Predicting Parkinson’s Disease Progression with Smartphone Data

Competition Synopsis

Predicting Parkinson's Disease Progression with Smartphone Data was a competition hosted on Kaggle seven years ago. The overarching goal of the competition was to use a large collection of raw signal data collected via smartphones as a proxy for measuring the symptoms of Parkinson's Disease [PD]. PD has no known cause and is typically diagnosed on the basis of specific symptoms presented, the primary ones being a general decrease in movement and increasing tremors. The primary stage of this illness consists of mild symptoms such as changes in posture or demeanor. The most severely impacted patients experience delusions, hallucinations, are usually bedridden, and require full-time assistance from a caretaker.

Via critical reasoning and experimentation, we attempted to use the data collected to optimally predict which participants had PD. As an extension of the experiment, we investigated the relationship between a participant’s audio, location (latitude and longitude), and diagnosis.

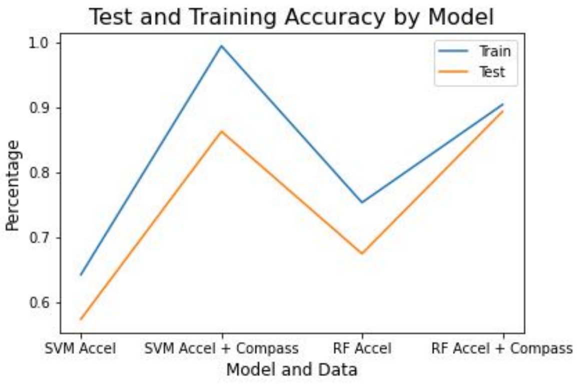
Data Pre-Processing

Preprocessing the data was undoubtedly the most time consuming part of our project. We struggled to download and preprocess the raw data efficiently on our local machines, given that the size of the data was incredibly large (20 GB). After trial and error, we wrote scripts that chunked the raw data files into halves, parallelized the pipeline to process multiple chunks at once, and concatenated the data frames together.

We utilized accelerometer data because we believed it would provide insight as to which participants lacked stability. Sharp recordings would indicate balance issues as well as tremors, both of which are presented in individuals with PD. We used the compass data to further these findings, as this information could also give insight as to whether an individual was prone to falls or frequently dropping their phone. Both of these characteristics would be frequently expected in an individual with PD. After conducting some preliminary research and exploration, we grouped our accelerometer and compass data by person, day, and sometimes hour. We also took the root means square of all the x, y, and z data in the accelerometer data frame, because many of the measurements were dependent on the participant’s phone’s position, which was different for everyone. This gave us enough data on the average amount of movement per hour for each participant. There were also gaps in the timestamps of the data, indicating that the participant on occasion forgot to carry their phone. We ended up dropping those null measurements and included the mean directions of the phone (roll.mean, azimuth.mean, pitch.mean) derived from the compass dataset.

Model

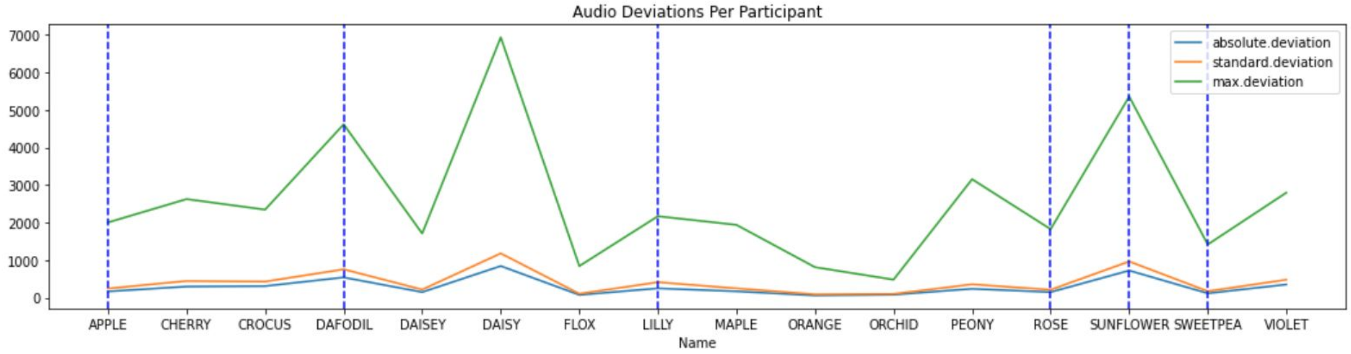
We initially trained a *SVM* classifier in Python on the pre-processed accelerometer data. This model fared poorly, with 65% training accuracy and 57% testing accuracy. It also consumed a lot of time and processing power. As a result, we shifted to a more powerful *Random Forest Classifier* and used random search (sklearn's *RandomizedSearchCV*) with *KFold* cross validation to tune the model's hyper-parameters. We got better results with an RFC which gave us 75% training accuracy and 67% testing accuracy on the accelerometer data. Although this made the testing accuracy higher, we were confident that we could do better. Realizing that we were still missing crucial features in our feature set, we added the phone's direction (found in the compass dataset), we were able to increase both our training and testing accuracy significantly to 91% and 90% in an RFC and 99% and 86% in a SVM.

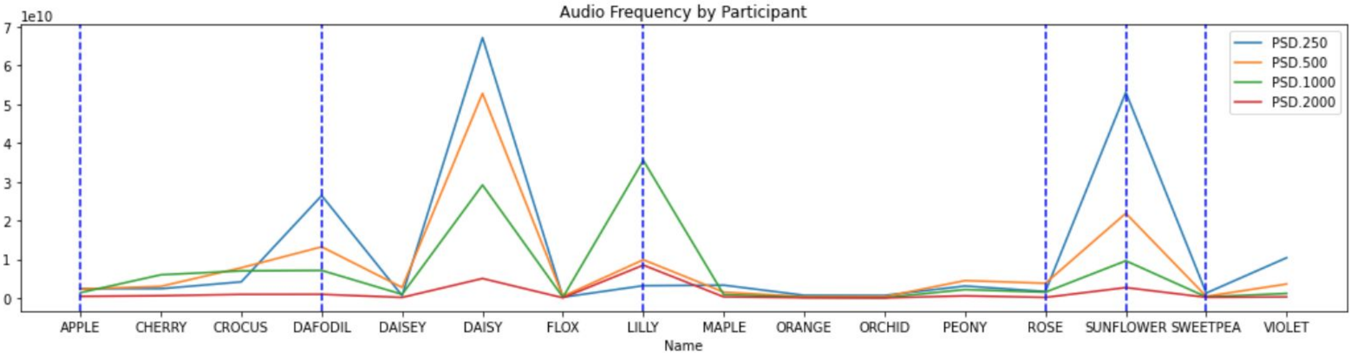


Exploratory Analysis

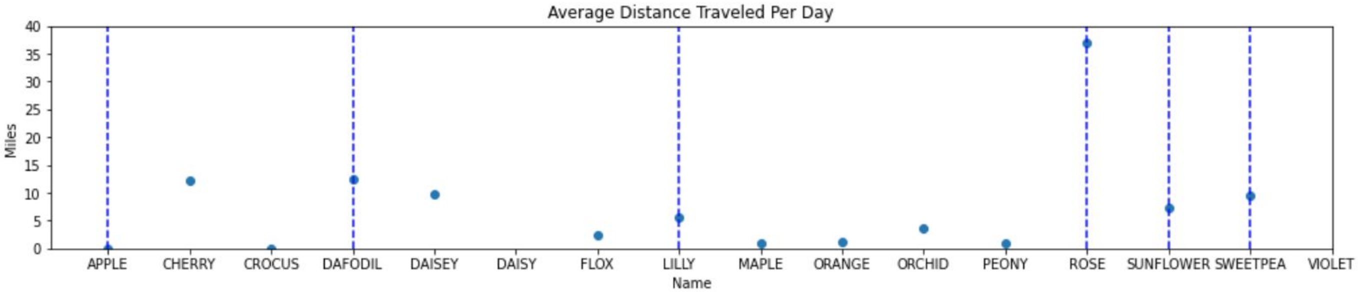
For our further exploratory analysis, we investigated the audio frequencies as well as GPS data. We read that a PD patient may have a more monotonous voice and experience a more restless sleep. We also thought that those with PD might travel less on average per day, given the degeneration of mobility that occurs. Thus, we also investigated the GPS data which could serve as a proxy for movement patterns.

After grouping the audio data by each participant, we plotted the mean frequency and deviations for each participant. The figures below show how there is no obvious correlation between these data points and diagnosis (participants with blue lines are the control group).





We also computed the average distance traveled each day for each participant using the longitude and latitude columns in the GPS dataset. We used geopy's *vincenty* function to compute the distance a participant traveled between points over intervals of an hour. We then summed it up and looked at the average distance traveled per day. While some PD patients traveled less on average, we realized that there are too many confounding variables-- such as varying methods of transportation, age, and occupation-- to definitively link distance traveled per day with patient outcome. In fact, adding this as a feature to our classification slightly lowered its accuracy. Interestingly, when plotting the GPS data on a map, using *GeoPandas*, we were easily able to determine the participants' "homes", daily routes, and how long they spent at each location, which poses serious privacy concerns with how participant data is handled and used.



Results and Conclusion

The challenges we faced in this project primarily arose while pre- processing the data. Moreover, we faced difficulties debugging our code and collaborating via a remote environment. The most difficult part of the data we were working with was the number of potentially confounding variables that could be present. For example, this study only consisted of 16 patients, all of whom were taking a wide variety of prescribed drugs. This in itself poses a problem, as we would not be able to discern what specific side effects patients are experiencing from the other drugs they are taking, and if those impacted the data that was collected.

Moreover, we are not aware of any of the living conditions the participants are in. This in itself can invalidate large amounts of data. For example, in the case of the audio readings, this data could mean a wide variety of things based on the individual's lifestyle. High levels of audio could imply that the individual simply had a busy social life. Or, these high levels of audio could be attributed to a caretaker and support system that has been put in place for the individual because they are experiencing heavy symptoms from PD. The range in meanings that could be attributed to this data was so large that it makes sense we were unable to derive any sound conclusions from our analysis.

Ideally, if this data could be collected again, this would be done via a longitudinal study with far more participants. As a result, we could gather a baseline of information regarding each participant's day-to-day lifestyle in order to note changes as the onset and progression of PD occurs.

References

"Stages of Parkinson's." Parkinson's Foundation, www.parkinson.org/Understanding-Parkinsons/What-is-Parkinsons/Stages-of-Parkinsons.