Project 4 Proposal

Title: Email Spam Detection

Industry Focus: Custom

Problem: Identifying email spam

Team Members: Colleen Cobb, Taylor Hill, Hannah Thelander, Gayatri Kotaru, Rilee Peebles, Stefanie Mendelsohn

Tools:

- **Python Pandas**
- **Python Matplotlib**
- **HTML/CSS/Bootstrap**
- **Database to Load:**

Roles:

- Machine learning models

Datasets to be used:

- https://www.kaggle.com/code/mfaisalqureshi/email-spam-detection-98-accuracy/data
  - Word count of spam email
  - Confusion matrix
- https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv
  - Count of common words in a spam email
- https://www.kaggle.com/datasets/ozlerhakan/spam-or-not-spam-dataset
  - 2500 ham and 500 spam email**s**
- https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/
- https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
  - 425 SMS spam messages was manually extracted from the Grumbletext Web site
  - A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC)
  - list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis available at [Web Link]
  - Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages and it is public available at: [Web Link]

.

Goal: Create a website where you input text and the machine learning can predict if the text is spam.