Applications of Machine Learning in Hydrology
Homework 2: Due Thursday, March 2nd at 11:59 PM

Please email me your code/any other documents at marc.berghouse@dri.edu

For this homework, we'll set up our machine learning pipeline for some streamflow data from the USGS "Machine-Learning-in-Hydrology/data/streamflow/usgs_stream_discharge". The data was collected from the NWIS (which we will go over next lecture), and includes discharge, site_id, date, temp, conductivity and pH from a number of sites. For a more complete machine learning analysis, we would want to substitute the site number for latitude and longitude, but for exploratory/educational analysis it's fine**. I will include most of the code required to set up the pipeline for a random forest regressor, but some of it will be missing!**

1. Get a machine learning model working for the prediction of discharge. You can do all the steps of the pipeline on your own, use parts of my code, or use my code in its entirety. If you use my code, some parts will not work, and you will have to debug to figure out what is missing.
    a. We first have to gain a basic understanding of the data. Write a couple sentences about what each of the variables represents, and what some of the hydrologic processes are that may control some of these variables (general is fine but specific is better). No more than 8 sentences, please.
    b. Part of the pipeline is exploratory data analysis, which I have done with Pandas profiler. What variables have significant correlations with each other? Write a couple sentences about why some variables are correlated and others aren't.
2. Get an R2 value of .85 or greater. Describe the steps you took to increase your score. If you have to abuse the random train-test split, comment about why this might be, providing at least one piece of evidence (for example: data points at a particular site/time are outliers, and removal of outliers provides a better score)
3. Generate R2 values and RMSE values for 3 different regression-based statistical or machine learning models. Try to pick things you think will work well, but don't worry too much about this. Create plots to visualize the results of each method (for example, scatter of y_pred vs y_true with line of best fit, R2 and RMSE displayed).
    a. Of the 3 methods you picked to graph, which one works best? Do a little bit of Google searching as to why this model outperformed your other model, and write 4-5 sentences about your findings (can be done in separate text document, or you can create a markdown cell in your notebook).
4. List 3 possible methods to improve the prediction of discharge for your best model, and attempt one of these methods. Did it work? Why or why not?
5. Can we predict any of the data better than discharge? Generate predictions/scores with each of the 3 models for each variable, and plot the results as you did in part 3. This time you will have a range of values for each model corresponding to the prediction of each variable, which will make visualization a bit trickier. One suggestion is to make 6 plots (one for each variable), but there are probably better ways to do it. Write a paragraph about why you think some variables are easier to predict, and why certain models might be better, the same, or worse, at predicting certain variables.