**Proj 1 Report:**

**Evan Lydon**

**Report:** Based on the results of the regression analysis, the unknown compound is most likely a signaling agent drug.

Data on the GI50 for 26 different drugs was read from csv files and stored as pandas dataframe objects. These dataframes were sorted into classes based on their reported mechanism of action. The dataframe for the unknown drug to be classified was added as a final column to each of the datasets.

Datasets were plotted by column into a pairplot using seaborn. A plot of the signaling agent GI50 values are shown at the end.

In order to preform linear regression on the datasets, the columns of each dataset had to be of equal length and represent the same type of data, however, the datasets for the GI50 values were of different length and had different types of data interspersed throughout. In order to align the datasests, the datasets had to be merged along the axes CellPanelName and CellLineName. This ensured that all data in a single row came from the same cell type and cell line and that regression would produce meaningful results.

Multiple linear regression was performed on each dataset with the GI50 values for the known compounds as regressors for the GI50 values of the unknown compound. In theory, a strong correlation between the known compound’s GI50 values and the unknown compounds GI50 values would suggest that they were in the same mechanistic class.

Statsmodels OLS regression on the whole datasets gave only one good fit for the unknown compound: signaling agent. The multiple linear regession gave an R^2 value of 0.761 with compounds 745750 and 718781 having the only two meaningful coefficients in the prediction equation. All of the other datasets produced R^2 values below 0.1 indicating poor fit and little to no relationship. The MSE was 2,3x lower for the signaling agent class compared to the other classes.

SKlearn train\_test\_split was used to split the datsets 0.8:0.2 for training and testing,respectively. Sklearn linear regression analysis was then performed and the r^2 values along with the mse were calculated and printed. Once again, the r^2 value for the signaling agent was significantly higher than any other class of molecule.

Changing the random seed in the train\_test\_split function did not change the conclusion, however, it did show that the large variability in the datasets could produce inaccurate regression models. 5 different random seeds were chosen for the train\_test\_split method. The r^2 values for the training sets were: [0.7296342039994141, 0.7704732144752366, 0.7576503814563453, 0.7907588102625975, 0.7797484559038739] and for the test sets: [0.8797803230319067, 0.6823737349284482, 0.7253315186850557, -2.216326810532933, 0.6463742231409122]. On the fourth fitting, the training set did not reflect the test set and thus the r^2 value for the test set was very low.

Similair results were obtained for cross\_validate, in which a large spread of r^2 values were observed. Again, for signaling agent class, the cross\_validate function returned the following array of r^2 test set scores using 5 fold cross validation which were significantly higher than any other mechanism class: [ 0.93603865, -1.26137933, -0.05130082, 0.60608167, 0.83616845]. The large spread in r^2 values is possibly due to the large variation in the data and possibly due to large clumps of data around specific values which can be seen in the pairplot below. The training set produced normal r^2 score values when using cross validate: [0.69709034, 0.77912238, 0.81536194, 0.79518689, 0.70702531].

Differences in the results using cross\_validate versus train\_test\_split methods likely is the result of the method each function uses to split the data into training and testing sets.

However, despite these inconsistencies, it seems that the clear choice for the unknown drug’s class is a signaling agent. This is because the unknown drug’s GI50 value in specific tissue cells and cell lines follows a trend found in drugs of the signaling agent class whereas the other drug classes do not show good linear regression fit with the unknown drug’s GI50 value.

