Stat 479 Course Project Guidelines

You may work on this project in groups of between 1 and 4 but each of **you must submit a final report individually**. You must let me know which group you are in and the dataset you plan to work with by **Wednesday November 25**. The final project report is due by midnight **Friday December 18**.

Tasks

1. Choose a dataset that has at least $p = 5$ variables and $n = 20$ samples. There are no restrictions on how you choose the dataset. It can be related to another research project or a particular interest you have. If you are having trouble finding a dataset, there are public data repositories (e.g. UCI Machine Learning repository). Before thanksgiving **(Nov. 25)**, let me know who is in your group and which dataset you plan to use.
2. For your dataset, understand (to the best of your knowledge) the underlying scientific question and how the data is collected. E.g. Were the samples drawn randomly?
3. Perform an exploratory data analysis (EDA) on your dataset. E.g. look at things like histograms of responses/covariates, scatterplots, e.t.c.
4. Based on the previous steps, determine suitable pre-processing of the data. E.g. normalization of the responses/covariates, transformation of the responses/covariates, encoding the covariates.
5. Determine at least 2 algorithms/methods that you can use from the class that suit the problem you are trying to solve.
6. Implement these methods in R (or some other software) and interpret the outputs (e.g. what are the most relevant covariates, which relationships between variables are significant, e.t.c.).
7. Validate your methods (e.g. cross-validated prediction error, stability, e.t.c).
8. Based on your results, connect the conclusions back to the original scientific goal.
9. Discuss (you do not need to implement) next steps/future work based on what you have found so far.

Assessment/tips

- Assessment will be based on the final report. The final report needs to address all the items listed above.
- Present the data/results in the clearest way. For example clearly labelled and annotated plots and a very clear description of each step taken is critical.
- All the results should be reproducible meaning all the details of what you did should be included such that anyone can take the report and re-produce the results.
- Clearly justify any decisions you made (e.g. why did you use a particular algorithm or transform the data in a particular way).
- Remember that since this project involves real data, there are many right answers and it is about justifying why you made the choices/decisions you did.
- Express your knowledge/understanding of both the scientific problem and the concepts in class in your report.
- If you work in groups, include a section at the end of your report explaining exactly what your contribution to the group project is.