MATH42815: Machine Learning- Assignment 2

1. Submission Information

Key information:

- **Submission deadline:** Monday February 12th 2024 at midday (12pm)
- **Submission format:** One PDF file containing the main report and appendix, and one supplementary file in R or ipynb format.
- **Submission location:** Blackboard Ultra
    - **Report submission link: Turnitin - Assignment 2**
    - **Supplement submission link: Ultra - Assignment 2 - Supplement**

Please note that:

- Support is available during office hours, extended office hours will be available, see Ultra for information. You can also ask questions during workshops.
- You may send questions about the assignment by email to eimear.dunne@durham.ac.uk or c.c.d.s.caiado@durham.ac.uk . While we can support you by answering questions on machine learning methods and helping with your code, we will not be able to provide feedback on whether or not your model or conclusions are correct.
- The report should not exceed **3000 words**. You may include figures and tables in your main report and they will not count towards your word count. However, all figures and tables must be suitably captioned and referenced within the text. You may include an Appendix, which does not count towards the word count, to include supplementary tabular and graphical output and other relevant information. A further supplementary file must be submitted with relevant R code.
- Further details on formatting can be found in the end of this brief.

2. Assignment Brief

The assignment is worth 63% of the overall mark for the module.

Your work should be presented as a coherent report, giving consideration to the tasks and marking scheme detailed in Sections 2.2.1 to 2.2.4. You do not need to comprehensively describe everything you have done to explore and model the data. However, you should provide a narrative which details and justifies the relevant features of your approach, in addition to reporting and interpreting your results in the context of the problem you are addressing.

There will also be marks for the academic writing, structuring and presentation of this report; see Section 2.2.5.

2.1. Data

In this assignment, you will choose **exactly one** of the following tasks:

a. **Wine Quality** – model **wine quality** based on physicochemical tests using **both** the red and white wine datasets. The data is available from UC Irvine Machine Learning Repository - Wine Quality

b. **Abalone Age** – model the **age** of abalone based on a series of measured and observed attributes. The data is available from UC Irvine Machine Learning Repository - Abalone.

c. **Pokemon** – model Pokemon **type** based on their stats and other relevant attributes. The dataset contains 802 Pokemon from the first Seven Generations. Some Pokemon may have more than one type, you may choose to model just the first type that appears in the dataset. The data is available via Kaggle – The Complete Pokemon Dataset.

d. **Temperature** – model the **oral temperature** of patients based on patient attributes, ambient measurements, thermal image readings, and other relevant measurements. The data is available from UC Irvine Machine Learning Repository – Infrared Thermography Temperature.

- All files can be accessed and downloaded from the links provided. Copies of the data files can also be found at: Github - MLRepo and in Ultra under **Assignment 2 – Datasets – [dataset name]**.
- You should still read the information on the links provided above as it will prove useful for you to understand the problem you are tackling.
- Not all datasets are in a nice format. Check your data before you load it to R. You may need to merge multiple datasets in some cases.

2.2. Report organization and Marking Scheme

Your goal is to model the variable highlighted in **one** of the four tasks above (**Section 2.1**).

It is part of the assignment to identify suitable modelling approaches. These are real datasets so there is no “correct” answer or “best” modelling approach.

You will be assessed on your ability to address the problem posed, present the data, justify your modelling choices, and explain your results. The sections below indicate the components your report should include and the number of marks attributed to each.

2.2.0. Title

Your title must start with the relevant short name for the problem you have chosen: **Wine Quality**, **Abalone Age**, **Pokemon,** or **Temperature**. The rest of the title is up to you. You may include a subtitle if you wish.

2.2.1. Introduction (5 marks)

In this section you should:

a. briefly describe the problem you have chosen,

b. introduce the dataset and relevant features, and

c. list the modelling approaches you will use.

2.2.2. Data Cleaning and Exploratory Data Analysis (15 marks)

Before starting any analysis, you should clean and prepare your data:

- Verify that all variables have been loaded with the type you expect to use in your models (e.g. factors, double, integer);
- Check for typos and missing values, outline the process you have taken to deal with the impurities in your data.

Consider some exploratory data analysis. For example:

- how might you summarise the data graphically and numerically?
- what does this tell you about the relationships between the response variable and predictor variables and about the relationships between predictor variables?

Remember to set your discussion in the context of the problem you have chosen in **2.1**.

2.2.3. Modelling (40 marks)

You will make use of supervised machine learning techniques to address the problem you have chosen. You must choose exactly **one** option from each of the two lists below to model the response variable of your chosen dataset:

List 1 (You must choose **one** option**):**

- **Option 1.1:** Ridge, lasso, or elastic net regression with suitable values for tuning parameters;
- **Option 1.2:** Classification and Regression Trees (CART) with appropriate pruning;

List 2 (You must choose **one** option**):**

- **Option 2.1:** Creation of a suitable classification or regression model using tree bagging or random forests;
- **Option 2.2:** Training a neural network with an appropriate number of layers to complete a classification or regression task.

Remember to provide an overview of each modelling technique and, where appropriate, a consideration of any modelling assumptions you are making. In each case, discuss what your results show.

2.2.4. Model Comparison (20 marks)

Compare the performance of your models using an appropriate approach to validation. Remember to provide an overview of the main ideas underpinning your model comparison.

2.2.5. Results and Conclusion (10 marks)

Write a short reflection on your results and how they may be used to address the problem you have chosen. Comment on the limitations of the data, and the approaches used. Outline one or two things that you would recommend as next steps; this can include data collection, modelling approaches, or extensions of the problem.

2.2.6. Report Writing and Presentation (10 marks)

You should present your work in the form of a report. This should be well structured, written and presented. The report:

- should include a Title as described in **2.2.0**;
- must have sections following the headings in Sections **2.2.1** to **2.2.5**;
- contain appropriately displayed and captioned graphs and tables;
- must not exceed **3,000 words** excluding figures, tables, captions, bibliography, and the appendix.

Appendix: You may include an Appendix, which does not count towards the word count, to include supplementary tabular and graphical output for example.

Supplement: You must upload relevant R code used for the tuning, fitting, and validation of your models to the **Assignment 2 – Supplement** submission area. You do not need to include all code you have produced. All code should be appropriately commented. All packages used must be clearly listed at the beginning of the file.

The accepted file types for the supplement are: Jupyter notebooks (.ipynb), and R files (.R). You do not need to upload datasets, you can link directly to the Github provided or, given that you have included an appropriate process for cleaning, the filenames from the original datasets.

Bibliography: You should include a short bibliography (APA style recommended). All items in the bibliography should be appropriately referenced/cited within the text. The data sources used should be part of the bibliography.

Formatting:

- **Font size:** minimum 11pt in headings and the main text; minimum 8pt in captions, figures, tables, and footnotes;
- **Font type:** Arial or other similar sans serif font. Do not use Arial Narrow or other fonts with compressed kerning. Do not use multiple font types unless necessary.
- **Spacing:** 1.5 spacing
- **Footnotes:** use them only when necessary.

File format(s):

- Submit your report including the appendix as a single PDF file.
- Submit your supplement containing your R code as a single .R file or .ipynb.

3. Marking Criteria

The marking criteria below will be used. The marks will be scaled to each section with the relevant parts of the criteria applied to each.

- The final mark for the report will be given on a scale of 0 to 100.
- For the modelling section, failure to use exactly one approach from each list as instructed in 2.2.3 will result on a penalty of 10 points.
- Marks in the range 80-100 are reserved for exceptional work.

Mark	Criteria
80-100	The report is exemplary, providing clear evidence of a complete grasp of both the statistical methods employed, and their interpretation in the given context. The report is exceptionally well-designed, concisely offering relevant information and strong commentary at all points. The English used is faultless.
70-79	The report is excellent, providing strong evidence of a complete grasp of the machine learning techniques employed, along with a thorough understanding of how to interpret them in the given context. The report is very well-designed, frequently offering relevant information and strong commentary. The English used is very strong.
60-69	The report is good, providing evidence of a strong grasp of the statistical methods employed, along with some understanding of how to interpret them in the given context. The report is well-designed, offering relevant information and satisfactory commentary. The English used is good.
50-59	The report is acceptable, and provides evidence of a reasonable grasp of the machine learning techniques employed, although with limited evidence of an ability to interpret them in context. The report is acceptable in design, though some information is not relevant, or is inappropriately placed. There are some flaws in the English used.
40-49	The report contains insufficient evidence of a reasonable grasp of the machine learning techniques employed, although there is some evidence present. The context of the data is under-served or presented inaccurately. The report is flawed in design, with limited examples of relevant information. The English used is flawed, and sometimes poor.
30-39	The report is unacceptable, containing little evidence of a reasonable grasp of the machine learning techniques employed. The context of the data is under-served or presented inaccurately. The report is badly flawed in design, difficult to read and with limited examples of relevant information and numerous errors. The English used is often poor.
20-29	The report is unacceptable, containing little evidence of a reasonable grasp of the machine learning techniques employed. The context of the data is ignored or presented with major errors. The report is extremely badly flawed in design, extremely difficult to read and with almost no examples of relevant information and numerous serious errors. The English used is very poor.
0-19	The report is completely unacceptable, containing no evidence of any grasp of the machine learning techniques employed. The context of the data is ignored or presented without any accuracy. The report is exceptionally badly flawed in design, almost impossible to read and with no examples of relevant information, and with numerous serious errors. The English used is essentially unreadable.

4. Academic Misconduct

The University policy on Academic Misconduct can be found in the Learning and Teaching Handbook 6.2.4 and in the Avoiding Plagiarism mandatory training (Oracle Learn) that you have completed.

5. Use of Generative AI tools

The rules for the use of generative AI tools in assignments submitted to the department of Mathematical Sciences are as follows:

5.1. The use of generative AI tools – including but not limited to ChatGPT, GrammarlyGO - to **generate** the content of work submitted for assessment is **prohibited** on the basis that it is **not** the student's own work.

5.2. Standard scholarly tools that use AI to check grammar and spelling or manage/suggest references are not affected and can continue to be used.

5.3. The use of AI tools in the writing process is only permissible when restricted to improving the readability and language of submitted work **if**:

- **a.** the tool provides suggestions for modifying work already written by the student;
- **b.** the submitted work includes a section in the Appendix entitled 'Declaration of use of AI tools in the writing process'. In that section, a list of **every** use of the AI tool is given, indicating all affected page(s), with line numbers or ranges, and specific details of the tool that was used, and the reason for using the tool.

5.4. Cases of suspected misuse of AI tools will be dealt with by the same procedures as suspected plagiarism (LTH 6.2.4.1 and related pages).