

# Projet de détection des spams


Groupe C





# INTRODUCTION

Dans notre quotidien, les messages spam représentent une vraie nuisance. Ils envahissent les boîtes de réception et peuvent contenir des arnaques ou des publicités non sollicitées. L'objectif de notre projet est donc de détecter automatiquement les spams dans les messages SMS à l'aide de techniques de machine learning.





# PRÉTRAITEMENT DES DONNÉES

Nous avons travaillé sur le dataset SMSSpamCollection, un fichier brut sans entête. La première étape a été de **nettoyer et organiser les données**. Nous avons ajouté les colonnes label et message, supprimé les espaces et les tabulations inutiles, et uniformisé les labels en minuscules pour éviter toute ambiguïté. Par exemple, un message était soit “ham” (normal) soit “spam”.

Ensuite, nous avons **créé des variables clés** pour mieux caractériser les messages.

Ces variables permettent au modèle de détecter des patterns typiques des spams, comme les messages longs avec beaucoup de majuscules ou de chiffres.

# SEPARATION DES DONNÉES

Nous avons divisé le dataset en **jeu d'entraînement (80%)** et **jeu de test (20%)** pour entraîner les modèles et évaluer leur performance sur des données inédites.

# CHOIX ET ENTRAINEMENT DES MODELES

Nous avons testé trois modèles :

- **Régression Logistique**
- **Arbre de Décision**
- **Random Forest**

Chaque modèle a été évalué avec plusieurs métriques : précision, rappel, F1-score, matrice de confusion et courbe ROC.



## RESULTATS ET ANALYSES

Après entraînement, la Random Forest s'est révélée être le modèle le plus performant, avec une **accuracy de 97 %**.

L'analyse des variables a montré que **le nombre de mots en majuscules** et la **longueur du message** sont les facteurs les plus déterminants pour classer un message comme spam.

Nous avons visualisé ces résultats avec des histogrammes, des boxplots, et des matrices de corrélation, ce qui nous a permis de comprendre le comportement des différentes variables. La courbe ROC a confirmé que la Random Forest dépasse les autres modèles et qu'elle est fiable pour détecter les spams.

## RECOMMANDATIONS METIERS

Sur la base de nos résultats, nous recommandons :

- De **déployer le modèle Random Forest** pour filtrer automatiquement les messages spam
- De prioriser les messages contenant beaucoup de majuscules ou de longueur élevée pour la détection
- De mettre en place un **monitoring régulier** afin de réentraîner le modèle avec les nouvelles données et maintenir sa performance
- Et de fournir **des rapports à l'équipe de modération** pour les messages suspects détectés



## **SAUVEGARDE ET EXPLOITATION DES DONNÉES**

Le dataset enrichi avec toutes les variables créées a été sauvegardé au format TSV (tabulation) pour faciliter l'exploitation future et l'intégration dans un système de détection en production.





## CONCLUSION

En combinant **nettoyage des données, création de variables pertinentes et modèles performants**, nous avons pu développer une solution fiable pour la détection de spams. Cette approche permet non seulement de filtrer efficacement les messages indésirables, mais aussi de fournir des recommandations claires pour les équipes métiers et les systèmes automatisés.



## LISTE DES PARTICIPANTS

- KIDIMBA LUNZI
- KALOMBO NEEMA
- SHAKO ONIA
- MAMPWO SEMETE
- MOKAMO NDOMBE
- MAKUKA NGIEDI
- KATSHAY MPENGO
- NSILULU MAVUNGU
- MALANDA TOKO
- BWEMA ISAAC
- AWASSO MAKAYA
- SENGI OSKA
- MBAWU NZUZI
- MUCHAIL KAMIN
- YAMAYAMA KIFAKIO