STAT-627 - Final Project

Due date: Sunday, June 26, 2022 by 9:00am in Canvas. (**Note the due date and time!**)

Instructions: For this project you are to apply, tune, assess, summarize, and provide reproducible code that implements and compares several statistical learning methods. Your analysis will focus on **both** prediction of a *quantitative* response variable and *classification* of categorical response variable from study data of your choosing. You are to implement **two** distinct methods for Regression and **two** distinct methods for Classification. Each method must depend on a tuning parameter.

1. **Choose your data sets**: Visit the UC Irvine Machine Learning Repository. Select one data set from the UCI repository whose *Default Task* is **Regression** *and* one data set whose *Default Task* is **Classification**. You can choose a single data set that has both Tasks. Cleary identify the response variables which correspond to the regression task vs. the classification task.

   Each data set must satisfy the following characteristics:

   (a) *Data Type* is Univariate or Multivariate (I recommend avoiding any that are also Time Series).

   (b) *Attribute Types* must include Real/Integer (but may also include Categorical). These are the predictor variable types.

   (c) # *Instances* ≥ 200 (after removal of missing data). This is the sample size.

   (d) # *Attributes* ≥ 8 (This is the number of predictor variables. If there is a particularly interesting data set with fewer than 8, make a case for using that data.)

   Clearly state which data set(s) you are using, provide a link to the UCI site, and submit the data with your final report. You may propose a different data set from a different source to use on this project or you may propose to design a *simulation study*. See me if you are interested in either of the latter two before starting your project to confirm it is a suitable data set.

For each of the Regression and Classification tasks, complete the following:

2. **Prepare your data set(s)**:

   (a) Briefly **define the variables** in the data set and the overall goal of your analysis. Clearly identify the response variable and predictor variables in your data set. Identify whether these are quantitative or categorical. State if this corresponds to a regression or classification setting and why.

   (b) **Remove** any observations that have **missing values** on *any* variables. Treat the remaining observations as your full (complete cases) data set. How many observations did you remove? What is the sample size of your remaining full data set?

   (c) *Randomly select* a **test data set** that is approximately 10% of your full data set. Separate this out from your full data set. Treat the remaining 90% of your data set as your **training data**.

3. **Identify and conduct your analysis on the training data**:

   (a) Conduct an **exploratory data analysis** on your training data and briefly summarize any interesting features of your data set.

   (b) **Identify the statistical learning methods** you will use to address the overall goal of the analysis. At least one method for each data set analysis must be selected from those we covered in Chapters 6 through 12. Provide formal representation of the methods (such as a mathematical expression for a model or a description of how the method works/is fit) and identify any important components in your representation. Your methods must depend on some type of *tuning parameter*. Identify the **tuning parameter** for your methods.

   (c) State the **assumptions** of your methods. Assess the assumptions and making any necessary adjustments. Clearly state what remedy you are applying to address issues with the assumptions.

   (d) **Implement** and **tune** your methods on the training data to select an optimal model/method for each method examined. State which method you are using to tune your model. Summarize your optimal models/methods and compare your results across the methods examined.

4. Apply each of your two fully tuned methods to get predictions for the **test hold-out** data set. Compare the results from each method and summarize your findings and what the methods suggest about the association between the predictors and the response. Use appropriate numerical and/or graphical displays to illustrate your results.

Submit the following items at the Project link in Canvas:

1. A **well-organized report** that addresses the above items. This should be no more than 15 pages (less is fine) and include *only* relevant figures and output. Judiciously select what is important to present. You may provide an appendix that includes additional supporting material that you reference in your final report. *Include only relevant R code and output* in the report.

2. An **R script file** that replicates your analysis in full and includes *all* data processing, decision points (for example, if you transformed a variable it should be clear why you transformed the variable), and analysis. You may use R Markdown here or just an R script. This file should run/compile without error and include a statement or statements to load (or generate) your data. You should use a random seed (with the `set.seed()` function) any time you randomize (such as when selecting the validation set and using cross-validation).

3. The original **data set** that you analyzed. (Not applicable if simulating data.)