
*Projet 6 — 24 mai 2021*

# Détection de faux billets

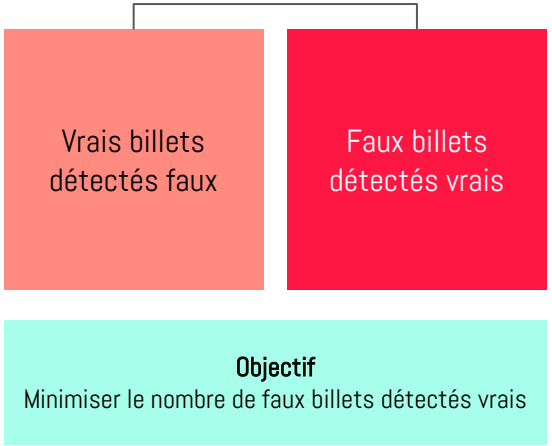
Algorithme de classification avec régression logistique

➔ 17 faux billets d'euros pour 1 million d'authentiques
- Dans la zone euro
- Niveau historiquement bas, en baisse chaque année
- 80% des contrefaçons sont des coupures de 20 et 50€
- 97% saisis dans la zone euro

➔ 30 à 40% émis depuis la France

➔ Caractéristiques discriminantes
- Impression sur coton pur
- Impressions en relief
- Mesures géométriques (adapté au machine learning)

2 types d'erreur de classification




Sources :
National Geographic
Banque centrale européenne

# Sommaire

# Exploration

01

## 170 billets, 7 variables

| | is_genuine | diagonal | height_left | height_right | margin_low | margin_up | length |
|---|---|---|---|---|---|---|---|
| 0 | True | 171.81 | 104.86 | 104.95 | 4.52 | 2.89 | 112.83 |

| | diagonal | height_left | height_right | margin_low | margin_up | length |
|---|---|---|---|---|---|---|
| count | 170.000000 | 170.000000 | 170.000000 | 170.000000 | 170.000000 | 170.000000 |
| mean | 171.940588 | 104.066353 | 103.928118 | 4.612118 | 3.170412 | 112.570412 |
| std | 0.305768 | 0.298185 | 0.330980 | 0.702103 | 0.236361 | 0.924448 |
| min | 171.040000 | 103.230000 | 103.140000 | 3.540000 | 2.270000 | 109.970000 |
| 25% | 171.730000 | 103.842500 | 103.690000 | 4.050000 | 3.012500 | 111.855000 |
| 50% | 171.945000 | 104.055000 | 103.950000 | 4.450000 | 3.170000 | 112.845000 |
| 75% | 172.137500 | 104.287500 | 104.170000 | 5.127500 | 3.330000 | 113.287500 |
| max | 173.010000 | 104.860000 | 104.950000 | 6.280000 | 3.680000 | 113.980000 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 169 | False | 171.96 | 104.00 | 103.95 | 5.63 | 3.26 | 110.96 |

170 rows × 7 columns

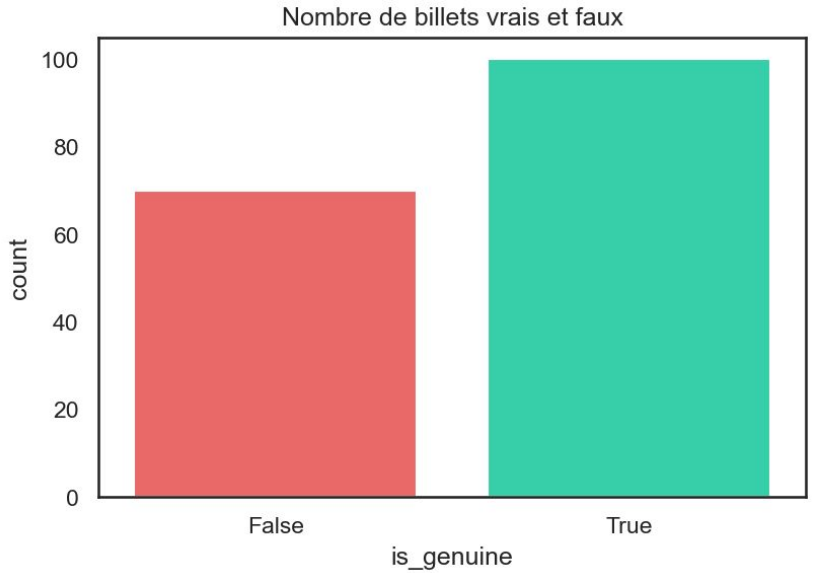
1 variable target (**vrai** ou **faux**)

6 mesures en millimètres

- ➔ Diagonale
- ➔ Hauteur à gauche
- ➔ Hauteur à droite
- ➔ Marge basse
- ➔ Marge haute
- ➔ Longueur

Aucun nettoyage requis

- ➔ Aucune valeur manquante
- ➔ Aucune valeur aberrante (outlier)
- ➔ Aucun doublon

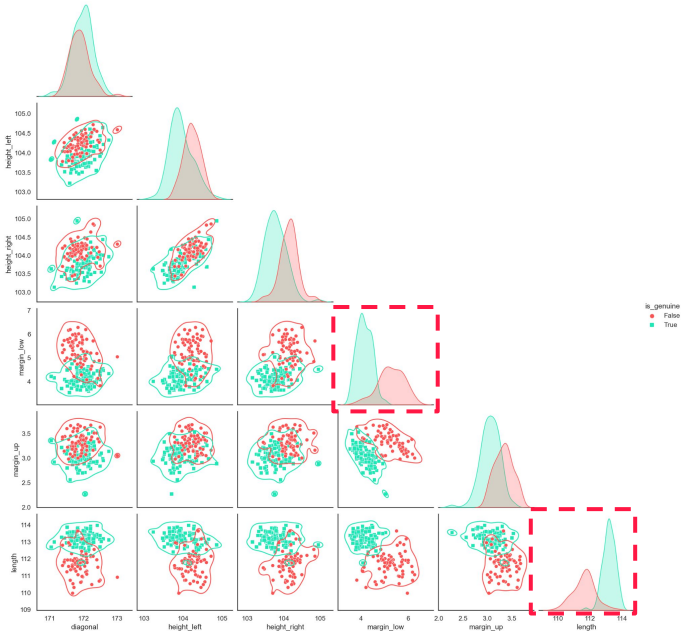



100 vrais billets — 70 contrefaçons

Oversampling inutile

# Distributions par authenticité





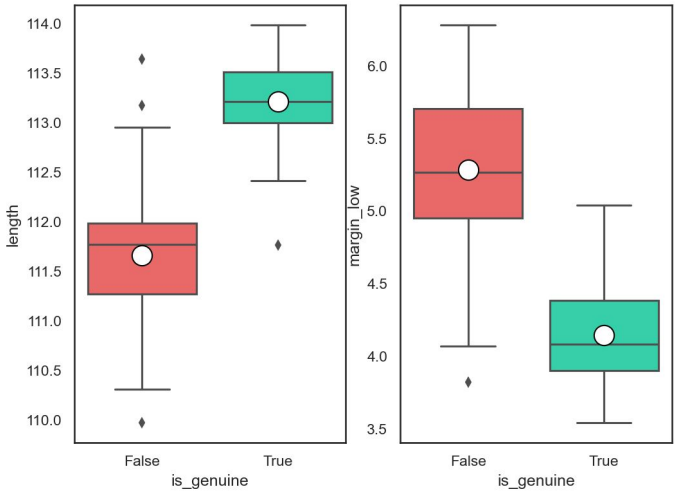

Distribution de chaque variable, par authenticité

➔ Fort impact de 2 variables sur l'authenticité

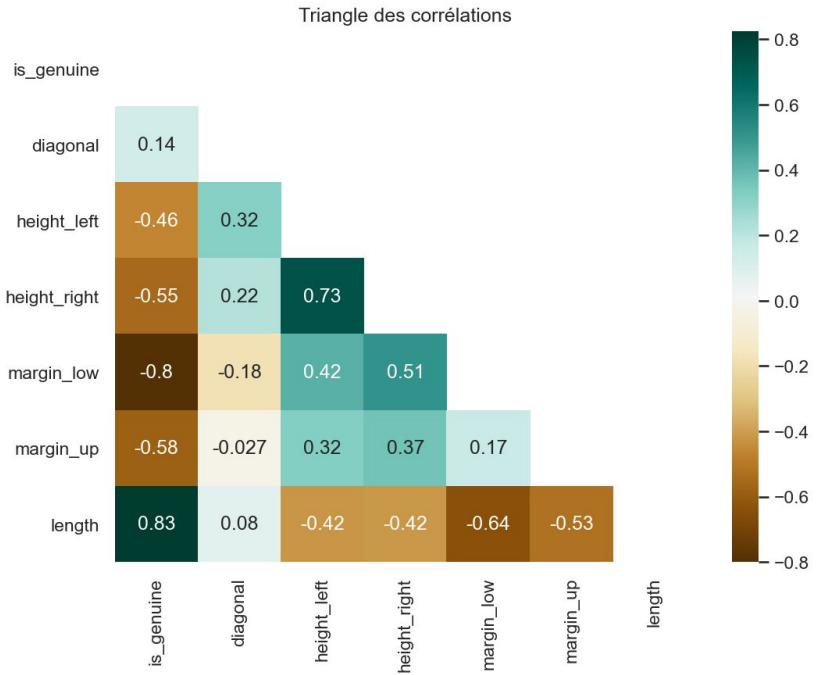
- longueur
- marge basse

➔ Les autres variables ont peu d'impact

- ➔ Vrais billets plus longs que les faux
- ➔ Marge basse plus courte chez les faux billets




Distributions de *length* et *margin_low*, par authenticité




## Matrice des corrélations

Résume les dépendances entre variables

- 0 pas de corrélation
- 1 corrélation positive
- -1 corrélation négative

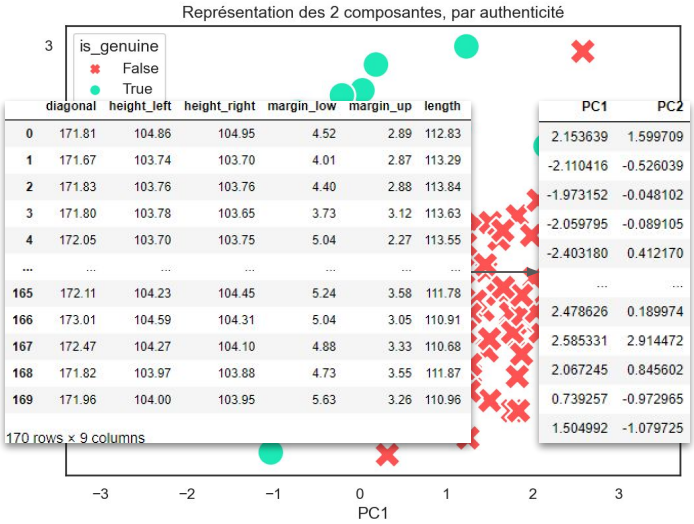
➔ Confirmation des fortes corrélations

➔ Toutes les variables apportent de l'information

# ACP

02

# Objectifs de l'ACP




| | diagonal | height_left | height_right | margin_low | margin_up | length |
|---|---|---|---|---|---|---|
| 0 | 171.81 | 104.86 | 104.95 | 4.52 | 2.89 | 112.83 |
| 1 | 171.67 | 103.74 | 103.70 | 4.01 | 2.87 | 113.29 |
| 2 | 171.83 | 103.76 | 103.76 | 4.40 | 2.88 | 113.84 |
| 3 | 171.80 | 103.78 | 103.65 | 3.73 | 3.12 | 113.63 |
| 4 | 172.05 | 103.70 | 103.75 | 5.04 | 2.27 | 113.55 |
| ... | ... | ... | ... | ... | ... | ... |
| 165 | 172.11 | 104.23 | 104.45 | 5.24 | 3.58 | 111.78 |
| 166 | 173.01 | 104.59 | 104.31 | 5.04 | 3.05 | 110.91 |
| 167 | 172.47 | 104.27 | 104.10 | 4.88 | 3.33 | 110.68 |
| 168 | 171.82 | 103.97 | 103.88 | 4.73 | 3.55 | 111.87 |
| 169 | 171.96 | 104.00 | 103.95 | 5.63 | 3.26 | 110.96 |

170 rows × 9 columns

| | PC1 | PC2 |
|---|---|---|
| | 2.153639 | 1.599709 |
| | -2.110416 | -0.526039 |
| | -1.973152 | -0.048102 |
| | -2.059795 | -0.089105 |
| | -2.403180 | 0.412170 |
| | ... | ... |
| | 2.478626 | 0.189974 |
| | 2.585331 | 2.914472 |
| | 2.067245 | 0.845602 |
| | 0.739257 | -0.972965 |
| | 1.504992 | -1.079725 |

➔ Réduction des variables à ***n*** dimensions

➔ Transformation linéaire

Préserve les rapports de colinéarité

➔ Perte d'informations

Valeurs obtenues non interprétables

➔ Valeurs projetables sur un plan à 2 ou 3 dimensions

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

➔ Sensible à la variance

➔ Différences d'échelle entre variables

Poids trop fort de certaines variables sur d'autres
Contributions inégales à l'ACP

➔ Objectif : harmoniser les **écarts type**

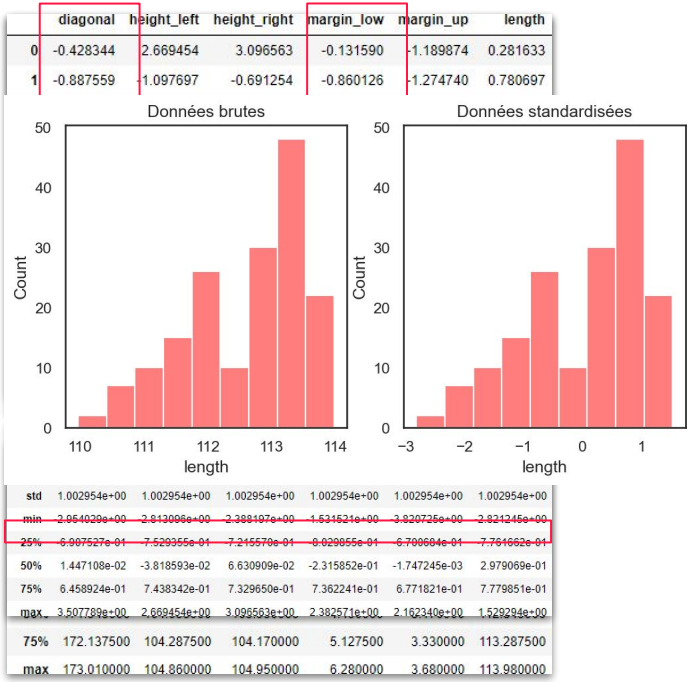
Racine carrée de la variance (moyenne des écarts à la moyenne)

➔ Standardisation

- Moyenne = 0 (centrage autour de cette valeur)
- Soustraction de la moyenne à toutes les observations
- Écart-type = 1
- Conserve la forme de la distribution

➔ Adapté aux algorithmes linéaires

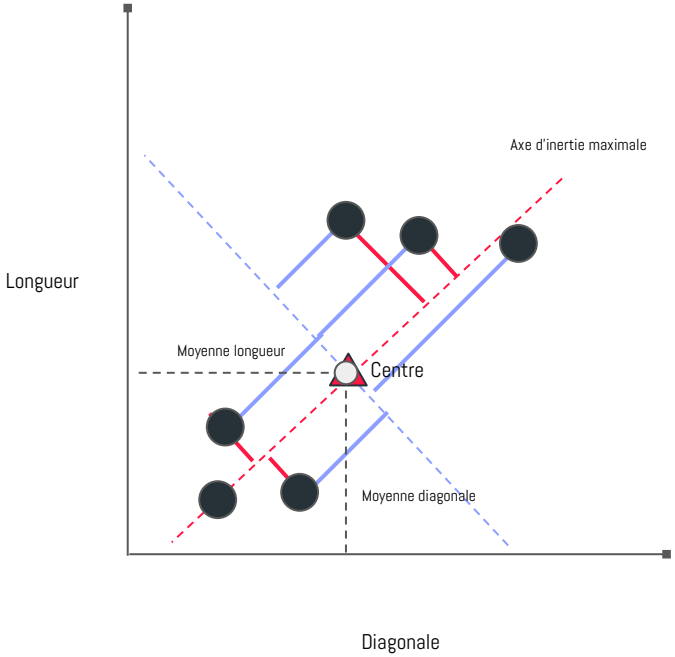
ACP, régression logistique …



| | diagonal | height_left | height_right | margin_low | margin_up | length |
|---|---|---|---|---|---|---|
| 0 | -0.428344 | 2.669454 | 3.096563 | -0.131590 | -1.189874 | 0.281633 |
| 1 | -0.887559 | -1.097697 | -0.691254 | -0.860126 | -1.274740 | 0.780697 |



| | | | | | | |
|---|---|---|---|---|---|---|
| std | 1.002954e+00 | 1.002954e+00 | 1.002954e+00 | 1.002954e+00 | 1.002954e+00 | 1.002954e+00 |
| min | -2.954029e+00 | -2.813096e+00 | -2.388197e+00 | -1.531521e+00 | -3.820725e+00 | -2.821245e+00 |
| 25% | -6.907527e-01 | -7.529355e-01 | -7.215570e-01 | -8.029855e-01 | -6.700684e-01 | -7.761662e-01 |
| 50% | 1.447108e-02 | -3.818593e-02 | 6.630909e-02 | -2.315852e-01 | -1.747245e-03 | 2.979069e-01 |
| 75% | 6.458924e-01 | 7.438342e-01 | 7.329650e-01 | 7.362241e-01 | 6.771821e-01 | 7.779851e-01 |
| max | 3.507789e+00 | 2.669454e+00 | 3.096563e+00 | 2.382571e+00 | 2.162340e+00 | 1.529294e+00 |
| 75% | 172.137500 | 104.287500 | 104.170000 | 5.127500 | 3.330000 | 113.287500 |
| max | 173.010000 | 104.860000 | 104.950000 | 6.280000 | 3.680000 | 113.980000 |




➔ **Objectif :** trouver l'axe exprimant le maximum de variance

- En partant du centroïde
- Conservation du maximum d'information
  (= variance : moyenne des distances au carré l'axe)

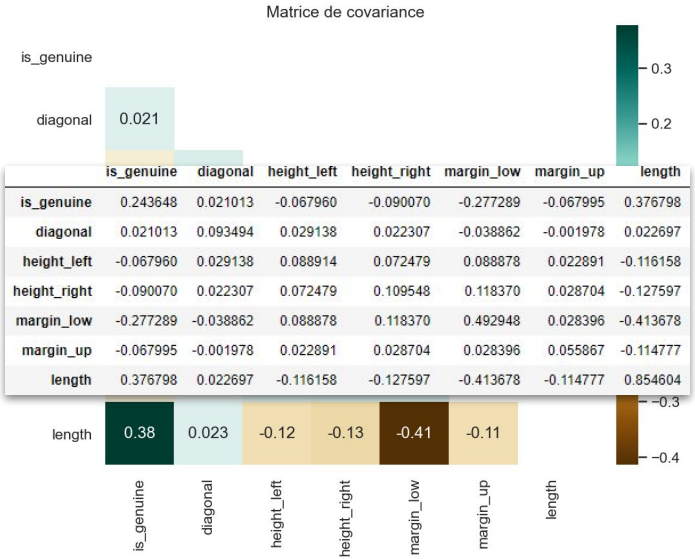
➔ Axes orthogonaux aux précédents

Composantes décorrélées

➔ Autant d'axes (composantes) que de variables

# Covariance






| | is_genuine | diagonal | height_left | height_right | margin_low | margin_up | length |
|---|---|---|---|---|---|---|---|
| **is_genuine** | 0.243648 | 0.021013 | -0.067960 | -0.090070 | -0.277289 | -0.067995 | 0.376798 |
| **diagonal** | 0.021013 | 0.093494 | 0.029138 | 0.022307 | -0.038862 | -0.001978 | 0.022697 |
| **height_left** | -0.067960 | 0.029138 | 0.088914 | 0.072479 | 0.088878 | 0.022891 | -0.116158 |
| **height_right** | -0.090070 | 0.022307 | 0.072479 | 0.109548 | 0.118370 | 0.028704 | -0.127597 |
| **margin_low** | -0.277289 | -0.038862 | 0.088878 | 0.118370 | 0.492948 | 0.028396 | -0.413678 |
| **margin_up** | -0.067995 | -0.001978 | 0.022891 | 0.028704 | 0.028396 | 0.055867 | -0.114777 |
| **length** | 0.376798 | 0.022697 | -0.116158 | -0.127597 | -0.413678 | -0.114777 | 0.854604 |

**Axe 7** = 0.37*is_genuine + 0.02*diagonal - 0.11*height_left - 0.12*height_right - 0.41*margin_low - 0.11*margin_up + 0.85*length

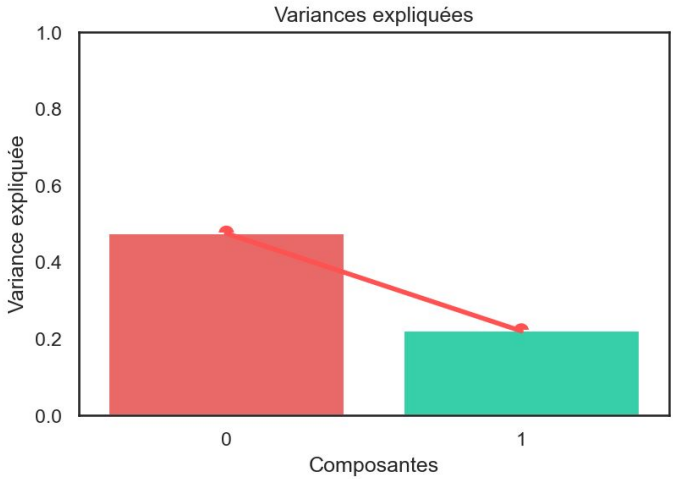
➔ 7 axes (vecteurs propres)

Combinaisons linéaires de chaque variables avec les autres

➔ Classement par somme des valeurs propres

La 1ère composante est celle cumulant le plus de variance

# Analyse des éboulis




Analyse des éboulis des 2 premières composantes

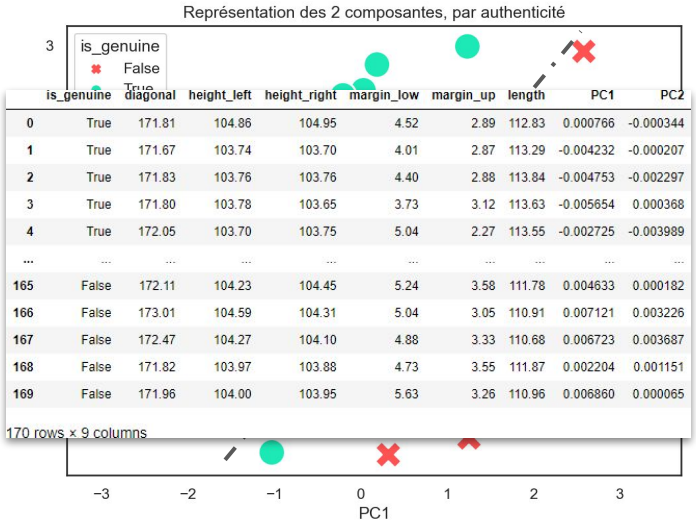
➔ Le maximum d'informations restantes conservé à chaque itération

➔ Essentiel de l'information stocké dans les 2 premières composantes

Variance expliquée : **69%**

- ➔ PC1 : 47% des données
- ➔ PC2 : 22% des données

- ➔ Jointure des 2 composantes avec les données
- ➔ Les données sont linéaires
  Les vrais et faux billets forment 2 groupes distincts

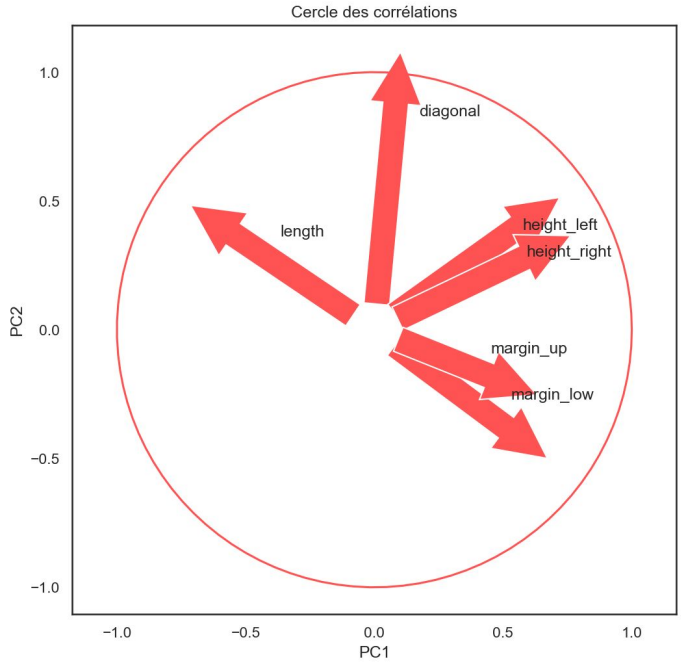



| | is_genuine | diagonal | height_left | height_right | margin_low | margin_up | length | PC1 | PC2 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | True | 171.81 | 104.86 | 104.95 | 4.52 | 2.89 | 112.83 | 0.000766 | -0.000344 |
| 1 | True | 171.67 | 103.74 | 103.70 | 4.01 | 2.87 | 113.29 | -0.004232 | -0.000207 |
| 2 | True | 171.83 | 103.76 | 103.76 | 4.40 | 2.88 | 113.84 | -0.004753 | -0.002297 |
| 3 | True | 171.80 | 103.78 | 103.65 | 3.73 | 3.12 | 113.63 | -0.005654 | 0.000368 |
| 4 | True | 172.05 | 103.70 | 103.75 | 5.04 | 2.27 | 113.55 | -0.002725 | -0.003989 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 165 | False | 172.11 | 104.23 | 104.45 | 5.24 | 3.58 | 111.78 | 0.004633 | 0.000182 |
| 166 | False | 173.01 | 104.59 | 104.31 | 5.04 | 3.05 | 110.91 | 0.007121 | 0.003226 |
| 167 | False | 172.47 | 104.27 | 104.10 | 4.88 | 3.33 | 110.68 | 0.006723 | 0.003687 |
| 168 | False | 171.82 | 103.97 | 103.88 | 4.73 | 3.55 | 111.87 | 0.002204 | 0.001151 |
| 169 | False | 171.96 | 104.00 | 103.95 | 5.63 | 3.26 | 110.96 | 0.006860 | 0.000065 |

170 rows × 9 columns

# Cercle des corrélations






**Projection des variables sur le plan factoriel**

Cosinus de l'angle entre 2 flèches = coefficient de corrélation entre les 2 variables

- 0° : corrélation positive
- 90° : absence de corrélation
- 180° : corrélation négative

➔ Corrélations préservées

➔ Bonne inertie
Longueurs des flèches homogènes et proches de 1

➔ Représentation fiable sur les 2 premières dimensions

# Clustering

03

➔ Regrouper les billets en ***K*** groupes homogènes

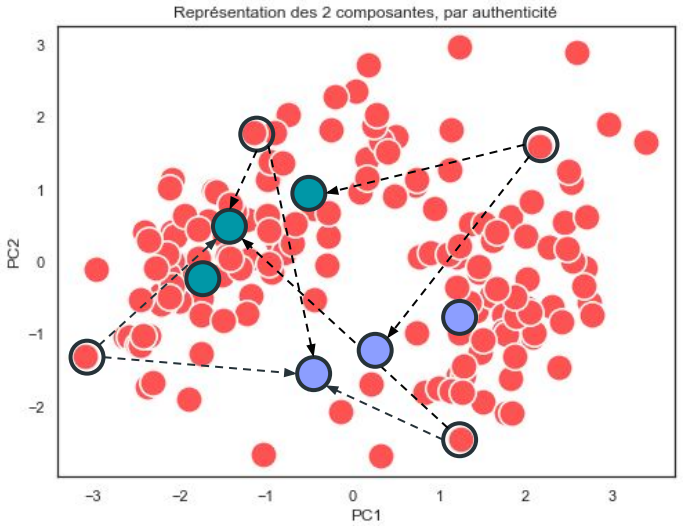
➔ Classification non supervisée

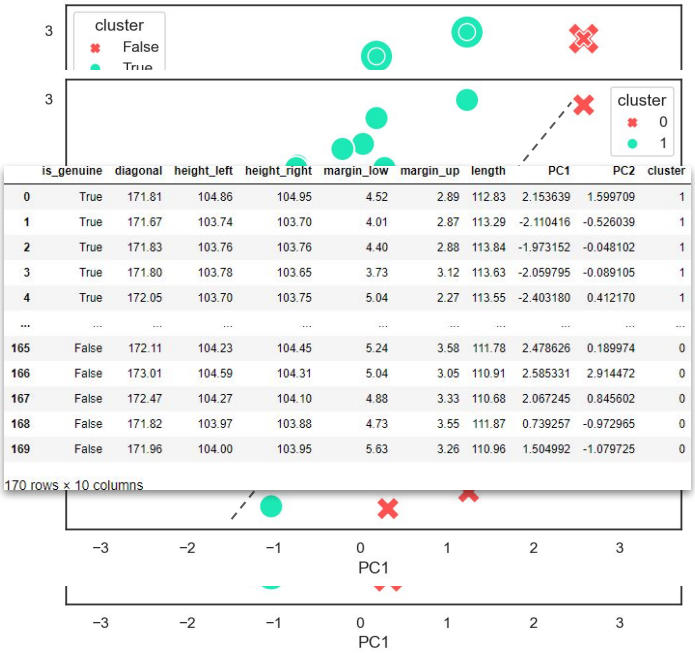
Postulat : on ne connaît pas les groupes auxquels appartiennent les billets

➔ Classification non hiérarchique

**Méthode des k-moyennes**

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





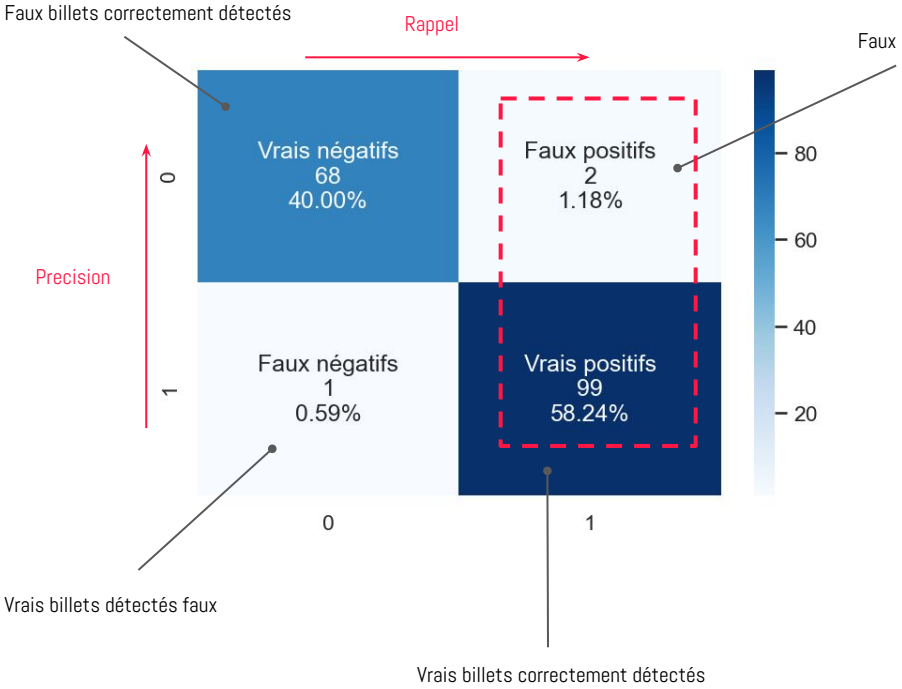

| | is_genuine | diagonal | height_left | height_right | margin_low | margin_up | length | PC1 | PC2 | cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | True | 171.81 | 104.86 | 104.95 | 4.52 | 2.89 | 112.83 | 2.153639 | 1.599709 | 1 |
| 1 | True | 171.67 | 103.74 | 103.70 | 4.01 | 2.87 | 113.29 | -2.110416 | -0.526039 | 1 |
| 2 | True | 171.83 | 103.76 | 103.76 | 4.40 | 2.88 | 113.84 | -1.973152 | -0.048102 | 1 |
| 3 | True | 171.80 | 103.78 | 103.65 | 3.73 | 3.12 | 113.63 | -2.059795 | -0.089105 | 1 |
| 4 | True | 172.05 | 103.70 | 103.75 | 5.04 | 2.27 | 113.55 | -2.403180 | 0.412170 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 165 | False | 172.11 | 104.23 | 104.45 | 5.24 | 3.58 | 111.78 | 2.478626 | 0.189974 | 0 |
| 166 | False | 173.01 | 104.59 | 104.31 | 5.04 | 3.05 | 110.91 | 2.585331 | 2.914472 | 0 |
| 167 | False | 172.47 | 104.27 | 104.10 | 4.88 | 3.33 | 110.68 | 2.067245 | 0.845602 | 0 |
| 168 | False | 171.82 | 103.97 | 103.88 | 4.73 | 3.55 | 111.87 | 0.739257 | -0.972965 | 0 |
| 169 | False | 171.96 | 104.00 | 103.95 | 5.63 | 3.26 | 110.96 | 1.504992 | -1.079725 | 0 |

170 rows x 10 columns

- ➔ Jointure des clusters
- ➔ 2 clusters distincts
  Données linéaires
- ➔ Clusters presque identiques aux classes réelles
- ➔ 3 erreurs de classification

# Matrice de confusion






| | precision | recall | f1-score | support |
|---|---|---|---|---|
| False | 0.99 | 0.97 | 0.98 | 70 |
| True | 0.98 | 0.99 | 0.99 | 100 |
| accuracy | | | 0.98 | 170 |
| macro avg | 0.98 | 0.98 | 0.98 | 170 |
| weighted avg | 0.98 | 0.98 | 0.98 | 170 |

➔ **Precision** : à optimiser pour réduire les faux positifs
billets correctement prédits / billets prédits dans la classe

➔ **Recall**
billets correctement prédits / nbre de billets dans la classe d'origine

➔ **F-score**
Moyenne harmonique de la précision et du rappel

# Modélisation

04

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

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

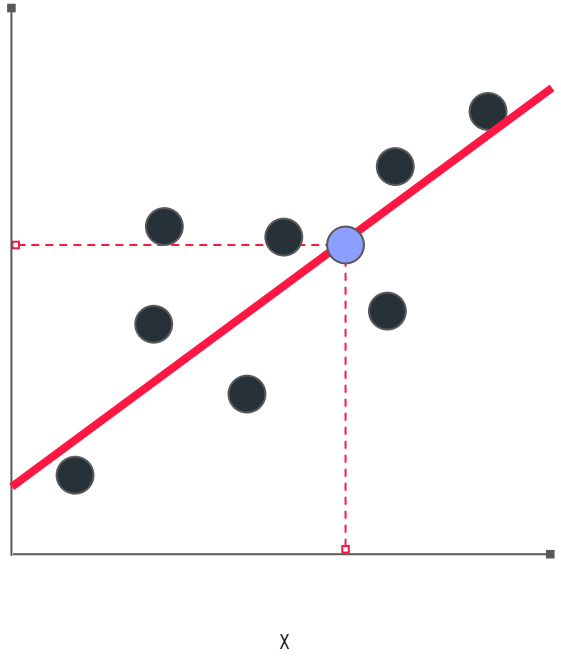
➔ Apprentissage supervisé

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

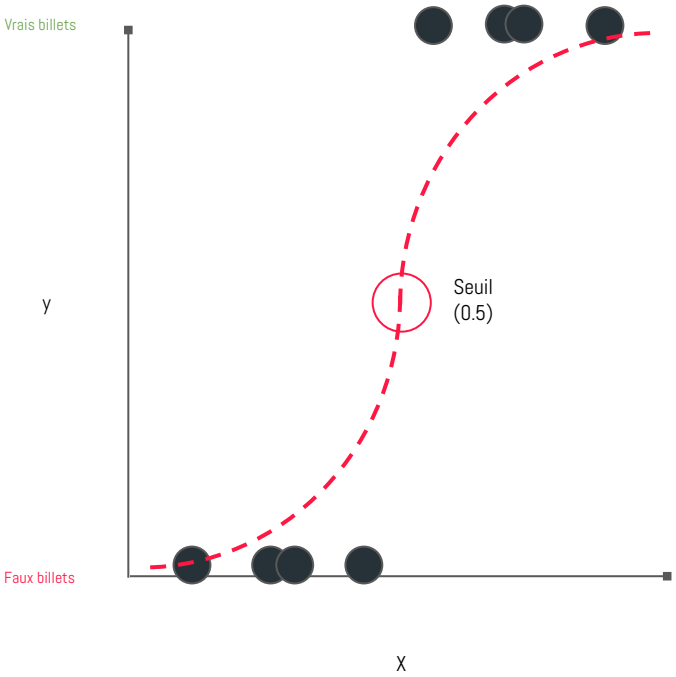
➔ Postulat : on peut aligner les points sur une droite

Relation linéaire décrivant le mieux la relation entre X et y
= f(**X**) = pente x **X** + 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 = logit (logarithme) 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

## train

| | diagonal | height_left | height_right | margin_low | margin_up | length |
|---|---|---|---|---|---|---|
| **27** | 172.02 | 104.23 | 104.26 | 4.92 | 2.89 | 113.49 |
| **78** | 172.16 | 104.39 | 103.85 | 3.77 | 3.32 | 112.55 |
| **147** | 172.25 | 104.52 | 104.22 | 4.65 | 3.43 | 110.48 |
| **38** | 172.21 | 104.27 | 104.01 | 4.23 | 2.79 | 113.78 |
| **41** | 171.81 | 104.10 | 103.69 | 4.29 | 2.95 | 112.72 |
| **...** | ... | ... | ... | ... | ... | ... |
| **71** | 172.17 | 103.93 | 103.62 | 4.06 | 3.08 | 113.10 |
| **106** | 172.22 | 104.17 | 104.07 | 4.52 | 3.67 | 112.13 |
| **14** | 172.04 | 103.94 | 103.76 | 3.81 | 3.24 | 113.41 |
| **92** | 171.86 | 103.47 | 103.59 | 4.04 | 2.97 | 113.22 |
| **102** | 171.94 | 104.21 | 104.10 | 4.28 | 3.47 | 112.23 |

113 rows × 6 columns

```
27      True
78      True
147    False
38      True
41      True
       ...
71      True
106    False
14      True
92      True
102    False
Name: is_genuine, Length: 113, dtype: bool
```

X

y

## test

| | diagonal | height_left | height_right | margin_low | margin_up | length |
|---|---|---|---|---|---|---|
| **139** | 171.60 | 104.37 | 104.20 | 5.82 | 3.08 | 112.84 |
| **30** | 172.19 | 104.05 | 103.81 | 3.90 | 3.22 | 113.52 |
| **119** | 171.51 | 104.13 | 103.90 | 4.99 | 3.60 | 111.23 |
| **29** | 171.84 | 103.75 | 103.38 | 4.08 | 2.70 | 113.72 |
| **144** | 171.56 | 103.80 | 103.87 | 5.66 | 2.98 | 112.95 |
| **163** | 171.78 | 104.07 | 104.16 | 5.77 | 3.30 | 111.27 |
| **166** | 173.01 | 104.59 | 104.31 | 5.04 | 3.05 | 110.91 |
| **51** | 172.22 | 104.48 | 104.06 | 4.59 | 2.91 | 112.82 |
| **105** | 171.99 | 104.18 | 104.20 | 5.26 | 3.23 | 111.83 |
| **60** | 172.11 | 103.67 | 103.43 | 4.19 | 2.98 | 113.09 |

```
139    False
30      True
119    False
29      True
144    False
163    False
166    False
51      True
105    False
60      True
Name: is_genuine, dtype: bool
```

➔ Séparation du dataset en 2 jeux

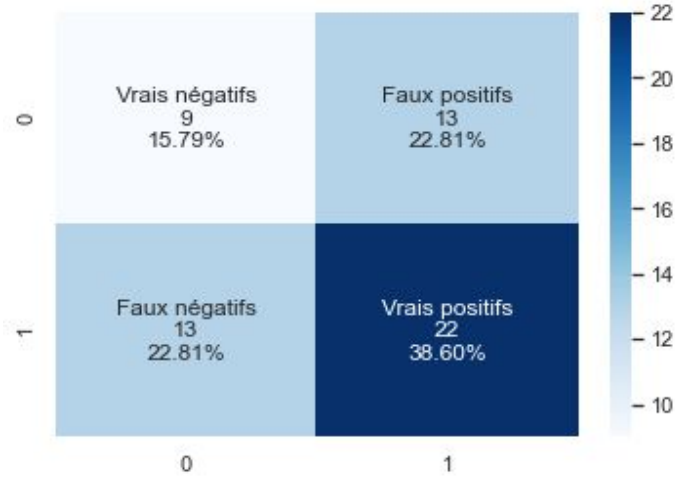
- **train** : 113 billets
- **test** : 57 billets

➔ Séparation de chaque jeu entre X et y

- **X** = explicatives (mesures)
- **y** = expliquée (is_genuine)

# Classifieur idiot






| | precision | recall | f1-score | support |
|---|---|---|---|---|
| False | 0.41 | 0.41 | 0.41 | 22 |
| True | 0.63 | 0.63 | 0.63 | 35 |
| accuracy | | | 0.54 | 57 |
| macro avg | 0.52 | 0.52 | 0.52 | 57 |
| weighted avg | 0.54 | 0.54 | 0.54 | 57 |

➔ Base de comparaison avec le futur modèle

➔ Prédictions aléatoires

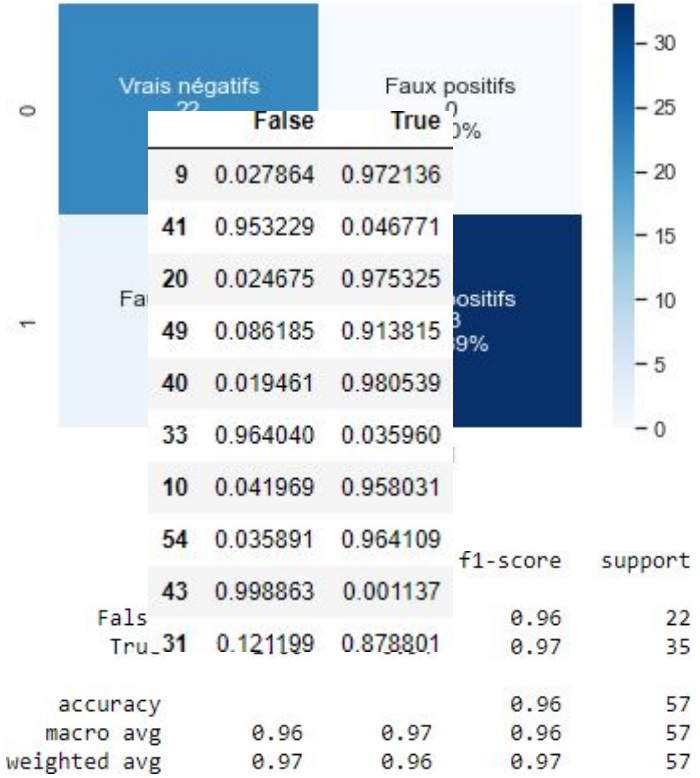
Donc médiocres

➔ Précision très faible

63 % de vrais billets correctement prédits / vrais billets prédits

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

Precision

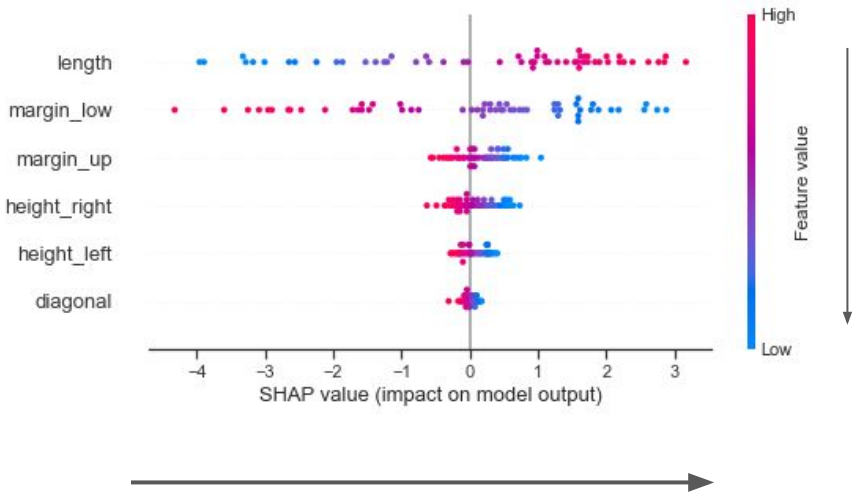



| | False | True |
|---|---|---|
| 9 | 0.027864 | 0.972136 |
| 41 | 0.953229 | 0.046771 |
| 20 | 0.024675 | 0.975325 |
| 49 | 0.086185 | 0.913815 |
| 40 | 0.019461 | 0.980539 |
| 33 | 0.964040 | 0.035960 |
| 10 | 0.041969 | 0.958031 |
| 54 | 0.035891 | 0.964109 |
| 43 | 0.998863 | 0.001137 |
| 31 | 0.121199 | 0.878801 |

| | | | f1-score | support |
|---|---|---|---|---|
| Fals | | | 0.96 | 22 |
| Tru | | | 0.97 | 35 |
| accuracy | | | 0.96 | 57 |
| macro avg | 0.96 | 0.97 | 0.96 | 57 |
| weighted avg | 0.97 | 0.96 | 0.97 | 57 |

# Contributions au modèle

Explication de la sortie du modèle




Classement par force de contribution au modèle

Force de contribution de la variable

Rouge : valeurs les + fortes

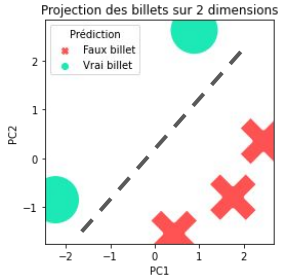
Bleu : valeurs les + faibles

- ➔ Fonction autonome
- ➔ Modèles enregistrées dans un fichier Pickle
  StandardScaler
  ACP
  Régression logistique
- ➔ Fichier CSV en entrée
- ➔ Résultats en tableau
- ➔ Contrôle des clusters sur 2 dimensions

Données en sortie

| | Prédiction | Probabilité de faux | Probabilité de vrai | id |
|---|---|---|---|---|
| 0 | Faux billet | 0.962899 | 0.037101 | A_1 |
| 1 | Faux billet | 0.994102 | 0.005898 | A_2 |
| 2 | Faux billet | 0.986890 | 0.013110 | A_3 |
| 3 | Vrai billet | 0.058722 | 0.941278 | A_4 |
| 4 | Vrai billet | 0.004059 | 0.995941 | A_5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | 172.49 | 104.55 | 104.34 | 4.44 | 3.03 | 113.20 | A_4 |
| 4 | 171.65 | 103.63 | 103.56 | 3.77 | 3.16 | 113.33 | A_5 |




# Résumé

- ➔ Données d'entraînement adaptées à la régression logistique
  - Variables à la même échelle
  - Individus séparables
  - Pas d'outliers
- ➔ Modèle à 100% de précision sur les vrais billets
  - Aucun faux billet classé comme vrai
  - 2 vrais billets détectés faux