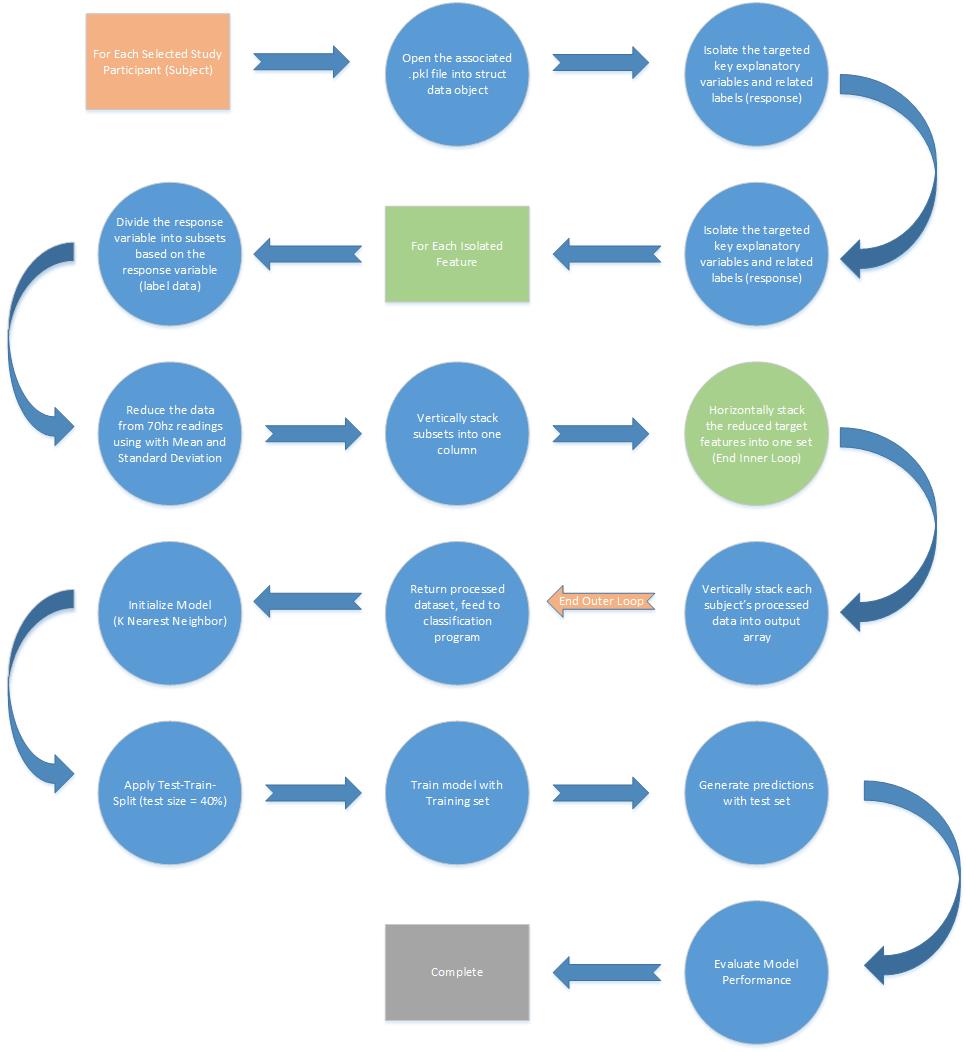
W

Data Ingestion – Processing – Classification – Model Evaluation



Main Techniques Applied:

We applied several course relevant techniques during each phase of our data mining project, to briefly summarize the key techniques applied;

During the data acquisition phase of our project we developed a program that first reads the data in from the source pickle (.pkl) files, and stores the data contained within the file in a user defined python class. For this, and each subsequent step until the classification or modeling step, will be performed in a loop - with each iteration working with the data from an individual subject from the study (there were 12 active participants in the study)

During the processing phase, we extract an individual raw feature and the label data (response) and process it by first subdividing the array based on the label data classes (1, 2, 3) using numpy methods. The result is 3 subarrays that we summarize by stepping through the array grabbing an n sized chunk of records at each step, calculating the standard deviation and mean, and storing this value in a new array. Once the subarray has been reduced, we vertically stack the subarrays to produce again a single feature array. We found that 700:1 or 10 second averages produced a high quality feature. Next, we performed this step for all of the features in the dataset, for both chest and wrist sensors, and evaluated the relationship or correlation to the response variable. From here we implemented a forward stepwise development process, we were able to identify 4 explanatory variables that consistently produced models scoring ~99.9% accuracy. We employed another loop here to assemble the individually processed feature arrays into one main table for analysis. It is also at this point that we vertically stacked the processed data from all queried individuals before passing it along to the mining or classification step.

During the classification step we apply a concept of breaking the dataset into testing and training subsets, through the random selection. This helps produce better performing models when it relates to ‘unseen data’ or testing the models performance with what could potentially be disparate characteristics in the test and training sets, as would be seen in a real life application of the model. For our exercise we use a 60/40 split, with 40% of the records allocated to testing. Once the data has been split we initialize a Scikit Learn model to classify our data, and then generate a set of predictions with our test set. Finally, we evaluate the results of the test predictions using a confusion matrix and simple pass/fail count rates (accuracy).

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 team spent a fair amount of time discussing how a basic sensor array and a model similar to this could be applied. A few of the ‘applications’ were more geared towards the product design perspective, “If we were going to build a product like this, how could we use our findings to build the \*best\* product”. An example of this would be; 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).

A few examples of where we saw that this figurative ‘product’ could be utilized are more intuitive, such as; 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. From another perspective, collecting emotional state feedback about some environment or situation. A product similar to what we have put together 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.