Data Science with Machine Learning

COMP4030

Coursework 2025 Coursework Brief

Assessment Name	Coursework – 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	Wednesday 12th February 2025		
Submission Date	Friday 2nd May 2025 by 3:00 PM		
Late Policy **(University of Nottingham default will apply, if blank)**	As per University of Nottingham's default late policy, work submitted after the deadline will be subject to a penalty of 5 marks for each late working day out of the total 100 marks. 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. Late submission deadline is Friday 9th May 2024 3:00 PM. Submissions after this date will only be accepted through the extenuating circumstances process.		
Feedback Mechanism and Date	We will aim to provide feedback as early as possible in Moodle (typically within 15 working days).		

1. Introduction

This lengthy document describes the main assessment of COMP4030, which is a **group coursework** weighted at 75% of the module grade.

2. Instructions

For this coursework assignment you will need be required to work in groups of three to conduct a data science project from start to finish. You will therefore need to:

(1) Find a dataset
(2) Find a research problem/research question to investigate/answer
(3) Build a machine learning model which aims to solve this problem
(4) Conduct one or more experiments to validate your findings
(5) Analyse the results of your experiment(s)
(6) Write up a joint academic paper to report on those findings, using the IEEE format (template available https://www.ieee.org/conferences/publishing/templates.html)

As part of your coursework, you will need to submit three deliverables.

(a) **A joint academic paper**, from 6 to 8 pages (including references).
(b) **A joint code submission**, in the form of a single Jupyter notebook file, and adhering to reproducibility guidelines (provided below).
(c) A document containing two things**:**
    a. **Part 1. An individual reflection report** of no more than 300 words (not including peer assessment), explaining your role in the project. This report will be marked for completion only and will be used to disentangle people's contributions in case of conflict.
    b. **Part 2. A peer assessment of your teammates**, across criteria that will be detailed in the section below.

There are more details about each of the deliverable in the following subsections.

2.1. The main report

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 and not just be used for decoration. Your paper should be organised into the following sections:

1. **Title**
2. **Abstract**
3. **Introduction** to the dataset and research question(s) and exploratory analysis
4. **Brief literature review**
5. **Methodology** – including an evidence-based justification for your approaches
6. **Results** – data analysis, pre-processing and prediction
7. **Discussion** – discussing results both within the study and with the literature
8. **Conclusion**s and recommendation for future research
9. **References**
10. **CRediT statement** – Please use the relevant sections from the Contributor Roles Taxonomy https://www.elsevier.com/en-gb/researcher/author/policies-and-guidelines/credit-author-statement

2.2. Source code of your work

Your code should be integrated as a single Jupyter Notebook with clear text sections showing attribution of each student for each section. We should be able to run this to generate your results in addition to the paper, so extra care should be given to reproducibility (reproducibility guidelines will be mentioned below and elaborated upon in the module). For example, make sure that all the packages you are using are declared early in your script. Testing will be done using whatever version of Python is available in A32 and/or the Virtual Desktop.

The ultimate aim of this coursework is to give you first-hand experience on working with a relatively real data set, your code needs 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, learning, and prediction.

2.3. Personal reflection

The reflection, worth 5%, will be **focused on completeness**[1]. To get full marks, it will need to contain **all** of the following three components:

- **A narrative description of the project** – what did your group do, how did you go about it, what did you try that may not have worked, etc.
- **A description of your contributions to the project** (even if they did not end up in the final report), with evidence provided in a zip folder.
- **An account of your usage of generative AI.** Explain whether you used any form of generative AI and in what way.
- **Peer assessment of your team** – an assessment of the participation of the members of your team across multiple criteria (overall effort, attendance to meetings, team spirit)

A template document will be given on Moodle later in the term.

[1] This means that it will be marked not based on its quality (how well it is written) but simply based on whether each element that we expect to see is present. This is partially to make your job easier, and partially to disincentivise the use of external tools such as LLMs to write it for you – use your own words instead.

3. Marking scheme

3.1. Report and code (95%)

This table contains three columns: Section, which refers to which section we expect to see this information in; Weight, which refers to the overall weight of the criteria; and Critical questions, which are not marking criteria, but the guidelines that the markers will need to bear in mind when evaluating your submission.

Section	Weight	Critical questions
Introduction of the project (Title, Abstract, Introduction, Literature)	15%	Are the title and abstract appropriately reflective of the content of the paper? 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? Have relevant papers been discussed and their approaches and results succinctly described?
Main body (Methodology, Results, Discussion)	50%	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? Have appropriate references been included and cited correctly? 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? Have the findings been interpreted in an appropriate manner? Have the results from different approaches been compared in a critical manner?
Conclusions, recommendations, teamwork	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? Was the project run smoothly, tasks evenly allocated based on strengths and weaknesses?
Code	20%	Is the code well commented and easy to follow? Does it follow reproducibility guidelines? Does it run without errors? 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 it gives the results as stated in the paper?

3.2. Reflection (5%)

- Marked on a scale of 0 to 5

Non-submitted / severely incomplete (0/5)	**Mostly incomplete (1/5)**	**Needs work (2-3/5)**	**Mostly complete (4/5)**	**Complete (5/5)**
There was either no personal reflection submitted, or it is so lacking in detail that it gives no useful information to understand the project.	While there is an attempt at providing a personal reflection, it is missing most of the information that is needed and does not inform the reader of your project or your role within it.	The reflection addresses most points with enough details that we can mostly tell what you have done, but it is missing some needed information.	The reflection addresses most points with enough details that we can mostly tell what you have done. You might be missing some details on specific points.	The reflection addresses all points with enough details that we can clearly tell what you have done.

3.3. Example

For example, if you had an excellent introduction and exploratory analysis (75%), a good methodology, results, and discussion (60%), a very good conclusion section (70%), and an excellent code submission (80%), the final grade would be 75*0.15 + 60*0.50 + 70*0.10 + 80*0.20 = 64.25 out of 95 points. This would correspond to approximately 67% if rescaled to 100. This group grade can then be scaled up or down by the person grading you based on your relative contribution to the project. If for example you contributed more than the average person on your project, you might end up with a personal grade of 73%, or 69/95. If you submitted a complete reflective report on top of that (5/5), that would yield a final grade of 74% for coursework 2.

4. Coursework policies

4.1. Reproducibility

It should be relatively straightforward to rerun your experiments straight from your Python notebook. In order to do so, please use the reproducibility checklist (adapted and simplified from the AAAI conference instructions) to check your work:

- Any code required for pre-processing data is included (yes/partial/no)
- All source code required for conducting and analysing the experiments is included (yes/partial/no)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes/partial/no/NA)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes/partial/no)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes/partial/no)
- This paper states the number of algorithm runs used to compute each reported result. (yes/no)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes/no)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank test). (yes/partial/no)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes/partial/no/NA)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes/partial/no/NA)

4.2. Academic judgment, marking fairness, and regrading

While we made the best attempt at providing clear guidelines that will be used to mark your work across a set of objective criteria, it remains a fact that some subjective assessment will have to be made. It is therefore important to understand that grades are a matter of academic judgment, which cannot be questioned. If you notice an arithmetic error however (something was marked wrong which should not be), you can request your coursework to be regraded. Note however that regrades can sometimes end up with lower grades than initial grading, since the second marker might notice errors that your first marker did not.

In some (hopefully rare) cases, such as:

- A strong conflict in the statements contained in group members' personal reflections (e.g., multiple members claiming to be solely responsible for the same piece of work).
- Statements made in a personal reflection/report that are not reflected in the code.

Your group may be asked to join an online meeting with a member of the teaching team in order to explain your work and help them attribute your efforts fairly.

4.3. Effective group work

Group work is difficult, and not everybody likes it, but since data science and machine learning typically sit in the middle of domain expertise and software professionals, being able to work as part of a team is a core competency that you will need to develop. For that reason, a set of rules will have to be followed:

- Teams will have to be declared by a specific deadline using a form that will be posted on Moodle.
- Teams can be made of 3 to 4 students.
- Once declared, team composition cannot be changed.
- Multiple teams cannot collaborate on the same project.

4.4. Extenuating circumstances

Due to the group nature of the project, extenuating circumstances will be dealt with through a dual submission system. Students granted an extension will submit their individual contribution as a "delta file." This file will only include the work completed **after** the original deadline, and it will aim to show their additional effort and avoid any unfair advantage over the rest of their group.

4.5. Resits and first sits

Resits and first sits (which are resits without a penalty) take place in the month of August and will take the shape of a smaller version of this group project, intended to be done individually.

4.6. Plagiarism and generative AI

Generative AI, mostly in the form of large language models, have become embedded in several applications and hard to avoid. For that reason, here is our generative AI policy:

Allowed	**Not allowed**

• Where it's unavoidable, e.g., as part of search (e.g., embedded in search engine) • As part of assisted coding (code completion, bug finding)	• To write the report for you • To write the code for you

Our plagiarism policy is simple: **do not plagiarise**. Plagiarism cases (whether based on AI, other work, or your own, uncited prior work) will be sent to the academic misconduct committee. How to avoid plagiarism?

- Cite the work you're getting your inspiration from using a consistent referencing system (we recommend IEEE for the sake of simplicity, but we are not attached to it)
- Reference the code you're using. Preferably in the comment where you're using it, but potentially in a text cell that explains it.