Mastère Spécialisé Big Data    
Janvier 2021

Préparé par :

Lauris DJILO  
Michele MBALLA MEKONGO

**Apprentissage statistique et data mining**

Rapport de Projet

*Détection des mails spams en utilisant des méthodes de classification et d’exploration des données*

**Notes préliminaires au rapport :**

Nous avons commencé ce projet en utilisant le langage Python et avons réussi à classifier les éléments de notre dataset.

Nous avons commencé par utiliser l’algorithme du k-means sous python qui est implementé en utilisant la librairie scikit-learn. Cependant, nous avons remarqué qu’il était très difficile de faire un choix de nombre de cluster, les résultats obtenus étant ainsi non concluants. Nous avons donc décidé d’utiliser d’autres méthodes de classification qui n’existent pas dans les librairies pythons officielles et migrer sous le logiciel R pour effectuer le clustering et les prédictions.

L’ analyse descriptive a été réalisée sous python à l’aide de la librairie matplotlib.