Milestone Report 2:

Predicting Neurodegeneration Diseases

Introduction

Objective

Researchers at Czech Technical University and Charles University in Prague collected vocal assessment data on patients with two, related neurodegenerative diseases: Parkinson's disease (PD) and eye movement sleep behavior disorder (RBD). A study involving 30 untreated, newly diagnosed patients with Parkinson's Disease, 50 people with eye movement sleep behavior disorder (RBD) and 50 healthy, control subjects was conducted. Participants were asked to perform 2 speaking tasks and one monologue task to analyze voiced speech, unvoiced speech, pause and respiration. (NCBI Report)

The goal of this project is to use the vocal assessment data in conjunction with the medical, motor, and demographic data collected to analyze and create a classification model that can predict a patient's likelihood of having either PD or RBD.

Target Audience

This model can provide benefits to both doctors and their patients. Doctors can use this data to detect neurodegeneration diseases in their patients earlier, that is cost-effective, non-invasive and scalable. Therapy and care for diseases like PD and RDB is more effective the earlier it is implemented and can mitigate a decline in the quality of life for the patient. Additionally, a classification model can help reduce the healthcare burden on society with early detection and implementation of preventative measures.

Data Source

The report published on the National Center for Biotechnology Information's website included a section of supplementary material that included the study's dataset. 130 unique patient observations are included with 64 features.

Data Wrangling

Raw Data Content & Structure

The dataset is in XLS format. The columns are grouped into categories. The columns in the clinical information and overview of motor examination categories do not include data for the control group since those subjects have not been diagnosed with PD or RBD and the feature does not apply. The columns included in the dataset are:

Participant Code

Each subject was given an ID, with a prefix corresponding with their status:

- PD - Subject with newly diagnosed Parkinson's disease
- RBD - Subject with rapid eye movement sleep behavior disorder
- HC - Healthy, control group

Demographic Information

- Age (years)
- Gender - M/F

Clinical Information

- Positive history of Parkinson disease in family - Yes/No
- Age of disease onset (years)
- Duration of disease from first symptoms (years)

Medication

- Antidepressant therapy - Yes/No, if yes includes name of medication
- Antiparkinsonian medication - Yes/No
- Antipsychotic medication - Yes/No
- Benzodiazepine medication - Yes/No, if yes includes name of medication
- Levodopa equivalent (mg/day)
- Clonazepam (mg/day)

Overview of Motor Examination

- Hoehn & Yahr scale - A commonly used scale for measuring the progression of Parkinson's disease. The scale includes stages 1, 1.5, 2, 2.5, 3, 4, and 5.
- UPDRS III total - Unified Parkinson's Disease Rating Scale: A motor evaluation scale scored by a clinician via observation. Provides additional insight on the progression of PD. Scores range from 0 to 108.

UPDRS III Motor Scale: Specific Items

Each item is rated based on a 0-4 scale where 0 = normal, 1 = slight, 2 = mild, 3 = moderate and 4 = severe. RUE = Right Upper Extremity, LUE = Left Upper Extremity, RLE = Right Lower Extremity, LLE = Left Lower Extremity

- Speech
- Facial Expression
- Tremor at rest - head

- Tremor at rest - RUE
- Tremor at rest - LUE
- Tremor at rest - RLE
- Tremor at rest - LLE
- Action or Postural Tremor - RUE
- Action or Postural Tremor - LUE
- Rigidity - neck
- Rigidity - RUE
- Rigidity - LUE
- Rigidity - RLE
- Rigidity - LLE
- Finger Taps - RUE
- Finger Taps - LUE
- Hand Movements - RUE
- Hand Movements - LUE
- Rapid Alternating Movements - RUE
- Rapid Alternating Movements - LUE
- Leg Agility - RLE
- Leg Agility - LLE
- Arising from Chair
- Posture
- Gait
- Postural Stability
- Body Bradykinesia and Hypokinesia

<u>Speech Examination: Speaking Task of Reading Passage / Speech Examination: Speaking Task of Monologue</u>

- Entropy of speech timing (EST) - Heterogeneity of speech in terms of the occurrence of voiced, unvoiced, pause and respiratory intervals
- Rate of speech timing (/min) (RST) - Speech rate with respect to quality of speech timing
- Acceleration of speech timing (/min2) (AST) - Acceleration of speech associated with parkinsonism
- Duration of pause intervals (ms) (DPI) - Quality of speech timing
- Duration of voiced intervals (ms) (DVI) - Fundamental phonatory mean
- Gaping in-between voiced intervals (/min) (GVI) - Deficits of phonatory onset and offset control
- Duration of unvoiced stops (ms) (DUS) - Measurement of stop consonants
- Decay of unvoiced fricatives (%/min) (DUF) - Temporal quality of articulation
- Relative loudness of respiration (dB) (RLR) - Audibility of respiration relative to loudness of speech
- Pause intervals per respiration (PIR) - Breath groups
- Rate of speech respiration (/min) (RSR) - Respiratory rate during speech
- Latency of respiratory exchange (ms) (LRE) - Measures pauses between expiration

Data Wrangling Steps

Note: Please reference my github account for a detailed overview of data wrangling code and steps (https://bit.ly/2VvQNTW)

1. Download Data
   The raw data set was downloaded from the supplementary section of the study's online report and stored in the project's local directory.
2. Raw Data Inspection
   The raw data was initially inspected in Google sheets and via a generated Pandas Profiling report. There are several columns with categorical data (e.g. Gender, Family History, Medication, etc.) that require encoding and specific data categories (e.g. Clinical History and Overview of Motor Examination) that only apply to patients with PD or RBD.
3. Import Data
   Pandas read_excel function was used to import the data and the header function was passed to indicate that the second row of the file was to be used as the header for the dataframe.
4. Inspect in Jupyter Notebook
   Further inspection was done using the head, tail and info methods. Many of the columns have a data type of 'object' due to empty cells and there are several trailing empty rows.
5. Cleaning Data Set
   a. The trailing empty rows were removed using df.dropna()
   b. The column names are long and have paragraph breaks and spaces in them that are not ideal for dataframe analysis and manipulation. All column names were explicitly updated using df.rename().
   c. During inspection, it was noted that several columns only had one value which would not be useful for analysis. Any column with a unique value = 1 was removed.
   d. The "id" column has a prefix indicating the observation's status. The prefix was extracted and the column name was then changed to 'status'.
   e. The categorical data (i.e. status, gender, medication, etc.) was encoded next.
   f. The columns that are not applicable to the control group (e.g. age of disease onset, etc.) had a '-' in the control group cells which caused the data type for the column to be set as 'object'. Those cells were replaced with NaN so that the correct data type for the column could be automatically set.
   g. Columns 'hy_scale' and 'age' had to have their data type set explicitly.
6. Feature Engineering
   a. In order to more efficiently analyze ages, a new "age_range" column was created based on the "age" column.
   b. The new column was then positioned next to the "age" column for coherency.
7. Export
   The cleaned dataframe was exported via Pandas to_csv and setting the Index parameter to False. This will ensure that when the cleaned .csv file is imported, it does not include an extra column.

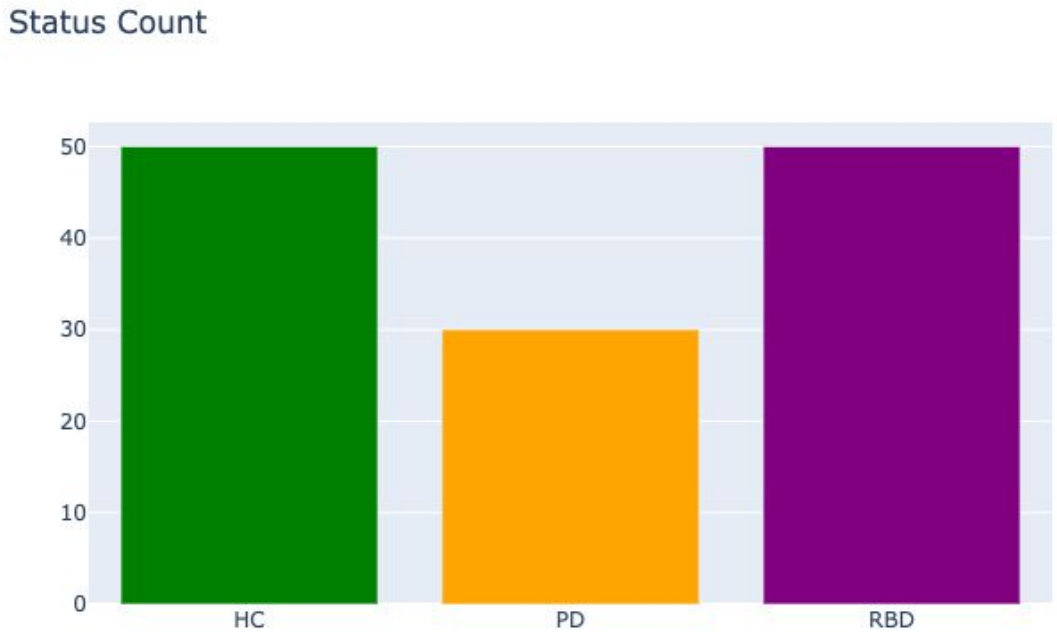
Exploratory Analysis

Visual Analysis

Note: Please reference my github account for a detailed overview of data visualization code and interactive graphs (https://bit.ly/3fieAPm)

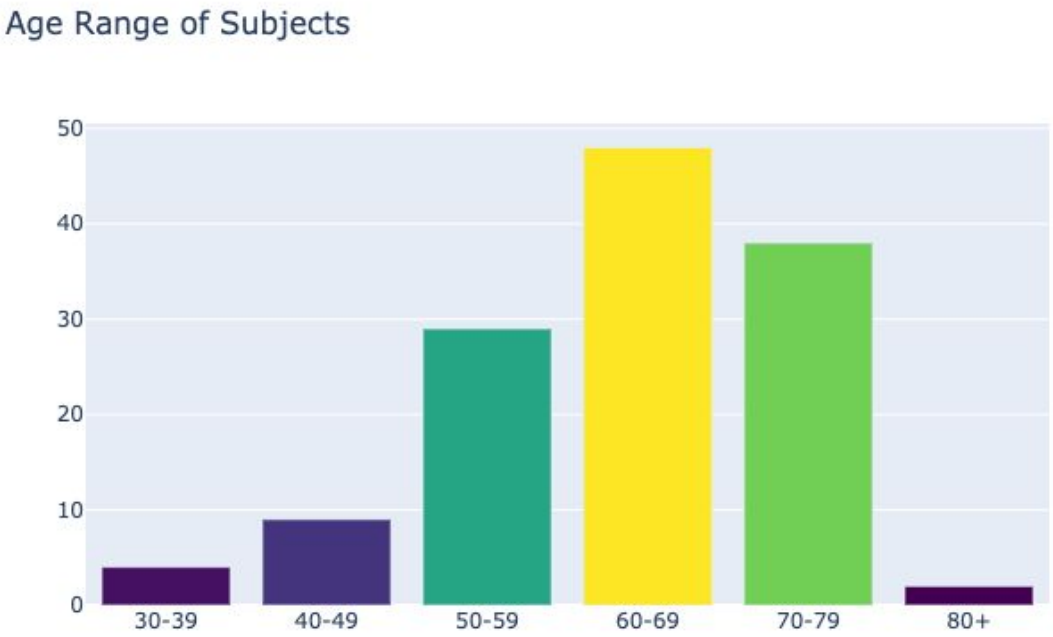
Status

PD (Parkinson's Disease) is the smallest sample group with only 30 subjects. There are 50 healthy, control subjects (HC) and 50 rapid eye movement sleep behavior disorder subjects (RBD).

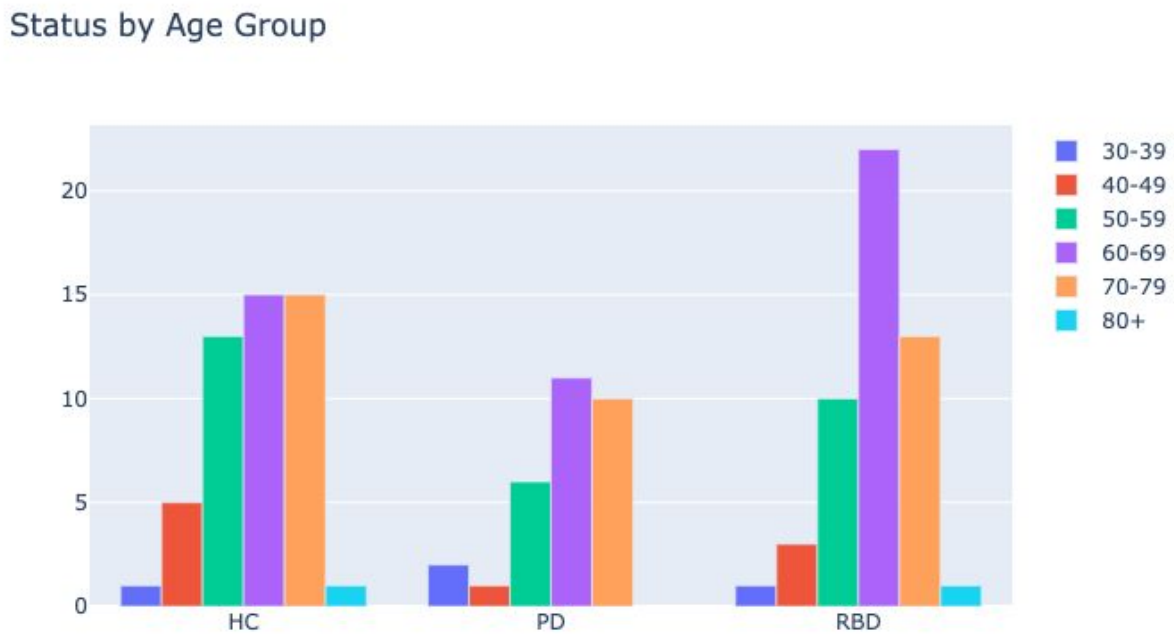


Age Ranges

The average onset age of PD and RBD subjects is 60.7 years old, which is also the largest age range group. The majority of subjects are between the ages of 50 and 79:

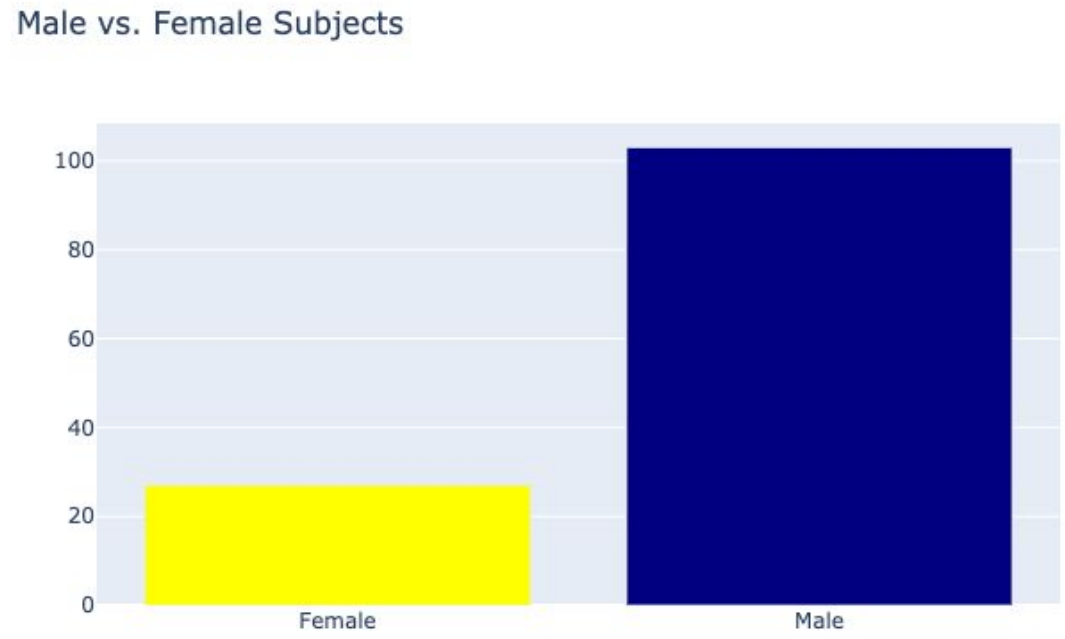


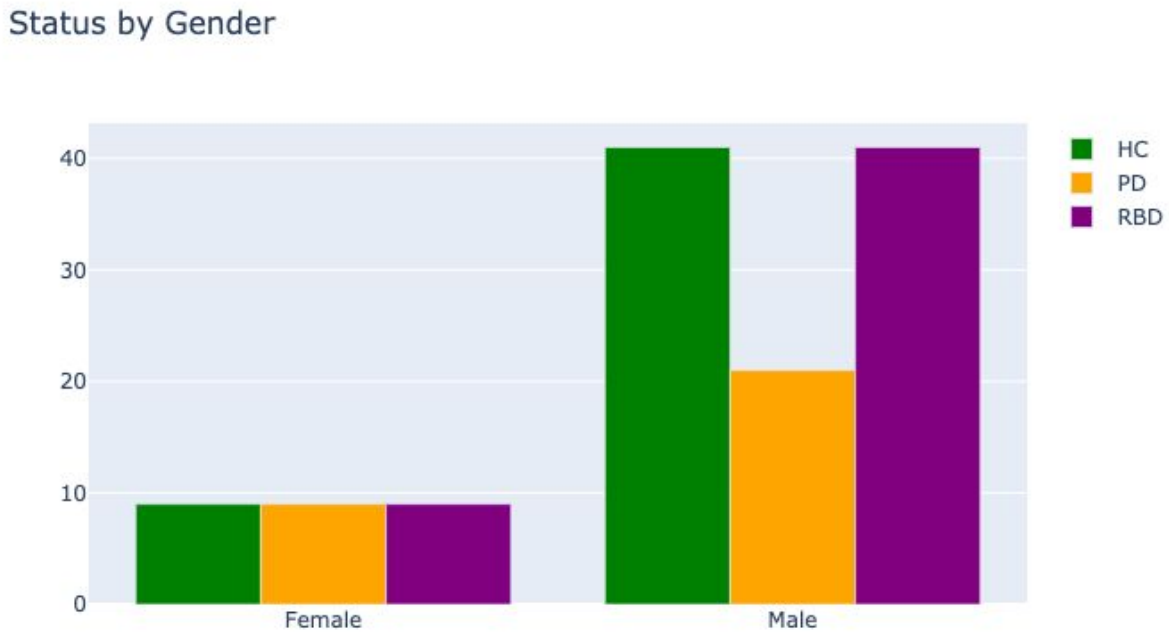
There is a small sample group in the 80+ age range and none of the subjects in that age range have a status of PD:

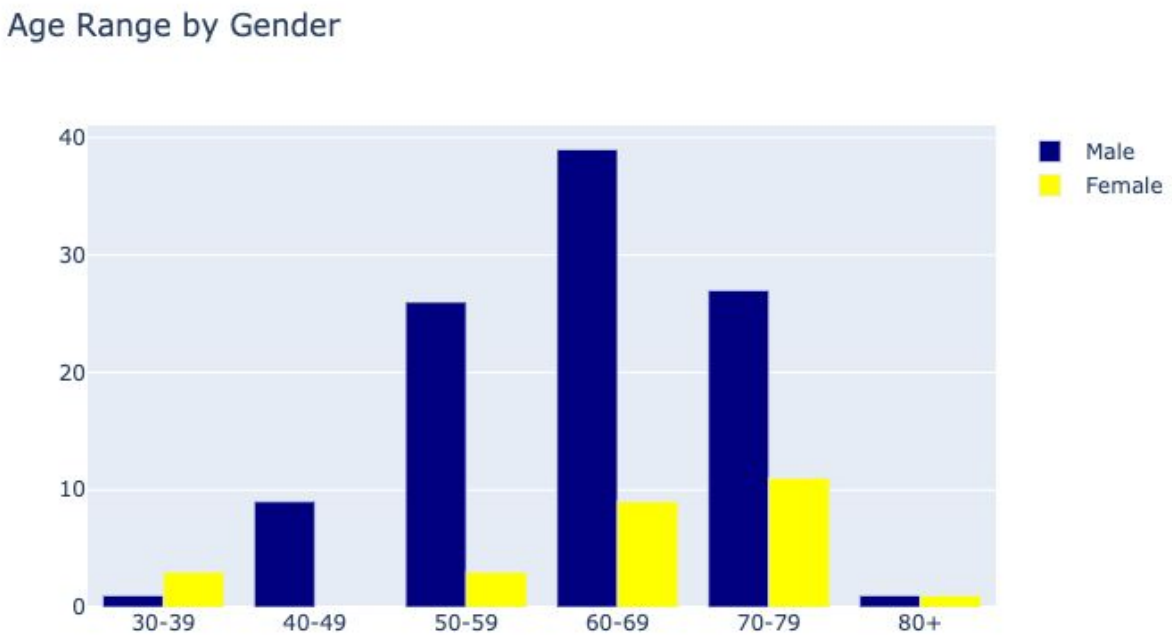


Gender

Over 79% of the subjects are male. Each status group has 9 female subjects each and there are no females in the 40-49 are range:

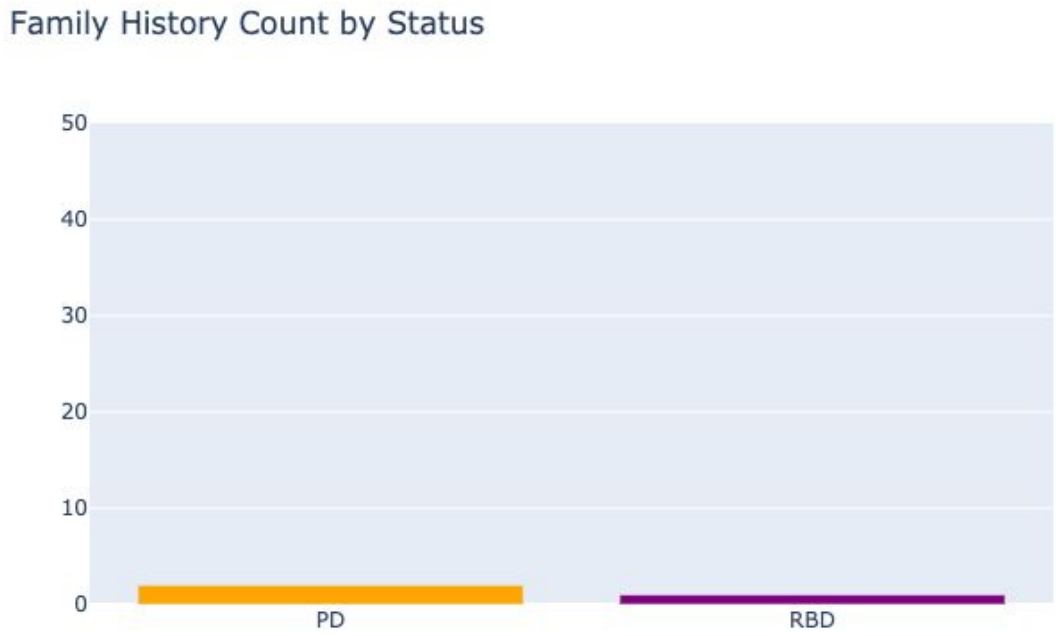






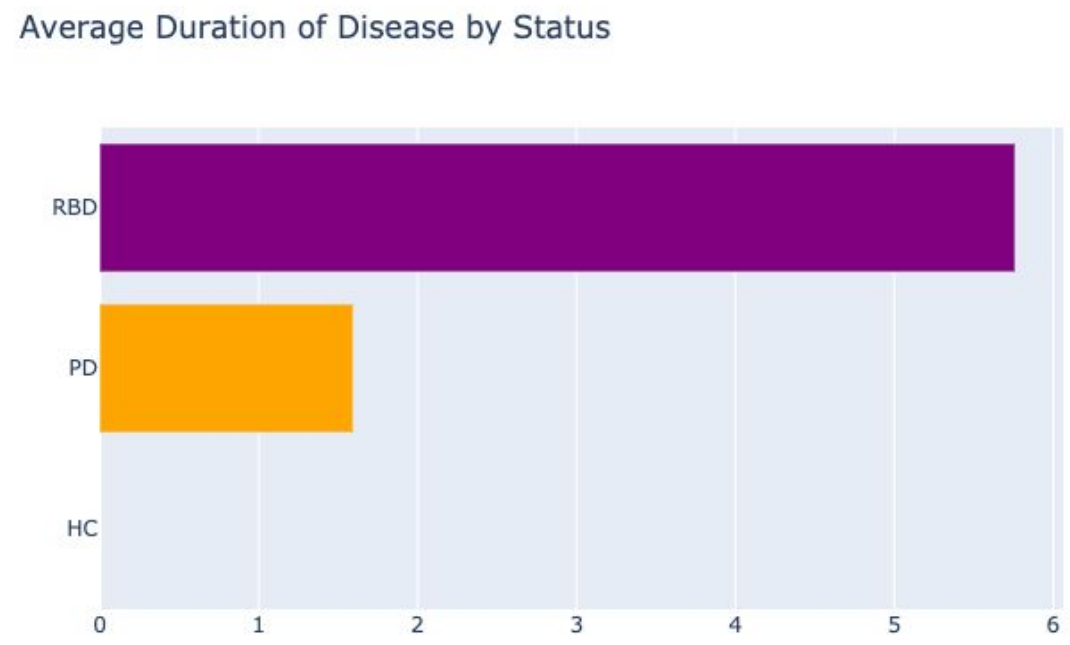
Family History

Very few subjects have a family history of Parkinson's disease:



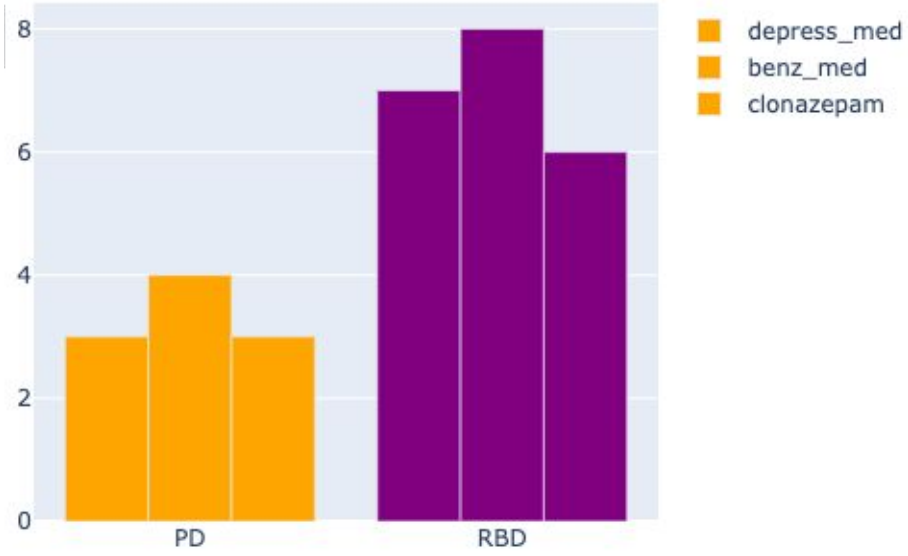
Duration of disease (in years) from first symptoms

RBD subjects have on average over 3 times longer duration of disease than PD patients:



Medication

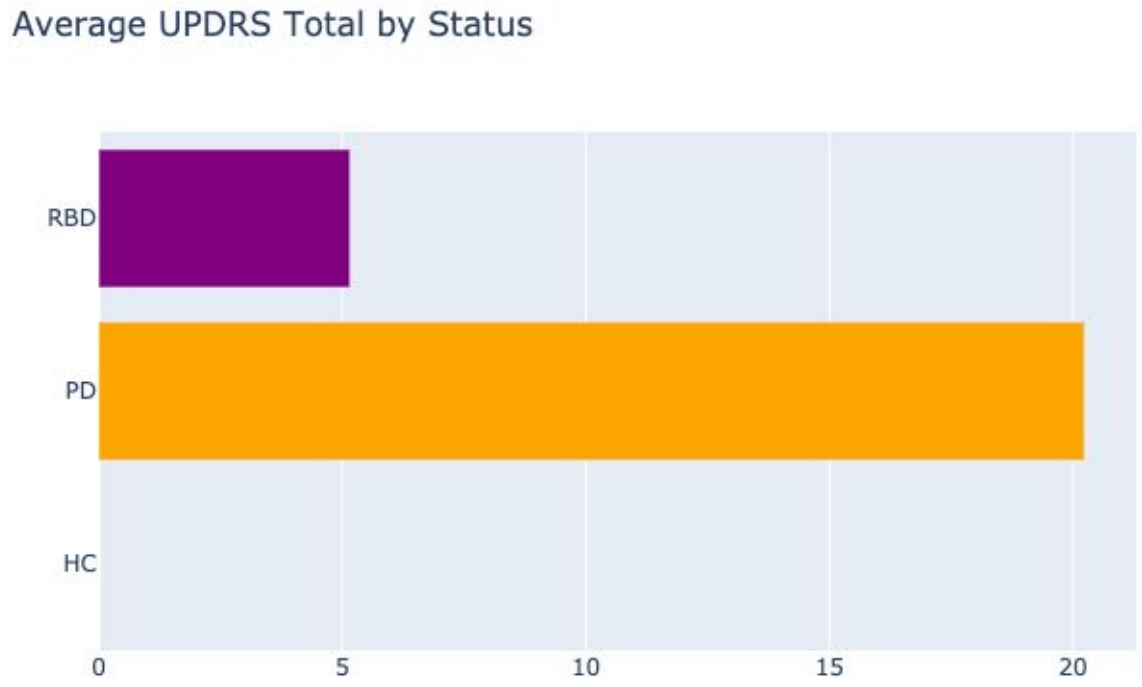
Although there is a small percentage of subjects on medication, twice as many RBD subjects are on medication than PD subjects.

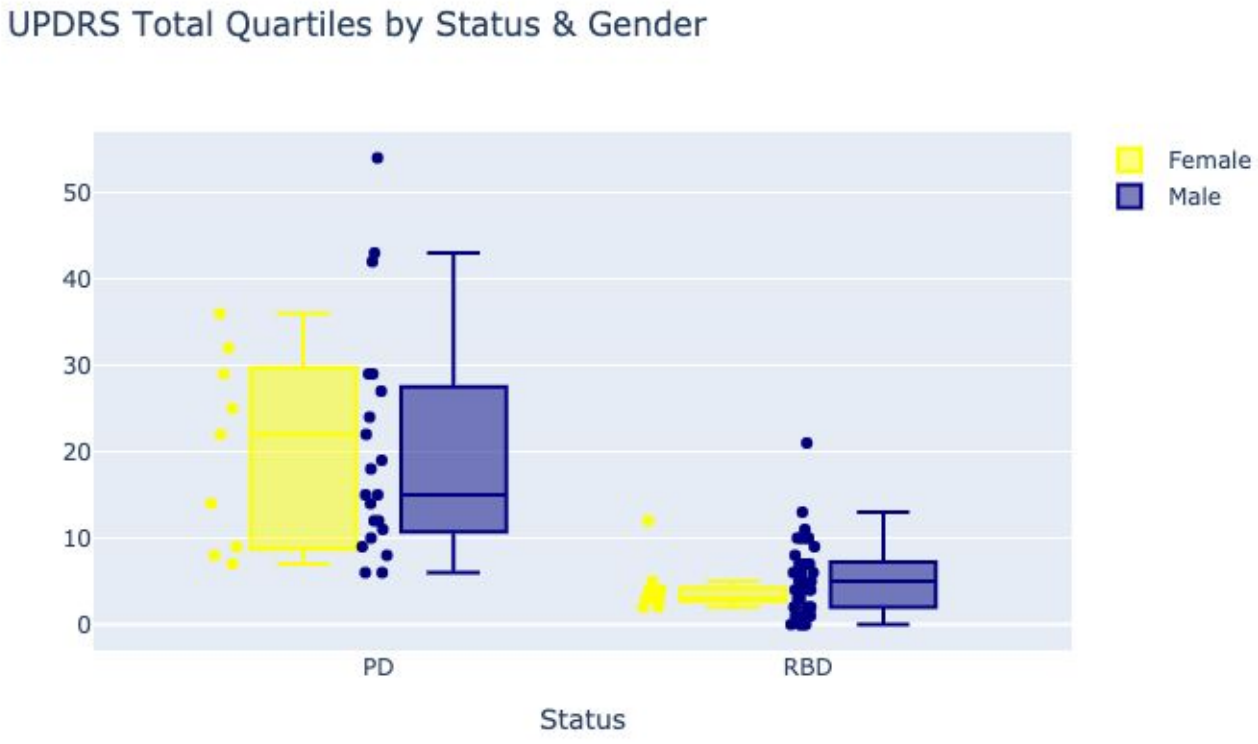


UPDRS

For each score item, a larger amount of RBD patients had lower scores. Click here to view graphs for each UPDRS item.

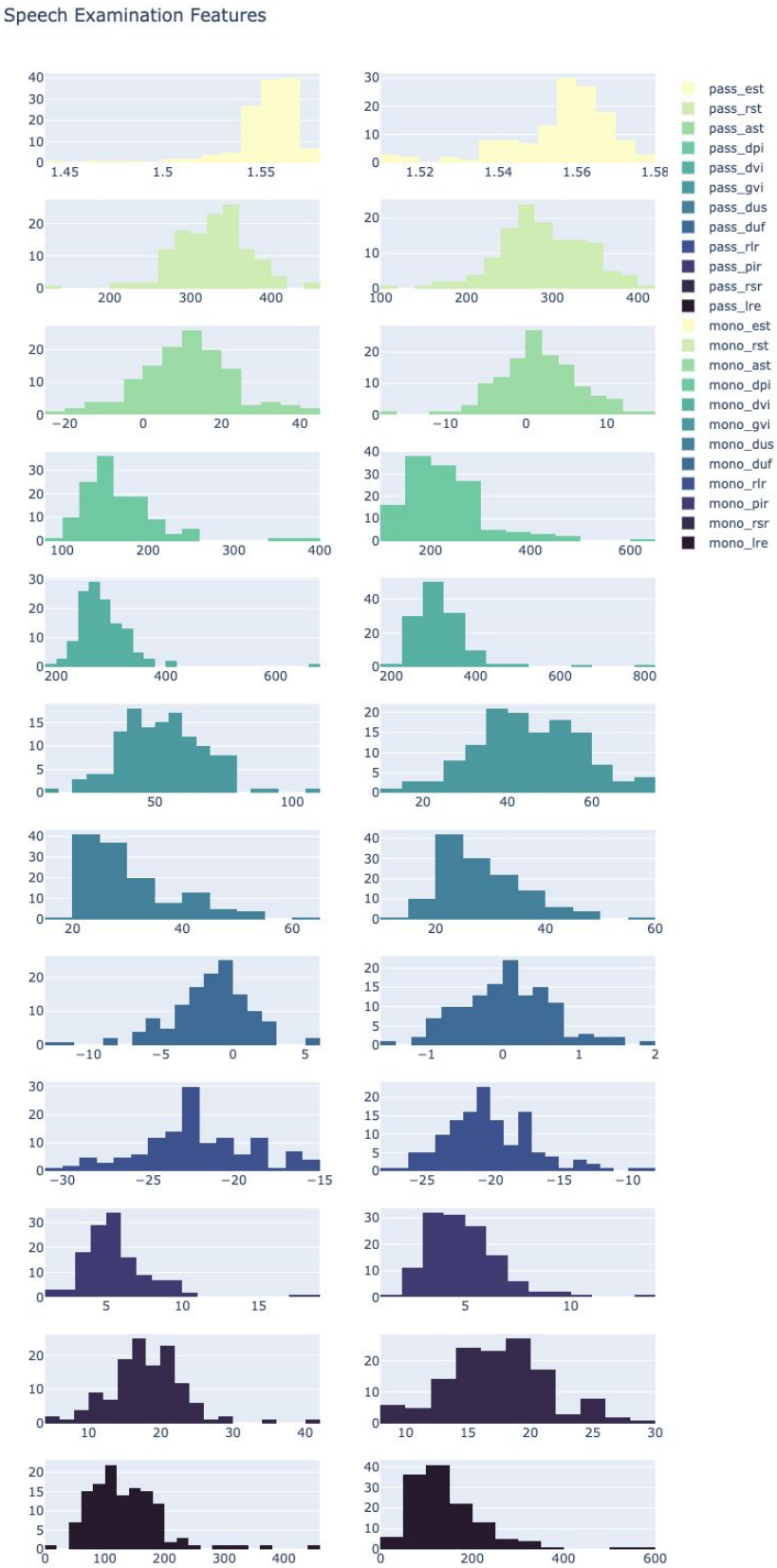
PD subjects had 4 times higher mean UPDRS total than RBD subjects. In both PD and RBD subjects, males had a wider range of UPDRS total scores:





Speech Examination

Over half of the Speech Examination features have a skewed distribution:



For many of the speech examination factors, there are not huge differences in average scores between the 3 statuses, with some features only having a difference within a few tenths of a point. Subjects with a PD status had significantly higher scores on PASS DPI and PASS DVI, while RBD status subjects

had higher mean scores on PASS AST, MONO RSR and lower average scores on PASS RBD an MONO AST. The control group had lower mean scores on MONO DPI, MONO DUF, and MONO LRE. Significantly higher average scores for the control were noted in PASS RST, PASS PIR, and MONO RST. Click here to view average score by status for each speech examination factor:

STATUS	PASS EST	PASS RST	PASS AST	PASS DPI	PASS DVI	PASS GVI	PASS DUS	PASS DUF	PASS RLR	PASS PIR	PASS RSR	PASS LRE
HC	1.553	343.652	9.472	149.819	274.346	52.467	29.170	-2.004	-22.823	6.025	18.178	146.991
PD	1.552	308.182	9.910	186.128	302.785	51.172	31.600	-1.512	-22.489	5.342	17.871	140.315
RBD	1.549	322.420	12.999	171.817	286.933	55.313	31.465	-1.256	-21.556	5.255	18.139	118.844
STATUS	MONO EST	MONO RST	MONO AST	MONO DPI	MONO DVI	MONO GVI	MONO DUS	MONO DUF	MONO RLR	MONO PIR	MONO RSR	MONO LRE
HC	1.557	311.837	1.746	198.337	303.153	46.554	25.030	-0.045	-19.887	4.980	16.354	132.886
PD	1.555	272.355	2.502	264.968	337.235	44.575	28.975	0.112	-20.749	4.283	17.871	165.754
RBD	1.554	274.361	0.845	238.247	330.109	43.935	30.160	0.097	-19.749	4.040	18.285	145.185

Statistical Analysis

Note: Please reference my github account for a detailed overview of the statistical analysis code and data (https://bit.ly/2SFsm4P)

Since the dataset contains both discrete and continuous variables, it was split into two groups. ANOVA analysis was used on the continuous variables (age and speech examination metrics) and chi-square was used on the discrete variables (gender). As was noted above, a majority of the speech examination features did not have a normal distribution. The speech examination features were normalized before conducting the ANOVA. Features were considered statistically significant and H_0 was rejected if the p-value was < 0.05.

Of the 24 speech examination factors, 10 were found to be statistically significant:

- PASS RST
- PASS DPI
- PASS RLR
- PASS LRE
- MONO RST
- MONO DPI
- MONO DVI
- MONO DUS
- MONO PIR
- MONO RSR

Chi-square was used to determine significance for gender and ANOVA was used for age. Gender had a p-value of 0.364 and Age had a p-value of .4495. Even though these two factors were not determined to be statistically significant, there are multiple studies and papers that have found age and gender to be factors of Parkinson's disease. Based on previous data, they will be included in the model.

Machine Learning

Note: Please see the accompanying jupyter notebook for a detailed overview of the code referenced below (https://bit.ly/3bKgZPI)

Preparing the Dataset

In order for the models to process the data, the target data was converted from a shape of (n,1) to (n,) using np.ravel. The data was then split into training and testing data using the train_test_split function.

Modeling

5 models were applied to the training data: K-Nearest Neighbors Classifier, C-Support Vector Classification, Nu-Support Vector Classification, Gaussian Naive Bayes, and Random Forest Classifier. Hyper-parameter tuning and cross-validation were completed using GridSearchCV. Precision, Recall, F1 Score, and Confusion Matrices were used to evaluate the models' performance. Accuracy scores were generated for both training and test data to check for overfitting.

Model Analysis

Models using the original dataset did not perform particularly well. None of the models performed particularly well when predicting if a patient has Parkinson's disease (status=1). KNN did not predict any cases for Status = 1. Several of the models' training accuracy score was significantly lower than the test data accuracy score. This is not an uncommon phenomenon for smaller datasets. The data is imbalanced since class 1 only had 30 samples compared to class 0 and 2 which had 50 samples. This can contribute to the low performance of the models. In order to mitigate this issue, RandomOverSampler was used to rebalance the class distribution.

Random Oversampling did improve the accuracy of the models, in particular the KNN model. Ensemble learning was used in order to improve accuracy. Voting Classifier and Bagging were used to experiment with improving the model. The Voting Classifier achieved an overall accuracy of 71%, although the training data has an accuracy of 81% which may be slightly overfitted.

Conclusion

Working with small datasets can be challenging, but can still provide significant insights and predictions. Considering the original study and its purpose, the recall score for the Parkinson's disease class would be the most important metric when selecting a model. In this case, the Voting Classifier model would be the best choice since it catches the highest percentage of positive Parkinson's disease cases.

The models' performance was somewhat disappointing, but not that surprising considering the small dataset. Interesting insights were gleaned from the data, especially during EDA. The data is also encouraging that a non-intrusive test like speech examination measurements have promise as a method to identify patients with early onset neurodegenerative diseases.

Next Steps

In order to increase model performance, more sample data would be needed as well as the inclusion of more female subjects and representation of all age ranges for each status, so that age and gender could have a stronger significance in the classification. If it is not possible to source additional data it may be worthwhile to investigate possibly reducing the number of classes to two where 0 = Healthy and 1 = Neurodegeneration Disease.