Coursework

Question 1 (40 pts)

Please download the `penguin.csv` dataset. This dataset has information about penguins of one of three types. You task is it explore the dataset and to predict the penguin type. The dataset is known as the Palmer Penguins and is found at:

`allisonhorst.github.io/palmerpenguins/`

This link contains information on how to cite the dataset.

You should consider how to visualize the data and which algorithms to try. Nothing you do will be completely successful, this coursework is not here to judge your final accuracy but the care you bring to your investigation. Here are some thing you should consider:

- The kind of algorithm to use, for example whether to classify, regress or cluster.
- The metric to use to measure the performance of the model.
- What sort of baseline to compare the model to.
- How to choose the hyperparameters of your model.

For good marks you should include some graphs that illustrate properties of the data and some exploration based on either unsupervised learning or regression, along with that you should compare two classification algorithms, both to each other and to a baseline model. The baseline model is just the performance you would get if you guessed without reference to anything but the sizes of the populations in each class. The algorithms you pick do not need to be unusual, for example $k$nn classification would be perfectly good, though, of course, for full marks this would include some consideration of how to pick k and how to measure the distance, even though, as you know, no approach to choosing k is ever going to be completely satisfactory. In addition, you should include either some exploratory regression or unsupervised learning; for regression you might regress two properties and examine whether the regression parameters are the same for each penguin type; unsupervised learning could use k-means, for example. You do not need to do both regression and unsupervised learning.

Thus, there are four elements expected:

1. A brief exploration of the data.
2. An unsupervised or regression approach.
3. A classification algorithm.
4. Another classification algorithm,

and, in the case of the (3) and (4), it is expected that you will compare the two classification approaches. More detail is provided by the marking scheme below.

You should make sure any assessment is not restricted to the data used in train models or decide on hyper-parameters: it is important to hold aside testing data. In your report you should explain your decisions. You code will not be marked for elegance, but it should run correctly; it is expected you will use Python, but any of Python, Julia or R is fine. Do not include screenshots of graphs, they should be imported directly; resize them to the correct size before importing them, if the labels are tiny the graphs will not be marked. Make sure figure captions are descriptive, it is better to have some overlap between figure captions and the main text than to have figure captions that are not reasonably self-contained.

As a rough guide to marking:

- Initial description of the data, including some graphs or other approaches to visualisation. 6 marks.
- Either unsupervised learning or regression. 6 marks.
- Two algorithms should be tested, if only one algorithm is included the 28 available marks will be halved.
- Overall presentation (3 marks), including use of appropriate sections, plots, diagrams, or tables to make your point. Do not include code snippets in the report. Instead, describe in words or equations what you are implementing. Format equations correctly.
- Suitable choice of algorithms (3 marks).
- Suitable choice of evaluation for algorithms (3 marks).

- Comparison with a suitable baseline (3 marks) and a justification for which baseline to use.
- A description of metaparameter selection (4 marks), if one algorithm has not metaparameter, then explain that and note why not and why this do or does not make it a better algorithm for these data.
- Describe and compare the results from your two classification algorithms, include a description of how you implemented the algorithms. (6 marks)
- There are some marks (6 marks) for something suprising and unusual. This is to allow us to give credit for exceptional work, but you should not expend too much time chasing these final marks, concentrate on the stardard requirements to get a good mark.

Question 2 (10 marks)

For two of these three types of ethical challenge or threat facing us in data science and AI:

1. The protection of data, of the people whose data they are and participants in any study.
2. Avoiding the amplification of biases and regressive values implicit in historic dataset.
3. The safety of AI systems and the possible of existential threats from machines.

describe what you think is a specific example that could arise or has arisen in the past; this could be a single example that relates to the two challenges you are considering, or it could be two examples, one for each challenge. Obviously the three broad types of challenge overlap, do not worry about the boundaries between these types, but do try to address different two different types of threat in your examples. Explain how the ethical problems could be addressed, or at least made more transparent.

Report

Your report should be no longer than five pages, excluding any references. It is expected that Question 2 would occupy about a fifth of this space; use an 11 or 12pt font and do not try tricks like expanding the margin to fit in more text, shorter is better than longer.

Your report must be submitted in pdf and should be prepared in LaTeX; overleaf is a good approach, but not required as long as LaTeX has been used[1]. As always when using LaTeX, give yourself over to defaults, our expectation of what a document should look like has been conditioned on LaTeX, so it is best not to try to override the look of the document. I have included a template but you need not use that.

Avoid code snippets in the report unless that feels like the best way to illustrate some subtle aspect of an algorithm; do always though consider a mathematical description if possible. You will be asked to submit code and it may be tested to make sure it works and matches your report. It will not, however, be marked in and of itself.

Submission

The deadline for report and code: 13h00 (GMT+1) on 22nd May, there will be a submission point on Blackboard under the "assessment, submission and feedback" link. Please upload the following two files:

1. Your report as a PDF with filename ¡student_number¿.pdf, where "¡student_number" is replaced by your student number, not your username. Upload this to the submission point "Introduction to AI Coursework (Turnitin)".

2. Your code inside a single zip file with filename ¡student_number¿.zip. Inside the zip file there should be a single folder containing your code, with your student number as the folder name. Please remove datasets and other large files to minimise the upload size - we only need the code itself. Upload this file to the submission point "Code for Introduction to AI Coursework".

[1] R-markdown and some other notebook-based environments typeset using LaTeX, this is acceptable

We may review your Python code by eye but your marks will be based on the contents of your report, with the code used to check how you carried out the experiments described in your report. We will not give marks for the coding style, comments, or organisation of the code. Code written in Julia or R is also acceptable as is the use of a standard notebook format. If you are particularly keen on another programming language let me know and I will consider this; I would accept other modern languages such as Rust, but outmoded or unsuitable languages like C++, Java or MATLAB would not be allowed.

Please do not include your name in the report text itself: to ensure fairness, we mark the reports anonymously.

Avoiding Academic Offences: Please re-read the university's plagiarism rules to make sure you do not break any rules. Academic offences include submission of work that is not your own, falsification of data / evidence or the use of materials without appropriate referencing. Note that sharing your report with others is also not allowed. These offences are all taken very seriously by the University and we have very little leeway within the framework the University has set out. Do not copy text directly from your sources - always rewrite in your own words and provide a citation. Work independently – do not share your code or reports with others; you can, of course, discuss your work with your classmates, but do not share text or code.

Suspected offences will be dealt with in accordance with the University's policies and procedures. If an academic offence is suspected in your work, you will be asked to attend an interview with senior members of the school, where you will be given the opportunity to defend your work. The plagiarism panel can apply a range of penalties, depending on the severity of the offence. These include a requirement to resubmit work, capping of grades and the award of no mark for an element of assessment. Again, we are not in a position to be lenient here, the academic offences procedure is not one we control.

Extensions and Exceptional Circumstances

If the completion of your assignment has been significantly disrupted by serious health conditions or personal problems, or other serious issues, you can apply for consideration in accordance with the normal university policy and processes. Students should refer to the guidance and complete the application forms as soon as possible when the problem occurs. Please see the guidance below and discuss with your personal tutor for more advice:

- `www.bristol.ac.uk/students/support/academic-advice/assessment-support/request-a-coursework-extension/`
- `www.bristol.ac.uk/students/support/academic-advice/assessment-support/exceptional-circumstances/`