Détectez des faux billets avec R ou Python

PROJET 10

Juan Luis Acebal rico

ONCFM

INTRODUCTION:

- Aujourd’hui on est ici pour travailler avec des données des billets
- Nous allons travailler avec models predictifs , et nous allons dire des vrai billets et les faux billets.

DONNEES

NOUS AVONS BILLETS.CSV

ON A DES MESURES DES BILLETS
ET S’ILS SONT VRAIS OU FAUSES

METHODOLOGIE UTILISÉE

IMPORTATION DES DONNEES

NETTOYAGE DES DONNEES

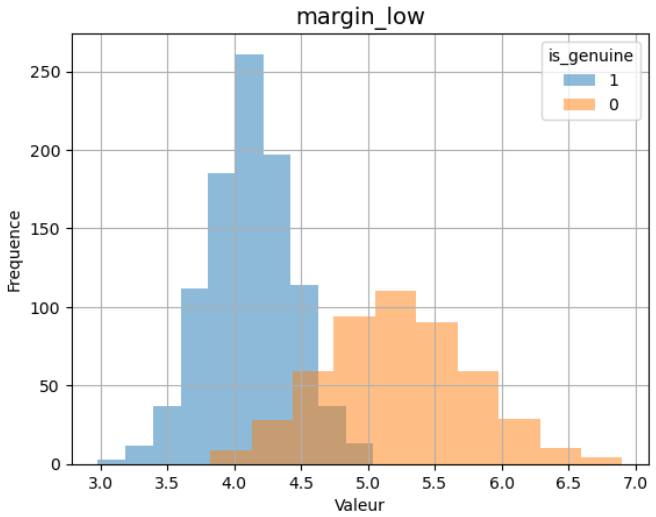
PRÉ PROCESSING DES DONNEES

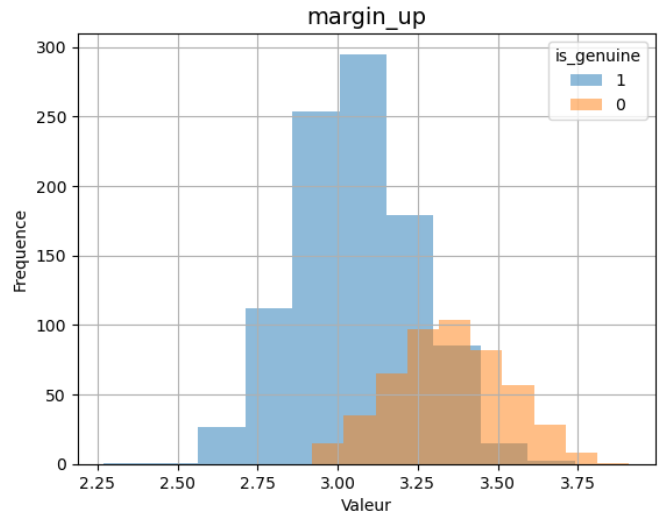
CHOISIR MODELE DE PREDICTION POUR CHAQUE CAS

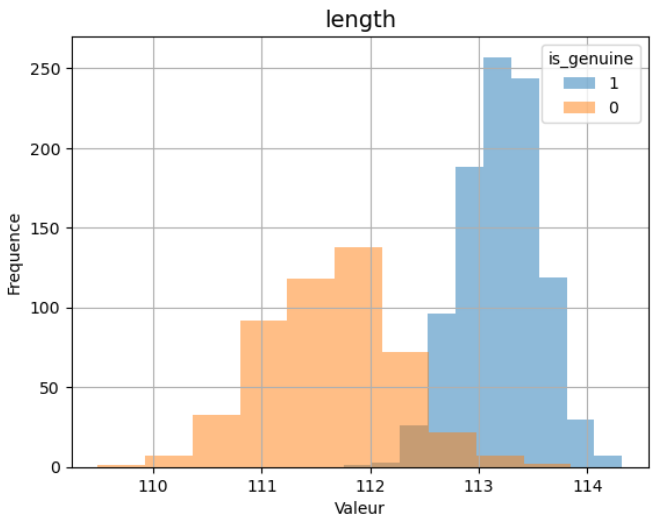
- REGRESION LINEARE POUR IMPUTER VALEURS
- REGRESION LOGISTIQUE POUR « IS_GENUINE »

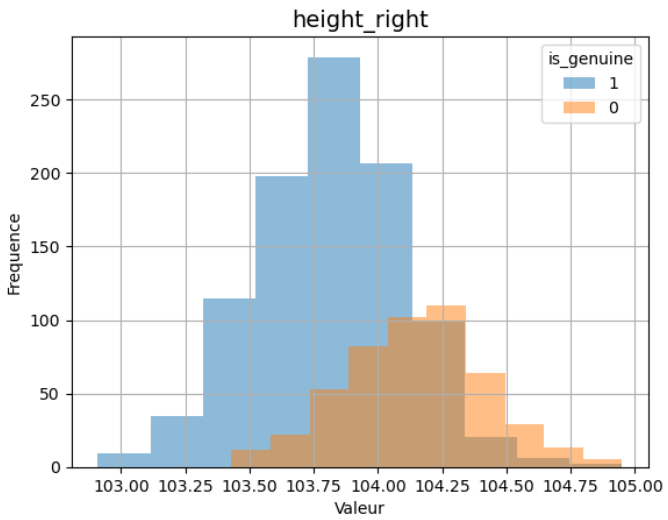
EVALUER LE MODELE

EDA histograms:

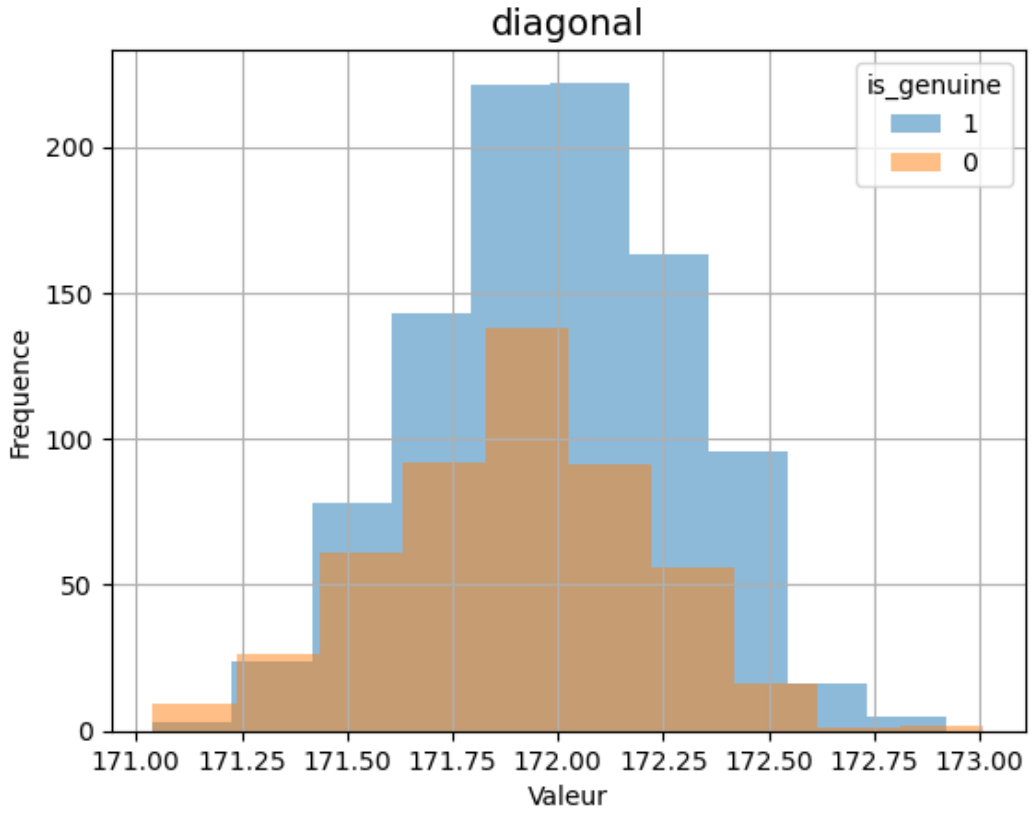


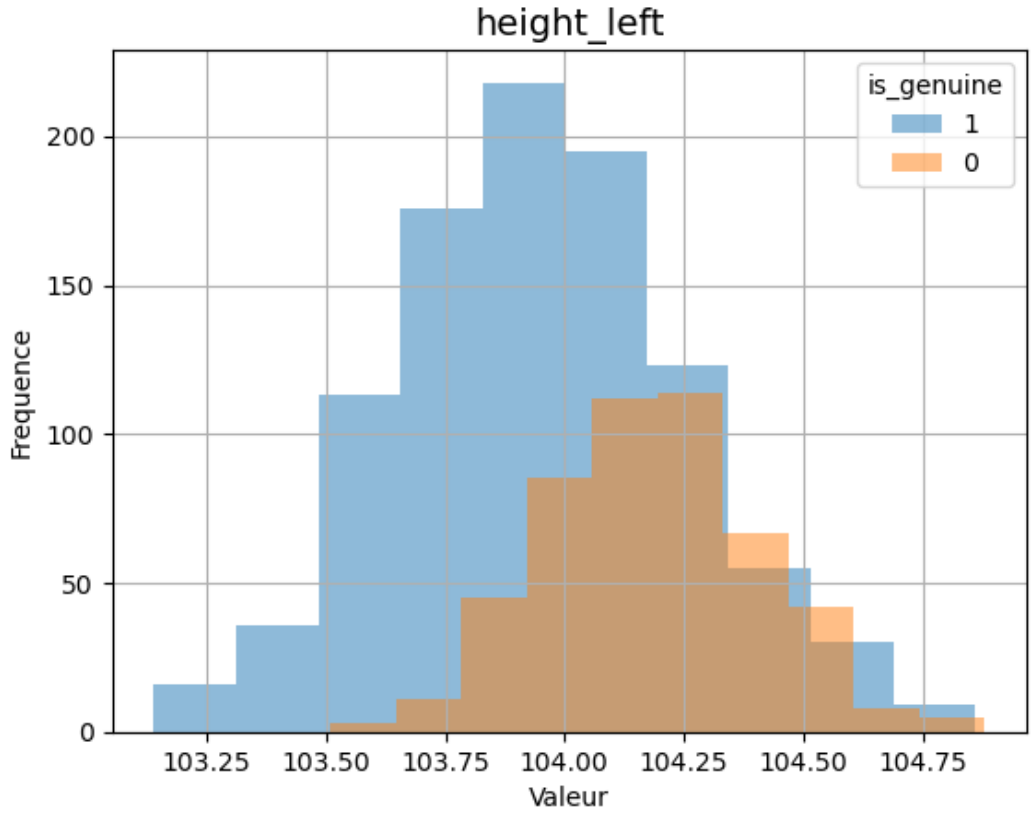




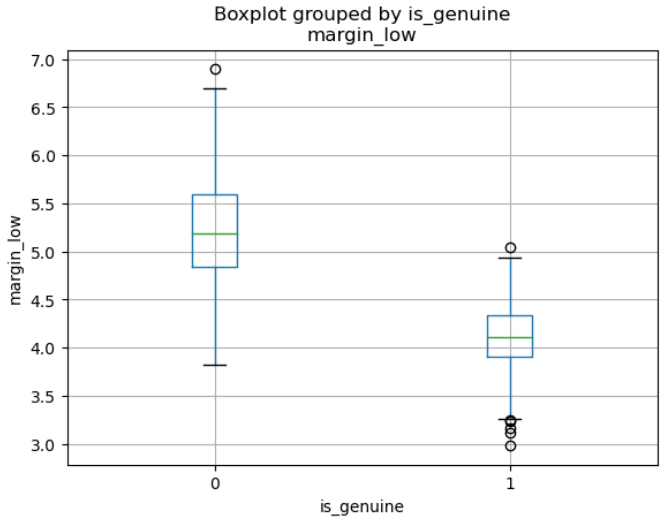


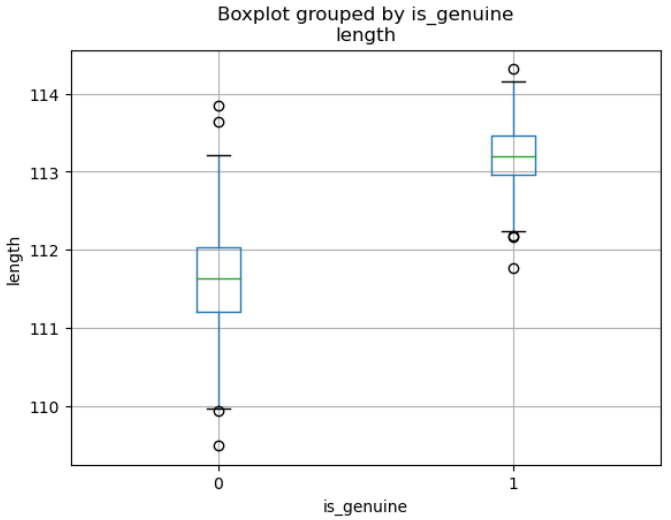
EDA histograms: variables non utilisés

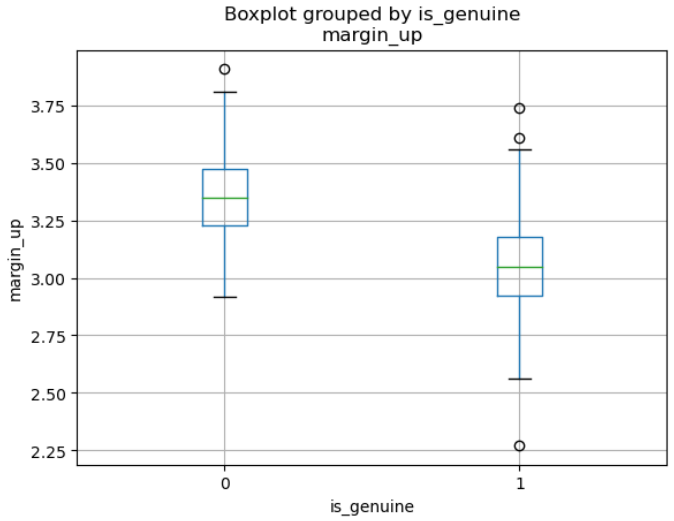


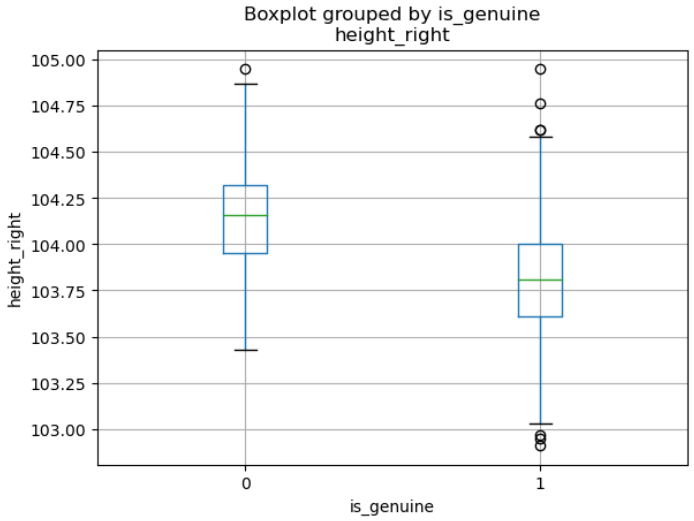


EDA boxplots:

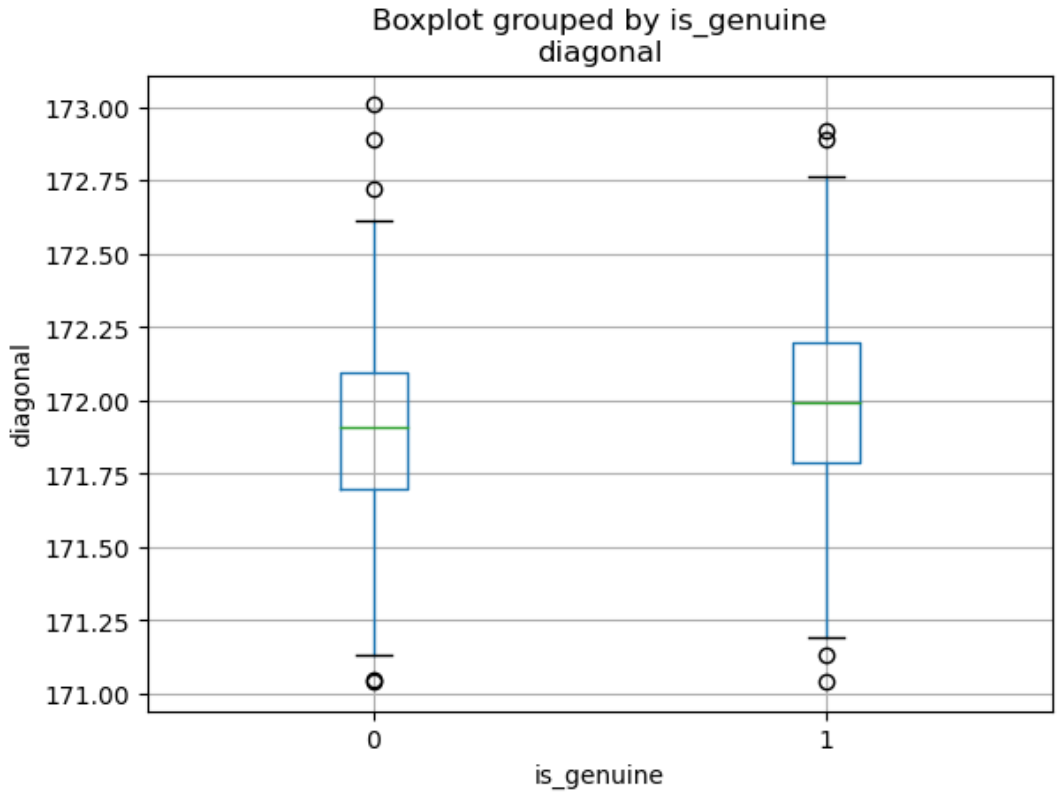


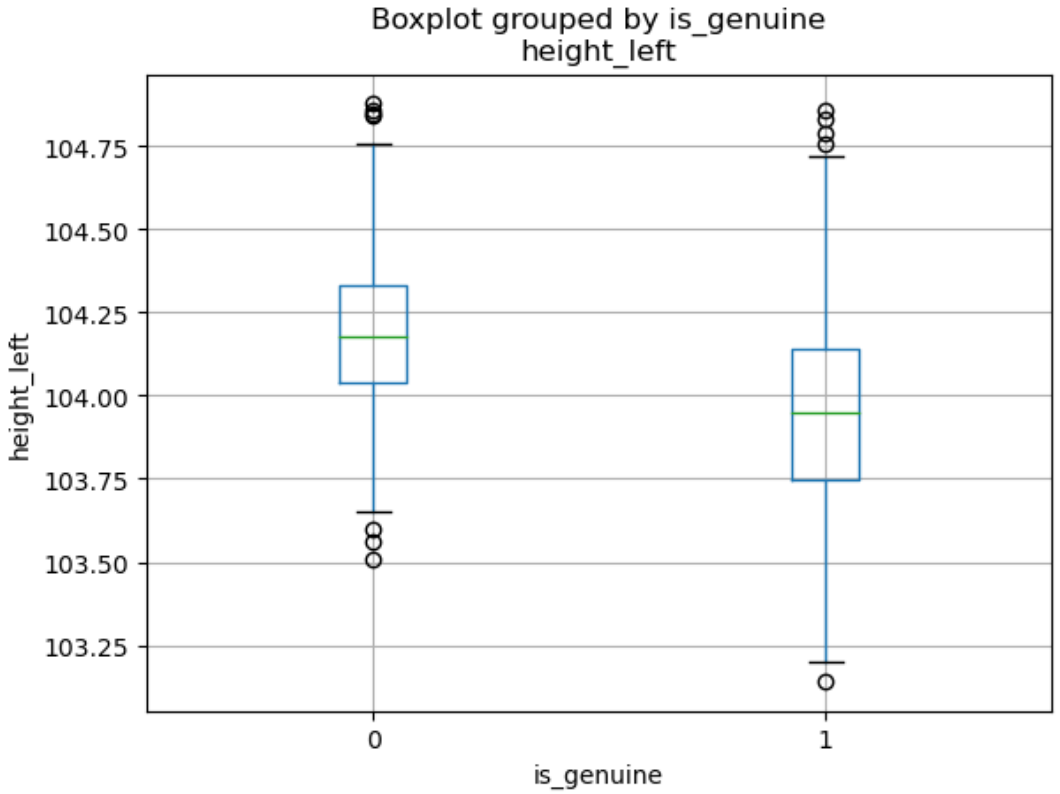




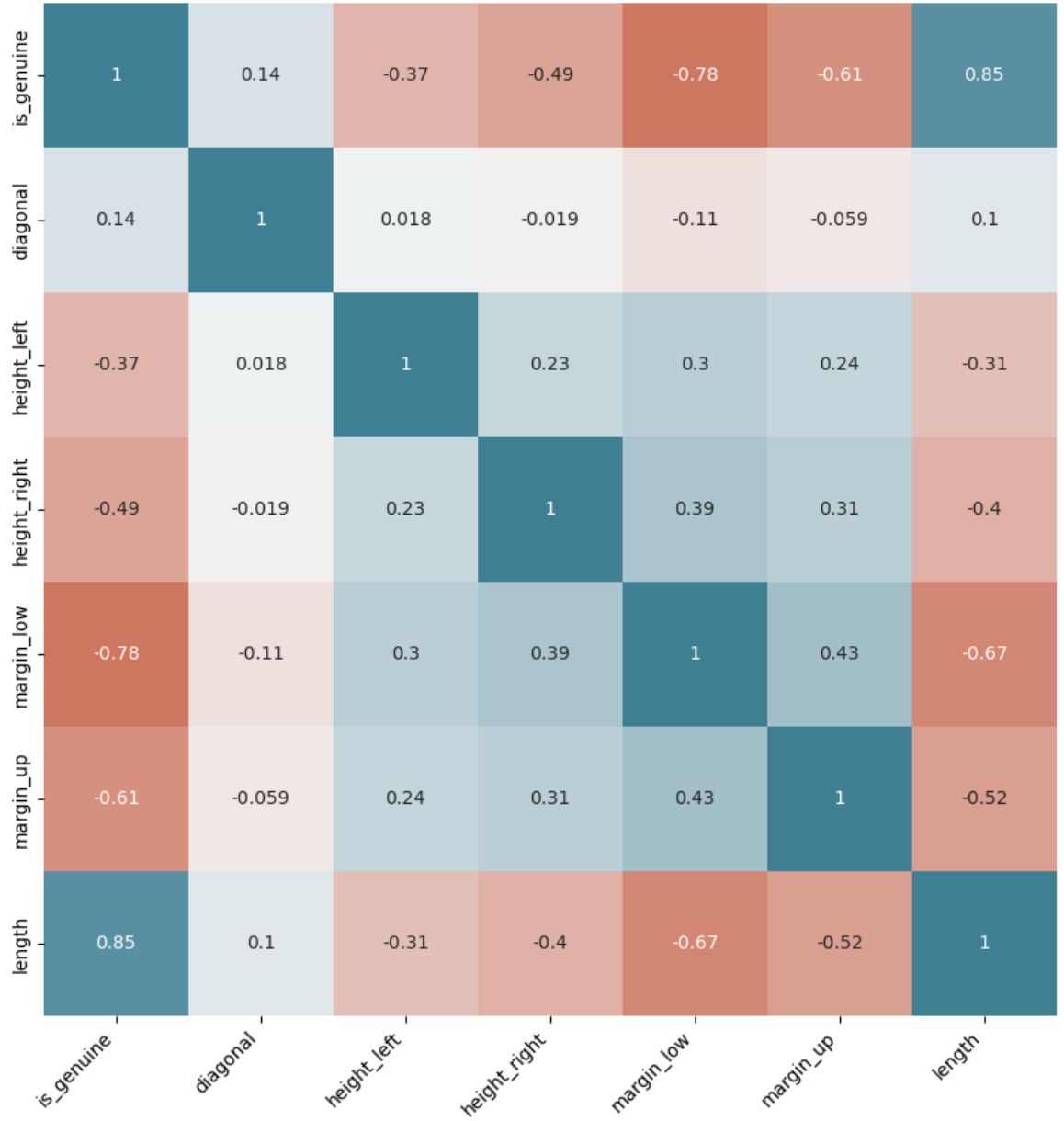


EDA boxplots:
variables non utilisés



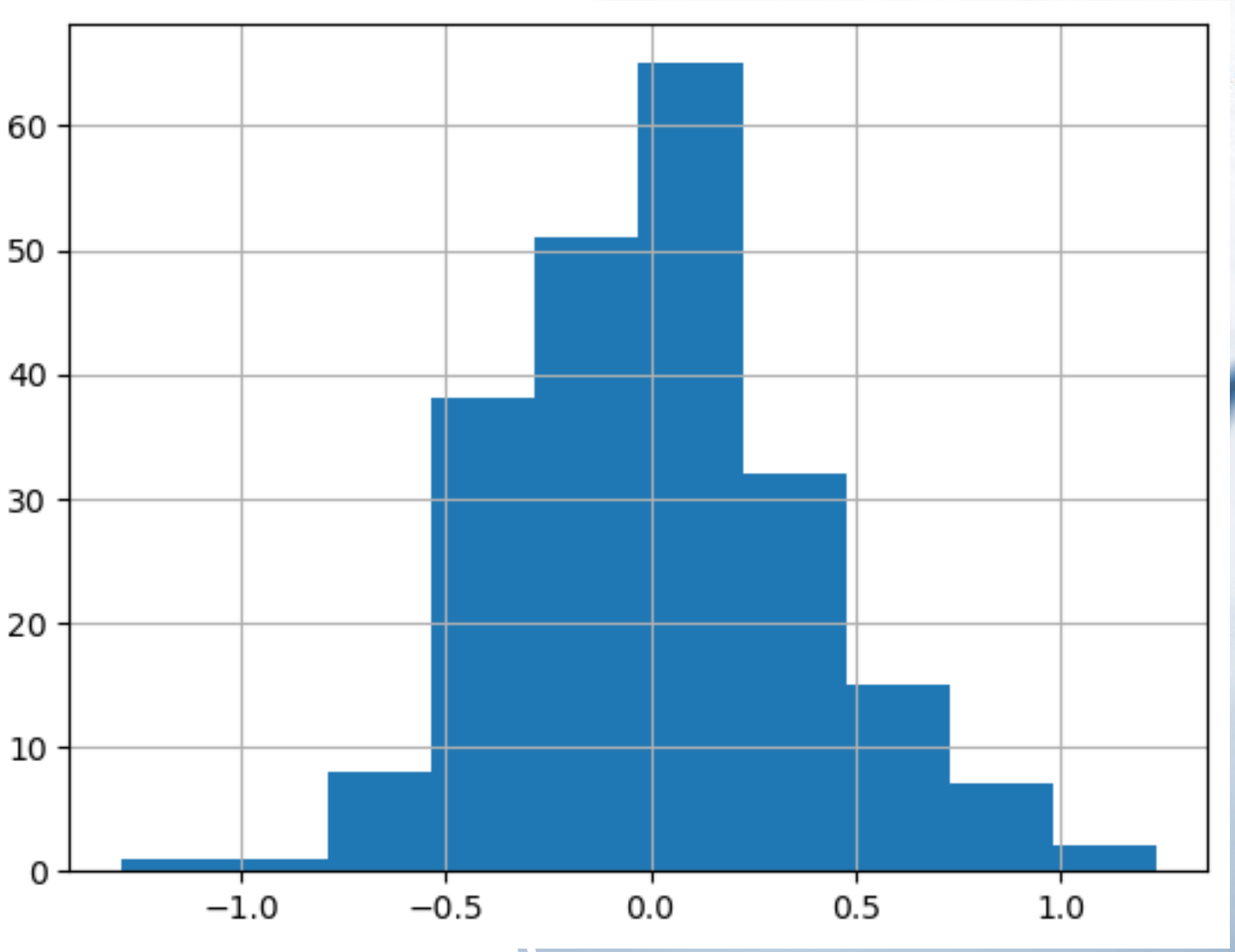


EDA corrélations :



CONSTRUCTION DU MODÈLE: regression linear

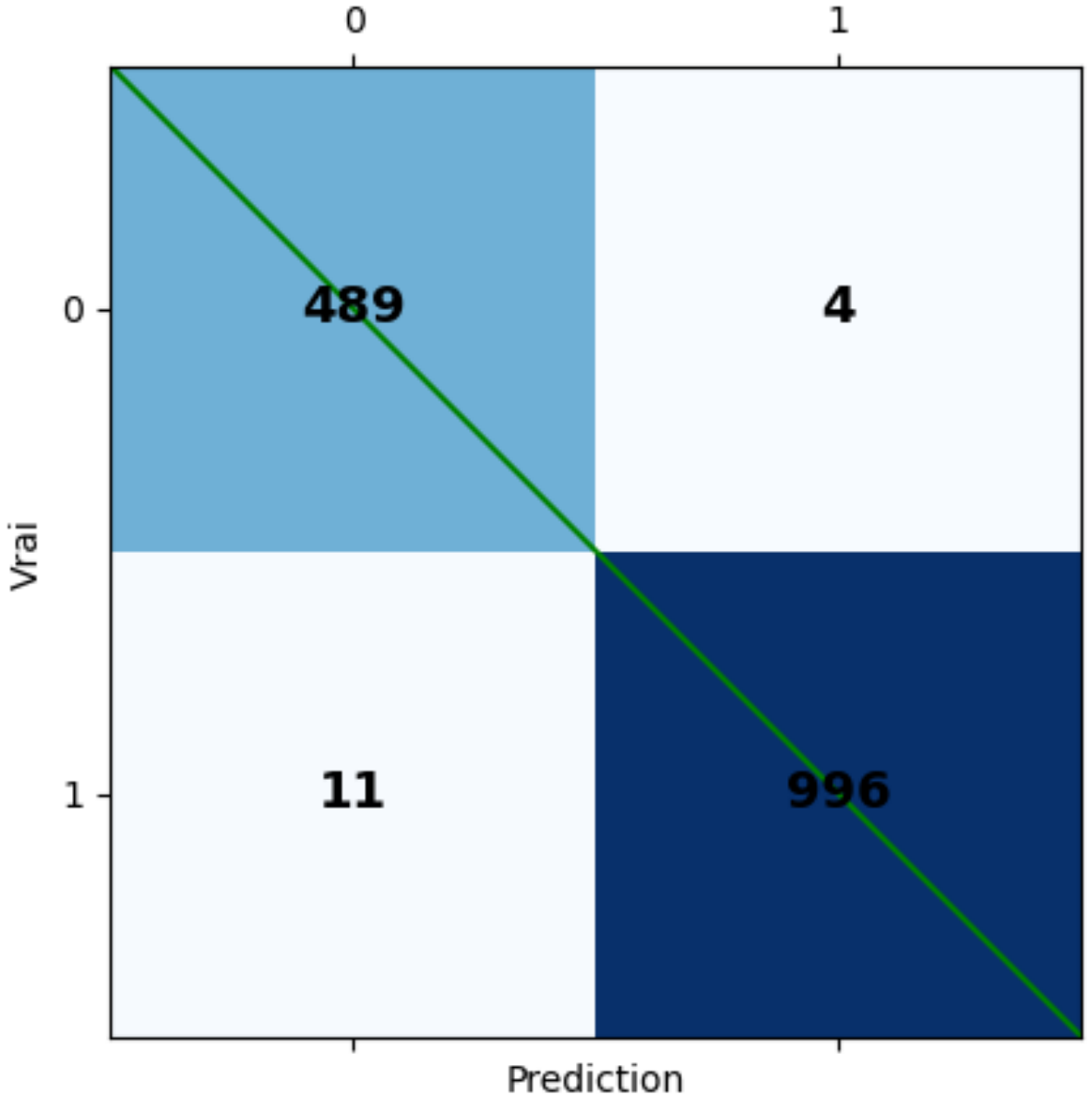
- j'ai utilisé la régression linéaire pour imputer les valeurs manquantes plutôt que d'utiliser d'autres méthodes telles que knnimputer (valeurs proches),
- Car elle présentait une distance plus faible par rapport à la moyenne (0.31 MAE).



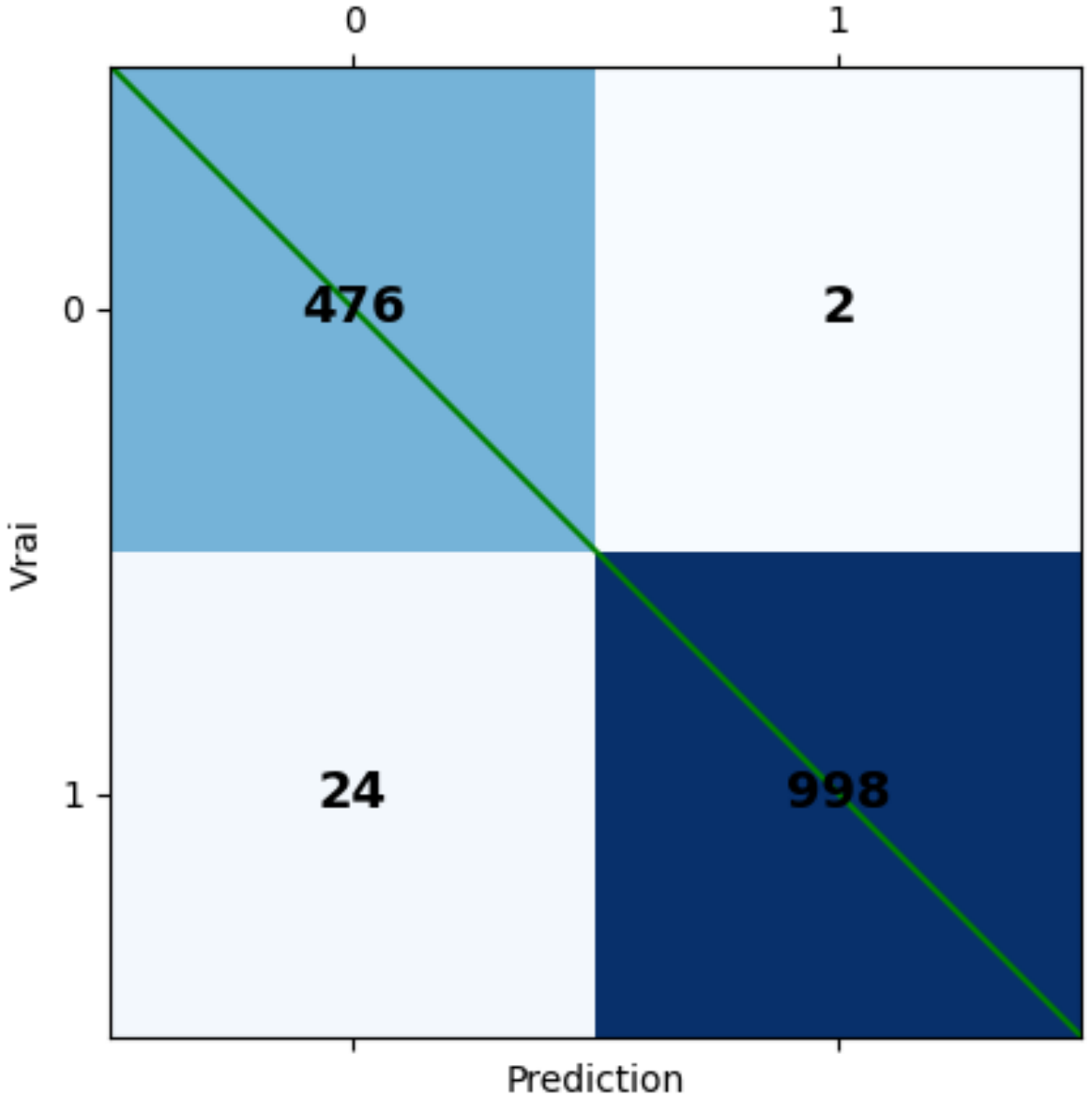
CONSTRUCTION DU MODÈLE:

- J'ai construit le modèle avec ces 4 variables:
  - margin_low
  - length
  - margin_up
  - height_right
- La régression logistique a été plus fiable que la méthode K-means (99% vs 98%)
- J'ai donc choisi la régression logistique avec 4 des 6 variables pour construire le modèle, parce que le P value de 2 variables était haut.

MATRICE DE CONFUSION en regression logistique:



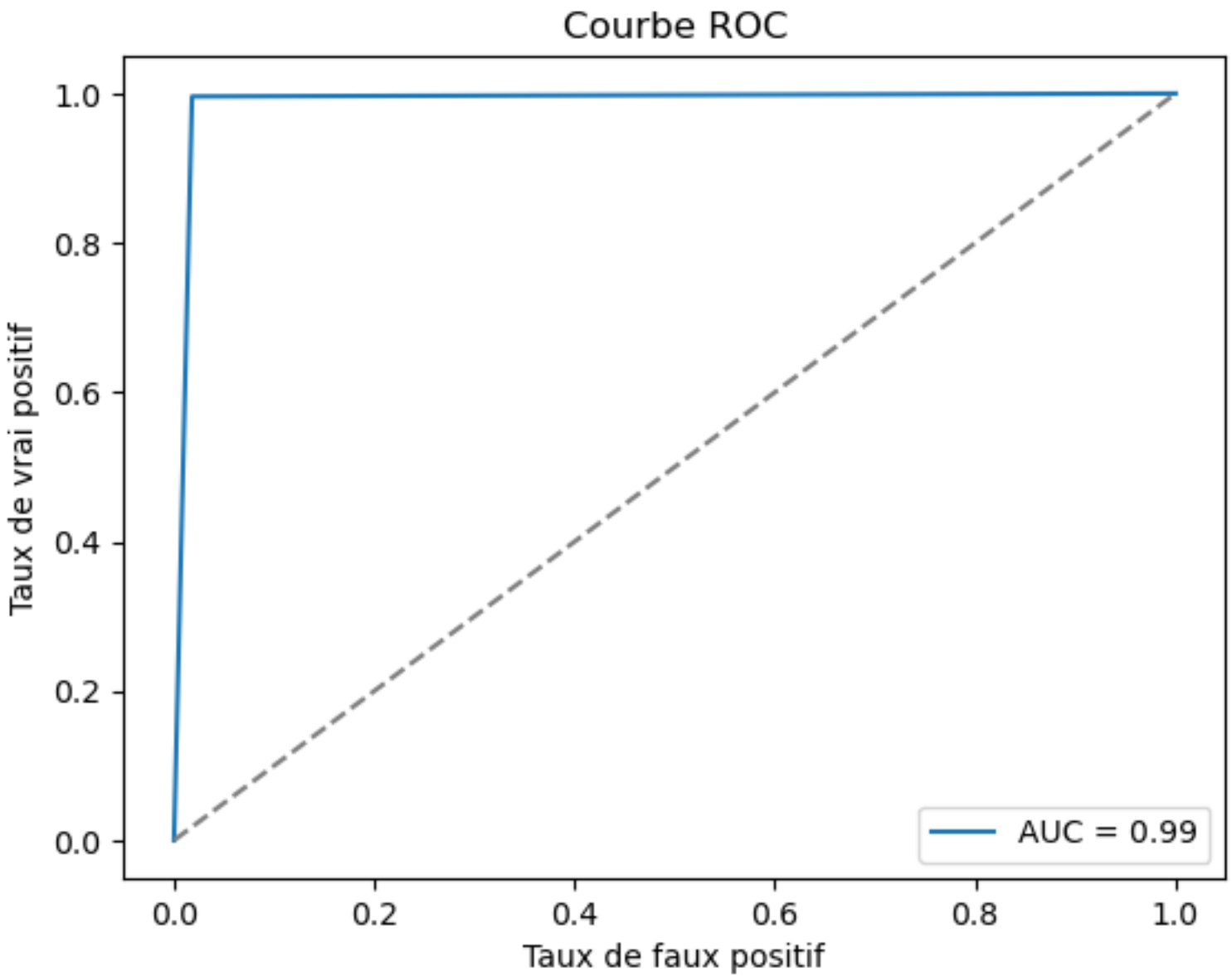
MATRICE DE CONFUSION en kmeans:



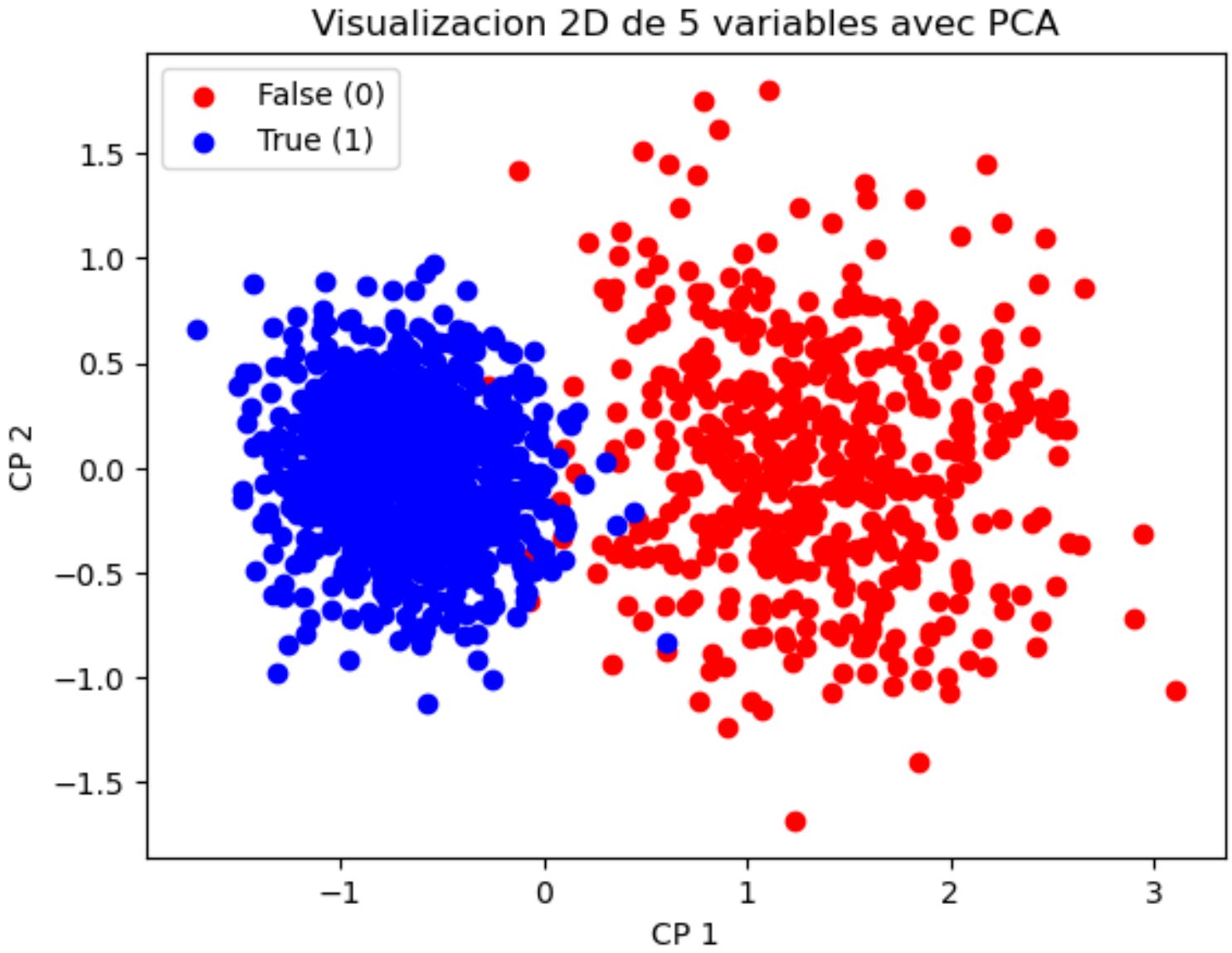
MODELE LOGISTIQUE. METRIQUES:

	precision	recall	f1-score	support
False	0.99	0.98	0.99	500
True	0.99	1.00	0.99	1000

MODELE LOGISTIQUE. COURBE ROC:



ACP:


Visualizacion 2D de 5 variables avec PCA
False (0)
True (1)
CP 2
CP 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2
3

CONCLUSIONS

- L'analyse en composantes principales (ACP) a été utile pour visualiser les données, réduire la dimensionnalité et identifier les variables les plus importantes.
- La régression linéaire a permis d'imputer les valeurs manquantes de manière satisfaisante et d'améliorer la qualité des données.
- La régression logistique a été efficace pour construire l'algorithme final et identifier les vrais et les faux billets.

Resumen

- Pour conclure, la combinaison de ces méthodes a permis de construire un modèle performant pour identifier les vrais et les faux billets avec une précision élevée. Il est important de souligner que d'autres méthodes pourraient également être explorées pour améliorer encore davantage la performance de l'algorithme, clustering, par example, mais moins performantes en une premiere regard que la regresion logistique.