CLEAN WATAI

Predicting Water Contamination Risk in Rural Communities

Group 7

Aluoch Phanela
Anthony Chege
Diana Mayalo
Lewis Mwaki
Margaret Kariuki

PROBLEM STATEMENT

💧 Background:

Clean water is a fundamental human right, yet over 2 billion people worldwide still lack access to safely managed drinking water.

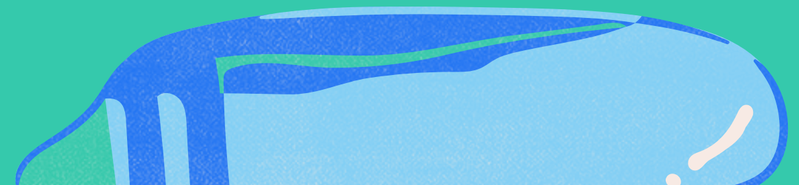
Project Goal:

Develop a machine learning model that predicts contamination risk in rural water points using environmental and infrastructure data, reducing reliance on costly and time-consuming laboratory testing.

🌍Impact:

Enable NGOs and governments to take proactive action, preventing public health crises and improving access to safe water.

.



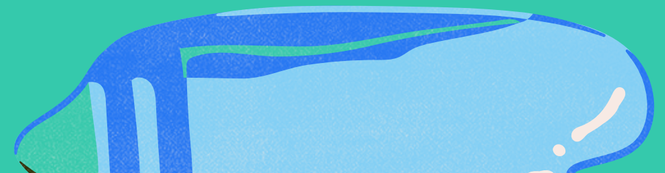
OBJECTIVES

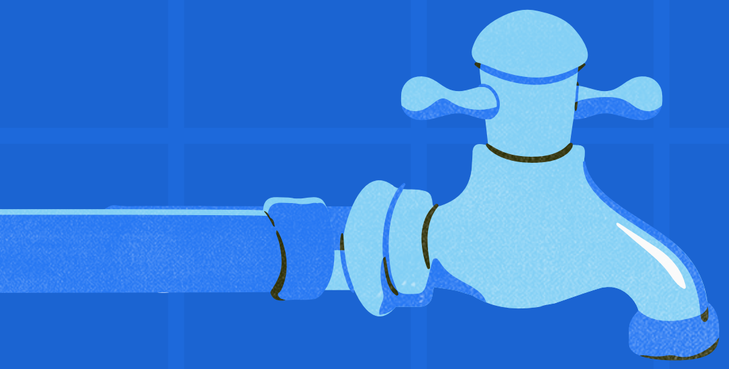
WHAT WE AIM TO ACHIEVE

- Collect and preprocess geospatial, weather, and textual data for Kenyan rural communities.
- Build and validate a contamination risk prediction model.
- Integrate NLP for early-warning signals.
- Develop a dashboard or app for actionable insights
- Enable actionable intervention planning.

DATA SOURCES

- GEMS/UNEP Water Quality Monitoring (GEMStat)
- WPDx Kenya– Water Point Metadata
- Remote sensing – GEE (Google Earth Engine) – vegetation indices (population, temperature,ndvi,chirps)
- World Bank WASH Infrastructure Data
- NGO field reports (simulated or public reports)





MODELING OVERVIEW

WPDx attributes merged with environmental variables from Google Earth Engine. (WPDx+GEE)

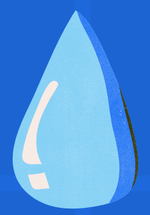
➢ Structured Data Models:

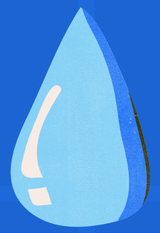
- Logistic Regression
- Random Forest
- XGBoost
- LightGBM

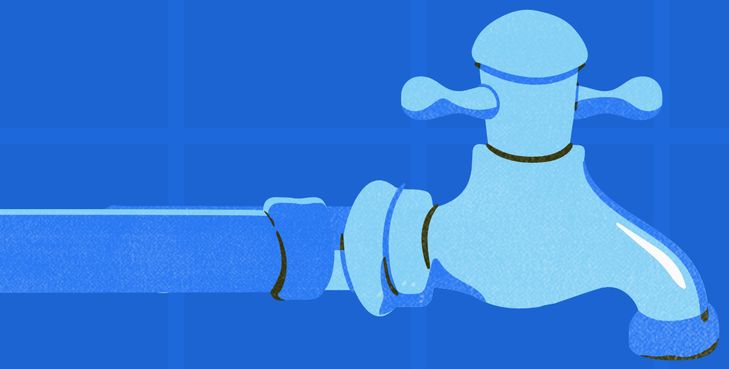
➢ NLP Models:

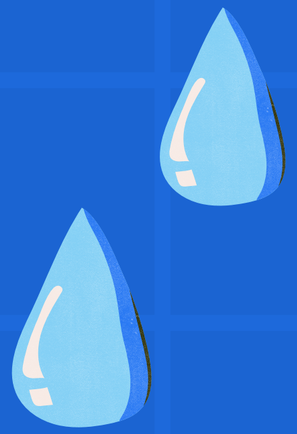
- TF-IDF + SVM
- Fine-tuned DistilBERT

➢ Interpretability: SHAP and LIME for feature impact analysis



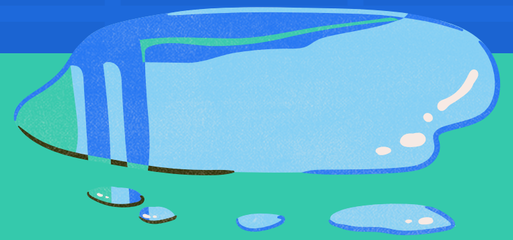


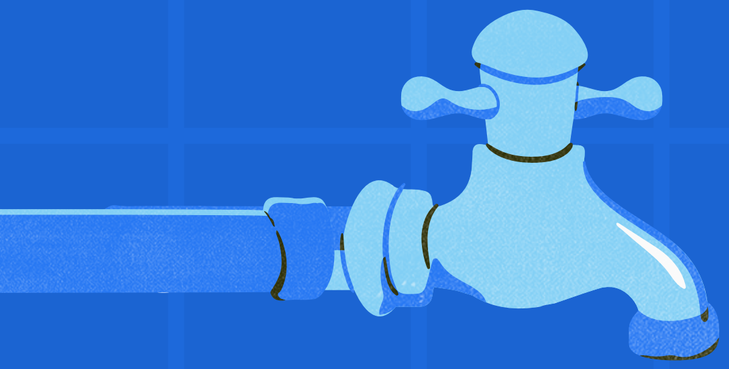


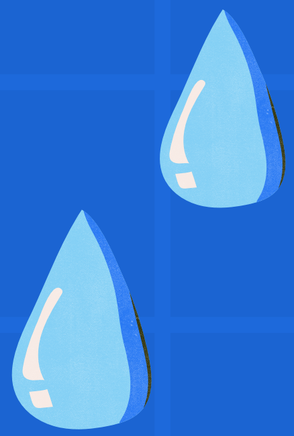


XGBOOST MODEL

- Risk score based on population, water pressure, and surface temp
- Risk levels: 0 (Low) to 3 (High)
- Balanced data using smart sampling
- Final model: XGBoost—high accuracy

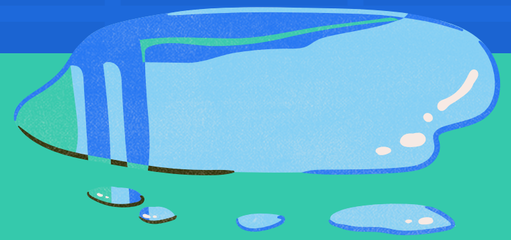


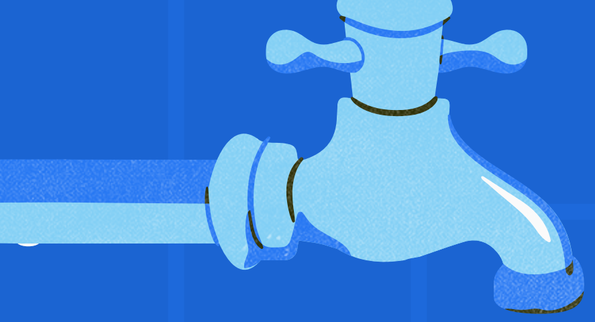




GEMS DATASET & ISOLATION FOREST

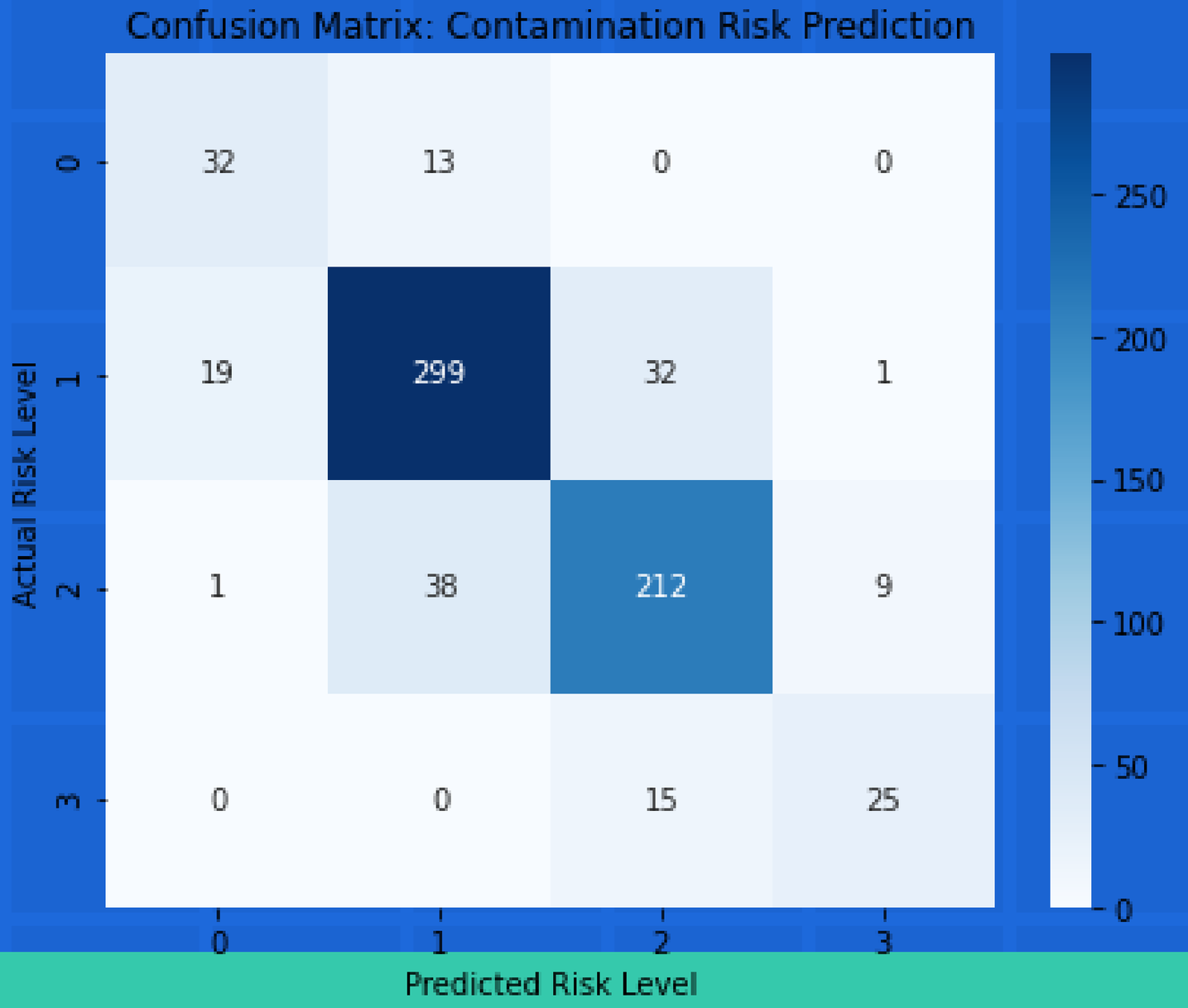
- GEMS – enhancing our data quality.
- Maintained the XGBoost model while adding GEMS to include chemical accuracy
- Trained an Isolation Forest: pH, temperature, conductivity.
- Enabled real-time anomaly detection even without expensive lab testing.



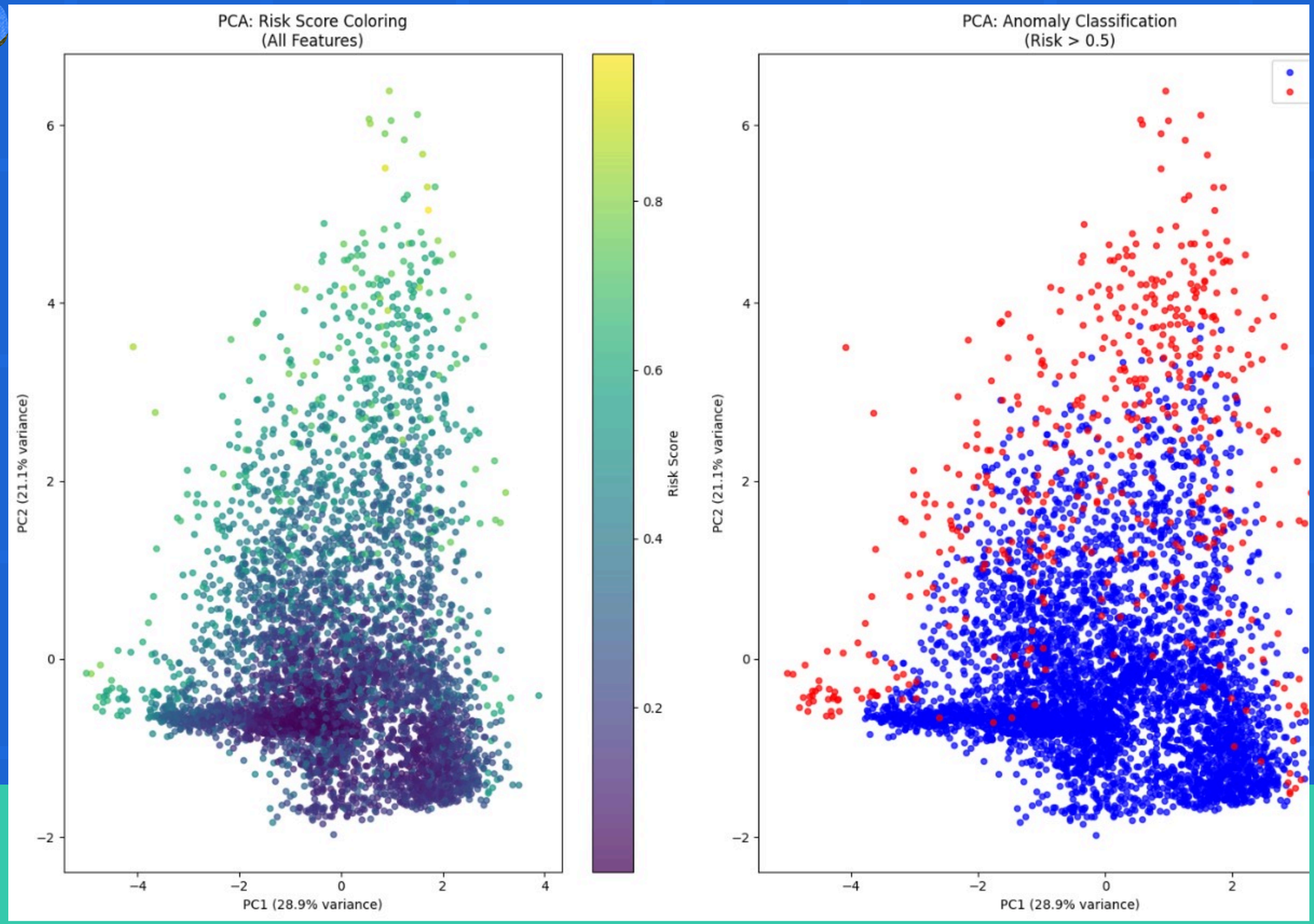


MODEL PERFORMANCE XGBOOST

- 81% overall accuracy
- 63% accuracy in predicting high-risk areas
- Supports targeted interventions



MODEL PERFORMANCE ISOLATION FOREST



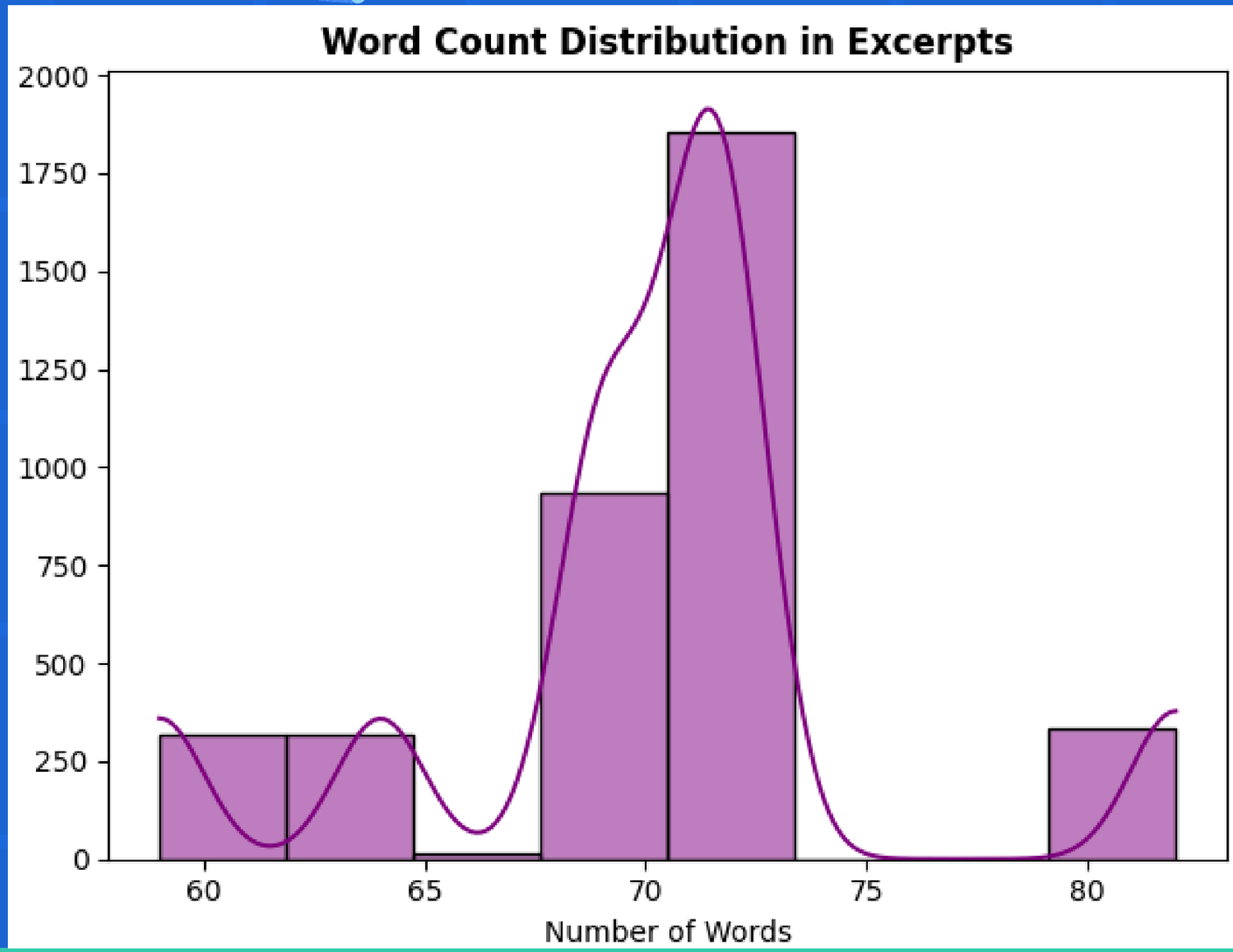
NLP MODEL FROM FIELD REPORTS

Text Preprocessing

- Cleaned, tokenized, lemmatized
- Removed noise (stop words, digits, punctuation)

Feature Extraction

- TF-IDF vectorization on critical terms: contaminated, sewage, outbreak
- Binary classification: Safe vs. Unsafe

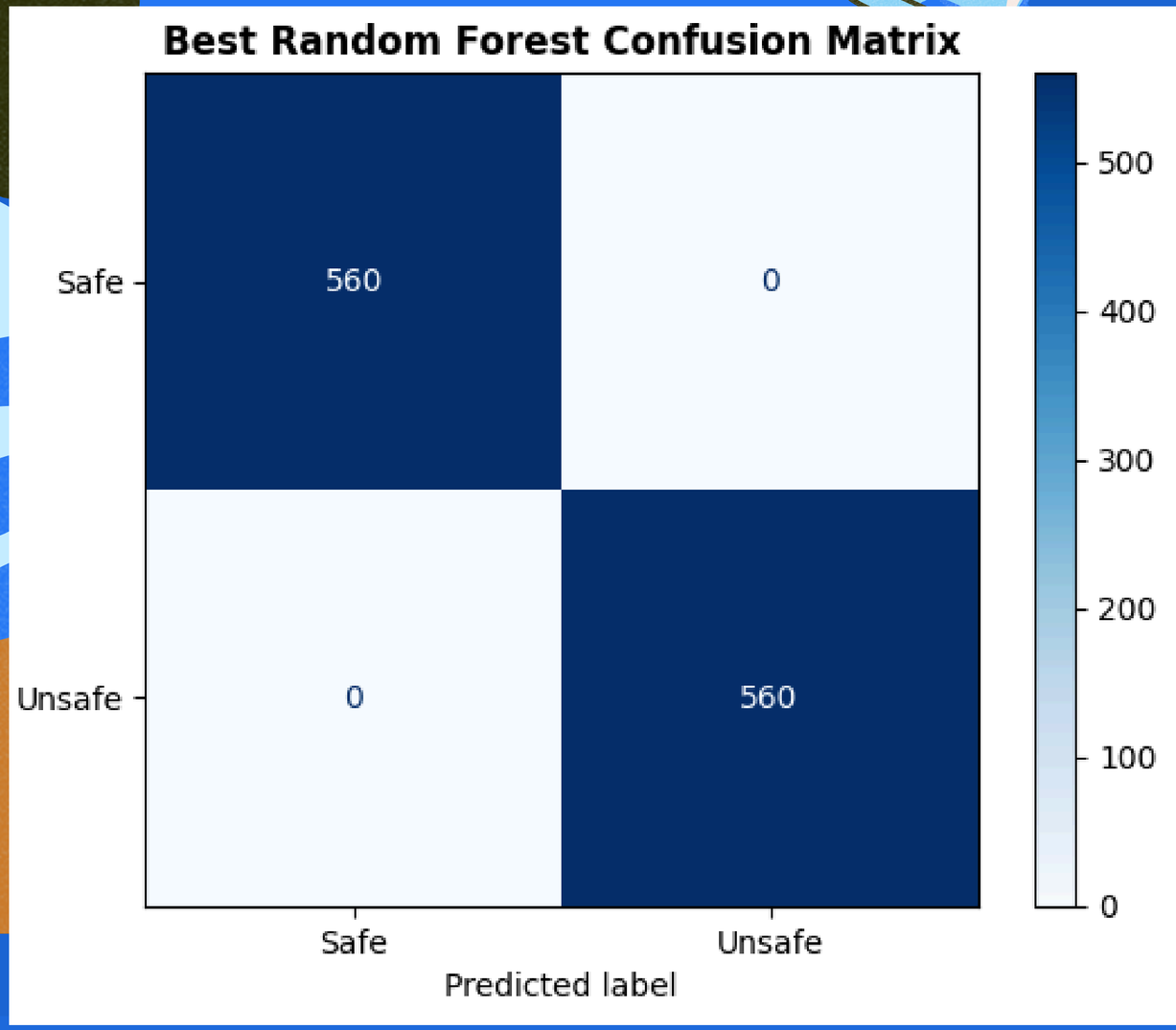


Key Insights

- Frequent in Unsafe: diarrheal, untreated, sewage
- “Safe” reports were fewer and textually repetitive

Results

- Accurately classified unseen water-related news
- Combined NLP + structured data = explainable, high-performing model



DEPLOYMENT

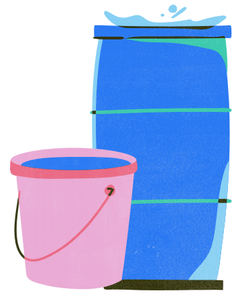
- Hosting: Deployed via Streamlit Cloud, enabling fast and accessible web-based interaction.
- Frontend: Developed using Streamlit, offering a lightweight and interactive interface for visualizing contamination risk scores by region and monitoring point.
- Backend: The model is loaded directly in the Streamlit app using joblib, eliminating the need for a separate FastAPI or Flask server.

Future iterations will integrate FastAPI + PostgreSQL for scalable backend and persistent storage.

NEXT STEPS

- Integrate real-time geospatial feeds
- Expand model to other East African regions

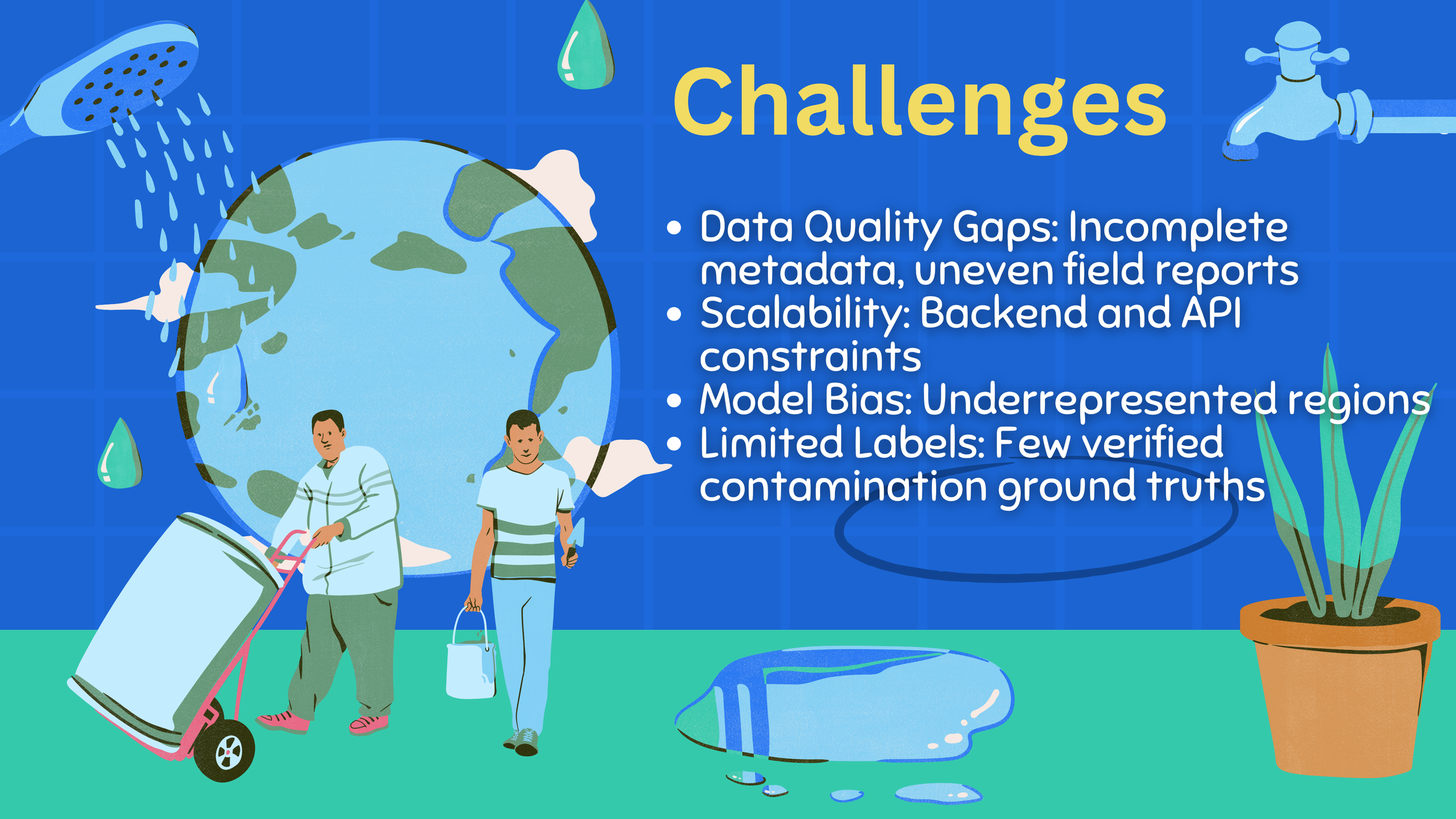
- Collaborate with local water boards for on-ground validation
- Integrate live water crisis feeds using web scraping or APIs
- Expand Streamlit app with multi-language support



- Fine-tune NLP with more field reports
- Incorporate user feedback into app

Challenges

- Data Quality Gaps: Incomplete metadata, uneven field reports
- Scalability: Backend and API constraints
- Model Bias: Underrepresented regions
- Limited Labels: Few verified contamination ground truths



EVERY DROP COUNTS!

Recommendations

- Collaborate with local NGOs for data collection
- Combine mobile + satellite data for more context
- Involve community feedback loops
- Prioritize mobile-first deployment

THANK YOU