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.

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 you should compare two algorithms, both to each other and to a baseline model. You should make sure any assessment is not restricted to the data used in train models or decide on metaparameters. 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.

As a rough guide to marking:

- Initial description of the data, including some graphs or other approaches to visualisation. 8 marks.
- Two algorithms should be tested, if only one algorithm is included the 32 available marks will be halved.
- Overall presentation (4 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 (6 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 (3 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 algorithms, include a description of how you implemented the algorithms. (10 marks)
- There are some marks (3 marks) for something suprising and unusual.

Question 2 (10 marks)

For two of these three types of ethical challenge 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 of a challenge that could arise or has arisen in the past. Obviously the three broad types of challenge overlap, do not worry about the boundaries between these types, but do try to address 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, including 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. 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.