|  |  |
| --- | --- |
| **Machine Learning, Vocals, and Parkinson’s Disease** | |
| Contact | Name: Daniella DeWeerd  Email: ddeweerd@bu.edu  Cell Phone: (616)-566-4293  Title:  Address: |
| Organization | Name of organization |
| Organization Description | Mission of Organization |
| Project Type | Data Science |
| Project Description | The goal of this project is to use vocal data from patients with and without Parkinson's Disease (PD) to classify whether a person has or does not have PD with above an 80% accuracy.  The current way that PD is diagnosed is through a neurologist. The neurologist will make the decision that a person has PD based on their medical history and symptoms. The neurologist may also try to rule out other diseases through an MRI, blood test, and/or DaTscan. They also may try to give them a PD medication and if it helps, it will help to convince that they have PD. No specific test is available which can lead to a wrong diagnosis. By being able to diagnose PD through other methods such as a person’s vocals, the neurologist will be more accurately able to diagnose PD. As PD progresses, the muscle of the vocal cord becomes thinner and less taught and changes the person’s voice as well as their intelligibility which is a change that could hopefully be noticed in the early stages of PD to help with diagnosis.  This model would be important to both the people trying to get diagnosed as well as the doctors trying to diagnose them. This diagnosis provides the doctor with more certainty a patient has PD which can ensure they can get treated properly. If a doctor is not sure enough, they may say its something else and mistreat you allowing the disease to progress enough that delaying progression isn’t possible. This method will only be able to function as a tool like any of the other tools mentioned above to help convince a neurologist that a patient does or doesn’t have PD. Since it would be combined with other tools, it wouldn’t have for sure effect a doctors decision to diagnose a patient, so it would be unlikely to be able to be misused and false positives or false negatives wouldn’t have a definite affect. Also, when a doctor diagnoses PD, they tell their patients a percentage of uncertainty since there is no definite way to diagnose, so the uncertainty of the algorithm would be included in the diagnosis. |
| Data Sets & Sources | https://archive.ics.uci.edu/ml/datasets/parkinsons |
| Suggested Steps | * Collect the data from the above source(s) * Clean/format the data * Display the ratios of PD vs Healthy Controls (HC) * Display each of the attributes and color by who does and doesn’t have PD (Discover if there is a visual difference) * Display the gender and age separation between HC and PD to see if there is any bias * Normalize the data * Run the following algorithms on the data (could add more if most have bad performance):   + Random Forest   + Logistic regression   + SVM   + Naive Bayes classifier   + Neural Network   + K-nearest neighbor   + PCA   + AdaBoost * Tweak parameters of each as necessary to achieve the best results |
| Questions to be answered in Analysis | * Is it possible to predict PD from vocals with machine learning? * What algorithm works best to predict Parkinson’s Disease (PD) vs the healthy controls? * How accurate can you get without overfitting? |
| Ideal Output + Final Deliverable | The ideal output would be a presentation on the data visualized, the final algorithm used, why that algorithm is the best, and why the results of this project are important. |
| Additional Information | The dataset was created by Max Little from the University of Oxford. The National Centre for Voice and Speech, Denver, Colorado recorded the speech signals. Each patient was recorded approximately 6 times and the following are the features that are shown in the data: Average vocal fundamental frequency, Maximum vocal fundamental frequency, Minimum vocal fundamental frequency, measures of variation in fundamental frequency (MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP), measures of variation in amplitude (MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA), two measures of ration of nois to tonal components in the voice, two nonlinear dynamical complexity measures, a signal fractal scaling exponent, nonlinear measures of fundamental frequency variation (spread1,spread2,PPE), and whether a patient has or does not have Parkinson’s.  This dataset has two papers connected to it which contain a bit more information on the dataset. The first is named [Suitability of dysphonia measurements for telemonitoring of Parkinson's disease](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3051371/). The other is named [Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection](https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-6-23). In the first paper, the specifics of the dataset can be found.  Some limitations of this dataset is that it is only made from 31 male and female subjects, though there are much more data points because there are 6 recordings for each. 23 of those have PD so the set definitely leans towards those with PD. Another thing to note is that the subjects range from 46 to 85 so I need to look into whether those with and without Parkinson’s lean towards a certain age group which could affect the predictability of the algorithm. Graphs of the age and sex will have to be included in the final presentation. Another limitation is that we can’t compare Parkinson’s to any other disease to see if the algorithm can specifically pick out Parkinson’s or if it is just finding those with a disease that affects the voice. |
| Deliverable 1 | After an analysis of the data on the worries I voiced in the previous paragraph, I found that the people with PD range more from 60s to 80s in age while people without range more in the 50s to 70s range. While this most definitely could affect the pitch and such of the voice, it is not included directly in the data we are using to predict, so it may be okay for now. As for age, this is another variable that isn’t directly included in the data, so it may be okay, but there is a larger male presence in those with PD and a larger female presence in the HCs.  In the quick code I wrote up, I used Logistic regression and random forest to try and answer my first question of whether it is possible to predict PD. Based on the results with RF pulling just above a 90% accuracy on average and LR pulling a middle to high 80%, I think it is possible. One thing I am noting for further exploration is the train test split and whether or not it is a fair split with its PD vs HC ratio. I also think a limit is going to appear with how small the data set is, so I would like to do research on if any other data was added to this set at a latter date. |
| Deliverable 2 | First, a note in regards to a previous question of if any additional data was added to the dataset at a later date. There wasn’t anything added to this particular data set that I can work with in this project, but the creator of the data created another data set in 2021 that is very similar. (A further description can be found in the following link to a paper: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8564663/>). This data is not publicly available based on what I could find, so even if the variables matched up, I couldn’t add it to what I have now.  Running all of the algorithms that I wanted to try in the beginning, I found that Random forest and K nearest neighbors worked the best with RF hovering around a 90% accuracy and KNN above 90%. I also think that with the proper changes to PCA, it may also be able to reach a better percentage than 87%.  The question I was able to answer partially with this deliverable is what algorithm(s) work best. The next steps are to fine tune them with a focus on KNN to see if I can increase the accuracy by a decent amount. I also want to include some cross validation if I can and write a whole new script for that.  I also can start working on the presentation of the data as I have reached some conclusions on the predictive ability of it when looking at the algorithms accuracy as a whole. |