

Are There Fecal Bacteria In Your Water?



Viet Dinh

Discovery: Class imbalance.

Much less samples with coliforms than without coliforms. As seen with right side of precision recall graph. Precision approaches fraction of samples with coliforms ($3.98\text{E-}3$).

Conclusion: Not many samples will have coliforms. Recall more important than accuracy.

Important to catch cases of possible contamination over being accurate; parallel to screening for brain tumors. Model is optimized for catching most true positives, and gets many false positives due to class imbalance. Conclusion is NYC water quality is top tier; contamination occurs rarely. Tap water is safe to drink.

