Ensemble learning

This task is based on a paper by Shin et al. (10.1002/bkcs.11835).

In addition to the identification of the compounds detected with non-targeted screening, it is also important to understand how important such compounds are. In the environmental samples, one of the main considerations is the toxicity of the compounds. However, for both know contaminants as well as true unknown-unknowns, the toxicity values are very often missing. In this task, we will develop an ensemble learning method to predict the toxicity of compounds based on their structure.

Task:

You are given two datasets. Firstly, a dataset of 143 compounds with their CAS numbers, SMILES codes, and NOAEC (No Observed Adverse Effect Concentration) values for repeated dose inhalation toxicity study on rodents. Secondly, molecular fingerprint descriptors for the same compounds.

(1) Combine the data into one dataset. Remove descriptors that do not show significant variation from compound to compound as well as correlated descriptors.

(2) Split the dataset into

(3) Develop MLR, kNN, decision tree, random forest, and supported vector machine models to predict the NOAEC values. For SVM try linear, polynomial as well as radial kernels. Look up suitable packages from the caret.

(4) Now, develop an ensemble learning method. Investigate how well are performing mean and median prediction. Compare with using one model only.

(5) How would you estimate the confidence limits for the predictions?

Anneli