# Constructing a classifier to predict activities based on Actigraph and RR data

This report shows the procedure taken to construct the classifier, including feature extraction. I hope this report can clearly explain all the decisions made throughout the process and I look for any feedback on it to improve.

Among the data files, *Actigraph.csv* and *RR.csv* are the two files that contain continuous data of each user’s movement and heartbeat to beat intervals throughout a roughly 24 hours timespan. In detail, *Actigraph.csv* contains the following features:

* Axis1: Raw Acceleration data of the X-axis expressed in Newton-meter.
* Axis2: Raw Acceleration data of the Y-axis expressed in Newton-meter.
* Axis3: Raw Acceleration data of the Z-axis expressed in Newton-meter.
* Steps: Number of steps per second.
* HR: Heartbeats per minutes (bpm).
* Inclinometer Off: values equal to 1 refer to no activation of the inclinometer. The values are reported per second.
* Inclinometer Standing: values equal to 1 refer to the standing position of the user, while 0 refers to other user positions. Values are reported per second.
* Inclinometer Sitting: values equal to 1 refer to the sitting position of the user, while 0 refers to other user positions. Values are reported per second.
* Inclinometer Lying: values equal to 1 refer to the lying position of the user, while 0 refers to other user positions. Values are reported per second.
* Vector Magnitude: vector movement derived from raw acceleration data expressed in Newton-meter.
* day: 1 and 2 refer to the first and second day of data recording, respectively.
* time: day time when the heartbeat happened (hours:minutes:seconds)

Using these features and beat to beat intervals, an attempt to construct a multi-class classifier to predict the output label from *Activity.csv* was made.

*Activity.csv* shows which activity out of the 12 different categories (from 1 to 12) the user was doing with start and end time for each activity. As mentioned earlier, these activity labels were used as output labels in order to perform a supervised learning.

\* However, it is important to note that *Activity.csv* is a log that was manually filled by each user, so it might not be as accurate as the data from *Actigraph.csv* and *RR.csv*.

## Feature extraction

1. A new column was added to *Actigraph.csv* to show activity labels from *Activity.csv*. For every data sample, the corresponding activity label was added. However, as *Activity.csv* has lots of gaps timewise between each activity there were lots of data points with missing activity label. They were assigned an activity label of “-1” for the time being.
2. Using the python library *hrvanalysis*, outliers in the RR intervals from *RR.csv* were replaced with linear interpolation. Then ectopic beats were also replaced with linear interpolation, resulting in NN intervals.
3. The transformed beat to beat interval data were then converted into short term Heart Rate Variability (HRV).

* Method: Root Mean Square of Successive Differences (RMSSD)
* Time duration: 5 minutes (a commonly used time interval for short term HRV)

One short term HRV sample was obtained every 5 minutes.

\* Ex. 03:15:00-03:20:00 short term HRV: 65.24665223703592 from 199 number of data

\* In order to make a fair comparison between users, the calculated short term HRVs were normalised with respect to the user’s average heart rate (calculated from the HR feature in *Actigraph.csv*).

1. To match with the short term HRV data which is just one sample per 5 minutes, the samples from *Actigraph.csv* were combined. For each attribute, mean values over every 5 minutes interval were calculated.

\* Similar to short term HRV, Axis1, Axis2, Axis3, and Vector magnitude were normalised with respect to their mean values throughout the entire dataset.

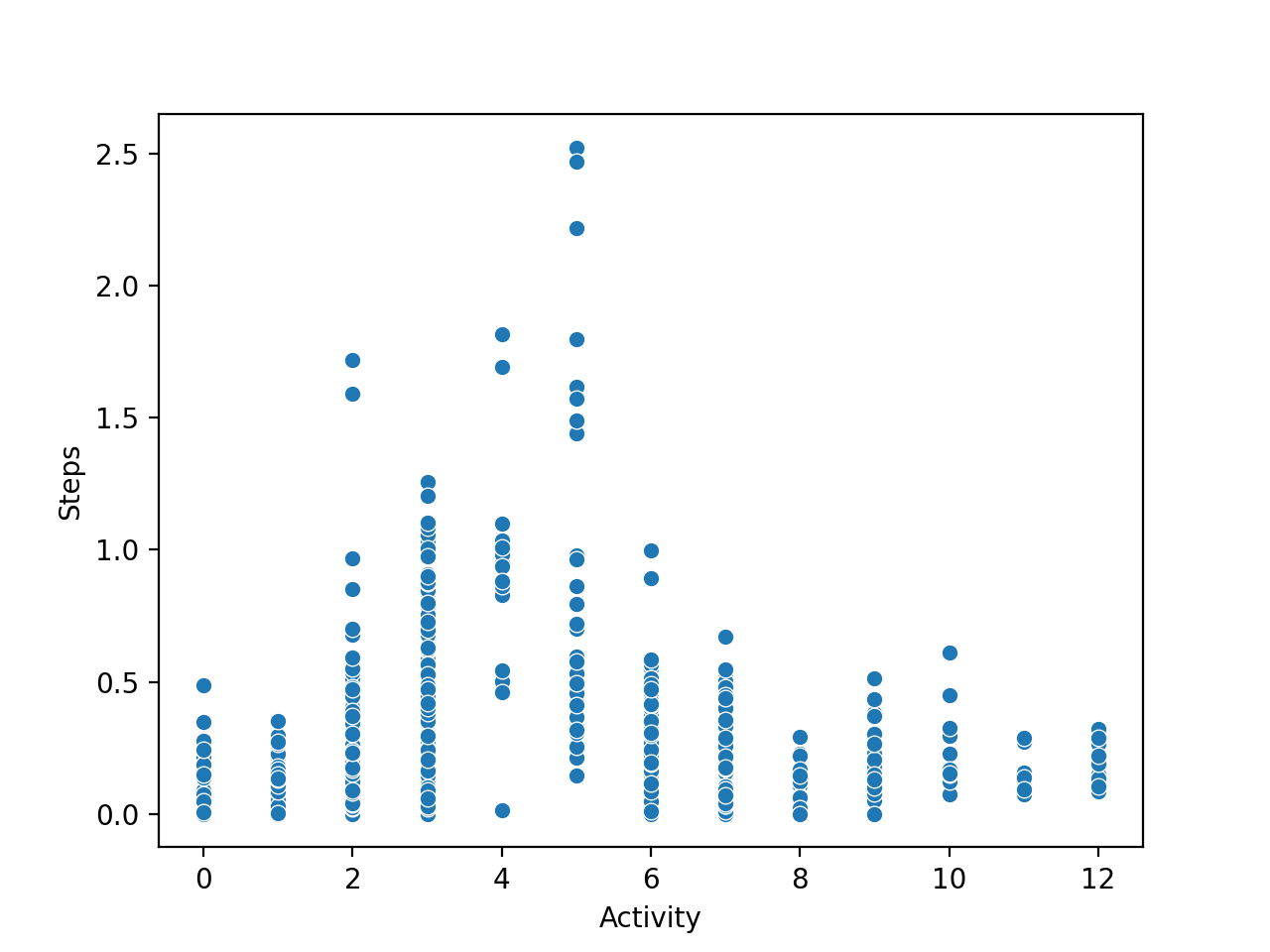
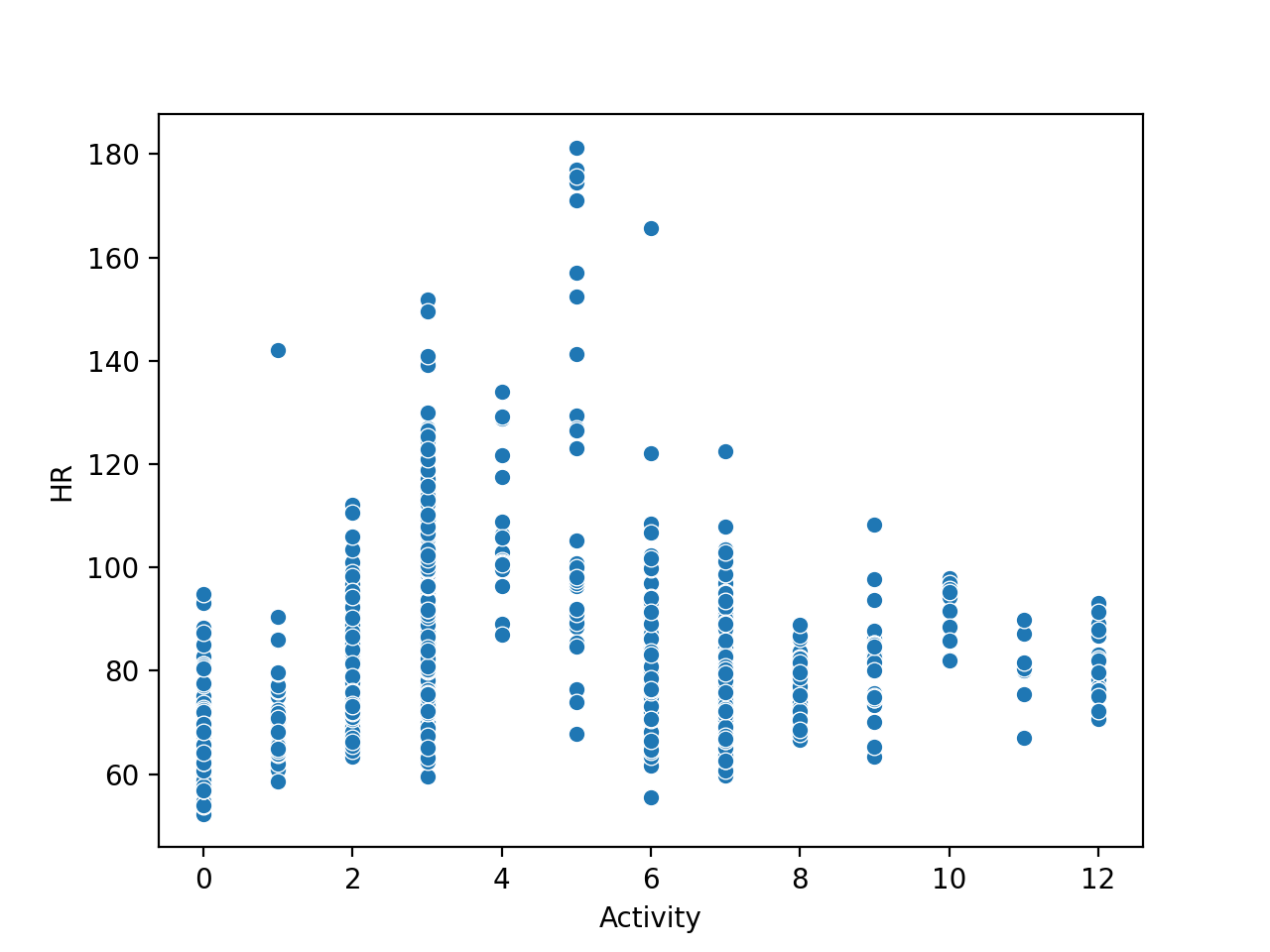
\* Unfortunately, combining samples over 5 minutes period into one caused a problem for the *Activity* feature. Sometimes, not all data points over the 5 minutes period had the same activity label so the most frequent activity label was chosen to represent that 5 minutes interval. For example, if the user was exercising from 10:37 to 10:40, the activity label from 10:35 to 10:40 was assigned exercising. This may have lowered the accuracy of the output label.

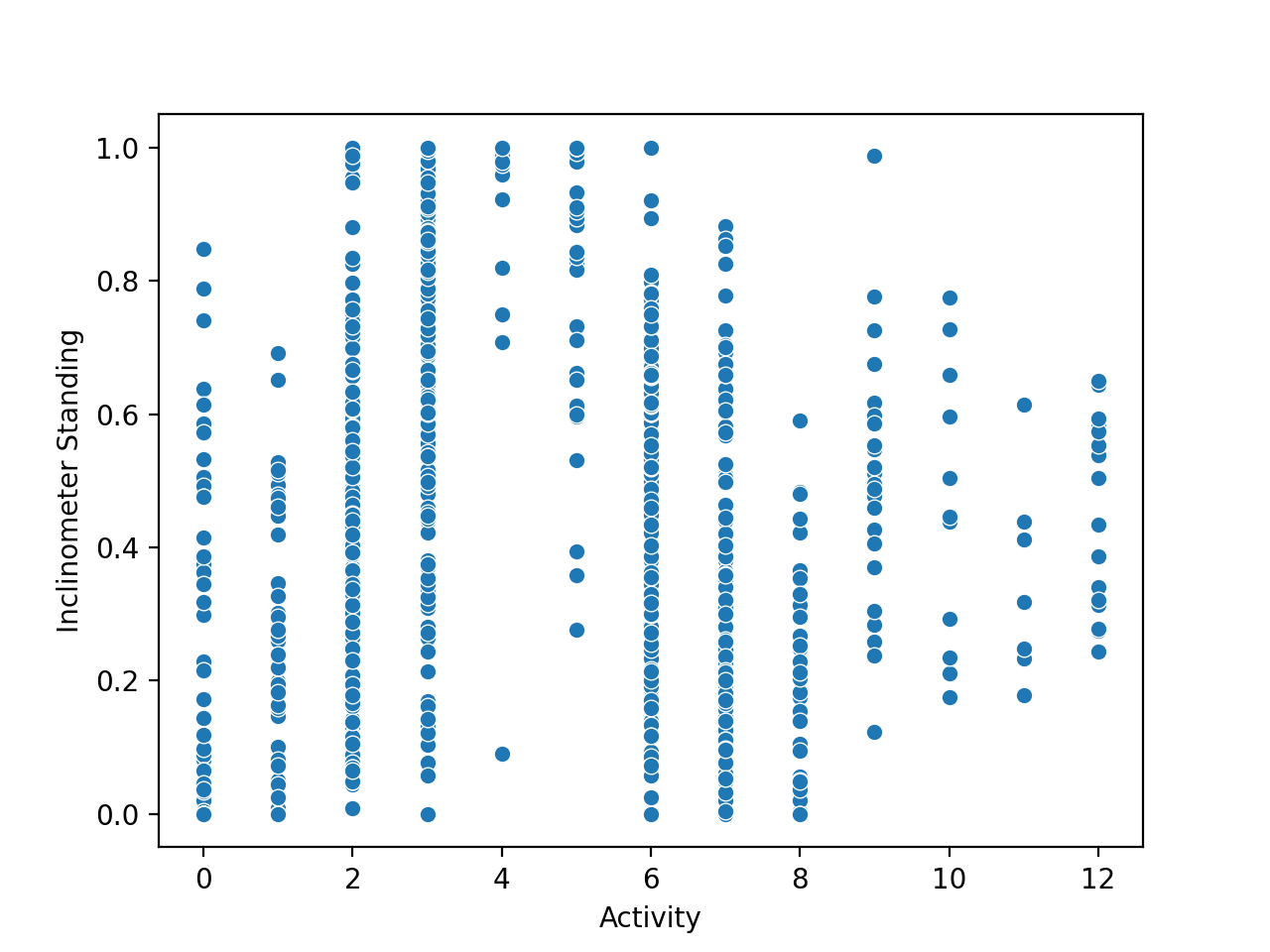
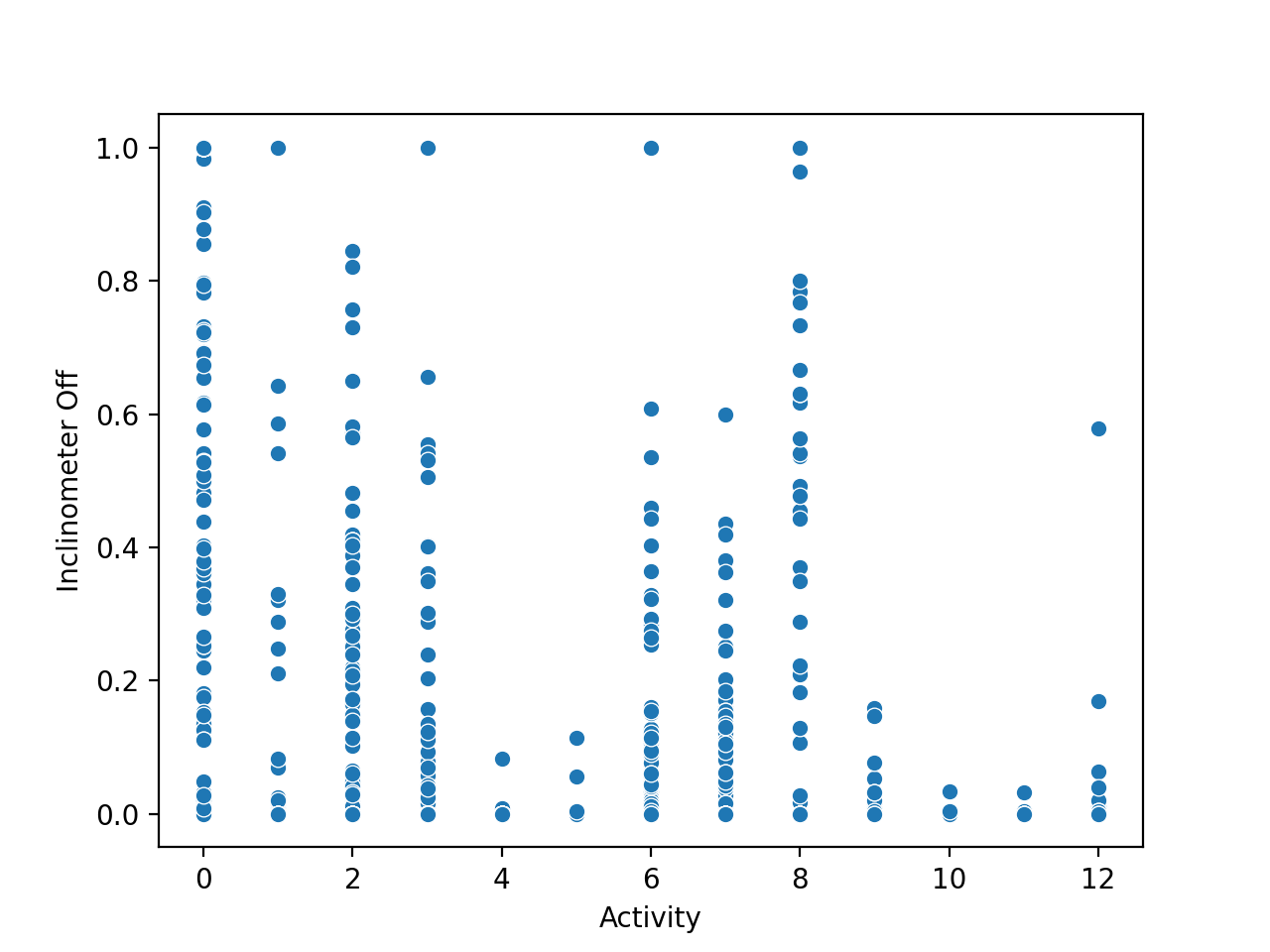
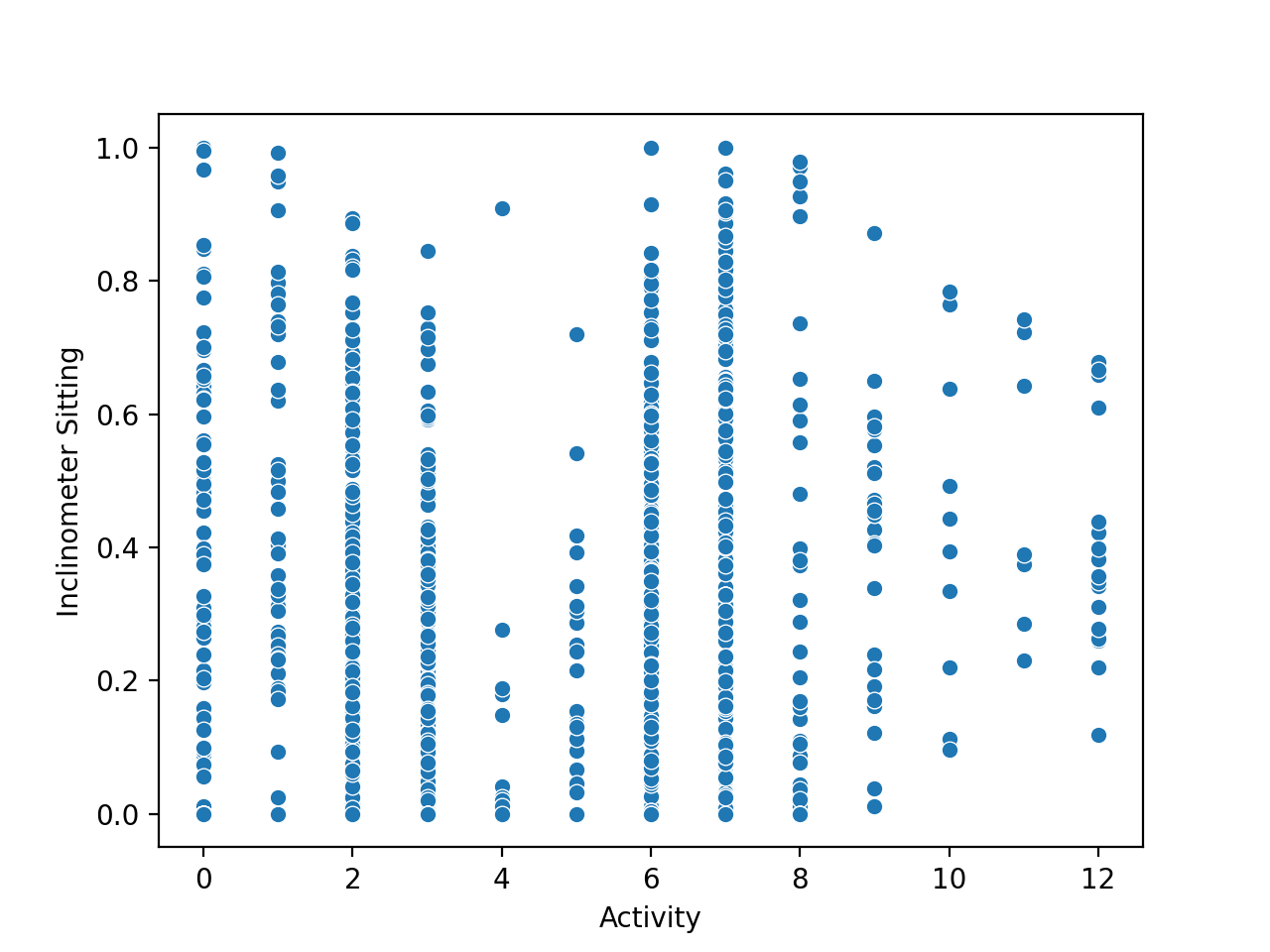
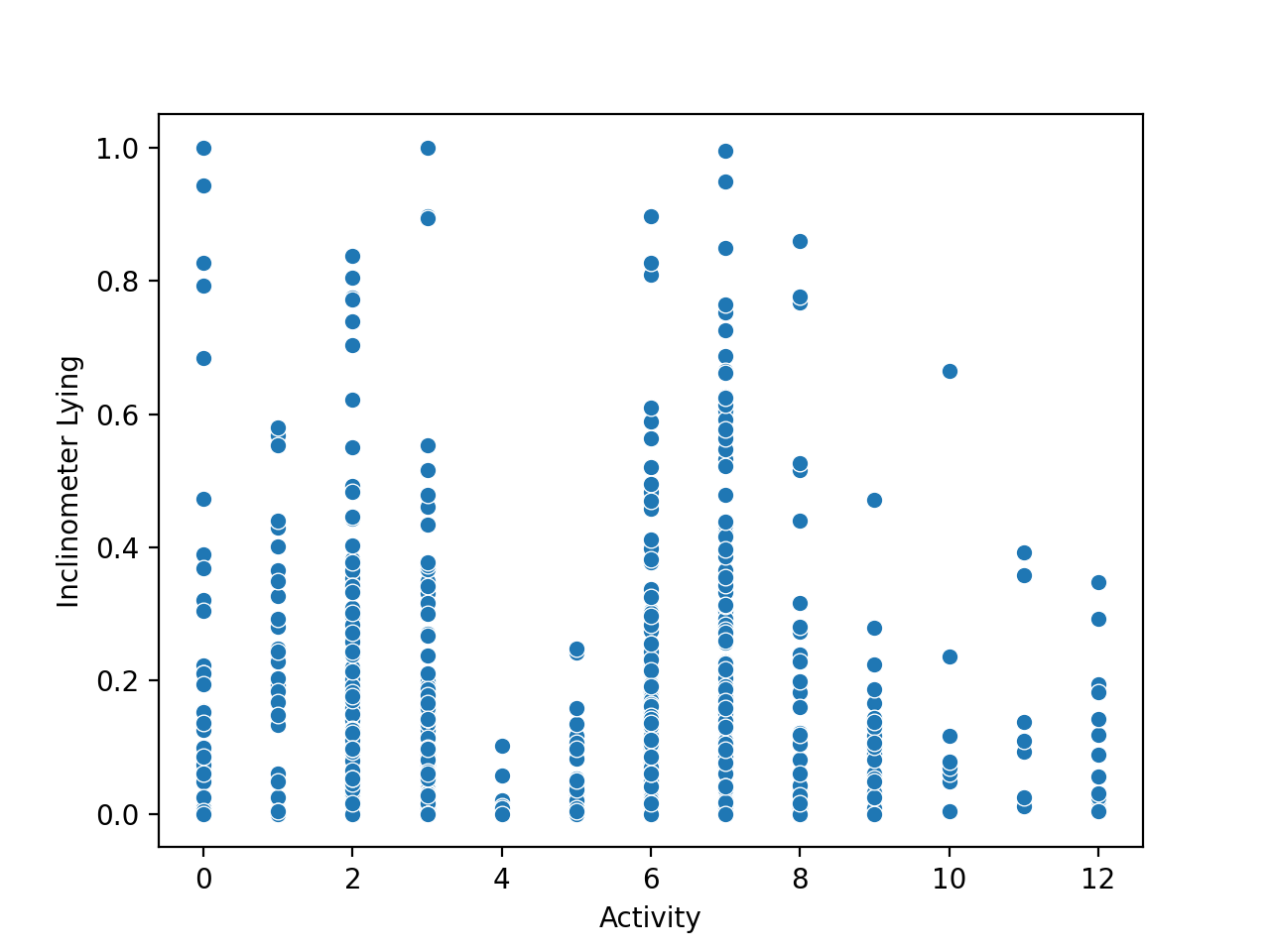
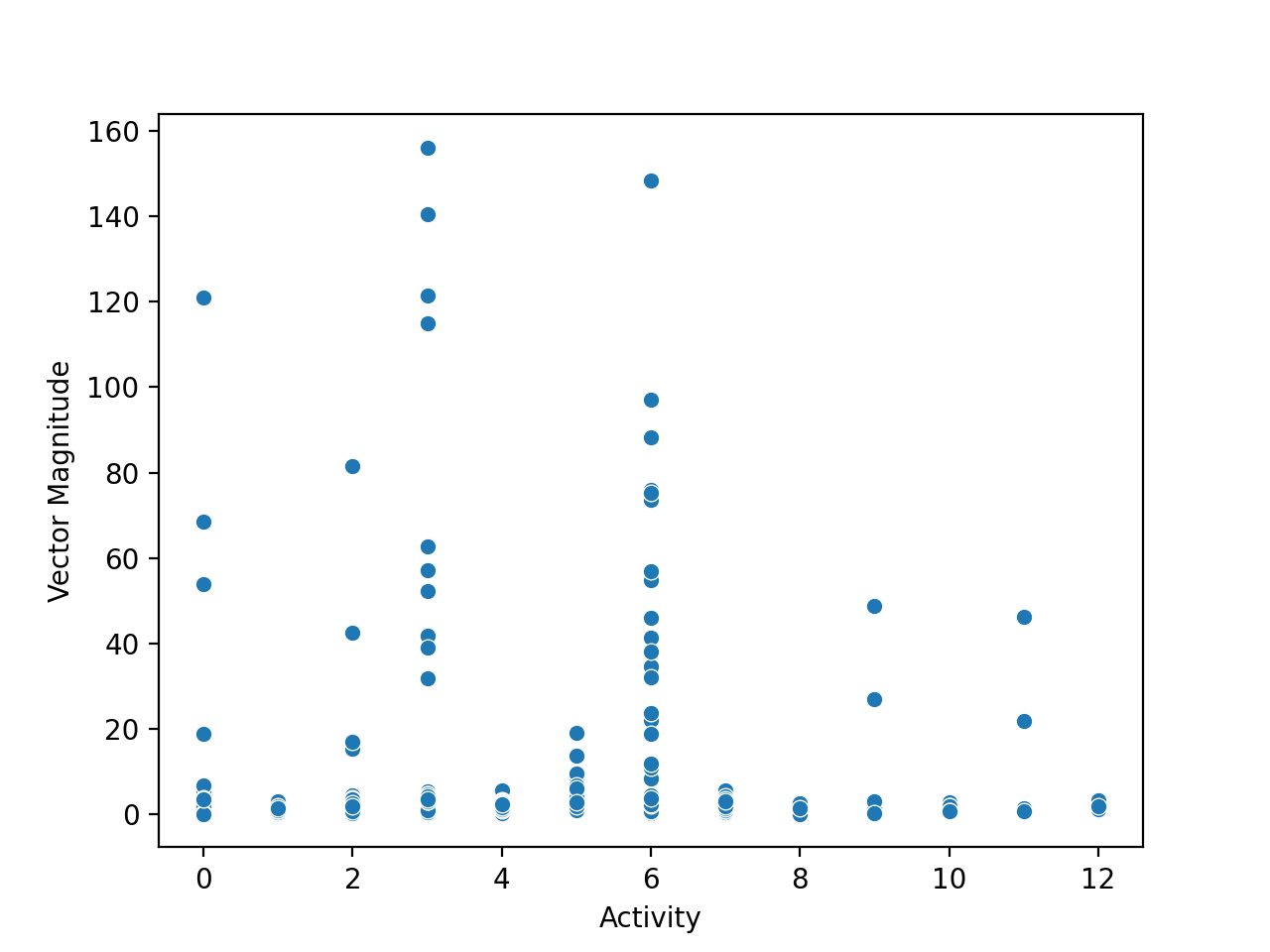
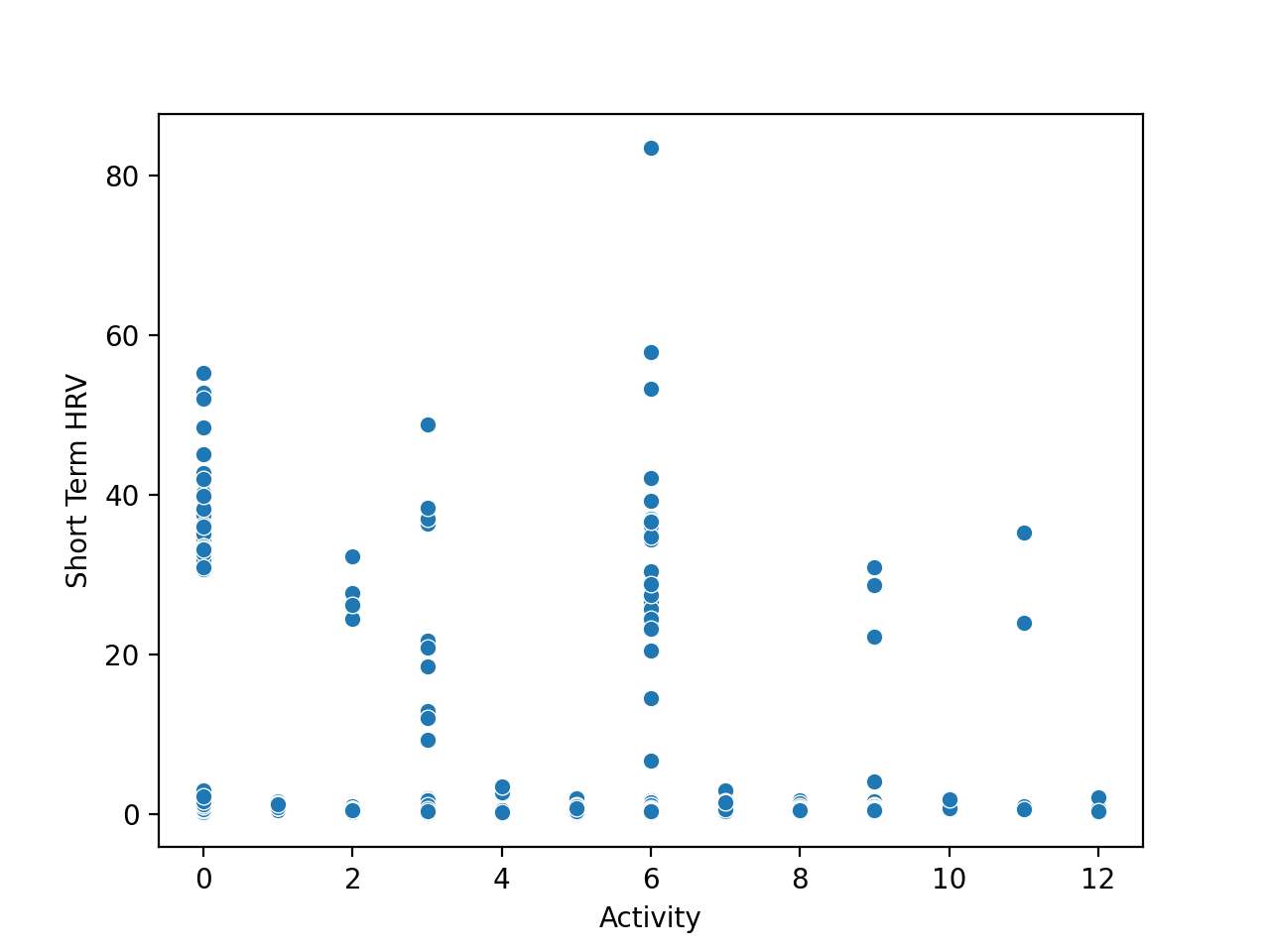
1. The calculated short term HRV and mean *Actigraph* values were combined into one single table. Still, lots of data points in that table had either missing activity label or missing *Actigraph* values due to the gaps in both *Actigraph.csv* and *Activity.csv*. Therefore, it was inevitable to remove all such data points.
2. After collecting all valid data samples from every user, in total there were 1045 samples.

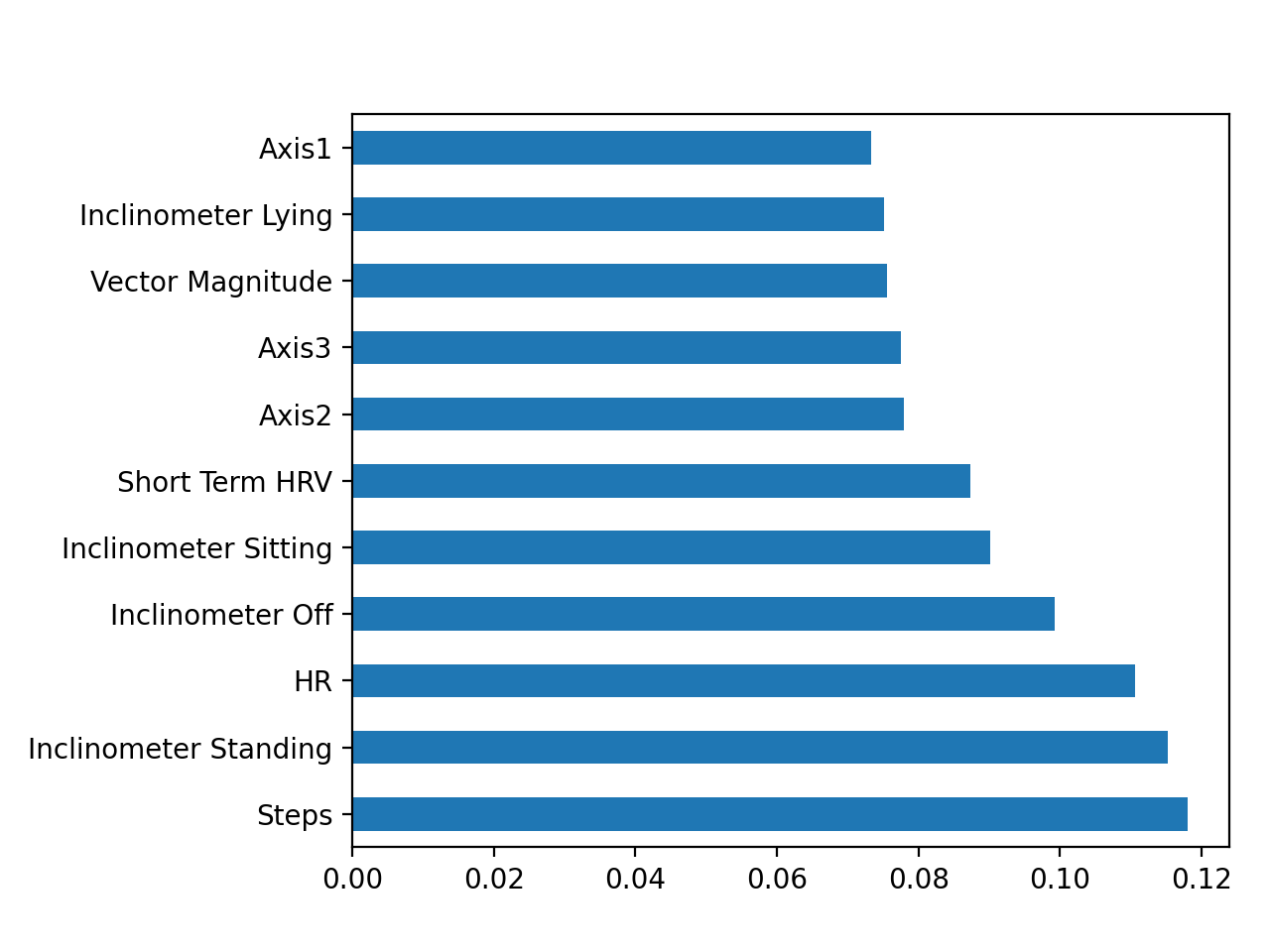
Lots of decisions were made during the feature extraction process. Although the decisions made seem reasonable, they resulted in a few number of samples for the classifier.

## Data overview

Quite large number of samples had activity labels of 0, which is not mentioned in the data description on Physionet at all. After removing these samples, the number of samples will be reduced even further. Also, there is a great imbalance among the activities as some activities were observed far more frequently than others. As it is not possible to collect more samples, some other methods to combat this imbalance would be necessary.





The figure shows the importance of each attribute with respect to the activity label. The importance of Short Term HRV is not as big as I thought considering that the entire dataset was converted into 5 minutes frame just to match with Short Term HRV. For the construction of the classifier, top 6 features were used: Steps, Inclinometer Standing, HR, Inclinometer Off, Inclinometer Sitting, Short Term HRV.

## Constructing the Classifier

Decision Tree Classifier from a python library ‘sklearn’ was used to construct a classifier. It can be found in ‘classifier.py’ on Github. Unfortunately, the accuracy of the classifier was around 40%. So hyperparameter tuning has been attempted using GridSearchCV from ‘sklearn’ but a meaningful increase in accuracy has not been seen yet. More attempts necessary.

## Possible improvements

* As the importance of short term HRV is not very big, try constructing a model with data from Actigraph only. In this case, data doesn’t need to be combined every 5 minutes.
* Not sure if this is valid but try using the raw Actigraph data (every second) with the calculated short term HRV. So in every 5 minutes interval, the Actigraph data will have the same short term HRV value calculated for that specific interval.
* Instead of decision tree classifier, K-Nearest Neighbours can be used to construct a classifier. But the imbalance in the output label might be a problem.