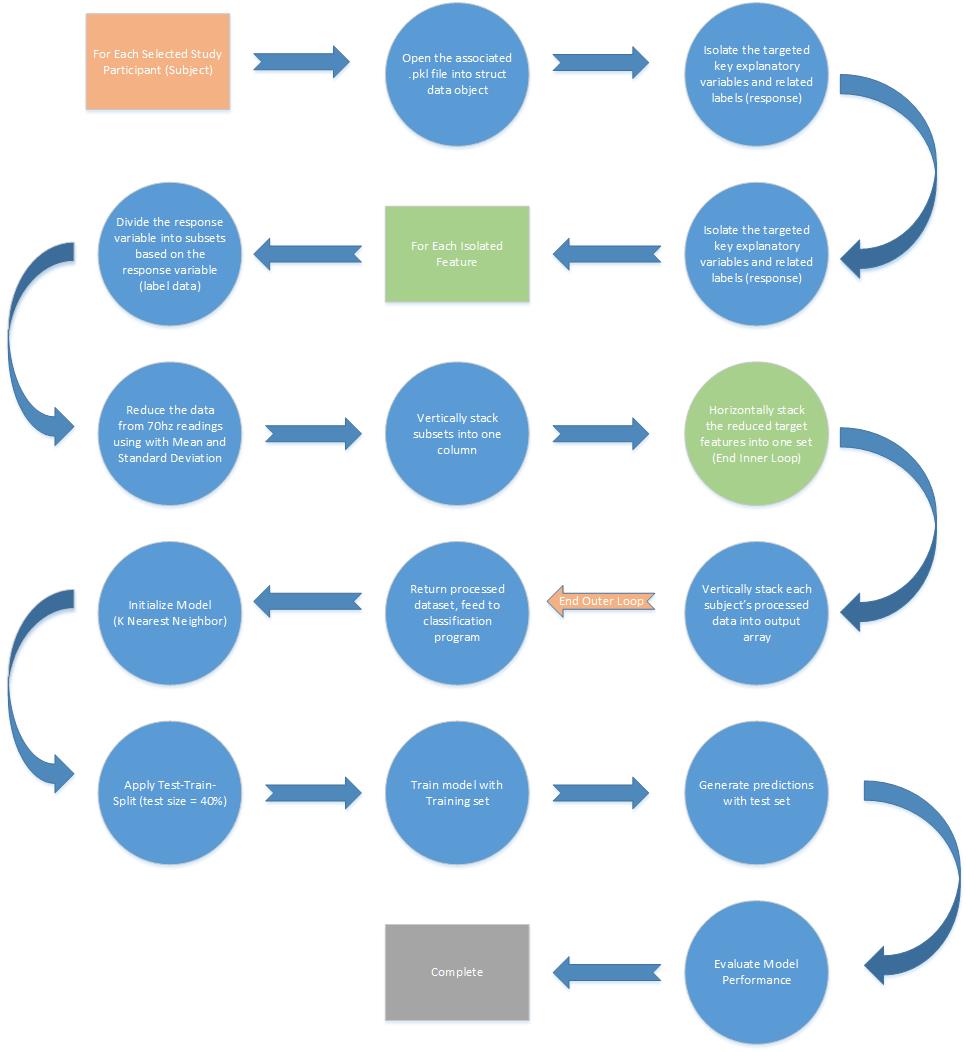
Data Ingestion – Processing – Classification – Model Evaluation



Main Techniques Applied:

The main techniques that we used in our classification process are:

* Reading in pickle .pkl files
* Storing the main file structure in our own python class
* Extracting individual raw features from the data stored in our class into numpy array
* Processing each feature by dividing into response specific subsets, with numpy methods
* Reducing each feature subset by stepping through and generating average and standard deviation, in 700 record increments with numpy methods
* Consolidating the three reduced subsets for each feature, into one column
* Consolidating all of the columns (targeted features and response data) into one table
* Performing test – train – split on the processed / reduced table, with a test size of 40%, with a scikit learn method
* Training a Scikit Learn K Nearest Neighbors model, with K Neighbors = 2, with the training feature set (explanatory variables)
* Generating predictions with the SKlearn KNN model, using the test set
* Evaluating the performance of the model with an accuracy measurement and displaying a confusion matrix

Key Results

Through this process we were able to develop the following conclusions:

* During our initial EDA we identified that analyzing the data in its raw state, 70hz readings, the data was too variable to generate strong predictions.
* We tested various reduction strategies to eliminate some of the noise or variability in the sensor data. We found that we were able to see a much stronger relationship as we approached a 700:1 ratio for reduction.
* Not all of the data available to us was necessary. We found the strongest relationship between ecg, temp, etc. sensor data and the response variable, emotional state.
* The Chest sensor readings, though supplying some of the same measurements, was more reliable.
* We tested several potential model approaches; linear regression, logistic regression, linear discriminant analysis, decision tree, and K Nearest Neighbors. We found the best results with K Nearest Neighbor (Tied with decision tree for performance)
* We were able to generate a model with an accuracy of ~100%, consistently, with just 4 explanatory features

Applications:

This type of model has applications in several domain, and could be used almost anywhere;

* The EDA portion of this study could be applied to hardware design and sensor specifications. We were working with data collected at 70hz. The definition was very high, and could likely be decreased considerably. This would impact the energy and computational resource requirements for any biofeedback hardware (some wearable device).
* This type of classification could be applied to wearable devices that would alert a user whether or not they are stressed, or inform them once they have reached a state of relaxation.
* This could be applied to design settings to test the response of participants. Such as, designing workspaces that maintain stress to a minimum, or entertainment spaces to optimize for amusement.