Homework 1. – Data 522, Due 02/06 Friday

Please turn in your Jupiter notebook files (*.ipynb) and a proper documentation.

An automated electrochemical setup in my collaborator's group has recently conducted 500-large experiments, whose data are provided here as "ExerciseData.csv". In each experiment, the automated platform chooses a synthetic recipe that include the concentration of multiple elements (V, Cr, Mg, Fe, Co, Ni, Cu, S, Se, P) as well as the experiment parameters ("voltage" and "time"). The synthetized materials were characterized for the overpotential η. While the η values are posted as negative numbers in the file, the smaller the absolute value of η (|η|), the better electrocatalyst.

Part 1: Preparing the training, validation, and test dataset.

1.1. The overpotential η can be considered as a function of 11 variables: V, Cr, Mg, Fe, Co, Ni, Cu, S, Se, P, voltage, time. Visualize how the values of η in such a 11-dimensional space for all 500-large data points. (Note: this is an open-ended question. Please search online and use your imagination and creativity to plot one).

1.2. Split the dataset randomly into the training, validation, and test dataset. The split should be 70% for training, 15% for validation, and 15% for test.

Part 2: Predicting "good" catalysts from "bad" ones using scikit learn.

Now, we define that any catalysts whose |η| is smaller than 0.6 is a "good" catalyst, while any ones no smaller than 0.6 a "bad" catalyst.

1.3. Within the 500-large dataset, find out the percentage of data points that are "good" catalysts, and the percentage of data points that are "bad" catalysts.

1.4. Using the training dataset and the default setting of any one of the algorithms in scikit-learn, train a model of random forest that predicts "good" catalysts. Against training, validation and test dataset, calculate the values of precision, recall, and F1 score for the trained model.

1.5. Conduct hyperparameter tuning for your proposed model. Try at least 4 different combinations of hyperparameters. For each combination of hyperparameters, record the combination of hyperparameters, calculate the averaged values of precision, recall, and F1 score against training, validation, and test dataset

1.6. Conduct the analysis of permutation feature importance (PFI) for your models. Compare the PFI results from models without and with input normalization. What is the ranking of top 5 most important variable?