# TP : Arbre de décision et Foret aléatoire

## Indications générales :

1. Le TP doit impérativement se faire en groupes de 2 à 3 personnes (selon la liste des groupes déjà fixée).
2. Le travail doit être démarré durant la séance de TP, à terminer chez soi pour être remis sur Campus avant le 21/10/2025 à 23h55.
3. Un compte rendu obligatoire en format PDF doit être soumis par chaque groupe avant le 21/10/2025 à 23h55.
4. Aucun dépôt en retard n'est autorisé, et les envois par email ne seront pas acceptés.
5. Dans le compte rendu vous présentez le code utilisé pour résoudre chaque partie (dans le cas où le code n'est pas donné) ainsi que les résultats obtenus et l'interprétation détaillée des résultats le cas échéant.
6. L'évaluation est principalement sur votre capacité d'analyser, de critiquer et d'interpréter les résultats. Ainsi, il est essentiel d'expliquer clairement vos conclusions.
7. Les codes sont donnés en R-Studio. Mais si vous êtes plus à l'aise avec un autre langage, n'hésitez pas à l'utiliser.

## Problème I : Arbres de décision, une application sur des données réelles :

Dans cet exercice, il s'agit d'expérimenter une implémentation de l'arbre de décision pour un problème de classification binaire.

1. A partir du répertoire en ligne "Échantillons de données" (voir sur Campus), choisissez un échantillon de données (Le choix de l'échantillon de données doit être communiqué à l'enseignant et validé avant de commencer le travail). Commencez par effectuer une analyse exploratoire descriptive de votre base de données (typologie des variables, valeurs manquantes, distributions des observations pour les différentes variables, boxplots, etc.).
2. Utilisez une implémentation de l'arbre de décision sous R (ou autre) pour construire un classifieur ayant pour objectif de prédire la classe de la variable dépendante dans votre base de données après avoir observé les variables indépendantes, en respectant les consignes suivantes :

   (a) Optimisez votre arbre de classification en passant par des techniques vues dans le cours (par exemple, l'optimisation des hyperparamètres par validation croisée, élagage, etc.).
   (b) Selon le mode opératoire, visualisez l'arbre optimal généré.
3. Vous serez évalué sur les résultats de votre modèle optimal **appliqué aux données test** :
   (a) Explorer le résultat de la classification :
      i. Detailed Accuracy By Class (Precision, Recall,...)
      ii. Confusion Matrix
      iii. Etc. (Je vous invite à calculer autant de métriques de performance que possible tout en justifiant les résultats obtenus à chaque fois.)
   (b) Veuillez inclure dans le compte rendu la courbe ROC et l'AUC de votre modèle optimal, ainsi que le code sous R correspondant.
   (c) Donnez une conclusion/interprétation globale par rapport aux résultats obtenus. N'hésitez pas à faire preuve de créativité et à aller au-delà des questions demandées. Toute idée d'analyse à valeur ajoutée sera fortement appréciée. ?
4. Enfin, comparez la performance de votre modèle avec un modèle de régression logistique classique, tel que développé dans le Problème I du TD, et interprétez les résultats.

## Problème II : Foret aléatoire, une application sur des données réelles :

Dans cet exercice, il s'agit d'expérimenter une implémentation d'une forets aléatoires de décision.

1. A partir du répertoire en ligne "Échantillons de données" (voir sur Campus), considérez le même échantillon de données du problème I. Utiliser différentes implémentations (au moins deux) du foret aléatoire sous R et pour chaque implémentation présentez-moi avec clarté les étapes de votre raisonnement menant au modèle optimal (Out-of sample estimation, Cross Validation, Élagage, optimisation des hyperparamètres, etc.)
2. Analysez et comparez les performances des différentes implémentations des forêts aléatoires (partie 1) à l'aide des indicateurs de classification tels que la précision, le rappel, la matrice de confusion, les courbes ROC et l'AUC, etc. Interprétez le résultat.

## Problème III : Détection des anomalies (Isolation Forest) :

Dans cet exercice, il s'agit d'implémenter une foret d'isolement pour calculer le score qu'une observation soit considérée comme anomalie :

1. Présenter le modèle de la foret d'isolement en quelques lignes.
2. À partir du répertoire en ligne "Échantillons de données"(voir sur Campus), l'échantillon de données intitulé « KPIs for telecommunication » est une base de données qui contient des informations sur les performances d'une cellule d'un réseau de télécommunications sur une période donnée. Cette base de données comprend dix variables numériques qui représentent des indicateurs clés de performance (KPI) pour évaluer divers aspects du réseau, tels que la qualité des appels, la stabilité de la connexion, le débit de données, etc. Ces variables permettent d'analyser et de surveiller la qualité et la fiabilité du réseau de télécommunications au fil du temps.
   (a) Analyse statistique : Faire l'étude statistique des 10 variables quantitatives de l'échantillon : moyenne – variance écart type – p-box ...
   (b) Diviser votre base de données en 70% pour l'apprentissage et 30% pour le test.
   (c) Penser à régler le problème des données manquantes.
   (d) Construisez une forêt d'isolement sur l'échantillon d'apprentissage et calculez le score d'anomalies pour les données de l'échantillon test. Vous êtes libre de choisir les paramètres de votre modèle afin d'optimiser la performance. Pour l'optimisation de la performance, je vous encourage à envisager des solutions astucieuses telles qu'une analyse graphique ou une analyse de la variance, etc.
   (e) Dans un premier tableau, donner les cinq observations avec les plus hauts scores d'anomalies et dans un second les cinq observations avec les scores les plus bas. Comparer, analyser et interpréter vos résultats.