Biomedical Data Science

Christopher Chiou, n10249435: Shivneil Deo, n8872198: Drashti Kanubhai Nayak, n10599568

Validity of Apple WatCH as a Sleep Tracking device: Classification Accuracy with ENMO readings

IFN 646

Contents

[Introduction: 2](#_Toc55936064)

[Approach: 3](#_Toc55936065)

[Findings: 4](#_Toc55936066)

[Discussion: 9](#_Toc55936067)

[Conclusion: 11](#_Toc55936068)

[References: 11](#_Toc55936069)

[Literature sources: 11](#_Toc55936070)

[Data Reproducibility: 11](#_Toc55936071)

[Tools: 11](#_Toc55936072)

[Appendix: 12](#_Toc55936073)

[Pre-processing: 12](#_Toc55936074)

[Random Forrest: 14](#_Toc55936075)

[K-nearest Neighbours 17](#_Toc55936076)

[Gaussian Naïve Bayes 19](#_Toc55936077)

[Multilayer Perceptron 20](#_Toc55936078)

# 

# Introduction:

Quality sleep is an essential part of a healthy lifestyle. Sleep deprivation can lead to physical and mental health problems, injuries, loss of productivity, and an even greater risk of death (Nhlbi, 2020). Increasing the proportion of adults who obtain an adequate amount of sleep is considered an important public health objective (Healthy People, 2020).

The added awareness surrounding good sleep has increased the number of available devices within the consumer sleep tracking (CST) market which allow the user to be mobile and continue a normal routine in his/her natural sleep environment while recording sleep/wake behaviour. A CST device is defined as a sleep tracking device that is made for consumer use. This may include smart watches, fitness bands, etc. The device is used to capture activity that can be transformed in sleep/wake recordings

The widespread usage of CST by the general population has contributed to the growing popularity within the scientific community. The low entry requirement (ie. Lower cost, low level of expertise to use the device) and the implementation of multisensory systems provide access to large sleep datasets which can be investigated for factors regarding sleep health and wellness.

Among wearable consumer devices, the Apple watch leads with a 50 percent share of the smartwatch market (Strategy Analytics, 2019). Apple positions its watch as a wellness device; capable of performing ECGs, able to measure oxygen saturation in blood and detect irregular heart rhythms (Apple, 2020). Apple’s integrated sleep app provides records of the duration of sleep, movement disturbances and heart rate. The data that supports the accuracy of these results remains undisclosed.

Despite the advantages of CST wearables, the lack of empirical data that supports the validity and reliability of CST hinder its recognition as a device considered for application in medical evaluation and treatment (AASM, 2016).

Previous research from Rookham et al. (2019) suggest that Apple watch ENMO values and “activity counts” derived from epochs can mimic the sleep/wake classification of a Phillips Actiwatch (an actigraphy device that is used by clinicians and researchers to monitor individuals which suffer from sleep disorders). Suggesting that additional research could be conducted with the Apple watch to compare its performance in Actigraphy with a Philips Actiwatch. The further optimization of classification algorithms could improve the accuracy of sleep/wake classifications and assist in providing additional information about the validity of using the Apple Watch as a device for medical evaluation.

ENMO values are captured as part of the motion detection system within the Apple watch. ENMO data represents Euclidean Norm Minus One which is used to measure raw acceleration data. Traditional methods of measurement require that the data is pre-processed to remove gravitational and noise components. The ENMO metric is computed from a resultant vector and therefore can be regarded as signifying dynamic acceleration (Bakrania, 2016). ENMO data can be summarised into 15 seconds of acceleration data which would reproduce activity counts. The Actiwatch uses “activity counts” which summarise the “activity” for a 15-second epoch. Proper use of ENMO data is crucial to the classification of sleep/wake readings.

From this we formulated a research question:

Can ENMO data and “activity counts” accurately classify sleep/wake data using Machine Learning techniques.

The research question addresses 2 challenges within the CST scientific community:

**Classification of sleep/wake data with algorithms.**

Algorithms used to classify sleep/wake data within the industry are often unvalidated, unstandardized and undisclosed. The research question aims to explore how different algorithms perform in classifying sleep/wake data with the Clinical Actiwatch sleep/wake results used as the ground reference point. A high accuracy in classification indicates that the Apple Watch could match the performance of the Actiwatch.

**Validity of a CST as a replacement to a clinical device:**

The research question aims to provide initial validation into whether a CST device can accurately measure actigraphy sleep/wake. A successful outcome would provide scope for further research into other forms of CST devices that could be used for monitoring sleep health.

# Approach:

As the main purpose of the work was to study the performance of machine learning classifiers to classify sleep/wake readings, we chose 4 well known supervised machine learning algorithms: These were **Random Forest**, **Gaussian Naïve Bayes**, **Multilayer Perceptron** and **K-Nearest Neighbour**. The machine learning algorithms and feature selection implemented were part of the Scikit-learn suite.

The four algorithms were selected because they each represent a different approach towards classification.  Random Forest is an ensemble learning method based on decision trees, K-Nearest Neighbours is a lazy learner, Naive Bayes is a probablistic classifier and a Multilayer Perception is a feedforward neural network.  Our goal was to find a specific class of classifier that would work best on this form of data.

Two experiments were conducted within this study to compare the outcomes of sleep/wake classification. The first experiment used all the features within the Apple watch dataset and the Philips Watch activity counts to generate models for classification. These models were compared to the Philips actiwatch sleep/wake data to determine the accuracy of using these algorithms. The second experiment involved looking at only Apple Watch data to determine the validity of using the Apple watch as an actigraphy device. The Actiwatch “activity counts” are included in the initial portion of the model building process to verify the performance of each dataset. After the initial accuracy score is computed the “activity counts” are removed. The remaining model will keep the data derived from ENMO, which will show how accurate the model is if it was built only from Apple Watch data. This process highlights the difference in accuracy if ENMO is used as the main data source for classification.

The data was first pre-processed and 66% of the total data set was used for training and 33% used for classification. The performance of the classifiers were evaluated for their accuracy in predicting sleep/wake compared to the Philips classification. Adjustments were made from the test-train-split if any results indicated that overfitting was an issue.

**Data pre-processing**

The dataset being examined contained 4 features:

- A datetime.

- A Euclidean Norm Minus One (ENMO) value for the last 15 seconds of acceleration data from the Apple Watch.

- An activity counts value, taken in the same 15 seconds as the ENMO value, measured from the Philips ActiWare.

- A Philips ActiWare classification of whether the user is asleep or awake (1 for awake, 0 for asleep).

The first step of our approach was pre-processing (Appendix 1, Jupyter notebooks).  The datasets were split apart in multiple CSVs and then merged for ease of use.  Secondly, the timestamps were modified to provide more anonymity and the dates were manipulated.  Without the dates only a time values exists and time is not so easily manipulated.  An easier value to manipulate, and to compare different parts of the data, would be a discrete epoch representing the second of the day.  For example, all rows representing "01:21:42", would become 4902 (1 \* 3600 + 21 \* 60 + 42), which is another way of saying 1:21AM and 42 seconds is the 4902 second in a day.  Intuitively this could also be useful to a sleep/wake classifier.  For most users, on most days, the 4902second in a day is likely one someone will spend asleep. NaN values are removed in the preprocessing stage.

# Findings:

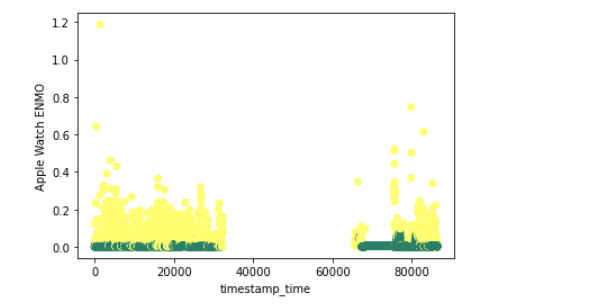
The results suggest that all the classifiers scored more than 90% accuracy when the full feature set is used. This accuracy decreases for all the classifiers by a minimal amount when the Philips activity count is taken away in step 2 of the experiment. The Random Forest and Multilayer Perceptron performed the best out of the classifiers on both the full feature data set and the data set without “activity counts”. This was followed by Naïve Bayes and then the K-nearest neighbours algorithm.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Full Feature set | Feature set (without activity count) | Feature set (without ENMO) |
| Random Forest | 97.48% | 96.25% | 97.44% |
| K-nearest Neighbour | 95.14% | 91.69% | 95.51% |
| Naïve Bayes | 97.32% | 92.18% |  |
| Multilayer Perceptron | 97.55% | 96.26% | 97.46% |

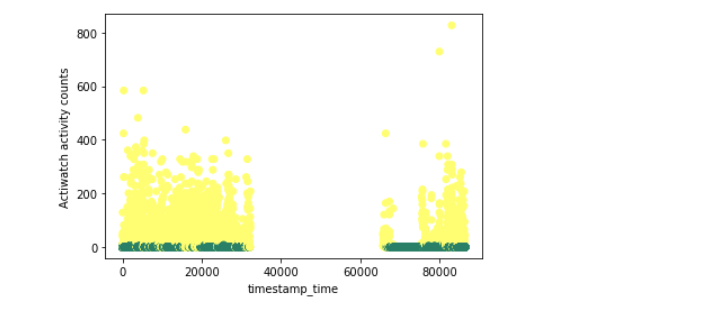
The data and results for each algorithm are presented separately below:

**Random Forest:**

The data was explored and plotted to show when the data was recorded.



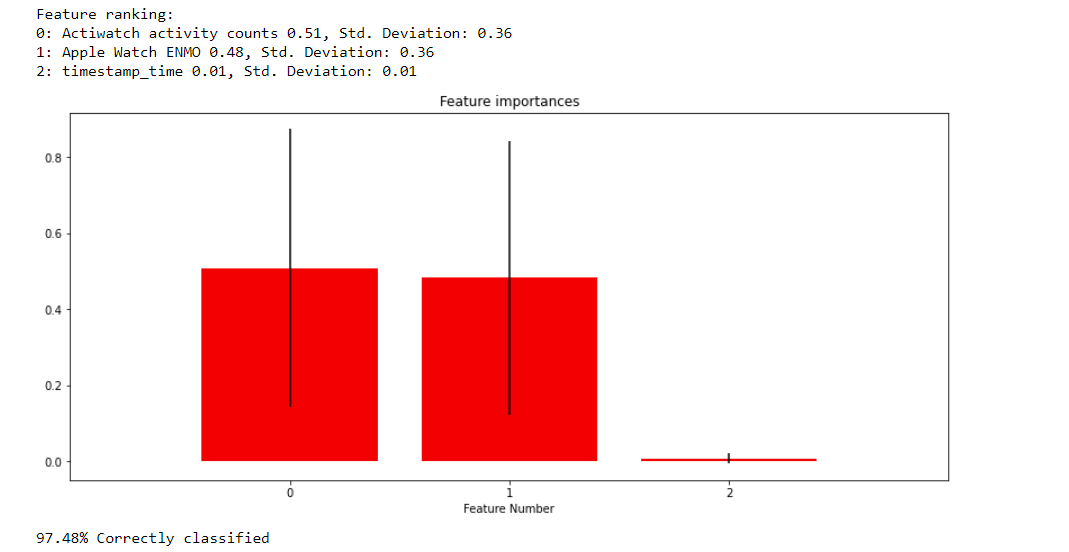
*Time of ENMO recording*



*Time of activity counts recording*

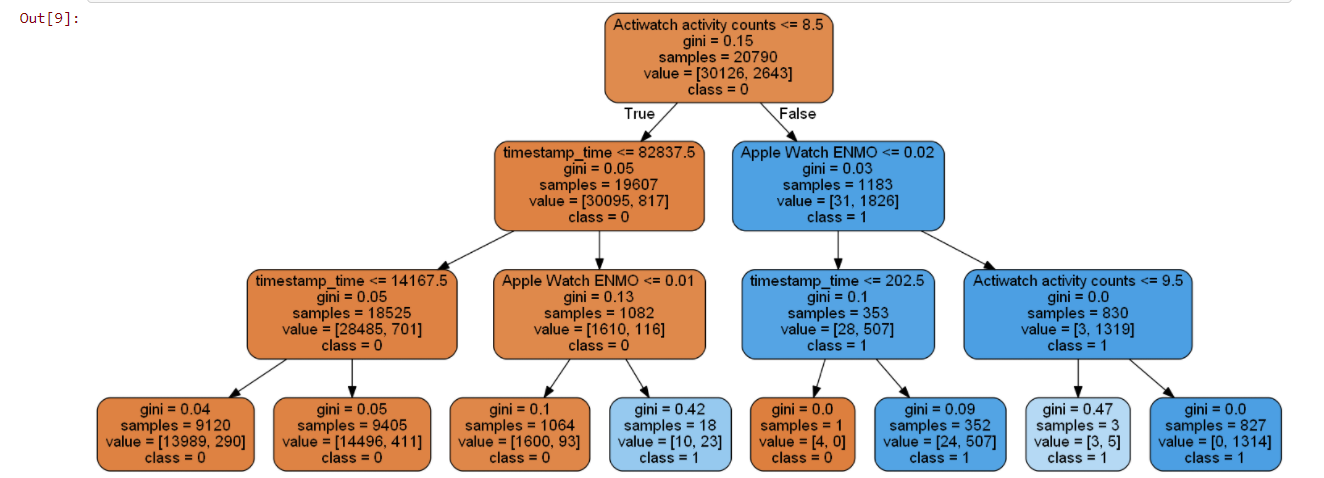
Both data-plots show that recordings stopped for a period. This indicates that during the day, users did not wear the watches or data was not collected during that timeframe. We can assume that the data that is being analysed is only from the time close to a patient’s bedtime.

Further analysis from the random forest shows which features were important in the classification.



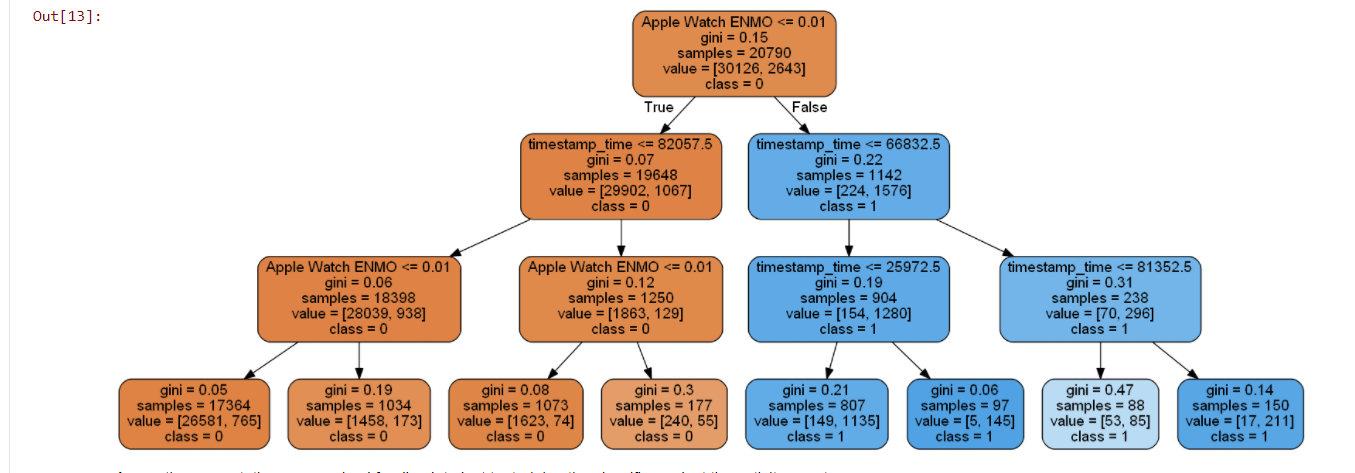
*Feature Rankings*

Both activity counts and ENMO were important features for the model.



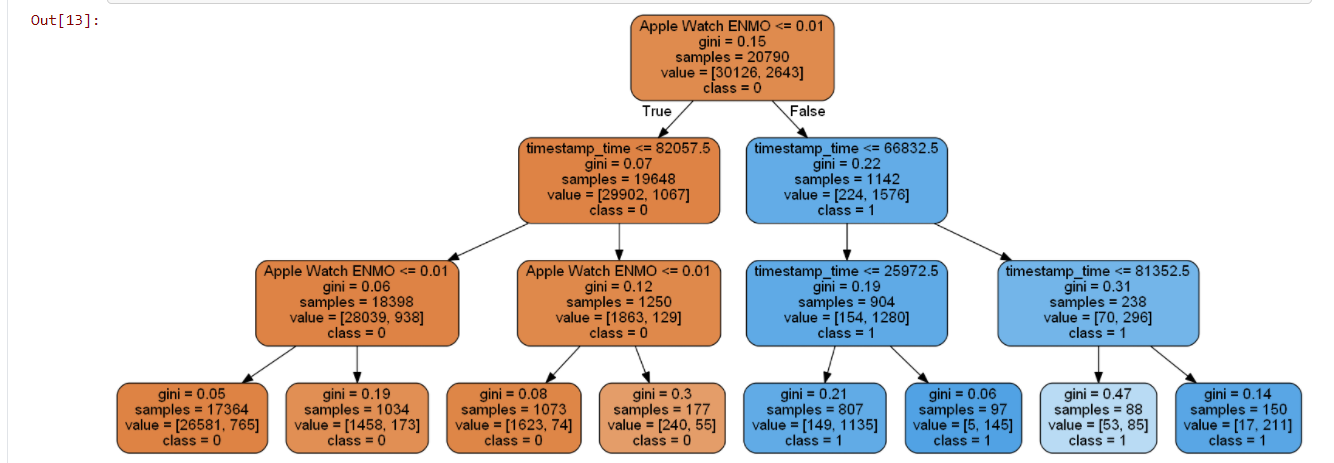
*Random Forest Tree (ENMO/ activity counts)*

There appears to be some signs of overfitting in this model as the gini scores approach 0 on the branches of the tree. When the activity counts data is removed then the signs of overfitting disappear.



*Random Forest Tree (ENMO only)*

The accuracy drops slightly to 96.25%. To scope the importance of ENMO in the classification ENMO is dropped from the dataset and a model is generated with only activity counts data.

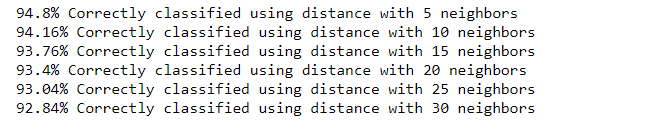


*Random Forest Tree (Activity counts)*

The accuracy drops to 97.44% when ENMO is excluded.

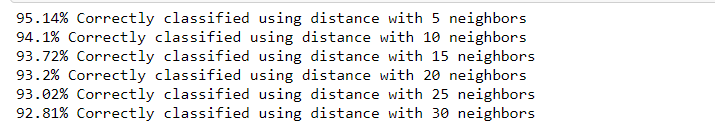
**K-Nearest Neighbour:**

The data is first classified with weight points that are determined by a distance measurement.



*K-Nearest Neighbour (Full dataset/weighted points)*

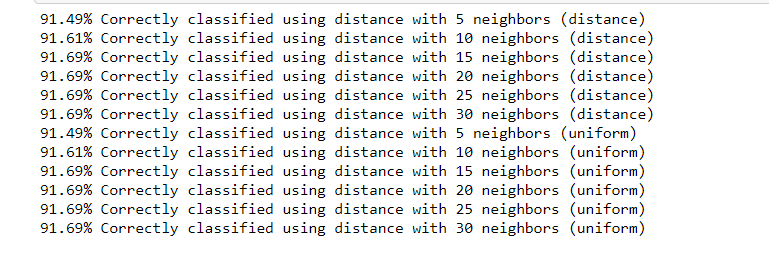
The data is then classified with equal weighting points as a comparison.



*K-Nearest Neighbour (Full dataset/uniform weight)*

The performance is like each other regardless of weight points.

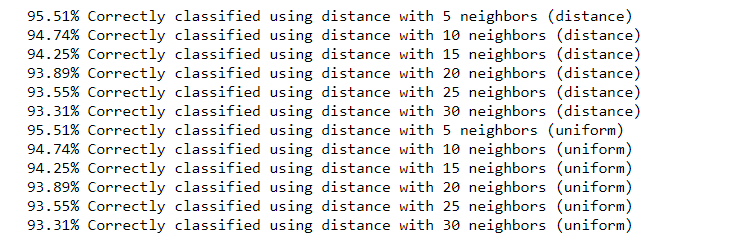
Further analysis is done to remove the Activity counts.



*Activity counts removed*

The performance decreases slightly here with the removal of Activity counts.

We can compare this result to a classifier with ENMO removed.

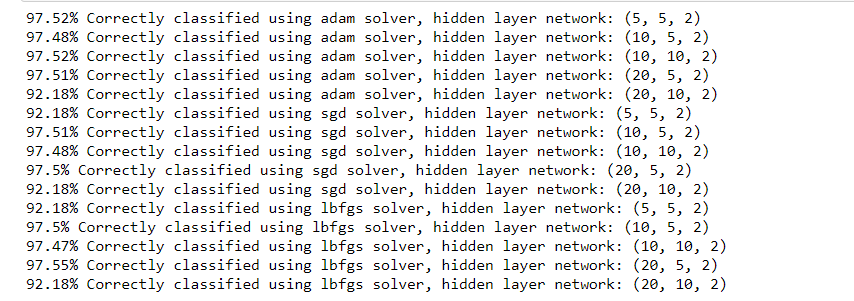


*ENMO removed*

The percentage classified correctly is higher when ENMO is removed. This would indicate that for the K-nearest neighbours’ algorithm ENMO would not be suitable as an accurate indication of sleep/wake data.

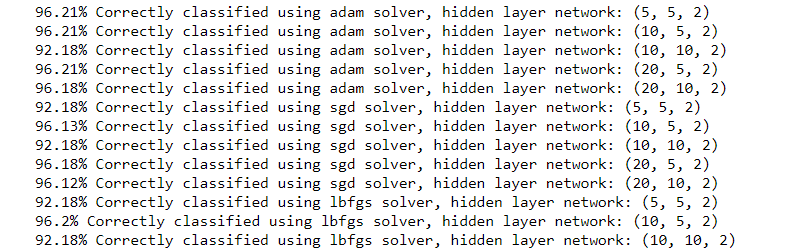
**Multilayer Perceptron:**

The data is first standardized for the Multilayer Perceptron so that an outcome can be converged for the network. 3 solvers are used for this classification.



*MLP (Full dataset)*

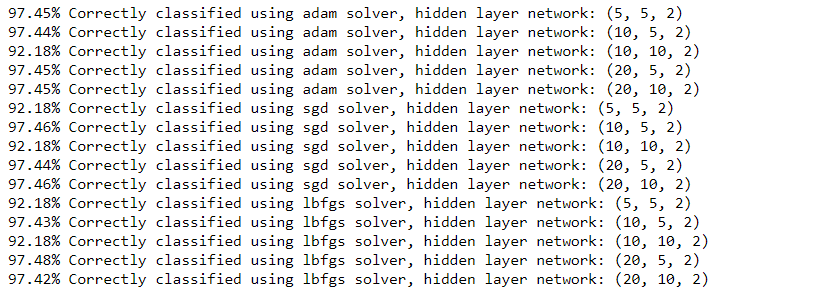
The activity count data is then removed from the classifier to see how much accuracy is maintained.



*MLP (Activity count removed)*

The accuracy drops however the change is not drastic.

This data is then compared to a model where ENMO is removed.



*MLP (ENMO removed)*

The model with activity counts is more accurate and better performing than the model with activity counts removed.

# Discussion:

The aim of the research is to understand whether Apple Watch ENMO data can accurately replicate the sleep/wake classification of the Philips Actiwatch. The first experiment with the full feature set was to explore the accuracy of using multiple machine learning algorithms to determine the closet the classifiers could get to the correct classification. The results that were returned provided a classification accuracy of more than 97% except in the K-nearest neighbour algorithm. The high accuracy results indicate that the tested classifiers would be suitable algorithms in classifying future data sets with full features as a high percentage of data would be classified correctly.

The feature sets without activity counts varied in performance. Random Forest and Multilayer Perceptron algorithms performed the best with results of over 96%. Performance of this level may be suitable for actigraphy results. Naïve Bayes and K-nearest Neighbour all performed lower with only Apple Watch data. This indicates that these algorithms may not be suitable when classifying sleep/wake data.

As activity counts are used to classify the sleep/wake recordings in the actiware software they should have the highest accuracy in classifying sleep wake data with other algorithms. This is due to the nature of having the sleep/wake recordings derived from activity count readings. The classifications where ENMO data is removed performed better than those with activity counts removed.

When classifications were done with ENMO data only, the accuracy was lower than those that were classified with both ENMO and activity counts data and activity counts data by itself. This would mean that ENMO data from the apple watch is unlikely to be as accurate as if the ENMO data was summarised into a different form. Ie. Into 15 seconds epochs like activity counts. Further processing of ENMO into different epoch timings may result in higher accuracy.

When both ENMO and activity counts were used, there were signs that the algorithm overfit. The activity counts performed better than ENMO data and signifies that the reduction in accuracy likely comes down ENMO being a worse fit than activity counts in classifying sleep/wake.

There are some limitations with the algorithms. In this paper we compare the direct accuracy of using raw ENMO data against an algorithm generated with both ENMO and activity counts data and then activity counts data by itself. This is done to understand how close ENMO data can compare to activity counts classifications. A limitation would arise with the algorithms used as we faced some blackbox issues with features being difficult to obtain from some of the classifications. The experiment was also limited by the pre-processing completed as NaN values were removed and ENMO was not changed into 15 epoch timings to mimic activity counts directly. Further research can be done through processing NaN values with a different method and with classifying ENMO data into different epoch timings. Different types of acceleration data could be collected so that further modelling can be explored.

We would recommend that further data collection is completed to get a better overview of whether CST devices would act as a good actigraphy device. The experiment directly compares the Apple watch as an actigraphy device rather than comparing the results to other methods of evaluating a patients’ sleep complaints. These may include clinical interviews, sleep diaries and polysomnography (PSG) results (Martin, 2008). A further understanding of how the Apple Watch compares with more traditional forms will give a better indication of performance in a clinical environment.

Using a CST as an actigraphy device is an area that requires additional exploration. Actigraphy did not gain clinical recognition until 2009 after the first recognition in 1995 that it was a useful research tool for sleep. This is the current perception among the CST scientific community in that CST devices are useful for sleep research but lack clarity on clinical usefulness.

Actigraphy infers sleep from the lack of movement. The data provided does not go into depth about the age, demographic and sleep disturbances of the patients. The accuracy of these results will vary in cases where subjects are awake but lie motionless; which incorrectly classifies the subject as awake.

The dataset also classifies 15 second epoch timings instead of the industry standard 30 second epoch timings as per various publications (Martin, 2008.  Hughes, 2014). An exploration of how the classification changes with 30 second timings should be explored.

We have shown which algorithms perform the best at classifying Apple watch data. These algorithms could be further explored by collecting data from other CST devices. These devices could be better suited than an Apple Watch due to the inherent limitations of Battery life and memory capacity.

# Conclusion:

The report used 4 machine learning algorithms as classifiers for sleep/wake data in the apple watch. The experiments were completed in 2 parts. The first part compared the results of using both Apple Watch data and Philips actiwatch activity counts as a reference point to see how accurate the ML classifiers could be. The second portion of the experiment evaluated the sleep/wake classifications of only Apple Watch data. The Random Forest Algorithms and the Multilayer perceptron performed the best. Using less features in classification led to large loses in accuracy for K-nearest neighbours and Naïve Bayes algorithms. Thus, Random Forest and MLP are recommended if further testing is too be completed. The Apple Watch shows promise that it could be used as an actigraphy device however further research is required to determine its performance as a tool for medical evaluation.

# References:

## Literature sources:

Bakrania K, Yates T, Rowlands AV, Esliger DW, Bunnewell S, et al. (2016) Intensity Thresholds on Raw Acceleration Data: Euclidean Norm Minus One (ENMO) and Mean Amplitude Deviation (MAD) Approaches. PLOS ONE 11(10): e0164045. <https://doi.org/10.1371/journal.pone.0164045>

Ibáñez, Vanessa & Silva, Josep & Navarro, Esther & Cauli, Omar. (2019). Sleep assessment devices: types, market analysis, and a critical view on accuracy and validation. Expert Review of Medical Devices. 16. 1-12. 10.1080/17434440.2019.1693890.

Luca Menghini, Nicola Cellini, Aimee Goldstone, Fiona C Baker, Massimiliano de Zambotti, A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code, Sleep, , zsaa170, <https://doi.org/10.1093/sleep/zsaa170>

Miller, C., & Miller, C. (2020). Apple Watch continues domination of smartwatch industry, new Strategy Analytics data shows - 9to5Mac. Retrieved 10 November 2020, from https://9to5mac.com/2019/08/06/apple-watch-smartwatch-industry-q2/

Roomkham S, Hittle M, Cheung J et al. Sleep monitoring with the Apple Watch: comparison to a clinically validated actigraph [version 1; peer review: 2 approved with reservations, 1 not approved]. *F1000Research* 2019, **8**:754 (<https://doi.org/10.12688/f1000research.19020.1>)

Sleep Deprivation and Deficiency. (n.d.). Retrieved from https://www.nhlbi.nih.gov/health-topics/sleep-deprivation-and-deficiency.

Sleep Health | Healthy People 2020. (2020). from https://www.healthypeople.gov/2020/topics-objectives/topic/sleep-health

## Data Reproducibility:

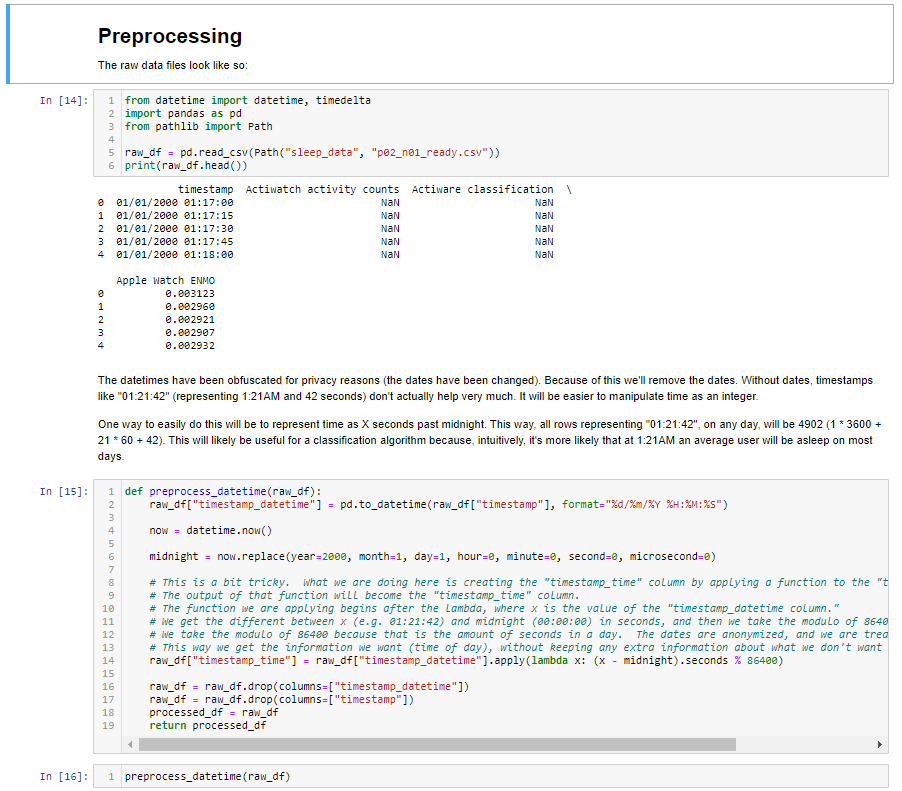
The dataset for the sleep recordings can be found on QUT research Data Finder under Sleep Data. The link is provided here: <https://data.researchdatafinder.qut.edu.au/dataset/sleep-data>. The dataset contains 27 csv files each corresponding to a single night’s sleep. Code for the project can be viewed in the Appendix section or on Github: https://github.com/shi-vy/ifn646\_project.git

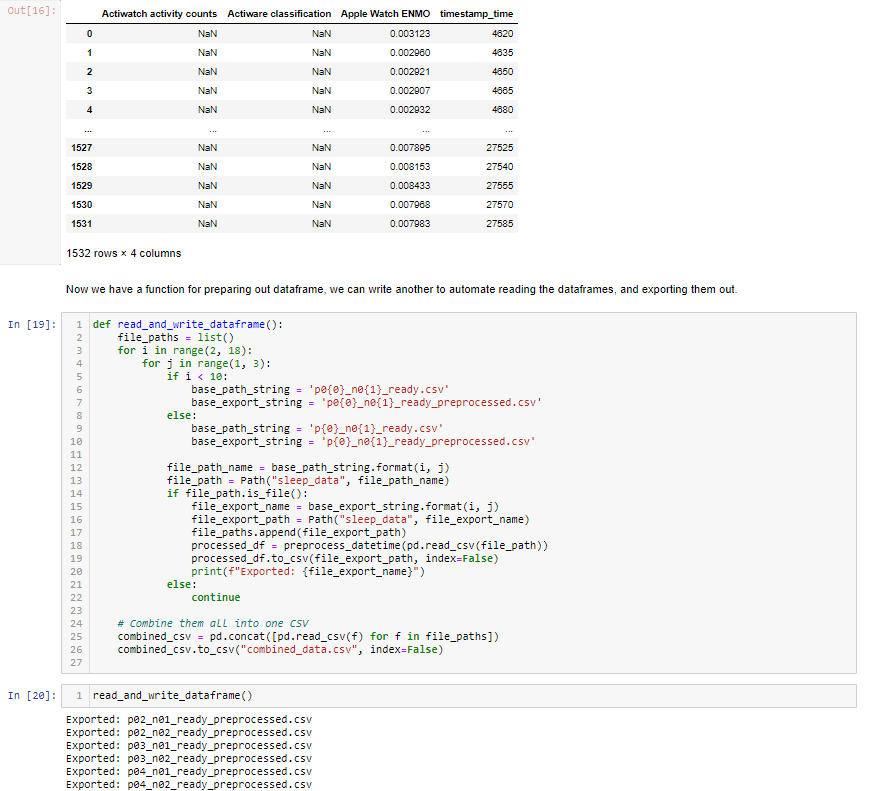
## Tools:

Scikit learn version 0.23.2. Python 3.7. Jupyter Notebook 6.1.5.

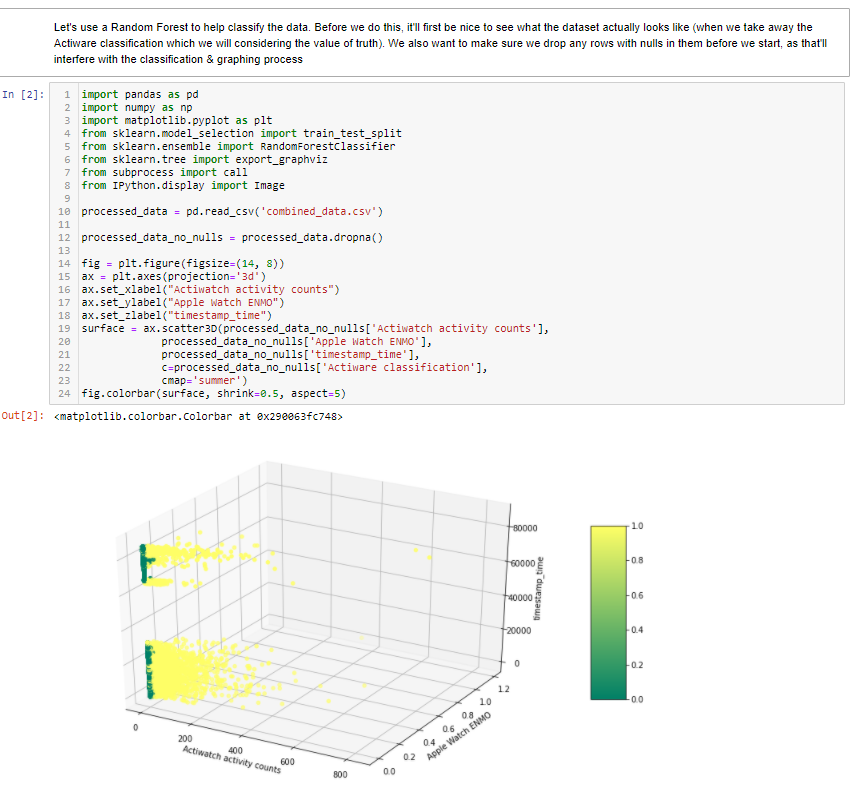
# Appendix:

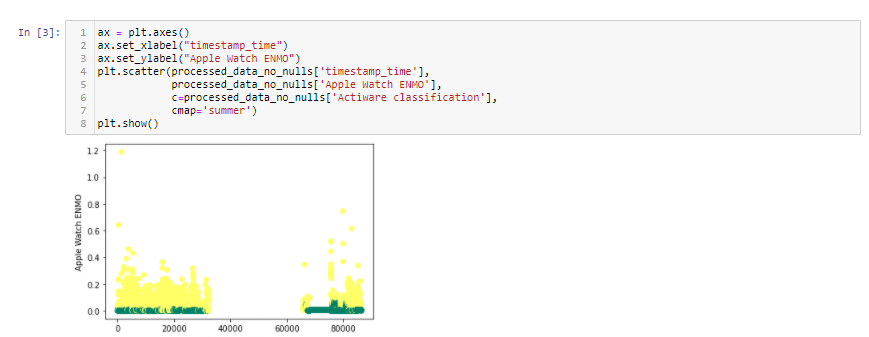
## Pre-processing:

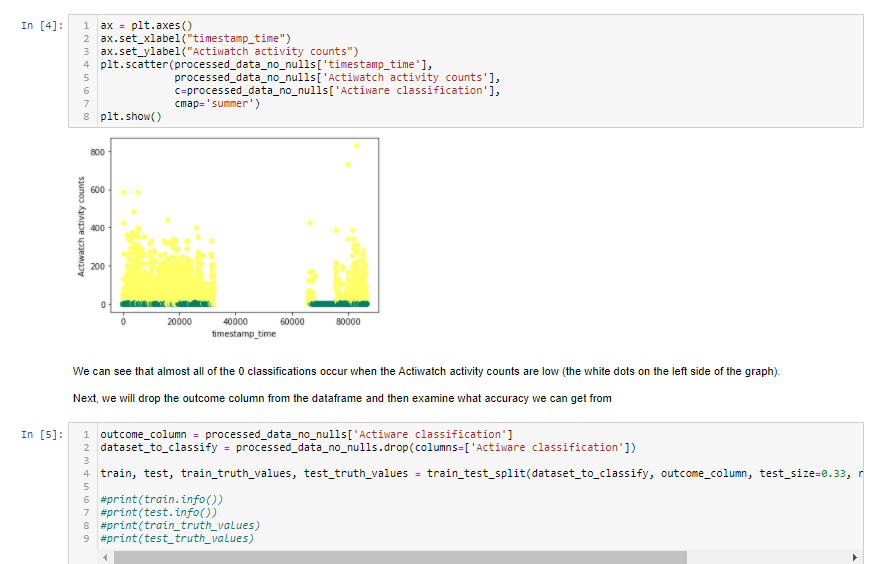




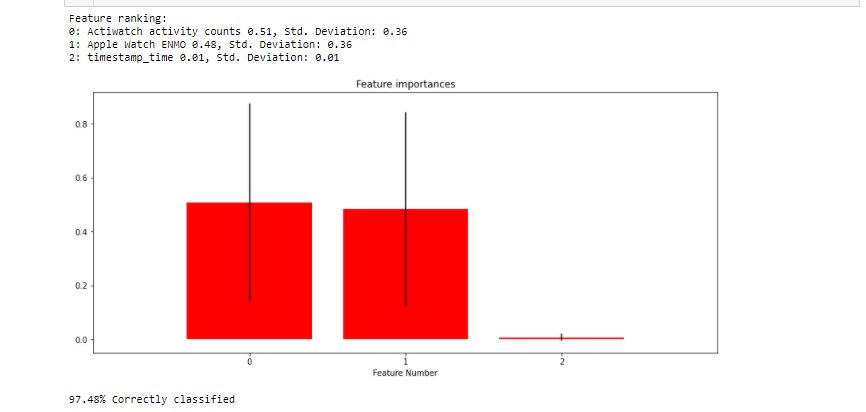
## Random Forrest:

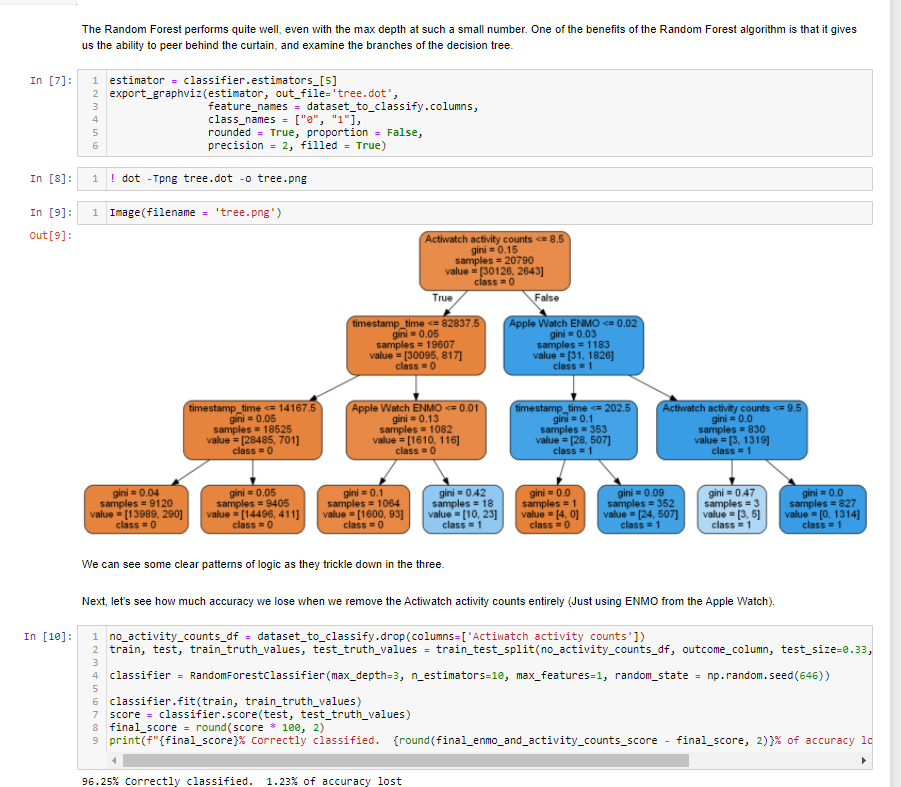


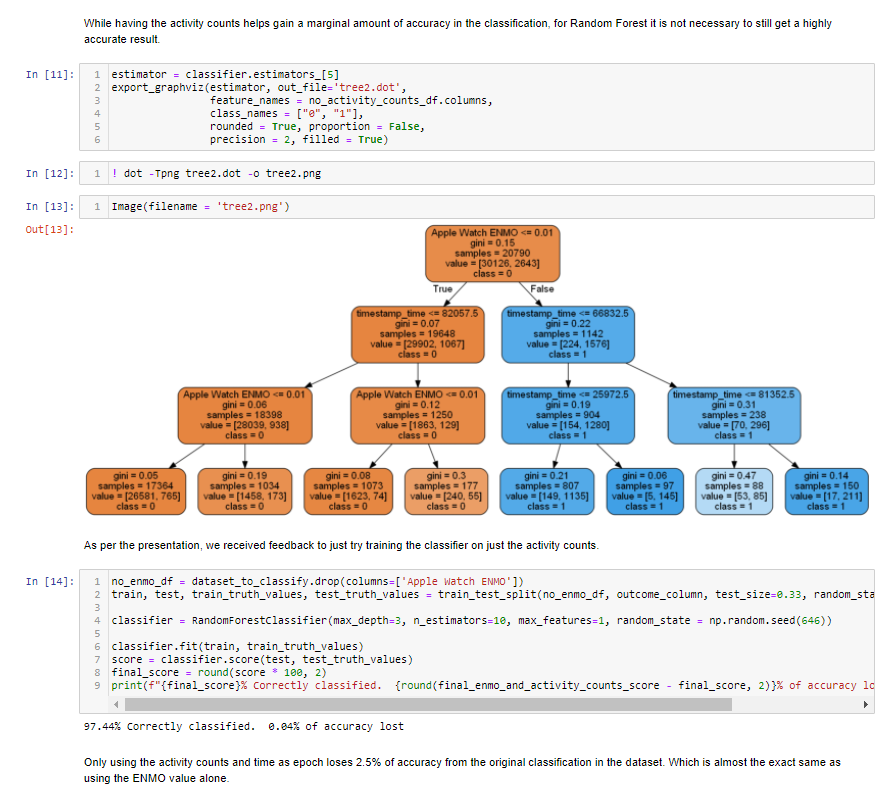






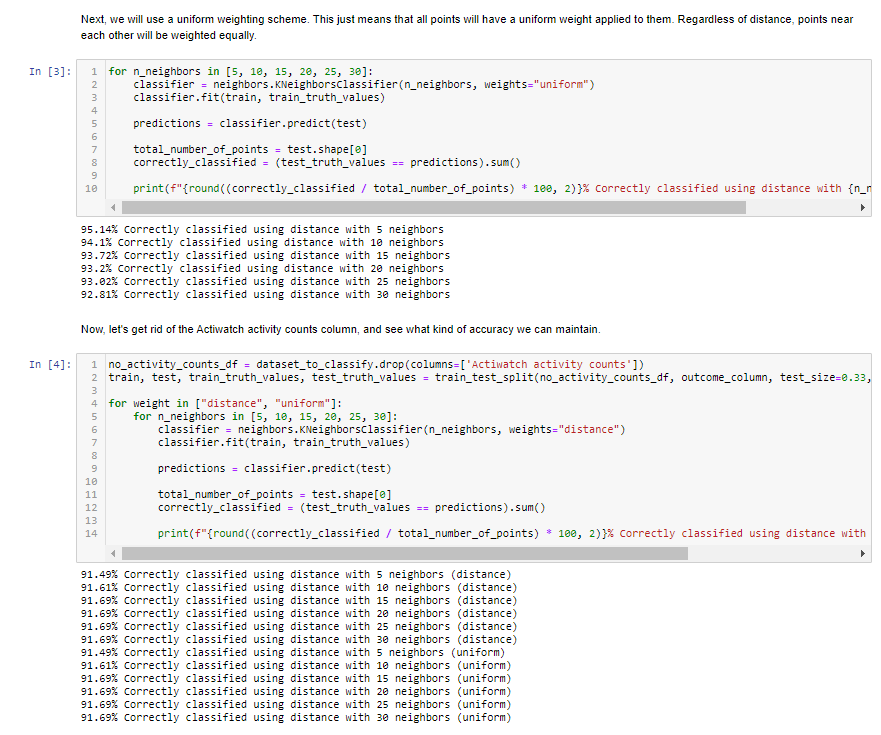




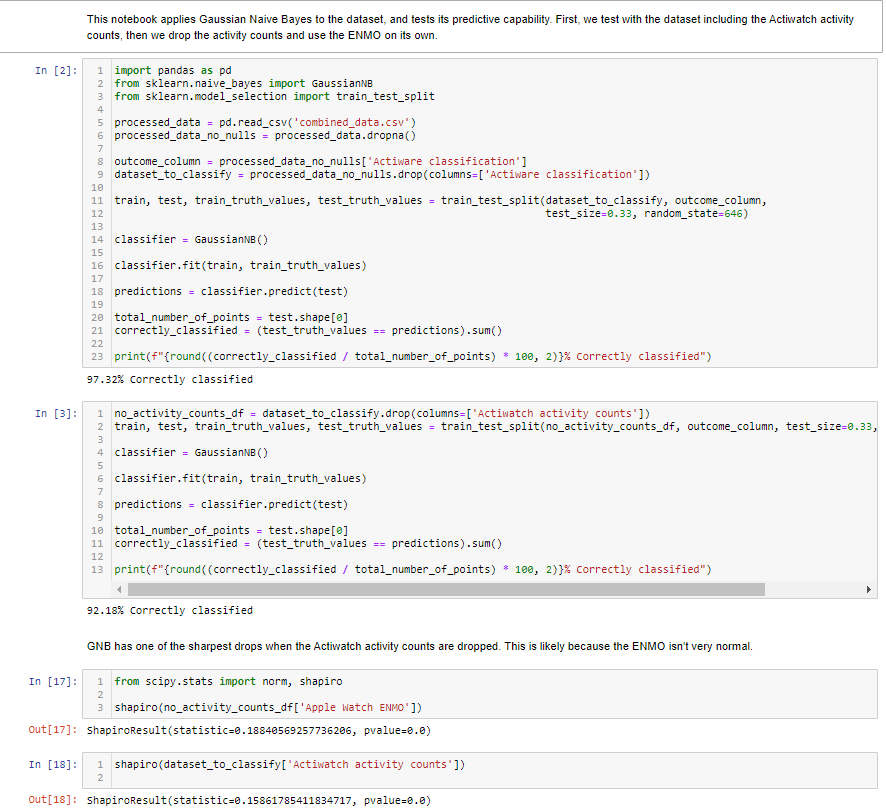


## K-nearest Neighbours



## Gaussian Naïve Bayes





## Multilayer Perceptron

