DATA-DRIVEN MACHINE LEARNING FOR ESTIMATION OF PFAS PARTITIONING ON VARIOUS SURFACE MATERIALS

Dr. Harsh V. Patel, PhD, E.I.T

Dr. Renzun Zhao, Ph.D., P.E. & Dr. Hyo-shin Park, Ph.D., P.E.

Department of Computational Data Science and Engineering

Department of Civil, Architectural, and Environmental Engineering

North Carolina Agricultural and Technical State University

1

Machine Learning

Recent application of machine learning for PFAS

Application	Model/Algorithm	Source
Predicting PFAS removal efficiency	XGBoost Model	Karbassiyazdi et al., 2022
Predicting contamination risk for GenX	Bayesian Network model	Roostaei et al., 2021
Source allocation for PFAS in environment	Extra Tree, Support Vector Machine, Neural Networks, and K-Neighbors	Kibbey et al., 2020
Mapping PFAS structure-function	t- Distributed Stochastic Neighbor Embedding (t-SNE) algorithm and PCA	(Su & Rajan, 2021)
Classification of PFAS bioactivity	multitask neural network (MNN)) and Graph Convolution Network models	(Cheng & Ng, 2019)
Predicting PFAS defluorination	Random Forest, Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Feed-forward Neural Network (FNN), and t-distributed Stochastic Neighbor Embedding (tSNE) algorithms	(Raza et al., 2019)

2

Importance of Machine Learning

Experimentally Determined Log K_d

- Cost: chemical, instrumental, analytical, human
- Time: days to months
- Low reproducibility

Computationally Determined Log K_d

- Cost: programming, human
- Time: minutes
- High reproducibility

3

Machine Learning

Objective of this study

"To **develop and use** ML models for estimating/predicting PFAS distribution in solid-liquid phase during adsorption"

4

Adsorption Isotherm

Distribution Coefficient

- the amount of sorbate adsorbed onto solid surface per unit amount in the aqueous phase at equilibrium

$$\log K = \log \frac{Q}{C} = f(properties\ of\ sorbate, properties\ of\ sorbent, properties\ of\ solution)$$

$$\text{Log } K_d = f \text{ (Log } C_e, \text{ E, S, A, B, V, } S_{area})$$

K_d is the distribution coefficient,

Q_e is the equilibrium concentration on solid phase

C_e is the equilibrium concentration in aqueous phase

E, S, A, B, V are the Abraham's Solvation Parameters

Machine Learning

Methodology

ML algorithms	Model Variables
• Linear Regression model • Decision Tree model • Ensemble Tree model • Support Vector Machine model • Gaussian Process Regression model • Artificial Neural Networks model	• Modeled Parameter: log K_d – Distribution Coefficient • Descriptors: 1. log C_e – equilibrium concentration 2. BET Surface area 3. Abraham solvation parameter: A, S, B, E, V.

Machine Learning

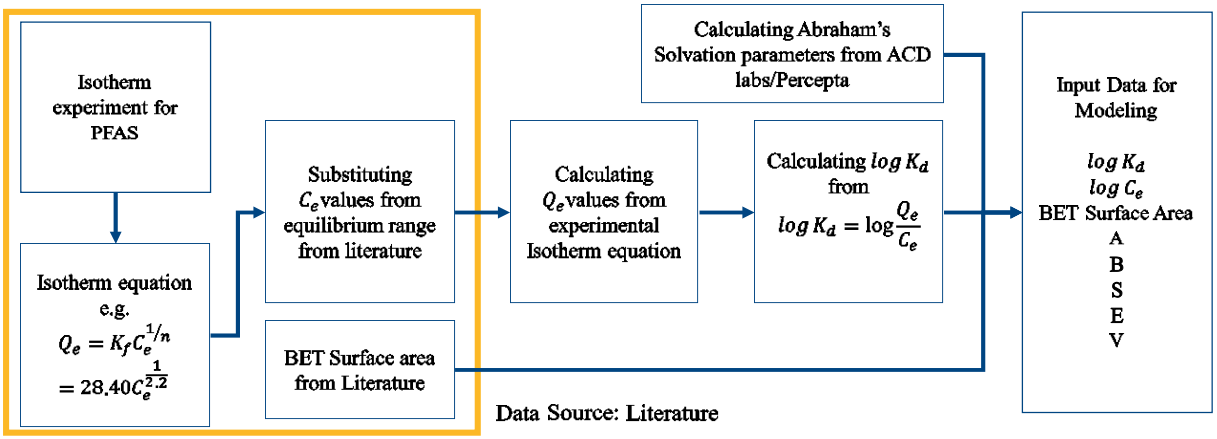
Data Mining Criteria

- An isotherm must contain at least 5 experimental data point.
- R^2 value of the isotherm should be at least 0.9.
- Isotherm with higher R^2 value should be used to data points.
- BET surface area should be measured and reported for adsorbent material.

7

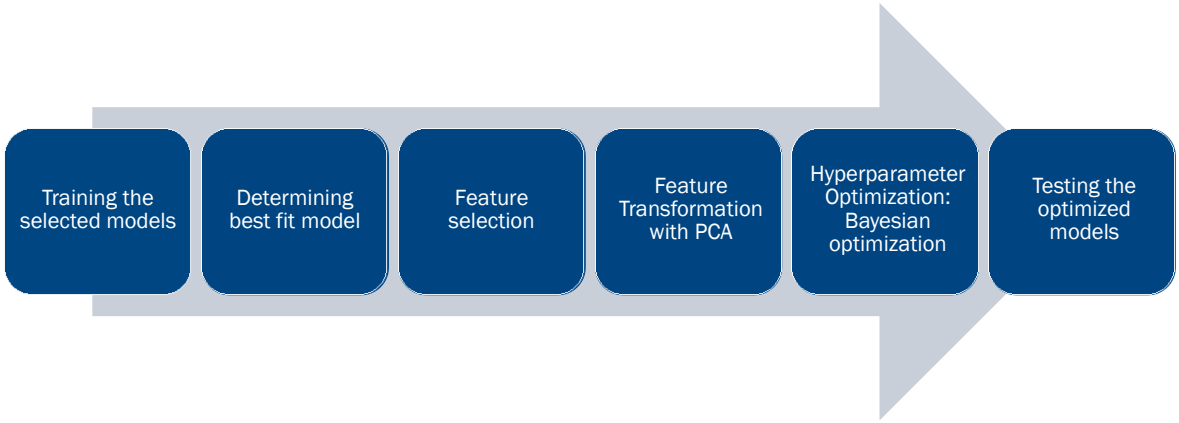
Machine Learning

Data collection process



8

MODELING WORKFLOW

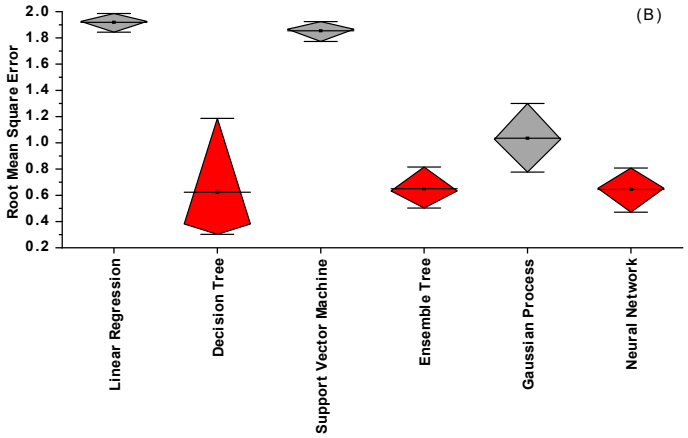


9

Carbon Based

Training Set: 70% = 651 data points

Best Fit Models
- Decision Tree,
- Ensemble Tree, and
- Neural Network



With cross validation k-fold = 2 to 10

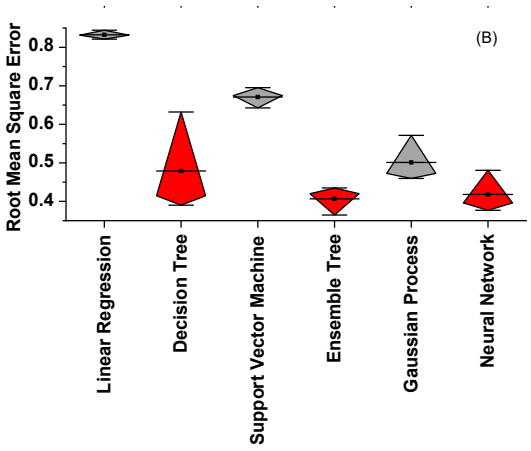
10

Mineral Based

Training Set: 70% = 284 data points

Best Fit Models

- Decision Tree,
- Ensemble Tree, and
- Neural Network



With cross validation k-fold = 2 to 10

Feature Selection

- Descriptors ranked based on f-test value
- Retrained models based on the f-test ranking

Descriptor	F-test		Rank
	Carbon-based	Mineral-based	
BET Surface Area	32.4523	164.5863	1
Log Ce	12.7671	84.4753	2
E	9.9886	50.5753	3
V	7.6174	44.5753	4
S	6.7755	42.9872	5
B	6.7755	21.2236	6
A	0.5268	12.0387	7

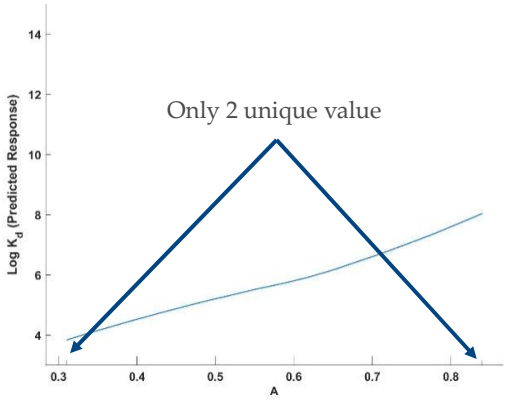
Feature Transformation with PCA

- Determining Principal components.
- Based on 95% explained variance.
- 100% variance explained by only descriptors
- Machine eliminates the rest

Descriptor	Explained Variance	
	Carbon-based	Mineral-based
BET Surface Area	85.16%	79.05%
Log C_e	14.84%	20.95%
E	0%	0%
V	0%	0%
S	0%	0%
B	0%	0%
A	0%	0%

Disadvantage of PCA in this case

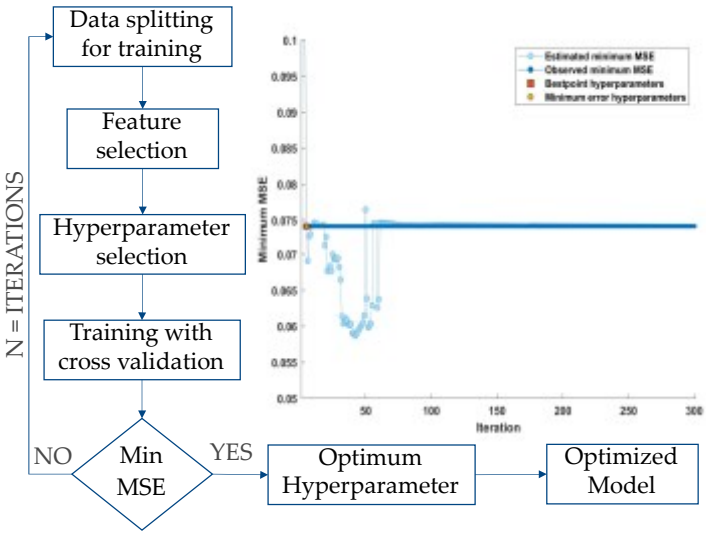
- Machine does not recognize 'A' as a PCA due to low variance from only 2 unique value
- Hence, the retrains the model with only 2 PCs
- i.e.,
  Log K_d = *f*(BET surface area, log C_e)
- However, PDP shows a relation between Log K_d and A
- Hence, PCA is disadvantageous in this case.



Partial Dependence Plot (Log K_d vs A)

Hyperparameter Optimization

- A hyperparameter is a parameter whose value is used to control the learning process.
- Process of choosing a set of optimal hyperparameters for a learning algorithm.

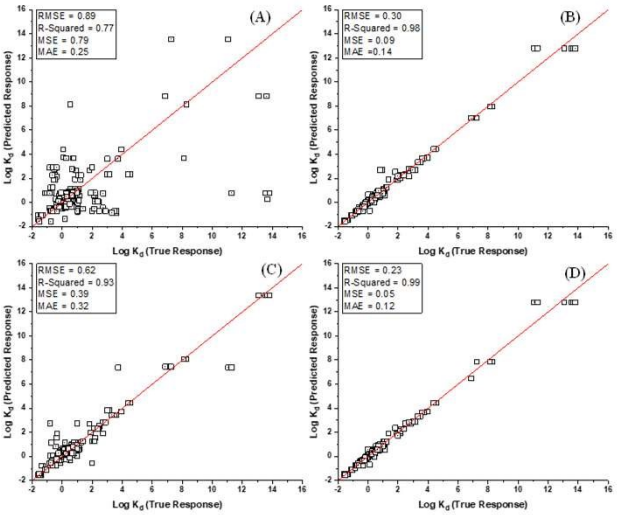


15

Carbon Based

Optimized Decision Tree Model

Regression Model	Optimizable Hyperparameter	Input Argument
Decision Tree	Minimum leaf size	4
	Surrogate decision splits	Off

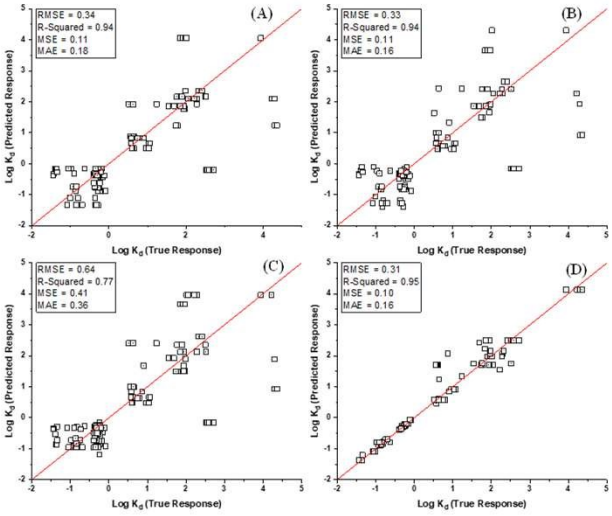


16

Mineral Based

Optimized Decision Tree Model

Regression Model	Optimizable Hyperparameter	Input Argument
Decision Tree	Minimum leaf size	4
	Surrogate decision splits	Off



Conclusion

Machine learning can be used for estimating PFAS distribution with high accuracy

Decision Tree, Ensemble Tree, and Neural Network were found to be the best models

PCA failed in this case due to lack of unique values for certain parameters

Only data for 6 PFAS was found relevant for this study out of 12,000.

More parameters may need to be considered to have universally applicable model

THANK YOU

QUESTIONS??

AGG IES DO

North Carolina Agricultural and Technical State University

19