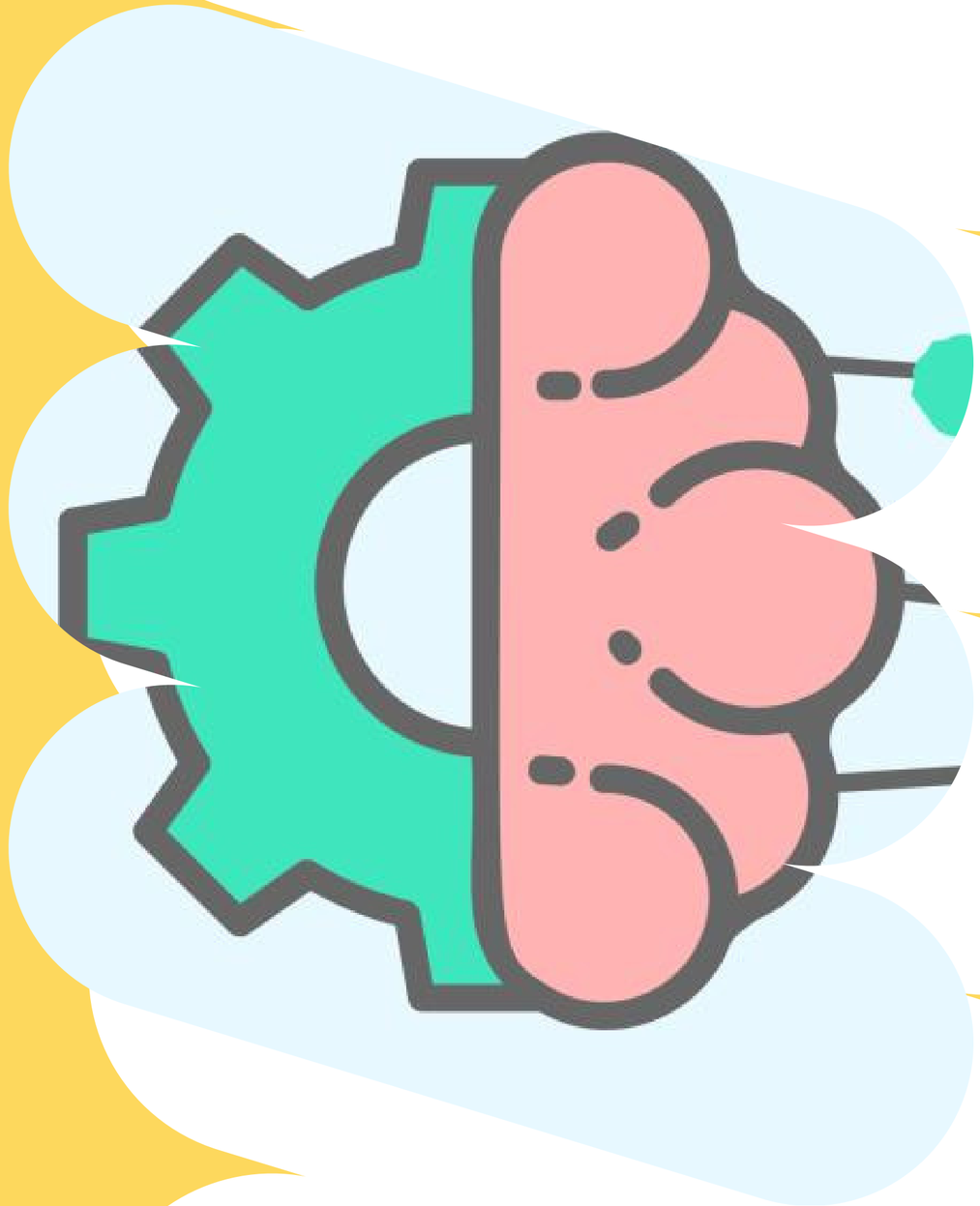
Mail Spam

Malek hentati
Hadil Masmoudi
Firas Barkia



LSI3 2.1

Plan

Introducrion

Affichage & interpritation des Donner

Verification des valeurs aberrantes

Selecion des parametre les plus pertinentes

Description des modèles &Evalution de la performance

Détermination des bon paramétré

Matrices de Confusion

Tableaux Récapitulatifs

INTRODUCTION

L 'accroissement du volume d'e-mails indésirables appelés spam a créé un besoin intense de développement de filtres antispam plus fiables et robustes. Ces méthodes d'apprentissage automatique sont récemment utilisés pour détecter et filtrer avec succès les spams. Nous présentons un revue systématique de certaines des approches populaires de filtrage du spam basé sur l'apprentissage automatique.

Affichage & Interprétation des Données

	Email No.	the	to	ect	and	for	of	a	you	hou	...	connevey	jay	valued	lay	infrastructure	military	allowing	ff	dry	Prediction
0	Email 1	0	0	1	0	0	0	2	0	0	...	0	0	0	0	0	0	0	0	0	0
1	Email 2	8	13	24	6	6	2	102	1	27	...	0	0	0	0	0	0	0	1	0	0
2	Email 3	0	0	1	0	0	0	8	0	0	...	0	0	0	0	0	0	0	0	0	0
3	Email 4	0	5	22	0	5	1	51	2	10	...	0	0	0	0	0	0	0	0	0	0
4	Email 5	7	6	17	1	5	2	57	0	9	...	0	0	0	0	0	0	0	1	0	0

5 rows × 3002 columns

Interprétation : les valeurs des données sont quantitatives

Vérification des valeurs aberrantes

```
dataset.isnull().any().sum()
```

```
0
```

il n' y a pas des valeurs aberrantes

Sélection des paramètres les plus pertinentes

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
X = dataset.iloc[:, 1:-1].values
y = dataset.iloc[:, -1].values
X.shape
```

```
(5172, 3000)
```

```
X_new = SelectKBest(chi2, k=2).fit_transform(X, y)
X_new.shape
```

```
(5172, 2)
```

Description des modèles & Evaluation de la performance

Random Forest Tree

La forêt aléatoire est une forêt flexible et facile à utiliser apprentissage automatique algorithme qui produit, même sans réglage hyper-paramétrique, un excellent résultat la plupart du temps. C'est également l'un des algorithmes les plus utilisés, en raison de sa simplicité et de sa diversité (, il peut être utilisé pour les tâches de classification et de régression).

Description des modèles & Evaluation de la performance

Random Forest Tree

Evaluation :

Accuracy Score of Random Forest Classifier : 0.9752513534416086

	precision	recall	f1-score	support
0	0.98	0.98	0.98	926
1	0.96	0.95	0.96	367
accuracy			0.98	1293
macro avg	0.97	0.97	0.97	1293
weighted avg	0.98	0.98	0.98	1293

Description des modèles & Evaluation de la performance

KNN

Evaluation :

```
Accuracy Score of KNN Classifier :  0.8654292343387471
              precision    recall  f1-score   support

           0       0.91      0.90      0.91       936
           1       0.75      0.76      0.76       357

    accuracy                           0.87      1293
   macro avg       0.83      0.83      0.83      1293
weighted avg       0.87      0.87      0.87      1293
```

Détermination des bons paramétres

Grid Search

Random Forest Tree

```
0.9886597001002669
{'max_depth': 21, 'min_samples_split': 5}
```

Détermination des bon paramétré

Random Forest Tree

```
0.9886597001002669
{'max_depth': 21, 'min_samples_split': 5}
```

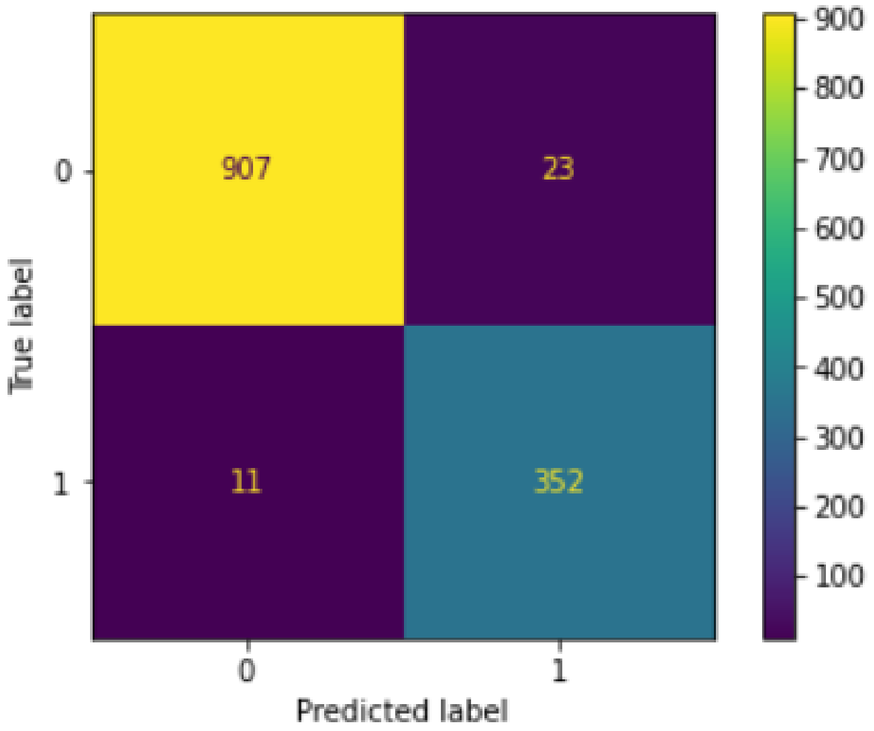
Détermination des bon paramétré

KNN

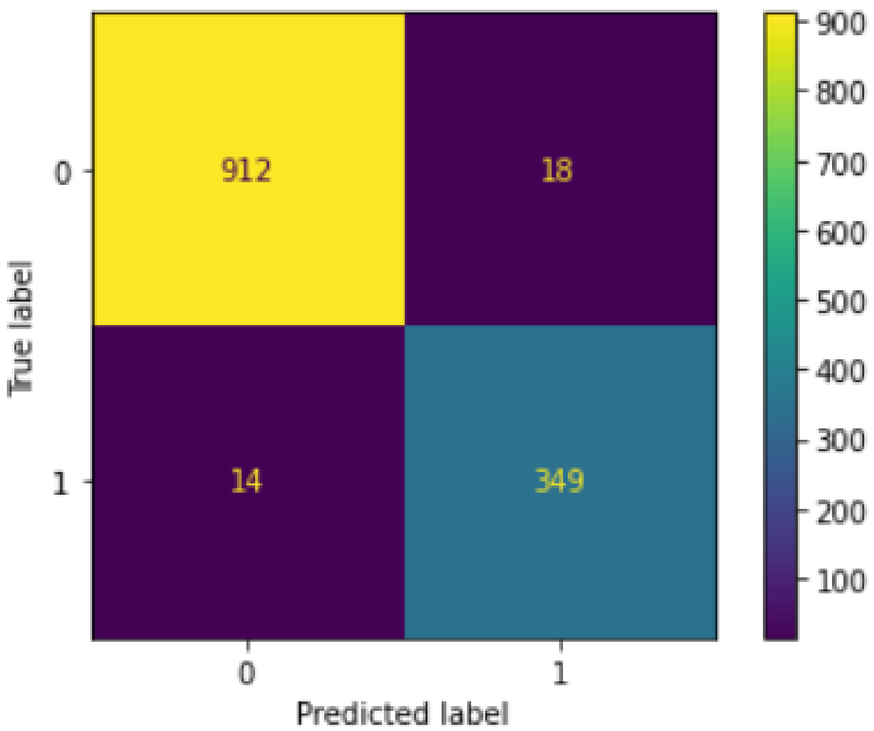
```
0.8690418435415403
{'metric': 'minkowski', 'n_neighbors': 4}
```

Matrices de Confusion

KNN



Random Forest Tree



Tableaux Récapitulatifs

	KNN	RANDOM FOREST TREE
Accuracy without GridCV	0.8654292343387471	0.9752513534416086
Accuracy with GridCV	0.8690418435415403	0.9886597001002669

MERCI POUR VOTRE ATTENTION !!!