**Project 1: Supervised Learning**

Due March 13, 2020 5:00 pm

For this project, you will demonstrate your knowledge of supervised learning techniques by performing two different data analysis tasks on specific datasets:

1. Regression Task
2. Classification Task

All group compositions must be communicated to Franz via e-mail ([fstoll@g.clemson.edu](mailto:fstoll@g.clemson.edu)) by Friday, 2/28 at 3pm.

A PPT should be including describing the algorithms used and the results you are obtaining (final results + sensitivity analysis)

**Regression Task**

Amazon Historical Stock Data - <https://www.kaggle.com/shawonstat/amazon-20-years-historical-stock-data> (file: AMZN.csv)

1. Create a time series of the data provided.
2. Create a statistical summary of the data set
3. Compare traditional linear regression with using 1) Ridge and 2) Lasso as the smoothing parameter. Which one these two methods give the best result?
4. Choosing the “best” ML algorithm from #3 and #4, perform a sensitivity analysis on the effects of training % vs. testing %.

**Classification Task**

Mental Health in Tech Survey - <https://www.kaggle.com/osmi/mental-health-in-tech-survey> (file: survey.csv)

1. Create a statistical summary of the data set.
2. Create an appropriate histogram for each of the features you consider relevant (at least 3).
3. Analyze the data using three different Machine Learning Algorithms to predict Mental Health Consequences.
4. Create a boxplot comparing the error from the three ML Algorithms you tested.
5. Make a summary results table showing the accuracy score, confusion matrix, and classification report for the methods compared.
6. Choosing the “best” ML algorithm from #5, perform a sensitivity analysis on the effects of training % vs. testing %.
7. Randomly create 10 instances based on the features used in the model, then use the model to predict Mental Health Consequence using the best ML algorithm obtained from #5.