# Travaux pratiques d'IA

## Série 4: Decision Trees

## Données

Les données à utiliser pour ce TP se trouvent dans les fichiers data.csv et data_test.csv. Les 5 premières colonnes spécifient les variables indépendantes tandis que la dernière colonne correspond à la variable dépendante (label). Le fichier data_test.csv sert à évaluer les performances des arbres développés à partir des données du fichier data.csv.

## 1 Entropie et gain d'information

1. Calculer l'entropie de la variable dépendante.
2. Calculer le gain d'information réalisé après l'application de trois critères de décision aléatoires.
3. Pour les mêmes critères de décision, calculer l'index Gini.
4. Quel est le critère de décision préférable selon le gain d'information ? Selon l'index Gini ?

## 2 ID3

1. Implémenter l'algorithme ID3 avec comme critères possibles le gain d'information et l'index Gini.
2. Comparer l'arbre obtenu à l'aide d'ID3 gain d'information avec celui produit par la démonstration.
3. Implémenter une procédure de génération de données à partir d'un arbre de décision.
4. À l'aide d'ID3 gain d'information, construire 5 arbres à partir d'échantillons aléatoires de 80% des données et utiliser comme prédiction finale un vote de majorité.
5. Comparer les performances du premier arbre obtenu avec celles de l'ensemble de 5 arbres selon les métriques suivantes: accuracy, precision, recall, F1 score.
6. Selon le F1 score, quel modèle devrait être privilégié ?

## 3 Aller plus loin avec des librairies (Optionnel)

1. Réaliser le tutoriel sur les arbres de décision de scikit-learn avec le jeu de données de votre choix (comportant également des variables continues).
2. Visualiser l'arbre de décision obtenu et calculer des métriques de performance sur des données d'évaluation.