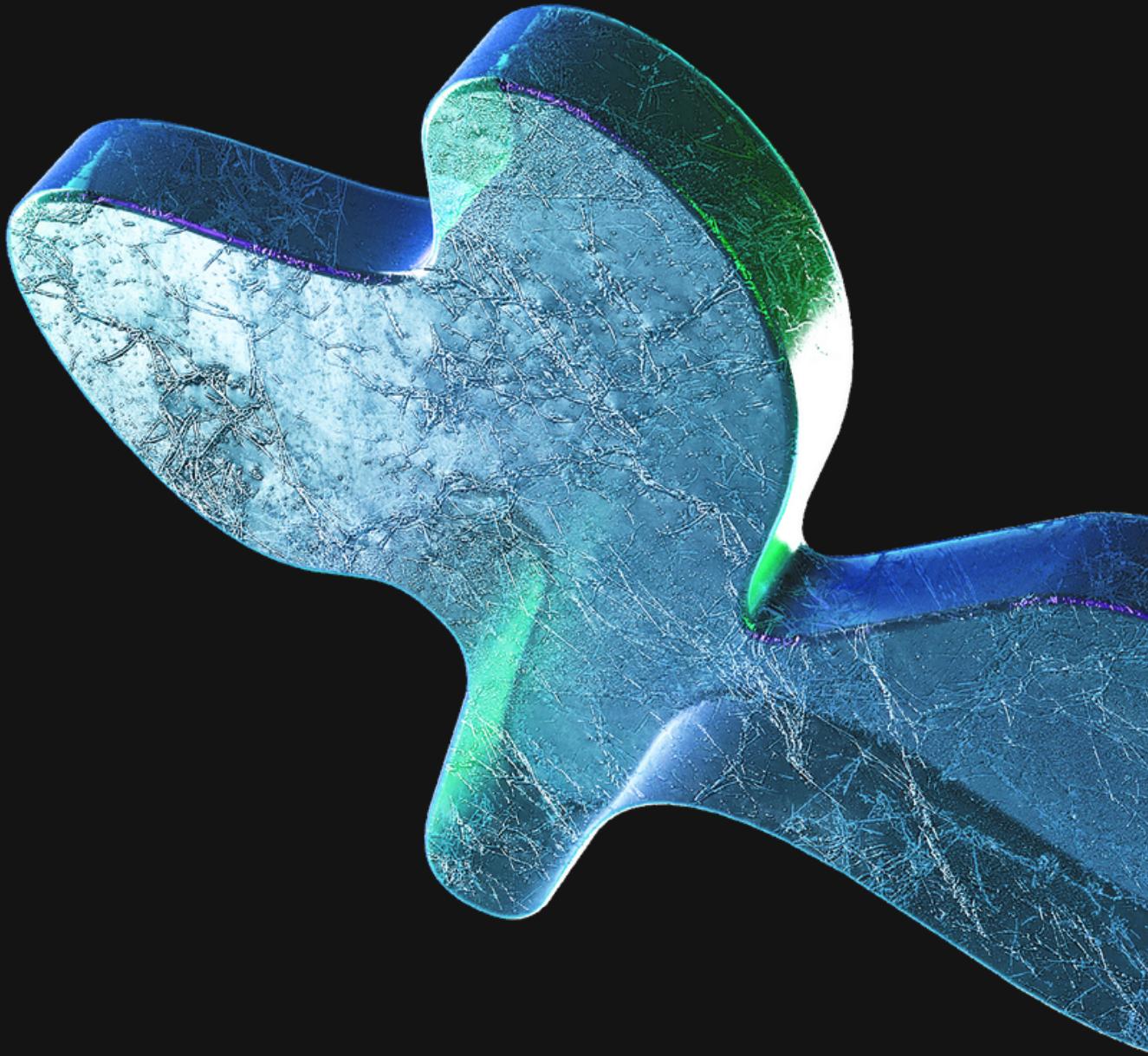


Détection de faux bILLETS

algorithme de classification avec régression logistique



Sommaire



Mise en contexte

Exploration des données / analyse

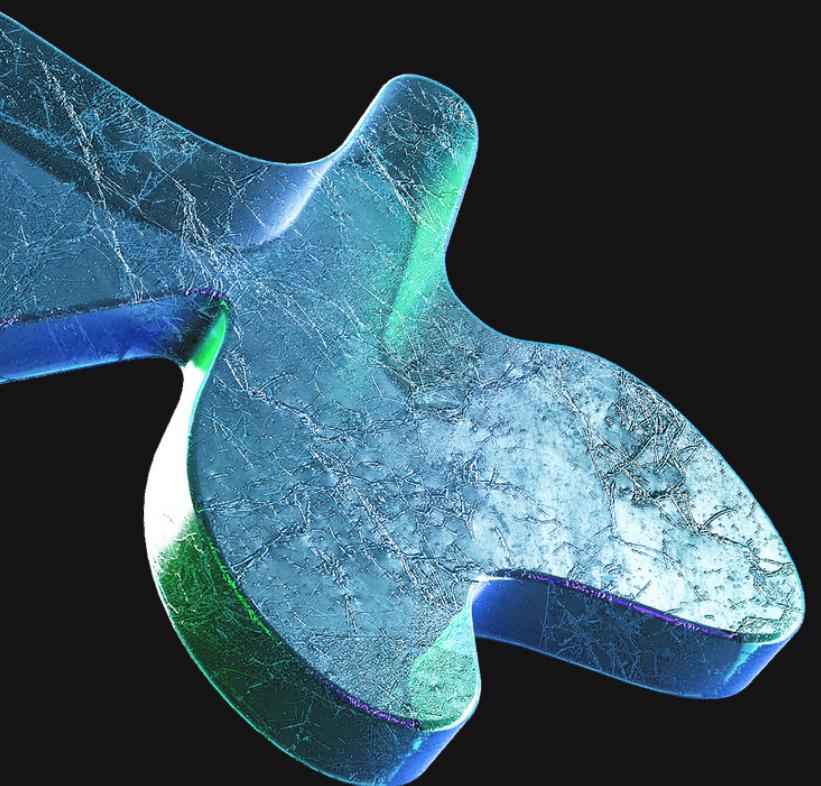
ACP

Clustering

Modélisation

programme de détection

Partie 1

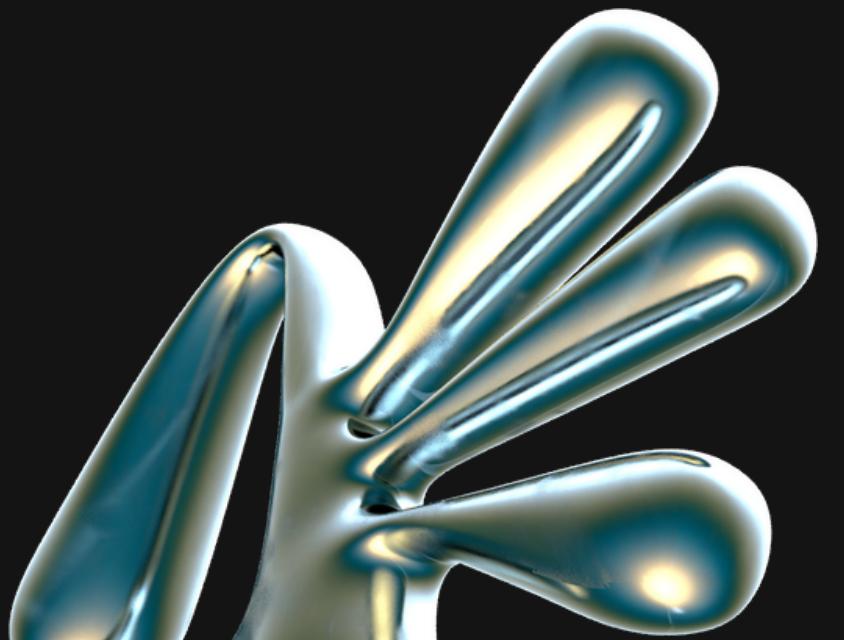


Mise en context

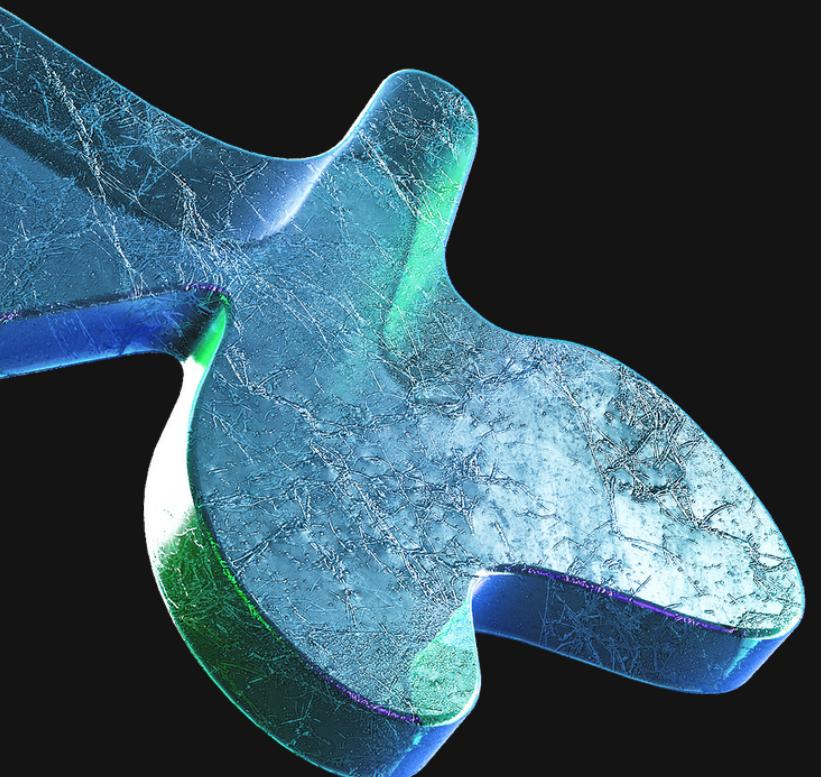
- Nous sommes une société de consulting informatique
- nous avons été engagés dans le cadre de la lutte contre la criminalité organisée, à l'Office central pour la répression du faux monnayage

Notre objectif

- Minimiser le nombre de faux billets détectés vrais
- Créer un algorithme de détection de faux billets.



Exploration des données



Exploration de donnée

170 billets, 7 variables

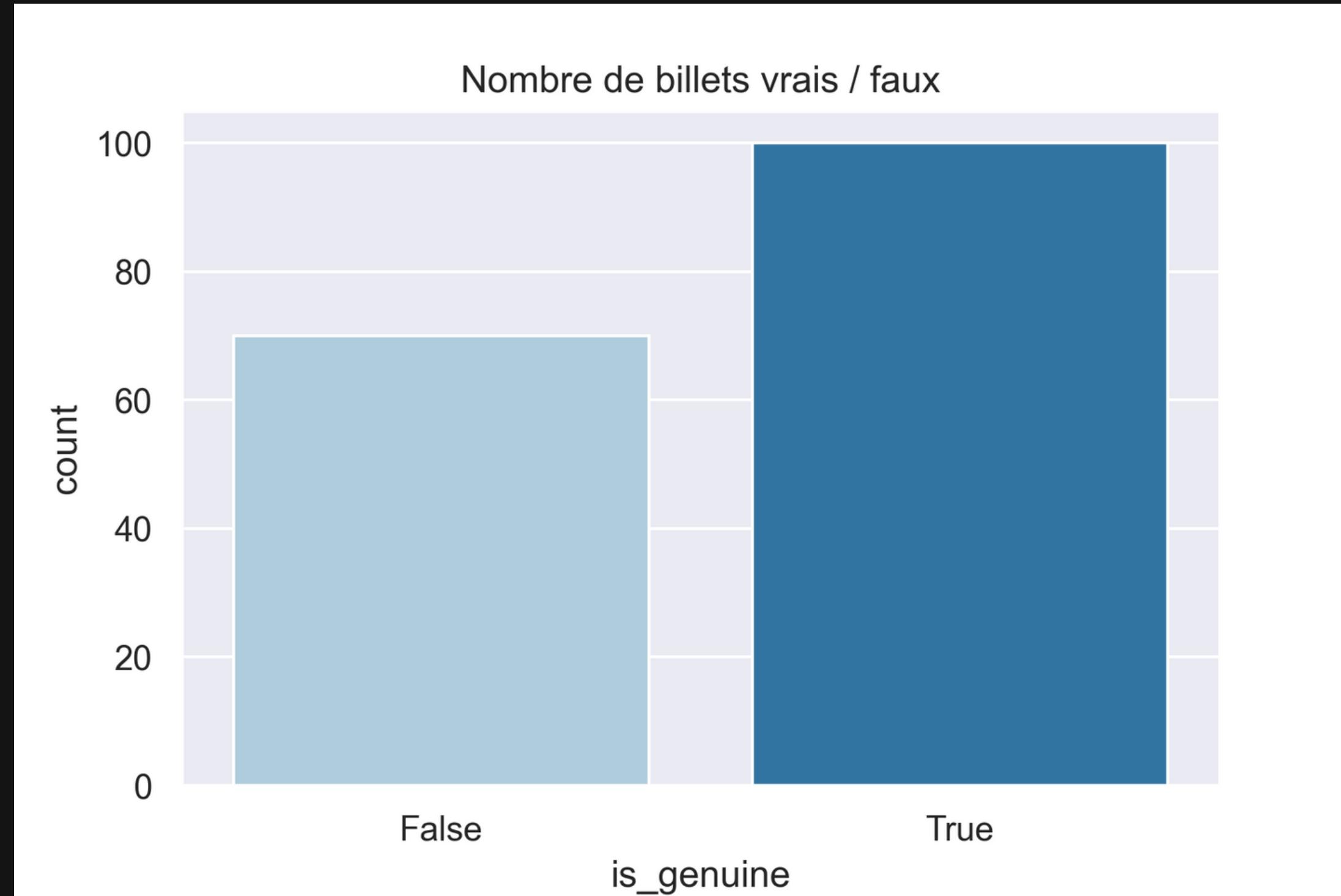
is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
False	172.5	104.07	103.71	3.82	3.63	110.74
False	171.78	104.07	104.16	5.77	3.3	111.27
False	171.51	104.13	103.9	4.99	3.6	111.23
True	172.03	103.87	103.4	4.29	3.01	113.09
False	171.43	104.14	103.95	5.34	3.14	111.76

- 6 variable de mesures en millimètres:
 - Diagonale
 - hauteur à gauche
 - hauteur à droite
 - marge basse
 - marge haute
 - longeur
- 1 variable d'indentification du billets
 - Vrai
 - Faux
- Aucune valeur manquante
- Aucune valeur aberrante
- Aucun doulon

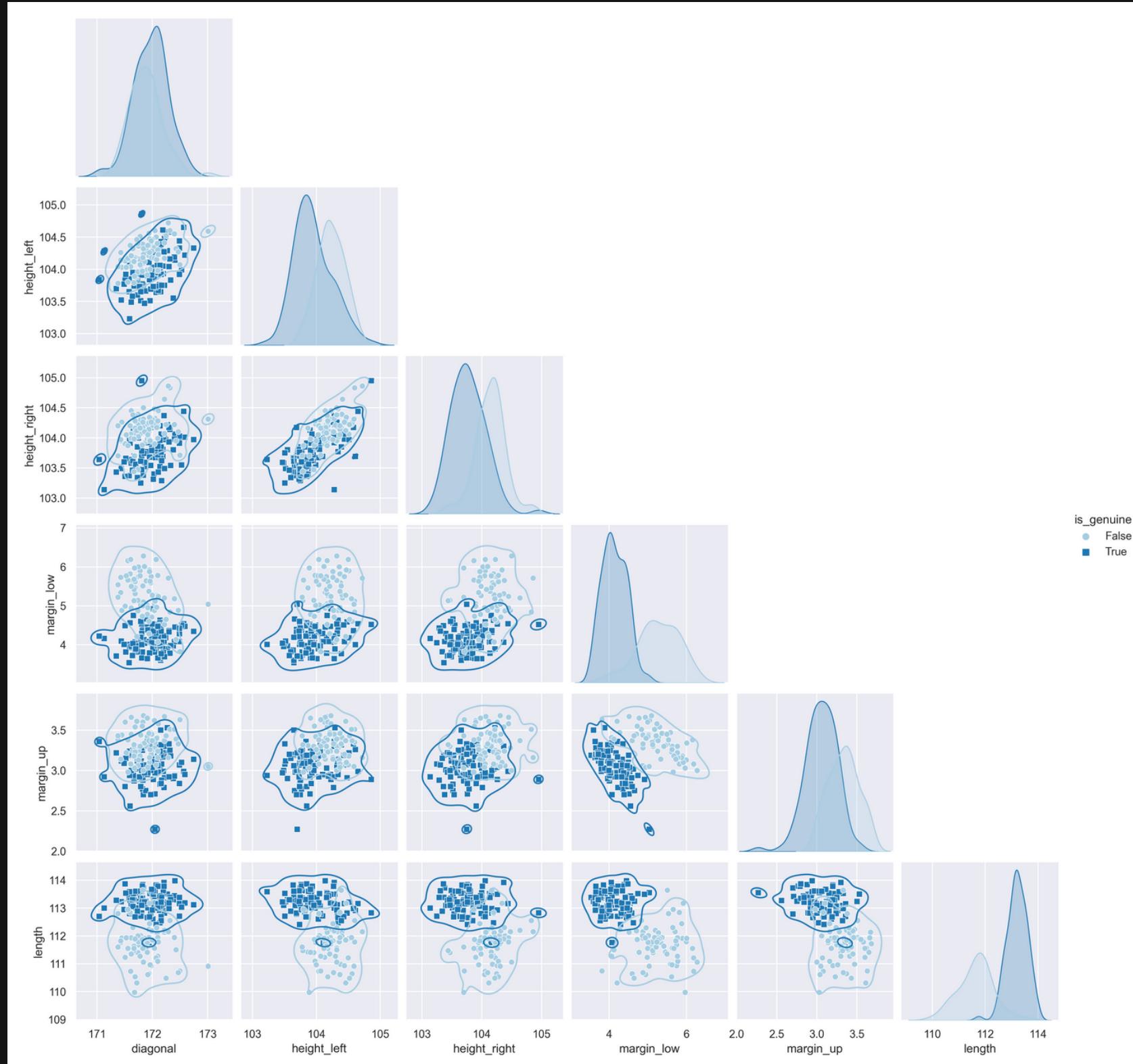
Exploration de donnée

100 vrais billets

70 faux billets

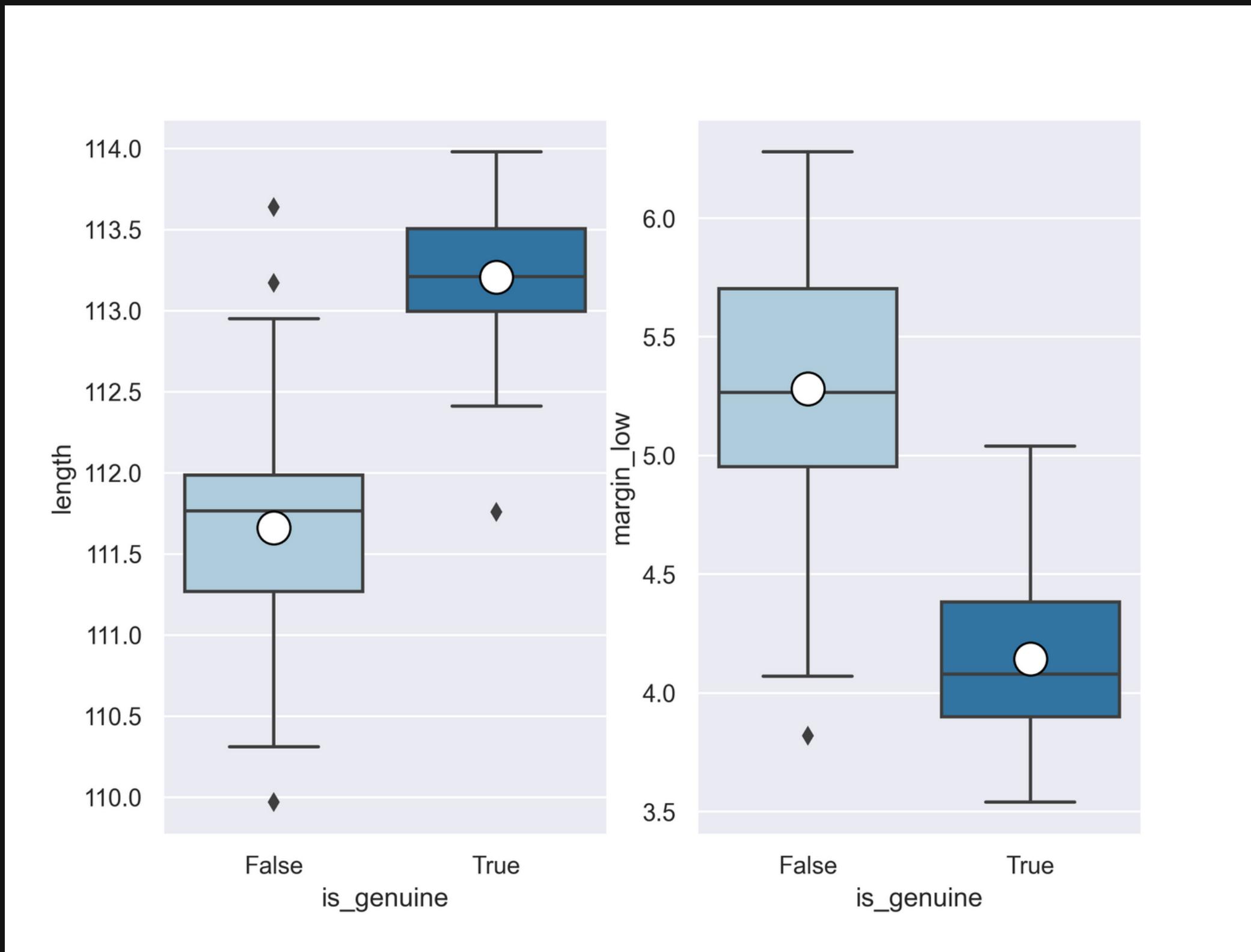


Exploration de donnée

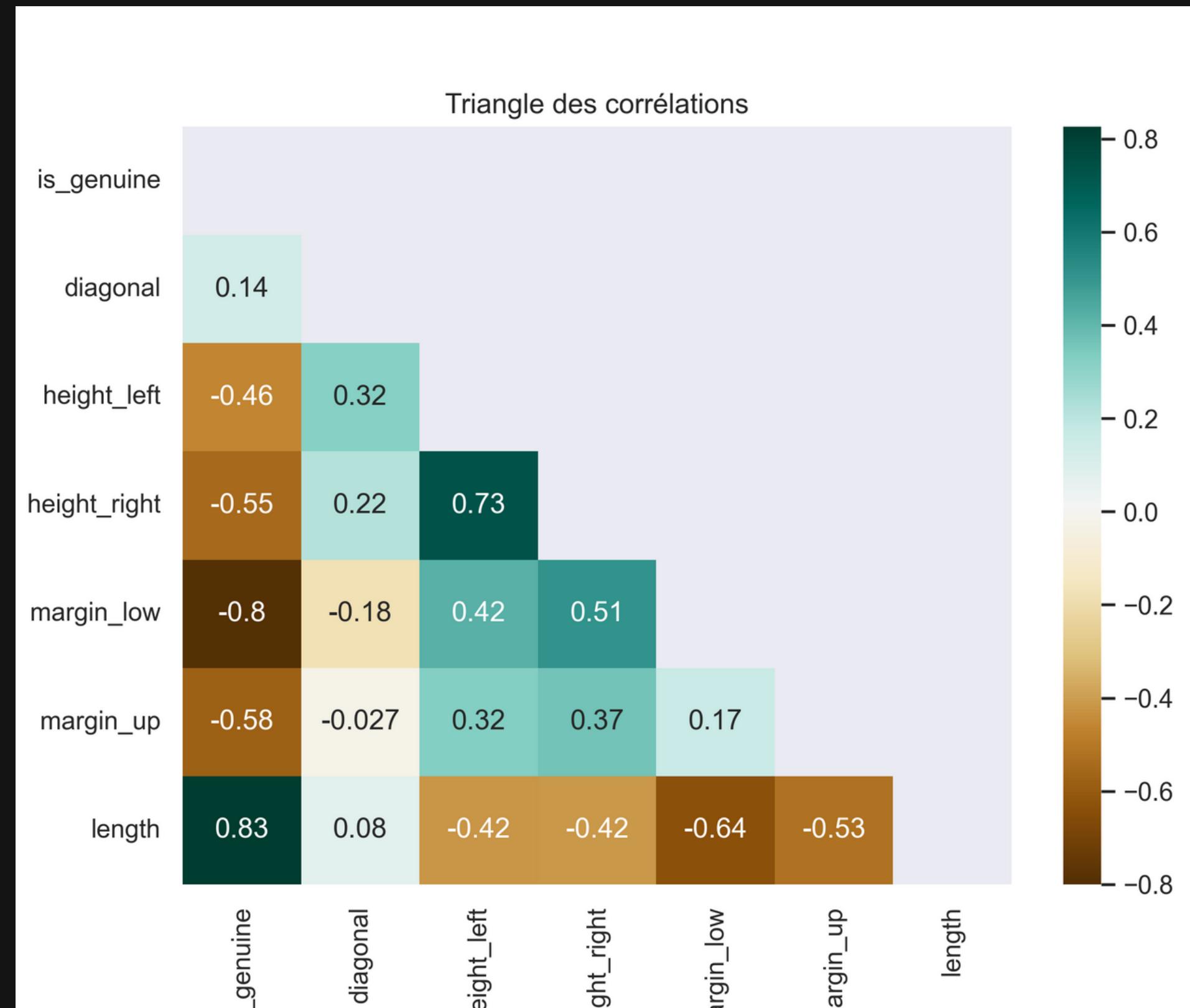


- 2 variables sortes du lots :
 - longueur
 - marge basse
- les autre variables ont peut d'impact

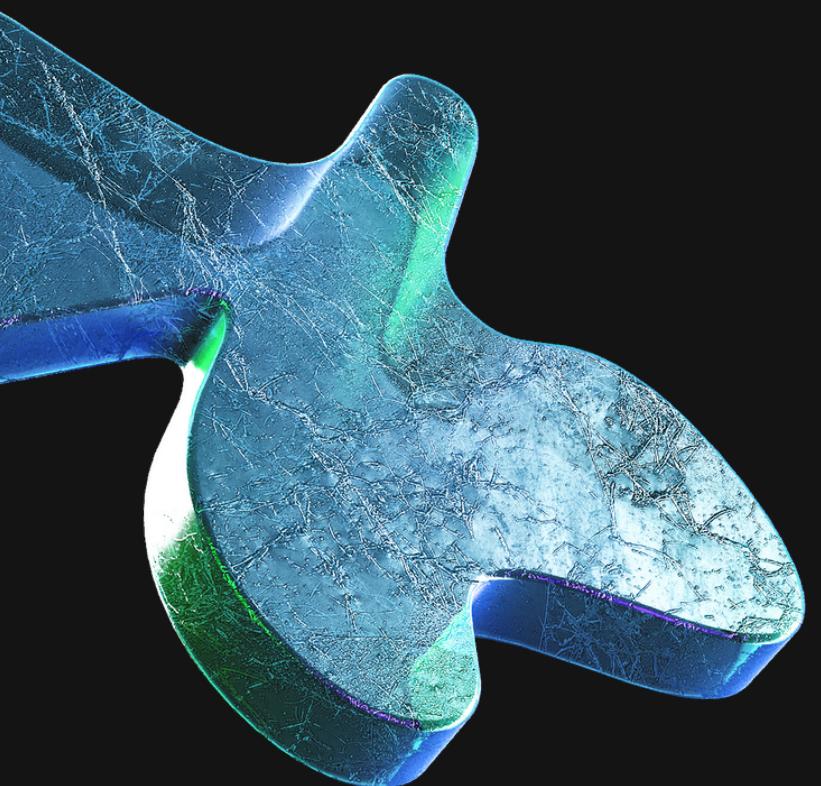
Exploration de donnée

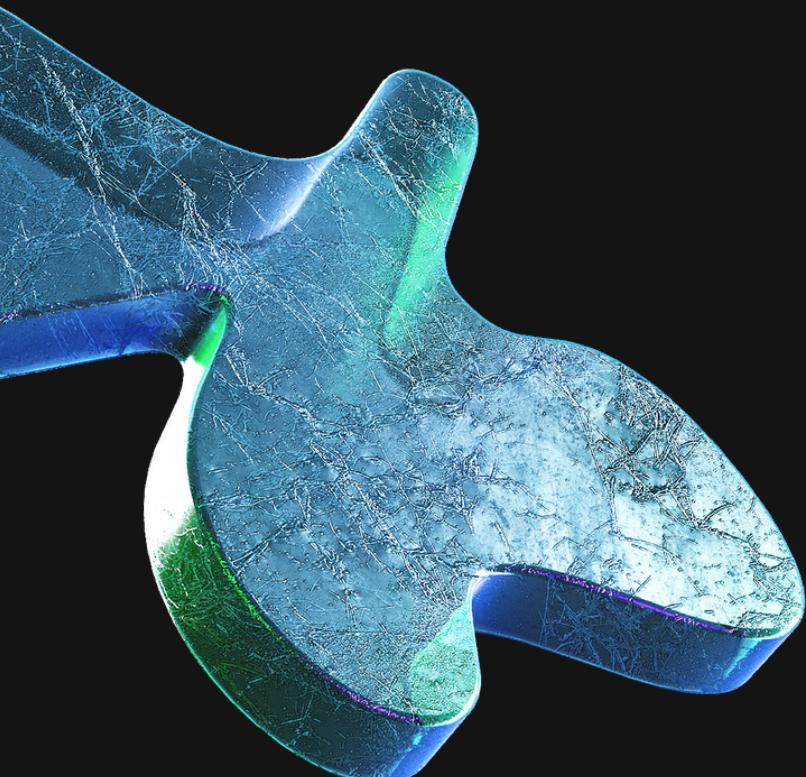


Exploration de donnée



Partie 2





ACP & algorithme de classification

ACP

objectifs de l'acp

Réductions des variables à n dimension

Transformation linéaire

Préserve les rapports de colinéarité

Perte d'informations

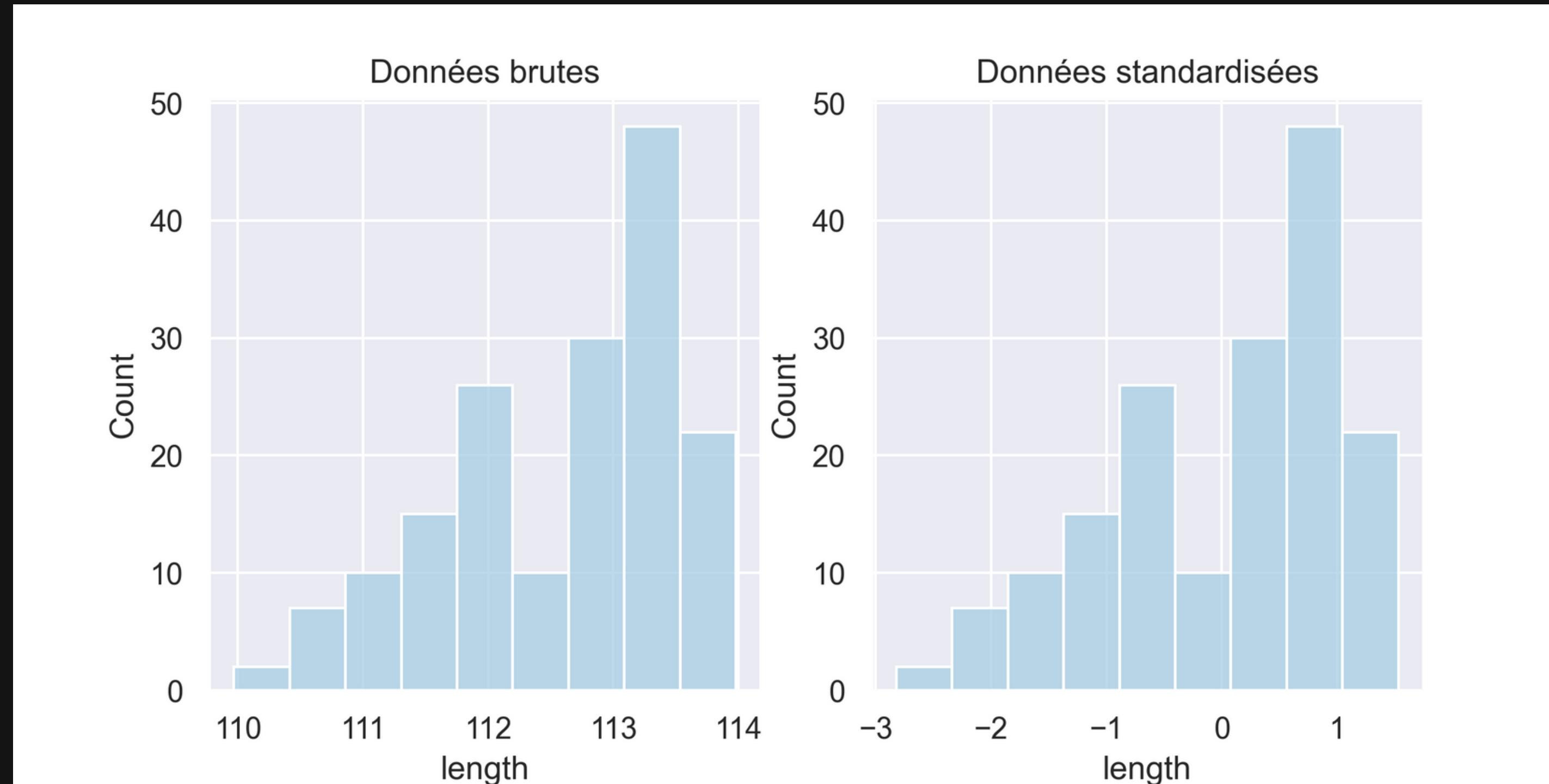
Valeurs obtenues non interprétables

Valeurs projetables sur un plan à 2 dimensions ou 3 dimensions

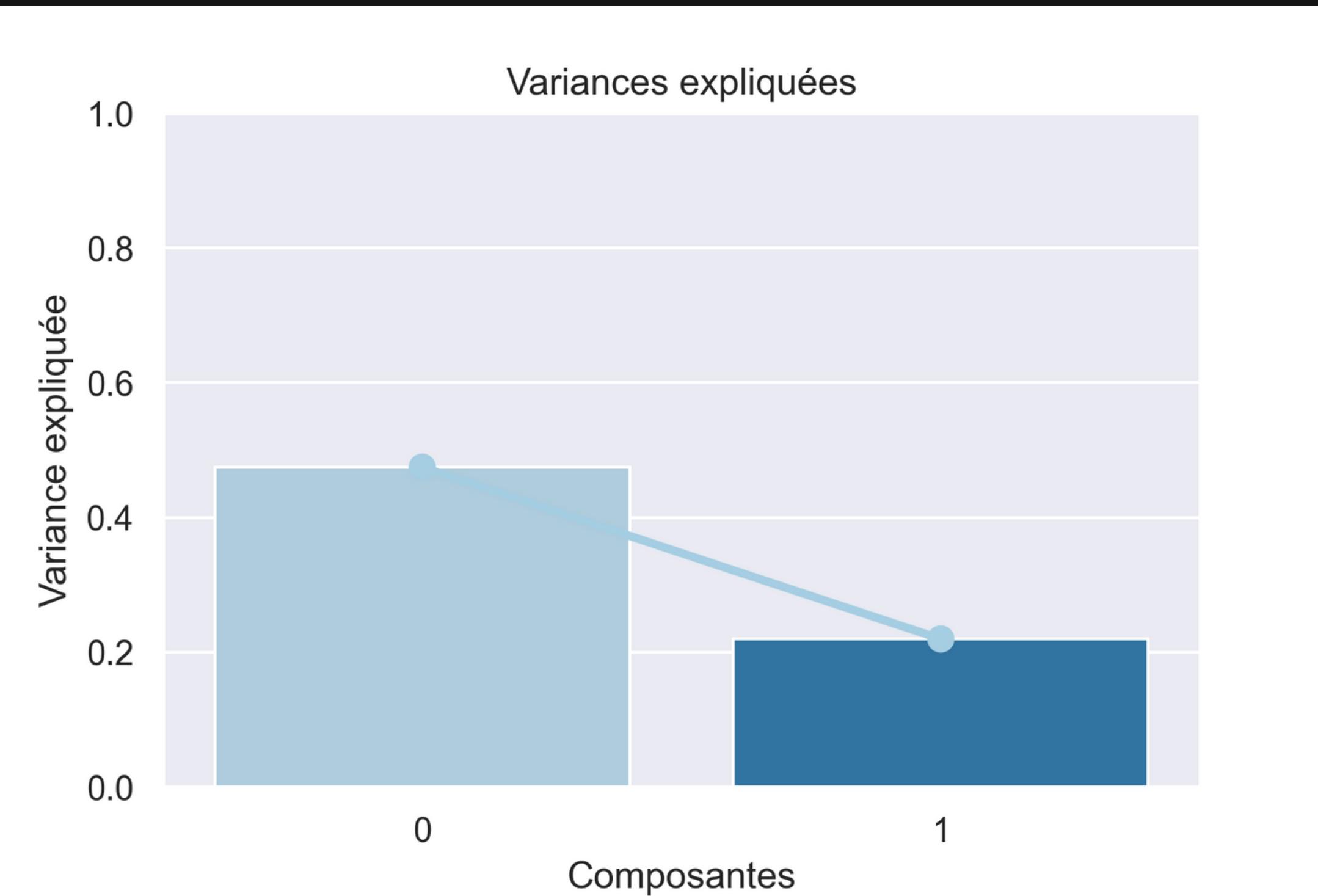
Requis : préserve le maximum d'information dans les 2 premières composantes
objectif : visualiser la ressemblance (variabilité) des individus et la linéarité

Sensible à la variance

Standardisation des donnée

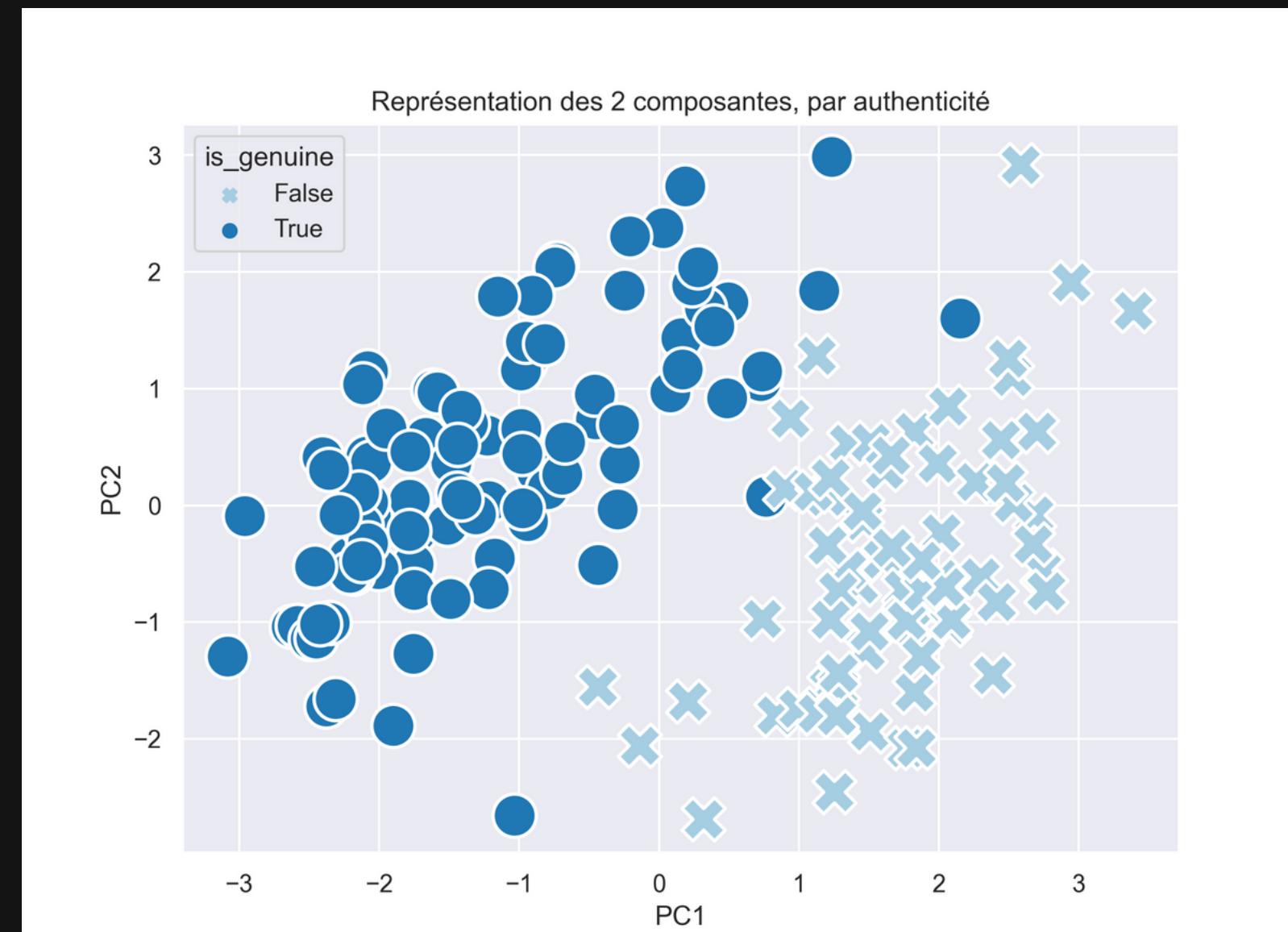


Analyse de l'éboulis

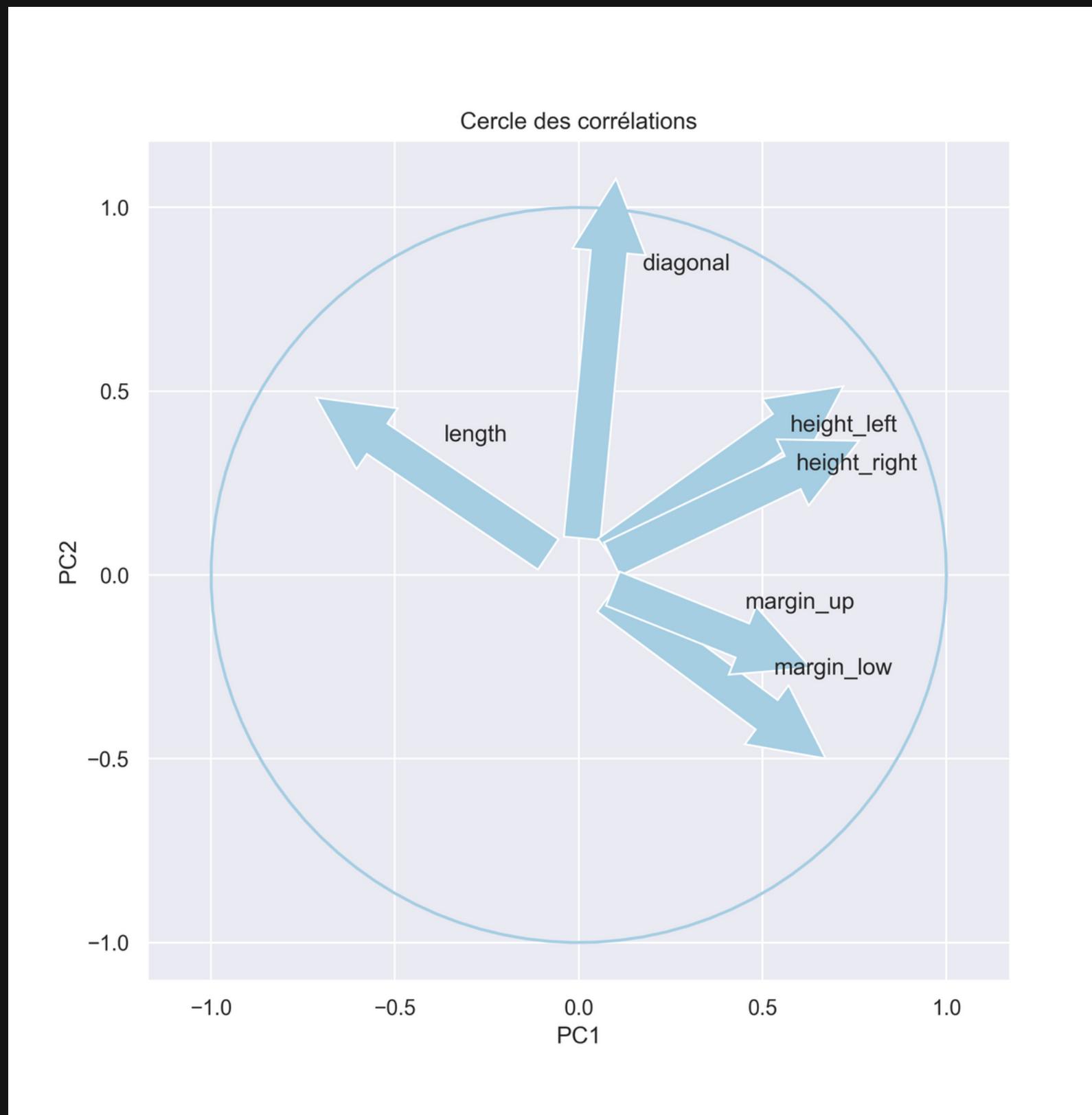


Représentation en 2 dimensions

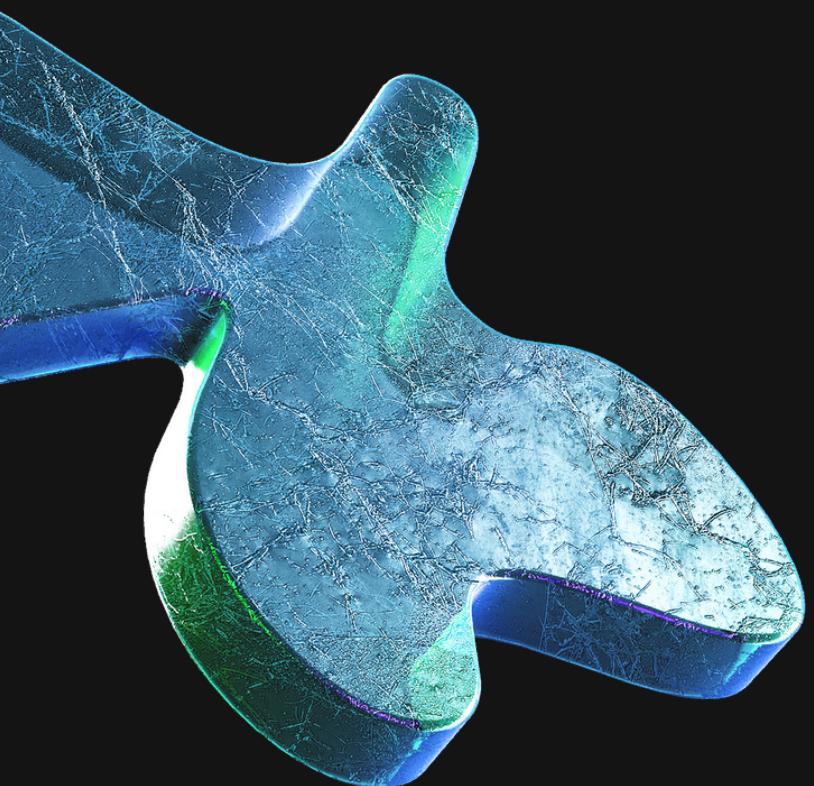
is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length	PC1	PC2
True	171.81	104.86	104.95	4.52	2.89	112.83	2.1536387504688634	1.5997094454160978
True	171.67	103.74	103.7	4.01	2.87	113.29	-2.1104158925555847	-0.5260389195126456
True	171.83	103.76	103.76	4.4	2.88	113.84	-1.9731524201989556	-0.04810178309059015
True	171.8	103.78	103.65	3.73	3.12	113.63	-2.0597950918526458	-0.08910521048714269
True	172.05	103.7	103.75	5.04	2.27	113.55	-2.403180089393602	0.41216977455743775



Cercle des corrélations



Clustering



K-MEANS

- Regrouper les billets en K groupes homogènes

- Classification non supervisée

Postulat : on ne connaît pas les groupes auxquels appartiennent les billets

- Classification non hiérarchique

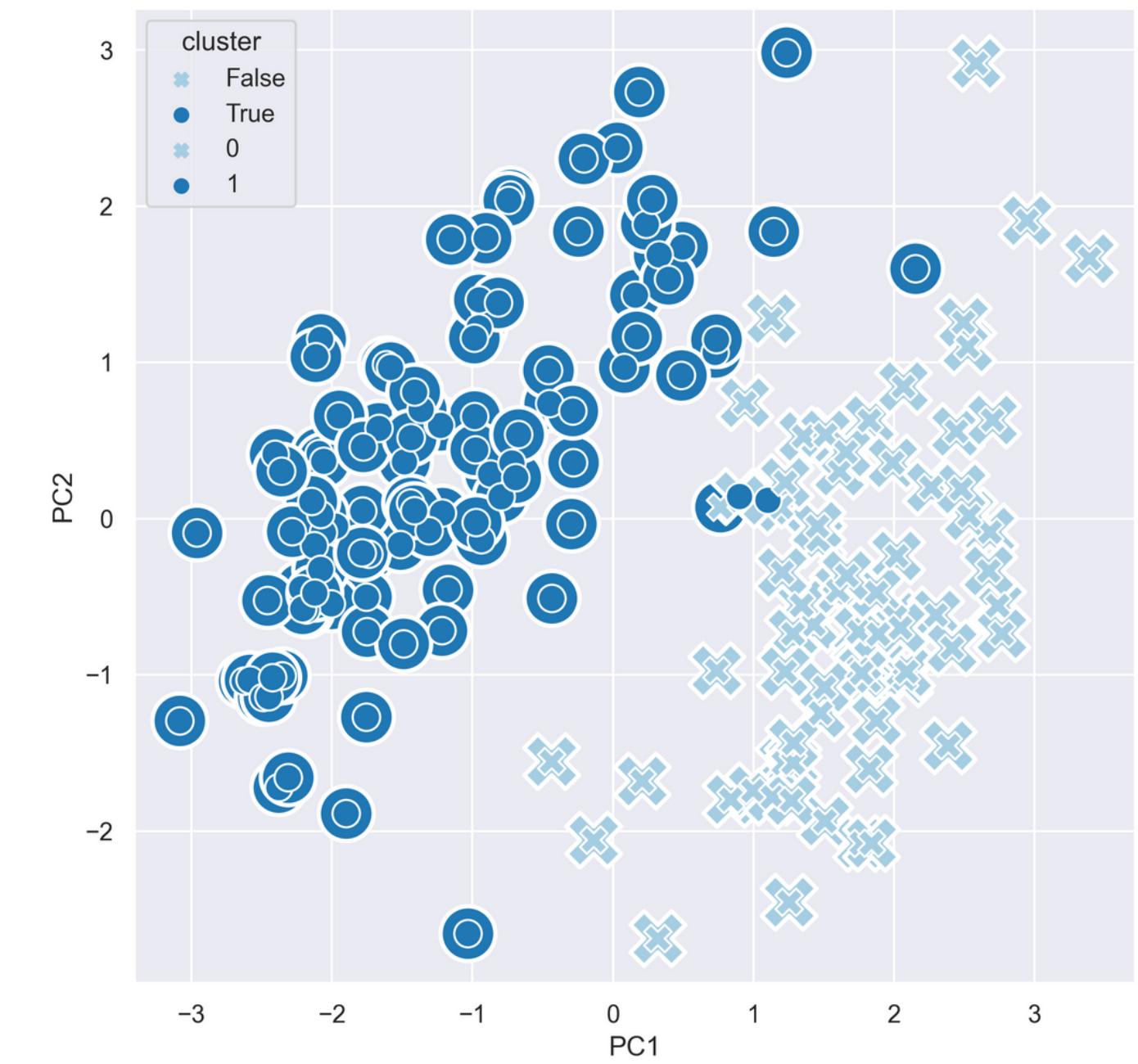
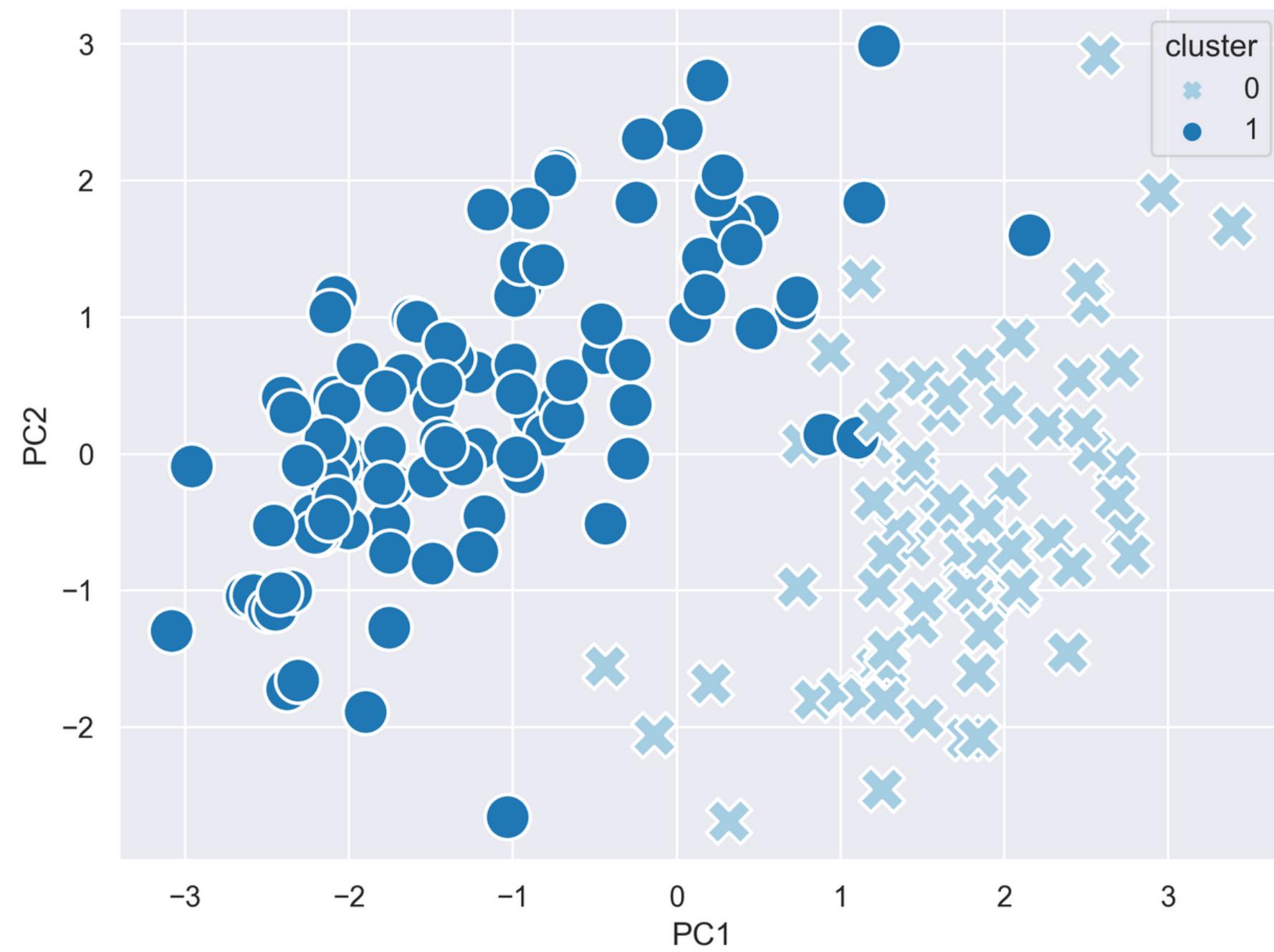
Méthode des k-moyennes

- Sélectionner le nombre clusters à identifier (K)
- Sélectionner K points aléatoires (clusters initiaux)
- Mesurer la distance euclidienne entre le 1er point et les K clusters
- Assigner le 1er point au cluster le plus proche
- Répéter pour tous les points
- Calculer le centroïde de chaque cluster
- Itérer jusqu'à ce que les centres ne bougent plus

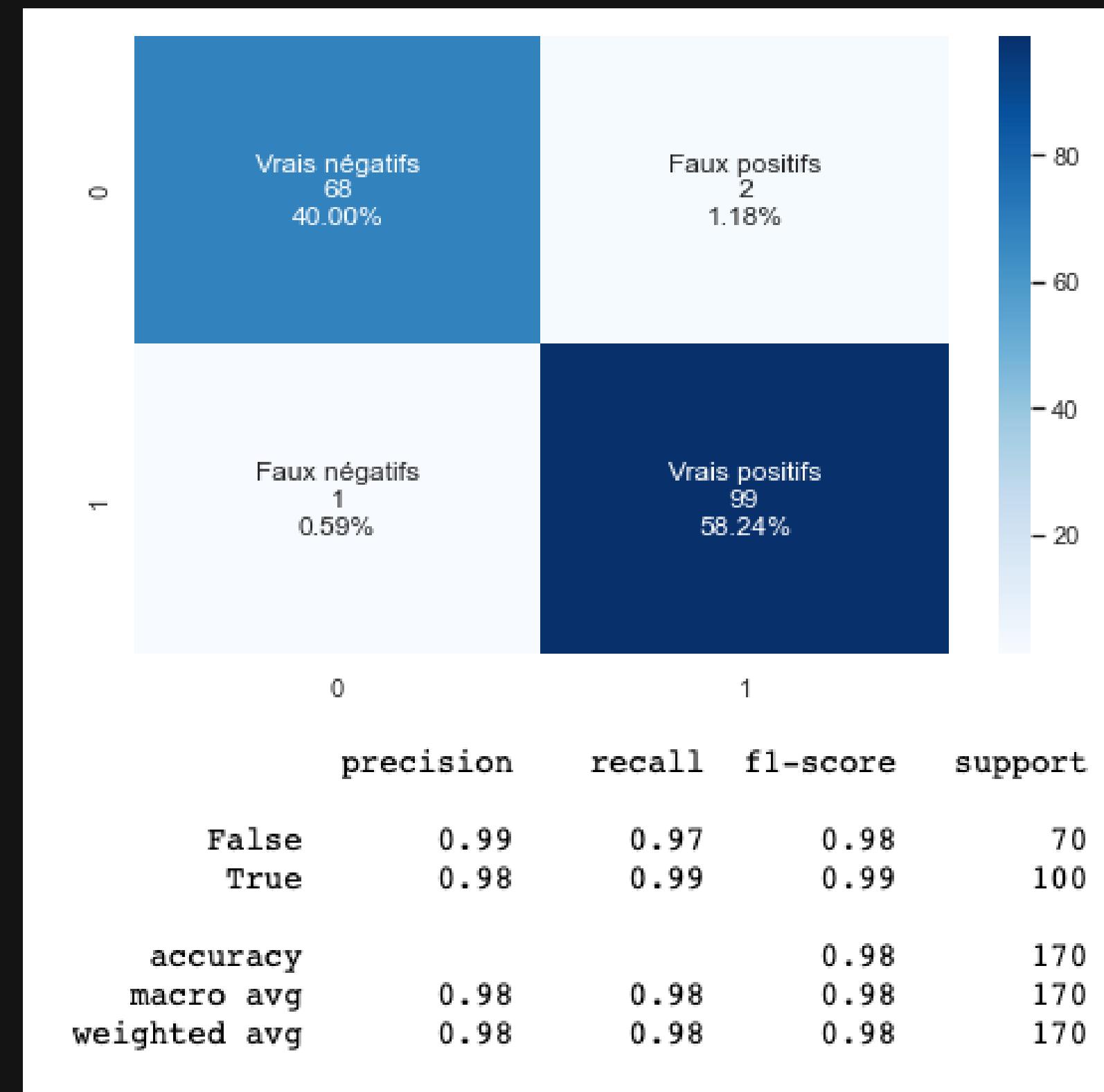
K-MEANS

is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length	PC1	PC2	cluster
True	171.81	104.86	104.95	4.52	2.89	112.83	2.1536387504688634	1.5997094454160978	1
True	171.67	103.74	103.7	4.01	2.87	113.29	-2.1104158925555847	-0.5260389195126456	1
True	171.83	103.76	103.76	4.4	2.88	113.84	-1.9731524201989556	-0.04810178309059015	1
True	171.8	103.78	103.65	3.73	3.12	113.63	-2.0597950918526458	-0.08910521048714269	1
True	172.05	103.7	103.75	5.04	2.27	113.55	-2.403180089393602	0.41216977455743775	1

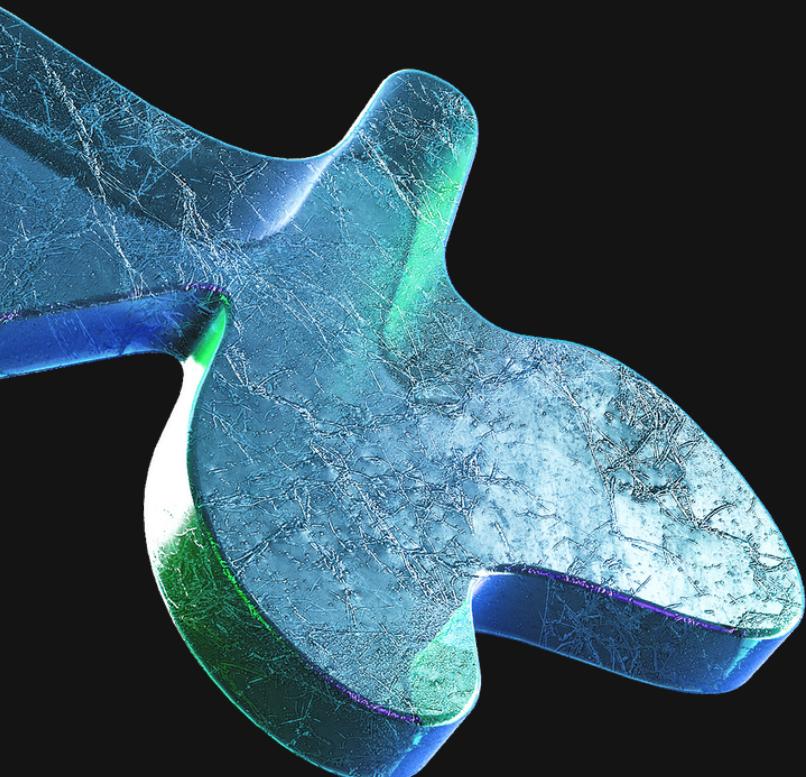
K-MEANS



Matrice de confusion



Modélisation



Régression linéaire

- Calcul d'une variable quantitative d'après d'autres quantitatives

x = variable(s) indépendante(s) explicatives (features, inputs, paramètres) y = variable dépendante expliquée

- Apprentissage supervisé

On connaît la valeur réelle de y pour chaque groupe d'explicatives

- Postulat : on peut aligner les points sur une droite

Relation linéaire décrivant le mieux la relation entre X et y

= $f(x) = \text{pente } x X + \text{constante}$ (point sur l'ordonnée quand $X = 0$)

- Prédit des données continues

Régression logistique

→ Quand la variable expliquée est qualitative

Nombre limité de valeurs possibles

→ Classification supervisée

Résultat de la variable dépendante déjà connu

→ Renvoie probabilités entre 0 et 1

Probabilité que l'individu appartienne à la classe True

Probabilité que l'individu appartienne à la classe False

→ Transformation logistique sur la fonction de régression linéaire

$S = \text{logit} (\text{logarithme}) \text{ sur } f(x)$

Seuil de probabilité fixé généralement à 0.5

Conditions pour de bons résultats

- Peu de features (risque d'overfitting sinon)
- Données facilement séparables (par une ligne)
- Suffisamment d'individus à disposition
- Pas de valeurs aberrantes
- Données normalisées

Split des données

```
] : y = df['is_genuine']

X_train, X_test, y_train, y_test = sk.model_selection.train_test_split(X_std, y, test_size=0.33, random_state=42)

print('X_train :', len(X_train), 'lignes',
      '\ny_train :', len(y_train), 'lignes',
      '\nX_test :', len(X_test), 'lignes',
      '\ny_test :', len(y_test), 'lignes')

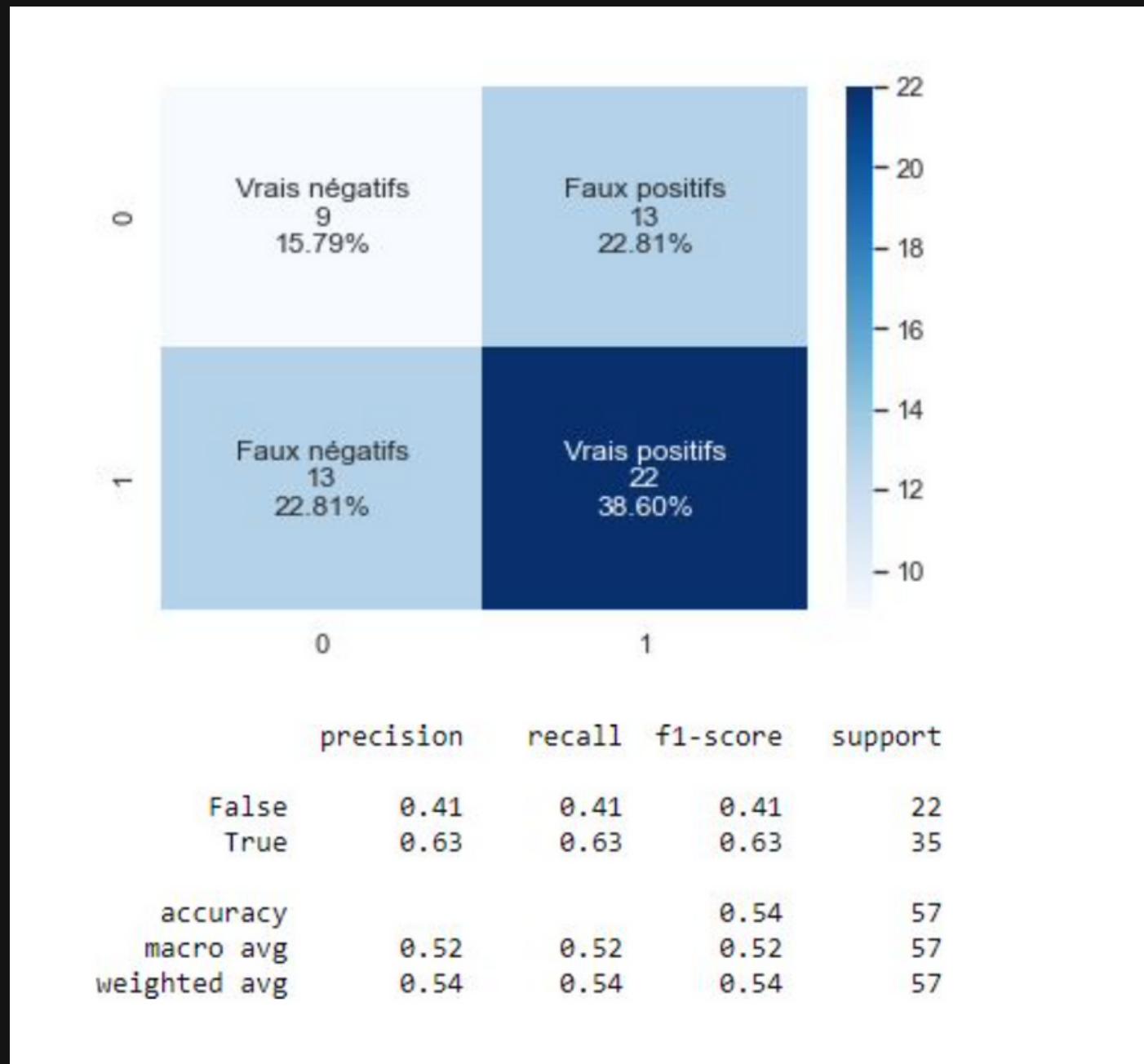
X_train : 113 lignes
y_train : 113 lignes
X_test : 57 lignes
y_test : 57 lignes
```

	diagonal	height_left	height_right	margin_low	margin_up	length
27	0.26	0.55	1.01	0.44	-1.19	1.00
78	0.72	1.09	-0.24	-1.20	0.63	-0.02
147	1.01	1.53	0.88	0.05	1.10	-2.27
38	0.88	0.68	0.25	-0.55	-1.61	1.31
41	-0.43	0.11	-0.72	-0.46	-0.94	0.16

	diagonal	height_left	height_right	margin_low	margin_up	length
139	-1.12	1.02	0.82	1.73	-0.38	0.29
30	0.82	-0.06	-0.36	-1.02	0.21	1.03
119	-1.41	0.21	-0.09	0.54	1.82	-1.45
29	-0.33	-1.06	-1.66	-0.76	-2.00	1.25
144	-1.25	-0.90	-0.18	1.50	-0.81	0.41

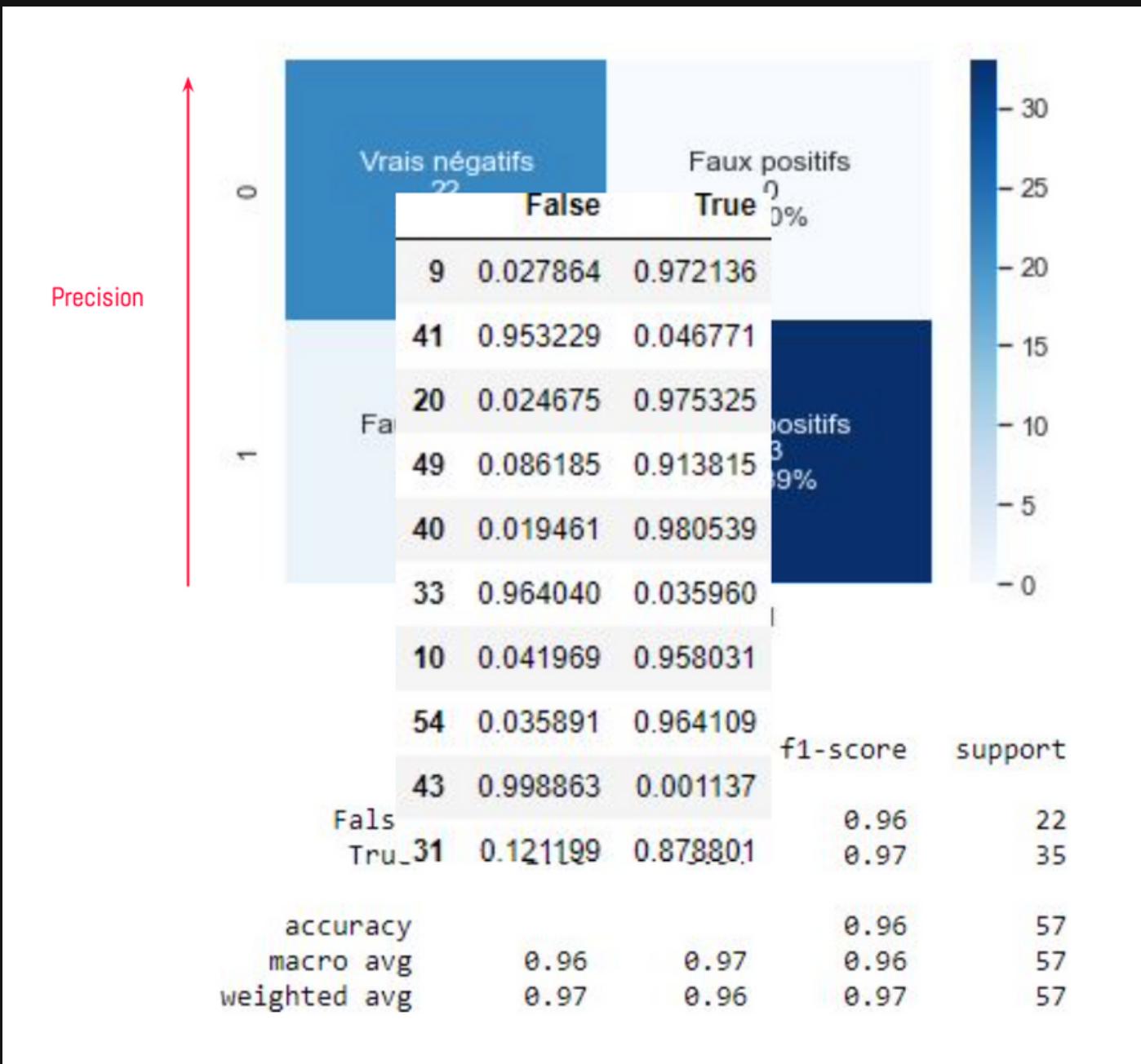
- Séparation du dataset en 2 jeux
 - train : 113 billets
 - test : 57 billets
- Séparation de chaque jeu entre X et y
 - X = explicatives (mesures)
 - y = expliquée (is_genuine)

Dummy classification



- Base de comparaison avec le futur modèle
- Prédictions aléatoires
 - Donc médiocres
- Précision très faible
 - 63 % de vrais billets correctement prédits / vrais billets prédit

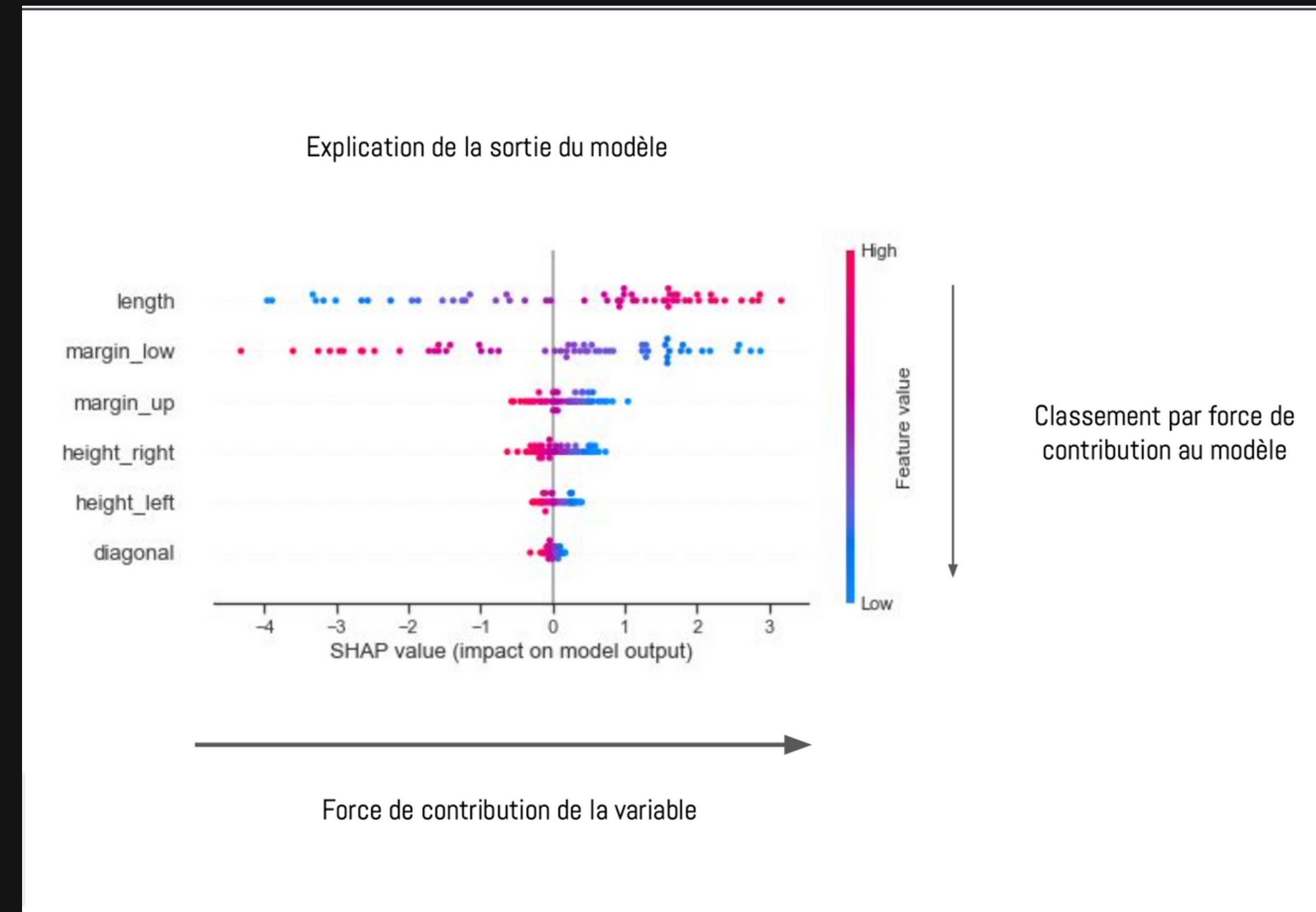
Résultats de la modélisation



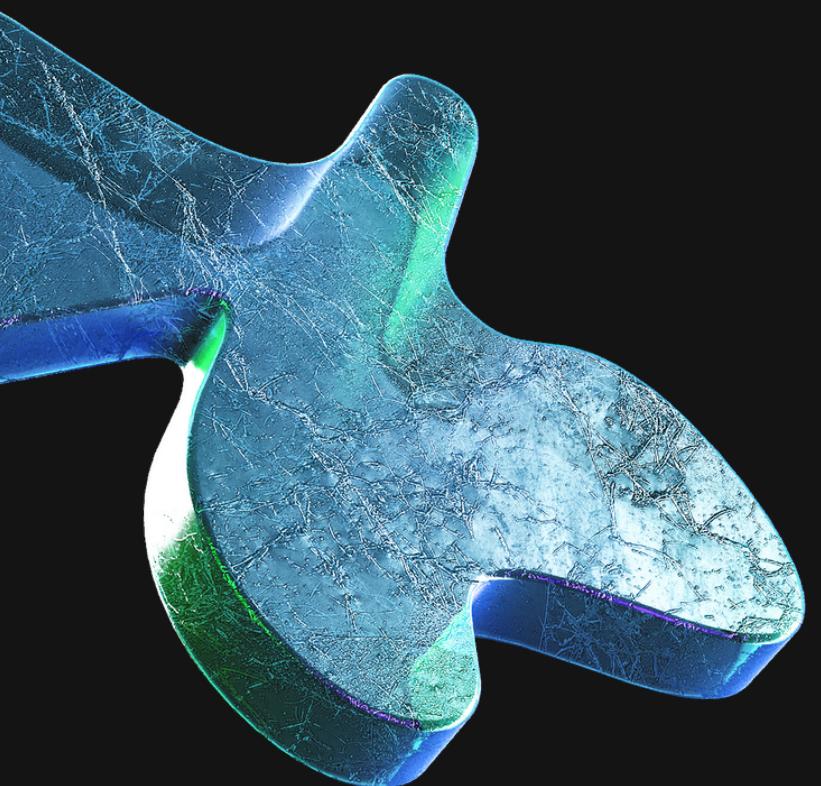
- Probabilités entre 0 et 1
 - Pour chaque classe et chaque billet
- Comparaison avec `y_test`
- Aucun faux billets détecté vrai
 - précision de 100 % des vrais billets
- 2 vrai billets détectés faux

Contribution au modèle

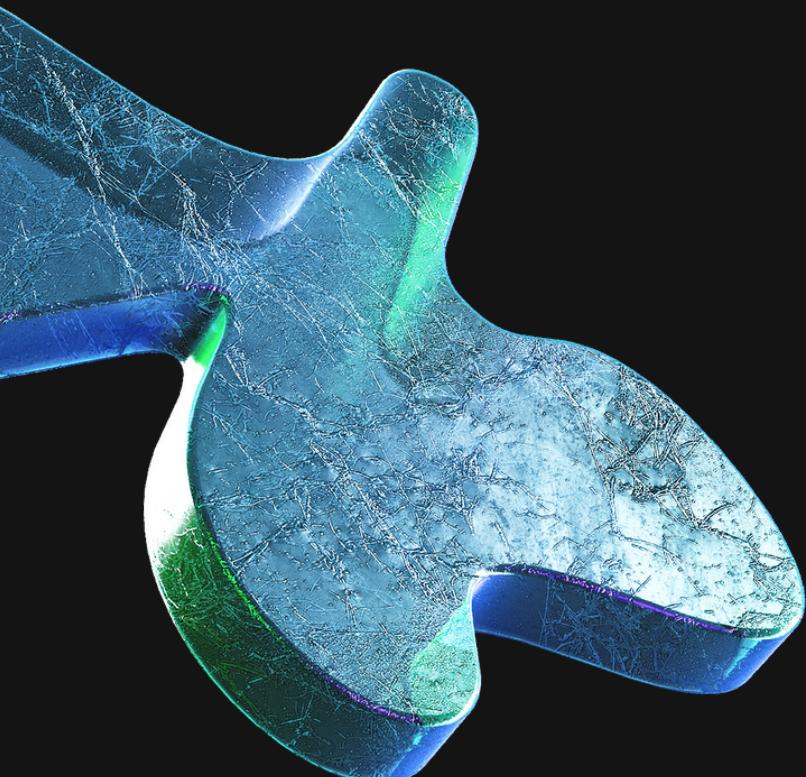
Rouge : valeurs les + fortes
Bleu : valeurs les + faibles



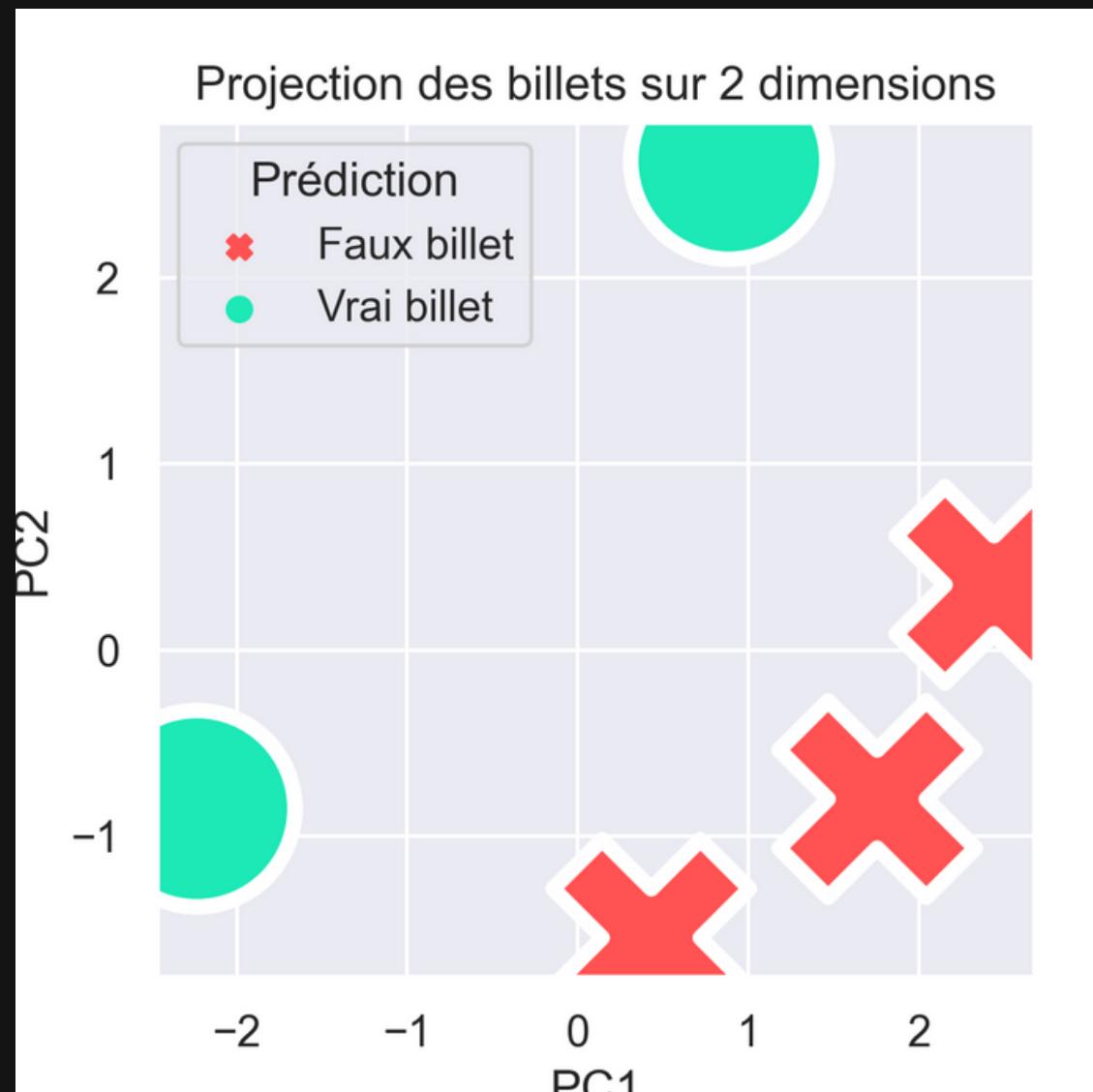
Partie 3



Programme de détéction



Programme de détection



- Fichier CSV en entrée
- Modèles enregistrées dans un fichier Pickle
 - StandardScaler
 - ACP
 - Régression Logistique
- Résultats sous forme de tableau
- Contrôle des clusters sur 2 dimensions

Prédiction	Probabilité de faux	Probabilité de vrai	id
Faux billet	0.9628985039063538	0.03710149609364621	A_1
Faux billet	0.9941019121655923	0.0058980878344077215	A_2
Faux billet	0.9868901541853136	0.013109845814686471	A_3
Vrai billet	0.05872220617394963	0.9412777938260504	A_4
Vrai billet	0.004059441447997858	0.9959405585520021	A_5

Avez-vous des
questions ?

