

# **Rapport du Modèle d'identification des SMS Spam basé sur la régression logistique :**

## **I- Analyse du programme :**

### **Jeu de données**

- Le programme lit des données à partir d'un fichier CSV nommé 'SMSSpamCollection.txt' qui contient probablement des messages SMS étiquetés comme spam ou non spam.

### **Entrée**

- Les données d'entrée consistent en des messages SMS, qui sont prétraités en utilisant des techniques telles que la suppression de la ponctuation et des mots vides.
- Le texte prétraité est converti en une matrice de comptage de jetons en utilisant le modèle de sacs de mots (BoW).

### **Sortie**

- La sortie est binaire, indiquant si un message est classé comme spam ou non spam.

### **Techniques Utilisées**

1. Prétraitement de Texte :
  - Suppression de la ponctuation.
  - Suppression des mots vides en utilisant la bibliothèque NLTK.
2. Extraction de Caractéristiques :
  - Modèle de sacs de mots (BoW) pour convertir les données textuelles en vecteurs numériques.
3. Modèle d'Apprentissage Automatique :
  - Classificateur bayésien naïf multinomial pour la classification du spam.
4. Évaluation du Modèle :
  - Rapport de classification, matrice de confusion et score de précision sont utilisés pour évaluer les performances du modèle.

## **II- Étapes :**

1. **Prétraitement des Données :**

- Suppression des doublons.
- Traitement des valeurs manquantes.
- Prétraitement du texte (suppression de la ponctuation et des mots vides).

## **2. Extraction de Caractéristiques :**

- Conversion des données textuelles en une matrice de comptage de jetons (BoW).

## **3. Entraînement du Modèle :**

- Division du jeu de données en ensembles d'entraînement et de test.
- Entraînement d'un classificateur bayésien naïf multinomial sur les données d'entraînement.

## **4. Évaluation du Modèle :**

- Évaluation des performances du modèle sur les ensembles de données d'entraînement et de test en utilisant des métriques de classification telles que la précision, le rappel, le score F1, la matrice de confusion et l'exactitude.