PHASE 3 - PROJECT

OVERVIEW

Predict water pump functionality across Tanzania to transform reactive maintenance into proactive interventions. Using Taarifa and Ministry of Water data, we classify pumps into three critical states: functional, needs repair, or non-functional. This addresses preventable water scarcity affecting millions.

Approach & Impact:

After rigorous feature selection and SMOTE balancing, we developed a tuned random forest classifier model achieving 76% accuracy with 80% recall for non-functional pumps. The solution enables:

- Targeted resource allocation
- Preventive maintenance prioritization

BUSINESS UNDERSTANDING

Tanzania's water crisis is exacerbated by unexpected pump failures, disrupting community health and development. Our predictive model identifies pump status (functional/needs repair/non-functional) to transform reactive maintenance into proactive solutions. This empowers:

- *Water Authorities* to prioritize high-risk pumps
- *NGOs* to prevent failures in vulnerable regions
- *Maintenance Teams* with actionable alerts

DATA UNDERSTANDING

We analyzed Tanzania's complete water pump ecosystem using **59,400 records** from the Ministry of Water and Taarifa's monitoring platform. The data reveals two main dimensions of pump operations:

Pump Environment & Usage

Geographic placement (GPS coordinates, regional basins)

Local demand indicators (surrounding population density)

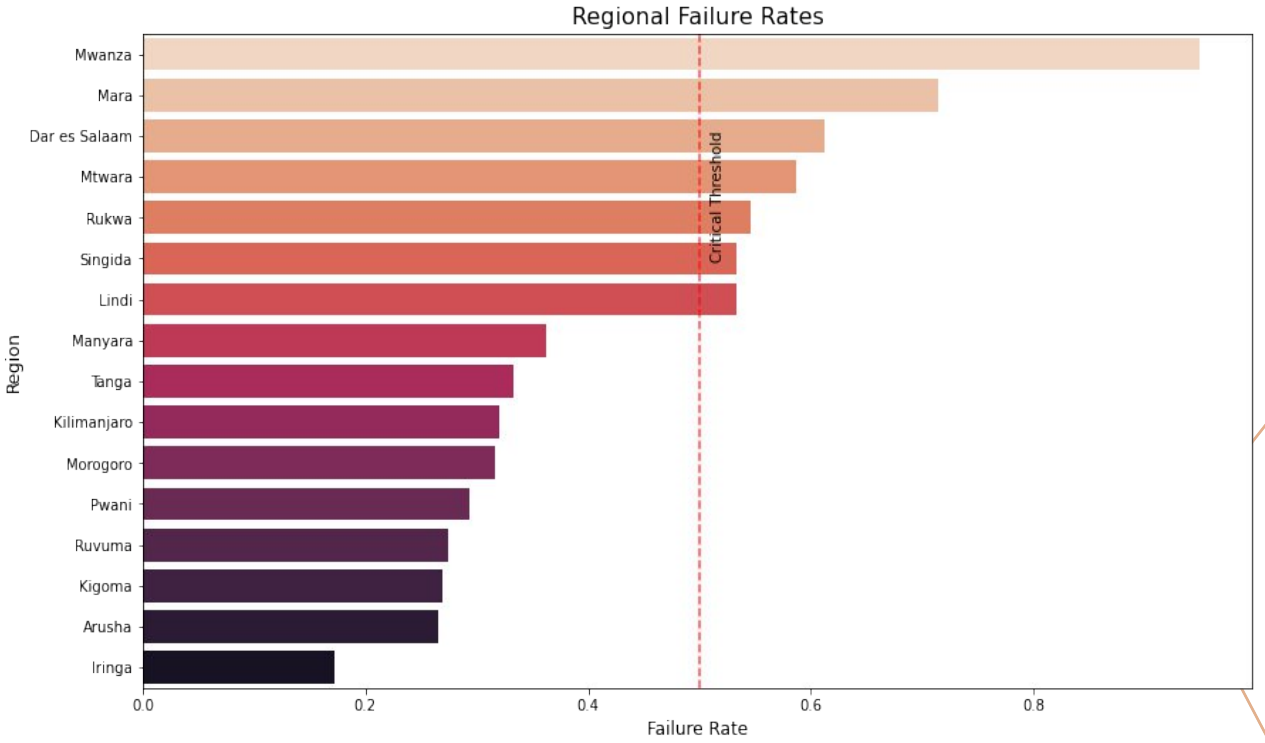
Water source characteristics (springs, boreholes, or lakes)

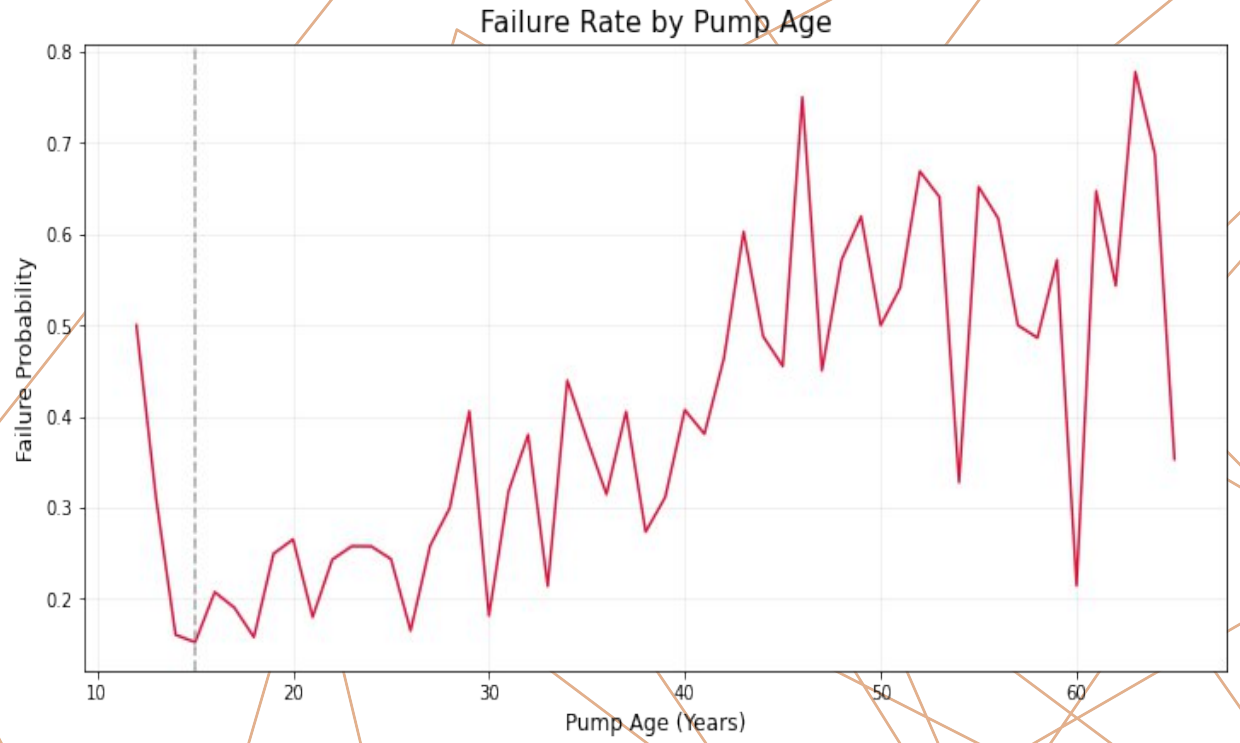
Pump Information Category Factors

Pump type and mechanical specifications (hand pumps vs. solar-powered)

Installation year (an increase in pump rate increases failure rates)

DATA UNDERSTANDING





From the graphs above one can see:

- The region vs pump failure rates.
- As the age of the pump increases so does the failure rate.

One can conclude that both the geography and age of pumps plays a factor on pump failure rates.

MODEL AND EVALUATION

Approach

Tested three predictive models to identify failing water pumps:

1. **Logistic Regression**: Baseline accuracy (61%)
2. **Decision Tree**: Improved detection (71% accuracy)
3. **Random Forest**: Optimal performance (76% accuracy)

Why Random Forest Wins

- Best at capturing complex patterns in pump data (has the highest accuracy)
- 80% recall for critical "non functional" pumps (meaning 4/5 times we correctly identify a non-functional pump)
- A high recall for critical “non-functional” pumps means that we can more accurately allocate manpower and ensure the water supply of high-risk areas.

RECOMMENDATIONS AND NEXT STEPS

1. **Model Deployment**

   - Implement **Random Forest** as the core prediction engine

   - Prioritize pumps flagged as "non functional" (80% recall)

   - Generate weekly repair priority lists for maintenance teams

2. **Resource Optimization**

   - Allocate 65% of maintenance budget to high-risk regions:

   - Pumps > 15 years old

   - Shift from reactive to preventive maintenance schedule

3. **Monitoring & Improvement**

   - Integrate with Taarifa's mobile platform for real-time reporting

   - Retrain model quarterly using new repair data

   - Track "water access days" per community as success metric