CW2 Review

Assessment Name	Coursework 2 – Data Science Study	Weight	75%
Description and Deliverable(s)	This assignment requires you to work in groups of three. You will need to analyse a data set using all the data science steps you have learnt to create and compare your trained models. 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. The joint paper should be submitted as a PDF document, using the IEEE template for formatting. The code should be submitted as a single Jupyter Notebook with clear comments showing attribution of each student for each section. 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 (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 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		

CW2 Review

Your paper should be organised as follows:

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.

Assessment Criteria

Section	%	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 pre-processing and data modelling/classification.	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?

Assessment Criteria

Section	%	Criteria
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?

Assessment Criteria

Section	%	Criteria
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? Were the data wrangling and pre-processing approaches for this dataset appropriate? Is there evidence of hyper-parameter tuning? Does the code give the results as stated in the paper?

Exploratory Data Analysis

- Why? To get a feel of the data, understand it and identify patterns, trends and relationships between variables
- EDA should help you identify missing data, outliers and distributions of variables.
- Based on your EDA you would be in a better position to know which would be the appropriate pre-processing methods to use

- How?
  - Summary stats, visualisation, correlation analysis, outlier detection

Pre-processing

- Why? To address the issues identified in the EDA
- How?
  - Feature engineering/transformation – create new features based on domain knowledge, or discretise continuous variables etc
  - Impute missing data
  - Scale or normalise variables
  - Encode categorical variables

Approach – depending on your RQ(s)

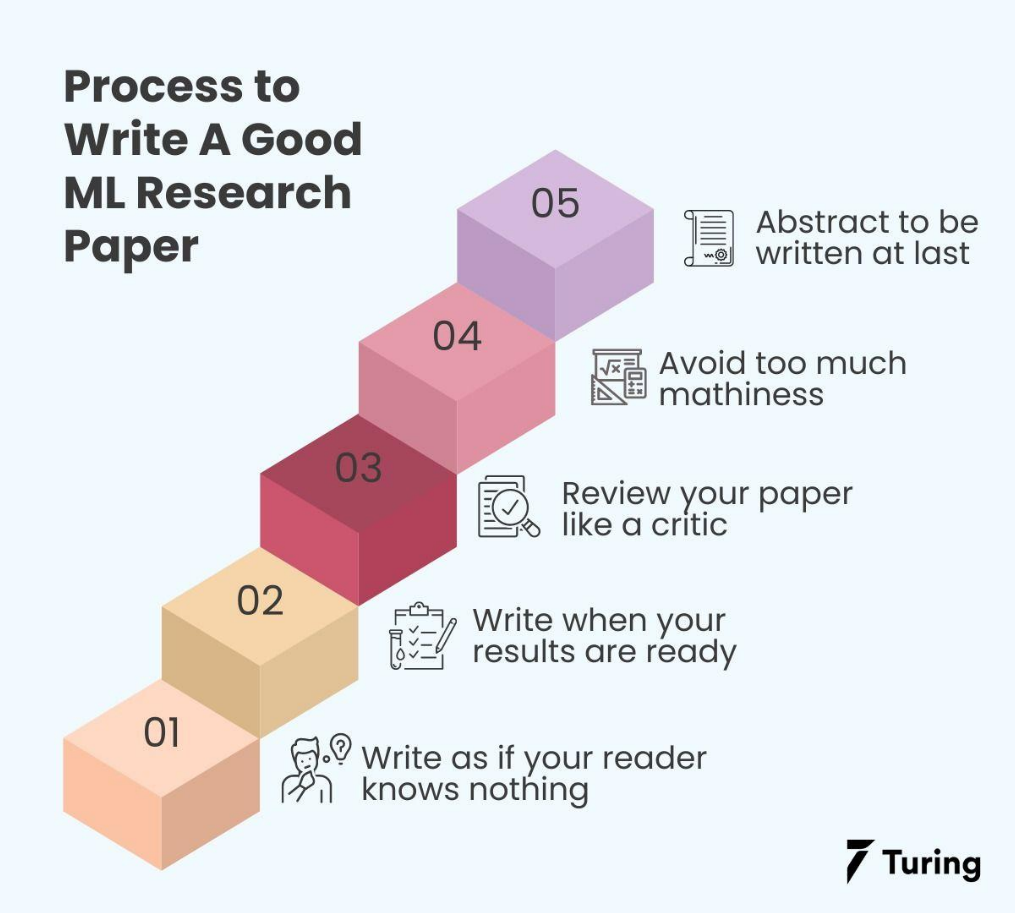
- Review research papers which use similar datasets
    - What ML methods were used?
    - Which proved to be most successful?
- Select suitable methods
- Evaluate performance (comparison metrics + train/test datasets)
- Go back to EDA and Pre-processing

- Compare and Critique

Writing up your work

- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8485210&tag=1 Essential Elements of Writing a Research/Review Paper for Conference/Journals

- https://www.turing.com/kb/how-to-write-research-paper-in-machine-learning-area


Process to Write A Good ML Research Paper
05
Abstract to be written at last
04
Avoid too much mathiness
03
Review your paper like a critic
02
Write when your results are ready
01
Write as if your reader knows nothing
Turing

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?

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

Data collection procedure - 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)
  - Moving your phone in a circle
  - Waving
  - Gesturing “come here”
  - Gesturing “go away”
- Do each gesture continuously for 15 repetitions (without stopping). Make 8 to 10 sets (files/recordings) for each gesture

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

- 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.

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. You can also use raw data if you use LSTM
- To calculate features, 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.

Pre-processing accelerometer data

Removing Movement Artifacts:

- To remove artifacts from the start and end of each gesture recording, you can employ heuristic methods such as trimming the first and last few seconds of data or using a threshold-based method where you only start considering data once the acceleration exceeds a certain threshold, indicating the start of a gesture.

Noise Reduction

- Apply a low-pass filter (e.g., Butterworth filter) to smooth the data and reduce high-frequency noise.

Pre-processing accelerometer data

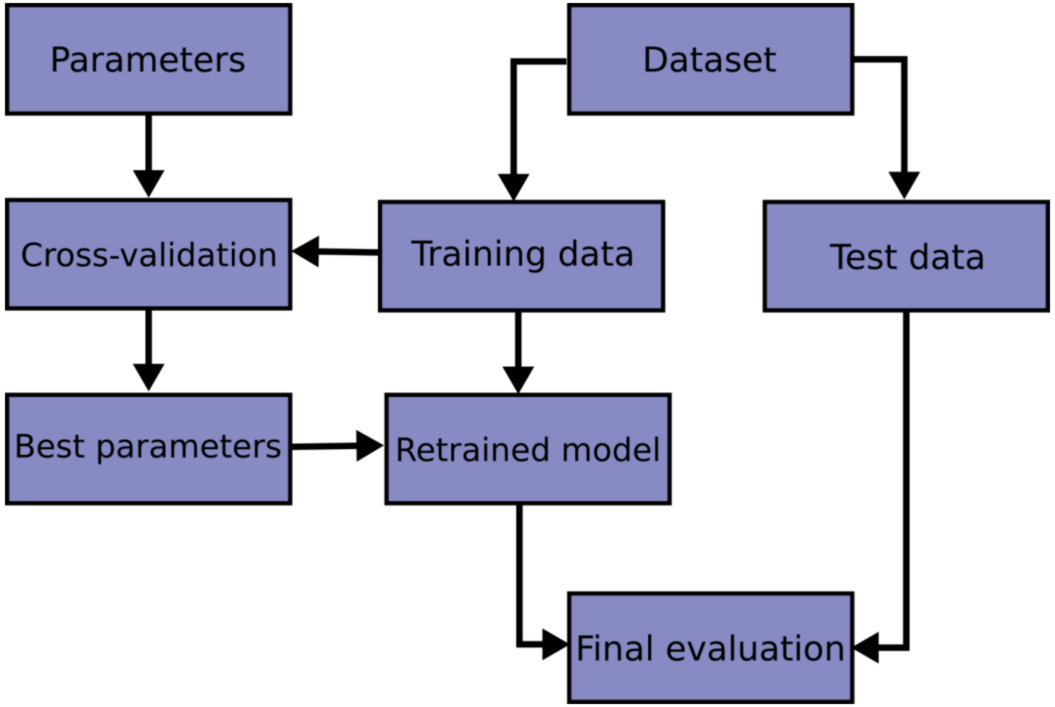
Feature Engineering

- Sliding Window Technique: Divide the time series data into overlapping windows (e.g., 2 seconds long with 50% overlap) to capture the temporal aspects of gestures.
- Calculate Descriptive Statistics: For each window, calculate features such as mean, standard deviation, min, max, skewness, and kurtosis for each axis of acceleration.
- Time-domain Features: Beyond basic statistics, calculate signal magnitude area (SMA), signal vector magnitude (SVM), and zero-crossing rate.
- Frequency-domain Features: Apply a Fast Fourier Transform (FFT) to each window and extract features such as the energy in specific frequency bands and the dominant frequency, which can help differentiate between different types of gestures.

Prepping for training

You will probably want to create a Combined Dataset

- Labelling: Ensure each gesture window is labelled appropriately based on the gesture it represents.
- Combining Data: Merge all the individual files into a single dataset with labels. This dataset should have the calculated features as columns and a label column indicating the gesture type.
- Splitting the Data
    - Train-Test Split: Use a stratified split to ensure that both the training and test sets have a proportional representation of each gesture type. A typical split ratio is 70% for training and 30% for testing. You should probably do 5-fold cross-validation



https://scikit-learn.org/stable/modules/cross_validation.html

Maybe start with using K-means clustering and Visualisation

- **Standardisation**: Before clustering, standardise the features to have zero mean and unit variance to ensure that all features contribute equally to the distance calculations.
- **K-means Clustering**: Perform k-means clustering on the training set to identify clusters within the data. The number of clusters k could be set to the number of gesture types, although additional analysis (e.g., the elbow method) might be needed to find the optimal k. This is part of exploratory data analysis.
- **Visualisation**: Use dimensionality reduction techniques (e.g., PCA, t-SNE) to visualize the clusters in two or three dimensions. Each point represents a window of accelerometer data, and its colour represents the cluster it belongs to.
- **Evaluate Clusters**: Analyse how well the clusters correspond to different gestures. Adjust preprocessing, feature engineering, or clustering parameters based on results.

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)

- The data consists 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)

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

FAQs on the CW2

- **From what we understand from the brief is that we should use different approaches at each stage e.g. Data Analysis/Pre-processing/Classification. Then throughout the paper we would compare and contrast our methods. However, we're just slightly confused at the end result.**

Our idea of stages:

1. Analysis – Should this be similar because it's data exploration.
2. Pre-processing – Each person uses different approaches – e.g. feature scaling, PCA, simple imputation and/or a combination of these
3. Classification – Each person tries a different ML approach - 1 uses random forest and person 2 uses decision trees, 3 uses ANNs

- Is this how it should be?

FAQs on the CW2

- Firstly, how many research questions should we come up with? And how different should they be?
- Secondly, say for example that me and my teammate decide that when pre-processing our data we decide that one person will deal with missing values by doing nothing and the other will impute using the mean. When it comes to marking the coursework, won't it look like the person that didn't do anything to deal with missing values did less work, and thus will receive less marks?

FAQs on the CW2

1. Does the Data Analysis stage refer to the Exploratory Data Analysis process? If so, I am slightly confused as to how my team-mates and I would come up with separate approaches for the stage

2. During the actual pre-processing stage, is it fine to use the same approaches as teammates for certain things that we agree on (for example, manually fixing data entry errors using common sense or label encoding for binary attributes) and only trying different methods where multiple methods have validity (imputation and encoding techniques)?

3. Are we expected to compare our pre-processing approaches and choose the best ones to carry forward to the modelling phase as one dataset, or are we meant to justify why each approach has been chosen individually and each use our own datasets in the modelling phase?

FAQs on the CW2

We are supposed to do the work together but with our individual findings, but kindly elaborate a bit more on the points below:

1. How are we supposed to adapt different approaches for pre-processing ? Do you want us to implement different solutions? is it ok if one finds out about the outliers and the other duplicates?
2. How many models are we supposed to implement? We have mutually decided to implement one each to get perfect results and appropriate findings .
3. How do you want us to write the report, what type of language and format should we use.
4. How do you want the Jupyter notebook to be , should we indicate who did what in the notebook as well?

FAQs on the CW2

1. Are the research questions the only topics we should focus in the coursework?

2. Do we need to focus on just one topic in the research question or can I focus on several ones?

FAQs on the CW2

- Should me and my team mates have different methods in each stage?
- Because I am not sure that if it make sense to have each stage to be in different approaches, as the results may not be fairly compared.
- We are currently thinking about using the same approaches in each stages except the feature selection part, so that we can answer the research questions that we newly came up with.