Machine Learning - Détecteur de Faux billets bancaires

Prologue

- Contexte d'étude :
  - Mission de consulting au sein de l'Organisation nationale de lutte contre le faux-monnayage (ONCFM).
- Objectif :
  - Créer un modèle prédictif capable d'identifier automatiquement les faux billets en analysant les dimensions géométriques et les caractéristiques constitutives.

Sommaire

- Données 4
- Analyse Exploratoire des Données 5
- Méthodes de classification – Supervisées 11
- Clustering – Non Supervisée 15
- Réduction de dimensionalité 18
- Conclusion 22

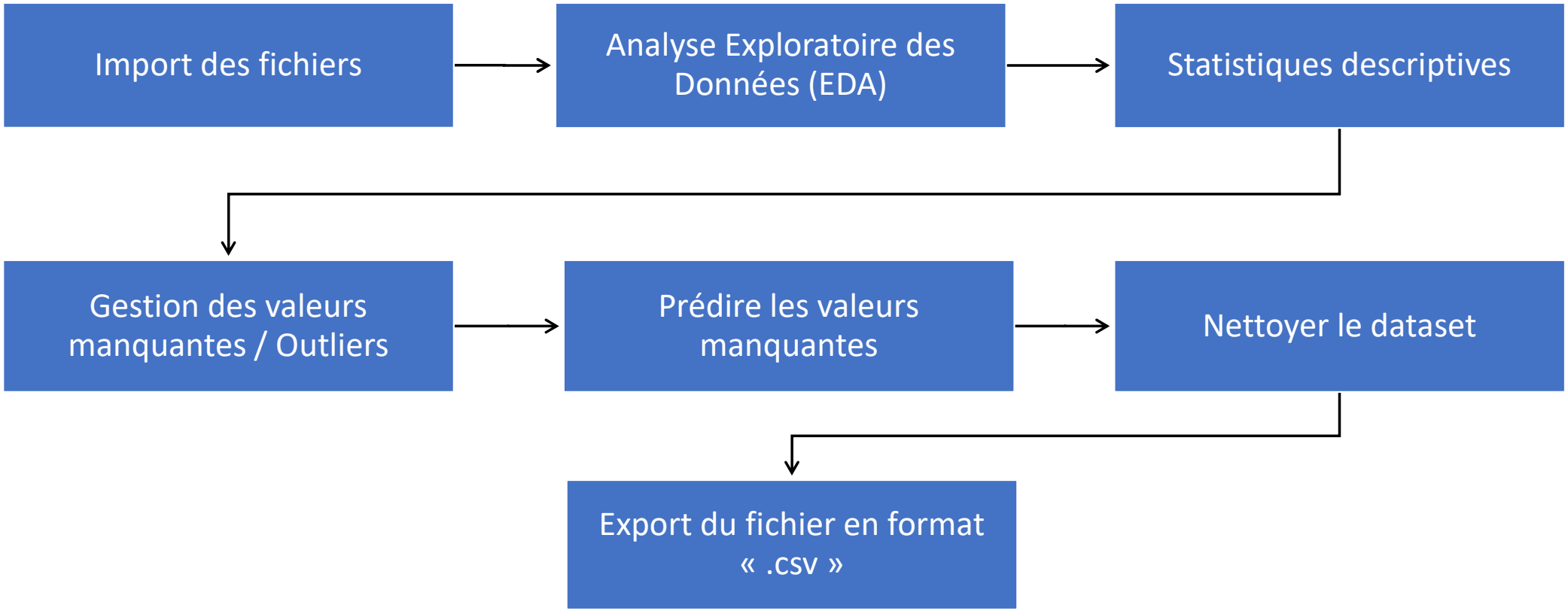
Données (ONCFM)

Composition du dataset :

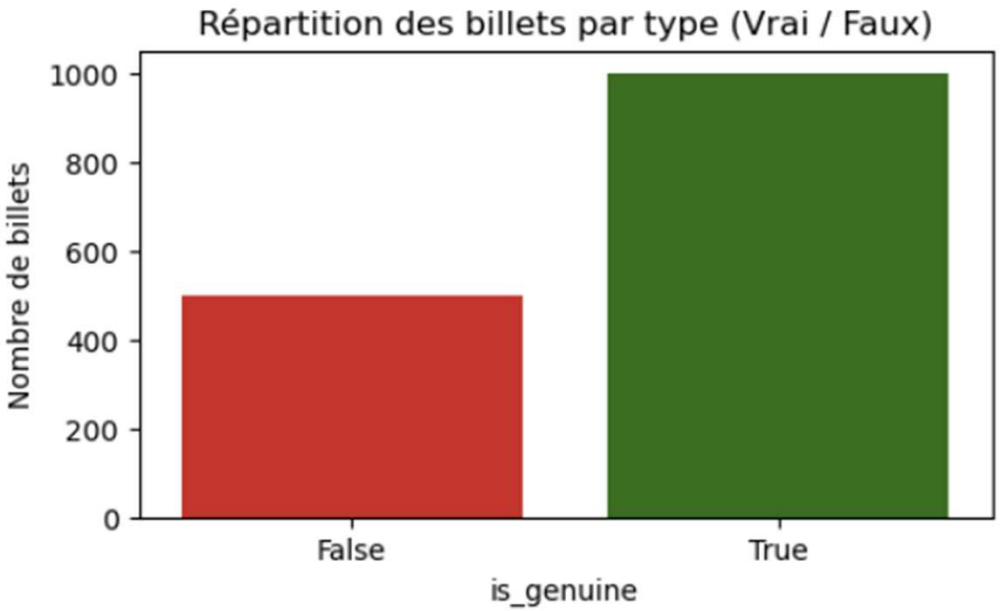
- 1 variable binaire (Vrai, Faux billet)
  - Is_genuine
- 6 variables géométriques quantitatives
  - Diagonal
  - Height_left
  - Height_right
  - Margin_low
  - Margin-up
  - Length

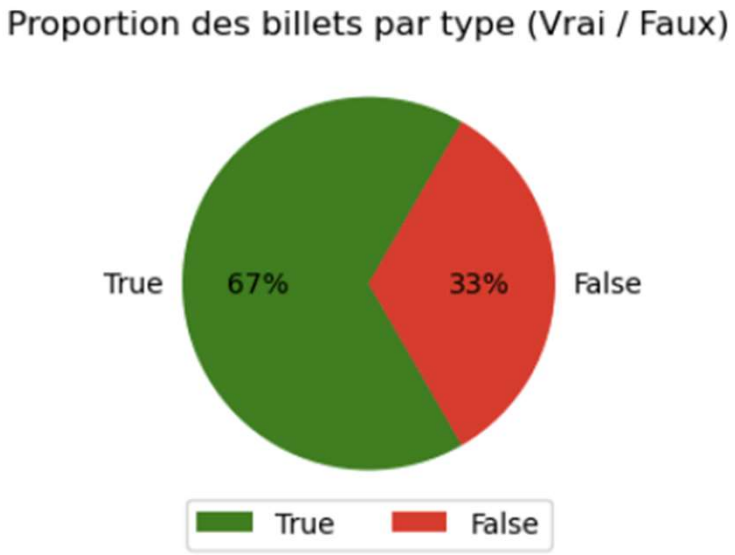
Analyse Exploratoire des Données (AED)

Process (EAD)


Import des fichiers
Analyse Exploratoire des Données (EDA)
Statistiques descriptives
Gestion des valeurs manquantes / Outliers
Prédire les valeurs manquantes
Nettoyer le dataset
Export du fichier en format « .csv »

Analyse Exploratoire des Données

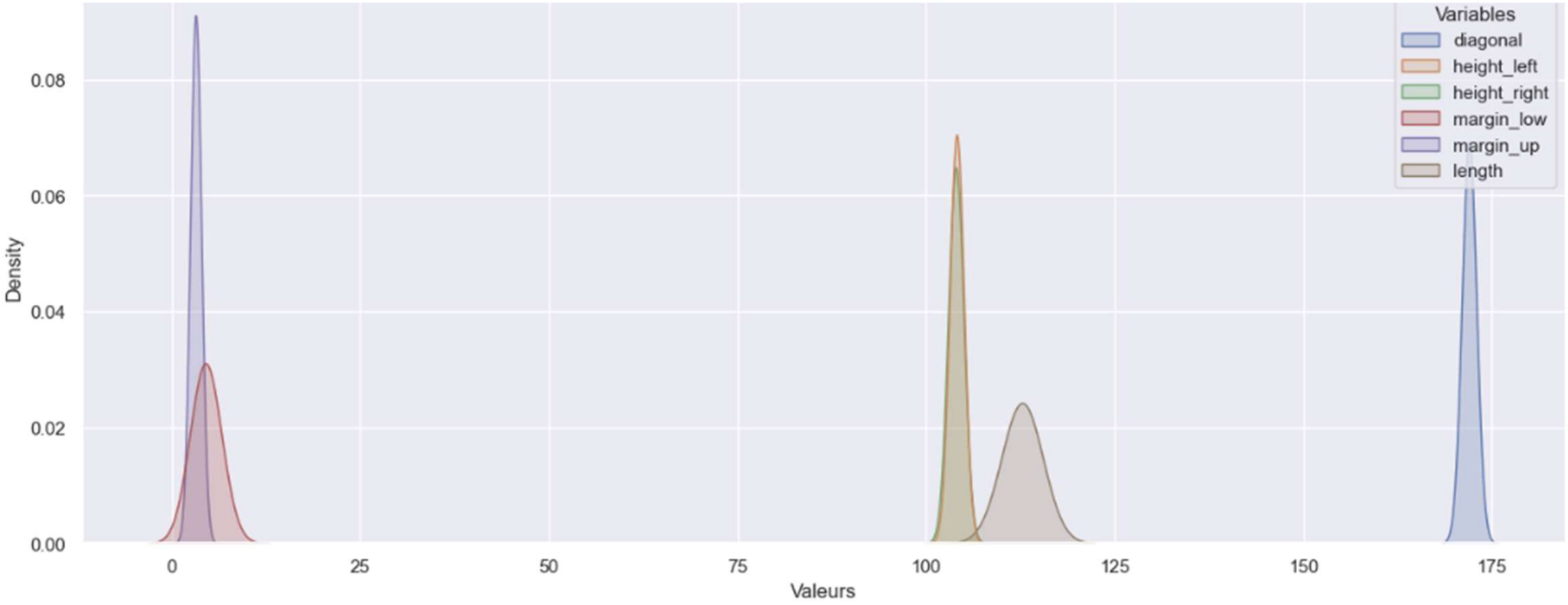




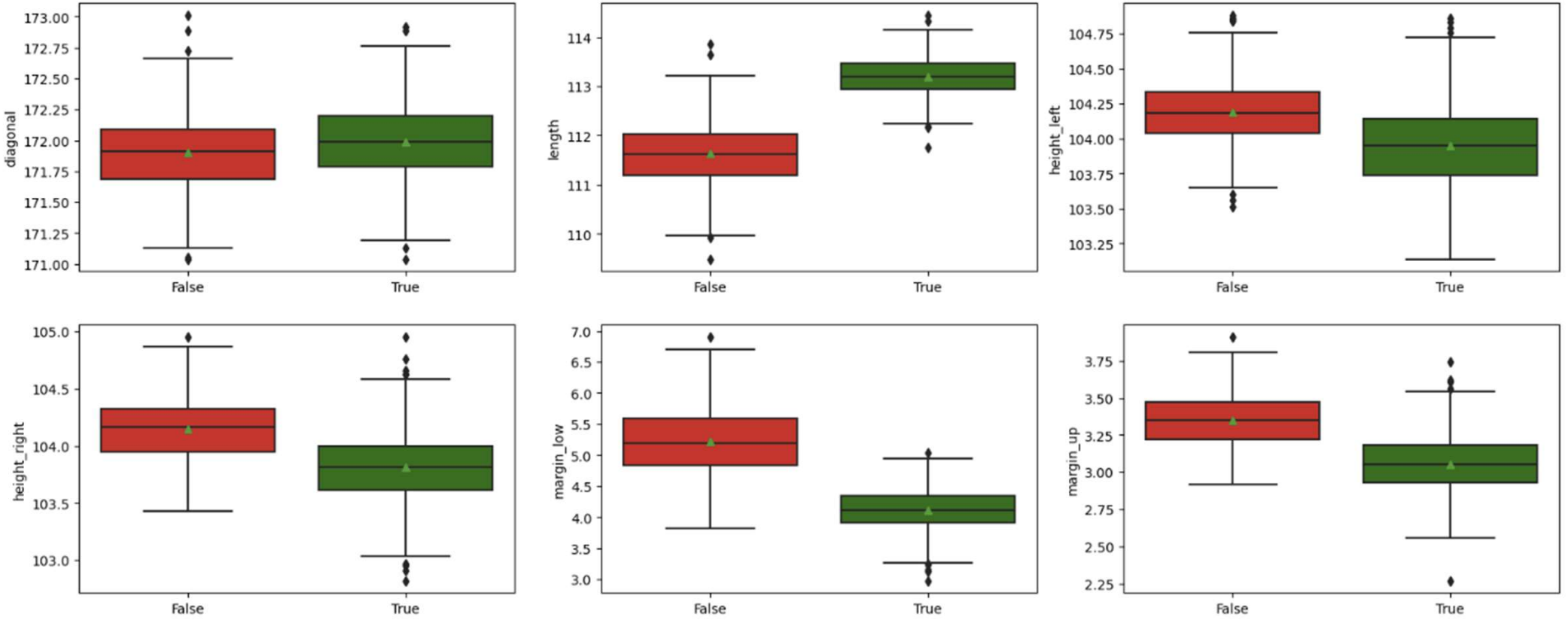
37 Valeurs manquantes

- 29 vrais billets
- 8 faux billets

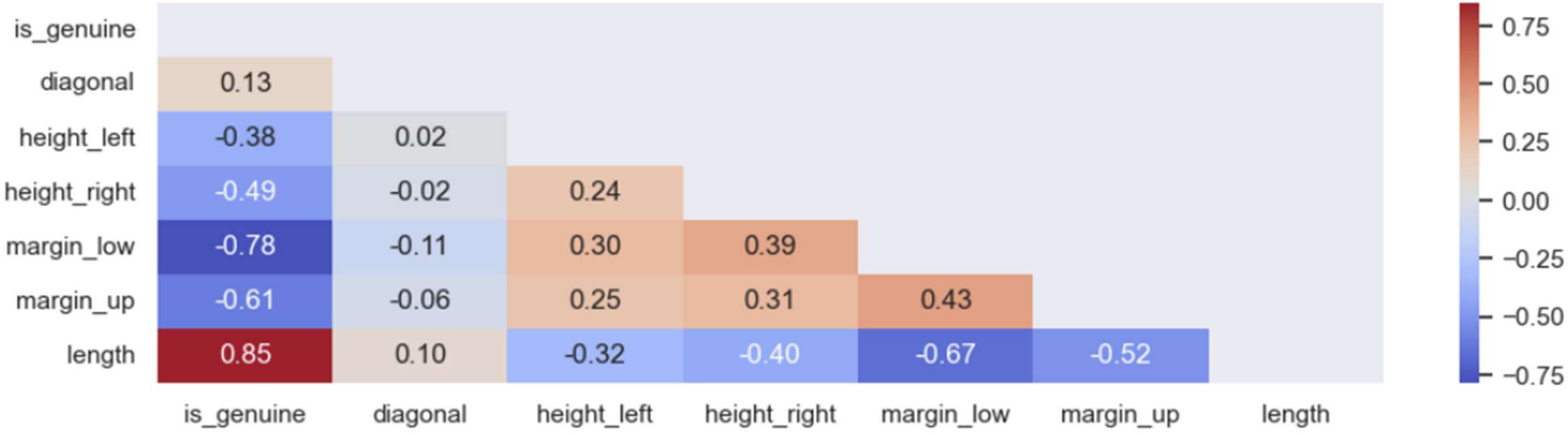
Ordre de grandeurs des variables



Distribution des Vrais / Faux billets



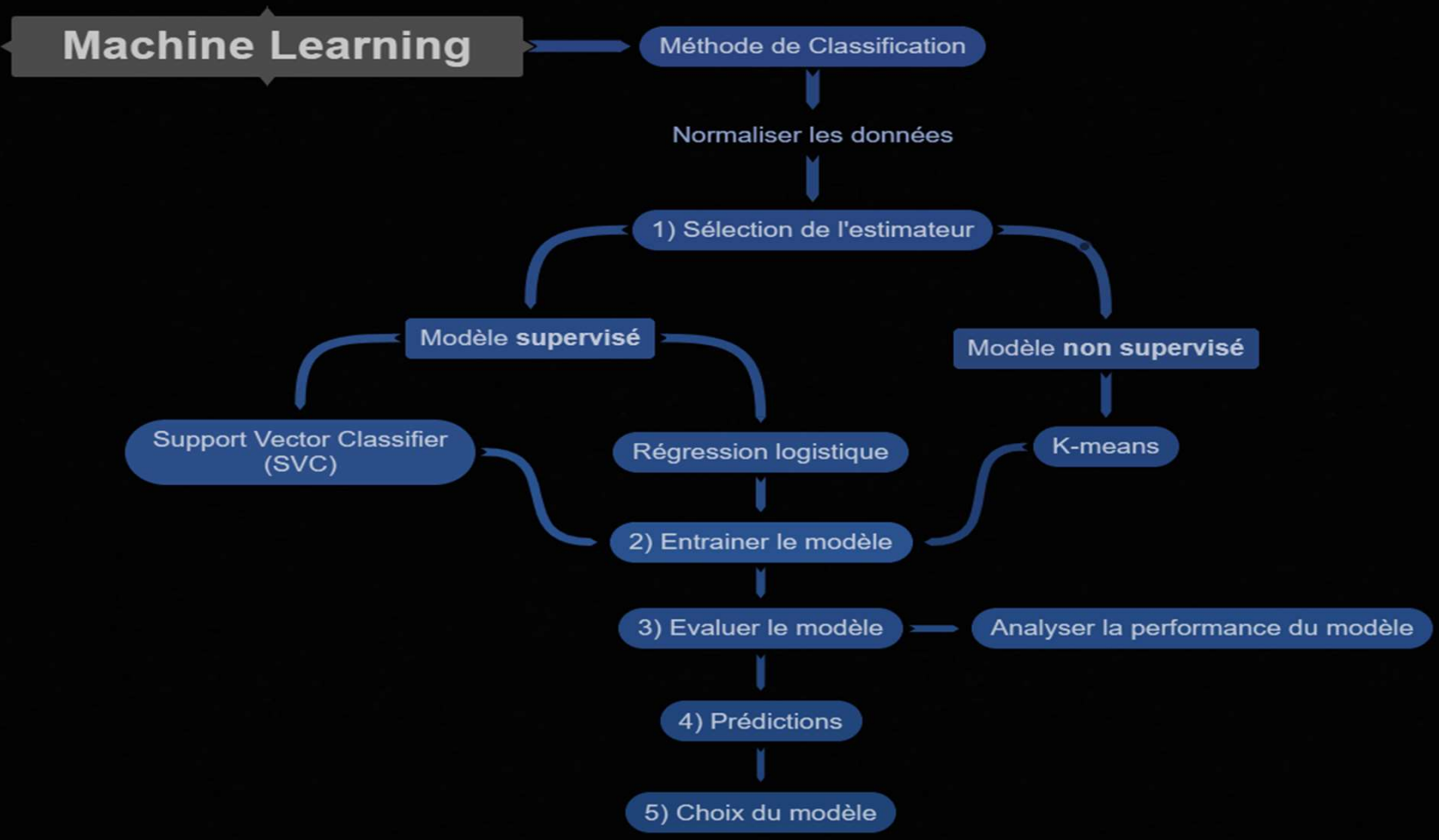
Corrélation



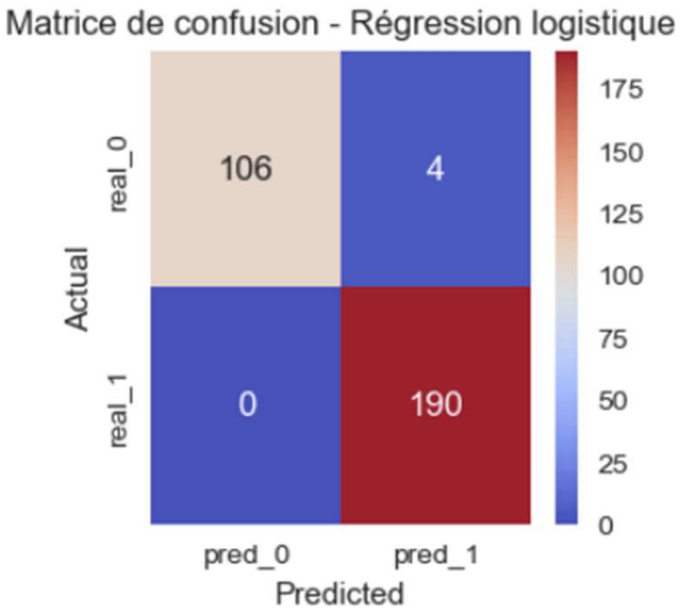
Méthodes de classification Supervisées

Support Machine Classifier (SVC) / Régression logistique

Process (Machine learning)



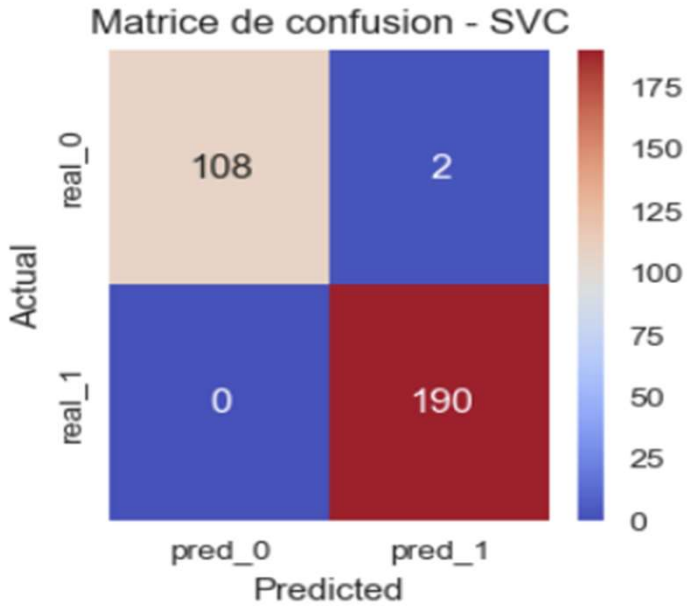
Evaluation des 2 Méthodes de Classifications



Score : 0,989

Marge d'erreur : 1,08%

4 Faux postif (FP)

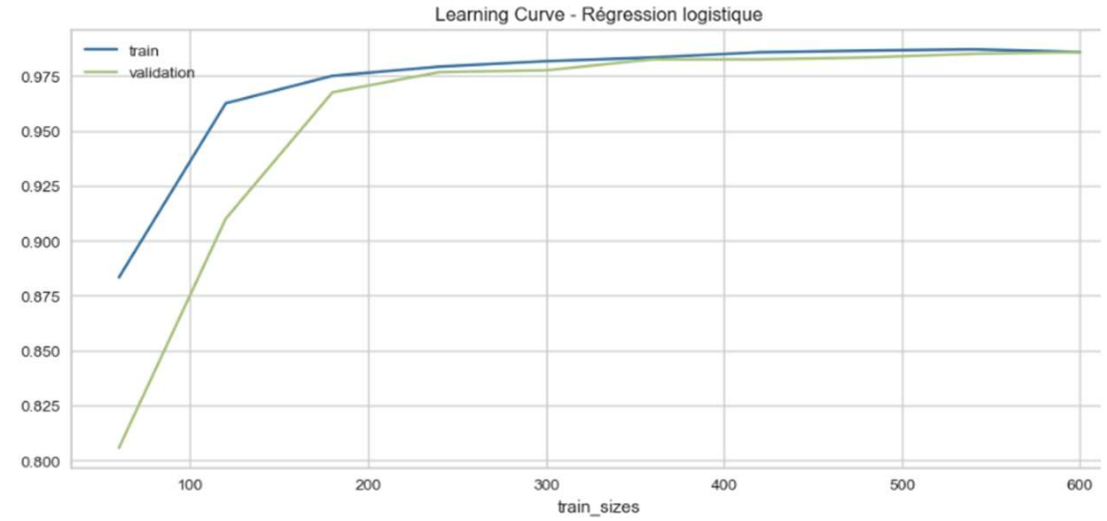


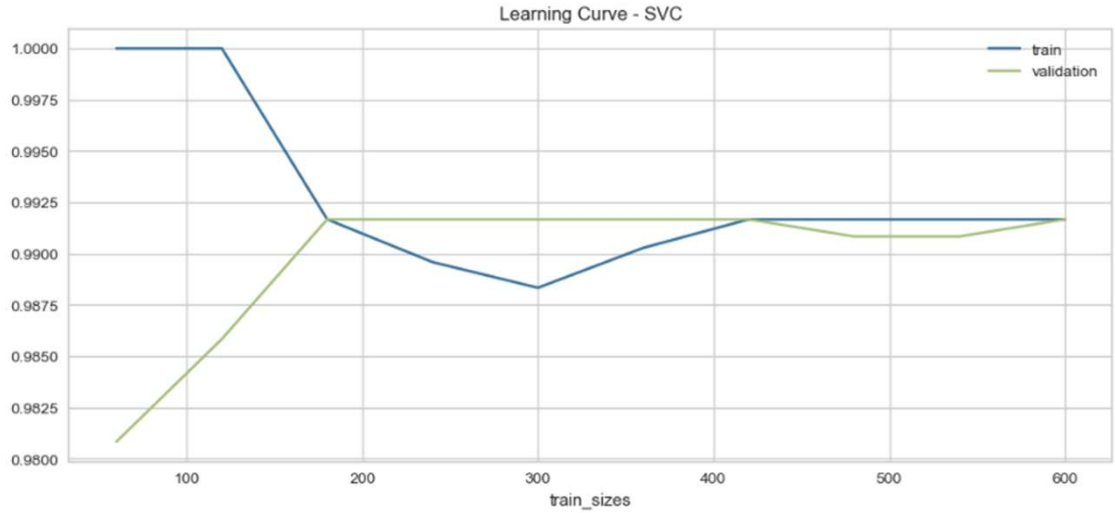
Score : 0,992

Marge d'erreur : 0,75%

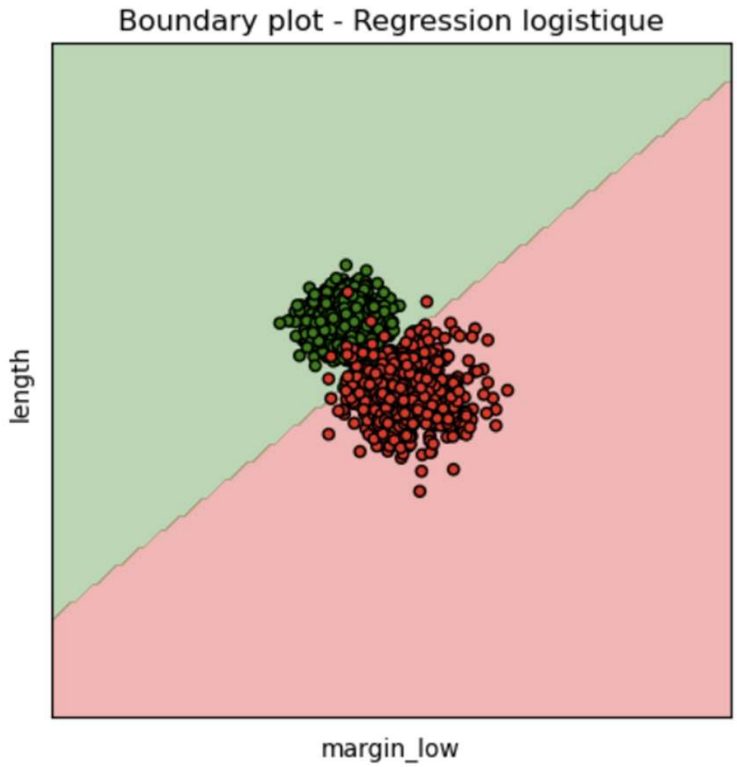
2 Faux postif (FP)

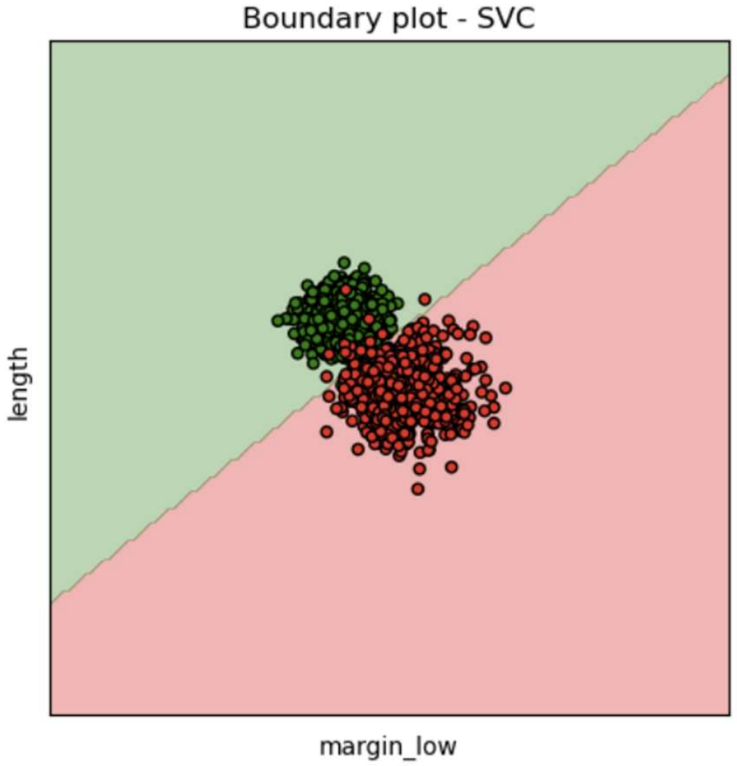
Learning Curve – Régression logistique vs SVC





Boundary decision plot



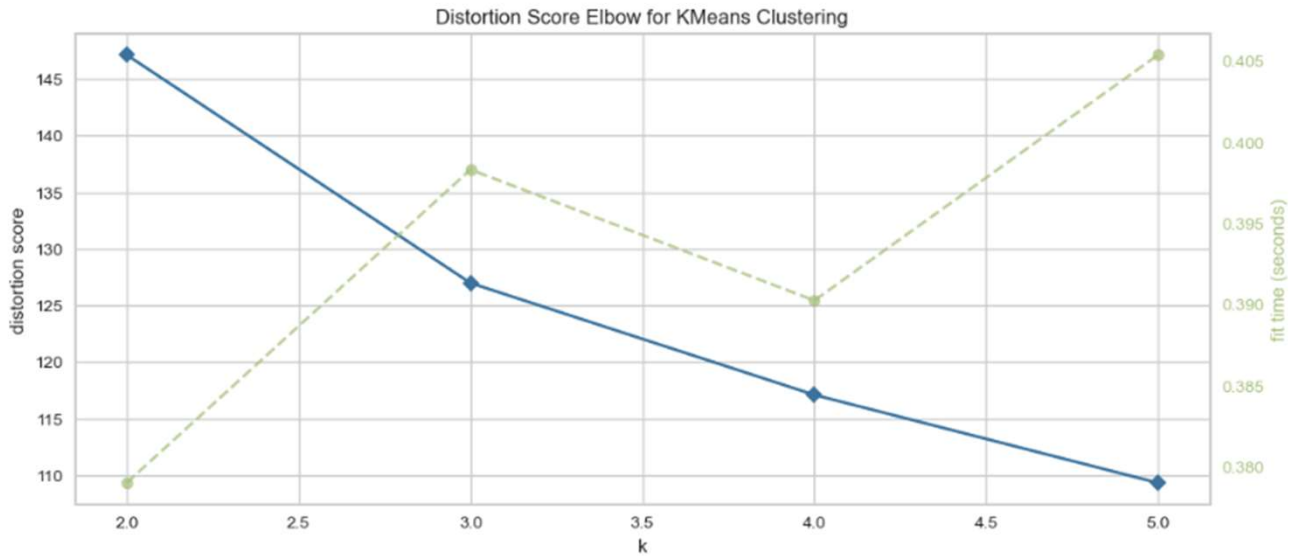


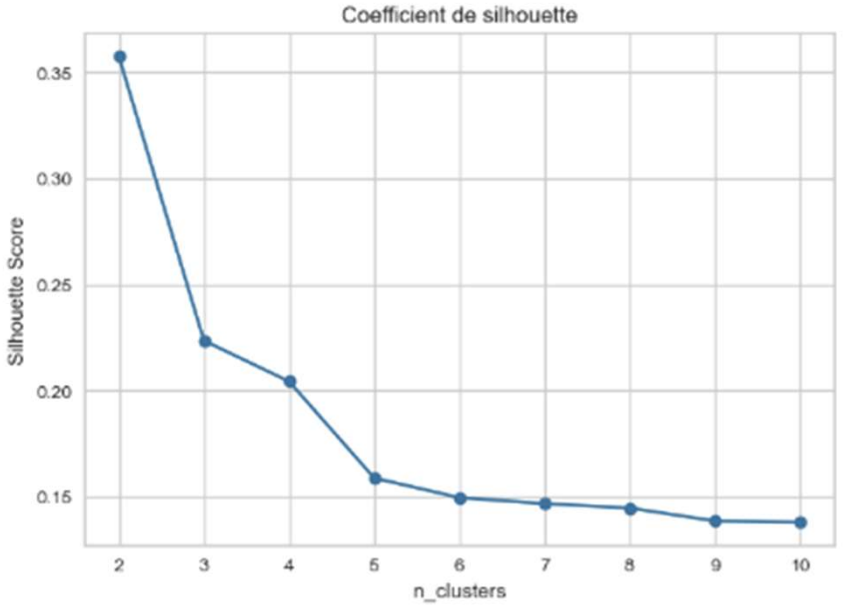
Clustering
Non - Supervisée

K-means

Clustering – K-means

Combien de clusters avons-nous ?

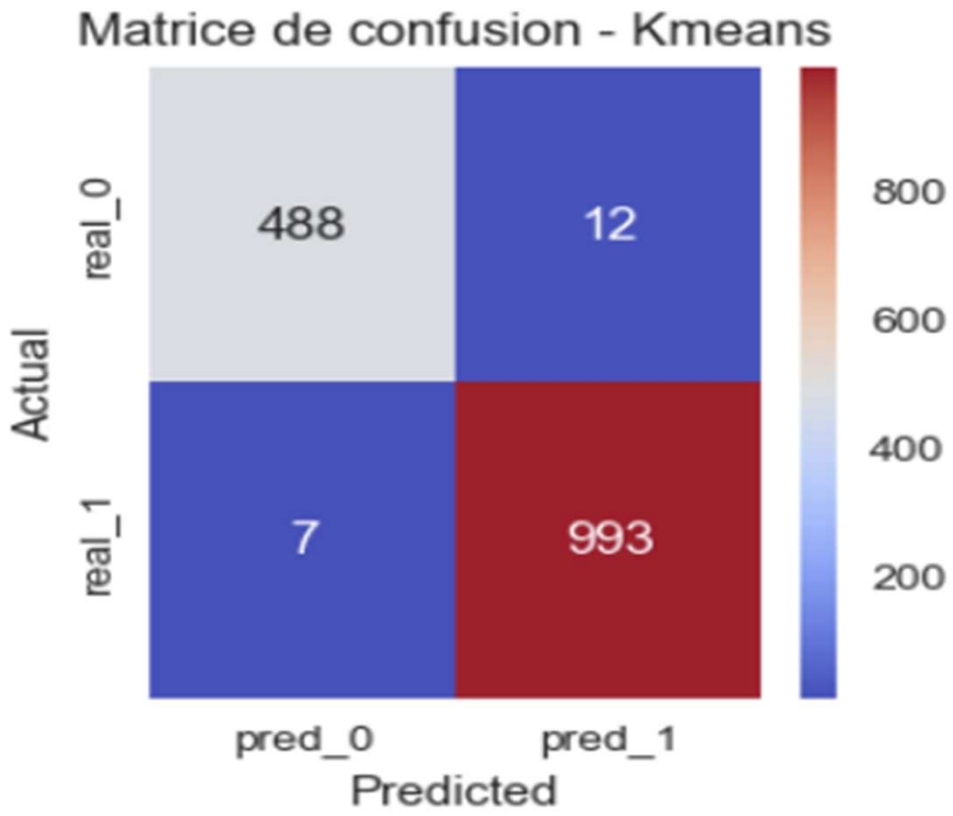




Bien que nous n'ayons de fait que 2 classes, vérifions que l'algorithme du K-means converge:

N_cluster = 2

Evaluation du modèle K-means



Score : 0,987

Marge d'erreur : 1,27%

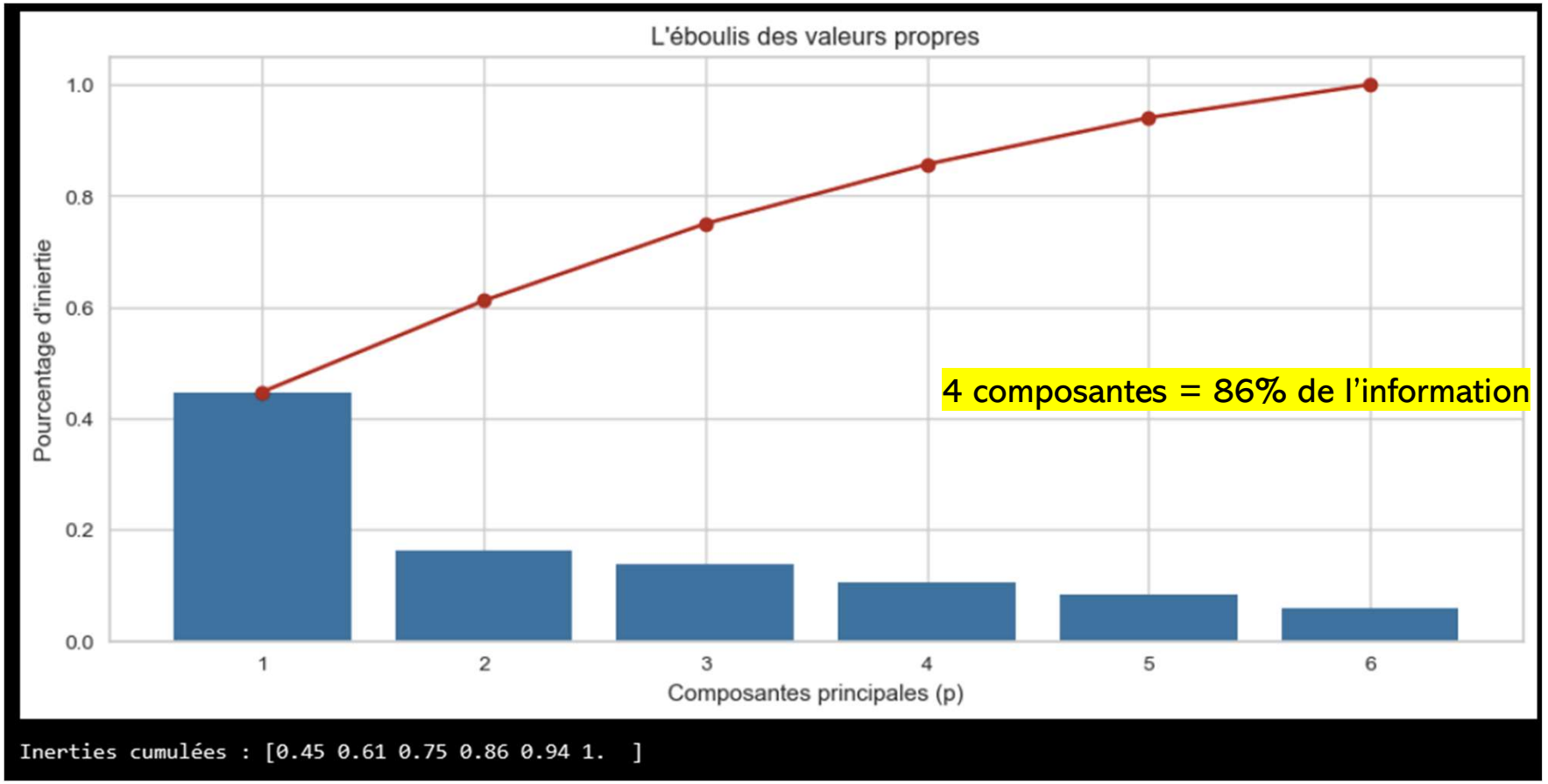
12 Faux postif (FP)

7 Faux négatif (FN)

Réduction de dimensionalité

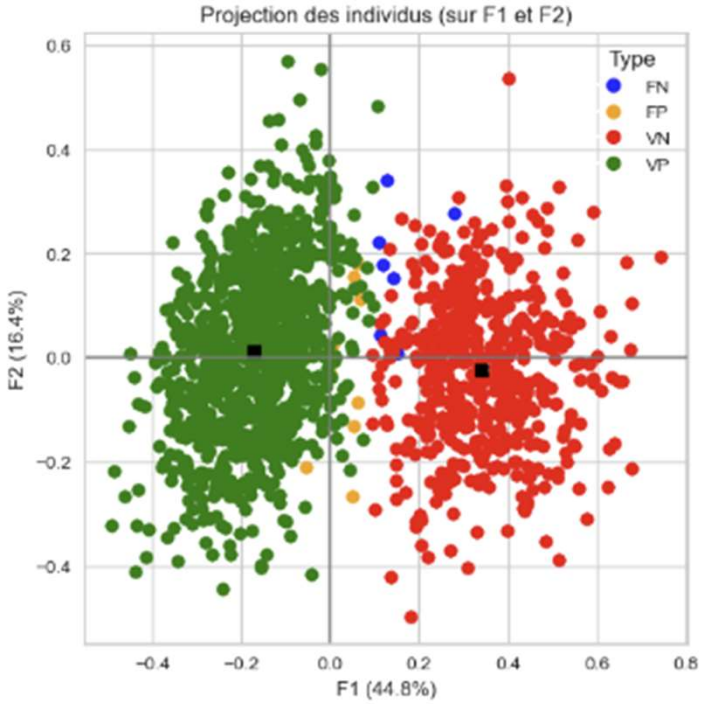
Analyse en Compoantes Principales (ACP)

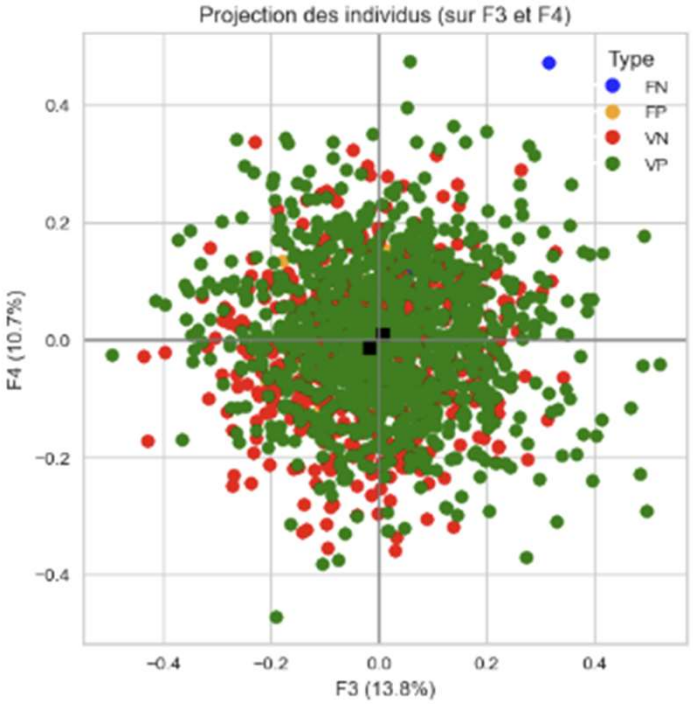
Analyse en Composantes Principales (ACP)

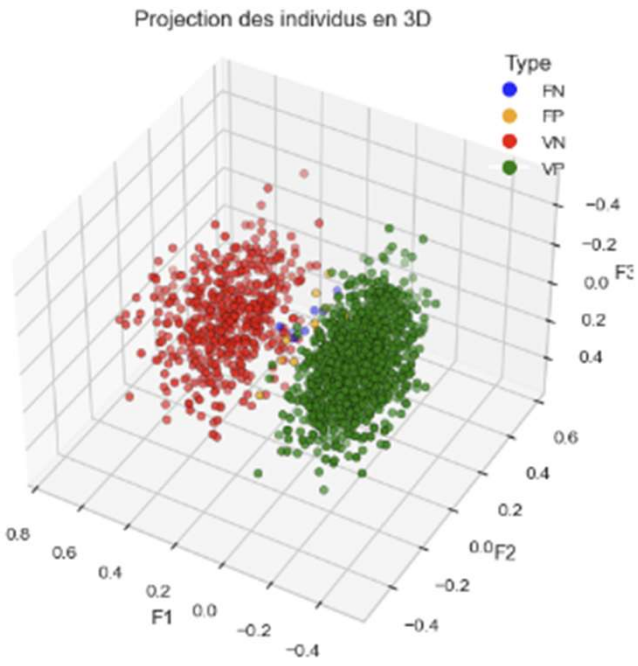



```
Inerties cumulées : [0.45 0.61 0.75 0.86 0.94 1.  ]
```

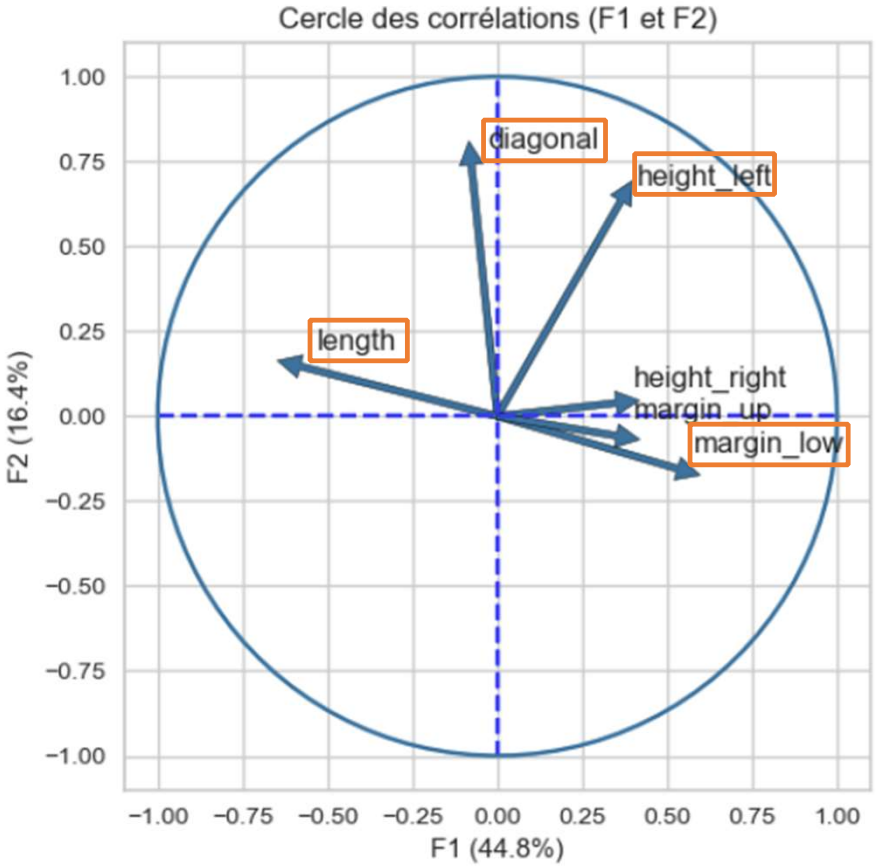
Visualisation des individus en 2D, 3D

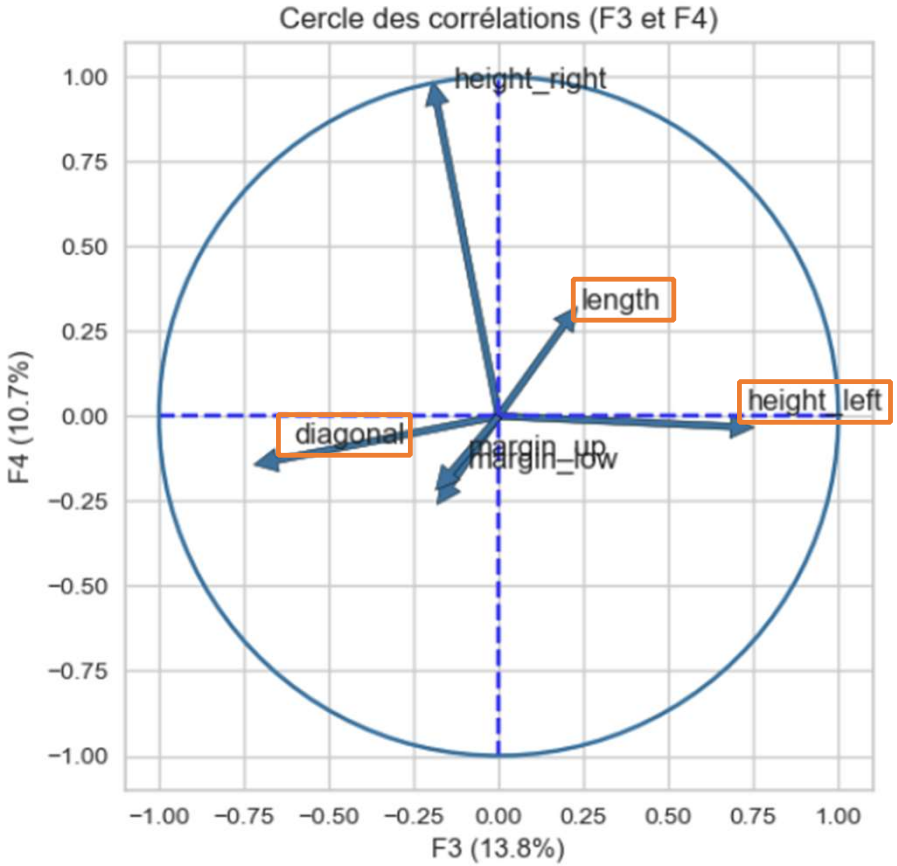






Plans factoriels





Conclusion – Choix du modèle

	SVC	Régression logistique	K-means
Score	0.992	0.989	0.987
Marge d'erreur	0.75%	1.08%	1.27%
Faux positif (FP)	2	4	12
Faux négatif (FN)	0	0	7
Performance	450	250	

Compararaison des scores :

- SVC plus performante que la Régression Logistique (différence minime de 0.33%).
- La classification faite avec le K-means est en revanche non adaptée.

Performance des modèles :

- Learning Curve en faveur de la régression logistique.

Le modèle de régression de **SVC** est celui qu'il faut prendre en compte.