DATA 252: Models and Machine Learning

Classification Project Milestone #1

Format of Submission: Please submit your assignment as an R Markdown file that has been knitted as a pdf. One submission per group please. Be sure to include all your group members' names.

Part 1: Get data approved!

- **(10 points)** Schedule time to chat with Professor Smalley early in Week 10
- I'll add extra office hours!

Part 2: Wrangle

- Do you need to wrangle any variables in your dataset so that you can analyze them in R?
    - Are there variables you thought would be numbers that R is seeing as strings?
    - Is your data tidy or do you need to clean it?
- **What to submit: (10 points)**
    - Please make a list of steps you need to take to tidy your data.
    - If you can, please show the steps that you took to tidy your data in the R Markdown that you submit.
    - If you need help wrangling data please reach out to me (Professor Smalley) and I will be happy to help!

Part 3: Questions of Interest

- **(30 points)** Write three well-defined questions of interest for your classification project. You will use these as a guide when putting your presentation together. These questions can change overtime, but it's good to have a starting point. These questions can pertain to…
    - Which variable(s) are significant in building a classification model
    - Relationships between the variables
    - Something about error rates (think overall error, false positives, false negatives, sensitivity, and specificity)
    - Something about which hyperparameters minimize error
    - Something about which model approach is best given its assumptions

IF YOU WANT, ONCE YOU HAVE YOUR DATA APPROVED… YOU CAN GET STARTED ON THE NEXT MILESTONE!

The next milestone won't be due until April 6th. I recommend starting this early.

There will be parts for knn, logistic, and classification trees. You can start on the KNN part now!

K-Nearest Neighbors

- **Step 0: (10 points)** Look at your data! Create a pairs plot with 3-5 variables. Please use `ggpairs` in `GGally`.

```
library(GGally)

ggpairs(YOUR_DATA, columns=c(COLUMNS YOU WANT),
        ggplot2::aes(colour = CATEGORICAL VARIABLE))
```

- **Step 1: (10 points)** Identify your response variable, a categorical feature, and a numeric feature (that you suspect might be related to your response). Describe the units for these variables and for the categorical variable describe the levels.

- **Step 2: (10 points)** Split your data randomly using **stratified splitting** into ***training and testing sets***. Be sure to set a seed so that your work is reproducible.

- **Step 3: (10 points)** Fit a ***k-nearest neighbors*** model on the training data set using k=3.

- **Step 4: (10 points)** Produce a **confusion matrix**. Compute the correct rate, error rate, false positive rate, false negative rate, sensitivity, and specificity.

- **Step 5: (10 points)** Perform a grid search to find the optimal number of neighbors to use in knn. State the best model and fit it.

- **Step 6: (10 points)** Produce a **confusion matrix for the best model**. Compute the correct rate, error rate, false positive rate, false negative rate, sensitivity, and specificity.

- **Step 7: (10 points)** What did you learn from this exercise? Please state in plain language without technical jargon.