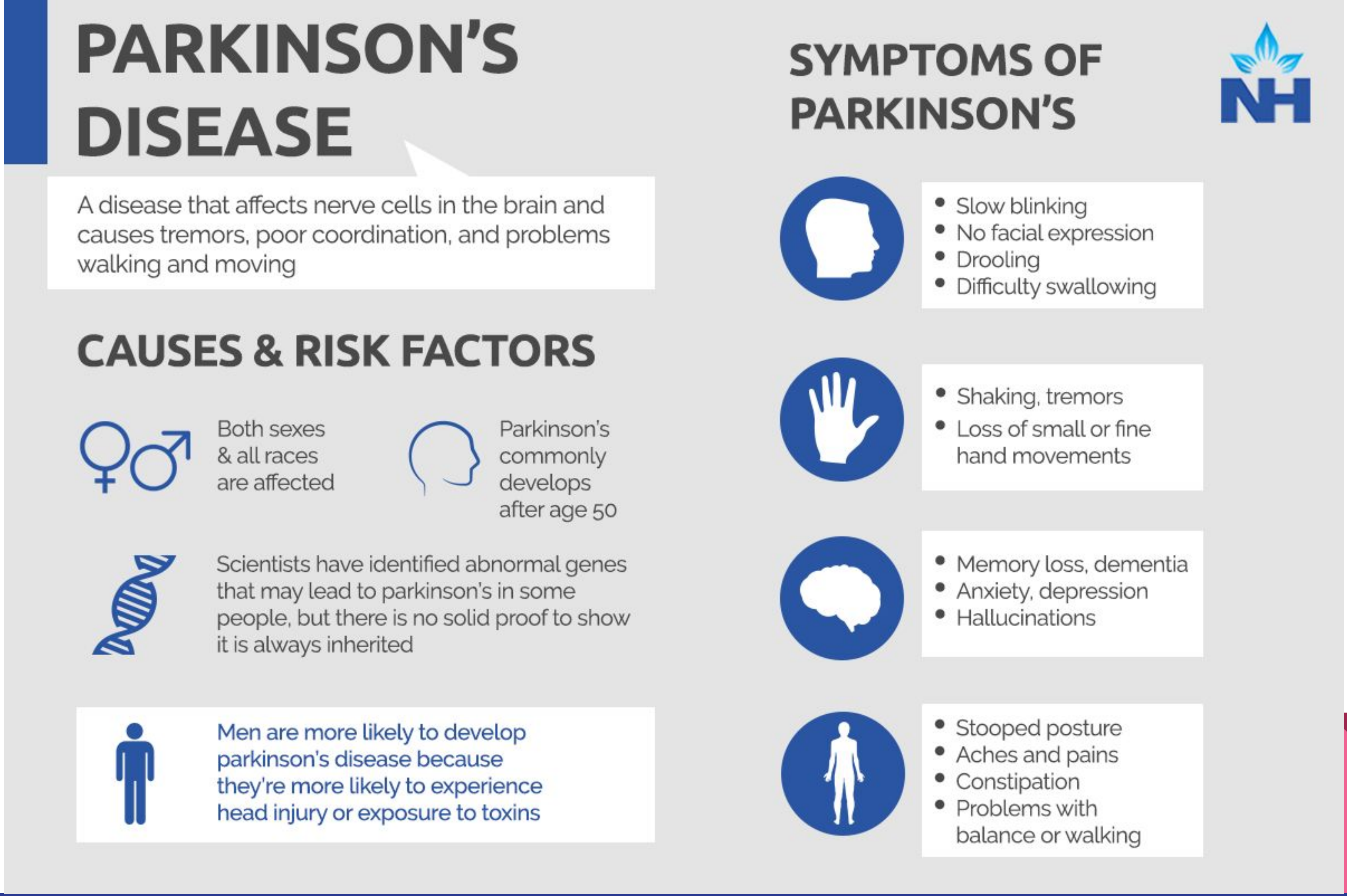
Predicting Parkinson’s Disease based on Vocal Recordings

Vilo Avila - Data Scientist

Problem Statement

Parkinson's disease is a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves.

My goal for this project is to find a model that can predict Parkinson's disease in a patient to the highest degree of accuracy based on features derived from vocal recordings, which could then be possibly used as an additional screening tool in a hospital or examination room setting.



PARKINSON'S DISEASE

A disease that affects nerve cells in the brain and causes tremors, poor coordination, and problems walking and moving

CAUSES & RISK FACTORS

Both sexes & all races are affected

Parkinson's commonly develops after age 50

Scientists have identified abnormal genes that may lead to parkinson's in some people, but there is no solid proof to show it is always inherited

Men are more likely to develop parkinson's disease because they're more likely to experience head injury or exposure to toxins

SYMPTOMS OF PARKINSON'S

- Slow blinking
- No facial expression
- Drooling
- Difficulty swallowing

- Shaking, tremors
- Loss of small or fine hand movements

- Memory loss, dementia
- Anxiety, depression
- Hallucinations

- Stooped posture
- Aches and pains
- Constipation
- Problems with balance or walking

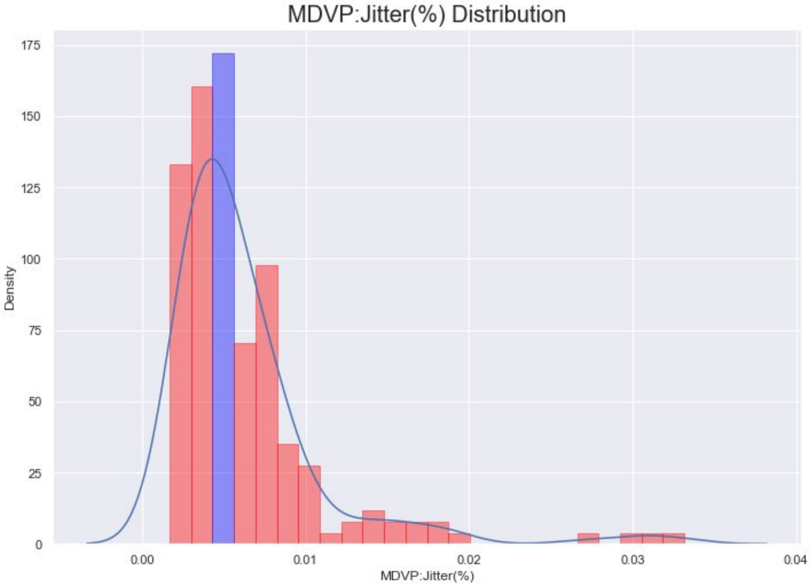
The Data

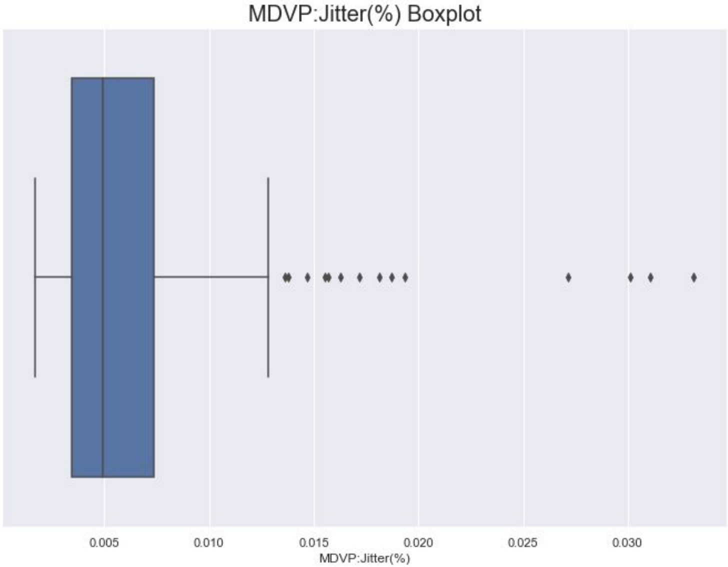
- **This dataset is composed of a range of voice measurements from 31 people, 23 with Parkinson's disease (PD)**
- **Each column in the table is a particular voice measure, and each row corresponds to one of 195 voice recording from these individuals.**
- **The dataset was downloaded from UC Irvine dataset repository.**

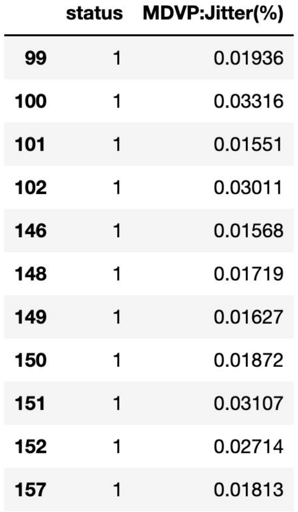
Cleaning

- Data set consists of 195 rows and 24 columns
- The dataset was cleaned by checking for null values, data types, and duplicates.
- Outliers were checked but I ultimately kept in the dataset.
- With keeping the outliers intact, I will look at models that are robust to outliers

Outliers and Distributions


MDVP:Jitter(%) Distribution

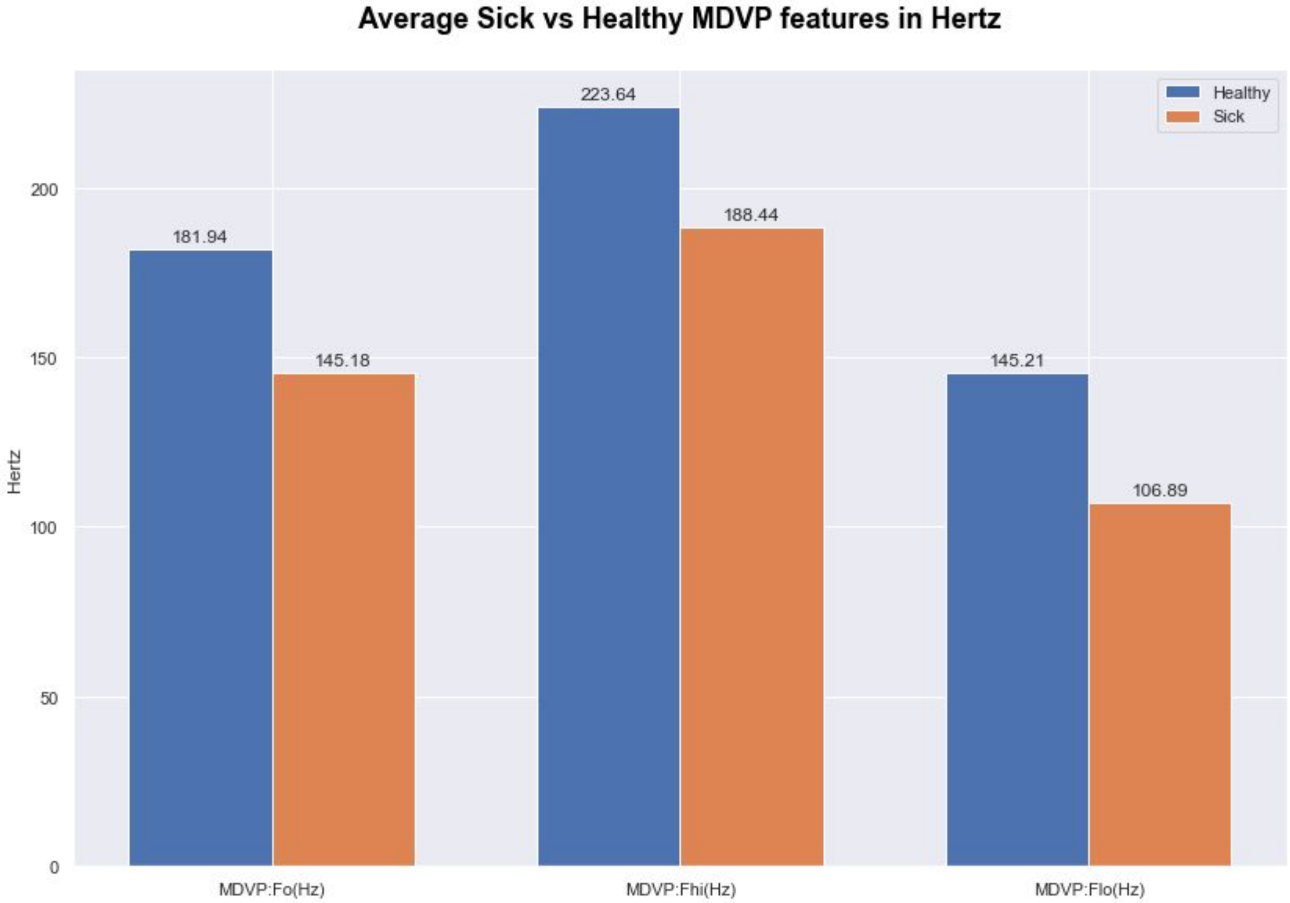

MDVP:Jitter(%) Boxplot



	status	MDVP:Jitter(%)
99	1	0.01936
100	1	0.03316
101	1	0.01551
102	1	0.03011
146	1	0.01568
148	1	0.01719
149	1	0.01627
150	1	0.01872
151	1	0.03107
152	1	0.02714
157	1	0.01813

Exploratory Data Analysis

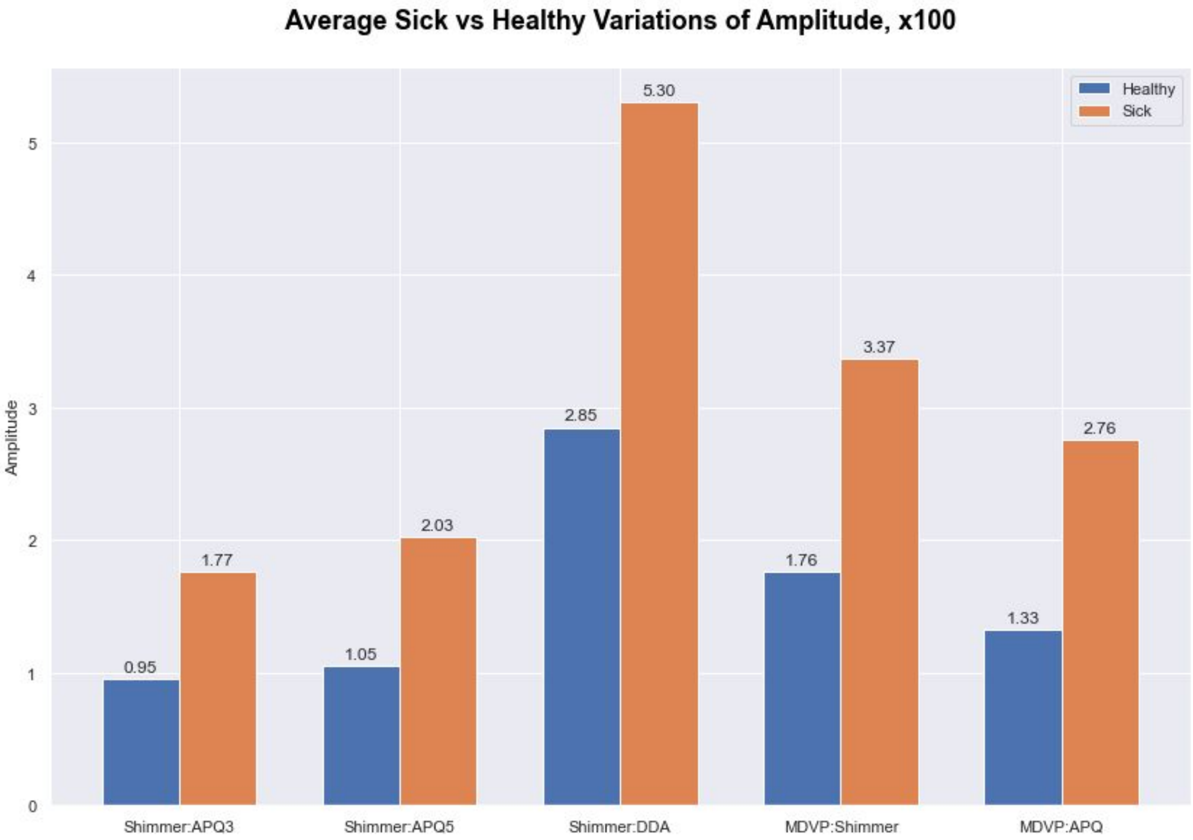
- **MDVP: Multidimensional Voice Program**
- **Fo - Average frequency**
- **Fhi - High Frequency**
- **Flo - Low frequency**



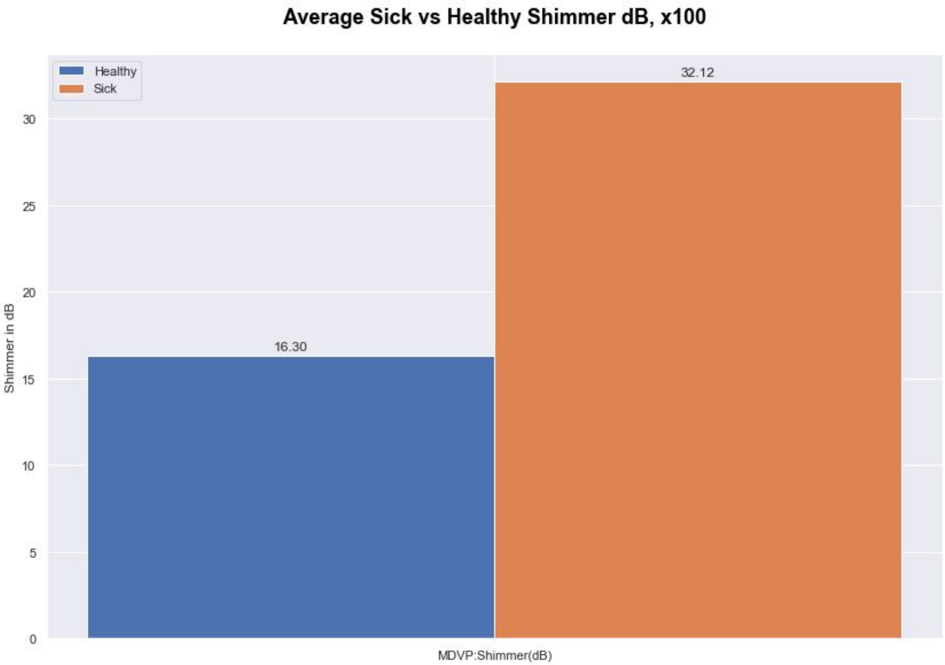
EDA continued

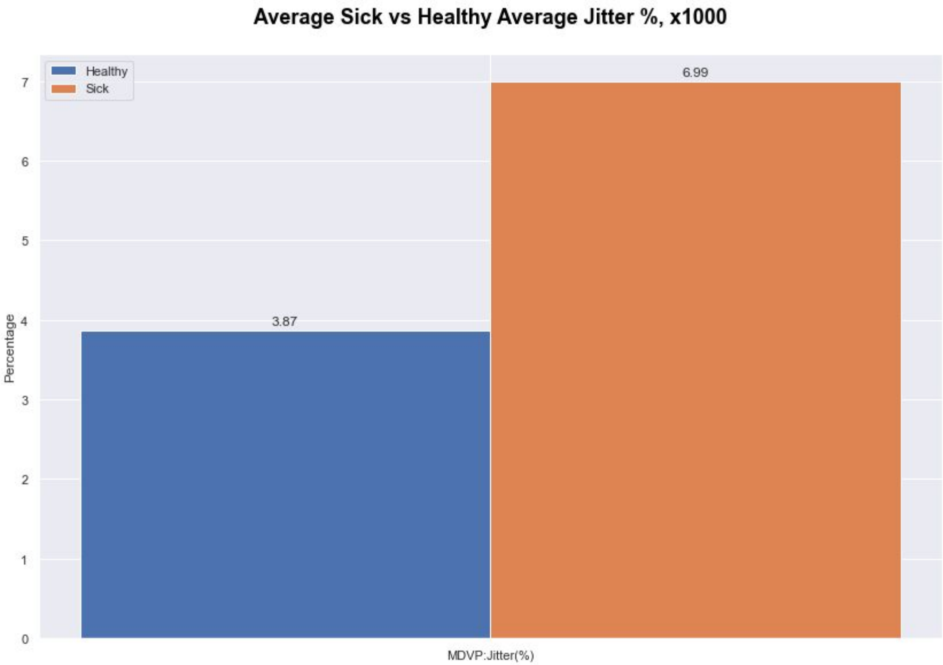
- **MDVP: Multidimensional Voice Program**
- **Shimmer - measurement of changes in the amplitude of the speech wave.**
- **APQ - Amplitude Perturbation Quotient**
- **DDA - Average absolute difference between consecutive periods**

Summary: Several Measures of variation in amplitude



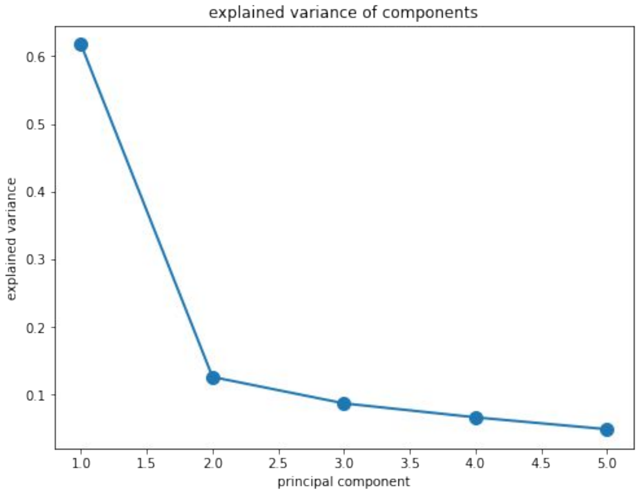
EDA continued



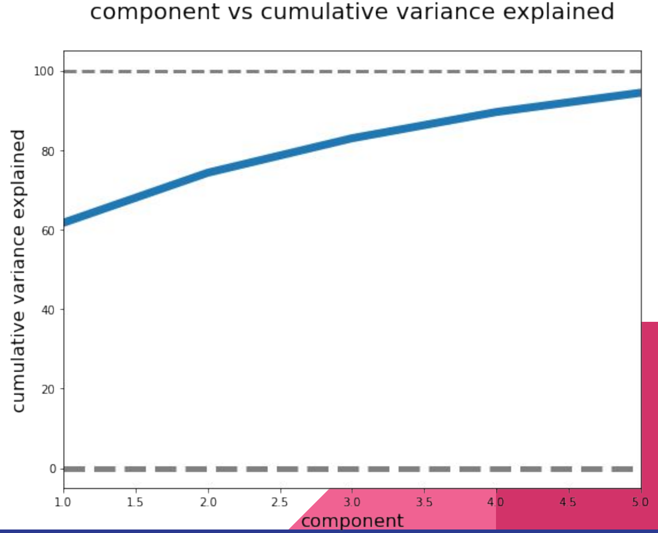


Pre-processing

- All features were numerical, there were no categorical or ordinal features that needed encoding
- Dataset was StandardScaled
- Out of 23 features, SelectKBest was used to select top 10 for model use
- PCA was used on the remaining features to capture signal into 5 components
- SelectKBest and PCA components combined for final dataframe



explained variance of components



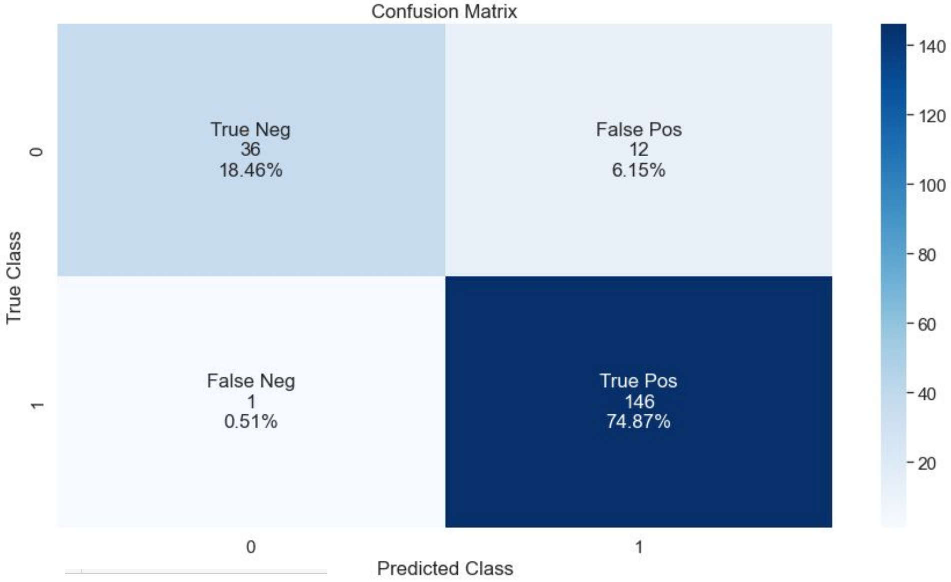
component vs cumulative variance explained

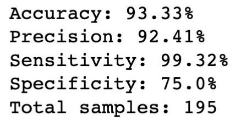
Machine Learning Models

- Random Forest
  - Accuracy: 93.33%
- Support Vector Machine
  - Accuracy: 97.95%
- Gradient Boosting
  - Accuracy: 98.46%

Confusion Matrices

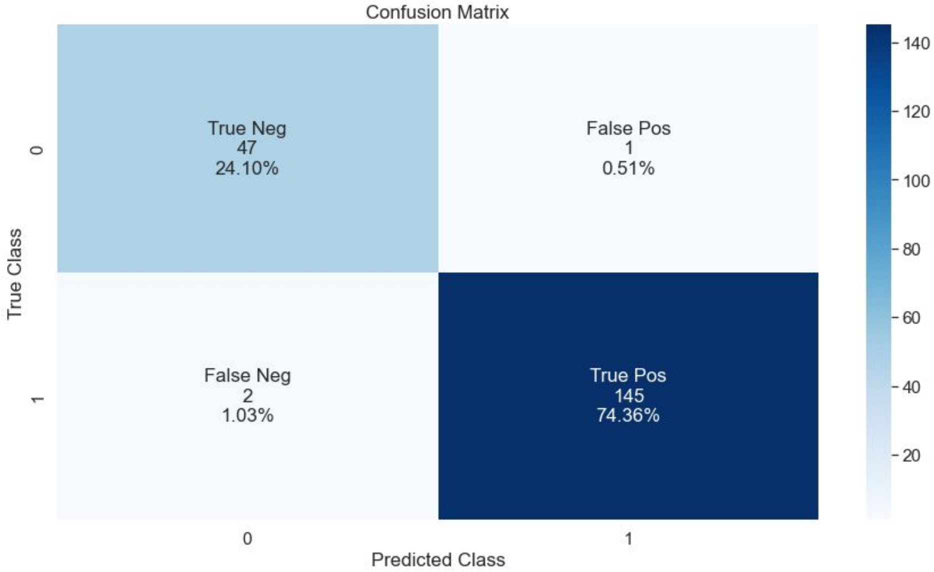
Random Forest

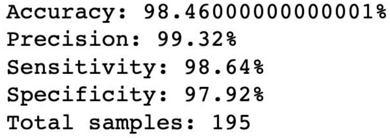




```
Accuracy: 93.33%
Precision: 92.41%
Sensitivity: 99.32%
Specificity: 75.0%
Total samples: 195
```

Gradient Boosting





```
Accuracy: 98.46000000000001%
Precision: 99.32%
Sensitivity: 98.64%
Specificity: 97.92%
Total samples: 195
```

Conclusions

- Given the outcome of our models, it would seem Parkinson's disease can be successfully predicted using recordings of a patient's voice. I feel this would be an excellent tool for a doctor to use while in early stages of diagnosing if a patient may have Parkinson's disease.
- The first model I would choose for production would be the random forest model. It performed the best at accurately predicting sick patients with only 1 false negative. It did falsely predict that 12 patients were sick when in fact they were healthy, but we could use this as a screener to get the patient examined again by other means. It is safer to tell a healthy patient that they may have the disease as opposed to telling a sick patient that they are completely healthy.
- With that note I would only use the random forest model until I can improve the gradient boosting model to the level of accuracy at predicting sick patients to that of the random forest one. The reason is that the gradient boosting performed better overall. It only had 2 false negatives and 1 false positive.

Recommendations

- With some further work I would be able to build an algorithm that could take in raw vocal data and feed it through a pipeline where everything gets processed and sent through to the final production model.
- I recommend investing time and money in my work in order to complete these goals so we could have a non-invasive tool that could quickly and accurately determine if a patient may have Parkinson's disease. This could be used in a doctor's office or hospital setting where they could ask the patient to say a predetermined phrase in a microphone and get a result in minutes.

Improvements going forward

- The main improvement would be to continue to work with the gradient boosting model by tweaking parameters to give us the best possible prediction.
    - One Caveat - The small degree of difference in accuracy could boil down to the random seed chose for the model. Ultimately more data is needed.
- If we are able to gather more samples it would improve our model overall.
- Once enough samples are gathered for a robust model, I would begin building the structure needed to deploy the model for testing purposes in real world scenarios.
- Funding for research and development