# Predicting Water Well Conditions in Tanzania

Machine Learning Approach for Maintenance Planning

# Problem Statement

- Water access is critical in rural Tanzania
- Many wells are functional or need repair .

*<u>Objectives</u>*

- Predict Water Well Condition: Build a machine learning model to classify whether a water well is functional or non-functional using features .
- Support Maintenance Prioritization: Provide actionable insights to help NGOs and government agencies identify wells that likely need repair.
- Discover Risk Patterns: Analyze the key factors contributing to well failure.

Data Overview

Approach and Methodology

Building Models and iterations

Evaluations and Conclusions

Recommendations

# Data Overview

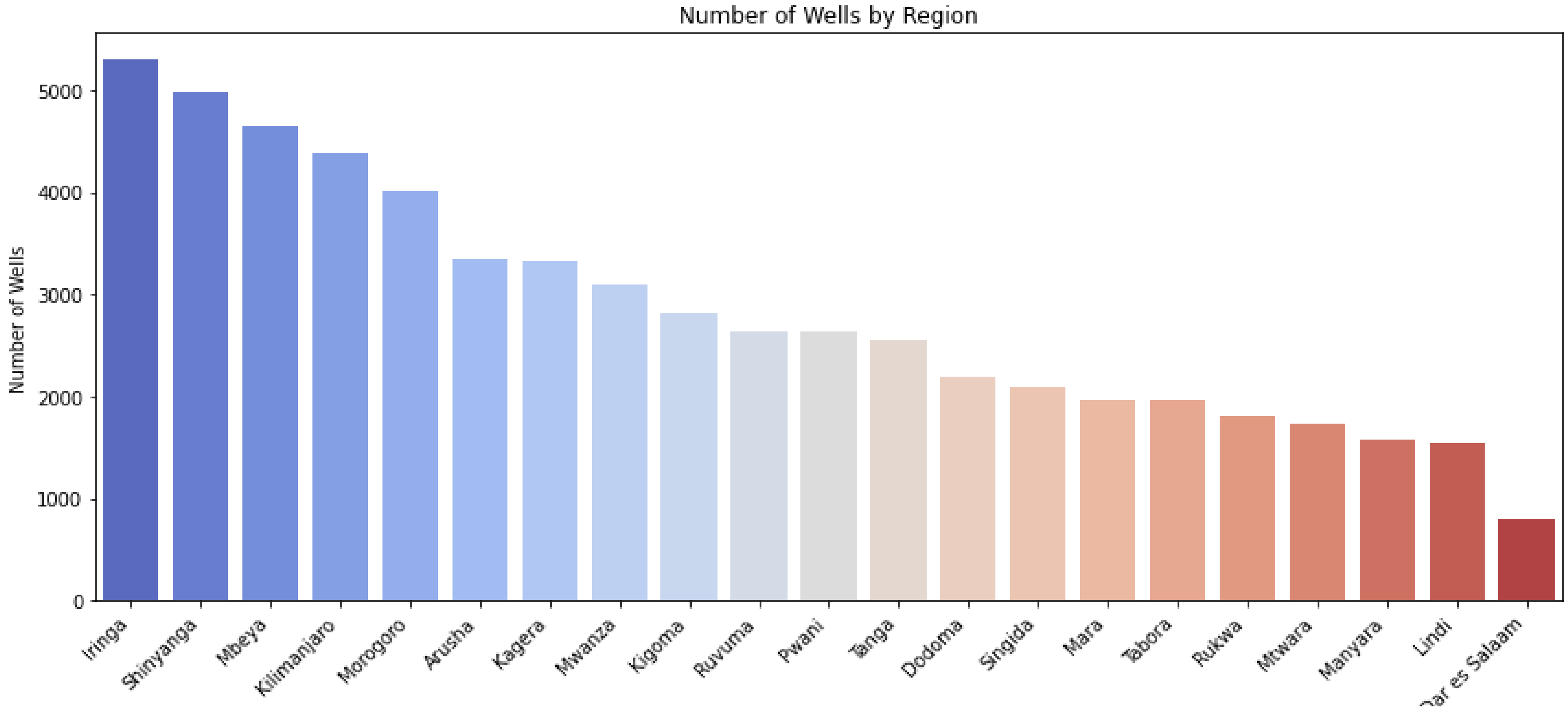
- • Data Source: Water well data from Tanzania (31,000+ records)
- • Key Features: Pump type, installation year, construction method, geolocation, etc.
- • Target Variable: Well status (functional, needs repair, non-functional)

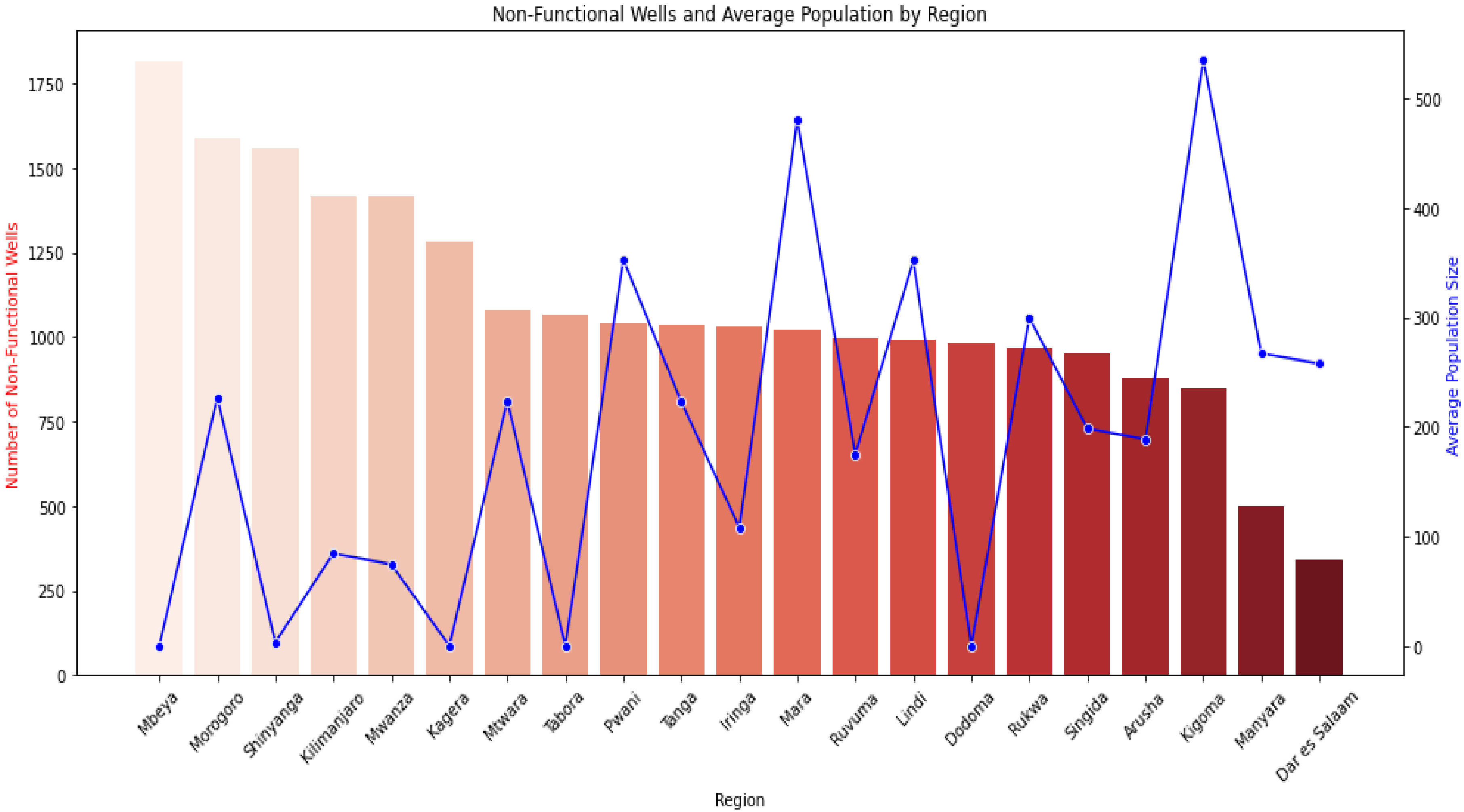
# Approach and Methodology

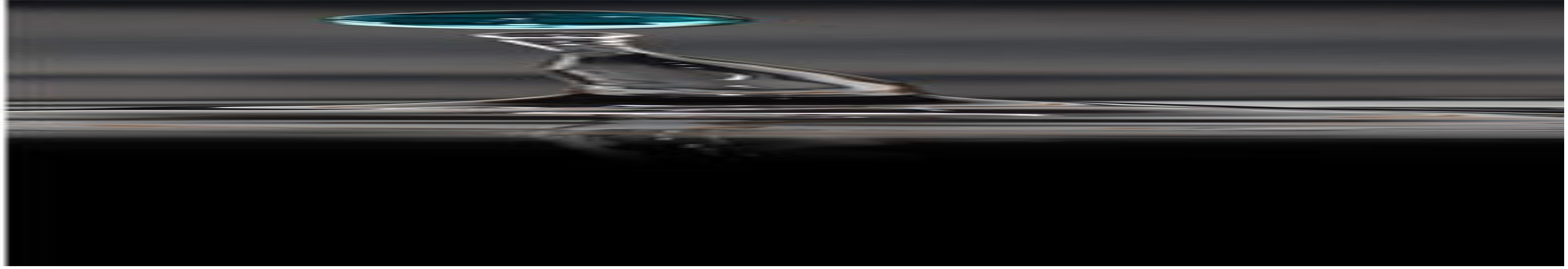
- Data Preprocessing: Cleaning, handling missing values, encoding categories

- Engineering: Converted ternary labels into binary for better clarity

- Modeling Techniques: Logistic regression, Decision Trees and Random Forest with hyperparameter tuning


Number of Wells by Region
Number of Wells
0
1000
2000
3000
4000
5000
Iringa
Shinyanga
Mbeya
Kilimanjaro
Morogoro
Arusha
Kagera
Mwanza
Kigoma
Ruvuma
Pwani
Tanga
Dodoma
Singida
Mara
Tabora
Rukwa
Mtwara
Manyara
Lindi

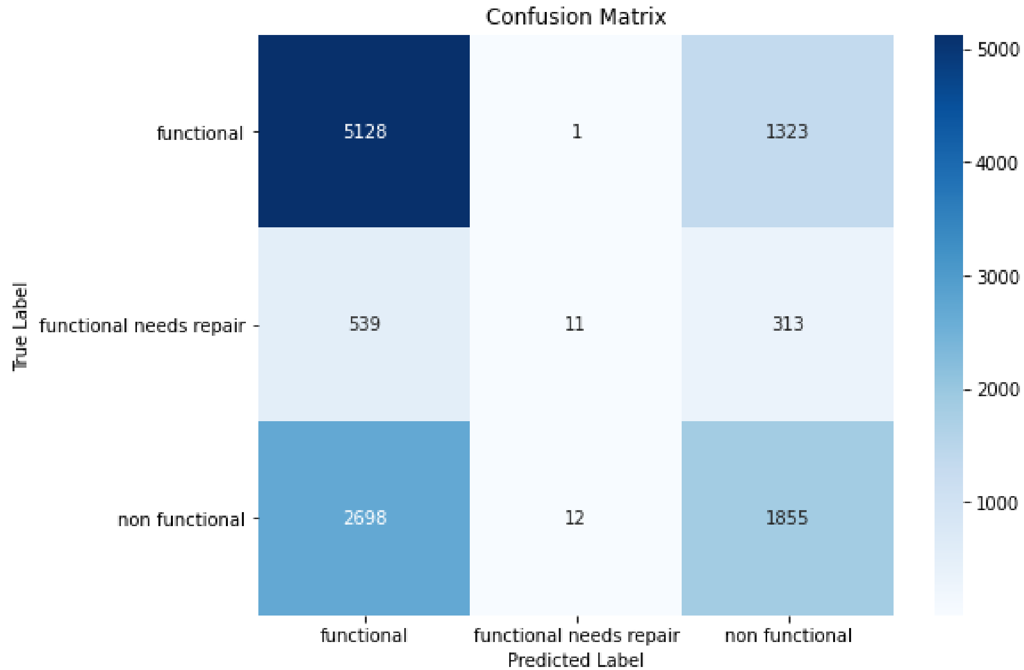

Non-Functional Wells and Average Population by Region
Number of Non-Functional Wells
0
250
500
750
1000
1250
1500
1750
Average Population Size
0
100
200
300
400
500
Mbeya
Morogoro
Shinyanga
Kilimanjaro
Mwanza
Kagera
Mtwara
Tabora
Pwani
Tanga
Iringa
Mara
Ruvuma
Lindi
Dodoma
Rukwa
Singida
Arusha
Kigoma
Manyara
Dar es Salaam
Region



- At least 3 models were built with varying complexity so as to improve the classification metrics
  Methods include:

1. Logistic regression
2. Decision Trees
3. Random Forest (with hyper parameter Tuning)

# Basic Regression Model




**Class 0 is mostly predicted correctly (5128), but misclassified as Class 2 (1323 times).**

**Class 1 is poorly predicted — only 11 correct predictions out of 863 (11+539+313), which is ~1.3% accuracy for Class 1.**

**Class 2 is also often misclassified as Class 0 (2698 times), despite 1855 correct predictions**

# Decision tree Model

- As per the basic model we learned the limitations of tertiary classification hence switched to binary and also induced parameters for better performance .

Class 1 (functional) is predicted slightly better than class 0 — higher recall and F1.It achieved moderate accuracy (67%), a good sign after class balancing and tuning.

# Random forest model

- Best Parameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}

Class 0 — Functional
Very strong recall (0.86): most functional pumps are correctly detected.

High precision (0.80): few false positives.

F1 = 0.83 → overall, your model handles Class 0 well. Class 1 — Non-

Functional
Recall (0.66) is lower: about 34% of non-functional pumps are missed.

Precision (0.75) is decent, so most predicted non-functionals are indeed correct.

- F1 = 0.70 → reasonable performance, but room to improve recall.

# Key Findings

**Model Performance**

- The **Random Forest classifier** achieved **78% accuracy** on the test data.

2. **High-Impact Features**

- The top predictors of a well's condition included:
  - **Pump Type**: Certain pump types (e.g., handpumps) were more likely to fail.
  - **Installation Year**: Older wells showed higher failure rates.
  - **Management Entity**: Wells managed by **private individuals or informal groups** had higher failure rates compared to those managed by government or NGOs.
- These findings support targeted investment in **infrastructure quality** and **management practices**.

3. **Class Imbalance Insights**

- Functional wells outnumber non-functional ones, which can **bias the model**.
- This was addressed via **resampling techniques** (e.g., down sampling/up sampling) and **class weights**, ensuring the model does not overlook at-risk wells.

# Recommendations

**Prioritize High-Risk Regions**

- Focus repair efforts in regions with high predicted failure (e.g., Shinyanga, Singida).
- Align maintenance with areas of high population and low well reliability.

**Target Risky Well Characteristics**

- Replace outdated pump types & tech (e.g., pre-2000 installations).

**Enable Proactive Monitoring**

- Use the model to flag at-risk wells.

**Inform Future Projects**

- Use failure patterns to guide new well placements.