Détecteur de faux billets

ONCFM

Contexte

ONCFM

Je suis data analyst dans une entreprise spécialisée dans le data.

Je réalisé une prestation au sein de l'ONCFM.

- Identification des contrefaçons des billets en euros
- suivies des indications données (cahier des charges + post-it)

Déroulement du projet

1. **Exploration et nettoyage des données**
2. **Enrichissement des données**
3. **Analyse des données**
4. **Programme de détection des faux billets**

Exploration et nettoyage des données

Mes données

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

```
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

```
data.isnull().sum()
```

```
is_genuine       0
diagonal         0
height_left      0
height_right     0
margin_low      37
margin_up        0
length           0
dtype: int64
```

```
data.info()
```

```
<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
```

```
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

1500 rows × 7 columns

Enrichissement des données

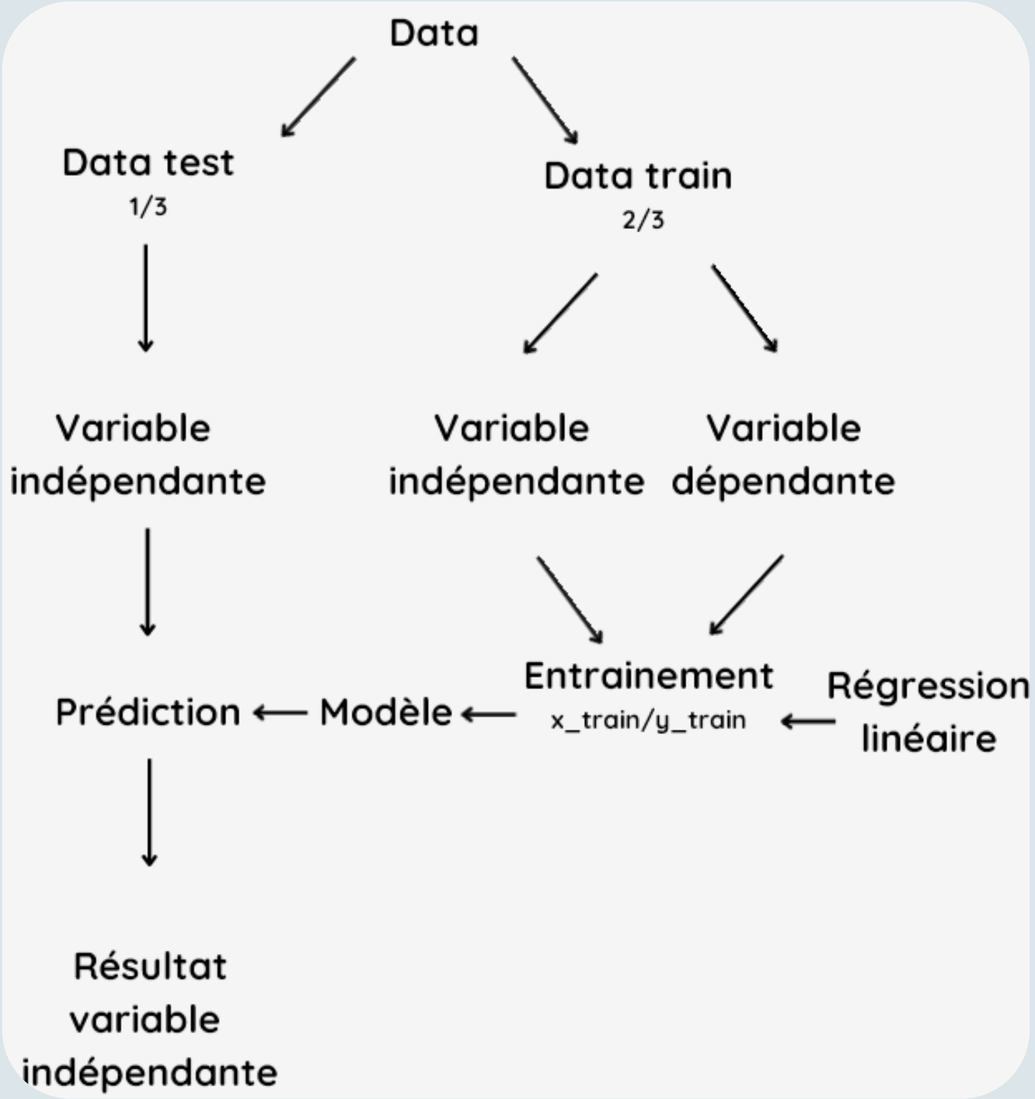
Définition

Régression linéaire :

- prédis un variable continue via un relation linéaire entre celle-ci et une (simple) ou plusieurs autres (multiple)

Ces condition de validité :

- Linéarité : relation entre les variables doit être linéaire
- Homoscedasticité : variance des résidus doit être constante
- Indépendance et normalité des résidus : il doivent être indépendants et suivre une distribution normale.



Enrichissement des données

Résultats

Régression linéaire simple :

- Performance du modèle relative.

Régression linéaire multiple :

- Performance du modèle relative
- Validation par l'étude des erreurs : MAE, MSE, RMSE
- Condition de validité

Simple

```
Constante : 60.71
Coef : [-0.5]
R² : 0.44
```

Multiple

```
Constante : 20.4
Coef : [-0.09  0.15  0.26  0.29 -0.4 ]
R² : 0.48
```

Erreurs

```
MAE est de : 0.34
MSE est de : 0.19
RMSE est de : 0.44
```

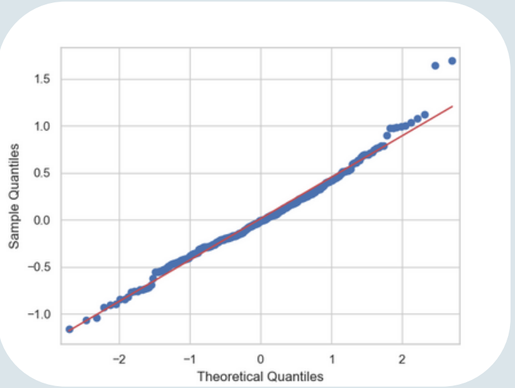
Enrichissement des données

Résultats

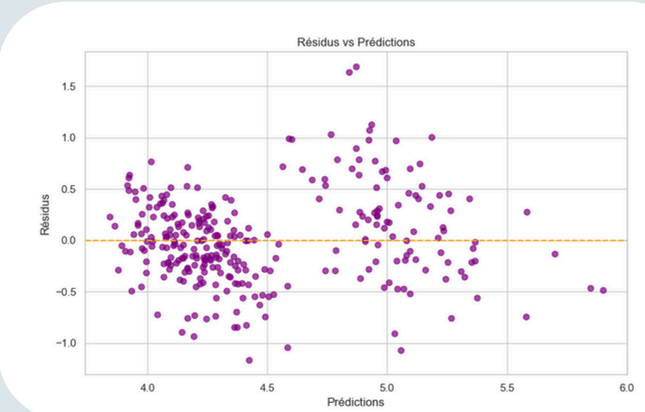
Régression linéaire multiple

- Normalité des résidus : Oui
  - Stat : 0.98
  - p-value : 0.004
- Homoscedasticité : Non
  - Stat : 85.57
  - p-value : 1.15
- Indépendance des résidus : Oui
  - Stat : 1.94

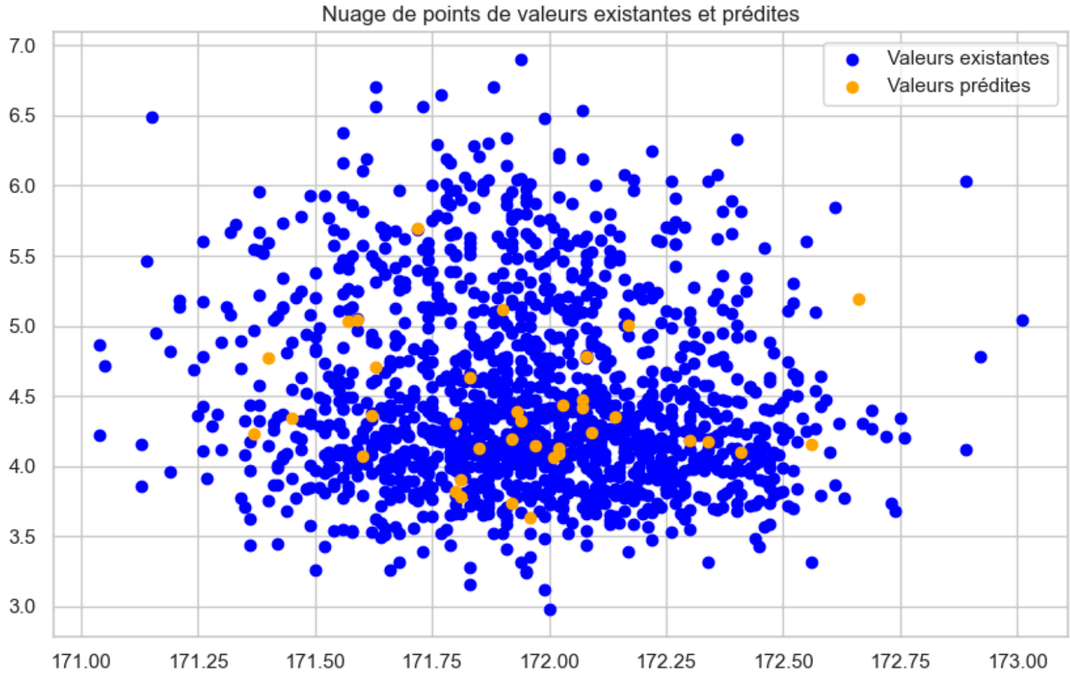
Normalité des résidus



Homoscédasticité des résidus



Valeurs prédites par rapport aux valeurs connues

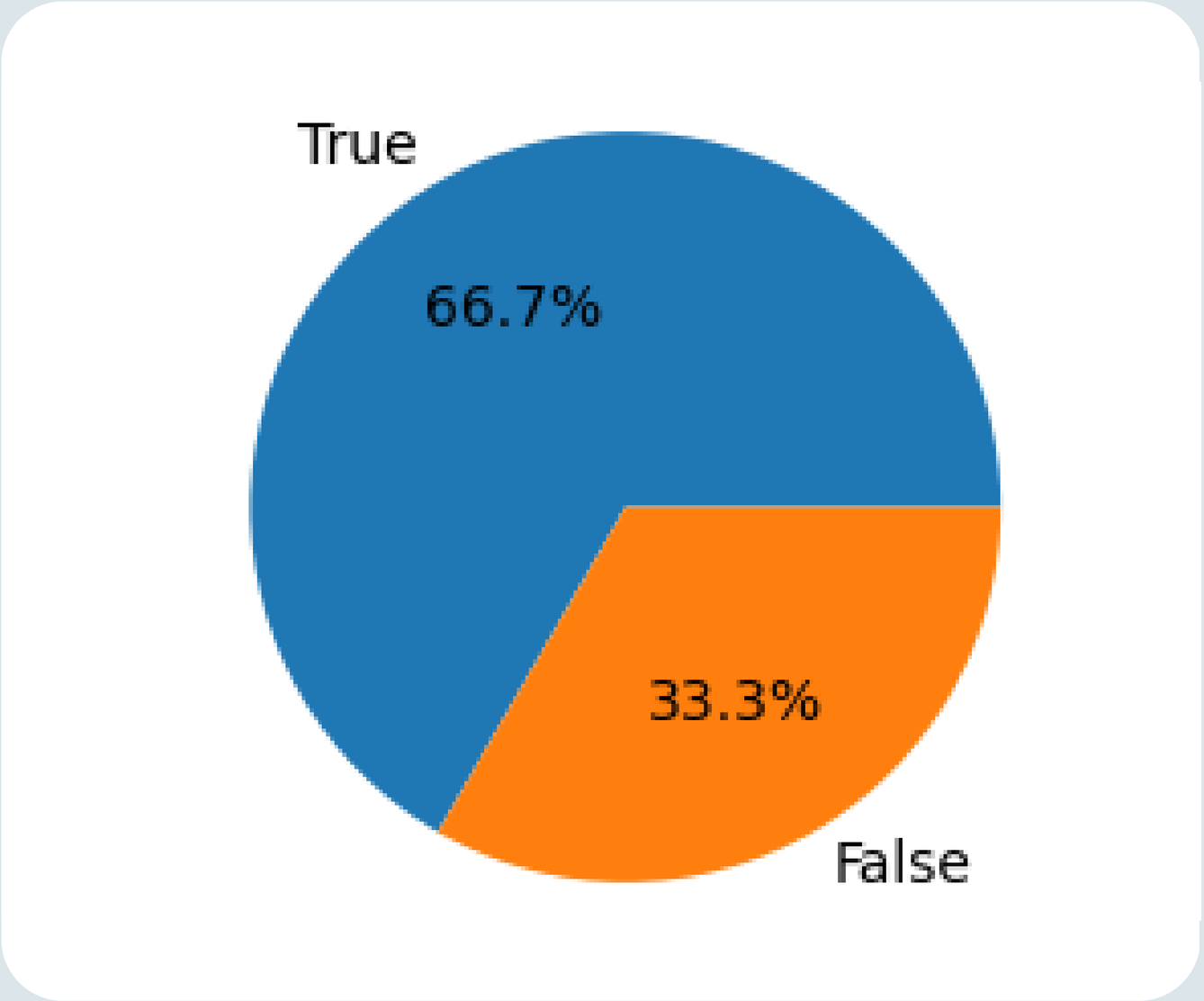


Analyse descriptive des données

Diagramme circulaire

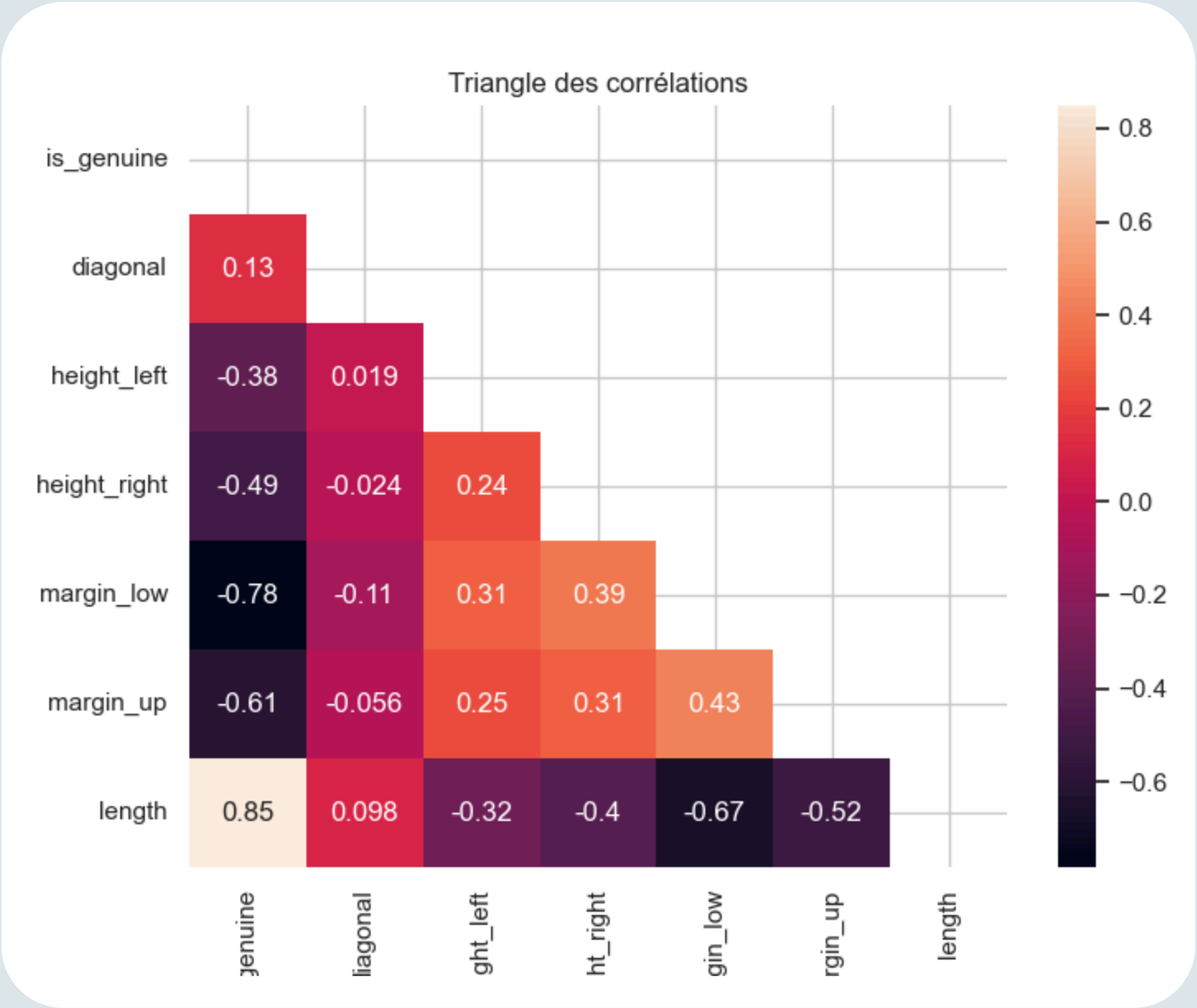
Répartition de is_genuine

- 1500 billets
- 500 faux billets
- 1000 vrais billets



Analyse descriptive des données

Carte thermique



Variables corrélées à “is_genuine”:

- length : 0.85
- margin_low : 0.78
- margin_up : 0.61

Analyse descriptive des données

Carte de nuage de points

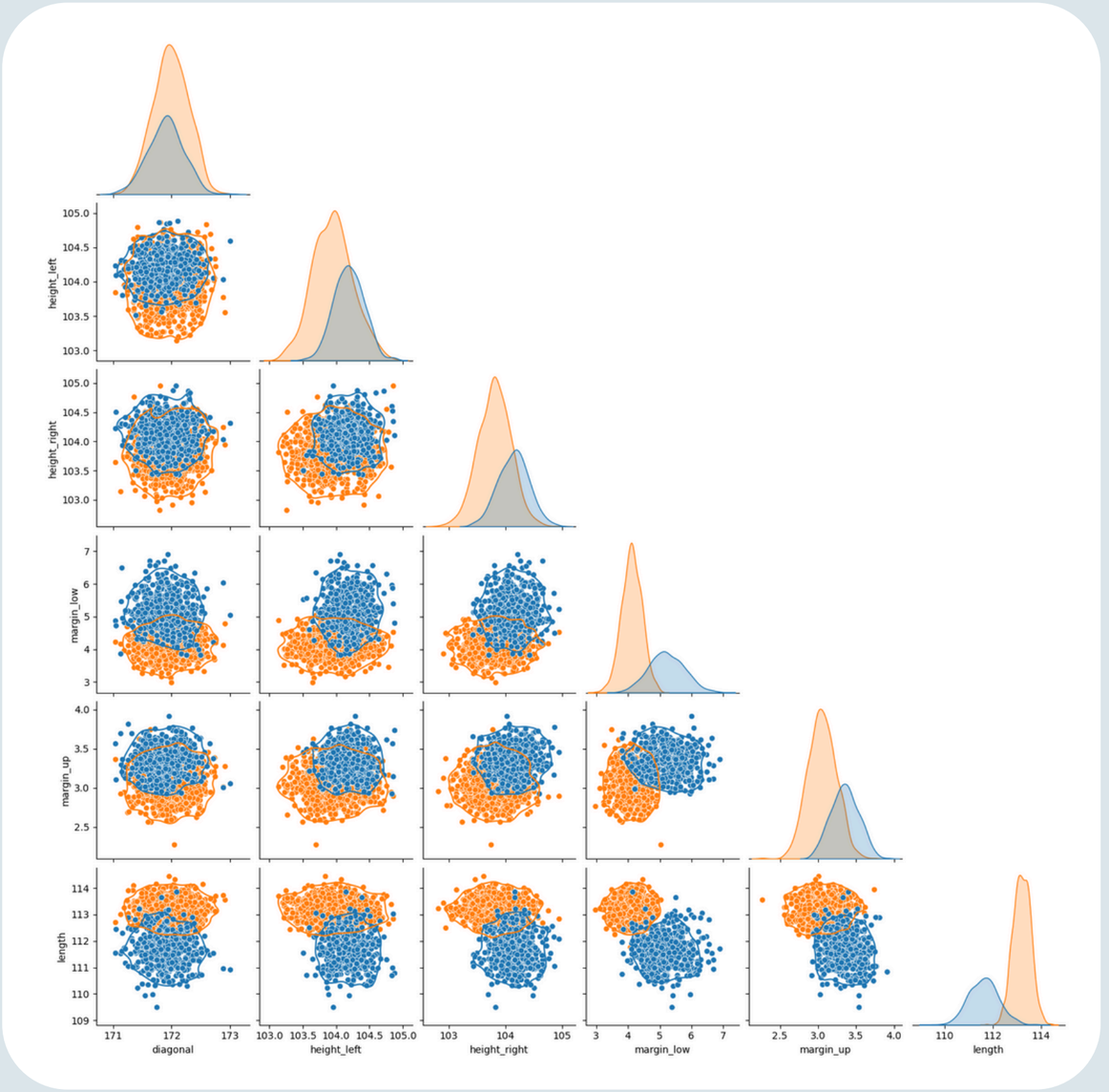
Vraies billets / **Faux billets**

Variables **distinguant mieux** les deux groupes :

- length
- margin_low

Variables **distinguant moins** les deux groupe :

- diagonal



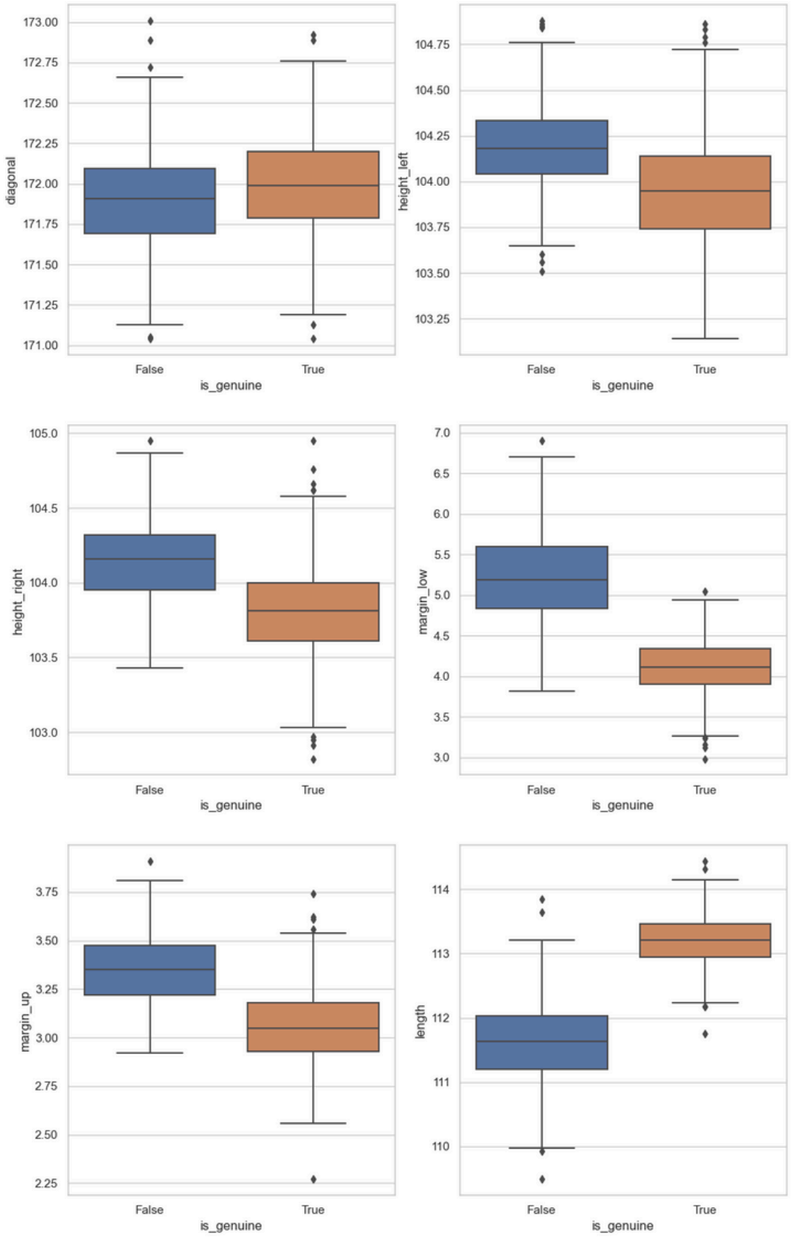
Analyse descriptive des données

Boîtes à moustaches

Vraies billets / **Faux billets**

Variables distinguant mieux les deux groupes :

- length
- margin_low



Analyse descriptive des données

Analyse de composantes principales - Définition

Théorique :

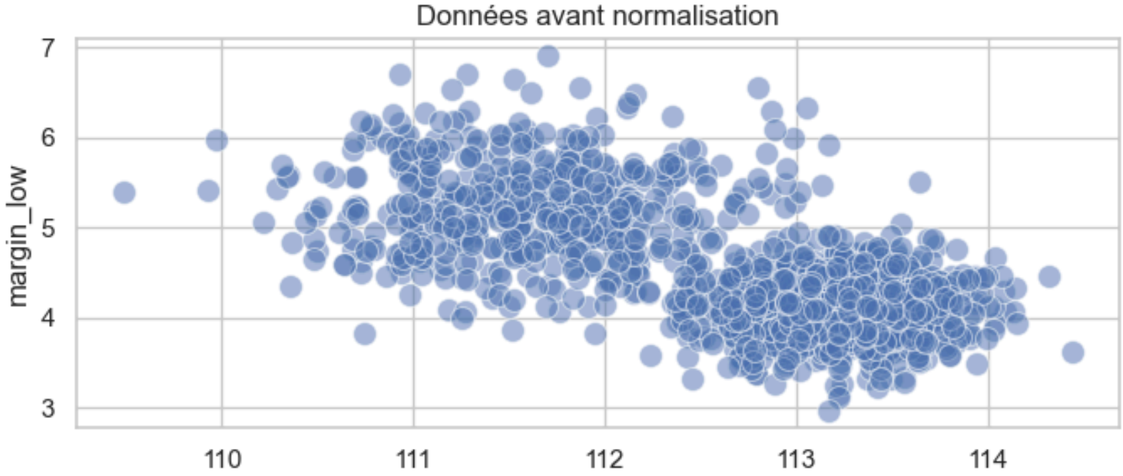
- méthode de réduction de dimensionnalité
- composantes principales
- capture de la variance

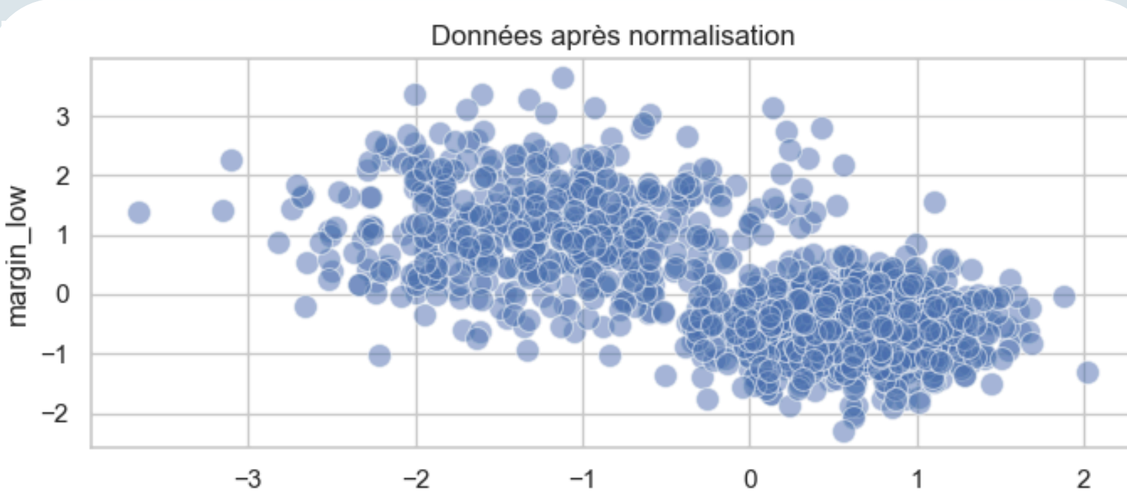
Les distances :

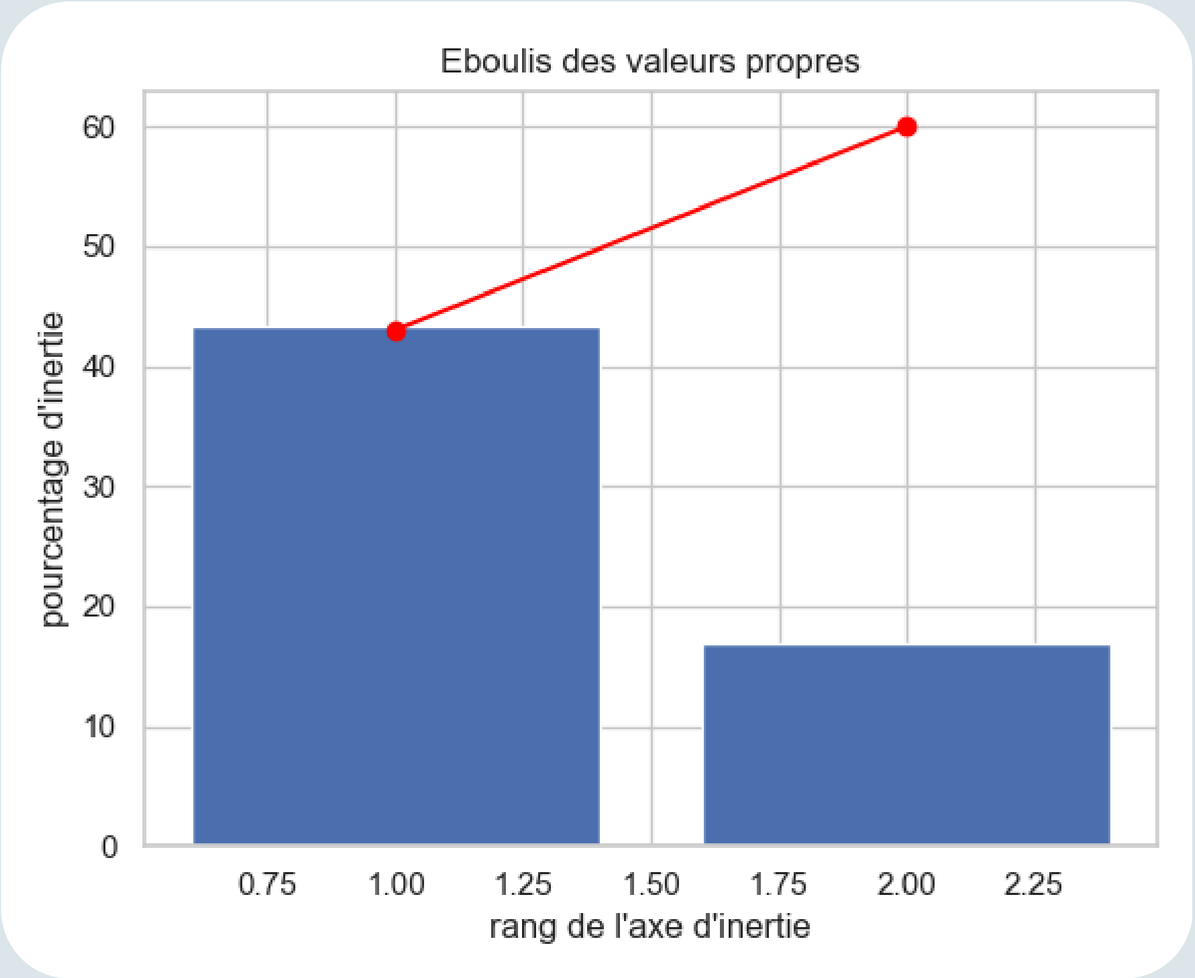
- distance indirectement
- entre variable et composantes principales

Analyse descriptive des données

Analyse de composantes principales - Résultats

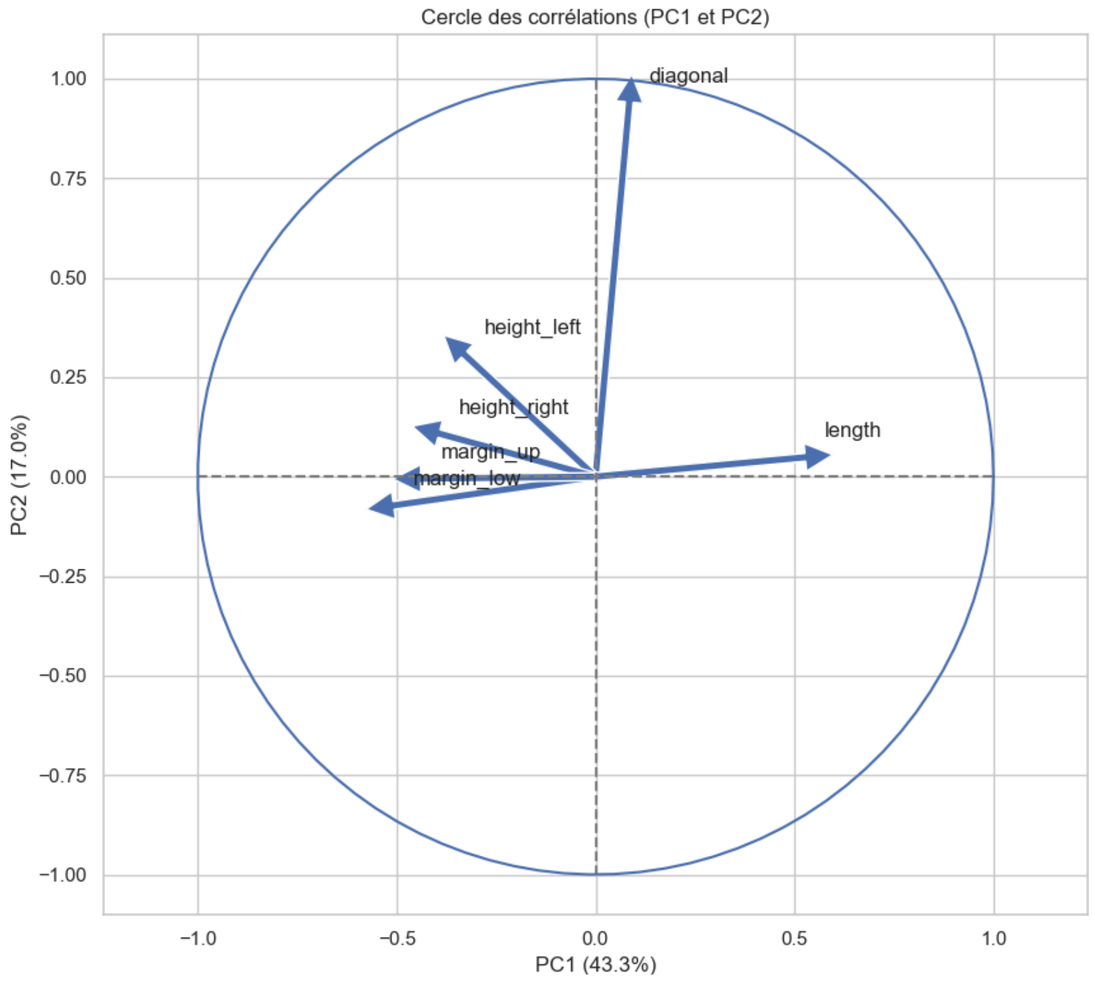






Analyse descriptive des données

Analyse de composantes principales - Résultats



Analyse en classification supervisé des données

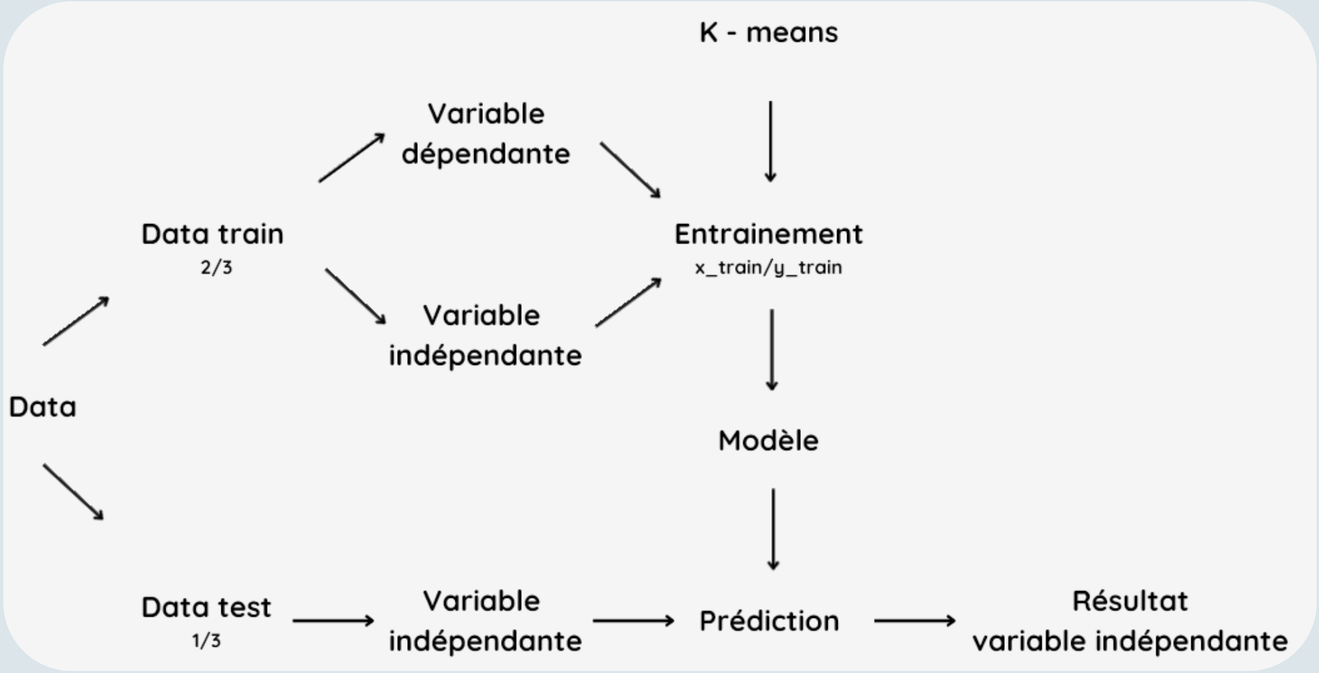
K-means - Définition

Théorique :

- un algorithme de clustering
- variables en clusters autour de centre de gravité
- pré définition du nombre de cluster

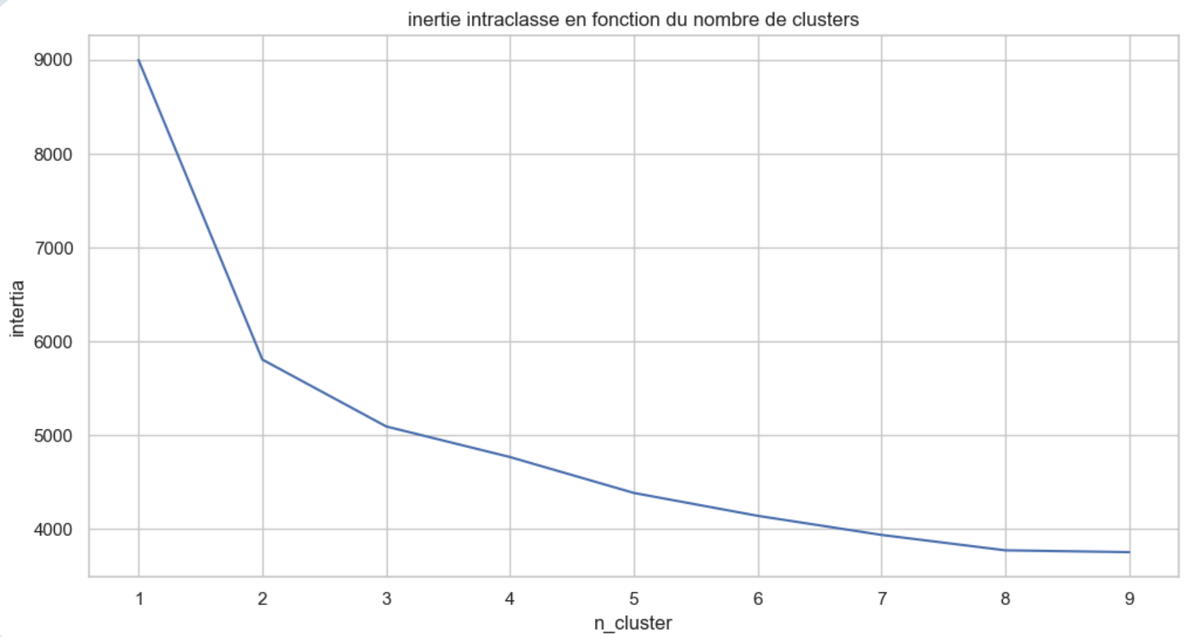
Les distances :

- distance euclidienne entre variables et centre de gravité
- minimisation des sommes des distance intracluster
- centre de gravité recalculés à chaque itération jusqu'à stabilisation



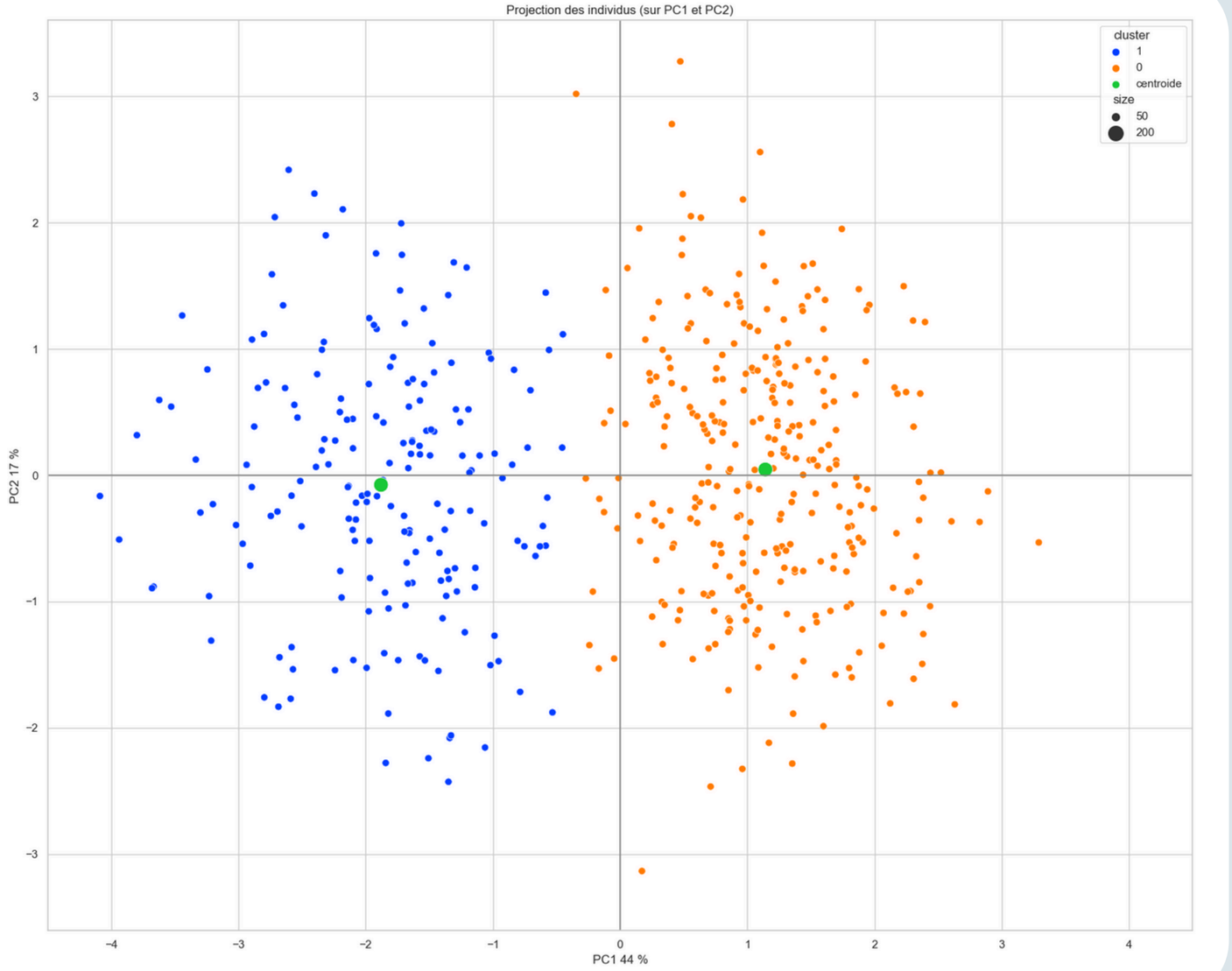
Analyse en classification supervisé des données

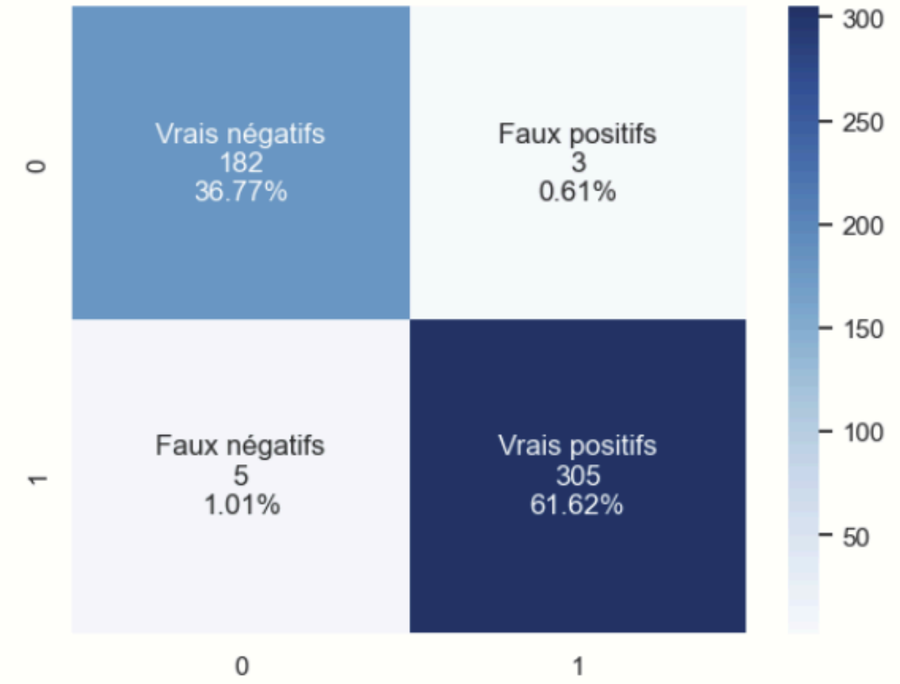
K-means - Résultats



Analyse en classification supervisé des données

K-means - Résultats





	precision	recall	f1-score	support
1	0.97	0.98	0.98	185
2	0.99	0.98	0.99	310
accuracy			0.98	495
macro avg	0.98	0.98	0.98	495
weighted avg	0.98	0.98	0.98	495

Analyse en classification supervisé des données

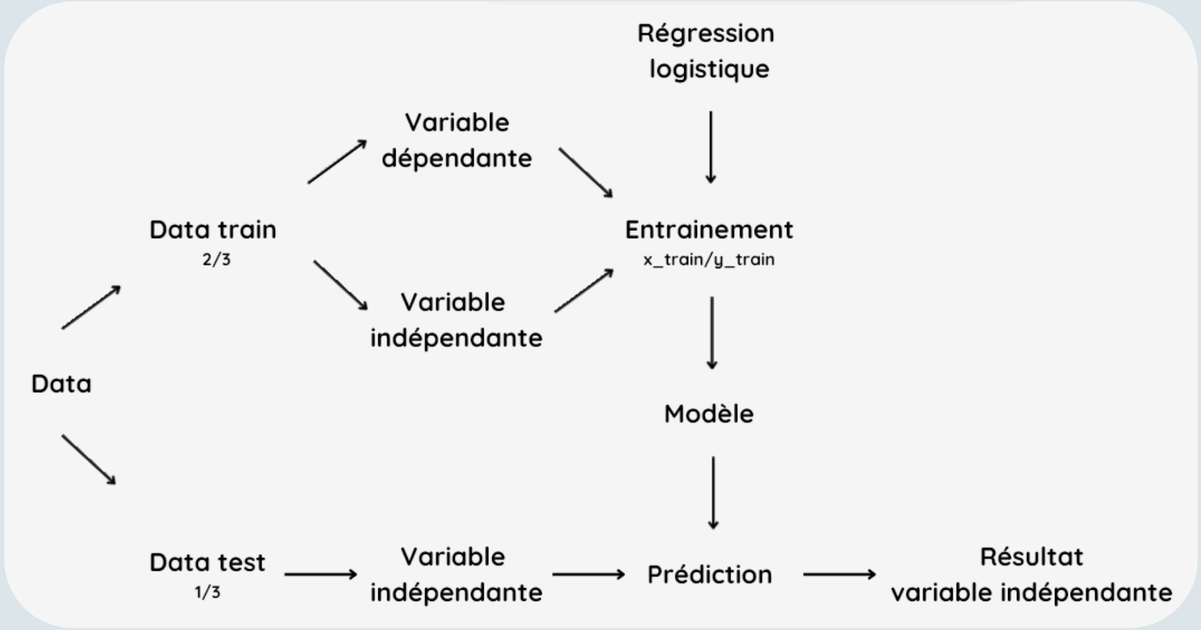
Régression logistique - Définition

Théorique :

- prédiction variable dépendente binaire
- calcul la probabilité d'appartenance à un classe

Condition d'application :

- linéarité entre prédicteurs et logit
- indépendance des observations
- significativité des coefficients



Analyse en classification supervisé des données

Régression logistiques - Résultats

Valeurs classiques

```
Constante : [-0.04]
R² : 0.9920398009950249
Coef :
           var       coef
0     diagonal  -0.492998
1  height_left  -1.281431
2 height_right  -4.345122
3   margin_low  -6.599565
4    margin_up -12.125349
5       length   6.571012
```

-23.92430076300218

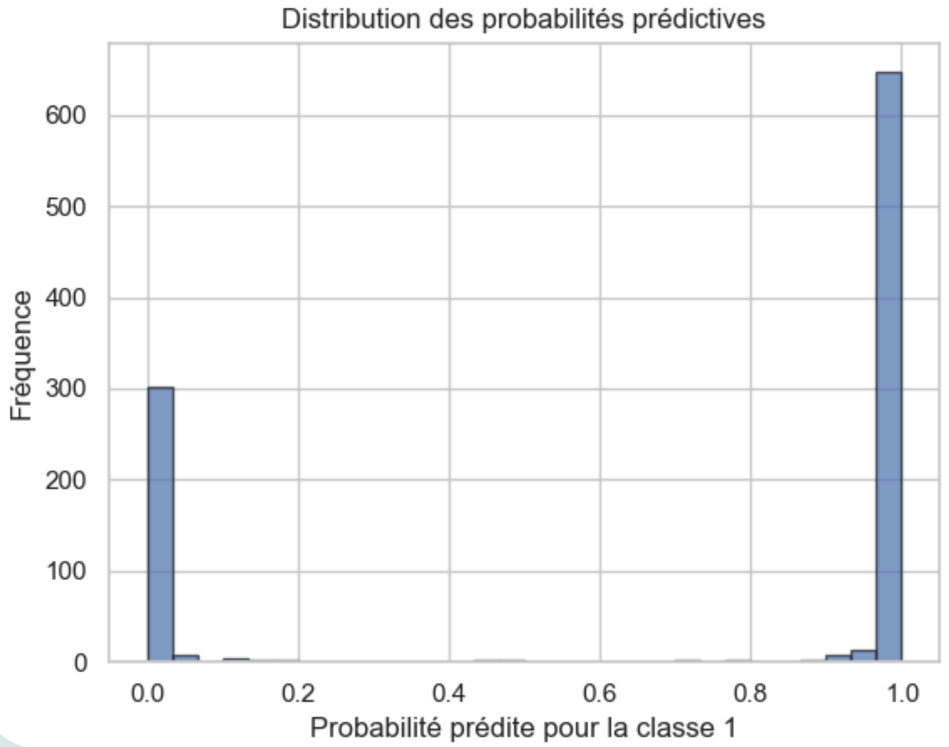
	Variable	Min	Max	Moyenne	Variance	Skewness	Kurtosis
0	diagonal	171.04	173.01	171.957592	0.093713	-0.018497	-0.135367
1	height_left	103.22	104.88	104.021244	0.086646	-0.083420	-0.230351
2	height_right	102.82	104.95	103.913423	0.112952	0.031918	-0.067470
3	margin_low	3.12	6.90	4.474756	0.421126	0.975721	0.612261
4	margin_up	2.27	3.77	3.151622	0.052888	0.133325	-0.227968
5	length	109.93	114.32	112.696448	0.745368	-0.870891	-0.157161

Valeurs normalisées

```
Constante : [2.95]
R² : 0.991044776119403
Coef :
           var       coef
0     diagonal  0.025602
1  height_left -0.434824
2 height_right -1.354770
3   margin_low -3.368295
4    margin_up -2.481555
5       length  4.228343
```

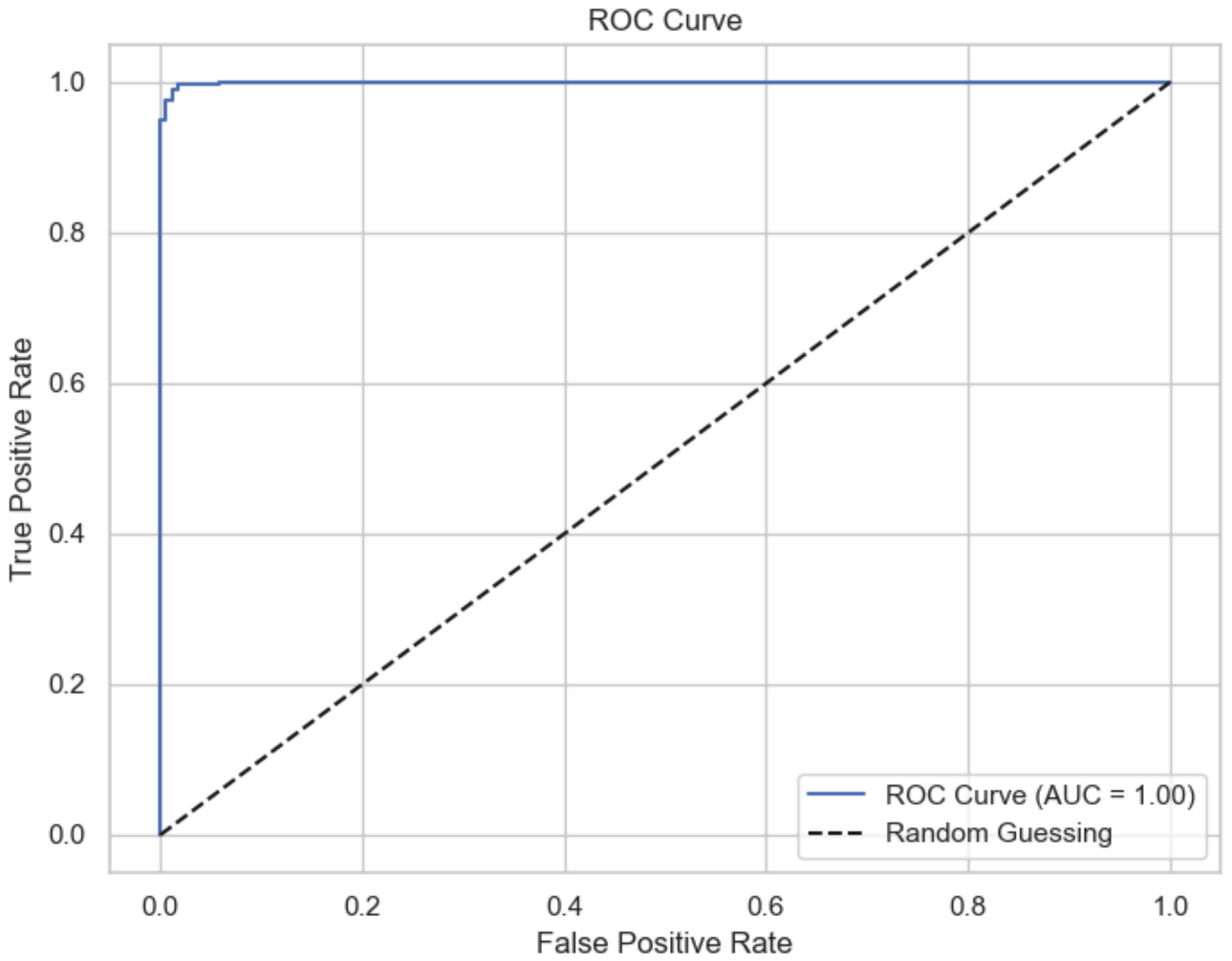
-23.654156859539878

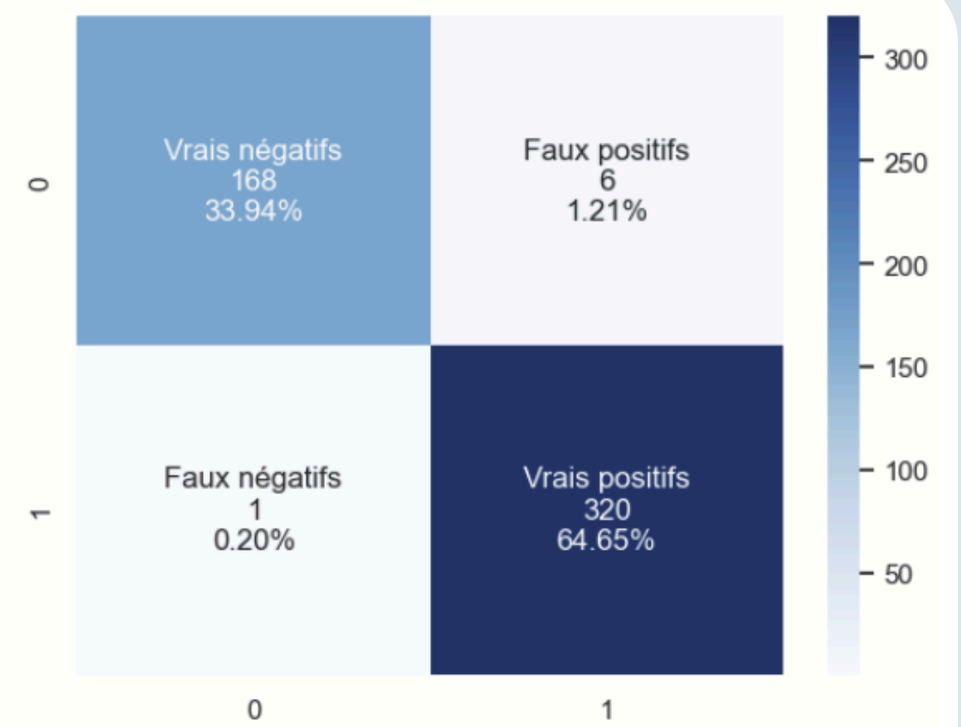
	Variable	Min	Max	Moyenne	Variance	Skewness	Kurtosis
0	diagonal	-2.997438	3.437832	-1.366292e-14	1.0	-0.018497	-0.135367
1	height_left	-2.722022	2.917406	4.660948e-15	1.0	-0.083420	-0.230351
2	height_right	-3.253424	3.084282	2.207632e-14	1.0	0.031918	-0.067470
3	margin_low	-2.087637	3.737225	7.070077e-17	1.0	0.975721	0.612261
4	margin_up	-3.833573	2.688905	-1.091443e-15	1.0	0.133325	-0.227968
5	length	-3.204328	1.880532	8.699730e-15	1.0	-0.870891	-0.157161



Analyse en classification supervisé des données

Régression logistiques - Résultats





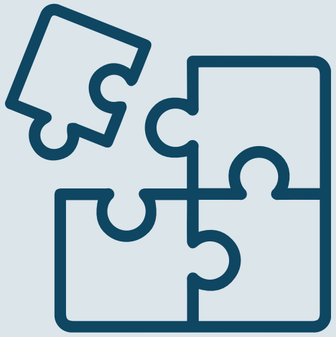
	precision	recall	f1-score	support
False	0.99	0.97	0.98	174
True	0.98	1.00	0.99	321
accuracy			0.99	495
macro avg	0.99	0.98	0.98	495
weighted avg	0.99	0.99	0.99	495

Programme de détection de faux billets

Construction

Modèles de normalisation des données

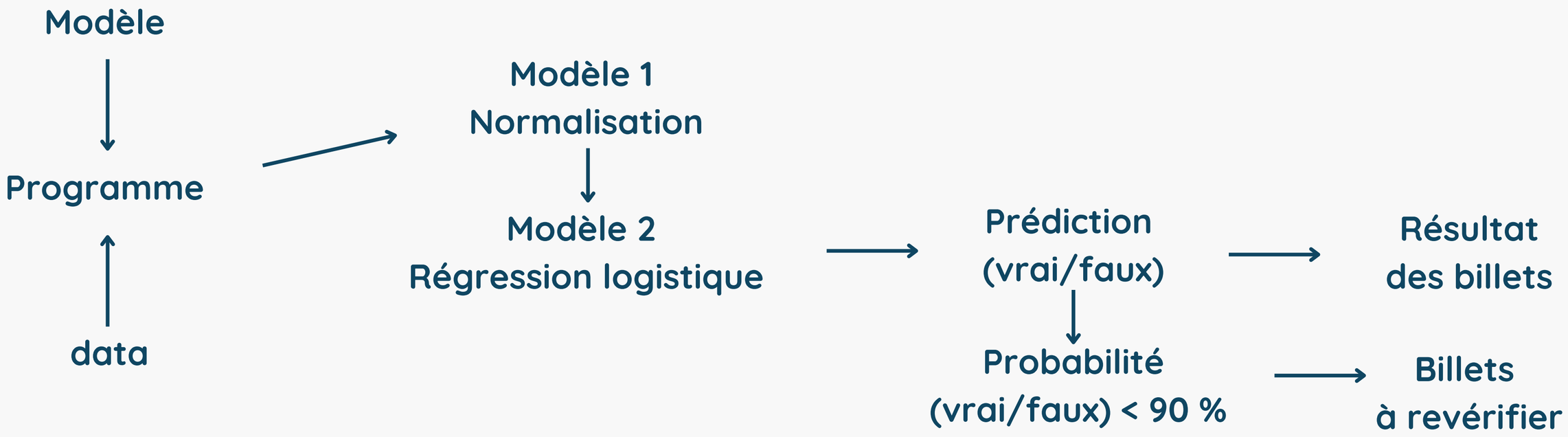
Enregistrement des modèles d’entraînement



Modèle de régression logistique

Programme de détection de faux billets

Fonctionnement



Merci