# Data Science with Machine Learning

# COMP4030

# Coursework 2024 CW2 Brief

| | | | |
|---|---|---|---|
| **Assessment Name** | Coursework 2 – Data Science Study | **Weight** | 75% |
| **Description and Deliverable(s)** | This assignment requires you to work in groups of three.<br>You will need to analyse a data set using all the data science steps you have learnt to create and compare your trained models.<br><br>You will write your work up as a joint academic paper, comparing and analysing your results of the data analysis and modelling pathway (6 to 8 pages including references and diagrams) as stated in this coursework specification.<br><br>The joint paper should be submitted as a PDF document, using the IEEE template for formatting.<br><br>The code should be submitted as a single Jupyter Notebook with clear comments showing attribution of each student for each section.<br><br>You will need to provide a peer assessment of the members of your group as part of an individual submission on Moodle. | | |
| **Release Date** | Tuesday 6th February 2024 | | |
| **Submission Date** | Thursday 9th May 2024 by 3pm | | |
| **Late Policy**<br>**(University of Nottingham default will apply, if blank)** | Work submitted after the deadline will be subject to a penalty of 5 marks (the standard 5% absolute) for each late working day out of the total 100 marks.<br><br>Due to the group nature of the project, ECs will need to be managed carefully. The team will need to submit their report at the set time, and the student with ECs will be able to send a revised version with a list of amendments and additions to show their work. Respective contributions will be taken into account during the marking.<br><br>Late submission deadline is Tuesday 16th May 2024 3pm. Submissions after this date will only be accepted through the extenuating circumstances process. | | |
| **Feedback Mechanism and Date** | Written feedback in Moodle on the 6th of June 2024 | | |

## Instructions

For this coursework assignment you will need be required to work in groups of three to analyse a data set (select one from the datasets provided or one you create as described) using all the data science steps you have learnt to create and compare your machine learning models.

You will write your work up as a joint academic paper with your coursework partners, comparing and analysing your results from the different stages of the data preprocessing, analysis, and modelling pathway.

You will need to present your paper in an IEEE format using a template from here: https://www.ieee.org/conferences/publishing/templates.html

Your paper should be between 6 to 8 pages (including tables, diagrams, and references as appropriate) and submitted as a PDF. The diagrams table and diagrams should add value to the writing. Diagrams are preferrable to tables.

Your paper should be organised into the following parts:

1. Title and Abstract (3%)
2. Introduction to the data set and research question(s) (5%)
3. Literature Review – covering a few key methods adopted by other researchers who used this or a similar dataset (5%)
4. Methodology – including a justification for your selected approaches for data analysis and pre-processing and data classification. (10%)
5. Results from each of the stages – data analysis, pre-processing and classification (20%) Please note that we expect to see a comparison of multiple approaches to solving the issue from different partners in the team.
6. Discussion - comparing and critiquing each other's results and also with other results from previous research on the dataset as noted in your literature review (25%)
7. Conclusions and recommendation for future research (10%)
8. References (2%)
9. Contributions – Please use the relevant sections from the Contributor Roles Taxonomy https://www.elsevier.com/en-gb/researcher/author/policies-and-guidelines/credit-author-statement

### Code Submission

Please include all your code as a single Jupyter Notebook with clear comments showing attribution of each student for each section.

We should be able to run this to generate your results (20% = each person in the group will be marked individually on this) in addition to the paper.

The ultimate aim of this coursework is to give you first-hand experience on working with a relatively large and real data set, your code need to reflect your learning for the entire process, from the first stages of data science: data preparation and pre-processing, exploratory data analysis, to the later stages of knowledge extraction and machine learning.

## Assessment Criteria

The total marks for CW2 will be out of 100 and scaled to represent the 75% weighting.

The main assessment criteria for the paper are:

| Section | Weighting % | Criteria |
|---|---|---|
| Title and Abstract | 3 | Are the title and abstract appropriately reflective of the content of the paper? |
| Introduction to the data set and research question(s) | 5 | Is there a statistical description that adequately highlights and summarises the key aspects of the dataset and are the research questions appropriate to the context of the dataset? |
| Literature Review – covering a few key methods adopted by other researchers who used similar datasets | 5 | Have relevant papers been discussed and their approaches and results succinctly described? |
| Methodology – including a justification for your selected approaches for data analysis and | 10 | Have appropriate approaches for each stage been selected? Have the selected approaches been clearly discussed and justified? Are they appropriate to the problem at hand? |

| | | |
|---|---|---|
| pre-processing and data modelling/classification. | | |
| Results from the different approaches applied at each of the stages – data analysis, pre-processing and modelling/classification | 20 | Were the techniques applied correctly? Have the results from alternative approaches been included at the different stages in an attempt to find the one that worked best? Have suitable diagrammatic representations of the results been included? |
| Discussion - comparing and critique the results | 25 | Have the findings been interpreted in an appropriate manner? Have the results from different approaches been compared in a critical manner? |
| Conclusions and recommendation for future research | 10 | Is there is a good summary of the work? Is there consideration of the shortcomings of the work? Are there any suggestions regarding how the techniques could be further combined in new and interesting ways? |
| References | 2 | Have appropriate references been included and cited correctly? |
| Python code (individually marked) | 20 | Is the code well commented and easy to follow? Is it consistent (i.e. consistent names for variables, functions, etc.)? does it use informative names for variables and functions? Are all the steps clearly marked up?<br>Were the data wrangling and pre-processing approaches for this dataset appropriate? Is there evidence of hyper-parameter tuning? Does the code it give the results as stated in the paper? |

## Module study expectations

| Activity | Per Week | Total Hours |
|---|---|---|
| **Lecture** – delivery key material | 2 × 12 | 24 |
| **Lab sessions** | 2 × 12 | 24 |
| **Self-study** – review lecture content and read associated background materials | 6 × 12 | 72 |
| **Coursework (75%)** | | **60** |
| **Lab submission Preparation (25%)** | 6.7 x 3 | 20 |
| **Total (20 credits)** | | **200** |

## References:

Jain, A., Bhandari, N.S. and Jain, N., 2018, February. Essential elements of writing a research/review paper for conference/journals. In *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)* (pp. 131-136). IEEE. (Paper on Moodle)

**Other resources for this coursework assignment:**

Reading list on Moodle

Materials covered and referenced in the lectures and lab sessions.

# Appendix 1 Datasets – More details will be provided for each dataset on Moodle

You can choose to work on any **one** of the following datasets:

## 1. Hand Gesture Recognition Data Set – You will need to collect this data yourself

The data that you will need for conducting hand gesture recognition needs to be collected by you and the other members of your group using https://phyphox.org/ app on your smartphone. You will find details of the data collection and expected classes/categories summarised below. We have recommended four different classes of gesture. You will need to perform these holding your phone in your hand. Note there will be variability in the dataset as you probably won't repeat the gesture exactly in the same way each time, and other members of your team might do the gesture a little differently. You might also need to down sample the accelerometer data. Can you create a classification model that can recognise the different gestures?

**Data collection procedure**

The aim of this Please download this app to your smartphone https://phyphox.org/

**You will need to collect gesture data - 4 classes/ categories** (note, you might want to have a general movement artefact category too)

1. Moving your phone in a circle
2. Waving
3. Gesturing "come here"
4. Gesturing "go away"

For recording the gesture data from the phyphox app, please use Acceleration (without g).

Do each gesture continuously for 15 repetitions (without stopping). Make 8 to 10 sets (files/recordings) for each gesture: Circle, Wave, Come Here, Go Away.

Each student should create a data set and then you can mix them up. Consider how you will split the data into train and test for creating your models. You will be provided with a set of gestures made by the teaching team for testing, against which you can evaluate your model.

**Features for conducting the classification**

Prior to performing classification, it is advised that you calculate some features from the raw accelerometer data signals that you collect. In order to do this, you will need to use experiment with different sized sliding time-windows, different overlaps for the sliding windows, and calculate features that describe the signal in each window. You will need to decide how many repetitions of each gesture you will use to indicate the gesture (2 or 3 repetitions) – this will help you determine the number of seconds for your sliding time-window.

You will be provided with material on Moodle as part of the lab sessions to support this.

## 2. Pump it Up: Data Mining the Water Table

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

The data comes from the Taarifa waterpoints dashboard, which aggregates data from the Tanzania Ministry of Water. Taarifa is an open-source platform for the crowd sourced reporting and triaging of infrastructure related issues.

Can you predict which water pumps are faulty? Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, and which don't work at all? Predict one of three classes based on a number of variables about what

kind of pump is operating, when it was installed, and how it is managed. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania. Think of it as a bug tracker for the real world which helps to engage citizens with their local government.

## 3. DengAI: Predicting Disease Spread

https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

Can you predict local epidemics of dengue fever?

Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world. In mild cases, symptoms are similar to the flu: fever, rash, and muscle and joint pain. In severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death.

Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. Although the relationship to climate is complex, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide.

In recent years dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. These days many of the nearly half billion cases per year are occurring in Latin America. Your goal is to predict the total_cases label for each (city, year, weekofyear) in the test set. There are two cities, San Juan and Iquitos, with test data for each city spanning 5 and 3 years respectively.

## 4. Recognition of different surface terrains using a 6-axis IMU

People with a mobility disability using an assistive walking device, such as a rollator, can experience a lot of difficulty, specially if the terrain is rough or uneven. This might cause them to trip and fall. In order to provide AI based assistance, the first step could be to detect what the surface they are walking/pushing the rollator on. One way in which this might be possible is through the use of inertial measurement units (IMU) which can measure acceleration and direction of movement.

For this study you will use accelerometer and gyroscope data from a IMU (an Axivity sensor - https://axivity.com/product/ax6 ) attached to a rollator (a walking assistance device with wheels) in the position shown in Figure 1. The data will consist of 3 axis of accelerometer reading and 3 axis of gyroscope readings, relating to movements (jitters) generated from walking with the rollator on different surfaces (Grass, Tarmac, loose stones, concrete with ridges and gaps)

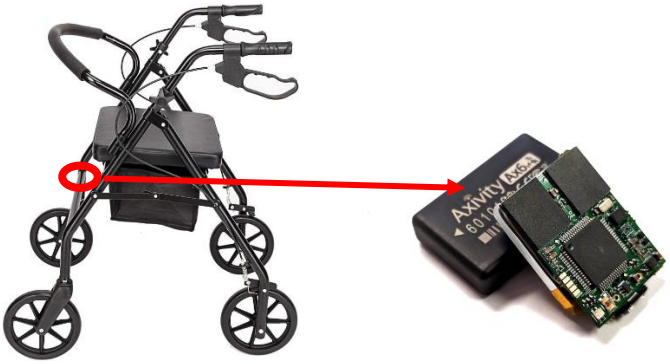

Figure 1. Rollator




Figure 2. Axivity 6-axis logging IMU

Your aim will be to build models based on the data from the Axivity sensors to determine which surface the rollator was being pushed.