UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
STA5073Z – Data Science for Industry 2025

Assignment 1
Neural networks

This project must be done in groups of three, four, or five
Due date 29 September 2025 12 pm (noon)

Problem background

Public avalanche forecasts are published daily during winter in many countries, including Scotland. These forecasts are made by a team of expert forecasters who combine daily field observations about the current snow conditions with weather and other information. In this project you will use a 15-year archive of avalanche forecasts produced by the Scottish Avalanche Information Service (`https://www.sais.gov.uk/`) to produce a predictive model of some aspect of avalanche forecasts.

You are provided with a dataset containing the following variables:

- *Date* = the date the forecast was made
- *Area* = one of six forecasting regions
- *FAH* = the forecast avalanche hazard for the following day
- *OAH* = the observed avalanche hazard on the following day (observation made the following day)
- *longitude:Incline*: position and topography at forecast location (predictor set 1)
- *Air.Temp:Summit.Wind.Speed* = weather in the vicinity of the forecast location at the time the forecast was made (predictor set 2)
- *Max.Temp.Grad:Snow.Temp* = results of a "snow pack test" of the integrity of snow at the forecast location (predictor set 3)

Assignment objectives and deliverables

Different groups will be assigned to different objectives. Your overall goal for this project will be one of the following:

Problem A Construct and evaluate a neural network model that predicts the forecasted avalanche hazard.

Problem B Construct and evaluate a neural network model that predicts the observed avalanche hazard.

Your group's work will be assessed in two ways:

1. a short scientific report (word limit: 4500) on this work that you will host on your own website. (one report/website per group; 80% of final mark)
2. a 20-minute group presentation targeted at management of the Scottish Avalanche Information Service. (20% of final mark)

Specific learning objectives are:

Analytical skills:

1. To be able to work with (i.e. read in, clean, manipulate) data.
2. To create appropriate test, validation, and training datasets and assess classifier performance correctly.
3. To be able to fit appropriate neural networks to the data.
4. To provide an informed assessment of which predictors are most important

Workflow skills:

1. To be able to work effectively in a group.
2. To be able to write up your work in the format of a short scientific paper. Please note the word limit of 4500.
3. To be able to host your paper on a GitHub Pages website, linked to your GitHub repository that you push to programmatically (i.e. not by dragging and dropping files).
4. To be able to work collaboratively on GitHub, for example with different group members working on the same file, using branches, using pull requests.
5. To demonstrate informed, creative use of a large language model such as ChatGPT to assist with the assignment and to critically assess its ability to do so (in terms of what did it do well or badly; prompt design, etc).
6. To produce and deliver an effective presentation appropriate to the target audience.

Submission guidelines

One member of each group must submit:

1. A link to a GitHub Pages website, where one page contains your scientific paper and another page contains your description of what you used the LLM for, how you used it, and a critical reflection on the performance of the LLM and what, if anything, you have learned about the use of LLMs for assisting with data science work such as this project. You may include a "landing" page and extra webpages for appendices and supplementary material if you have any.
2. A link to a GitHub repository from which your website was built. This should contain the html source files for your website as well as all Quarto markdown (.qmd) files that contain your written text and code.
3. On Amathuba, please submit three files, using your **STUDENT ID as the files' names** – e.g. `ABCXYZ001.txt`, `ABCXYZ001.html` and `ABCXYZ001.qmd`, where

(a) `ABCXYZ001.txt` contains the links to the GitHub Pages website and GitHub repository.

(b) `ABCXYZ001.html` is your rendered scientific paper in html format. This should be the same as what is on your webpage. Do not include appendices, etc.

(c) `ABCXYZ001.qmd` is the .qmd file used to generate the paper. Use 'embed-resources: true' in your YAML to generate a single .html file from the .qmd.

Please indicate all students in your group in the author section of .qmd YAML.

Note that I primarily use points 1 (written report) and 2 (code) to grade your work. Point 3 (the Amathuba submission) is to verify the website hasn't changed after the submission date, and for me to be able to rerun your analysis if need be.

Use of AI

Use of an LLM is encouraged and is an important part of the assignment. See the separate document on the course policy on LLM use for assignments for further details.

Scientific paper writing style

This is an important part of the assignment.

- Scientific papers have a distinctive structure and style. The best way to familiarise yourself with these is to read a few papers yourself and read some of the many online guides on paper writing (to be provided separately).
- The structure is usually: Abstract, Introduction, Literature review, Data and methods, Results, Discussion, Conclusion. Each of these sections has a clear purpose (see references Sometimes, especially in short papers, the literature review can be very short and is weaved into the introduction – this is probably appropriate for your assignment. Similarly the discussion and conclusion are often combined. Sometimes Data and Methods are separate sections.
- Your code should **not** be displayed in the final typeset document (use `echo = FALSE`).
- Reference any material you use, whether it be a blog post, a figure, code, etc.
- A reader should be able to understand all the steps you followed in your analysis, including any cleaning, to the point where they could replicate your analysis if they had to. Your work should be fully reproducible. A common mistake is to not provide enough detail, or to not provide those details in a concise way (i.e. identifying what is really essential). Things like whether you read in a csv file or a xls file or whether you reshaped the data from wide to long are not relevant, but things like if you deleted any observations, what you did with missing data, parameter settings for models etc are all relevant.
- Pay attention to the display and formatting of figures and tables. Use appropriate captions, titles, font sizes, etc. Don't just rely on whatever ggplot or base R output gives as a default. See published papers for examples of what is required.

Here are some useful resources. Note that these are aimed at people writing a paper to be submitted to a peer-reviewed journal for possible publication. We're not aiming quite this high, but the sources are still useful in telling you about the expected structure and content of a paper.

- https://jarisaramaki.fi/2017/04/28/why-can-writing-a-paper-be-such-a-pain/. The link is to the first of a series of blog posts; at the bottom of each post you'll see a link to the next one in the series. The whole series is something like 18 posts. I find the ones on the introduction (around part 4-6) the most useful, but it is all excellent.
- https://www.nature.com/scitable/topicpage/scientific-papers-13815490
- https://www.turing.com/kb/how-to-write-research-paper-in-machine-learning-area
- https://conservationbytes.com/2012/10/22/how-to-write-a-scientific-paper/
- https://spie.org/news/photonics-focus/janfeb-2020/how-to-write-a-scientific-paper?SSO=1
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3474301/
- https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/bes2.1258

Other points

- On group work: it is perfectly fine and probably desirable to divide the work between group members so that some members focus on certain parts of the project. Your report should include a declaration stating what tasks each person in the group was responsible for/-participated in. See http://journals.plos.org/plosone/s/submission-guidelines#loc-author-contributions and http://journals.plos.org/plosone/s/authorship#loc-author-contributionsfordetails. However, all group members must be familiar with all aspects of the project. If you are working on a topic (say sentiment analysis), you are responsible for making sure that the rest of the group understands that topic, and what you have done, to the extent that they could explain the work to someone else.
- Doing your project in R is **recommended** but Python submissions will be accepted. Your report and website must be rendered from a **Quarto Markdown** document.
- Anyone else should be able to run the code in your .qmd to completion. Use `set.seed()` to set a random seed so your final results do not change.
- You may not share any coding or write-up with any other group. Please sign the plagiarism declaration provided and append it to your submission.
