**README .md**

**Overview**

Kaggle seizure prediction competition.

**Data**

**Data Source**: https://www.kaggle.com/c/melbourne-university-seizure-prediction/data

**Data Size**: 30GB Training data.(approximate)

30GB Test data. (approximate)

**Total number of files:**

Training- 6041 .mat files

Testing – 6261 .mat files

Due to the very large number of files and size of the data, I decided to compress the files and store them in npz format. This helps through severe reduction of time for running the model. By storing data in this format after feature extraction many processing steps can be avoided (need to be performed only one time) such as PCA and tsne that take a very large amount of time.

3 npz files are loaded into the **Main.ipynb**. The 3 files represent data from 3 people.

**Dependencies**

Should work with all the packages in anaconda.

**Files**

1) **preprocess.ipynb**- Visualize raw signals, t-sne, eigenvalue of correlation matrix in frequency domain

2) **preprocess2.ipynb**- Blind source separation using FastICA.

3) **Main.ipynb**. Obtain features from Wavelet Transformation and run predictive modelss.

**First Attempt**

This is my first attempt at making predictions and first end to end model.

Data----->Remove/Predict missing values--->Wavelet--> PCA--> Oversampling-->Supervised Model

**Additions**

* ICA
* t-SNE
* Spectogram visualization
* Eigenvalue of correlation matrix in frequency domain