MA429 – Algorithmic Techniques in Machine Learning

Summative Group Project Description

This summative group project is similar in nature to the earlier formative one. The difference is that here you have a wide choice of datasets, and you are expected to go into greater depth.

Dataset choice and options

- A list of some sources of data is posted on Moodle separately. You are also very welcome to suggest other datasets. It would be great to see choices beyond the well-trodden UCI Repository and such, tackling for example data from the UK government or the World Bank (see links provided). I really encourage you to look for data that interests you.
- In any case you will need my **approval for your dataset choice**. Please ask me by **Friday 5th December**, especially as discussion may be needed. I will try to reply quickly. Please include **key information about the dataset in your email** (e.g., what it is, dataset size), a **link to the dataset** in case I want to see more (don't send the data itself), and **briefly explain your interest in it**, and your thoughts (if you have any yet) on your approach to it. You may list two or three choices, in your preferred order, if you wish. Please use email subject line "**MA429 group X dataset choice**" and **copy your groupmates** on the email.
- You can simply email me on n.olver@lse.ac.uk. This will compromise anonymity, but I'm not sure that there is a realistic alternative. I will not be making any particular effort to record the link between dataset and group, other than keeping track of what datasets have been "claimed" so far. If you have a serious worry about this, you can email Rebecca Batey, and ask her to forward your request anonymously (and I will then send my approval, or not, back through her).

While there is not a specific numerical requirement, and it's not just a matter of size, I would expect to see at least thousands of instances and 20+ features. Watch out for datasets with many features, but where most of them are (probably) useless; if there are only a few features that are likely to be relevant and useful, that's an issue.

- In previous years, there have been some "popular" choices. I will consider allowing two groups to use the same dataset, but definitely not more than that.
- Students often ask if more challenging data or more sophisticated methods would lead to higher marks. Indeed, if you pick a challenging problem and make good progress on it (maybe even get good results on it), that will earn you more recognition than doing an equivalent job on an easy problem. Likewise, it will be recognised if you use more challenging methods like neural networks or (with reason) use methods we haven't covered, work in a domain like clustering that we spent less time on, or work with data that has special challenges (timeseries data, many missing or erroneous entries, etc).

However, you should choose a dataset you are happy with. Most projects use manageable data and the techniques we've spent the most time on, and they can be excellent when done well.

Structure

Similar to the formative except for the following.

- The title page must include the group member's exam candidate numbers (but not names or registration numbers, and not the group number).
- After the title page, your report must include an *executive summary* of at most one page. This should summarise the setting of the problem investigated and your main conclusions. It should focus on the problem domain and should not contain any technical detail. Emphasis on "executive"; imagine your target reader as someone very important with little time.
- You should have a table of contents after the executive summary.
- The main body of the report, *excluding* title page, executive summary, table of contents, bibliogrpahy and appendices, should not exceed **16 pages**. This is an upper bound, not a lower bound. There are no page limits for the appendices, but you should still only include things that are relevant and of interest.
- There is scope for a more substantial literature review, if appropriate. You are encouraged to look at any publications related to your dataset, including those linked at the dataset-hosting sites, using and citing them as appropriate.

Other than this, as before: after the main report and bibliography, you *must* include a generative AI statement; after this, optionally, appendices.

> **Consult the formative project description and remind yourself of the suggestions there on what to put in the report, how to structure it, etc.**

Generative AI policy and statement

The policy here is the same as the formative. I include it again, particularly to emphasise that you *must* include a generative AI statement, and that I will absolutely perform interviews if I have any concerns that the policy is not being adhered to.

I will allow the use of large language models such as ChatGPT, Copilot, Claude, Gemini, etc., **but** I expect responsible usage, and you are *required* to document your use of it.

You are ultimately responsible for the work you produce, and you need to be able to stand behind it. This means that:

- You must understand the code that is being used to obtain your results. If an LLM generated the code, you plugged it in, and got some (seemingly) plausible results, that's no good if you don't understand what's being done, why it's being done that way, what choices are implicit in the code being run, why the code is doing what you *think* it's doing, etc.

- You should be extremely careful about using LLMs for generating prose in your report. The moment you start using it to generate any content that you do not already understand well, things tend to go off the rails very quickly. Again, you must be able to stand behind what's written. For example: if the text discusses some existing literature in comparison to the project, and you've not actually looked at the article, and haven't understood how and why it relates to the project, that's a very serious problem.
- As discussed elsewhere in this project description, clarity is very very important. If the text is poorly targeted – if it blithely uses a whole lot of jargon, or tools from statistics or elsewhere that I, or another MA429 student is not likely to know – that's poor, mistargeted writing.

Your report *must* include a detailed statement regarding your use of LLMs (after the bibliography). This is irrespective of whether you have made any use of LLMs or not — of course, if you have not, then your statement will be very brief. If you have, then discuss in detail what you have used it for and how.

As discussed in the departmental statement on generative AI use, the department reserves the right to conduct interviews after the submission of any assessment, which may be random or targeted based on marking, in order to determine whether plagiarism has occurred. Uncited use of generative AI tools is considered plagiarism.

Team work

The groups will remain the same as with the formative project.

You can divide the work however you like, as long as it is fair. Since the whole group is judged on the one project, it would be wise for everyone to look over everything, as it progresses and at the end. Budget time for this.

I strongly recommend meeting in person (or by Zoom if that's not possible). It is generally recommended that any meeting is summarised in print right away, especially any "action items". I recommend email over other text communication, using your LSE email account and leaving a paper trail should it be needed for any reason.

Professionalism is expected. If there are issues that you are unable to resolve yourselves, you can discuss it with me. Our experience so far has been that most groups work well, and we expect it to continue so.

Contribution statement. You will each individually (not as a group) submit a form on Moodle explaining the role and level of contribution of each team member (including yourself). Be specific, without going into unnecessary detail. E.g., not: 'we all wrote some code'; what specific aspects did you code? What parts of the report were you primarily responsible for? Etc. Experience suggests it is best if you organise this by group member: for each person, say what they contributed.

This will be submitted using an online form in Moodle. I recommend writing this in a text editor or Word, and copying this into the form.

Marking

Same as the formative:

30 Context, modelling and originality: How well is the context of the problem, and the results, explained? As seen in the executive summary, introduction, literature review and conclusion.

Appropriate (but relevant!) exploratory data analysis and understanding of data, intelligent modelling choices, and appropriate evaluation metrics.

The extent to which the project goes beyond standard existing approaches, or takes a novel perspective.

40 Analysis: How well is the main work of analysing the data done? This includes pre-processing, choice of learning methods, correct use of methods, correct methodology (e.g., train/validate/test), parameter tuning, interpretation of results.

Credit for handling challenges (e.g., using more difficult methods).

20 Organisation and presentation: In the large: logical and structured discussion, clarity of expression. **Must be targeted correctly**; reports written as if the target reader is an expert in statistics, machine learning (beyond what is covered in MA429) or some specific domain, will score poorly.

In the small: layout, typesetting, table of contents, bibliography, clarity and legibility of figures, absence of typos – overall polish.

10 Code quality: Well-structured and easy-to-understand code; should be well-commented. It should be possible for me to reproduce your results from your code without too much difficulty.

Individual adjustment: This will take into account information contained in the contribution statements. In the hopefully most common case of (roughly) equal contribution, each team member will receive (roughly) the same mark. If there is a serious discrepancy about who did what in the contribution statements, I will interview group members individually.

In extreme cases where a team member has made no serious contribution, this will be treated the same as the lack of a serious attempt at an exam, and the student will have to perform an alternative, individual piece of work at a later date.

Submission & deadlines

- **Dataset choice** (see below) to me by **Friday 5th December 2025**. Some discussion may be required.
- **Project submission deadline: 18:00 Monday 30th March 2026.**
- **Extensions**, if needed, should be requested by submitting a completed extension-request form – https://info.lse.ac.uk/current-students/services/assets/documents/Extension-Request-Form.pdf – to the Exam board chair for your

MSc. For anonymity, I should not be copied, but you could pass along my email address for the chair's convenience. I cannot grant an extension informally.

Use only your **candidate examination numbers** in naming the files, and within the files, including code comments. In the contribution statement, refer to yourself and your teammates by examination number. *Your name and student ID number must not appear anywhere.* All submissions are to be made via Moodle links provided, as usual.

Each **group** should submit the following, just *one copy per group*:

- A softcopy of the report, in pdf format. Please name the submission with one group member's examination number, e.g., `54321 report.pdf`.
- An electronic appendix including your code, computational results, and any additional data you deem important, in a single zip file. Please name this by the same group member's examination number, e.g., `54321.zip`.

Each **individual** should submit:

- A completed contribution statement via the Moodle form. This should be a detailed account of what parts of the project you were responsible for, and what parts your teammates were responsible for. This deadline will be a few hours later – midnight on the same day as the project deadline.