

# TP : Kaggle

## 1 Travail demandé

### 1.1 Choix du jeu de données

La première étape consiste à choisir un jeu de données parmi les jeux de données disponibles sur le site Kaggle. Ce jeu de données doit présenter un problème de classification, c'est-à-dire prédire un label parmi une liste définie de labels. Le jeu de données doit faire au minimum 50mb, il n'y a pas de limite sur la taille maximale. Si le jeu de données est trop gros, vous pouvez utiliser un extrait (à faire vous même). Exemples (non limitatifs) de jeux de données :

- <https://www.kaggle.com/shayanfazeli/heartbeat>
- <https://www.kaggle.com/flo2607/traffic-signs-classification>
- <https://www.kaggle.com/malekzadeh/motionsense-dataset>
- <https://www.kaggle.com/mmoreaux/audio-cats-and-dogs>
- <https://www.kaggle.com/kinguistics/heartbeat-sounds>
- <https://www.kaggle.com/birdy654/eeg-brainwave-dataset-mental-state>

### 1.2 Analyse du jeu de données

L'objectif est d'écrire un script Python afin de traiter les données et de tester différentes méthodes vues en cours à l'aide du package scikit-learn, au minimum :

- Plus proche(s) voisin(s)
- Arbres de décision
- Classifieur bayésien
- SVM
- Réseaux de neurones

D'autres méthodes disponibles dans scikit-learn peuvent être utilisées. Pour chaque méthode, vous êtes invités à utiliser différents paramètres (par exemple la valeur de k dans le K plus proches voisins). Votre jeu de données doit être décomposé en jeu d'entraînement (train) et jeu de test (test). Si ce découpage est disponible sur le jeu de données choisi sur Kaggle, le conserver, sinon le créer vous même. L'objectif est d'obtenir la meilleure précision de classification sur les données.