
ONCFM

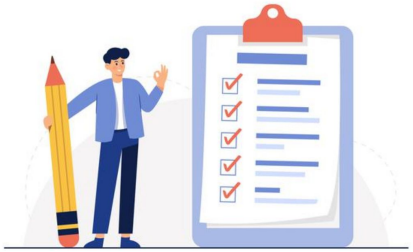
Contexte

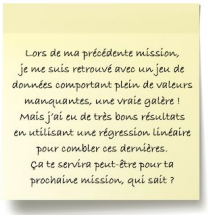
- Data Analyst dans une entreprise spécialisée dans la data

- Prestation en régie au sein de l'ONCFM

- Identification des contrefaçons des billets en euros

- Lecture cahier des charge + post-it





Déroulement du projet

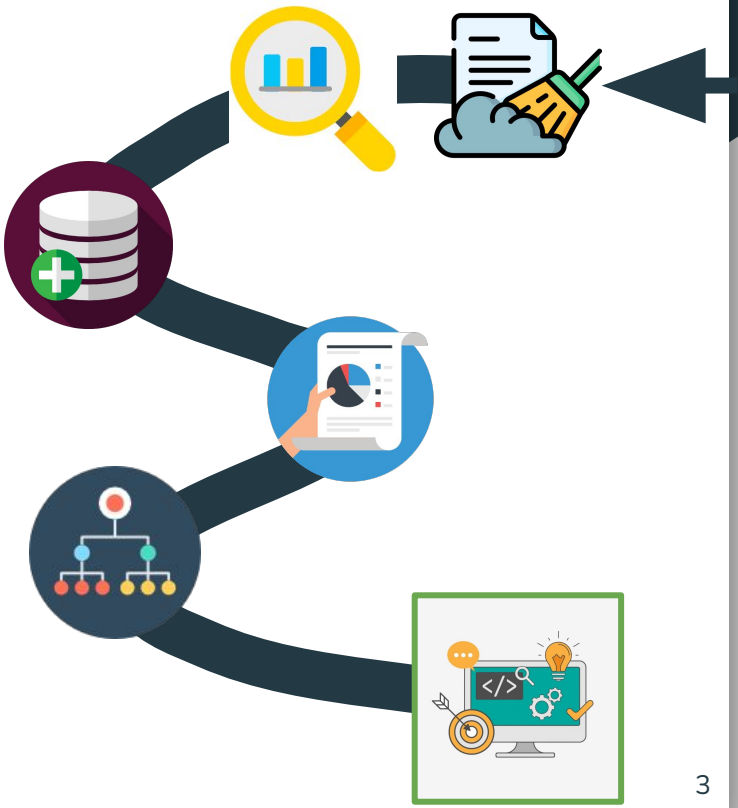
1 - Exploration et nettoyage des données

2 - Enrichissement des données
- Régression linéaire multiple

3 - Analyse descriptive des données
- ACP

4 - Classification supervisée
- K-Means
- Régression logistique

5 - Programme de détection des faux billets



Outils

Langage de programmation :

Logiciel :

Exploration et nettoyage des données

Exploration et nettoyage des données

Data

	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
1	True	171.46	103.36	103.66	3.77	2.99	113.09
2	True	172.69	104.48	103.50	4.40	2.94	113.16
3	True	171.36	103.91	103.94	3.62	3.01	113.51
4	True	171.73	104.28	103.46	4.04	3.48	112.54
...	...	...	...	...	...	...	...
1495	False	171.75	104.38	104.17	4.42	3.09	111.28
1496	False	172.19	104.63	104.44	5.27	3.37	110.97
1497	False	171.80	104.01	104.12	5.51	3.36	111.95
1498	False	172.06	104.28	104.06	5.17	3.46	112.25
1499	False	171.47	104.15	103.82	4.63	3.37	112.07

Data.drop_dublicates()

```
# Suppression des doublons

print('Nombre de doublons supprimer :',len(df) - len(df.drop_duplicates()))
```

```
Nombre de doublons supprimer : 0
```

- 1500 billets
- 1 variable qualitative
- 6 variables quantitatives
- Pas de valeur aberrante !
- Pas de doublons !
- 37 valeurs null (margin_low)

Data.info()

```
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)
```

Data.describe()

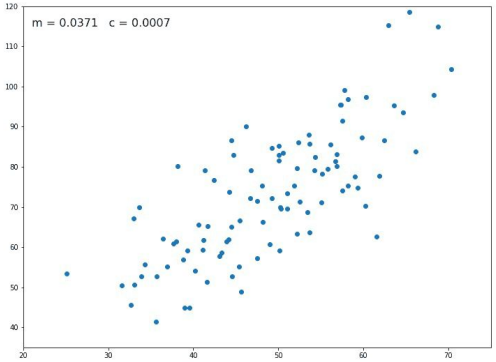
	diagonal	height_left	height_right	margin_low	margin_up	length
count	1500.000000	1500.000000	1500.000000	1463.000000	1500.000000	1500.00000
mean	171.958440	104.029533	103.920307	4.485967	3.151473	112.67850
std	0.305195	0.299462	0.325627	0.663813	0.231813	0.87273
min	171.040000	103.140000	102.820000	2.980000	2.270000	109.49000
25%	171.750000	103.820000	103.710000	4.015000	2.990000	112.03000
50%	171.960000	104.040000	103.920000	4.310000	3.140000	112.96000
75%	172.170000	104.230000	104.150000	4.870000	3.310000	113.34000
max	173.010000	104.880000	104.950000	6.900000	3.910000	114.44000

Enrichissement des données

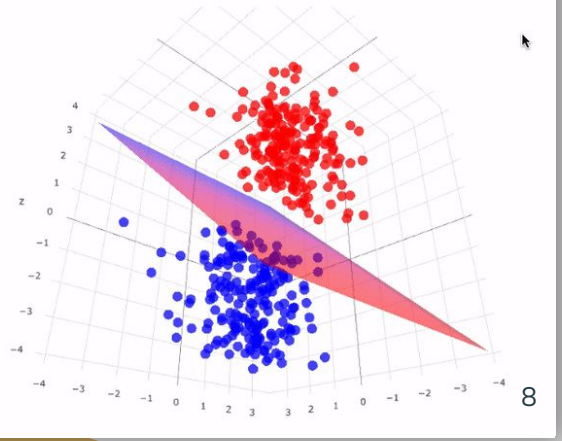
Enrichissement des données

Régression linéaire multiple (Définition)

La régression linéaire multiple est une méthode d'analyse statistique utilisée pour prédire une variable dépendante continue à partir de deux ou plusieurs variables indépendantes !

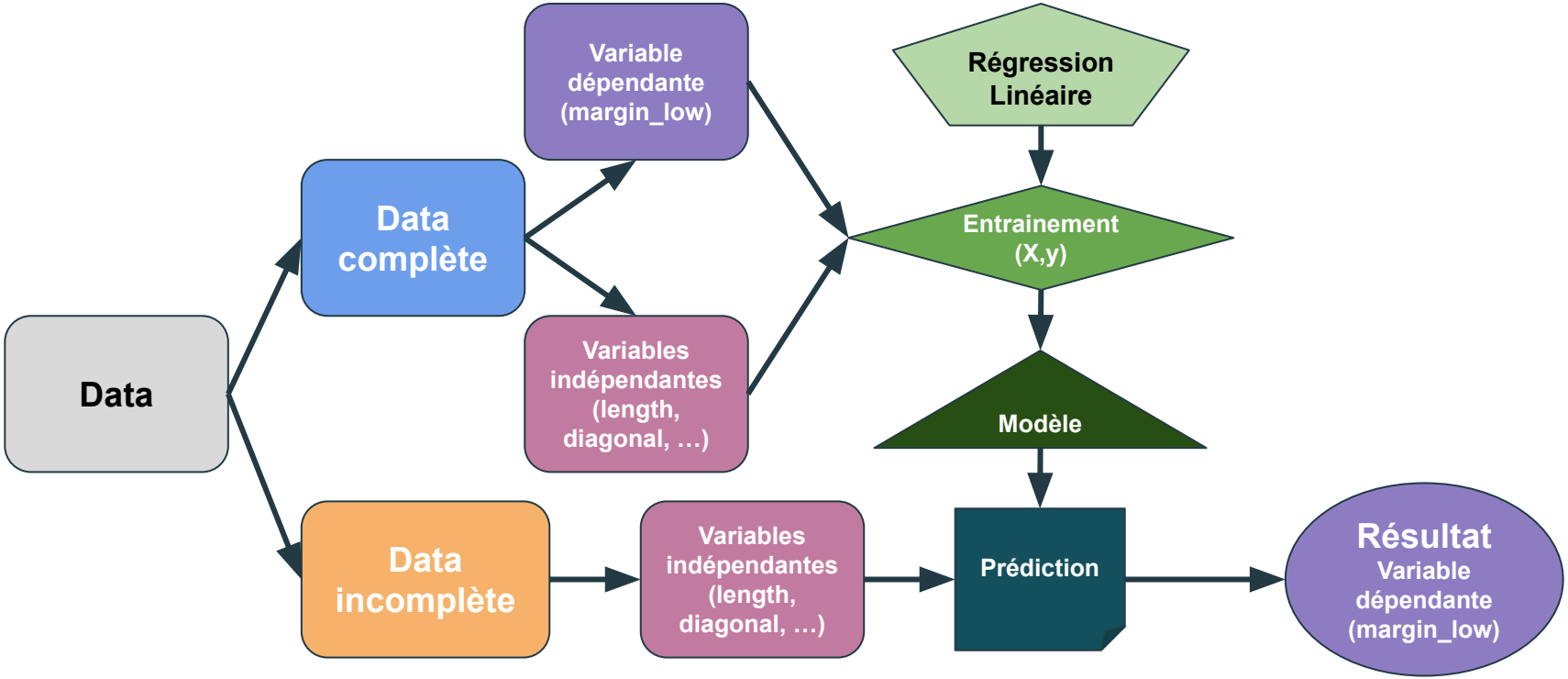


Lors de ma précédente mission, je me suis retrouvé avec un jeu de données comportant plein de valeurs manquantes, une vraie galère ! Mais j'ai eu de très bons résultats en utilisant une régression linéaire pour combler ces dernières. Ça te servira peut-être pour ta prochaine mission, qui sait ?



Enrichissement des données

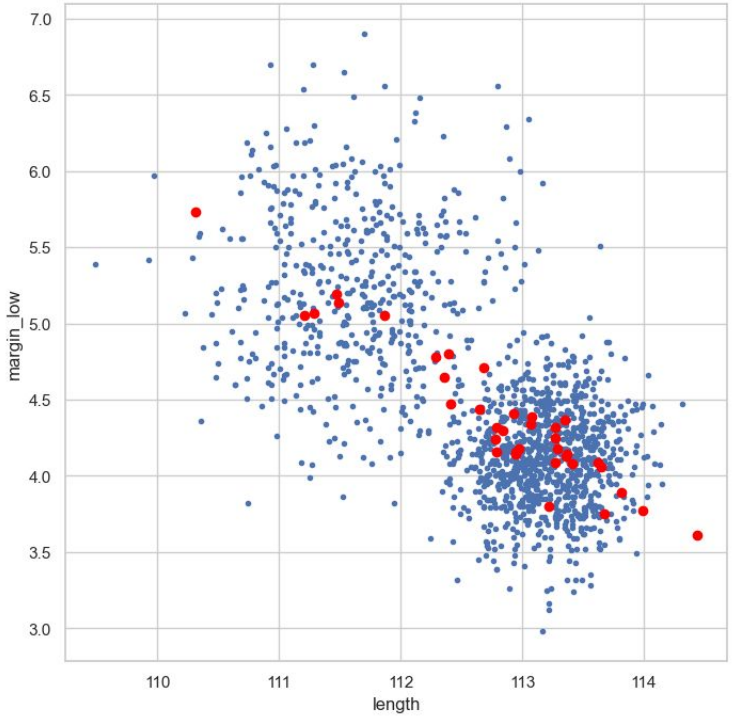
Régression linéaire multiple (Cheminement)



Enrichissement des données

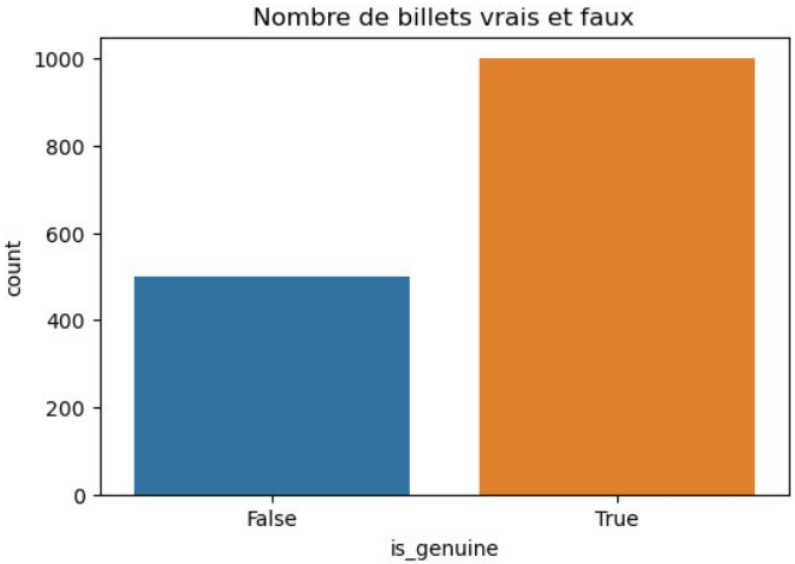
Régression linéaire multiple (Evaluation / Visualisation)

- Coefficient de détermination (R^2) : 0,47
- Test de normalité des résidus : p-value < 0,05
- Test d'homoscédasticité : p-value < 0,05
- Erreur quadratique moyenne (RMSE) : 0,47
- Coefficients de corrélation (r) :
  - diagonal : -0,11
  - height_left : 0,18
  - margin_up : 0,25
  - height_right : 0,25
  - length : -0,40



Analyse descriptive des données

Analyse descriptive des données



- 1500 billets

- 500 faux billets

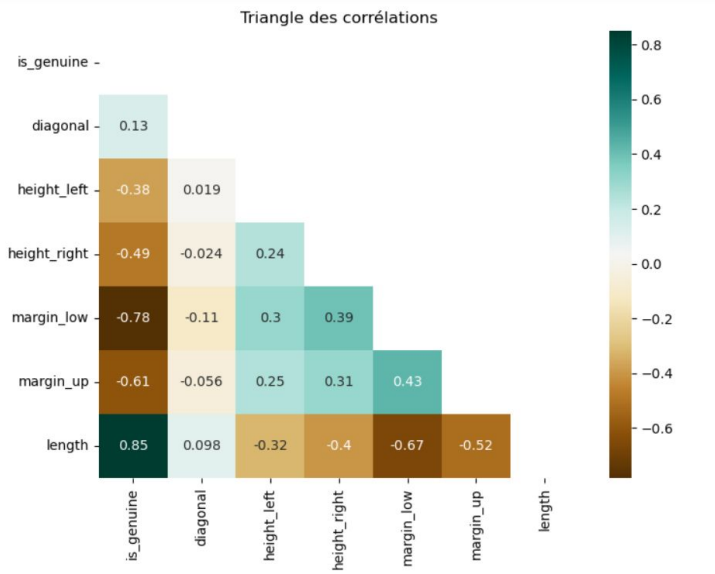
- 1000 vrais billets

Analyse descriptive des données

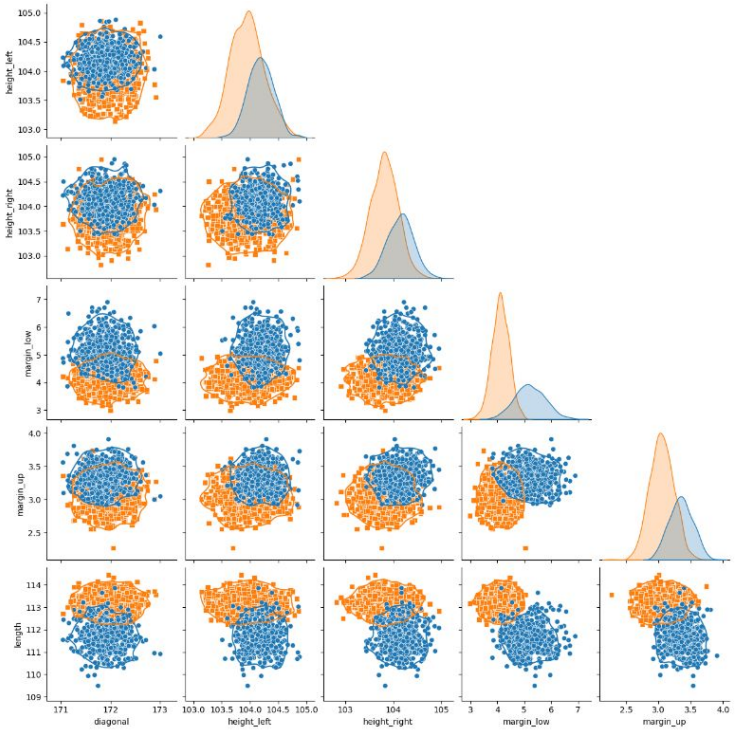
- Variables corrélées à ‘is_genuine’:
  - length : 0,85
  - margin_low : 0,78
  - margin_up : 0,61

- Variable à forte corrélation avec les autres :
  - length

- Variable à faible corrélation :
  - diagonal



Analyse descriptive des données



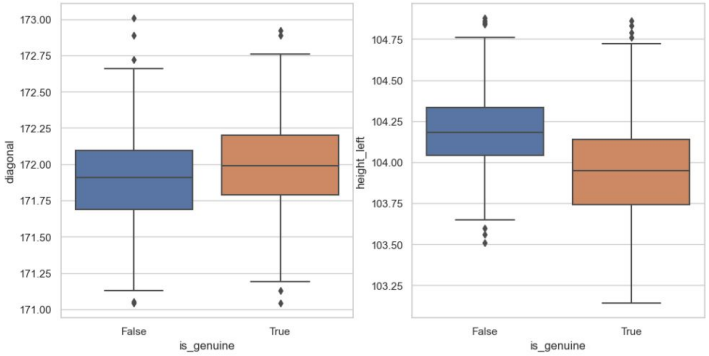
Vrais billets / **Faux billets**

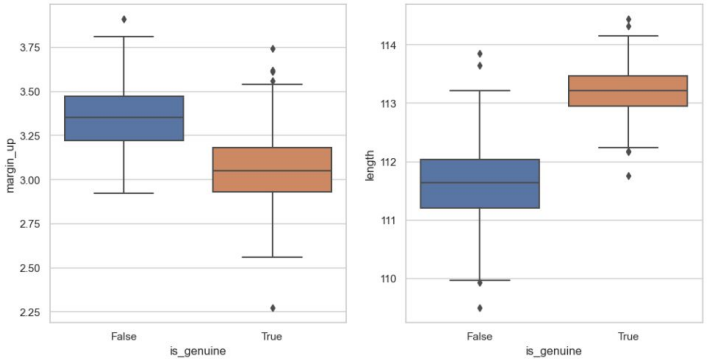
- Variables distinguant le mieux les 2 groupes :
  - length
  - margin_low

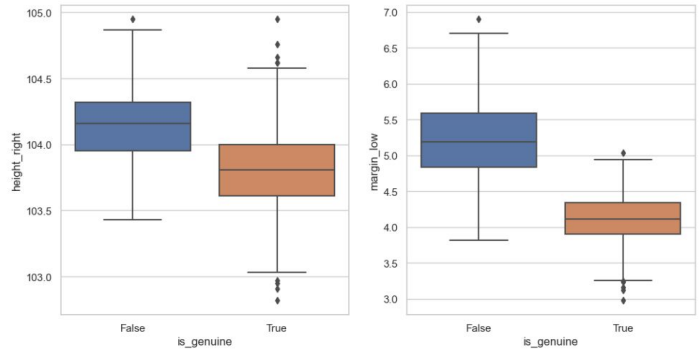
- Variable distinguant le moins les 2 groupes : :
  - diagonal

Analyse descriptive des données

- Variables distinguant le mieux les 2 groupes :
  - length
  - margin_low



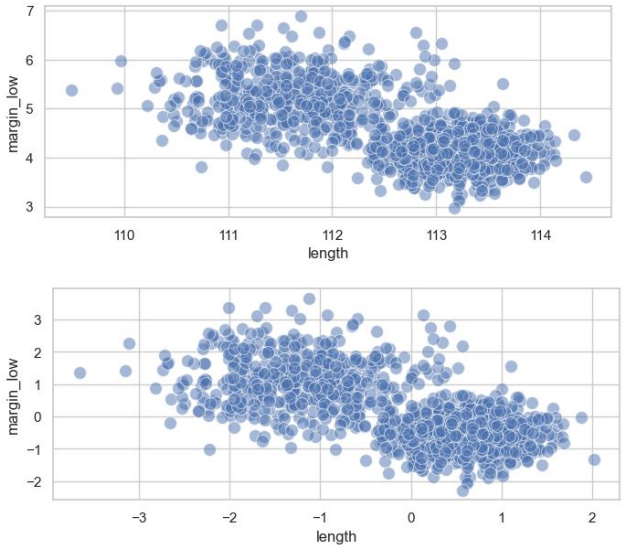




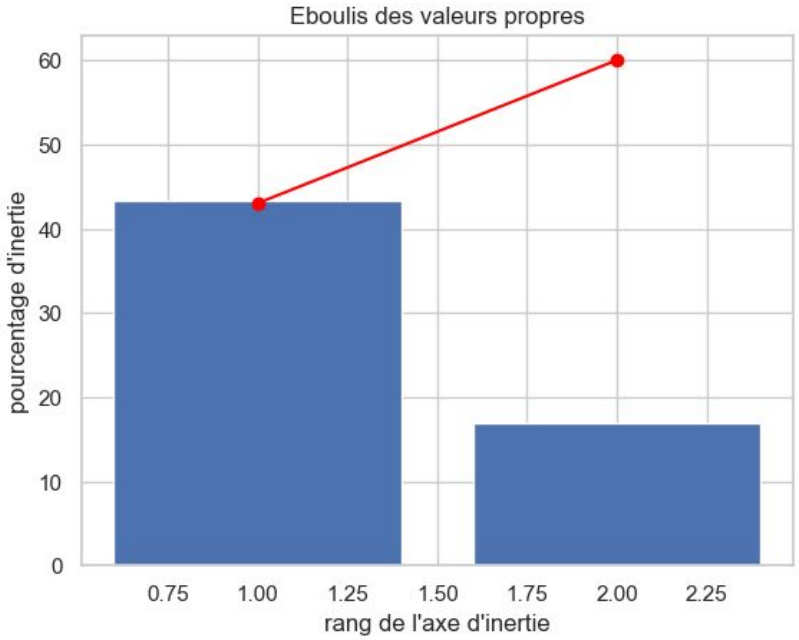
Analyse descriptive des données

Analyse en Composantes Principales

Normalisation des données

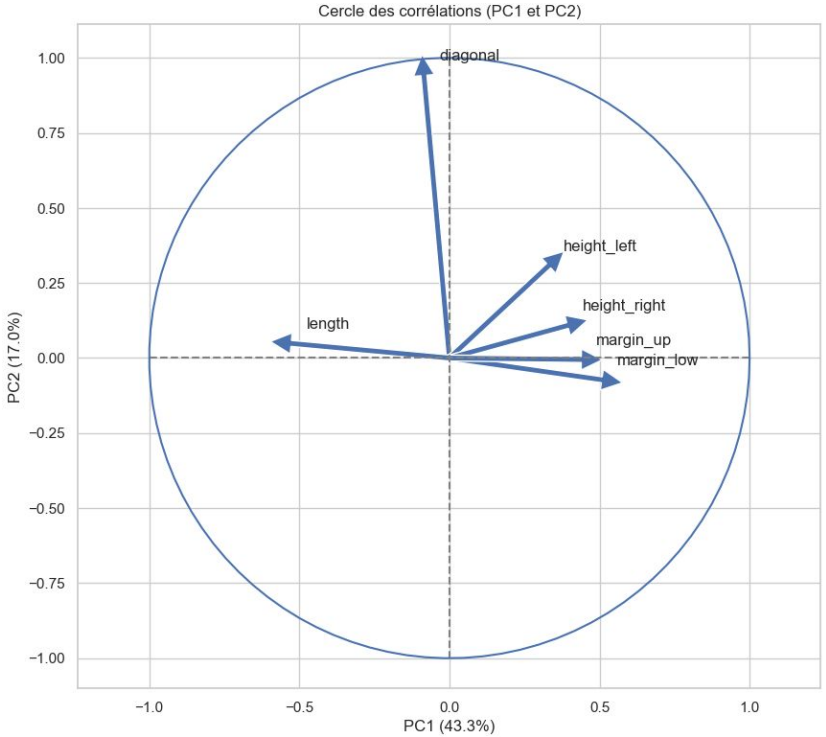


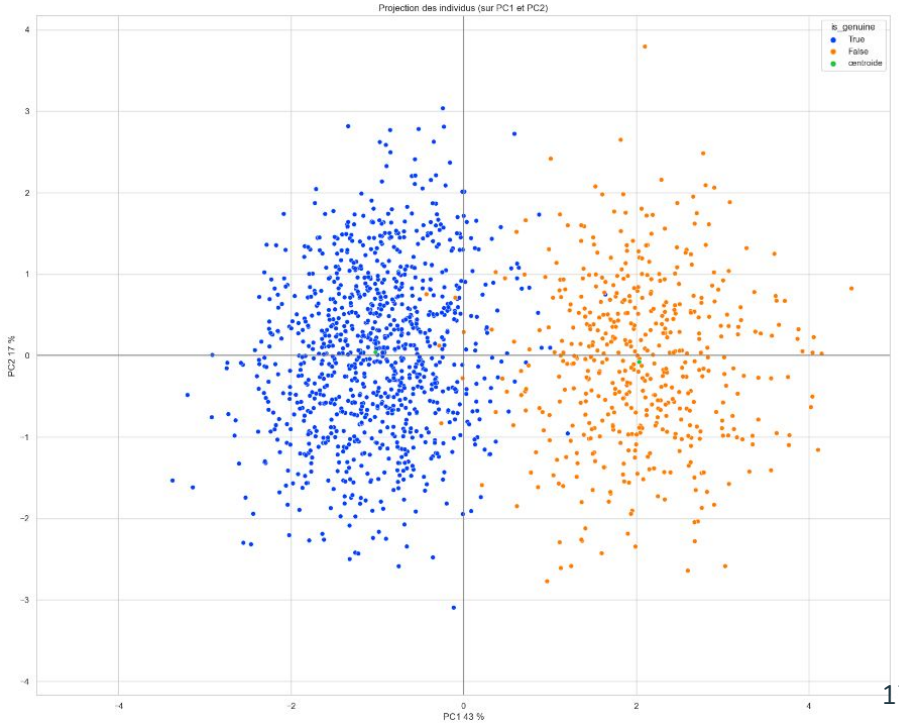
Variance expliquée : 60,2%



Analyse descriptive des données

Analyse en Composantes Principales

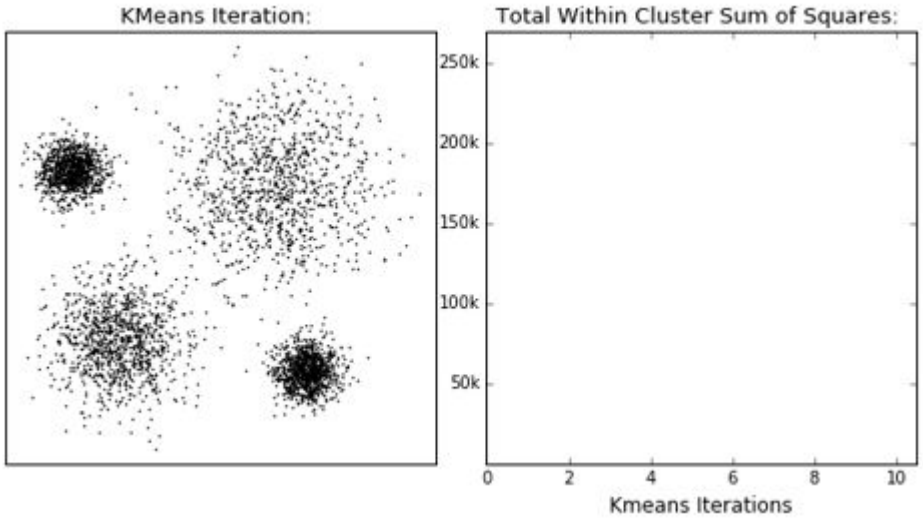




Classification supervisée

Classification supervisée

K-Means (Définition)

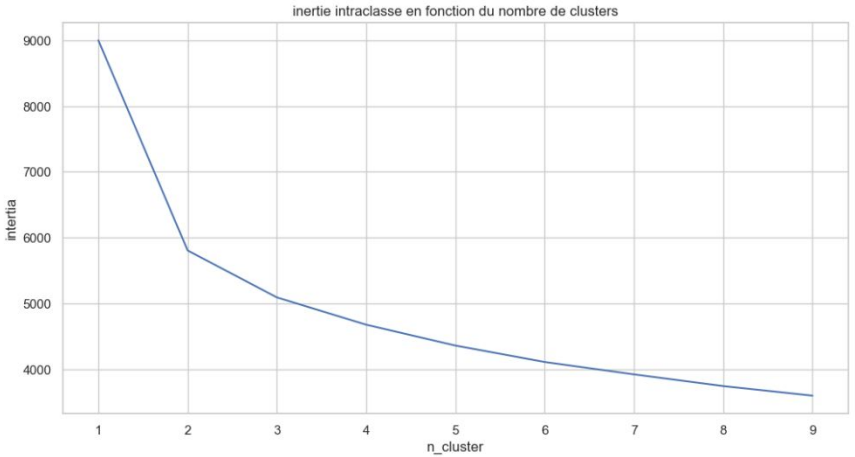


Méthode de clustering,
qui permet de regrouper des données similaires en groupe grâce aux centroïdes !

Classification supervisée

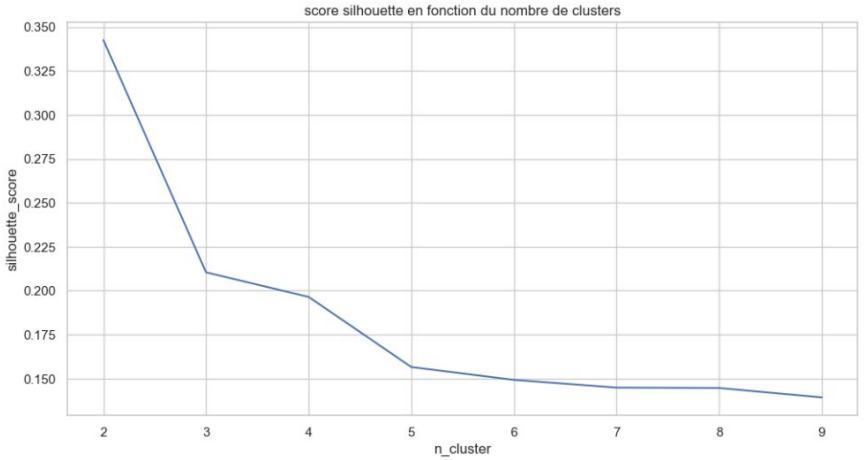
K-Means

Méthode du coude



- Cassure au niveau de **2 clusters**

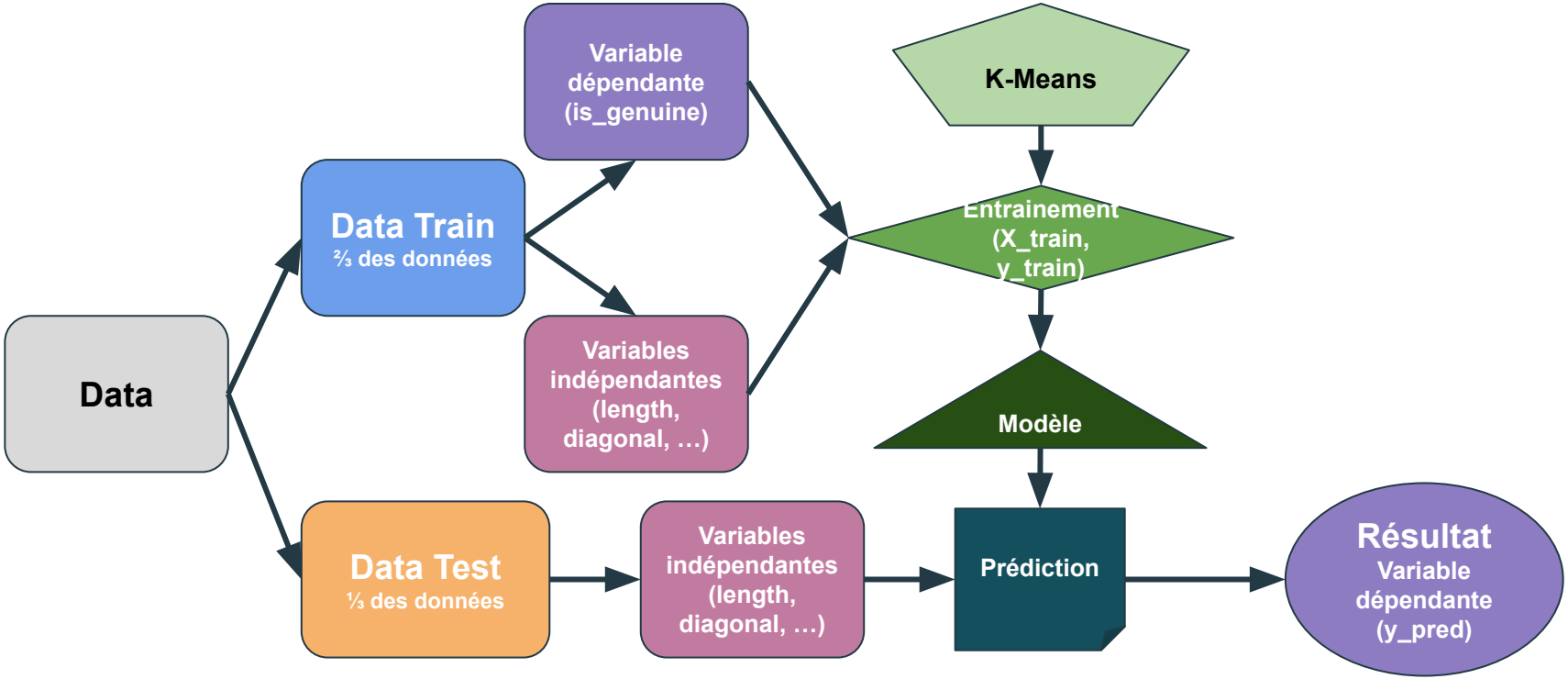
Score silhouette



- Meilleur score au niveau de **2 clusters**

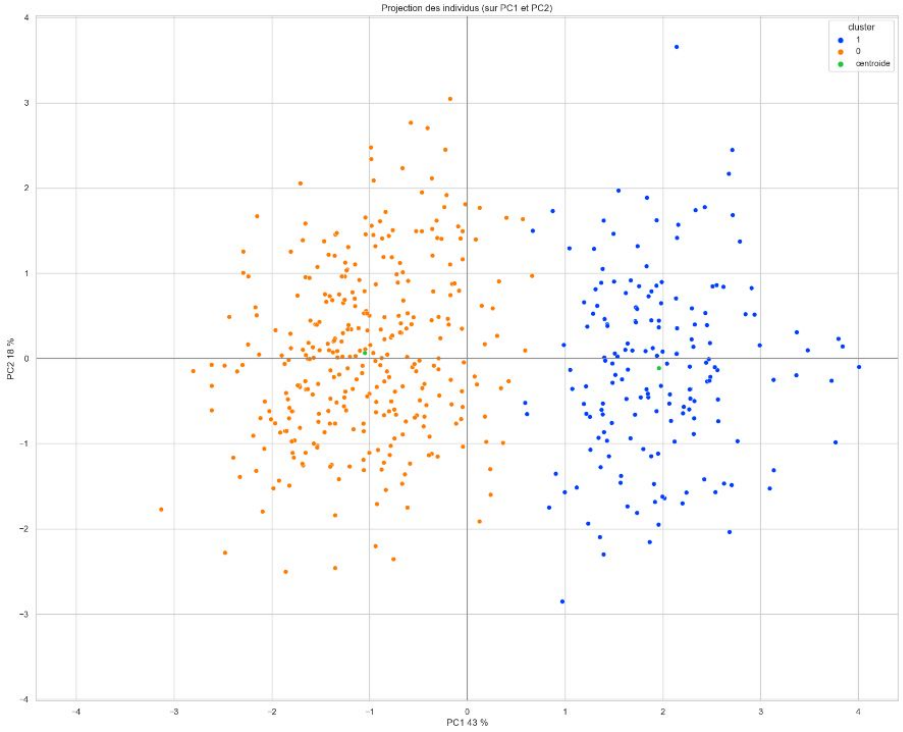
Classification supervisée

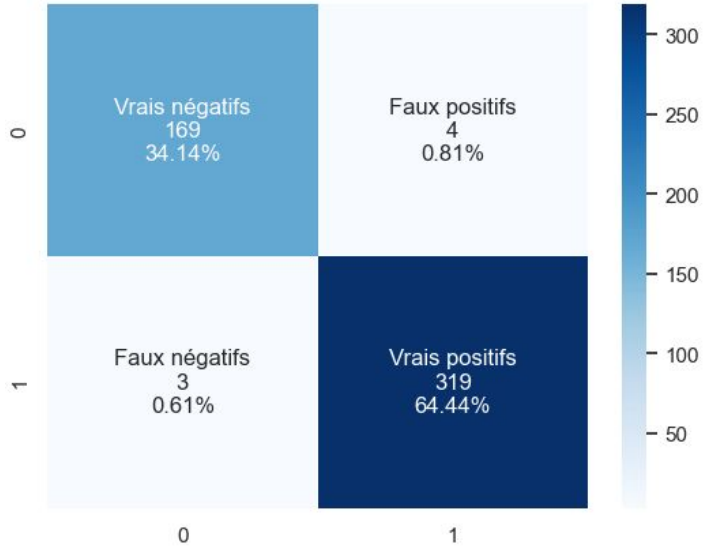
K-Means (Cheminement)



Classification supervisée

K-Means (Visualisation / Evaluation)



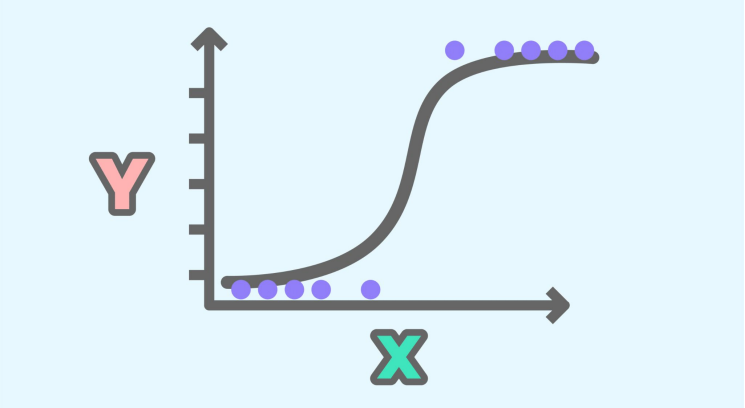


Précision : 0,98
Sensibilité : 0,98

Classification supervisée

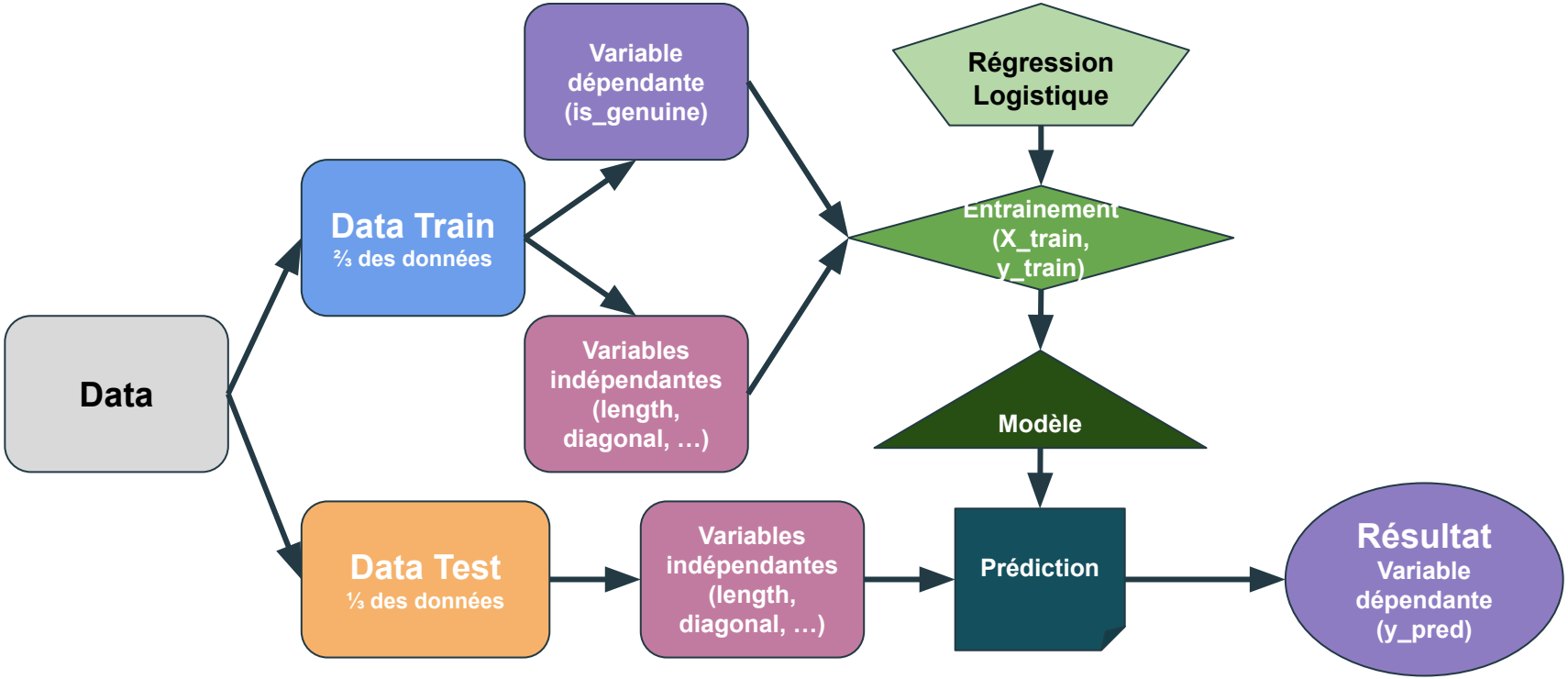
Régression Logistique (Définition)

Méthode d'analyse statistique utilisée pour prédire une variable binaire à partir d'une ou plusieurs variables indépendantes !



Classification supervisée

Régression Logistique (Cheminement)



Classification supervisée

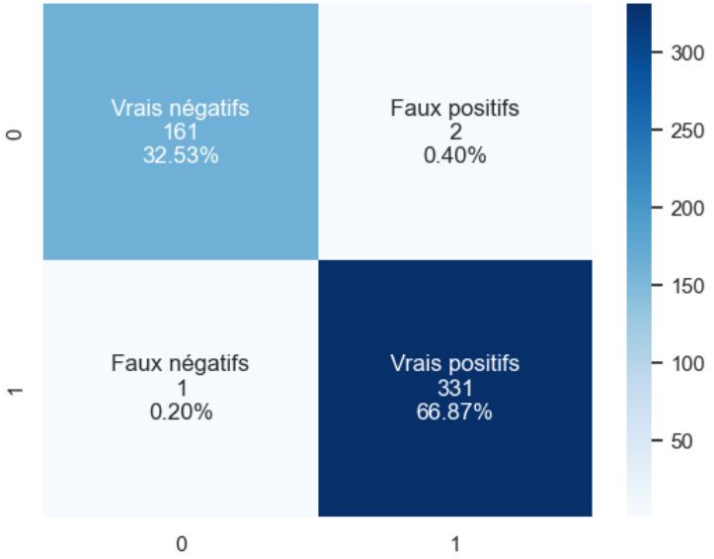
Régression Logistique

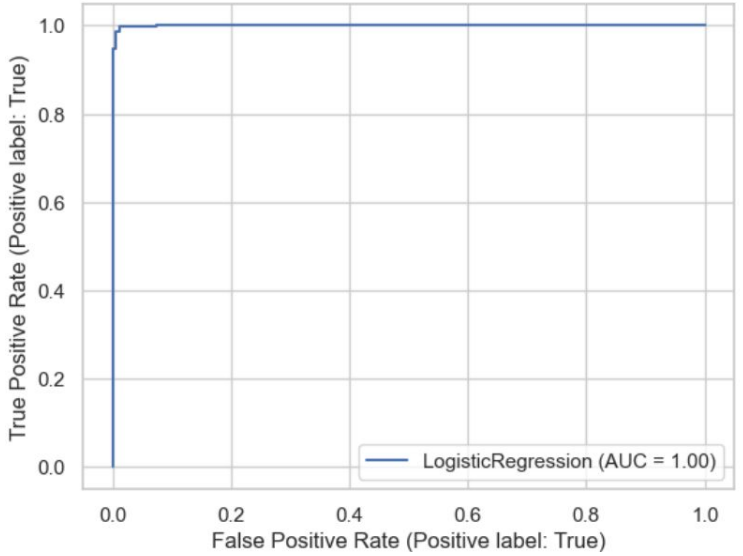
	False	True
0	0.996883	0.003117
1	0.053609	0.946391
2	0.993031	0.006969
3	0.999834	0.000166
4	0.104806	0.895194
5	0.002506	0.997494
6	0.006800	0.993200
7	0.001976	0.998024
8	0.000158	0.999842
9	0.000528	0.999472

- Coefficient de corrélation à la variable dépendante :
  - diagonal : 0,17
  - height_left : -0,44
  - height_right : -0,67
  - margin_up : -1,60
  - margin_low : -2,61
  - length : 3,32

Classification supervisée

Régression Logistique (Evaluation)





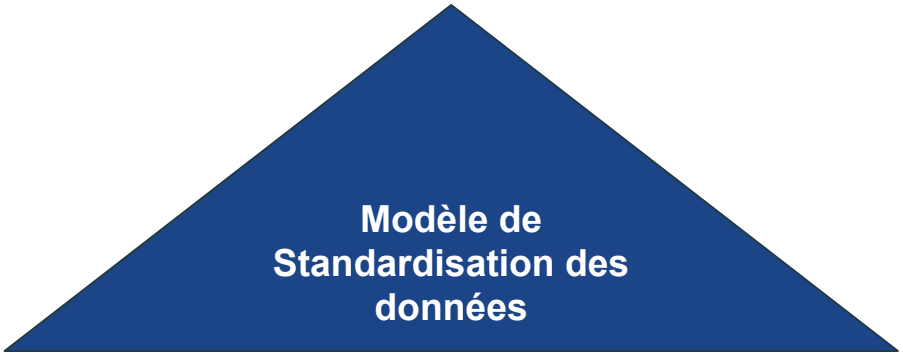
Précision : 0,99 proche de 1
Sensibilité : 0,99 proche de 1

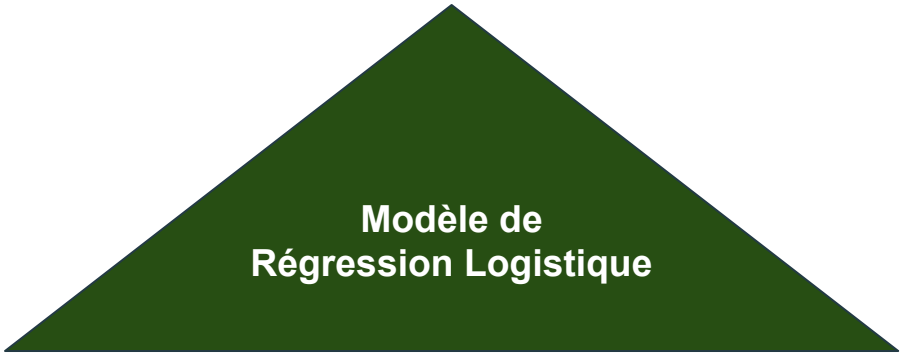
Programme de détection de faux billets

Programme de détection

Prérequis

Enregistrer les modèles d'entraînement dans un fichier !





Programme de détection

Fonctionnement