PFAS Toxicity Differentiation With Machine Learning

Isabela Yepes | January 2022

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 & Refinement

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 data was requested via email. To eliminate missing values, it was refined to 7,761 compounds. There were a total of 35 columns of numerical data. Explanations and names of these columns are included in Table 1. Of these, 22 columns were enhanced chemical descriptors generated using Excel formulas which calculated numerical values from the string/text columns: molecular formula and preferred name.

Table 1. 35 columns for the 7,761 data points.

Name	Explanation
MONOISOTOPIC_MASS	Molecular mass determined by the sum of the masses of the most abundant naturally occurring stable isotope of each atom in the molecule
Number of Carbons	Number of Carbons in the molecule
Number of Fluorines	Number of Fluorine in the molecule
Contains N	1 if nitrogen is in the molecule, 0 if not

Contains O	1 if oxygen in the molecule, 0 if not
ATMOSPHERIC_HYDROXYLATION_RATE_(AOH)_CM3/MOLECULE*SEC_OPERA_PRED	Rate at which the molecules enter the first step of decomposition in air, consists of an oxygen making an alcohol group
BIOCONCENTRATION_FACTOR_OPERA_PRED	The ratio of the concentration of a substance in an organism to the equilibrium concentration in water, a higher value indicates a higher bioaccumulation potential
BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED	Biodegradation half-life in days for compounds containing only carbon and hydrogen (i.e. hydrocarbons). Where the half life is the number of days for half of the molecules to degrade into environmentally acceptable products.
BOILING_POINT_DEGC_OPERA_PRED	The boiling point of a molecule in degrees Celsius
HENRYS_LAW_ATM-M3/MOLE_OPERA_PRED	the ratio of a compound's pressure in air to the concentration of the compound in water. Units of atmospheres for the air and of moles per liter for water
OPERA_KM_DAYS_OPERA_PRED	Half life in days, it is the whole body primary biotransformation rate constant for organic chemicals in fish.
OCTANOL_AIR_PARTITION_COEFF_LOGKOA_OPERA_PRED	the partitioning behavior of organic compounds between air and solid environmental matrices such as soil, vegetation, and aerosol particles
SOIL_ADSORPTION_COEFFICIENT_KOC_L/KG_OPERA_PRED	Ratio between the concentration of the substance in the soil and the concentration of the substance in the aqueous phase
OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED	Greater than one if a substance is more soluble in fat-like solvents such as n-octanol, and less than one if it is more soluble in water.
MELTING_POINT_DEGC_OPERA_PRED	The melting point of said molecule in degrees Celsius
VAPOR_PRESSURE_MMHG_OPERA_PRED	Pressure of vapor emitted by substance in mm of Mercury

WATER_SOLUBILITY_MOL/L_OPERA_PRED	Amount of chemical substances that can dissolve in water in moles per liter
Acid Group	1 if the preferred name contains “acid”, ”ic” or “ous”, 0 if not
Hyde Group	1 if the preferred name contains “hyde”, 0 if not. An aldehyde is a COH group on the end of the chain in molecule
Carbon Ring Group	1 if the preferred name contains “cyclo”, ”phen”, “benz” or “oxaspiro”, 0 if not.
Alcohol Group	1 if the preferred name contains “ol”, 0 if not. Meant to indicate a OH group in molecule
Ether Linkage	1 if the preferred name contains “oxy”, 0 if not. Meant to indicate an oxygen atom that connects two alkyl or aryl groups
Linear Vs. Branched	1 if the preferred name contains “n-”, 0 if not. N- groups signifies a standard non branched isomer molecule.
ide	1 if the preferred name contains “ide”, 0 if not. Non metal compounds/ salts
ate	1 if the preferred name contains “ate”, 0 if not. Polyatomic ion of oxygen with most oxygens
ite	1 if the preferred name contains “ite”, 0 if not. Polyatomic ion of oxygen with least oxygens
sodium	1 if the preferred name contains “sodium”, 0 if not. Sodium in molecule
polymer	1 if the preferred name contains “poly”, 0 if not. A polymer is made up of many repeating monomer units.
telomer	1 if the preferred name contains “telom”, 0 if not. Small fluorocarbon polymers which are usually precursors to other PFAS.
nitrile	1 if the preferred name contains “nitril”, 0 if not. A molecule with a Carbon triple bonded with a nitrogen
Bis	1 if the preferred name contains “bis”, 0 if not. It indicated if two identical but separated complex groups are in one molecule.

Amide	1 if the preferred name contains “amid”, 0 if not. A carbonyl group is linked to a nitrogen atom, or a compound that contains this functional group (CO=NH2).
Amine	1 if the preferred name contains “amin”, 0 if not. Nitrogen group where only 1 carbon atom is bonded directly to nitrogen atom.
Si	1 if the molecular formula contains “Si”, 0 if not. Silicon is in the molecule.
Sulfur	1 if the preferred name contains “sulf” or “thi”, 0 if not. Sulfur is in the molecule

TEST and OPERA are two prediction functionalities from the database that allow for data columns with the respective suffix labels (Kmansouri 2016). Due to TEST columns having increased levels of missing values, only OPERA is used in the features. Here are the definitions for both (EPA 2021):

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).

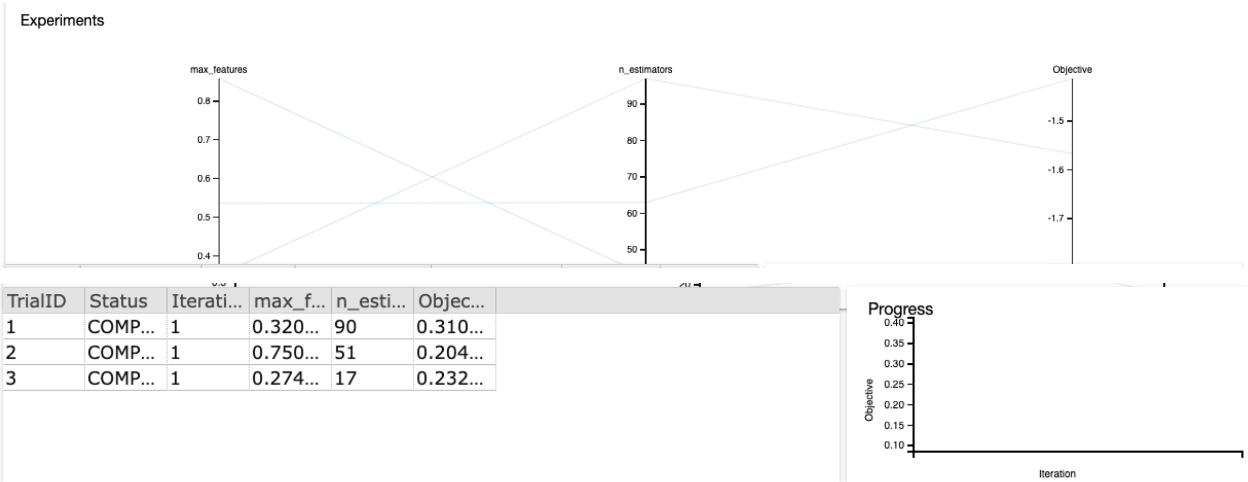
Within decision tree algorithms, boosted trees, such as in an XGBoost model, are traditionally more efficacious than non boosted trees, such as a Random Forest model (Sarkar 2021).

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. Additionally, as a verification a neural network model is tested for performance comparison. The toxicity will be proxied by the prediction of the feature: “biodegradation half life days opera predicted”, measured in days and then by the prediction of the feature: “bioconcentration factor opera predicted”, a numerical ratio. SInce for both predictions have a continuous scale, they are both regression problems.

Parameters

For comparing XGBoost Reg and RF(Random Forest) Reg default parameters were initially used. Parameter optimization for the Random Forest regression uses RandomizedSearchCV, GridSearchCV and GPyOpt Bayesian Optimization with Sherpa. Parameter optimization for the XGBoost regression uses RandomizedSearchCV. The parameter optimization results are included in the results section.

For the GPyOpt Bayesian Optimization with Sherpa, three trials were completed as shown in Figure 1 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. It happened when attempting to predict RF(Random Forest) Reg for the biodegradation half life variable and for the bioconcentration factor variable.



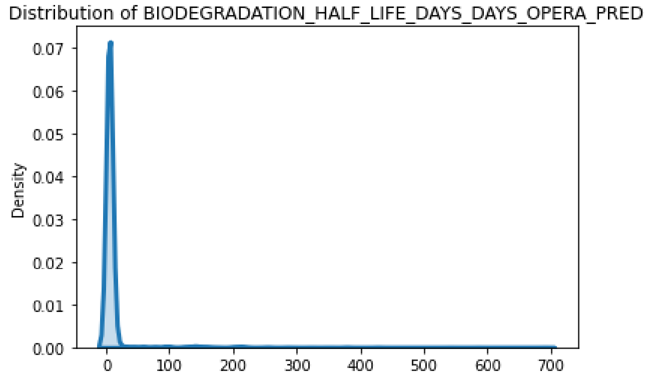
TrialID	Status	Iterati...	max_f...	n_esti...	Objec...
1	COMP...	1	0.320...	90	0.310...
2	COMP...	1	0.750...	51	0.204...
3	COMP...	1	0.274...	17	0.232...

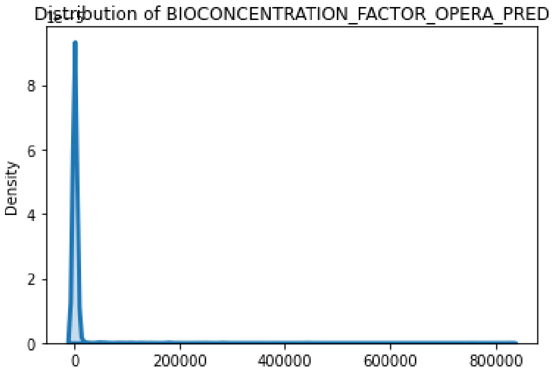
Figure 1. Sherpa Dashboard showing the three completed trials for predicting the bioconcentration factor.

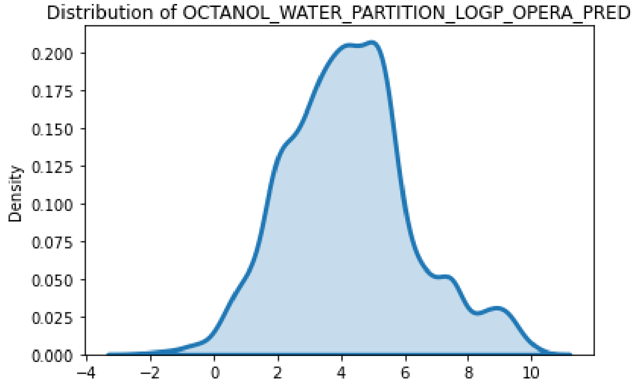
Results

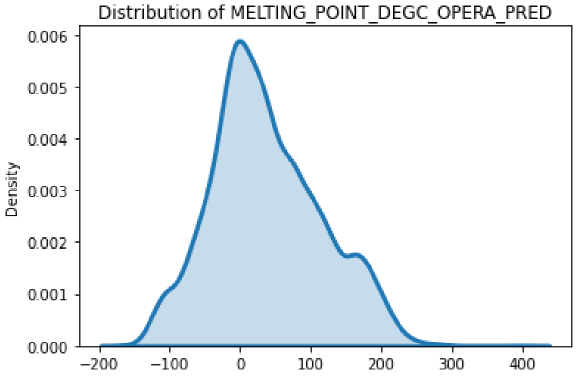
Distributions

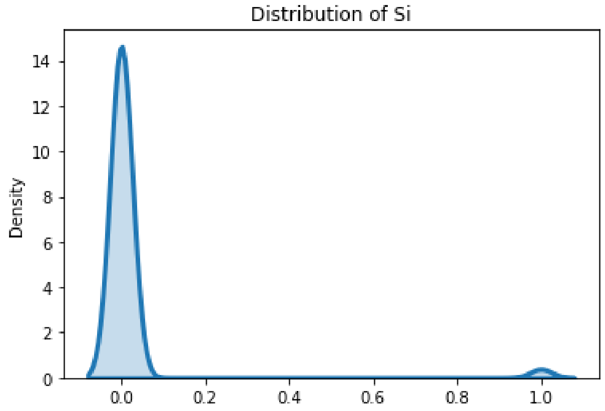
Figure 2 shows the distributions of the top features in RFReg and XGBReg.

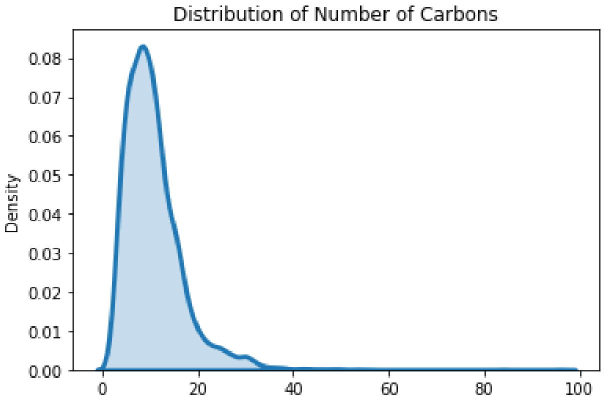

Distribution of BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED


Distribution of BIOCONCENTRATION_FACTOR_OPERA_PRED


Distribution of OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED


Distribution of MELTING_POINT_DEGC_OPERA_PRED


Distribution of Si


Distribution of Number of Carbons

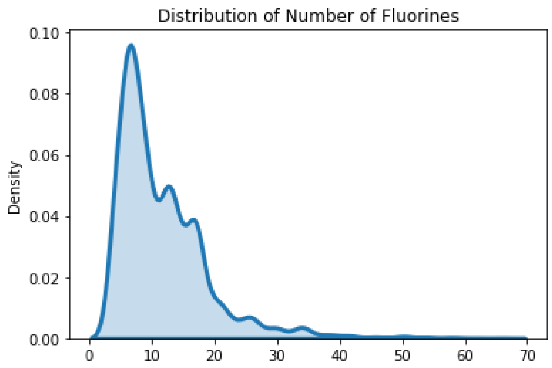


Figure 2. 7 distributions of a total 35 columns for the refined 7,761 data points.

Predicting Biodegradation Half Life

Random Forest Regression

Table 2. Biodegradation Half Life with Random Forest

	# features	*# rows*	*prediction variable*
	34	7761	BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED
Results	*Default Parameters*	*RandomizedSearchCV*	*GridSearchCV*
RMSE Test	17.306	22.558	22.558
RMSE Train	8.456	27.225	27.225
MSE Test	299.499	508.844	508.844
MSE Train	71.497	741.199	741.199
R2 Test	0.620	0.354	0.354
R2 Train	0.924	0.213	0.213

Table 3. Parameter Recommendations

Parameter	min_samples_split	max_features	max_depth
RandomizedSearchCV	0.6	‘auto’	6
GridSearchCV	0.6	‘auto’	13

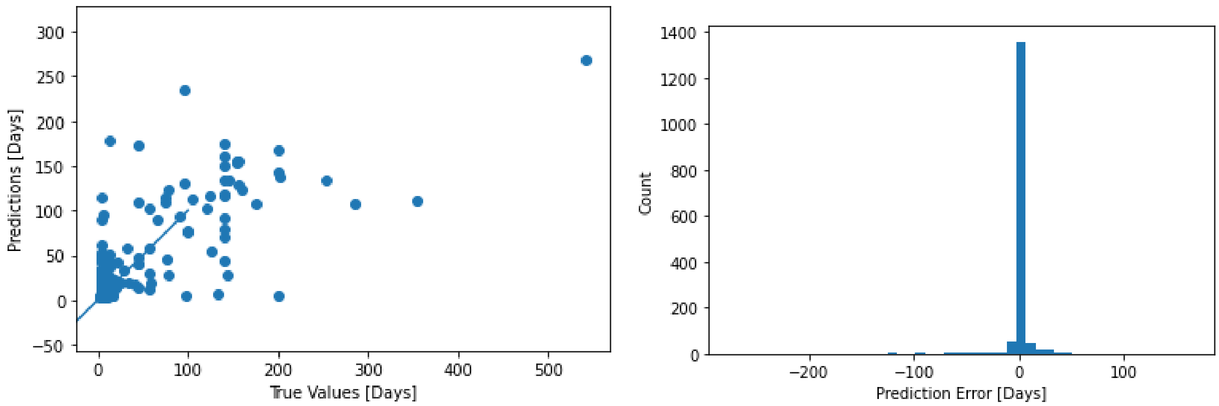


Figure 3. For the RF(Random Forest) Reg with default parameters, on the left, its scatter plot which shows the correlation between the predicted and the true values and on the right, the prediction error which shows minimal bias.

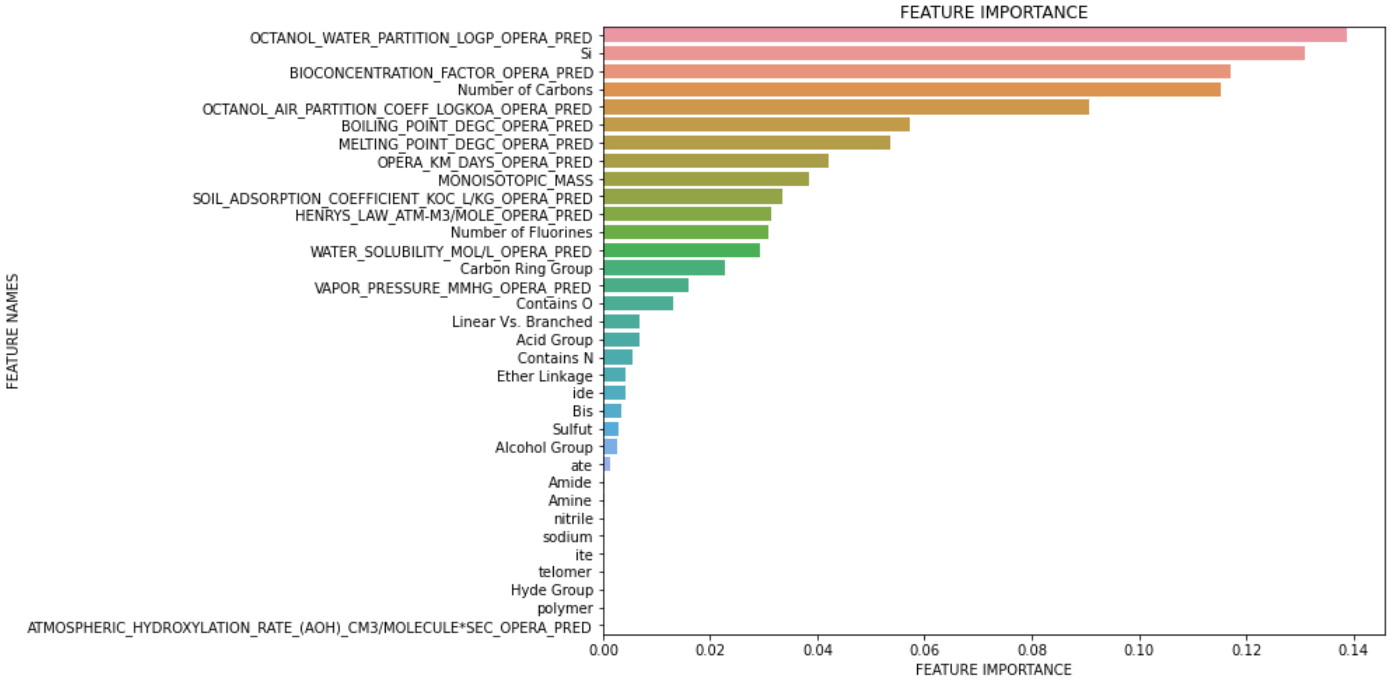


Figure 4. Feature importance chart for the RF Reg with default parameters which has the highest test R^2 of 0.620.

XGBoost Regression

Table 4. Biodegradation Half Life with XGBoost

	# features	*# rows*	*prediction variable*
	34	7761	BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED
Results	*Default Parameters*	*RandomizedSearchCV*	
RMSE Test	19.369	17.844	
RMSE Train	1.260	5.026	
MSE Test	375.152	318.419	
MSE Train	1.588	25.261	
R2 Test	0.523	0.595	
R2 Train	0.998	0.973	

Table 5. Parameter Recommendations

Parameter	objective	learning_rate	max_depth
RandomizedSearchCV	'reg:squarederror'	0.3	4

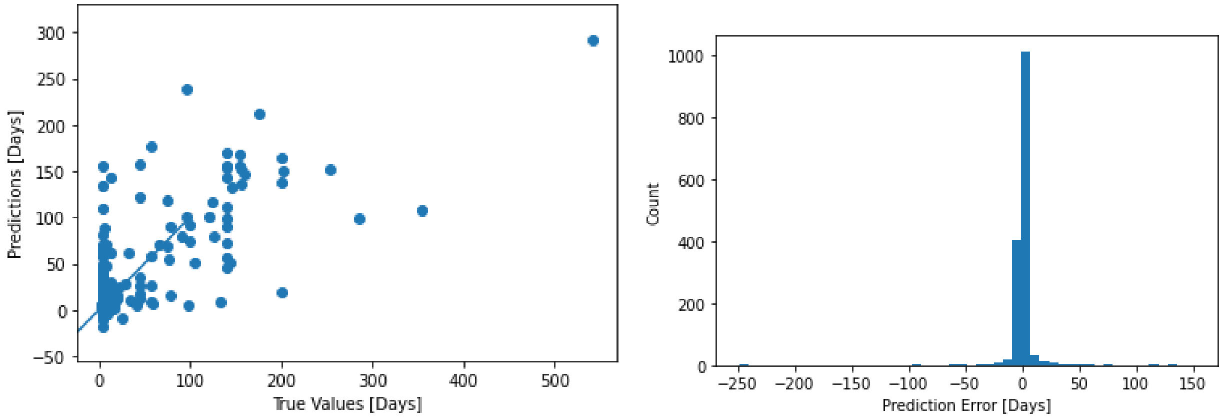


Figure 5. For the XGBoost Reg with RandomizedSearchCV parameters, on the left, its scatter plot which shows the correlation between the predicted and the true values and on the right, the prediction error which shows minimal bias.

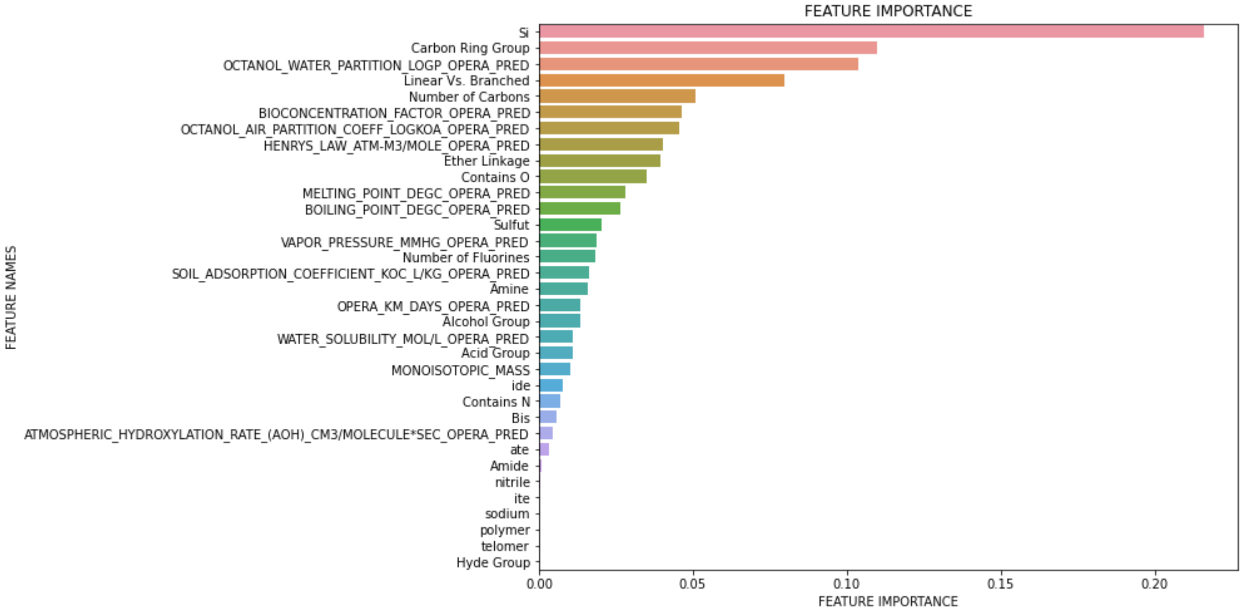


Figure 6. Feature importance chart for the XGBoost Reg with RandomizedSearchCv parameters which has the highest test R^2 of 0.595.

Neural Network Regression

The neural network is sequential with three dense hidden layers and one dense output layer. There were 37,633 total params. Figure 7 and table 6 show the results of training:

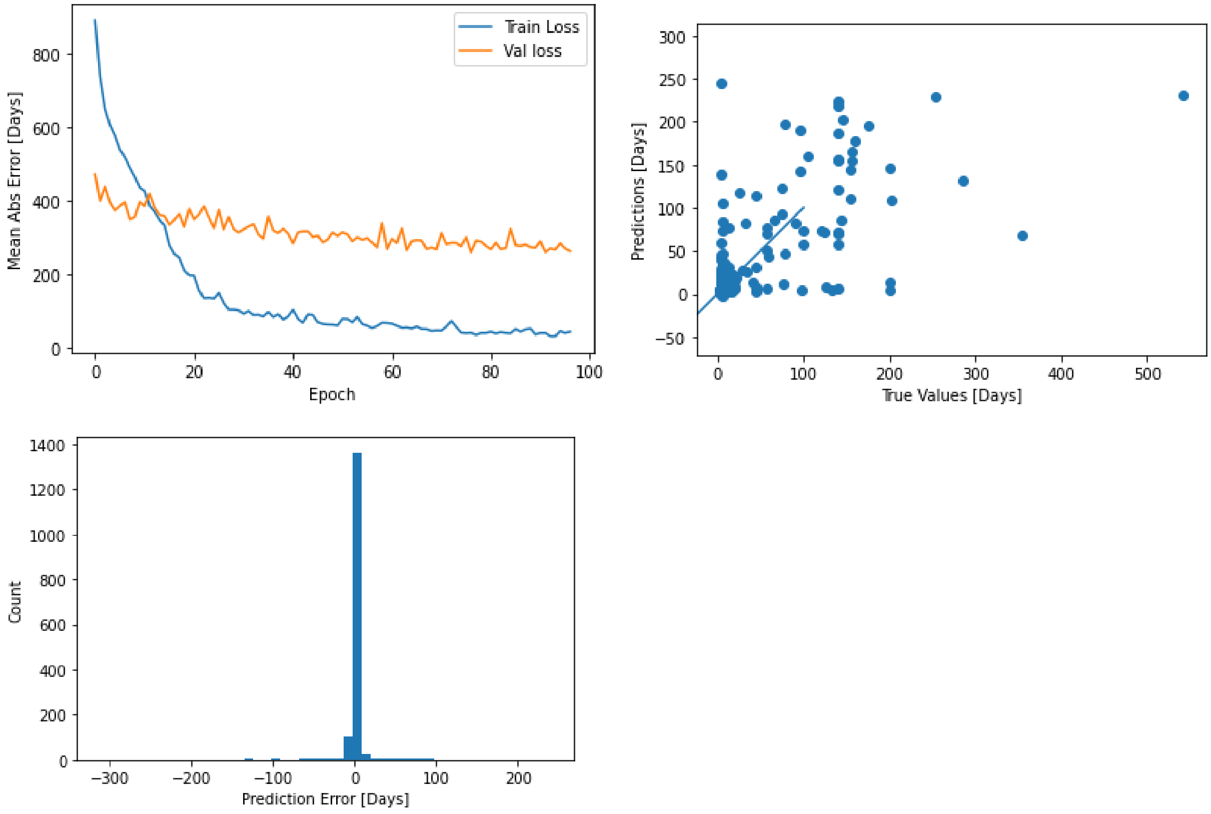


Figure 7. On the top left, a plot of history for neural network. On the top right, a scatter plot which shows the correlation between the predicted and the true values. On the bottom, the prediction error which shows minimal bias.

Table 6. Results of the neural network model.

Testing set Mean Abs Error	4.12 Days
corrcoef	0.732
R_2 score	0.525

Predicting Bioconcentration Factor

Random Forest Regression

Table 7. Bioconcentration Factor with Random Forest

	# features	*# rows*	*prediction variable*
	34	7761	BIOCONCENTRATION_FACTOR_OPERA_PRED
Results	*Default Parameters*	*RandomizedSearchCV*	*GridSearchCV*
RMSE Test	17375.424	26348.781	22001.663
RMSE Train	7467.059	21944.543	22001.663
MSE Test	301905367.364	694258251.777	691821873.003
MSE Train	55756974.037	481562976.484	484073191.613
R2 Test	0.604	0.090	0.094
R2 Train	0.898	0.121	0.117

Table 8. Parameter Recommendations

Parameter	min_samples_split	max_features	max_depth
RandomizedSearchCV	0.1	‘sqrt’	8
GridSearchCV	0.1	‘log2’	3

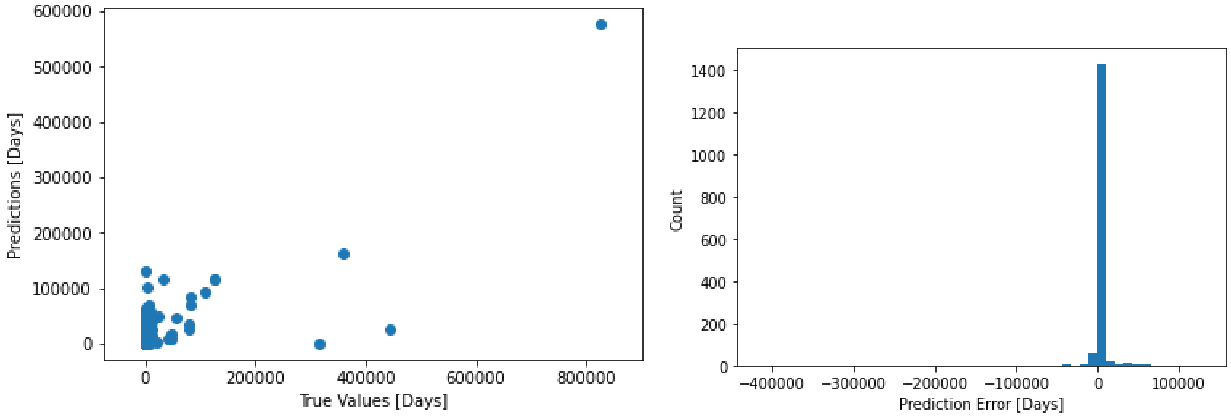


Figure 8. For the RF(Random Forest) Reg with default parameters, on the left, its scatter plot which shows the correlation between the predicted and the true values and on the right, the prediction error which shows minimal bias.

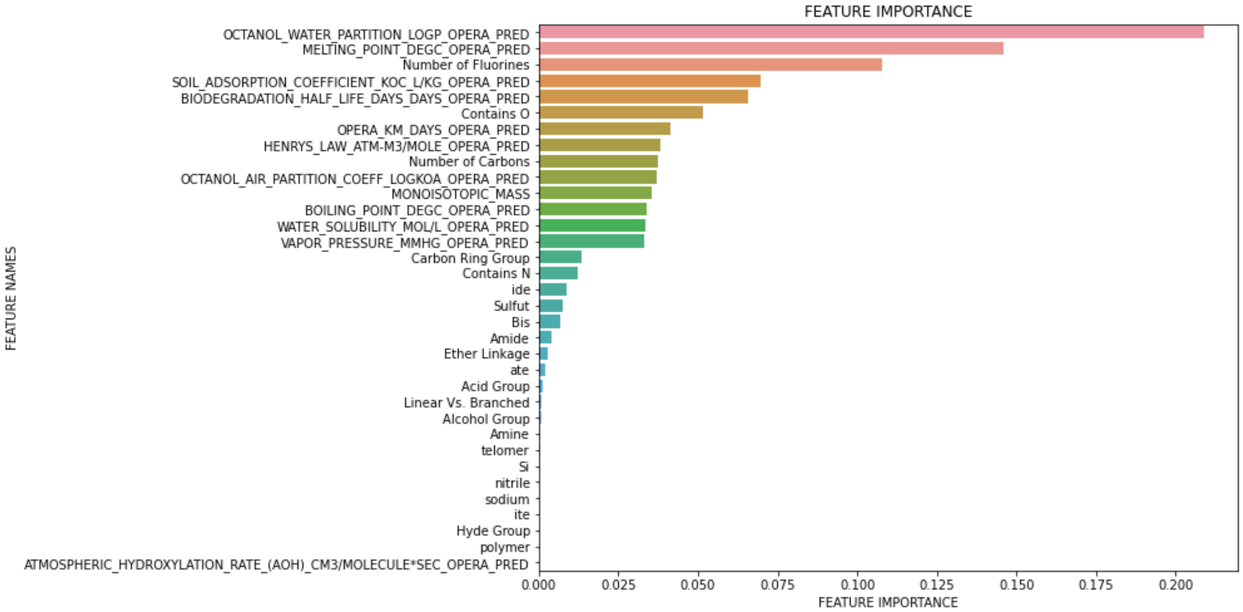


Figure 9. Feature importance chart for the RF Reg with default parameters which has the highest test R^2 of 0.604.

XGBoost Regression

Table 9. Bioconcentration Factor with XGBoost

	# features	*# rows*	*prediction variable*
	34	7761	BIOCONCENTRATION_FACTOR_OPERA_PRED
Results	*Default Parameters*	*RandomizedSearchCV*	

RMSE Test	16727.022	27707.648	
RMSE Train	325.719	23510.214	
MSE Test	279793263.415	767713749.248	
MSE Train	106092.604	552730156.010	
R2 Test	0.633	-0.006	
R2 Train	0.9998	-0.009	

Table 10. Parameter Recommendations

Parameter	objective	learning_rate	max_depth
RandomizedSearchCV	'reg:squaredlogerror'	0.9	11

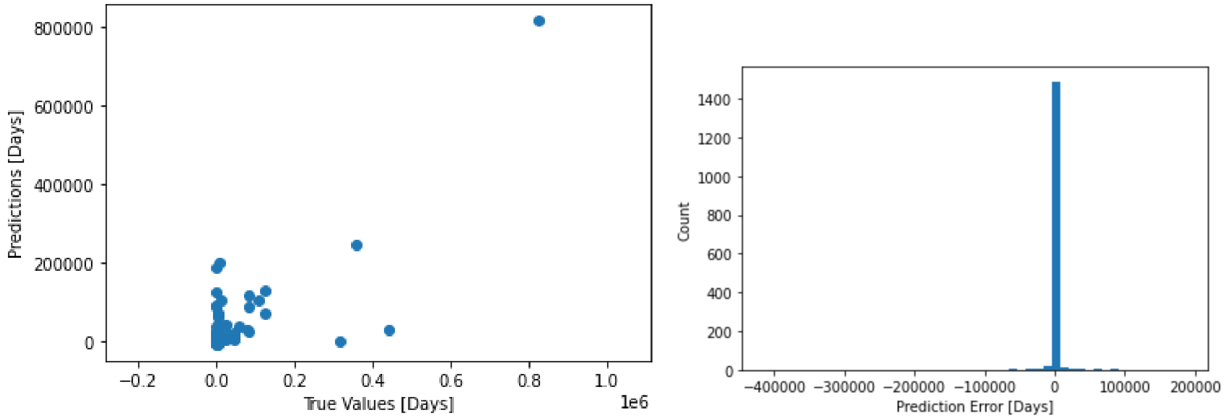


Figure 10. For the XGBoost Reg with default parameters, on the left, its scatter plot which shows the correlation between the predicted and the true values and on the right, the prediction error which shows minimal bias.

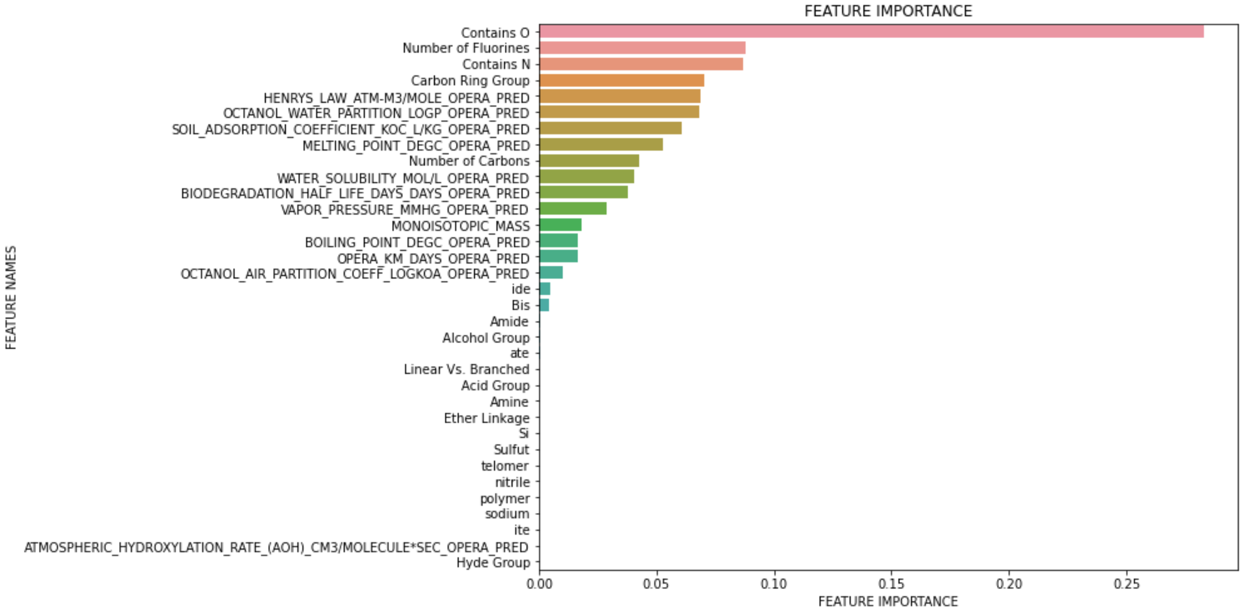


Figure 11. Feature importance chart for the XGBoost Reg with default parameters which has the highest test R^2 of 0.633.

Neural Network Regression

The neural network is sequential with three dense hidden layers and one dense output layer. There were 37,633 total params. Figure 12 and table 11 show the results of training:

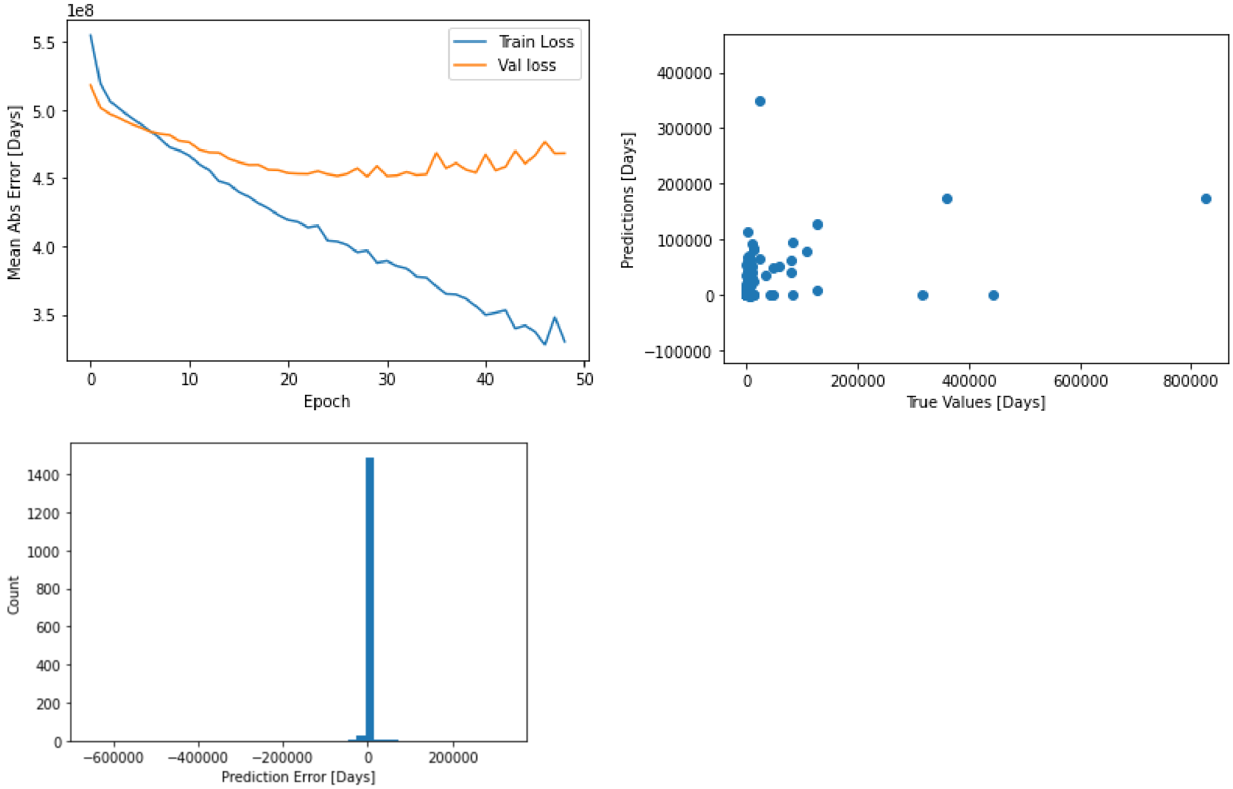


Figure 12. On the top left, a plot of history for neural network. On the top right, a scatter plot which shows the correlation between the predicted and the true values. On the bottom, the prediction error which shows minimal bias.

Table 11. Results of the neural network model.

Testing set Mean Abs Error	2915.86 Days
corrcoef	0.438
R_2 score	0.187

Best Models

The best model for predicting the biodegradation half-life in days is a RF(Random Forest) Reg with default parameters due to its R^2 of 0.620. The best model for predicting the bioconcentration factor is an XGBoost Reg with default parameters due to its R^2 of 0.633.

Table 12. Feature Importance ranking for the best model for predicting the biodegradation half-life in days.

```
Variable: OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED Importance: 0.14
Variable: Si                 Importance: 0.13
Variable: Number of Carbons    Importance: 0.12
Variable: BIOCONCENTRATION_FACTOR_OPERA_PRED Importance: 0.12
Variable: OCTANOL_AIR_PARTITION_COEFF_LOGKOA_OPERA_PRED Importance: 0.09
Variable: BOILING_POINT_DEGC_OPERA_PRED Importance: 0.06
Variable: MELTING_POINT_DEGC_OPERA_PRED Importance: 0.05
Variable: MONOISOTOPIC_MASS    Importance: 0.04
Variable: OPERA_KM_DAYS_OPERA_PRED Importance: 0.04
Variable: Number of Fluorines  Importance: 0.03
Variable: HENRYS_LAW_ATM-M3/MOLE_OPERA_PRED Importance: 0.03
Variable: SOIL_ADSORPTION_COEFFICIENT_KOC_L/KG_OPERA_PRED Importance: 0.03
Variable: WATER_SOLUBILITY_MOL/L_OPERA_PRED Importance: 0.03
Variable: VAPOR_PRESSURE_MMHG_OPERA_PRED Importance: 0.02
Variable: Carbon Ring Group    Importance: 0.02
Variable: Contains N          Importance: 0.01
Variable: Contains O          Importance: 0.01
Variable: Acid Group          Importance: 0.01
Variable: Linear Vs. Branched  Importance: 0.01
```

Table 13. Feature Importance ranking for the best model for predicting the bioconcentration factor.

```
Variable: Contains O          Importance: 0.2800000011920929
Variable: Number of Fluorines  Importance: 0.09000000357627869
Variable: Contains N          Importance: 0.09000000357627869
```

```
Variable: HENRYS_LAW_ATM-M3/MOLE_OPERA_PRED Importance:
0.07000000029802322
Variable: OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED Importance:
0.07000000029802322
Variable: Carbon Ring Group    Importance: 0.07000000029802322
Variable: SOIL_ADSORPTION_COEFFICIENT_KOC_L/KG_OPERA_PRED Importance:
0.05999999865889549
Variable: MELTING_POINT_DEGC_OPERA_PRED Importance: 0.05000000074505806
Variable: Number of Carbons    Importance: 0.03999999910593033
Variable: BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED Importance:
0.03999999910593033
Variable: WATER_SOLUBILITY_MOL/L_OPERA_PRED Importance: 0.03999999910593033
Variable: VAPOR_PRESSURE_MMHG_OPERA_PRED Importance: 0.029999999329447746
Variable: MONOISOTOPIC_MASS    Importance: 0.019999999552965164
Variable: BOILING_POINT_DEGC_OPERA_PRED Importance: 0.019999999552965164
Variable: OPERA_KM_DAYS_OPERA_PRED Importance: 0.019999999552965164
Variable: OCTANOL_AIR_PARTITION_COEFF_LOGKOA_OPERA_PRED Importance:
0.009999999776482582
```

Correlations

The top 7 features of the best models were evaluated for correlations with the predicted variable. Table 14 shows the correlations with biodegradation and table 15 shows the correlations with bioconcentration. Table 14 and 15 have correlations with absolute values greater than or equal to .5 underlined, indicating they have strong correlation.

Table 14. Correlations of features and BIODEGRADATION_HALF_LIFE_DAYS_DAYS_OPERA_PRED. The ** indicates the result is statistically significant at the alpha=0.01 level.

	OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED	Si	Number of Carbons	BIOCONCENTRATION_FACTOR_OPERA_PRED	OCTANOL_AIR_PARTITION_COEFF_LOGKOA_OPERA_PRED	BOILING_POINT_DEGC_OPERA_PRED	MELTING_POINT_DEGC_OPERA_PRED
Pearson’s	0.210**	0.377**	0.245**	0.075**	0.137**	0.167**	0.016**
Spearman’s	0.281**	0.146**	0.201**	0.296**	0.039**	0.049**	-0.074**
Kendall’s	0.194**	0.119**	0.151**	0.199**	0.034**	0.042**	-0.047**

Table 15. Correlations of features and BIOCONCENTRATION_FACTOR_OPERA_PRED. The ** indicates the result is statistically significant at the alpha=0.01 level, * at only the 0.05 level.

	Contains O	Number of Fluorines	Contains N	HENRYS_LAW_ATM-M3/MOLE_OPERA_PRED	OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED	Carbon Ring Group	SOIL_ADSORPTION_COEFFICIENT_KOC_L/KG_OPERA_PRED
Pearson’s	-0.104**	0.216**	-0.031**	-0.003	0.170**	0.062**	0.107**
Spearman’s	-0.404**	0.409**	-0.162**	0.170**	0.655**	0.169**	0.676 **
Kendall’s	-0.330**	0.289**	-0.132**	0.108**	0.467 **	0.138**	0.483 **

Discussion

Data

Data from the CompTox Chemicals Dashboard was obtained by contacting an EPA toxicologist. The obtained data included 12,039 annalytes. For the machine learning models to run, rows and columns with missing values had to be deleted. Eventually the list was refined to 7,761 analytes which all had no missing values.

Models & Parameters

The toxicity was proxied by the prediction of the features: “biodegradation half life days opera predicted”, measured in days and then by the prediction of the feature: “bioconcentration factor opera predicted”, a numerical ratio. Since both predictions have a continuous scale, they are both regression problems. Since the data is structured tabular data it was expected that tree based algorithms would outperform neural networks. However, as a verification a neural network model was tested for performance comparison. As expected, the tree-based algorithms performed better in terms of R^2. Notably, the neural network performed better at predicting the biodegradation half life (R^2 of 0.525) than at predicting the bioconcentration factor (R^2 of 0.187).

The best model for predicting the biodegradation half-life in days is a RF(Random Forest) Reg with default parameters due to its R^2 of 0.620. The best model for predicting the bioconcentration factor is an XGBoost Reg with default parameters due to its R^2 of 0.633. The prediction errors for these models shown in Figure 3 and Figure 9 show a mostly normal distribution so there is likely no bias. Likewise the prediction errors for the other models also show likely no bias. This can be interpreted to reflect the random representation from splitting into testing and training samples. All model runs in this report were made with all 35 variables, 34 as features and 1 as the predicted variable.

Adding the 22 columns of chemical binary descriptors generated using Excel formulas greatly increased R^2 values. Without the 22 columns, the test R^2 for the RF(Random Forest) Reg with default parameters was 0.477. While still a moderate correlation without them, the addition of the 22 columns enhanced the ability to differentiate the PFAS toxicity, resulting in a stronger correlation.

Parameter optimization was attempted for the RF(Random Forest) Reg and the XGBoost Reg. Default parameters yielded the highest R^2 for predicting both variables. Parameter optimization for the Random Forest regression used RandomizedSearchCV and GridSearchCV. Parameter optimization for the XGBoost regression used RandomizedSearchCV. Only in the case of XGBoost used for predicting biodegradation half-life in days did the RandomizedSearchCV tuning yield a higher R^2 (R^2 of 0.595) than the default parameters (R^2 of 0.523).

Correlations

Bioconcentration factor had a statistically significant positive Spearman’s correlation of 0.655 with Octanol water partition and bioconcentration factor had a statistically significant positive Spearman’s correlation of 0.676 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 there is a direct relationship with Octanol water partition since high Octanol water partition coefficients tend to bioaccumulate in the fatty tissue of organisms.

Interpretation of Feature Importance

The feature importance rankings for the best models are shown in table 12 and 13 for predicting biodegradation and bioconcentration, respectively. The higher importance of number of carbons (0.12) for predicting the biodegradation half life aligns with the present knowledge that longer chain PFAS have longer half lives than shorter chain PFAS. Interestingly, though bioconcentration had a higher correlation with octanol water partition, octanol water partition is of higher feature importance for predicting the biodegradation half life (0.14) than it is for predicting the bioconcentration factor (0.07). Reasons for the high importance of features such as Contains O (0.28), Number of Fluorines (0.09), Contains N (0.09) for predicting the bioconcentration factor can be further investigated. Reasons for the high importance of features such as Si (0.13) and Octanol air partition (0.09) for predicting the biodegradation half-life can also be further investigated. It is possible the boiling point importance (0.06) for predicting biodegradation half life can reflect that higher boiling point molecules contain bonds that require more energy to break and thus have a longer half life. Interestingly, despite the bioconcentration factor and the biodegradation half life having a low correlation (0.296) as shown in Table 14, the bioconcentration factor had a high importance (0.12) in predicting the biodegradation half life. Meanwhile, the biodegradation half life had a low importance (0.04) in predicting the bioconcentration factor.

Conclusion

PFAS toxicity was proxied by the prediction of the features: “biodegradation half life days opera predicted”, measured in days and then by the prediction of the feature: “bioconcentration factor opera predicted”, a numerical ratio. The model data set consisted of 7,761 PFAS analytes, 35 columns, 34 features and 1 prediction variable. Random Forest Regression, XGBoost Regression and Neural Networks were used as models. Parameters were hypertuned with RandomizedSearchCV and GridSearchCV. The best model for predicting biodegradation half life used a Random Forest regression with default parameters (R^2 of 0.620). The best model for predicting bioconcentration factor used an XGBoost regression with default parameters (R^2 of 0.633). Statistically significant correlations were found between data columns. For example, bioconcentration factor had a statistically significant positive Spearman’s correlation of 0.655 with Octanol water Partition and bioconcentration factor had a statistically significant positive Spearman’s correlation of 0.676 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

Kmansouri 2016. OPERA Github, Open Source. https://github.com/kmansouri/OPERA

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/