Shreya Mishra

shreyamishra414@gmail.com

Report on Parkinson Prediction Model

Table of Contents

[Parkinson’s disease 2](#_Toc145191163)

[Citation for the dataset 2](#_Toc145191164)

[License 2](#_Toc145191165)

[Dataset 2](#_Toc145191166)

[Dataset Description 2](#_Toc145191167)

[Attribute Information 2](#_Toc145191168)

[Target Variable 3](#_Toc145191169)

[Models Used 3](#_Toc145191170)

[Analysis of the models and techniques used 3](#_Toc145191171)

[Distribution of Variables among people who have Parkinson vs those who do not 5](#_Toc145191172)

[Boxplot showing overall distribution of data variables 7](#_Toc145191173)

[Logistic Regression 8](#_Toc145191174)

Report on Parkinson’s Prediction Model

This project focuses on a classification model. We have applied different Machine Learning models to predict the presence of Parkinson’s disease in a patient.

# Parkinson’s disease

Parkinson’s disease is a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves. Symptoms start slowly. The first symptom may be a barely noticeable tremor in just one hand. Tremors are common, but the disorder may also cause stiffness or slowing of movement. Although Parkinson’s disease can’t be cured, medications might significantly improve your symptoms. Occasionally, your health care provider may suggest surgery to regulate certain regions of your brain and improve your symptoms.

# Citation for the dataset

'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)

# License

[GNU Free Documentation License 1.3](http://www.gnu.org/licenses/fdl-1.3.html)

# Dataset

The dataset is available at Kaggle

https://www.kaggle.com/datasets/gargmanas/parkinsonsdataset

# Dataset Description

The dataset consists of 24 columns and 195 records.

The dataset contains 23 attributes and 1 target variable.

# Attribute Information

1. name - ASCII subject name and recording number

2. MDVP:Fo(Hz) - Average vocal fundamental frequency

3. MDVP:Fhi(Hz) - Maximum vocal fundamental frequency

4. MDVP:Flo(Hz) - Minimum vocal fundamental frequency

5. MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP - Several measures of variation in fundamental frequency

6. MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA - Several measures of variation in amplitude

7. NHR, HNR - Two measures of ratio of noise to tonal components in the voice

8. status - Health status of the subject (one) - Parkinson's, (zero) - healthy

9. RPDE, D2 - Two nonlinear dynamical complexity measures

10. DFA - Signal fractal scaling exponent

11. spread1, spread2, PPE - Three nonlinear measures of fundamental frequency variation

12. Tonnetz - The set of pitch classes used to characterize each note

# Target Variable

status - Health status of the subject (one) - Parkinson's, (zero) - healthy

# Models Used

1. Logistic Regression

2. Decision Tree

3. Pruned Decision Tree

4. Random Forest

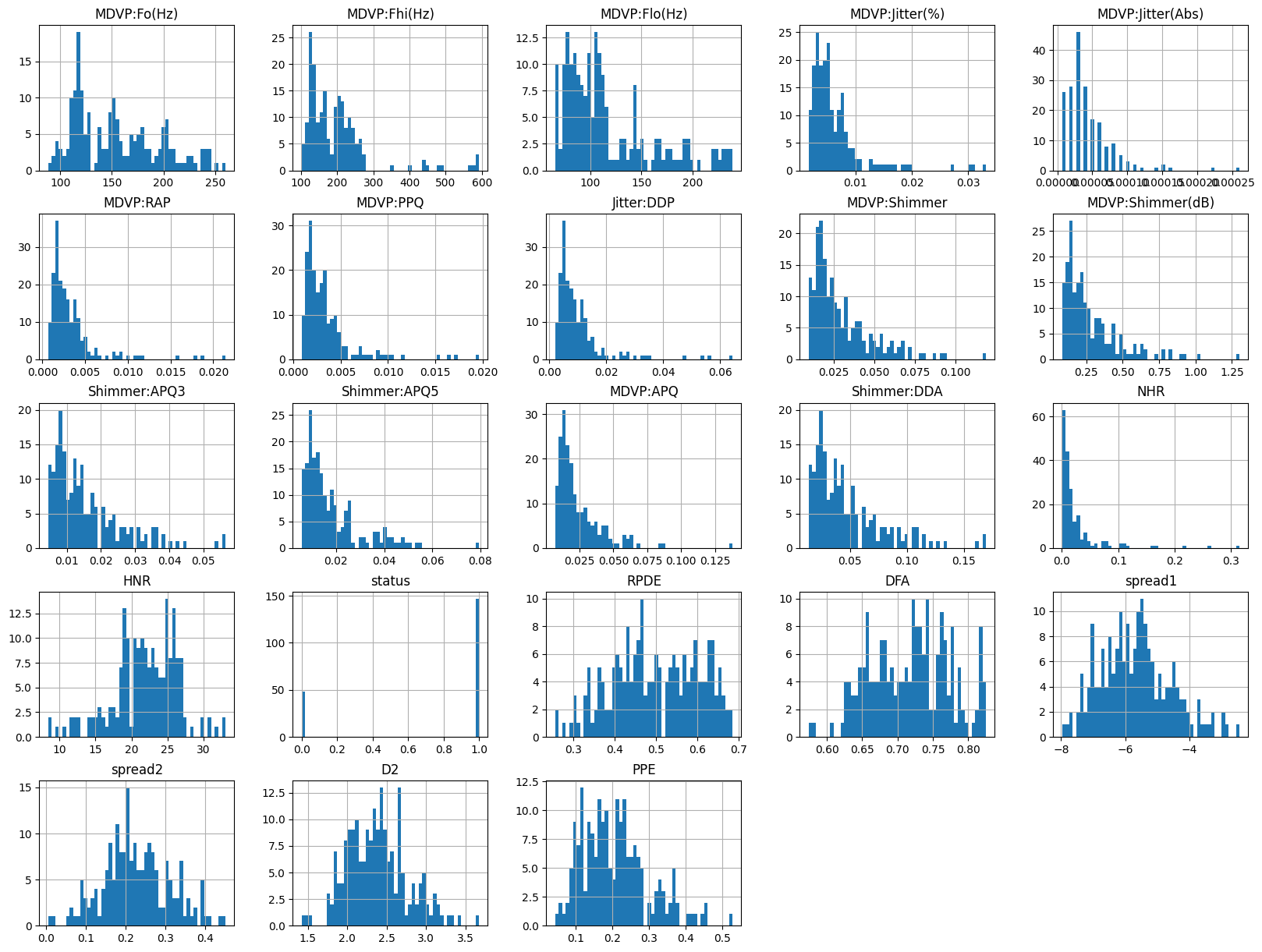
5. XGBClassifier

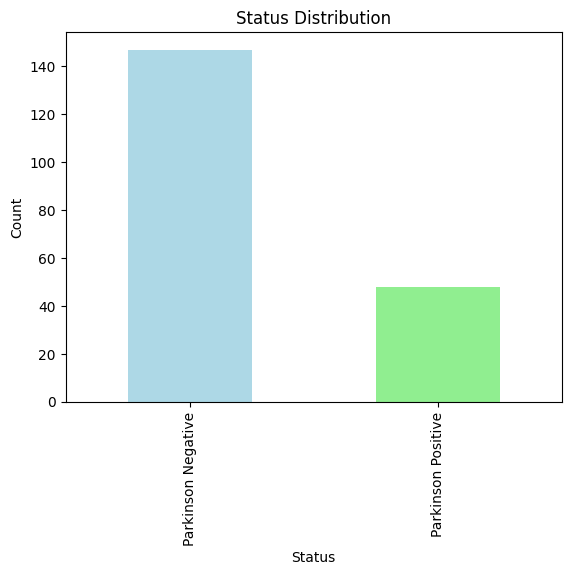
6. Support Vector Machine

# Analysis of the models and techniques used

We have 'name', 'MDVP:Fo(Hz)', 'MDVP:Fhi(Hz)', 'MDVP:Flo(Hz)', 'MDVP:Jitter(%)', 'MDVP:Jitter(Abs)', 'MDVP:RAP', 'MDVP:PPQ', 'Jitter:DDP', 'MDVP:Shimmer', 'MDVP:Shimmer(dB)', 'Shimmer:APQ3', 'Shimmer:APQ5', 'MDVP:APQ', 'Shimmer:DDA', 'NHR', 'HNR', 'status', 'RPDE', 'DFA', 'spread1', 'spread2', 'D2', 'PPE' as our columns in parkinsons.csv file. We have dropped the ‘name’ column from the dataframe. We see that all the variables except ‘status’ are continuous numerical variables. The ‘status’ is a categorical variable with values 1 and 0.

1. Parkinson Positive
2. Parkinson Negative





From the above diagram, we can conclude that our dataset is imbalanced. The people with Parkinson positives are only 48 in number out of 195. Thus, when evaluating the model's performance, accuracy alone may not be an appropriate metric.

Imbalanced datasets can lead to a bias in the trained model towards the majority class. In this case, the model may perform well at identifying individuals without Parkinson's disease but poorly at identifying those with the disease, which could be bad.

We might need to employ resampling techniques to address the class imbalance. These techniques include oversampling the minority class, undersampling the majority class, or using more advanced methods like Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples. We have used SMOTE technique later in the model training.

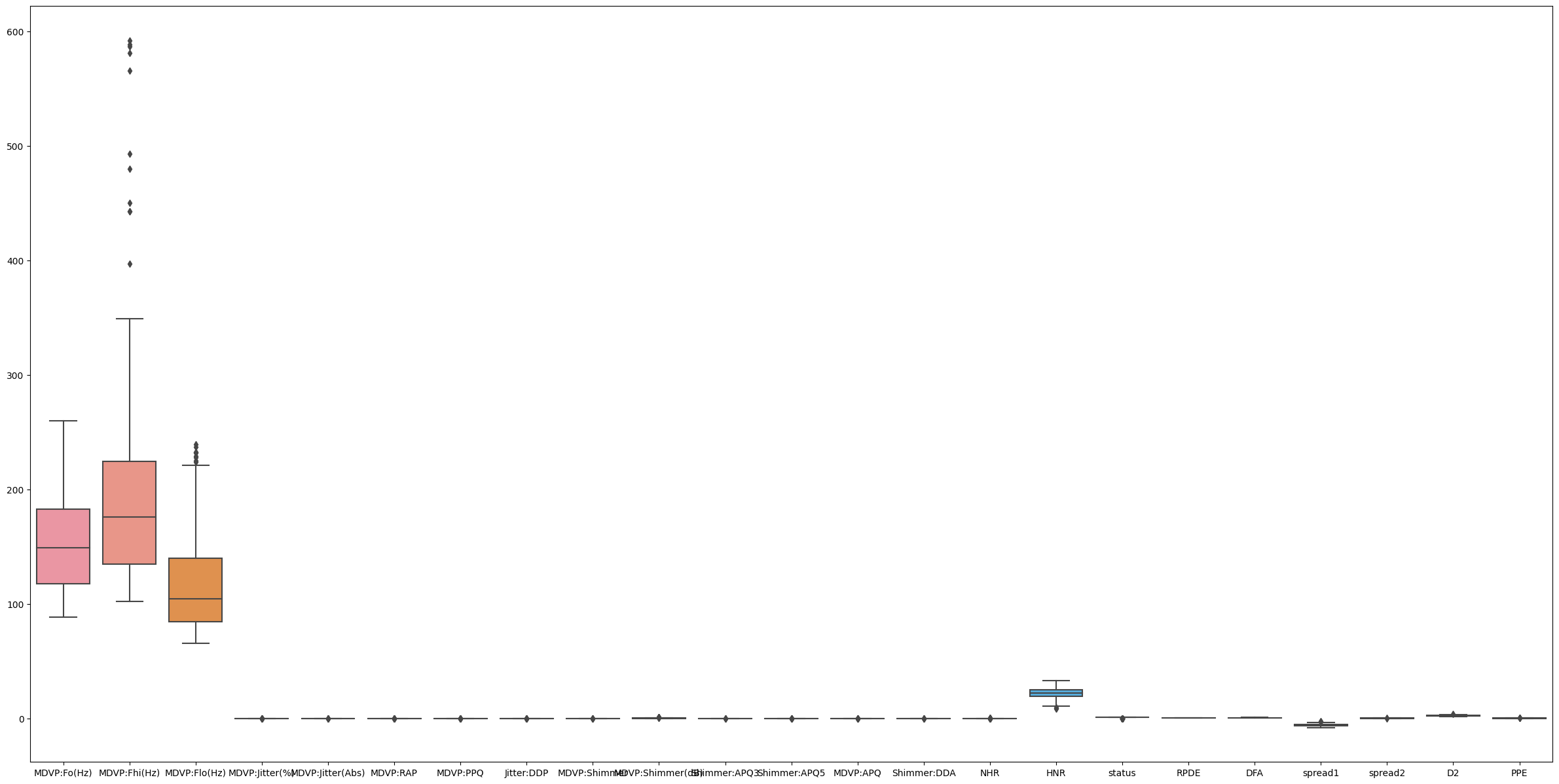
Machine learning algorithms may be more sensitive to class imbalance than others. For instance, decision trees can struggle with imbalanced data, while ensemble methods like Random Forests or Gradient Boosting can handle it better.

### Distribution of Variables among people who have Parkinson vs those who do not

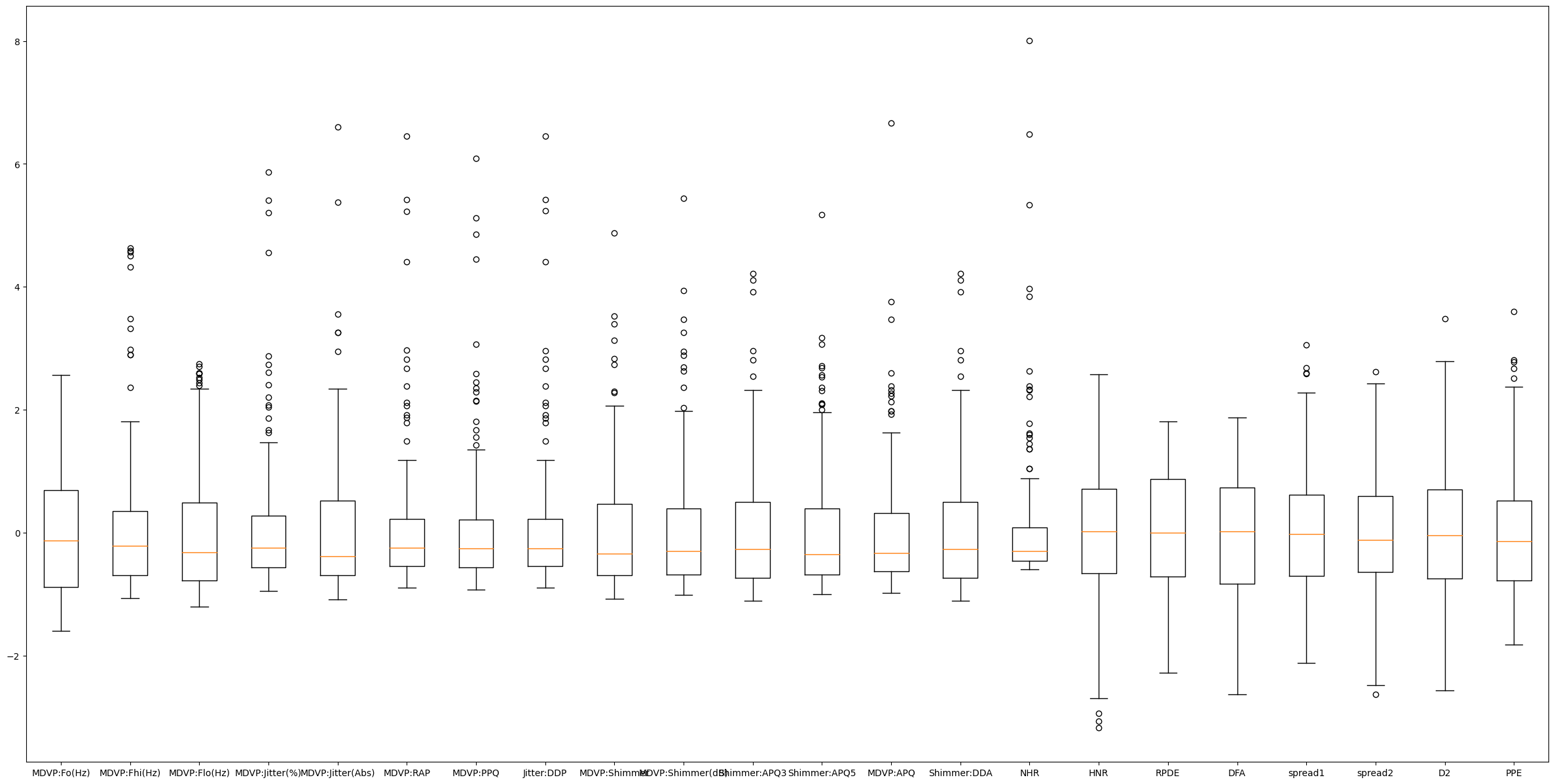
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

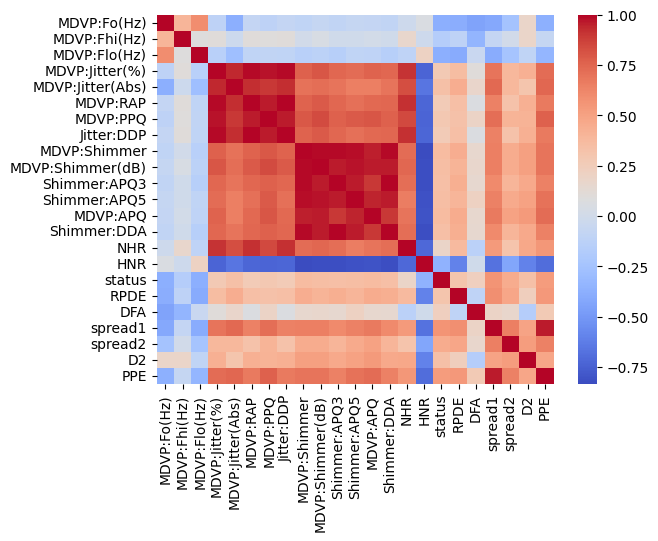
From the above visualizations, we can see that the variables are skewed and of different summary statistics when one has Parkinson vs when one does not. The difference in ranges of values such as mean, median, min value, max value, etc can help us determine whether a person has Parkinson or not.

### Boxplot showing overall distribution of data variables



The boxplot displayed above demonstrates the requirement for standardisation of our data. We used the Standard Scalar Library for our standardisation technique, standardising the independent variables using the Z score.





The correlation matrix above shows that we have a few variables which are strongly correlated among each other.

Removing highly correlated variables, also known as feature selection or dimensionality reduction, can be beneficial for several reasons in data analysis and modeling:

1. Avoid Multicollinearity
2. Reduced Overfitting
3. Makes data difficult to interpret the impact of individual features on the target variable
4. Reduces Noise
5. Avoids Curse of Dimentionality

We define a threshold for correlation as 0.80. Any variable having correlation greater than 0.80 and less than -0.80 are removed

The columns which are dropped are(Feature Selection): -

HNR  
Jitter:DDP  
MDVP:APQ  
MDVP:Jitter(Abs)  
MDVP:PPQ  
MDVP:RAP  
MDVP:Shimmer(dB)  
NHR  
PPE  
Shimmer:APQ3  
Shimmer:APQ5  
Shimmer:DDA

## Logistic Regression