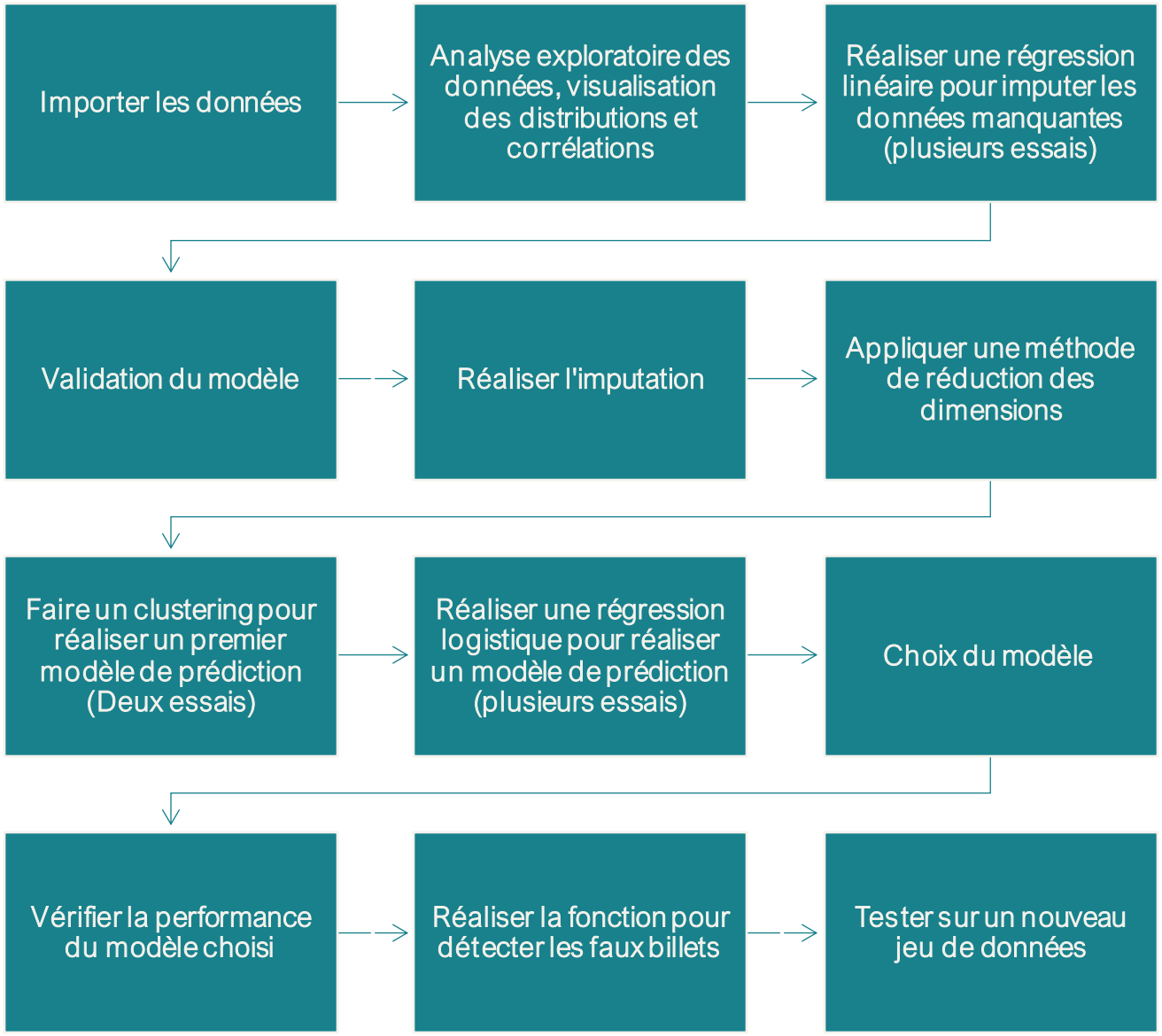
Détecter des faux billets

- Pauline Felten
- Consultante Data Analyst

Objectif & Démarche

Mettre en place une modélisation capable d'identifier automatiquement les vrais des faux billets à partir de certaines dimensions qui les composent



Analyse exploratoire

Résultats

Variables quantitatives (dimensions) : 6

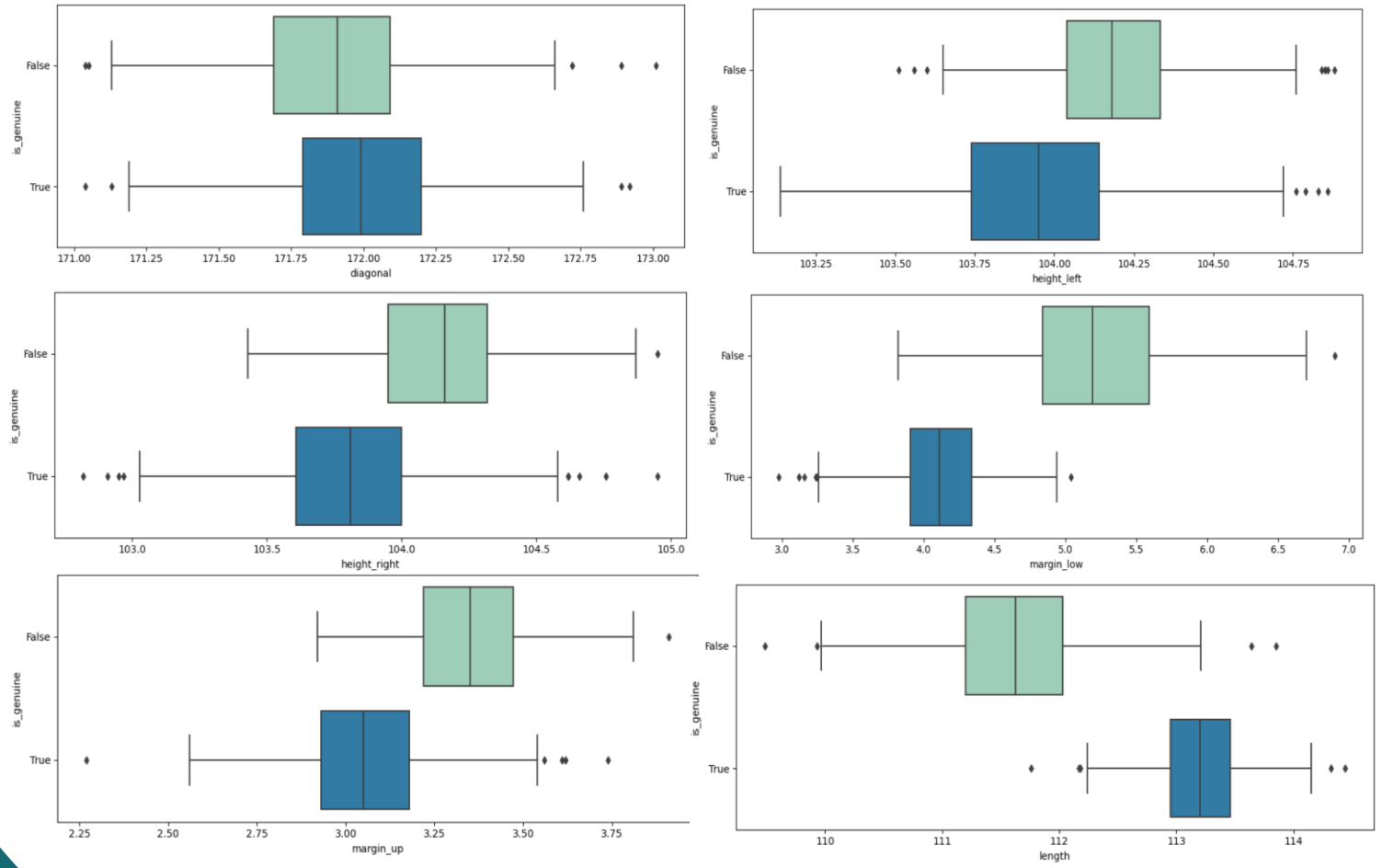
Variable qualitative (binaire) : 1 (vrai ou faux billet)

Données manquantes : 37 lignes (2.5% dataset) sur la variable "margin_low" (vrais et faux billets)

Dataset : 2/3 vrais billets pour 1/3 de faux billets

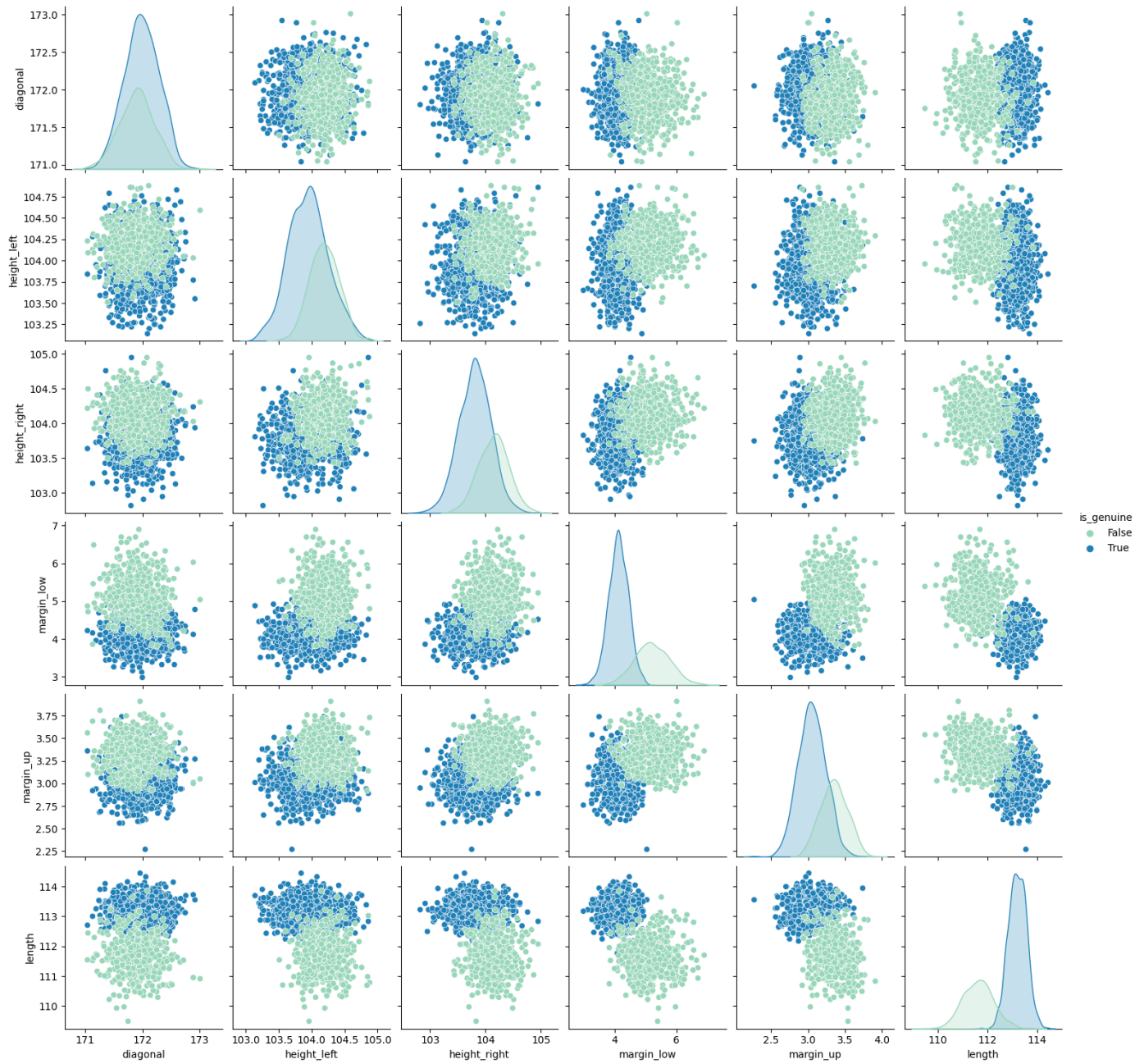
Pas d'outliers

Distribution des données



Les faux billets sont moins longs, plus larges et les marges hautes et basses sont plus élevées que les vrais.

Distribution des données



Imputation des données manquantes

Matrice des corrélations

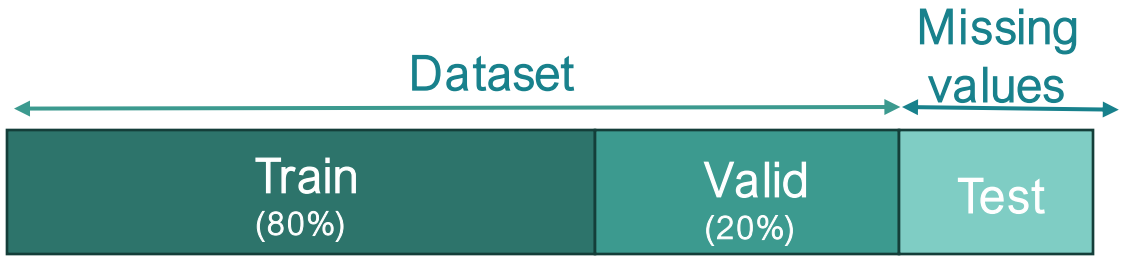
Régression linéaire

- Prédire la valeur de data inconnue (var à expliquer, y) en fonction de data apparentées et connues (var explicative(s), x)
- Determiner une droite passant au plus près des points du nuage de points allongé

Doit répondre à **4 hypothèses** :
- Existence d'une relation linéaire
- Indépendance résiduelle
- Normalité des résidus
- Homocedasticité

Analyse des résultats :
- **Coefficient de determination** : mesure la qualité de la prédiction
- **MSE** : différence entre valeurs prédites et valeurs observées
- **RMSE** : racine carrée de MSE
- **MAE** : différence moyenne entre VP et VO

Régression linéaire simple



- Résultats :

Mean Squared Error (MSE) sur l'ensemble de test: 0.2050

Mean Absolute Error (MAE) sur l'ensemble de test: 0.3443

Root Mean Squared Error (RMSE) sur l'ensemble de test: 0.4527

Coefficient de détermination R2: 0.5136

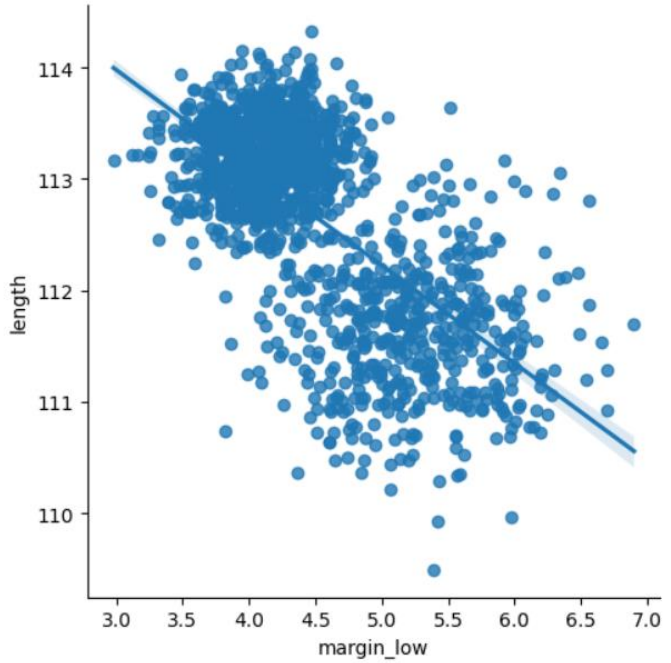
- Validation du modèle :

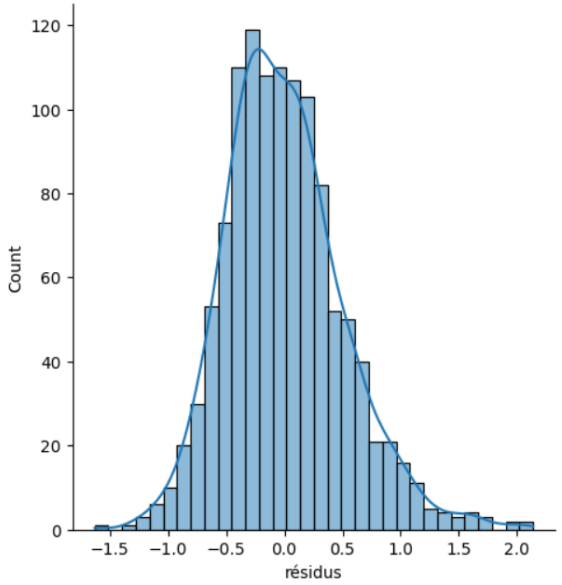
Kolmogorov et Shapiro: Les résidus ne suivent pas une distribution normale.

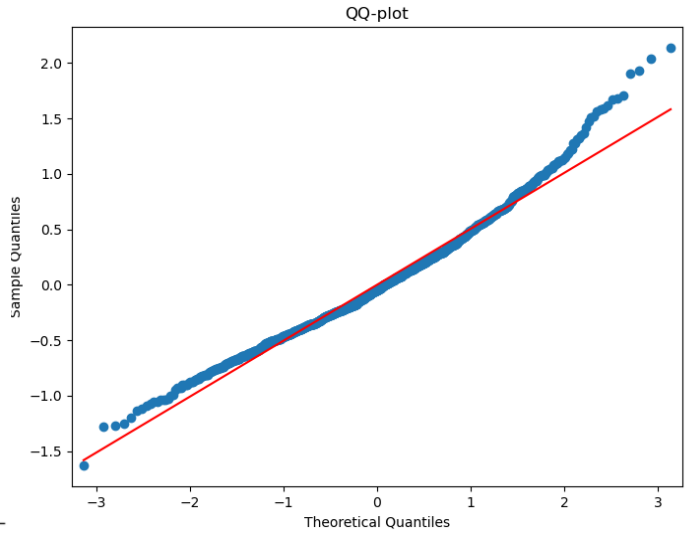
Test de Breusch-Pagan : On constate une hétéroscédasticité.

Test de Durbin Watson: 1.9695 H0 : pas d'auto corrélation

Facteurs d'inflation de la variance (VIF) : 1.0







Régression linéaire multiple

- Essais :

Essai 4 : Les scores sont significativement à l'introduction de la variable is_genuine comme variable explicative

Mean Squared Error (MSE): 0.1942 => 0.1396

Mean Absolute Error (MAE) : 0.3366 => 0.2887

Root Mean Squared Error (RMSE) : 0.4406 => 0.3737

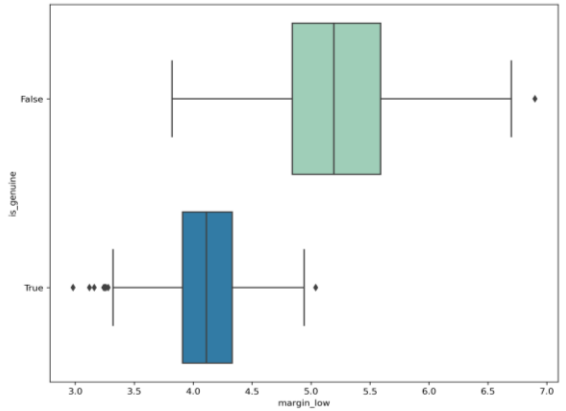
Coefficient de détermination R2: 0.5392 => 0.6686

Essai 6 : meilleur que le 5ème on exclu la variable diagonale qui est très peu correlée.

Mean Squared Error (MSE): 0.1373 => 0.1372
Mean Absolute Error (MAE) : 0.2894 => 0.2894
Root Mean Squared Error (RMSE) : 0.3706 => 0.3705
Coefficient de détermination R2: 0.6740 => 0.6742

- Imputation

Prédire les données manquantes selon essai 6 et concatener avec notre dataset.



- Validation du modèle :

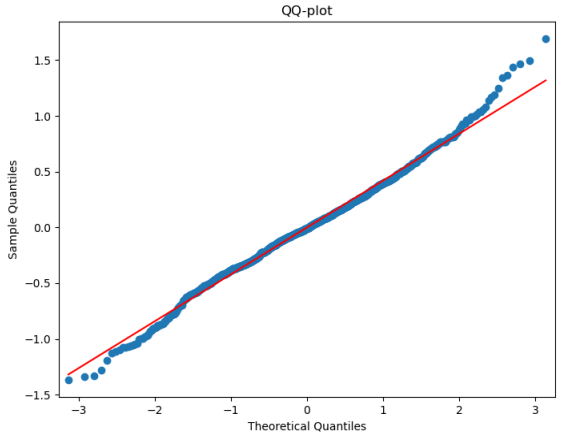
Kolmogorov et Shapiro: Les résidus ne suivent pas une distribution normale.

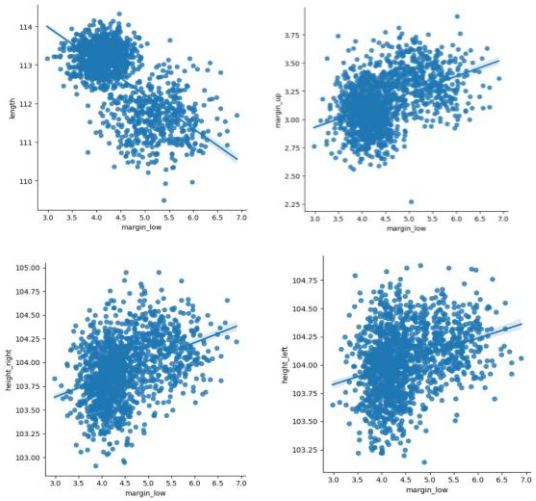
Test de Breusch-Pagan : On constate une hétéroscédasticité.

Test de Durbin Watson: 2.02304 H0 : pas d'auto corrélation

Facteurs d'inflation de la variance (VIF) :

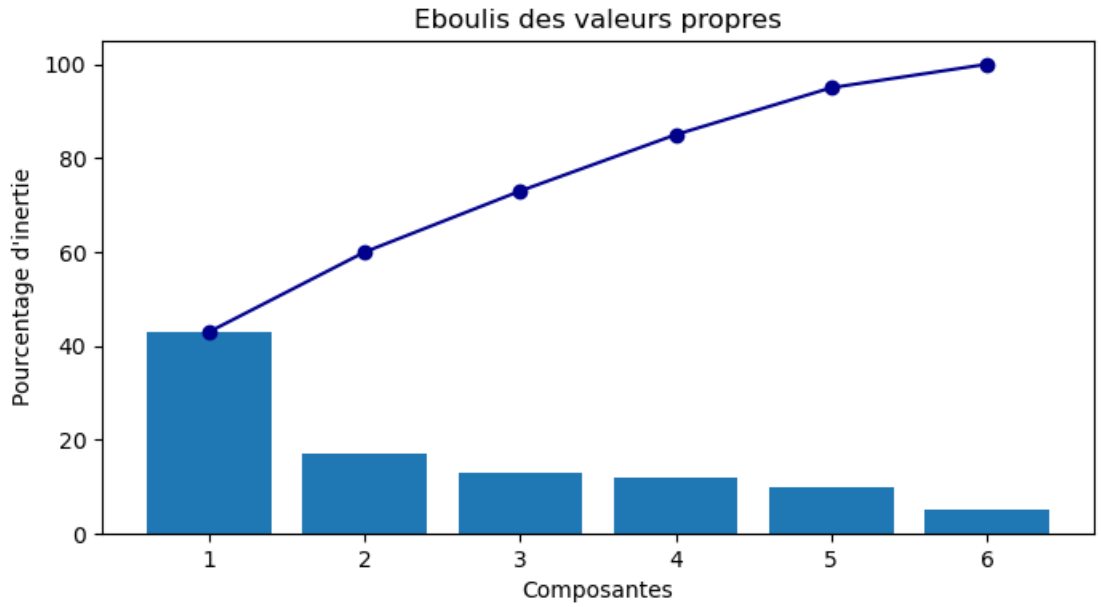
0	1.171731
1	1.317546
2	1.645780
3	3.478213
4	4.562063





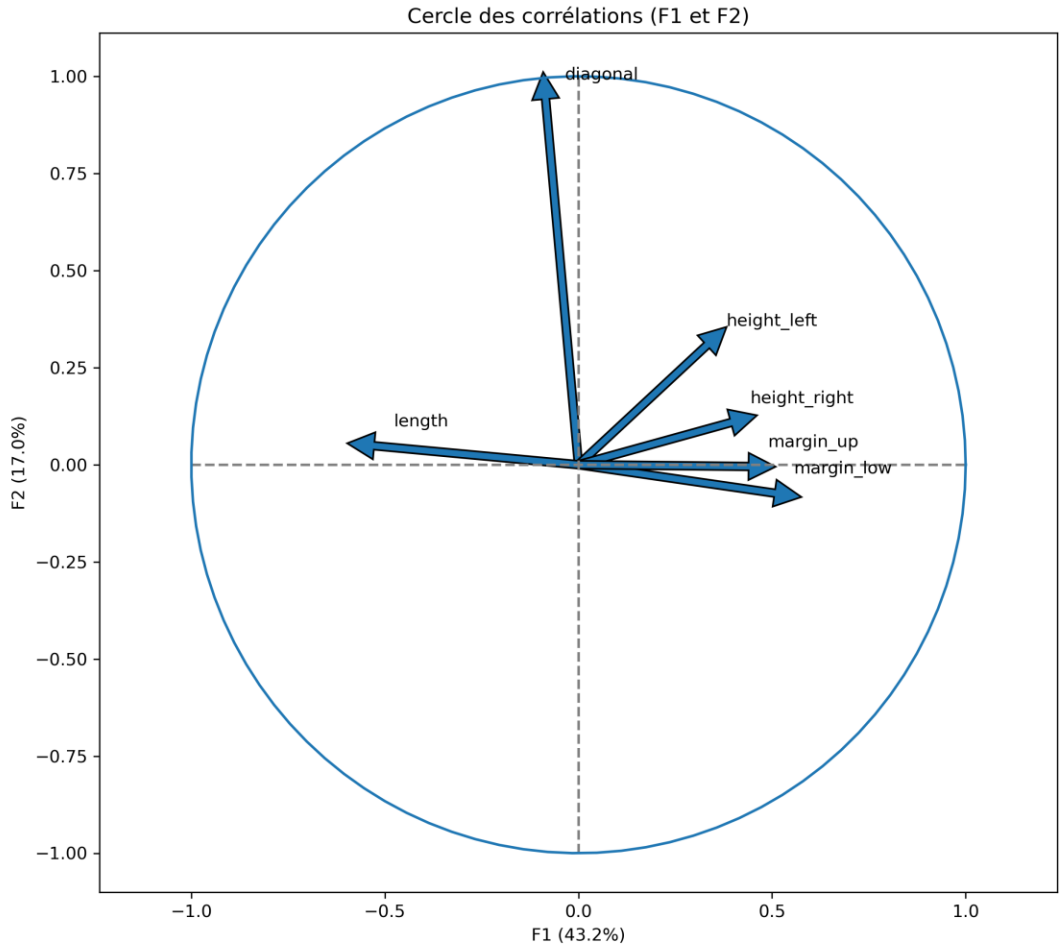
ACP

Eboulis des valeurs propres

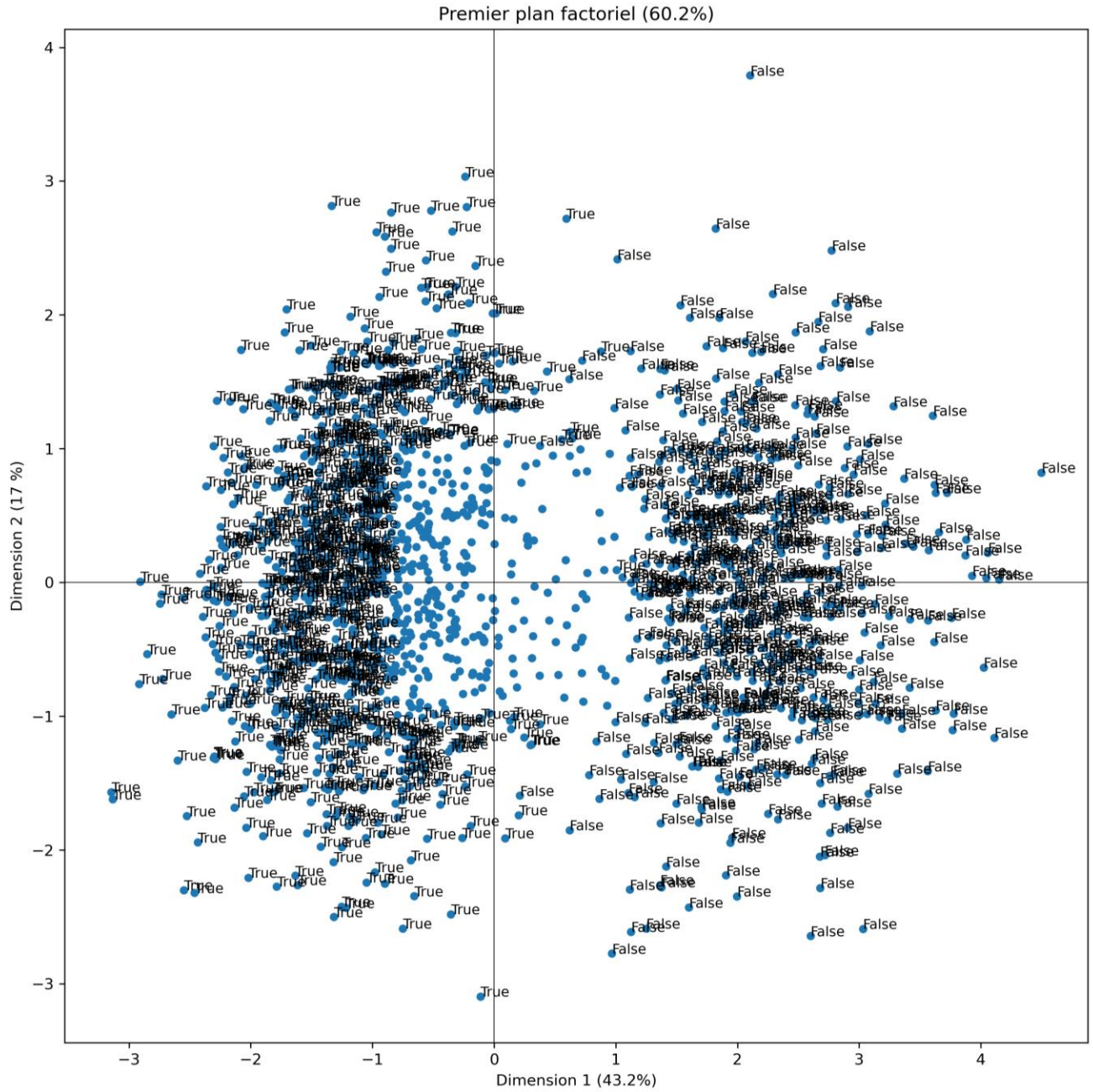


	F1	F2	F3	F4	F5	F6
diagonal	-0.085	0.941	-0.287	-0.103	-0.117	0.008
height_left	0.331	0.307	0.885	-0.047	0.104	0.006
height_right	0.394	0.108	-0.166	0.866	0.234	0.004
margin_low	0.507	-0.073	-0.106	-0.090	-0.571	0.627
margin_up	0.439	-0.004	-0.271	-0.444	0.710	0.181
length	-0.528	0.049	0.150	0.177	0.302	0.758

Cercle des corrélations

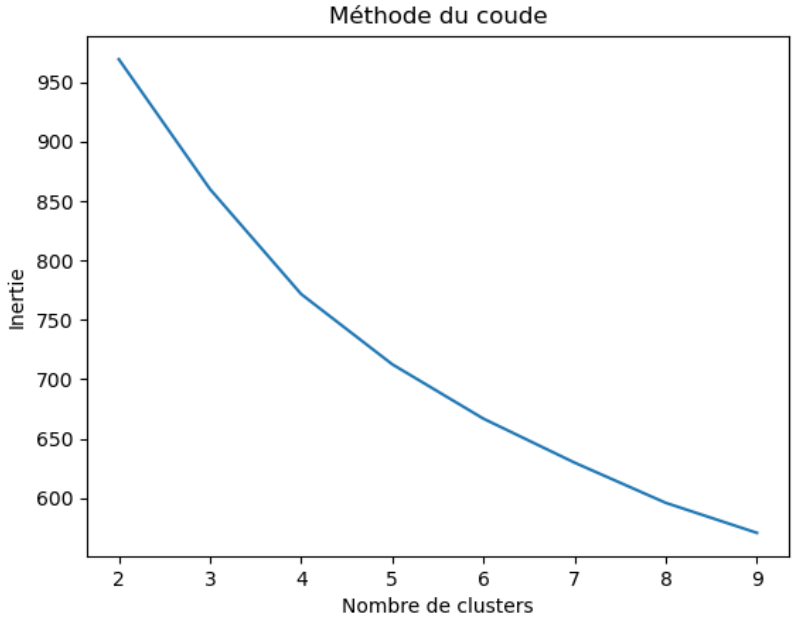


Premier plan factoriel



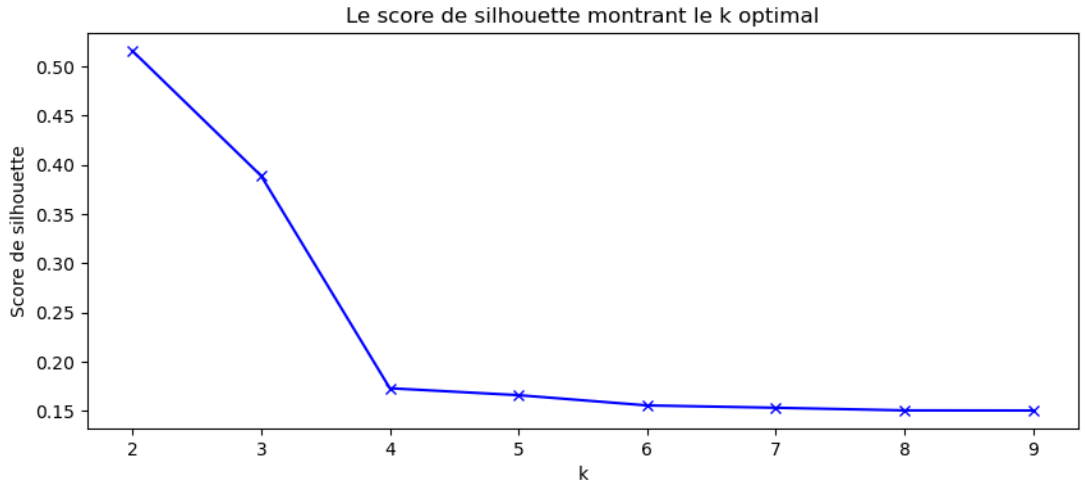
KMeans

KMeans sur données d'origine



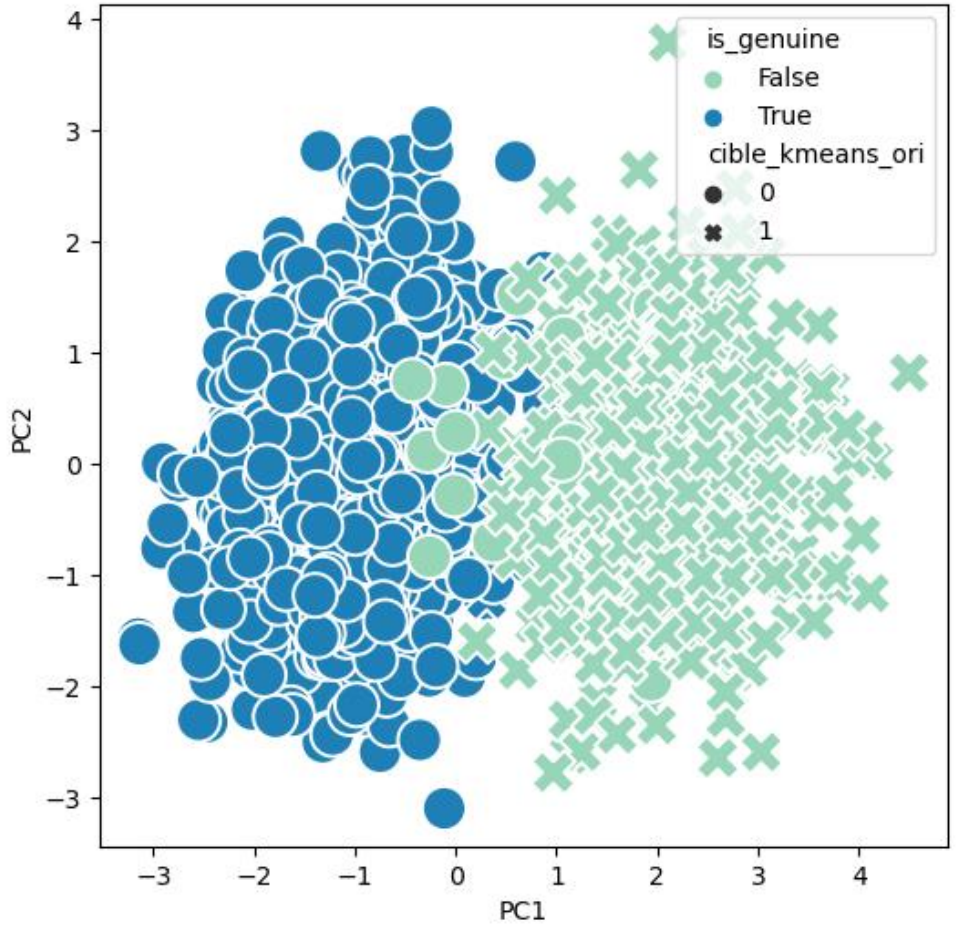
cible_kmeans_ori	diagonal	height_left	height_right	margin_low	margin_up	length
0	171.99	103.95	103.81	4.12	3.06	113.20
1	171.90	104.19	104.15	5.24	3.35	111.59

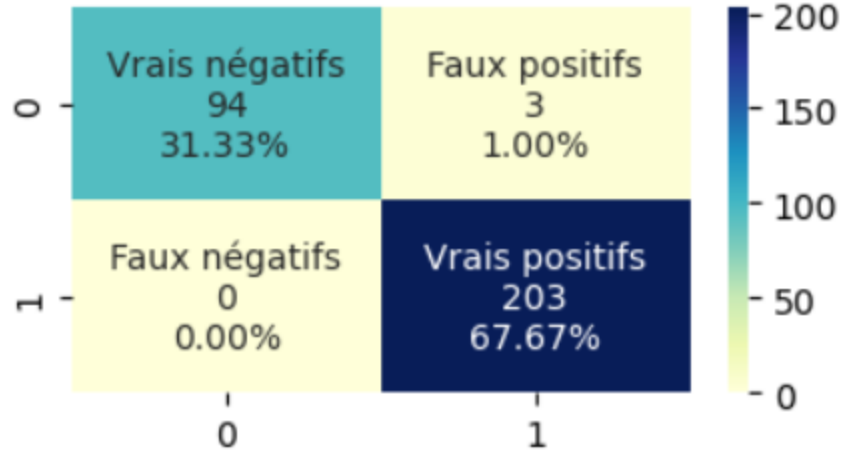
Homogeneity score : 0.917
Calinski Harabasz Score : 394.908



cible_kmeans_ori is_genuine	False	True
False	481	19
True	2	998

KMeans sur données d'origine

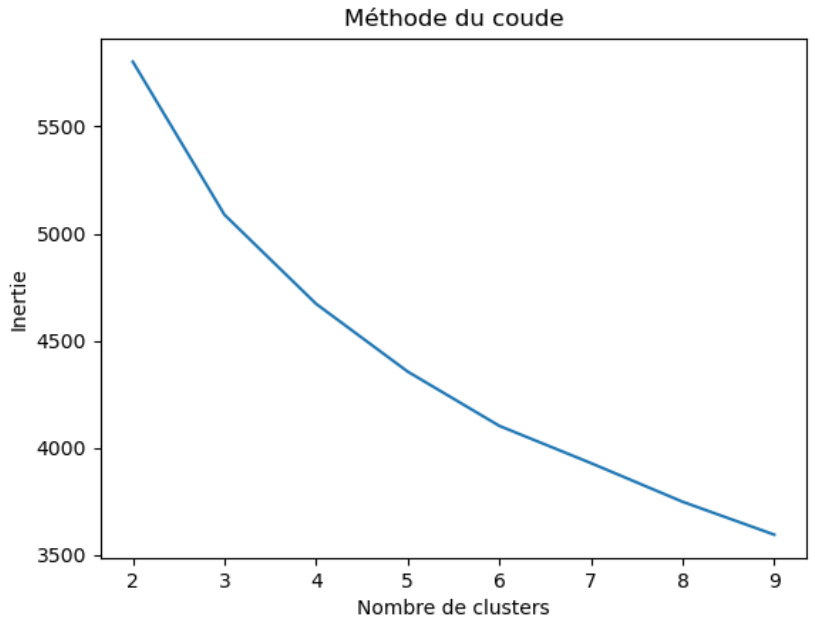




	precision	recall	f1-score	support
False	1.00	0.97	0.98	97
True	0.99	1.00	0.99	203
accuracy			0.99	300
macro avg	0.99	0.98	0.99	300
weighted avg	0.99	0.99	0.99	300

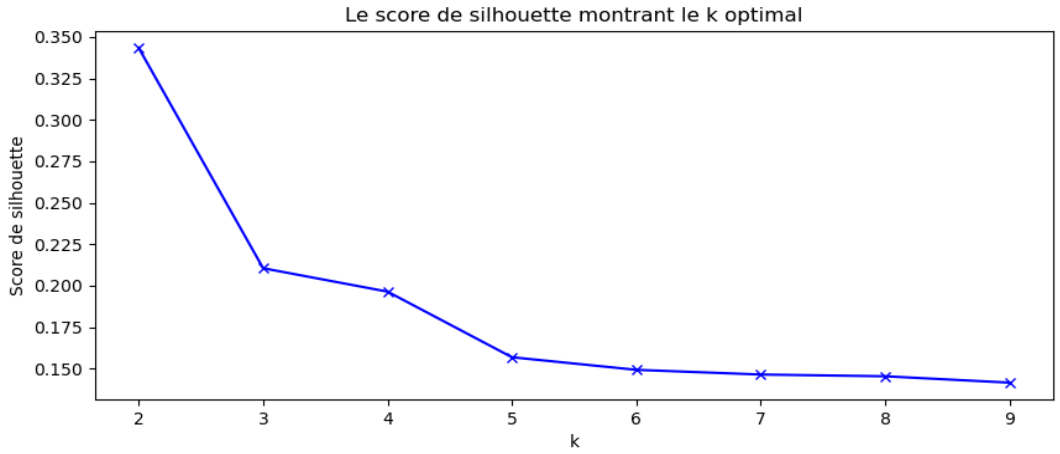
Precision : % des prédictions correctes positives sur le nombre total des prédictions positives
Recall sensibilité : % des prédictions correctes positives sur le nombre total des données positives
F-1 Score : moyenne harmonique entre precision et recall
Accuracy : mesure le taux de prédictions correctes sur l'ensemble des individus

KMeans sur ACP



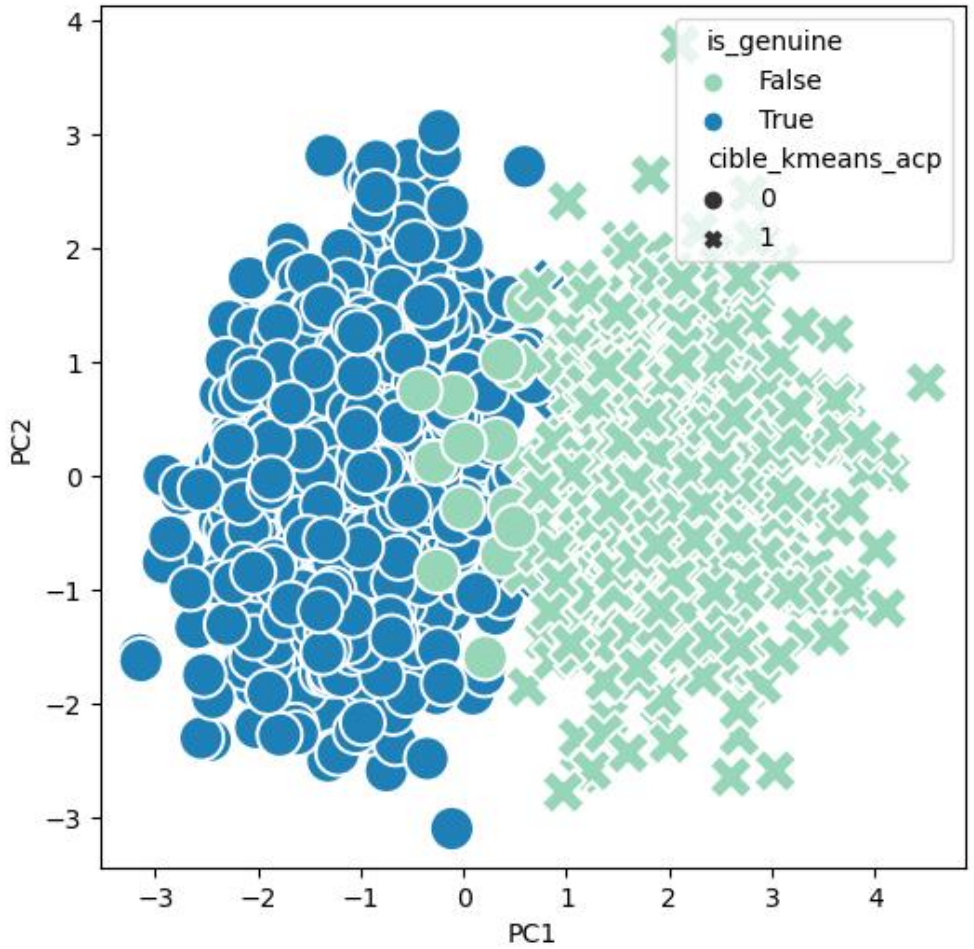
cible_kmeans_acp	diagonal	height_left	height_right	margin_low	margin_up	length
0	171.99	103.95	103.81	4.12	3.05	113.20
1	171.90	104.20	104.15	5.22	3.35	111.63

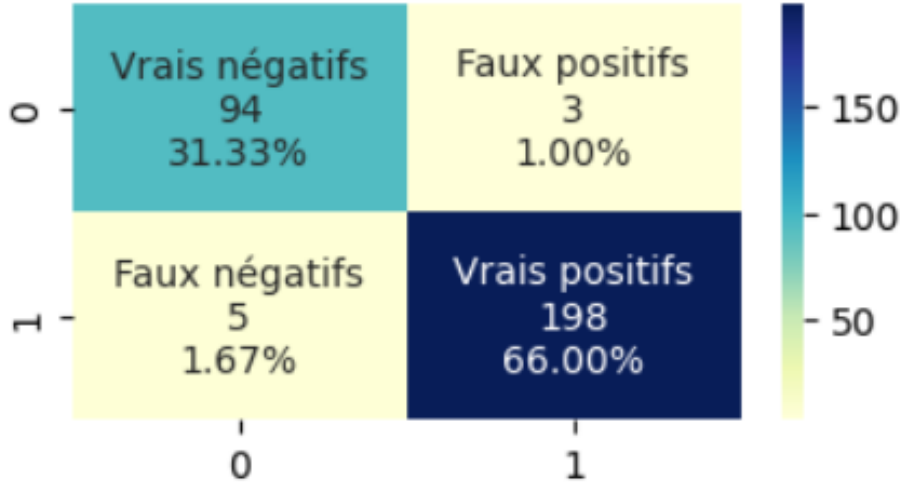
Homogeneity score: 0.813
Calinski Harabasz Score : 147.579



cible_kmeans_acp	False	True
is_genuine		
False	486	14
True	10	990

KMeans sur ACP





	precision	recall	f1-score	support
False	0.95	0.97	0.96	97
True	0.99	0.98	0.98	203
accuracy			0.97	300
macro avg	0.97	0.97	0.97	300
weighted avg	0.97	0.97	0.97	300

=> Qualité des predictions degradées

Régression logistique

Régression logistique

- Expliquer une variable binaire par des observations réelles et nombreuses.
- Permets de prédire avec le plus de précision possible les valeurs prises par une variable binaire.

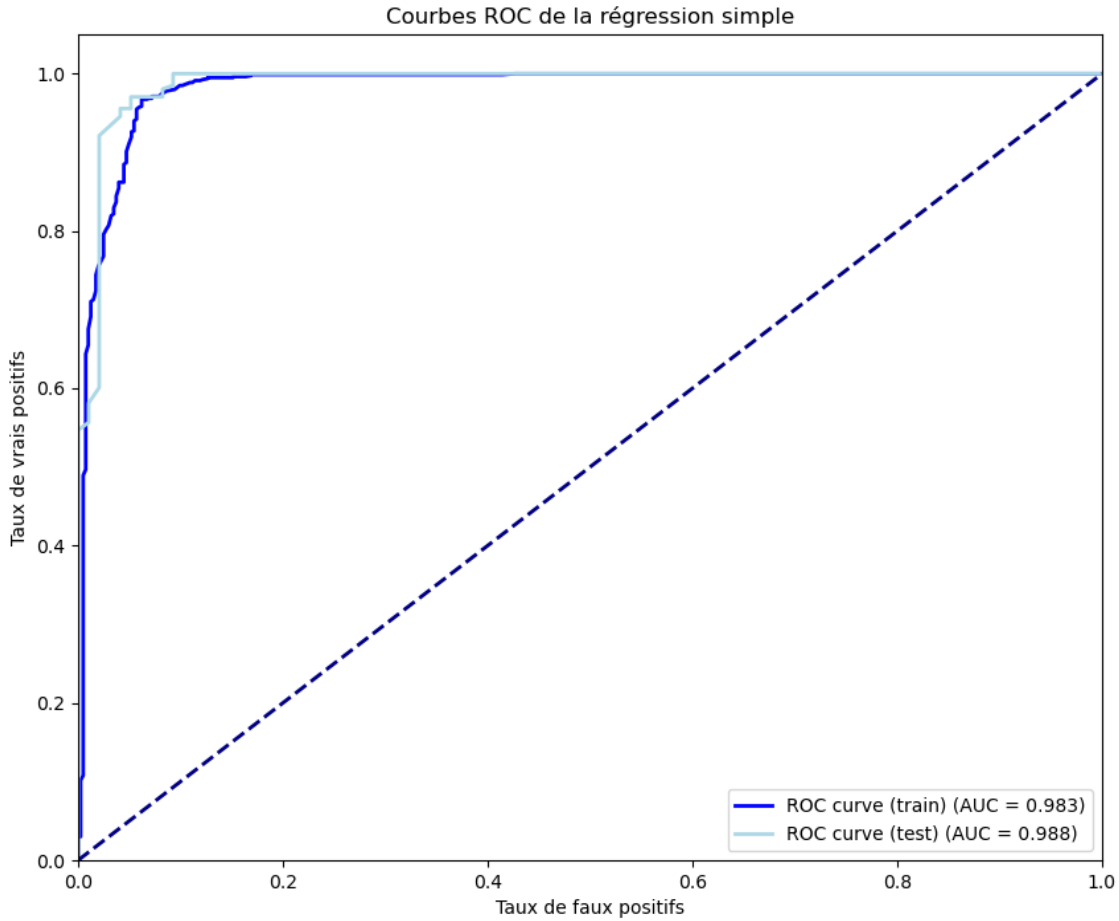
Evaluation du modèle :

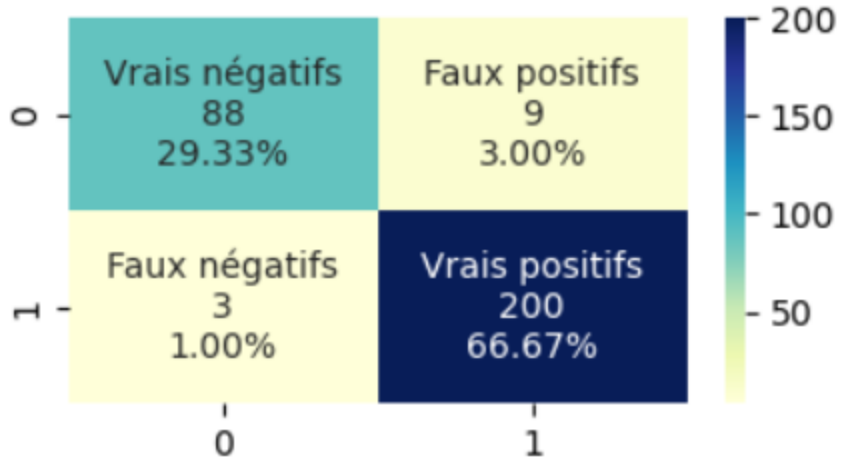
- **Matrice de confusion :** confronter données prédites avec les vraies valeurs
- **Courbe ROC :** evaluer l'exactitude des prédictions d'un modèle en tracant la sensibilité par rapport à la spécificité -
- Métrique associée **AUC :** Aire sous la courbe ROC

Analyse des résultats :

- Précision (PPV),
- Recall (TPR),
- Spécificité (TNR) : proportion d'individus négatifs effectivement bien détectés par le test
- Accuracy (ACC)
- F-1 Score

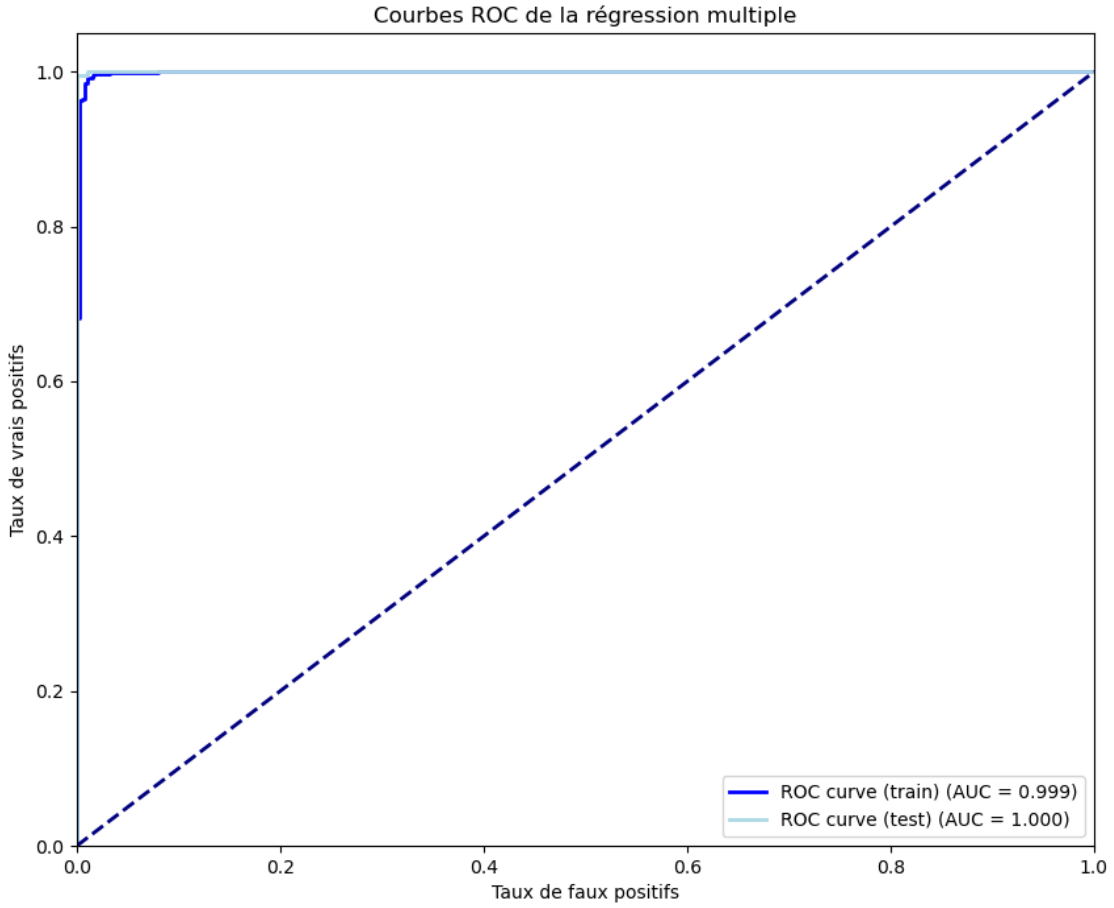
Régression Logistique Simple

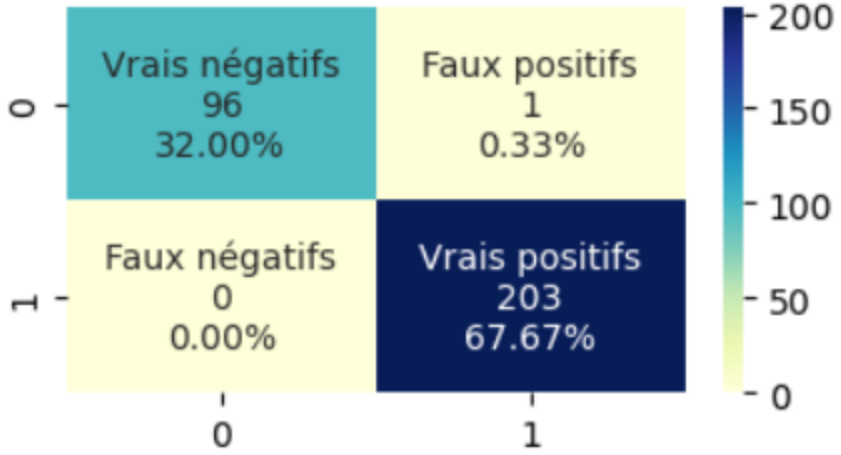




	precision	recall	f1-score	support
False	0.97	0.91	0.94	97
True	0.96	0.99	0.97	203
accuracy			0.96	300
macro avg	0.96	0.95	0.95	300
weighted avg	0.96	0.96	0.96	300

Régression Logistique Multiple

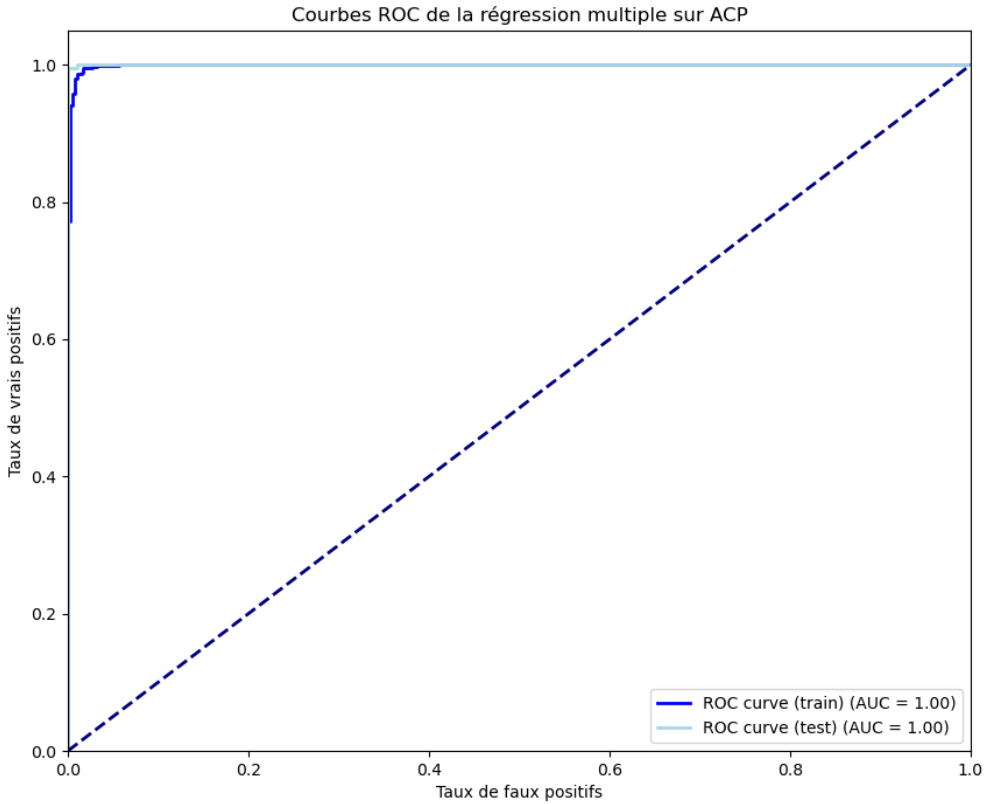


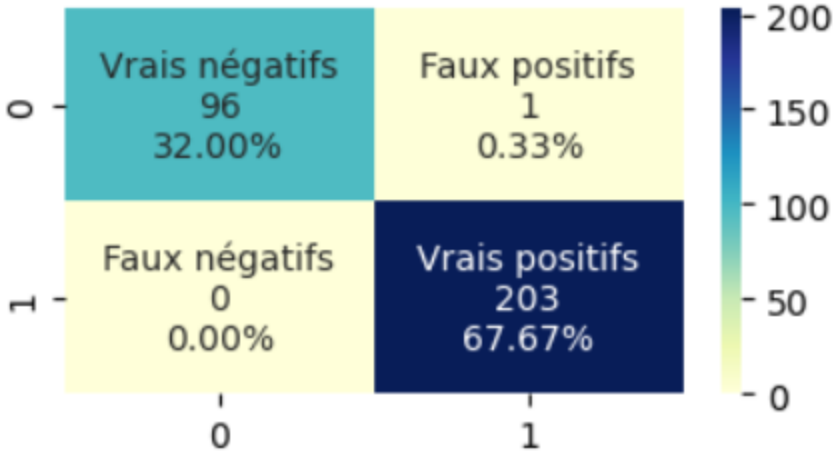


	precision	recall	f1-score	support
False	1.00	0.99	0.99	97
True	1.00	1.00	1.00	203
accuracy			1.00	300
macro avg	1.00	0.99	1.00	300
weighted avg	1.00	1.00	1.00	300

Régression Logistique Multiple sur ACP

=> Retrait des composantes 2 et 6 non significatives
=> Supprimer la corrélation entre les variables



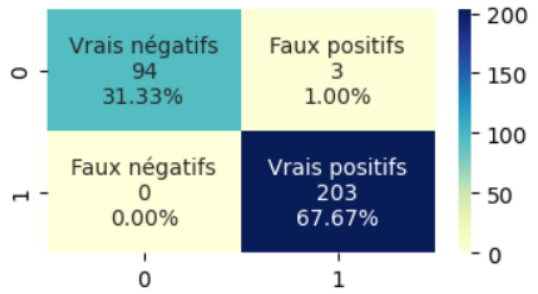


	precision	recall	f1-score	support
False	1.00	0.99	0.99	97
True	1.00	1.00	1.00	203
accuracy			1.00	300
macro avg	1.00	0.99	1.00	300
weighted avg	1.00	1.00	1.00	300

Choix du modèle

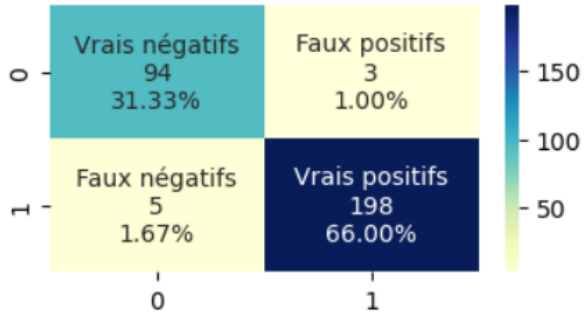
KMeans

Données d'origine :



	precision	recall	f1-score	support
False	1.00	0.97	0.98	97
True	0.99	1.00	0.99	203
accuracy			0.99	300
macro avg	0.99	0.98	0.99	300
weighted avg	0.99	0.99	0.99	300

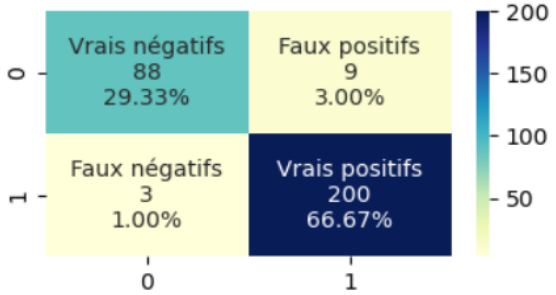
ACP :



	precision	recall	f1-score	support
False	0.95	0.97	0.96	97
True	0.99	0.98	0.98	203
accuracy			0.97	300
macro avg	0.97	0.97	0.97	300
weighted avg	0.97	0.97	0.97	300

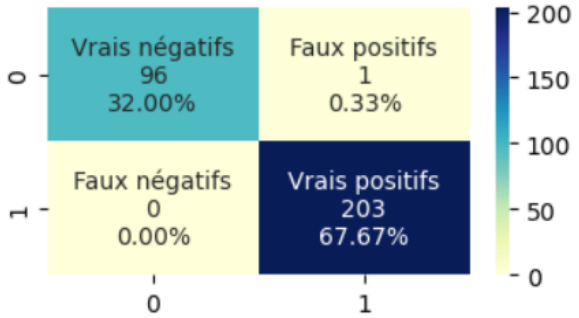
Régression logistique

Simple :



	precision	recall	f1-score	support
False	0.97	0.91	0.94	97
True	0.96	0.99	0.97	203
accuracy			0.96	300
macro avg	0.96	0.95	0.95	300
weighted avg	0.96	0.96	0.96	300

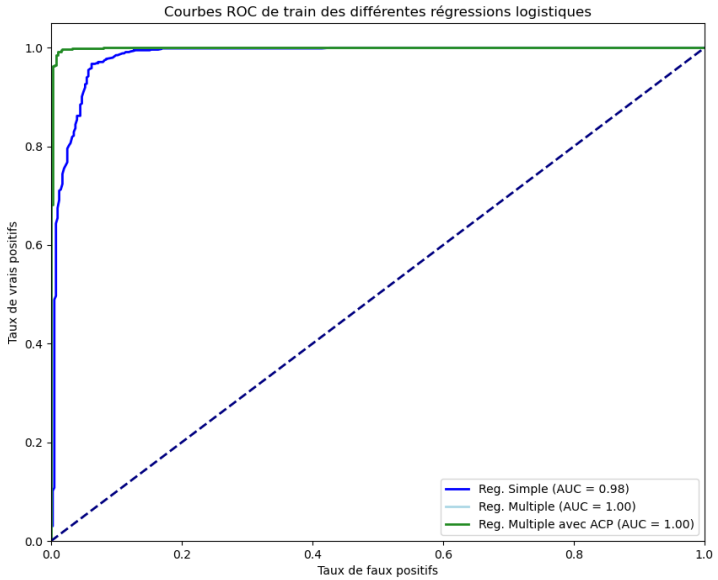
Multiple & ACP :

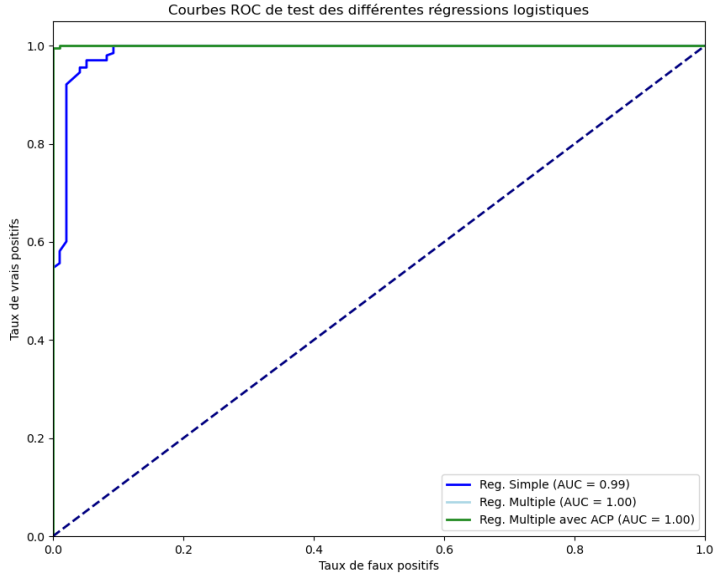


	precision	recall	f1-score	support
False	1.00	0.99	0.99	97
True	1.00	1.00	1.00	203
accuracy			1.00	300
macro avg	1.00	0.99	1.00	300
weighted avg	1.00	1.00	1.00	300

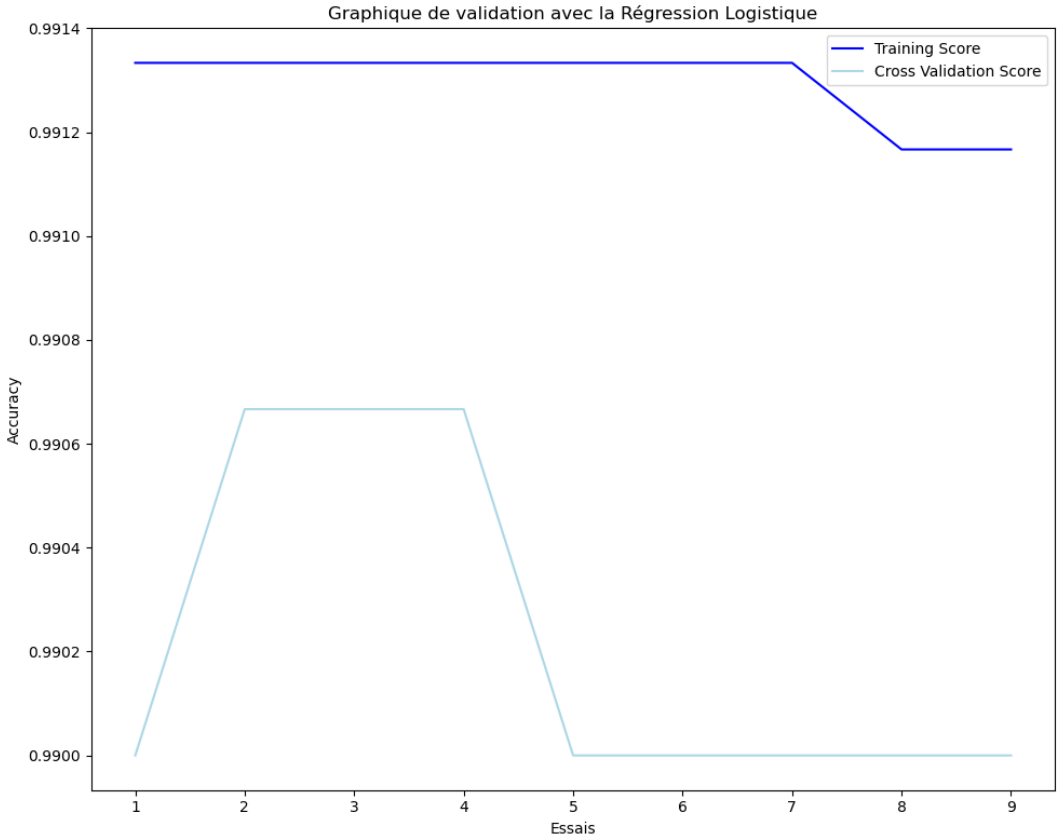
Performance du modèle choisi

Apprentissage





Validation croisée



Permet d'évaluer la fiabilité d'un modèle

Méthode des k folds : diviser l'ensemble de données de manière aléatoire en K folds.

Cross Validation Score : indique le score de chaque fold de test

Merci de votre attention !

Démonstration sur nouveau dataset