Faulty Water Pump Detection in Tanzania

Project Done By: Yvonne Kamari
October 23, 2023

Business Problem

Tanzania has over 60 million people and is struggling to provide them with clean and safe drinking water.

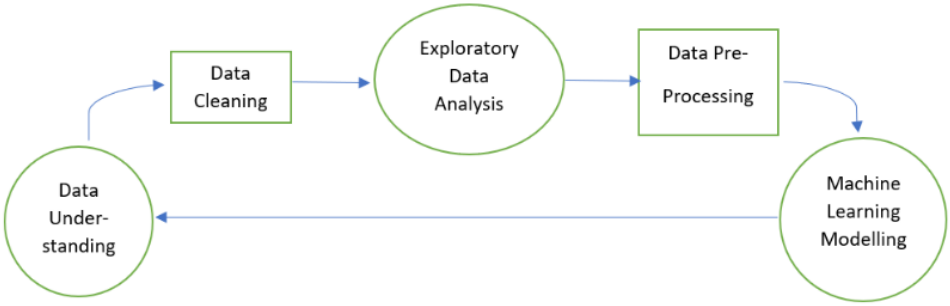
Waterborne diseases like cholera and dysentery are prevalent due to poor water quality and inadequate sanitation practices.

Project Objective

To develop a predictive machine learning model that can classify the operational status of water pumps across Tanzania.

The target audience is a non-governmental organization (NGO) focused on well maintenance and repair. The NGO will utilize the model's predictions to prioritize wells in need of attention.

Methodology



Data Understanding

The dataset is sourced from Taarifa and the Tanzanian Ministry of Water, aiming to facilitate the identification of waterpoints functionality.

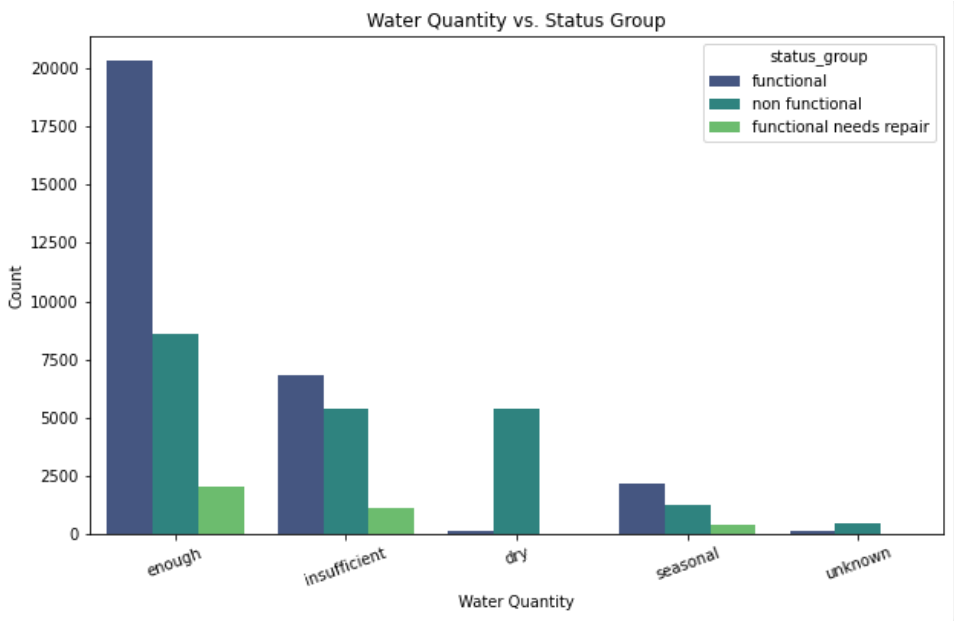
The dataset contains information on 59,400 waterpoints characterized by 39 features.

Data Cleaning

The dataset presented several issues, notably the presence of missing values in various features, with "scheme_name" having the highest count of missing data at 28,166 entries. I also cleaned and transformed features such as "funder", "installer", "public_meeting", "scheme_management", "permit", longitude, and latitude.

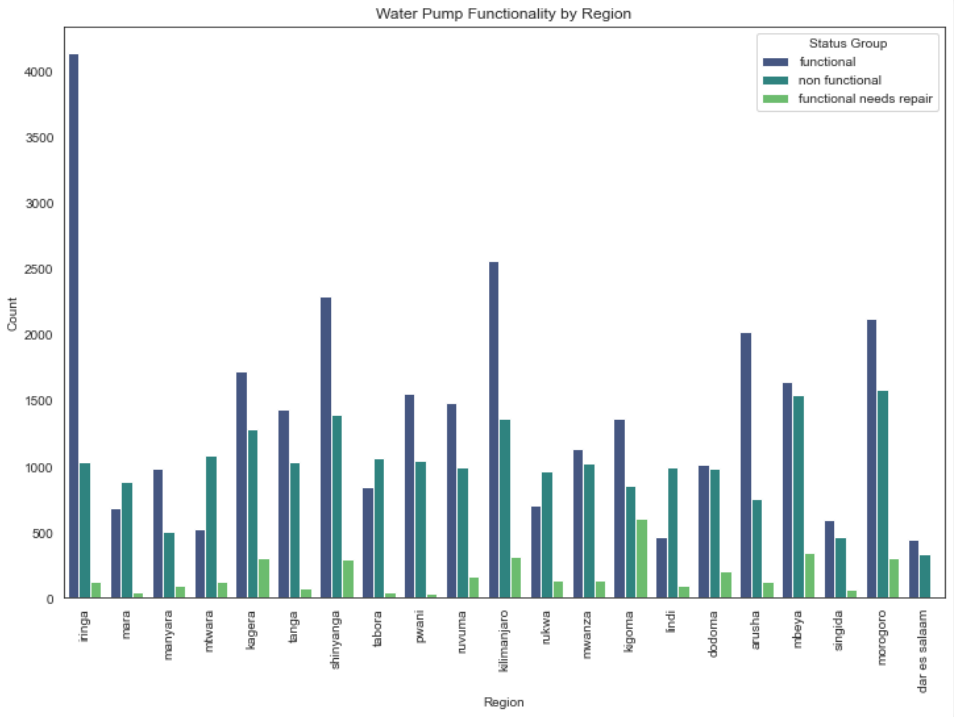
Exploratory Data Analysis

1. Relationship between Water Quantity and Water Well Functionality



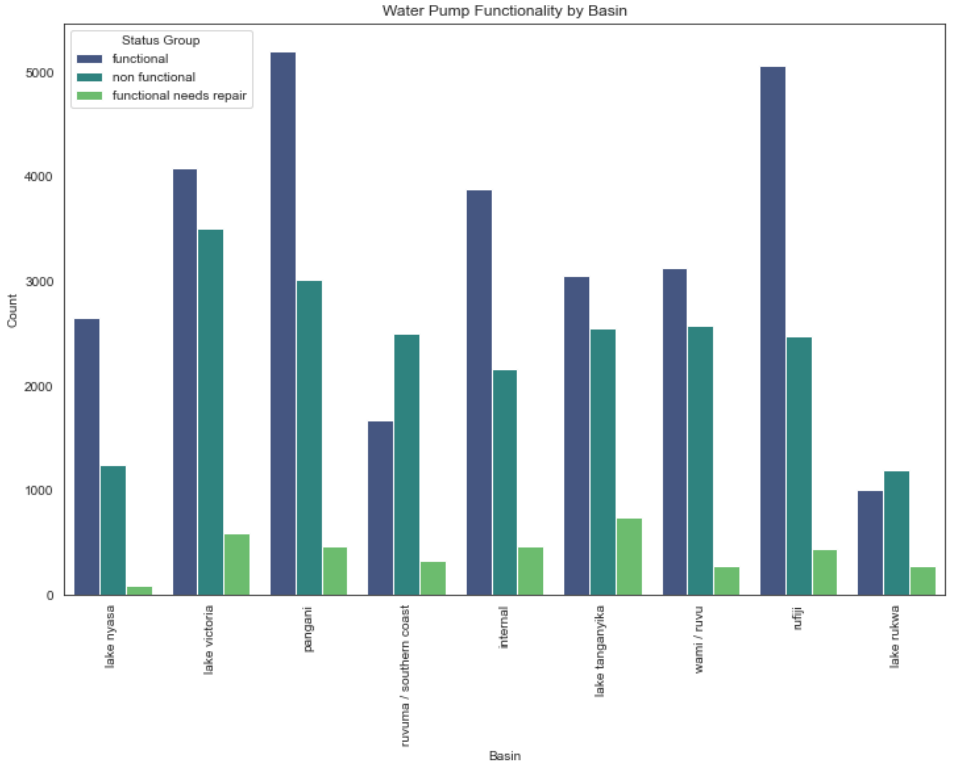
There is a clear positive correlation between water availability and water pump functionality.

2. Relationship between Regions and Water Well Functionality



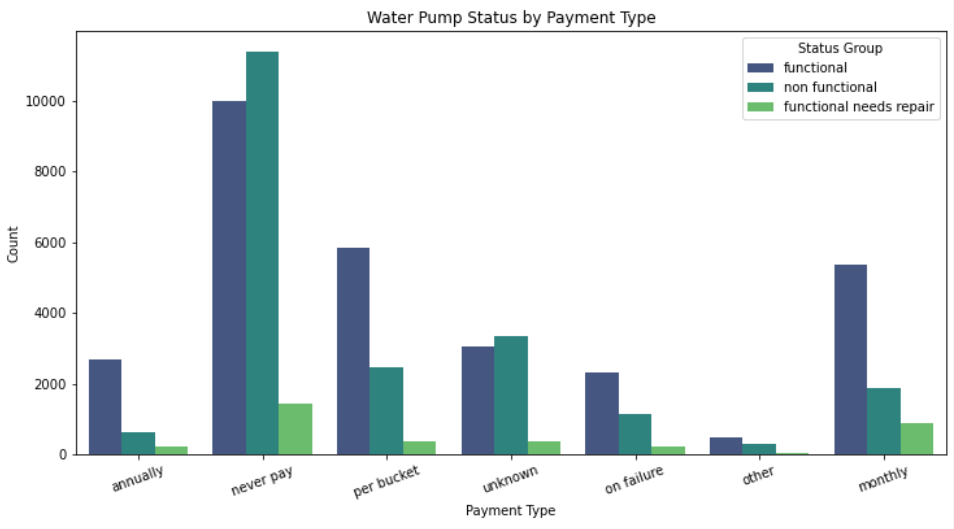
Iringa leads with a notable 4,138 functional pumps, followed by Kilimanjaro, Morogoro, and Arusha, which each have over 2,000 functional pumps.

3. Relationship between Water Basin and Water Well Functionality



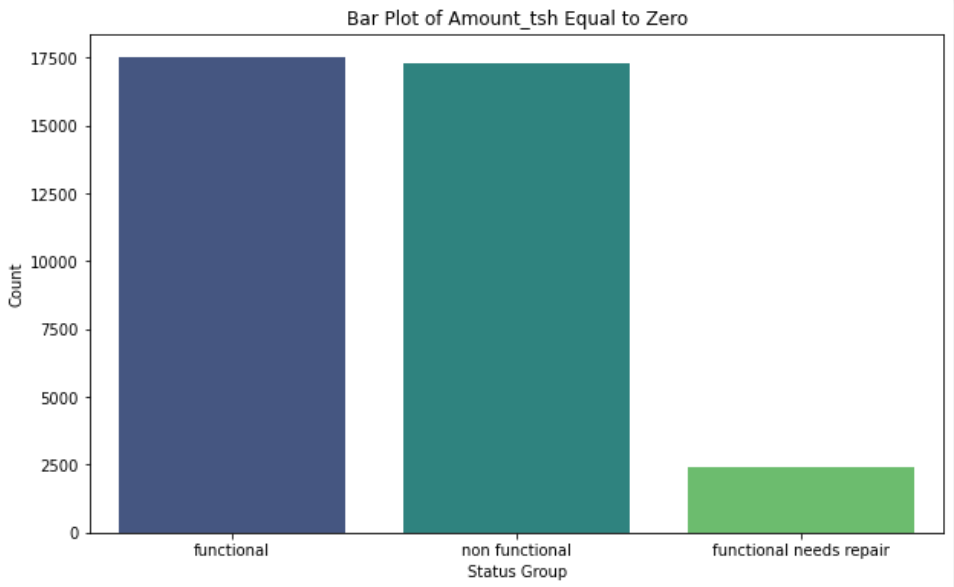
Some water basins in Tanzania emerge with a substantial number of functional water points, notably Pangani, Lake Victoria, and Rufiji. Lake Victoria stands out with a significant count of functional water points, suggesting a relatively robust water infrastructure, albeit with a notable need for repairs.

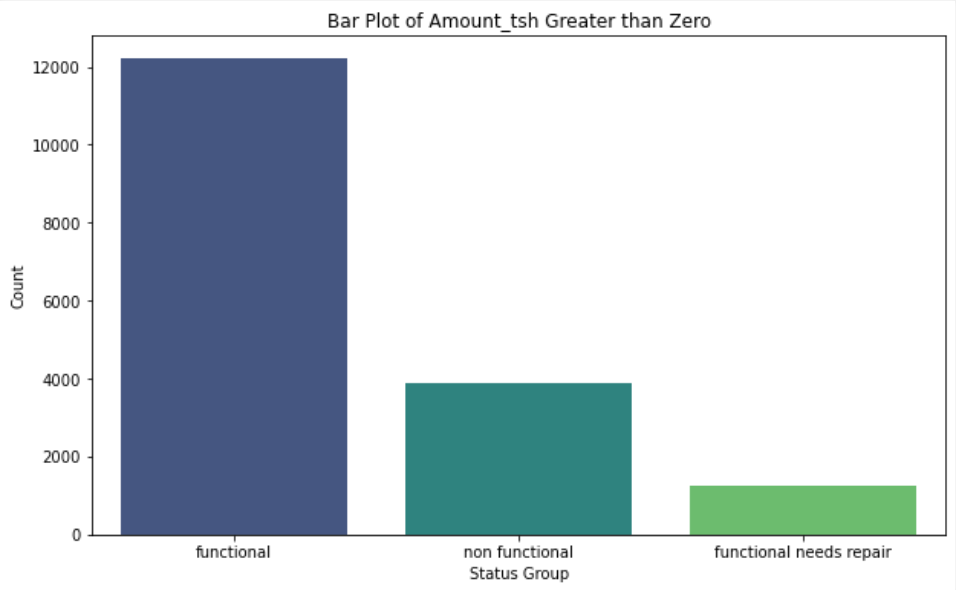
4. Relationship between Payment Status and Water Well Functionality



Water pumps marked as "Never pay" are widespread in Tanzania, but they exhibit a higher count of non-functional water points compared to the functional ones

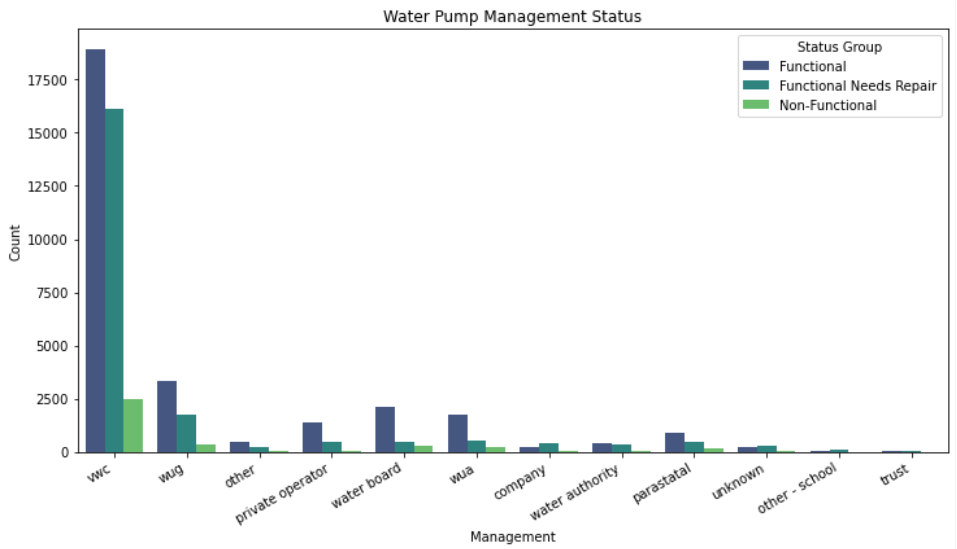
5. Relationship between Total Static Head and Water Well Functionality





The above bar plots reveal that when "amount_tsh" registers as zero, there is an almost equal likelihood of the pump being functional or non-functional. However, as the "amount_tsh" value increases, there is a corresponding rise in the likelihood of pump functionality.

6. Relationship between Management and Water Well Functionality



Water well management is primarily conducted by VWC, they oversee approximately 19,000 functional wells, around 16,000

in need of repair, and fewer than 2,500 that are non-functional.

MODELING

F1 score is used as the performance metric, benefiting the project by considering both precision and recall, thus balancing the model's ability to correctly classify pump functionality.

I tested the following models to identify the best classifier:

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **K Nearest Neighbor**
4. **Random Forest Classifier**
5. **eXtreme Gradient Boosting (XGBoost)**
6. **eXtreme Gradient Boosting (XGBoost) with SMOTE**

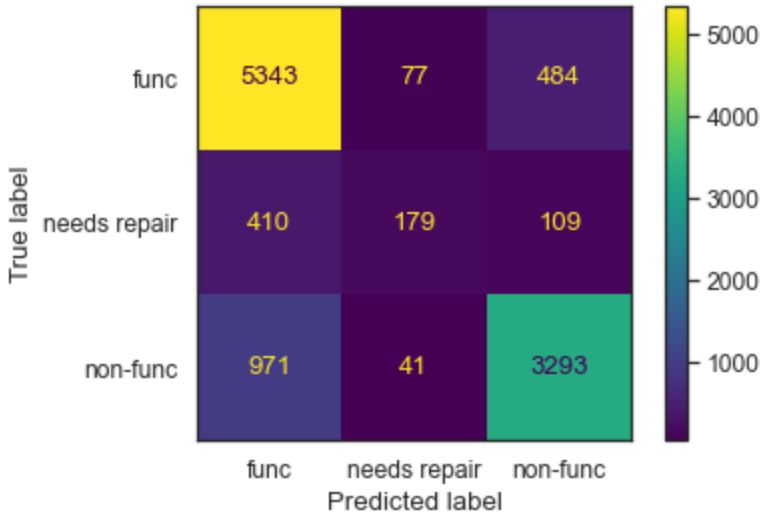
MODEL PERFORMANCE RESULTS

	Model	Accuracy	Precision	Recall	F1-Score
0	Linear Regression	0.741725	0.725382	0.741725	0.717297
1	Decision Tree	0.773907	0.767194	0.773907	0.765795
2	KNN	0.788118	0.781324	0.788118	0.782129
3	Random Forest	0.808197	0.803168	0.808197	0.798445
4	XG Boost	0.810030	0.806290	0.810030	0.802104

Random Forest Classifier

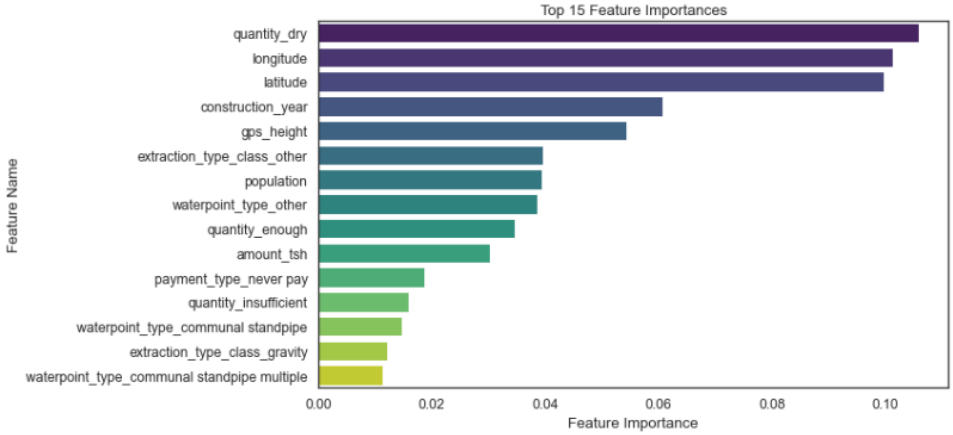
	precision	recall	f1-score	support
functional	0.79	0.90	0.85	5904
functional needs repair	0.60	0.26	0.36	698
non functional	0.85	0.76	0.80	4305
accuracy			0.81	10907
macro avg	0.75	0.64	0.67	10907
weighted avg	0.80	0.81	0.80	10907

```
Overall accuracy score 0.8081965710094435
Overall precision score 0.803168218299176
Overall recall score 0.8081965710094435
Overall F1-score 0.7984451084672377
```



The random forest model demonstrates good overall performance with an **accuracy score of 81%.** It excels in **correctly identifying non-functional water pumps, achieving an impressive F1-score of 0.80 for this class**.

However, **it struggles with the "functional needs repair" category, as indicated by the lower precision and recall scores**, suggesting that the model faces challenges in accurately predicting this class.



- **"quantity_dry"** emerges as the most critical factor, contributing approximately **10.57%** to the model's predictive power.

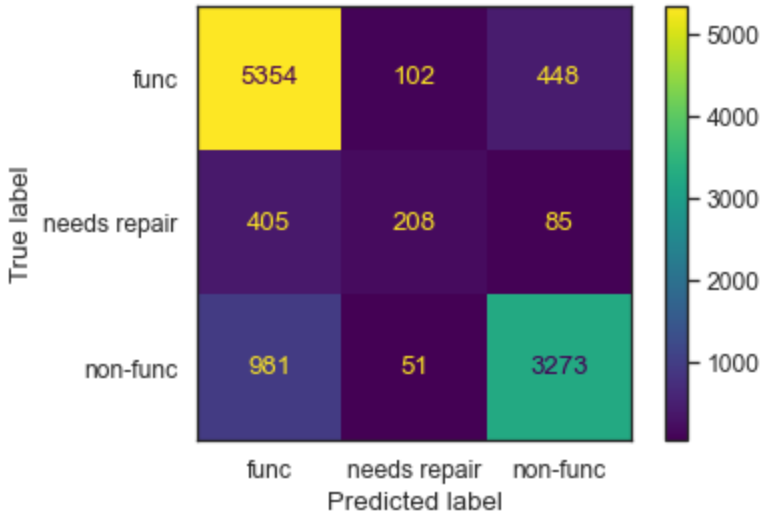
- **"longitude" and "latitude" closely follow, each contributing around 10.12% and 9.96%.**

- **"construction_year" plays a vital role at 6.07%, while "gps height" contributes approximately 5.43%.**

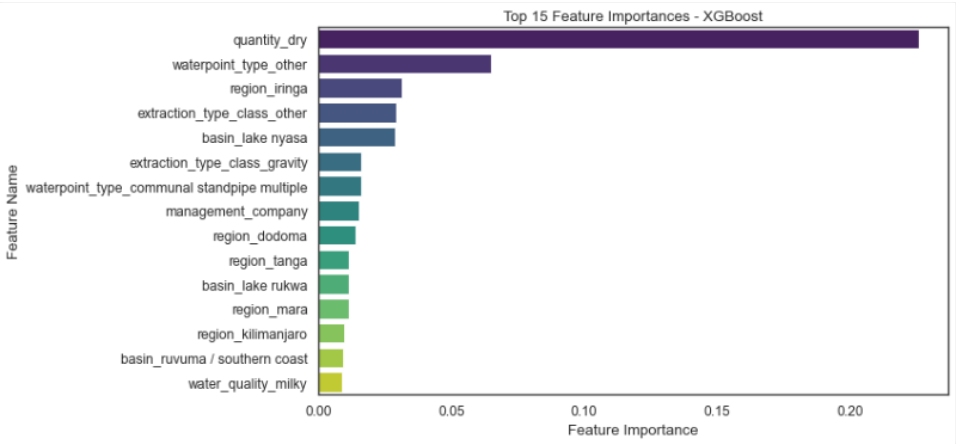
eXtreme Gradient Boosting (XGBoost)

	precision	recall	f1-score	support
functional	0.79	0.91	0.85	5904
functional needs repair	0.58	0.30	0.39	698
non functional	0.86	0.76	0.81	4305
accuracy			0.81	10907
macro avg	0.74	0.66	0.68	10907
weighted avg	0.81	0.81	0.80	10907

```
Overall accuracy score 0.8100302557990281
Overall precision score 0.8062899166465556
Overall recall score 0.8100302557990281
Overall F1-score 0.8021044089672305
```



The XG Boost model stands out as the top-performing model in various aspects. It achieves the **highest accuracy, reaching 81%, and has the best F1-Score of 80.21%**, showcasing its overall effectiveness in classifying water pump conditions.



- **"quantity_dry" notably emerges as the most influential factor, accounting for approximately 22.58% of the model's predictive power**, highlighting the substantial impact of water scarcity on functionality.

- **"waterpoint_type_other" and "region_iringa" play a key role contributing 6.50% and 3.12% importance** respectively underscoring the significance of water point types and region in water well functionality.

Summary

I recommend that the NGO utilizes the XG Boost model as the primary tool for predicting water point functionality in Tanzania.

The model achieved the **highest accuracy of 81% and the highest F1-Score of 80.21%** although the Random Forest Classifier was not far behind as it achieves an accuracy of 80.82% and F1-Score of 79.84%.

The model excelled in predicting "functional" and "non-functional" water pump conditions but similarly to the other models, it struggled with the "functional needs repair" class.

The NGO should also take into consideration the following findings identified through the exploration of the data:

1. A positive correlation between water availability and functionality is evident.
2. Regions like Iringa stands out with the highest number of functional water pumps, but Morogoro, Mbeya, and Shinyanga face substantial challenges with non-functional pumps.
3. "Never pay" water pumps are prevalent in Tanzania however they have a higher number of non-functioning water points compared to functioning ones.
4. The status of water points varies across different water basins, with some basins, like Lake Rukwa, experiencing a high number of non-functional points.
5. Total Static Head measurement, may serve as a valuable predictor for pump functionality.
6. Water points with zero recorded populations tend to have a higher proportion of non-functional pumps. However, there is a clear correlation between population and water point functionality.
7. There is a distinct correlation between the age of water pumps and their functionality.

Study Limitations

1. Data-related challenges included missing values in various features, with "scheme_name" having the highest count of missing data, and other features like "funder," "installer," "public_meeting," "scheme_management," and "permit" also containing substantial gaps.
2. Additionally, certain features had a high prevalence of zero values, raising concerns about data quality and making it challenging to handle these during modeling.
3. Class imbalance in the target variable was addressed using SMOTE, but it had limited impact on model performance.

Recommendations for Future Study

1. Implement rigorous data quality assurance measures to reduce the presence of missing values, ensure consistent data entry practices, and address zero values accurately.
2. Evaluate a broader range of machine learning algorithms, including ensemble methods, and gradient boosting, to identify the most suitable algorithms for this classification task.
3. Collect more diverse and comprehensive data, especially concerning well characteristics, water quality, geological factors, and socioeconomic information of the regions. This additional data can contribute to a better understanding of water pump functionality.