Jason Pan

5/7/25

FWE 458

Machine Learning of PFAS in Environments

**Introduction**

Per- and polyfluoroalkyl substances (PFAS) are a group of man-made chemicals that pose significant environmental and health risks due to their persistence and bioaccumulation in both humans and animals (1). When exploring the proliferation of PFAS in ecosystems, researchers often have to turn to time consuming and costly sampling. As data is often missing in a number of regions and are oftentimes limited to specific groundwater reservoirs and waterways, a machine learning approach can be beneficial to predict the risk of PFAS occurrence (2). This paper seeks to explore various modeling approaches of determining the average concentration across various provinces in China. Understanding and predicting PFAS concentrations can allow policymakers to focus their resources on regulatory decision-making in high-risk areas. In this paper, I evaluate multiple machine learning models for both regression and classification tasks to analyze PFAS data.

**Dataset Source**

A map of china with different colored dots

Description automatically generatedThis analysis utilizes a dataset derived from an extant database downloaded as a csv of a variety of pollutant concentrations drawn from literature sources like CNKI, Web of Science, and PubMed, containing over 55,000 data points collected between 2007 and 2017 (3). The dataset details PFAS concentrations across time, min, max and average concentrations in wildlife, sampling locations, and protocols with both categorical and numerical values. Figure 1 shows a representation of the geographic distribution.

**Figure 1:** A representation of PFAS collection sites within the dataset with a color gradient representing average concentration in µg/kg.

A graph of blue bars

Description automatically generated with medium confidence

**Figure 2:** A bar graph showing the counts of each PFAS pollutant in the dataset.

**Dataset Pre-Processing**

A graph with blue dots

Description automatically generatedA graph with numbers and lines

Description automatically generatedThe dataset, although standardized across its features, contained numerous missing values (NaNs) across multiple features as well as numerous pollutants not of interest. All pollutants not identified as PFAS in the pollutant dictionary were removed from the database, leaving only pollutant IDs 1-17. Each Pollutant ID was then converted from a numerical value to a categorical value which is shown in Figure 2. To ensure consistency, the data was cleaned by formatting cells into a uniform date format (YYYY), excluding averages derived from date ranges (e.g., 2010–2017). Only measurements in micrograms per kilogram (ug/kg) were also retained with rows with no data or unfit data deleted. Other Missing data (NaNs) were kept, but specific NaNs including non-numerical ‘<0.1’ concentration were converted to 0. Several one off inconsistencies such as organ method classification and errors in inputs were also corrected or excluded. Specific pre-processing of NaNs was conducted for each machine learning model. After initial exploration of concentrations over time in figure 3A, outliers were identified as having a large impact on the distribution of average concentration across time. Therefore, a separate dataset with outliers were filtered out using the interquartile range (IQR) method with quartiles set at 0.25 and 0.75 was created with the distribution shown in figure 3B. Finally, all city level and geographic data was removed from the dataset due to overfitting concerns, only the province was kept for reference in merging datasets later.

**Figure 3A:** Time Series Analysis with Raw Data in Average Concentration (µg/kg) showing significant outliers. **Figure 3B:** Time Series Analysis without Outliers utilizing IQR in Average Concentration (µg/kg).

**Selecting Target Variable**

Although the most prevalent and consistent variable was the average concentration, there also existed a significant number of min-max concentration values. To determine the best value to measure as our target variable as well as any potential relationship between them a K-Nearest Neighbor (KNN) analysis with an optimal K of 4 calculated in Figure 4A was undertaken utilizing the IQR dataset.

A graph of different colored dots

Description automatically generatedA diagram of different colored dots

Description automatically generatedA graph of a graph with colored dots

Description automatically generated with medium confidenceA graph with a line

Description automatically generated

**Figure 4A:** Elbow Method to determine optimal K. 4. **Figure 4B:** 3D Visualization of K Means clustering at K=4 (µg/kg**). Figure 4C:** Max Concentration clustering versus average concentration (µg/kg). **Figure 4D:** Min Concentration clustering versus average concentration (µg/kg).

The cluster summary reveals distinct patterns across the four clusters with a silhouette score of 0.513 suggests a moderate level of cluster separation. This indicates that the data points are somewhat distinctly grouped but may still have some overlap which is shown in the visualization in figure 4B. Based on our results, cluster 0 has the lowest average concentration (mean = 0.48 ± 0.54), with relatively low minimum (0.19 ± 0.29) and maximum (0.97 ± 1.05) values. Cluster 1 shows higher average concentrations (6.30 ± 1.43), with moderate minimum (2.97 ± 0.87) and notably higher maximum (12.18 ± 5.09) values. Cluster 2 falls in between (mean = 2.47 ± 0.90), while Cluster 3 exhibits the highest concentrations (mean = 10.30 ± 2.17), with the largest min (4.80 ± 1.01) and max (20.42 ± 5.68) values. The KNN allows us to conclude that there is a good relationship between the min-max concentration and average concentration values as indicated in figure 4A and 4B and therefore we can postulate that using the variables with the most values as our target variable would be the best method.

**Dataset Merging**

As the original dataset was now lacking in features due to the removal of several variables, additional features were acquired from National Bureau of Statistics in China (4). Datasets were downloaded in the csv format and cleaned, with missing values substituted for 0 where appropriate. Those with a significant number of NaNs were not included. These additional features included industrial, environmental, economic, and demographic information specific to provinces in China. Utilizing the province and year of collection columns in our original dataset, these datasets were merged according to their year and the location. After merging the original provinces column any extra added columns were deleted. This process occurred in both the original as well as the dataset with outliers excluded.

**Machine Learning Regression Algorithms**

Initially regression tasks were selected to determine the overall ability of our features to predict average concentration. A random forest and decision tree regressor were selected to the dataset without outliers removed. Imputation was utilized for any missing NaNs as mean values would guarantee the most consistent data initially. We then utilized the dataset with the outliers removed to train two further models of random forest regression and XG Boost. For this random forest, all features with over 40% NaN were removed from the dataset and missing numerical values excluded which meant that imputation was not necessary. As for XG Boost, we utilized the built in function to predict our numerical values. A missing label was utilized for any absent categorical values. We randomly split the dataset into 80% train and 20% test to determine target avg concentration for all models as shown in figure 5.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Task** | **Target** | **Data Handling** | **Key Metrics** |
| Random Forest Regressor | Regression | avg | Imputation (Mean for numerical, 'missing' for categorical NaNs), '<0.1' -> 0 | MSE: 66028825.69, RMSE: 8125.81, R2: 0.8945 |
| Decision Tree Regressor | Regression | avg | Imputation (Mean for numerical, 'missing' for categorical NaNs), '<0.1' -> 0 | MSE: 65133196.09, RMSE: 8070.51, R2: 0.8960 |
| Random Forest Regressor (Without Outliers) | Regression | avg | Outlier removal in avg (IQR), '<0.1' -> 0, All features with >40% NaN removed | MSE: 4.797, RMSE: 2.190, R2: 0.431 |
| XG Boost (Without Outliers) | Regression | avg | XG Boost Prediction for numerical, 'missing' for categorical NaNs, ‘<0.1' -> 0 | RMSE: 1.06 MAE: 0.75  R2: 0.3462 |

**Figure 5:** A data table showing regression models (avg: average concentration, NaN = Not a Number)

**Analysis of Regression Models**

Both the initial Random Forest Regressor and the Decision Tree Regressor, when trained on the dataset with missing feature values handled by imputation, achieved a high R-squared value (around 0.89). This indicates that these models were able to explain a large proportion of the variance in the avg column but since the MSE and RMSE were relatively high we cannot conclusively determine their accuracy. This trend is influenced by the presence of outliers in the avg column, which likely significantly increased the error.

The Random Forest Regressor after outlier removal in avg showed a dramatic decrease in MSE and RMSE compared to the model including outliers. However, the R-squared value also dropped significantly (to 0.431). This highlights that while outliers inflate error metrics, they also contribute significantly to the total variance in the target variable, still we can trust the outputs of this particular model to a greater extent. When outliers are removed, the remaining variance is smaller, and the model explains a smaller proportion of this reduced variance, even if its predictions on the non-outlier data are more accurate in absolute terms. The XG Boost utilizing its built-in predictor of NaNs had a worse R-squared (0.3462) but had the lowest MSE and RMSE which indicates that it may be able to accurately predict those missing values, but its predictive power is overall lower.

|  |  |
| --- | --- |
| Feature | Importance |
| cat\_\_organ\_leaf | 0.180017 |
| cat\_\_poid\_PFHxA | 0.138394 |
| num\_\_PollutantEmission | 0.130348 |
| num\_\_IndustryValue\_Overall | 0.113913 |
| num\_\_IndustryValue | 0.113317 |

**Figure 6:** Top ranked feature importance attribution (PollutantEmission = Total Waste Water Discharge; IndustryValue\_Overall = Net Value of Accounts Received of Industrial Enterprises over Designated Size (100 million yuan); Industry Value = Value-added of Industry (100 million yuan))

Delving deeper into the random forest model without the outliers, we are able to derive a feature importance chart. We observe in figure 6 that the highest predictors included the “organ” leaf in the original dataset, which showed a large distribution based on concentration in initial data preprocessing and therefore was expected. PFHxA has the highest importance of the pollutants and is unsurprising given it has the highest counts in the data, likely allowing it to have an outsized effect on the modeling. The amount of pollutant released and industries in a province based on our added datasets also play a significant part in the model which matches previous research that adjacency to industry is one of the primary indicators of PFAS (1).

**Machine Learning Classification Algorithms**

As the effect of outliers impacted our results greatly under the regression tasks, a classification task was therefore implemented to conclusive determine we could categorize the dataset with outliers so. This was to see if there was a way to keep our high R squared score in our regression models but reduce the large error involved. Imputation was utilized as the outlier dataset contained more NaNs. The Logistic Regression, Support Vector Machine (SVM), Gaussian Naive Bayes models and a random forest classification model were used to predict the categorized avg concentration levels that correlated to 'low', 'medium' and 'high' shown in figure 7.

**Analysis of Classification Models**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Task** | **Target** |  | **Data Handling** | **Key Metrics** |
| Logistic Regression | Classification | avg (Low/Medium/High) |  | Imputation (Mean for numerical, 'missing' for categorical NaNs), '<0.1' -> 0 | Accuracy: 0.4075, High F1: 0.33, Low F1: 0.49, Medium F1: 0.30 |
| Random Forest Classifier | Classification | avg (Low/Medium/High) |  | Imputation (Mean for numerical, 'missing' for categorical NaNs), '<0.1' -> 0 | Accuracy:  0.6649, High F1: 0.75, Low F1: 0.67, Medium F1: 0.57  Cross Validation: 0.7164 |
| Random Forest Classifier | Classification | avg (Low/High) |  | Imputation (Mean for numerical, 'missing' for categorical NaNs), '<0.1' -> 0 | Accuracy:0.84, High F1: 0.85, Low F1: 0.82  Cross Validation: 0.8365 |
| Support Vector Machine | Classification | avg (Low/Medium/High) |  | Imputation (Mean for numerical, 'missing' for categorical NaNs), '<0.1' -> 0 | Accuracy: 0.4075, High F1: 0.33, Low F1: 0.49, Medium F1: 0.30 |
| Gaussian Naive Bayes | Classification | avg (Low/Medium/High) |  | Imputation (Mean for numerical, 'missing' for categorical NaNs), '<0.1' -> 0 | Accuracy: 0.3902, High F1: 0.39, Low F1: 0.49, Medium F1: 0.10 |

**Figure 7:** A data table showing classification models (avg: average concentration, NaN = Not a Number)

The random forest model performed best with an accuracy of 66.49% with a good cross validation of 0.7164, indicating to us a reasonable model for classifying average concentration. All models struggled to predict the 'medium' category effectively, as indicated by low recall and F1-scores for this class. The 'low' category generally had higher recall, meaning the models were better at identifying actual 'low' instances, but the precision was lower, suggesting they also incorrectly classified other instances as 'low'. The 'high' category often had higher precision especially SVM meaning when the model predicted 'high', it was often correct, but the recall was low, indicating it missed many actual 'high' instances.

As such a further random forest model classifier was run with only with a binary classification of high or low concentration to eliminate this shortfall and achieved the best accuracy of 84%. This was supported by an excellent cross validation of 0.8365 which gives us the best model so far. This binary model however would likely have less utility than the 3 classifier models and therefore some limitations in use. Still, in determining if PFAS presented as a yes or no question, it performs well. In comparison the boundaries between 'low', 'medium', and 'high' based on quantiles appear less separable in the feature space when looking at the other models.

**Limitations**

A significant challenge in this study was the prevalence of data limitations, which directly impacted model performance and interpretability. The dataset contained numerous missing values (NaNs) across key features, requiring either imputation or exclusion with bias being inevitable. Substituting NaNs with mean values or zero (for values like "<0.1") may have artificially skewed distributions and proved a detriment to model accuracy, particularly for pollutants with sparse measurements or min-max concentration values. Additionally, the geographic scope chosen which was province-level aggregation constrained the analysis of finer trends near the collection site itself, such as localized contamination hotspots or the presence of specific industries nearby. Inconsistencies in reporting of the dataset such as randomly inputted numerical values, varying units, complicated wording and overall inefficiency of the dataset led to an overcomplicated preprocessing process. These issues underscore the need for standardized PFAS monitoring protocols which can include consistent uniform units, and systematic standardized documentation across a national database.

**Conclusion**

This paper demonstrates that machine learning has potential in predicting PFAS concentrations, though model performance heavily depends on data handling and methodology and especially the quality of the dataset. The Random Forest regressor using imputation achieved a high R-squared value (0.89), indicating strong explanatory power, but its high MSE and RMSE suggest outlier sensitivity. In contrast, the outlier-free Random Forest model showed lower variance explanation (0.431) but more reliable error metrics (RMSE = 2.19), making it better for practical applications despite its reduced scope. Classification tasks proved more challenging, with the Random Forest classifier reaching 66.5% accuracy for low/medium/high categorization but struggling with the "medium" class. Simplifying to binary classification (low vs. high) however improved accuracy to 84%, offering a robust tool for binary risk assessment.

Both regression and classification approaches can be utilized to approach the lack of concrete PFAS data. Regression models should make precise concentration predictions, while classification models can better categorize overall risk levels. Ultimately, machine learning provides a valuable framework for PFAS risk assessment, enabling policymakers to target high-risk areas efficiently.

**Note:** I actually ran the classification models before the regression models and didn’t plan on including it but since I already had the code, I decided to include it in the report anyhow. Combining datasets would also not be possible without the assistance of Gemini.

**References**

1. Dong, J., Tsai, G., & Olivares, C. I. (2023). Prediction of 35 target per-and polyfluoroalkyl substances (PFASs) in California groundwater using multilabel semisupervised machine learning. ACS ES&T Water, 4(3), 969-981.
2. Hu, X. C., Ge, B., Ruyle, B. J., Sun, J., & Sunderland, E. M. (2021). A Statistical Approach for Identifying Private Wells Susceptible to Perfluoroalkyl Substances (PFAS) Contamination. Environmental science & technology letters, 8(7), 596–602. <https://doi.org/10.1021/acs.estlett.1c00264>
3. Liu, X. B., Lin, C., Wu, Y. Y., Huang, H. N., Zhu, L. T., Jiang, R., & Huang, Q. (2022). A dataset of multi-pollutants in multi-species in China. Environmental Science and Pollution Research, 29(33), 49985-49997.
4. National Bureau of Statistics China (NBS) https://data.stats.gov.cn/