International Research Journal of Public and Environmental Health
Vol.11 (1), pp. 18-29 January 2024
ISSN: 2360-8803(e) ISSN: 2814-3817(p)
Available online at https://www.journalissues.org/IRJPEH/
DOI: https://doi.org/10.15739/irjpeh.24.003

Original Research Article

Evaluating the relationship between socioeconomic disparity and potential PFAS contamination in the United States with machine learning: Implications for public health and environmental justice

Received 3 October, 2023 | Revised 26 November, 2023 | Accepted 6 December, 2023 | Published 20 December, 2023

Sameer Menghani[1*], Jonathan Yang[1], Lena Haefele[1], Louise Carroll[1], Sathvik Samant[1], Robert Lee[1], Anthony Sapp Guadarrama[1], Andrew Noviello[1] and Alexander Noviello[1]

[1]The Lawrenceville School, Department of Mathematics. Lawrenceville, New Jersey, USA

*Corresponding Author
Email: smenghani06@gmail.com

PFAS (Per-and Polyfluoroalkyl Substances), or synthetic "forever chemicals," pose a dangerous threat to public health, contaminating food, water, and air. PFAS testing regulations vary widely across states, remaining inadequate and inconsistent. This study sought to investigate the relationship between the presence of PFAS-handling industries and socioeconomic factors to identify areas with the highest vulnerability to PFAS contamination and prioritize testing at those locations. A Random Forest classifier was developed for predicting proximal PFAS industry presence based on six socioeconomic features and uncovered non-linear and non-monotonic relationships using partial dependence plots and Shapley Additive Explanations analysis. With the notable exception of direct income level, it was found that more disadvantaged socioeconomic conditions generally yielded a higher likelihood of PFAS contamination in communities. More specifically, for most features, it was determined that individuals of generally middle to lower socioeconomic conditions, not the lowest, may be at the greatest risk of PFAS exposure, contrary to traditional expectations. Our results therefore reveal a necessity for greater nuance in the identification of particularly vulnerable communities for more effective prioritization of PFAS testing areas. This study hopes to quantitatively inform the implementation of consistent and targeted PFAS testing to advance public health across the United States.

Keywords: PFAS, socioeconomic, public health risk, variations, contamination, exposure, testings, eco-friendly

INTRODUCTION

PFAS (Per- and Polyfluorinated Substances), or "forever chemicals", are a group of more than 9,000 artificial compounds characterized by synthetic carbon-fluorine bonds that are highly resistant to natural breakdown (What are PFAS?, 2022). PFAS are used in a wide variety of industrial and consumer products, including non-stick cookware, water-repellent clothing, and stain-resistant fabric (*Understanding PFAS*, 2022; *Per- and Polyfluoroalkyl*, 2022). As a result of their widespread use and breakdown-resistance, PFAS chemicals have been found in the blood of 97% of Americans, proliferated primarily through seeping into drinking water (*Perfluoroalkyl and Polyfluoroalkyl*,

2023). Exposure to PFAS has been linked to increased risk of prostate, kidney, and testicular cancers, fertility issues, increased cholesterol levels, and other gastrovascular diseases (PFAS Explained, 2023).

Despite the grave risks of PFAS to public health across the United States, research and testing surrounding the full effects of these chemicals on individuals is still in early phases (Our Current, 2023). Furthermore, the specific levels of exposure at which negative health consequences may arise remain unclear (National Academies of Sciences, Engineering, and Medicine., 2022). Additionally, given the many identified variations of PFAS compounds, using a singular testing method may be unreliable or fail to produce complete results (*Perfluoroalkyl and Polyfluoroalkyl*, 2023). This limitation is primarily due to the variety of sources of PFAS chemical contamination (Foss, 2023). Currently, the EPA uses three methods to analyse PFAS: Methods 533, 537, and 537.1 to analyse 29 different types of PFAS contamination in drinking water (PFAS Analytical, 2023).

Federal PFAS contamination regulations are only just developing. The Environmental Protection Agency first released a framework to combat the industrial release of PFAS (Sokolowski, 2023). In the absence of a coordinated federal response, PFAS regulation and testing has been largely left to state governments, resulting in inconsistent protocols and proposed solutions to the crisis (Evans et al., 2020). Given these inconsistencies, statistical conclusions from performing quantitative analysis on aggregated national PFAS testing results may be compromised due to confounding differences in testing protocols. As such, the study leveraged a substitute measure of PFAS contamination: proximal density of facilities that may potentially manufacture, process, use, or release PFAS chemicals.

Supporting the validity of this substitution, past studies into the relationship between industrial sites and PFAS detection in the U.S. have demonstrated that these industries, including manufacturing sites, military bases, and water treatment plants, are all indicators and effective tools for measurement as to PFAS levels in local public water supplies (Hu et al., 2016; Salvatore et al., 2022). In other words, a greater number of PFAS handling industries strongly alludes to heightened levels of contamination.

Thus, this study seeks to quantify the relationship between PFAS contamination as measured by industry density and socioeconomic disparity, in order to inform future priority testing areas nationwide. Previous research has revealed that community water systems with higher proportions of Hispanic/Latino and non-Hispanic Black residents had significantly increased likelihood of PFAS presence of around 6% (Liddie et al., 2023). Another study found that there were 22% more people of color and 15% more low income households than expected living within 5-mile radii of 73 non-military PFAS-contaminated sites (Desikan et al., 2019). As such, the authors hypothesized that increased PFAS contamination and subsequent human exposure would positively correlate with disadvantaged socioeconomic conditions.

This study used quantitative techniques, including both statistical analyses and machine learning algorithms, to investigate the relationships between each of the designated socioeconomic factors and proximal PFAS industry density, in an effort to generate actionable insights for public health officials.

MATERIALS AND METHODS

Tools

The study employed Jupyter Notebook and Google Colaboratory for Python programming. The following packages were used for analysis: Pandas, Geopandas, MatPlotLib, SciPy, SHAP, and Sklearn (Pandas, 2023; GeoPandas, 2023; MatPlotLib, 2023; SciPy, 2023; SHAP, 2023; Scikit-learn, 2023). TIGER/Line shapefiles, geographical data sets produced by the U.S. Census Bureau for public use, were used to produce maps of the mainland United States with designations for state boundaries (TIGER/Line Shapefiles, 2021).

Quantifying PFAS Testing Scarcity

To investigate the scarcity of PFAS testing across the United States, all state drinking water testing site geolocations collected by EPA were plotted in Figure 1a (PFAS Analytic Tools, 2023). The resulting plot offers a qualitative visualization of the lack of comprehensive PFAS testing across the country. To quantify this demonstration, the number of PFAS testing sites with geolocations within 25 square miles of each of the approximately 30,000 major city locations in our cities dataset was calculated. This dataset was compiled from the U.S. Geological Survey and U.S. Census Bureau (SimpleMaps, 2023). The aforementioned area was chosen to accommodate computing RAM limitations and refrain from capturing surrounding suburbs for most cities. In order to calculate distances and radii, Latitude and Longitude distances were converted to mileage using basic arithmetic. One degree of Longitude is approximately 54.6 miles and one degree of latitude is approximately 69 miles (How Much, 2023). Using this conversion, longitude and latitude values of 5 miles were calculated, ultimately allowing a 5x5 mile (25 sq. mile) square boundary to be set around each city. Summary statistics (mean, median, variance, minimum, maximum, Q1, and Q3 values) for the distribution of PFAS testing site quantities near each major city are displayed in the "Testing Sites" column of Table 1.

Evaluating PFAS Industry Density

Similarly, in order to investigate the prevalence of PFAS industries across the United States and the practical viability of substituting industries for PFAS testing results, the process of creating Figure 1a was emulated to plot the

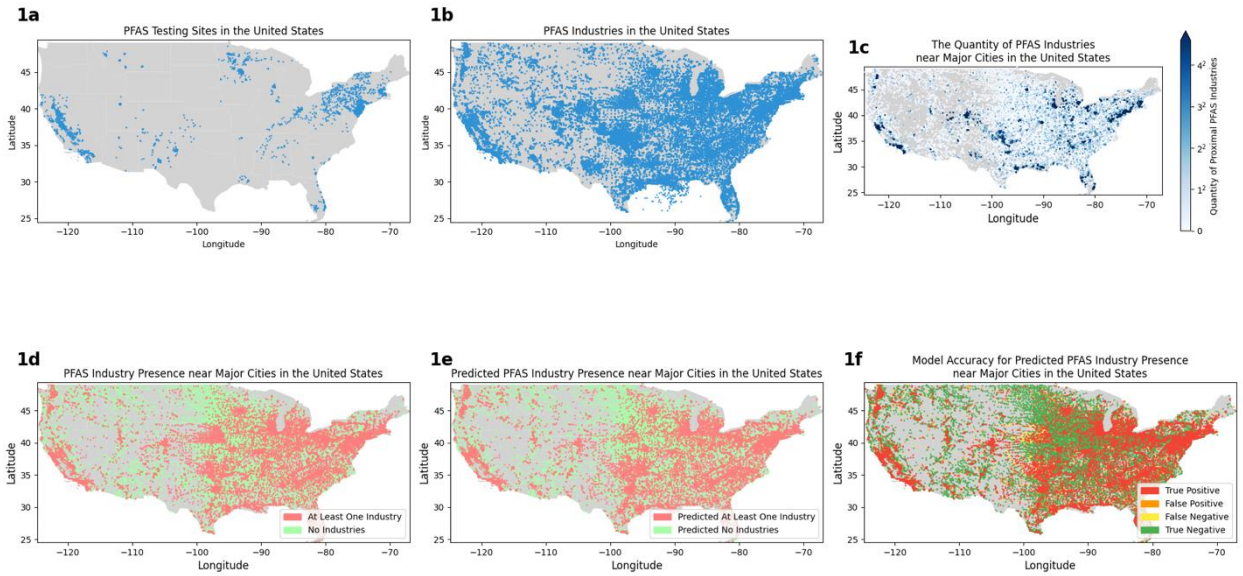


Figure 1: Tiger/Line shapefile maps with geolocation points, visualizing PFAS industries and testing sites across the mainland United States. (1a) Map of PFAS testing sites from the US Water Council database. (1b) Map of PFAS Handling Industries geolocations from EPA database. (1c) Map of major US cities geolocations with gradient shades of blue corresponding to the magnitude of proximal PFAS handling industries, with darker blue indicating more industry presence. (1d) Map of binary color-coded major US cities with (red points) and without (green points) at least one proximal PFAS handling industry. (1e) Map of binary color-coded major US cities with (red points) and without (green points) Random Forest predicted proximal PFAS handling industries. (1f) Map of color-coded major US cities based on Random Forest model accuracy for predicting the presence of proximal PFAS handling industries. Red points: Cities correctly predicted to have proximal industries (True Positive); Orange points: Cities incorrectly predicted to have proximal industries (False Positive); Yellow points: Cities incorrectly predicted to be absent of proximal industries (False Negative); Green points: Cities correctly predicted to be absent of proximal industries (True Negative).

Table 1. Measuring sparsity with aggregated proximal site statistics (total, mean, median, variance, Q1, Q3, min and max) for the quantity of testing sites and industry sites within 25 square miles of major cities. Industry sites outnumber testing sites by close to 50:1.

Statistical Measure	Testing Sites	Industry Sites
Total	3006	143246
Mean	0.0975	4.64
Median	0	1
Variance	0.306	247
Q1	0	0
Q3	0	3
Min	0	0
Max	22	662

geolocations of PFAS industries in Figure 1b. Figure 1c offers an alternative visualization of the density distribution of PFAS industries around major cities, indicating severity of potential contamination. Qualitative observations reveal that PFAS industry locations are far more widespread and numerous across the United States than PFAS testing locations. Additionally, the process by which EPA gathers data on the presence of these industries is conducted uniformly on a national scale, allowing for comparison of site density across state lines, without potentially confounding procedural differences. Summary statistics for the distribution of industry density near major cities are displayed in the "Industry Sites" column of Table 1.

The large difference in the total number of testing sites (3,006) compared to industrial sites (143,246) highlights the inadequacy of using testing sites for analysis. Furthermore, the testing site data median (0) and Q3 (0) signify a lack of testing sites in a majority of cities nationwide, thus preventing comprehensive assessment.

The low variance (0.306) and limited range (22) across all testing site densities would further inhibit the ability of machine learning algorithms to effectively classify areas for priority testing.

Feature Engineering: Shift to Binary Classification

Contrastingly, to ensure that unnecessarily high variance of the proximal density of industries to cities across the nation did not hinder our machine learning models' ability to generalize effectively, PFAS industry counts were converted to a binary measure of industry presence. This resulted in a shift from regression to classification models to determine the binary presence or absence of potential PFAS contamination. A regression analysis would provide unreliable additional insight, due to unknown differences of size and pollution contributions amongst specific industries (i.e. "unnecessarily high variance"). For example, one high-polluting industry may output more PFAS than three smaller industries. Therefore, binary classification predictions offered more consistent and reliable outputs, yielding more insightful suggestions for resource allocation and targeted testing to address the risks associated with PFAS contamination nationwide. Figure 1d represents a visualization of the ultimate binary dataset.

Compiling Socioeconomic Data

Socioeconomic data was collected from EJScreen, a public EPA application programming interface. We scraped EJScreen for socioeconomic data within a 25 mile circle area (2.82 mile radius) in each of the top populated city locations. The following socioeconomic measures were scraped as input features for analysis (Overview of Socioeconomic, 2023).

1. Low Income: National percentile for households whose income is less than twice the federal poverty level.
2. Minority Rate: National percentile for proportion of people who identify their race as other than white alone.
3. Less than High School Education Rate: National percentile for proportion of 25 year olds that did not graduate high school.
4. Unemployment Rate: National percentile for proportion of people who did not have a job at the time of questioning but were available to work and had attempted to acquire a job.
5. Under Age 5 Rate: National percentile for proportion of people under the age of 5.
6. Over Age 64 Rate: National percentile for proportion of people over the age of 64.

Calculating Correlations Between Socioeconomic Features

Pearson correlation coefficients assume a bivariate normal distribution and a linear relationship between variables, although may be empirically robust to violations (Havlicek and Peterson, 1976). Kendall and Spearman correlation coefficients assume monotonic relationships between variables, not normality nor linearity (Liu et al., 2016). All three correlation measures yielded similar results when calculated between socioeconomic features and proximal PFAS industry presence. Pearson correlations were chosen for further analysis.

Testing Linear Classification Models for Predicting PFAS Presence

Two linear machine learning models were initially proposed to solve the binary classification problem of predicting PFAS presence: Logistic Regression and Support Vector Machines.

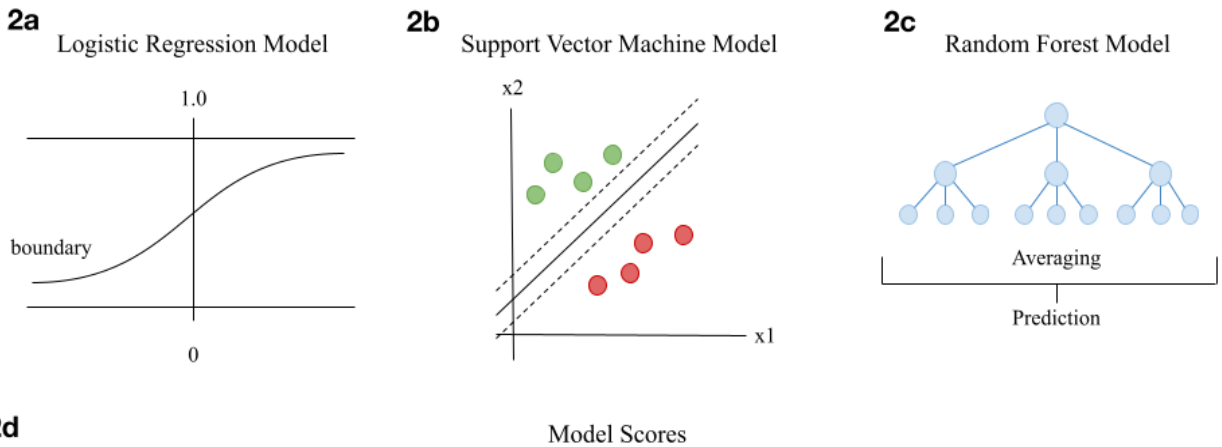
Logistic Regression is a generalized linear model for binary classification that estimates probabilities of target classes (Figure 2a). To find the probability score, a logit transformation is applied to the probability of PFAS presence divided by the probability of PFAS absence. The model is subsequently trained through maximum likelihood estimation to find the best-fitting parameters that minimize the log-loss, which penalizes incorrect predictions. A probability score less than 0.5 will predict 0, while a probability greater than 0.5 will predict 1 (IBM, 2023a).

Support Vector Machine (SVM) models transform data into a higher-dimensional feature space, allowing for categorizing data points even when they are not linearly separable (Figure 2b). It searches for a separator between the categories and the data is restructured to allow the separator to be represented as a hyperplane. As a result, SVM can predict the class to which a datapoint belongs based on its characterization around the hyperplane (IBM, 2023b).

However, preliminary results suggested that the linear models failed to adequately encapsulate the complexity of the relationships between features. Thus, a nonlinear model, Random Forest, was evaluated. Figure 2d contains results comparing the three considered model architectures after optimization. The Random Forest model ultimately proved most successful in handling the complexity and multi-collinearity of the data and was thus chosen for predictions and further analysis. More in-depth explanations of the Random Forest's function, optimization, and deployment for this study are detailed below.

Random Forest Classifier Technical Overview and Development

Random Forest models make predictions based on the majority result derived from an ensemble of decision trees that each perform classification by recursively splitting data into subgroups to maximize information gain (Figure 2c) (Speiser et al., 2019; Wang et al., 2012; Xu et al., 2017). Each tree in the Random Forest model randomly samples a subset of training data with replacement through bootstrapping, or "bagging." Bootstrapping allows the model to increase its capacity for pattern encapsulation while reducing overfitting (Ao et al., 2019). The ability of



2d

Model Scores

Metric	Logistc Regression Model Score	SVM Model Score	Random Forest Model Score
Precision	63.6%	64.5%	72.3%
Accuracy	62.6%	62.6%	71.2%
F1	62.3%	61.2%	70.7%

Figure 2: Machine learning model diagrams and respective performance scores. (2a) Diagram of Logistic Regression model classification (2b) Diagram of Support Vector Machine optimization (2c) Diagram of Random Forest model training and architecture (2d) Model precision, accuracy, and F1 scores when predicting binary PFAS industry presence based on the study's 6 socioeconomic indicators.

Random Forest to produce reliable predictions by aggregating the outputs of multiple decision trees makes it well-suited for analysing the impacts of many socioeconomic features.

Prior to model training, we removed missing values in the dataset. For model training and hyperparameter optimization of the Random Forest, we used an exhaustive grid search technique. A grid search iterates through predefined hyperparameters with incremental value increases to identify the combination for maximal model accuracy and precision, within computing constraints.

The implemented grid search used three hyperparameters: n_estimators, max_depth, and min_samples_split.

1. N_estimators: Number of decision trees used to predict final outcome. Tested values of 100, 125, 150, and 175. Optimal: 150.
2. Max_depth: Maximum number of layers that a decision tree can grow, splitting at each node. Larger max_depth allows for more nuanced pattern discovery; too large a max_depth value may lead to overfitting. Tested values of 10, 12, 14, and 16 for max_depth. Optimal: 16.
3. Min_samples_split: Minimum number of samples required to split a node in each decision tree. Higher values result in fewer splits and may lead to underfitting; lower values capture more detailed relationships but may result in overfitting. Tested values of 8, 10, and 12. Optimal: 12.

We split the dataset into training and testing sets using a 9:1 ratio (90% training data, 10% testing data). The training set was used to fit the model, while the testing set was employed to assess its out-of-sample performance. Following model training, we made predictions on the testing set using the optimized Random Forest, calculating accuracy, precision, and F1 as shown in Figure 2d.

Investigating Feature Importance

Feature importance values for the trained Random Forest Classifier were calculated through the Gini importance method, which evaluates the purity increase associated with each feature across all the trees in the ensemble. A greater purity increase improves prediction certainty, thus increasing a feature's importance score (Menze et al., 2009). These values quantify the relative influence of each socioeconomic feature in making predictions on PFAS presence.

Random Forest Deployment

After training the Random Forest, the model was deployed to make predictions for industry presence in major cities across the United States based on socioeconomic data. We

visualized the model's predictions in Figure 1e. Figure 1f represents a visualization of the model's performance against true values.

Partial Dependence Plots

Partial dependence plots (PDPs) were developed to visualize and interpret complex relationships between each of the socioeconomic indicators and proximal PFAS industry presence (Figure 4). To generate a PDP, a range of all possible values is set for each indicator (1-99 in this study, as all features represent percentiles). Next, a specific indicator is incremented while keeping all other variables constant. At each increment, the average of the probability of proximal PFAS industry presence is plotted. Although PDPs illuminated important and previously unknown trends, they assume feature independence, an unfulfilled condition in this study, thereby necessitating the use of additional methods to corroborate results.

SHAP Calculations

Shapley Additive Explanations (SHAP) values reveal a given sample's vector contribution to the classification model's prediction (Lundberg and Lee, 2017). SHAP values are calculated by assessing how the prediction probabilities change when varying the value of a specific target feature, while also acknowledging inter-variable relationships. Thus, SHAP plots mitigate the greatest caveat of PDPs in their assumption of feature independence.

The process of calculating a SHAP value starts by creating 2 vectors. The values in the two vectors are taken partially from an original identified sample and partially from a random row plucked from the rest of the data. In vector 1, the original values for all features to the right of the target feature are replaced by values from the random row. In vector 2, the original values for all features to the right and including the target feature are replaced by values from the random row. Through this method, the danger of creating improbable data scenarios through the assumption of no intervariable relationships is effectively mitigated because the data in the vectors is changed together in batches from the original and random sample (Lundberg and Lee, 2017).

After running the 2 vectors through the Random Forest model, the predicted probabilities of industry presence, 2 prediction vectors, are subtracted to determine the effect of changing the target feature on the model's prediction. The difference is recorded, and a new permutation of features is taken. These steps are repeated until all possible iterations are complete. The final SHAP value is the average of the differences between the 2 prediction vectors for all possible iterations, measuring the average effect of changing the target feature on the model's probability of predicting PFAS industry presence for a specific input value. This entire process is repeated for every value of every feature resulting in a comprehensive interpretation of each feature's impact on the model's output.

Pearson Correlations, PDPs, and most other visualizations of feature importance lack the ability to convey each indicator's vector (+, -) relationship with PFAS presence in order of importance. However, SHAP plots encapsulate both feature importance rankings and vector (i.e. directional) importance.

SHAP Visualizations

SHAP values were calculated for each sample and a bee swarm plot was generated for each feature representing its vector relationship to PFAS industry presence in Figure 4a. Summary statistics for the distributions of sample impacts for each feature, as displayed on the SHAP bee swarm plots, are shown in Figure 4b.

Scatterplots were also generated which compare raw percentile values (independent) against raw SHAP values (dependent) for each socioeconomic feature (Figure 5).

RESULTS

Pearson Correlation Matrix and PDP Analysis

Pearson correlation coefficients between socioeconomic indicators and PFAS industry presence demonstrated generally minimal relationships. The minority rate feature had the highest correlation of 0.24 with PFAS industry presence, indicating a low positive relationship. The unemployment rate, under age 5, and less than high school education rate features had negligible positive correlations of 0.13, 0.10, and 0.02, respectively. The over age 64 and low income rate features had small negative correlations of -0.17 and -0.039, respectively. This trend suggests that areas with (1) less people over the age of 64 and, contrary to the hypothesis of this study, (2) higher income, had a generally greater likelihood of PFAS industry presence. Nevertheless, all 6 socioeconomic conditions demonstrated showed weak relationships with PFAS industry presence based on correlations alone. However, calculated correlations revealed several strong relationships between the socioeconomic factors themselves, often portraying somewhat obvious relationships, including a correlation of 0.68 between lower income and less education. See Appendix 1 for the full correlation matrix. These intervariable relationships revealed why utilizing Pearson correlations between socioeconomic indicators and PFAS industry presence directly was inadequate, as statistical assumptions for use were not fulfilled.

All PDPs demonstrated generally downward concave trends with peak partial dependency scores around the 60-70th percentiles, revealing non-linear and non-monotonic relationships between the respective features and PFAS industry presence, failing to meet Pearson, Spearman, and Kendall correlation assumptions. The most notable examples of this trend are seen in Figures 3a, 3b, and 3c that have peak partial dependence scores at around the 60th, 68th, and 62nd percentiles, respectively.

Figure 3a, whose graphs showed the minority percentile

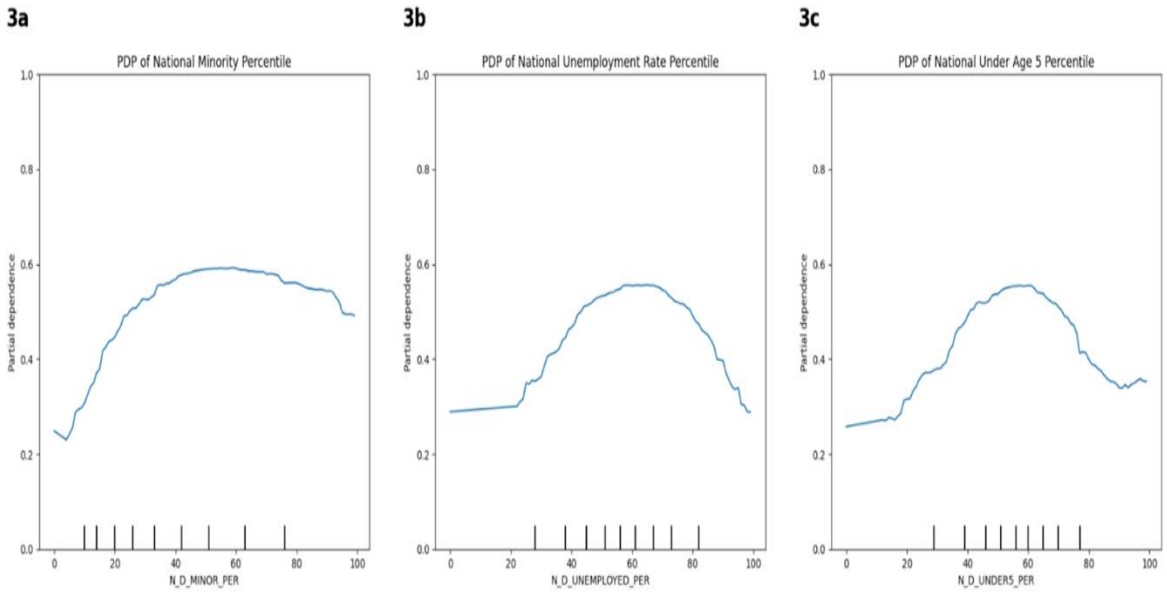


Figure 3: Partial Dependence Plots (PDPs) for the Random Forest model on three socioeconomic features, revealing the complex, non-linear, and non-monotonic relationships between socioeconomic disparity and PFAS industry presence. These plots do not account for inter-variable relationships. PDPs of the Minority (3a), Unemployment (3b), and Under Age 5 (3c) features are displayed due to their demonstration of clear trends, whereas results from the other three features' PDPs are less conclusive and interpretable.

indicator's impact, has the highest peak partial dependence score of 0.6, revealing that the racial demographic had the greatest impact on the model's prediction of PFAS industry presence. This result is analogous to other feature importance scores discussed below. The PDPs for the other features had less obvious concave trends and lower peak partial dependence scores (around 0.5), reflecting less specific relationships between their respective features and PFAS industry presence. These PDPs are provided in Appendix 2.

Random Forest Scores and Feature Importance

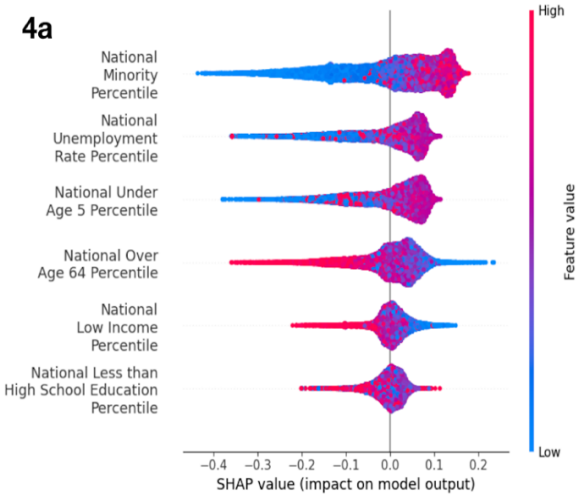
The optimized Random Forest Classifier achieved a precision score of 72.3%, an accuracy score of 71.2%, and an F1 score of 70.7%, on testing data (Figure 2d). This model performance demonstrates its ability to effectively predict proximal presence or absence of PFAS industries based on socioeconomic factors.

The national minority rate percentile holds the highest feature importance score at around 0.22, suggesting that the presence of a minority population plays a notable role in industry presence. Next, the percentiles for national unemployment rate, national under age of 5 percent, and national over age of 64 percent all had similar feature importance scores of 0.18, 0.18, and 0.17, respectively. The feature importance scores decline significantly for the remaining features, indicating their comparatively lesser impact. Features such as the national less than high school education percentile and national low income percentile both had a low feature importance scores of 0.13, indicating minimal influence on predicting PFAS industry presence.

SHAP Plot Analysis

The SHAP bee swarm plots in Figure 4a are displayed in descending order of their corresponding feature importance, with regard to inter-variable relationships. The SHAP-calculated feature importance order mirrors the Gini Importance computation ranking, except the least important two features (education and low income).

The minority feature has a primarily positive tail extending to about SHAP value 0.19 and a primarily negative tail extending to about SHAP value -0.41, indicating that higher minority rates generally yield positive industry presence predictions and vice versa. A similar trend is also clearly seen in the unemployment and under age 5 features. Both of these features have generally positive tails extending to about 0.13 and negative tails extending to about -0.38, signifying positive relationships between both unemployment and population youth with PFAS industry presence. The under age 5 feature plot clearly demonstrates a greater concentration of mixed values, most evident in the negative tail. The over age 64 feature plot corroborates the trend of the under age 5 plot with generally higher risk for PFAS industry presence associated with generally younger populations. The feature had a primarily positive tail extending to roughly SHAP value -0.37 and a primarily negative tail toward around SHAP value 0.28.



4b SHAP Value Distribution

Feature	Mean	Q1	Median	Q3	Variance	Percentage of Positive
National Minority Percentile	0.00132	-0.0736	0.0271	0.0918	0.0129	59.9%
National Unemployment Percentile	-0.000239	-0.0432	0.0294	0.0625	0.00694	64.1%
National Under Age 5 Percentile	0.000327	-0.0437	0.0266	0.0602	0.00616	63.0%
National Over Age 64 Percentile	-0.00167	-0.0279	0.0141	0.0452	0.00485	60.1%
National Low Income Percentile	0.00489	-0.0116	0.0047	0.0235	0.00127	57.6%
National Less than Highschool Education Percentile	-0.00228	-0.0186	0.00197	0.0203	0.00101	53.1%

Figure 4: (4a) SHAP bee swarm plots displaying each sample's vector importance toward Random Forest predictions in descending order based on SHAP-calculated feature importance and summary statistics. Bee swarm plots. Red and blue points represent higher and lower percentile sample values, respectively. (4b) Summary statistics were included to provide a quantitative distribution of the sample contributions to model classifications.

The low income feature's positive and negative tails have SHAP values of -0.23 and 0.16, respectively, with lower variability on both sides. This unexpected trend indicates that areas of high income may actually be more likely to have proximal PFAS industries, which is consistent with the relative trend revealed by Pearson correlations. The national less than high school education feature lacks any clear trend with a range of percentile values scattered across the plot.

Summary statistics in Figure 4b explain the quantitative distributions of the sample contributions to the model classifications, as communicated visually in the bee swarm plots. This may provide a basis of comparison for future analysis, further investigating the quantitative complexity of the relevant relationships.

Analysing the SHAP value scatterplots yielded more specific insight into the nature of the relationships between the socioeconomic features and the presence of PFAS industries. For example, several of the regression graphs displayed parabolic trends, peaking around the 70th percentile. For these features, Q2 and Q3 experienced the most significant risk of PFAS industry presence, as opposed to the most socioeconomically disadvantaged quartile. Variability also seemed to increase near the extremes of specific plots, most notably the low income (Figure 5d) and education features (Figure 5c). The potential reasons for these trends and more are explained in greater detail in the discussion.

DISCUSSION

Consistent, comprehensive, and targeted nationwide PFAS testing in water supply represents the first and most critical step toward addressing the public health dangers of "forever chemicals." Testing is important everywhere, but resource constraints make prioritizing specific locations necessary. While testing in the areas of densest PFAS industry concentrations should be the top priority, it is also vital to consider the vulnerabilities of specific communities in resource allocation. In turn, a comparative analysis of socioeconomic indicators with PFAS industry presence, such as this study, is necessary to further illuminate areas of highest public health concern. Socioeconomically disadvantaged communities may lack access to sufficient public health resources, and/or effective social, political, and economic representation to mitigate the harmful effects of PFAS contamination. As such, our study's prediction tools offer a quantitative approach to guiding resource allocation for PFAS testing, informed by both industry concentration and community vulnerability through the analysis of their relationship.

The Random Forest SHAP analysis, as visualized in the scatterplots of Figure 5, provides the most complete quantitative insights into the specific relationships between each socioeconomic feature and PFAS industry presence, beyond the other statistical tools leveraged in the study. Figure 5a reveals a fractional-power relationship between minority concentration and potential PFAS contamination, with minimal diversity generally associated with lower likelihood of PFAS industry presence. Diverse communities may lack the influence of predominantly white communities to oppose the development of PFAS industries and enforce stricter environmental regulations.

The unemployment rate SHAP scatterplot in Figure 5b demonstrates a parabolic trend. Areas with extremely low rates of unemployment are often financially well off and have the economic and public health resources to prohibit PFAS industries. Moreover, these local economies may be better suited financially and infrastructurally to support

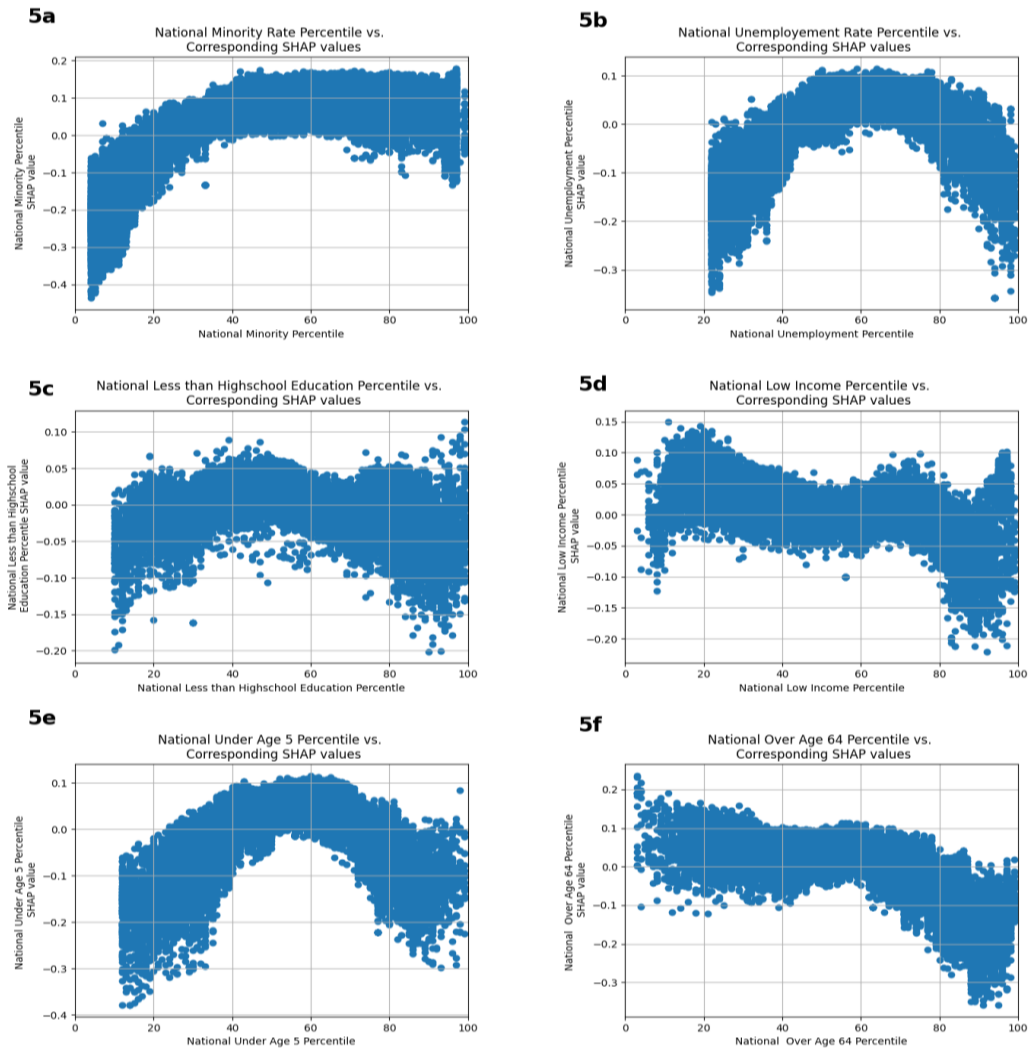


Figure 5: Scatterplots of feature sample values on x-axis and corresponding SHAP values on y-axis.

white-collar jobs and office work, as opposed to heavy industrial operations linked to PFAS. Extremely high rates of local unemployment also present slightly decreased risks of contamination, potentially as a result of a dearth of local economic activity and an effective workforce to support PFAS industries. Thus, areas with moderate-high unemployment appear to face the highest risk of PFAS industry concentration, perhaps due to their more readily available and more adequately suited economies and workforces for heavy PFAS-related industries.

The education feature SHAP scatterplot in Figure 5c displays a less pronounced parabolic relationship. This may relate to the same trends discussed previously, in which PFAS industries require some critical mass of educated workers, engineers, or others, and local economies in the least educated places in the U.S. simply cannot support these facilities. However, variability increases toward both extremes, which have been proposed due to conflicting economic and social/political circumstances in the most and least socioeconomically advantaged areas. In the most advantaged areas, economic conditions are ripe for industry, but individuals understand and are empowered to oppose PFAS industries in their localities. The opposite is true in underprivileged areas where economic conditions are poor, but social/political opposition is minimal. This dichotomy results in increased variability at extremes.

According to Pearson correlations and initial SHAP analyses, low income areas demonstrate a lesser likelihood of proximal PFAS industry presence, potentially due to a lack of industrial and transportation infrastructure. This positive relationship between income and potential PFAS contamination has been corroborated in other literature (Buekers et. al., 2018). The low income SHAP scatterplot in Figure 5d provides more specific insight, as, once again, high variability appears in percentile extremes, possibly due to aforementioned conflicting influences.

The under age 5 SHAP scatterplot in Figure 5e shows a clear parabolic relationship, revealing that areas with

median young populations generally have the highest concentrations of PFAS industries. This may be attributed to the negative correlation between fertility rates and socioeconomic status (and, by extension, access to healthcare) (Doepke et al., 2023). Areas with lower populations under age 5 have less PFAS industry concentration, potentially due to greater healthcare/public health resources and awareness. Areas with the highest populations under age 5 may simply not have workforces or economies able to operate PFAS industries efficiently. The over age 64 plot Figure 5f shows a negative linear trend, suggesting that older populations face decreased risk of PFAS exposure, as would be expected due to a relative lack of potential heavy industry workers.

Our results reveal a necessity for greater nuance in the identification of particularly vulnerable communities and the prioritization of PFAS testing areas. By implementing comprehensive and targeted testing protocols guided by quantitative research, public health needs can be more effectively met within resource constraints, minimizing the impact of PFAS contamination on communities across the country.

Synopsis

Through quantitative analysis and machine learning, the study reveals the need for additional and prioritized PFAS testing across the U.S with implications for public health.

Author Information

Sameer Menghani and Jonathan Yang are the first authors, that have executed the analyses of this study and compiled the manuscript. Lena Haefele, Louise Carroll, Sathvik Samant, Robert Lee, and Anthony Sapp Guadarrama each assisted immensely with the execution of the study analyses. Andrew Noviello and Alexander Noviello are senior authors, having guided the authorship team, contributed to the study design, and edited the manuscript.

ACKNOWLEDGEMENTS

The authors would like to thank Ajanae Bennett for her mentorship, George Negroponte for his sponsorship, and Stephen Laubach for his advocacy on behalf of the authorship team.

REFERENCES

Ao Y, Li H, Zhu L, Ali S, Yang Z (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering, 174:776-789. https://doi.org/10.1016/j.petrol.2018.11.067.

Buekers J, Colles A, Cornelis C, Morrens B, Govarts E, Schoeters G. (2018). Socio-Economic Status and Health: Evaluation of Human Biomonitored Chemical Exposure to Per- and Polyfluorinated Substances across Status. International Journal of Environmental Research and Public Health. *15*(12). https://doi.org/10.3390/ijerph15122818

Desikan A, Carter J, Kinser S, Goldman G (2019). Abandoned Science, Broken Promises: How the Trump Administration's Neglect of Science Is Leaving Marginalized Communities Further Behind. Union of Concerned Scientists.

Doepke M, Hannusch A, Kindermann F, Tertilt M (2023, June 22). The New Economics of Fertility. IMF. https://www.imf.org/en/Publications/fandd/issues/Series/Analytical-Series/new-economics-of-fertility-doepke-hannusch-kindermann-tertilt

Evans S, Andrews D, Stoiber T, Naidenko O (2020, January 23). PFAS Contamination of Drinking Water Far More Prevalent Than Previously Reported. EWG. https://www.ewg.org/research/national-pfas-testing

Foss J (2023). Addressing PFAS Analysis Challenges and EPA Compliancy. PerkinElmer. https://resources.perkinelmer.com/lab-solutions/resources/docs/app-addressing-pfas-analysis-challenges.pdf

Fox EW, Hill RA, Leibowitz SG, Olsen AR, Thornbrugh DJ, Weber, MH (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environmental Monitoring and Assessment, *189*(7). https://doi.org/10.1007/s10661-017-6025-0

GeoPandas 0.13.2. (2023). GeoPandas. https://geopandas.org/en/stable/index.html

Havlicek, LL, Peterson, NL (1976). Robustness of the Pearson Correlation against Violations of Assumptions. Perceptual and Motor Skills, 43(3_suppl):1319–1334. https://doi.org/10.2466/pms.1976.43.3f.1319

How Much Distance Does a Degree, Minute, and Second Cover on Your Maps? (2023). United States Geological Survey. https://www.usgs.gov/faqs/how-much-distance-does-a-degree-minute-and-second-cover-your-maps/

Hu XC, Andrews DQ, Lindstrom AB, Bruton TA, Schaider LA, Grandjean P, Lohmann R, Carignan CC, Blum A, Balan SA, Higgins CP, Sunderland EM. (2016). Detection of Poly- and Perfluoroalkyl Substances (PFASs) in U.S. Drinking Water Linked to Industrial Sites, Military Fire Training Areas, and Wastewater Treatment Plants. Environmental Science & Technology Letters, *3*(10):344-350. https://doi.org/10.1021/acs.estlett.6b00260

IBM. (2023a). What is logistic regression? IBM. https://www.ibm.com/topics/logistic-regression

IBM. (2023b). How SVM Works. IBM. https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works

Liddie JM, Schaider LA, Sunderland EM (2023). Sociodemographic Factors Are Associated with the Abundance of PFAS Sources and Detection in U.S. Community Water Systems. Environmental Science & Technology, *57*(21):7902-7912. https://doi.org/10.1021/acs.est.2c07255

Liu J, Tang W, Chen G, Lu Y, Feng C, Tu XM (2016). Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Archives of Psychiatry, *28*(2):115–120. https://doi.org/10.11919/j.issn.1002-0829.216045

Lundberg SM, Lee SI. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

MatPlotLib. (2023). MatPlotLib: Visualization with Python. https://matplotlib.org

Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, & Hamprecht FA (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, *10*(1). https://doi.org/10.1186/1471-2105-10-213

National Academies of Sciences (2022, July 28). Guidance on PFAS exposure, testing, and clinical follow-up. The National Academies Press. https://nap.nationalacademies.org/catalog/26156/guidance-on-pfas-exposure-testing-and-clinical-follow-up

Our Current Understanding of the Human Health and Environmental Risks of PFAS. (2023, June 7). United States Environmental Protection Agency. https://www.epa.gov/pfas/our-current-understanding-human-health-and-environmental-risks-pfas

Overview of Socioeconomic Indicators in EJScreen. (2023, January 30). United States Environmental Protection Agency. https://www.epa.gov/ejscreen/overview-socioeconomic-indicators-ejscreen

Pandas. (2023). *Pandas 2.0.3*. https://pandas.pydata.org

Per- and Polyfluoroalkyl substances (PFAS). (2022, September 15). National Institute for Occupational Safety and Health. https://www.cdc.gov/niosh/topics/pfas/default.html

Perfluoroalkyl and Polyfluoroalkyl substances (PFAS). (2023, June 2). National Institute of Environmental Health Sciences. https://www.niehs.nih.gov/health/topics/agents/pfc/index.cfm

PFAS Analytical Methods Development and Sampling Research. (2023, February 16). United States Environment Protection Agency. https://www.epa.gov/water-research/pfas-analytical-methods-development-and-sampling-research

PFAS Analytic Tools. (2023). United States Environmental Protection Agency. https://echo.epa.gov/trends/pfas-tools

PFAS Explained. (2023, April 10). United States Environmental Protection Agency. https://www.epa.gov/pfas/pfas-explained

Salvatore D, Mok K, Garrett KK, Poudrier G, Brown P, Birnbaum LS, Goldenman G, Miller MF, Patton S, Poehlein M, Varshavsky J, Cordner A (2022). Presumptive Contamination: A New Approach to PFAS Contamination Based on Likely Sources. Environmental Science & Technology Letters, *9*(11):983-990. https://doi.org/10.1021/acs.estlett.2c00502

Scikit-learn. (2023). Scikit-learn Machine Learning in Python. Scikit-learn. https://scikit-learn.org/stable

SciPy *1.11.1*. (2023). SciPy. https://scipy.org

SHAP. (2023). Welcome to the SHAP Documentation. SHAP. https://shap.readthedocs.io/en/latest

SimpleMaps. (2023, January 31). United States Cities Database. *US Cities Database*. https://simplemaps.com/data/us-cities

Sokolowski T (2023, July 6). EPA Announces New Framework for New PFAS in Industry. PFAS Project Lab - Northeastern University. https://pfasproject.com/2023/07/06/epa-announces-new-framework-for-new-pfas-in-industry

Speiser JL, Miller ME, Tooze J, Ip E (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems With Applications, 134:93-101. https://doi.org/10.1016/j.eswa.2019.05.028

TIGER/Line Shapefiles. (2022, December 5). United States Census Bureau. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

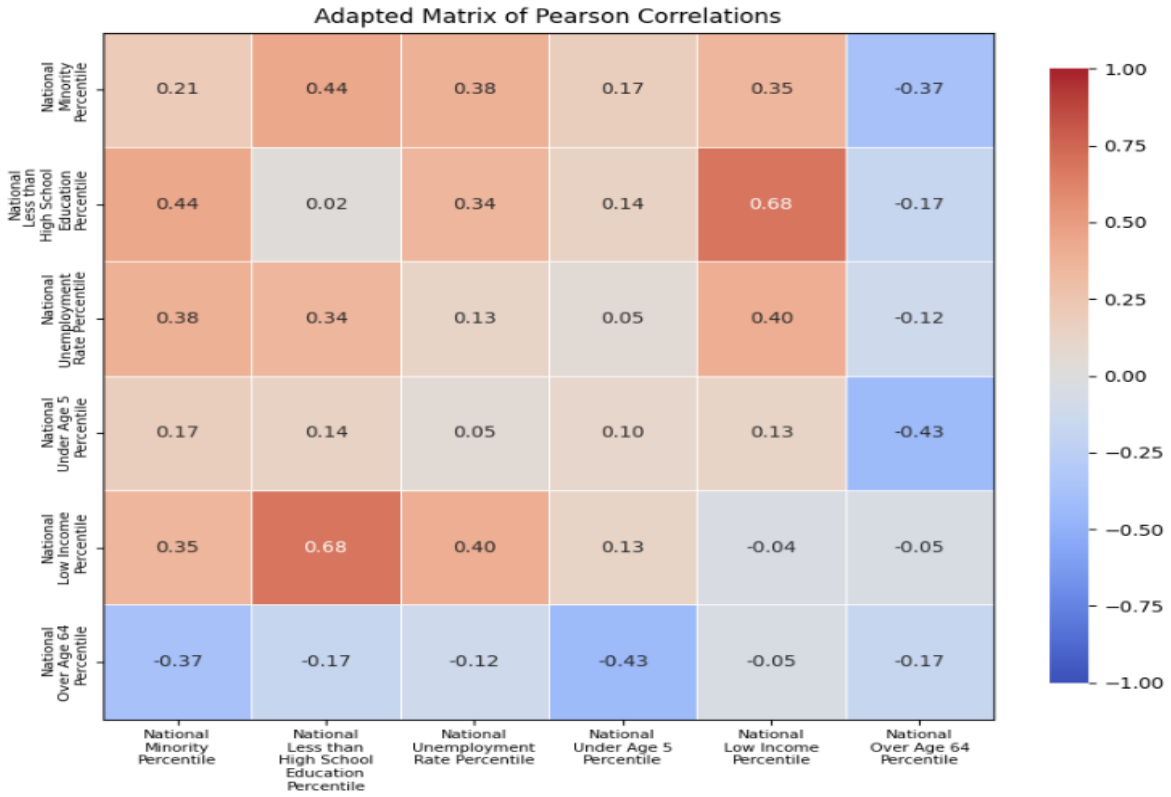
Understanding PFAS Exposure and Your Body. (2022, November 1). Agency for Toxic Substances and Disease Registry. https://www.atsdr.cdc.gov/pfas/health-effects/PFAS-exposure-and-your-body.html

Wang L, Liu ZP, Zhang XS, Chen L (2012). Prediction of hot spots in protein interfaces using a random forest model with hybrid features. Protein Engineering, Design, and Selection, 25(3):119-126. https://doi.org/10.1093/protein/gzr066

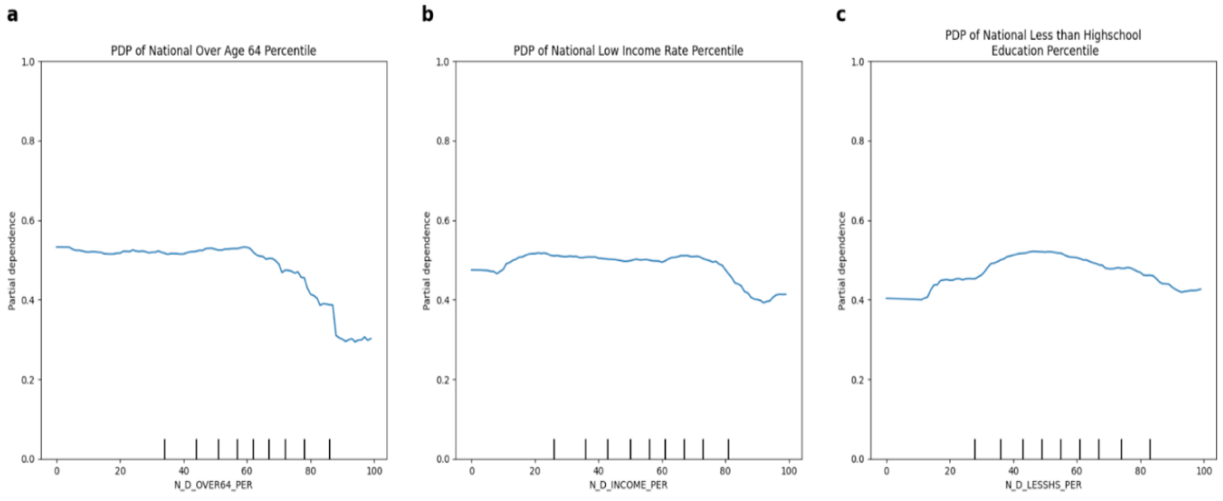
What are PFAS? (2022). Agency for Toxic Substances and Disease Registry. https://www.atsdr.cdc.gov/pfas/health-effects/overview.html

Xu W, Zhang J, Zhang Q, Wei X (2017). Risk prediction of type II diabetes based on random forest model [Conference session]. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, India. https://ieeexplore.ieee.org/abstract/document/7972337

Appendix



Appendix 1: Pearson correlation matrix for each of the six socioeconomic features. Principal diagonal elements represent correlations between corresponding features and binary proximal PFAS industry presence.



Appendix 2: Partial Dependence Plots (PDPs) for the Random Forest model on three socioeconomic features (over age 64, low income, and education features), revealing the complex, non-linear, and non-monotonic relationships between socioeconomic disparity and PFAS industry presence.