Project 1

Note: The project report should be up to approximately 3 pages long including figures or tables. Please provide details about methodological aspects, parameters used, rationale, references, and interpretation of the results. In addition to the report, upload in Canvas your python code(s) and their output. Partial credit will be given to incomplete solutions of each problem.

Identification of anticancer agents with similar growth inhibition in the NCI60 Panel of Human Tumor Cell Lines

The study described in the attached paper analyzes patterns of growth inhibition by anticancer agens in the 60 tumor cell lines (NCI60) utilized by the Developmental Therapeutics Program at the National Cancer Institute. Results highlight similar patterns of growth inhibition for compounds of similar mechanism. The reference for this paper is: Susan L. Holbeck, Jerry M. Collins, and James H. Doroshow, "Analysis of Food and Drug Administration–Approved Anticancer Agents in the NCI60 Panel of Human Tumor Cell Lines" *Mol. Cancer Ther.* (2010), **9**, 1451-1460.

Perform statistical analysis of a dataset obtained from these studies to characterize the similarity of the growth inhibition patterns, and to assign an "unknown" compound to the appropriate group. **The dataset to be used for this project is available in Canvas.**

A. (30 points) Build datasets comprising the GI50 value (i.e. the concentration at which the compound inhibits growth by 50%) of "known" compounds with similar mechanism. The mechanism groups are shown in Table 1 in the paper and the GI50 is indicated in the "logValue" column of each data file. Include the "unknown" compound data as an additional column in each dataset. Print information about the data contained in the dataset, such as the header, description or summary.

B. (20 points) Graph the dataset using a pair plot representation.The diagonal plots should represent probability distributions. Include the regression line in each plot. The upper triangle should use the scatter plot representation and the lower triangle kernel density estimates.

C. (50 points) Build input datasets comprising only data for compounds in the same group. Assign the "unknown" compound data to the output variable. For each group, perform a **multiple linear regression** of the "unknown" compound data on the predictor dataset.

- Use statsmodels ordinary least squares (OLS) regression model to perform the linear regression. Print the statistics using the summary table (use the summary() function in statsmodels). Using these results explain how good the statistical prediction is.
- Split the data into training and test sets using a randomized approach.
- Determine the regression coefficients and obtain an assessment of the fit using the Mean Squared Error (MSE) and the R^2 statistic for the **training** and **test** sets. Do the values of MSE and R^2 indicate that the model fits well the measured values?
- Use the **cross validation** (cv) method to estimate MSE and R^2. How do the cv values compare with those obtained using the entire dataset? Explain why the results are better or worse than for the previous calculation.
- Repeat the above calculations for a different random seed number used in the data splitting procedure. How robust are the results? Explain these findings.

- On the basis of these results, assign the “unknown” compound to one of the groups. Justify your answer.