**Lama Ayash**

**Title:**

Parkinson's disease diagnoses prediction by biomedical voice measurements

**Introduction:**

The goal of this proposal is to share my experience of exploring and visualizing the data before starting a predictive analytics project. I hope to inspire you to get insights into data encouraged statisticians to pay more attention to this approach.

My work will be with the voice recording dataset from UCI machine learning repository. The project will be performed to identify Parkinson's disease (PD) patients from the healthy individuals by using voice recording data with machine learning algorithms.

This approach may have significant implications for early stage diagnosis of PD in a cost-effective manner.

**Problem:**

Parkinson's disease (PD) is an age-related neurodegenerative disorder affecting millions of elderly people world-wide. The early and accurate diagnosis of PD with available treatment might delay neurodegeneration and prevent disabilities. The existing diagnosis method such as brain scanning is an expensive process. The use of speech recognition for the diagnosis of PD patients could be less expensive.

**Objectives:**

1- Predict motor and total UPDRS scores (Unified Parkinson Disease Rating Scale)

2- Enhances prediction accuracy for diagnosing PD

3-Detecte PD early to have better experiences of care and lower treatment morbidity.

4- Reduce the budget of PD existing diagnosis method.

**Literature Review:**

There are countless studies that concern about diagnosis of PD. Yet, some of them needs a very long time or needs more budget. My project will enhance prediction accuracy for diagnosing PD in a short time and in a small-budget.

**Reaches Methodology**

The dataset is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. The recordings were automatically captured in the patient's homes.   
The dataset contains subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. Each row corresponds to one of 5,875 voices recording from these individuals.

**Tools:**

1-Python (Pandas, NumPy, Scikit-Learn, Statsmodels, Matplotlib, Seaborn)

2-Jupyter Notebook

3-Linear Regression ,Random Forest, Cross Validation and Grid Search

**Attribute Information:**

subject# - Integer that uniquely identifies each subject   
age - Subject age   
sex - Subject gender '0' - male, '1' - female   
test\_time - Time since recruitment into the trial. The integer part is the number of days since recruitment.   
motor\_UPDRS - Clinician's motor UPDRS score, linearly interpolated   
total\_UPDRS - Clinician's total UPDRS score, linearly interpolated   
Jitter(%),Jitter(Abs),Jitter:RAP,Jitter:PPQ5,Jitter:DDP - Several measures of variation in fundamental frequency   
Shimmer,Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,Shimmer:APQ11,Shimmer:DDA - Several measures of variation in amplitude   
NHR,HNR - Two measures of ratio of noise to tonal components in the voice   
RPDE - A nonlinear dynamical complexity measure   
DFA - Signal fractal scaling exponent   
PPE - A nonlinear measure of fundamental frequency variation

**References:**

The dataset was created by Athanasios Tsanas (tsanasthanasis '@' gmail.com) and Max Little (littlem '@' physics.ox.ac.uk) of the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals. The original study used a range of linear and nonlinear regression methods to predict the clinician's Parkinson's disease symptom score on the UPDRS scale.