# TP D'ÉVALUATION

## HETIC - DATA SCIENCE APPROFONDISSEMENT - MD4 - 2020-S1

**Sujet** : Prédiction de la probabilité d'attrition (churn ou désabonnement) des services de téléphonie et d'Internet à domicile d'une entreprise de télécommunications

L'objectif de ce TP est d'aider une entreprise de télécommunications à identifier ses clients qui ont une forte probabilité de se désabonner des services de téléphonie et d'internet à domicile.

Pour ce faire, on dispose d'une base de données avec les informations suivants :

- ***customerID*** : ID client
- ***gender*** : sexe du client (homme ou femme)
- ***SeniorCitizen :*** indique si le client est un senior (plus de 65 ans) ou non (1, 0)
- ***Partner*** : indique si le client a un conjoint ou non (Oui, Non)
- ***Dependents*** : indique si le client a des personnes à charge (enfants, parents, grand-parents) ou non (Oui, Non)
- ***Tenure*** : Nombre de mois pendant lesquels le client est resté abonné avec l'entreprise
- ***PhoneService*** : indique si le client a un abonnement téléphonique ou non (Oui, Non)
- ***MultipleLines*** : indique si le client a souscrit à plusieurs lignes ou non (Oui, Non, Pas de service téléphonique)
- ***InternetService :*** indique si le client a souscrit à un abonnement Internet (DSL, Fibre optique, Non)
- ***OnlineSecurity :*** indique si le client dispose d'une sécurité en ligne ou non (Oui, Non, Pas de service internet)
- ***OnlineBackup* :** indique si le client dispose d'une sauvegarde en ligne ou non (Oui, Non, Pas de service internet)
- **DeviceProtection :** Indique si le client souscrit à un abonnement de protection supplémentaire pour son équipement Internet fourni par la compagnie (Oui, Non, Pas de service internet)
- **TechSupport :** Indique si le client souscrit à un programme d'assistance technique supplémentaire de l'entreprise avec des temps d'attente réduits (Oui, Non, Pas de service internet)

- **StreamingTV :** Indique si le client utilise son abonnement Internet pour regarder des programmes en streaming de télévision provenant d'un fournisseur tiers : Oui, Non. L'entreprise ne facture pas de frais supplémentaires pour ce service.
- **StreamingMovies :** Indique si le client utilise son abonnement Internet pour regarder des films en streaming depuis un fournisseur tiers : Oui, Non.
  L'entreprise ne facture pas de frais supplémentaires pour ce service.
- **Contract :** Indique le type de contrat actuel du client : Mois par mois, un an, deux ans.
- **PaperlessBilling :** Indique si le client a opté pour la facturation électronique : Oui, Non
- **PaymentMethod :** Indique le mode de paiement de la facture par le client : Chèque électronique, Chèque postal, Virement bancaire (prélèvement automatique), Carte de crédit (prélèvement automatique)
- **MonthlyCharges :** Indique le montant mensuel total que le client paie actuellement pour tous ses services auprès de la société.
- **TotalCharges :** Indique le montant total des facturations du client
- **Churn :** Yes = le client s'est désabonné de l'entreprise, Non : le client est resté abonné dans l'entreprise (variable à expliquer)

## Tâches à faire :

### I. Data management

1. Faites une brève description de la base de données (nombre de lignes, de colonnes, % de churn)
2. Détectez et traitez les valeurs manquantes ou aberrantes (s'il y en a)
3. Enrichissez la base de données en créant des variables pertinentes au regard du churn

### II. Exploration des données

1. Proposez des graphiques (ou tableaux) pour regarder la corrélation/liaison entre le taux de churn (attrition) et les variables explicatives (démographiques et celles liées aux abonnements téléphoniques et internet)
2. Interprétez les résultats de ces analyses descriptives

### III. Modélisation

1. Construction de la base de modélisation (conversion de variables catégorielles en numérique, découpage de variables, transformation de variables, ….)

2. Scindez la base de modélisation en échantillon d'apprentissage (70%) et de test (30%)
3. Implémentez les modèles de machine learning suivants et optimisez-les avec une des méthodes d'optimisation vu en cours.
    a. Random forest
    b. XGBoost
    c. ANN (en optimisant le pas de descente de gradient)
4. Évaluez la performance des différents modèles sur le jeu de données test
5. Comparez les modèles et choisissez-en le meilleur au regard de la performance sur la base test
6. Effectuez une prédiction de la probabilité de churn (d'attrition) d'un client sur le jeu de données **Evaluation** fourni avec votre meilleur modèle

Bonne chance et bon courage ☺