# Travail Pratique 1

**Forage de données**
**8INF418 – Groupe 1**

**25 janvier 2019**

## 1. Description Générale

L'objectif général de ce travail pratique 1 (TP1) est de vous familiariser avec le traitement des données qui peut avoir lieu avant l'étape d'exploitation des algorithmes d'apprentissage machine pour l'extraction des connaissances ou la classification par exemple. Ainsi, ce TP1 vous permettra de prendre en main le langage de programmation Python et de ses bibliothèques pour la science des données (ex. : numpy, pandas, matplotlib, …).

De plus, dans ce premier travail, vous aurez à produire un rapport scientifique qui analyse les données et qui décrit les étapes de traitement que vous aurez appliqué.

Vous aurez besoin des chapitres 1 à 5 du cours 8INF418 afin de réaliser ce TP1 et d'une réflexion de votre part.

## 2. Formalités

Le travail est à réaliser en groupe de **deux ou individuellement** et se composera d'un rapport scientifique avec les scripts Python que vous aurez réalisé.

La date limite pour le rendu de votre travail est le **8 février 2019 à 8h00**. Après cette date, il y aura une **pénalité de 10% par jour de retard**.

Vous déposerez votre travail sous le format d'un dossier .zip (ou .rar) sur Moodle (du cours 8INF418) dans la sous-section « **TP1** » de la section « **Remise des Travaux Pratiques** ». Le nom du fichier prendra la forme de **Prenom1_Nom1_Prenom2_Nom2_TP1.zip**. N'oubliez pas d'inclure votre code permanent sur la page de garde de ce document. Le nom respect des règles citées ci-dessus entraînera une pénalité de 5%.

## 3. Ce qui est attendu

Le rapport du TP1 devra comprendre :

- la description de votre ensemble de données
    - Par exemple :
        - dans quel(s) objectif(s) les données ont été collectées (s'il y a lieu) ;
        - comment ont été collectées les données ? (s'il y a lieu) ;
        - quelles sont les variables ? ;
        - le nombre d'instances ;
        - le nombre de classes ;
        - …
- la vérification et le pré-traitement des données
    - Par exemple
        - combien de valeurs manquantes possède votre ensemble de données ;
        - quel(s) moyen(s) avez-vous utilisé pour gérer ces valeurs manquantes ;
        - un résumé du nombre d'instances par classes ;
        - …
- une étude statistique des données et analyses de cette étude statistique
    - Par exemple
        - étude statistique pour chaque variable par rapport aux différentes classes ;
        - corrélation possible entre deux ou plusieurs variables ;
        - …
    - Dans cette partie, n'hésitez pas à utiliser des outils de visualisation des données.

Ainsi, la programmation avec Python doit être implémentée pour atteindre les objectifs du rapport.

J'attends également que vous programmiez **une** autre méthode de traitement, ou statistique ou de visualisation que nous n'avons pas vue en cours.

## 4. Précisions

En ce qui concerne l'ensemble de données (dataset), il n'y a aucune restriction. Vous allez chercher sur le Web un ensemble de données dans un domaine qui vous intéresse (ex. : bio-informatique, marketing, commerce, …). Ainsi, adapter votre code pour qu'il puisse traiter tout type de format de fichier commun (ex. : json, csv, …).