**Final Report**

Upon examining the dataset we were required to classify activities for, it was immediately apparent that there are 2 separate datasets included ‘Protocol’ and ‘Optional’ – our research focused on the more robust ‘Protocol’ dataset.

**Exploratory Data Analysis and Preprocessing:**

After further inspection we noticed that some data was either missing or flawed. To combat this issue we approached each scenario on a case-by-case basis:

* Transient activities (denoted as 0 in the activity ID) do not describe any specific activity, but rather unspecified or unknown activity or collection of activities. As such, we removed these samples from our dataset entirely.
* The ‘orientation’ columns were specified to be invalid in this data collection, as such, these were removed and have not been taken into consideration.
* The ‘3D-acceleration data (ms-2), scale: ±6g, resolution: 13-bit’ was specified to be less reliable than its counterpart with a scale of ±16g. Therefore, we decided to remove these measurements from our research.
* The sampling frequency of the heart rate monitor was specified to be ~9Hz, therefore, roughly every 0.1 seconds a value was measured, while all the other sensors yielded results every 0.01 seconds. We modified all missing samples that were tagged as NaN to the value of their previous real measurement.
* IMU sensory data had occasional missing values, in order to avoid losing potentially valuable data by dropping said rows, we decided to impute the missing data (i.e. infer them from the known part of the data).

Now that we have our dataset modified we are ready to perform some further analysis:

After inspecting the data given in ‘PerformedActivitiesSummary.pdf’ and after displaying the amount of times each activity was performed we can see that ‘rope jumping’ is significantly under-sampled as compared to most of the other activities. From this inspection alone we can expect our predictions on this activity and others that are not as well represented, to be less accurate.

We inspected the data more closely by plotting graphs from the available data and gathered the following insights:

* Activities that apply low physical strain on the body (such as lying, sitting, standing) have similar measurements in data such as heart rate or 3D-acceleration data. These types of activities are therefore expected to be more difficult to learn.
* Activities that contain different inclinations of the body parts measured have more apparent differences as part of the magnetometer measurements, this will allow the model to learn the difference between otherwise difficult activities to differentiate between, such as lying and sitting.

**Validation Strategy:**

We’ve decided to validate our models using the Group K-fold validation strategy.

Since our test set consists of new and unseen subjects the model won’t be privy to during the training process, we want to assess its ability to generalize the input given.

**Naïve Solution:**

Our naïve solution yielded surprisingly relatively good results. Thanks to our exploratory data analysis we managed to extract relevant features from the raw data and process and iterate over them well.

**Classical Machine Learning Algorithm:**

The Decision Tree Classifier Algorithm we used yielded similar results to our naïve solution. We can deduce from this result that the features we decided to eventually use and impute were satisfactory but choosing a different algorithm may have yielded better results.

**1st Model:**

Our initial model implementation and training resulted in over fitting with moderately high validation accuracy.

In order to improve upon this result we suggested various ways to tackle the problem and ended up implementing the 2nd version using, among others, a larger window size and stride which will allow us to take a more representative portion of the measurements into account when making a prediction.

**2nd Model:**

Due to the lower learning in this version we added more epochs to allow our model more time to converge. The result was that our 2nd model yielded generally better predictions across the folds and in average.

**Final Model:**

Our final model’s final performance was more similar to that of our 1st model, despite using the specs of the 2nd model. The reason for this could potentially be the fact that the tested subjects were different, and that the model did not generalize the features as well as we had been led to believe by the results of our 2nd model’s validation performance.

**Self-Supervised Pretraining:**

We decided to pretrain our model on the classification of subjects. This allows us to initialize our model’s weights to a good starting position from which we can start fine tuning by training the model for our initial objective.

The result from the pretraining wasn’t as decisive as we had anticipated. This could potentially have occurred due to a sub-optimal learning rate which made our model converge on a sub-optimal minimum.