Forecasting Urban Water Escherichia coli Contamination Using Machine Learning Models

Iyer, Vidhatri (School: University High School of Indiana)

The state of Indiana ranks first in the nation for water recreation impairments. Over 24,000 miles of water streams are polluted and potentially dangerous for human bodily contact. Routine water sampling and analytical methods to measure fecal coliform bacteria, Escherichia coli (E. coli) are time-consuming and only provide retrospective analysis. Thus, real-time alerts for elevated E. coli levels in urban streams is a necessary initiative to protect public health and safety. This research tested the hypothesis that weather parameters play a vital role to accurately predict E. coli levels in water streams. E. coli data was collected for water streams from the Marion County, IN watershed project for a period of 2003-2022. Daily temperature and precipitation data were obtained from the National Oceanic and Atmospheric Administration site. These two sources of data were combined, and initial exploratory data analysis was performed to understand the correlation of parameters to E. coli levels. Next, additional calculated values such as Cumulative Degree Days and Max Precipitation in 10 days or 15 days were included as input for six machine learning models (Logistic Regression, Random Forest, Extra Trees, Decision Tree, Gradient boosting and XGB Classifier). Finally, feature importance analysis and overall accuracy scores across these six machine learning models were compared to identify the best model. XGB classifier consistently had ROC value of above 85% based on thresholds of cumulative degree days and precipitation. This study showed that local weather parameters are critical in predicting E. coli bursts in water streams.

Awards Won: