Parkinson's Disease Prediction

Table of contents

01

Parkinson Disease

Introduction

Parkinson's is a neurodegenerative progressive disorder disease. It affects the nervous system and the part of the body controlled by it.

Parkinson's disease is caused by the disruption of the brain cells that produce the dopamine which allows the brain to communicate to each other and control the fluency of the movement.

Symptoms

A person suffering from the Parkinson's can have following symptoms:

- Tremor
- Slowed movement
- Rigid muscle
- Impaired posture
- Loss of automatic movement
- Speech change
- Writing change
- Impaired posture

Also, Parkinson's disease can have other symptoms that include:

- Depression
- Anxiety
- Sleeping and memory-related issues
- Loss of sense of smell along with balance problems

Symptoms

The symptoms of Parkinson's disease can be different from patient to patient. Early sign of the disease can be mild and can go unnoticed. Symptoms often begin on one side of your body and usually remain worse on that side.

Sometimes it is difficult to detect whether there is Parkinson's disease present in the patient's body. Parkinson's disease if detected in the early stage will be curable, and will be time and cost effective, but there is no effective treatment in the advanced stage.

Project Introduction

.Info

The dataset was created by the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The dataset has 24 signal features. The target or independent variable is "status" with binary values of 0 and 1.

Status values for healthy person and PD person are 0 and 1 respectively. The goal of this project is to develop the best machine learning model to predict the Parkinson's disease so that we can treat the patient in timely manner.

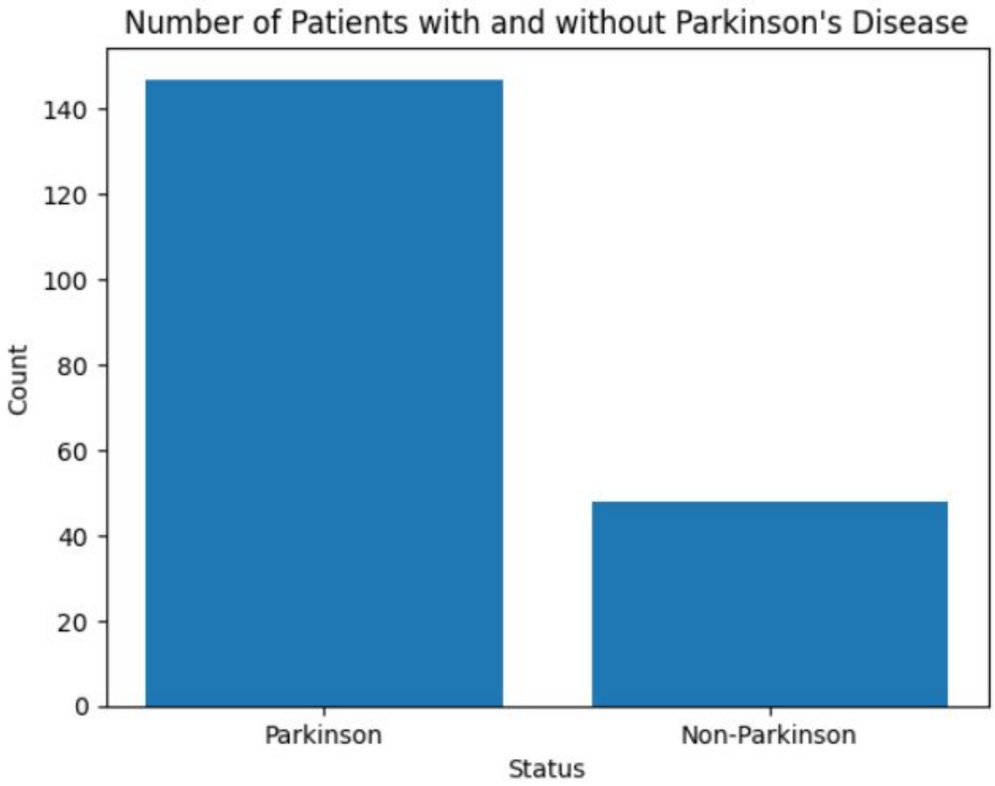
Citation - 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

Features Information:

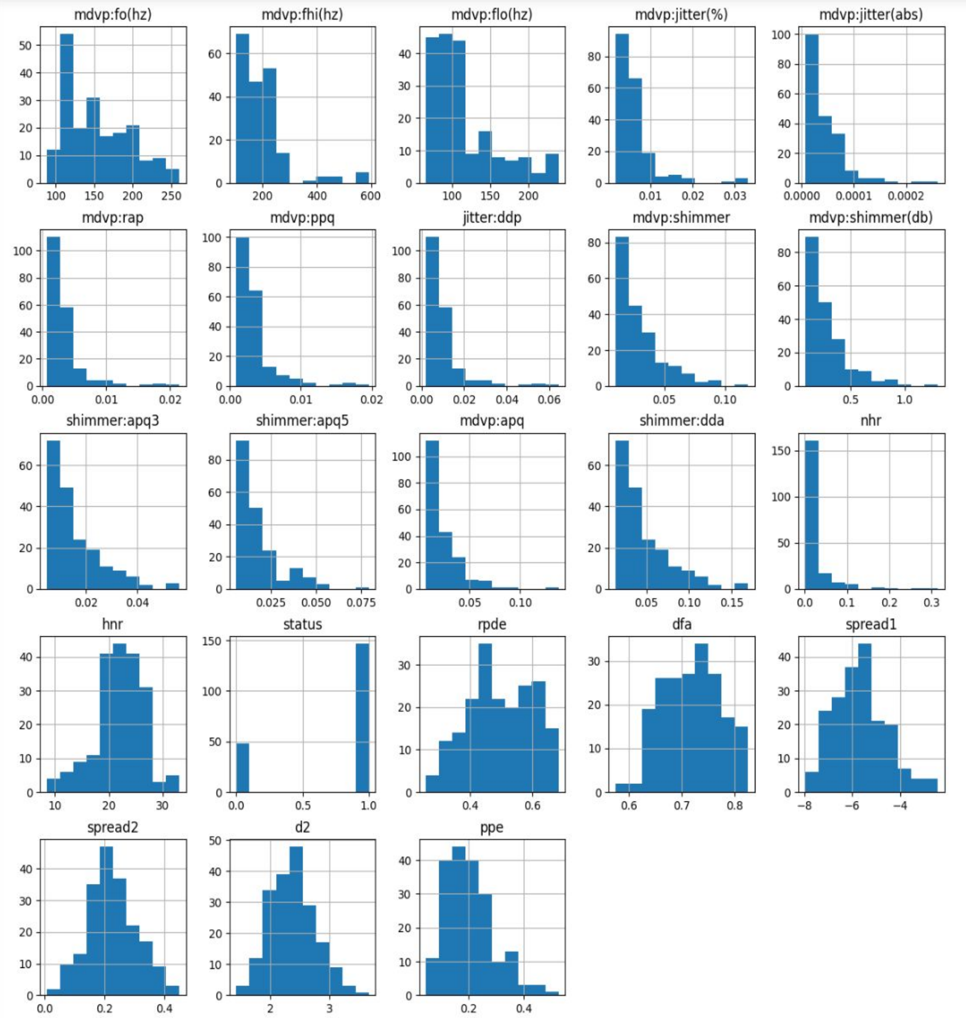
- name - ASCII subject name and recording number
- MDVP:Fo(Hz) - Average vocal fundamental frequency
- MDVP:Fhi(Hz) - Maximum vocal fundamental frequency
- MDVP:Flo(Hz) - Minimum vocal fundamental frequency
- MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency
- MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,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

	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	0.01109	0.04374
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	0.01633	0.05233
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425
...	...	...	...	...	...	...	...	...	...	...
190	phon_R01_S50_2	174.188	230.978	94.261	0.00459	0.00003	0.00263	0.00259	0.00790	0.04087
191	phon_R01_S50_3	209.516	253.017	89.488	0.00564	0.00003	0.00331	0.00292	0.00994	0.02751
192	phon_R01_S50_4	174.688	240.005	74.287	0.01360	0.00008	0.00624	0.00564	0.01873	0.02308
193	phon_R01_S50_5	198.764	396.961	74.904	0.00740	0.00004	0.00370	0.00390	0.01109	0.02296
194	phon_R01_S50_6	214.289	260.277	77.973	0.00567	0.00003	0.00295	0.00317	0.00885	0.01884

195 rows × 24 columns



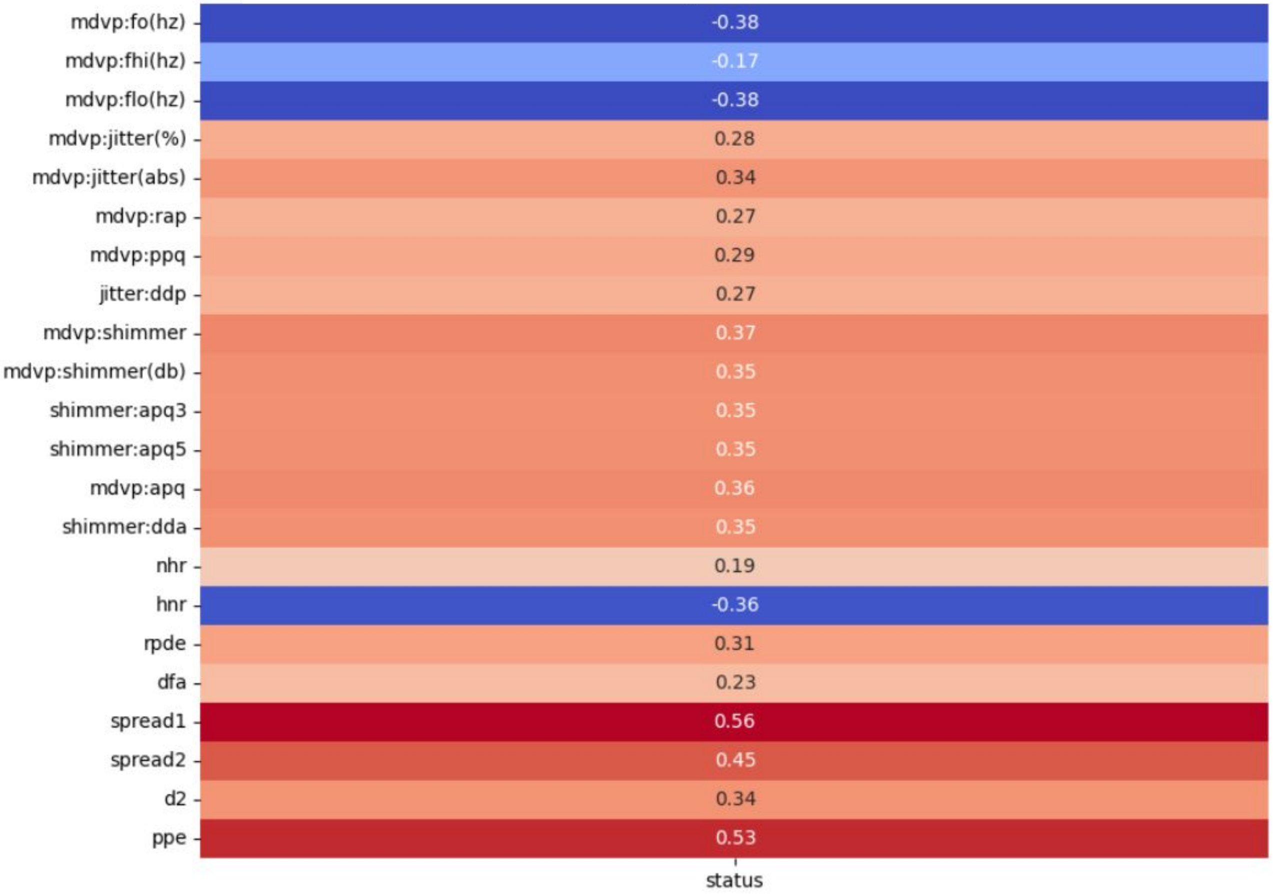
Here, we can see that we have a case of class imbalance which would be treated on later before modelling the data



This set of histograms shows the distribution of each numerical variable individually.

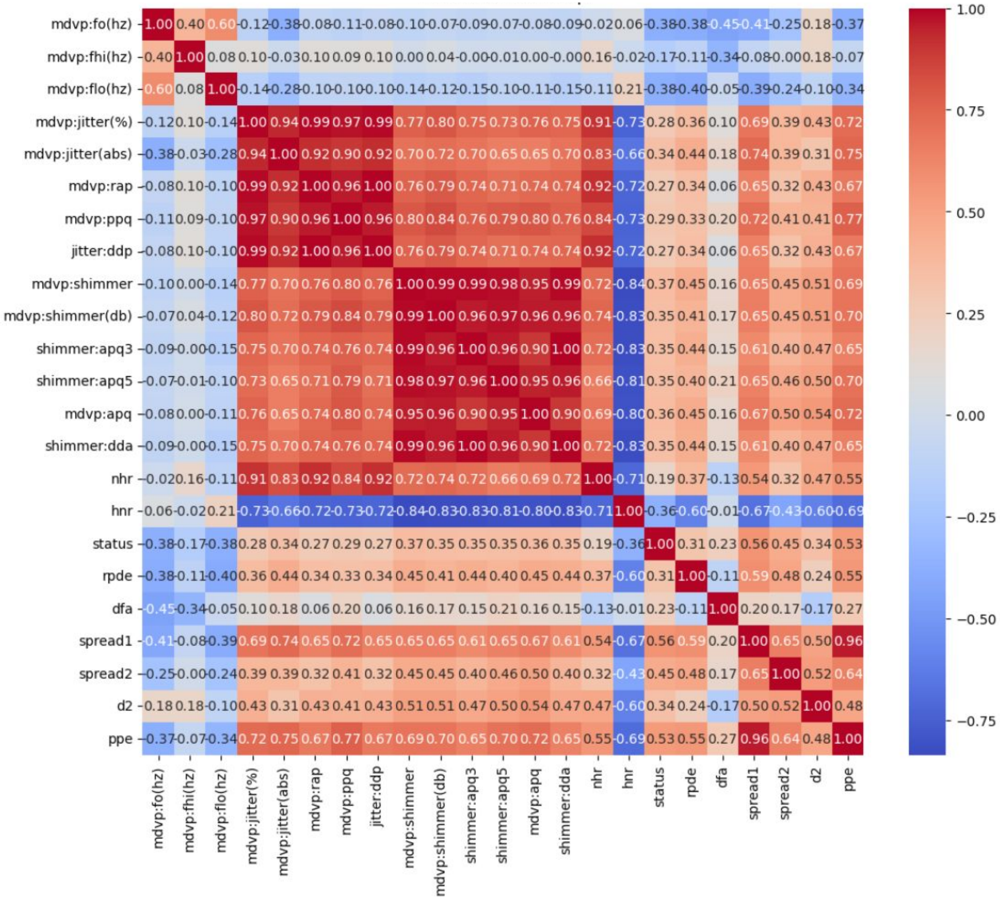
It helps understand the shape and spread of the data for each variable. Some columns are normally distributed and most of the attributes are right skewed

Correlation Heatmap: Status vs Other Columns



This heatmap shows the correlation between status and other columns.

Correlation Heatmap



This heatmap displays the correlation between each pair of numerical variables using color intensity.

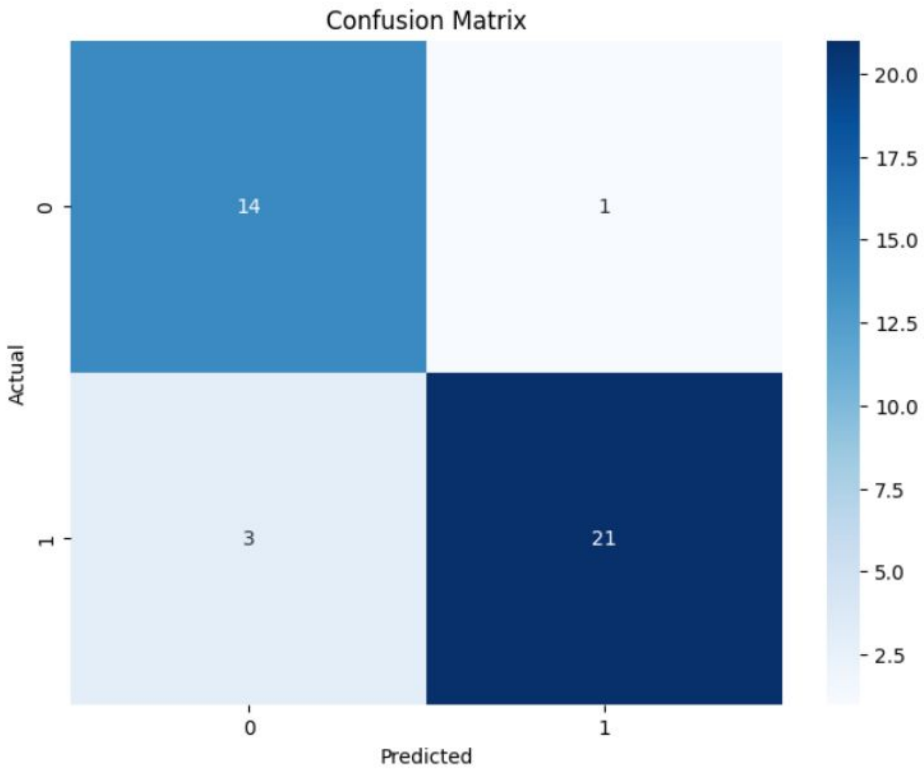
It provides insights into the strength and direction of the relationships between variables.

MACHINE LEARNING

STEPS TAKEN

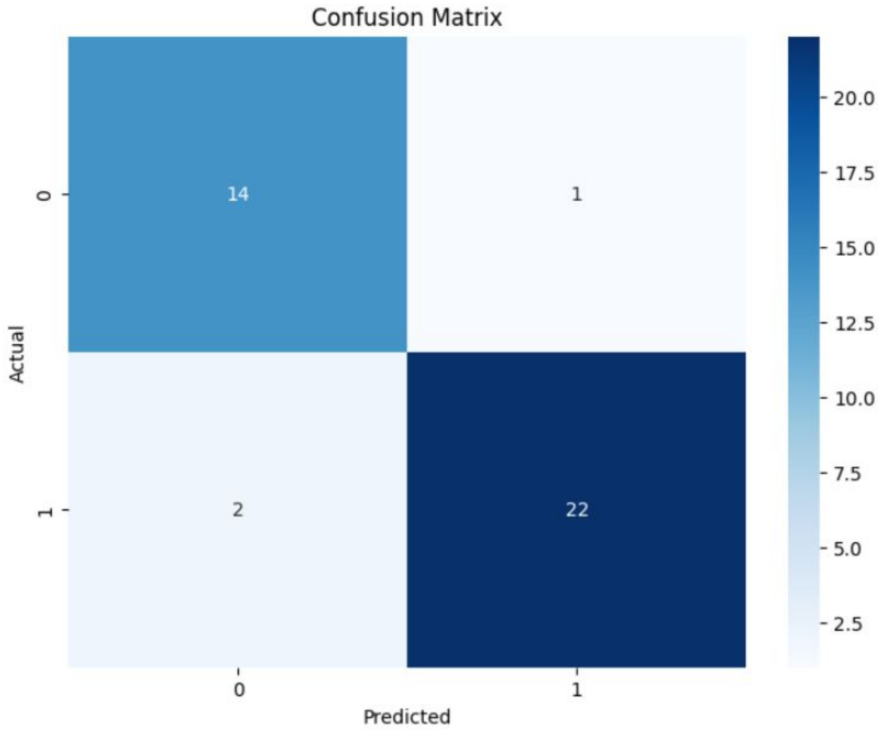
- Exploratory Data Analysis
- Feature Engineering
- Train Test Split
- Scaling the Data
- Balancing the Dataset
- Model Evaluation Function using Accuracy score, Confusion Matrix and Kappa score
- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- XG boosting

XGBoost



Accuracy: 0.9
Kappa Score: 0.79

Gradient Boosting



Accuracy: 0.92
Kappa Score: 0.84

Conclusion

The evaluation of the models was done using the evaluate_classification_model function, which provided a confusion matrix, accuracy score, and kappa score.

The results showed that both Gradient Boosting and XG Boost achieved high accuracy and kappa scores, with over 88% accuracy and a kappa score of over 80%.

THANK YOU!