**Capstone Project Idea Proposal**

Model p53 transcriptional activity (active vs inactive) using features from biophysical simulations

**Background and Problem**

Over 6 million people worldwide die of cancer each year. Cancer is caused by the accumulation of genetic mutations in two critical pathways: normal cell growth and programmed cell death (apoptosis). Defects in the cell growth pathway can result in uncontrolled cellular proliferation - tumors. Tumor suppressor proteins such as p53 normally trigger cell death in affected cells and destroy the tumor. Mutations in the p53 gene occur in more than 50% of human cancers.

Studies have shown that p53 mutants can be reactivated through second site suppressor (‘cancer rescue’) mutations. Reactivated p53 holds great therapeutic promise because animal models have shown that reintroduction of active p53, even in advanced tumors, leads to tumor regression. There have been many efforts to find small molecule drugs that mimic the cancer rescue effect. Despite some promising discoveries, studying a large and more diverse collection of cancer rescue mutations that reactivate p53 cancer mutants is required to understand the p53 mutation-structure-function relationship.

Unfortunately, testing all possible mutation combinations to determine their cancer rescue effects experimentally is infeasible due to time and expense. Therefore, it would be very desirable to have a computer model to run in silico experiments on virtual mutants. Such a model could narrow down the list of likely cancer rescue mutants to a number that reasonably could be tested in the laboratory.

**Data set**

The data set for this project is from the UCI Machine Learning Repository.

<https://archive.ics.uci.edu/ml/datasets/p53+Mutants>

Biophysical models of mutant p53 proteins yield features which can be used to predict p53 transcriptional activity. All class labels were determined experimentally.

The data consists of 16,772 instances and 5409 attributes per instance. Attributes 1-4826 represent 2D electrostatic and surface based features. 4827-5408 represents 3D distance based features. The 5409th attribute is the class attribute which is either active or inactive. The class labels are to be interpreted as follows: ‘active’ represents transcriptionally competent, active p53 whereas the ‘inactive’ label represents cancerous, inactive p53.

**Solution and Client**

This is a binary classification problem – classifying the mutations as active or inactive. Identifying the features that play a crucial role in this classification will also help in understanding the mutations.

This will help cancer researchers in the biotech or pharma industry working specifically on protein engineering, drug development or basic science research.

**Relevant Papers**

Danziger, S.A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G.W., Kaiser, P., and Lathrop, R.H. (2009) Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning, PLOS Computational Biology, 5(9), e1000498  
  
Danziger, S.A., Zeng, J., Wang, Y., Brachmann, R.K. and Lathrop, R.H. (2007) Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants, Bioinformatics, 23(13), 104-114.  
  
Danziger, S.A., Swamidass, S.J., Zeng, J., Dearth, L.R., Lu, Q., Chen, J.H., Cheng, J., Hoang, V.P., Saigo, H., Luo, R., Baldi, P., Brachmann, R.K. and Lathrop, R.H. (2006) Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants, IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 3, 114-125.