Waterborne Disease Prediction Report

DataVerse Africa

PRESENTED BY : DA - 14

Problem Statement

- Waterborne diseases remain a significant public health concern, especially in underserved communities.
- This project aims to predict total waterborne disease cases and assess community risk levels based on water quality indicators.

Data Collection & Preparation

- **Dataset**: 10,400 records, including water quality indicators and disease case counts.
- **Columns**: Date, Month, Region, Region Code, Community, Country, Turbidity(NTU), E. coli Count(CFU/100ml), Nitrate(mg/L), pH, Cholera Cases, Typhoid Cases, Diarrhea Cases.
- **Missing Dates & Months**: 41 missing values each, handled using forward-fill after sorting by date.
- **Engineered Features**:

  i. Country: Added based on community research.

  ii. Total Waterborne Cases = Cholera + Typhoid + Diarrhea.

  iii. Risk Level: High (≥10), Medium (5–9), Low (<5).

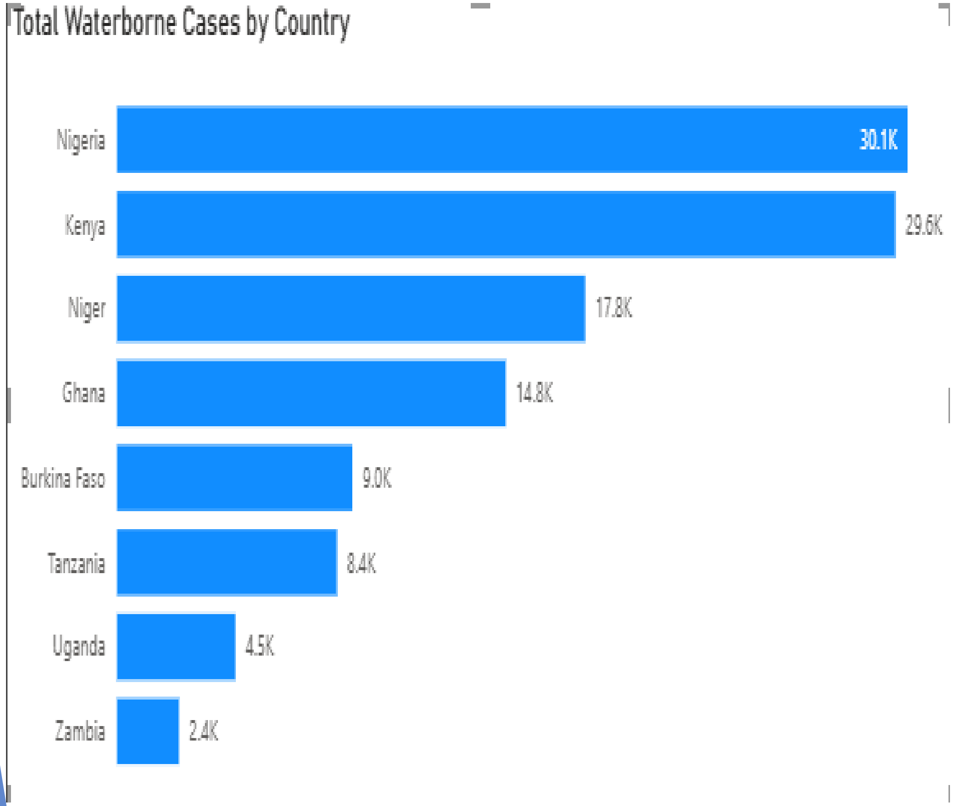
Outlier Analysis

- **Initial features included:** Turbidity, E. coli Count, Nitrate, pH, Cholera Cases, Typhoid Cases, Diarrhea Cases.
- **Strong multicollinearity was found:** Total Waterborne Cases was a perfect sum of Cholera, Typhoid, and Diarrhea cases.
- To avoid data leakage, disease case columns were removed from predictors, leaving only water quality features.

Model Insights

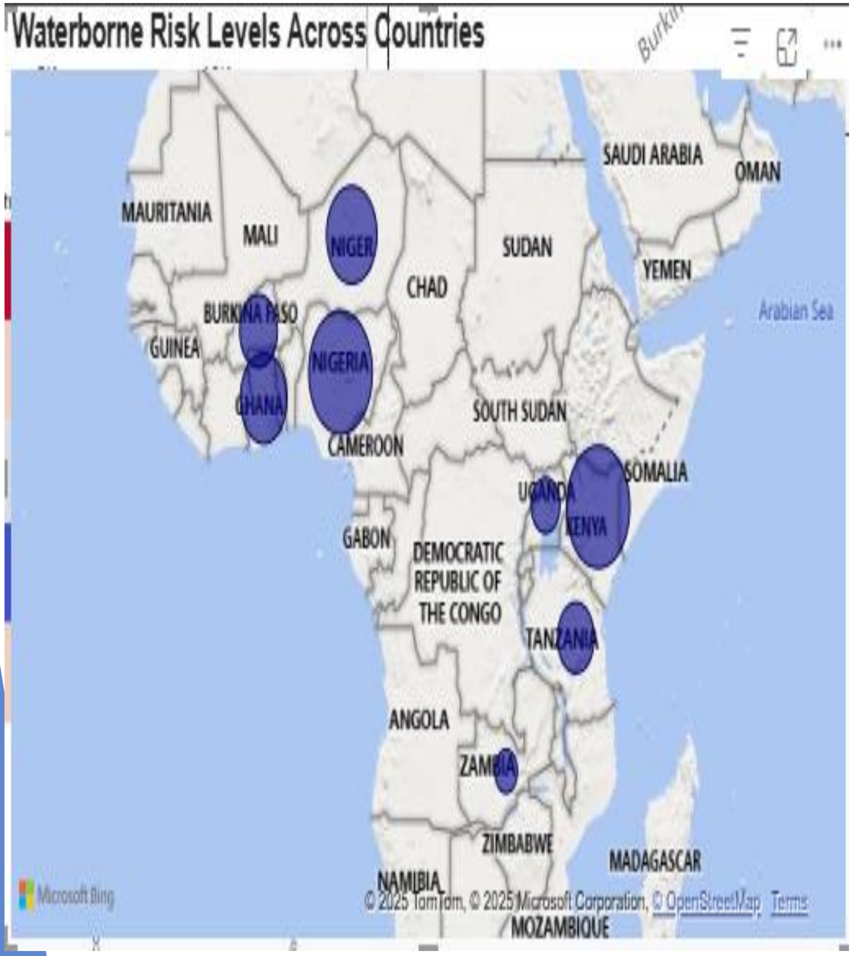
- Health Impact on Waterborne Cases across African Communities.
- Spatial Risk Level across African Communities.
- Correlation Heat Map.

Health Impact on Waterborne Cases across African Communities



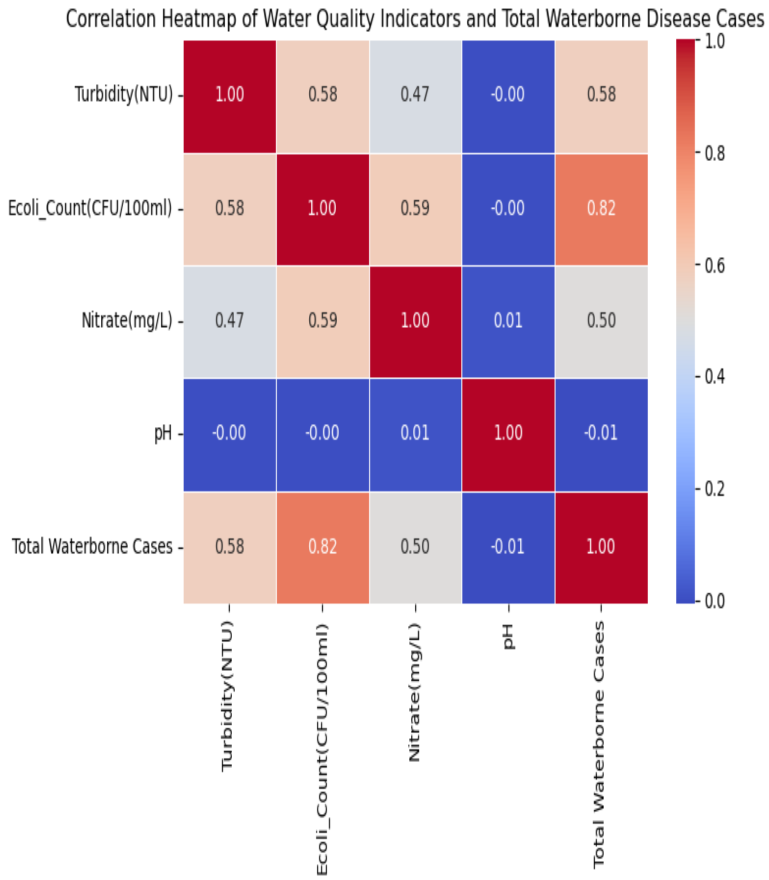
Total waterborne cases were reported to be highest in Nigeria, representing approximately 27% of total cases.

Spatial Risk Level across African Communities



Nigeria and Kenya were reported to have more high-risk cases than the other countries.

Correlation Heat Map



- **The E coli count was the most significant predictor of the disease cases, with a correlation coefficient of 0.82, implying a strong positive correlation.**
- **On the other hand, the pH was seen to have a weak and negative relationship with the disease cases, with a correlation coefficient of -0.01**

Modeling Approach

- Regression (Primary Task):
- Target: Total Waterborne Cases
- Models:
  - Linear Regression (baseline)
  - Random Forest Regressor
  - XGBoost Regressor
  - LSTM (optional time series enhancement)
- Classification (Secondary Task):
- Target: Risk Level
- Models:
  - Logistic Regression
  - Random Forest Classifier
  - XGBoost Classifier

Regression Results

- **Linear Regression:** MSE=10.99, R^2=0.694
- **Random Forest:** MSE=12.41, R^2=0.655
- **XGBoost:** MSE=11.73, R^2=0.674
- **LSTM:** MSE=17836.80, R^2=0.674

Classification Results

- **Logistic Regression:**

  Accuracy=0.78, Macro F1=0.72, Weighted F1=0.77

- **Random Forest Classifier:**

  Accuracy=0.76, Macro F1=0.72

  Precision: High=0.83, Low=0.75, Medium=0.56

- **XGBoost Classifier:**

  Accuracy=0.76, Macro F1=0.72

  Precision: High=0.84, Low=0.76, Medium=0.56

Policy Recommendations

- Invest in sensors and IoT solutions for continuous water quality tracking in high-risk communities.
- Use the risk classification to prioritize vaccination, health education, and sanitation initiatives in vulnerable zones.
- Promote hygiene and water treatment education, particularly in areas flagged as high-risk by the model.
- Conduct regular water quality monitoring focusing on turbidity, E. coli, nitrate, and pH.
- Launch awareness campaigns in high-risk areas.
- Integrate predictive analytics into public health planning to allocate resources more effectively.

Impact of Data-Driven Insights on Waterborne Disease Control

- Provide actionable guidance to NGOs, health ministries, and WASH (Water, Sanitation, and Hygiene) programs for strategic deployment of water sanitation units in vulnerable and at-risk communities.
- Enable early prediction of disease outbreaks to reduce response times, allowing for timely intervention before situations escalate.
- Support evidence-based budgeting and optimal resource allocation by leveraging data-driven insights, ensuring funds and efforts are directed where they are most needed.

THANK YOU FOR LISTENING