# PFAS Differentiation With Decision Tree Algorithms

Isabela Yepes | imy2103 | EAEE E4000 | December 2021

https://github.com/isabelayepes/EAEE4000pfas

# Table of Contents

# Introduction

## Context

Teflon is the first PFAS and was accidentally discovered in 1938 (Miller 2020). In the 1950s, Teflon was first used in consumer and industrial products. By the 2000s there was a global distribution of certain PFAS. There are now more than 9,000 known synthetic C-F compounds at a global scale (Sneed 2021). Today PFAS and its related chemicals are found in stain and water resistance items, nonstick cookware, waterproof apparel, cleaning products, firefighting foam, takeout containers and carpets and textiles. Companies began to be aware of the harmful health effects but continued with production. DuPont began facing lawsuits as the effects of the unregulated chemical increasingly came to light. PFAS chemicals can be found in drinking waters near factory discharge areas. However the problem is not limited to areas with factories. Sadly, PFAS chemicals have been found in 99% of the humans tested and are known to cause a long list of cancers, birth defects, infertility, thyroid disease and more (Sprout n.d.). PFAS can be expensive to remove from contaminated environments due to their carbon-fluorine bonds; these compounds are known as “forever chemicals,” because they are very resistant to thermal, chemical, and biological degradation (Miller 2020). Further, even if PFAS are successfully removed through reverse osmosis, storing or destroying the waste byproduct from reverse osmosis is a main problem with PFAS removal efforts.

Further emphasizing the global reach of PFAS and its health effects, the EU stated in 2019: “With more than 4700 known PFAS, undertaking substance-by-substance risk assessments and comprehensive environmental monitoring to understand exposure would be an extremely lengthy and resource-intensive process. As a result, complementary and precautionary approaches to managing PFAS are being explored.” The 4,700 is in fact an understatement as there are more than 9,000 known PFAS as per the US EPA CompTox Chemical Dashboard listings.

This also identifies a main problem with PFAS, little is known about relative toxicities among all 9,000 compounds. One known toxicity difference is that “Long-chain PFAS half-life, such as of PFOS and perfluorohexane sulfonic acid (PFHxS) in the human body is upwards of 5 years. Alternatively, the half-life of PFBA, a short-chain PFAS with 4 carbons, is 3 to 4 days.” (American Water Works Association 2019). Hence why there is a current direction to phase out long chain PFAS for short chain PFAS.

Stopping all PFAS production is sadly unrealistic: “In cases where the uses of PFAS are seen as “necessary for health, safety or is critical for the functioning of society” but no functional alternatives with favourable hazard properties are currently available, certain uses of PFAS will probably continue, at least in the short term (Cousins et al. 2020). Since essential uses of PFAS are likely to continue, differentiating PFAS by toxicity could inform selection of essential PFAS. The use and application of a machine learning tool to differentiate PFAS by toxicity could further inform phase out and selection of essential PFAS substitutes.

## Literature

Cheng and Ng 2019 used various machine learning models to classify a yes/no for bioactivity of nearly 5000 PFAS. They stated that one shortcoming of the study was that it did not predict “intensity of biological effect or dose−response” (Cheng and Ng 2019). Additionally, certain chemical traits, such as the head group of PFAAs, is known to influence its bioaccumulation potential (Cousins et al. 2020).

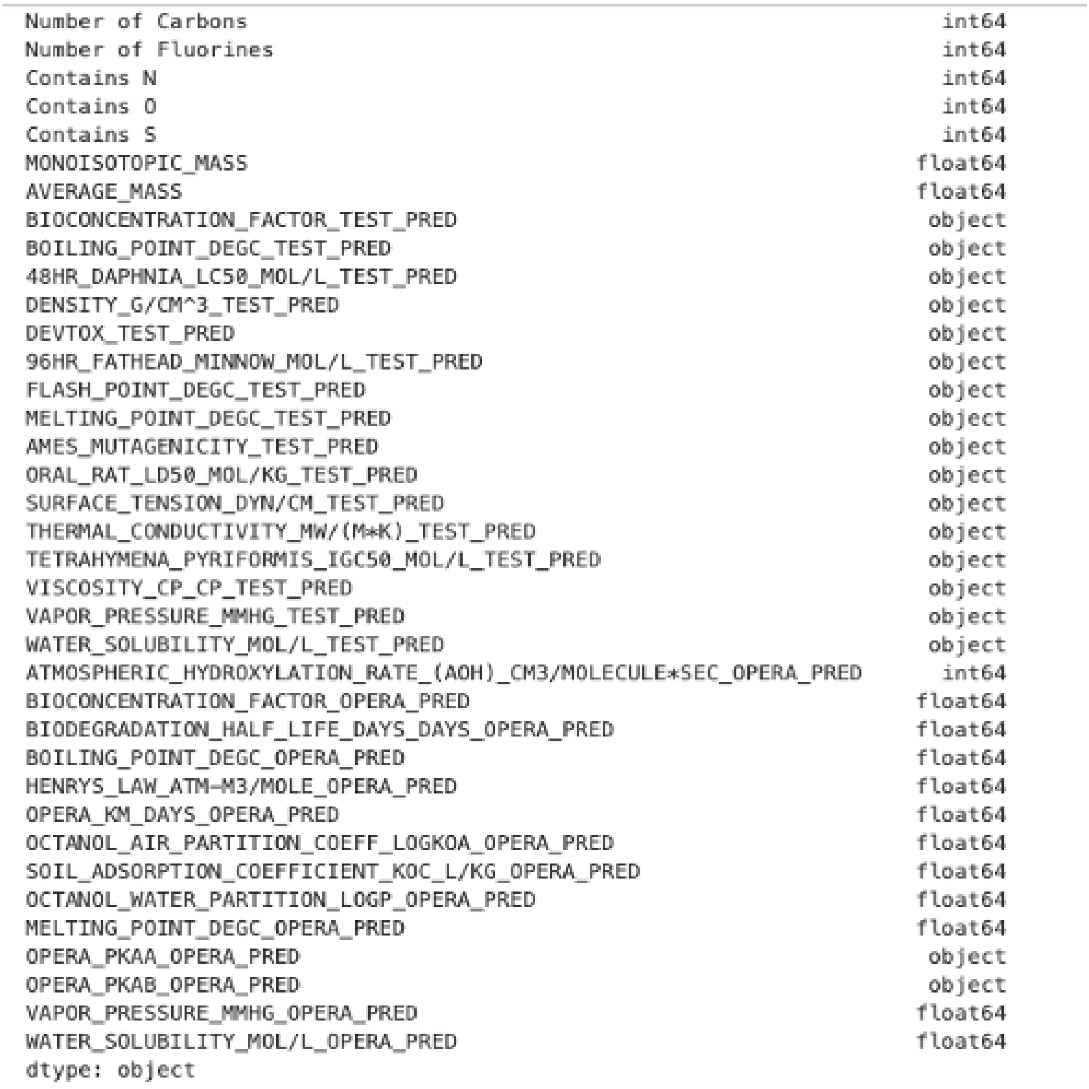
## Goal

Machine learning could create toxicity predictions based on chemical similarities, which should be easier than individually assessing toxicity of 9,000+ PFAS compounds. Hopefully improving selection of essential PFAS towards lower toxicity, combined with systemic regulation towards prevention of future PFAS-like disasters, regulation on PFAS containment and improved remediation PFAS methods, can help lower polluted drinking water health effects like those seen from PFAS.

# Methods

## Data Source

The CompTox Chemicals Dashboard is an online database from the U.S. Environmental Protection Agency. It contains lists with around 12,000 synthetic PFAS compounds. Though data is viewable online, the database has technical difficulties with data downloads. For this reason the size of downloaded data was limited to 74 PFAS compounds. TEST and OPERA are two prediction functionalities from the database that allow for data columns with the respective suffix labels. A total of 37 numerical columns shown in Figure 1 describe each PFAS compound. The non numeric data types in figure 1 are due to missing values in the downloaded data.



```
Number of Carbons                                                      int64
Number of Fluorines                                                    int64
Contains N                                                             int64
Contains O                                                             int64
Contains S                                                             int64
MONOISOTOPIC_MASS                                                    float64
AVERAGE_MASS                                                         float64
BIOCONCENTRATION_FACTOR_TEST_PRED                                     object
BOILING_POINT_DEGC_TEST_PRED                                          object
48HR_DAPHNIA_LC50_MOL/L_TEST_PRED                                     object
DENSITY_G/CM^3_TEST_PRED                                              object
DEVTOX_TEST_PRED                                                      object
96HR_FATHEAD_MINNOW_MOL/L_TEST_PRED                                   object
FLASH_POINT_DEGC_TEST_PRED                                            object
MELTING_POINT_DEGC_TEST_PRED                                          object
AMES_MUTAGENICITY_TEST_PRED                                           object
ORAL_RAT_LD50_MOL/KG_TEST_PRED                                        object
SURFACE_TENSION_DYN/CM_TEST_PRED                                      object
THERMAL_CONDUCTIVITY_MW/(M*K)_TEST_PRED                               object
TETRAHYMENA_PYRIFORMIS_IGC50_MOL/L_TEST_PRED                          object
VISCOSITY_CP_CP_TEST_PRED                                             object
VAPOR_PRESSURE_MMHG_TEST_PRED                                         object
WATER_SOLUBILITY_MOL/L_TEST_PRED                                      object
ATMOSPHERIC_HYDROXYLATION_RATE_(AOH)_CM3/MOLECULE*SEC_OPERA_PRED       int64
BIOCONCENTRATION_FACTOR_OPERA_PRED                                   float64
BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED                        float64
BOILING_POINT_DEGC_OPERA_PRED                                        float64
HENRYS_LAW_ATM-M3/MOLE_OPERA_PRED                                    float64
OPERA_KM_DAYS_OPERA_PRED                                             float64
OCTANOL_AIR_PARTITION_COEFF_LOGKOA_OPERA_PRED                        float64
SOIL_ADSORPTION_COEFFICIENT_KOC_L/KG_OPERA_PRED                      float64
OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED                              float64
MELTING_POINT_DEGC_OPERA_PRED                                        float64
OPERA_PKAA_OPERA_PRED                                                 object
OPERA_PKAB_OPERA_PRED                                                 object
VAPOR_PRESSURE_MMHG_OPERA_PRED                                       float64
WATER_SOLUBILITY_MOL/L_OPERA_PRED                                    float64
dtype: object
```

Figure 1. 37 columns for the 74 data points.

## Data Definitions (EPA 2021)

Bioconcentration factor opera pred - The ratio of the concentration of the substance in a specific genus to the exposure concentration at equilibrium ; the ratio of the concentration of a substance in an organism to the concentration in water
Biodegradation half life days opera pred - The days it takes for half of the molecules to degrade into environmentally acceptable products
Opera - Predictive model for chemical compound that utilizes Quantitative structure-activity relationship (QSAR) a computational modeling method for revealing relationships between structural properties of chemical compounds and biological activities.
Toxicity Estimation Software Tool (TEST) - allows users to easily estimate the toxicity of chemicals using Quantitative Structure Activity Relationships (QSARs) methodologies.

## Model Selection

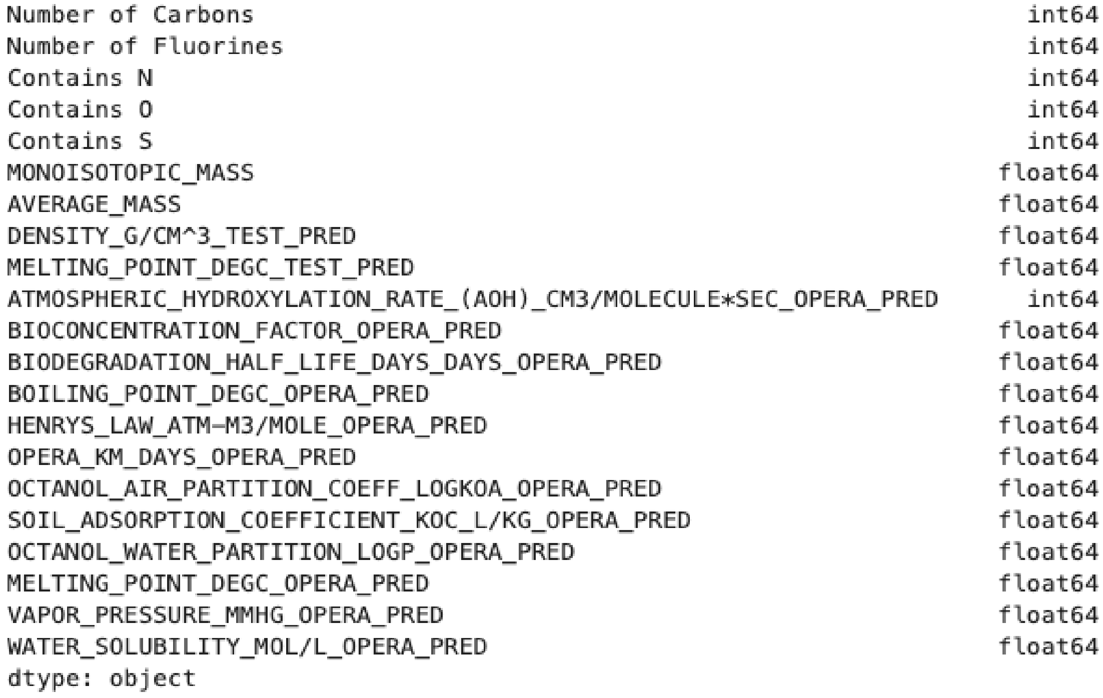
Two main classes of algorithms were considered: decision trees and neural networks. Neural networks are best suited for “unstructured data like images, text, videos and audio” (Sarkar 2021). For structured tabular data, tree-based models are preferred. Both classes of algorithms can be applied to two problem categories: regression or classification which differ based on if the end result is desired to be a prediction on a continuous scale (regression) or a prediction in a discrete grouping (classification). Additionally to evaluate models for relative success, accuracy score is only for classification problems. For regression problems the metrics used are R2 Score, MSE (Mean Squared Error) and RMSE (Root Mean Squared Error).

Within decision tree algorithms, boosted trees, such as in an XGBoost model, are traditionally more efficacious than non boosted trees, such as a Random Forest model (Sarkar 2021).

The project data is structured because it is tabular. For this reason, decision tree algorithms were selected. XGBoost and Random Forest are two popular models which will be used. Since the toxicity will be proxied by the prediction of the feature: “biodegradation half life days opera predicted”, measured in days, the problem’s prediction has a continuous scale and therefore it is a regression problem.

## Data Refinement

Rows and columns with missing values in the downloaded data were eliminated through data refinement in google sheets. Figure 2 shows the 21 numeric refined data columns for 65 data points now.



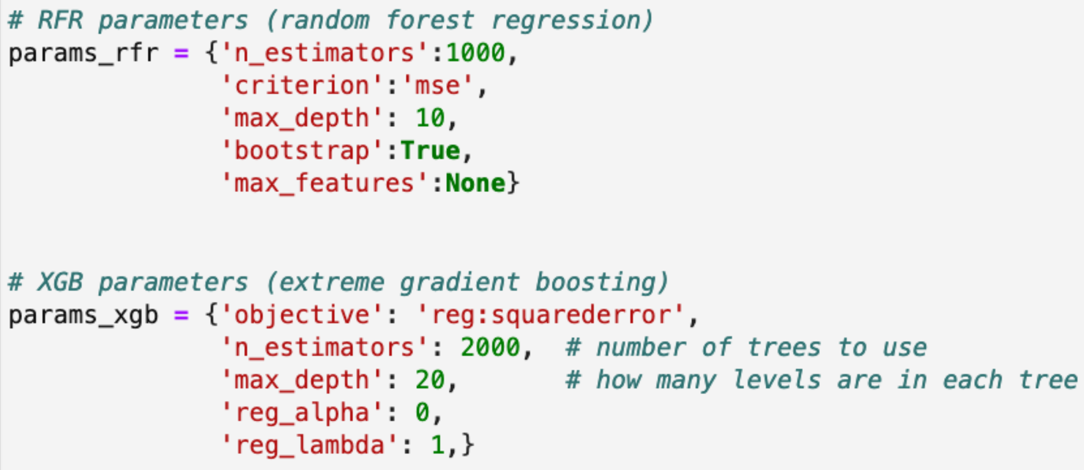
```
Number of Carbons                                                   int64
Number of Fluorines                                                 int64
Contains N                                                          int64
Contains O                                                          int64
Contains S                                                          int64
MONOISOTOPIC_MASS                                                 float64
AVERAGE_MASS                                                      float64
DENSITY_G/CM^3_TEST_PRED                                          float64
MELTING_POINT_DEGC_TEST_PRED                                      float64
ATMOSPHERIC_HYDROXYLATION_RATE_(AOH)_CM3/MOLECULE*SEC_OPERA_PRED    int64
BIOCONCENTRATION_FACTOR_OPERA_PRED                                float64
BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED                     float64
BOILING_POINT_DEGC_OPERA_PRED                                     float64
HENRYS_LAW_ATM-M3/MOLE_OPERA_PRED                                 float64
OPERA_KM_DAYS_OPERA_PRED                                          float64
OCTANOL_AIR_PARTITION_COEFF_LOGKOA_OPERA_PRED                     float64
SOIL_ADSORPTION_COEFFICIENT_KOC_L/KG_OPERA_PRED                   float64
OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED                           float64
MELTING_POINT_DEGC_OPERA_PRED                                     float64
VAPOR_PRESSURE_MMHG_OPERA_PRED                                    float64
WATER_SOLUBILITY_MOL/L_OPERA_PRED                                 float64
dtype: object
```

Figure 2. Refined 21 columns for now 65 data points.

## Parameters

Parameter optimization for the Random Forest regression using RandomizedSearchCV, GridSearchCV and GPyOpt Bayesian Optimization with Sherpa is attempted in the model section.

For comparing XGBoost Reg and RF Reg in tables 1 and 2 the parameters in Figure 3 were specified, if a parameter was not specified the default was used. These were also the parameters used to generate the graphs in the feature importance section.



```
# RFR parameters (random forest regression)
params_rfr = {'n_estimators':1000,
              'criterion':'mse',
              'max_depth': 10,
              'bootstrap':True,
              'max_features':None}


# XGB parameters (extreme gradient boosting)
params_xgb = {'objective': 'reg:squarederror',
              'n_estimators': 2000,  # number of trees to use
              'max_depth': 20,       # how many levels are in each tree
              'reg_alpha': 0,
              'reg_lambda': 1,}
```

Figure 3. Specified Parameters for table 1 and table 2.

## KNeighborsClassifier

To generate figures 4 and 5, an instance of KNeighborsClassifier was made. The following arguments were passed through the classifier: n_neighbors=15 and weights='uniform'. Additionally since the data is a regression problem due to its continuous prediction/output variable, a categorical prediction variable extrapolated from the continuous "biodegradation half life days opera predicted" was created. The following half-life categories were selected, informed by the distribution shown in Figure 4: low (between 0 and 3.9), medium (greater than 3.9 up to 4.5) and high (greater than 4.5 up to 9). This was the resulting category distribution:

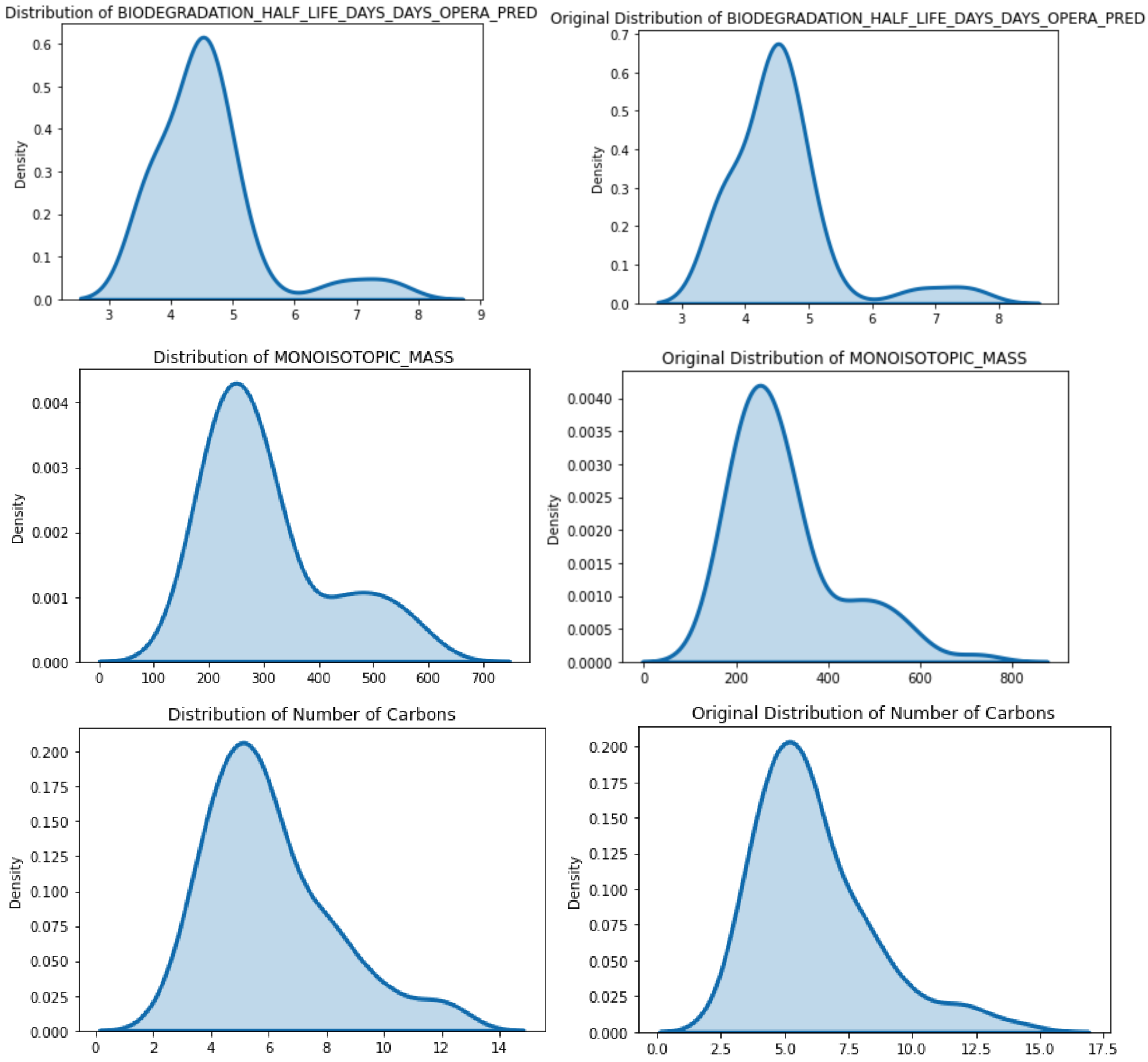
| | |
|---|---|
| medium | 0.415 |
| high | 0.338 |
| low | 0.246 |

Low halflife is pink, the medium halflife is green and the high halflife is purple in Figure 7.

# Results

## Distributions

To see if the refinement altered data distributions, figure 4 shows that the distributions are similar in the 65 data points and 74 data points for biodegradation, for monoisotopic mass and for the number of carbons the distribution is slightly shifted to the left after refinement. The distribution of octanol air partition and boiling point for example, could not be shown in the original data set using the matlab code because of the missing values keeping the data type from being numeric for these columns. Overall, comparing these distributions visually shows that the refinement was successful.



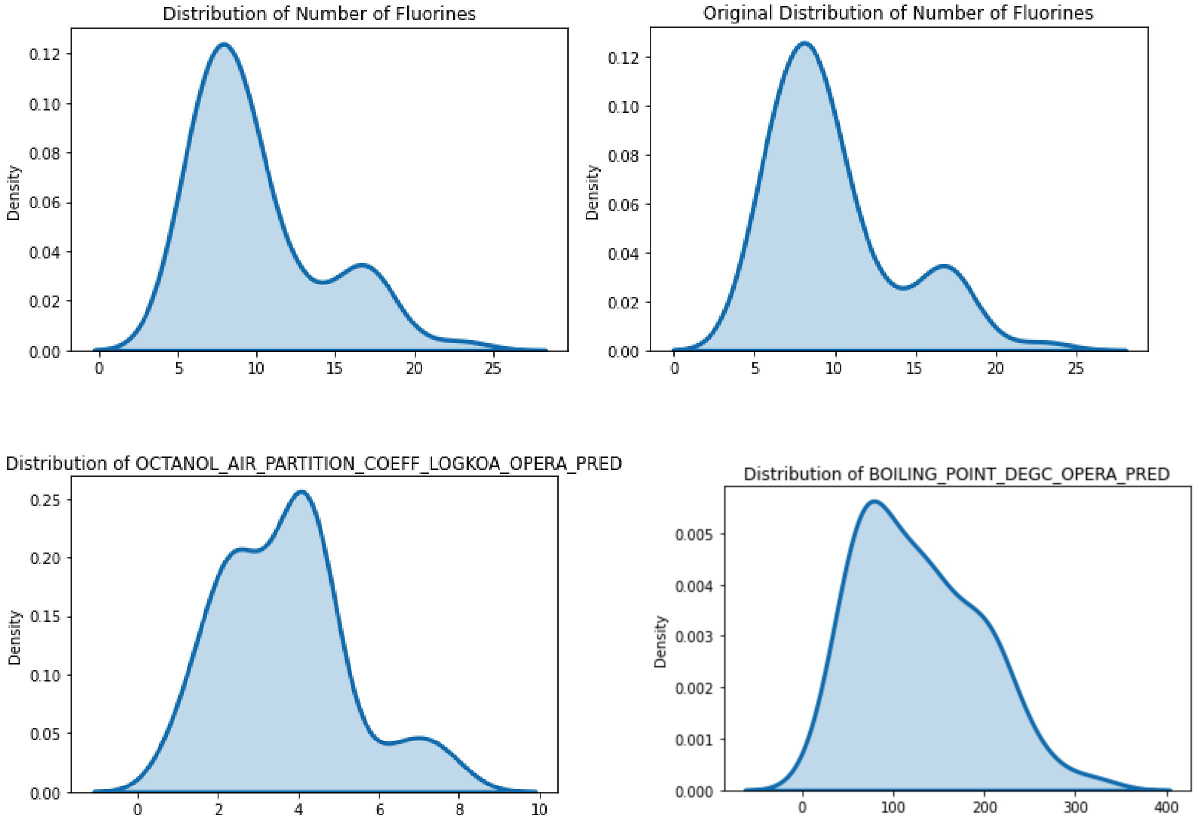


Figure 4. 6 distributions of a total 21 for the refined 65 data points compared to the original distribution if available in the 74 data points.

## Effect of Data Refinement

Table 1. Performance of RF Reg and XGB Reg, (rows, columns).

| Refined | # of Features | RF Reg MSE | XGB Reg MSE |
|---|---|---|---|
| (65,21) | 6 | 0.31 | 0.38 |
| | 18 | 0.42 | 1.3 |
| Not Refined | | | |
| (74,37) | 6 | 0.62 | 0.74 |
| | 16 | 0.99 | 0.81 |

Table 1 shows MSE scores for 4 different model scenarios. Random Forest had a lower MSE in 3/4 scenarios. All 4 models in Table 1 predicted biodegradation half-life [days]. Data refinement led to results with lower MSE. The RF reg with only 6 features had the lowest MSE and was further assessed. Bioconcentration factor & atmospheric hydroxylation were omitted as features, the latter because it had a value of 0 for every row and the former in case of omitted variable bias since like the predicted variable it is a biological factor.

## Predicting Bioconcentration Factor

It was attempted to try predicting the bioconcentration factor in place of biodegradation half life. Table 2 shows the results of this attempt. Because the MSE values are so high, the focus stayed on predicting the biodegradation half life and optimizing the model. Notably, unlike predicting biodegradation half life where RF Reg performed better, the XGB Reg performed better for predicting the bioconcentration factor.

Table 2. Performance of RF Reg and XGB Reg, (rows, columns).

| Refined | # of Features | RF Reg MSE | XGB Reg MSE |
|---|---|---|---|
| (65,21) | 6 | 1384.8 | 1221.1 |
|  | 18 | 458.8 | 261.2 |

## Feature Importance Overview

### Predicting Biodegradation Half Life

Refined 65 rows, 6 features’ importance is shown in Figure 5 and Table 3. As mentioned in Table 1 this model had an MSE of 0.307.

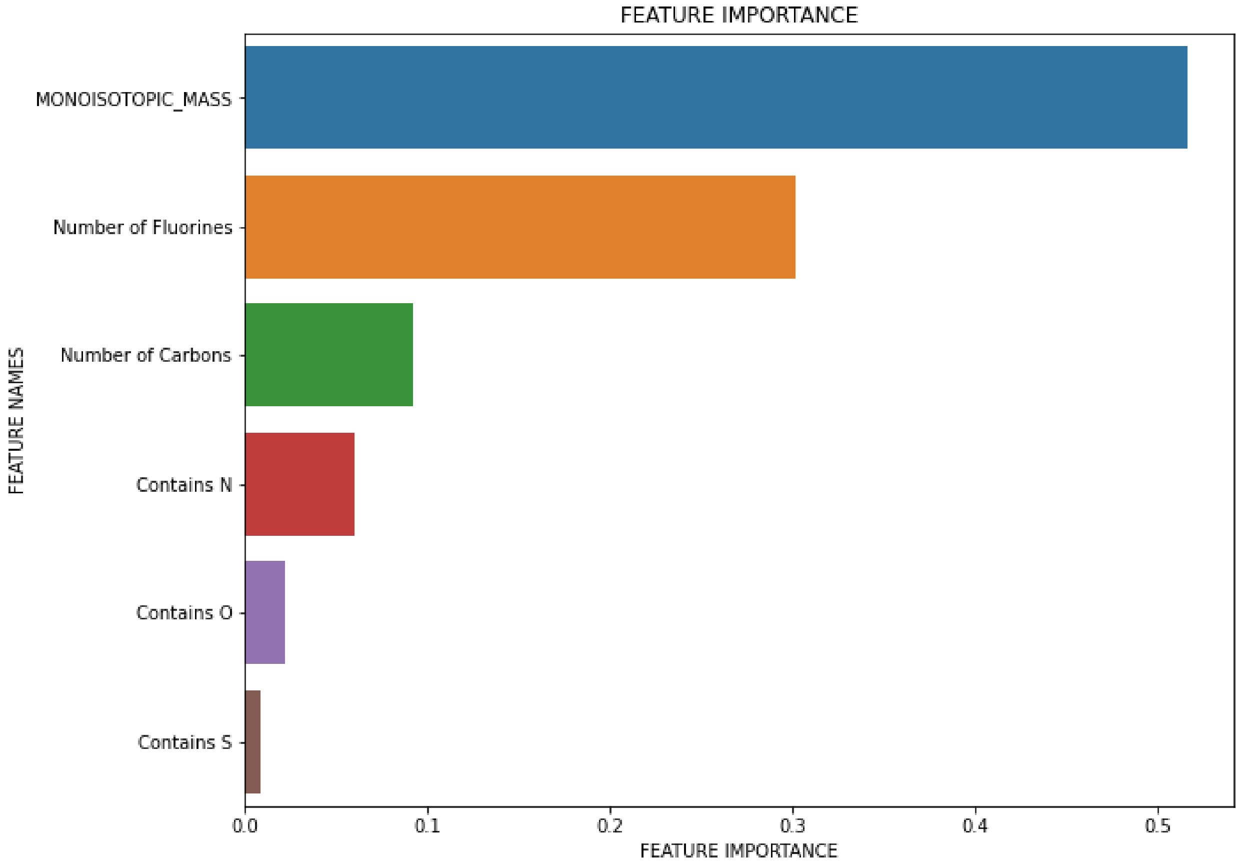


Figure 5. Refined RF Reg Feature Importance with 6 total features.

Table 3. Refined RF Reg Feature Importance with 6 total features.

```
Variable: MONOISOTOPIC_MASS    Importance: 0.52
Variable: Number of Fluorines  Importance: 0.3
Variable: Number of Carbons    Importance: 0.09
Variable: Contains N           Importance: 0.06
Variable: Contains O           Importance: 0.02
Variable: Contains S           Importance: 0.01
```

Refined 65 rows, 18 features' importance is shown in Figure 6 and Table 4. As mentioned in Table 1 this model had an MSE of 0.42.

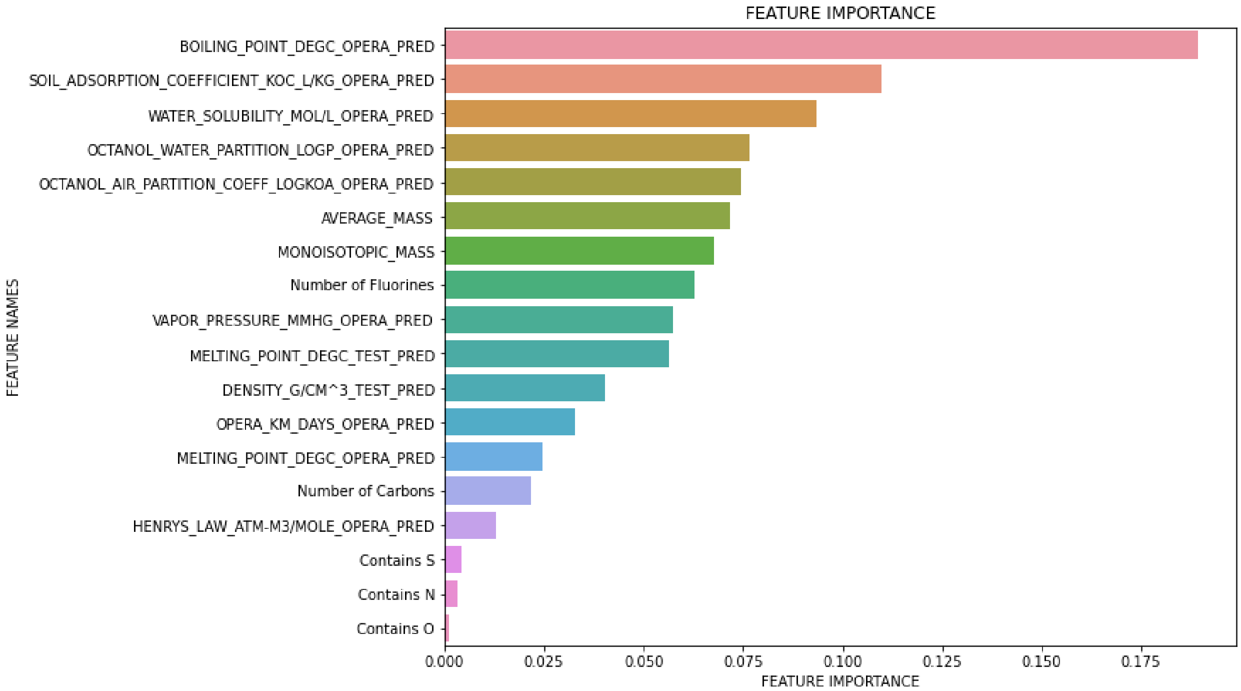


Figure 6. Refined RF Reg Feature Importance with 18 total features.

Table 4. Refined RF Reg Feature Importance with 18 total features.

```
Variable: BOILING_POINT_DEGC_OPERA_PRED Importance: 0.19
Variable: SOIL_ADSORPTION_COEFFICIENT_KOC_L/KG_OPERA_PRED Importance: 0.11
Variable: WATER_SOLUBILITY_MOL/L_OPERA_PRED Importance: 0.09
Variable: OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED Importance: 0.08
Variable: MONOISOTOPIC_MASS    Importance: 0.07
Variable: AVERAGE_MASS         Importance: 0.07
Variable: OCTANOL_AIR_PARTITION_COEFF_LOGKOA_OPERA_PRED Importance: 0.07
Variable: Number of Fluorines  Importance: 0.06
Variable: MELTING_POINT_DEGC_TEST_PRED Importance: 0.06
Variable: VAPOR_PRESSURE_MMHG_OPERA_PRED Importance: 0.06
Variable: DENSITY_G/CM^3_TEST_PRED Importance: 0.04
Variable: OPERA_KM_DAYS_OPERA_PRED Importance: 0.03
Variable: Number of Carbons    Importance: 0.02
Variable: MELTING_POINT_DEGC_OPERA_PRED Importance: 0.02
Variable: HENRYS_LAW_ATM-M3/MOLE_OPERA_PRED Importance: 0.01
Variable: Contains N           Importance: 0.0
Variable: Contains O           Importance: 0.0
Variable: Contains S           Importance: 0.0
```

# Correlations

For the following analysis, spearman's, pearson's and kendall's correlation coefficients were considered. As expected, there is a high correlation between average and monoisotopic mass. When both included monoisotopic mass has 0.29 feature importance and average mass 0.28. For this reason of duplicity, only monoisotopic mass will be included as a feature.

## Biodegradation and Bioconcentration

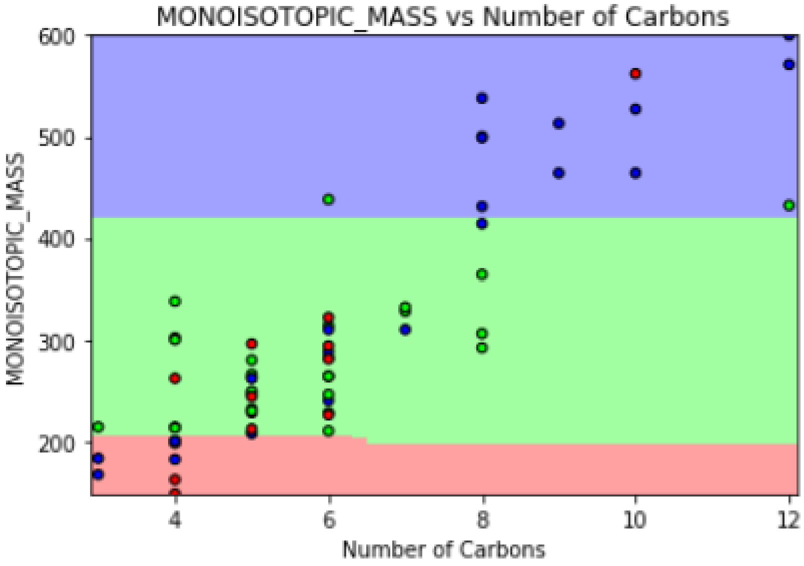
Biodegradation half-life and bioconcentration's correlations with the other columns were analyzed for a high absolute value of correlation of 0.6 and above. While biodegradation did not have any correlations meeting such conditions, bioconcentration did and is shown in Table 5.

Table 5. Results shown if absolute value is 0.6 and above. * on p-value if <=0.05 alpha.

| | BIOCONCENTRATION_ FACTOR_OPERA_PRED | Spearman's Rho (p-value) | Pearson's R (p-value) |
|---|---|---|---|
| WATER_SOLUBILITY_MOL/L_OPERA_PRED | x | -0.7751 (3.51971e-14)* | |
| SOIL_ADSORPTION_COEFFICIENT_KOC_L/KG_OPERA_PRED | x | 0.6657 (1.44345e-09)* | 0.63381 (1.4516e-08)* |

Monoisotopic mass' correlation with the other columns were analyzed for a high absolute value of correlation of 0.6 and above. The results in Table 6 include only the correlations which satisfy that condition. This feature was of interest because in the RF Reg of least MSE, it had the highest feature importance. Figure 7 shows the results from the KNeighborsClassifier; visually the respective 0.87 and 0.96 linear correlations are visible. Additionally, as the feature importance for these variables were highest in the RF Reg of least MSE, the half life categorical clustering: low (between 0 and 3.9 days) in pink, medium (greater than 3.9 up to 4.5 days) in green and high (greater than 4.5 up to 9 days) in purple is also visible.

## Monoisotopic Mass



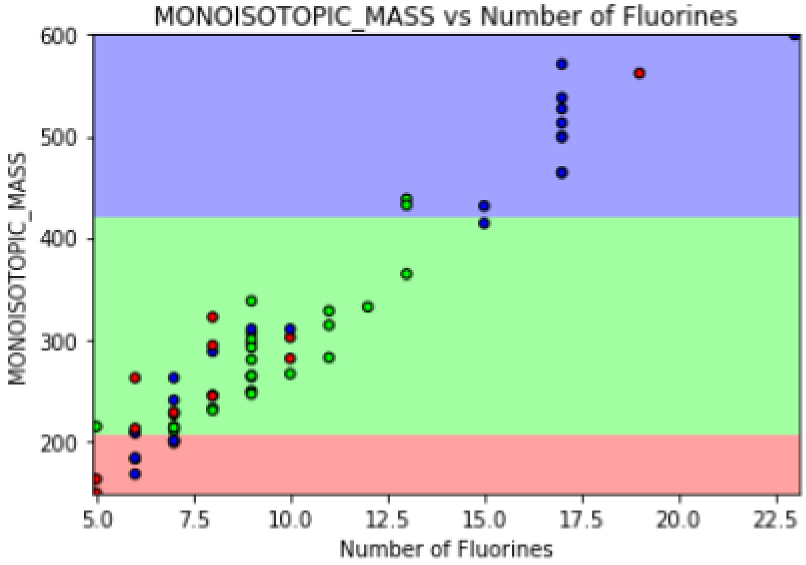


Figure 7. Top: Monoisotopic Mass vs Number of Carbons. Bottom: Monoisotopic Mass vs Number of Fluorines.

Table 6. Monoisotopic mass if absolute value is 0.6 and above. * on p-value if <=0.05 alpha.

| | MONOISOTOPIC _MASS | Spearman's Rho (p-value) | Pearson's R (p-value) | Kendall's Tau (p-value) |
|---|---|---|---|---|
| Number of Carbons | x | 0.80747 (4.459e-16)* | 0.8652 (1.507e-20)* | 0.6752 (1.111e-13)* |
| Number of Fluorines | x | 0.9281 | 0.964 | 0.8166 |

| | | | | |
|---|---|---|---|---|
| | | (9.91e-29)* | (4.683e-38)* | (5.399e-20)* |
| DENSITY_G/CM^3_TEST_PRED | x | 0.7572 (2.924e-13)* | 0.733 (3.976e-12)* | |
| BOILING_POINT_DEGC_OPERA_PRED | x | 0.6861 (2.838e-10)* | 0.641 (8.72e-09)* | |
| HENRYS_LAW_ATM-M3/MOLE_OPERA_PRED | x | -0.636 (1.271e-08)* | | |
| OPERA_KM_DAYS_OPERA_PRED | x | 0.6316 (1.682e-08)* | 0.8087 (3.707e-16) | |
| OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED | x | 0.601 (1.214e-07)* | | |
| MELTING_POINT_DEGC_OPERA_PRED | x | 0.66996 (1.04e-09)* | 0.644 (6.95e-09)* | |
| VAPOR_PRESSURE_MMHG_OPERA_PRED | x | -0.6471 (5.74e-09)* | | |

Further exploring possible omitted variable bias between bioconcentration factor and biological degradation,  low correlation indicates perhaps that omitting the bioconcentration factor was unnecessary. Bioconcentration factor and biological degradation Pearson's correlated 0.191 (p value 0.128), as shown in Figure 8, with Spearman's 0.129 (p value 0.31), Kendall's 0.095 (p value 0.265). In all cases, Monoisotopic mass is included in Figure 8 because it was the feature of top importance 0.5 when included, without the average mass feature, in the RF reg model. Interestingly, as Figure 5 shows, despite being a top feature, monoisotopic mass had low correlations with biodegradation half life. Monoisotopic mass and biodegradation Pearson's correlated 0.1699 (p value 0.18), Spearman's 0.345 (p value 0.0049) and Kendall's 0.259 (p value 0.0023). Monoisotopic mass with bioconcentration also had low correlations; Pearson's correlated 0.426 (p value 0.00041), Spearman's 0.258 (p value 0.038) and Kendall's 0.177 (p value 0.037).

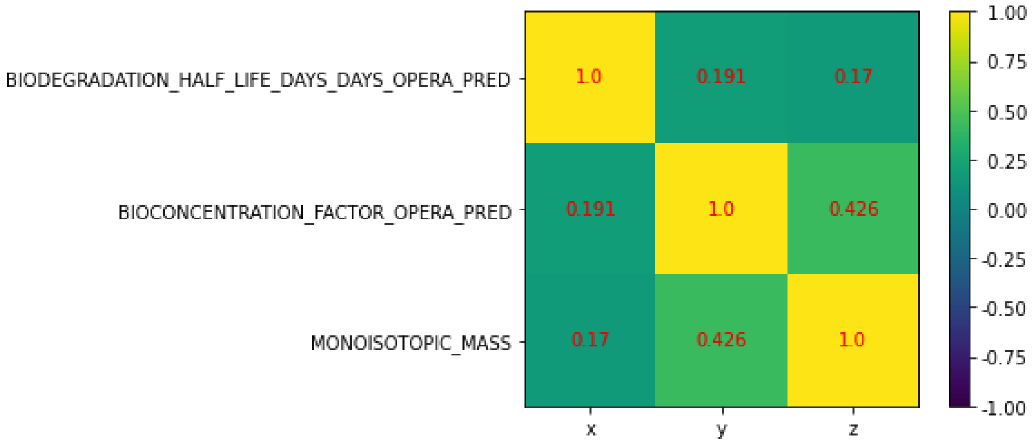


Figure 8. Pearson's R correlation values between monoisotopic mass, bioconcentration and biodegradation half-life.

## Model Optimization

Table 7. Performance of model predictions for different parameter optimization methods. The best results are highlighted.

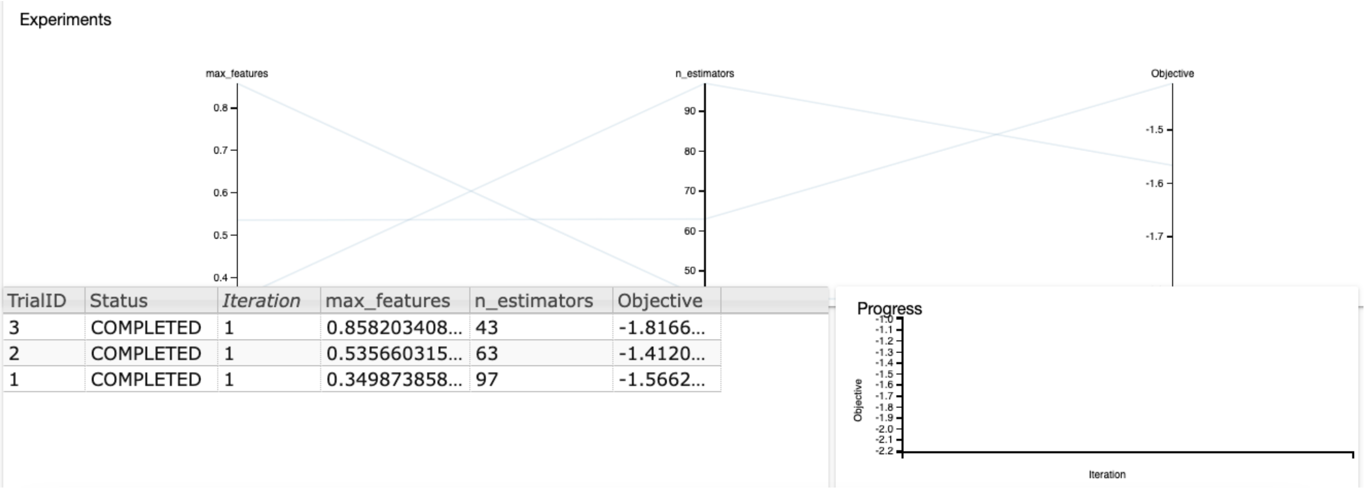
| | *# features* | *# rows* | *prediction variable* |
|---|---|---|---|
| | 6 | 65 | BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED |
| *Results* | *Default Parameters* | *RandomizedSearchCV* | *GridSearchCV* |
| *RMSE Test* | 0.550 | 0.604 | 0.604 |
| *RMSE Train* | 0.428 | 0.925 | 0.925 |
| *MSE Test* | 0.303 | 0.365 | 0.365 |
| *MSE Train* | 0.183 | 0.855 | 0.855 |
| *R2 Test* | 0.120 | -0.061 | -0.061 |
| *R2 Train* | 0.786 | 0.000 | 0.000 |

Table 8. Parameter Recommendations

| Parameter | min_samples_split | max_features | max_depth |
|---|---|---|---|
| *RandomizedSearchCV* | 1.0 | sqrt | 4 |

| | | | |
|---|---|---|---|
| *GridSearchCV* | 0.8 | auto | 4 |

GPyOpt Bayesian Optimization with Sherpa was attempted. Three trials were completed as shown in Figure 9 but for the fourth trial multiple runs led to an error, a score of not available nan. It was inconclusive as to why this occurred as of the date of this report.



| TrialID | Status | *Iteration* | max_features | n_estimators | Objective |
|---|---|---|---|---|---|
| 3 | COMPLETED | 1 | 0.858203408... | 43 | -1.8166... |
| 2 | COMPLETED | 1 | 0.535660315... | 63 | -1.4120... |
| 1 | COMPLETED | 1 | 0.349873858... | 97 | -1.5662... |

Figure 9. Sherpa Dashboard showing the three completed trials.

## Best Model

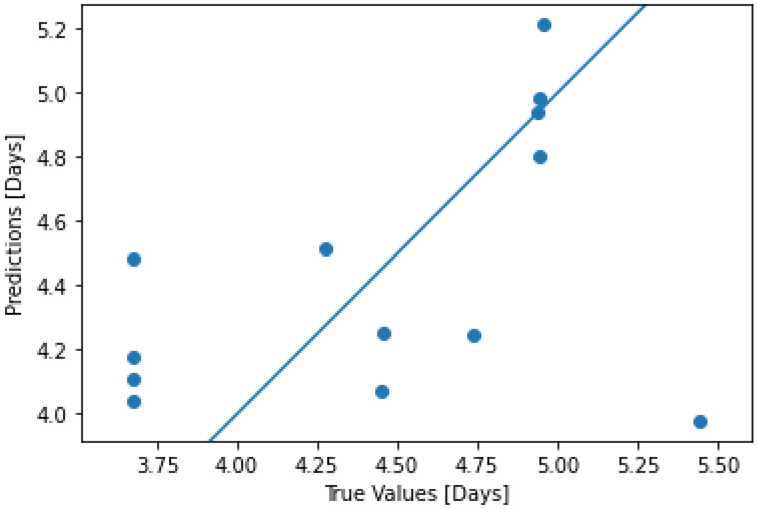


Figure 10. Default parameters model’s predictions [days] vs true values [days]

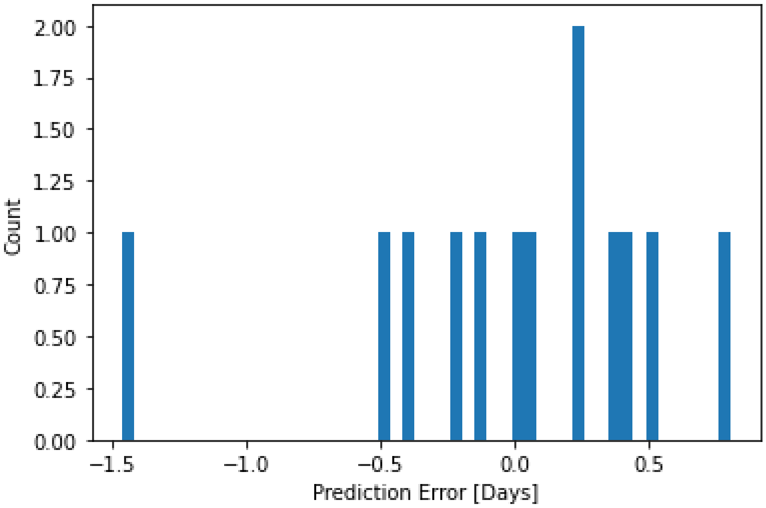


Figure 11. Default parameters model’s prediction error.

Table 9. Feature Importance ranking for the best model.

| | |
|---|---|
| Variable: MONOISOTOPIC_MASS | Importance: 0.51 |
| Variable: Number of Fluorines | Importance: 0.31 |
| Variable: Number of Carbons | Importance: 0.1 |
| Variable: Contains N | Importance: 0.05 |
| Variable: Contains O | Importance: 0.03 |
| Variable: Contains S | Importance: 0.01 |

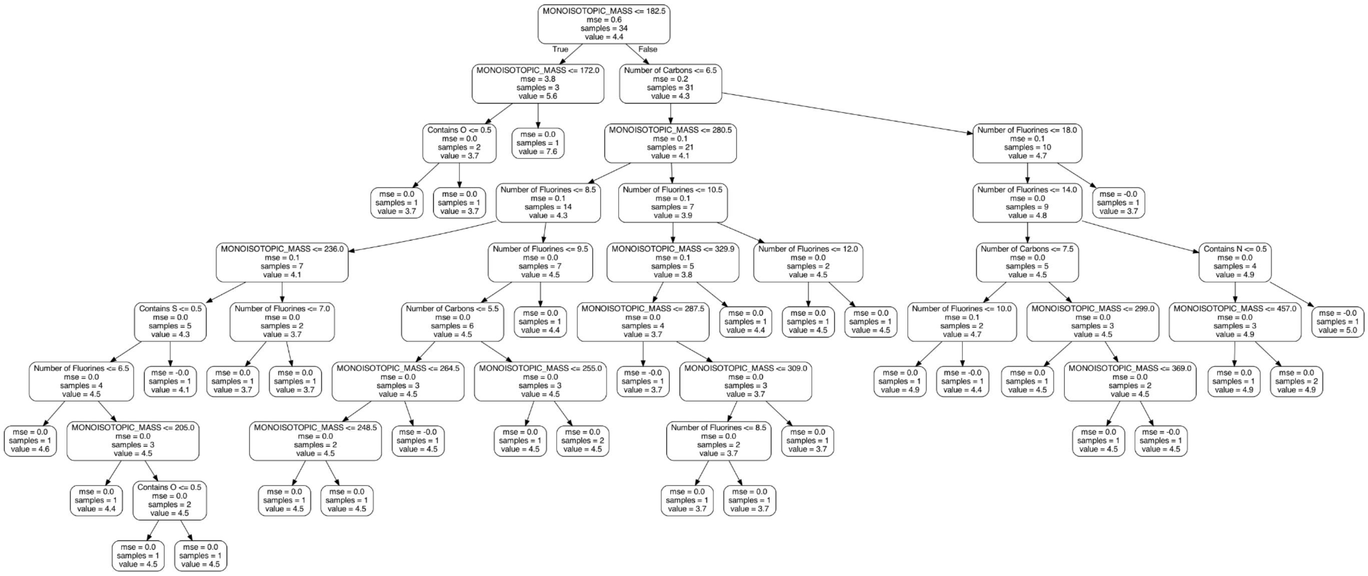

MONOISOTOPIC_MASS <= 182.5 mse = 0.6 samples = 34 value = 4.4
True
False
MONOISOTOPIC_MASS <= 172.0 mse = 3.8 samples = 3 value = 5.6
Number of Carbons <= 6.5 mse = 0.2 samples = 31 value = 4.3
Contains O <= 0.5 mse = 0.0 samples = 2 value = 3.7
mse = 0.0 samples = 1 value = 7.6
MONOISOTOPIC_MASS <= 280.5 mse = 0.1 samples = 21 value = 4.1
Number of Fluorines <= 18.0 mse = 0.1 samples = 10 value = 4.7
mse = 0.0 samples = 1 value = 3.7
mse = 0.0 samples = 1 value = 3.7
Number of Fluorines <= 8.5 mse = 0.1 samples = 14 value = 4.3
Number of Fluorines <= 10.5 mse = 0.1 samples = 7 value = 3.9
Number of Fluorines <= 14.0 mse = 0.0 samples = 9 value = 4.8
mse = -0.0 samples = 1 value = 3.7
MONOISOTOPIC_MASS <= 236.0 mse = 0.1 samples = 7 value = 4.1
Number of Fluorines <= 9.5 mse = 0.0 samples = 7 value = 4.5
MONOISOTOPIC_MASS <= 329.9 mse = 0.1 samples = 5 value = 3.8
Number of Fluorines <= 12.0 mse = 0.0 samples = 2 value = 4.5
Number of Carbons <= 7.5 mse = 0.0 samples = 5 value = 4.5
Contains N <= 0.5 mse = 0.0 samples = 4 value = 4.9
Contains S <= 0.5 mse = 0.0 samples = 5 value = 4.3
Number of Fluorines <= 7.0 mse = 0.0 samples = 2 value = 3.7
Number of Carbons <= 5.5 mse = 0.0 samples = 6 value = 4.5
mse = 0.0 samples = 1 value = 4.4
MONOISOTOPIC_MASS <= 287.5 mse = 0.0 samples = 4 value = 3.7
mse = 0.0 samples = 1 value = 4.4
mse = 0.0 samples = 1 value = 4.5
mse = 0.0 samples = 1 value = 4.5
Number of Fluorines <= 10.0 mse = 0.1 samples = 2 value = 4.7
MONOISOTOPIC_MASS <= 299.0 mse = 0.0 samples = 3 value = 4.5
MONOISOTOPIC_MASS <= 457.0 mse = 0.0 samples = 3 value = 4.9
mse = -0.0 samples = 1 value = 5.0
Number of Fluorines <= 6.5 mse = 0.0 samples = 4 value = 4.5
mse = -0.0 samples = 1 value = 4.1
mse = 0.0 samples = 1 value = 3.7
mse = 0.0 samples = 1 value = 3.7
MONOISOTOPIC_MASS <= 264.5 mse = 0.0 samples = 3 value = 4.5
MONOISOTOPIC_MASS <= 255.0 mse = 0.0 samples = 3 value = 4.5
mse = -0.0 samples = 1 value = 3.7
MONOISOTOPIC_MASS <= 309.0 mse = 0.0 samples = 3 value = 3.7
mse = 0.0 samples = 1 value = 4.9
mse = -0.0 samples = 1 value = 4.4
mse = 0.0 samples = 1 value = 4.5
MONOISOTOPIC_MASS <= 369.0 mse = 0.0 samples = 2 value = 4.5
mse = 0.0 samples = 1 value = 4.9
mse = 0.0 samples = 2 value = 4.9
mse = 0.0 samples = 1 value = 4.6
MONOISOTOPIC_MASS <= 205.0 mse = 0.0 samples = 3 value = 4.5
MONOISOTOPIC_MASS <= 248.5 mse = 0.0 samples = 2 value = 4.5
mse = -0.0 samples = 1 value = 4.5
mse = 0.0 samples = 1 value = 4.5
mse = 0.0 samples = 2 value = 4.5
Number of Fluorines <= 8.5 mse = 0.0 samples = 2 value = 3.7
mse = 0.0 samples = 1 value = 3.7
mse = 0.0 samples = 1 value = 4.5
mse = -0.0 samples = 1 value = 4.5
mse = 0.0 samples = 1 value = 4.4
Contains O <= 0.5 mse = 0.0 samples = 2 value = 4.5
mse = 0.0 samples = 1 value = 4.5
mse = 0.0 samples = 1 value = 4.5
mse = 0.0 samples = 1 value = 3.7
mse = 0.0 samples = 1 value = 3.7
mse = 0.0 samples = 1 value = 4.5
mse = 0.0 samples = 1 value = 4.5

Figure 12. The shape of the best model decision tree. Monoisotopic mass appears on top; the right branch has a number of fluorines on top.

# Discussion

## Data

Access to data could be improved by python web scraping methods to obtain more data from the CompTox Chemicals Dashboard. An EPA supervising toxicologist was contacted with a data request and in the future this could improve results. As of the date of this paper, the data has not been received.

## Parameters

A higher MSE when hypertuning parameters compared to no hypertuning was observed in Table 7. This can be lowered in a number of ways. First, trying an alternative hypertuning algorithm, second, increasing the parameter space for hypertuning by adding more parameters, third optimizing the range options for tuning of each parameter, and fourth increasing the number of trials, though this takes more processing power. The third way might require previous research examples or may be specific to the dataset. A further implementation with the first way could use the hyperopt framework, with the tool hyperas for simpler syntax.

## Correlations

Bioconcentration factor had a statistically significant negative Spearman's correlation of -0.78 with water solubility and a statistically significant positive Spearman's correlation of 0.67 with soil adsorption coefficient at the 1% level and therefore also satisfying the 5% and 10% levels. Returning to the definition of bioconcentration factor as "the ratio of the concentration of a substance in an organism to the concentration in water" (EPA 2021), it is perhaps expected that water solubility and bioconcentration would have an inverse relationship.

## Best Model

The best model was the default parameter Random Forest regression. On testing sets it had the lowest MSE 0.303 and the highest R2 0.120 of all the models examined in the report. The correlation is lower than ideal as is visually shown in Figure 10. The prediction error in Figure 11 shows a mostly normal distribution so there is likely no bias. The feature importance for the best model in table 9 is almost exactly the same as the RFReg model in Table 3 when a couple other parameters were specified in place of the default. The decision tree in Figure 12 frequently mentions monoisotopic mass which is not surprising given its high feature importance.

Interestingly the model's R2 in the testing data 0.120 is lower than the correlations between monoisotopic mass and biodegradation half life: Spearman's 0.345 (p value 0.0049) and Kendall's 0.259 (p value 0.0023). However the model's R2 in the training data 0.786 is higher than the correlations between monoisotopic mass and biodegradation half life.

It is also important to note that only the same 6 features identified as yielding a lower MSE in table 1 were used and given as options during parametrization optimization. It would be interesting to run the parameter optimization with all features except for the predicted variable to see if the MSE is lowered. Also it would be similarly interesting to see if the parameter optimization would help predicting the bioconcentration factor which was attempted in Table 2 and disregarded for the rest of the report.

## Interpretation of Feature Importance

The feature importance ranking for the best model is shown in table 9. The higher importance of mass and length of the carbon chain on half life prediction aligns with the present knowledge that longer chain PFAS have longer half lives than shorter chain PFAS.

# Conclusion

The model data set consisted of 65 PFAS analytes, 6 features shown in Table 9: MONOISOTOPIC_MASS, Number of Fluorines, Number of Carbons, Contains N, Contains O, Contains S and the predicted variable: Biodegradation Half Life Days Predicted Opera. The best model used default parameters and a Random Forest regression. The prediction performance had an MSE of 0.303 and an R2 of 0.120 in the testing data. The correlation is lower than ideal as is visually shown in Figure 10. The prediction error in Figure 11 shows a mostly normal distribution so there is likely no bias. In Table 9, the higher the importance of mass and length of the carbon chain on half life prediction aligns with the present knowledge that longer chain PFAS have longer half lives than shorter chain PFAS. A higher MSE when hypertuning parameters compared to no hypertuning was observed in Table 7. Access to data is the main factor identified for future improvement. Statistically significant correlations were found between data columns. For example, bioconcentration factor had a statistically significant negative Spearman’s correlation of -0.78 with water solubility and a statistically significant positive Spearman’s correlation of 0.67 with soil adsorption coefficient at the 1% level.

# References

American Water Works Association 2019. *Per- and Polyfluoroalkyl Substance (PFAS) Overview and Prevalence*. https://www.awwa.org/Portals/0/AWWA/ETS/Resources/Per-andPolyfluoroalkylSubstances(PFAS)-OverviewandPrevalence.pdf?ver=2019-08-14-090234-873

Anonymous 2018. *Hyperparameter Tuning in Random forest*. Stack overflow. https://stackoverflow.com/questions/53544996/hyperparameter-tuning-in-random-forest

Cheng and Ng 2019. *Using Machine Learning to Classify Bioactivity for 3486 Per- and Polyfluoroalkyl Substances (PFASs) from the OECD List*. Environmental Science & Technology. https://pubs.acs.org/doi/pdf/10.1021/acs.est.9b04833

Cousins et al. 2020. *Strategies for grouping per- and polyfluoroalkyl substances (PFAS) to protect human and environmental health*. Environmental Science: Processes & Impacts. https://pubs.rsc.org/en/content/articlelanding/2020/EM/D0EM00147C

EPA Accessed 2021. *CompTox Chemicals Dashboard*: https://comptox.epa.gov/dashboard/

European Union 2019. *Emerging chemical risks in Europe — 'PFAS*'. https://www.eea.europa.eu/publications/emerging-chemical-risks-in-europe

Makarow n.d. *Per- and polyfluoroalkyl substances (PFAS)*. Department of Ecology State of Washington. https://ecology.wa.gov/Waste-Toxics/Reducing-toxic-chemicals/Addressing-priority-toxic-chemicals/PFAS

Miller, Mark 2020. *Nothing Lasts Forever, Except PFAS: What Are They and What Can You Do About Them?*. https://www.kimley-horn.com/what-is-pfas/

Sarkar, Tushar 2021. *XBNet: An Extremely Boosted Neural Network*. KJ Somaiya College of Engineering, Mumbai. https://github.com/tusharsarkar3/XBNet/blob/master/Research_Paper/XBNET_paper.pdf

Hertel, Lars et al. 2021. *Sherpa: Robust Hyperparameter Optimization for Machine Learning*. SoftwareX. https://github.com/sherpa-ai/sherpa

Sneed 2021. *Forever Chemicals Are Widespread in U.S. Drinking Water*. Scientific American. https://www.scientificamerican.com/article/forever-chemicals-are-widespread-in-u-s-drinking-water/

Sprout n.d. *PFAS Crisis – The "Forever Chemicals" Found in 99% of Humans*. https://www.sproutsanfrancisco.com/get-educated/pfas-chemicals-crisis/