# Rubric - final report

The goal of this assignment is to practice your writing skills and to document your work so you can share it with others and easily catch up with it at some later point.

Your report should be no more than 2000 words (excluding references); we will deduct 1 point for each additional 50 words. Please feel free to use your midterm report but revise it if necessary based on the feedback you received.

Please submit your pdf to gradescope by 6pm on December 7th (8pm with the grace period)!

Formatting requirements:

1. Your report should start with the title, your name, your affiliation, and a link to your Github repository.
2. Make sure all of your work has been pushed to your repository and that your repository is well-organized (more on this below).
3. Your text should be concise and clear. No code should be included in the report, and the reader should be able to understand your methods and results without looking at your code.
4. All figures and tables should have captions. Be aware of your overall word count as you include captions.
5. All axes should be labeled and the visualization type should fit the data you plot.
6. Add a references section to cite publications, data sources, any previous work that you mention in your report. References do not count towards the total word count.

The report will be graded on a scale of 50 points.

# Report sections

Please include the following sections in your report and **do not deviate from this structure**.

### Introduction

This should be similar in structure to the midterm report. In your introduction, make sure to motivate your problem, describe your dataset, and the previous work. **5 points**

### EDA

This should be largely the same as the EDA you presented in your midterm presentation. But feel free to update your figures and generate new ones if necessary. **5 points**

### Methods

In this section, please explain your splitting strategy, the data preprocessing, and ML pipeline you developed. Try at least four different ML algorithms on your dataset, describe what parameters you tune and the values you try. What metric do you use to evaluate your models' performance and why? Measure uncertainties due to splitting and due to non-deterministic ML methods you use (e.g., random forest). In general, explain what considerations went into each step of the pipeline. **10 points**

### Results

Discuss how your scores compare to a baseline score, how many standard deviations above the baseline your model is (in classification: what is the baseline accuracy, f_beta score, etc.; in regression: calculate what is the baseline MSE/RMSE or R2 score). Which ML model was the most predictive? Summarize the performance of the ML models in a table or using a figure. Calculate at least three different global feature importances and discuss your findings. Also calculate SHAP values for local feature importance. Discuss the results of model interpretations in the context of the problem. Which features are the most and least important? Did you find something that’s unexpected/surprising/interesting? **15 points**

### Outlook

The outlook is the place to describe what else you could do to improve the model or the interpretability, and what are the weak spots of your modeling approach. How would you improve this model? What additional techniques could you have used? What additional data could you collect to improve model performance? **5 points**

### References

List the publications, data sources, and any previous work you found. **5 points**

## Github repo

Additionally, **5 points** will be given for your github repository. The first thing people inspect in your repo is the readme file. The readme file should give an overview of the project, it states what python version and package versions were used to develop the code so others can run it locally and reproduce your results (add a yaml file for ease of use). There should be a licence file to let people know what they can and cannot do with your code. Github offers a couple of licence options, check those out and decide what’s best for you. The repository should have the following directory structure:

```
.
├── data/
├── figures/
├── results/
├── report/
├── src/
├── .gitignore
├── LICENSE
└── README.md
```

All (raw and preprocessed) data files are in /data, all your generated figures are in /figures, your results (predictions, saved models, etc) should be in /results, the pdf version of your final report should be in /report, and all of your source codes (ipython notebooks or python files) should be in /src.