# Predictive Analysis of Water Pump Functionality in Tanzania

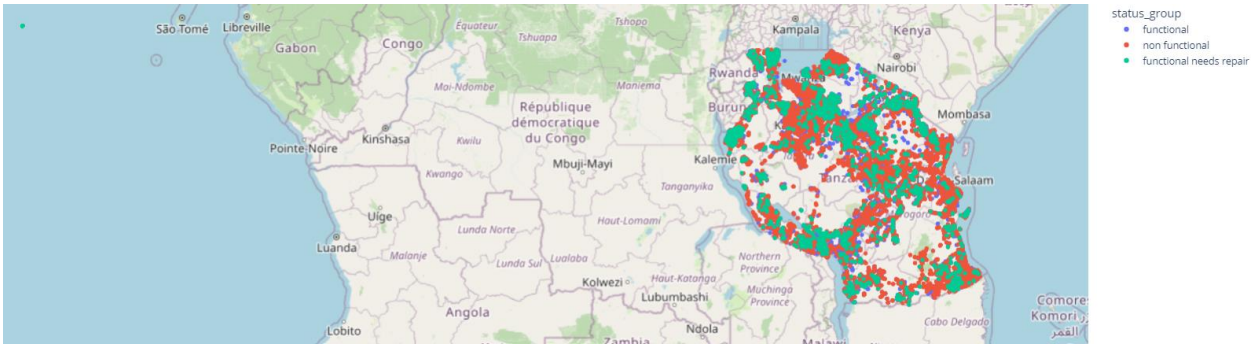
## Overview

This report details the process and outcomes of a data modeling project aimed at predicting the functionality of water pumps in Tanzania. The dataset used for this project includes information about various factors influencing the status of water pumps, such as location, construction details, and management practices.

## Data Exploration

### Initial Exploration

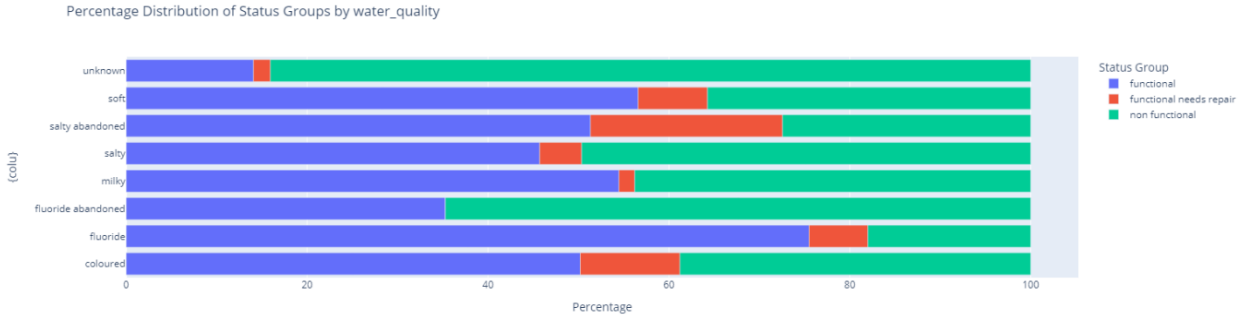
The dataset was loaded and explored to understand its structure, features, and the distribution of the target variable, 'status_group.' Descriptive statistics, visualizations using Plotly Express and Seaborn, and geographical mapping were employed to gain insights into the dataset.

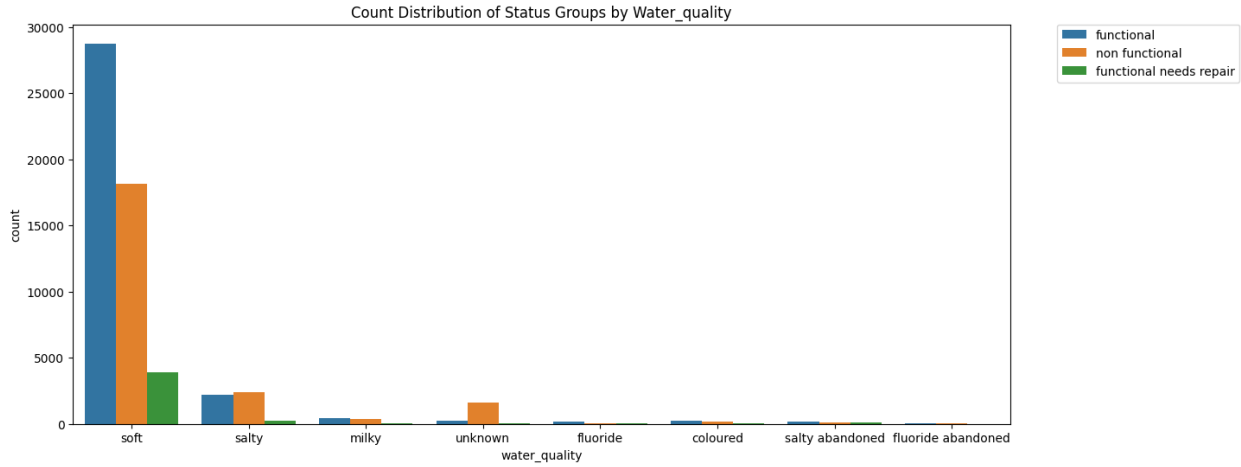



- Pumps location showed on world map. All of pumps are located in Tanzania. But 1802 pumps location was not correct longitudes and latitudes are around zero. They are showed on left corner of map.

### Exploratory Data Analysis

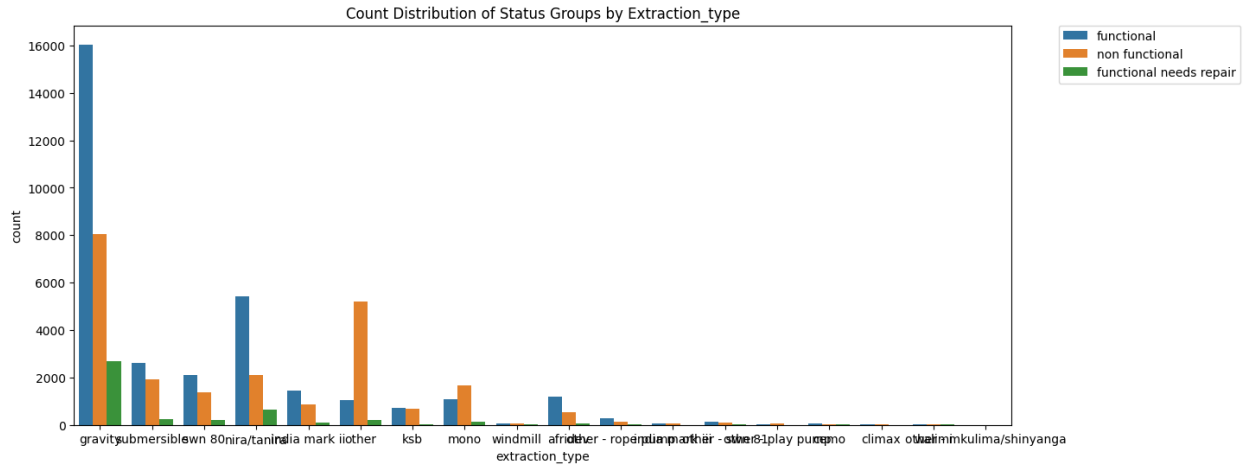
A detailed exploration of various categorical features, such as 'management,' 'payment,' and 'water_quality,' was conducted. Horizontal bar plots and count distributions were visualized to understand the relationships between these features and the target variable.

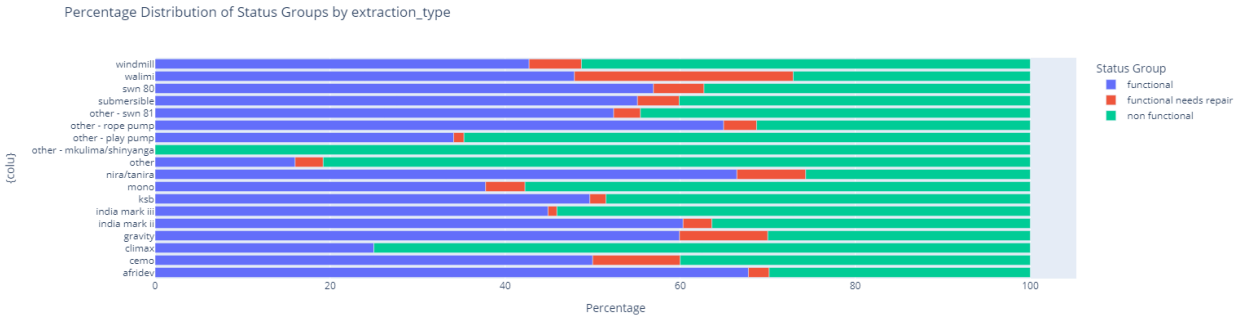



Most interested column is water quality, because I think water is most impact to pumps. From above chart, more than 75% pumps of fluoride water are functional. Also 56% of soft water pumps are functional. On the other hand more that 80% of water quality unknown pumps are non functional.

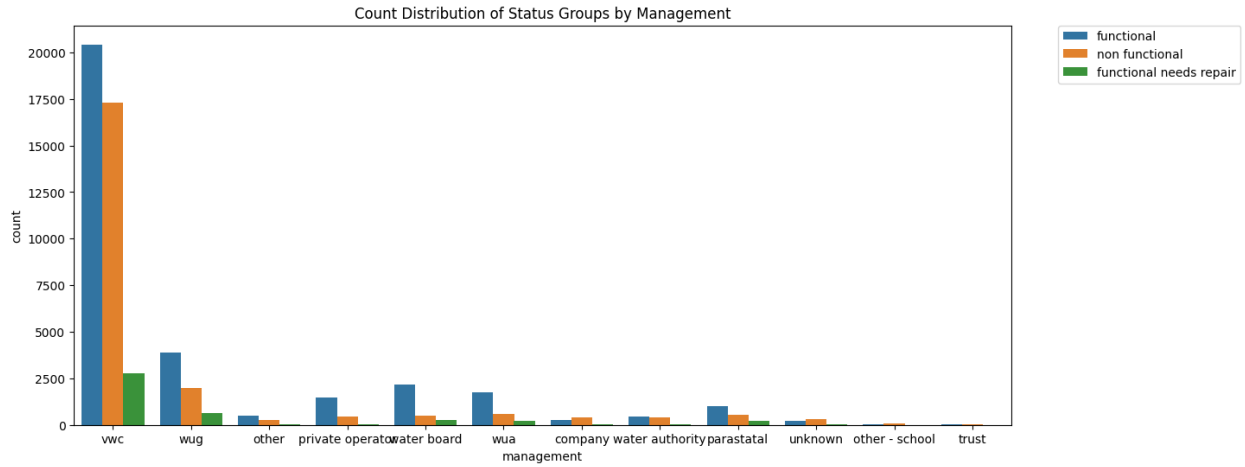
In the dataset, most of water quality was soft.

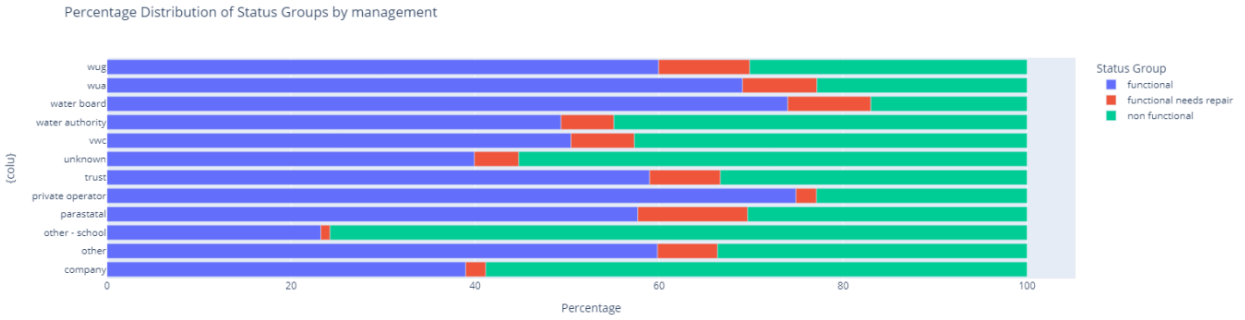


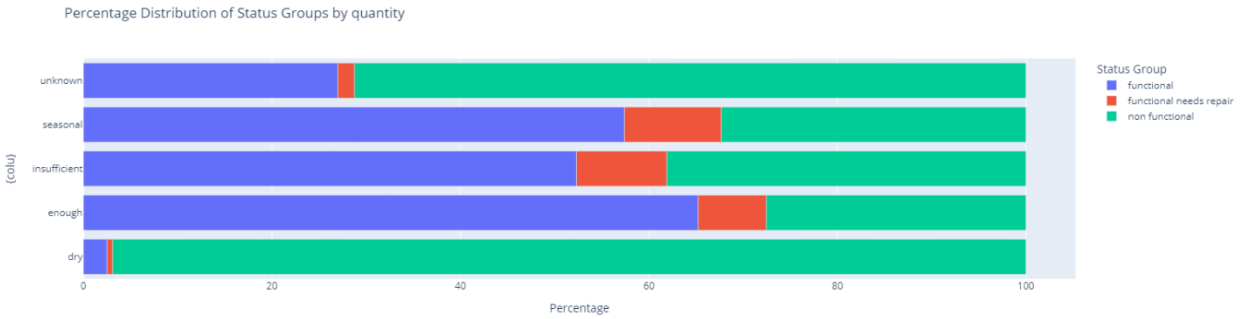

All pumps of the "Mkulima Extraction" type are non-functional. This indicates a critical issue with this specific pump type, requiring thorough investigation and potential redesign or replacement.

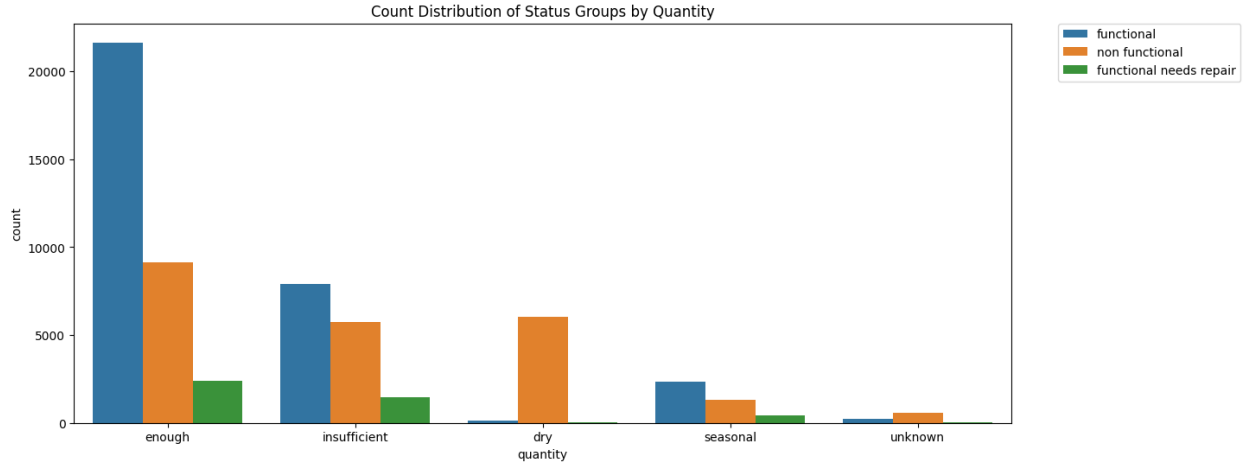
Rope Pump, Nira/Tanira, Indian Mark II, Gravity, Afridev:  More than 60% of pumps belonging to these types are functional. This suggests that these pump types exhibit a higher level of reliability and effectiveness in providing water.



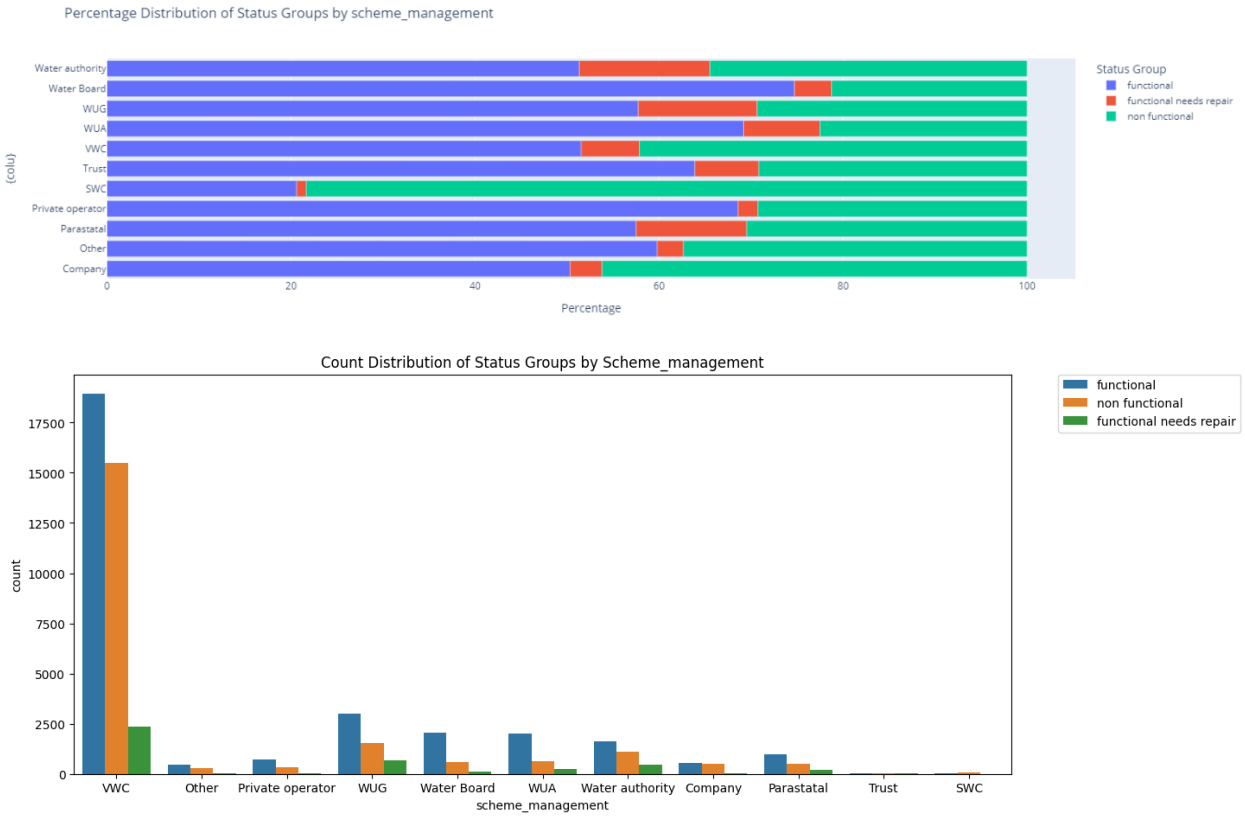




Management - Other School: Approximately 75% of pumps under the "Other School" management category are non-functional. This indicates a significant challenge within this management classification, requiring attention and intervention.

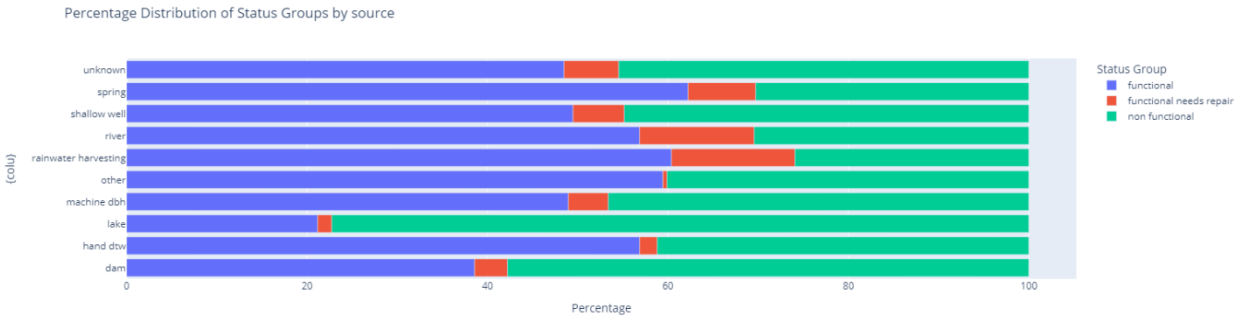


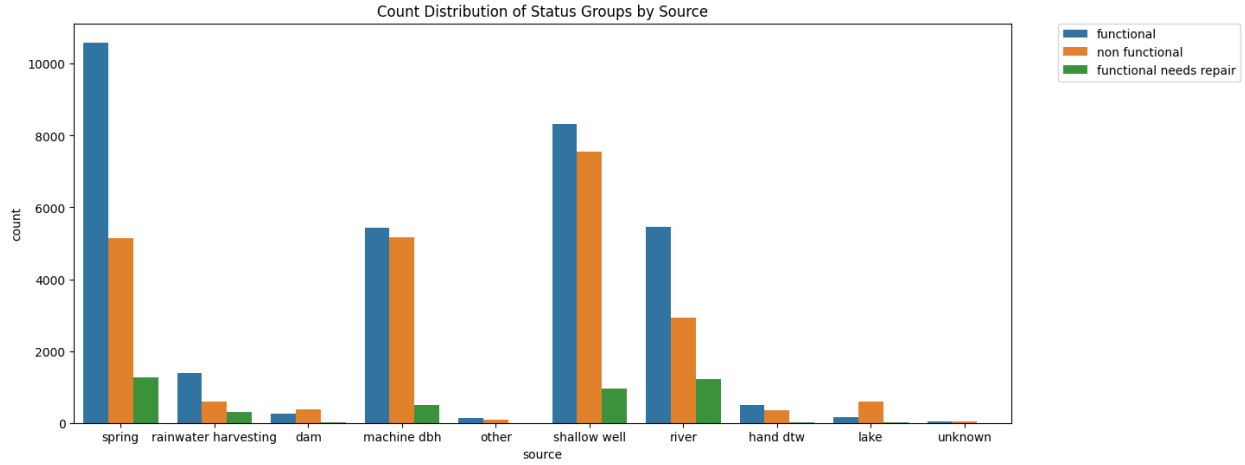




Water Quantity - Dry: Looking at the "Dry" water quantity, it's striking that a big 96% of pumps linked to these dry conditions are not working. This high percentage tells us that when there's not enough water (dry conditions), the pumps are more likely to be non-functional.




Scheme Management - SWG: A substantial 78% of pumps associated with the "SWG" scheme management category are non-functional. This high percentage indicates a notable correlation between SWG scheme management and non-functionality.




Water Source - Lake: When it comes to pumps connected to lakes as a water source, over 75% of them are not working. This high percentage indicates a clear link between using lake water as a source and pump non-functionality.

## Data Preprocessing

### Feature Removal

Unnecessary features, such as 'wpt_name,' 'scheme_name,' and 'recorded_by,' were removed to streamline the dataset.

- Dropped 'wpt_name' column, because most of them are unique.
- Dropped 'scheme_name' column, because most of them are missing.
- Dropped 'recorded_by' column, because all same.

| Name of Column | Missing value | Unique value |
|---|---|---|
| 'wpt_name' | 2 | 37399 |
| 'scheme_name' | 28810 | 2695 |
| 'recorded_by' | 0 | 1 |

Data cleaning from the map graph, dropped longitude is 0 values. Dropped values 1812.

### Handling Missing Values

Missing values were addressed using forward filling, ensuring that relevant information was retained without compromising the dataset's integrity.

### Feature Engineering

New features, including 'age' and 'year_recorded,' were created to enhance the dataset. This involved calculating the age of water pumps and extracting the year from the recorded date.

### Encoding Categorical Variables

Label Encoding was applied to convert categorical variables into a numeric format. This step is crucial for machine learning algorithms to process categorical data effectively.

## Modeling

### Model Selection

Several classification models, including Random Forest, Decision Tree, K-Nearest Neighbors, Stochastic Gradient Descent, Logistic Regression, Linear SVC, Gaussian Naive Bayes, and Perceptron, were trained and evaluated.

### Model Evaluation

The accuracy of each model was assessed, providing an initial understanding of their performance. For future iterations, additional metrics like precision, recall, and F1-score could be incorporated for a more comprehensive evaluation.

| Classification models | Accuracy |
|---|---|
| Random Forest | 80.82 |
| Decision Tree | 74.55 |
| K-Nearest Neighbors | 63.27 |
| Stochastic Gradient Descent | 60.05 |
| Logistic Regression | 56.93 |
| Linear SVC | 58.81 |
| Perceptron | 43.87 |
| Gaussian Naive Bayes | 53.89 |

- The Random Forest model demonstrated the highest accuracy among the tested models. This indicates its effectiveness in predicting the functionality of water pumps based on the provided features.

- The Decision Tree model showed commendable accuracy, making it a valuable option for predicting water pump functionality. Its performance suggests good decision-making based on the input features.

- While the K-Nearest Neighbors model performed reasonably well, there is room for improvement. Further optimization or consideration of alternative models may enhance accuracy.

## Conclusion

In our project to predict water pump functionality, we've achieved meaningful progress. One standout performer is the Random Forest model, boasting an impressive accuracy of 80.82%. This means it's quite good at helping us make decisions about managing water supply systems.