Applying Data Mining to Parkinson’s Disease Data

Joshua Gearhart

Senior in Computer Science and Cybersecurity

Slippery Rock University

1. Introduction

The goal of this paper is to analyze Parkinson’s Disease, and to use data mining techniques to analyze patterns, and classify patients possibly with Parkinson’s disease. Parkinson’s disease otherwise known as PD is a progressing disorder, it affects the patient’s nervous system and slowly affects them[CITATION!!!!]. The point of this data mining paper is to take data from the Parkinson’s Disease data set [CITATION] and to analyze the patients recorded speech data to be able to formulate an opinion on if they have PD. In this paper topics that will be discussed will be as follows:

Literature review- We will discuss what references were used in this type of paper, how they came in handy and what we gathered from them.

Methods – The methods of which that were used to come to our results, and how this translated into our results.

Results – The findings of this research compiled into a section. This section will talk about our findings and present them to you as the reader. We will talk about was used and how they compared to each other.

And the paper will end with a discussion and references used in the paper.

1. Literature Review

The literature review in this paper was integral to understanding the domain and aspects of Parkinson’s disease(referred to as PD in this paper). Many sources were re-viewed throughout the time that was taken to compile this project, not only did many scientific articles come in handy for establishing an appropriate flow to the paper but throughout this project. Non-scientific articles were also looked at, these articles typically gave information on the domain of PD. It’s important in data science to compile a complete and broad understanding of the domain you will be analyzing before doing so. This gives researchers a credibility and confidence in the information that they are presenting is not only credible, but that they have a knowledge to comprehend the data itself. In this paper I started with researching well what is Parkinson’s Disease, I’ve heard the name, but I wasn’t sure on the specifics.

1. Method

In data mining often the first step before even pre-processing is to get an understanding of the domain you are working with. This was the first step taken in this report to ensure that the researcher understands different data values, what the data set means, and how it can be correlated to solving the problem. The first step was to retrieve the data set, the Parkinson’s Disease data set was obtained through Kaggle. Often these datasets come with some information, reading about the data we see this paragraph.

“This dataset is composed of a range of biomedical 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 one of 195 voices  
recording from these individuals ("name" column). The main aim of the data  
is to discriminate healthy people from those with PD, according to "status"  
column which is set to 0 for healthy and 1 for PD” [CITATION!!!!!]

From this paragraph we gather some important information, the first being that 1. This data set is dealing with voice measurements, we also can gather that there are thirty-one different subjects in this dataset, more importantly though we have twenty-three positive PD patients. Next, we can get a little idea about the shape of the dataset, we see this in that there are 195 voice recordings, which means we should have 195 rows of data. We also get clarification on a what could be considered more of an abstract column which is the “status” column, in this paragraph it is indicated that zero equivalates to a healthy person and one is for a Parkinson’s Disease positive patient. Reading on from the Kaggle source, we see attribute information, This attribute information is highly integral to understanding the domain of PD. Because we can now understand each column, and what to expect out of those values. See figure 1 for information on each column.

|  |  |
| --- | --- |
| Name | ASCII subject name and recording Number |
| MDVP:Fo(Hz) | Average vocal fundamental freq. measured in Herts |
| MDVP:Fhi(Hz) | Maximum vocal fundamental frequency measured in Herts |
| MDVP:Flo(Hz) | Minimum vocal fundamental frequency measured in Herts |
| MDVP:Jitter(%) | MDVP jitter in percentage |
| MDVP: Jitter(ABS) | Kay Pentax absolute jitter in microseconds |
| MDVP:RAP | Relative amplitude perturbation |
| MDVP:PPQ | Five-point period perturbation quotient |
| JITTER:DDP | Average absolute difference of differences between cycles divided by average period |
| MDVP: Shimmer | Local shimmer |
| MDVP:Shimmer(dB) | Local shimmer in decibels(dB) |
| Shimmer: APQ3 | 3-point amplitude perturbation quotient |
| Shimmer:APQ5 | 5-point amplitude perturbation quotient |
| MDVP:APQ | Elven point amplitude perturbation quotient |
| Shimmer:DDA | Average absolute difference between consecutive differences between amplitude of consecutive periods |
| NHR | Noise of harmonic ratio |
| HNR | Harmonics to noise ratio |
| STATUS | Boolean Value 0 – indicates healthy 1 – PD positive |
| RPDE | Recurrence Period density entropy |
| D2 | Correlation Dimension |
| DFA | Detrend fluctuation analysis |
| Spread1 | Nonlinear measure of fundamental frequency |
| Spread2 | Same as spread 1. |
| PPE | Pitch Period Entropy |

Figure – 1. Attributes and their explanations

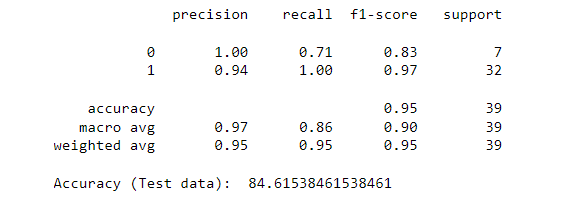
Note: MDVP stand for Multi-Dimensional Voice Program. [CITATIONS]

The next step for preparing for fitting models was checking for null values a simple code line was applied to the data, and zero nulls were returned, this makes the job of pre-processing a lot easier as now we don’t have to deal with null values within our data. The data set also had proper data types when loaded into python, this was checked by using the Pandas info method. This isn’t a huge issue to fix, but the process of converting values is thankfully not needed. Next it was time to check for the split of positive and negative patients. “data['status'].value.counts()” was used to show that we have 147 instances of positive PD patients and 48 negative patients, This is important to note for a couple of reasons that will be discussed at the end of the paper. As per standard practice to pre-pare for data modeling we scaled the data before modeling. We will discuss this in the results as well later. This ensures that outliers are standardized instead of removing them, they help refine the model.

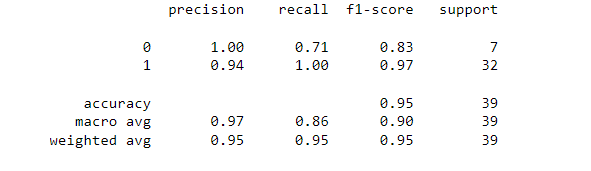
1. Results

In this section of the paper, we will talk about the results of our implemented methods. In this project we implemented two ML algorithms to try and predict PD in patients.

Implementation of SVM



Implementation of XBGC Classifier



The goals of this paper were to predict Parkinson’s with increasing accuracy using different methods of data predictions. This method would allow a comprehensive analysis of model predictions of varying models on the PD dataset, accessing what model is the best is integral for predicting PD, as if you were to implement this on patients, you’d very much want to be accurate.

1. Ending Discussion

The goals of this paper were to predict Parkinson’s with increasing accuracy using different methods of data predictions. This method would allow a comprehensive analysis of model predictions of varying models on the PD dataset, accessing what model is the best is integral for predicting PD, as if you were to implement this on patients, you’d very much want to be accurate.

A lot more positive patients than negative, analytics could be improved by having more negative patients for more information for models to learn form.

References