A machine learning algorithm (using Linear Regression, Random Forest Regression, Extra Trees regression and XGBoost to predict sulfur concentration at sulfide saturation

Updated 15/03/23:

\*\*\*\*\*\*

**The code**

The codes are provided as python scripts (.py).

Kmeans.py: splits the whole dataset based on K means clustering

SCSS\_predict.py is the main algorithm.

\*\*\*\*\*\*

**Input files:**

Dataset: Data\_0 to Data\_7 in folder "Dataset", based on the output of Kmeans.py

#stratified random sampling is done on the datasets individually, followed by merging them into train and test data.

#prediction dataset: loaded in "to predict\_dataset" in the same folder as SCSS\_predict.py

# choose "Regress\_Mode" to decide the algorithm to be used: #0-linear; 1-extratree; 2-Random forest; 3-XGBoost

**Output files:**

#test data is saved as file1; train data: file2; prediction for test data: file4 ; prediction for train data: file3.

# The image outputs are stored in "Image" folder

# The prediction output is stored in "predict" folder