USING MACHINE LEARNING TO DIAGNOSE THE PATIENTS AFFECTED BY PARKINSON’S DISEASE

By
Gadidasu Pothana Shiva Prasad
LILIVERPOOL JOHN MOORES UNIVERSITY
Under supervision of
Archana Nanade

INTRODUCTION & PROBLEM STATEMENT

- Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects more number of people around the globe everyday.
- Patients diagnosed with PD have wide-ranging clinical characteristics, including motor and non-motor symptoms. Movement disorders including stiffness, resting tremor are some of the motor symptoms. Hallucinations, cognitive impairment, and impulse control disorders are some of non-motor symptoms.
- There will be a near doubling of the number of elderly Europeans and North Americans affected by PD by the year 2050. While prevalence rates were generally lower in Asia than elsewhere, this also made PD a critical concern in Asia, where half of the global ageing population.
- The precise cause of PD is still unknown, but various studies have shown that both genetic and environmental factors are involved
- It is essential to diagnose PD as early as possible to serve the general public in a cost-effective manner through latest techniques such as machine learning algorithms

AIM AND OBJECTIVE

- The primary aim of this study is to provide an improved algorithm based on machine learning techniques that aids clinicians in accurately diagnosing patients by that can distinguish between Parkinson's patients and healthy people.

- To treat missing and duplicate values accordingly. For this study, there are no missing or duplicate values.

- To bring numerical columns with various range to same range between 0 to 1 by using MinMax scalar which doesn’t affect the distribution of data.

- To understand and estimate the performance of model in terms of Sensitivity, Specificity, F1 score and Accuracy.

- Dealing with Imbalanced data by using ADASYN (Adaptive Synthetic Technique) to adaptively oversample minority class (class “0” or healthy) according to complexity of data.

LITERATURE REVIEW

2018

Machine learning for EEG-based biomarkers in Parkinson's disease by M. Isabel Vanegas, M. Felice Ghilardi and team

2019

Preliminary Study on Machine Learning Application for Parkinson's Disease Diagnosis by Ahmad Habbie Thias, Isca Amanda, Jessika, Navila Akhsanil Fitri and team

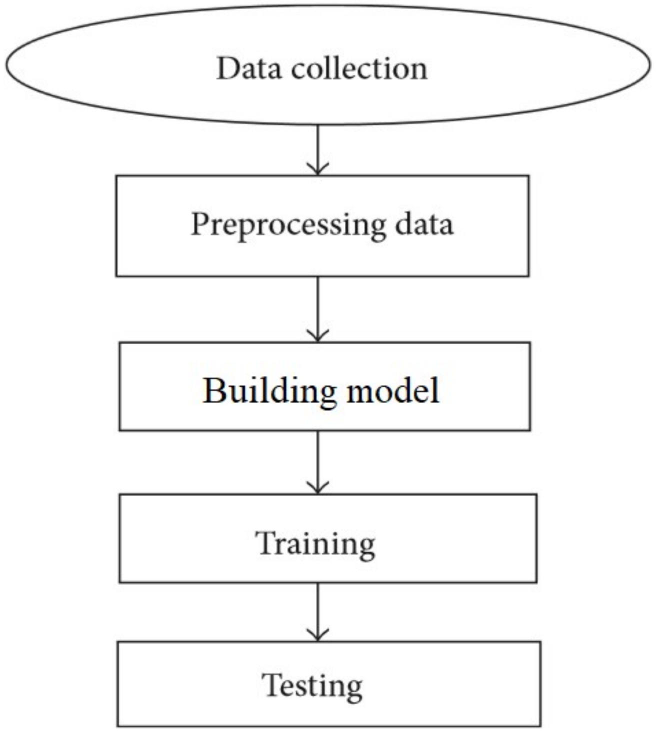
2020

Comparative Survey of Machine Learning Techniques for Prediction of Parkinson's Disease by Merry Saxena, Sachin Ahuja

2021

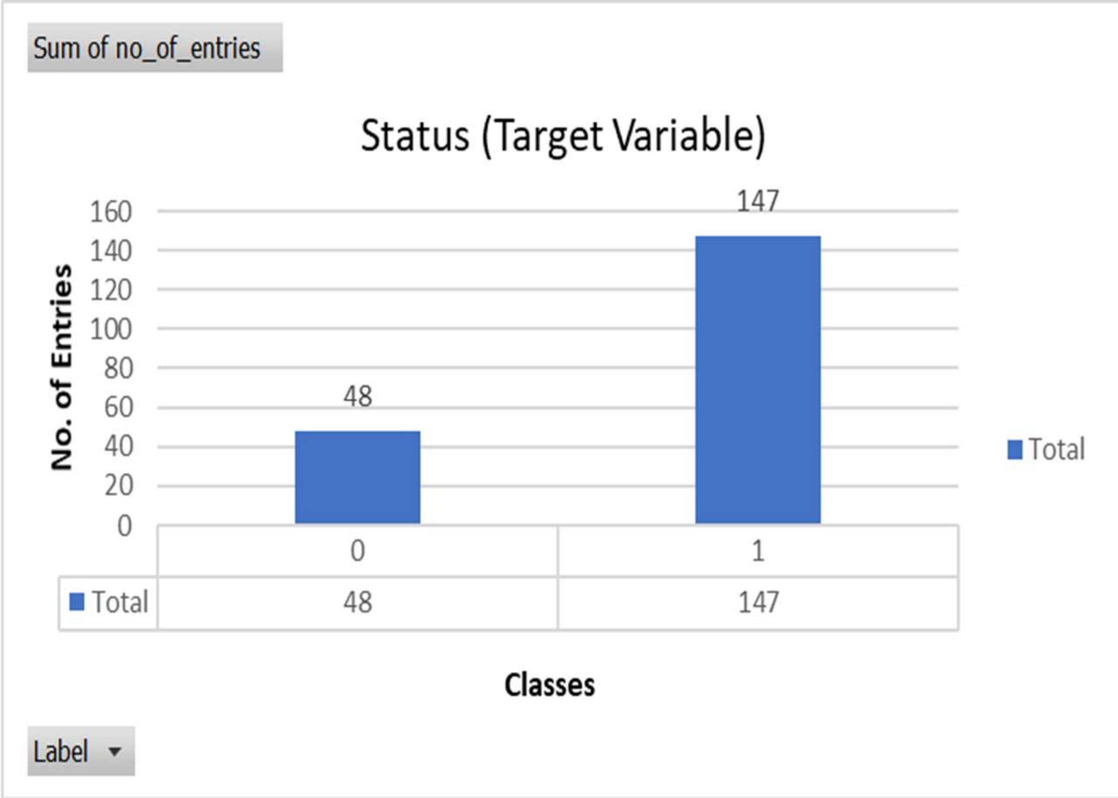
Prediction of Parkinson's disease and severity of the disease using Machine Learning and Deep Learning algorithm by Dr.Pooja Raundale, Chetan Thosar and team

MODEL DATA FLOW



EDA AND ANALYSIS

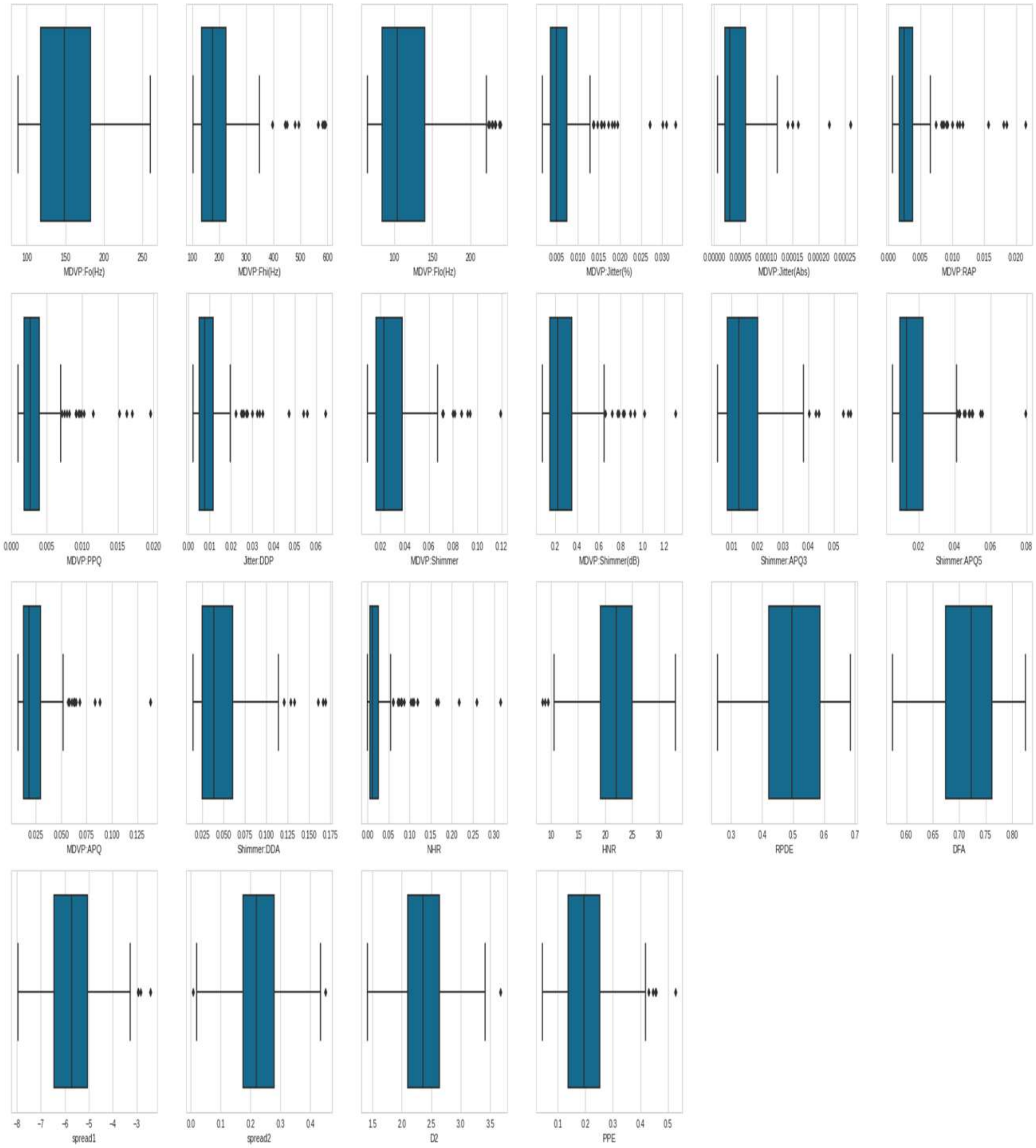
Label or Class	Class Description	No. of entries in Status (Target Column)
0	Healthy	48 (24.6%)
1	Parkinson's patient	147 (75.4%)



ANALYSIS (Univariate)

The columns with outliers

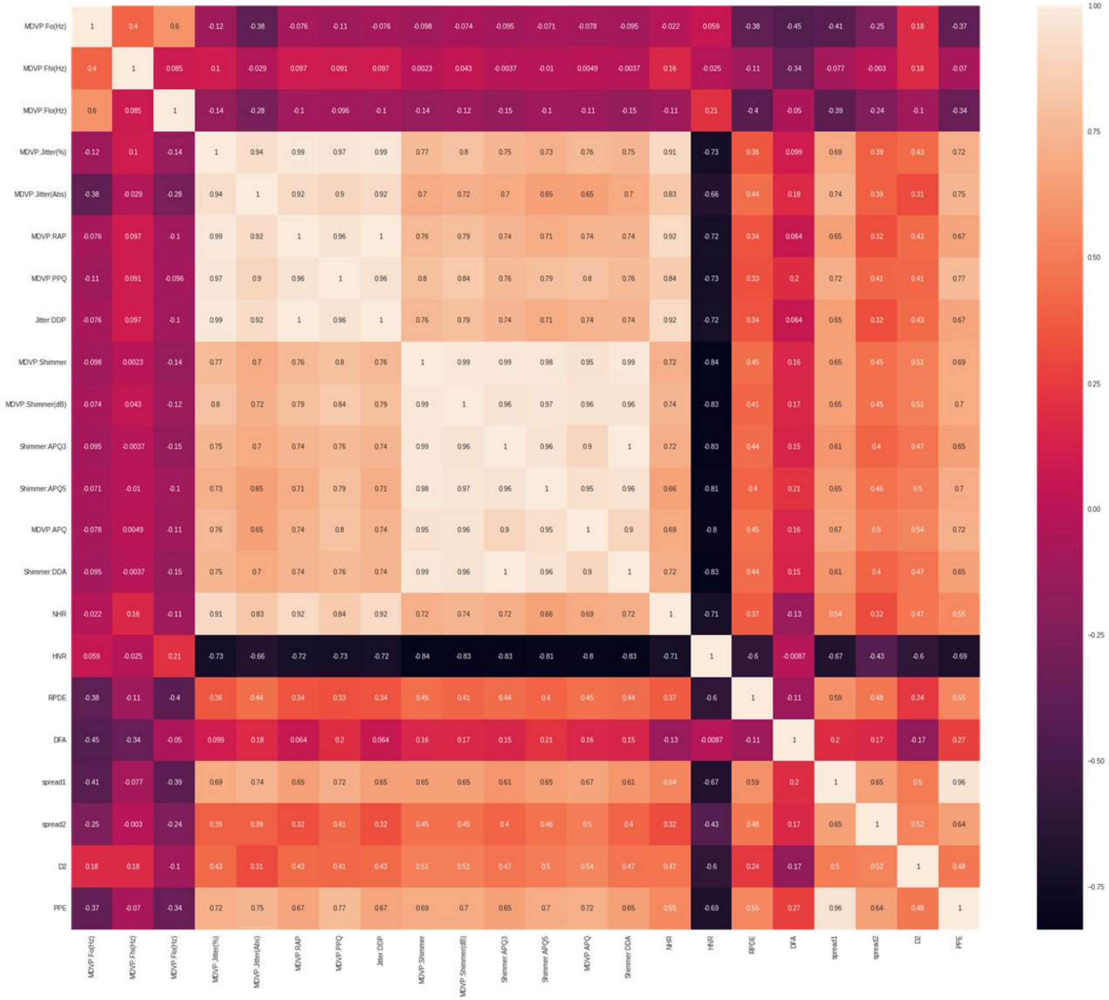
D2	MDVP:PPQ
PPE	Jitter:DDP
MDVP:Fhi(Hz)	MDVP:Shimmer
MDVP:Flo(Hz)	MDVP:Shimmer(dB)
MDVP:Jitter(%)	Shimmer:APQ3
MDVP:Jitter(Abs)	Shimmer:APQ5
MDVP:RAP	MDVP:APQ
Shimmer:DDA	NHR
spread1	HNR
spread2	



ANALYSIS (Bivariate Analysis)

The top 10 column pairs with highest positive and negative correlation (5 each)

Column A	Column B	Correlation Value
Shimmer:APQ3	Shimmer:DDA	1
MDVP:RAP	Jitter:DDP	1
MDVP:Jitter(%)	Jitter:DDP	0.99
MDVP:Jitter(%)	MDVP:RAP	0.99
MDVP:Shimmer	Shimmer:DDA	0.99
Column A	**Column B**	**Correlation Value**
MDVP:Shimmer	HNR	-0.835271
MDVP:Shimmer(dB)	HNR	-0.827805
Shimmer:DDA	HNR	-0.82713
Shimmer:APQ3	HNR	-0.827123
Shimmer:APQ5	HNR	-0.813753



Adaptive Synthetic (ADASYN)

- To overcome the Imbalanced data, we have two options
  - Under sampling the majority class
  - Oversampling the minority class
- ADASYN is an algorithm which tackles class imbalance by generating synthetic data k-Nearest Neighbors of each minority example.
- The advantage of using ADASYN over other algorithms is that it will not just copy the same minority data multiple times, ADASYN generates more data for complex parts of data which is tricky for ML algorithm to train on.

Label	No. of entries (Imbalanced Training Data)	No. of entries (ADASYN Training Data)
0	37	119
1	119	118

EVALUATION METRICS

- Sensitivity (PD patients correctly predicted as PD)
- Specificity (Healthy patients correctly predicted as healthy)
- F1 score
- Accuracy

TEST DATA MODEL RESULTS (IMBALANCE)

Model	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity	ROC
Random Forest	0.9	0.88	1	93.33	1	0.64	0.93
Logistic Regression	0.79	0.79	0.96	87.1	0.96	0.36	0.87
Decision Tree	0.92	0.93	0.96	94.74	0.96	0.82	0.89
SVM (Linear)	0.72	0.84	0.75	79.25	0.75	0.64	0.86
KNN	0.82	0.84	0.93	88.14	0.93	0.55	0.9
LightGBM	0.92	0.9	1	94.92	1	0.73	0.94
XGBoost	0.92	0.9	1	94.92	1	0.73	0.94

TEST DATA MODEL RESULTS (ADASYN)

Model	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity	ROC
Random Forest	0.87	0.87	0.96	0.9153	0.96	0.64	0.9
Logistic Regression	0.69	0.86	0.68	0.76	0.68	0.73	0.9
Decision Tree	0.82	0.84	0.93	0.8814	0.93	0.55	0.7
SVM (Linear)	0.74	0.85	0.79	0.8148	0.79	0.64	0.8
KNN	0.79	0.92	0.79	0.8462	0.79	0.82	0.9
LightGBM	0.9	0.88	1	0.9333	0.96	0.64	1
XGBoost	0.92	0.93	0.96	0.9474	0.98	0.87	0.9

CONCLUSION & FUTURE RECOMMENDATION

- There are no missing values and duplicate entries. Most of the data is numerical data. Hence, no need to label encode the data and it is easy for data pre-processing and model building.
- Used MinMax scaling to bring range of all numerical data to one range which is 0 to 1 without affecting the distribution of the data. ADASYN is used to tackle class imbalance
- Used various Classification machine learning algorithms and finally got an XGBoost algorithm on ADASYN data which gave highest sensitivity (0.98) and then highest specificity (0.87) which helps in predicting Parkinson's effectively.

FUTURE RECOMMENDATIONS

- The primary focus of future work should be on improving the performance of classification algorithms and also in using various other approaches from feature selection methods in order to bring out improved results.
- The current work only focused on the sensitivity of the model but the future work can also focus on improving the efficiency of model building both in terms of cost and time.