# BIOS 735 Project Plan & Proposal

## **Choice of dataset**

* Web Page Phishing data (classification)
  + https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset

## **Proposal**

### Web Phishing dataset summary

* 1. Cross-sectional dataset on web page phishing. It contains the extracted feature from URLs and a binary outcome variable (whether it is a phishing website). Original URLs are not available in this data. It comes from Kaggle: https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset
  2. 100,077 observations and 20 variables.
  3. Response variable: Binary, is this webpage phishing or not?
  4. Explanatory variables: 19 features extracted from the original URL including:
     1. “url\_length”: The length of the URL
     2. “n\_dots”: The number of dots in the URL
     3. “n\_hypens”: The number of hyphens in the URL

### Research question(s) of interest

* 1. Model the probability of phishing using extracted features of URLs
     1. Comparing likelihood-based approach (GLM) with random forest (ML)
        1. For GLM, utilize various optimization methods to find coefficients (ie. gradient descent, nelder mead, etc.)
     2. Feature selection/importance under both methods
        1. For GLM, utilize feature selection methods such as lasso regression, LRT for nested models, etc.
        2. Interpretation of results in the context of the problem/real-world application

### Specific tasks

|  | Coding | Report writing | Who to claim |
| --- | --- | --- | --- |
| Data exploratory analysis | Summarization (summary statistics & figures, univariate and bivariate)  Preprocessing (missing data, transformation, etc.) | Introduction | TBD |
| Likelihood-based model | - Model implementation using optimization method from Model 2 (TBD)  - Feature selection | Method (writing out likelihood equations, optimization algorithm)  Result (Figures) | TBD |
| Random Forest (ML model) | - Model implementation with cross-validation  - Tuning hyperparameters  - Feature selection | Method  Result (Figures) | TBD |
| Result gathering | Rmarkdown writeup  Package creation  Repository management | Result  Discussion | TBD |

### To-do next

* 1. Job splitting - assign specific tasks to specific people
     1. Code
        1. 2 people for GLM
        2. 2 people for ML
        3. 1 person for EDA and package creation/repository management
     2. Report writing
        1. TBD
  2. Start working on data preprocessing, EDA, and likelihood based method (from scratch)
  3. Schedule next meeting
  4. Create a timeline for project progression and completion

## **Weekly meeting time**

* Next meeting: week of March 25th or April 1st
* Future meetings: as necessary (weekly date/time TBD)