**Parkinsons Disease Data Set Case Study**

1. Brief description of the data set and a summary of its attributes

The dataset was created by Max Little (University of Oxford) in collaboration with the National Centre for Voice and Speech (Denver, Colorado), who recorded speech signals.

It contains biomedical voice measurements from 31 people, 23 with Parkinson’s disease. The attributes are particular voice measures, and the examples correspond with the 195 voice recordings.

**Attributes:**

**name** - ASCII subject name and recording number

**MDVP:Fo(Hz)** - Average vocal fundamental frequency

**MDVP:Fhi(Hz)** - Maximum vocal fundamental frequency

**MDVP:Flo(Hz)** - Minimum vocal fundamental frequency

**MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP** - Several

measures of variation in fundamental frequency

**MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA** - Several measures of variation in amplitude

**NHR,HNR** - Two measures of ratio of noise to tonal components in the voice

**status** - Health status of the subject (one) - Parkinson's, (zero) - healthy

**RPDE,D2** - Two nonlinear dynamical complexity measures

**DFA** - Signal fractal scaling exponent

**spread1,spread2,PPE** - Three nonlinear measures of fundamental frequency variation

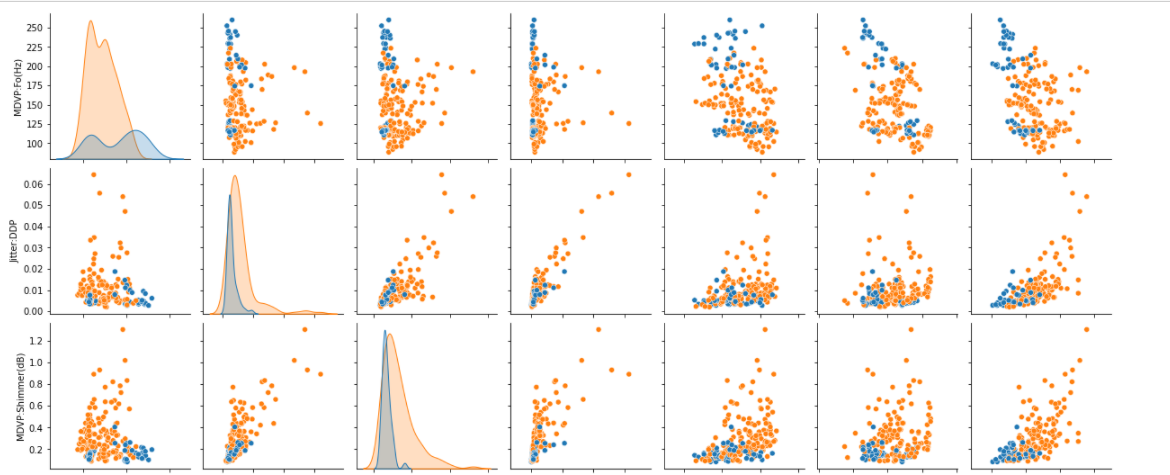
(Attribute information from Parkinson.names file in the UCI Repository).

All attributes but name (pacient name) and status (pacient has the disease (1) or not (0)) are real values. The pacient name is not important as it shouldn’t have any descriptive insight to whether a person has Parkinson or not, so we will later remove that attribute. This makes all our predictors real values.

1. Initial plan for data exploration

The initial plan is to explore attribute distributions, correlations and detect outliers within the distributions.

First, I have decided to plot (sns pairplot) one of each type of attribute so as to have a clearer visualization than if we had called sns pairplot on the entire dataset. By type of attribute im referring to choosing only one predictor that models a certain characteristic. For example, I only plot NHR instead of plotting NHR and HNR in order to reduce the number of graphs created.



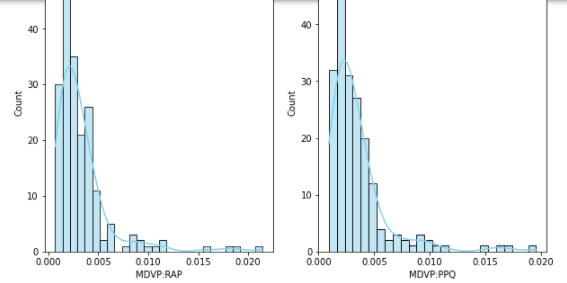
By checking at some pairs of attributes, we can notice some degree of separation between both classes. For example, we can see that when comparing NHR and spread1, patients who have parkinsons disease have higher values of spread 1 than patients who don’t have the disease. The outlier points in the majority of the plots tend to belong to the parkinsons class too.

The highest correlations between attributes and status are 0.56 (spread1) and 0.53 (PPE).

1. Feature engineering and data cleaning

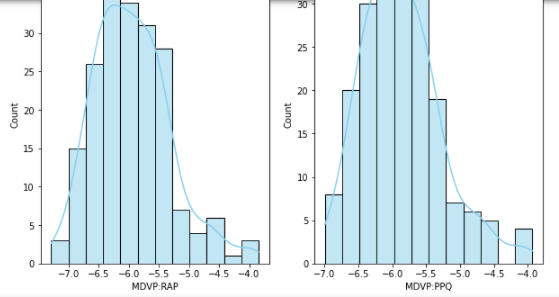
Regarding Data cleaning, not much is required for the selected dataset. We don’t have highly correlated columns and there are no null values. Thus we don’t need to think about which would be the best strategy (imputation, drop column, drop row, etc.). The only cleaning made was to drop the “pacient name” attribute.

As for feature engineering, I have noticed the dataset has some attributes whose distribution is heavily skewed.

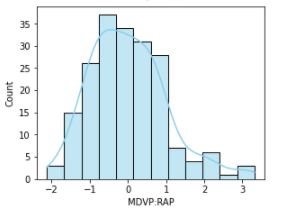
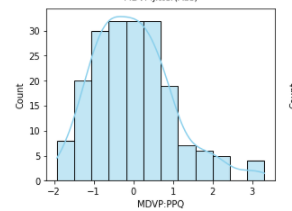


Due to this fact, I have decided to use a log transformation for attributes with a skew value greater than 1. This makes the distribution of the transformed predictors to resemble a gaussian distribution, which is a prerequisite for many ML algorithms.

Attributes’ distribution after transformation:



The next step is to normalize the data. I chose to use z transformation, but Min-Max normalization is another option

1. Key findings and insights

By comparing the distributions of parkinson patients’ attributes with non- parkinson patients’ attributes, we can identify some insights which were previously mentioned. The range of values in almost all of the attributes is undeniably greater than the one In attributes of non- parkinson patients. This could be a key insight when trying to predict a parkinson patient. For example, if spread1 is greater than -4, we could infer that the patient has parkinson. Also, the higher spread1 is, the more likely a given patient has parkinson. Also, outlier points tend to correspond with patients who have the disease.

1. Formulate hypothesis and
2. Conduct a formal significance tests, discussing results

1-Null hypothesis: The mean of spread1 for parkinsons patients is the same than the mean of spread1 for non-parkinsons patients.

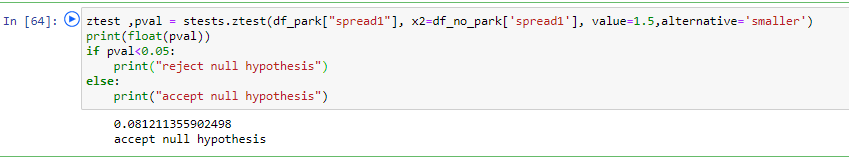
Alternative hypothesis: The mean is different



Thus, we reject the null hypothesis by a great margin. The means are different.

2-Null hypothesis: The mean of spread1 for parkinsons patients is larger than the mean of spread1 for non-parkinsons patients by at least 1.5 standard deviations of the attribute’s distribution.

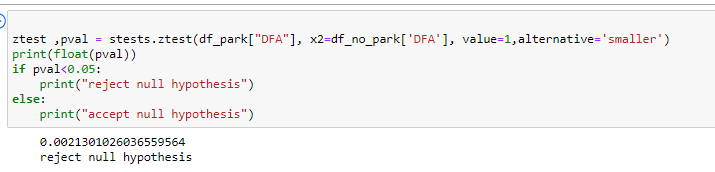
Alternative-hypothesis: The mean of spread1 for parkinsons patients is not larger than the mean of spread1 for non-parkinsons patients by at least 1.5 standard deviations of the attribute’s distribution.



The null hypothesis is accepted, so we can confirm with 95%confidence that the null hypothesis is correct.

3-Null hypothesis: The mean of DFA for parkinsons patients is larger than the mean of DFA for non-parkinsons patients by at least 1 standard deviation of the attributes distribution.

Alternative-hypothesis: The mean of DFA for parkinsons patients is not larger than the mean of DFA for non-parkinsons patients by at least 1 standard deviation of the attributes distribution.



The null hypothsis is rejected, so we can’t confirm with a 95% confidence that the mean of DFA for parkinsons’ patients is larger than the mean of DFA for non-parkinsons’ patients by at least 1 standard deviation of the attributes distribution.

1. Suggestions for next steps in analyzing this data.

I would suggest to analyze and compare the distributions of attributes within the same attribute types (attributes that measure the same thing). It is important to separate the distributions by status, so we can check the distribution for parkinsons patients and non-parkinsons patients.

Also, we could implement Principal component analysis to check which attributes have the greatest variance.

1. Summary of quality of data

I believe the quality of the dataset is decent. There are no missing values and you can visually have an idea of which attributes are going to have a greater effect on predictions (such as spread1).

However, the dataset only has 195 rows, so we could be building this whole EDA on a biased small group of individuals. If more data could be generated with patients from different parts of the world, different age ranges and social status, I would be more confident that the model generated with this data would be a good one.

**References:**

<https://archive.ics.uci.edu/ml/datasets/parkinsons>

Required Citations:  'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)