Early Biomarkers for Morbus Parkinson

Analysis of the data from Jan Hlavnicka et al. (2017)

Kilian / 18.07.2021

This is the report for the project "Early Biomarkers for Morbus Parkinson" that I have conducted to introduce myself to the python data-science libraries. The data stems from a research paper on automatic speech processing from Hlavnicka et al. (2017) The data was collected between 2014 ans 2016 in the Czech Republic. The study included 30 recently diagnosed and untreated Morbus Parkinson patients, idiopathic RBD (which is a strong predictor of Morbus Parkinson) and 50 healthy controls. All probands participated in two speeking tasks (reading, monologue). The goal of the study was to automate the analysis of recordings of the patients, which worked out very well. Hlavnicka et al. (2017) published the dataset within their appendix and it was reposted on Kaggle. The research paper and the kaggle dataset can be accessed via the following links:

-https://www.nature.com/articles/s41598-017-00047-5
-https://www.kaggle.com/ruslankl/early-biomarkers-of-parkinsons-disease

Table of Contents

Introduction

From Kaggle: The dataset includes 30 patients with early untreated Parkinson's disease (PD), 50 patients with REM sleep behavior disorder (RBD), which are at high risk developing Parkinson's disease or other synucleinopathies; and 50 healthy controls (HC). All patients were scored clinically by a well-trained professional neurologist with experience in movement disorders. All subjects were examined during a single session with a speech specialist. All subjects performed reading of standardized, phonetically- balanced text of 80 words and monologue about their interests, job, family or current activities for approximately 90 seconds.
Inspiration: Predict a pattern of neurodegeneration in the dataset of speech features obtained from patients with early untreated Parkinson's disease and patients at high risk developing Parkinson's disease.

Hoehn-and-Yahr Scale - A Measure for Morbus Parkinson

The state of the Morbus Parkinson disease can be examined with different clinical assessments:The Hoehn-and-Yahr scale is used to classify the severity of PD based on symptoms. It is an easy-to-perform clinical instrument to assess the underlying movement disorders. The Hoehn and Yahr scale was developed in 1967. It divides the disease into 5 stages. So-called modified stages according to Hoehn and Yahr (1987) are also frequently used (https://flexikon.doccheck.com/de/Hoehn-und-Yahr-Skala). The 5 states are:

1) Unilateral symptomatology
1.5) Unilateral symptoms and axial involvement
2) Bilateral symptoms; no postural instability
2.5) Mild bilateral symptoms; patient can regain balance on pull test (compensates on pull test)
3) Mild to moderate bilateral symptomatology; mild postural instability; independence maintained
4) Severe disability, but patient can still walk and stand without assistance
5) Patient is wheelchair bound or bedridden without assistance from others

UPDRS III total - A More General Measure

The UPDRS III total scale is a more general scale to measure the state of the Morbus Parkinson disearse. The Hoehn-and-Yahr Scale is integrated into the UPDRS III total, as a sub-examination. The UPDRS scale refers to Unified Parkinson Disease Rating Scale, and it is a rating tool used to gauge the course of Parkinson's disease in patients. The UPDRS scale has been modified over the years by several medical organizations, and continues to be one of the bases of treatment and research in PD clinics. The UPDRS scale includes series of ratings for typical Parkinson's symptoms that cover all of the movement hindrances of Parkinson's disease. The UPDRS scale consists of the following five segments:
1) Mentation, Behavior, and Mood, 2) ADL, 3) Motor sections, 4) Modified Hoehn and Yahr Scale, and 5) Schwab and England ADL scale.
Each answer to the scale is evaluated by a medical professional that specializes in Parkinson's disease during patient interviews. Some sections of the UPDRS scale require multiple grades assigned to each extremity with a possible maximum of 199 points. A score of 199 on the UPDRS scale represents the worst (total disability) with a score of zero representing (no disability).
(Copied from: https://www.theracycle.com/resources/links-and-additional-resources/updrs-scale/)

Descriptive Statistics

The dataset consists of 130 observations, including 30 patients with recently diagnosed Morbus Parkinson (PD), 50 Patients with idiopathic RBD and 50 healthy controls. The variable names were tidied up to be available in an accessible, easy-to-work-with format. The original variable names and the new variable names were stored in a .xlsx file called dictionary. The autoprocessed speech task variables were encoded either with the subscript "speaking" for the speaking task, or with the subscript "monologue", for the monologue task. A new variable called "parkinson" was added to the dataframe which indicates whether a subject was diagnosed with an early state of parkinsonism or not (Yes = 1, No = 0). The patients with idiopathic RBD were labeled as diseased as well, since well over 80% of patients with the diagnosis will end up with PD (some research to back up this claim: https://pubmed.ncbi.nlm.nih.gov/29170501/ , https://pubmed.ncbi.nlm.nih.gov/19768814/ , https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4649803/ , https://pubmed.ncbi.nlm.nih.gov/22136899/). Further, variables containing Yes/No strings were recoded according to the following scheme: Yes = 1, No = 0. Some variables also included mediactions, which were deleted since they did not add additional value to the analysis (This might be a limitation of this analyis). Last, variables only containing "No" answers were dropped from the dataset.

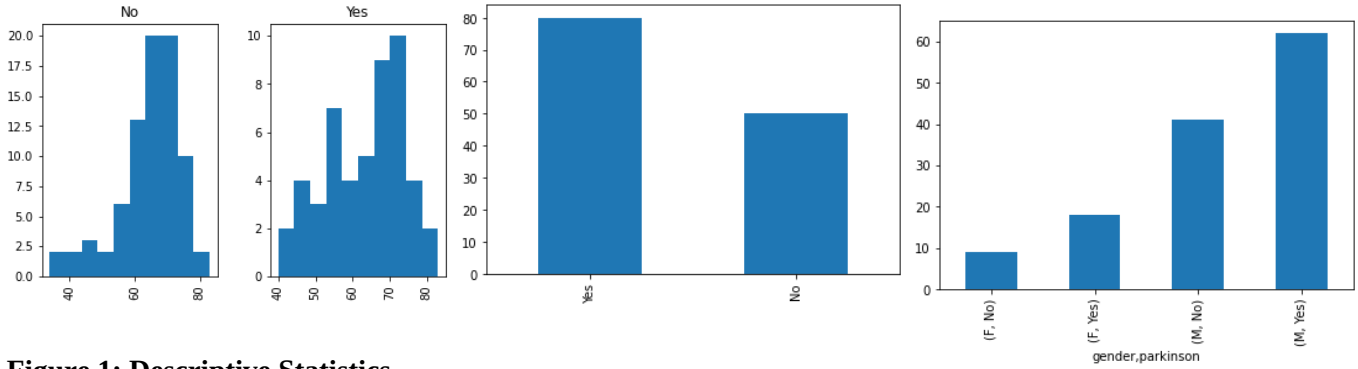

Figure 1: Descriptive Statistics.

Left: Age by disease,
Middle: Balance of the criterion "Parkinson"
Right: Cases separated by Gender and whether they are diseased

Mean Analysis

To get a good understanding of the dataset, a mean analysis was conducted. Many variables were not collected for the control group (e.g., Parkinson's stage), so these were excluded from the mean comparisons. Mean and standard deviation were calculated separately for the control group and the diseased group. Further, the difference of the mean values of the two groups was calculated. Significance tests (especially the T-test) were not used because they do not give a good picture of reality without post-hoc contrast analyses (e.g. Bonferroni correction). Since more than 30 comparisons are made, a Bonferroni correction, for example, would result in hardly any result being significant (e.g. for a given significance level of 0.05 the new alpha for a single comparison would be 0.05/32=0.001). To still get a good understanding of the differences in means between the two groups, effect sizes were calculated. Effect sizes express the difference in means relative to the common standard deviation. However, this imposes the strong limitation that some effect sizes cannot be interpreted properly because the standard deviations are very different. Therefore, caution is advised. Some variables (e.g. clonazepam) follow a binomial distribution. The effect size for these variables is therefore invalid. Their mean values can be interpreted as proportions.

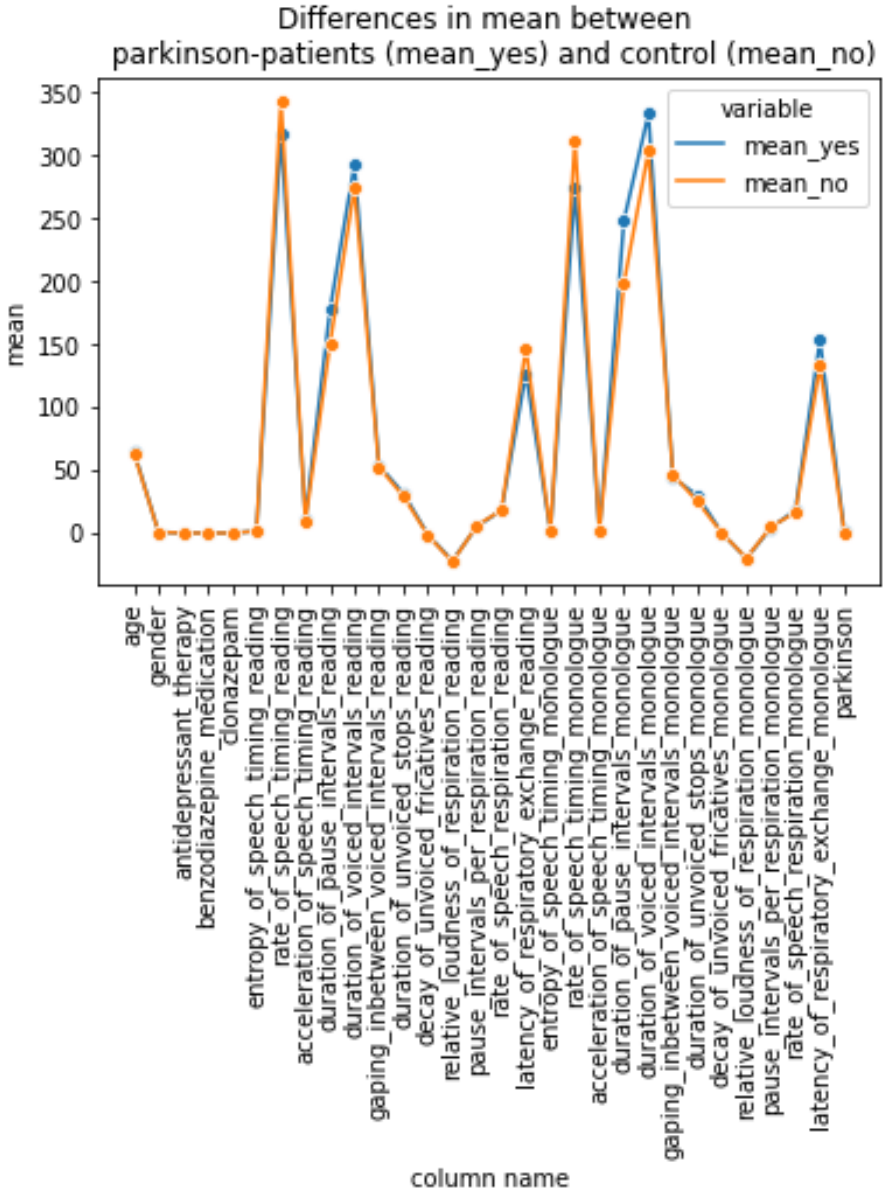

Figure 2: Mean Comparisons.

Comparison of the means between patients with early-state parkinson (blue) and the control group (orange). The lineplot further visualizes the relationship between the reading task and the monologue task.

Classification

The basic idea behind the classification task is that one can use the recordings of the two speech tasks evaluated automatically by Hlavnicka et al. to be able to make a diagnosis of PD at an early stage. Such a diagnostic procedure would be interesting in many respects. The execution of the task, as well as the analysis of the data could be carried out inexpensively and without a large expenditure of time. It would also be conceivable, for example, that the patient would no longer have to come to the clinic for the diagnosis, but could read out the text at home and record it for example using a smartphone. This would also make it possible to carry out nationwide screenings, although this would make less sense in the case of PD. In addition, the objectivity of the diagnosis will be very high, since the automatic evaluation of the audio data means that human interpretation is no longer necessary.
The criterion contains 50 No-labels (38.5%) and 80 Yes-labels (61.5%). The data was not further balanced, since the dataset is rather small and wrong labeling (through upsampling) could have a rather big impact on the final model. Only the automatically analized variables from the two speech-tasks were taken as predictors for the criterion. The dataset was split into training (80%) and testing (20%) datasets to detect overfitting. The assignment to training vs. testing was randomized and stratified. Bot training and testing data were then scaled, since all variables were assumed to be equally important for the prediction of the criterion.

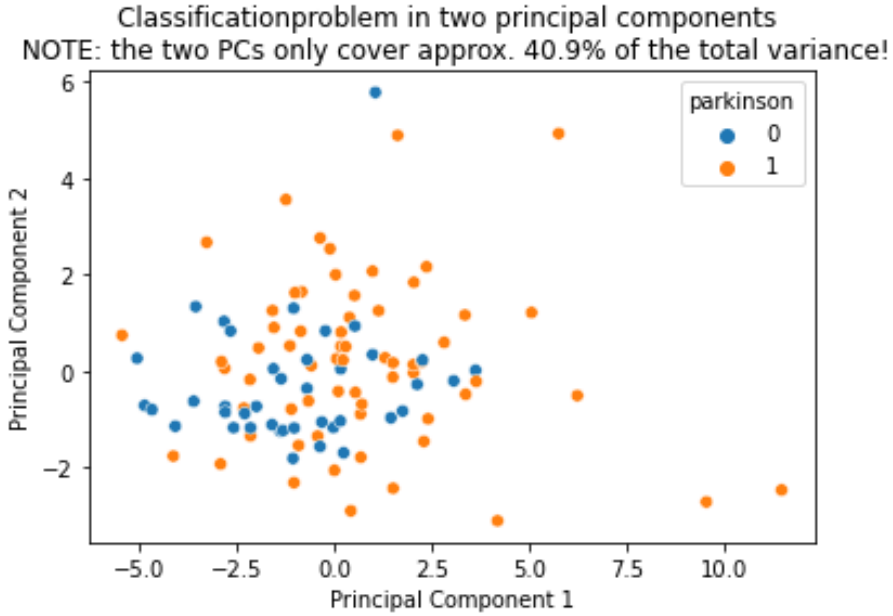

Figure 3: Plot of the classification-problem at hand.

Principal Component Analysis (PCA) was performed on the predictors. The first two PCs were then plotted and colored according to the criterion. The first two PCs only covered 40.9% of the total variance, which is a strong limitation of this plot. Caution with interpretation is advised!

For the prediction-task, three models were used for comparison: Logistic Regression, Quadratic Discriminant Analysis (QDA) and RandomForrest. The Logistic Regression model (with regularization) was used as a baseline model for the other models to beat (https://pubmed.ncbi.nlm.nih.gov/19523085/)QDA was used since it can model non-linear boundaries, but is still a vectorbased method (which is recommended for small data: e.g. https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-019-0160-3) The method is generally suggested with large datasets, which is why it is being used with regularization for this analyis. RandomForest was used, since many sources recommend it particularly for smaller datasets (https://www.sciencedirect.com/science/article/abs/pii/S1386505620306511?via%3Dihub). The combined "voting" of multiple trees together often leads to better accuracy than single algorithms.

The model performance was measured using 10-fold, 5-times repeated, cross validation. The parameters were tuned using grid search. For the Logistic Regression and the QDA, only the regularization-parameter was tuned. For the Random Forest, the max. depth and the amount of trees were tuned. Since Random Forests are not that sensitive to overfitting, more trees likely mean better prediction. As will be described in more detail later, the analysis struggled with overfitting. Hyperparameter tuning is therefore very dangerous, because the accuracy may well be inflated by artifacts in the data. For this reason, as few parameters as possible were tuned and the tuned parameters are all considered to be insensitive to overfitting. (A more in-depth discussion of hyperparameter tuning with small datasets can be found here: https://booksc.org/book/46424573/ed5526).As expected, the in-sample logistic regression just missed the 70% mark. Out-of-sample it could not distinguish itself from the a priori probability (60%). The quadratic regression made it to about 72% accuracy in-sample and slightly less out-of-sample. Surprisingly, the overfitting remained within limits. The same is true for the random forest classifier. It also achieved about 72% in-sample and similar values out-of-sample.

During the analysis, it was noticed that the accurracy and the hyperparameters were very different for train-test-splits with different random states. Furthermore, the models can be poorly generalized and show out-of-sample very poor accuracy values. Often the out-of-sample predictions cannot even beat the a priori probability (60%). These phenomena are a very strong indicator for overfitting!

The question now is whether the predictors contain enough information to accurately predict the disease. To find this out, the labels were attempted to be recovered from the predictors alone. The KMeans algorithm (unsupervised classification algorithm) was used for this purpose (K = 2). The confusion matrix (true labels vs. recovered labels) indicates that the algorithm fails to recover the labels. The main diagonal holds only about 50% of all observations (accuracy 52%). This is identical to randomness.

```
[40, 52]
[10, 28]
```

Conclusion

Overall, unfortunately, one cannot claim that a finished diagnostic instrument can be produced with this data set. It is not (yet) comprehensive enough. This is indicated in particular by the inability of the predictors to recover the criterion. The models trained in this analysis can only inadequately predict the criterion and are inconsistent with different random states in the train-test-split. To become a good diagnostic tool, a model would probably need to achieve an accuracy of 99% (at 50% a priori) or more. However, this requires more data. Since the analysis method of Hlavnicka et al. is very efficient and could even be done from home, data acquisition should be very possible.

Appendix

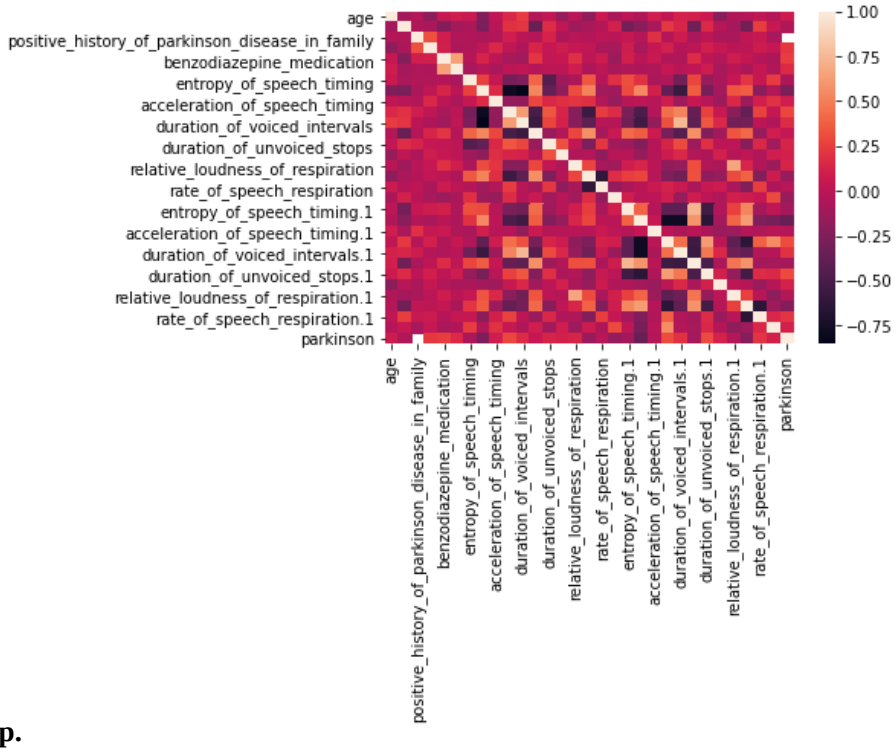

Figure 4: Heatmap.

Heatmap that visually represents the intervariable correlations. The darker the color the more negative the correlation; The brighter the color the more positive the correlation