A Novel Machine Learning Approach for Assessing PFAS Pollution in North Carolina Using Water Contamination Sources

Lakshmanan, Aadarsh (School: Ardrey Kell High School)

Per- and poly-fluoroalkyl substances (PFAS) are chemical compounds widely used in consumer products and industrial applications, this contamination has ties to severe health risks such as cancer liver damage, and ulcerative colitis (UC), as identified by CDC studies. North Carolina (NC) faces significant PFAS contamination from various sources including landfills, military sites, and chemical plants. In order to test the contamination of various sites it requires significant resources and cost therefore, the development of a reliable predictive machine learning (ML) model for identifying PFAS presence and concentration level is crucial. Using the data procured from multiple data sources including the EPA and NCDEQ, which utilizes PFAS (20+ sources) transportation trends in order to predict water source contaminations. With a total of 477 test results across NC which were manually annotated and 66 parameters derived and linked using latitude and longitude. The final model is able to predict contamination sites greater than 100 parts-per-trillion (PPT) with 95% accuracy and identifies 0 PPT or no contamination levels with 83% accuracy. Parameters such as distance between landfills and waterways, pre-regulatory landfill sites, and spill data significantly influence PFAS concentration (1-10PPT). Additionally, the count and proximity of chemical factories correlate strongly with high PFAS concentration (>100PPT). Integrating waterway contamination data increases the model's predictive performance, increasing the ROC curve from the lower 70s to 80+ as well as boosting accuracy for lower-level detection from 60% to 94%. These parameters emerge as crucial for accurate PFAS contamination prediction.

Awards Won: