DÉTECTER DES FAUX BILLETS AVEC PYTHON

ONCFM

Mame Diarra **DABO**

06/11/2023

PROBLÉMATIQUE

Consultante Data Analyst

ONCFM

L'objectif est « de mettre en place une modélisation qui serait capable d'identifier automatiquement les vrais des faux billets. »

<u>Mission:</u> Demandes de Marie

- Voir les traitements et analyses réalisés en amont;
- Les différentes pistes explorées pour la construction de l'algorithme;
- Ainsi que le modèle final retenu.

Plan

- Importer le jeu de données
- Analyse exploratoire
- Nettoyage des données
- Répondre aux demandes: Modélisation
- Tester le modèle final
- Conclusion: Pistes d’amélioration

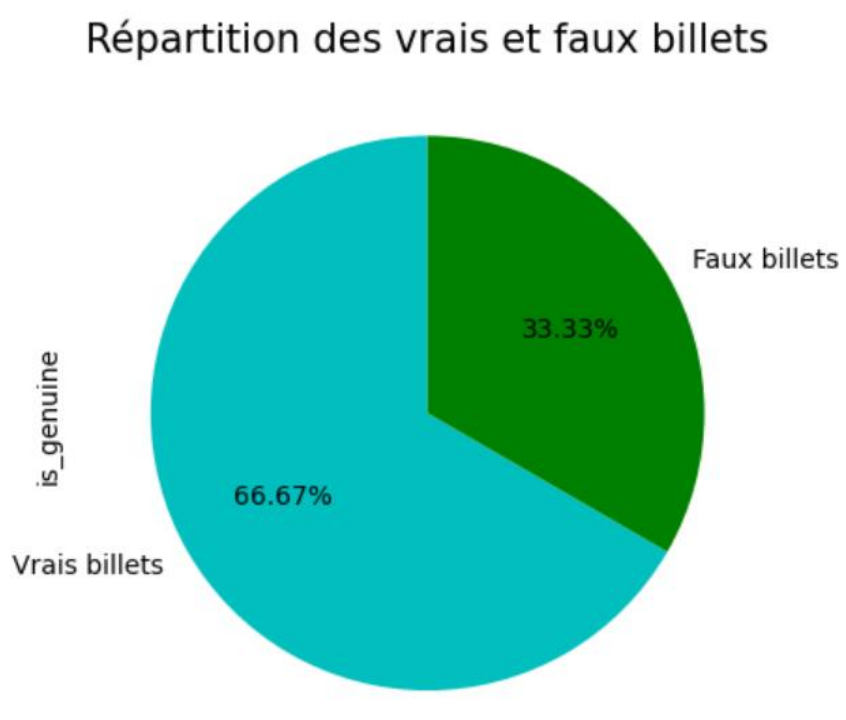
Le jeu de données

- *Fichier billets : (1500, 7)*

- Parcourir le jeu de données :

  => les informations géométriques
  => les types de données
  => Vérifier les doublons
  => Vérifier les valeurs manquantes
  => Vérifier les valeurs aberrantes

Analyse Exploratoire: statistique descriptive

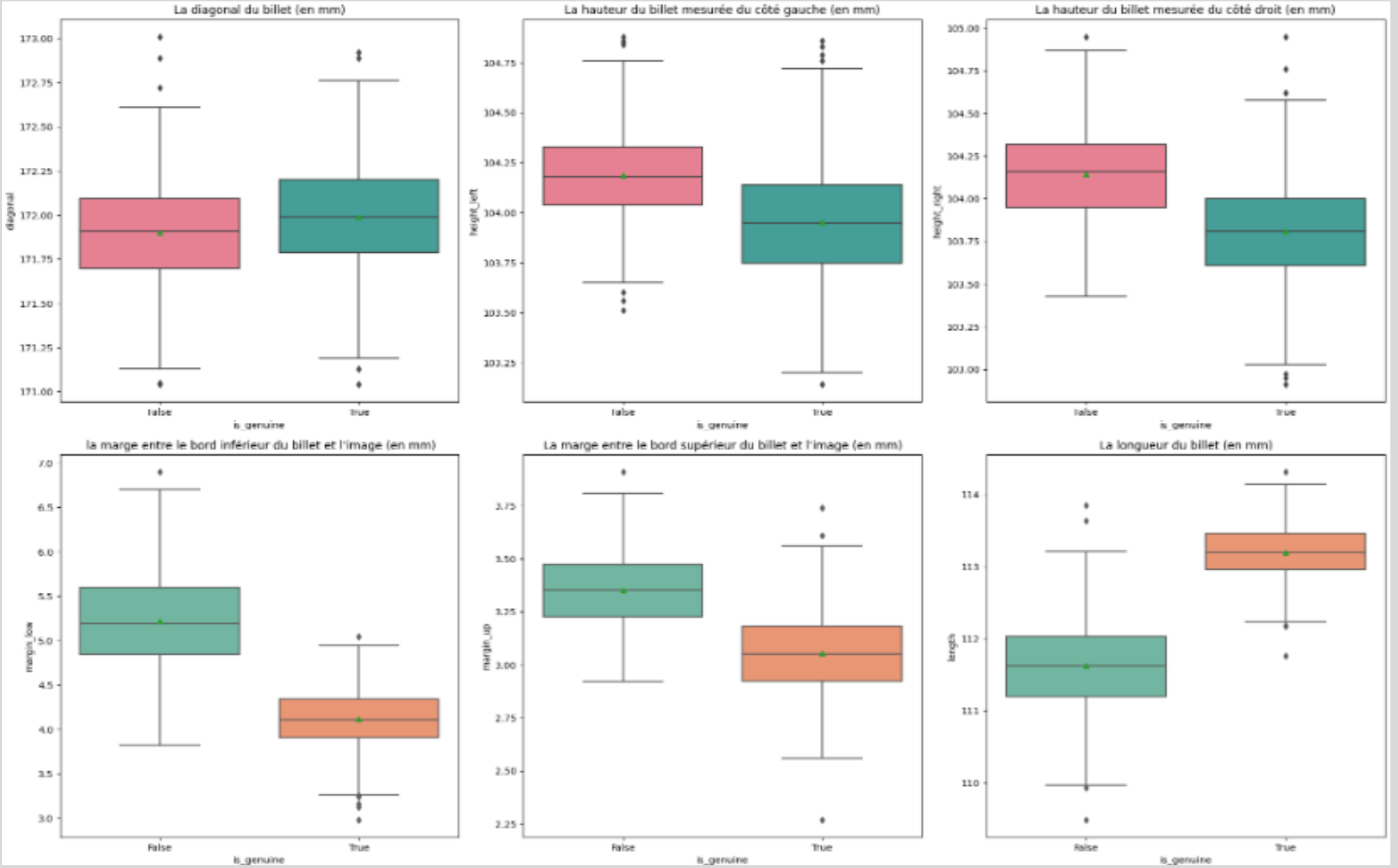


	count	mean	std	min	25%	50%	75%	max
diagonal	1500.0	171.958440	0.305195	171.04	171.750	171.96	172.17	173.01
height_left	1500.0	104.029533	0.299462	103.14	103.820	104.04	104.23	104.88
height_right	1500.0	103.920307	0.325627	102.82	103.710	103.92	104.15	104.95
margin_low	1463.0	4.485967	0.663813	2.98	4.015	4.31	4.87	6.90
margin_up	1500.0	3.151473	0.231813	2.27	2.990	3.14	3.31	3.91
length	1500.0	112.678500	0.872730	109.49	112.030	112.96	113.34	114.44

Sur les 1500 billets, nous avons 66.67% qui représentent les vrais billets (les 1000 billets) et 33.33% (les 500 billets) qui représentent les faux billets .

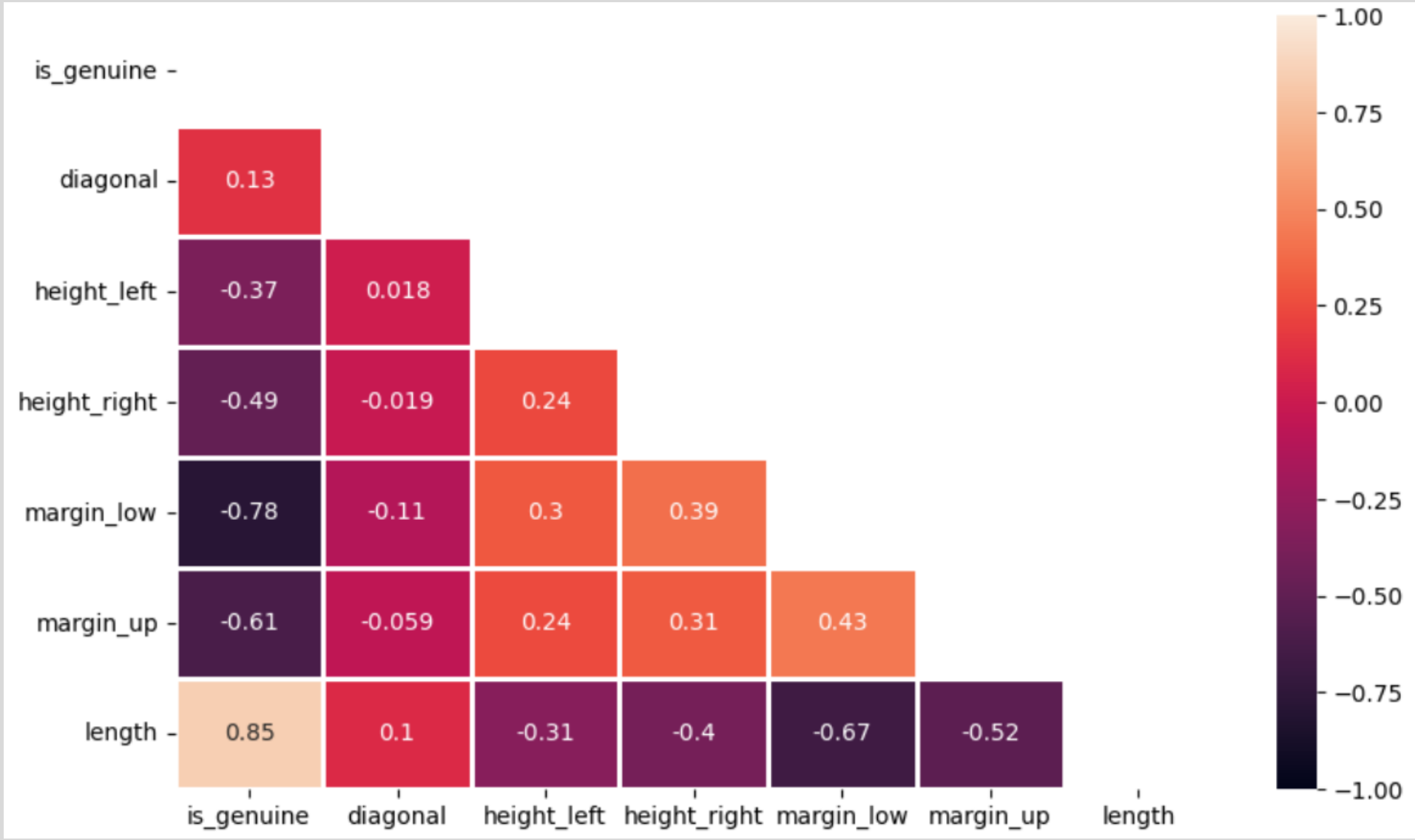
Analyse exploratoire: Types de billets

Les types de billets (nature des billets) avec différentes dimensions.



Analyse exploratoire: relation entre les dimensions

Observer les relations qui existent entre les différentes variables.



Nettoyages des données: Traitement

Afficher chaque jeu de données

- **Gestion des valeurs**

  => Convertir au bon format de données
  => Remplacer les valeurs manquantes avec le modèle de la régression logistique
  => Traitement des valeurs aberrantes par la médiane

- **aberrantes :**

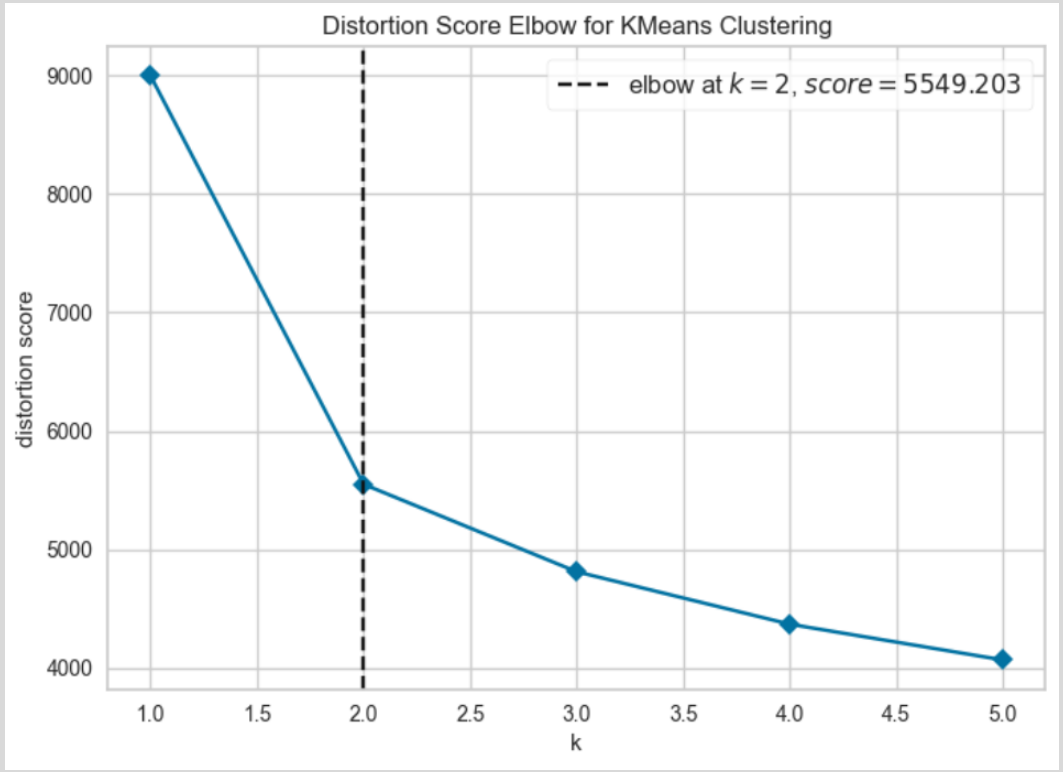
  => aucun outliers semblent être erronés

Fichier final : 1500 billets

7 caractéristiques

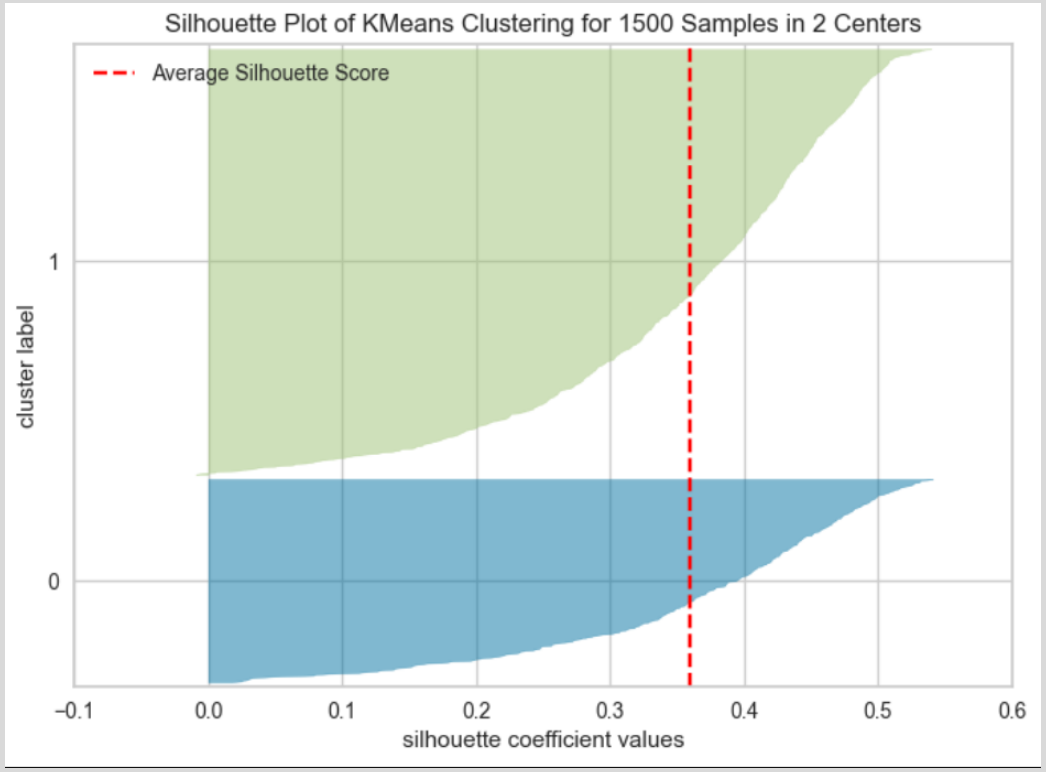
Modélisation: K-means Clusters

Distorsion score Elbow



Grâce à la méthode du coude basée sur le score de distorsion *(somme moyenne des carrés des distances aux centres)*, une segmentation en K=2 clusters serait la meilleure option.

Silhouette



Silhouette : *rapport moyen entre la distance intra-cluster et la distance du cluster le plus proche*.

Modélisation: Interprétation K-means Clusters

Interprétation des clusters

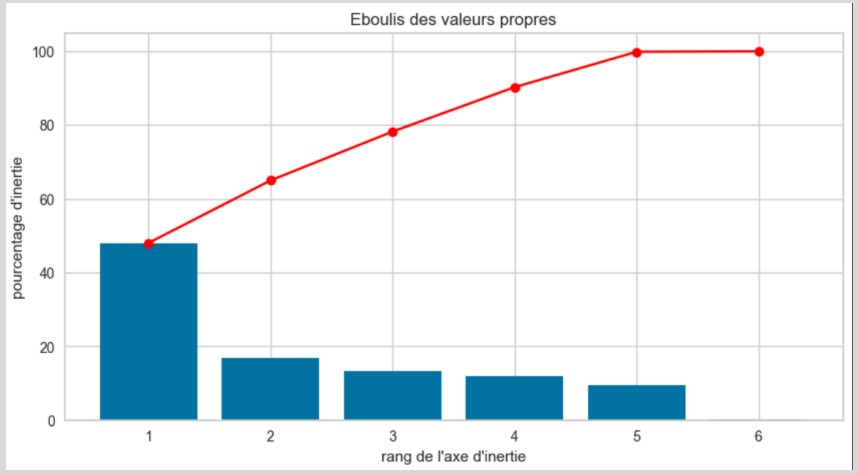
```
Nbre de clusters avec le K-means:
 1    1015
0     485
```

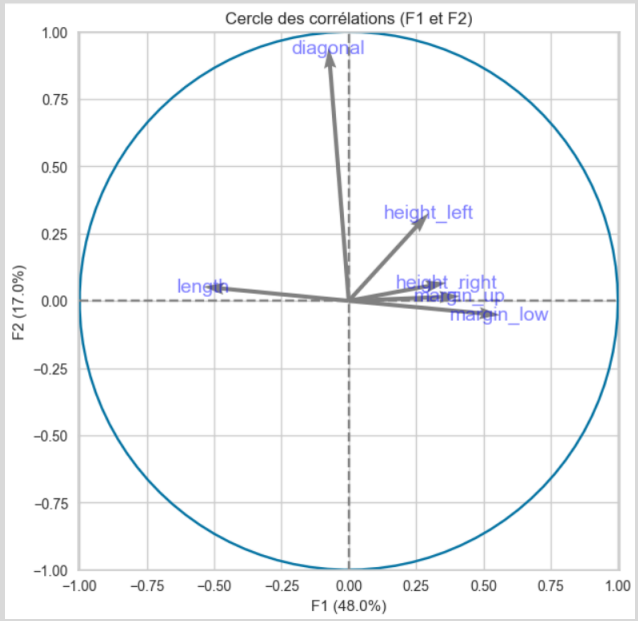
	Cluster_Billets	diagonal	height_left	height_right	margin_low	margin_up	length
0	0	-0.181261	0.558728	0.711546	1.276478	0.885465	-1.244885
1	1	0.086613	-0.266979	-0.340000	-0.609943	-0.423104	0.594847

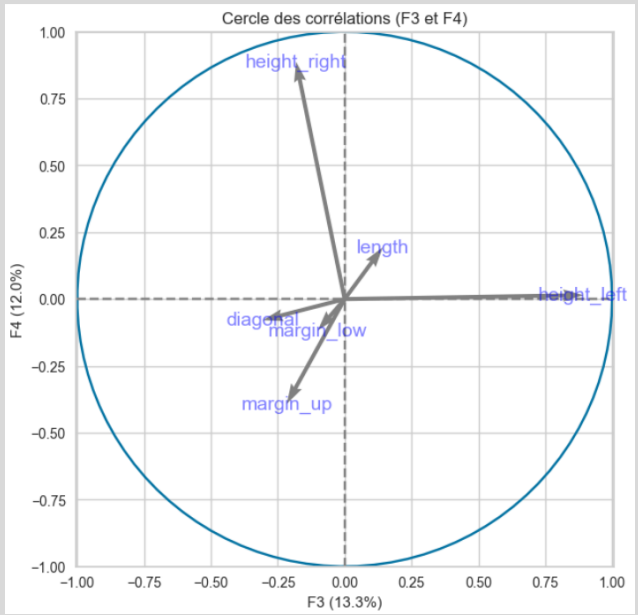
Cluster 0: Les faux billets
Nous observons que les faux billets ont des hauteurs plus longs en moyenne;
En plus ils ont des longueurs et diagonales moins longs.
Cluster 1: Les vrais billets
Pour les vrais billets c'est tout à fait le contraire avec les différentes dimensions.

Modélisation: Réduction de dimensions (PCA)

PCA





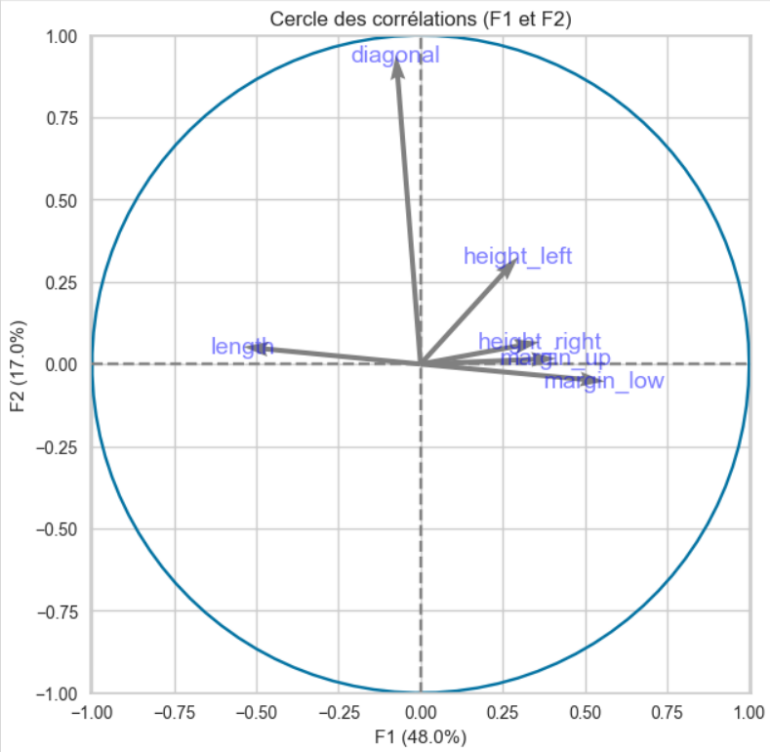


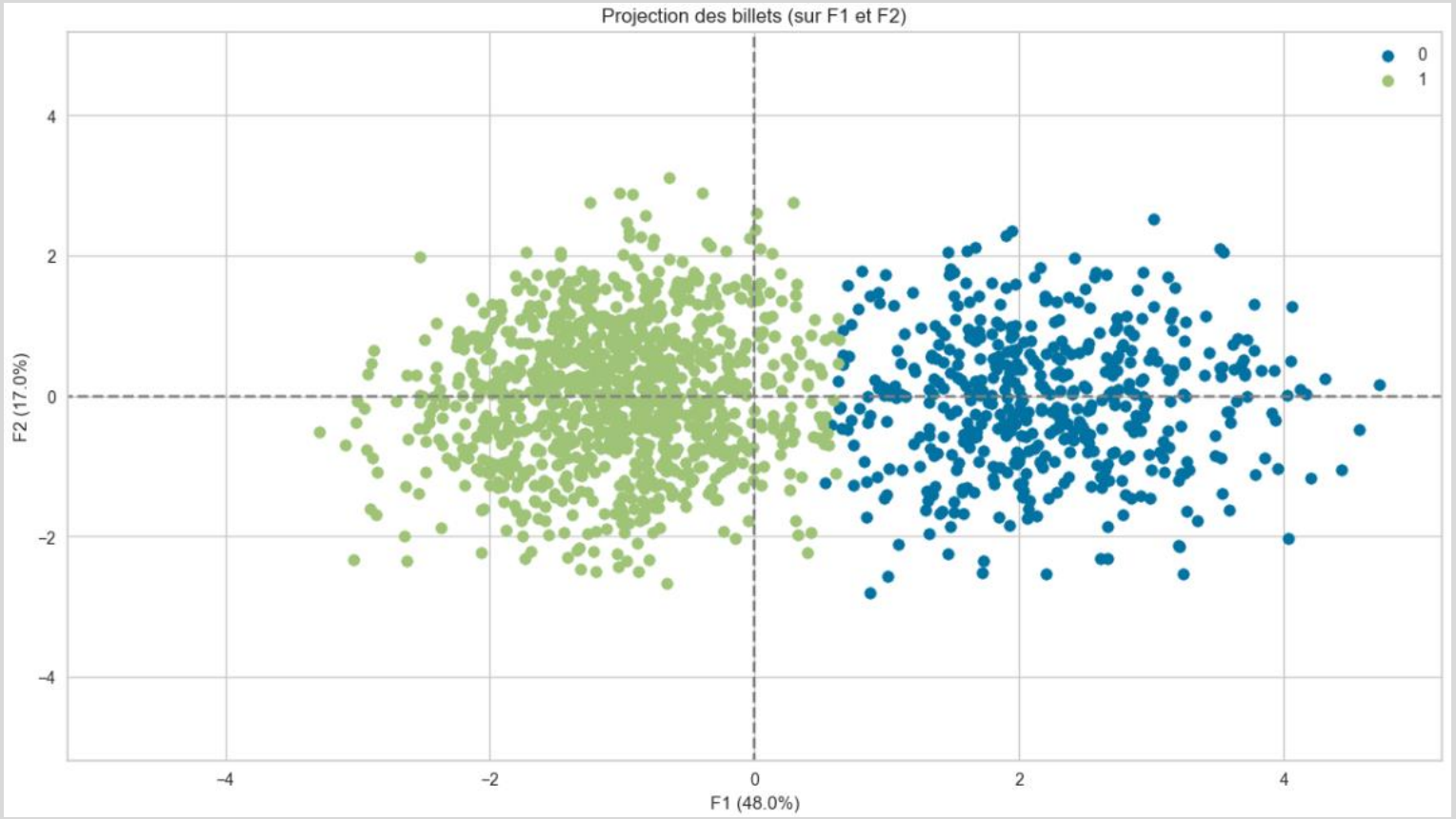
Interprétation

Les variables corrélées positivement à F1
height_left
height_right
margin_up
margin_low
Les variables corrélées négativement à F1
length
diagonal
Les variables corrélées positivement à F2
diagonal

Les variables corrélées positivement à F3¶
length
height_left
Les variables corrélées négativement à F3
diagonal
margin_low
Les variables corrélées positivement à F4
height_right
Les variables corrélées négativement à F4
diagonal
margin_low
margin_up

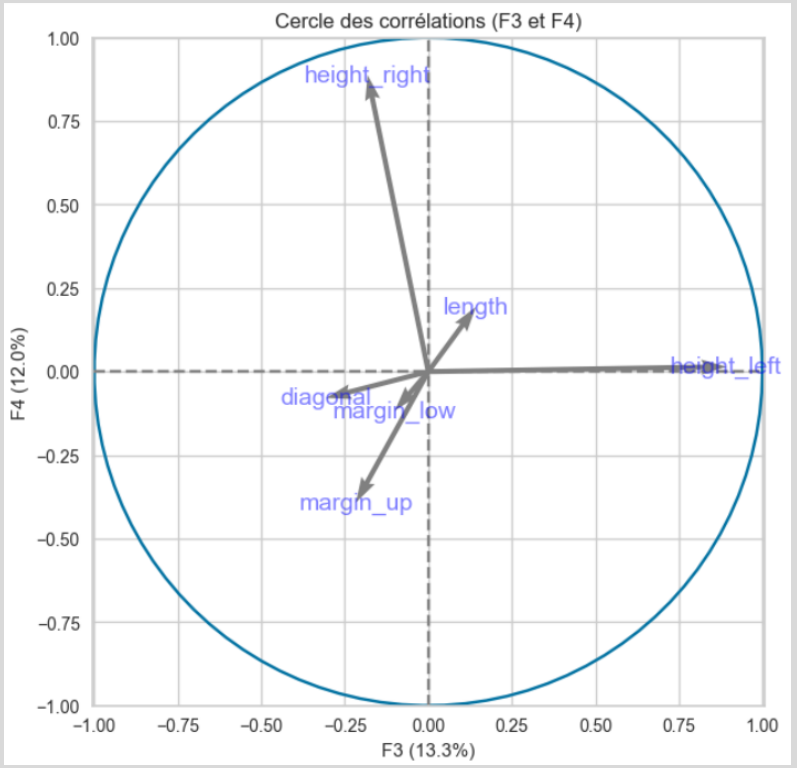
Modélisation: Projection des billets avec le K-means

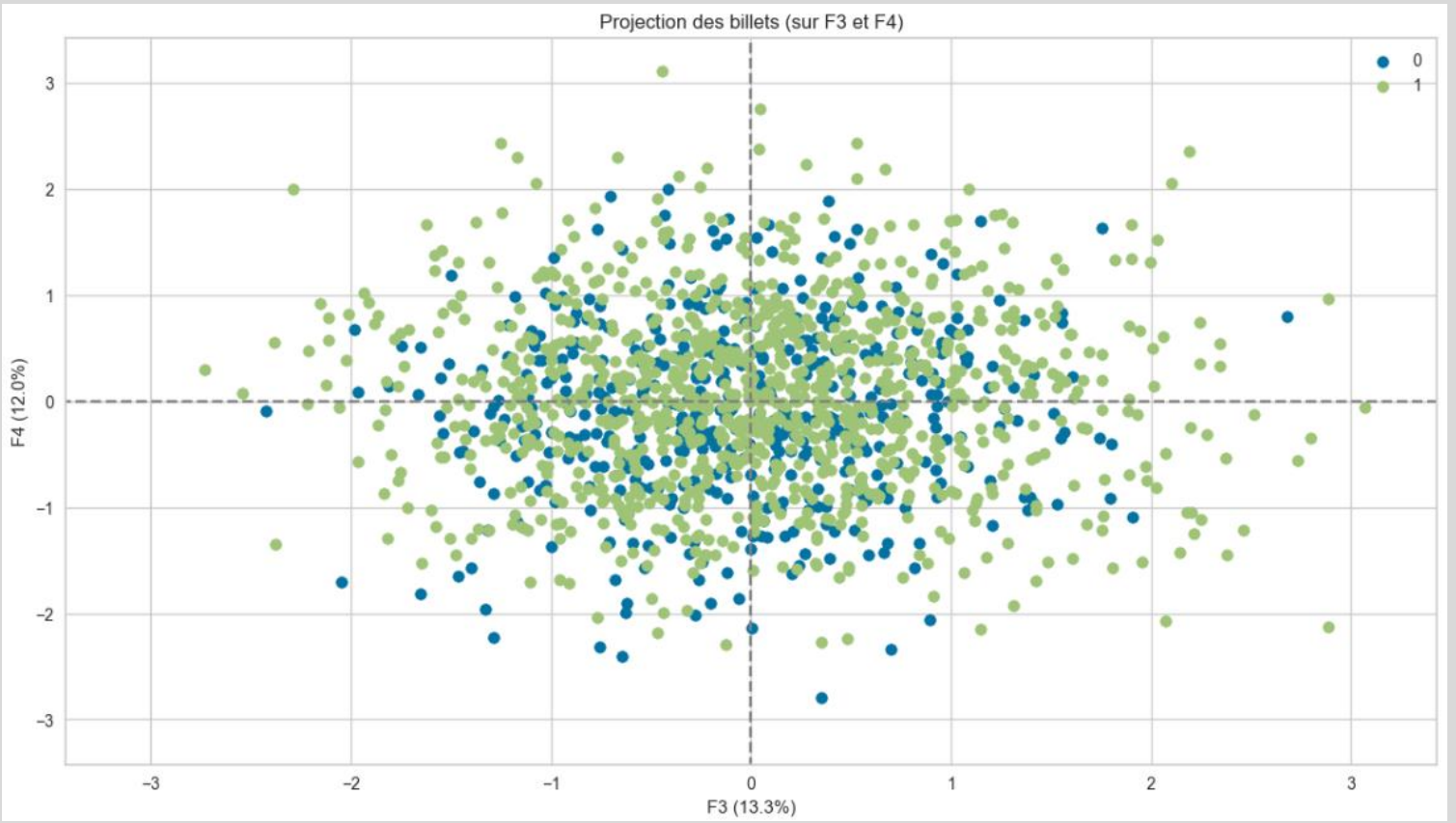




Nous avions F1 qui est corrélée aux variables suivantes: height_left, height_right, margin_up, margin_low, donc il y'a de grande chance que ces billets aient aussi une grande valeur pour ces variables.

Modélisation: Projection des billets avec le K-means

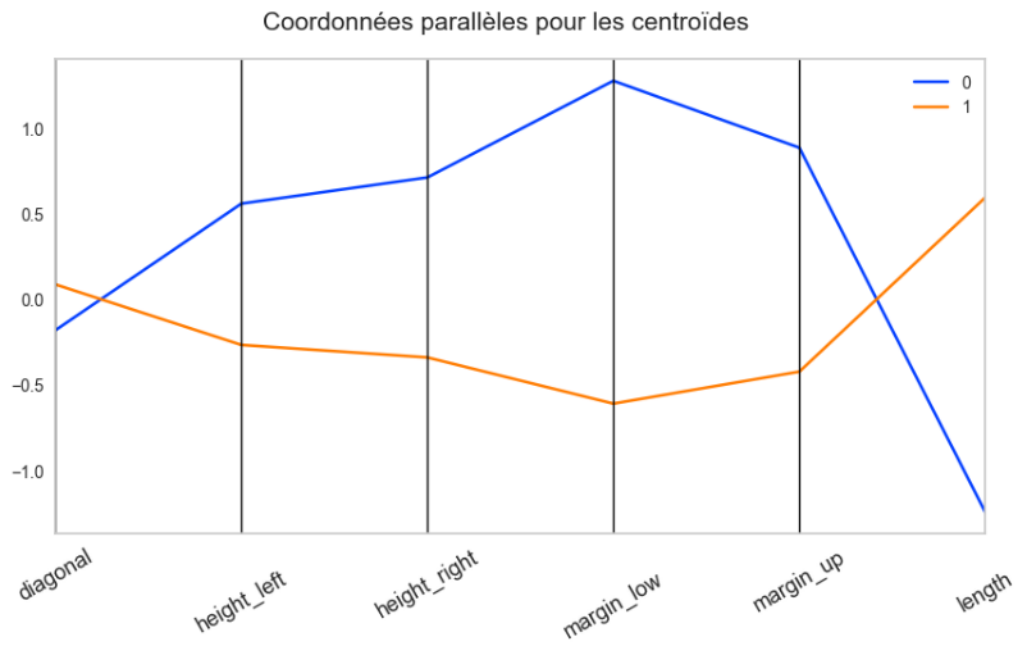




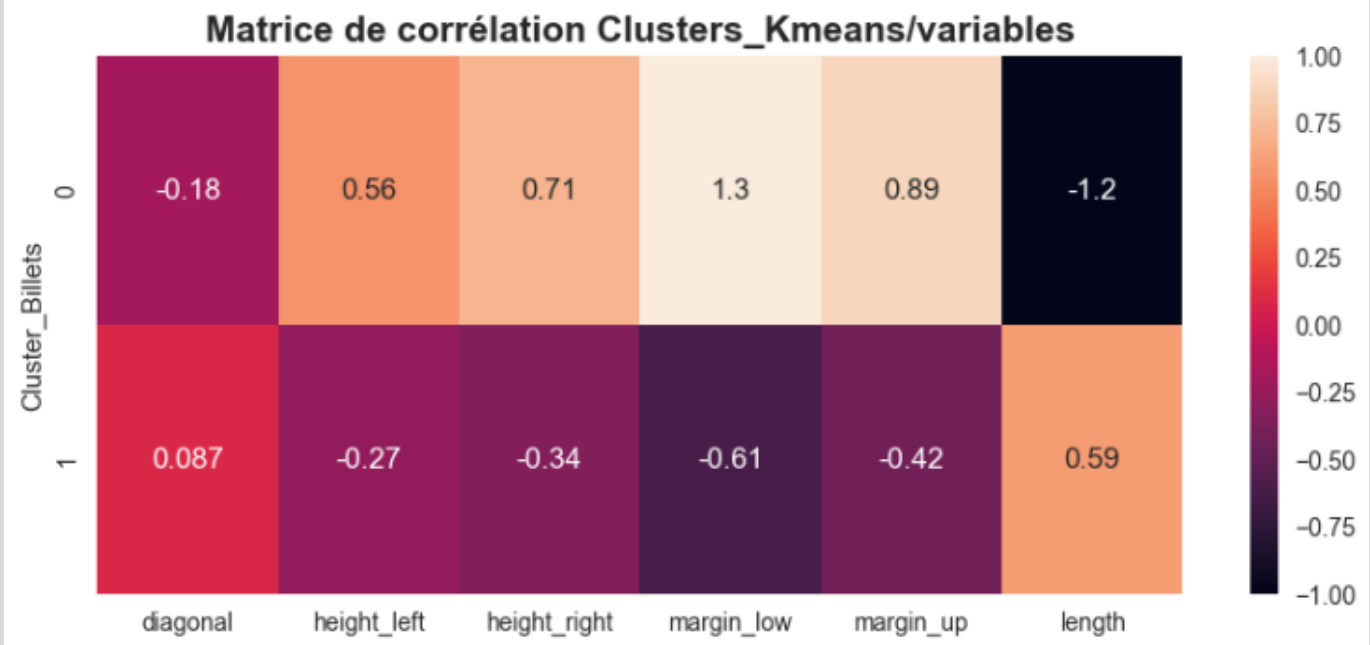
Sur le graphique de la projection des billets, nous avions des groupes qui sont assez homogène puis se distinguent de certains groupes.

Modélisation: Projection des billets avec le K-means

Coordonnées parallèles pour les centroides

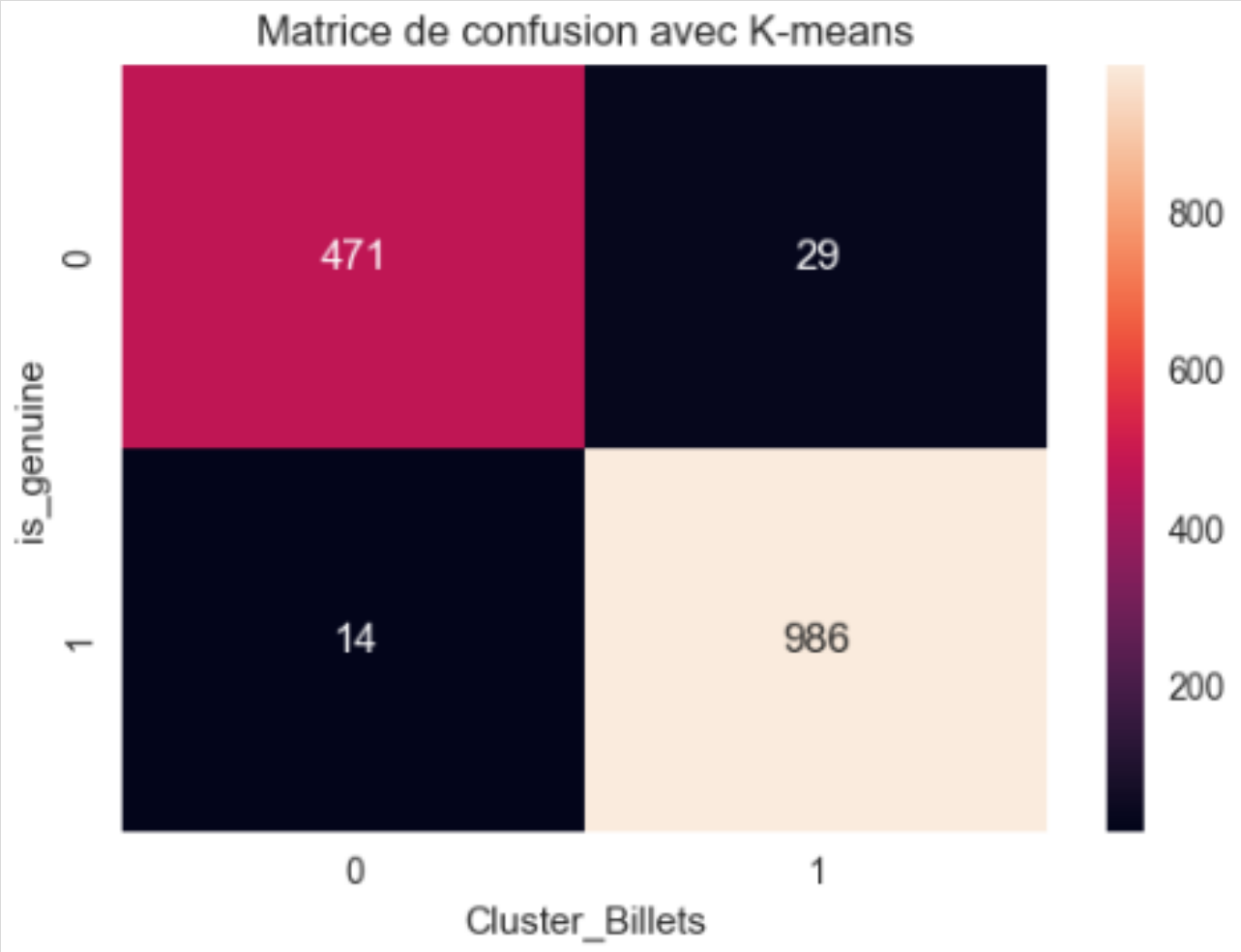


Matrice des corrélations avec les clusters



Cluster 0 : représente les billets où toutes les dimensions sont moyennement élevés à l'exception de la longueur du billet.
Cluster 1 : représente les billets avec un diagonal et longueur de billets moyennement élevés par rapport aux autres dimensions qui sont moyennement faibles.

Modélisation: Evaluation du modèle

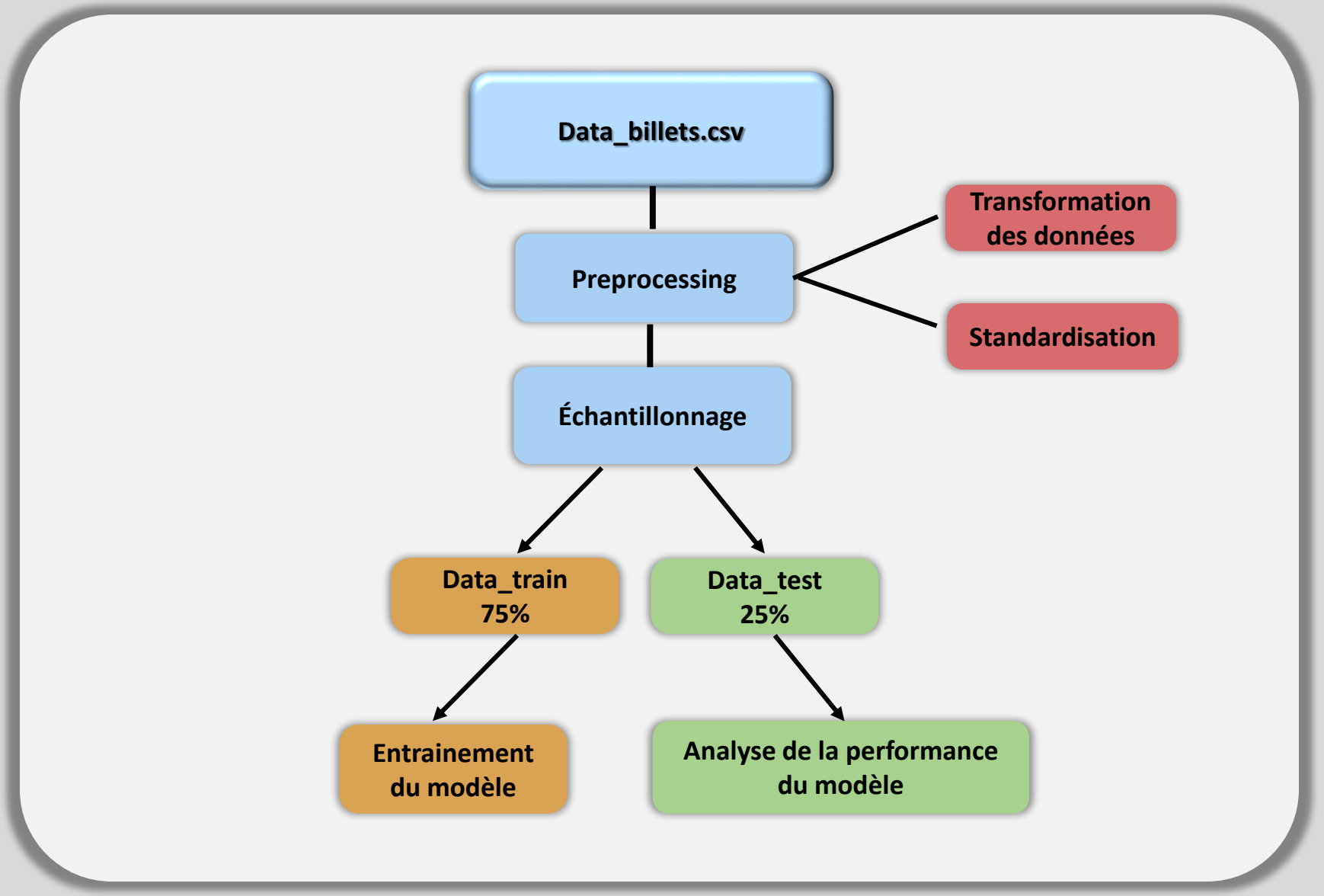


Interprétation

- Sur 500 faux billets, le modèle a correctement détecter 471 faux billets (TN) qu'il l'a prédit comme faux billets mais aussi le modèle a détecter 29 faux billets (FN) comme des vrais billets alors que le modèle les a prédit comme des vrais billets.

- Sur 1000 vrais billets, nous avons 986 vrais billets (TP) qui ont été détectés comme vrais billets puis prédit comme des vrais billets, mais 14 billets sont détectées comme des vrais billets (FP) alors que le modèle les a prédit comme des faux billets.

Modélisation: Régression Logistique

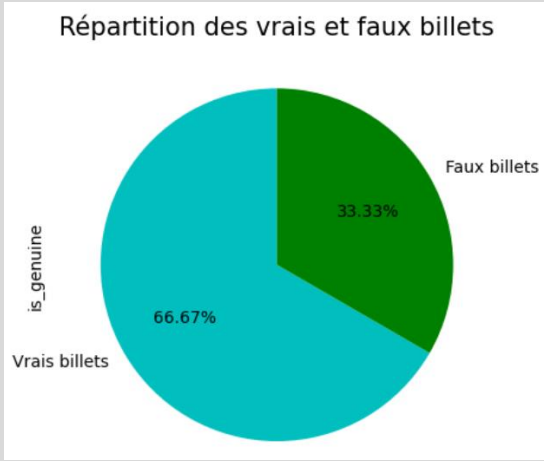


Modélisation: Analyse des résultats de la RL

Explication des targets / Déséquilibre de la classe:
Nous avons à faire à un problème de classification binaire où la classe est fortement déséquilibrée.

Explication des targets :
- **Target = 0 :** Vrais billets
- **Target = 1 :** Vrais billets

Il y a 66.67% de vrais billets et 33.33% dans notre classe.



Métriques pour un modèle de classification :

1. **Accuracy :** La précision du modèle
2. **Precision :** Performance du modèle quand celui-ci déclare une classe 1.
3. **Rappel :** Pourcentage de détection des classes 1.
4. **F1_score :** Moyenne harmonique de la précision et du rappel.

$$Accuracy = \frac{TN + TP}{(TN + FN + FP + TP)}$$

$$\text{Précision} = \frac{VP}{VP + FP}$$

$$\text{Rappel} = \frac{VP}{VP + FN}$$

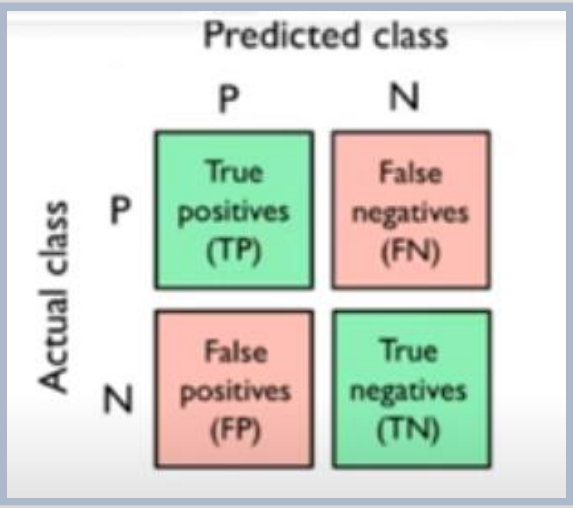
Analyse de notre modèle sur des données déséquilibrées :
On constate que le modèle ne prédit que des 1. Son accuracy est très bonne mais nous cherchons à prédire si une Target sera égale à 0.
Nous focaliserons donc notre performance de modèle sur la précision et le rappel.

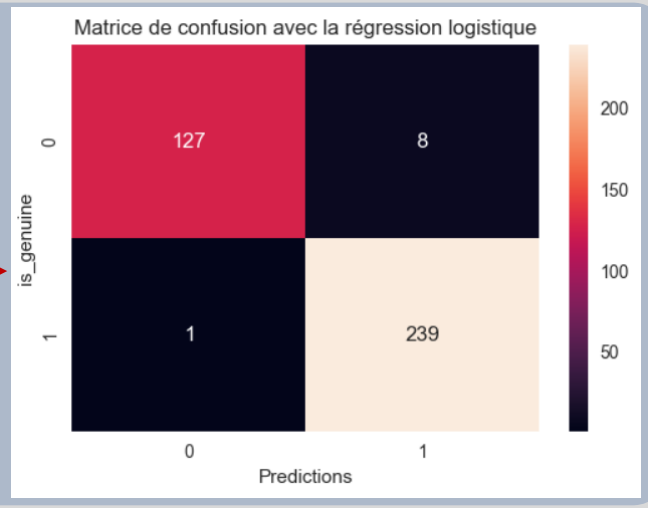
	precision	recall	f1-score	support
0	0.99	0.94	0.97	135
1	0.97	1.00	0.98	240
accuracy			0.98	375
macro avg	0.98	0.97	0.97	375
weighted avg	0.98	0.98	0.98	375

Modélisation: Analyse des résultats de la RL

La matrice de confusion :

La matrice de confusion consiste à compter le nombre de fois où des observations de la classe 0 ont été rangées dans la classe 1. Par exemple, si nous voulons connaître le nombre de fois où le classifieur à bien réussi à classer une classe 1, on examinera la cellule à l'intersection de la ligne 1 et de la colonne 1.



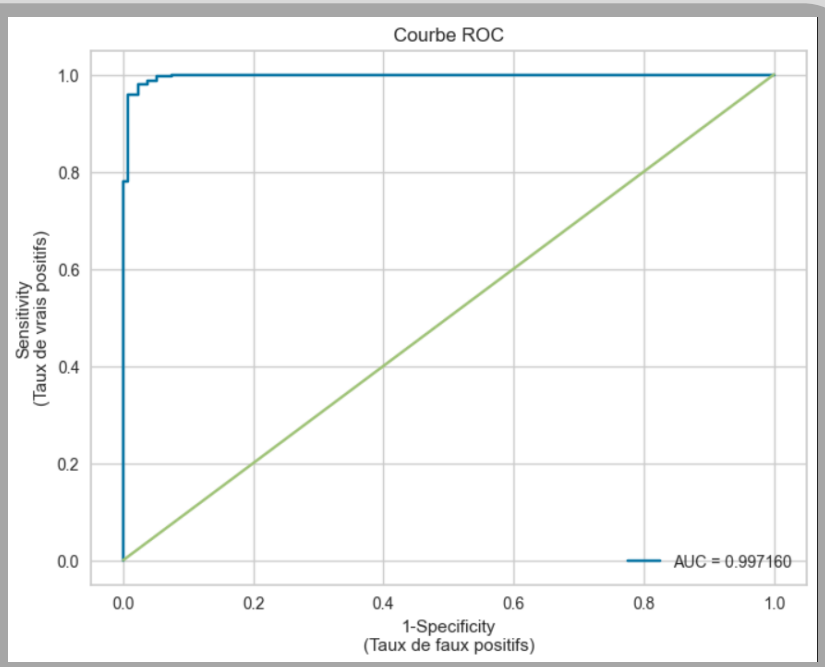


La courbe ROC et le score AUC :

La courbe ROC (Receiver Operating Characteristic) est un outil communément utilisé avec les classifieurs binaires. Elle croise le taux de TP avec le taux de FP.

Un bon classifieur aura sa courbe qui s'approche le plus possible du coin supérieur gauche du graphique.

Une autre façon de comparer des classifieurs consiste à mesurer l'aire sous la courbe (Area Under the Curve ou AUC). Un classifieur parfait aurait un score AUC égal à 1, tandis qu'un classifieur purement aléatoire aurait un score AUC de 0.5.



Modèle final: essai de l'algorithme

Jeu de données

billets_production.csv

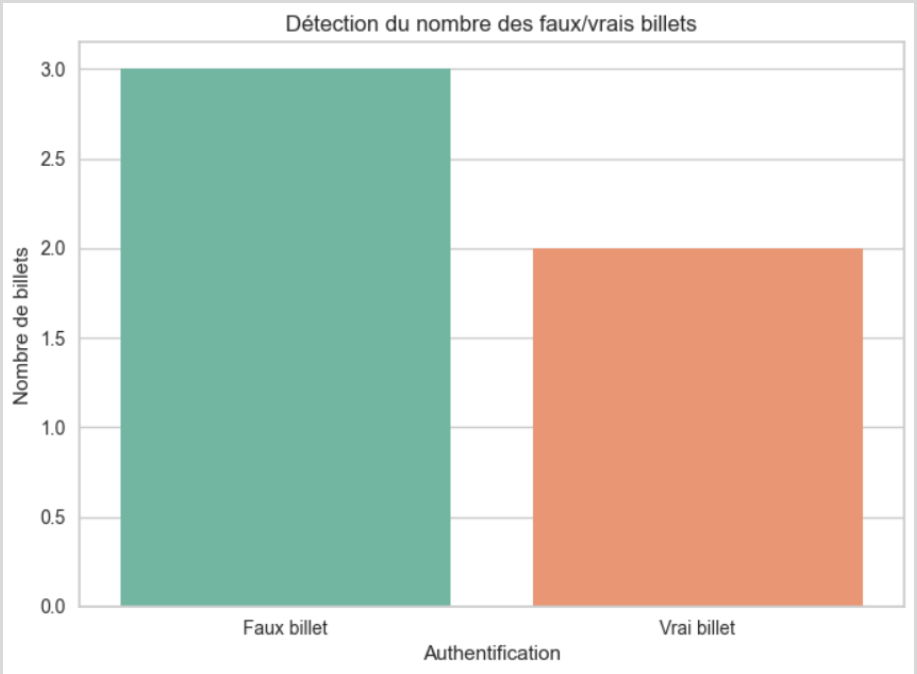
Fichier billets production : (5, 7)

Standardisation les données

Modèle Régression Logistique

Prédictions

```
Nombre total de billets :  5
Nombre de vrais billets :  2
Nombre de faux billets :  3
```



AUTRES

Piste à améliorer

- Optimisation du modèle avec des hyper paramètres
- Tester plus de modèles
- Tester une technique d’équilibrage des données (**Ex: La méthode SMOTE**)
- Tester d’autres performances de prédiction plus approfondie (**Ex: réseaux de neurones**)

BILAN

- Données

- Explore et Analyse

- Réponses