Parkinson’s Diagnosis Predictive Modelling

# Background:

Parkinson’s Disease (PD) is a degenerative neurological disorder marked by decreased dopamine levels in the brain. It manifests itself through a deterioration of movement, including the presence of tremors and stiffness. This is generally marked by effect on speech, including dysarthria (difficulty articulating sounds), hypophonia (lowered volume), and monotone (reduced pitch range). Additionally, cognitive impairments and changes in mood can occur, and risk of dementia is increased.

Traditional diagnosis of Parkinson’s Disease involves a clinician taking a neurological history of the patient and observing motor skills in various situations. Since there is no definitive laboratory test to diagnose PD, diagnosis is often difficult, particularly in the early stages when motor effects are not yet severe. Monitoring progression of the disease over time requires repeated clinic visits by the patient.

An effective screening process for the Parkinson’s diagnosis, particularly one that doesn’t require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and non-invasive tool for diagnosis. We want to build a supervised learning model that could be applied to a voice recording dataset to accurately diagnosis PD, this would be an effective screening step prior to an appointment with a clinician.

# Objective:

To identify if a patient is suffering from Parkinson’s from the voice samples data in order to ease the screening process by avoiding clinical visits.

# Deliverables:

* Perform basic data pre-processing (if needed), univariate and bivariate analysis. Use relevant visualizations to understand the features at hand. Which features are strongly correlated to the target variable? - 15
* Build a pruned decision tree model and present the evaluation metrics. – 15
* Build all the ensemble models taught as a part of the curriculum and compare the models. - 20
* Present the model comparison in a data frame. - 5
* Comment on the codes and provide detailed explanation of the steps followed. – 5

# Data Attribute Information:

* 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,Shimm er: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