شعيب.

WATER WELLS

CLASSIFIER

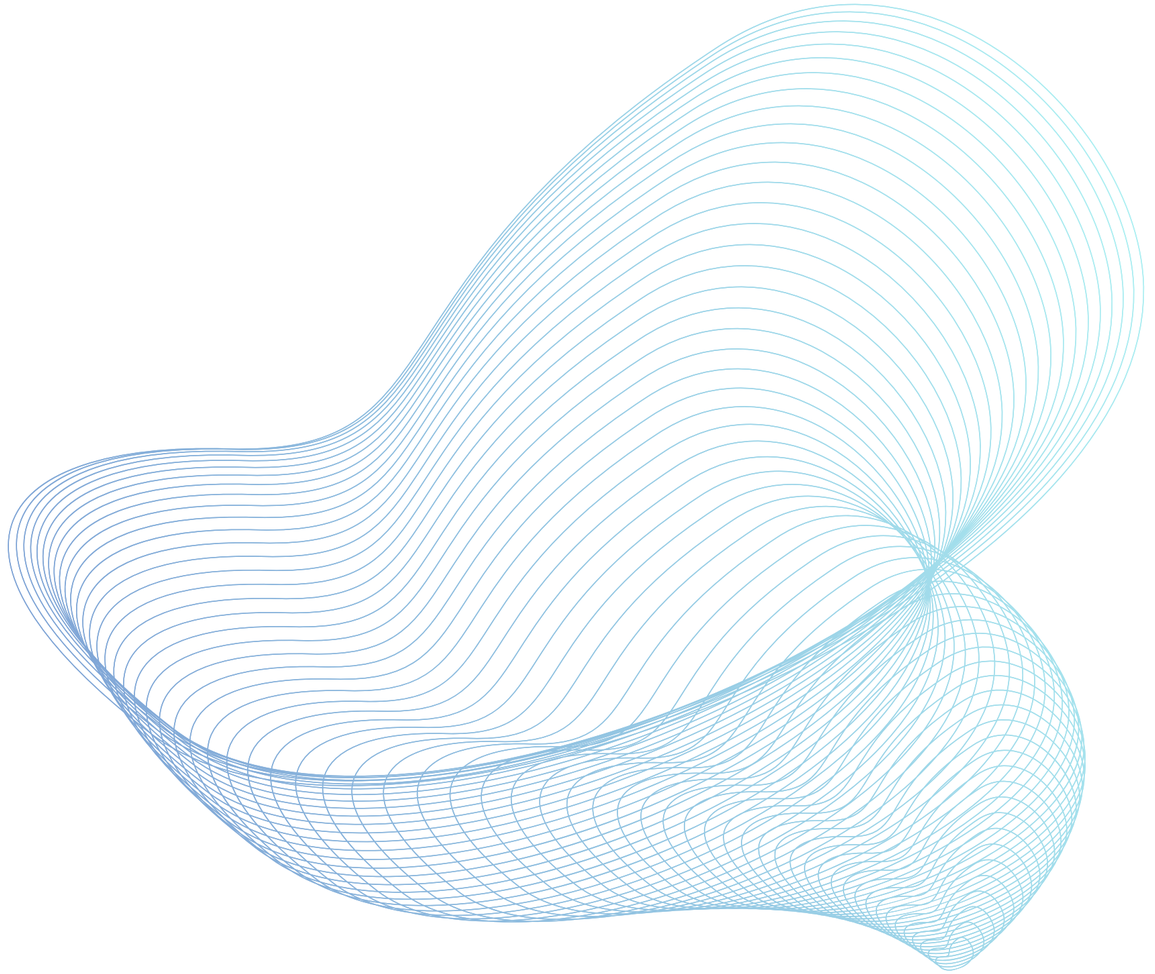


TABLE OF CONTENT

BUSINESS OVERVIEW

Tanzania, with over 60 million people, struggles to provide clean water despite efforts. Around 61% have access, leaving many vulnerable to diseases due to contaminated sources. Collaborative projects aim to improve infrastructure, but challenges like non-functional wells persist, impacting public health, especially among vulnerable groups like children and the elderly. Waterborne diseases continue due to contamination, causing preventable deaths and ongoing health crises.

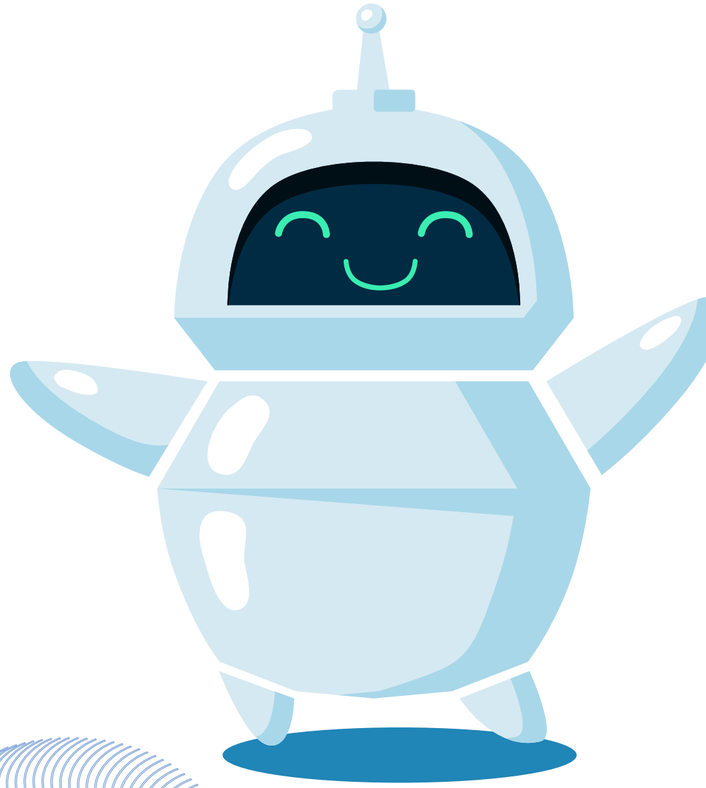
PROBLEM STATEMENT

The Ministry of Water and WHO in Tanzania are partnering to improve clean water access. Despite past efforts, 31,000 annual deaths are linked to inadequate water services, with over 10% preventable.

An initiative is underway to assess water wells and pumps. I lead the effort to predict pump functionality patterns. This aims to guide maintenance and resource allocation, enhancing water accessibility and reducing preventable deaths, aligning with Tanzania's development goals.

MAIN OBJECTIVE

Develop a robust machine learning model capable of forecasting water pump performance utilizing diverse attributes, including pump type, installation date and well conditions.



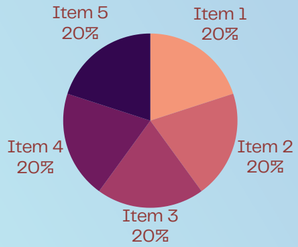
SPECIFIC OBJECTIVES

A) Determine key features affecting pump functionality via statistical analysis and feature importance methods.

B) Use diverse machine learning algorithms (e.g., Random Forest, Gradient Boosting, Logistic Regression) to construct a predictive model.

C) Create data-driven suggestions to enhance clean water source accessibility.

DATA UNDERSTANDING



1. **Scope of Data:**

- The dataset appears to encompass diverse information related to water supply and infrastructure in Tanzania.
- It likely includes details about geographical locations, water sources, population, infrastructure details, and possibly water quality indicators.

2. **Attributes:**

- The dataset contains multiple attributes such as 'id', 'amount_tsh', 'funder', 'gps_height', 'installer', 'longitude', 'latitude', 'population', 'construction_year', 'status_group', and many more.
- These attributes seem to cover various aspects, including financial details, geographical coordinates, demography, and project status.

3. **Target Variable:**

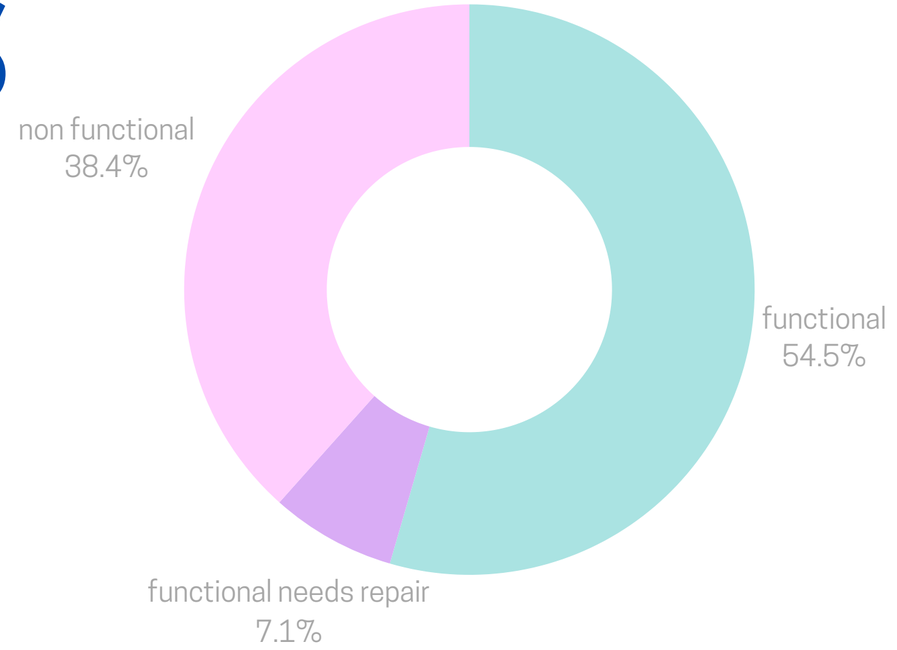
- 'status_group' might be a crucial target variable indicating the condition or status of water points or wells.
- Understanding this variable is critical to assess the functionality of water sources, which could be vital in addressing waterborne diseases.

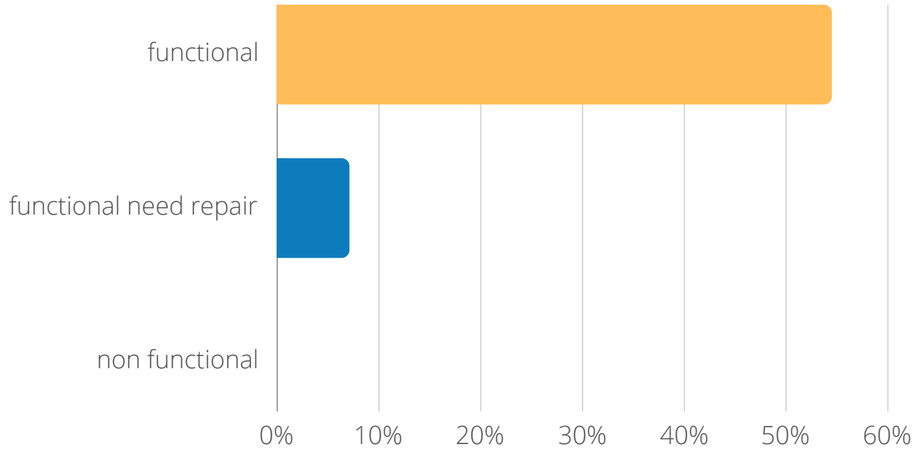
DATA ANALYSIS

Understanding the distribution helps in predicting or classifying water point statuses. Our machine learning aims to predict whether a water point is functional, needs repair, or is non-functional based on its attributes.

Target Variable: "status_group" Distribution

- Functional (54.5%): Water points in operational condition, providing clean water.
- Functional Needs Repair (7.1%): Operational but requiring maintenance for optimal functionality.
- Non-Functional (38.4%): Water points not operational, failing to provide clean water.





MODELS

Random Forest (hyperparameter tuning)

Yielded an accuracy of 0.679 with slightly enhanced precision, recall, and F1-score after tuning.

Random Forest (hyperparameter tuning)

Emerged as the top performer with an accuracy of 0.706, showcasing the best precision, recall, and F1-score among the models after tuning.

Logistic Regression

Achieved an accuracy of 0.655 with notable precision and recall for class 1 but struggled with classification for class 2.

Decision Tree

Demonstrated an accuracy of 0.71 with improved precision and recall for multiple classes compared to Logistic Regression.

KNN

Showed balanced performance with an accuracy of 0.672, maintaining moderate precision, recall, and F1-score.

CONCLUSION

The models were evaluated based on their performance metrics, including accuracy, precision, recall, and F1-score.

While all models showed varying levels of performance, Gradient Boosting stood out as the most effective model after tuning, showcasing superior predictive capability and overall better handling of the classification task compared to other algorithms evaluated. This suggests that for this particular dataset and task, Gradient Boosting might be the most suitable choice among the models considered.

RECCOMENDATIONS

1. **WHO-Government Collaboration**: Mobilize funds efficiently.

2. **DWE for Pump Installation**: Consider engaging DWE.

3. **Key Pump Indicators**: 'amount_tsh' & 'water quantity' crucial.

4. **Lake Victoria Issue**: WHO exploration needed.

5. **Water Access & Urbanization**: Majority away from wells; urbanization affects.

6. **Expanded WHO Initiative**: Include urban water access for SDG 2030.