**Machine Learning**

**Projet Covid-19**

**Démarche de travail**

1. Définir un objectif mesurable :
   1. **Objectif**: Prédire si une personne est infectée en fonction des données cliniques disponibles.
   2. **Métrique**:
      1. Précision : permet de réduire au maximum le taux de **Faux Positifs.**
      2. Recall : permet de réduire au maximum le taux de **Faux Négatifs.**
      3. Score F1 : permet de résumer les valeurs de la précision et du recall en une seule métrique.

Dans notre cas, nous allons choisir comme métrique un score F1 de 50% et Recall de 70%.

1. EDA (Exploratory Data Analysis)
   1. **Objectif** : Comprendre au maximum les données dont on dispose pour définir une stratégie de modélisation.
   2. **Checklist de base**:
      1. **Analyse de la forme**:
         1. Identification de la target
         2. Nombre de lignes et colonnes
         3. Types de variables
         4. Identification des valeurs manquantes
      2. **Analyse du fond**:
         1. Visualisation de la target (Histogramme / Boxplot)
         2. Compréhension des différentes variables (Internet)
         3. Visualisations des relations features-target (Histogramme / Boxplot)
         4. Identification des outliers
2. Pre-processing
   1. **Objectif**: transformer le data pour le mettre dans un format propice au Machine Learning.
   2. **Checklist de base**:
      1. Création du Train set/ Test set
      2. Elimination des NaN : dropna(), imputation, colonnes « vides »
      3. Encodage

* Premier modèle : Evaluation - Diagnostique
  + 1. Suppression des outliers néfastes au modèle
    2. Feature Selection
    3. Feature Engineering
    4. Feature Scaling

Suivre cette démarche : idée -> code -> évaluation -> repeat

1. Modelling
   1. **Objectif**: Développer un modèle de machine learning qui réponde à l’objectif final.
   2. **Checklist de base**
      1. Définir une fonction d’évaluation
      2. Entrainement de différents modèles
      3. Optimisation avec GridSearchCV
      4. (Optionnel) Analyse des erreurs et retour au Preprocessing / EDA
      5. Learning Curve et prise de décision

Pratiquez sur **Kaggle**