Projet 6 : Détecteur de faux billets

Sommaire

- Contexte et données
- Analyse de la distribution des classes dans les variables
- Valeurs des variables
- Analyse ACP
- Méthode de clustering Kmeans
- Analyse des poids des variables suite à la régression logistique
- Modèle prédictif
- Conclusion

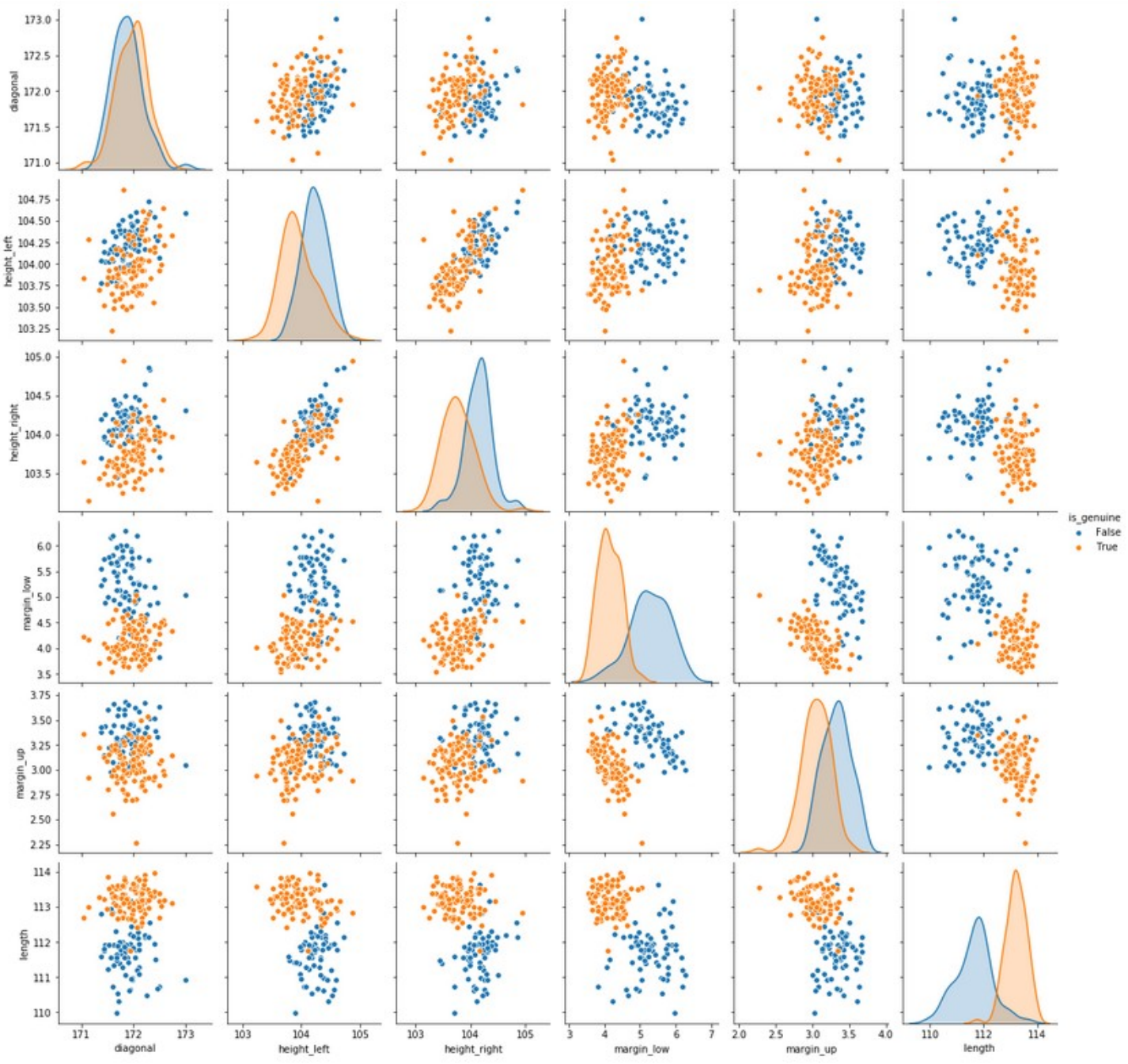
Contexte et données

Créer un détecteur de faux billet à partir d'un modèle prédictif

Données :

Table	Dimension	Description
Donnees	170X7	Chaque ligne correspond à un billet, on a une colonne booléenne qui définit sur le billet est vrai ou faux, le reste des colonnes sont quantitatives et définissent différentes dimensions du billet

Analyse de la distribution des classes dans les variables



On distingue deux variables discriminantes :

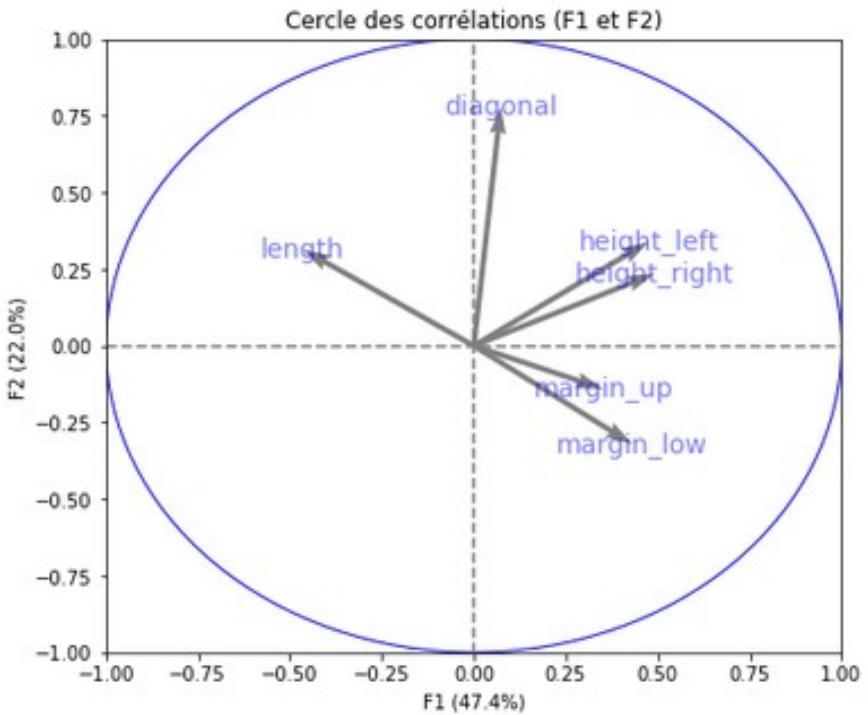
Length
Margin_low

Valeurs des variables

is_genuine	diagonal		height_left		height_right		margin_low		margin_up		length	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
False	171,89	171,88	104,23	104,22	104,15	104,17	5,28	5,27	3,33	3,34	111,66	111,77
True	171,98	172,01	103,95	103,92	103,78	103,76	4,14	4,08	3,06	3,07	113,21	113,21
% écart True/False	0,05 %	0,08 %	-0,27 %	-0,29 %	-0,35 %	-0,39 %	**-21,55 %**	**-22,51 %**	**-8,37 %**	**-7,95 %**	**1,38 %**	**1,29 %**

Confirmation du pairplot sur : « lenght » et « margin_low » + nouvelle : « margin_up »

Analyse ACP



ACP sur les variables :

- Composante 1 caractérisée par lenght, height_right, height_left, margin_low, lenght
- Composante 2 caractérisée par diagonal

Margin_up étant mal représenté par les2 autres composantes on en récupère une troisième

:

	id	CORR_1	CORR_2	CORR_3	COS2_1	COS2_2	COS2_3	CTR_1	CTR_2	CTR_3
0	diagonal	0.073275	0.779639	-0.088894	0.015286	0.800781	0.006749	0.005369	0.607837	0.007902
1	height_left	0.475502	0.339250	0.122760	0.643685	0.151624	0.012871	0.226102	0.115091	0.015070
2	height_right	0.491821	0.235543	0.153830	0.688626	0.073091	0.020211	0.241888	0.055480	0.023664
3	margin_low	0.431027	-0.320537	0.512808	0.528904	0.135358	0.224597	0.185784	0.102744	0.262972
4	margin_up	0.352540	-0.141120	-0.821149	0.353822	0.026236	0.575888	0.124284	0.019915	0.674285
5	length	-0.465373	0.314536	0.126913	0.616553	0.130337	0.013757	0.216572	0.098933	0.016107

Analyse ACP individus

Analyse de l'ACP sur individus :

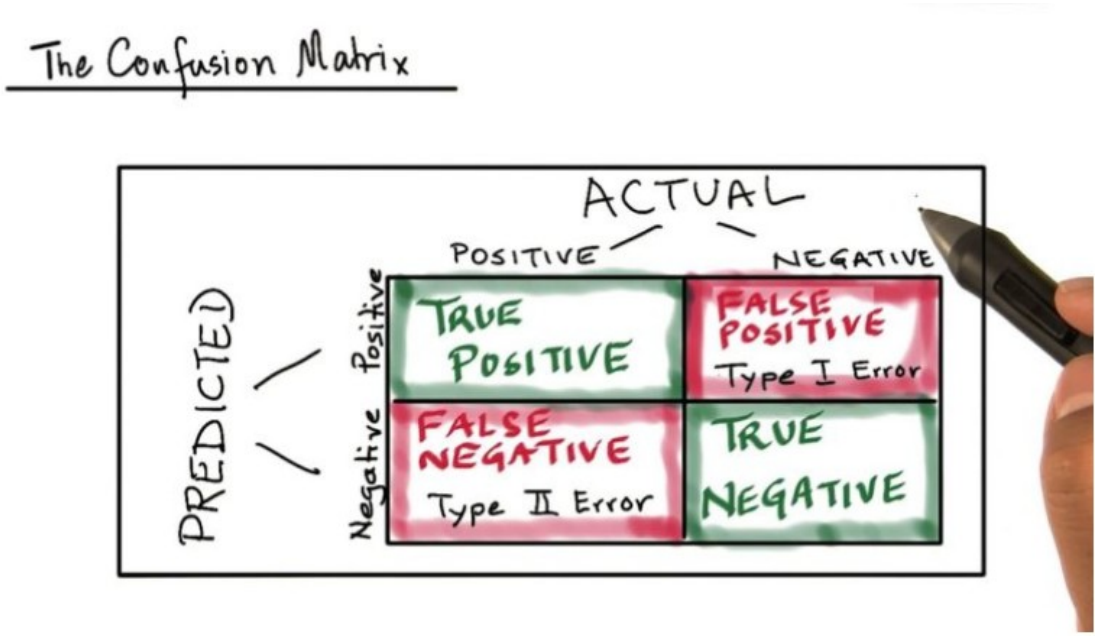
- Le premier plan factoriel arrive bien a dissocié deux amas, par analyse croisée du cercle des corrélations :
  - Une longueur plus élevée caractériserait les vrais billets
  - Une marge haute et basse plus élevées caractériseraient les faux billets
- Comparaison 5 plus fortes contributions PCA1 vs 5 plus faibles : height_left, height_right, margin_low inférieurs et lenght supérieure
- Comparaison 5 plus fortes qualités représentation sur PCA1 : height_left, height_right, margin_low inférieurs et lenght supérieure

CTR :

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length	Coord_C1	Coord_C2	Coord_C3	CTR	Cos2_1	Cos2_2	Cos2_3	Cum_Cos2_1	CTR_1	CTR_2	CTR_3	rank_CTR_1	rank_Cos2_1
49	True	171,59	103,23	103,64	4,01	2,94	113,59	-3,081	-1,295	0,125	12,918	73,509	12,981	0,120	86,490	0,020	0,007	0,000	2	34
29	True	171,84	103,75	103,38	4,08	2,7	113,72	-2,959	-0,092	1,051	10,118	86,527	0,083	10,913	86,610	0,018	0,000	0,008	3	12
7	True	171,58	103,65	103,37	3,54	3,19	113,38	-2,624	-1,040	-1,069	9,344	73,681	11,578	12,232	85,259	0,014	0,005	0,008	10	32
40	True	171,51	103,85	103,36	4,49	2,8	113,87	-2,582	-1,032	1,152	9,977	66,804	10,681	13,291	77,485	0,014	0,005	0,009	12	53
44	True	171,79	103,51	103,25	4,05	3,08	112,71	-2,488	-1,145	-0,579	8,783	70,504	14,921	3,823	85,425	0,013	0,006	0,002	16	42

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length	Coord_C1	Coord_C2	Coord_C3	CTR	Cos2_1	Cos2_2	Cos2_3	Cum_Cos2_1	CTR_1	CTR_2	CTR_3	rank_CTR_1	rank_Cos2_1
122	False	172,29	104,72	104,86	5,71	3,16	112,15	3,391	1,665	1,385	16,791	68,479	16,503	11,425	84,982	0,024	0,012	0,013	1	47
112	False	172,32	104,6	104,83	4,84	3,51	112,55	2,947	1,908	-0,489	14,422	60,235	25,249	1,658	85,484	0,018	0,016	0,002	4	68
158	False	171,84	104,32	104,5	6,28	3	111,06	2,768	-0,737	2,008	12,725	60,231	4,265	31,695	64,496	0,016	0,002	0,028	5	69
101	False	171,97	104,38	104,18	5,59	3,47	110,98	2,737	-0,557	-0,308	8,249	90,831	3,761	1,151	94,592	0,015	0,001	0,001	6	3
147	False	172,25	104,52	104,22	4,65	3,43	110,48	2,702	0,631	-0,931	10,500	69,529	3,793	8,262	73,322	0,015	0,002	0,006	7	44

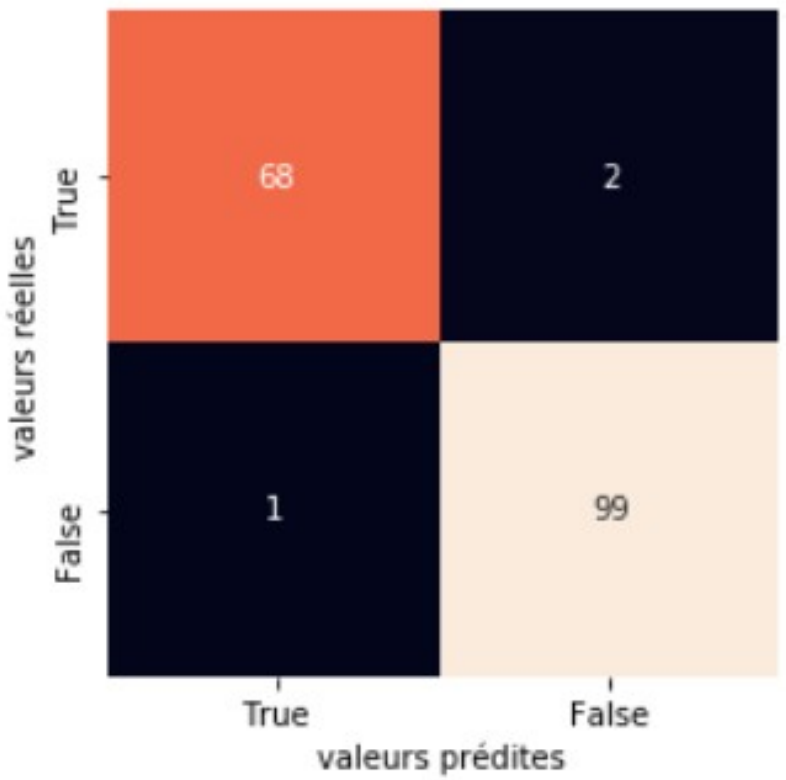
Matrice de confusion



- La matrice de confusion permet de vérifier nos modèles prédictifs
- Compare nos valeurs prédites, aux valeurs réelles
- Nous comparons 4 valeurs : vrai positif, faux positif, faux négatif et vrai négatif

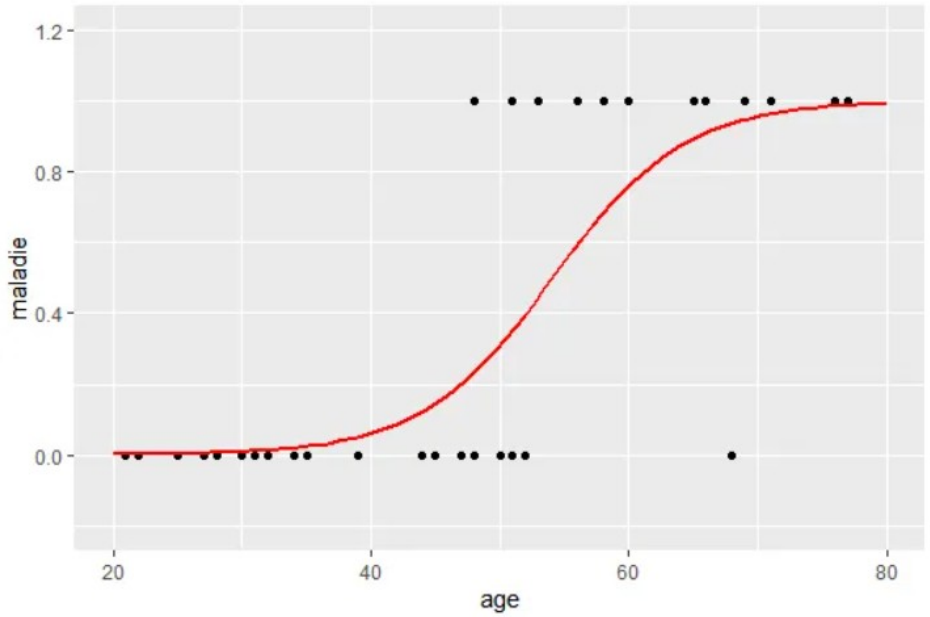
Méthode de clustering Kmeans

Projection du Kmeans sur le premier plan factoriel, très proche de l'original



La matrice de confusion à partir du Kmeans confirme la projection. Les clusters du Kmeans reflètent la classification des vrais vs faux billets.

Déterminer un modèle à l'aide d'une régression logistique



- Méthode de classification binaire à partir de variables explicatives
- Permet de calculer la probabilité d'appartenance à une classe parmi 2 classes binaires

Analyse des poids des variables suite à la régression logistique

	coef
constante	-0.247670
diagonal	-3.361121
height_left	-17.240598
height_right	5.039367
margin_low	-65.162074
margin_up	-100.007209
length	21.989445

Nous avons deux variables ayant un faible poids :

- « Diagonal »
- « Height_right »

J’ai décidé de ne pas prendre en compte ces variable de par leur faible représentation

Modèle prédictif

```
Meilleur(s) hyperparamètre(s) sur le jeu d'entraînement:
{'logit__C': 0.1}
Résultats de la validation croisée :
accuracy = 0.994 (+/-0.024) for {'logit__C': 0.1}
accuracy = 0.994 (+/-0.024) for {'logit__C': 1.0}
accuracy = 0.994 (+/-0.024) for {'logit__C': 10}
accuracy = 0.988 (+/-0.029) for {'logit__C': 100}
```

Nous avons utiliser une régression logistique avec l'algorithme « libelinear »

Ajustement avec Gridsearch et Pipeline :
Régularisation à 0,1 pour une précision de 0,994

Conclusion

- En moyenne, les vrais billets sont identifiables par une longueur plus élevée et une marge inférieure plus faible
- Cette différence entre les deux classes se retrouve dans le Kmeans
- 4 variables sur 6 ont vraiment un poids représentatif
- Modèle prédictif élaboré efficace et vérifié via la méthode kfolds avec une précision de 0,99