This project was genomic based, using raw gene nucleotide sequences. Genes can be sorted into classes based off of nucleotide sequence similarity and function. The goal of our model is to take raw nucleotide sequences and be able to predict the numerical class to which they belong based off of the gene sequence. We looked to take these sequences of varying lengths and accurately classify them into one of seven classes. We initially looked to run a linear regression on an encoding of these sequences, but found that linear regression and logistic regression only accept data of the same length. To get around this, we found out which sequence length appeared the most in the 4000 sequences and ran the regressions on only these sequences. The major issue with this however was the clear waste of data. We next looked to implement a neural network that would accept all of our data. However, once again, we had an issue with our encoded data. Attempting to pass the encoded data through tensors would not work for us because of a type error, so we looked for another method.

Eventually, with the advice of Professor Tristan, we looked into string kernels and found that we would not need to encode our data or worry about the varying lengths of the sequences. We looked into the parameters of the SVC model, and found that using the gama parameter as 'auto' produced the lowest mse. Similarly, we experimented with varying the degrees parameter but found no significant change in mse across all six degrees. In the future, we would like to implement an LSTM model to further investigate accuracy across different models.

Inspiration and data from:

https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723#c401