Sonde Data Science Challenge

This exercise is based on Physionet's VOICED dataset, which contains recordings from 208 individuals. Each recording consists of a vocalization of the vowel 'a' five seconds in length without any interruption of sound. In addition, various demographic, clinical, and lifestyle features were collected for each individual. Your goal in this exercise is to explore if these individual-level features can be predicted from the voice recording.

We used a standard tool called openSMILE to extract features from the recordings, using the INTERSPEECH 2010 Paralinguistic Challenge feature set. Please refer to section 2.5.6 of the openSMILE documentation (download from here) for additional information about these features.

The result is a csv file with one row consisting of 1582 features for each recording. A second csv contains the individual-level features. Both csvs are included with this exercise.

This exercise will require data exploration, preparation, and modeling. Please **do not spend more than 4 hours working on it**. We **do not expect any background** in speech analysis. Also, you do not need to explore all the individual-level features, but please document which ones you do.

Our goal is to evaluate your thought process, rigor, practical data science skills, and how you communicate your work. The performance of your models is less important. We encourage you to prioritize a rigorous process and an end-to-end solution over sophisticated modeling.

Please submit your work in a working Jupyter notebook. If your code depends on any non-standard packages, please provide information about their installation.

In addition, please submit a short report (no more than 2 pages) explaining your thought process, approach, and results. This is also a good place to outline additional ideas that you didn't have time to implement.