Milestone 1

The dataset for the application project is Parkinsons Telemonitoring Data Set. The dataset was created by Athanasios Tsanas ([tsanasthanasis@gmail.com](mailto:tsanasthanasis@gmail.com)) and Max Little ([littlem@physics.ox.ac.uk](mailto:littlem@physics.ox.ac.uk)) of the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals1.

There are 5875 instances in the dataset without any missing data. The actual number of features is 20 and there are two different label associated with each instance. The main aim of the data is to predict the motor and total UPDRS scores (Unified Parkinson Disease Rating Scale) from the 16 voice measures (the other four additional features are – subject ID, subject age, subject sex and subject testing time, which are not very relevant).

After preprocessing, the raw data is analyzed by Principal Component Analysis SVD method and among the sixteen voice measures, one particular feature explained 99% of the data variance due to the fact that features may have not been normalized.

This dataset measure the extent to which Parkinson affects the voice and hence is very useful. One of the major application is in the Parkinson diagnosis field as the voice measurements can be used to predict UPDRS scores.