# Travail Pratique 3

**8INF436 – Forage des données**

***Professeur Julien Maitre, Ph.D.***
***Hiver 2021***

## 1. Description Générale

L'objectif général de ce travail pratique 3 (TP3) est de vous familiariser un peu plus avec les méthodes non supervisées et plus particulièrement avec les auto-encodeurs. En réalité, ce TP3 est l'exact même TP que le TP1. En d'autres termes, vous devrez détecter des anomalies par la reconstruction des instances après leur réduction de dimensionnalité.

D'une autre manière, vous ferez le même travail que dans le TP1, mais cette fois ci avec des auto-encodeurs. Vous pouvez utiliser le même ensemble de données, si vous le souhaitez, afin de comparer les performances.

## 2. Formalités

Le travail est à réaliser en groupe de **deux (il peut y avoir des exceptions)** et se composera d'un rapport scientifique et des scripts Python que vous aurez réalisé.

La date limite pour le rendu de votre travail est le **09 avril 2021 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 8INF436) dans la section « **TP3** ». Le nom du fichier prendra la forme de **Nom1_Nom2_TP3.zip**. N'oubliez pas d'inclure votre code permanent sur la page de garde du rapport scientifique. **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 (***il doit absolument posséder 2 classes***)
    - 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 de classes et leur nom/ou valeur;
        - le nombre d'instances total et le nombre d'instances par classe;
- 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;
        - y a-t-il des variables avec des valeurs catégoriques ?
            - *si oui : qu'avez-vous fait avec ces valeurs catégoriques ?*
        - est-ce que les classes sont débalancées, si oui qu'avez-vous fait ?
- une classification des instances du *testing dataset* selon la méthodologie du Tutoriel 1 ou du Chapitre 6
    - Vous testerez différents ratios (60-20-20 ;70-15-15) de division pour le ***dataset*** d'entrainement et celui de test.
        - Pour une comparaison, toujours avoir le même ***random seed*** et l'option ***stratify***.
    - Vous exploiterez les différentes architectures d'auto-encodeur.
    - Vous présenterez les différents résultats de la classification.
    - Vous tenterez d'expliquerer la raison des bonnes ou mauvaises performances.
- une conclusion de l'étude
    - Vous résumerez les informations essentielles de votre étude.
- une conclusion générale
    - Vous résumerez ce que vous avez apprécié, appris, moins apprécié par rapport aux données, les outils utilisés ou encore le travail demandé dans ce TP3.

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

De surcroît, j'aimerais observer une démarche reposant sur des hypothèses (issues de la nature des données ou et des résultats) et de validation des hypothèses en fonction des résultats que vous obtenez. ***Tout ceci est réalisé de manière itérative.***

Je veux également que votre code soit constitué d'un fichier `main` et d'autres fichiers où vous créerez des fonctions et/ou objets.

# 4. Précisions

En ce qui concerne l'ensemble de données (***dataset***), il doit y avoir au minimum 2 classes et avoir des variables/attributs avec des valeurs numériques (au mieux). Vous allez chercher sur le Web un ensemble de données dans un domaine qui vous intéresse pour plus de « fun » (ex. : bio-informatique, marketing, commerce, …). ***Néanmoins, j'interdis l'utilisation d'un jeu de données traitant de la COVID-19 et du domaine des jeux vidéo***. Voici un échantillon de liens Web qui donnent accès à des ensembles de données :

- https://www.data.gov/
- https://www.reddit.com/r/datasets/
- https://www.reddit.com/r/data/
- https://registry.opendata.aws/
- https://rs.io/100-interesting-data-sets-for-statistics/
- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/ml/datasets.php
- https://datasetsearch.research.google.com/
- etc.

# Annexe A

**Grilles des barèmes**

Bien que le TP3 ne représente que 20% de la notes finale, celui-ci est noté sur 100 points.

| **Rapport Scientifique**<br>**Points notés** | **Barème** |
|---|---|
| Structure (Plan) de votre rapport | 5/100 |
| Description de votre ensemble de données | 5/100 |
| Vérification et pré-traitement des données | 5/100 |
| Classification non supervisé | 60/100 |
| Conclusion de l'étude | 5/100 |
| **Total** | **80/100** |

| **Scripts Python**<br>**Points notés** | **Barème** |
|---|---|
| Structure de votre code | 15/100 |
| Niveau des commentaires | 5/100 |
| **Total** | **20/100** |

En ce qui concerne la qualité du Français, ce sont des points de pénalité pouvant aller jusqu'à 10/100.