Machine Learning Unveils the Hidden Link: Public Water Quality, Cancer Incidence, and Energy Waste in DFW Texas

Zima, Stepan

Dr. Athitsos, Vassilis, Dr. Wilson, Gabriela

Department: CSE & CONHI Project Type: REU

Executive Summary

1. **Water Quality and Cancer Incidence**:
   - A study spanning two decades (2002–2023) collected water quality data from utilities in Dallas, Fort Worth, and Arlington, Texas.
   - Simultaneously, cancer incidence data for 2022 was obtained.
   - Despite efforts, no significant correlation was found between cancer cases and water quality in specific zip codes. Limitations include missing tap water consumption data and complex factors influencing cancer development.

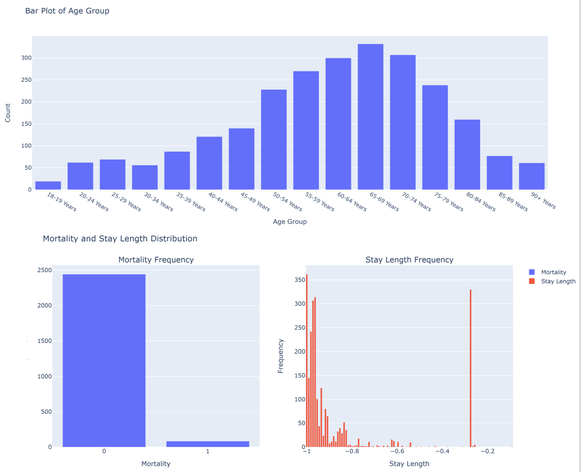
2. **Energy Waste and Wastewater**:
   - New York City and Washington, D.C. utilize wastewater to power water treatment plants, generating excess energy for the city grid.
   - Unfortunately, raw data supporting these claims was inaccessible for analysis.

Background

Data Collection:
- Dallas Water Utilities (DWU)
- Fort Worth Water Department (FWWD)
- Arlington Water Utilities (AWU)
- Texas Health Care Information Collection (THCIC)

Water Quality Assessment:

DWU, FWWD and AWU analyzed:
- 220-250 unique contaminants and characteristics
- 2002 to 2023
- 43-174 distinct locations.
- undisclosed **locations** due to security measures.

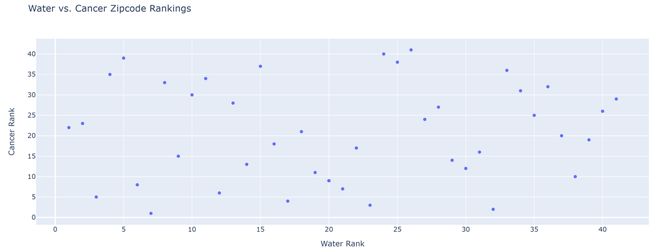


Experimental Setup

Data Preparation
- Cleaned and consolidated water quality and cancer incidence data
- Removed unnecessary columns and consolidated data from three cities

Standardization
- Scaled continuous columns to a range of -1 to 1

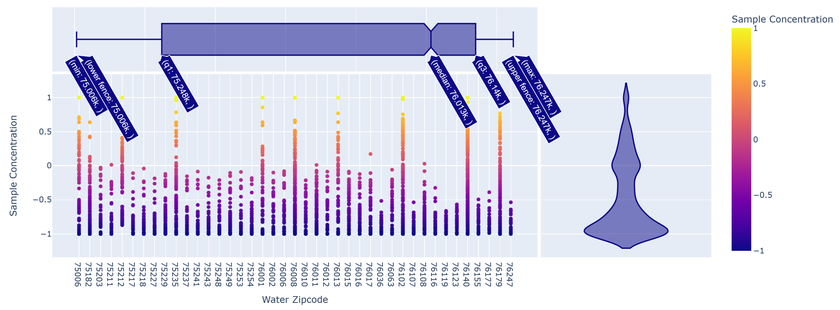
Experimental Design
- Merged water and cancer data
- Conducted exploratory data analysis, correlation analysis, regression modeling, and sensitivity analysis
- Aimed to understand environmental factors influencing cancer risk and inform public health policies



Experimental Test Plan

Key Highlights
- Evaluated 7 machine learning models including eXtreme Gradient Boosting, Random Forest, and Neural Networks
- Aimed to predict patient mortality and link public water contamination to cancer incidence
- Employed statistical tests like t-tests and ANOVA to validate the significance of water data

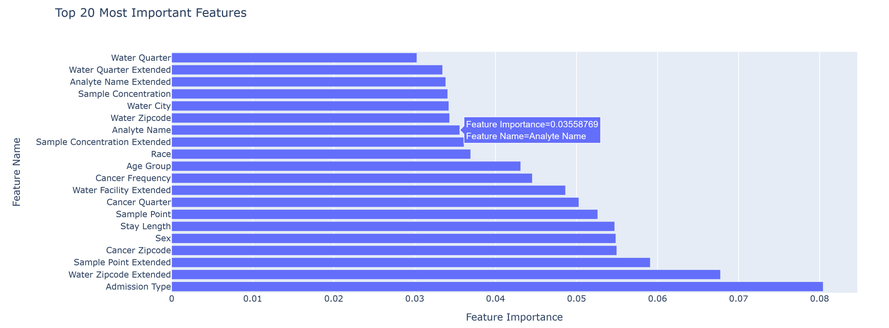
The study meticulously accounted for data **size** differences and excluded duplicate values to ensure **robustness**. Additionally, to calculate energy requirements for water **distillation** and explore **alternative** water purification methods to minimize waste.



Experimental Results

Model	Accuracy
Cancer Predict Cancer (Mortality)	95.06%
Merged Predict Merged (Mortality)	94.86%
T-statistic	-0.58
P-value	0.5636

Predicted Column	F-Statistic	P-Value	Cancer Predict Cancer	Water Predict Cancer	Merged Predict Merged
Mortality	0.1836	0.8323	0.0601	0.0561	0.0541
Admission Type	0.8594	0.4236	1.5252	1.6647	1.5639
Stay Length	1.2915	0.2752	0.0078	0.0077	0.0082
Cancer Frequency	0.0016	0.9984	0.0000	0.0035	0.0000



Conclusions

Cancer vs. Water + Cancer: Both datasets predicted cancer patient mortality with **96% accuracy**, but the Mortality column had the same distribution ratio in both. Water-related data didn't improve predictions.

Distillation Costs: Clean water treatment via distillation costs cities significantly more:
- **Dallas**: $70 million
- **Fort Worth**: $53 million
- **Arlington**: $20 million

Promising Alternative: Wastewater biogas production, used by facilities like Newtown Creek and Blue Plains, eliminates costs and supplies surplus energy back to the grid. Fort Worth already generates **30% of its electricity** this way.

References

1."DEP, EPA and National Grid Celebrate Innovative Project That Converts Wastewater into Renewable Ener." *The Official Website of the City of New York*, 2023, www.nyc.gov/site/dep/news/23-026/dep-epa-national-grid-celebrate-innovative-project-converts-wastewater-renewable#/0.

lgayhardt. "Build Classical Machine Learning Models with Supervised Learning - Training." *Learn.microsoft.com*, learn.microsoft.com/en-us/training/modules/introduction-to-classical-machine-learning/. Accessed 8 Apr. 2024.

2. "TCEQ - Drinking Water Watch." *Dww2.Tceq.texas.gov*, dww2.tceq.texas.gov/DWW/JSP/AnalyteList.jsp?tinwsys_is_number=5796&tinwsys_st_code=TX&wsnumber=TX2200012%20%20%20&DWWState=TX. Accessed 8 Apr. 2024.

3. "The Largest Advanced Wastewater Treatment Plant in the World | DC Water." *Www.dcwater.com*, www.dcwater.com/about-dc-water/what-we-do/wastewater-treatment/blue-plains.