Water Wells Functionality Prediction in Tanzania

Presentation By: Group 5

Group Members: Darvin (Team Leader); Mary; Shadrack; & Marvin

Period: May 2023

PROJECT OVERVIEW

The project aims to develop a model to classify the functionality status of water wells in Tanzania using data sourced by Taarifa and the Ministry of water.

BUSINESS PROBLEM

- Tanzania is a developing country which is in the midst of a water crisis & struggles to provide clean water to its population of over 57 million people despite the fact that the country has many established water points. It aims at resolving the water crisis in the country by maintaining and repairing the water pumps in time.

- In order for its Engineers to achieve the objectives faster, they need to be know in advance which water pumps are likely to fail and understand the causes of failure.

- The model created will enable the Tanzanian Ministry of Water to improve the maintenance operations its water pumps.

OBJECTIVES

To Develop a machine learning classifier model that can be used to predict the functionality of water wells in Tanzania

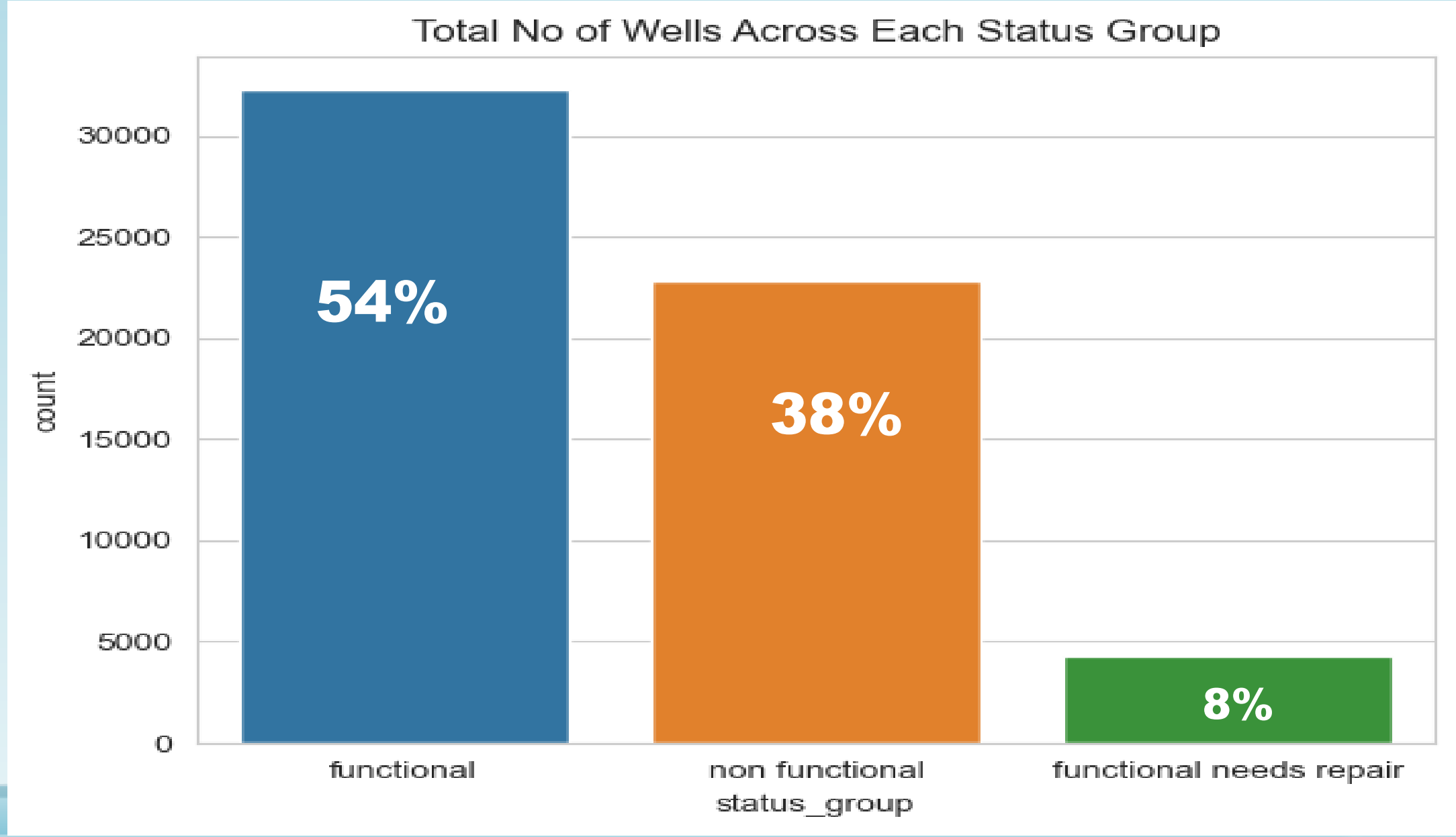
Wells Classification:

1. Functional (0)
2. Functional but needs repair (1)
3. Non functional (2)

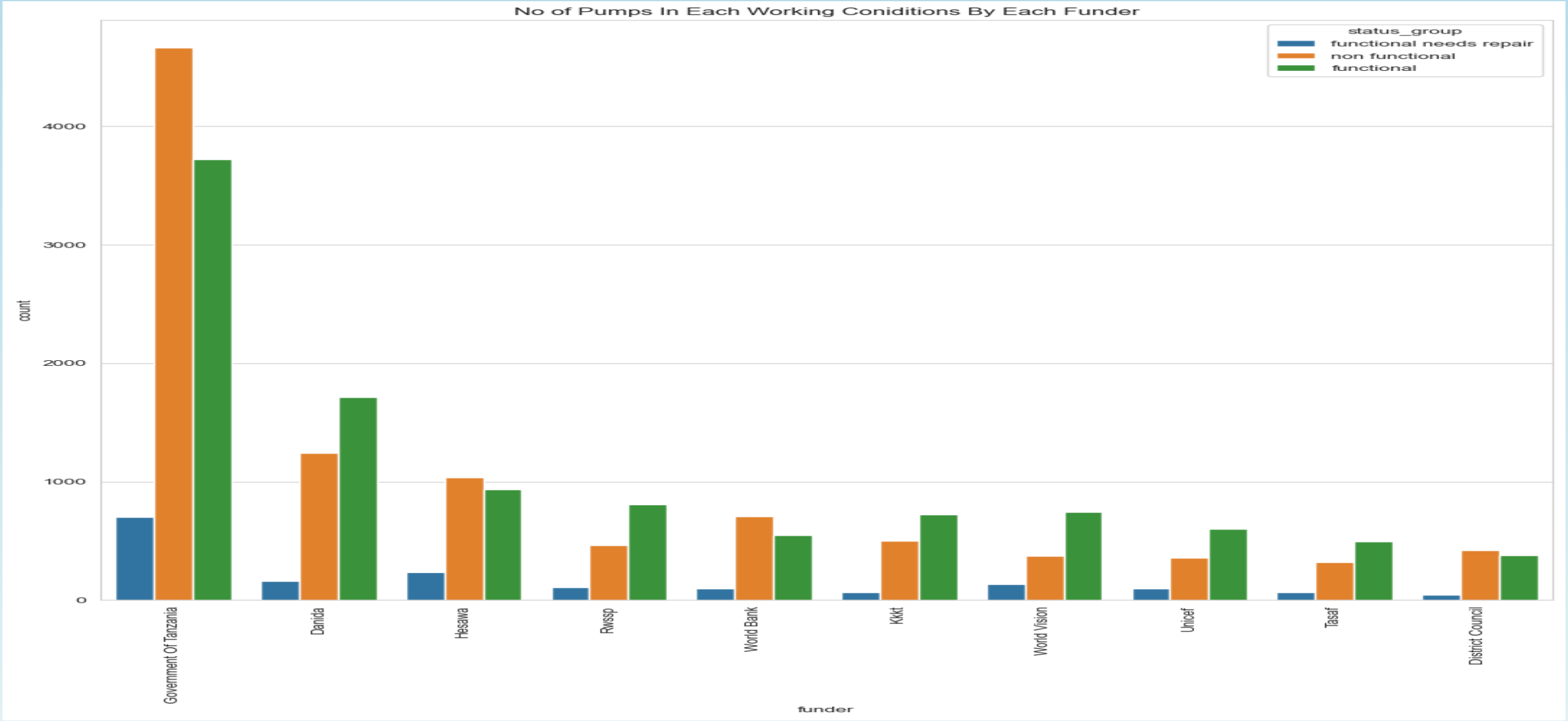
The expectation is that by predicting the functionality wells:

1. Ensure water wells are functioning properly
2. Improved access of water across Tanzania
3. Efficient Resource allocation by Ministry (Repairs & Servicing) – Improve Maintenance Operations

Existing Wells Distribution

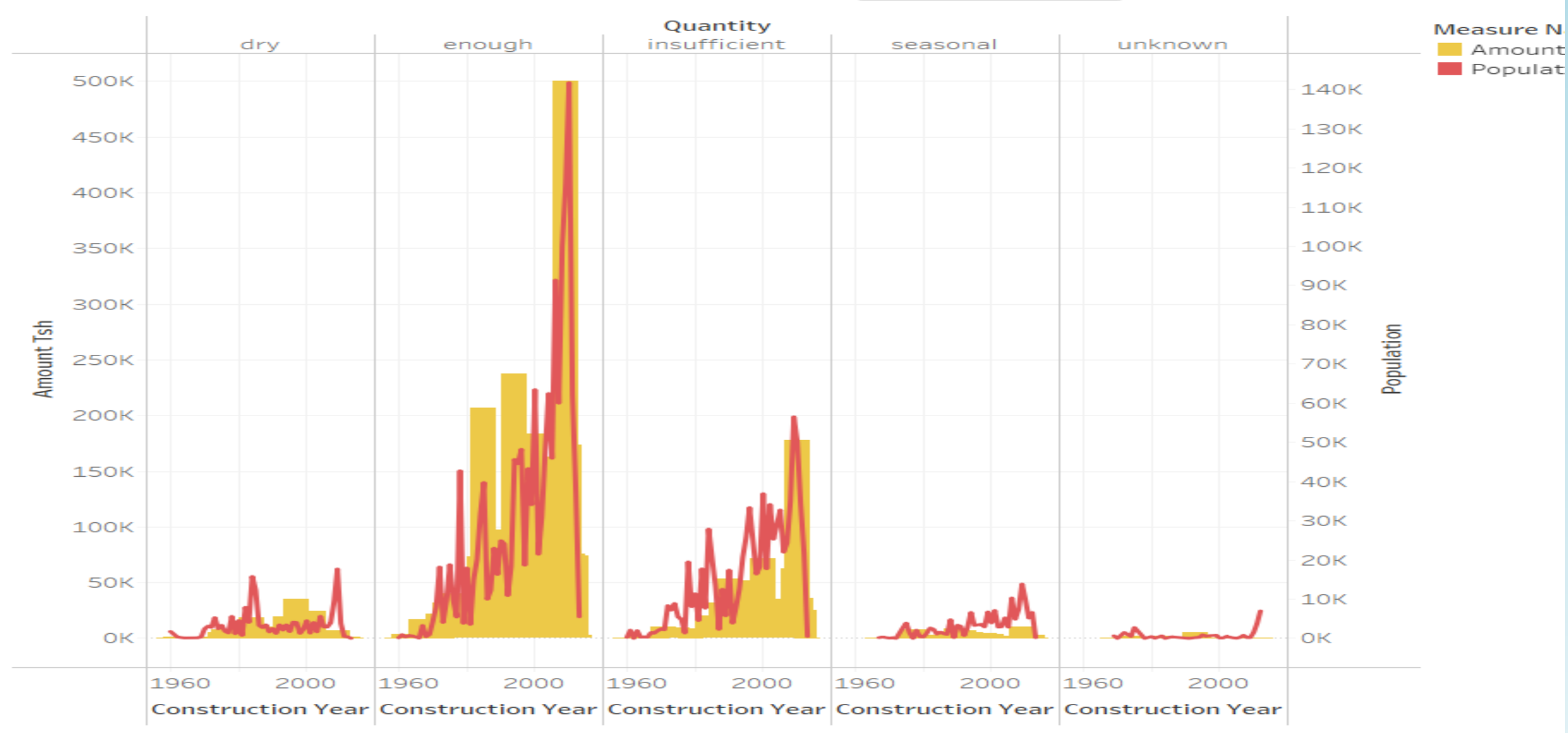


Status Group vs Funder



Major Funder Government of Tanzania

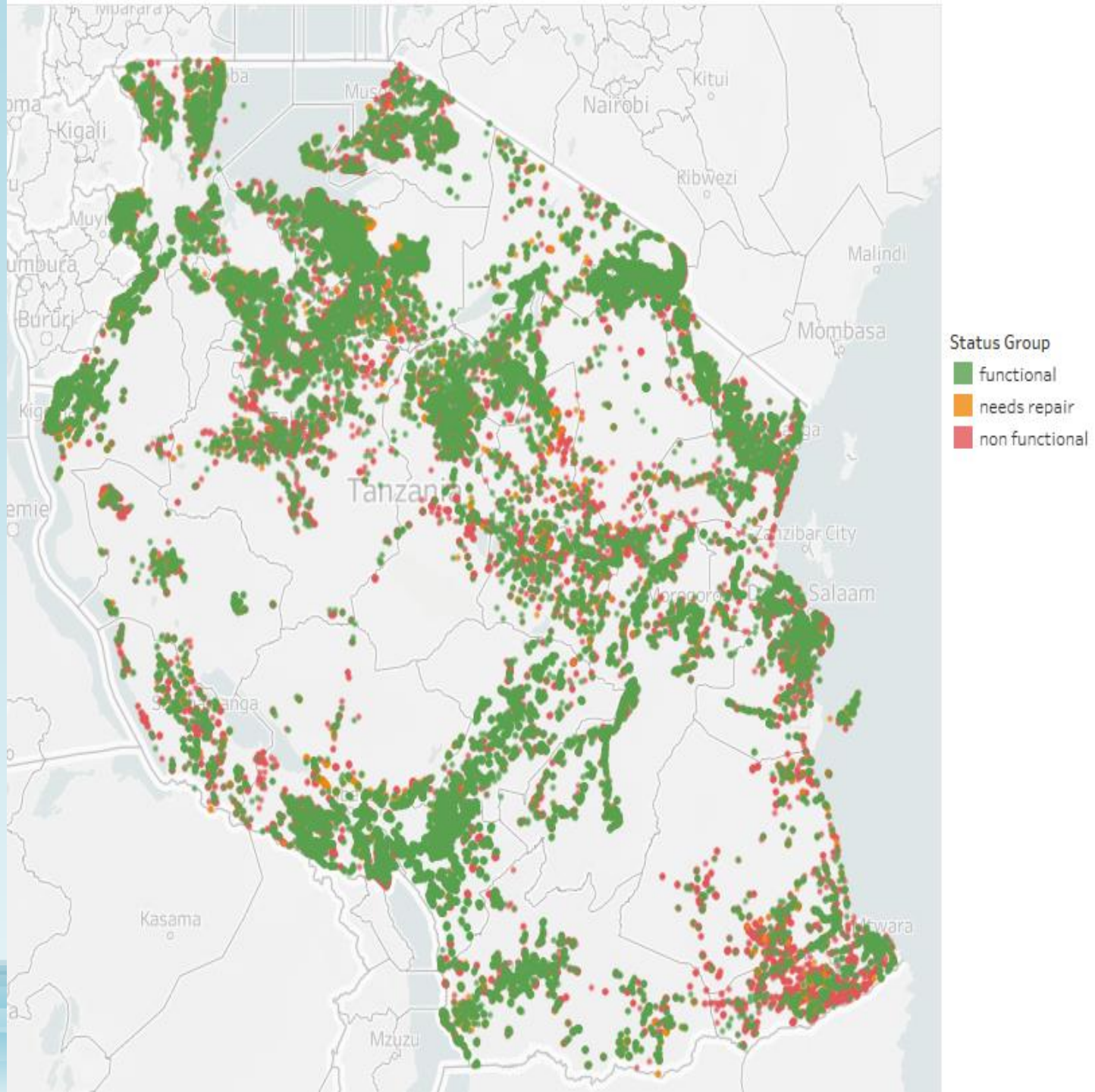
Population vs Water Quantity vs Construction Year



Population, Level of Investments & Water Quantity are Directly Proportional

Visualizing the Water Pumps

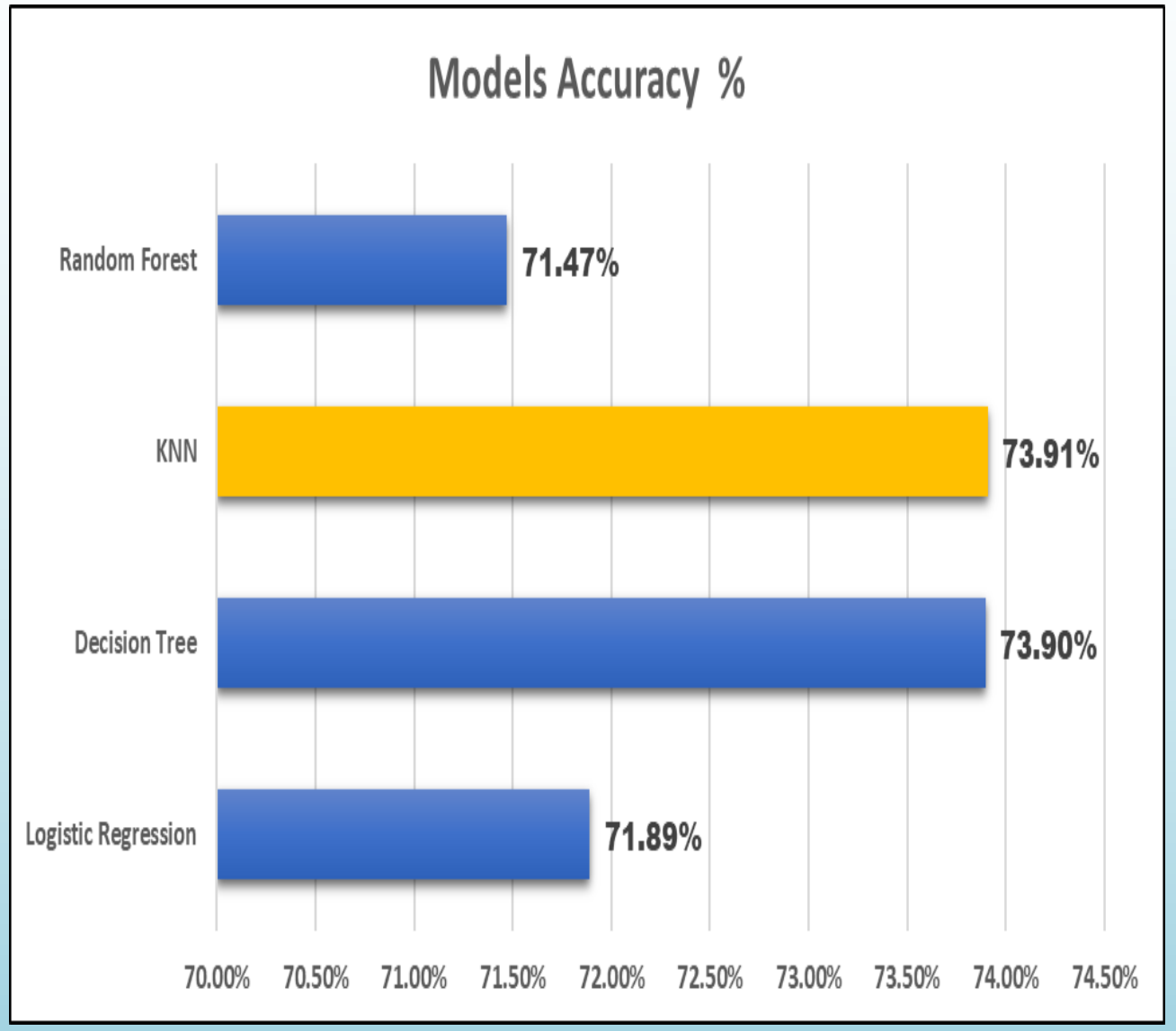
- Most Nonfunctional Wells spread across south & boarder point
- Middle Region No Wells – ASAL, Forest, Wildlife Reserve?
- Functional Wells mainly spread across Water Bodies/Basins



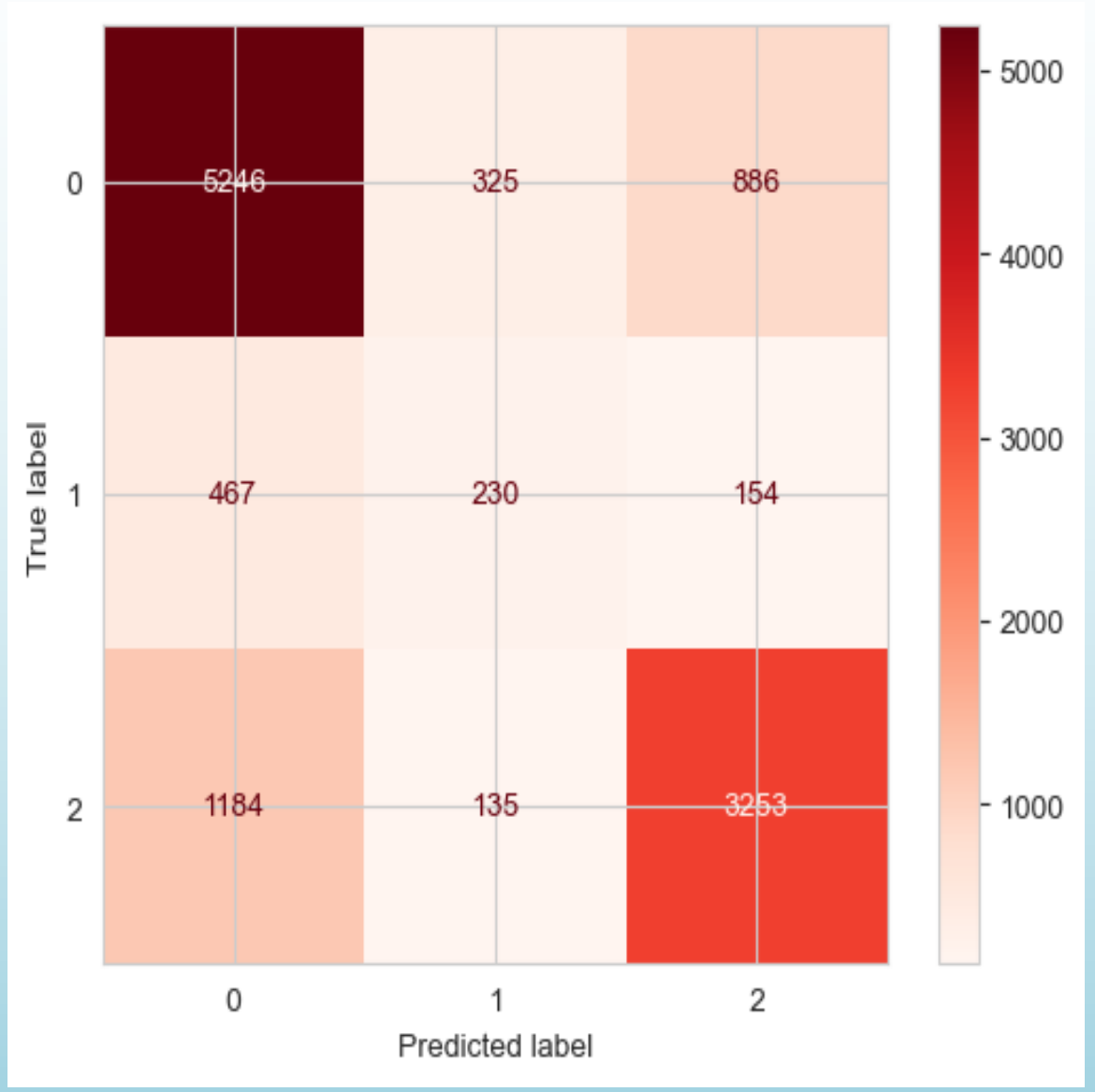
Features used in our Predictive Model

- **gps_height**: The altitude of the well location
- **installer**: The organization or entity that installed the water well
- **population**: The population around the well.
- **construction_year**: The year when the water well was constructed
- **extraction type:** The kind of extraction the waterpoint uses
- **water quality:** The quality of the water
- **quantity group:** The quantity of water
- **source type:** The source of the water
- **waterpoint type:** The kind of waterpoint
- **Funder:** Who funded the well
- **basin** - Geographic water basin

MODEL SELECTION – KNN (Based on Accuracy)

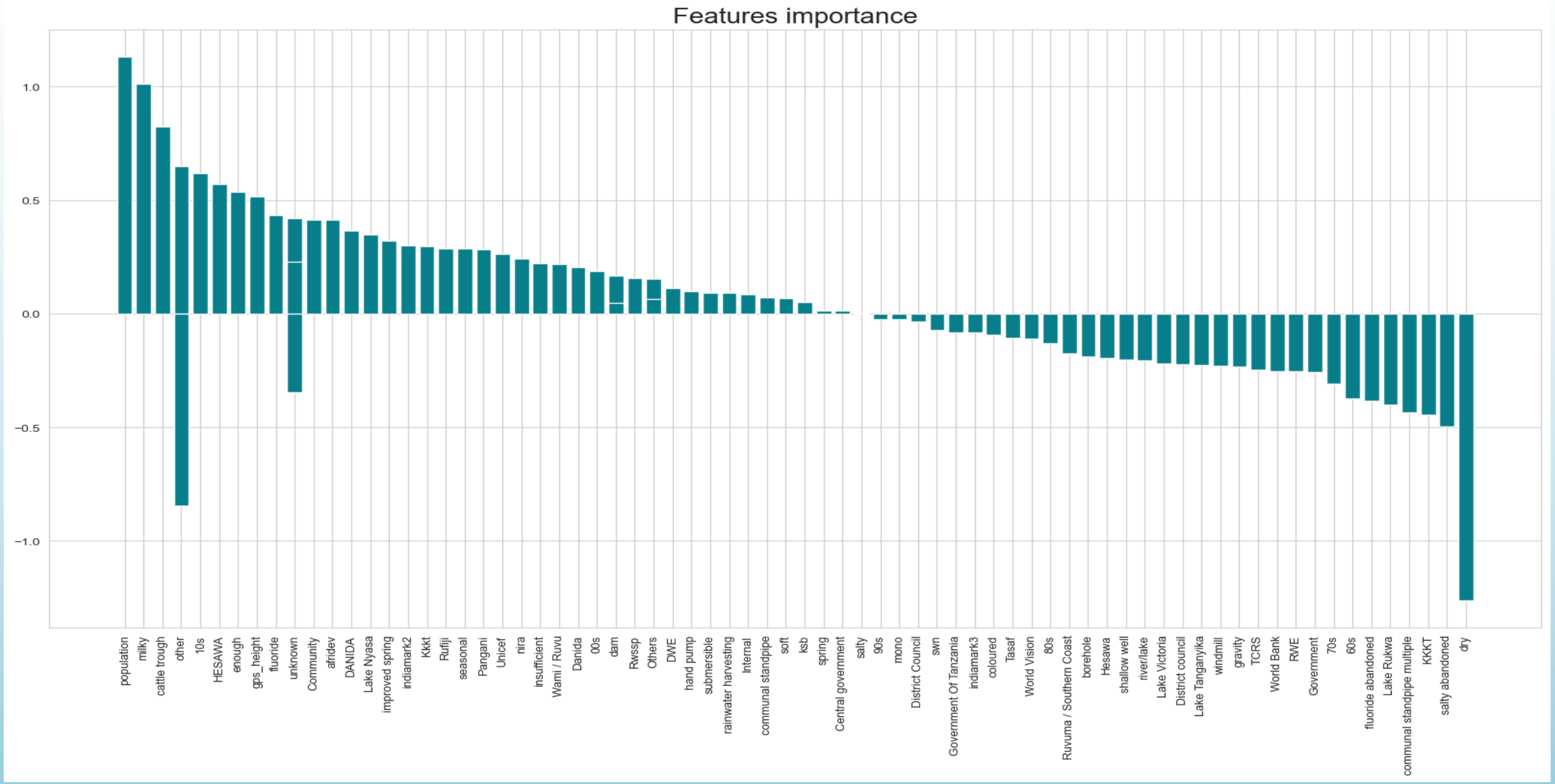


KNN Highest Accuracy Compared to Other Classifiers



Diagonal Elements Correctly Predicted Samples
8729 Samples Correctly Predicted (74%)

Key Estimators – Visualization (Features Importance Score)



Recommendations

- For large complex datasets like the Tanzania wells data, which had 59,400 observations, simple data analysis using Excel would be difficult, so Machine Learning is ideal because it is used for large complex datasets allowing the visualization of patterns and relationships that would ideally be missed by simple data analysis methods

- The Tanzania well dataset has both categorical and numerical datasets, so ML was able to provide algorithms to identify patterns and relationships in the data such as in places where wells in a sparse population, the wells was rarely used and needed repair. Such wells need to be identified if there is need for repair or maintenance

- For large dataset, data cleaning is key to ensure data quality prior to applying ML algorithms, in this data we dropped variables like latitude, longitude, district codes etc. which were not necessary for our model.

- The characteristic of the dataset is determined by the data type and goal of analysis, e.g. For the Tanzania wells data set we used Decision tress KNN was the best model it had 73.9% accuracy score meaning the model correctly predicted the classification levels

- The analysis showed that data collected should highlight more on non-functional pumps or those in need of repair so that they can plan resource allocation for repairs and maintenance

Areas of Improvement – To Increase Model Prediction Power

1. Further Tuning: More fine tuning to improve model accuracy.
2. Try More Classifiers & Consider using more complex ensemble models
3. Consider using more features provided in Original dataset


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
dank je
gracias
tapadh leat
bedankt
hvala
mauruuru
dziękuję
obrigado
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
arigaló
dakujem
terima kasih
grazie
merci
감사합니다
мерси