RAPORT

DU PROJET MACHINE LEARNING

Réalisé par :

- LEGSIR Imane
- ZAOUI Noussaiba
- LAOUAJ Kaoutar
- DOBLI Hajar
- ANIBA Doha
- BENDALI Mana

Encadré par :

Pr.HAJA Zakaria

2024

SOMMAIRE

1

DÉFINITION D'OBJECTIF DU PROJET

- **Notre projet vise à utiliser le Machine Learning pour la détection des e-mails spam et ham (non-spam).**

- L'objectif principal est de développer un modèle de **classification des e-mails** qui minimise les erreurs de classification, en particulier en évitant que des e-mails importants soient incorrectement identifiés comme spam.

- Pour atteindre cet objectif, nous mettons en œuvre un modèle de Machine Learning avancé capable de **distinguer efficacement les e-mails spam des e-mails légitimes**. Une des caractéristiques clés de notre projet est l'intégration d'une fonctionnalité d'évaluation continue pour les e-mails envoyés. Cette fonctionnalité permet de vérifier en temps réel si des e-mails cruciaux risquent d'être marqués comme spam, et d'ajuster le modèle en conséquence pour améliorer sa précision.

- Notre approche combine des techniques de Machine Learning sophistiquées avec des mécanismes d'évaluation en temps réel pour assurer une classification précise des e-mails, **garantissant ainsi que les communications importantes ne soient pas perdues dans les filtres ham.**

2

SOURCE DES DONNES

Le dataset spam_ham_dataset.csv de Kaggle est un excellent choix pour effectuer une prédiction des spams en machine learning en raison de plusieurs raisons clés :

1.Qualité et Taille du Dataset. Ce dataset est bien structuré et contient suffisamment de données pour entraîner un modèle de machine learning performant. Il est crucial d'avoir suffisamment d'exemples de spam et de ham pour permettre au modèle d'apprendre la définition unique des caractéristiques de chaque classe.

2. Équilibrage des Classes: Les données du dataset sont généralement équilibrées entre les classes du spam et du ham.

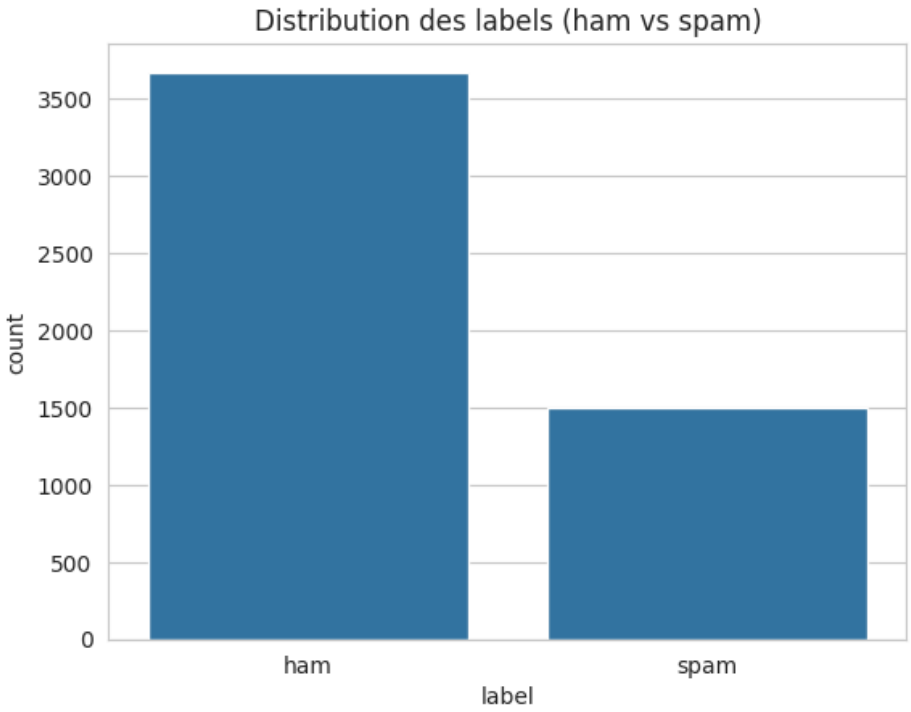
3. Variété du Contenu Textuel :
Il est possible trouver différents types du spam et de hams dans les emails due dataset – des publicités malveillantes et du phishing, ainsi que de l'information corrélatif ou coronique. Ce basson du variety permet le modèle apprendre d'identifier le spam et le ham comme des types du documents.

3

ANALYSE EXPLORATOIRE DES DONNÉES

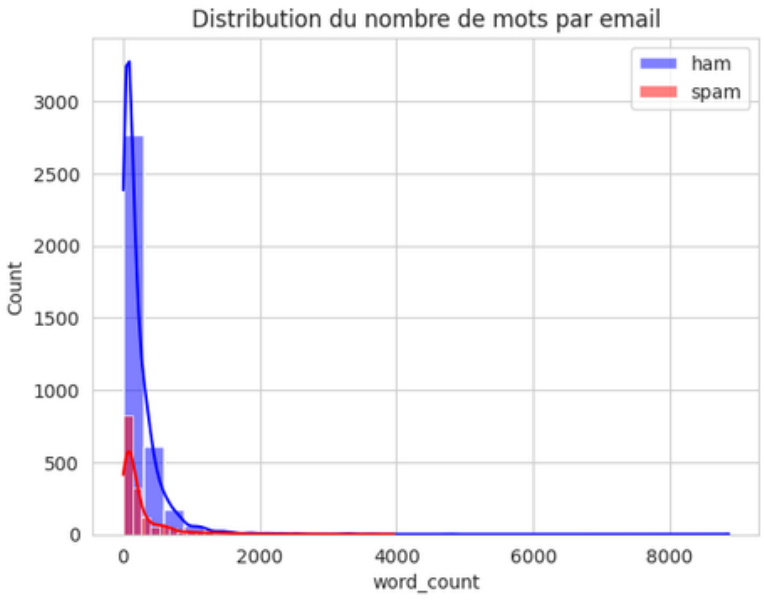
1. Distribution des Labels

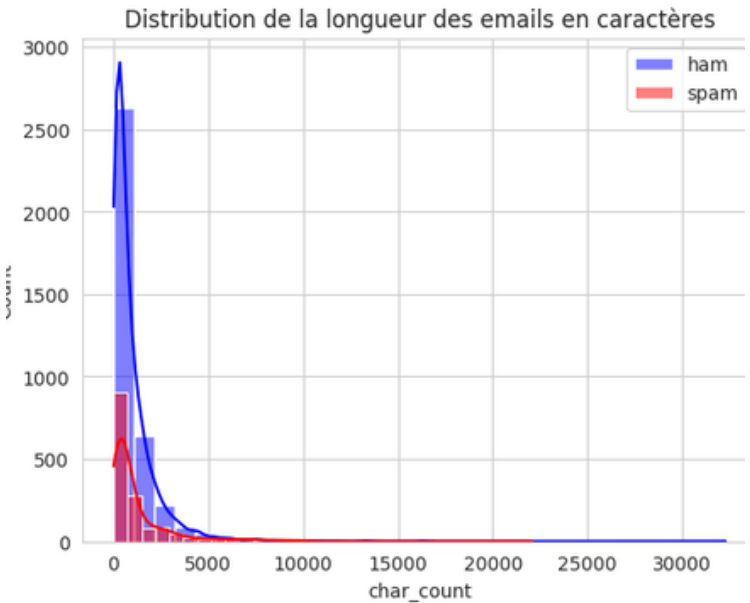
Le graphique montre la distribution des emails "ham" et "spam" dans notre dataset. On observe un déséquilibre notable avec environ 3500 emails "ham" et 1500 emails "spam".



3. Distribution du Nombre de Mots par Email

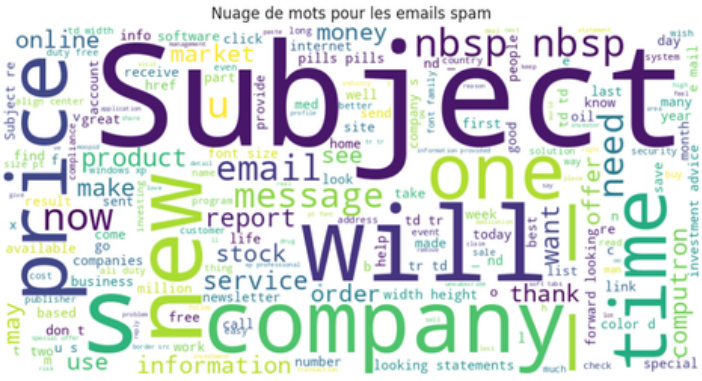
Les histogrammes ci-dessous montrent la distribution du nombre de mots pour les emails "ham" et "spam". Les emails "spam" ont tendance à être plus courts que les emails "ham".

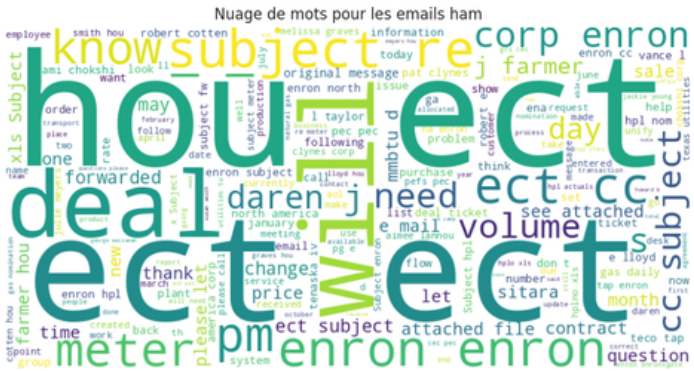




Nuages de Mots :

Les nuages de mots suivants montrent les termes les plus fréquemment utilisés dans les emails "ham" et "spam".





Statistiques Descriptives :

Les statistiques descriptives des colonnes "word_count" et "char_count" fournissent des informations sur la longueur des emails :
Ces valeurs montrent une grande variabilité dans la longueur des emails.

STATISTIQUE	NOMBRE DE MOTS	NOMBRE DE CARACTÈRES
Moyenne	228	1048
Écart type	336	1529
Min	1	11
25%	51	244
50%	121	540
75%	282	1237
Max	8862	32258

4 PRÉ-TRAITEMENT DES DONNÉES

La phase de prétraitement des données, souvent appelée prétraitement ou nettoyage des données, est une étape essentielle dans le processus d'analyse de données et de construction de modèles. Elle vise à préparer les données brutes en vue d'une analyse ultérieure ou d'une utilisation dans des modèles d'apprentissage automatique :

- **Chargement des données :**
- Les données sont chargées à partir d'un fichier CSV contenant des informations sur les e-mails, y compris le texte et les étiquettes de spam/ham.

- **Exploration des données :**
- Une analyse exploratoire est effectuée pour comprendre la structure et la distribution des données. Cela comprend des visualisations telles que des histogrammes pour les étiquettes et des nuages de points pour les variables numériques.

- **Nettoyage des données :**

  - **Suppression des colonnes inutiles :** Certaines colonnes, comme "Unnamed: 0", sont supprimées car elles ne contribuent pas à la classification.
  - **Conversion des étiquettes en valeurs numériques :** Les étiquettes de spam et de ham sont converties en valeurs numériques (0 pour ham, 1 pour spam) pour permettre l'entraînement des modèles.
  - **Vérification des valeurs manquantes :** Il est vérifié s'il y a des valeurs manquantes dans le jeu de données. Dans ce cas, aucune valeur manquante n'est détectée.

- **Vectorisation du texte :**
- Le texte des e-mails est transformé en vecteurs numériques à l'aide d'une technique de vectorisation (ici, CountVectorizer). Cela permet de représenter les données textuelles sous une forme numérique que les algorithmes d'apprentissage automatique peuvent comprendre.
- Division des données : Les données sont divisées en ensembles d'entraînement et de test pour évaluer les performances des modèles.

5 LES ALGORITHMES UTILISÉS

Pour notre projet de classification des e-mails en spam ou non-spam, nous avons exploré plusieurs algorithmes de machine learning. :

- **Arbre de décision :**
  Un arbre de décision est un modèle prédictif qui utilise une structure en arbre pour prendre des décisions basées sur des règles dérivées des données d'entraînement.
  - **Avantages** : Facile à comprendre et à interpréter, capable de gérer des données catégorielles et numériques.
  - **Inconvénients** : Peut être sujet à l'overfitting, surtout avec des arbres profonds.
- **Régression logistique :**

Un modèle de classification qui utilise une fonction logistique pour prédire la probabilité qu'une instance appartienne à une classe particulière.
  - **Avantages :** Simple et efficace pour les problèmes de classification binaire, bien interprétable.
  - **Inconvénients** : Peut avoir des performances limitées avec des relations non linéaires complexes.
- **Machines à vecteurs de support (SVM) :**
  Un algorithme de classification qui cherche à trouver l'hyperplan qui sépare le mieux les classes dans un espace de caractéristiques.
  - **Avantages :** Efficace dans des espaces à haute dimension, robuste face à l'overfitting grâce à l'utilisation de la marge de séparation.
  - **Inconvénients** : Peut être computationalement coûteux, surtout avec de grands jeux de données.
- **Forêt aléatoire (Random Forest) :**
  Un ensemble de nombreux arbres de décision entraînés sur différentes parties des données avec des sous-ensembles de caractéristiques, puis combinés pour améliorer la précision et la robustesse.
  - **Avantages** : Réduit le risque d'overfitting, fonctionne bien avec des données hétérogènes et manquantes.
  - **Inconvénients** : Peut être plus lent et moins interprétable que des arbres de décision individuels.
- **K plus proches voisins (KNN) :**
  Un algorithme de classification qui attribue une classe à une instance en fonction des classes de ses k voisins les plus proches dans l'espace des caractéristiques.
  - **Avantages** : Simple à implémenter, non paramétrique, ce qui le rend flexible.
  - **Inconvénients :** Peut être lent pour des jeux de données de grande taille, sensible aux caractéristiques redondantes ou non pertinentes.

Chacun de ces algorithmes a ses propres forces et faiblesses. En combinant plusieurs approches, nous pouvons sélectionner le modèle le plus performant en fonction des caractéristiques spécifiques de notre jeu de données et de notre problème de classification.

6

CONCLUSION

Notre projet de classification des e-mails en spam ou non-spam a exploré plusieurs algorithmes de machine learning :

- Arbre de Décision,
- Régression Logistique,
- Machines à Vecteurs de Support (SVM),
- Forêt Aléatoire et K Plus Proches Voisins (KNN).

Chaque algorithme présente des avantages spécifiques, tels que l'interprétabilité pour les Arbres de Décision et la Régression Logistique, et une grande précision et robustesse pour les SVM et les Forêts Aléatoires.

Le dataset **spam_ham_dataset.csv** de Kaggle a été utilisé pour entraîner nos modèles en raison de sa qualité, de l'équilibrage des classes, et de la diversité de son contenu textuel. Cette diversité permet au modèle d'apprendre à identifier différents types de spam et de ham.

En combinant les forces de ces algorithmes et en utilisant des techniques d'optimisation, nous avons pu développer un modèle performant et robuste pour la classification des e-mails. Les prochaines étapes incluront l'optimisation des hyperparamètres et l'exploration de techniques d'ensemble pour améliorer encore les performances de notre système.