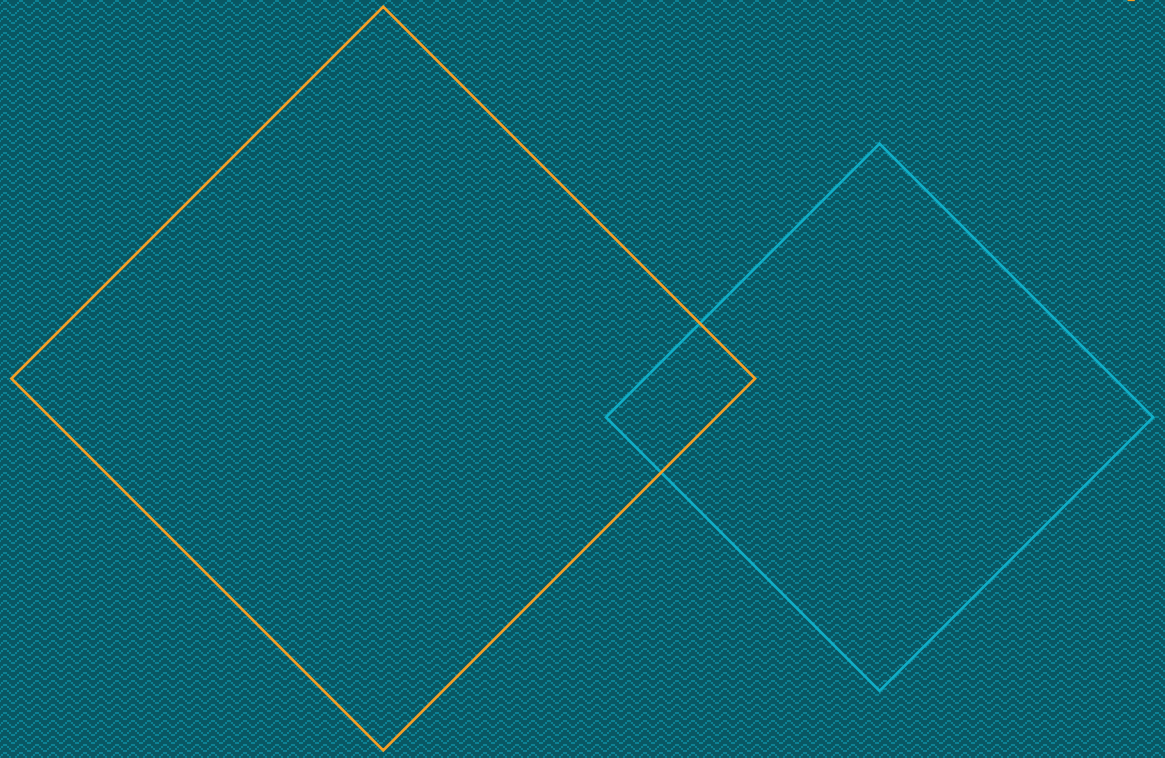
Détection de faux billets

Création d'un programme



Marc Sellam / DA / Projet 6 / juin 2020

Analyse du projet

Introduction ?

?

Introduction

La PJ nous a transmis un jeu de données contenant les caractéristiques géométriques de billets de banque.

Nous allons créer un programme capable d'effectuer une prédiction sur un billet, c'est-à-dire de déterminer s'il s'agit d'un vrai ou d'un faux billet. Pour chaque billet, votre programme devra donner la probabilité que le billet soit vrai si cette probabilité est supérieure ou égale à 0.5,

Les données

Description des données

Le jeu de données fourni est un tableau de 170 lignes, chacune représentant un billet. Chaque ligne est composées de 7 colonnes:

-6 colonnes numériques, contenant les caractéristiques géométriques des billets.

-1 une colonne contenant la nature Vrai/Faux des billets.

	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	171.67	103.74	103.70	4.01	2.87	113.29
2	True	171.83	103.76	103.76	4.40	2.88	113.84
3	True	171.80	103.78	103.65	3.73	3.12	113.63
4	True	172.05	103.70	103.75	5.04	2.27	113.55
...	...	...	...	...	...	...	...
165	False	172.11	104.23	104.45	5.24	3.58	111.78
166	False	173.01	104.59	104.31	5.04	3.05	110.91
167	False	172.47	104.27	104.10	4.88	3.33	110.68
168	False	171.82	103.97	103.88	4.73	3.55	111.87
169	False	171.96	104.00	103.95	5.63	3.26	110.96

170 rows × 7 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 170 entries, 0 to 169
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   is_genuine    170 non-null    bool
 1   diagonal      170 non-null    float64
 2   height_left   170 non-null    float64
 3   height_right  170 non-null    float64
 4   margin_low    170 non-null    float64
 5   margin_up     170 non-null    float64
 6   length        170 non-null    float64
dtypes: bool(1), float64(6)
memory usage: 8.3 KB
```

Aucune valeur manquante.

Les valeurs

- ‘Length’ représente la longueur du billet (en mm) :
`moyenne: 112.57 médiane: 112.84 écart-type: 0.92 , Min:109.97 Max:113.98.`

- ‘height_left’ représente la hauteur du billet (côté gauche, en mm) :
`moyenne: 104.07 médiane: 104.06 écart-type: 0.3 , Min:103.23 Max:104.86.`

- ‘height_right’ représente la hauteur du billet (côté droit, en mm) :
`moyenne: 103.93 médiane: 103.95 écart-type: 0.33 ,Min:103.14 Max:104.95.`

- ‘margin_up’ représente la marge entre le bord supérieur du billet et l'image de celui-ci (en mm):
`moyenne: 3.17 médiane: 3.17 écart-type: 0.24 ,Min:2.27,Max:3.68.`

- ‘margin_low’ représente la marge entre le bord inférieur du billet et l'image de celui-ci (en mm):
`moyenne: 4.61 médiane: 4.45 écart-type: 0.7 , Min:3.54 Max:6.28.`

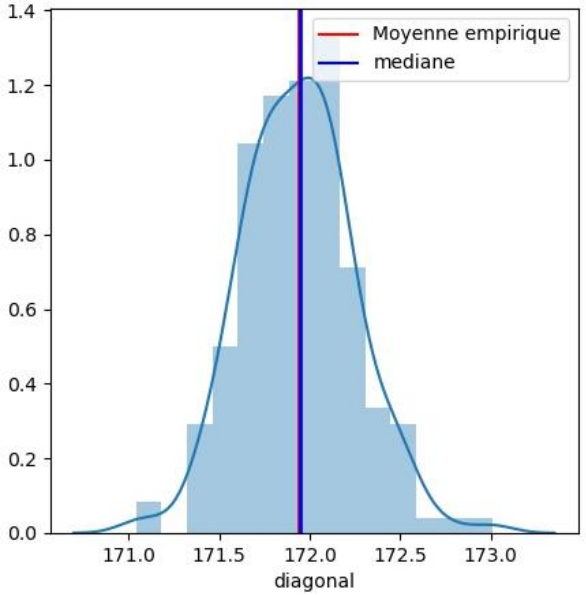
- ‘diagonal’ représente la diagonale du billet (en mm) :
`moyenne: 171.94 médiane: 171.94 écart-type: 0.31 , Min:171.04 Max:173.01.`

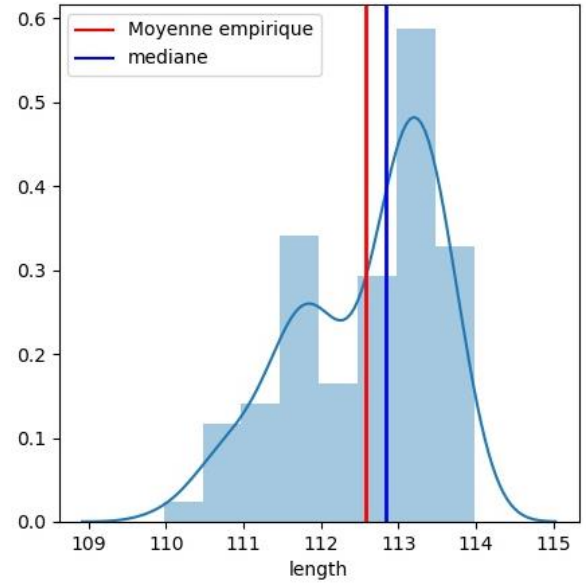
- ‘is_genuine’ représente la colonne donnant la nature Vrai/Faux du billet:
`‘True’: 100 ‘False’: 70`

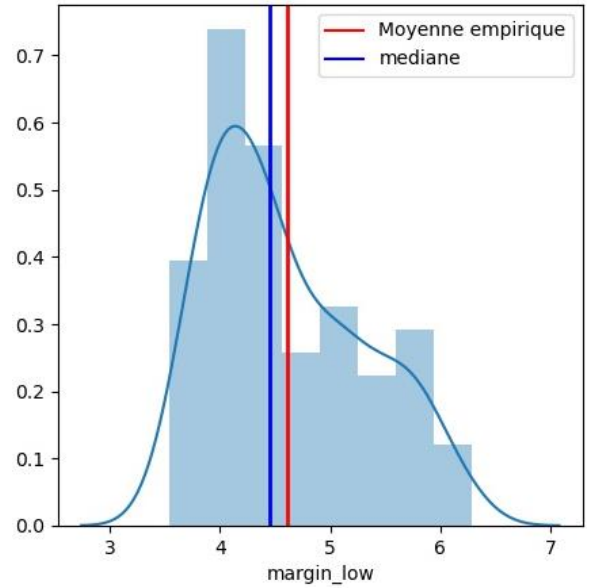
Aucune valeur aberrante.

Analyses graphiques

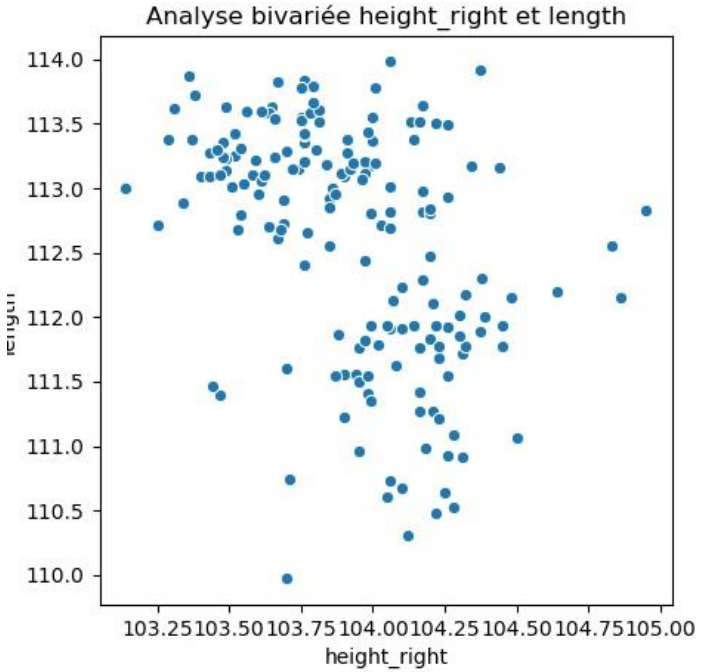
Analyse univariée :

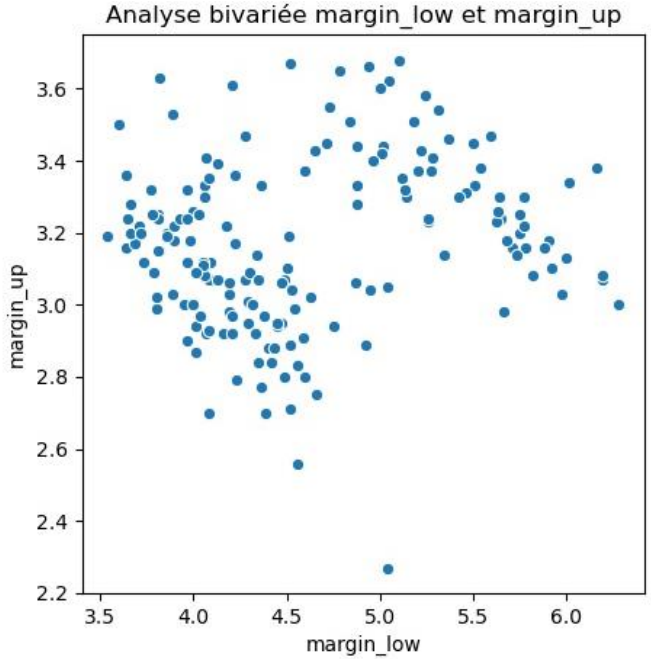






Analyse bivariée :



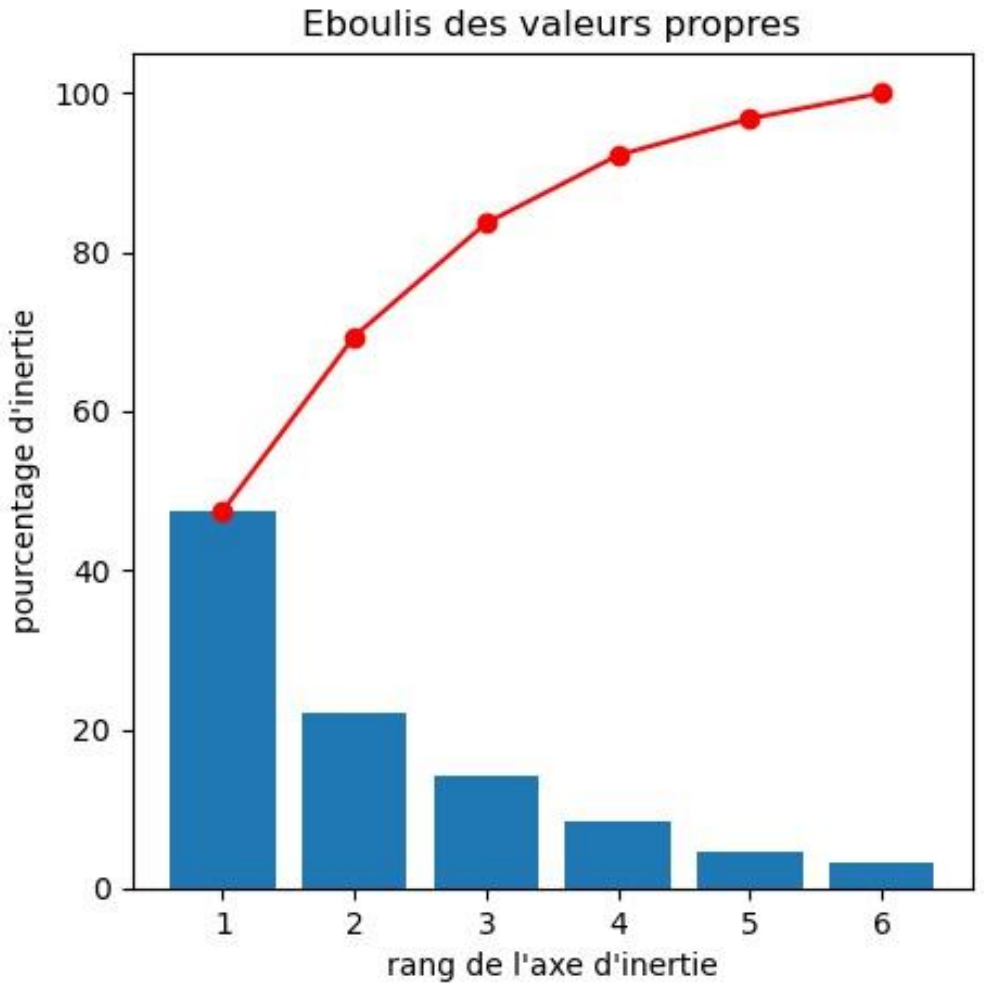


Deux groupes se distinguent sous certaines conditions.

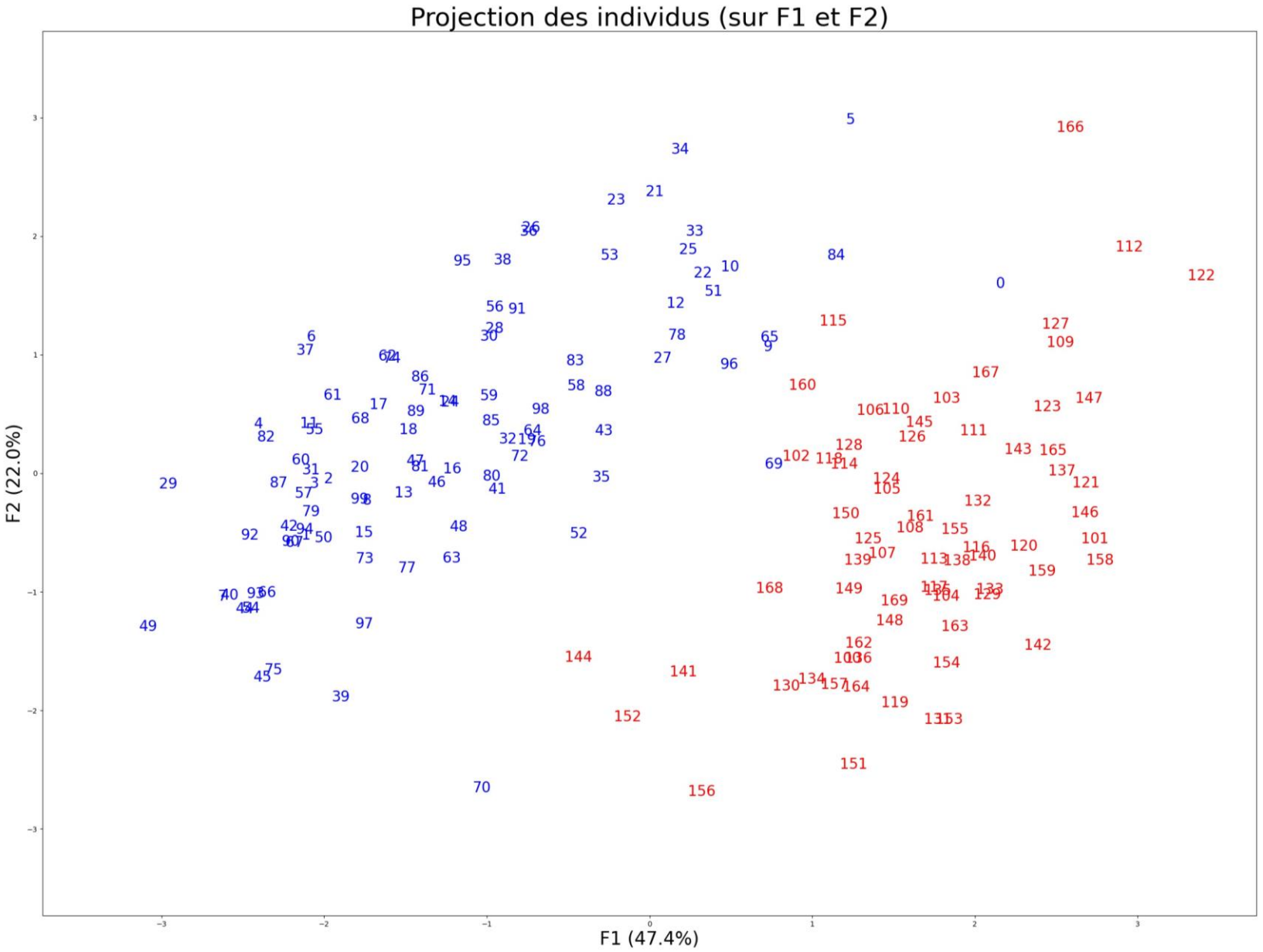
Analyse en composantes principales

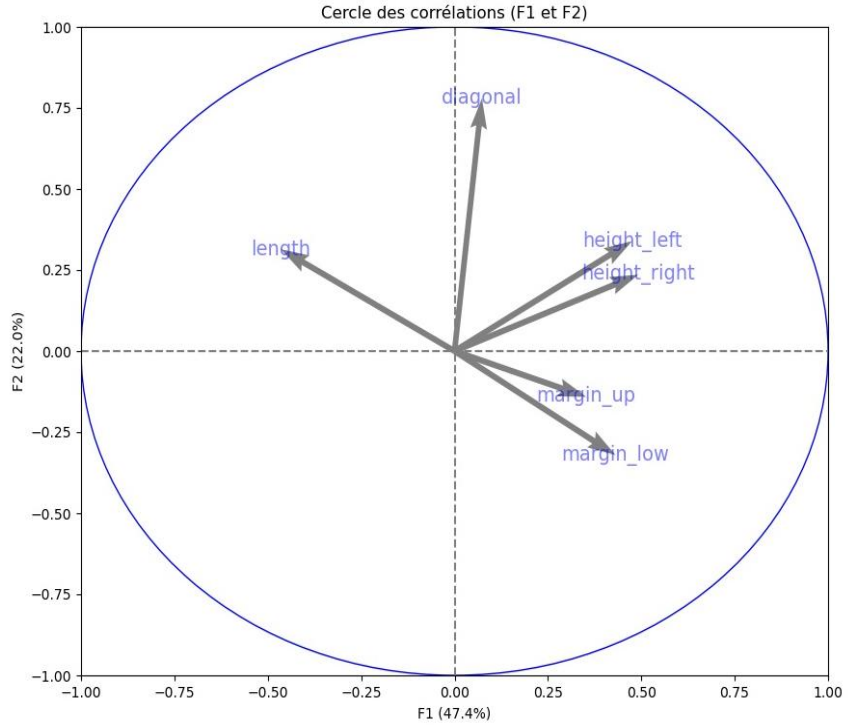
Analyse des données

Le cumul de la variance expliqué nous indique que nous disposons de 47.44 % de l'information disponible sur la première composante et de 69.39% avec les deux premiers facteurs



Cercle des corrélations



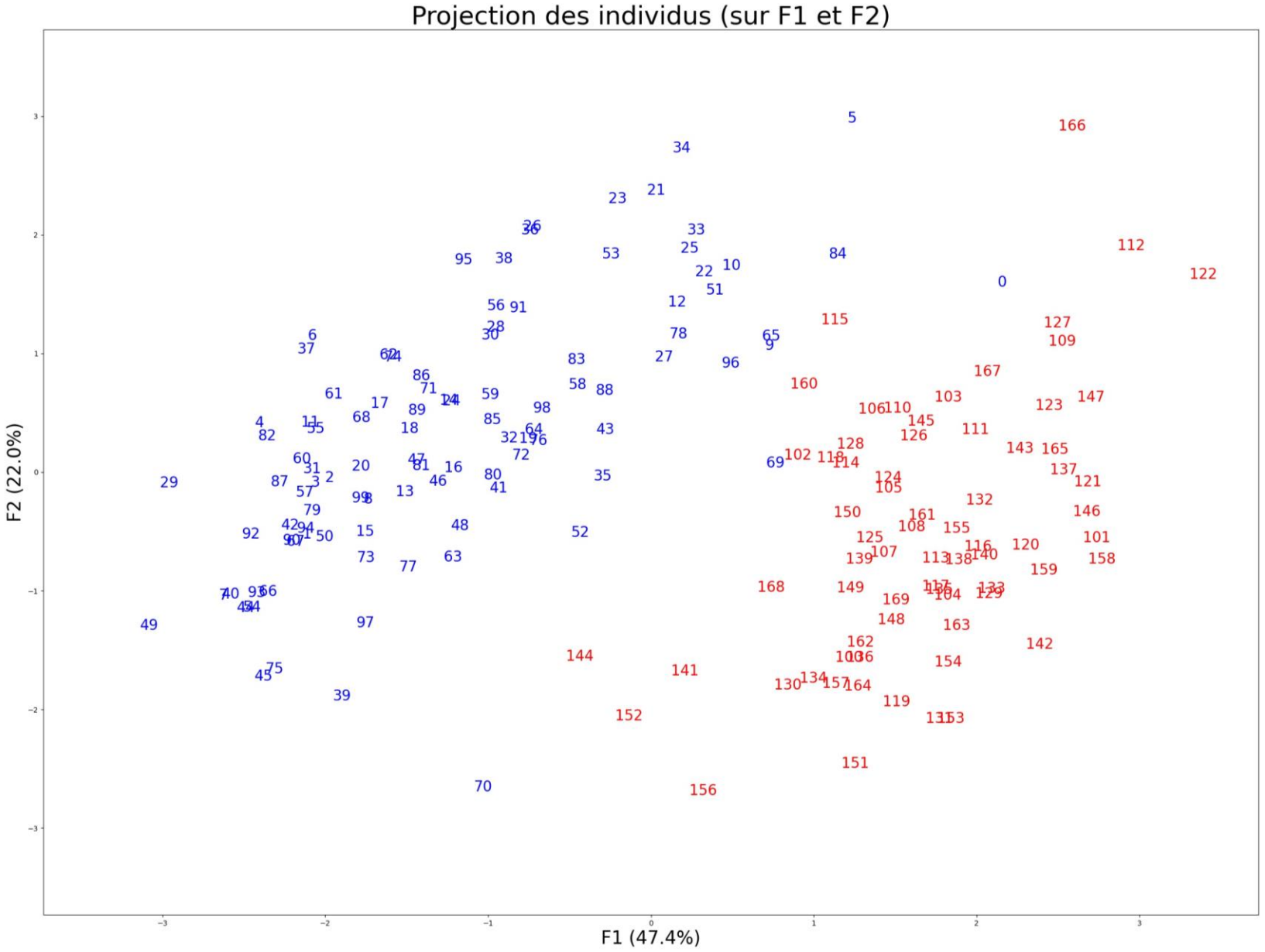


Les variables les plus corrélées a F1 sont ‘height_left/height_right/margin_up /margin_low (positivement) et lenght(correlee négativement a F1).

La variable ‘diagonal’ est corrélée à F2.

On remarque ‘Length’ et ‘margin low’ en opposition(correlation negative),.

Contribution des individus à l'inertie totale



	contrib_it	is_genuine
64	0.813929	True
98	0.878319	True
85	1.431664	True
76	1.435202	True
59	1.484454	True
...	...	...
39	15.052608	True
122	16.790944	False
4	18.039567	True
0	18.410598	True
166	20.625650	False

Les valeurs les plus hautes sont aux extrémités alors que les plus basses sont plutôt vers le centre des vrais billets

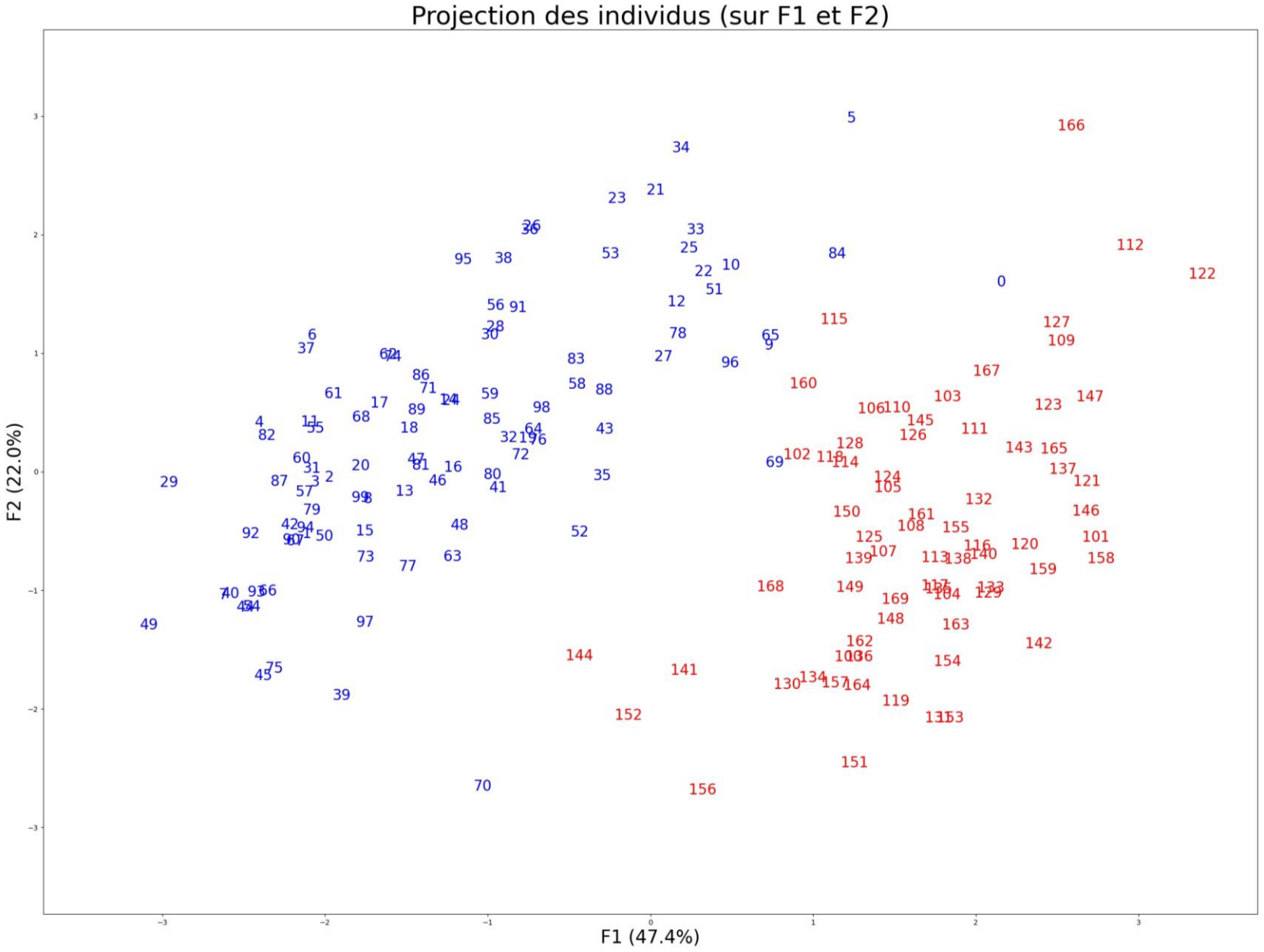
Qualité de représentation des individus par axe

	COS2_1	COS2_2	COS2_F1F2	is_genuine
35	0.024786	0.000346	0.025132	True
43	0.038755	0.063289	0.102043	True
160	0.068742	0.042901	0.111643	False
88	0.023273	0.135767	0.159040	True
52	0.076673	0.106134	0.182807	True
...	...	...	...	...
154	0.542732	0.417831	0.960564	False
45	0.631308	0.331305	0.962613	True
46	0.962296	0.003343	0.965639	True
143	0.973456	0.007584	0.981040	False
148	0.578193	0.410807	0.989001	False

69.45 % est le pourcentage moyen de la qualité de représentation des faux billets.

68.18 % est le pourcentage moyen de la qualité de représentation des vrais billets.

Contribution des individus à l'inertie d'un axe



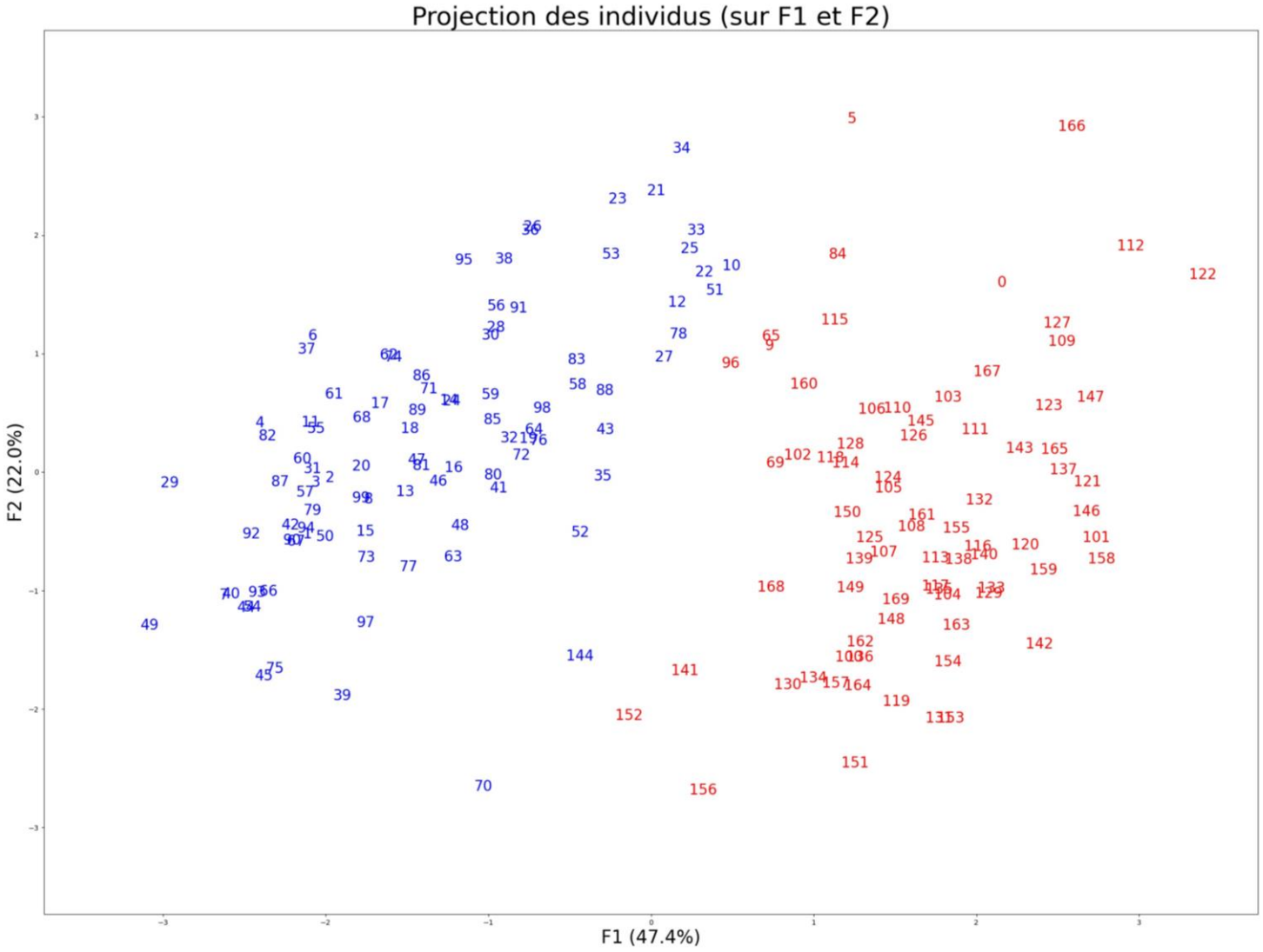
	CTR_1	is_genuine
21	0.000002	True
27	0.000014	True
152	0.000038	False
12	0.000053	True
78	0.000059	True
...	...	...
158	0.015743	False
112	0.017845	False
29	0.017982	True
49	0.019505	True
122	0.023618	False

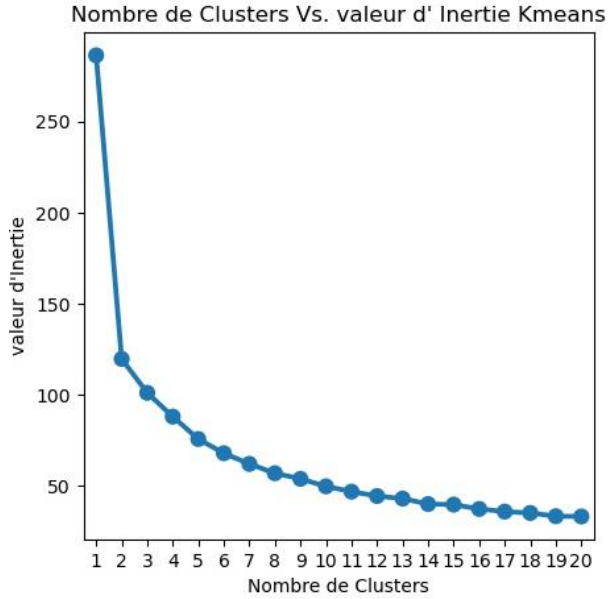
	CTR_2	is_genuine
137	0.000002	False
80	0.000003	True
31	0.000003	True
35	0.000005	True
16	0.000006	True
...	...	...
70	0.031331	True
156	0.032074	False
34	0.033106	True
166	0.037704	False
5	0.039503	True

Les 5 valeurs les plus fortes sont déterminants pour le premier axe, on les retrouve aux extrémités de celui-ci.
Pareil pour le second axe.

Classification par l'algorithme KMeans

Classification par KMeans

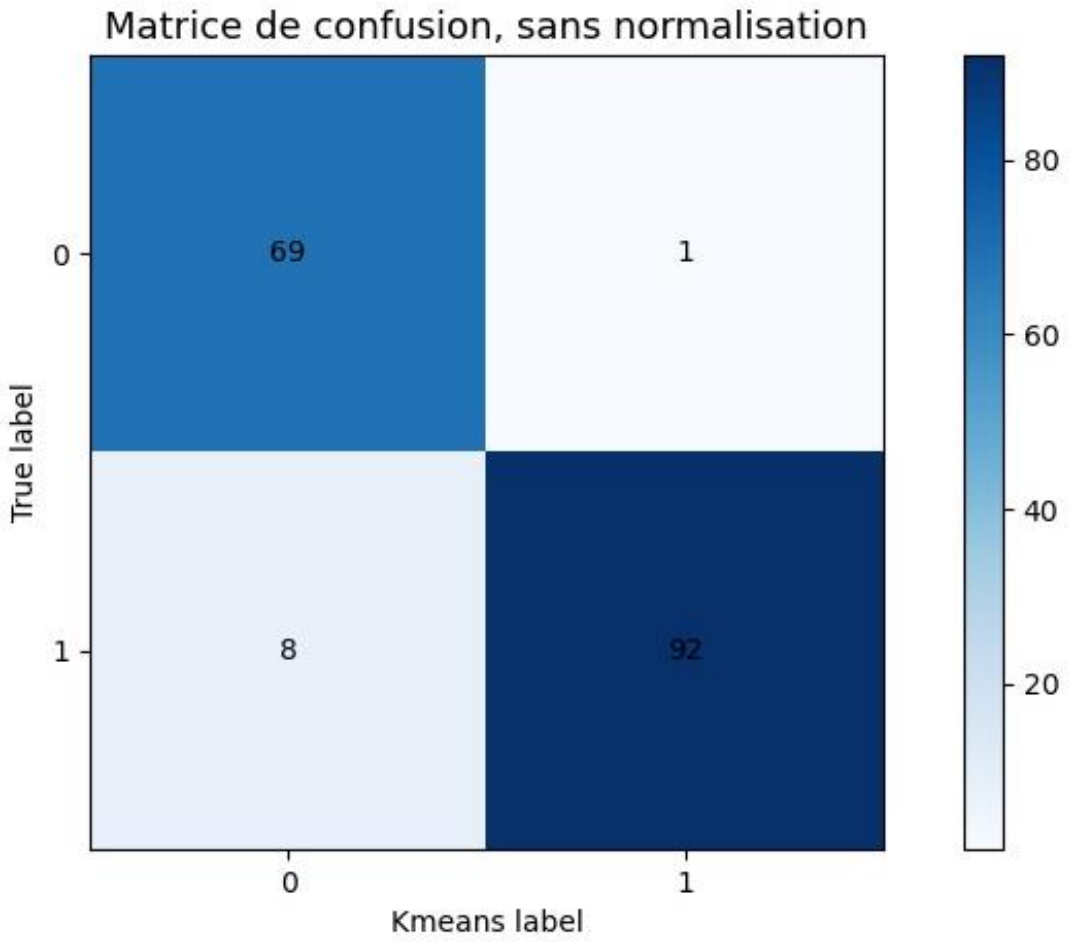




Le coude nous confirme qu'une clustérisation par K = 2 est le nombre de groupes idéal.

Résultats de l'algorithme KMeans

Pourcentage de similitude de KMeans comparé aux résultats réels : 94.70 %



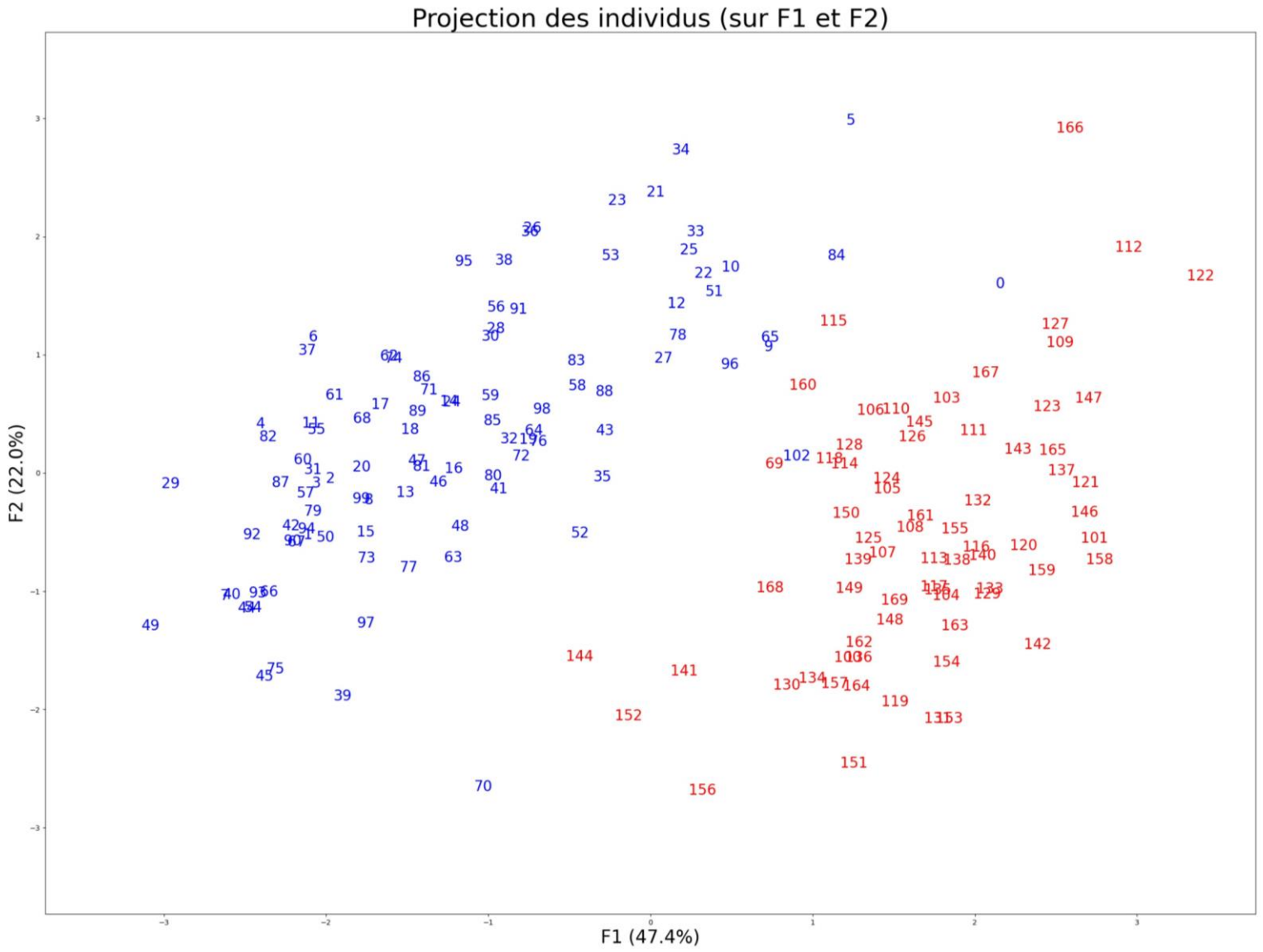
Les 9 erreurs de l'algorithme Kmeans :

	diagonal	height_left	height_right	margin_low	margin_up	length
0	171.81	104.86	104.95	4.52	2.89	112.83
5	172.57	104.65	104.44	4.54	2.99	113.16
9	172.14	104.34	104.20	4.63	3.02	112.47
10	172.27	104.29	104.22	3.89	3.53	113.50
65	172.16	104.43	104.06	4.51	3.19	112.69
69	171.94	104.11	104.16	4.08	3.35	111.76
84	172.30	104.58	104.17	4.36	3.33	112.98
96	172.00	104.32	104.26	4.53	3.04	112.93
144	171.56	103.80	103.87	5.66	2.98	112.95

L'algorithme KMeans découpe clairement la projection des individus en deux parties distinctes.

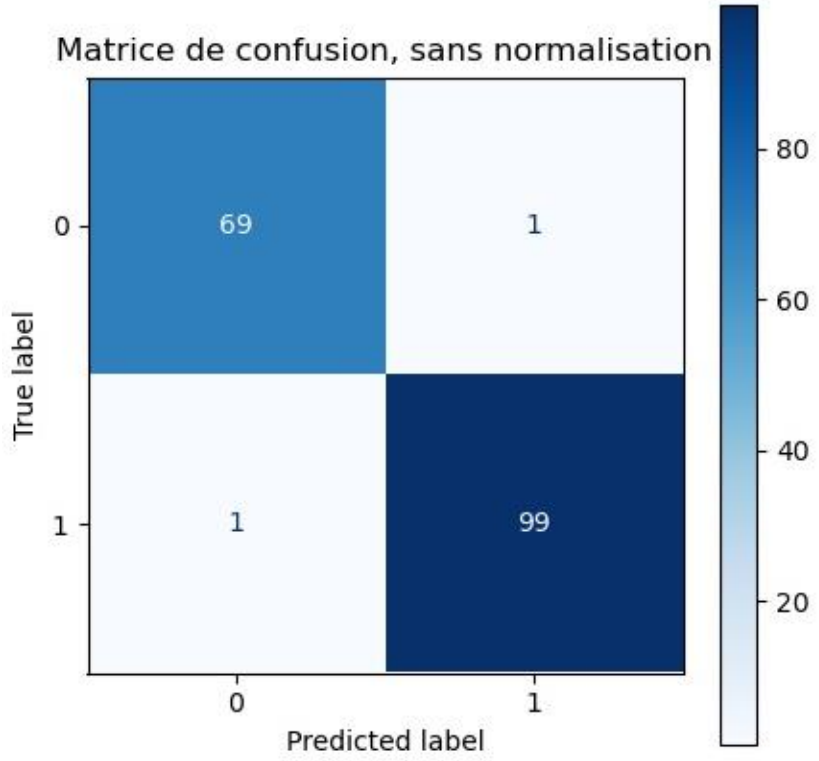
Projection des individus classifiés par régression logistique

Création d'un algorithme de classification



Résultats de l'algorithme de régression logistique

La précision après test de l'algorithme de régression logistique sur les données d'entrainement : 98.82 %



Les 2 erreurs de l'algorithme de régression logistique :

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
69	1	171.94	104.11	104.16	4.08	3.35	111.76
102	0	171.94	104.21	104.10	4.28	3.47	112.23

Le billet 69 prédit comme étant faux est une erreur commune aux 2 algorithmes

Test sur le fichier ‘example.csv’.

Test de l’algorithme de classification

```
Nom du fichier d'entrainement (sans l'extension .csv) : notes
Pourcentage de précision après test de l'algorithme sur les données d'entrainement :  98.82 %.
Nom du fichier à traiter (sans l'extension .csv) : example
Ajout de la colonne 'prediction', 1 pour un vrai billet, 0 pour un faux billet.
Résultats de la verification de vos billets:
40.0 pourcents des billets sont vrais.
```

	diagonal	height_left	height_right	margin_low	margin_up	length	id	vrai_%	prediction
0	171.76	104.01	103.54	5.21	3.30	111.42	A_1	4.00	0
1	171.87	104.17	104.13	6.00	3.31	112.09	A_2	0.91	0
2	172.00	104.58	104.29	4.99	3.39	111.57	A_3	2.77	0
3	172.49	104.55	104.34	4.44	3.03	113.20	A_4	87.74	1
4	171.65	103.63	103.56	3.77	3.16	113.33	A_5	99.64	1

Conclusion

Apres une analyse des données fournies, il s'est avéré que les vrais billets avaient tendance a êtres plus long que les faux, et les faux billets comme ayant une marge entre le bord inférieur du billet et l'image supérieure plus importante que les vrais billets.

Deux algorithmes aux caractéristiques différentes ont été utilisés afin de déterminer si un billet est vrai ou faux à partir des données fournies :

Un algorithme de classification non supervisé, Kmeans, qui nous a fourni un pourcentage de similitude de 94.70 % lors d'un essai sur les données fournies.

Un algorithme de classification supervisé , la régression logistique, qui nous a fourni une précision de 98.82 % lors d'un essai sur les données fournies.

Un programme de prédiction a été mis au point à partir de la modélisation des données fournies à l'aide de la régression logistique, programme qui a partir de caractéristiques géométriques prédéfinies déterminera la probabilité qu'un billet soit vrai.

Merci