**EXERCISE: METHODS PRACTICE**

**Instructions**

1. Upload your short discussion (≤ 1000 words) as a PDF on Canvas.
2. Upload any code or code notebooks that you used on Canvas.

**Description**

The goal of this exercise is to give you a chance to experiment with interpretability or explainability techniques in a problem that you choose yourself. For example, if your research is to develop a certain type of machine learning model, then you could apply some of the methods we discuss in the course to help explain that model. The most important aspect of the entire exercise is that you have a chance to get hands-on experience with the topics discussed in this course, so that you can develop a more critical eye towards what does and does not work well in the problems that you care about. In this sense, the project could serve as a starting point for new interpretability and explainability research that’s tailored to your problem context.

You can divide the exercise into three parts: **problem formulation**, **method application**, and **discussion** of the results. In the problem formulation, describe the problem context, a representative dataset, and why interpretability/explainability is important. What types of people would be the main beneficiaries of effective interpretation/explanation, and what exactly would they be able to get out of it? Finally, search for any relevant literature on the topic and briefly summarize some of the main approaches that you learned about. Feel free to quote or include screenshots from references. If you include those references, please provide proper citations.

In the **method application**, you will either apply an interpretable model to the dataset you described above or explain a model that has been trained to that data (you do not have to have trained it yourself). Determine whether the output is a local or a global explanation. Include 1 – 3 figures resulting from your analysis, and summarize how to read them. Make sure the figures are readable. Report the runtime of the steps that are related to interpretability/explainability. Finally, you should provide the main takeaways of the analysis – what would you share in a report to one of the beneficiaries you identified above?. You should not need to write a lot of code for this exercise, but you should think carefully about how well the method fits to your problem. In the **discussion**, you will critically analyze the results of the application. Are the results very different from your initial expectations? Which results are useful, and which are irrelevant? Finally, briefly describe how you could potentially sanity check your analysis. What type of additional study/experiments would help you convince your readers that your analysis is trustworthy?

You are free to vary the choice of data, models, and techniques however much you would like from submission to submission. For example, you may choose the apply a new interpretability technique to the same dataset from an earlier submission. Alternatively, you could gather more outputs and deepen your understanding of an interpretability technique that you considered in an earlier submission. The important thing is that in each submission, you are not blindly applying methods and that you learn something new and interesting about your problem area and about the use of interpretability/explainability techniques in the real world.

**Rubric**

*Problem Formulation* [5 points]: The focus of the application is on a broader, independently interesting problem domain. All data are reported within context, rather than assuming prior familiarity (with specific variable names or data collection methods, for example). The questions asked do not have obvious answers, and the demo could potentially find an audience beyond the course.

*Methods Application* [5 points]: Techniques from the interpretability and explainability literature are appropriately used. Their outputs are presented clearly and described correctly.

*Discussion* [5 points]: The report insightfully discusses the strengths and limitations of the methods application within the chosen problem context.