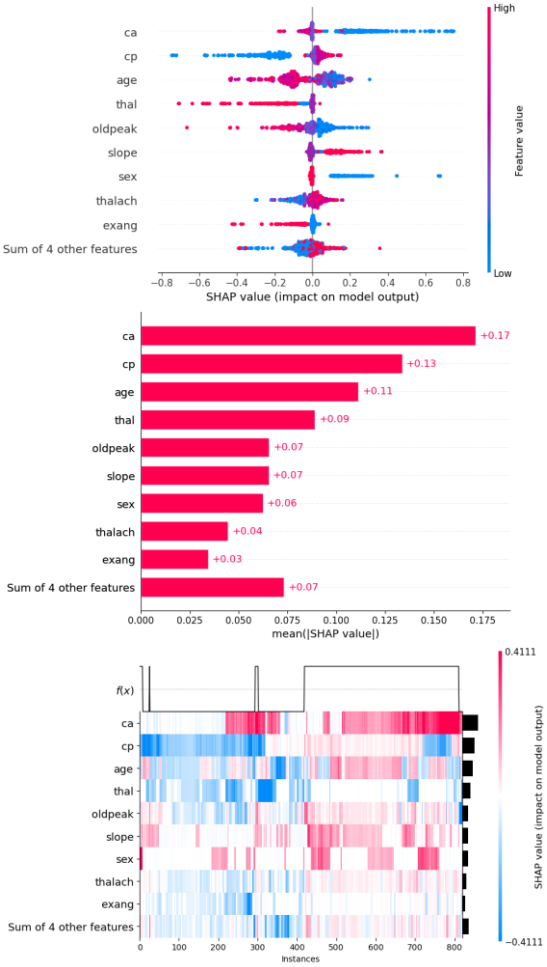
Report

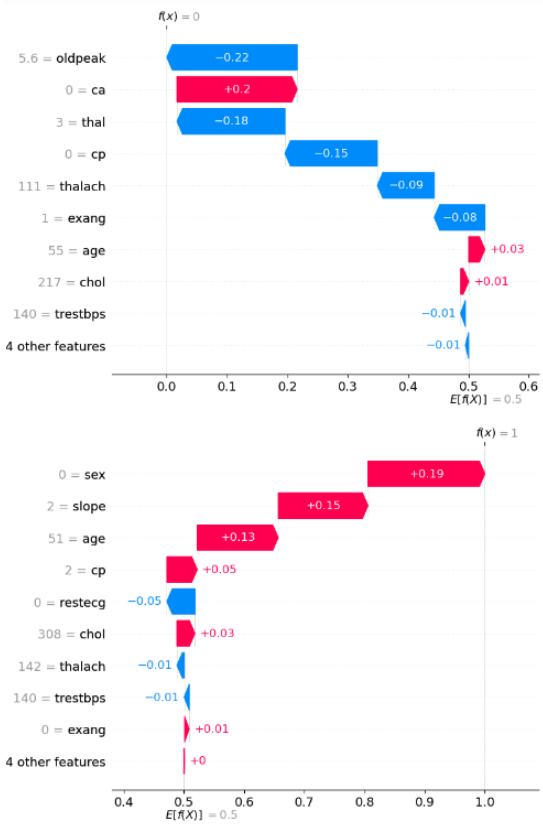
Part 1:

1. The model has predicted the positive class. (class 1)
2. Ca, cp and age are the most important factors.
3. Low Ca and Sex feature values push the impact on the model output (SHAP value) more to the positive side, while high Thal feature values push the impact on the model output to the negative side.
4. Ca, cp and thal are predominantly on one side of the SHAP value chart, meaning that those features consistently influence the outcome. It seems like ca has a big impact on the outcome.
5. 

6. The Bee Swarm plot looks very similar to the 20% data plot. (it shows the same directions of SHAP values, as well as having the same general shape.) The bar plot shows almost the same order of importance, especially in the first 4 features. I think you could therefore consider them similar too. The heatmap is difficult to interpret, because the vertical columns are a bit small to see. The SHAP value 'influencing instances' are quite the same, since they generally give the same impact on the output when generalized from 20% to 100% of the data. In general, the most substantial

change is that sex, oldpeak and slope are switched around when you go from 20- to 100%. I would not consider that a large change, but it is a change.

7. I have chosen instance 10 and 480, based on a big difference in ‘cp’:



8. Instance 0’s features with high impact:
    a. Positive: None.
    b. Negative: Age, slope, oldpeak, cp …

    Instance 10 vs instance 0:

    - Age and slope became irrelevant compared to instance 0.
    - Reversed impact direction: oldpeak, cp, thalach.
    - The predicted output changed significantly, from -0.5 to 0.5

    Instance 480 vs instance 0:

    - Oldpeak became irrelevant compared to instance 0. Sex was irrelevant, and now is relevant.
    - Reversed impact direction: restecg (small amount).
    - No change in predicted output.

Part 2:

1. The aspects that caught attention:
    a. “Theories of human cognition and behaviours can offer conceptual tools to inspire new computational and design frameworks for XAI.” they mention that they create frameworks for XAI with theories in mind, basically stating that XAI

algorithms are developed based on theories of human cognition, which was not something that was apparent to me from the start of the course.

b. “Second, XAI algorithms were often not developed with specific usage contexts in mind,” it sounded strange to me that you would design an algorithm without having a context to apply it in, like it was discovered on accident. (which is possible of course).
c. In the ‘How to Be That’ section of the table it mentions:
‘Highlight feature(s) that if changed (increased, decreased, absent, or present) could alter the prediction to the alternative outcome, with minimum effort required’, caught my attention because it mentions minimum effort required, which basically means that the computation time would be low for these methods, which would be really handy for larger datasets. It could also naturally identify the most influential features for the outcome.
d. In the ‘performance’ part of the table, it mentions ‘Provide performance information of the model’ which would also be a good tool to use when creating models that explains data, because you can identify in what ways you can improve the model to become more efficient. Also, you can identify strengths and weaknesses. I chose this because of it’s usability in the future to me.
e. ‘A pitfall robustly found in recent work is that explanations can lead to unwarranted trust or confidence in the model.’ This part highlights the dangers of using technologies without a clear understanding of how people interact with them, which is very relevant because this means that people who do not inherently know the things that are necessary for using XAI should first acquire knowledge to be able to use it in a meaningful way, which is at the core of the conversation of XAI. I chose this because it captures a major pitfall that many people might have, and I thought about the relevance it might have when explaining someone what XAI does and how they could use it.