Tanzanian Water Wells Prediction

DSF-FT2 Phase 3 Project

Authored by Patricia Ngugi

---

1. Business Understanding

1.1 Overview

Water, covering about 70% of the Earth's surface, is crucial for all living creatures. However, 96.5% of this water is oceanic, leaving a limited supply of freshwater. Tanzania, with a population exceeding 57 million, faces significant challenges in meeting its demand for safe drinking water due to limited extraction resources. Although water pumps are available, many are non-functional or in need of repair. The World Health Organization (WHO) reports that three out of ten people in Tanzania still lack access to clean water, highlighting a critical need for investment in water infrastructure. Water scarcity contributes to serious health issues, high infant mortality, poor education, economic struggles, and unproductive agriculture.

1.2 Problem Statement

Tanzania struggles to meet its potable water needs. Despite existing water wells, many are either non-functional or require repair. WHO aims to address this by drilling new wells and improving existing ones. Our role as data scientists is to identify patterns in non-functional wells to guide the construction of new wells and predict which existing wells need intervention. This will ensure that Tanzanians have access to clean, potable water.

1.3 Project Justification

Climate change, increasing water scarcity, and population growth present significant challenges to water supply systems. WHO's mission includes providing quality drinking water to water-stressed countries like Tanzania. This project will help in building new wells, ensuring the sustainability of these wells, and maintaining existing ones.

1.4 Specific Objectives

- Identify trends and patterns between functional and non-functional wells.
- Use analysis to predict the functionality of wells based on available data.

1.5 Research Questions

- How can Machine Learning and classification methods predict the functionality of wells in Tanzania?
- What features are consistent across functional and non-functional wells?
- How can we improve the construction and maintenance of water wells?

1.6 Success Criteria

1.6.1 Business Success Criteria

- Ensure new wells provide good quality water.

- Accurately identify the functionality and viability of wells.
- Predict which wells are likely to dry up and guide the construction of new wells based on area-specific data.
- Determine which wells need repair or better management and ensure new wells are durable.

1.6.2 Project Success Criteria

Develop a model with at least 75% accuracy in predicting the quality status of wells in Tanzania.

1.7 Project Plan

- Utilize the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.
- Maintain a GitHub repository for project code and documentation.
- Create presentation slides summarizing findings and recommendations.
- Use a Trello Board for project collaboration and tracking.

1.8 Scope of the Study

This study will leverage machine learning models and data from an open-source platform to analyze factors affecting the condition of water wells in Tanzania.

---

2. Data Understanding

2.1 Overview

The dataset for this project contains information about existing water wells in Tanzania, sourced from the DrivenData competition.

2.2 Data Description

The dataset includes 59,400 records and 40 columns, categorized into:

2.2.1 Location-Based Data

- **Numerical:** longitude, latitude, gps_height
- **Categorical:** region, region_code, district_code, lga, ward, subvillage, basin

2.2.2 Time-Based Data

- **Numerical:** construction_year
- **Categorical:** date_recorded

2.2.3 Monetary-Related Data

- **Categorical:** payment, payment_type

2.2.4 Technical Data

- **Numerical:** amount_tsh

- **Categorical:** extraction_type, extraction_type_class, extraction_type_group, water_quality, quality_group, quantity, quantity_group, source, source_type, source_class, waterpoint_type, waterpoint_type_group, status_group

2.2.5 Non-Technical Data

- **Numerical:** population
- **Categorical:** installer, funder, wpt_name, public_meeting, scheme_management, scheme_name, permit, management, management_group

2.2.6 Miscellaneous Data

- **Numerical:** num_private
- **Categorical:** recorded_by

2.3 Data Quality Verification

The dataset required detailed cleaning for completeness and consistency. Missing values were found in several columns, which were addressed by removing or imputing data. Duplicate records were retained, and outliers were analyzed but not removed.

---

3. Data Preparation

3.1 Data Selection

Columns with redundant information (e.g., payment and payment_type) were removed. Data on location and basin proximity were retained for analysis.

3.2 Data Cleaning

- Addressed missing values in columns like funder, installer, and permit.
- Ensured data validity by checking for duplicates and outliers.
- Standardized column names and corrected typographical errors in installer and funder names.

3.3 Exploratory Data Analysis (EDA)

- **Univariate Analysis:** Assessed the distribution of categorical and numerical features.
- **Bivariate Analysis:** Investigated relationships between features and the target variable (well functionality).

3.4 Data Quality Report

Post-cleaning, the data exhibited good validity, accuracy, completeness, consistency, and uniformity.

---

4. Modeling

The dataset's class imbalance was noted, with 54.1% functional wells, 38.8% non-functional wells, and 7% wells needing repair. Various models were tested:

- **Logistic Regression:** Achieved 67% accuracy. This baseline model provided a starting point for comparison.
- **Decision Tree:** Achieved 70% accuracy, indicating an improvement over the baseline.
- **Random Forest:** Also achieved 70% accuracy, similar to the decision tree model.
- **K-Nearest Neighbors (KNN):** Encountered errors and could not be executed.

---

5. Evaluation

Models were evaluated using pipelines for scaling and cross-validation to avoid overfitting. Despite not reaching the desired 75% accuracy, the achieved 70% accuracy is a solid foundation and falls within an acceptable range.

---

6. Conclusion and Recommendations

Conclusion

Both the Random Forest and Decision Tree models achieved 70% accuracy. While further feature engineering could improve this, the current models meet our objectives for predicting well functionality.

Recommendations

- **Basin Area:** Consider Lake Rukwa for future well construction due to its high number of non-functional wells.
- **Region:** Focus on Dodoma, which has a higher number of non-functional wells.
- **Permits:** Ensure wells are permitted as they tend to be more functional.
- **Payment Schemes:** Implement affordable payment options to reduce misuse and improve functionality.
- **Proximity to Lake Victoria:** Wells near Lake Victoria have shown to be more durable.