PARKINSON DISEASE PREDICTION

IDS 506 | HEALTH INFO MANAGEMENT

Presented By Team 3

MARCH 14, 2022

Literature Review

Research has shown that changes in voice measures can occur years before the onset of clinical symptoms of Parkinson's disease. For example, one study found that individuals who later developed Parkinson's disease had lower pitch variability and increased pauses in their speech compared to healthy individuals. These changes were detected up to 10 years before clinical diagnosis.

Reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2784698/

All the data present is cleaned and contains different voice measures.

We didn't remove any variables duirng the literature study.

About the Dataset

- The dataset comprises biomedical voice measurements.
- Each column in the table is a particular voice measure
- Each row corresponds to one of the 195 phonetic recordings from individuals.
- All are continuous variables.

Attributes [All from Data set]

- **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**

Data QA and Info

Dataset Shape and Columns

```
df.shape
```

```
(195, 23)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   MDVP:Fo(Hz)       195 non-null    float64
 1   MDVP:Fhi(Hz)      195 non-null    float64
 2   MDVP:Flo(Hz)      195 non-null    float64
 3   MDVP:Jitter(%)    195 non-null    float64
 4   MDVP:Jitter(Abs)  195 non-null    float64
 5   MDVP:RAP          195 non-null    float64
 6   MDVP:PPQ          195 non-null    float64
 7   Jitter:DDP        195 non-null    float64
 8   MDVP:Shimmer      195 non-null    float64
 9   MDVP:Shimmer(dB)  195 non-null    float64
 10  Shimmer:APQ3      195 non-null    float64
 11  Shimmer:APQ5      195 non-null    float64
 12  MDVP:APQ          195 non-null    float64
 13  Shimmer:DDA       195 non-null    float64
 14  NHR               195 non-null    float64
 15  HNR               195 non-null    float64
 16  status            195 non-null    int64  
 17  RPDE              195 non-null    float64
 18  DFA               195 non-null    float64
 19  spread1           195 non-null    float64
 20  spread2           195 non-null    float64
 21  D2                195 non-null    float64
 22  PPE               195 non-null    float64
dtypes: float64(22), int64(1)
memory usage: 35.2 KB
```

- Each column in the table is a particular voice measure
- Each row corresponds to one of the 195 phonetic recordings from different individuals

```
df.isnull().sum()
```

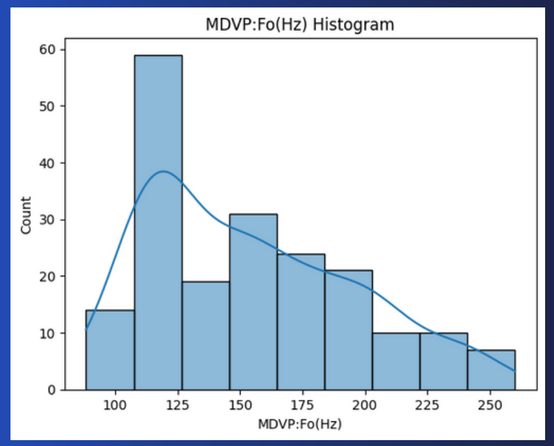
```
name              0
MDVP:Fo(Hz)       0
MDVP:Fhi(Hz)      0
MDVP:Flo(Hz)      0
MDVP:Jitter(%)    0
MDVP:Jitter(Abs)  0
MDVP:RAP          0
MDVP:PPQ          0
Jitter:DDP        0
MDVP:Shimmer      0
MDVP:Shimmer(dB)  0
Shimmer:APQ3      0
Shimmer:APQ5      0
MDVP:APQ          0
Shimmer:DDA       0
NHR               0
HNR               0
status            0
RPDE              0
DFA               0
spread1           0
spread2           0
D2                0
PPE               0
dtype: int64
```

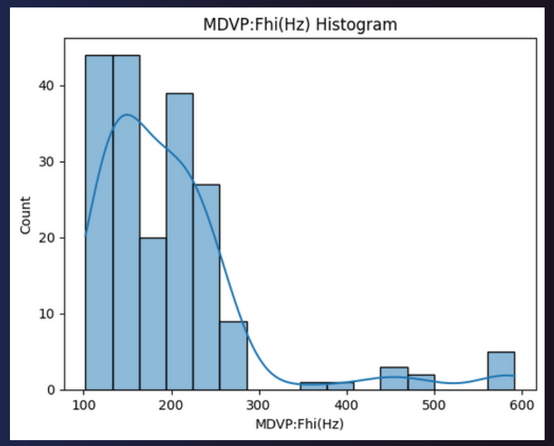
Univariate Analysis

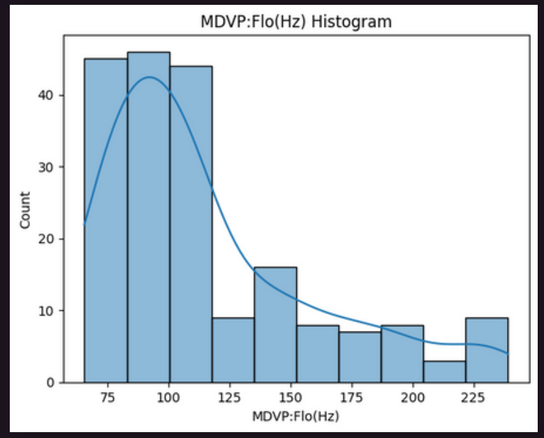
index	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:Shimmer(dB)	Shimmer:APQ3	Shimmer:APQ5
count	195.0	195.0	195.0	195.0	195.0	195.0	195.0	195.0	195.0	195.0	195.0	195.0
mean	154.22864102564102	197.10491794871797	116.32463076923077	0.006220461538461538	4.395897435897436e-05	0.003306410256410257	0.003446358974358974	0.009919948717948717	0.0297091282051282	0.2822512820512821	0.015664153846153845	0.017878256410256407
std	41.39006474907147	91.49154763503036	43.52141318199365	0.004841336926025 6	3.482190859976326e-05	0.002967744162016884	0.002758766469679313	0.008903344355858987	0.01885693185894681	0.19487729006053414	0.010153161595709018	0.012023705538741727
min	88.333	102.145	65.476	0.00168	7e-06	0.00068	0.00092	0.00204	0.00954	0.085	0.00455	0.0057
25%	117.572	134.8625	84.291	0.00346	2e-05	0.00166	0.00186	0.004985	0.016505	0.1485	0.008245	0.00958
50%	148.79	175.829	104.315	0.00494	3e-05	0.0025	0.00269	0.00749	0.02297	0.221	0.01279	0.01347
75%	182.769	224.2055	140.01850000000002	0.007365	6e-05	0.003835	0.003955	0.011505000000000001	0.037885	0.35	0.020265	0.02238
max	260.105	592.03	239.17	0.03316	0.00026	0.02144	0.01958	0.06433	0.11908	1.302	0.05647	0.0794

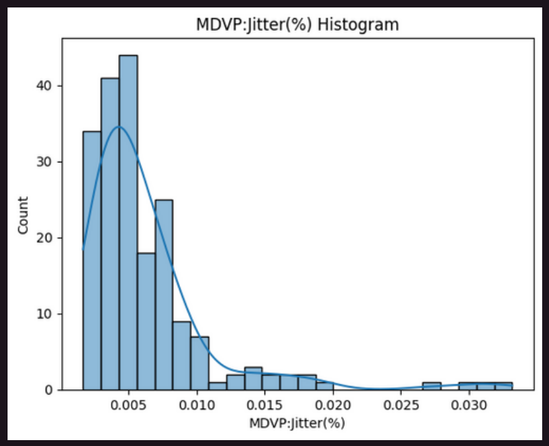
MDVP:APQ	Shimmer:DDA	NHR	HNR	status	RPDE	DFA	spread1	spread2	D2	PPE
195.0	195.0	195.0	195.0	195.0	195.0	195.0	195.0	195.000000	195.000000	195.000000
0.02408148717948718	0.04699261538461539	0.02484707692307692	21.885974358974355	0.7538461538461538	0.4985355384615385	0.7180990461538461	-5.684396743589745	0.226510	2.381826	0.206552
0.016946736247029432	0.030459119431240397	0.04041844855606928	4.425764269063427	0.43187803371226474	0.10394171413073468	0.0553358303465968	1.090207763740309	0.083406	0.382799	0.090119
0.00719	0.01364	0.00065	8.441	0.0	0.25657	0.574282	-7.964984	0.006274	1.423287	0.044539
0.01308	0.024735	0.005925	19.198	1.0	0.421306	0.6747575	-6.450096	0.174351	2.099125	0.137451
0.01826	0.03836	0.01166	22.085	1.0	0.495954	0.722254	-5.720868	0.218885	2.361532	0.194052
0.0294	0.060795	0.02564	25.075499999999998	1.0	0.5875625	0.7618815	-5.046192	0.279234	2.636456	0.252980
0.13778	0.16942	0.31482	33.047	1.0	0.685151	0.825288	-2.434031	0.450493	3.671155	0.527367

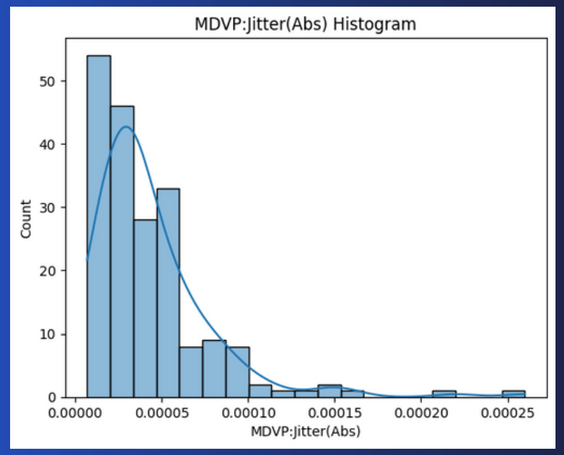
Univariate Analysis

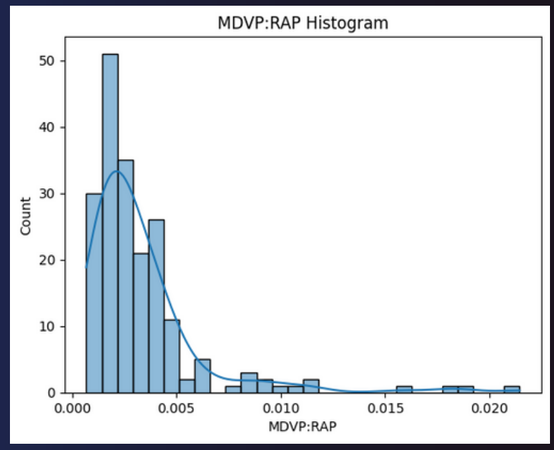


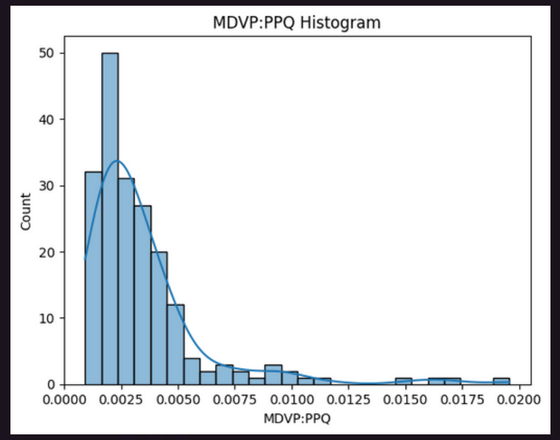


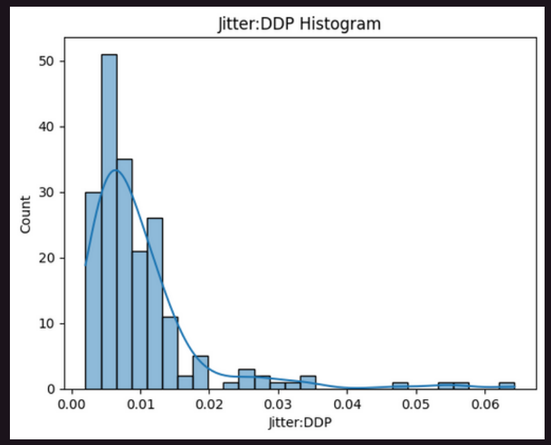




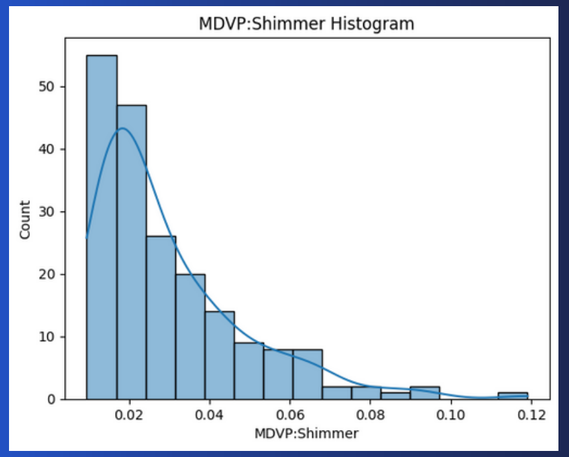


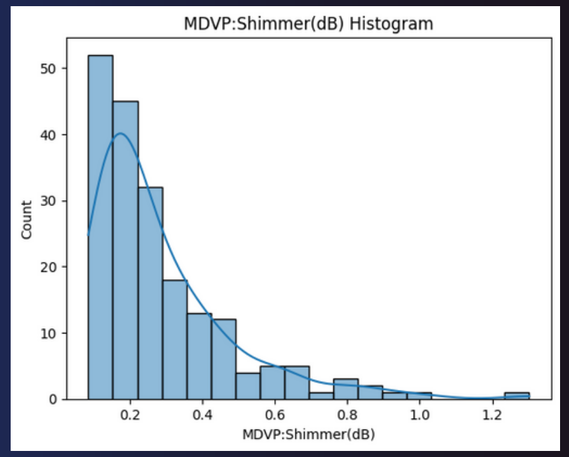


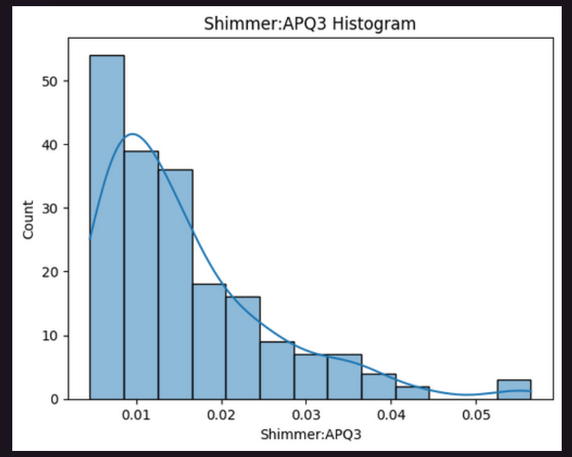


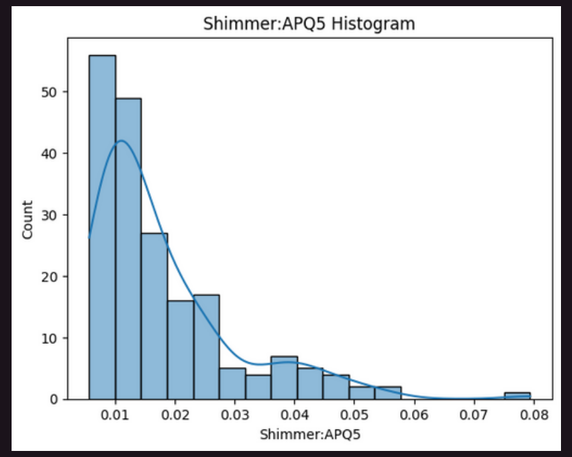


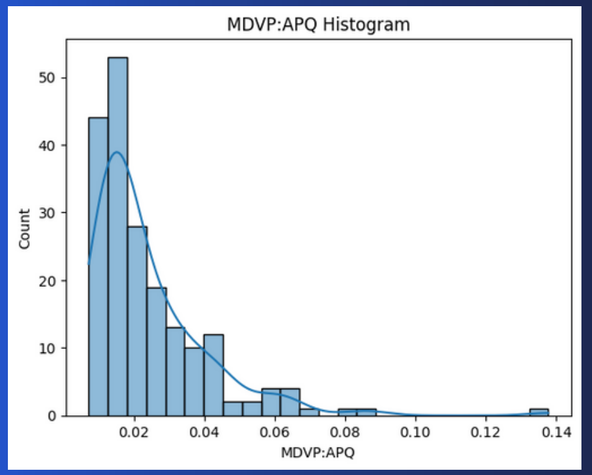
Univariate Analysis

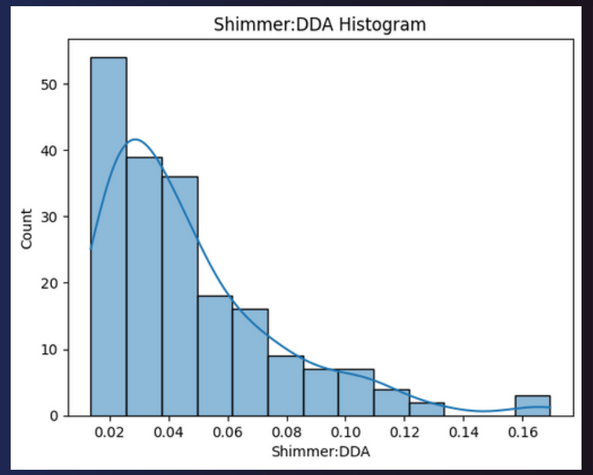


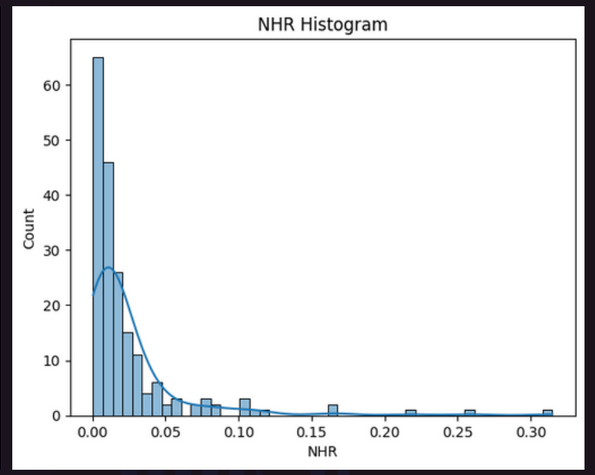


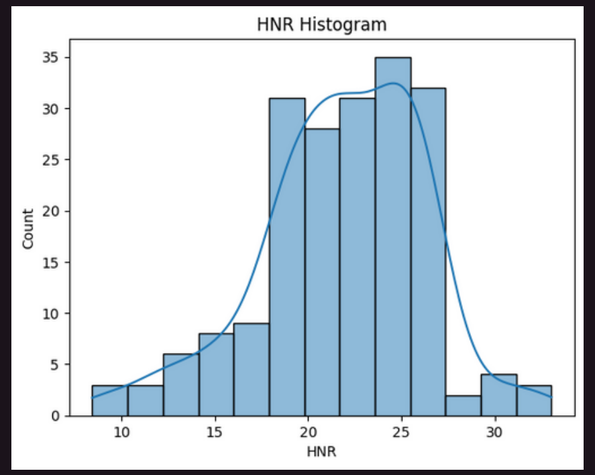




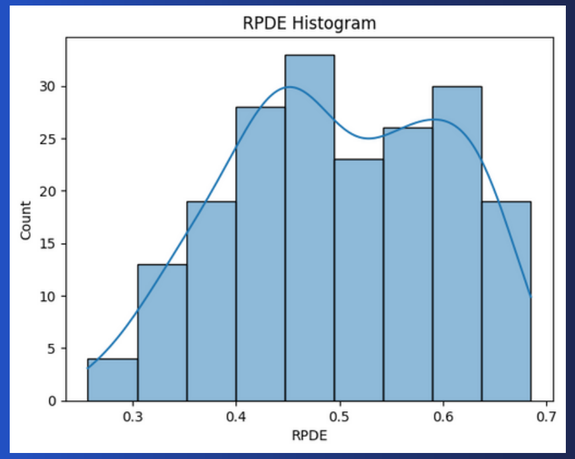


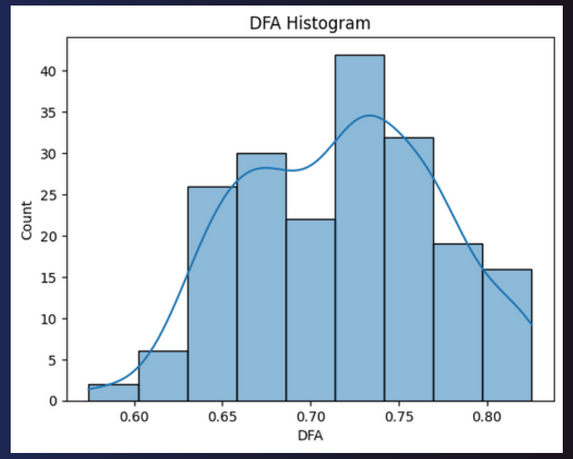


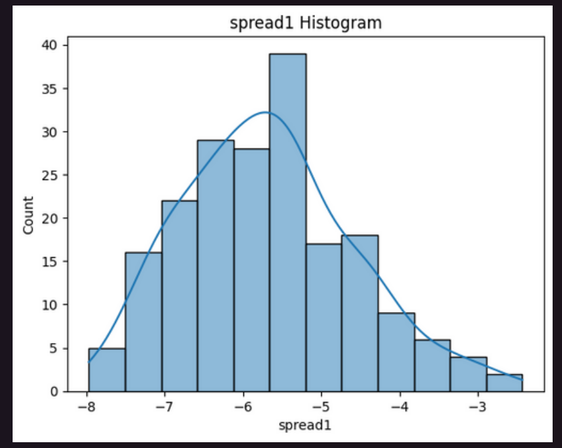


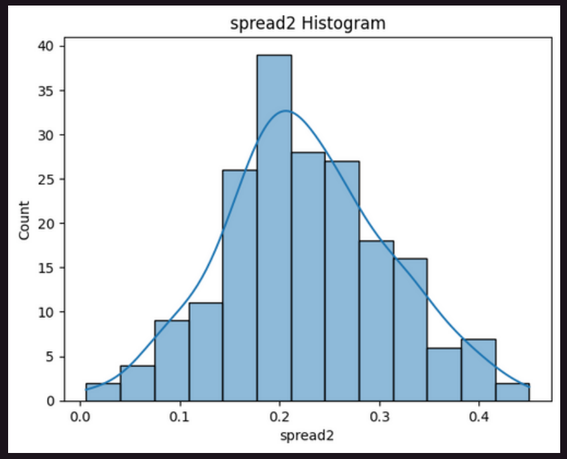


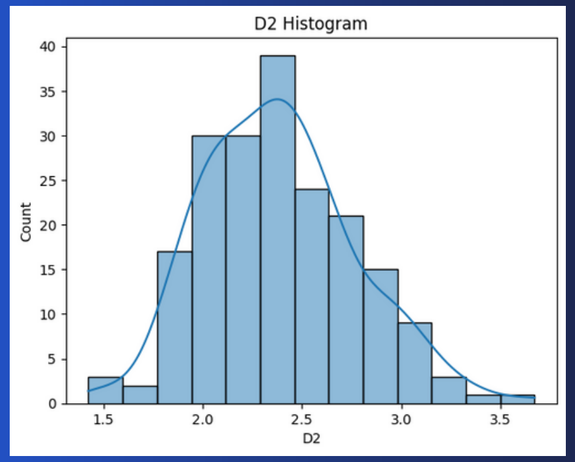
Univariate Analysis

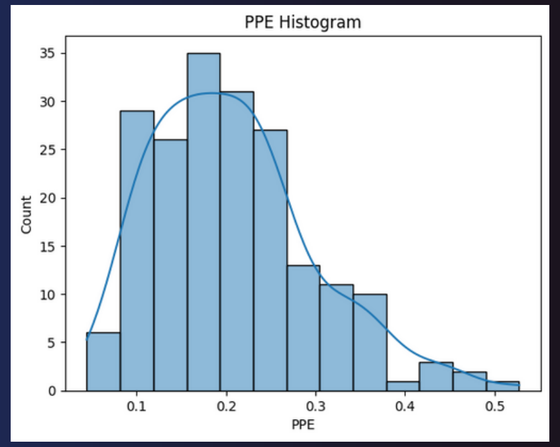


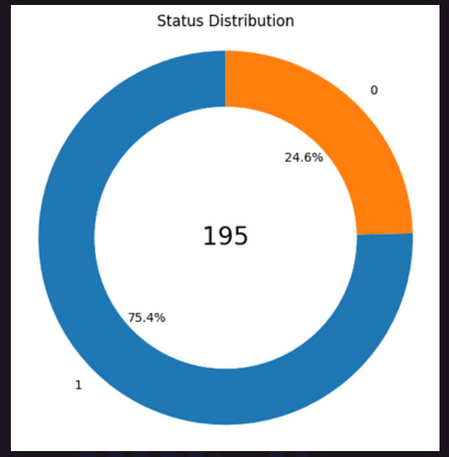




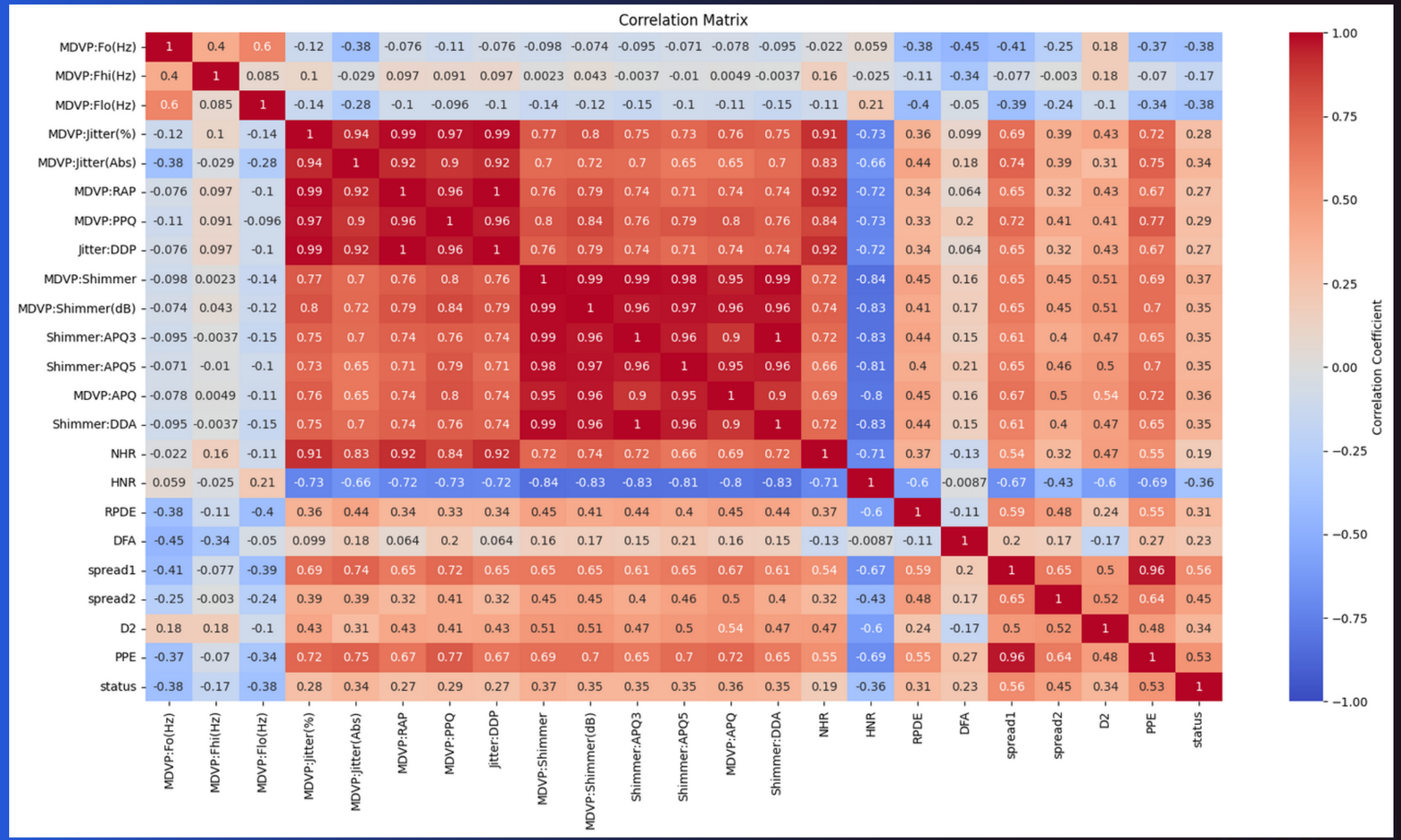


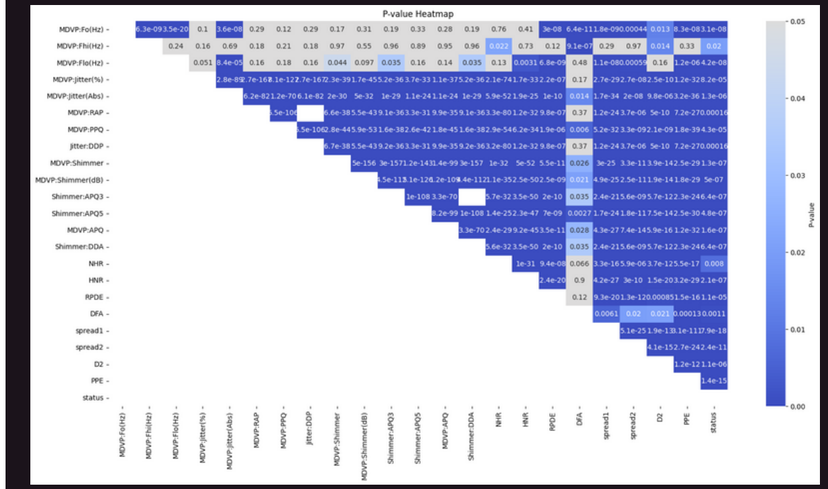






Attributes Correlation Matrix





Highly correlated data. We need to deal with this before building any models. We used Principal Component Analysis to create new dimensions and then build models.

T Test - Summary

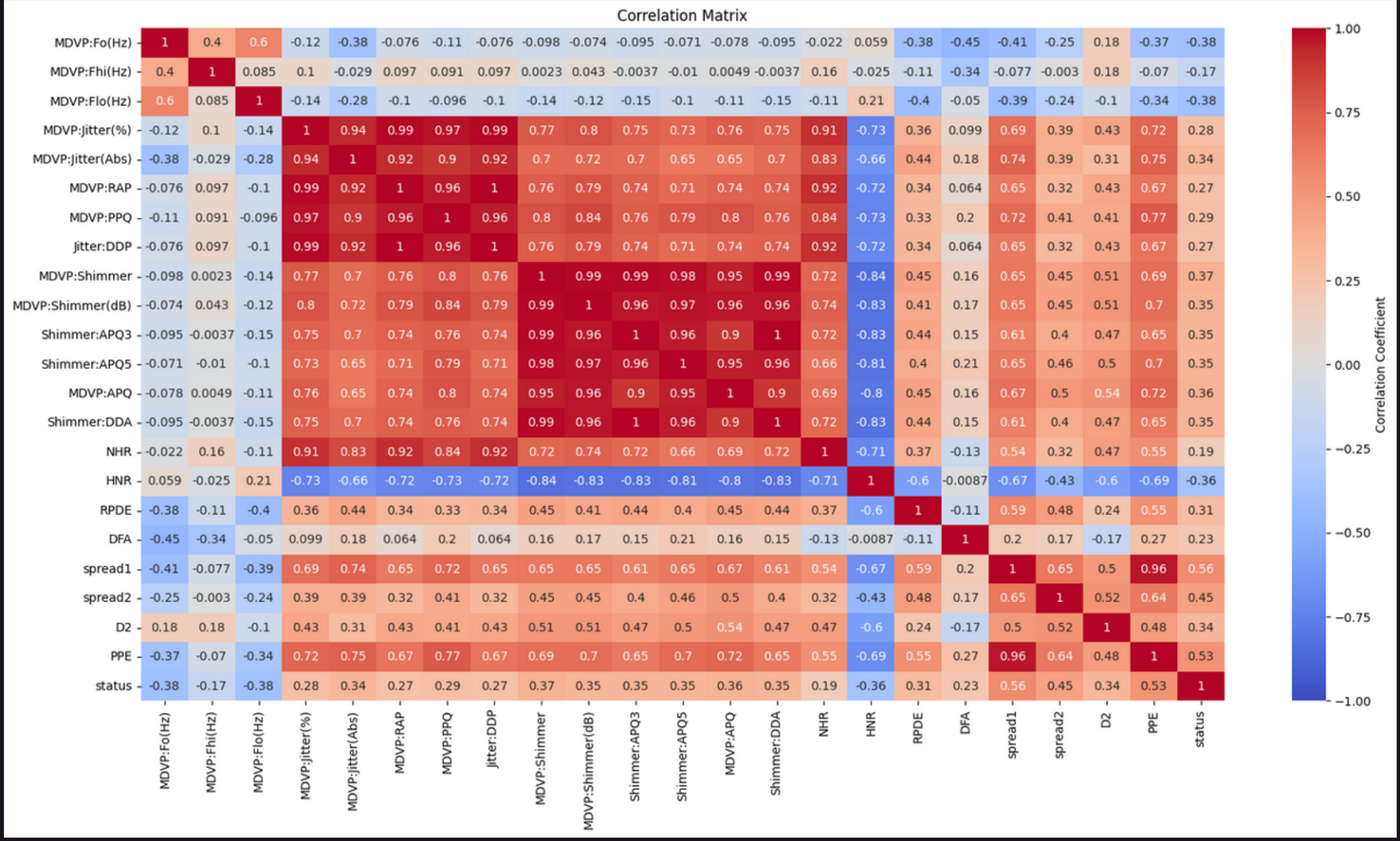
```
T-Test for MDVP:Fo(Hz): T-statistic = 5.769, P-value = 0.000, Significance = Yes
T-Test for MDVP:Fhi(Hz): T-statistic = 2.341, P-value = 0.020, Significance = Yes
T-Test for MDVP:Flo(Hz): T-statistic = 5.711, P-value = 0.000, Significance = Yes
T-Test for MDVP:Jitter(%): T-statistic = -4.024, P-value = 0.000, Significance = Yes
T-Test for MDVP:Jitter(Abs): T-statistic = -5.000, P-value = 0.000, Significance = Yes
T-Test for MDVP:RAP: T-statistic = -3.844, P-value = 0.000, Significance = Yes
T-Test for MDVP:PPQ: T-statistic = -4.189, P-value = 0.000, Significance = Yes
T-Test for Jitter:DDP: T-statistic = -3.844, P-value = 0.000, Significance = Yes
T-Test for MDVP:Shimmer: T-statistic = -5.488, P-value = 0.000, Significance = Yes
T-Test for MDVP:Shimmer(dB): T-statistic = -5.202, P-value = 0.000, Significance = Yes
T-Test for Shimmer:APQ3: T-statistic = -5.150, P-value = 0.000, Significance = Yes
T-Test for Shimmer:APQ5: T-statistic = -5.210, P-value = 0.000, Significance = Yes
T-Test for MDVP:APQ: T-statistic = -5.435, P-value = 0.000, Significance = Yes
T-Test for Shimmer:DDA: T-statistic = -5.150, P-value = 0.000, Significance = Yes
T-Test for NHR: T-statistic = -2.680, P-value = 0.008, Significance = Yes
T-Test for HNR: T-statistic = 5.387, P-value = 0.000, Significance = Yes
T-Test for RPDE: T-statistic = -4.507, P-value = 0.000, Significance = Yes
T-Test for DFA: T-statistic = -3.310, P-value = 0.001, Significance = Yes
T-Test for spread1: T-statistic = -9.509, P-value = 0.000, Significance = Yes
T-Test for spread2: T-statistic = -7.095, P-value = 0.000, Significance = Yes
T-Test for D2: T-statistic = -5.027, P-value = 0.000, Significance = Yes
T-Test for PPE: T-statistic = -8.707, P-value = 0.000, Significance = Yes
```

This statistical analysis provides strong evidence to suggest that there is a meaningful distinction between the two groups. [Groups are done by using the status column]

Principal Component Analysis

As the data is highly correlated, we decided to implement PCA

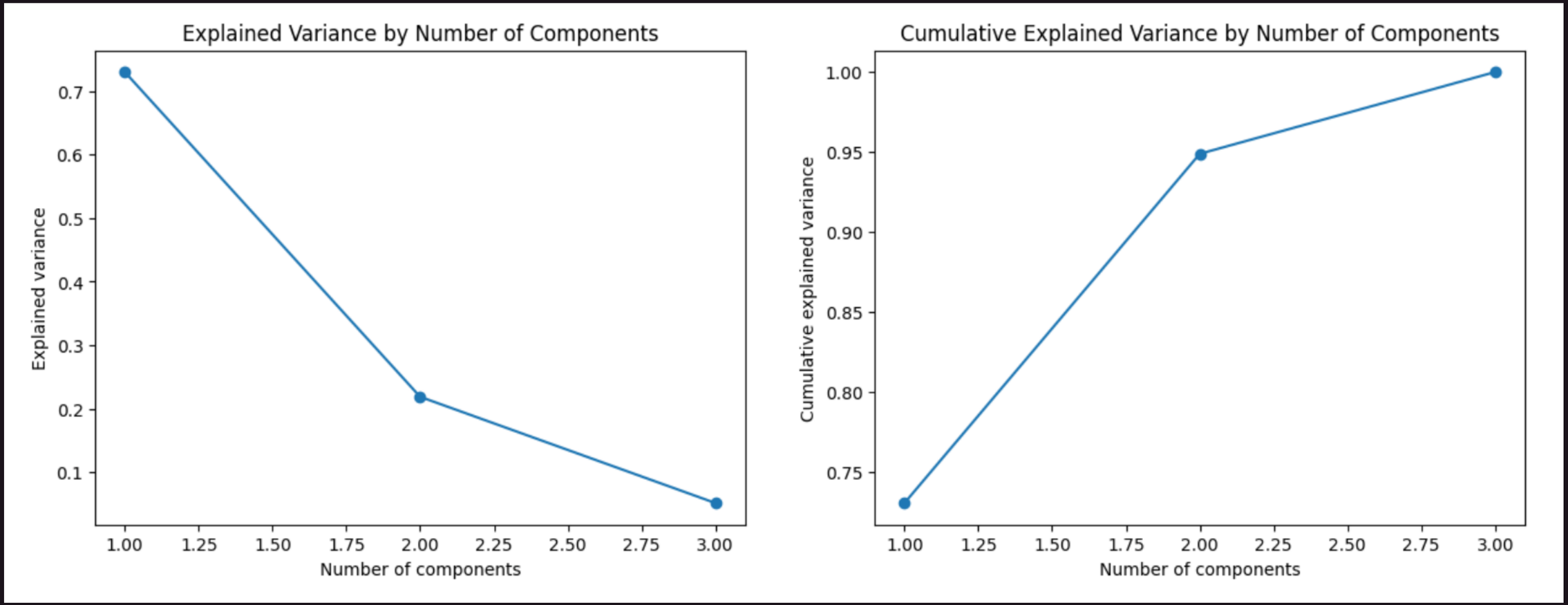
We approached this problem by conducting a principal component analysis in 3 steps.



- There are four red blocks of high correlation in the matrix.
- The first red block is related to all the frequency variables.
- The second block is related to all the shimmer variables.
- The third block is related to all the jitter variables.
- The final fourth block is related to the spread1, spread2, D2, and PPE variables.

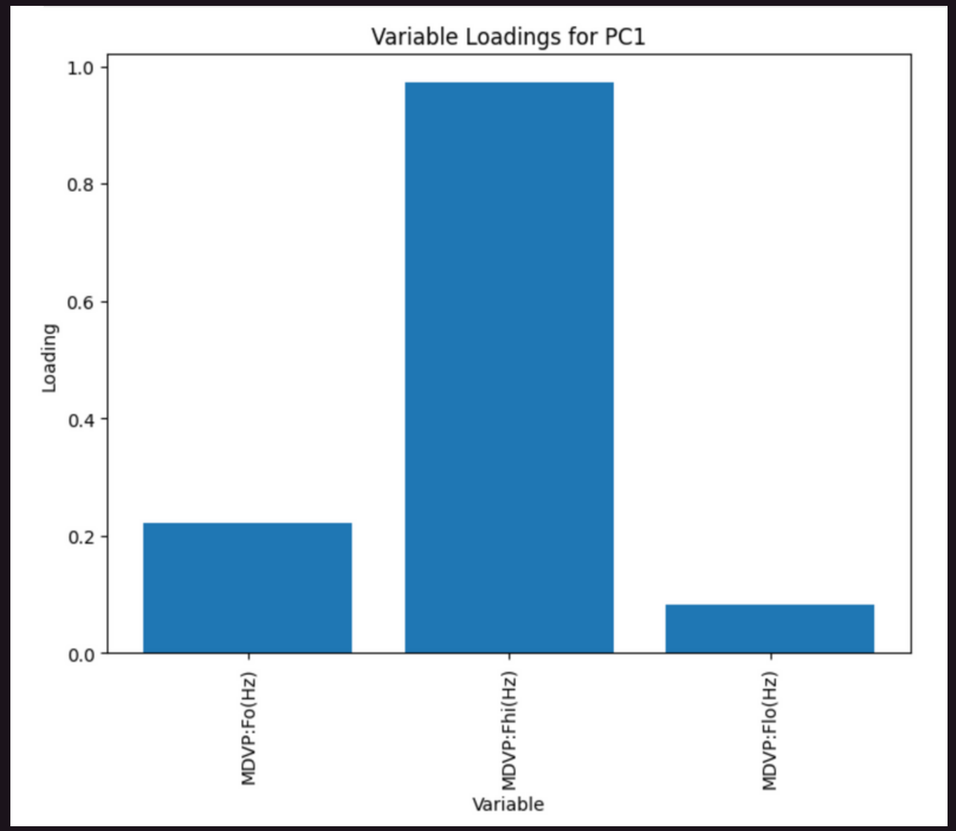
The goal is to reduce the dimensions to identify the variable that best represents the data.

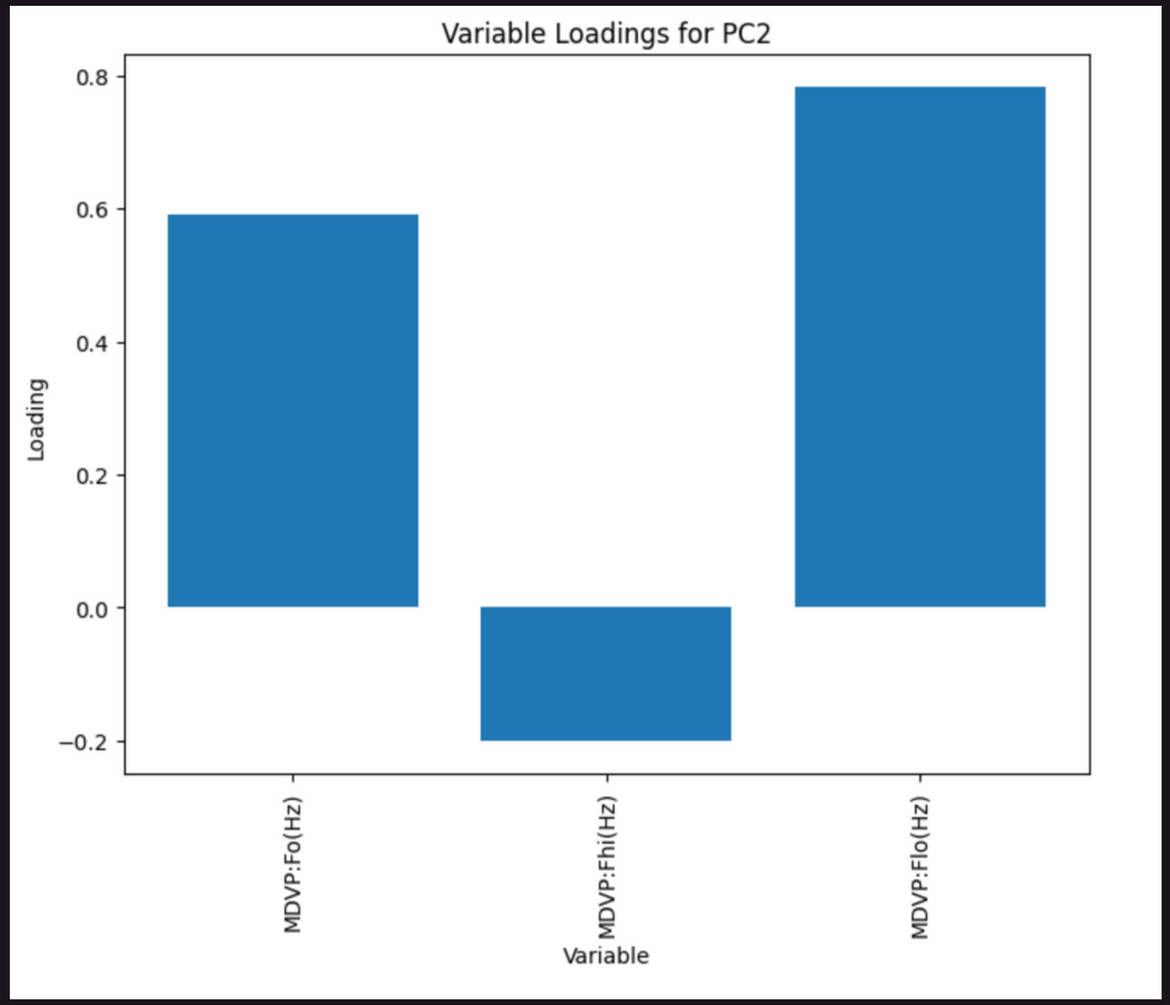
Applying PCA to Identify Key Variables in Frequency Data



The number of components to explain 80% of the variance: 2

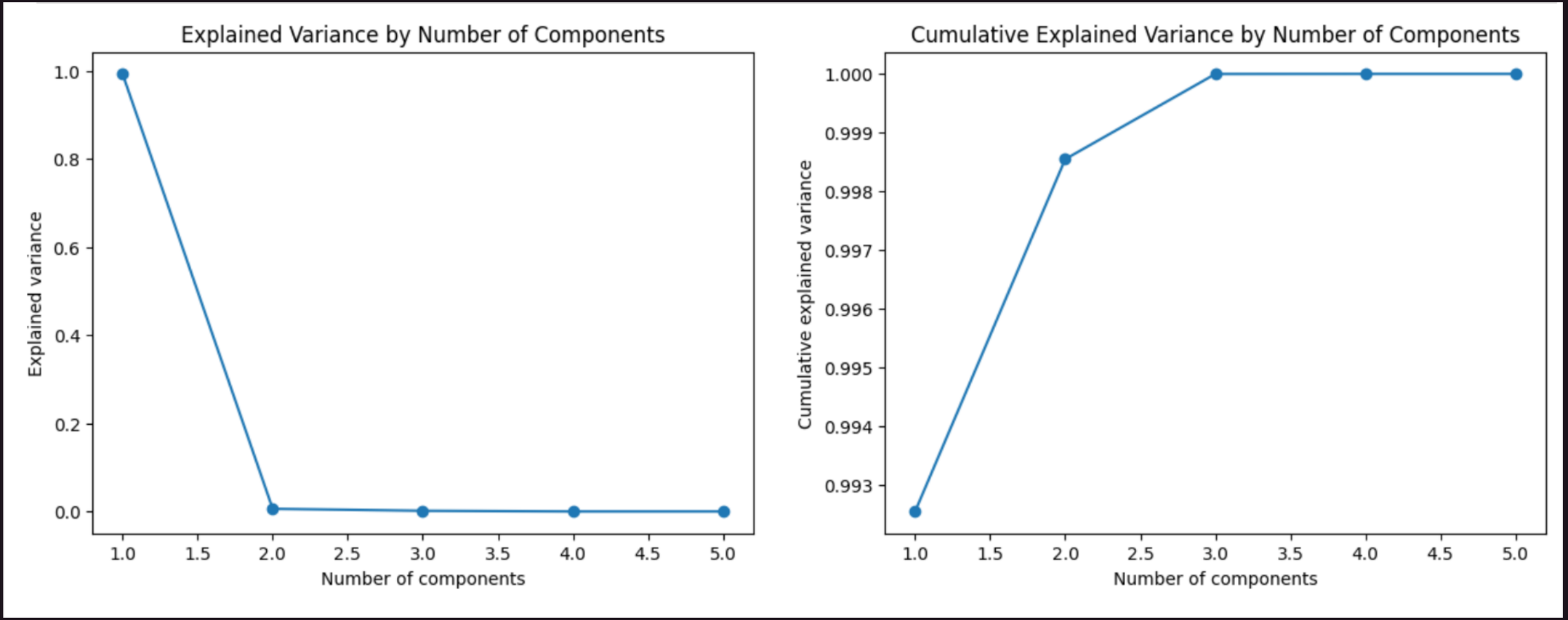
Applying PCA to Identify Key Variables in Frequency Data





We are going with MDVP:Fhi(Hz) for final modeling.

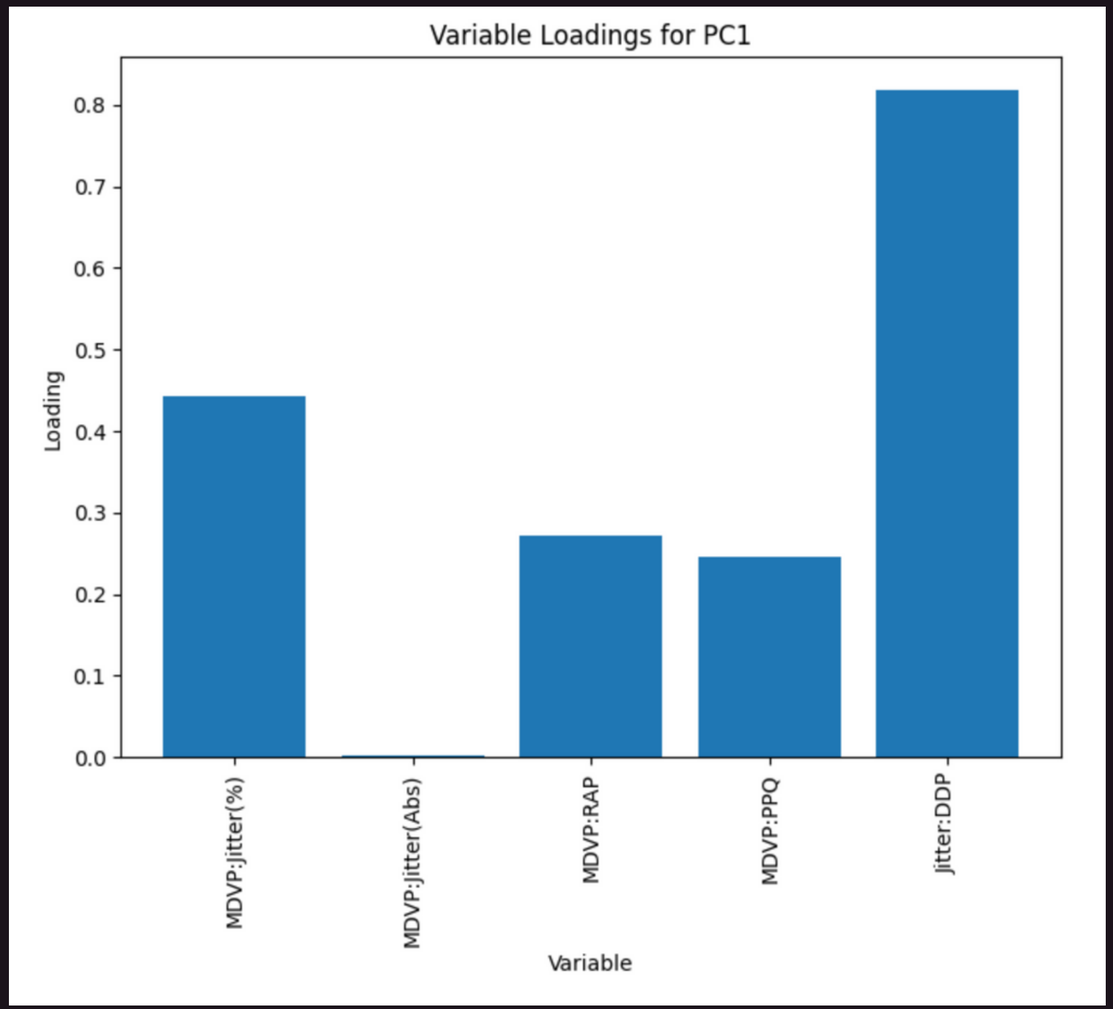
Applying PCA to Identify Key Variables in Jitter Data



The number of components to explain 80% of the variance: 1

Please note that the cumulative variance graph starts at 0.993

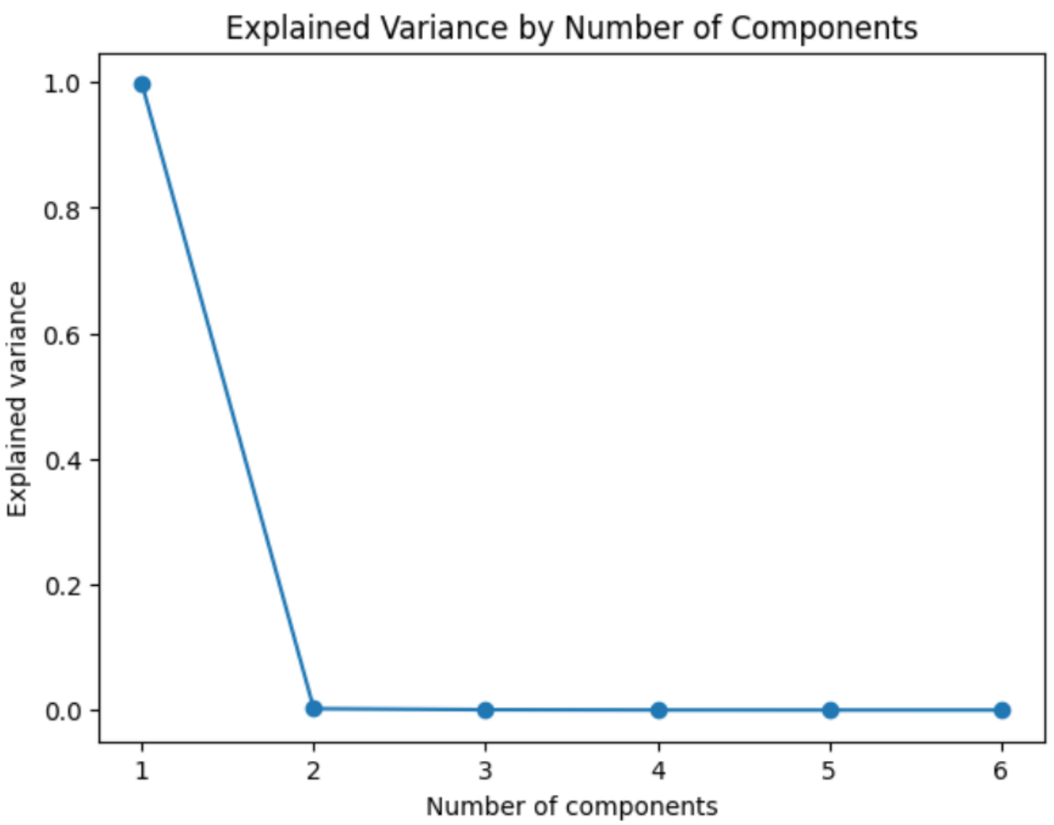
Applying PCA to Identify Key Variables in Jitter Data

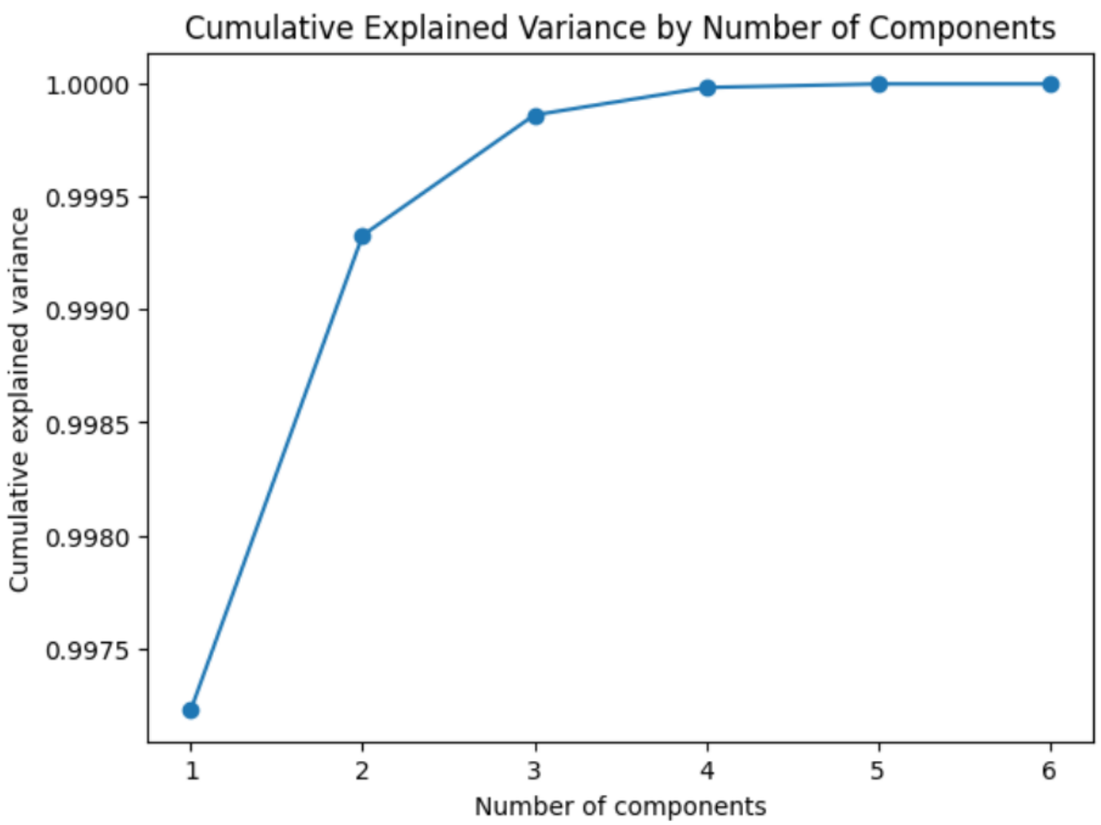


The variable sticking out here is Jitter:DDP.

We will be going with this for our final modeling.

Applying PCA to Identify Key Variables in Shimmer Data

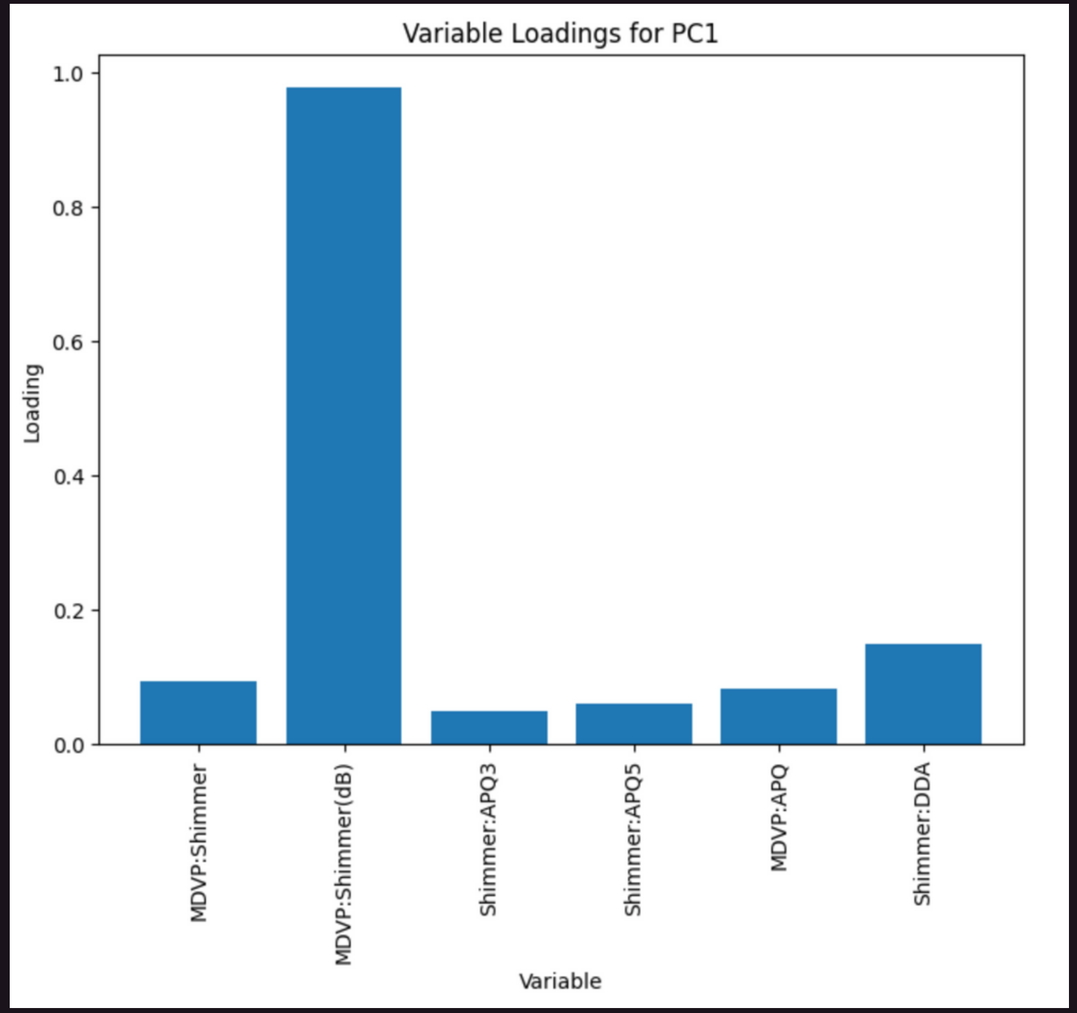




The number of components to explain 80% of the variance: 1

Please note that the cumulative variance graph starts at 0.9975.

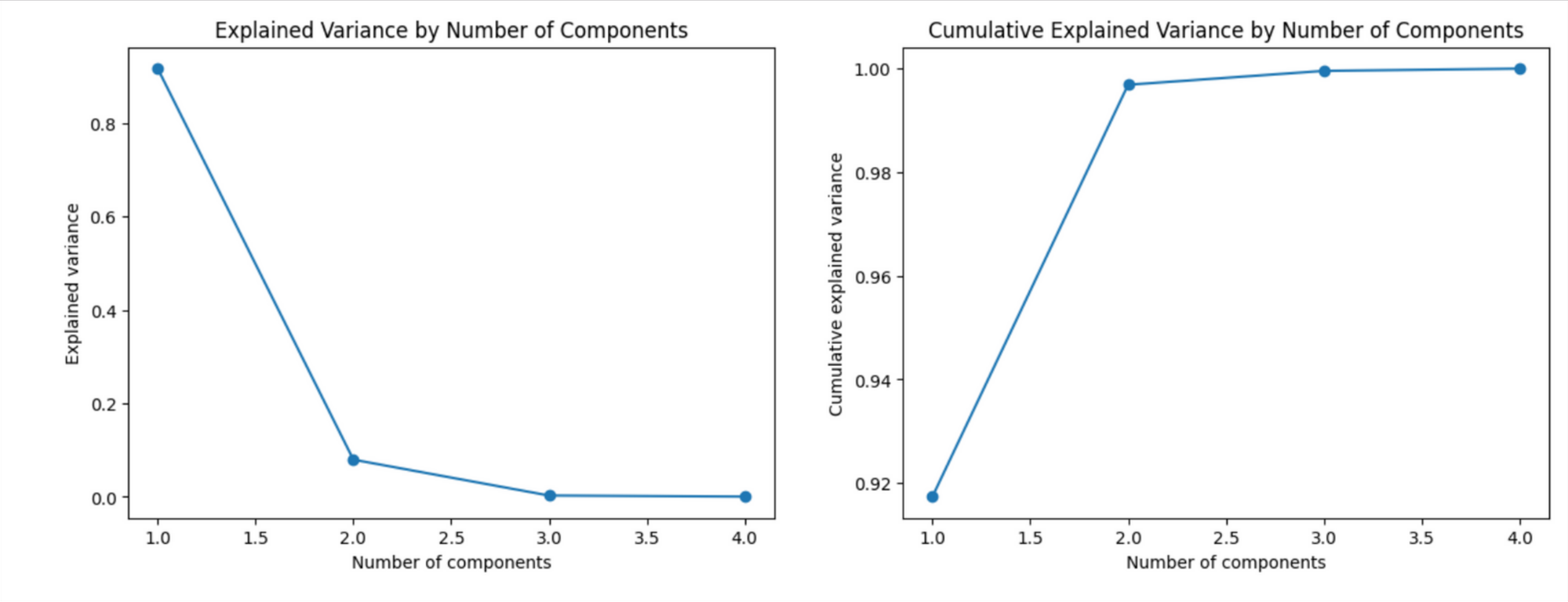
Applying PCA to Identify Key Variables in Shimmer Data



The variable sticking out here is MDVP: Shimmer.

We will be going with this for our final modeling.

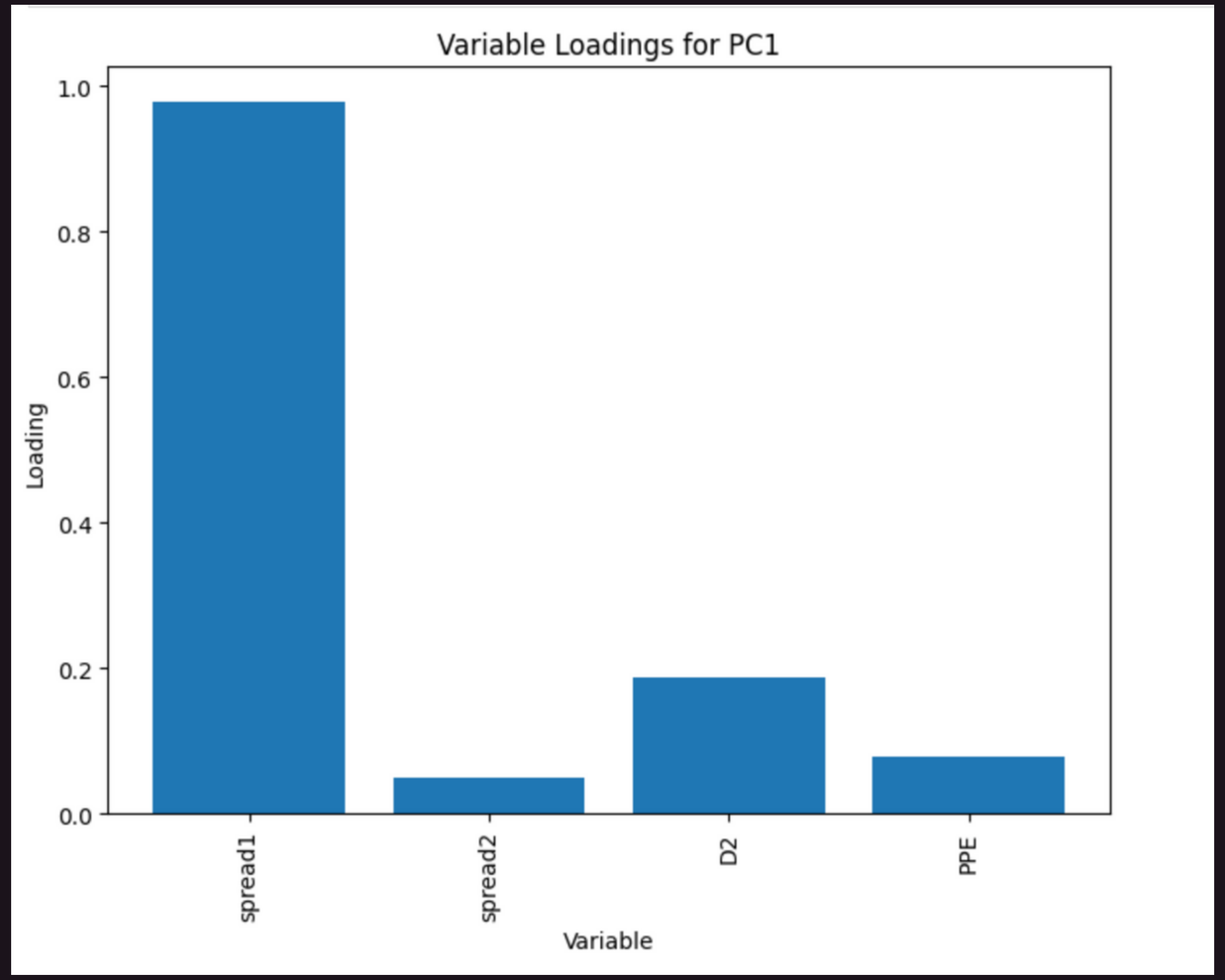
Applying PCA to Identify Key Variables in spread1, spread2, D2, PPE



The number of components to explain 80% of the variance: 1

Please note that the cumulative variance graph starts at 0.92

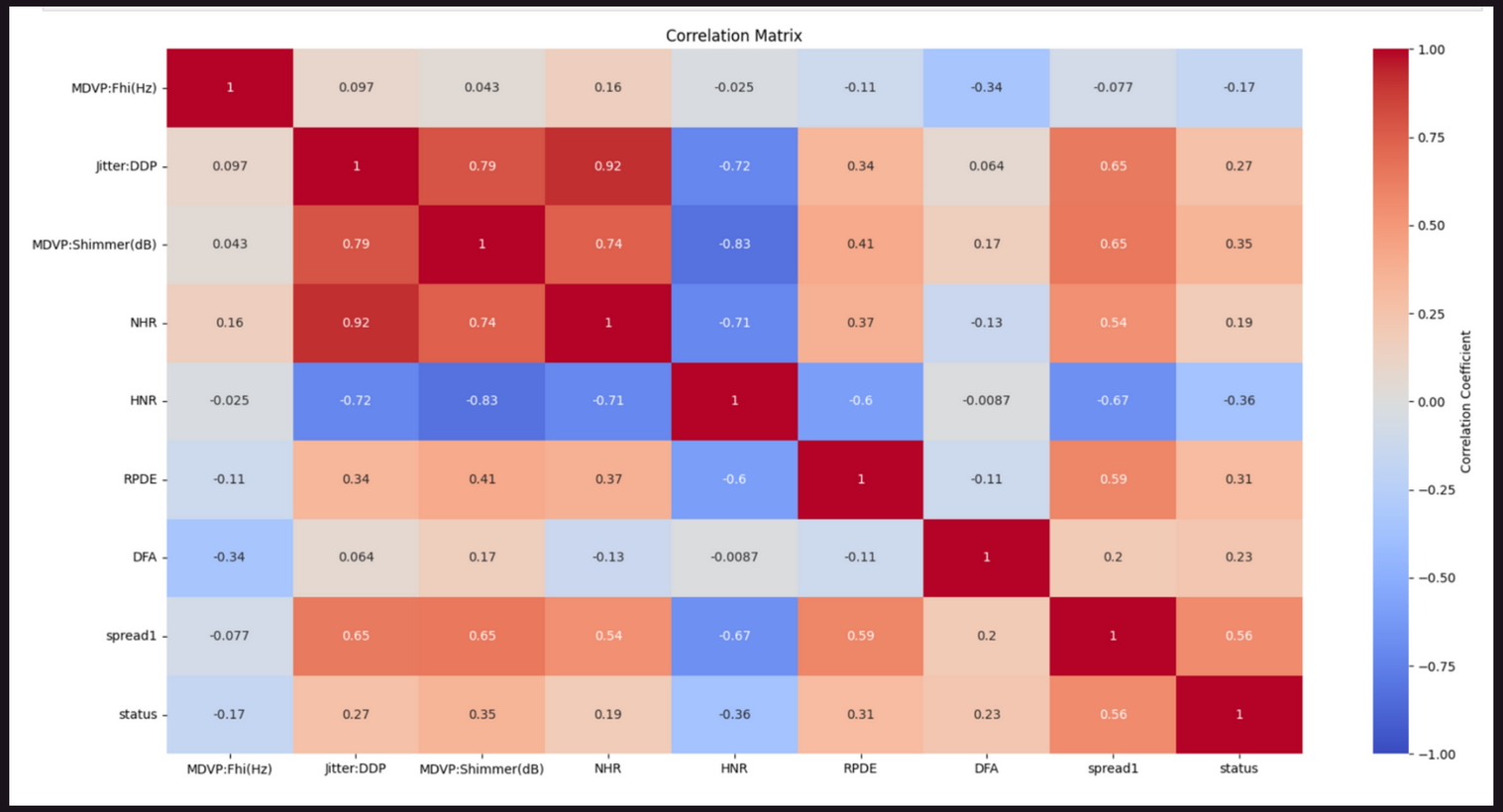
Applying PCA to Identify Key Variables in spread1, spread2, D2, PPE



The variable sticking out here is Spread1.

We will be going with this for our final modeling.

Final Correlation Matrix after PCA

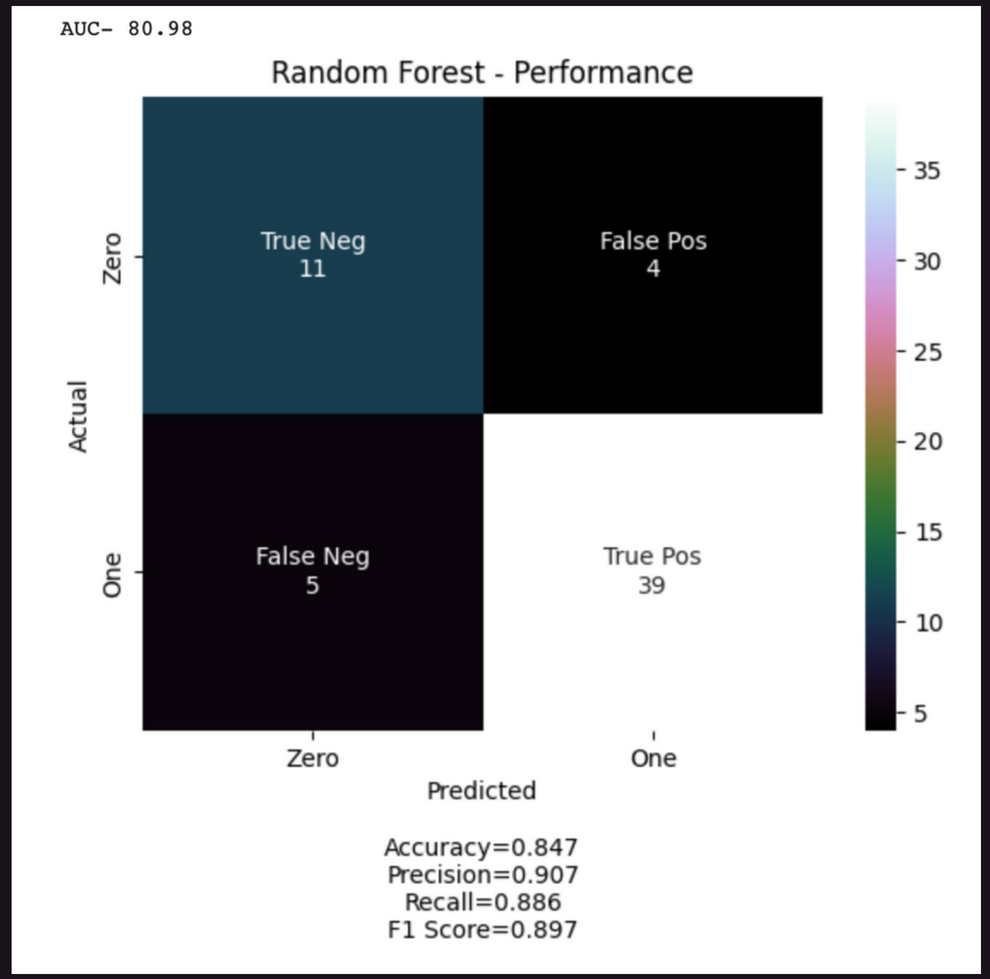


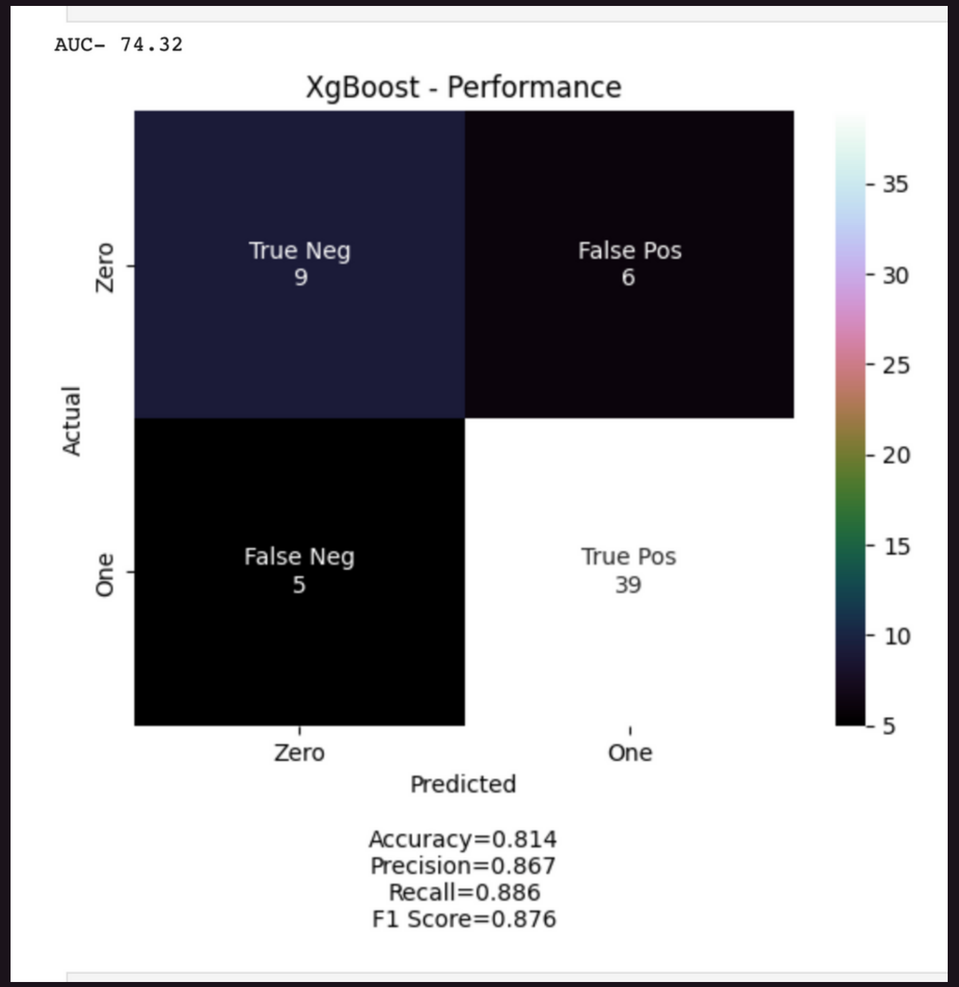
- We are not considering NHR or HNR as it highly correlates with Jitter and shimmer.
- For the final modeling, we will build 4 models.
  - With shimmer as the main variable from the red block.
    - Random Forest
    - xgboost
  - With jitter as the main variable from the red block.
    - Random Forest
    - xgboost

Modeling using Jitter

Test Train Split is done. [0.7/0.3]

Model testing on 30% of the unseen data.

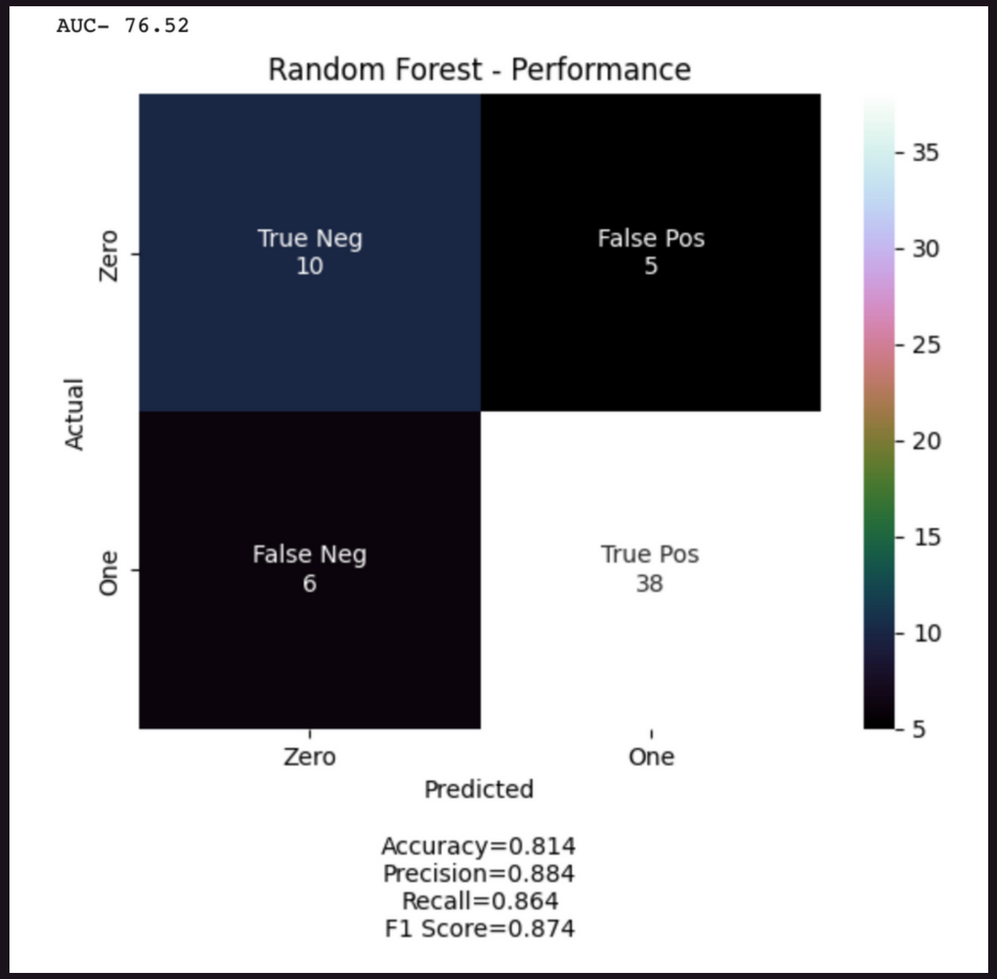


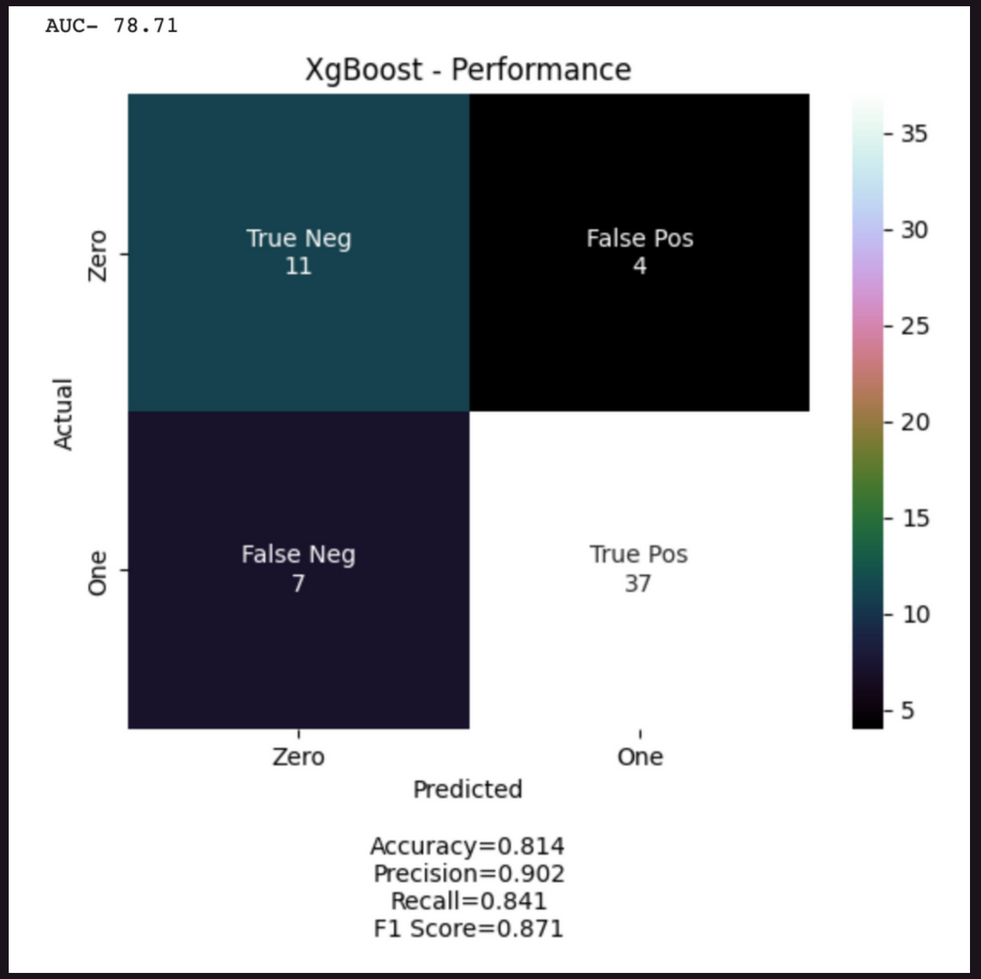


Modeling using Shimmer

Test Train Split is done. [0.7/0.3]

Model testing on 30% of the unseen data.





Thank you. Any feedback?