**Capstone Project 2 Outline**

**Introduction (Problem Idnetification):**

For my final project I want to explore the potential for AI to predict druggability of proteins. The first step is to recreate the model using stacked auto-encoders and biased support vector machines demonstrated in Wang et al (1). This is a model that was able to identify drug target proteins at a higher-rate then previous models by accounting for biochemical properties of proteins in a more advanced way. The data set included is generated the European Institute for Bioinformatics’ EMBOSS PepStat Calculator. This calculator generates data based on the amino acid sequences of a set of proteins. This set of proteins consists of a smaller subset of known drug target proteins (DTPs) and a larger contaminated set of NDTPs. This paper looks to build a model using deep learning and machine learning techniques to able to pick out DTPs from the contaminated list.

**Data Wrangling/ Preprocessing/ Exploratory Analysis:**

Utilizing the statistical tests mentioned in the paper (Z-score, Kolmogorov-Smirnov (K-S) test) I will determine what features are most important in determining druggability of a protein from this data set. First the data must be partitioned into two subcategories, DTPs and contaminated NDTPs. Utilizing the K-S Test we will determine the difference in each of the features as they relate to the two sub categories. A KS test histogram will be plotted to show the significant difference between each of the features of DTPs and NDTPs. There are several continuous and nominal features embedded within the data set. The continuous variables are normalized in z-space and the categorical variables will be one-hot encoded to give the features discreate binary values in ordinance with the value of the datapoint in question. Other data visualizations can be made to demonstrate the differences on how the features differ based on the z score and the categorization of DTP or NDTP. Figures showing the importance of each feature as it relates to DTP or NDTP will be show to emphasize feature importance.

**Preprocessing/Modeling:**

The data will then be passed through stacked auto encoders before being trained and tested to a series of SVMs and other machine learning models. We will then look to extract DTPs from the contaminated NDTPs and retrain the model for higher accuracy and do another one pass to obtain the best model possible.

**Goals:**

I want to recreate the work that has been done in this paper, and then build upon their findings. This paper was written in 2017 and since then major advancements in protein modeling have been achieved. Deep Mind’s Alphafold 2.0 is a company that has developed a deep learning algorithm to model proteins 3 dimensionally with impressive accuracy. I talked with the senior data scientist on our team about possible ways to incorporate date from Alphafold’s platform to bolster the significance of the findings from this exercise.

**Overview of project: Flow Chart**

Diagram

Description automatically generated

(Wang et al., 2018)

**Works Cited:**

Wang, Q., Feng, Y. H., Huang, J. C., Wang, T. J., & Cheng, G. Q. (2017). A novel framework for the identification of drug target proteins: Combining stacked auto-encoders with a biased support Vector Machine. *PLOS ONE*, *12*(4). https://doi.org/10.1371/journal.pone.0176486