Mastère Spécialisé Big Data    
Janvier 2021

Préparé par :

Lauris DJILO  
Michele MBALLA MEKONGO

**Apprentissage statistique et data mining**

Rapport de Projet

*Détection des sites de phishing ou hameçonnage à l’aide 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 implémenté 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.

# Objectif de l’analyse

## Motivation et positionnement du projet

  Le Phishing (Hameçonnage) est une forme d’escroquerie sur internet où un fraudeur s’attaque à des utilisateurs non averti afin de leur soutirer des informations sensibles tels que les coordonnées bancaires (numéro de compte, code personnel) et des adresses de résidence. Ils utilisent les données soutirées afin d’usurper l’identité de ces utilisateurs.

Traditionnellement pour se protéger du phishing, on utilise une Blacklist contenant adresses de site web considérées comme malicieux ou suspect et en mettant à jour cette liste au fur et à mesure que d’autre sites sont identifiés. L’inconvénient de cette méthode est que les blacklists ne peuvent pas être exhaustives et ne peuvent pas détecter des adresses/ sites nouvellement conçus.

Pour palier à ça, il convient donc d’utiliser des méthodes de classification ou de Machine Learning afin de pouvoir dynamiquement identifier les sites web suspect ou malicieux.

Nous allons proposer dans ce projet un programme