MSDS696\_X70\_Data Science Practicum II – Project Introduction:

Classification of Escherichia coli (E.coli) proteins.

The project was inspired by a paper (<https://www.aaai.org/Papers/ISMB/1996/ISMB96-012.pdf>) authored by Horton and Nakai (1996), titled “A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins”. In the paper, the authors defined a model of “classifying proteins into their various cellular localization sites based on their amino acid sequences (Horton and Nakai, 1996, p. 109).” Horton and Nakai classified 336 E.coli proteins into 8 classes, achieving an accuracy of 81%. They selected a probabilistic system to do the classification due to uncertainty in the field of computational biology and a probabilistic tool is designed to handle the uncertainty.

The Data:

The data is from UCI Machine Learning Repository.

<https://archive.ics.uci.edu/ml/machine-learning-databases/ecoli/>

The data file is ecoli.data and has 336 rows with 9 columns. The ecoli.names file describes the data structure, attribute Information, and the class distribution (the class is the localization site).

The data did not have any missing attributes.

The project:

I selected the project for my data science practicum since E.coli is a pathogen that exists in our gastrointestinal tract and variants of E.coli cause illnesses that can infect the gastrointestinal tract, the urinary tract, the bloodstream, and the central nervous system. The illnesses cause death worldwide and they can be a major public health concern.

There have been recent advances in understanding the intestinal pathotypes of E.coli. The localization site of a protein within a cell is primarily determined by its amino acid sequence. The location of a protein can provide valuable information about its function. With the increase of sequenced genomic data, there is need for accurate prediction of localization of proteins, as these advances help in novel approaches to the development of vaccines and treatments that prevent illnesses and complications caused by the E.coli bacteria.

I wanted to do a classification project that would compare the results to those attained by Horton and Nakai. I compared a Support Vector Machine (SVM) model (one with a linear and the other with a radial kernel) and compared with a random forest model.