**PROJET MACHINE LEARNING : Analyse des Signaux Corporels Liés au Tabagisme**

#### 1. Introduction :

Le tabagisme demeure une menace majeure pour la santé publique mondiale, causant chaque année des millions de décès. L'analyse des signaux corporels associés au tabagisme peut jouer un rôle crucial dans l'identification des fumeurs et le développement de programmes d'intervention pour la prévention et l'arrêt du tabac. Ce projet vise à utiliser des techniques d'apprentissage automatique pour analyser et comprendre ces signaux corporels.

#### 2. Objectif du Projet :

L'objectif principal de ce projet est de développer des modèles prédictifs pour identifier les facteurs corporels associés au tabagisme. En analysant divers paramètres physiologiques, nous souhaitons déterminer les indicateurs significatifs du tabagisme et créer des modèles capables de prédire si une personne est fumeuse ou non.

#### 3. Source des Données :

Les données utilisées proviennent d'un jeu de données extrait de la plateforme Kaggle, spécifiquement d'une enquête sur le tabagisme menée auprès d'une population donnée. Le jeu de données inclut des variables telles que le sexe, l'âge, le niveau d'hémoglobine,entre autres. Ces informations ont été recueillies dans le cadre d'une étude clinique visant à explorer les corrélations entre les paramètres physiologiques et le comportement tabagique.

#### 4. Analyse Exploratoire des Données :

La première étape de notre analyse a consisté à explorer les données pour comprendre la distribution des variables et identifier les tendances et corrélations potentielles.

##### *4.1 Visualisation des Données*

Nous avons utilisé un outil de visualisation pour examiner les distributions des variables et leurs relations avec le tabagisme :

* **Diagrammes en boîte (box plots)** : pour comparer la distribution des variables entre les fumeurs et les non-fumeurs en fonction de la quantitee d’himoglobine

##### *4.2 Résultats de l'Analyse Exploratoire*

L'analyse exploratoire a révélé plusieurs tendances intéressantes :

* **L'âge** : Les fumeurs ont tendance à se concentrer dans certaines tranches d'âge.
* **Niveau d'hémoglobine** : Une différence notable a été observée entre les niveaux d'hémoglobine des fumeurs et des non-fumeurs.

#### 5. Description de la Phase de Pré-Processing des Données

La phase de prétraitement des données est cruciale pour préparer les données brutes en vue de l'analyse par les modèles d'apprentissage automatique. Cette phase comprenait plusieurs étapes :

##### *5.1 Suppression des Variables Inutiles*

Nous avons éliminé les variables qui n'étaient pas pertinentes pour l'analyse ou qui présentaient trop de valeurs manquantes.

##### *5.2 Traitement des Valeurs Manquantes*

Les valeurs manquantes ont été traitées de différentes manières, selon la nature des données :

* **Imputation** : Remplacement des valeurs manquantes par la moyenne, la médiane ou le mode de la variable concernée.
* **Suppression** : Élimination des enregistrements contenant trop de valeurs manquantes.

##### *5.3 Normalisation des Données*

Les variables numériques ont été normalisées pour garantir que toutes les variables aient une importance équivalente dans les modèles prédictifs.

##### *5.4 Transformation des Variables Catégorielles*

Les variables catégorielles ont été transformées en variables numériques en utilisant des techniques comme le codage à chaud (one-hot encoding) pour les intégrer dans les modèles prédictifs.

#### 6. Modèles Prédictifs

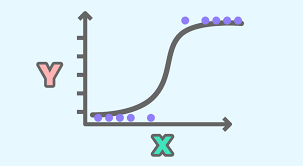
Nous avons développé plusieurs modèles prédictifs en utilisant divers algorithmes d'apprentissage automatique. Chaque modèle a été évalué pour déterminer sa capacité à prédire le statut de fumeur ou de non-fumeur.

## Les algorithmes utilises : phase de prediction

Nous allons construire trois modèles d'apprentissage automatique différents :

1. **Régression logistique**
2. **Forêt aléatoire**
3. **Gradient Boosting**

# Regression Logistique



La régression logistique est une méthode de classification qui prédit une catégorie

(0 ou 1) pour une variable cible en fonction de plusieurs variables explicatives.

**Entrée** : Variables explicatives (age, height, weight, waist, eyesight (left), eyesight (right), hearing (left), hearing (right), hemoglobin )

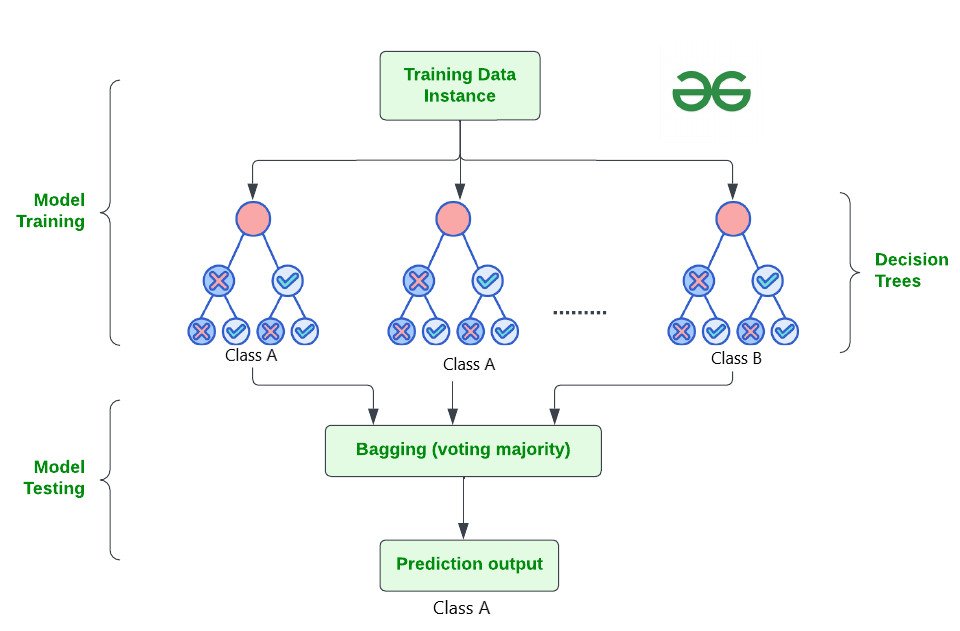
**Sortie** : Probabilité (entre 0 et 1) pour la variable cible (smoking)

**Décision** : Si la probabilité est supérieure à 0,5, la prédiction est 1, sinon 0

**Étapes** :

1. Créer la matrice de caractéristiques (X) et la variable cible (y).
2. Diviser les données en ensembles d'entraînement (80%) et de test (20%).
3. Entraîner le modèle sur l'ensemble d'entraînement.
4. Tester sur l’ensemble des tests

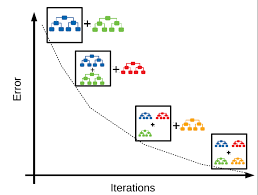
# Forêt Aléatoire



Le modèle de forêt aléatoire combine les prédictions de plusieurs arbres de décision pour faire des prédictions précises.

**Principe** : Créer un groupe d'arbres de décision où chaque arbre est entraîné sur un sous-ensemble aléatoire de données et de caractéristiques

# 3. Gradient Boosting



Le Gradient Boosting combine plusieurs modèles faibles pour créer un modèle prédictif fort. Il entraîne à plusieurs reprises de nouveaux modèles pour corriger les erreurs des modèles précédents.

**-Optimisation du Gradient Boosting** :

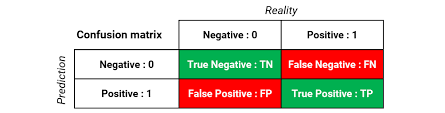
Pour optimiser les performances du modèle, un processus de réglage du modèle est effectué.

**Note** : Le Gradient Boosting est plus sensible aux valeurs aberrantes que la forêt aléatoire.

### Évaluations:

Pour évaluer les modèles, nous utiliserons les métriques suivantes :

**Matrice de confusion** : Permet de visualiser les performances du modèle en comparant les prédictions aux valeurs réelles.



**Score de précision** : Proportion de prédictions correctes parmi les prédictions faites.

Formule : Precision score= TP/TP+FP

**Score de rappel** : Capacité du modèle à identifier toutes les instances positives.

Formule : rappel score= TP/TP+FN

**Score F1** : Moyenne harmonique du score de précision et du score de rappel.

Formule : score F1= 2 × precision× Rappel/ precision+Rappel

#### 8. Résultats et Discussion

Les résultats de nos analyses ont montré que l'âge, le sexe, et le niveau d'hémoglobine sont des prédicteurs significatifs du tabagisme. Les performances des modèles étaient acceptables, avec des différences notables entre les différents algorithmes :

* **Régression Logistique** : Ce modèle a montré une bonne capacité à identifier les fumeurs, avec une précision raisonnable.
* **Forêt Aléatoire** : Ce modèle a offert une meilleure performance globale, en particulier en termes de rappel, grâce à sa capacité à capturer des interactions complexes entre les variables.
* **Gradient Boosting** : Bien que sensible aux valeurs aberrantes, ce modèle a présenté une très bonne précision et un excellent score F1 après optimisation des hyperparamètres.

#### 9. Conclusion

Ce projet a démontré l'efficacité de l'apprentissage automatique pour analyser les signaux corporels liés au tabagisme. Les modèles développés peuvent être utilisés pour identifier les fumeurs potentiels, contribuant ainsi à la mise en place de programmes de prévention et de cessation du tabac. Des travaux futurs pourraient explorer des sources de données supplémentaires et améliorer les modèles existants pour accroître encore leur précision et leur robustesse.

#### 10. Références

[DataSet: KAGGLE.]

RÉALISÉ PAR :

ACHOUR Oumaima

BOUSHABA Imane