COE 379L Final Project Proposal

By: Antonio Jimenez and Anna Victoria Lavelle

Introduction:

For our final project, we would like to investigate how the use of XGBoost and Support Vector Machines might improve our cancer recurrence predictions from Project 2. For some datasets and projects, increasing accuracy and recall by even only a few percentage points can make a major difference. This is especially true for the dataset from Project 2 since it predicts cancer recurrence. In Project 4 we seek a model that increases both recall and accuracy. Given the stakes of a false negative it is important that we increase recall. While maximizing recall is important, we also want to increase our accuracy in order to also reduce the number of false positives saving false positive patients hundreds dollars. Given the large number of models we generated during Project 2, it is easy to draw comparisons once we implement XGBoost and SVM.

Data Sources:

We will be using the dataset from Project 2 that was provided through the class GitHub. The data can be found here.

Methods, Techniques, and Technologies:

We are planning to investigate the effects of XGBoost and Support Vector Machines on our predictions. We also hope to use our MLOps from Project 3 to create an interactive, automated prediction system for users.

Deliverables:

Our deliverables will include a containerized inference server containing our new model using XGBoost and SVM. We will also produce a report showing our process for creating the new model and how it compares to models previously created during Project 2. Within the report we hope to produce visual aids that depict the effect of the advanced algorithms. Lastly, we will record a video displaying the use of our inference server. The deliverable will be a github containing the data, report, models, Dockerfiles, README, jupyter notebook and other files necessary for the completion of the Project 4.