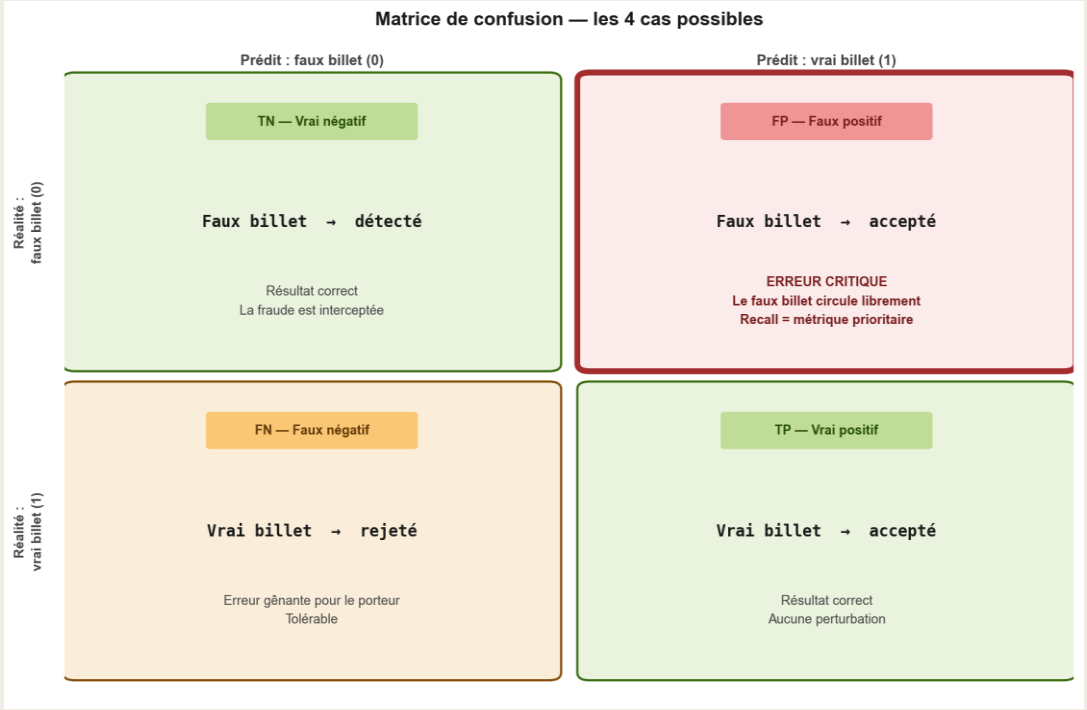
ATOUT

Robuste aux outliers · fournit l'importance des variables.

Chaque modèle entraîné sur X_train · évalué sur X_test · même protocole pour tous

Matrice de confusion & métrique prioritaire

LES 4 CAS POSSIBLES



MÉTRIQUES D'ÉVALUATION

Accuracy

Parmi tous les billets examinés, combien ont été correctement classés?

Précision

Parmi les billets détectés faux, combien le sont ?

F1-score

Un seul chiffre qui résume le compromis entre ne pas rater de faux billets(Recall) et ne pas accuser de vrais billets (precision)

MÉTRIQUE PRIORITAIRE

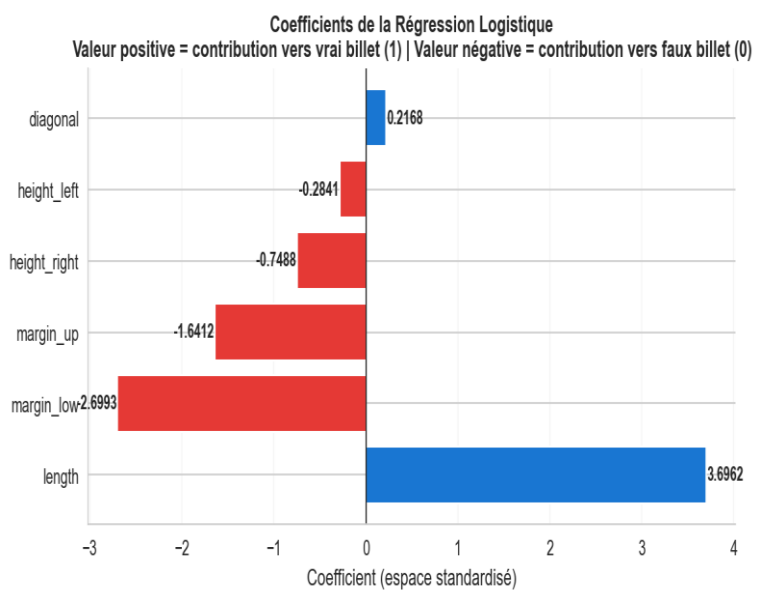
Recall

Parmi tous les faux billets réels, combine le modèle en a-t-il détectés?

Recall prioritaire · Accuracy insuffisante sur données déséquilibrées · AUC pour comparer globalement

Régression Logistique — résultats

COEFFICIENTS — INFLUENCE DES VARIABLES

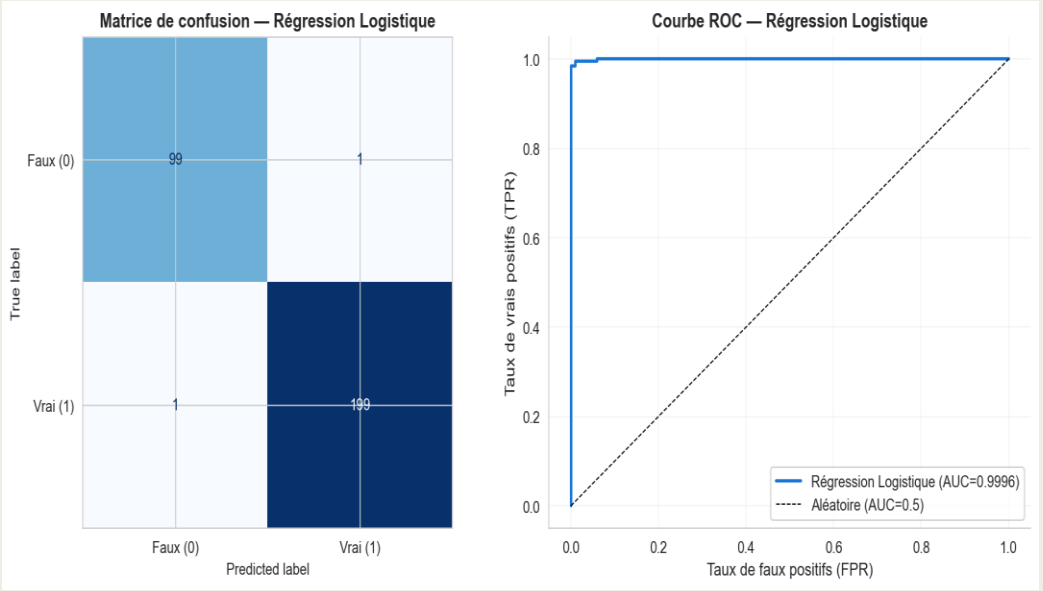


length = variable la plus discriminante · confirme l'EDA

+3.696 length

−2.699 margin_low

MATRICE DE CONFUSION + COURBE ROC

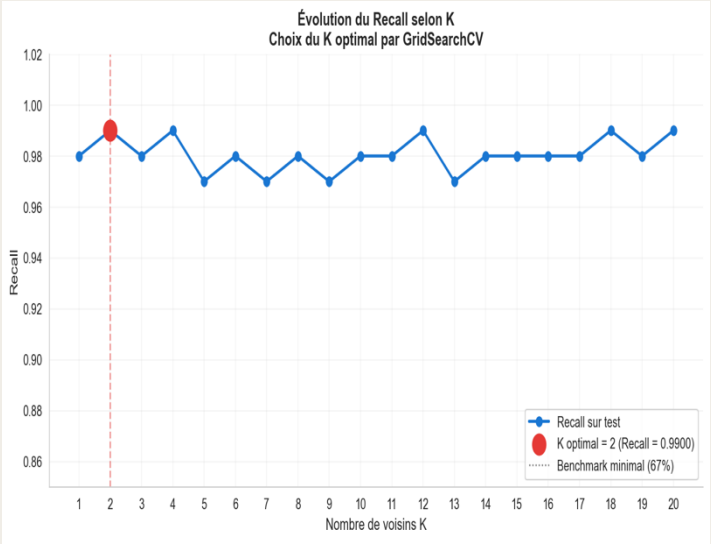


1 seul faux billet non détecté sur 100

Recall	AUC	FN
0.990	0.9996	1

KNN — résultats

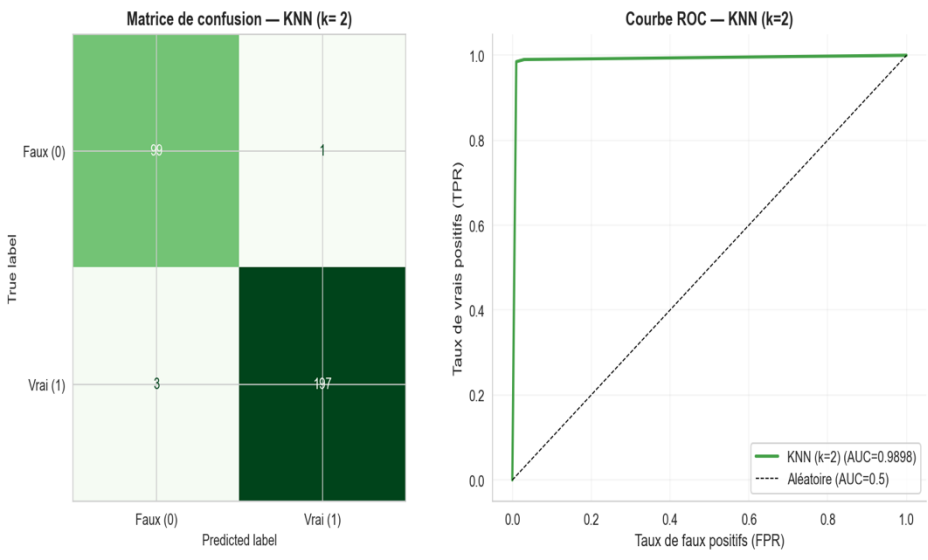
CHOIX DU K OPTIMAL — GRIDSEARCHCV



K=2
optimal

0.990
Recall

MATRICE DE CONFUSION + COURBE ROC



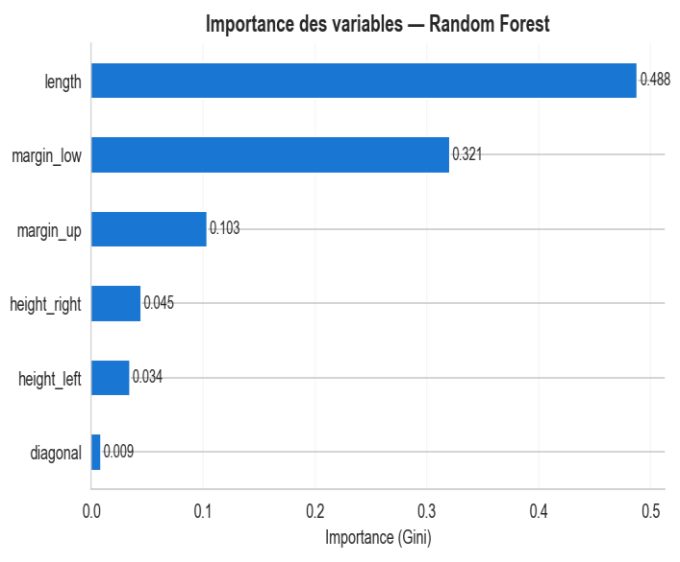
0.990
Recall

0.989
AUC

1
FN

Random Forest — résultats

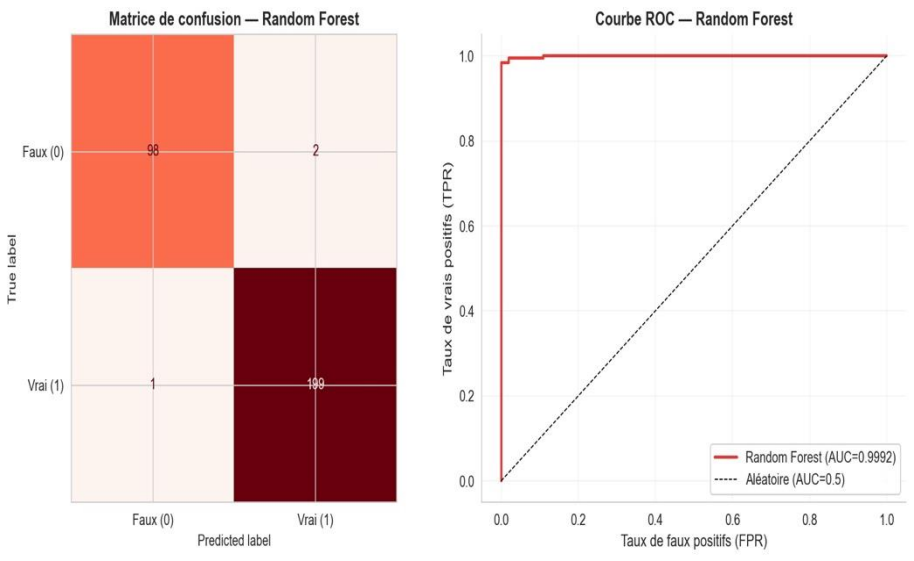
IMPORTANCE DES VARIABLES



0.488 length

0.321 margin_low

MATRICE DE CONFUSION + COURBE ROC

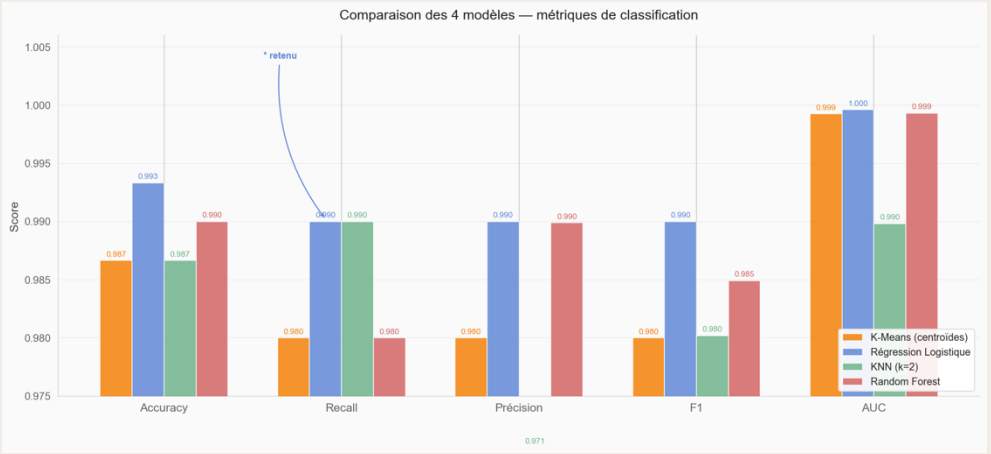


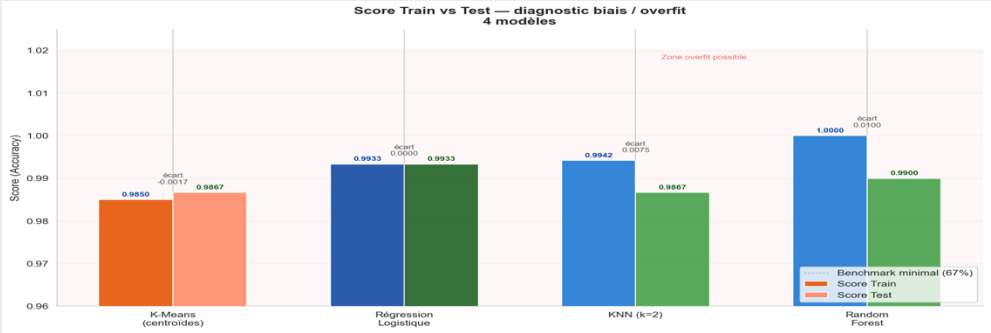
0.980 Recall

0.999 AUC

2 FN

Comparaison & modèle retenu





SYNTHÈSE COMPARATIVE

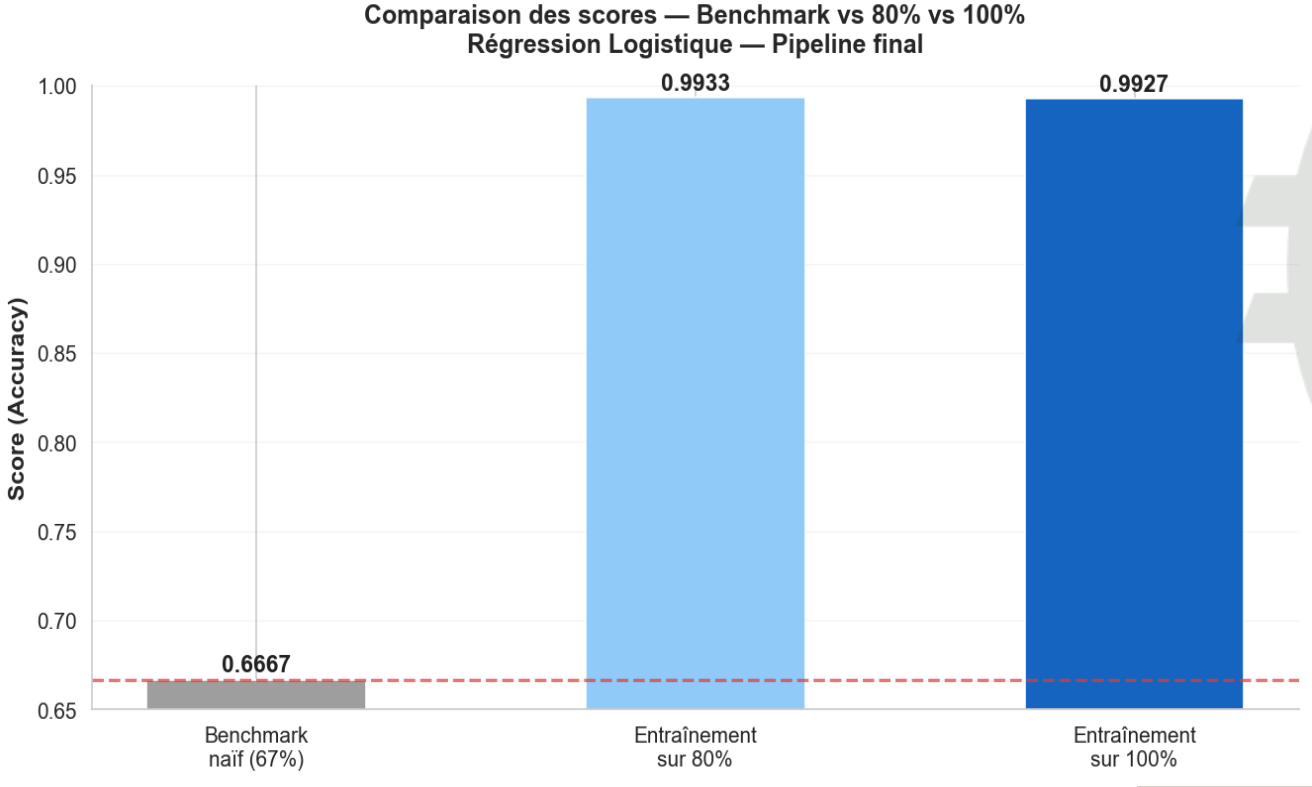
Synthèse comparative — 4 modèles

Recall et AUC prioritaires — minimisation des faux négatifs (billets non détectés)

Modèle	Type	FN	Recall	AUC	Recommandé
K-Means (centroïdes)	Non supervisé	2	0.9800	0.9992	Non retenu
Régression Logistique	Supervisé	1	0.9900	0.9996	**Retenu**
KNN (k=2)	Supervisé	1	0.9900	0.9898	Non retenu
Random Forest	Supervisé	2	0.9800	0.9993	Non retenu

MODÈLE RETENU — Régression Logistique

Validation du pipeline final



Benchmark dépassé

+0.33 au-dessus du seuil 66,7 %

Stabilité 80 % → 100 %

écart négligeable (0.9933 vs 0.9927)

Pipeline prêt production

ré-entraîné sur 1 500 billets · sauvegardé

Le modèle final écrase le benchmark et reste stable sur 100 % des données — pipeline validé pour la production

PROJET 12 · ONCFM

Détection automatique de faux billets

Disponible pour vos questions

Démonstration disponible

Application fonctionnelle · prédiction de la nature d'un billet à partir de ses 6 mesures géométriques