A Brief report on Parkinson's Disease detection – Exploratory Data Analysis and Modelling

Submitted by-

Saurav Mobarsha – 214103438

Ashish Sharma – 214103331

Abstract:

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. There is commonly a marked 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, 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. If machine learning algorithms 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.

Introduction:

From the given dataset we have performed the EDA and plotted various histograms and graphs to determine the key features of any person with PD and tried to model that data in a Multi-layer perceptron.

Source:

The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

Data Set Information:

This dataset is composed of a range of biomedical voice measurements from people. Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). **The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.**

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column.

Research paper used for reference –

- **'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)**
- **Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering (to appear).**

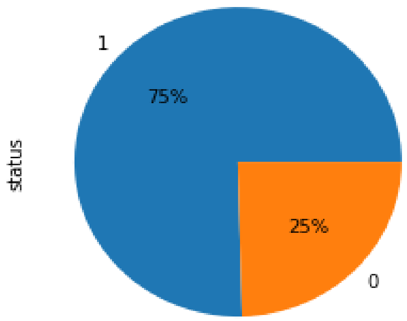
Attribute Information:

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
status - Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE,D2 - Two nonlinear dynamical complexity measures
DFA - Signal fractal scaling exponent
spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

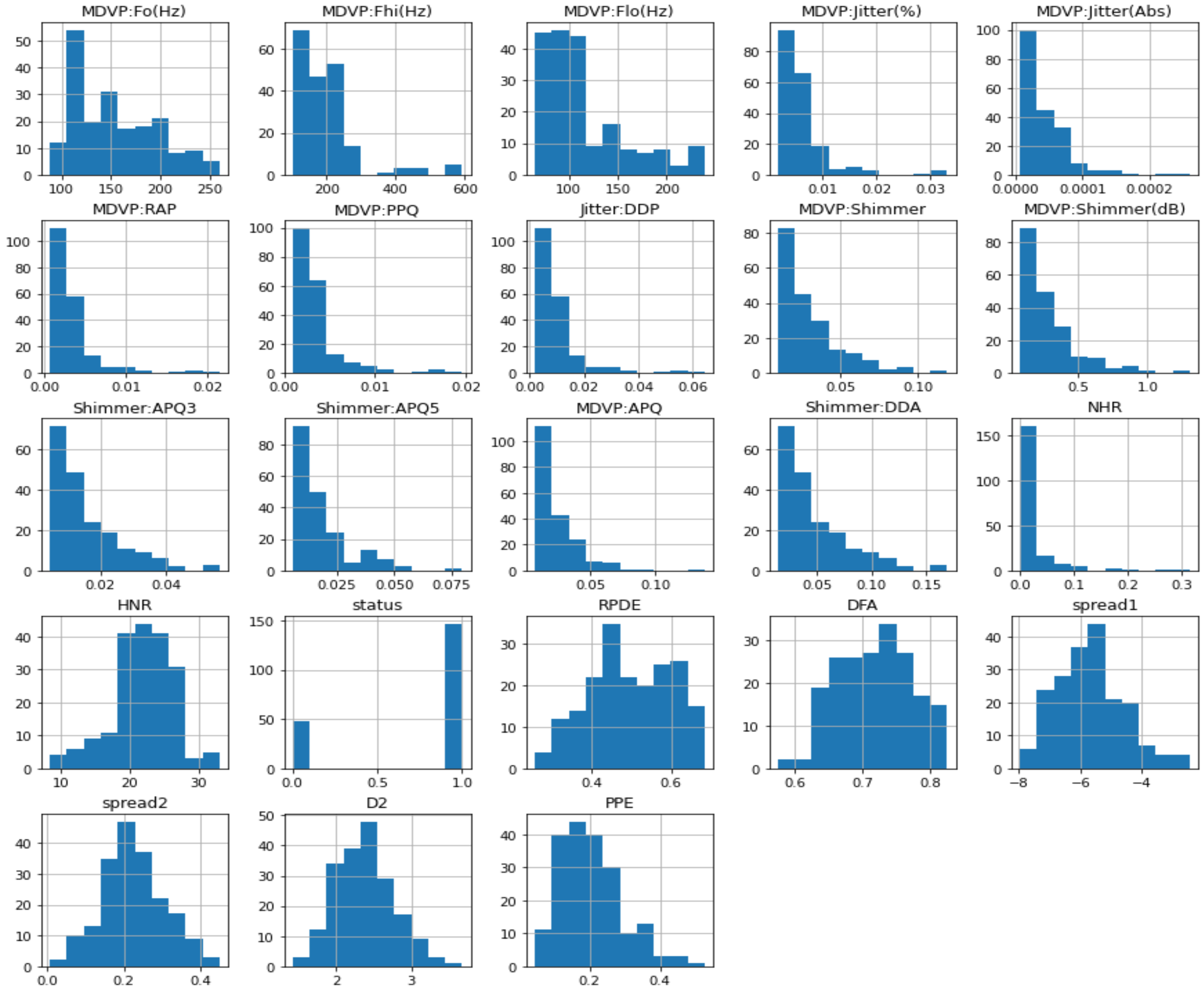
Observations & Results:

From EDA:

- There are 195 rows and 24 columns (22 float columns, 1 integer and 1 object column).
- There are some outliers as we can see some attributes have huge difference in their 75 percentile value and maximum value.
- From the pie chart we can observe that the target column i.e. status is imbalanced as 75% is for 1 and rest 25% is for 0.

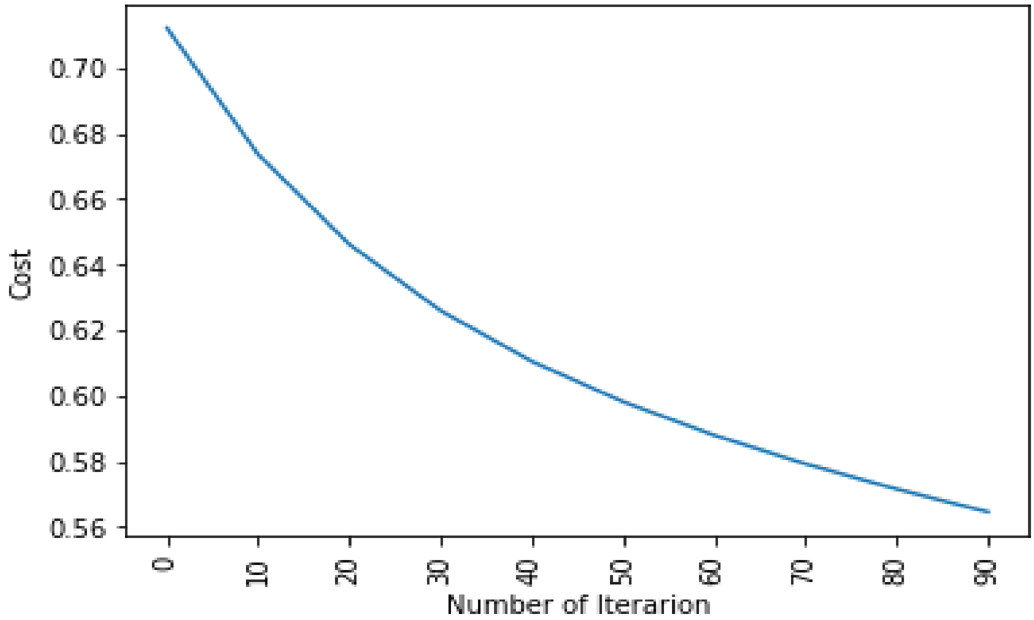


- In this we have 48 healthy patients and 147 parkinson disease affected patients.
- From the histograms we can see some of the data is normally distributed and most of the attributes are right skewed.

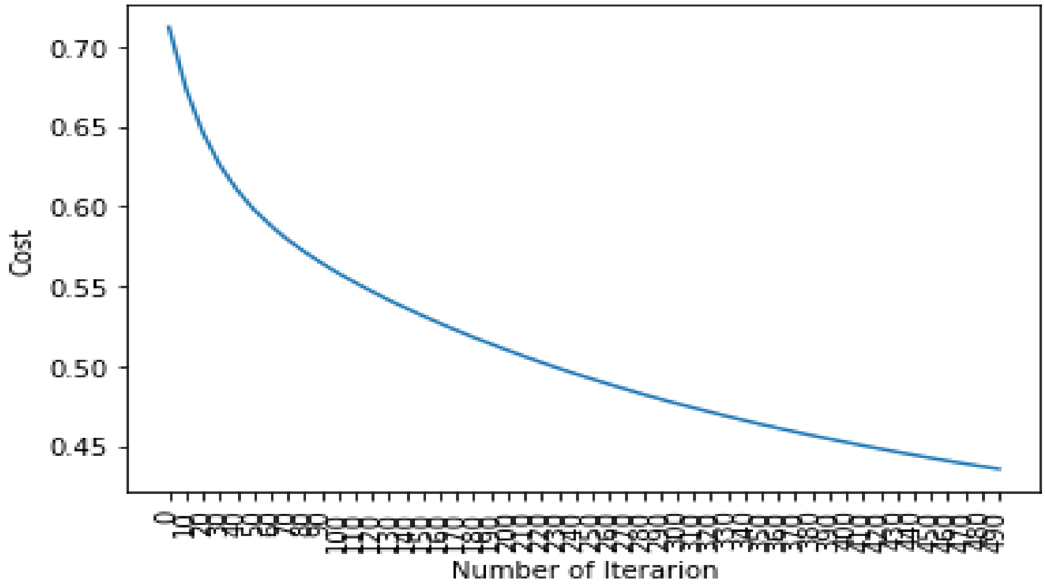


From MLP Modelling:

- Using 80% of the data to train the model we’ll use 156 datapoints to minimise the cost function and later remaining 20% will be used to test the model accuracy.
- Using sigmoid as activation function and learning rate as 0.01.
- After **100 iterations**:
  Cost: 0.564
  Train accuracy: 74.35897435897436 %
  Test accuracy: 79.48717948717949 %



- After **500 iterations**:
  Cost: 0.435
  Train accuracy: 85.25641025641025 %
  Test accuracy: 82.05128205128204 %



It can be seen that after increasing the iteration numbers the accuracy went upto 10%.

Conclusion:

Our final multi-layer ANN model can successfully predict the health status of the subject(patient)

(one) - Parkinson's

(zero) – Healthy

with 85% accuracy.