

TP 4 : Détection de la maladie de Parkinson

L'ensemble de données sur la maladie de Parkinson a été créé par Max Little de l'Université d'Oxford, en collaboration avec le National Center for Voice and Speech, Denver, Colorado, qui a enregistré les signaux vocaux. L'étude originale a publié les méthodes d'extraction de caractéristiques pour les troubles généraux de la voix.

Informations sur l'ensemble de données :

Ce dataset est composé d'un ensemble de mesures biomédicales de la voix de 31 personnes, dont 23 atteintes de la maladie de Parkinson (PD). Chaque colonne du tableau est une mesure de voix particulière, et chaque ligne correspond à l'un des 195 enregistrements vocaux de ces individus. L'objectif principal des données est de discriminer les personnes en bonne santé de celles atteintes de la maladie de Parkinson, selon la colonne "statut" qui est définie sur 0 pour les personnes en bonne santé et 1 pour les malades.

Informations sur les attributs :

Matrix column entries (attributes):

name - ASCII subject name and recording number

MDVP:Fo(Hz) - Average vocal fundamental frequency

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency

MDVP:Flo(Hz) - Minimum vocal fundamental frequency

MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency

MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude

NHR,HNR - Two measures of ratio of noise to tonal components in the voice

RPDE,D2 - Two nonlinear dynamical complexity measures

DFA - Signal fractal scaling exponent

spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

status - Health status of the subject (one) - Parkinson's, (zero) - healthy

Exercice:

- 1- Étudier l'objet **naive_bayes.GaussianNB** de la librairie Python, scikit-learn.

https://scikit-learn.org/stable/modules/naive_bayes.html

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

- 2- Utiliser une **validation croisée** pour calculer l'erreur de classification (**ex.** moyenne de 10 validations en retenant à chaque fois 10% de données pour la validation).