DSC672 Data Science Capstone – Project Report Fox Insight Parkinsons Data

Group 4: Aswin Guvvala (aguvvala@depaul.edu), Wei Tong Su (wsu6@depaul.edu), Michael Ruggeri (mruggeri@depaul.edu)

ABSTRACT

The Fox Insight Parkinson's Dataset provides a vast window into the struggles of patients fighting Parkinson's disease. This demanded a multi-faceted approach to the analysis in order to mine value from the extensive catalog of questionnaires from both one-time studies, as well as longitudinal data capture. In this analysis, we employed both supervised and unsupervised methods to reveal subtypes of the disease, understand what factors were statistically significant in treating the disease, as well as to use a broad array of data to predict the diagnosis of Parkinson's disease. Clustering techniques allowed us to define subtypes of Parkinson's disease, which mainly serves as a step towards future work in understanding the disease. Our study of cannabis use in treating Parkinson's disease suggested that it helped patients sleep.

INTRODUCTION

Parkinson's disease is a degenerative brain disorder that causes involuntary shaking and physical imbalance. It is a progressive disease which causes ambulatory or verbal difficulty. It is a disease that does not often cause mortality, but the condition often functions as a comorbity to other fatal diseases. Parkinsons patients can expect to have near-normal life expectancy. It is characterized nerve cell degeneration in the substantia nigra, which governs dopamine production. Genetics may play a role in the disease, though it is not believed to be hereditary. Research suggests that Parkinson's disease results from genetic and environmental exposures, including head injury, alcohol consumption, and toxic chemicals perhaps as a result of hazardous working conditions.

LITERATURE REVIEW

Our clustering approach was mainly based on previous work, particularly in Fereshtehnejad, Seyed-Mohammad et.al. [1]. Their dataset contained similar data, including genetic markers, which was also incorporated into our analysis. Also similar were the elements of longitudinal study of the disease.

We investigated methods of perhaps reducing the dimensionality of our data, such as in Maaten and Hinton [2]. Their description of t-SNE may be useful in future work for both visualization and further modeling.

DATA

One of the challenges of the dataset was the questionnaire format. There seemed to be too many gradients to each answer category. For instance, for the environmental exposure questionnaire for smoking, one question asked, "During the time that you regularly smoked, on average, how much did you smoke per day?" The multiple-choice answers included the following: "1. Packs per day; 2. Cigarettes per day; 3. Don't Know; 4. Prefer Not to Answer." In this case, "prefer not to answer was treated as "no". We know that the patient smoked to some extent, so answers of "Don't Know" were given a value of 1.5, halfway between "cigarettes per day" and "packs per day." It would have better served the analysis not to introduce noise with "I don't know" responses.

In this study, we used data from three sources: the PD-PROP 2.0 dataset, the Cannabis Use in PD dataset and a general dataset containing information about whether a patient has Parkinson's Disease or not. The PD-PROP 2.0 dataset contains 14 domains where patients were asked to rank the botherness caused by particular symptoms. Patients could rank any number of symptoms in a particular rank.

The Cannabis Use in PD dataset contains information about THC and CBD dosage with different categories for the amount taken per day. This data was transformed such that THC and CBD use were coded as binary variables (1 for use and 0 for non-use). This transformed data was combined with the symptom data from the PD-PROP 2.0 dataset to analyze any potential relationships between cannabis use and symptom botherness.

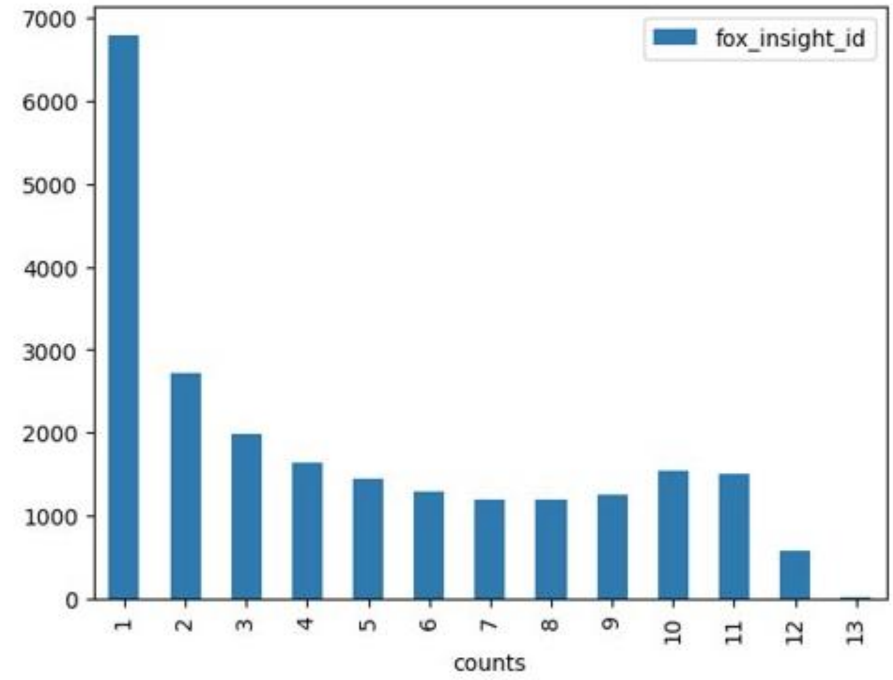
The dependent variable in this analysis is symptom botherness as measured by the PD-PROP 2.0 domains while the independent variables are cannabis use (THC and CBD), age and whether or not a patient has Parkinson's Disease.

For the Environmental Exposure: Caffeine questionnaire, patients were questioned about intervals in their life when they may have stopped and then restarted caffeine consumption. This seems to be an attempt to make the one-time survey more longitudinal. Realistically, this made the responses difficult to deal with both for patients and analysts. Unlike the actual longitudinal part of the dataset, these responses lacked a regular time variable. Each question could represent a variable number of years of exposure, or weeks, or just days. Since very few patients even bothered to record this historical data, and since trying to account for the cumulative time led to negative time of exposure, it made more sense to avoid this level of granularity. Instead, we relied on the breadth of the questionnaires to create a more holistic picture of the patients.

METHODOLOGY

For the longitudinal portion of the data, alignment of the data points was an issue. Patients were scheduled to give data on different days, and some skipped questionnaires on their schedule of activities. For example, the Brief Motor Screen Questionnaire may have 13 data points for a given participant, or it may have just one response for the entire length of the study. To add even more complexity, each questionnaire falls on different days for patients, at different intervals, such as each study visit, every three months, six months, yearly, etc. This makes merging the datapoints from all questionnaires difficult, since the data streams in erratically over time on different days depending on the questionnaire, in different amounts for each participant.

Number of longitudinal data points for participants taking the Brief Motor Survey.



Most participants only completed the survey on one occasion, which is useful for static analysis, though not conducive to any time-series modeling.

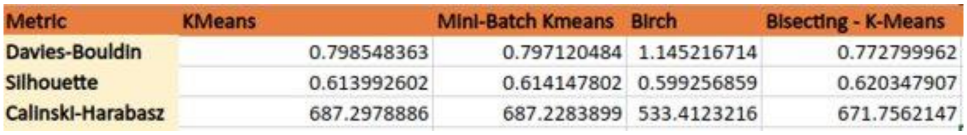
The goal of this data analysis project was to investigate the effects of cannabis (both THC and CBD) and age on the botherness of patients with Parkinson's disease (PD). To achieve this goal, data from multiple CSV files were merged and cleaned. The code reads in three CSV files: Cannabis_Use_in_PD.csv, PDPROP2.csv, and General.csv. These files are merged together using the fox_insight_id column as the key. After merging the data, a subset of columns is selected and sorted by fox_insight_id and age_x. The resulting DataFrame is then cleaned by replacing any NaN values with 0 using the . Finally, two additional columns (CannCDBdose and CannTHCdose) are added to the DataFrame from the original cannabis DataFrame. The values in these columns are then replaced in-place with 1 if they smoke and 0 if they do not smoke.

The final DataFrame is further processed by sorting it again by fox_insight_id and age_x and selecting the last row for each fox_insight_id. The CannCDBdose and CannTHCdose columns are then converted to numeric values by removing the 'mg' string and stripping any whitespace. Next, several new columns are created by checking if any of the values in a set of related columns are equal to 1. If any of these values are equal to 1, then the corresponding new column is set to 1. Otherwise, it is set to 0. Finally, all of the original related columns used to create these new columns are dropped from the DataFrame.

Several logistic regression models are then fit to the data in the final DataFrame. Each model is fit using a different response variable (Pdprop2Cog,'Pdprop2OthMot, Pdprop2Pain,Pdprop2Slp,Pdprop2Brady, Pdprop2Trem, Pdprop2PI).

RESULTS AND CONCLUSION

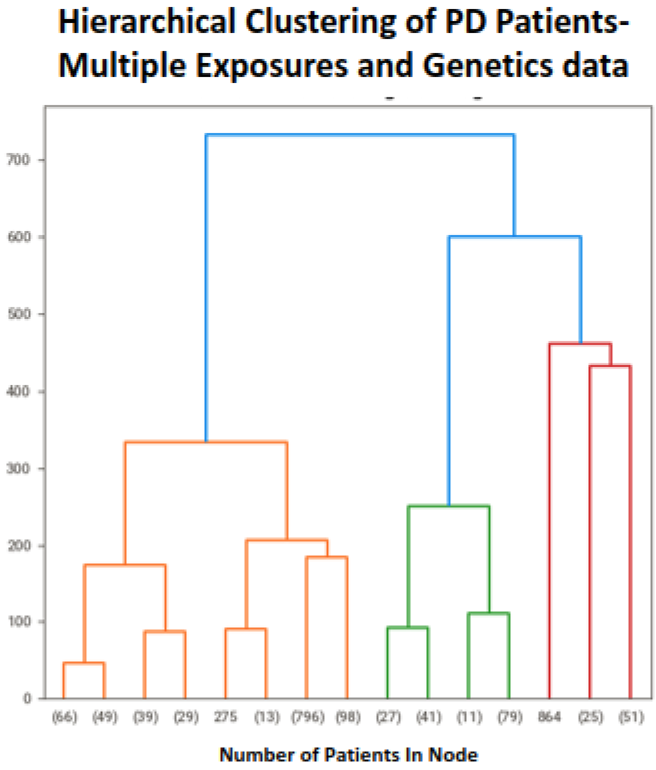
The results of our regression model were in line with our expectations. The fact that cannabis aided patient sleep was an expected result, since it is a substance known to be a depressant and induces relaxation. We had no expectation of which model produced the best clustering results. Truly, the scores were mixed, but mini-batch k-means clustering produced the best model for subtyping the disease.



Metric	KMeans	Mini-Batch Kmeans	Birch	Bisecting - K-Means
Davies-Bouldin	0.798548363	0.797120484	1.145216714	0.772799962
Silhouette	0.613992602	0.614147802	0.599256859	0.620347907
Calinski-Harabasz	687.2978886	687.2283899	533.4123216	671.7562147

Results of clustering for subtypes of parkinsons

If we had more time, we would deal with the longitudinal portion of the data set. Particularly, we believe that it would be most appropriate to create an LSTM learning model. The preparation necessary to make this model made it out of scope, however it may be possible to cleverly arrange the data to train the model on multiple patients' data, even though they have varying numbers of data points. Aligning the data by study visit number would not be adequate, since it appears patients answered different questions during each study visit.



Resulting subtypes of Parkinsons Disease determined by clustering.

We would also like to finish our work towards making the data more ergonomic for non-analysts. We were successfully able to produce a term-frequency matrix from the questionnaire files. We observed several non-sensical tokens in our matrix, such as repeated letters, such as “mmmmm” and “nnnnn”. These anomalies seemed to come from a single document. Next we could derive the inverse document frequencies for each term. This would give us the necessary backbone for an information retrieval system.

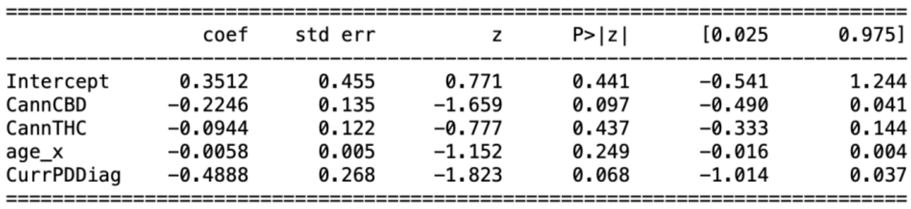
The results of the five statistical analyses indicate that age has a consistent negative relationship with all five dependent variables (cognition bother, motor symptoms bother, pain bother, sleep bother and Bradykinesia bother). This means that as age increases, the likelihood of reporting that these outcomes bother individuals decreases.

CurrPDDiag has a positive relationship with motor symptoms botherness but has no significant relationship with any other dependent variables. This suggests that individuals with a current Parkinson’s disease diagnosis are more likely to report that motor symptoms bother them than those without a diagnosis.

CannCBD has a positive relationship with sleep botherness but no significant relationship with any other dependent variables. This indicates that individuals who use CBD may be more likely to report that sleep bothers them than those who do not use CBD.

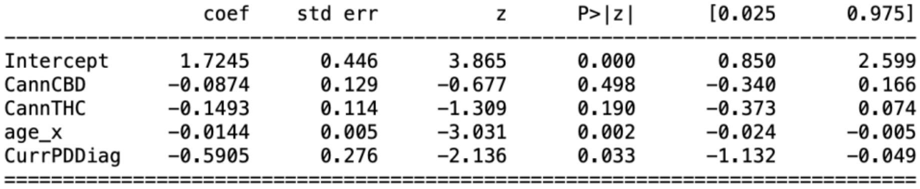
CannTHC does not have a significant relationship with any of the dependent variables. This suggests that the use of THC may not have a significant impact on cognition botherness, motor symptoms bother, pain botherness, sleep botherness or Bradykinesia botherness.

Overall, these results suggest that age and current Parkinson’s disease diagnosis may be important factors in predicting whether these outcomes will be reported as bothersome. The use of CBD may also have an impact on sleep bother.



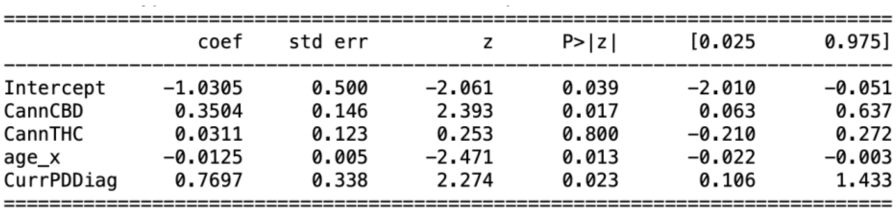
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.3512      0.455      0.771      0.441      -0.541       1.244
CannCBD       -0.2246      0.135     -1.659      0.097      -0.490       0.041
CannTHC       -0.0944      0.122     -0.777      0.437      -0.333       0.144
age_x         -0.0058      0.005     -1.152      0.249      -0.016       0.004
CurrPDDiag    -0.4888      0.268     -1.823      0.068      -1.014       0.037
==============================================================================
```

Results for pain as dependent variable



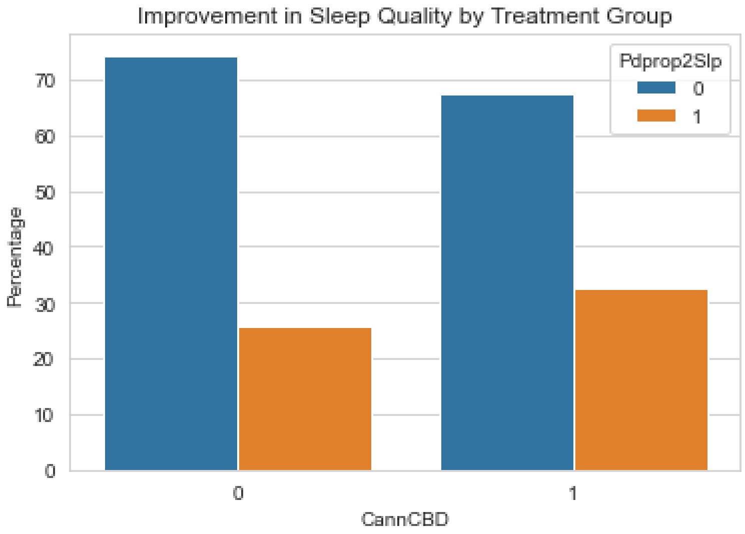
```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.7245      0.446      3.865      0.000       0.850       2.599
CannCBD       -0.0874      0.129     -0.677      0.498      -0.340       0.166
CannTHC       -0.1493      0.114     -1.309      0.190      -0.373       0.074
age_x         -0.0144      0.005     -3.031      0.002      -0.024      -0.005
CurrPDDiag    -0.5905      0.276     -2.136      0.033      -1.132      -0.049
==============================================================================
```

Results for Motor Symptoms as dependent variable



```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.0305      0.500     -2.061      0.039      -2.010      -0.051
CannCBD        0.3504      0.146      2.393      0.017       0.063       0.637
CannTHC        0.0311      0.123      0.253      0.800      -0.210       0.272
age_x         -0.0125      0.005     -2.471      0.013      -0.022      -0.003
CurrPDDiag     0.7697      0.338      2.274      0.023       0.106       1.433
==============================================================================
```

Results for sleep as dependent variable



CONTRIBUTIONS

Michael and Wei both worked on the clustering analysis. Aswin produced the regression model for cannabis use. Michael produced the term frequency matrix of the questionnaire text files.

REFERENCES

1. Fereshtehnejad, Seyed-Mohammad et.al. Clinical Criteria for Subtyping Parkinson’s Disease: Biomarkers and Longitundinal Progression. Brain, Journal of Neurology. Multicenter Study; 2017
2. Maaten, Laurens van der, Geoffrey Hinton. Visualizing Data Using t-SNE. In: Yoshua Bengio, editor. Journal of Machine Learning Research 9 2579-2605; 2008.
3. Parkinson’s disease: https://www.nia.nih.gov/health/parkinsons-disease
4. Parkinson’s disease: https://www.nhs.uk/conditions/parkinsons-disease/causes/