ISyE 6740 - Fall 2024
Final Project

Team Member Names: Daniel Solon

Project Title: Comparing Water Source Site Types and Predicting Contamination of Tap Water exposed to Per- and Polyfluoroalkyl Substances (PFAS) in the United States

1 Problem Statement

According to the Environmental Working Group, more than 50 percent of thousands of Americans surveyed say their tap water is unsafe, and 40 percent would not or cannot drink it. This means there is still a majority of people drinking from the tap. Different contaminants including PFAS can be found in tap water, filtered or unfiltered. Even bottled waters are not safe from these contaminants because there is no government requirement for PFAS testing of bottled water, no public information about potential PFAS contamination of water supplies that manufacturers use for production of bottled water, and no guarantee that the levels of PFAS in bottled waters are lower than those of tap water.

Exposure to PFAS is inevitable because they are extremely persistent, lasting thousands of years. Our water source, whether it is from a public supply or a private well, are not exempt from being contaminated by these forever chemicals. However, there may be a difference in the types and amounts of PFAS found in these sites. This project will compare these site types in terms of PFAS contamination and predict if tap water from a site exceeds the Maximum Contaminant Levels (MCLs) of PFAS in accordance to the U.S. Environmental Protection Agency (EPA) standards.

2 Data Source

PFAS data can be found from the United States Geological Survey's (USGS) website providing different types and levels per state for years 2016 to 2021 [3]. In addition, USGS has an interactive map showing PFAS types, its corresponding amount found in tap water, and surrounding PFAS facilities at a point of interest anywhere in the United States.

Studies with respect to PFAS is limited and most research focuses on its effect to people such as contracting kidney cancer [2] or testicular cancer [1] when exposed to high levels of combinations of different PFAS types. No known studies have explored classifying site types contaminated by PFAS nor predicting if its contamination in a site exceeds EPA standards (Table 1). These standards are used to derive labels for the dataset used in this project which determines whether a site exceeds contamination levels with respect to certain PFAS types.

Compound	MCL
PFOA	4.0 ppt
PFOS	4.0 ppt
PFHxS	10 ppt
PFNA	10 ppt
HFPO-DA (commonly known as GenX Chemicals)	10 ppt
Mixtures containing two or more of PFHxS, PFNA, HFPO-DA, and PFBS	1 Hazard Index

Table 1: Maximum Contaminant Levels (MCLs) for select PFAS Compounds

The Hazard Index in Table 1 can be calculated using a formula provided by the EPA as follows:

$$\text{Hazard Index (unitless)} = \left(\frac{\text{HFPO-DA}_{\text{ppt}}}{10\,\text{ppt}}\right) + \left(\frac{\text{PFBS}_{\text{ppt}}}{2000\,\text{ppt}}\right) + \left(\frac{\text{PFNA}_{\text{ppt}}}{10\,\text{ppt}}\right) + \left(\frac{\text{PFHxS}_{\text{ppt}}}{10\,\text{ppt}}\right) \quad (1)$$

3 Methodology

This section will explain the steps taken to preprocess the data and the methods used to answer the problems in this study.

3.1 Data Preprocessing

The dataset had missing values labeled as "NA" (not applicable) and some values are labeled "nd" (not detected). These values were handled as follows:

- Subset 1: replaced "NA" and "nd" with zeros
- Subset 2: removed columns with "NA" and replaced "nd" with zeros
- Subset 3: removed a column if it contained more than 50 counts of "NA" and removed rows with "NA"; "nd" was replaced with zeros

The best performing subset, in terms of the metric used in subsection 3.3, will be chosen for analysis. Next, outliers were removed from the data which are more than three standard deviations away from the mean to prevent skewed results; this method is feasible if the distribution of the data is normal. The data has approximately 200 data points, large enough to assume normality in accordance to the Central Limit Theorem. Next, PFAS types that have all zero values were excluded, meaning there is no significant information with respect to that type. Finally, the data is standardized through normalization to bring different variables to a similar scale, ensuring that certain variables do not dominate others due to their magnitudes.

3.2 Comparing Water Source Site Types

To compare the two water source site types, the data was plotted, and each type was labeled with a unique color. Given that the preprocessed data has approximately ten PFAS types, Principal Component Analysis (PCA) was applied, then the first six principal components were plotted to see if it is visually possible to split the data into two groups. The results are shown in Figure 1.

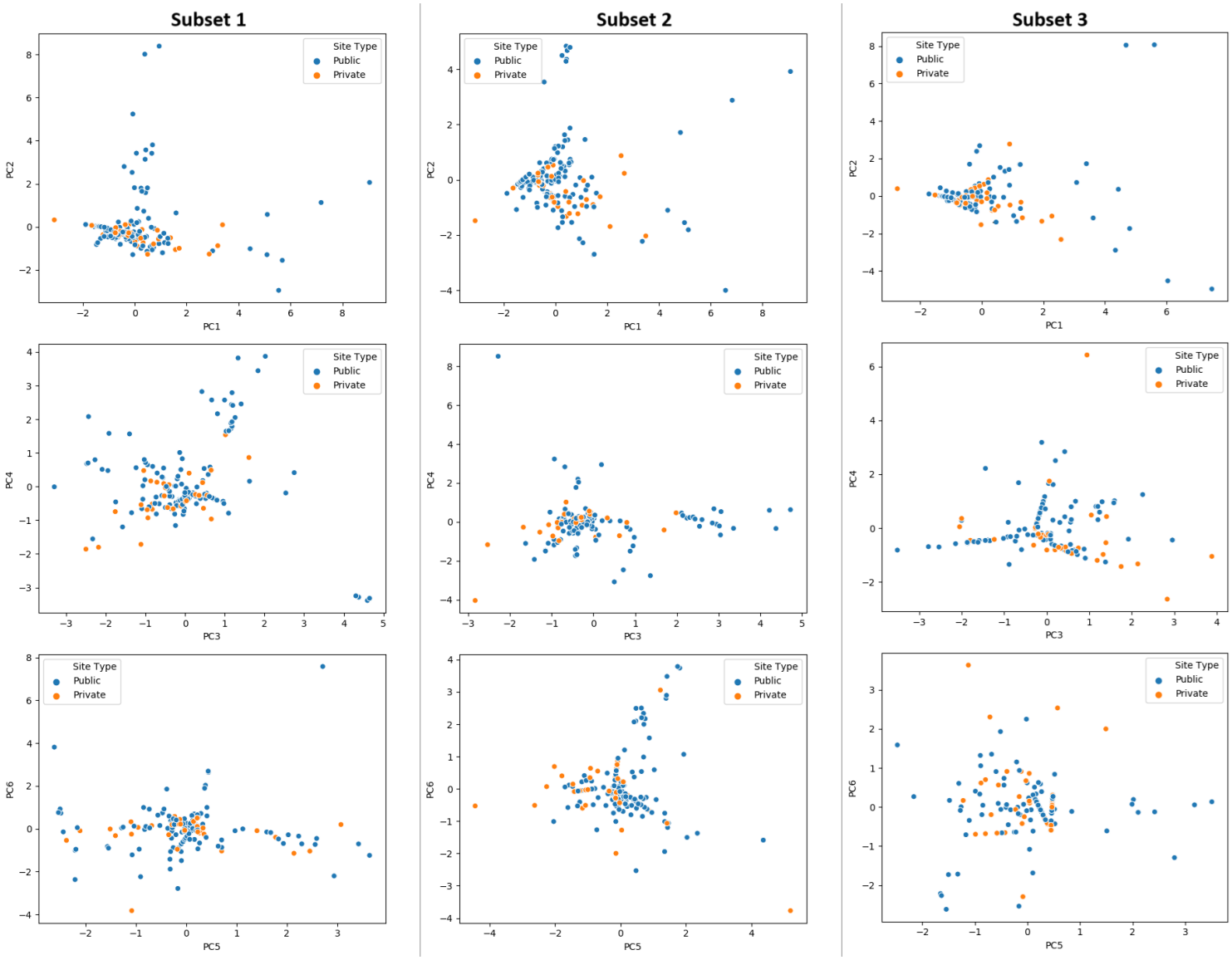

Figure 1: PCA applied on preprocessed data showing the top six Principal Components (PCs)

Based on these results, the site types cannot be separated when using PCA on all three subsets. Therefore, there is no significant difference in the amounts and types of PFAS found among site types. To support this result, a two sample t-test was used to determine whether these two groups are indeed similar given data distribution normality and equal variance assumptions. Let us define the following null and alternative hypothesis respectively:

$$H_0 : \mu_{public} = \mu_{private}$$

$$H_A : \mu_{public} \neq \mu_{private}$$

where μ is the mean of each group

For this experiment, the first principal component from the PCA results of subset 2 was used. Below are the formulas to calculate the test statistic t, pooled variance s_p^2, and the degrees of freedom df:

$$t = \frac{\bar{X}_{public} - \bar{X}_{private}}{\sqrt{s_p^2 \left(\frac{1}{n_{public}} + \frac{1}{n_{private}}\right)}}$$

$$s_p^2 = \frac{(n_{public} - 1)s_{public}^2 + (n_{private} - 1)s_{private}^2}{n_{public} + n_{private} - 2}$$

$$df = n_{public} + n_{private} - 2$$

where $\bar{X}$ is the sample mean and n is the number of samples in each group

After calculating the test statistic and the critical value with df and a significance level of 0.05, we fail to reject H_0 since the test statistic is less than the critical value. Therefore, there is no significant difference in site types.

3.3 Predicting PFAS Contamination in Tap Water

The dataset does not have labels as to whether PFAS in tap water from a site exceeds or is below the maximum contamination levels mandated by the EPA. Using the MCL thresholds from Table 1 and the formula to calculate the Hazard Index from equation (1), binary labels were generated to determine whether a site exceeds these limits.

PCA was applied on the subsets, and the first two principal components were plotted to see if the groups, above or below MCL, are separable. After visually confirming that it is indeed separable, the subset that is visually the easiest to separate was chosen. Then, different classifier models were applied on the first two principal components, and on all the PFAS types (which will also be referred to as features). Using all the features, PCA was applied on the chosen subset to remove correlation between them.

4 Evaluation

In this section, the results obtained from classifying whether PFAS found in tap water from a site exceeds MCL will be explained. On Figure 2, all the classifier models used on the first two principal components of subset 2 are shown. The blue circles represent sites with PFAS below

MCL, and the orange circles represent sites with PFAS above MCL. Circles with lower opacity represents the training data points. The boundary lines of the classifiers are also shown in the background as contours. The cross-validated (CV) accuracy score of the classifiers are summarized on Table 2.

Classifier	**CV Score**
Nearest Neighbors	0.8968
Linear SVM	0.8597
RBF SVM	0.8925
Gaussian Process	0.9025
Logistic Regression	0.8751
Decision Tree	0.8682
Random Forest	0.9021
Neural Net	0.8941
AdaBoost	0.8744
Naive Bayes	0.8270
QDA	0.8227

Table 2: Classifier cross-validated (CV) accuracy scores based on the first two Principal Components

From this table, Gaussian Process obtained the highest CV score followed by Random Forest when applied on the first two principal components resulting from applying PCA on subset 2. For Logistic Regression, L1 regularization was used to benefit from its capability of feature selection. This time, let us see how the same classifiers perform when using all the features for training.

Classifier	**CV Score**
Nearest Neighbors	0.9026
Linear SVM	0.8883
RBF SVM	0.8497
Gaussian Process	0.9512
Logistic Regression	0.9416
Decision Tree	0.9212
Random Forest	0.9021
Neural Net	0.9610
AdaBoost	0.9319
Naive Bayes	0.8411
QDA	0.2552

Table 3: Classifier cross-validated (CV) accuracy scores

From Table 3, Neural Networks obtained the highest CV score, followed by Gaussian Process when using all the features of subset 2. QDA, in contrast, performed poorly because of collinear features.

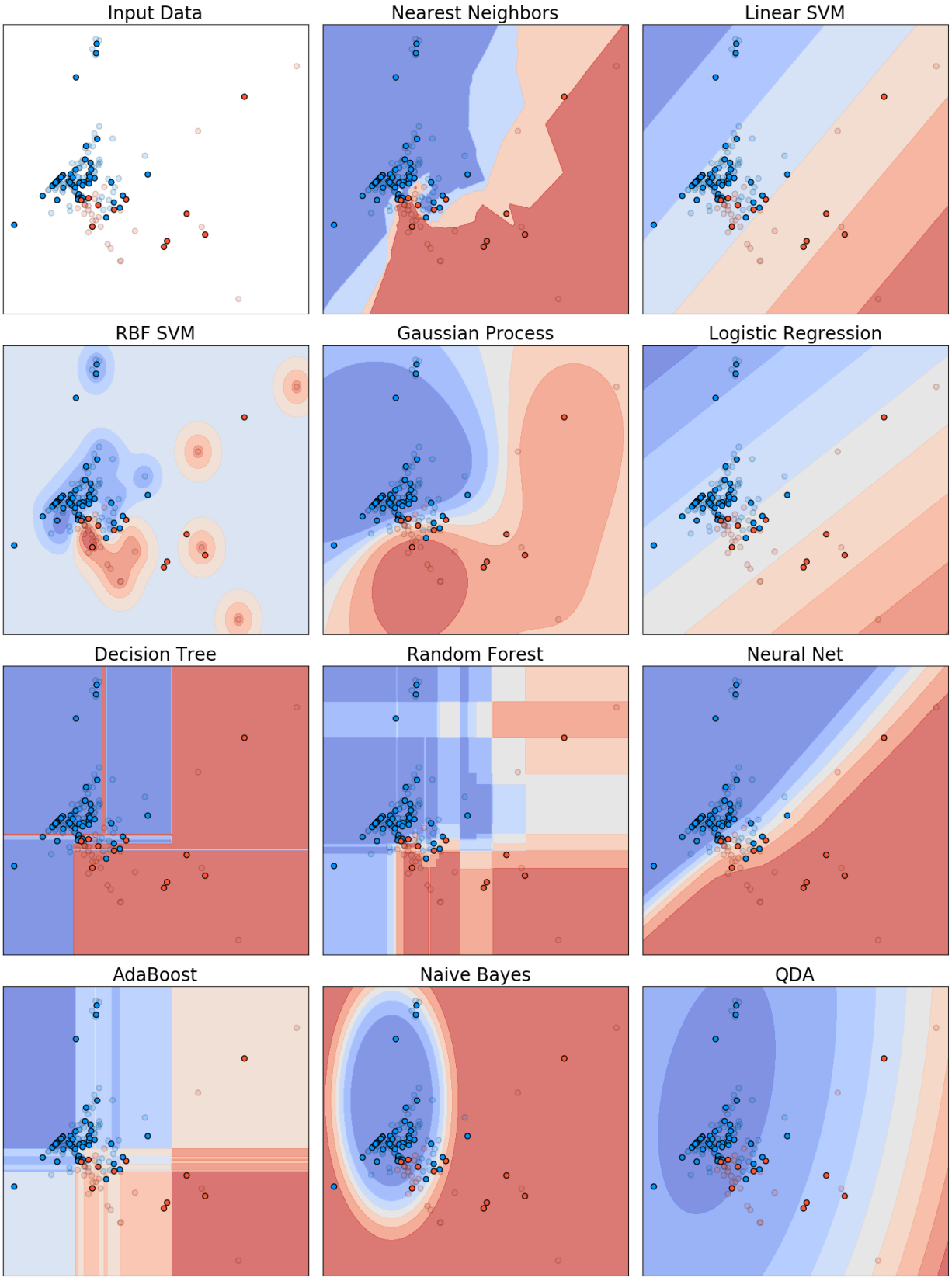

Figure 2: Classifier models on the first two Principal Components of Subset 2

5 Conclusion

Public supply and private wells were compared in terms of PFAS contamination, resulting in insignificant difference between the two. Classifier models, such as Neural Networks and Gaussian Process, were found effective in predicting whether PFAS found in tap water is above maximum contamination levels based on EPA standards.

References

[1] Mark P. Purdue, Jongeun Rhee, Hristina Denic-Roberts, Katherine A. McGlynn, Celia Byrne, Joshua Sampson, Julianne Cook Botelho, Antonia M. Calafat, and Jennifer Rusiecki. 2023. A Nested Case–Control Study of Serum Per- and Polyfluoroalkyl Substances and Testicular Germ Cell Tumors among U.S. Air Force Servicemen. *Environmental Health Perspectives* 131, 7 (2023), 077007. https://doi.org/10.1289/EHP12603 arXiv:https://ehp.niehs.nih.gov/doi/pdf/10.1289/EHP12603

[2] Joseph J Shearer, Catherine L Callahan, Antonia M Calafat, Wen-Yi Huang, Rena R Jones, Venkata S Sabbisetti, Neal D Freedman, Joshua N Sampson, Debra T Silverman, Mark P Purdue, and Jonathan N Hofmann. 2020. Serum Concentrations of Per- and Polyfluoroalkyl Substances and Risk of Renal Cell Carcinoma. *JNCI: Journal of the National Cancer Institute* 113, 5 (09 2020), 580–587. https://doi.org/10.1093/jnci/djaa143 arXiv:https://academic.oup.com/jnci/article-pdf/113/5/580/37798495/djaa143.pdf

[3] Kelly L. Smalling, Kristin M. Romanok, Paul M. Bradley, Mathew C. Morriss, James L. Gray, Leslie K. Kanagy, Stephanie E. Gordon, Brianna M. Williams, Sara E. Breitmeyer, Daniel K. Jones, Laura A. DeCicco, Collin A. Eagles-Smith, and Tyler Wagner. 2023. Per- and polyfluoroalkyl substances (PFAS) in United States tapwater: Comparison of underserved private-well and public-supply exposures and associated health implications. *Environment International* 178 (2023), 108033. https://doi.org/10.1016/j.envint.2023.108033