For my Capstone project I have selected three datasets from the UCI Machine Learning Repository. All three datasets contain frequency based features extracted from voice recordings of Parkinson’s disease patients and for two of the datasets also a set of control samples. Research clearly indicates that there are Parkinson’s disease-related voice changes such as that the voice becomes overly quiet and soft-edged, can lose crispness of articulation, and can be reduced in pace.

The first dataset contains two sets of data: a) a training set consisting of 26 frequency based features from several voice recordings from 20 Parkinson’s disease cases and 20 healthy controls and b) a test set consisting of the same 26 frequency based features from additional 28 Parkinson’s patients. The main aim of the analysis will be to identify features from the voice recordings that best differentiates Parkinson’s disease patients from healthy individuals. The training data set also includes the UPDRS score as measure of Parkinson’s disease progression. Therefore, I will also aim to identify features that predict the UPDRS score. However, as the dataset only consist of data from a single time point the prediction model would be very limited. Therefore, this part of the dataset may be better suited to serve as independent test dataset for the third dataset which is available.

The second dataset consists of 23 features obtained from 197 voice recording from 23 Parkinson’s disease patients and 8 controls. For this dataset, the main aim is to identify features that best differentiate between Parkinson’s patients and healthy controls. If the available features overlap with the features used in the model for the first data set, this dataset will be used foremost as independent test data set for the first identified classification model. However, in addition I also will build a classification model specifically for this dataset, as this dataset is closely related to the third available dataset and therefore offers the possibility to compare the resulting classification and predictive models.

The third dataset is composed of 16 biomedical voice measurements from 42 patients with early-stage Parkinson’s disease that were measured over a period of six-month giving a total of 5,875 voice recordings. The main of the data is to obtain a model that best predicts the overall and motor-specific UPDRS score based on the 16 voice features. If there should be overlap between the selected features and features available in the first dataset, I will use the training data from the first dataset as independent test set for the identified model.

In addition, I also plan to compare the resulting models to determine if they are based on similar features or if they differ strongly.

There is no direct client available for this project. However, if this should produce interesting results, I will communicate them to the scientists that originally produced the datasets. The last dataset is particularly interesting because the data was collected with an automatic telemonitoring device provided by Intel. Therefore, based on the results Intel may become interested to introduce their device as general option to monitor progression of Parkinson’s disease in the patients’ home.

In terms of approach, I will first visualize the available raw data first on the level of individual voice recordings that based on mean and median data for each individual to better understand the available data, to identify potential outliers and identify missing data. For each dataset separately I will try to identify the best model for classification or prediction using different clustering methods as well as machine learning algorithms such as random forests. Next I will use another of the available datasets as test set for the resulting model if that is an option. In addition, I will compare the features that are part of each model.

As deliverables, I will produce an iPython notebook with the applied code and the results of the analysis and I will create a presentation of the final results.