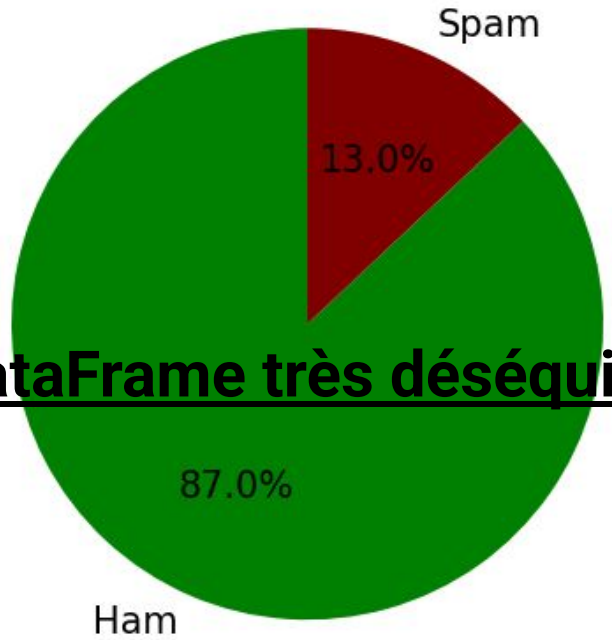
Spam Classifier

Mieux vaut déclasser le classé ou classer le déclassé ?

Chevallier Dylan et Proust Maximilien

5574 messages

4827 messages



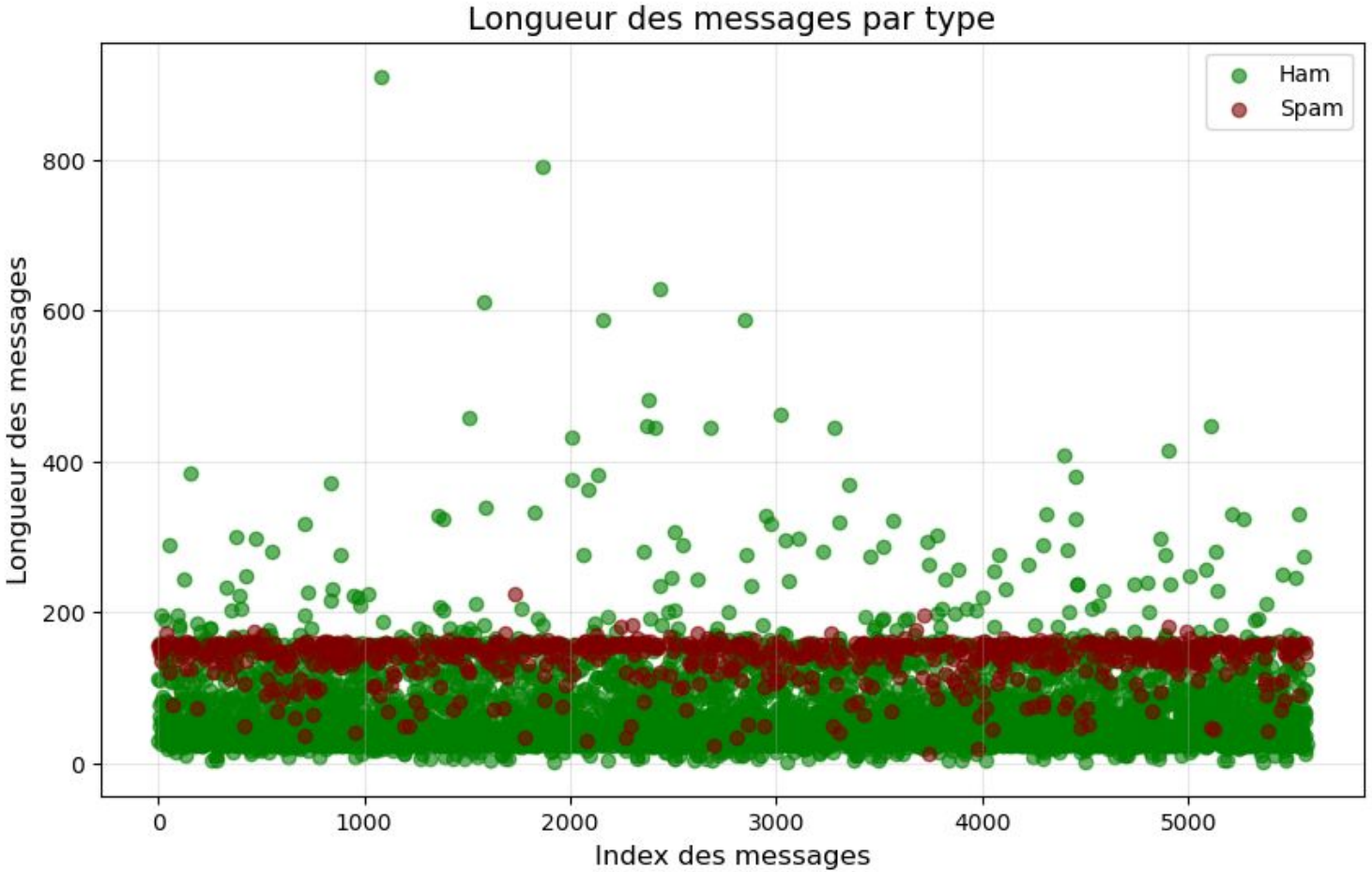
747 messages

DataFrame très déséquilibré !

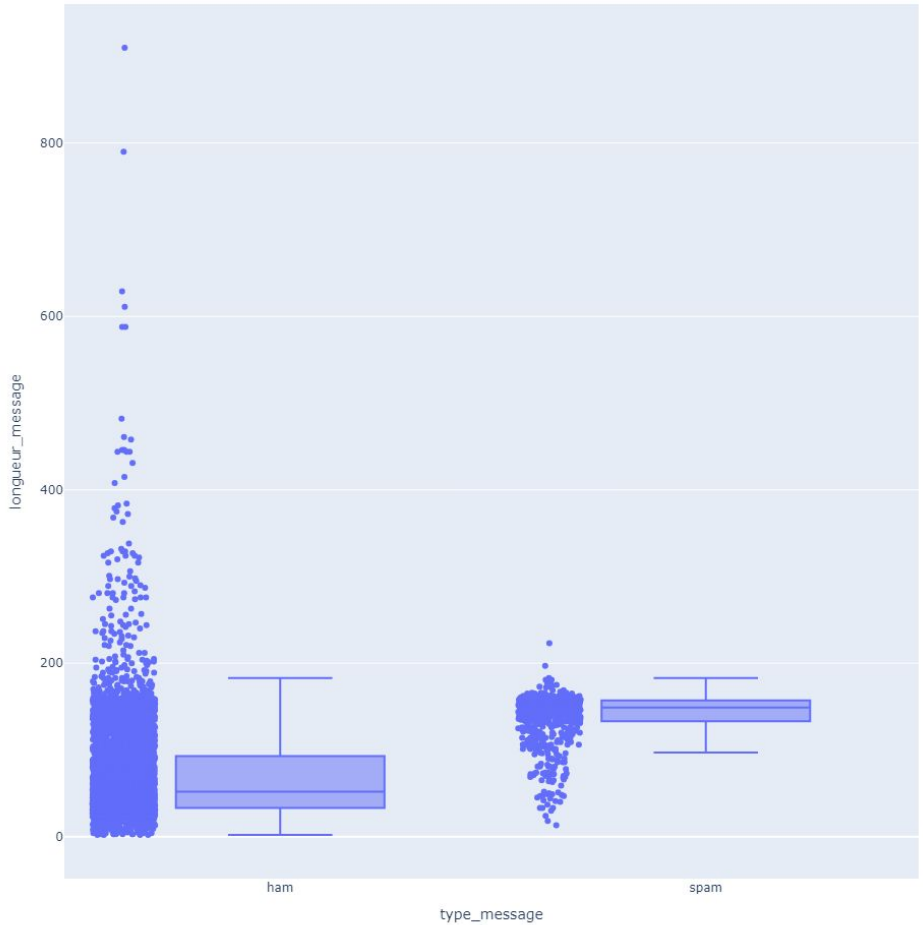
Quelles caractéristiques pour notre projet ?

- Nombre de caractères par message
- Nombre de mots par type de message
- Redondance des mots types dans les spam
- La présence des chiffres dans les messages
- Le nombre de chiffre par message
- Le nombre de caractères spéciaux par type de message

1) Le nombre de caractère par message



1) Le nombre de caractère par message



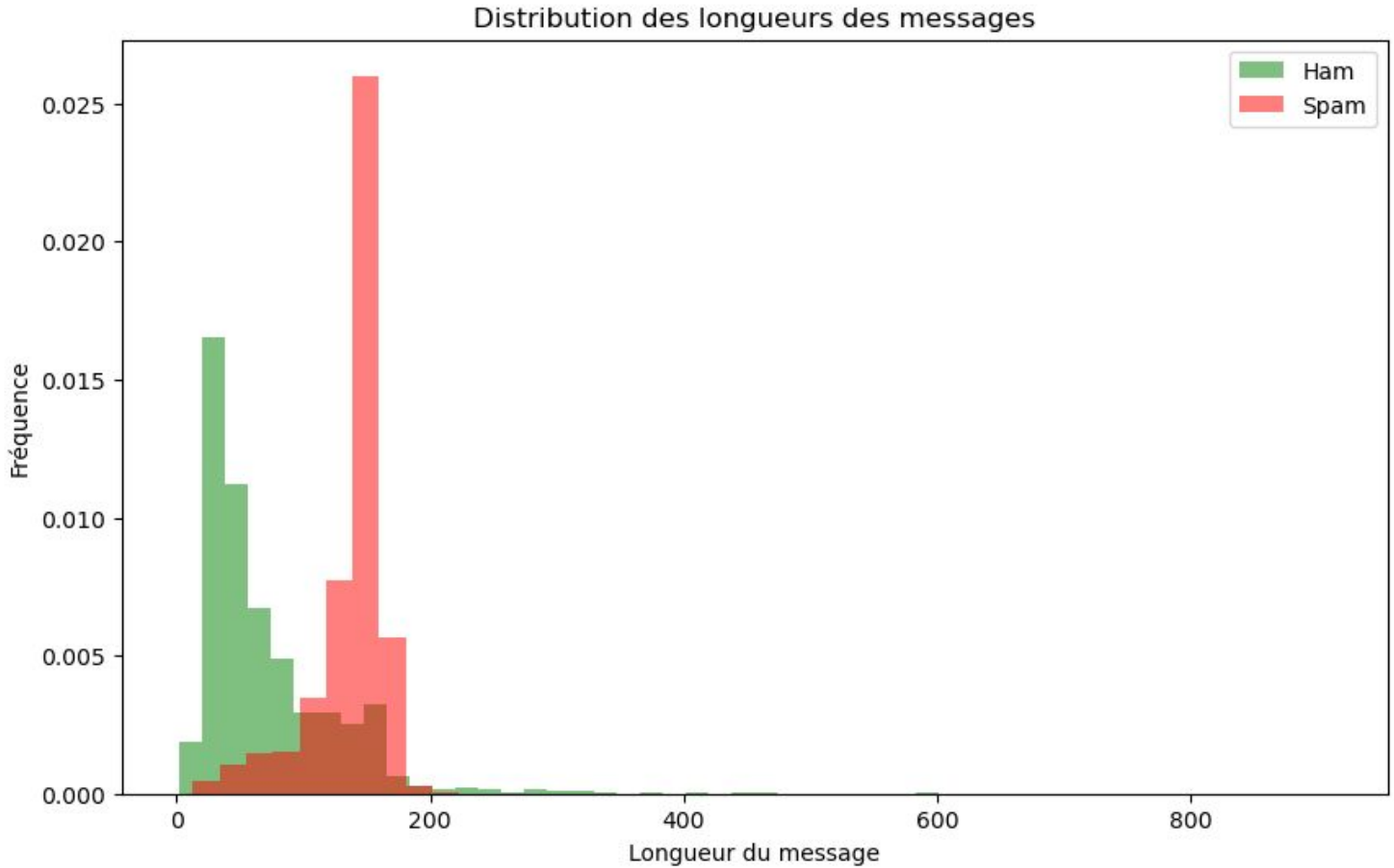
1) Le nombre de caractère par message

1) Le nombre de caractère par message

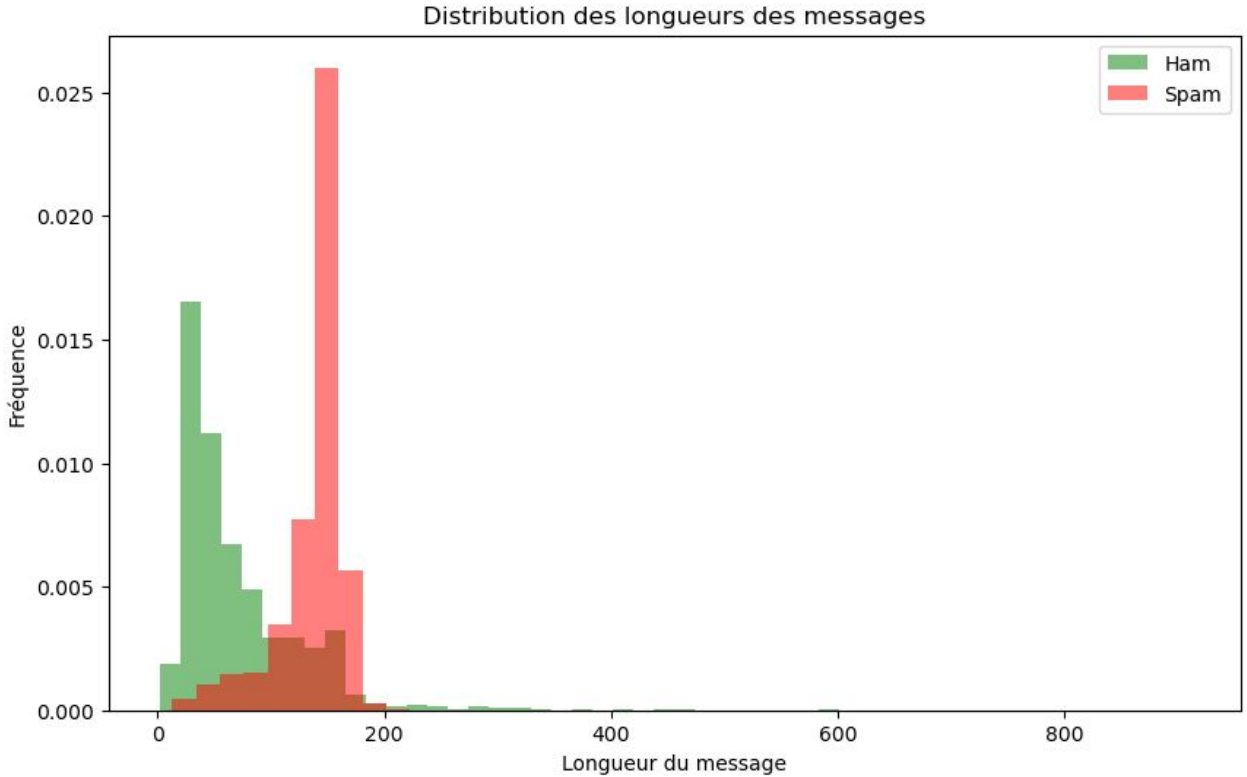
Quelques statistiques supplémentaires...

type_message	count	mean	std	min	25%	50%	75%	max
ham	4827.0	71.47	58.33	2.0	33.0	52.0	93.0	910.0
spam	747.0	138.68	28.87	13.0	133.0	149.0	157.0	223.0

1) Le nombre de caractère par message

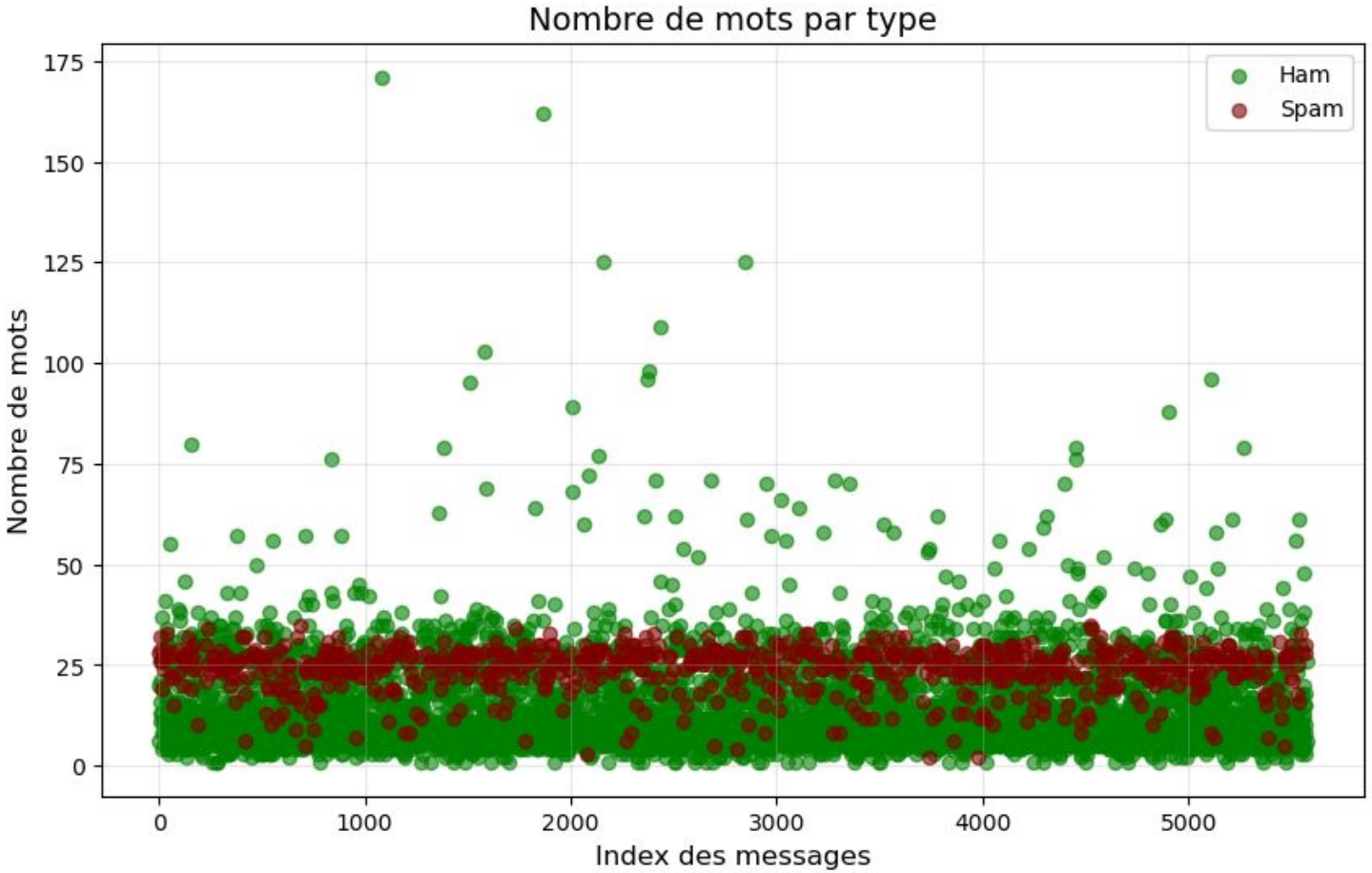


1) Le nombre de caractère par message

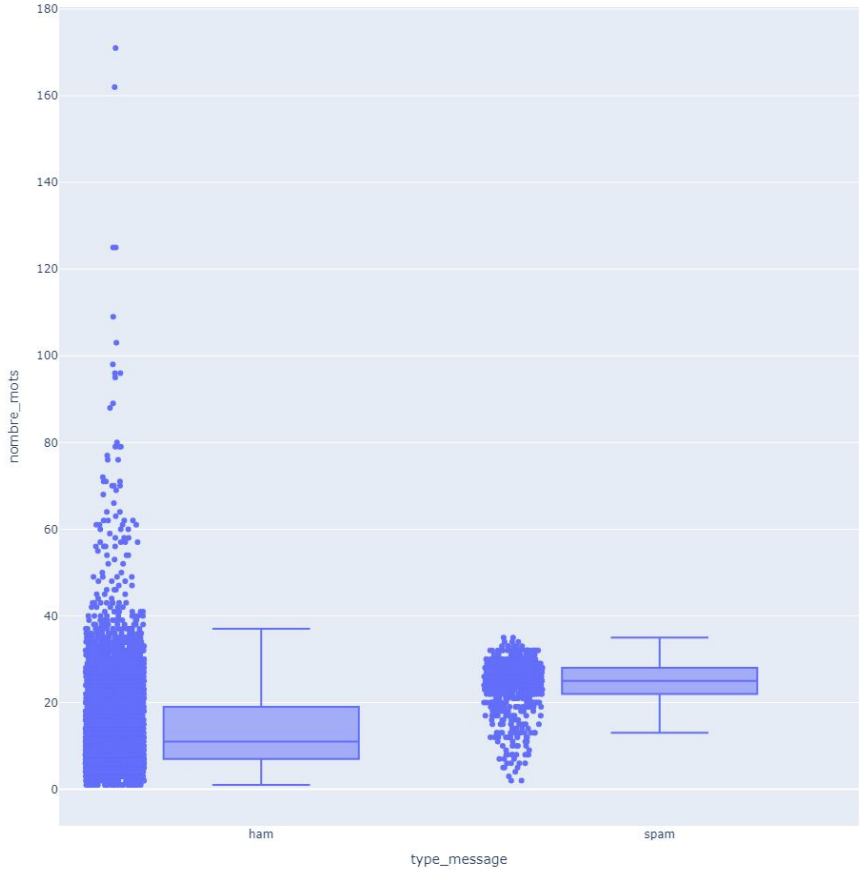


Caractéristique intéressante, on conserve !

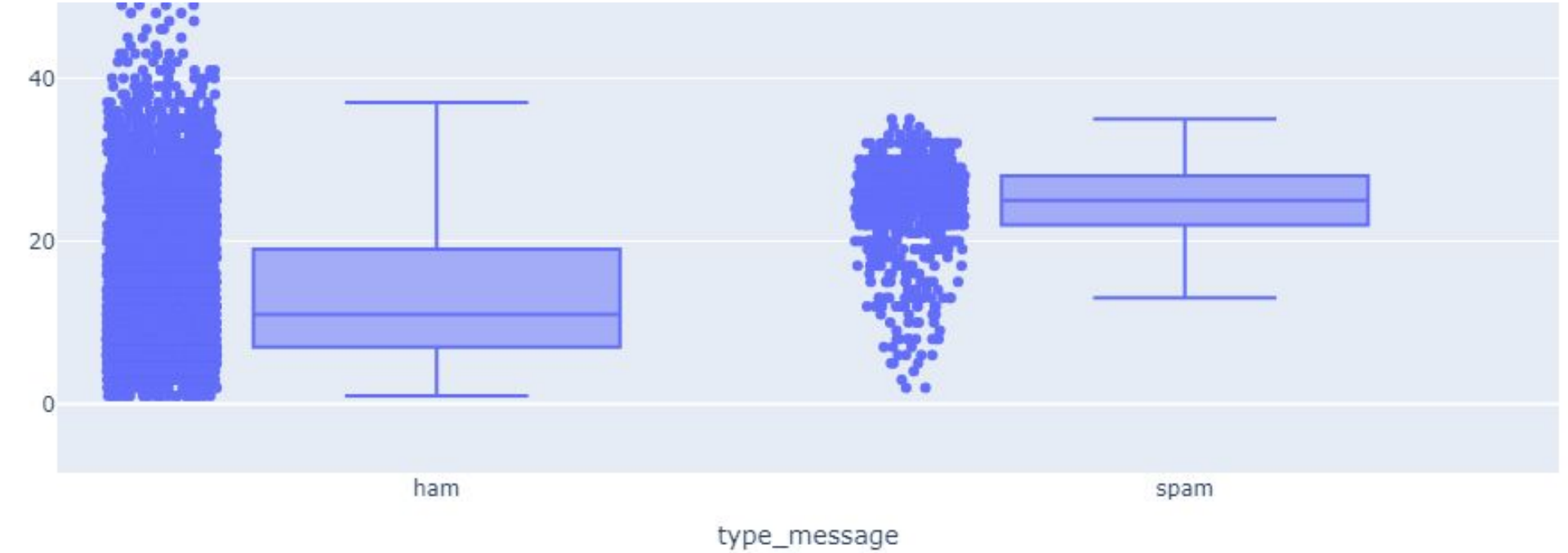
2) Le nombre de mots par message



2) Le nombre de mots par message



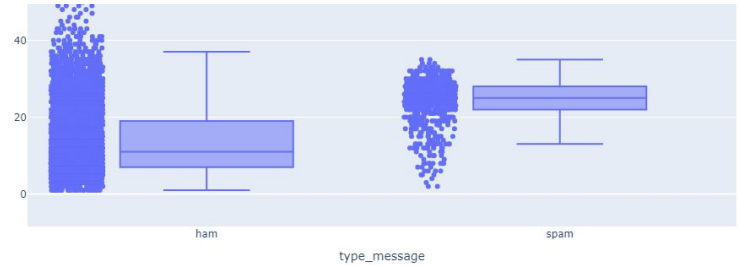
2) Le nombre de mots par message



2) Le nombre de mots par message

Quelques statistiques supplémentaires...

type_message	count	mean	std	min	25%	50%	75%	max
ham	4827.0	14.30	11.50	1.0	7.0	11.0	19.0	171.0
spam	747.0	23.91	5.78	2.0	22.0	25.0	28.0	35.0



2) Le nombre de mots par message

Nombre de mots = Nombre de caractère ?

Pour le vérifier … Test de corrélation !

2) Le nombre de mots par message

Nombre de mots = Nombre de caractère ?

Résultat : 0,97

Quasiment identique !

Choix : Ignorer cette caractéristique pour la suite

3) La redondance des mots présent dans les spams

	Mot	Fréquence
33	to	1538
87	you	1462
29	I	1439
83	the	1029
98	a	977
73	i	742
72	and	739
5	in	736
19	u	651
41	is	645

Pertinent ??

39	my	621
48	me	541
226	of	499
62	for	481
126	that	399
192	it	376
52	your	374
237	on	352
96	have	349
100	at	334

3) La redondance des mots présent dans les spams

	Mot	Fréquence
33	to	1538
87	you	1462
29	I	1439
83	the	1029
98	a	977
73	i	742
72	and	739
5	in	736
19	u	651
41	is	645

Pertinent ??

Pas du tout !

→ **Utilisation des stop words**

39	my	621
48	me	541
226	of	499
62	for	481
126	that	399
192	it	376
52	your	374
237	on	352
96	have	349
100	at	334

3) La redondance des mots présent dans les spams

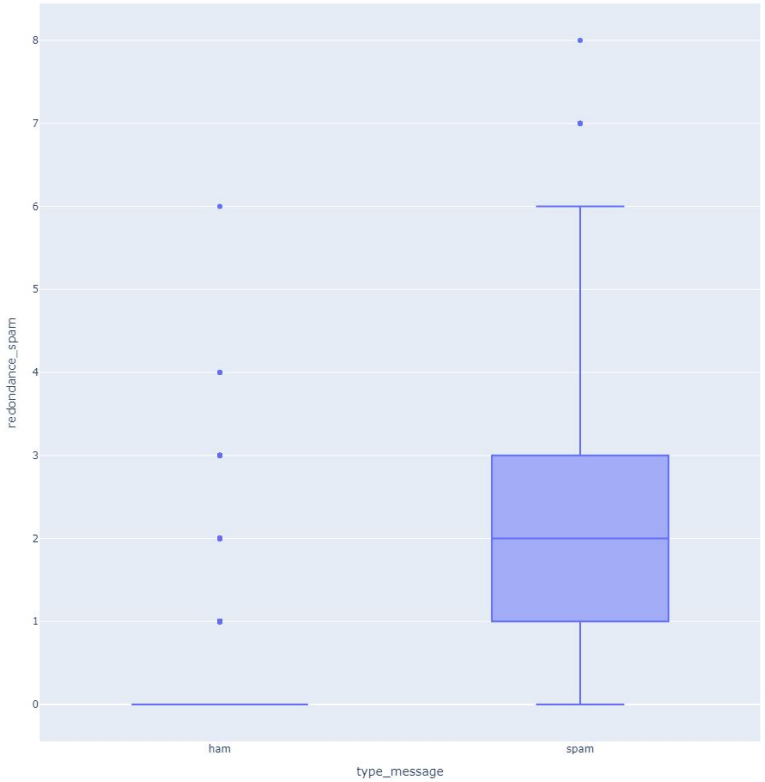
	Mot	Frequence
38	call	187
51	Call	138
53	FREE	115
43	mobile	95
37	claim	78
66	Txt	75
165	text	73
35	prize	73
13	txt	71
265	STOP	63
104	free	62
98	reply	58

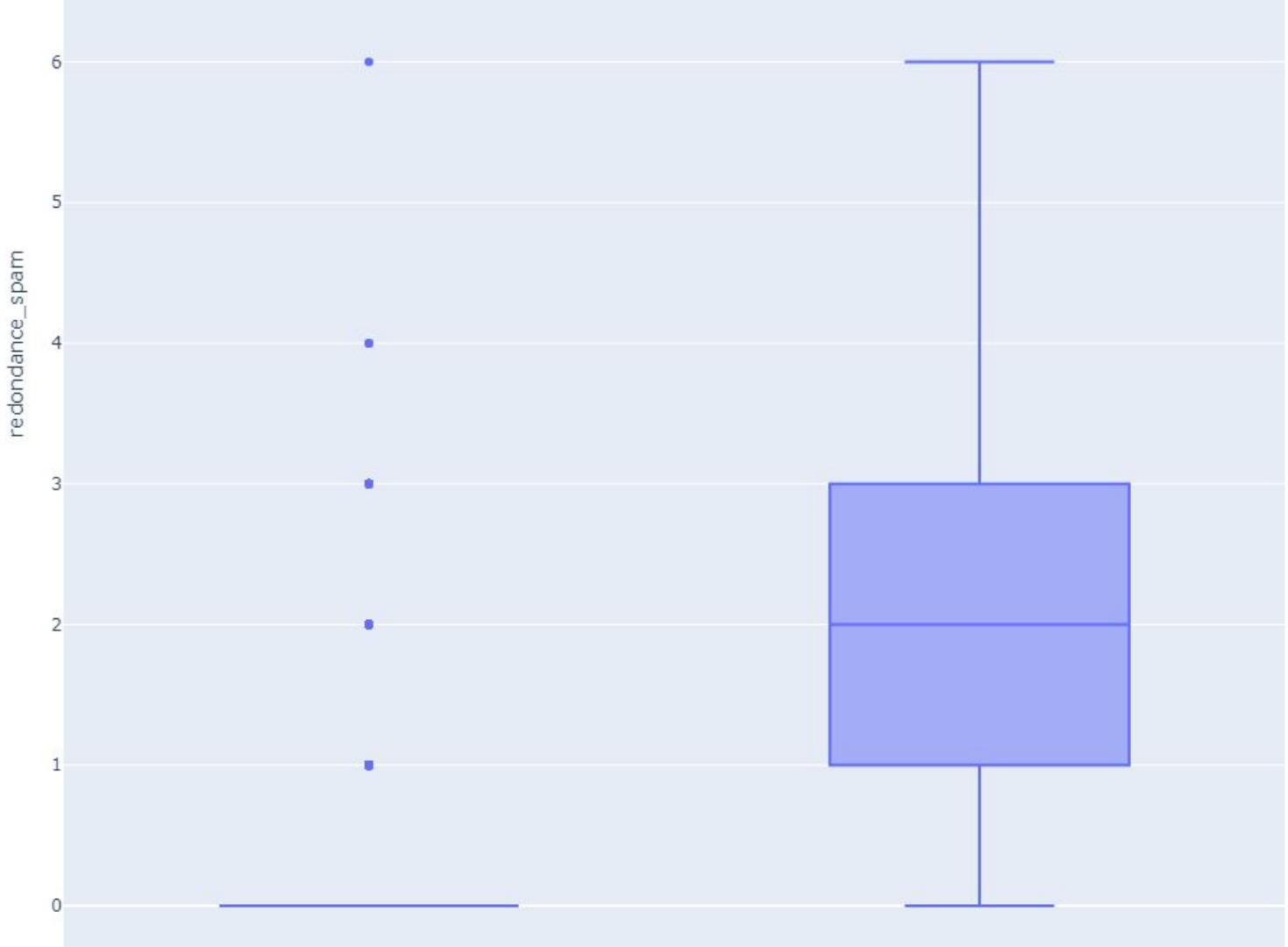
C’est mieux !

Maintenant comptons les !

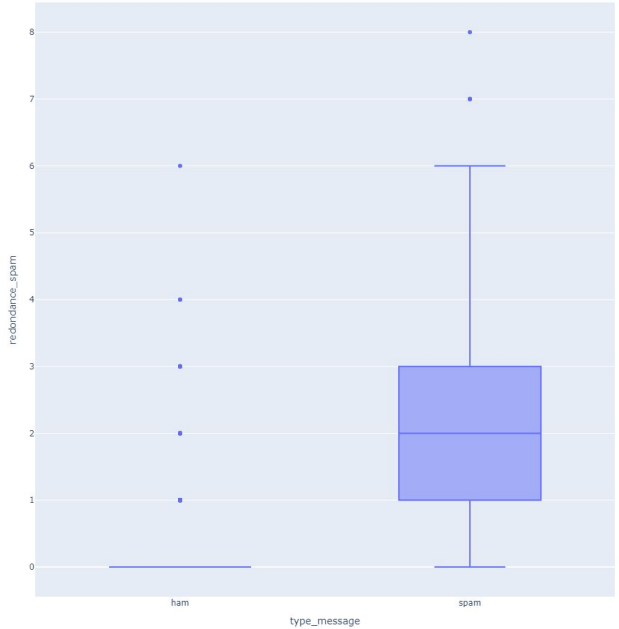
176	contact	56
19	week	52
157	service	49
27	send	47
257	per	46
427	Nokia	46
226	get	45
343	stop	44
60	Reply	44
160	cash	43
0	Free	42
151	new	42
62	URGENT	41
9	Text	40

3) La redondance des mots présent dans les spams



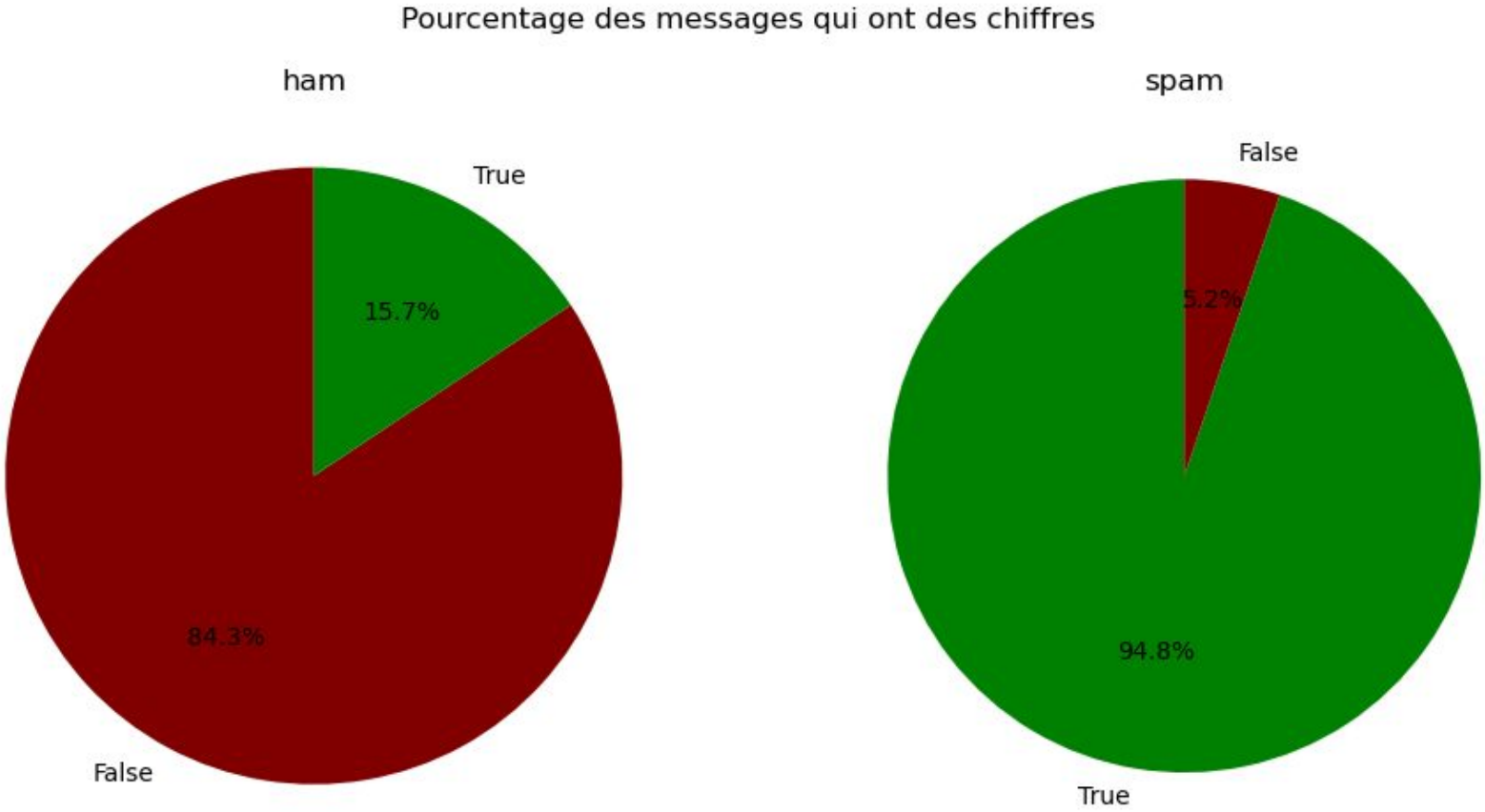

redondance_spam
6
5
4
3
2
1
0

3) La redondance des mots présent dans les spams

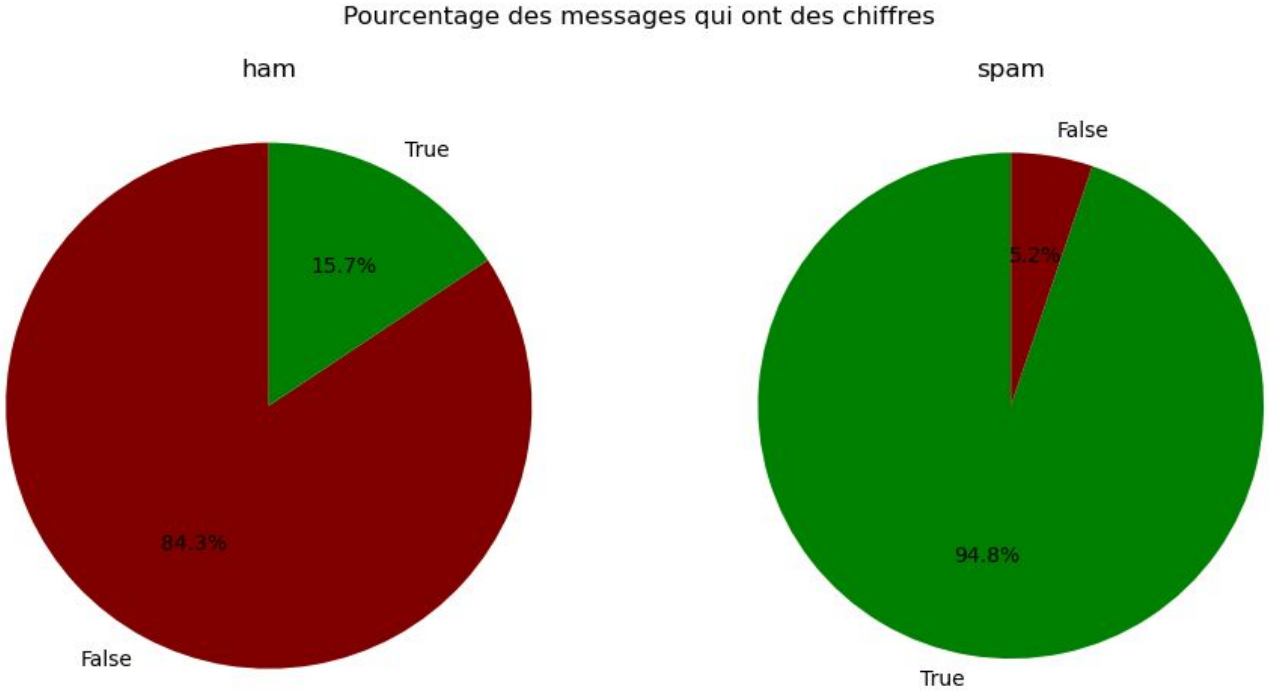


Caractéristique intéressante, on conserve !

4) La présence des chiffres dans les messages

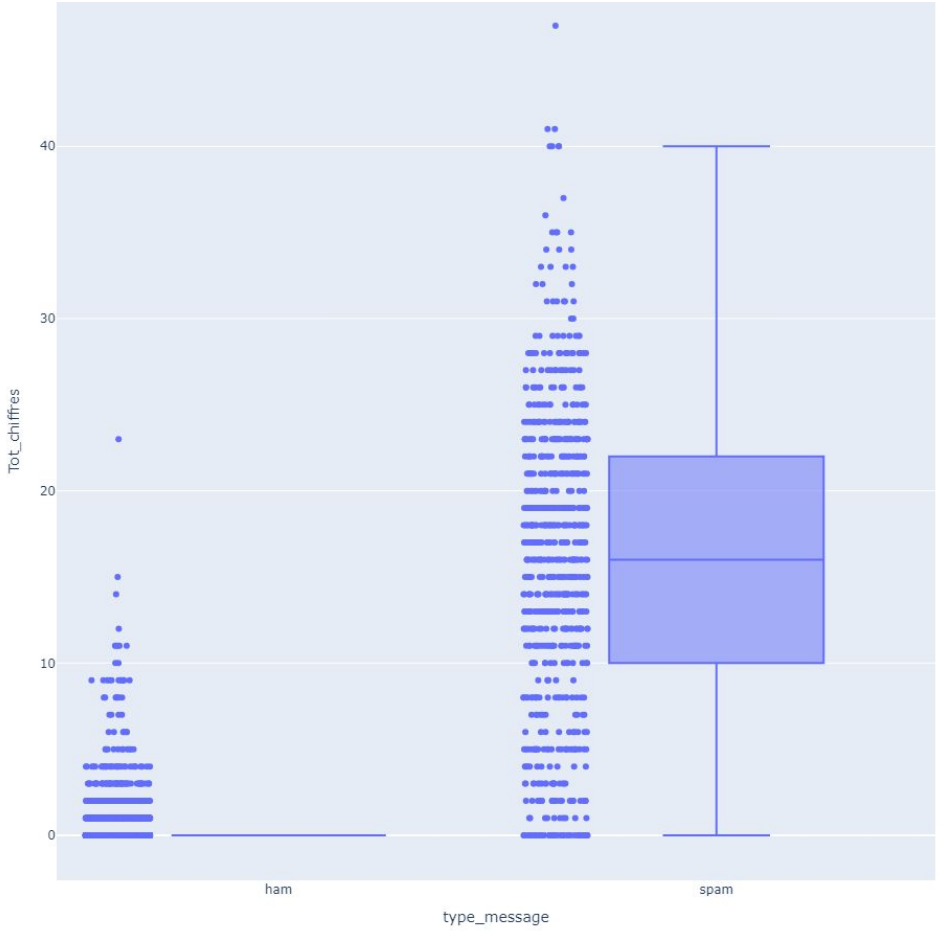


4) La présence des chiffres dans les messages

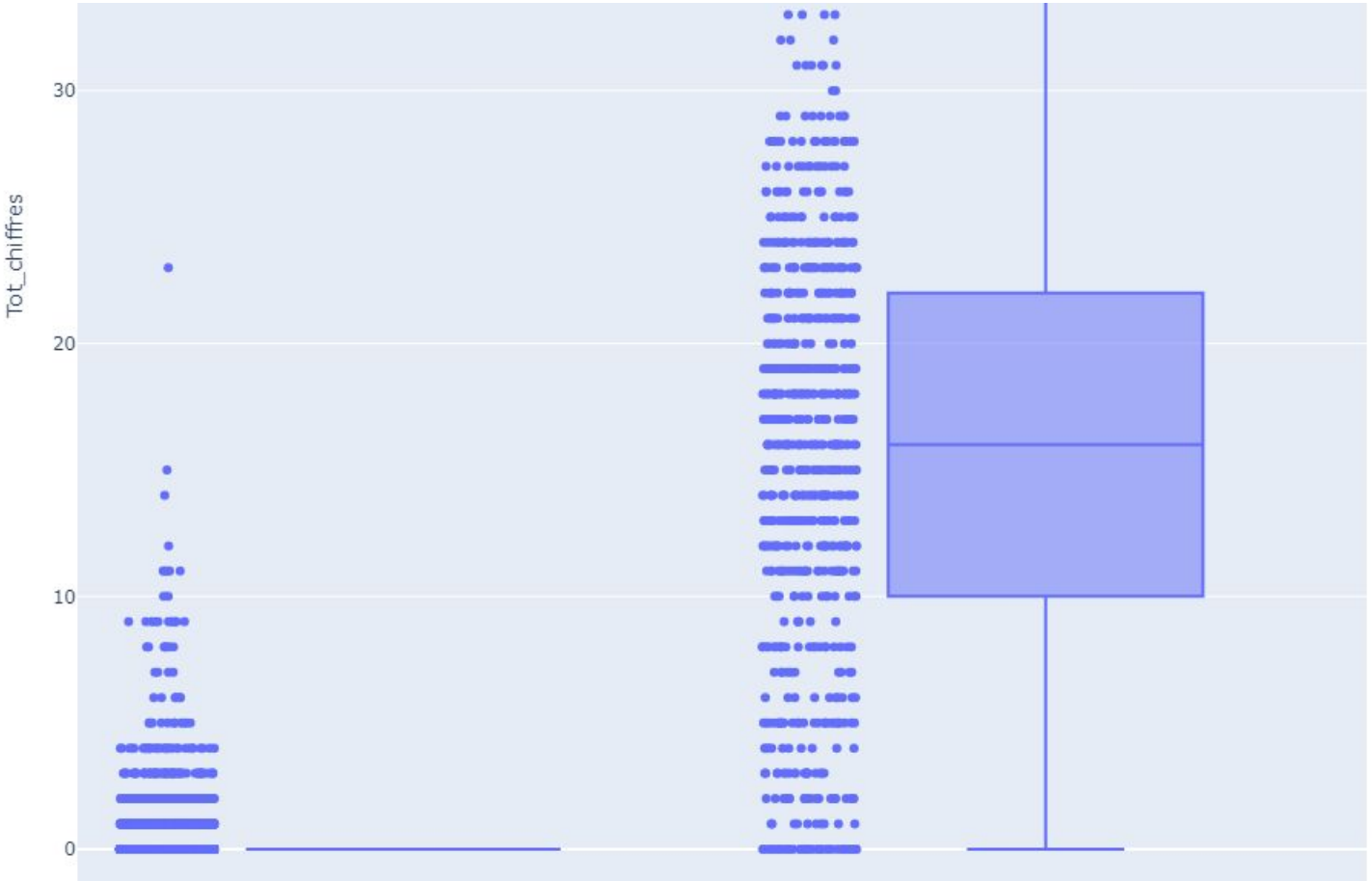


Insatisfait ! On en veut plus !

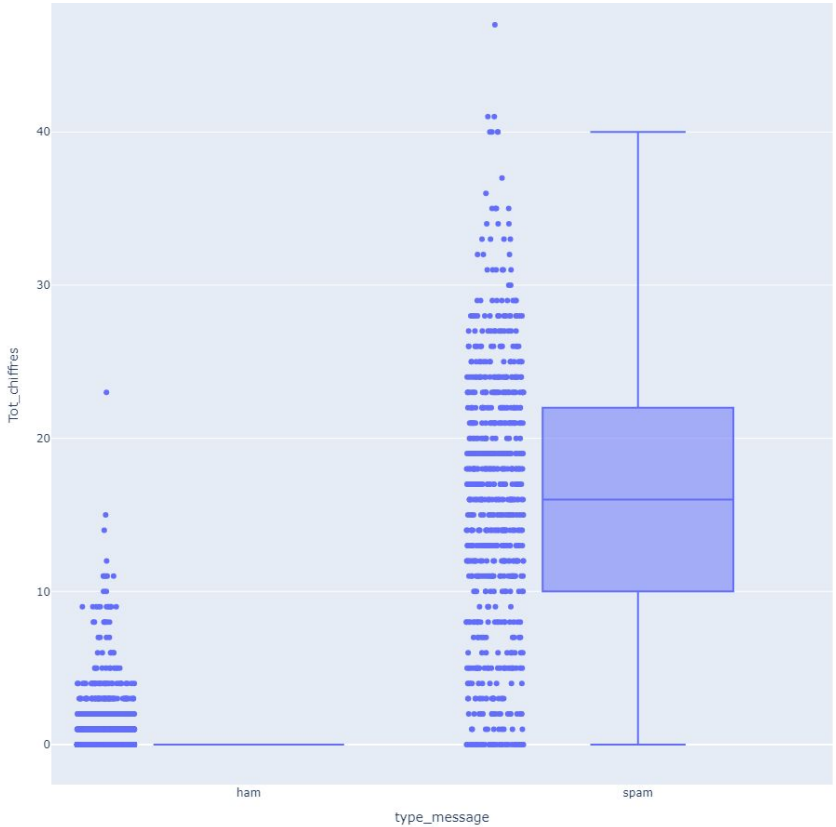
6) Le nombre de chiffre(s) par message



6) Le nombre de chiffre(s) par message



6) Le nombre de chiffre(s) par message



Caractéristique intéressante, on conserve !

7) Les caractères spéciaux

Qu’est ce ?

```
def caracspec(text):
    caraspe = re.findall(r'[^a-zA-Z0-9\s!"\'(),-.:;?]', text)
    return len(caraspe)
```

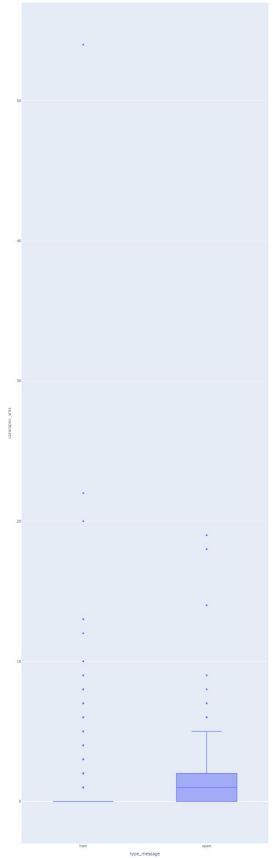
7) Les caractères spéciaux

Quelques statistiques supplémentaires...

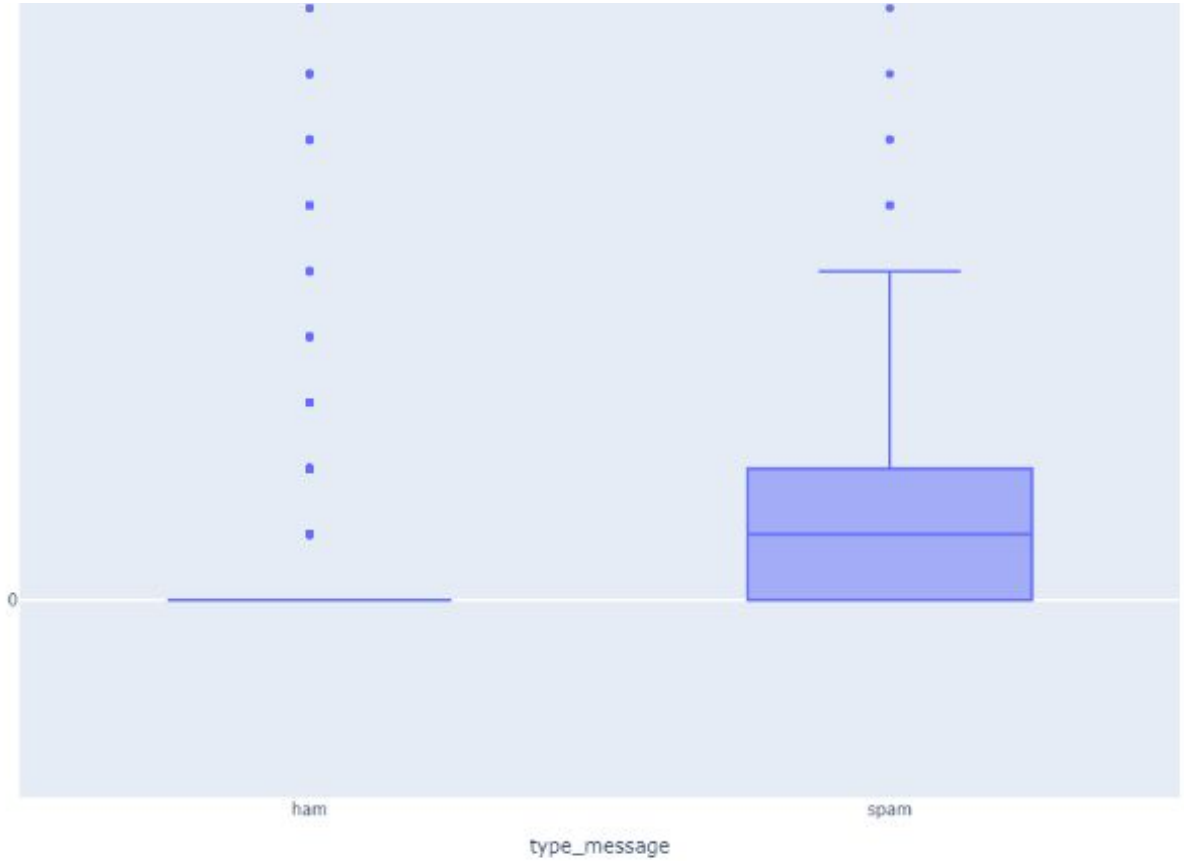
	count	mean	std	min	25%	50%	75%	max
type_message								
ham	4827.0	0.357986	1.375094	0.0	0.0	0.0	0.0	54.0
spam	747.0	1.613119	1.862999	0.0	0.0	1.0	2.0	19.0

7) Les caractères spéciaux

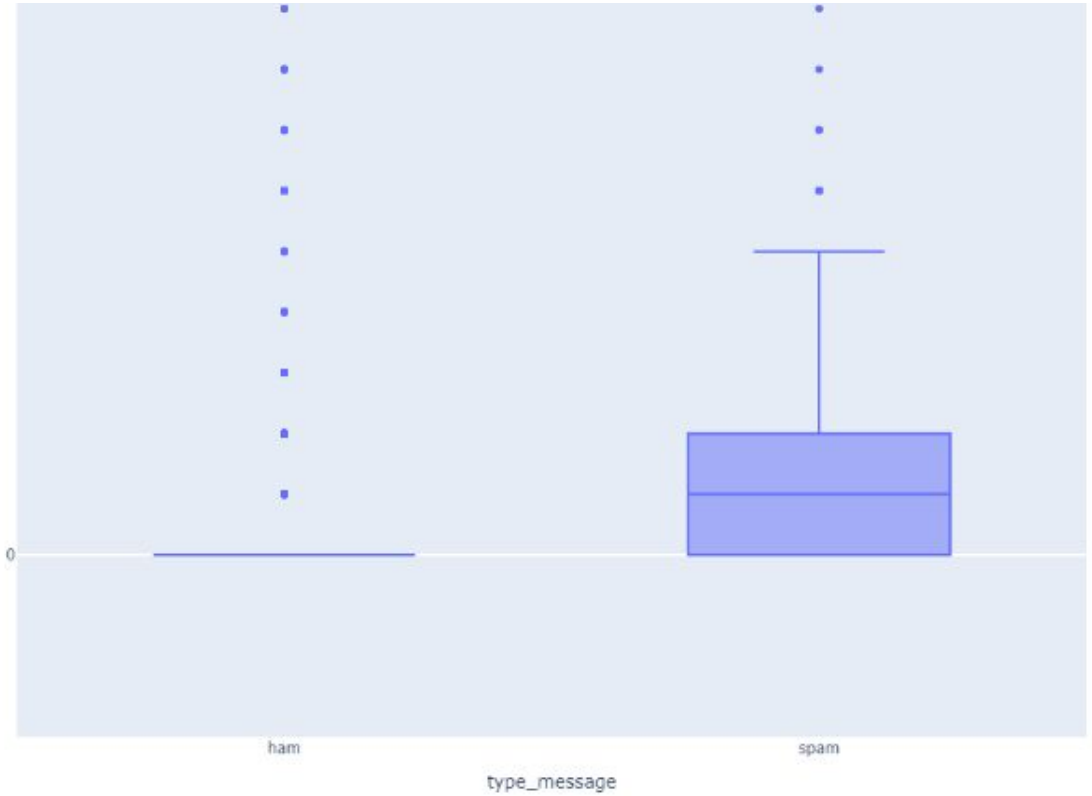
Sacré distribution !



7) Les caractères spéciaux

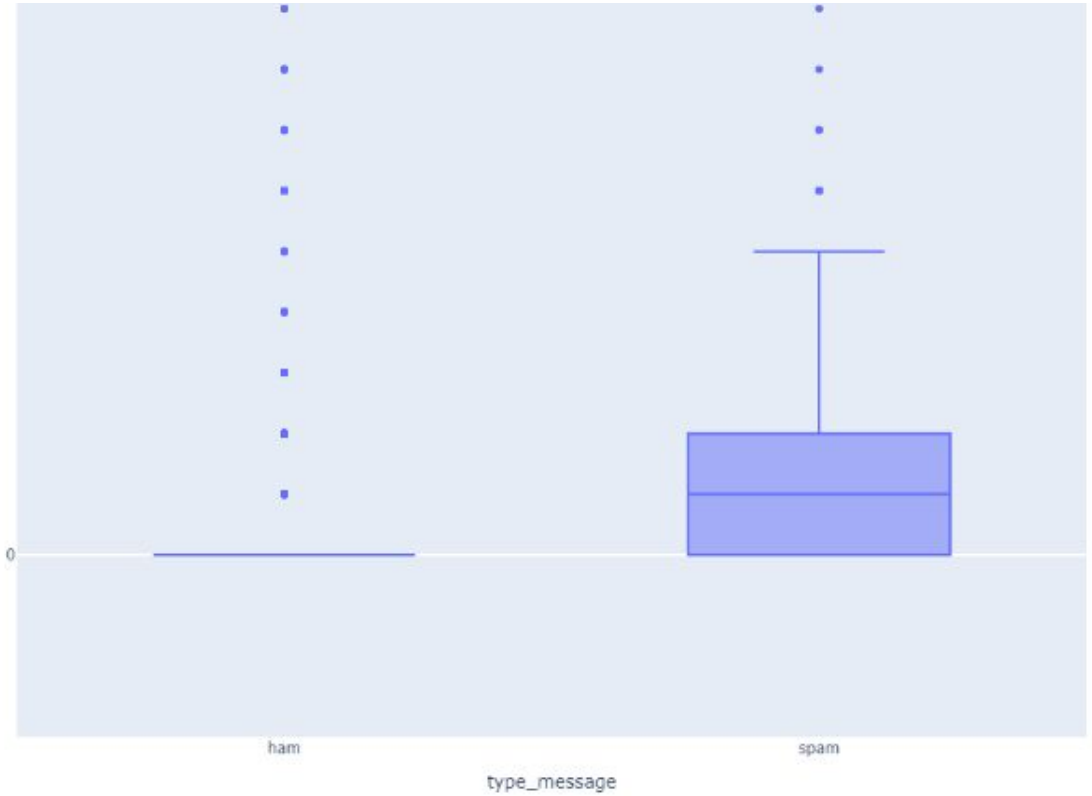


7) Les caractères spéciaux



Caractéristique intéressante, on conserve !

7) Les caractères spéciaux

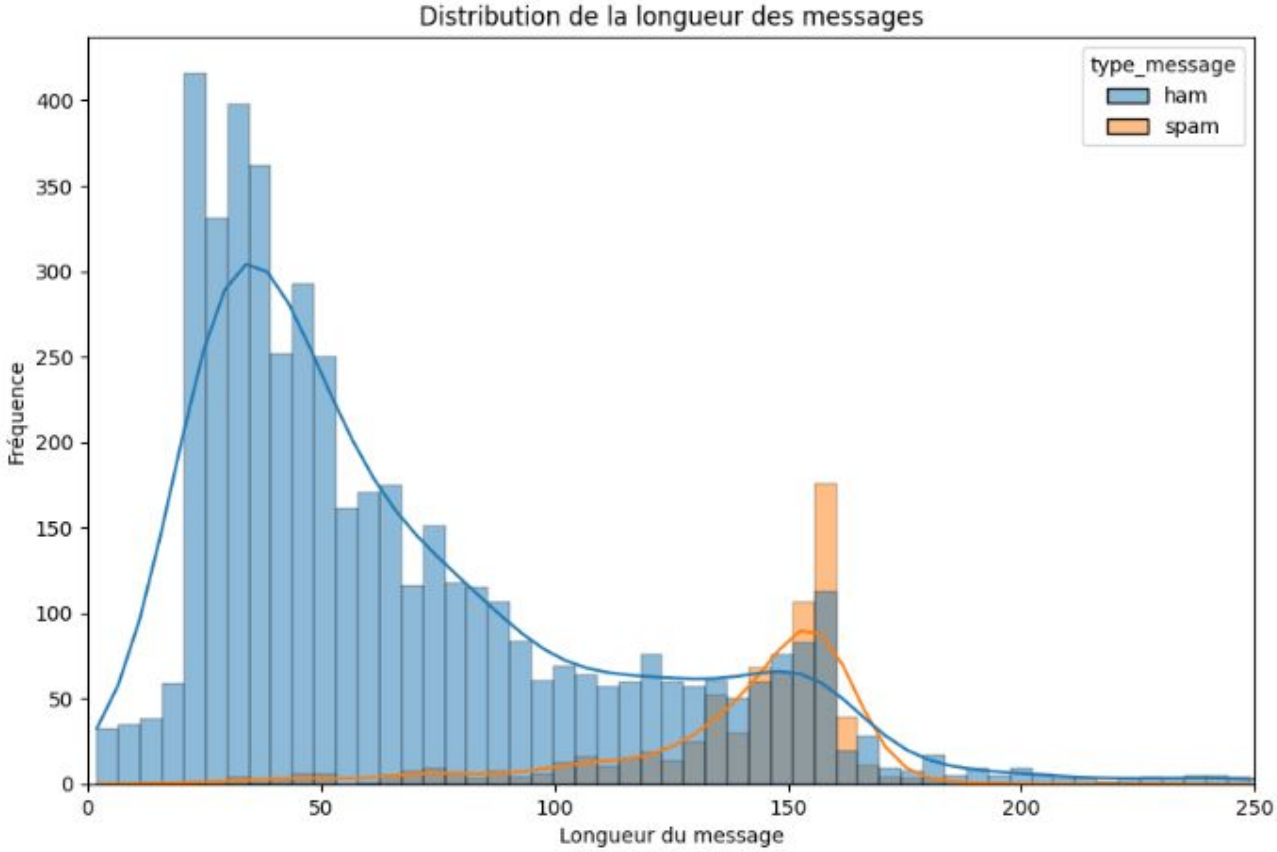


Caractéristique intéressante, on conserve !

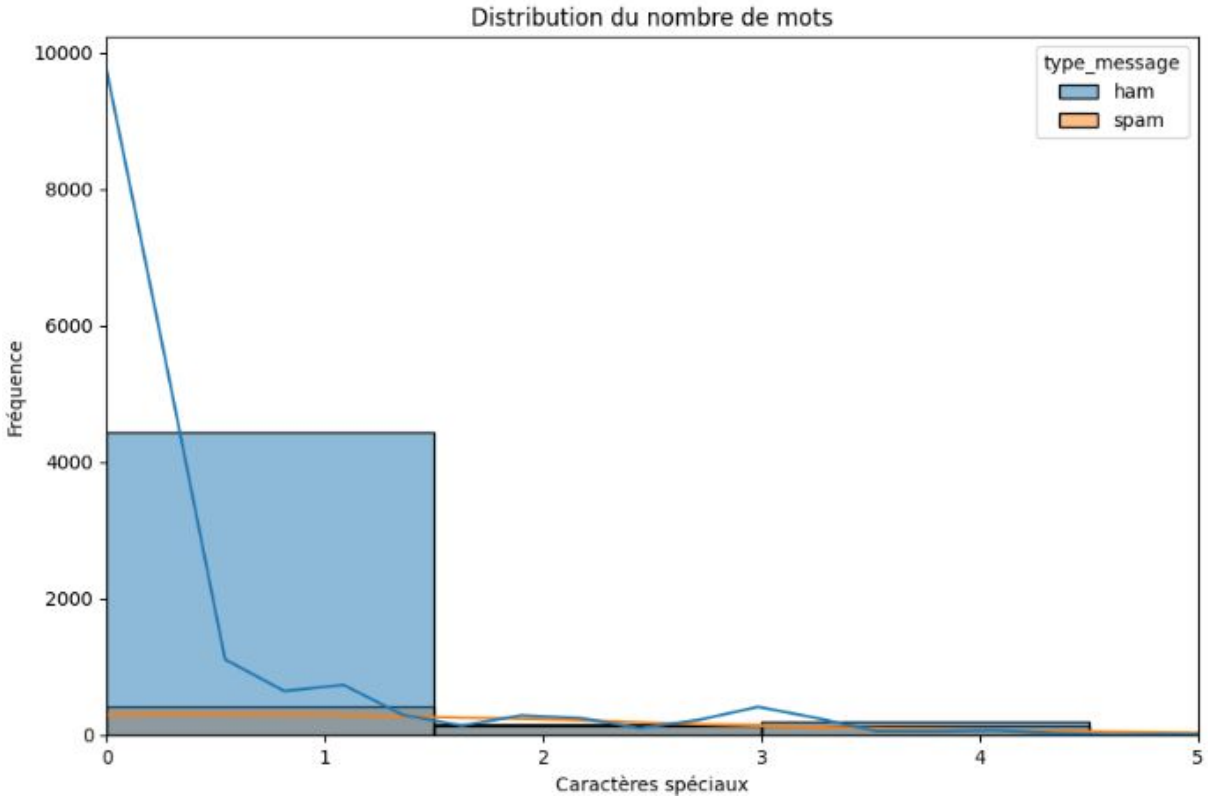
Et finalement, qu’est ce que l’on conserve ?

- La longueur des messages
- Les mots redondants des spams
- Les chiffres totaux
- Les caractères spéciaux

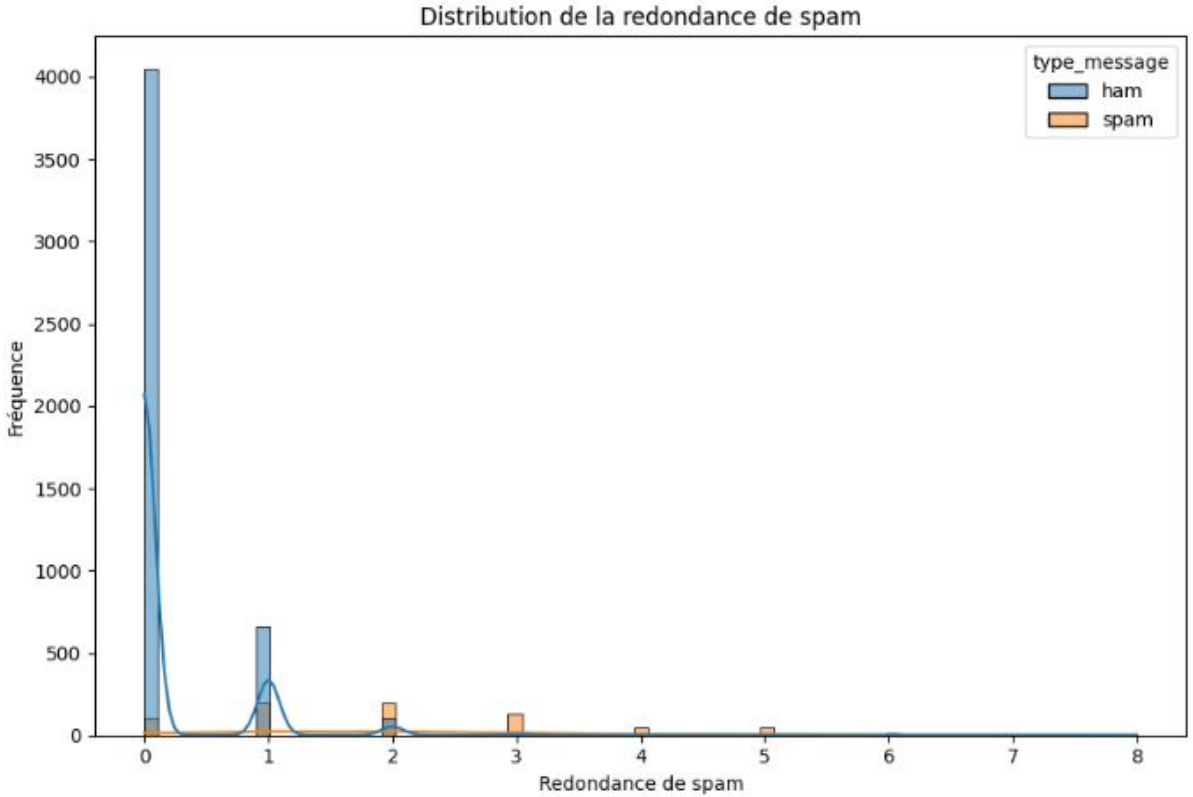
La distributivité de nos données pour choisir le modèle de normalisation



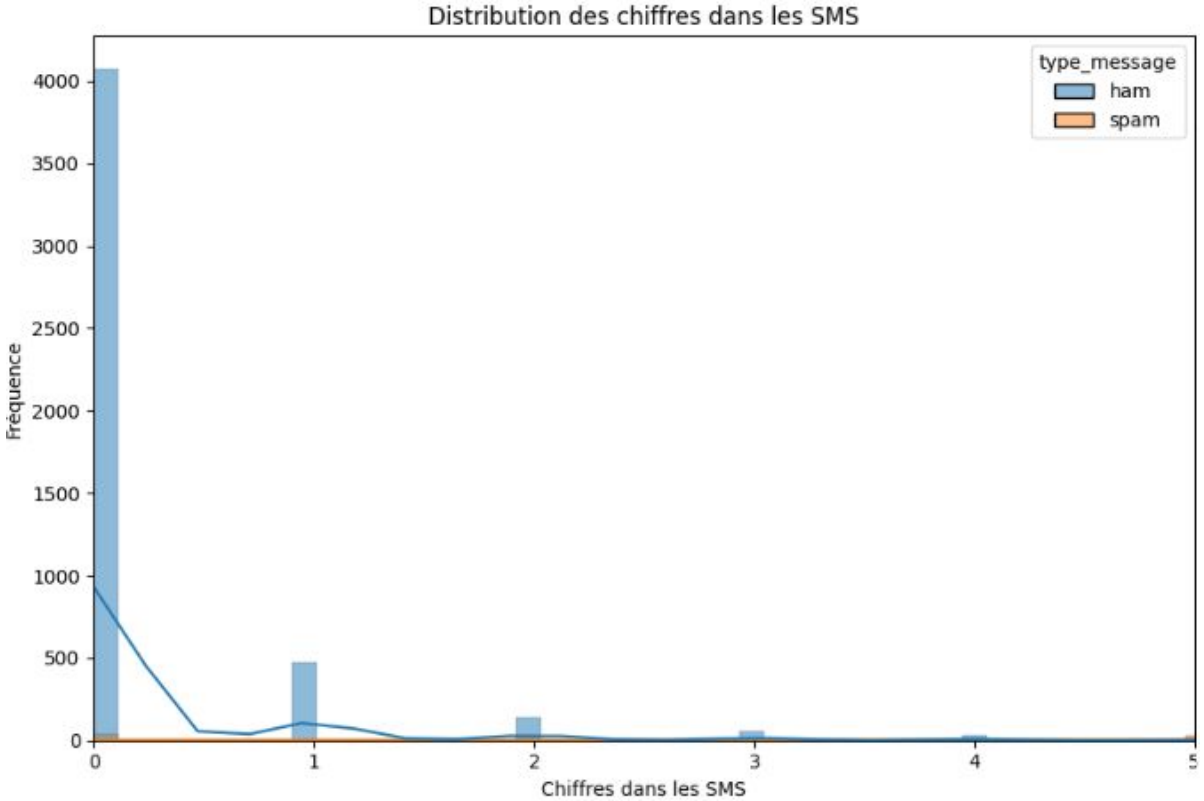
La distributivité de nos données pour choisir le modèle de normalisation



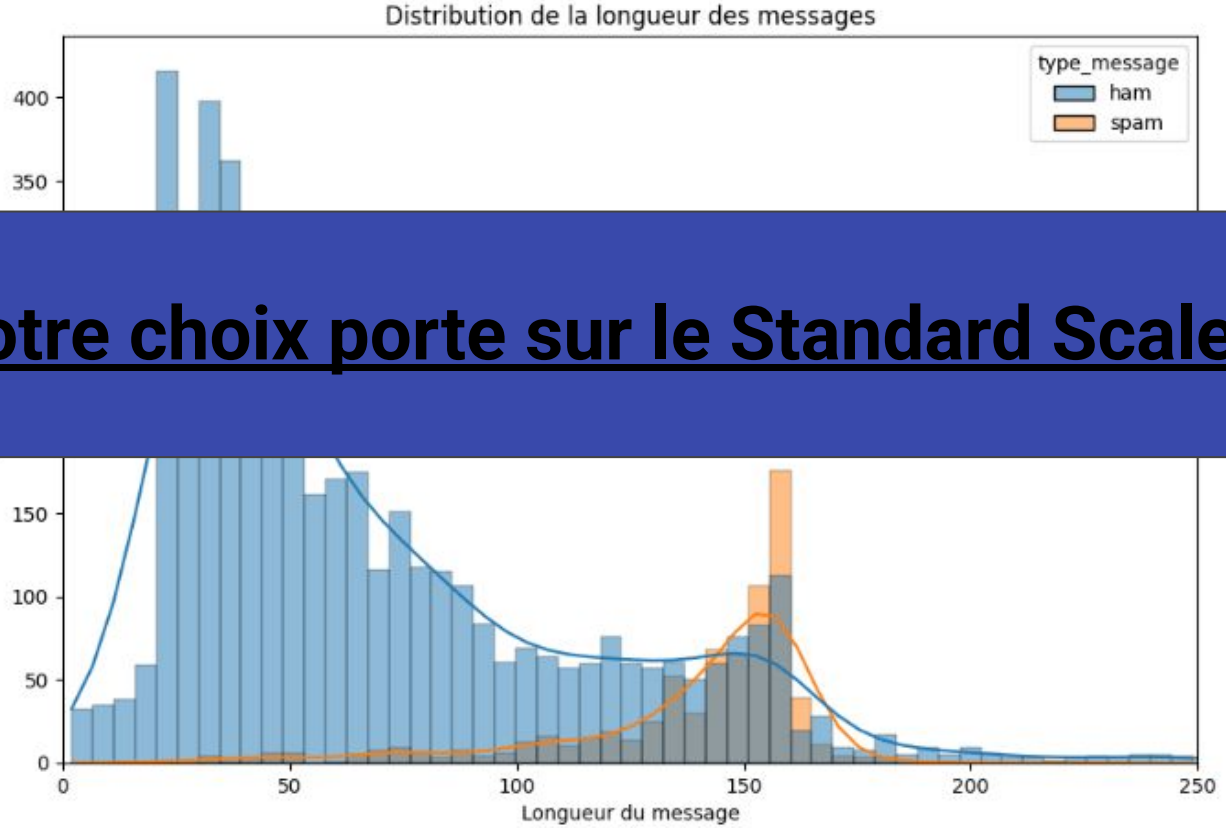
La distributivité de nos données pour choisir le modèle de normalisation



La distributivité de nos données pour choisir le modèle de normalisation



La distributivité de nos données pour choisir le modèle de normalisation



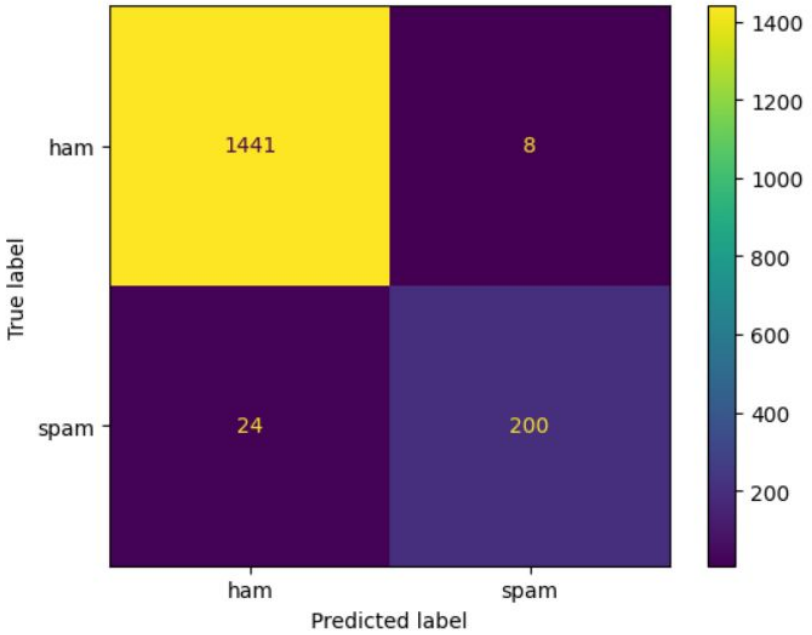
Notre choix porte sur le Standard Scaler

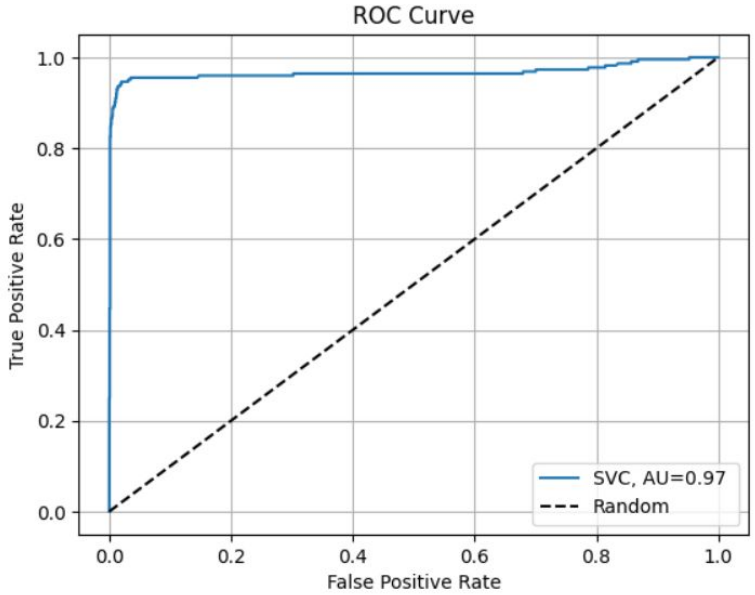
Les modèles utilisés

- Modèle SVC
- Modèle Logistic Regression
- Modèle Naive Bayes Gaussien
- Modèle Random Forest Classifier

Modèle 1 : SVC

F1 Score : 0.923
Test de précision : 0.961
AUC : 0.97

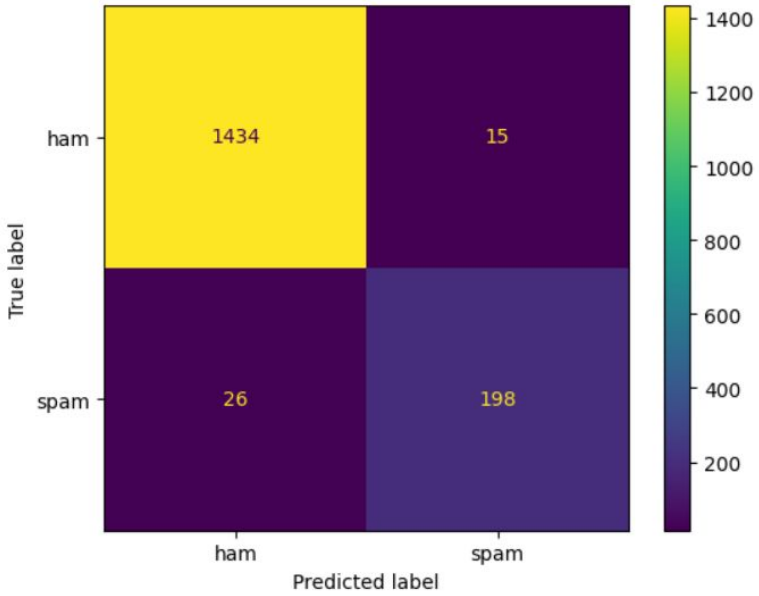


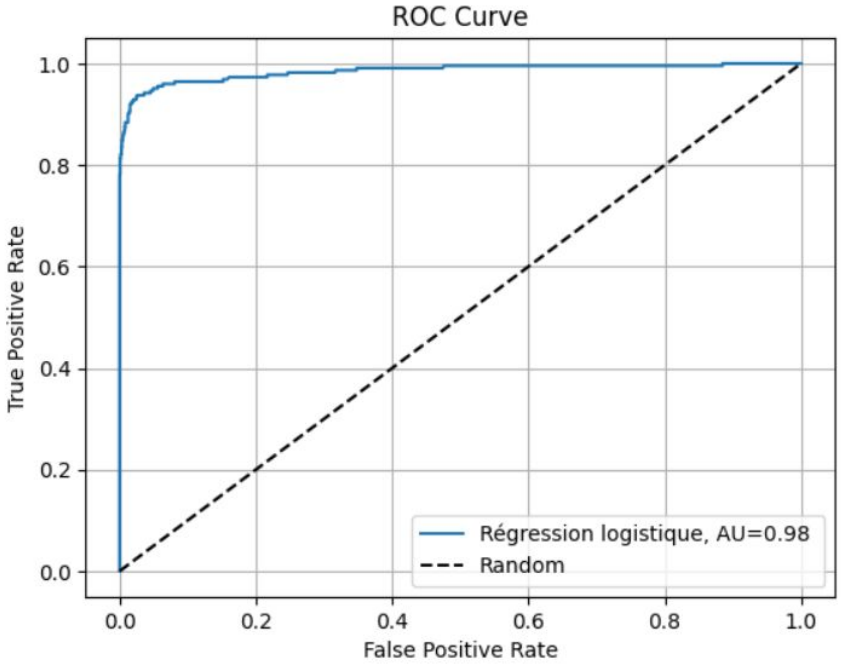


Modèle 2 : Logistic Regression

F1 Score : 0.90
Test de précision : 0.930

AUC : 0.98

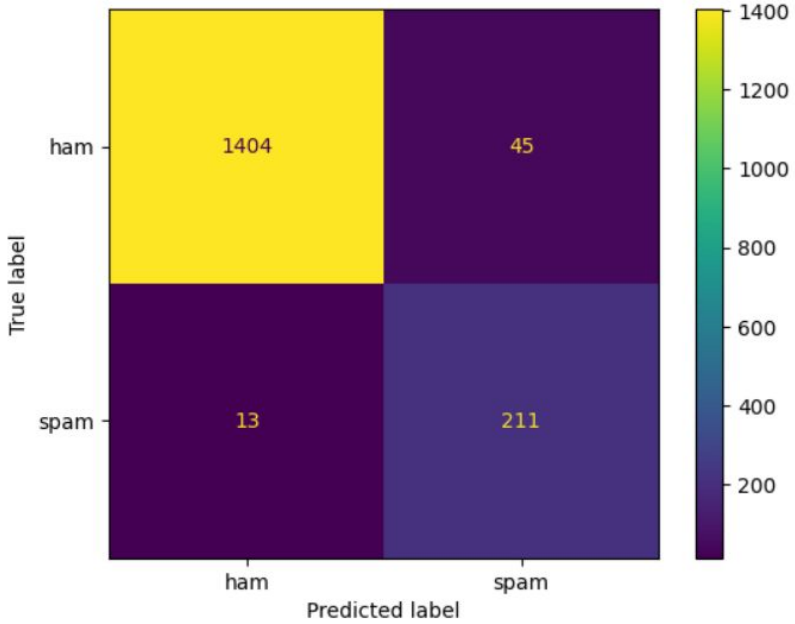


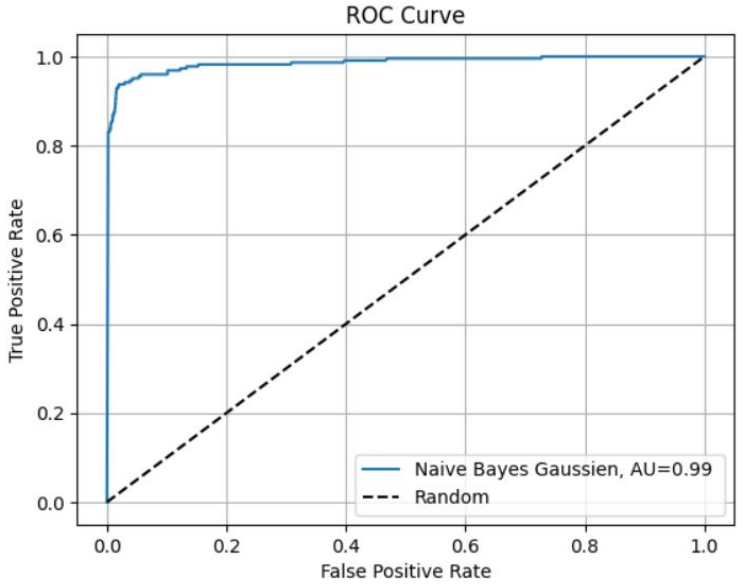


Modèle 3 : Naive Bayes Gaussien

F1 Score : 0.879
Test de précision : 0.824

AUC : 0.99

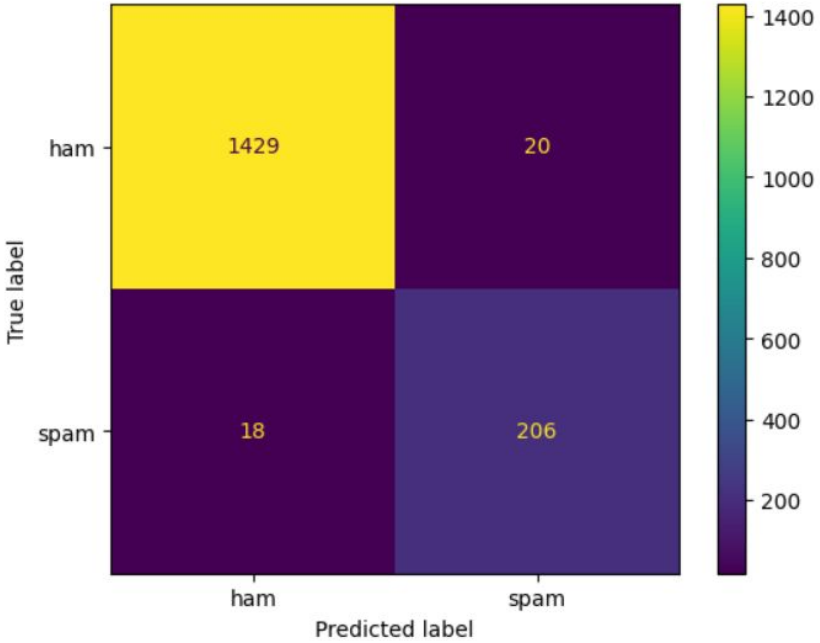


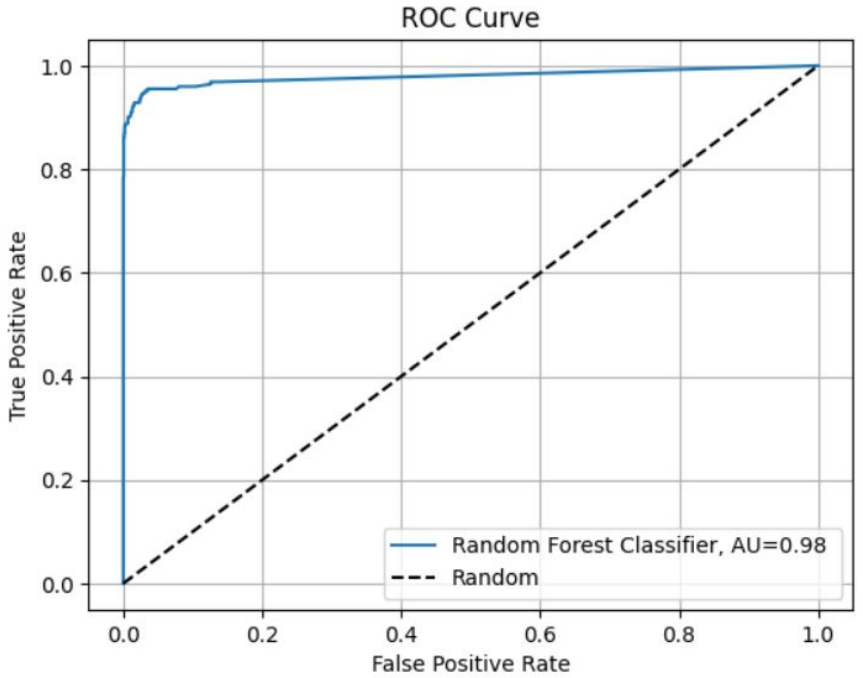


Modèle 4 : RandomForestClassifier

F1 Score : 0.916
Test de précision : 0.912

AUC : 0.98





Modèle choisi :

Modèle SVC

F1 Score : 0.923
Test de précision : 0.961
AUC : 0.97

Modèle Logistic Regression

F1 Score : 0.90
Test de précision : 0.930
AUC : 0.98

Modèle Naive Bayes

F1 Score : 0.879
Test de précision : 0.824
AUC : 0,99

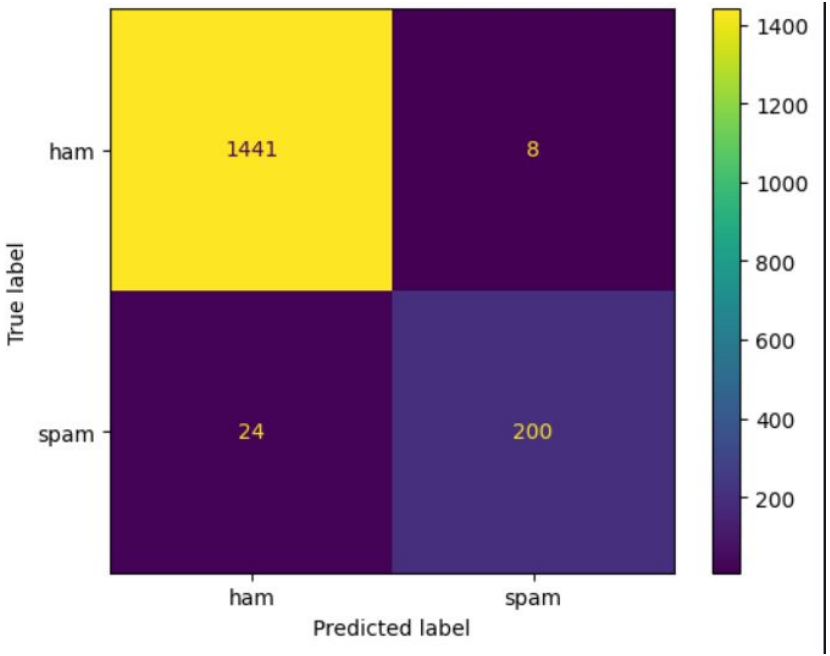
Modèle Random Forest Classifier

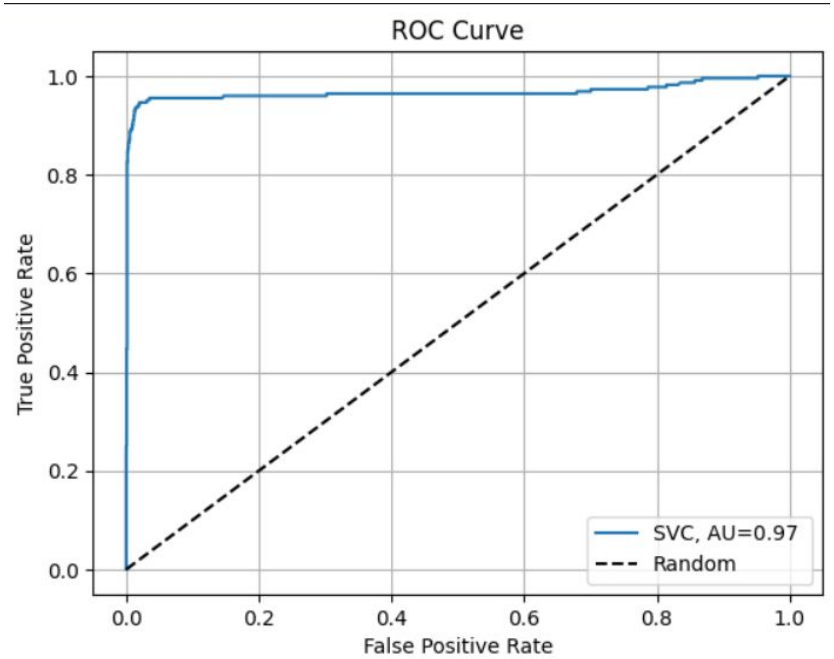
F1 Score : 0.916
Test de précision : 0.912
AUC : 0.98

Validation du Modèle choisi
Modèle SVC

F1 Score : 0.926
Test de précision : 0.962

AUC : 0.97





1) Application Spamlit

Interface Utilisateur Initiale

1) Application Spamlit

Test Application avec un message Spam

1) Application Spamlit

Test Application avec un message Ham

Merci pour votre attention, avez-vous des questions ?