

Aide au choix d'un algorithme de détection.

Projet de détection de la qualité « vrai » - « faux ».

Sommaire.

Contexte de l'étude.

Collecte des données, visualisation, nettoyage.

Analyse descriptive, signifiance des variables.

Méthode de l'analyse en composantes principales.

- Détermination du nombre de composantes.
- Visualisation des nuages de points projetés.
- Représentation des variables synthétiques, « heatmap », synthèse.

Méthode des K-Means.

- Etude de l'effet de la qualité des données au travers de trois cas.
- Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
- Synthèse.

Méthode de régression logistique.

- Cas 1 et 2: Données d'entrée, matrices de confusion.
- Synthèse.

Conclusion.

Sommaire.

- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

Contexte de l'étude.

- Tout objet possède un attribut de valeur. Soit par la somme des couts engagés pour le produire, soit par sa capacité d'échange.
- La valeur d'un billet de banque, c'est sa capacité d'échange. La valeur du bien que le billet permet d'obtenir en contre partie.
- Cette valeur est aussi induite par le contrôle des autorités financières exercé sur la masse monétaire mise en circulation.
- On peut légitimement penser que cette valeur serait remise en question si une quantité significative de billets était émise hors du contrôle des autorités financières. Ce qui aurait pour conséquence de freiner l'achat de biens de consommations courants, les ménages n'ayant plus confiance dans leur monnaie.
- La question de l'authenticité de la monnaie se pose donc.
- L'analyse qui suit étudie différents scenarii de détection de vrais - faux billets. Et propose un modèle performant.

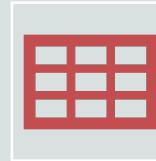
Sommaire.

- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

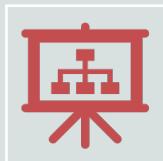
Collecte des données, visualisation, nettoyage.



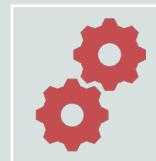
Après collecte des données, on obtient un objet de type »dataframe», 1500 lignes, 7 colonnes.



Données de type numérique pour 6 colonnes, booléen pour la 7ème.



Existence de valeurs manquantes: 37, soit 2.47% des 1500 lignes.



Remplacement des valeurs manquantes par méthode de régression linéaire.

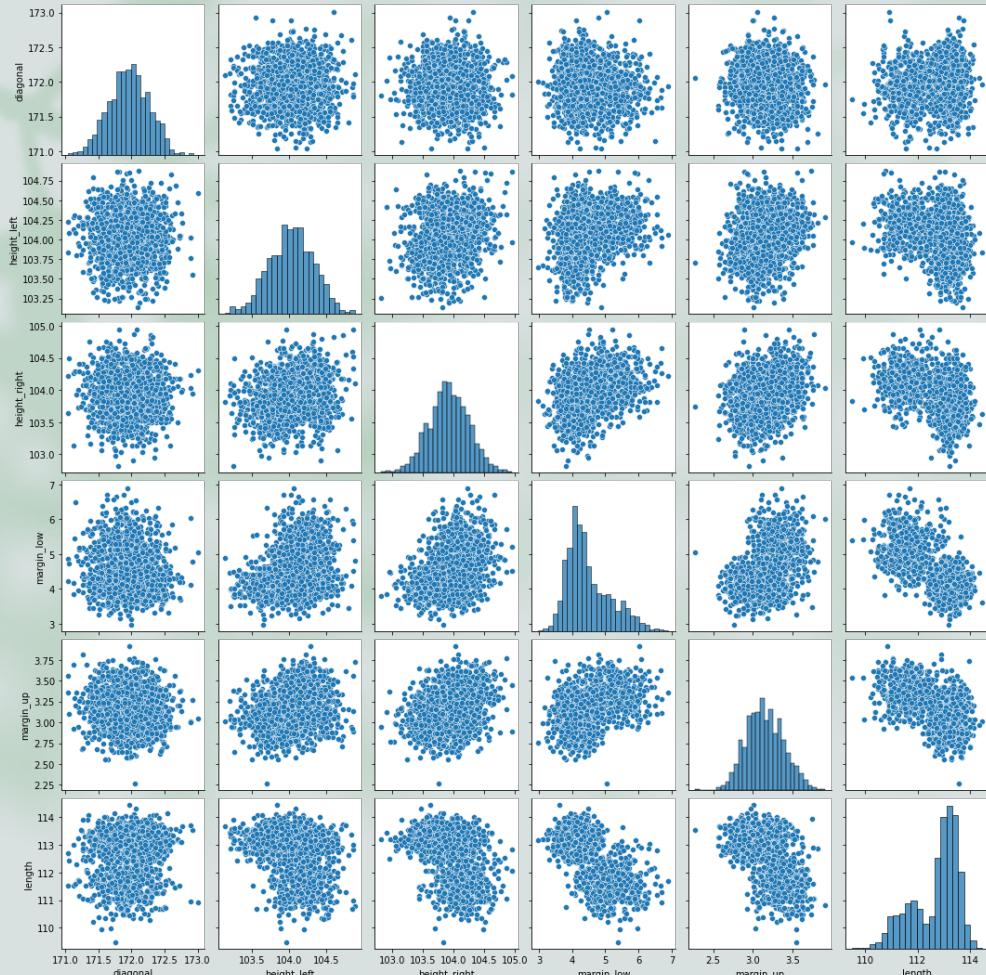
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   is_genuine  1500 non-null   bool   
 1   diagonal    1500 non-null   float64 
 2   height_left 1500 non-null   float64 
 3   height_right 1500 non-null  float64 
 4   margin_low   1463 non-null  float64 
 5   margin_up    1500 non-null  float64 
 6   length       1500 non-null  float64 
dtypes: bool(1), float64(6)
memory usage: 71.9 KB
```

is_genuine	False
diagonal	False
height_left	False
height_right	False
margin_low	True
margin_up	False
length	False
dtype: bool	

Sommaire.

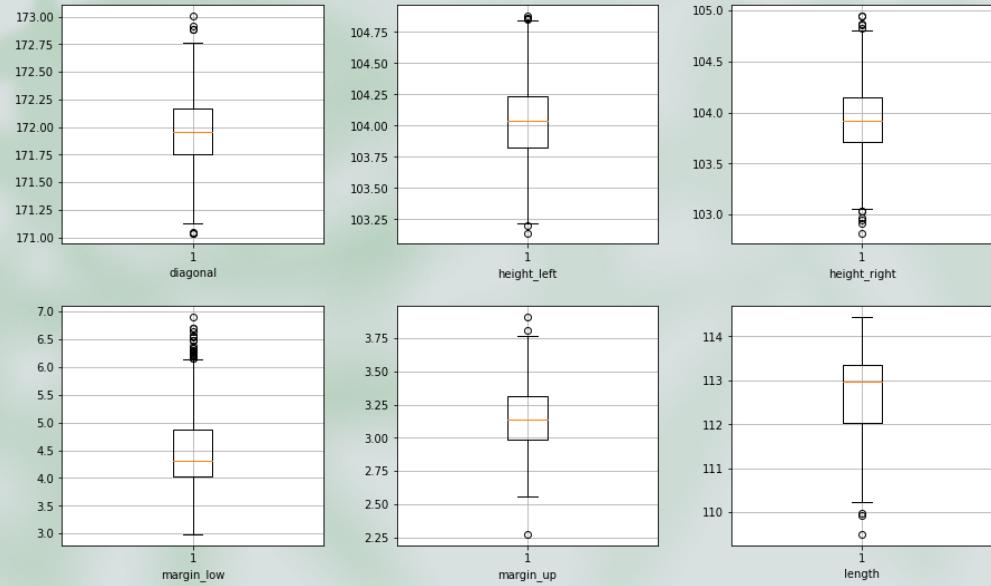
- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

Analyse descriptive, significance des variables.



- les variables margin_low et margin_up sont inversement corrélées avec length mais positivement corrélées avec les autres variables.
- Pour cette raison, les variables margin_low et margin_up et length peuvent être différentiantes.
- A l'exception de length les densités de distribution sont assez proches de la loi normale.

Analyse descriptive, significance des variables (suite).



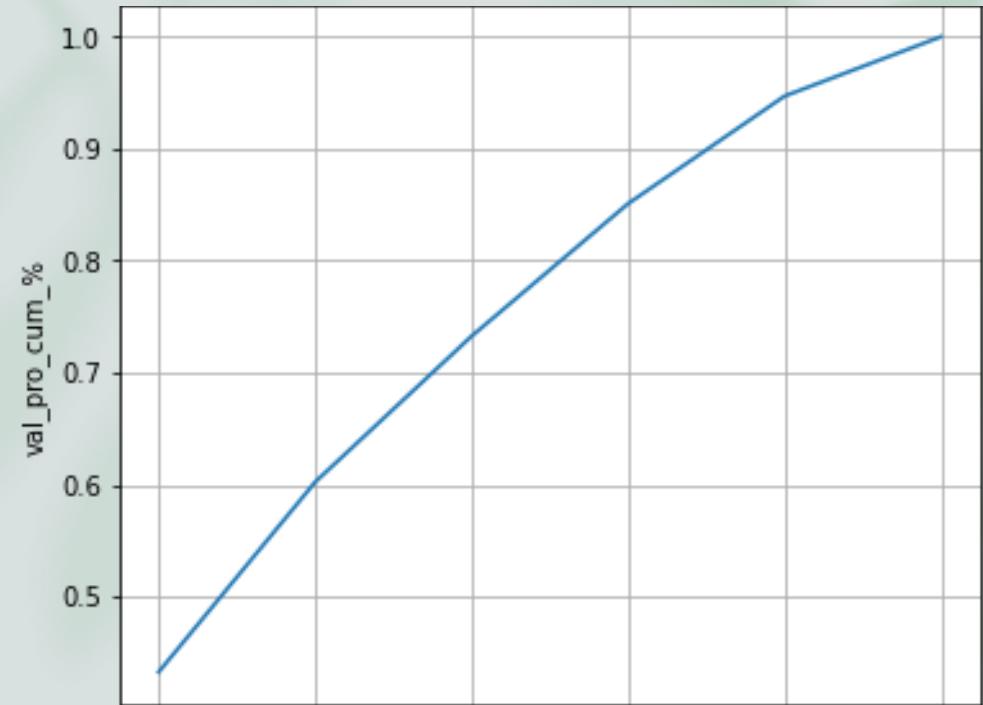
- Les valeurs sont regroupées autour de la médiane pour les variables 'diagonal', 'height_left', 'height_right', et 'margin_up'.
- Pour les autres variables, 'margin_low' et 'length' les valeurs ne sont pas distribuées de façon homogène, selon une cadence régulière.
- Le test statistique ANOVA montre que l'on peut rejeter l'hypothèse nulle énonçant que les variables sont identiques ou très proches.
- Le test statistique de Tukey (test par paires de variables) confirme le rejet de l'hypothèse nulle (les variables prises deux à deux sont identiques au seuil de 5%).

Sommaire.

- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

Détermination du nombre de composantes.

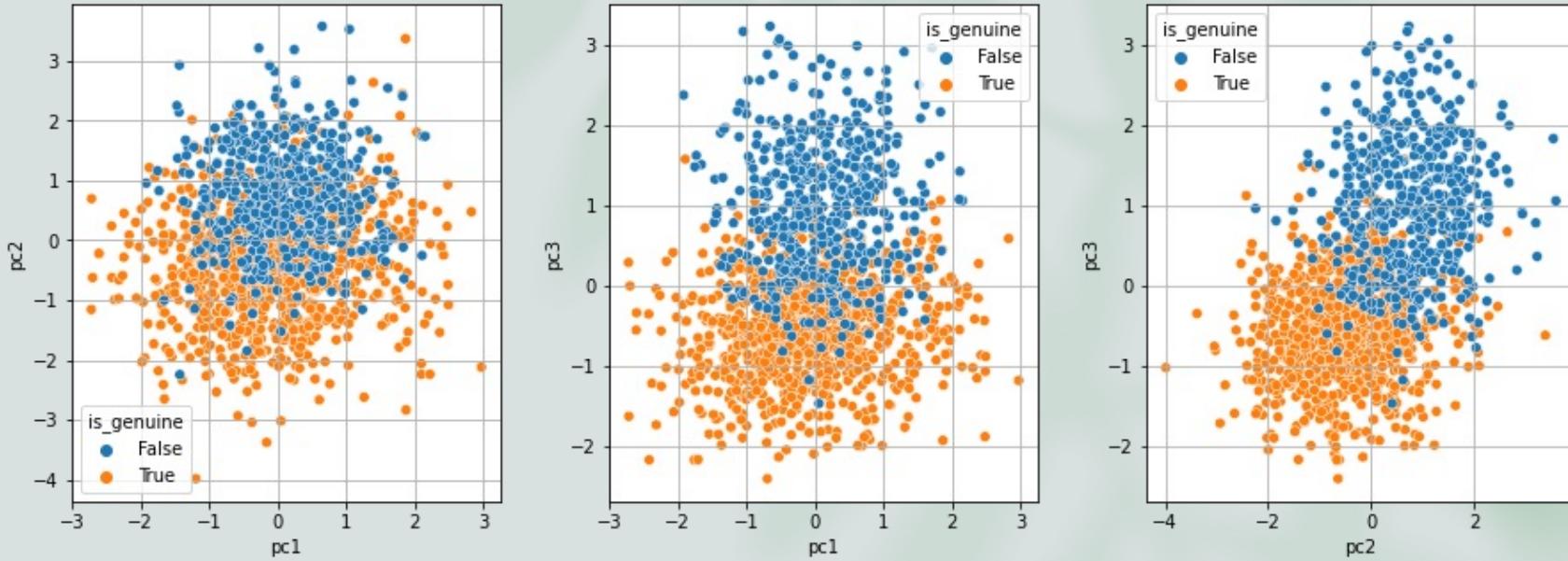
- L'ACP permet de réduire le nombre de variables en exprimant de nouvelles variables synthétiques par une fonction linéaire des variables originelles.
- En retenant 3 variables synthétiques, l'information captée est de 73%.
- Dans un premier temps, nous allons utiliser ces 3 variables et calculer une projection des individus successivement sur 3 plans factoriels.



Sommaire.

- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

Visualisation des nuages de points projetés.

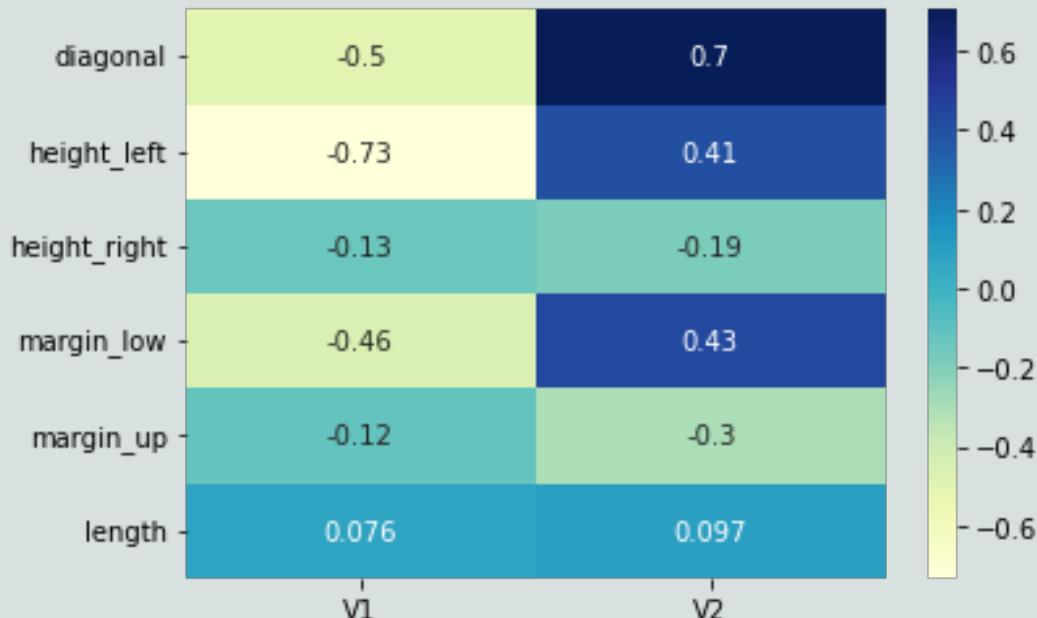


- A l'issue de la projection des individus sur les vecteurs, on visualise assez bien deux groupes d'individus mais la méthode ACP reste inopérante à départager nettement les deux groupes d'individus.
- Ce n'est pas une méthode que l'on peut utiliser pour réaliser cette détection.

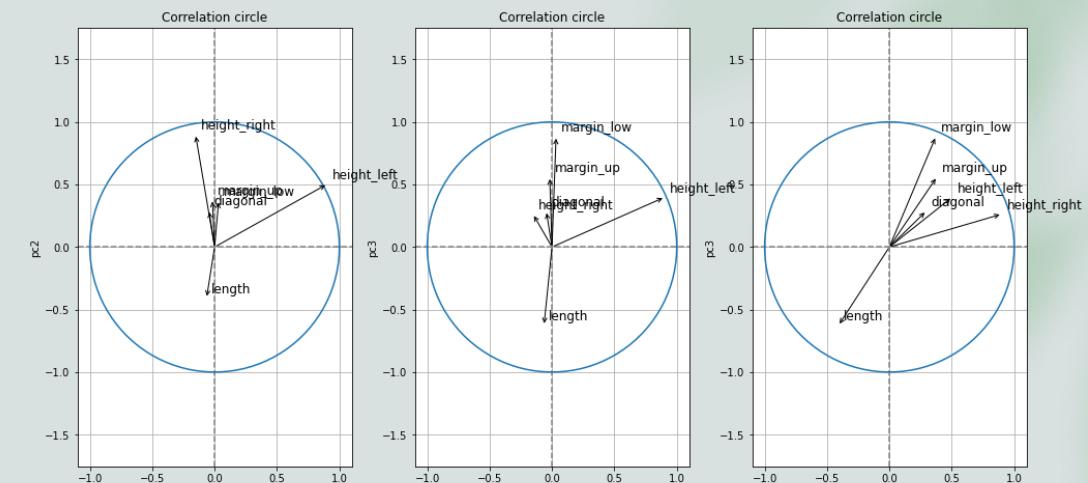
Sommaire.

- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

Représentation des variables synthétiques, « heatmap ».



- Bien que l'ACP soit inopérante dans ce cas, la représentation des variables synthétiques donne une information sur les variables discriminantes.
- Sur les axes 2 et 3, on note que les variables « margin_low », et « height_right » sont inversement correlées à « lenght ».
- C'est une information que l'on utilisera par la suite.

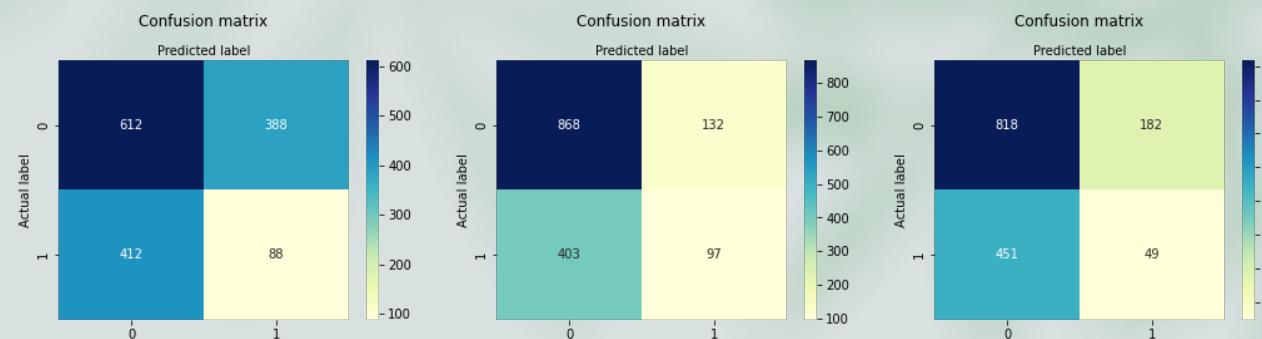
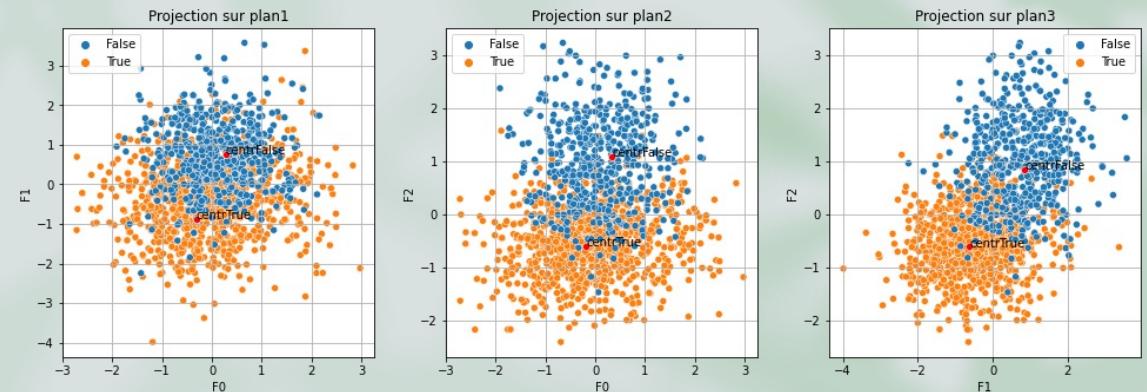


Sommaire.

- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

Etude de l'effet de la qualité des données au travers de trois cas.

- **Cas 1:** Dans un premier temps, pour se rendre compte de l'effet de la qualité des données d'entrée, nous allons appliquer un K-Means sur une partie seulement de l'information originelle disponible en utilisant les 3 plans factoriels identifiés précédemment.
- Les matrices de confusion rendent compte de résultats de détection inappropriés.
- La combinaison de K-Means et de cette qualité de données ne correspond pas aux attentes.



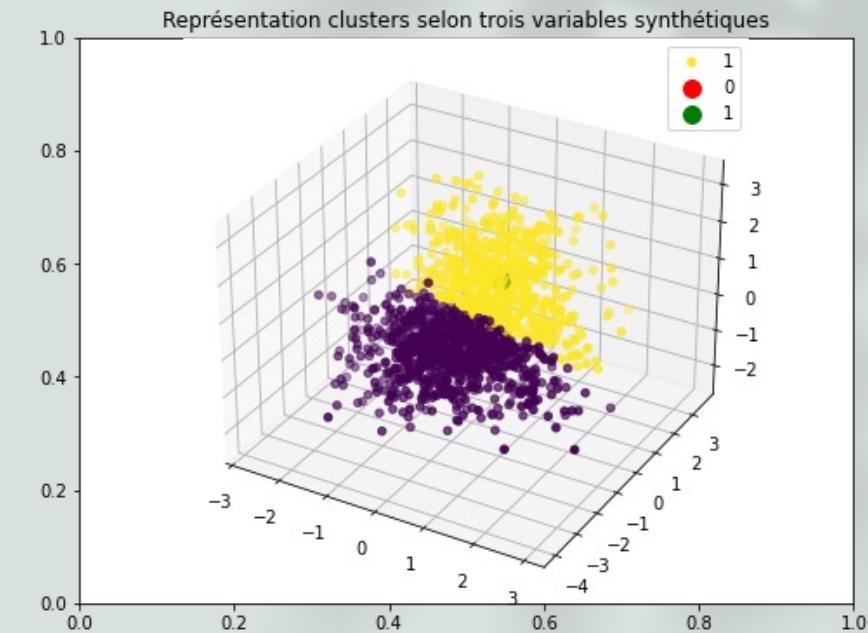
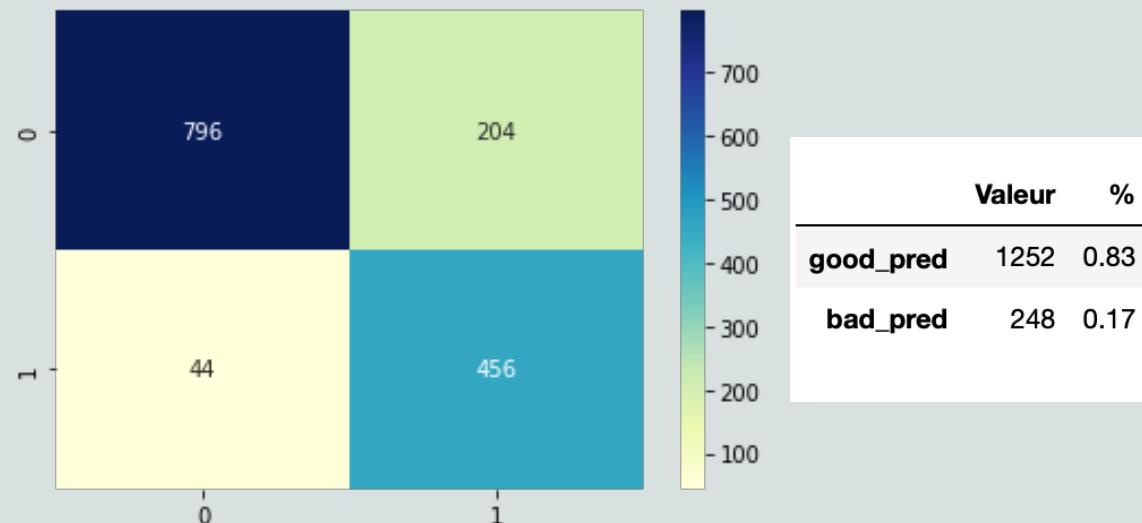
	plan1	plan2	plan3
good_pred	0.466667	0.643333	0.578
bad_pred	0.533333	0.356667	0.422

Sommaire.

- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

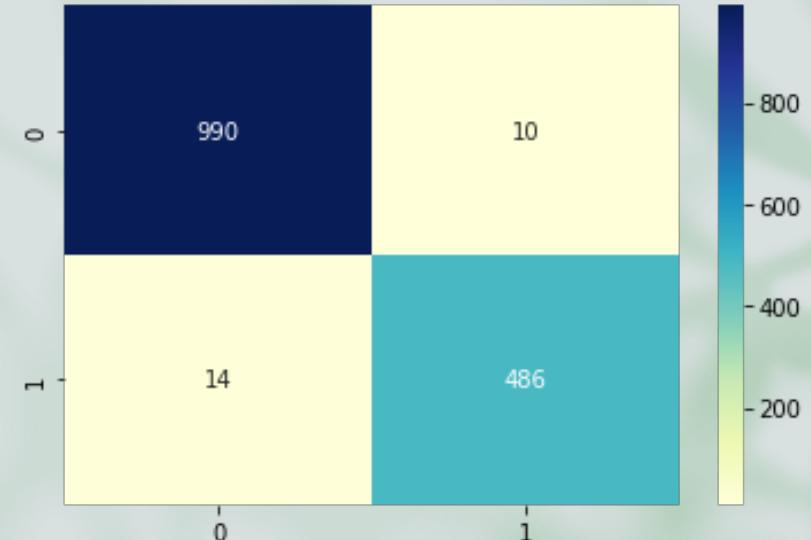
Cas 1, 2 et 3: Données d'entrée, matrices de confusion.

- Dans cette **2ème situation**, on utilise simultanément les 3 composantes factorielles lors de la projection des individus.
- On a une richesse d'information de 73%, supérieure au regard de la situation précédente.
- On note une amélioration des résultats obtenus (83% versus 64%).



Cas 1, 2 et 3: Données d'entrée, matrices de confusion (suite).

- **3ème cas:** utilisation de 100% de l'information disponible.
- Les résultats obtenus sont améliorés par rapport au cas précédent (98% versus 83%).
- En synthèse, on note que la méthode du K-Means donne des résultats intéressants, mais non suffisants.



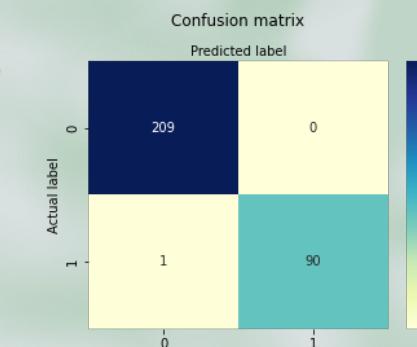
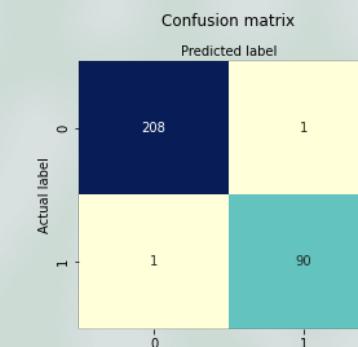
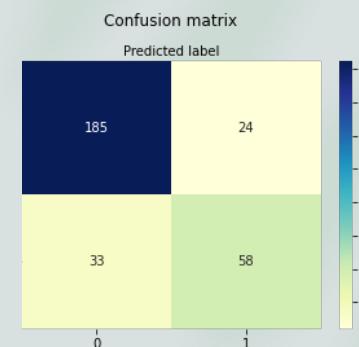
	Valeur	%_KM
good_pred	1476	0.984
bad_pred	24	0.016

Sommaire.

- Contexte de l'étude.
- Collecte des données, visualisation, nettoyage.
- Analyse descriptive, signifiance des variables.
- Méthode de l'analyse en composantes principales.
 - Détermination du nombre de composantes.
 - Visualisation des nuages de points projetés.
 - Représentation des variables synthétiques, « heatmap », synthèse.
- Méthode des K-Means.
 - Etude de l'effet de la qualité des données au travers de trois cas.
 - Cas 1, 2 et 3: Données d'entrée, matrices de confusion.
 - Synthèse.
- Méthode de régression logistique.
 - Cas 1 et 2: Données d'entrée, matrices de confusion.
 - Synthèse.
- Conclusion.

Cas 1 et 2: Données d'entrée, matrices de confusion.

- **Cas 1:** Appliquons un modèle de régression logistique, successivement, sur les variables originelles suivantes:
- (height_right et height_left), (margin_low, height_left, lenght), (margin_low, height_right, lenght).
- Pour ce modèle de régression, on a choisi de scinder les données en 80-20 (80% des données pour entraîner le modèle, 20% pour tester le modèle).
- On a un « support » de 300 (20% * 1500 lignes).
- Obtention des matrices de confusion: avec l'amélioration de la qualité des données d'entrée, on obtient une qualité des résultats de détection croissante.



Cas 1 et 2: Données d'entrée, matrices de confusion (suite).

- Les métriques, "precision", "recall" et "F1-score", montrent une amélioration sensible entre les modèles basés sur 2 variables d'entrée et 3 variables.
- Entre la 2ème et la 3ème matrice, l'erreur de prédiction, faux billet alors qu'il est vrai est corrigée.
- Subsistait une erreur de prédiction: vrai billet alors qu'il est faux.

	precision	recall	f1-score	support
0	0.85	0.89	0.87	209
1	0.71	0.64	0.67	91
accuracy			0.81	300
macro avg	0.78	0.76	0.77	300
weighted avg	0.81	0.81	0.81	300

Coefficients du modèle linéaire: [[1.21881723 0.91208126]]

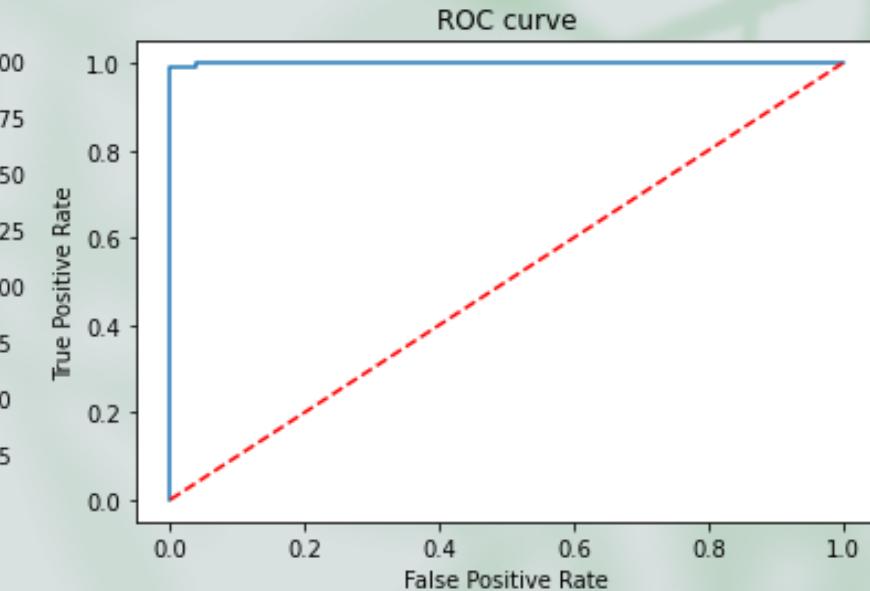
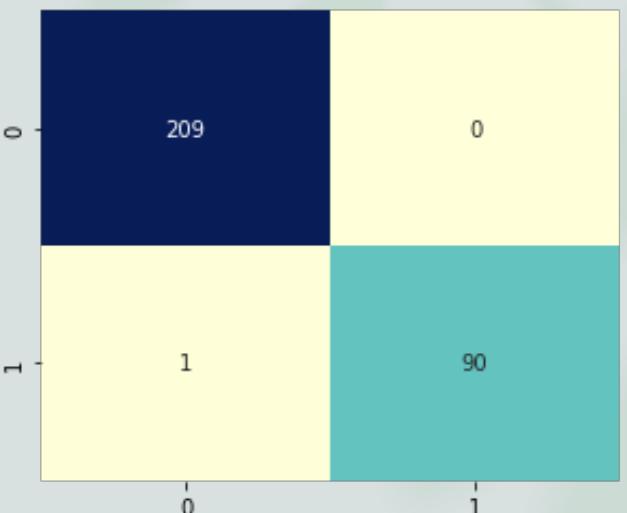
	precision	recall	f1-score	support
0	1.00	1.00	1.00	209
1	0.99	0.99	0.99	91
accuracy			0.99	300
macro avg	0.99	0.99	0.99	300
weighted avg	0.99	0.99	0.99	300

Coefficients du modèle linéaire: [[2.83245543 0.6535214 -4.07399049]]

	precision	recall	f1-score	support
0	1.00	1.00	1.00	209
1	1.00	0.99	0.99	91
accuracy			1.00	300
macro avg	1.00	0.99	1.00	300
weighted avg	1.00	1.00	1.00	300

Cas 1 et 2: Données d'entrée, matrices de confusion (suite).

- **Cas 2:** Appliquons maintenant le modèle de régression logistique sur 100% des variables originelles.
- Obtention de la matrice de confusion.
- Proche des résultats obtenus avec la méthode de RL avec trois variables (margin_low, height_right, lenght).
- La courbe ROC indique que le modèle est performant à classer les deux groupes distinctement.



	precision	recall	f1-score	support
0	1.00	1.00	1.00	209
1	1.00	0.99	0.99	91
accuracy			1.00	300
macro avg	1.00	0.99	1.00	300
weighted avg	1.00	1.00	1.00	300

Coefficients du modèle linéaire: [[-0.19588424 0.50098615 0.56057366 2.63139143 1.7606656 -3.49235102]]

Constante du modèle linéaire: [-1.97366763]

Cas 1 et 2: Données d'entrée, matrices de confusion.

	%_RL	%_KM
good_pred	0.997	0.984
bad_pred	0.003	0.016

- Nous avons mené les analyses avec des données d'entrée de qualité plus ou moins importante et des algorithmes différents.
- En synthèse, lorsque l'on calcule le ratio des prédictions réussies et ratées, pour les algorithmes de régression linéaire et de K-Means (avec les données d'entrées correspondant aux variables originelles), le tableau ci-contre indique que pour ce cas précis, c'est l'algorithme de régression logistique qui est le plus performant.

FIN