Final Project/Report 4

Due 12/05/2022 (Report 4) or 12/14/2022 (Final Project)

The final project represents an integration of statistical learning techniques covered in the entire course. We will follow the same guidelines as we have been all semester, with the final project including five sections (Introduction, Methods, Results, Discussion, References) and written as an R markdown report.

The final project report should meet the following guidelines:

1. **Methods:** Compare at least *THREE* models based on techniques discussed throughout the semester. **At least two of these models should be from the second half of the course (GAMs, splines, decision trees, PCA, clustering, SVMs, etc.)**. You should provide a brief description of each technique *in your own words*.
2. **Methods:** Describe your rationale and all modeling choices made (e.g. if you choose a specific form of a relationship between a predictor, x_1, and the response, y, for a GAM, what motivated your choice? How did you choose how many principal components to retain, or how many degrees of freedom to use, or a particular value of λ?).
3. **Results:** Include at least two figures that directly supports modeling choices made (see #2 above) or other key claims/statements made in your report.
4. **Results:** Utilize a cross-validation technique described in Chp. 5 to evaluate performance of your three models on a hold-out test set. You should present the results of your cross-validation analysis in a well-formatted table.
5. **Discussion:** Provide a recommendation of which modeling technique is the best choice for this dataset among those compared. You should justify your recommendation with a discussion of the main trade-offs associated with different modeling techniques (in terms of trade-offs between bias and variance, accuracy vs. interpretability, etc.).
6. **Discussion & Intro:** Interpret the results of your model in the context of a compelling research question. If you set out to understand a relationship between a predictor or set of predictors and a response variable, how does your model address this question?