

# PHASE 3 PROJECT – PETE NJAGI

NON- TECH PRESENTATION

**DRIVEN****DATA**

A Moring Student Data Science  
Project





# EXECUTIVE SUMMARY

"Blue Life NGO", (A hypothetical NGO) in partnership with the Government of Tanzania, is dedicated to enhancing the nation's access to clean and reliable water supplies.

This report outlines the recent machine learning initiative to analyze water well functionality, identify maintenance needs, and inform new well construction strategies.



# PROBLEM

- Tanzania, as a developing country, struggles with providing clean water to its population of over 57,000,000. There are many water points already established in the country, but some are in need of repair while others have failed altogether.

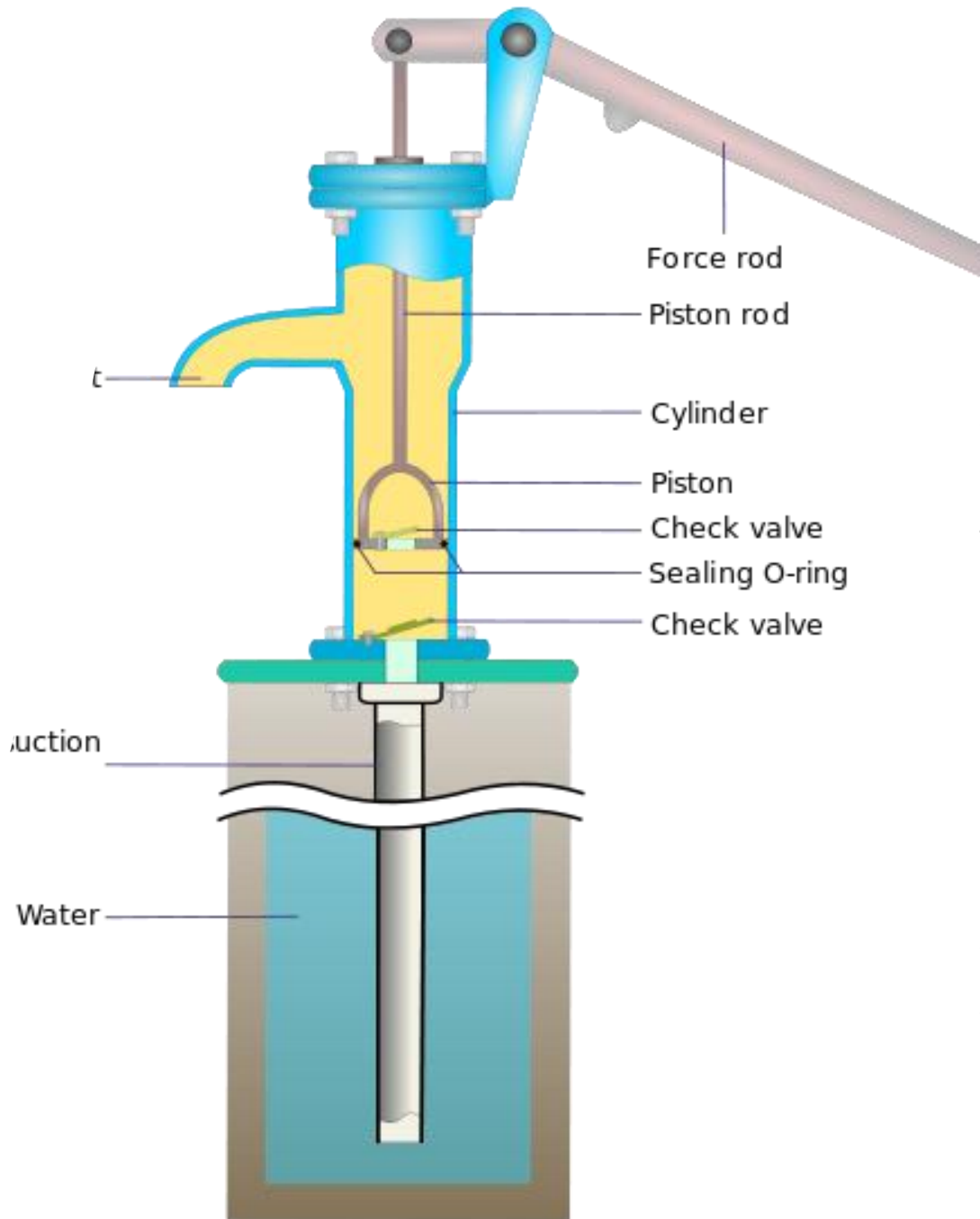
## Solution

- We built a classifier to predict the condition of a water well, using information about the sort of pump, when it was installed, etc. Our audience is an NGO focused on locating wells needing repair, partnered with the Government of Tanzania looking to find patterns in non-functional wells to influence how new wells are built





# PROJECT INTRODUCTION



Access to clean water is vital to health, agriculture, and overall development. In Tanzania, thousands of water wells serve as the primary water source for many communities.

However, with many of these wells falling into disrepair, there's a critical need to identify and prioritize maintenance to ensure sustainable water access. Additionally, understanding the factors influencing well functionality can guide the construction of new, more reliable wells.



# DEVELOPMENT PROCESS

This Tanzanian Water Wells project employs machine learning to predict the operational status of water pumps across various regions in Tanzania. The aim is to identify which pumps are functioning, which require repairs, and which are non-functional. This predictive insight is a crucial enabler for effective resource allocation and operational planning for water supply maintenance.

## A DATA SCIENCE APPROACH

Our team utilized a dataset of Tanzanian water wells, comprising numerous features such as geographical location, construction details, and operational status. We employed machine learning models to predict the current status of wells, categorizing them into functional, non-functional, or in need of repair.





# BUSINESS OBJECTIVE

## OUR BUSINESS OBJECTIVES ENCOMPASS:

- Crafting a predictive model to forecast well statuses for preemptive action.
- Guiding the placement of new wells based on historical data trends.
- Assisting NGOs and government agencies in data-driven decision-making.
- Enhancing the efficiency of repair and maintenance operations.
- Contributing to the overall health and well-being of Tanzanian communities.
- Strengthening partnerships for international support based on actionable data.

These goals underscore our commitment to leveraging data for improving the accessibility and reliability of water resources.



# GOALS FOR THE PREDICTIVE PROJECT

## Hypotheses



### SOFTWARE OPPORTUNITIES

- **Null Hypothesis (H0):** Water quality has no significant effect on well functionality.
- **Alternative Hypothesis (H1):** Poor water quality is associated with non-functional wells.
- ii. Hypothesis on Geographic Location and Well Status:
  - **Null Hypothesis (H0):** The geographic location of wells does not significantly affect their functionality.
  - **Alternative Hypothesis (H1):** Wells in certain regions are more likely to be non-functional or in need of repair.
- iii. Hypothesis on Installation Company and Well Durability:
  - **Null Hypothesis (H0):** The company that installed the well has no significant impact on its durability.
  - **Alternative Hypothesis (H1):** Wells installed by specific companies have higher functionality rates.

# VALUE PROPOSITION

To Governmental and Non-Governmental Organizations  
(NGOs) Managing Water Resources

UNIQUE

VALUE

PROPOSITION

The ability to accurately forecast the functional status of water wells across Tanzania is invaluable. It translates to:

- Proactive Maintenance: Machine learning models can predict which wells are likely to need repairs soon, allowing for proactive maintenance before they fail.
- Optimized Resource Allocation: With predictive insights, resources can be allocated more efficiently, focusing on areas with the highest need, which is particularly important where resources are limited.
- Enhanced Planning: Long-term planning of water resource management can be informed by trends and patterns identified through data analysis, leading to more sustainable practices.
- Community Impact: Ensuring the functionality of water wells directly impacts community health and livelihood, as access to clean water is critical.



# DATA UNDERSTANDING

From Governmental and Non-Governmental  
Organizations (NGOs) Managing Water Resources

The analysis is based on data extracted from multiple sources, consolidated into two primary datasets:

- **WaterPoint Data Exchange:** Contains detailed records of water well functionality across Tanzania.
- **Tanzania Ministry of Water:** Provides official records of water infrastructure and maintenance.

## Data Highlights:

- **Geospatial Data:** Offers insights into patterns of well distribution and functionality.
- **Construction Details:** Helps correlate well types and ages with functionality status.
- **Operational Status:** Key target variable for predicting maintenance needs and guiding new constructions.

An abstract graphic on the left side of the slide. It features a large green shape with a white dotted pattern inside, and a smaller orange shape with a white plus sign pattern. There are also several small black wavy lines and a dotted line scattered around the shapes.

# MODEL DEVELOPMENT

We developed several models, starting with a simple logistic regression to establish a baseline, progressing to more sophisticated models like decision trees and Random Forest classifiers. Through iterative training and evaluation, we fine-tuned these models to achieve higher predictive accuracy.

- Logistic Regression
- Decision Tree
- Random Forest Model

An abstract graphic on the left side of the slide. It features several organic, blob-like shapes in orange and green. The orange shapes are filled with a pattern of small white dots or crosses. The green shape is solid. The background is white with scattered black dots and small black wavy lines. A horizontal bar with a color gradient from orange to dark blue is located at the bottom right of the graphic area.

## MODEL DEVELOPMENT - WHAT ARE WE PREDICTING?

The goal is to predict the **status\_group** of each water point, which can be one of the following:

- **Functional:** The water point is operational and working correctly.
- **Non-functional:** The water point is not operational and needs repairs or replacement.
- **Functional needs repair:** The water point is operational but has issues that need fixing.



An abstract graphic on the left side of the slide. It features a large green shape with a white dotted pattern, a smaller green shape with a white dotted pattern, and a large orange shape with a white plus sign pattern. There are also several small black wavy lines and a dotted line scattered around the shapes.

# MODEL DEVELOPMENT –

## HOW ARE WE MAKING THESE PREDICTIONS?

- 1. Model Training:** We use historical data with known outcomes to train a machine learning model. This involves feeding the model with features (like **amount\_tsh**, **gps\_height**, **water\_quality**, etc.) and the known status (**status\_group**) of each water point.
- 2. Model Learning:** The model learns patterns from this data. For example, it might find that wells with a certain feature profile are more likely to be non-functional.
- 3. Making Predictions:** Once trained, the model can take features from new, unseen data (without known outcomes) and predict the status group based on what it has learned.

# MODEL DEVELOPMENT - HOW DO WE KNOW THE RESULTS?



**1. Model Evaluation:** We evaluate the model's accuracy by making predictions on a separate set of data where you already know the outcome (the validation set). By comparing the model's predictions against the actual outcomes, we assess the model's performance using various metrics (like accuracy, precision, recall, and F1-score).

**2. Validation:** Through cross-validation or by using a test set, we ensure that the model's performance is reliable and not just due to chance or overfitting to the training data.

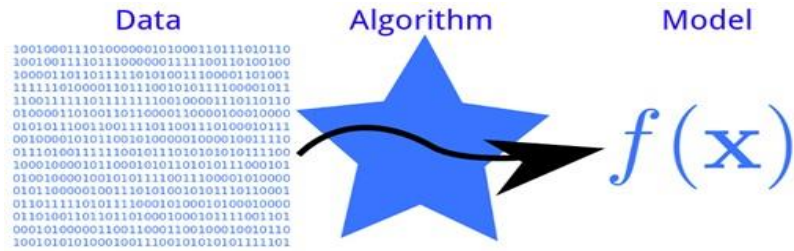
**3. Metrics:** Our choice of metric depends on what's important for the problem: Functional prediction

1. **Accuracy** is the overall correctness of the model but can be misleading if the classes are imbalanced.
2. **Precision** and **Recall** are used when the cost of false positives and false negatives are high respectively.
3. The **F1-score** provides a balance between precision and recall, useful when you need a single metric to reflect overall performance.

**4. Real-World Testing:** Ultimately, we'd test the model's predictions in the real world. For example, if the model predicts that a well is likely to need repairs, technicians can check the well to see if the prediction is accurate.

**5. Feedback Loop:** Any discrepancies can be used as feedback to improve the model, either by adjusting the model itself or by collecting more data that might be missing from the current dataset. We tuned our model to hit 81% accuracy score





# MACHINE LEARNING MODELS

## Logistic Regression Model

- Accuracy: 56.54%
- This model correctly predicts the status of water wells approximately 56.54% of the time.
- Logistic regression is a linear model that may struggle with complex patterns and interactions in the data, which might explain the lower accuracy.
- Given its simplicity and interpretability, this performance sets a baseline. However, the accuracy suggests that many wells' statuses are misclassified, highlighting the challenge of the classification task with linear models.

## Decision Tree Model

- Accuracy: 74.89%
- The decision tree model shows a significant improvement over logistic regression, correctly predicting the well status about 74.89% of the time.
- Decision trees can capture non-linear patterns through branching decisions based on feature values, which likely contributes to better handling of the complexities in the dataset.
- However, despite being more accurate than logistic regression, there's still room for improvement, as approximately 25% of predictions are incorrect. This misclassification rate might impact resource allocation for water well maintenance and repairs if used in practical applications.

## Random Forest Model

- Accuracy: 81%
- High Accuracy: The Random Forest model achieved an accuracy of approximately 81%, making it a reliable indicator for predicting the operational status of water wells across Tanzania.
- Complex Patterns Identification: Random Forests are capable of uncovering complex nonlinear patterns in the data, which can be crucial for understanding the various factors that contribute to well functionality and failure. These factors included quantity, date\_recorded, population.. Etc



# CONTEXTUALIZING THE RESULTS:

---



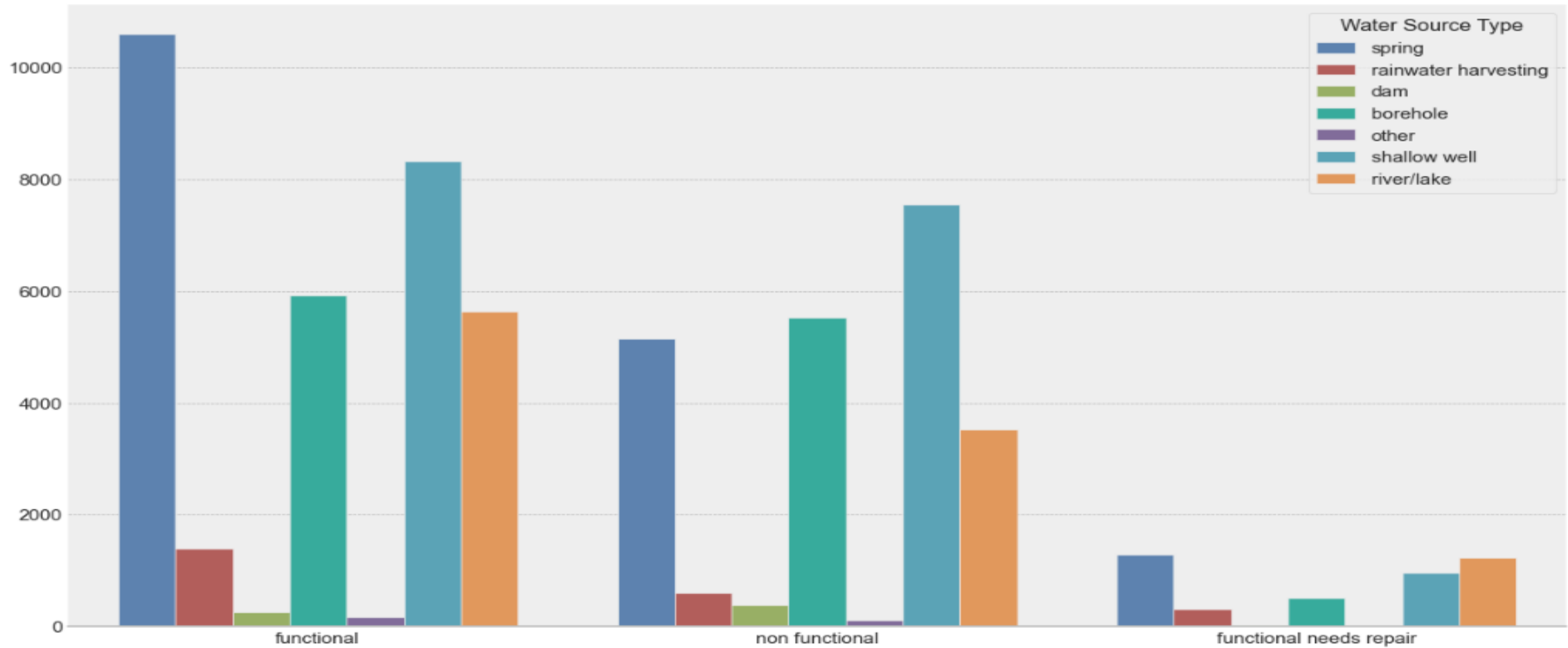
- Implications for Water Well Maintenance: Higher accuracy means better identification of wells needing repair or replacement, which is crucial for maintaining water access. The decision tree's better performance suggests it could be a more reliable model for prioritizing interventions on the ground.
- Model Selection and Improvement: While the decision tree model outperforms logistic regression, neither model perfectly classifies all water wells, indicating the problem's complexity and possibly the need for more complex modeling techniques or additional feature engineering.
- Further Steps: Given these results, exploring more complex models like the Random Forest (which we did) or Gradient Boosting might improve prediction accuracy. Additionally, considering the class imbalance and exploring techniques to address it (such as SMOTE or class weights) could enhance model performance, especially for minority classes like "functional needs repair."

# VISUALIZATIONS

16

## Tanzanian Water Pump Status Based On Water Source Type

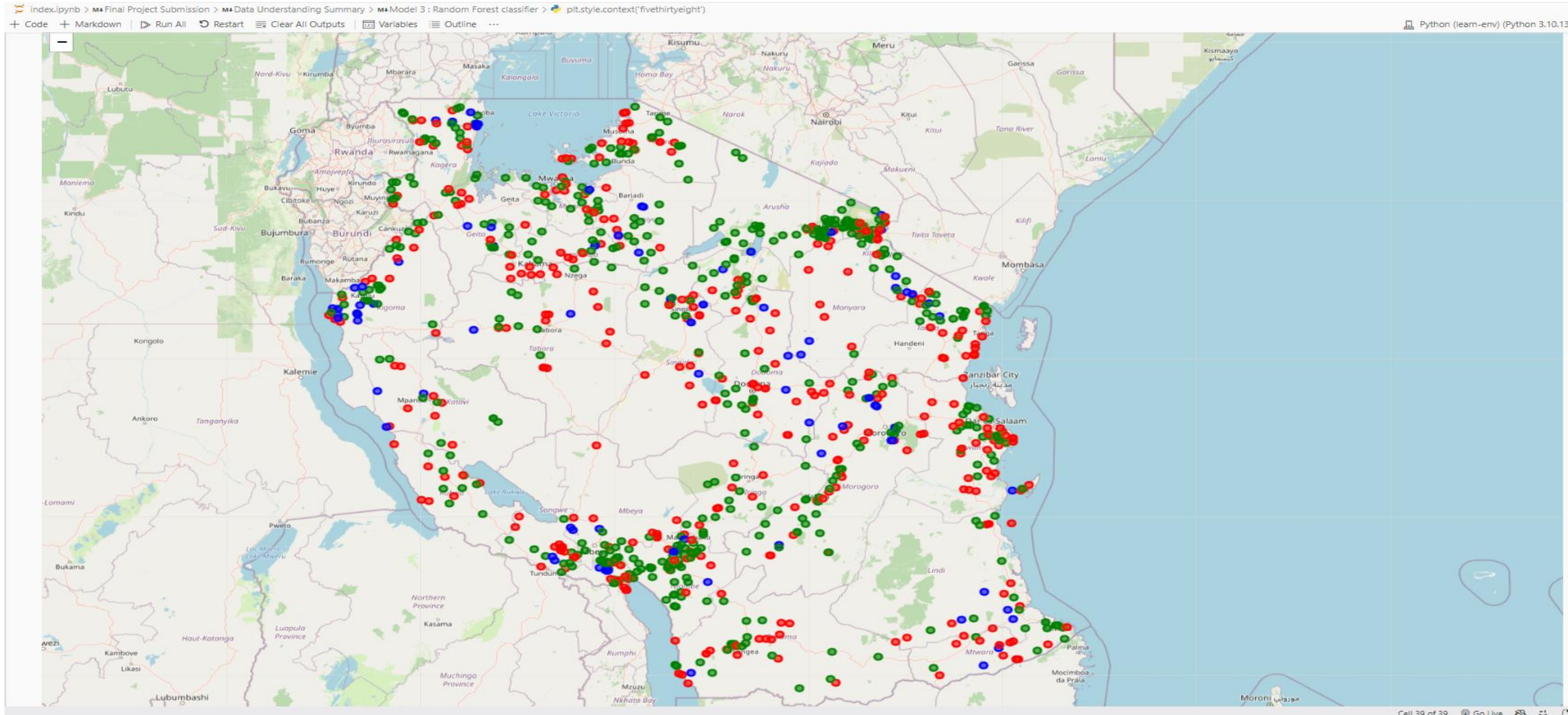
*Visualizing The Spread Of Water Sources vs Functionality*



This graph shows the water source type: The diagram shows that spring water is the most abundant water source, followed by shallow wells and boreholes.

Caption





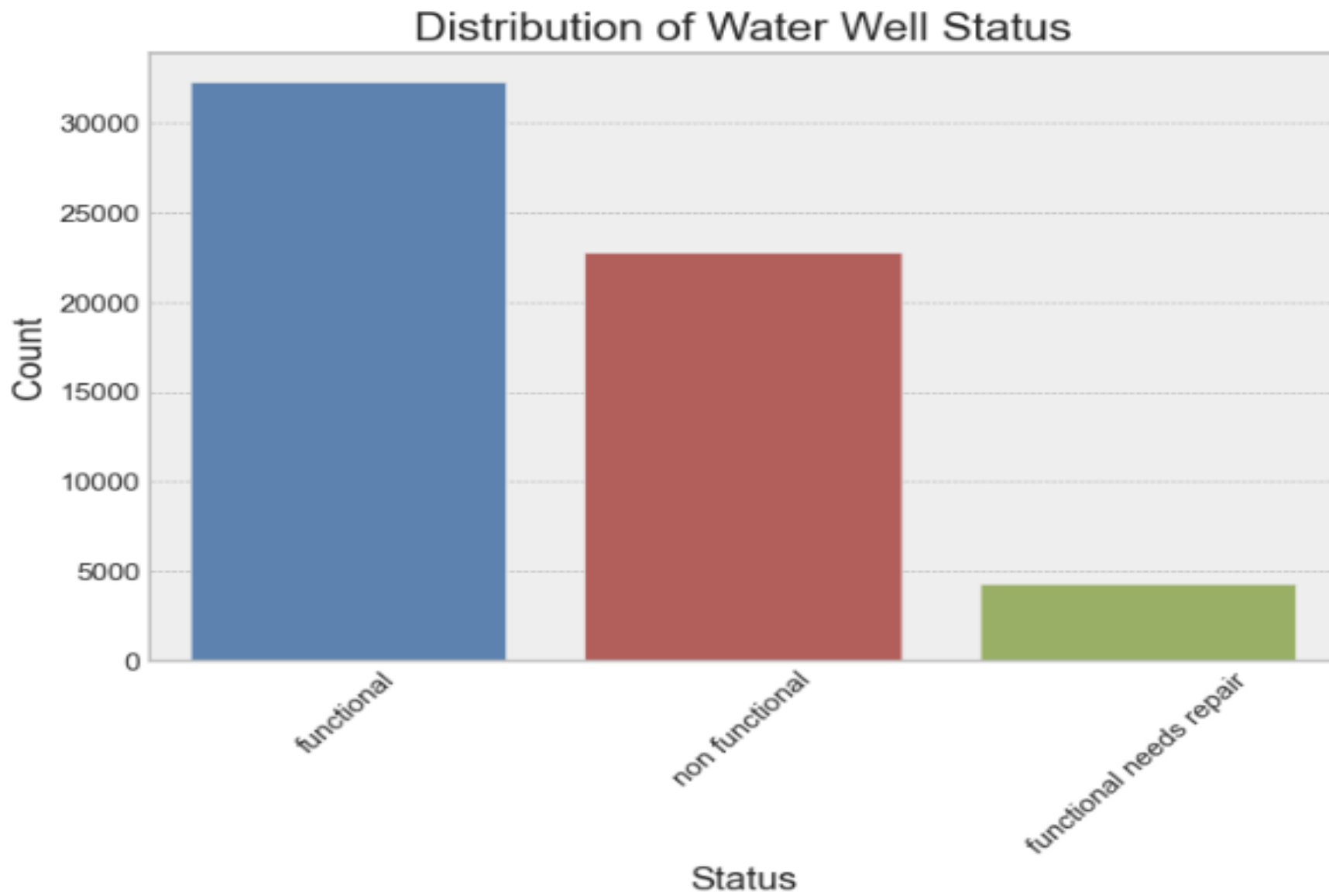
**Geospatial Distribution:** Since we have latitude and longitude data, we can plot the locations of the wells on a map and color-code them by their functionality status. This reveals geographic patterns of well functionality.

Caption

**RED – NON FUNCTIONAL**

**BLUE –FUNCTIONAL NEEDS REPAIR**

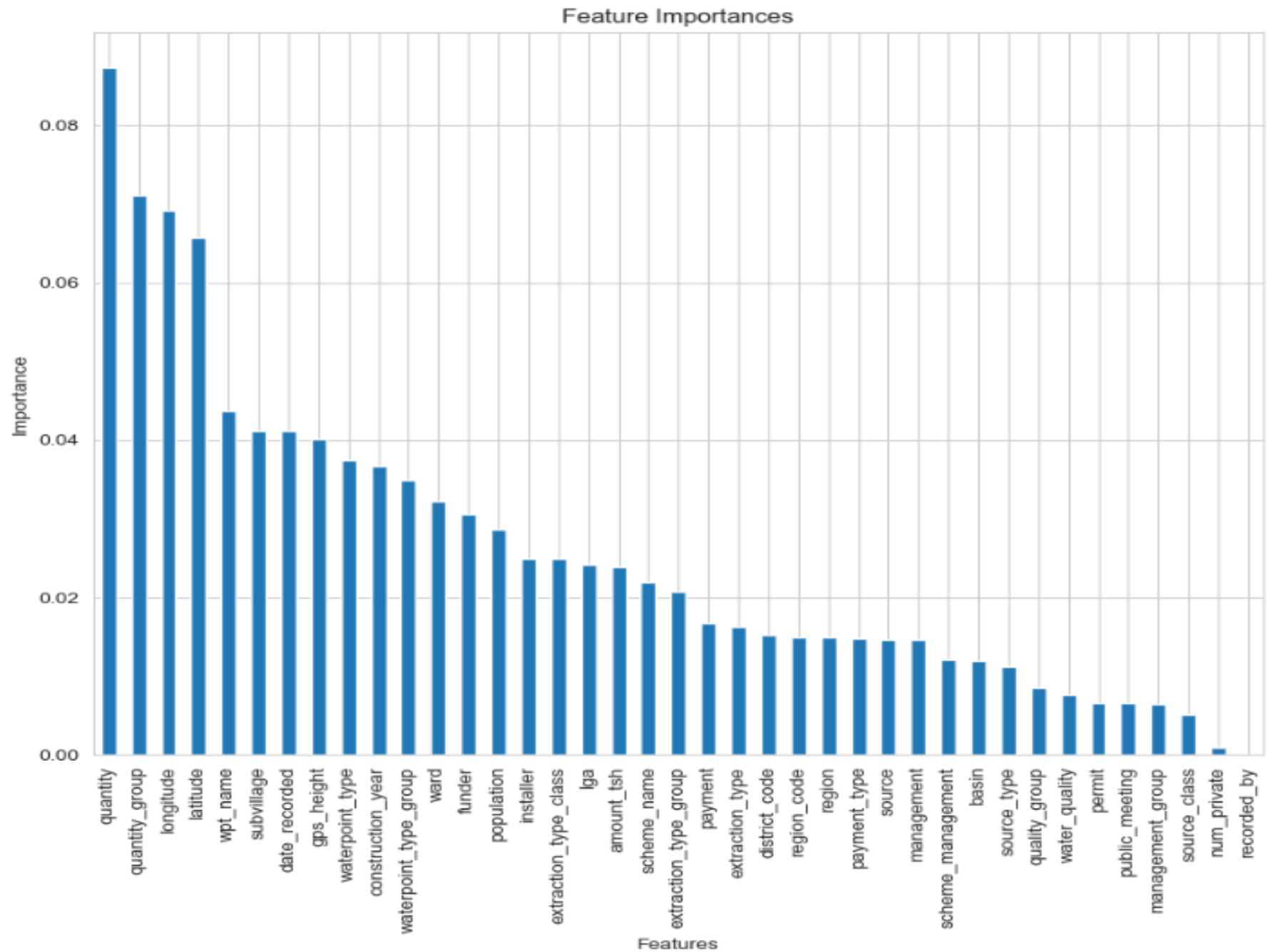
**GREEN- FUNCTIONAL**



Caption

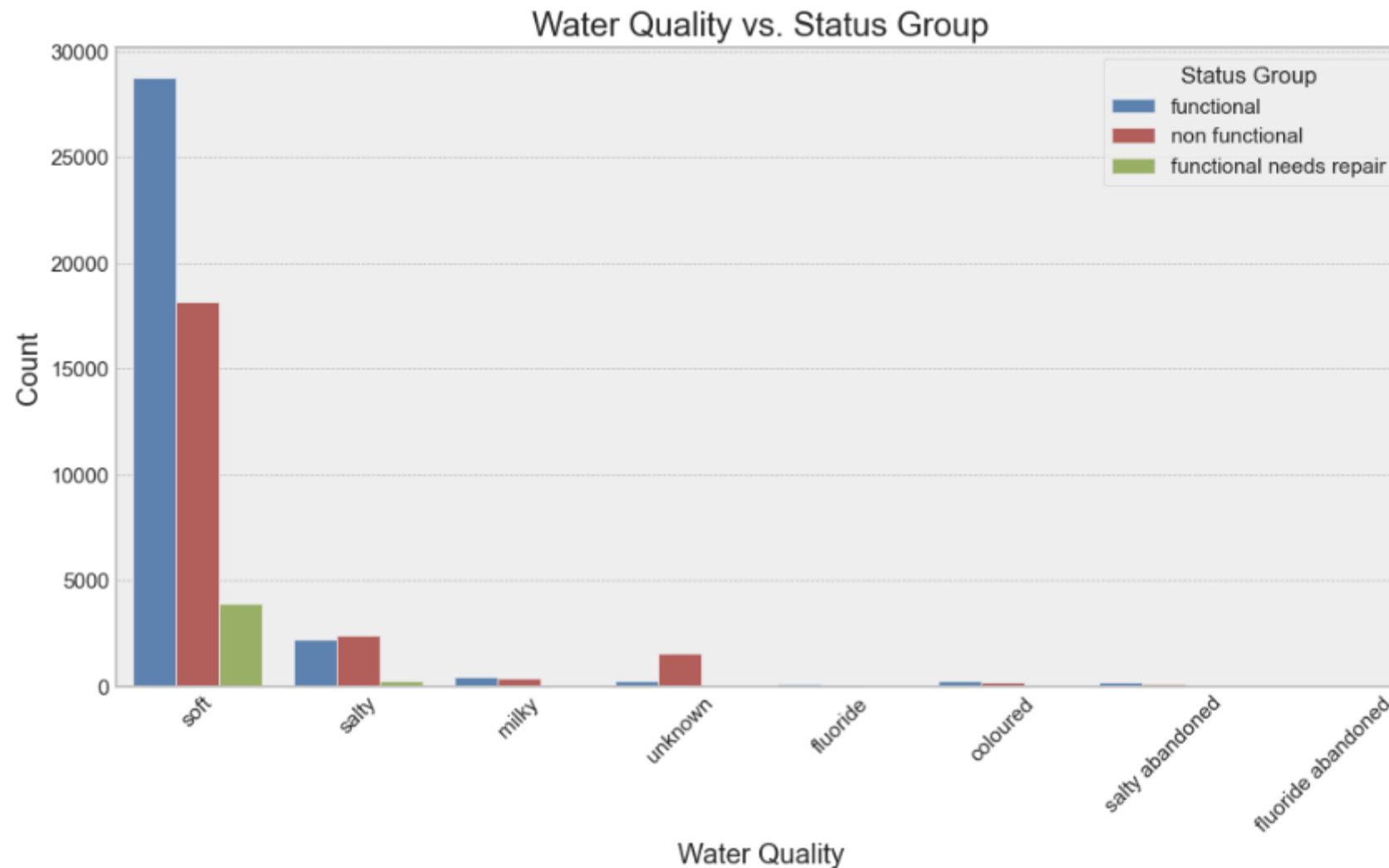
This graph shows the status count of the water wells by status

This graph shows the factors that influenced our predictive models the most..



Caption



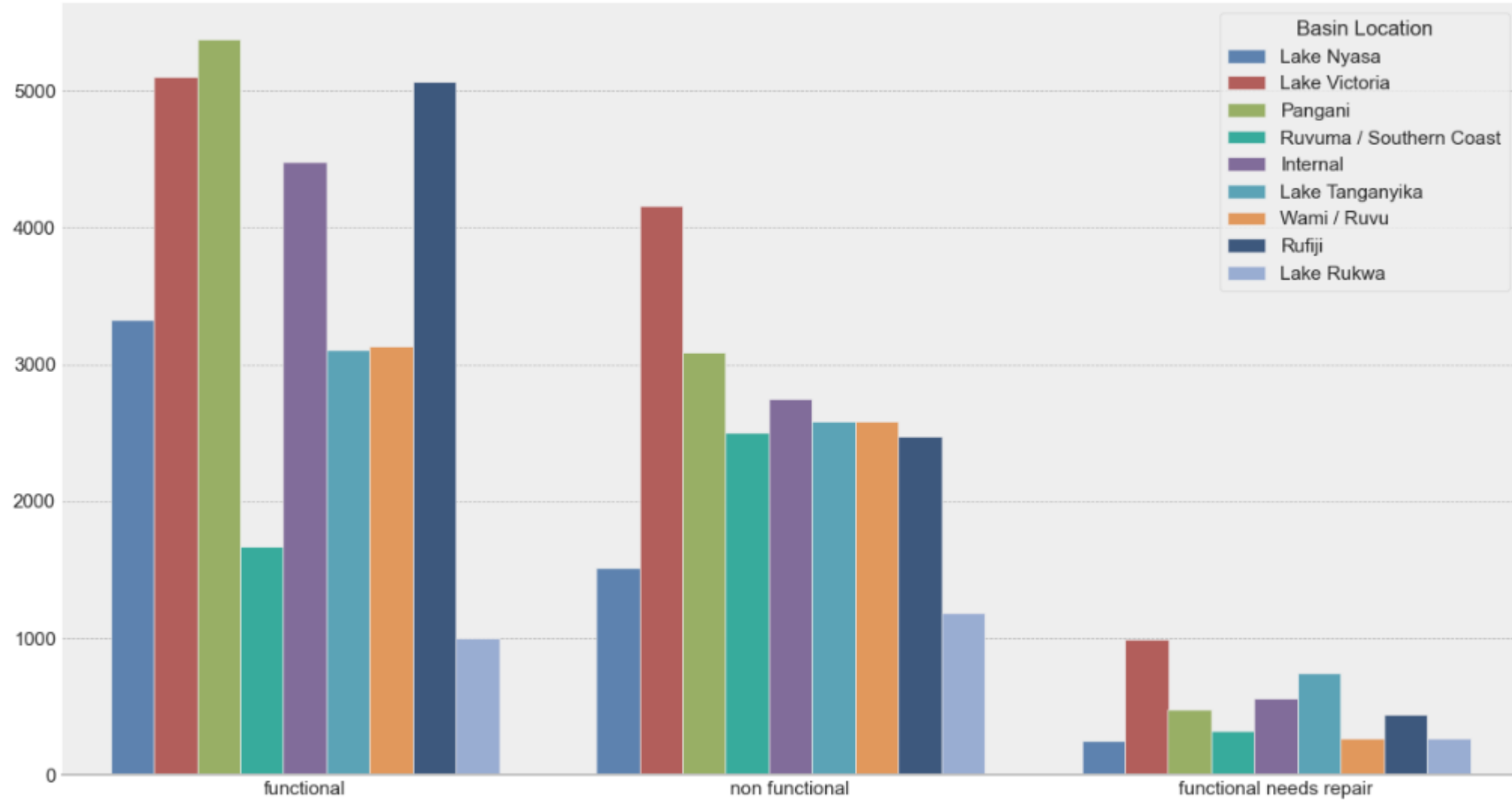


Caption

This graph shows the water quality against the status group functionality of the well

## Tanzanian Water Pump Status Grouped By Basin Location

*A Breakdown Of Pump Functionality Across The Basins*



Caption

This graph shows the status of water pumps grouped by basin locations



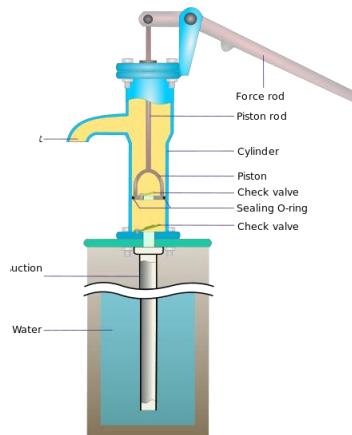
# EVALUATION AND RESULTS

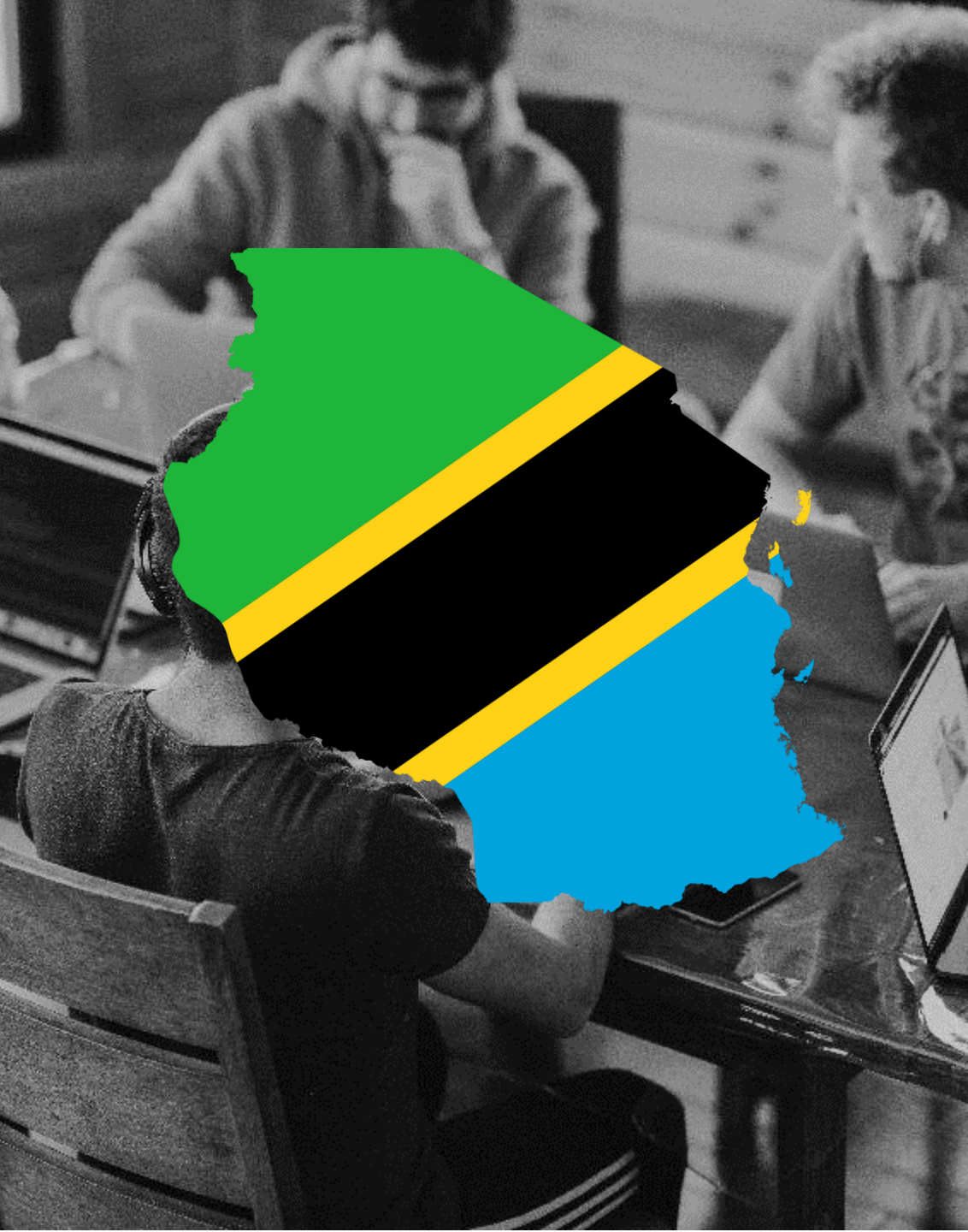
## Results

Our Random Forest model demonstrated the highest predictive accuracy, correctly identifying the status of water wells in approximately 81% of cases. This performance indicates a significant potential to prioritize well repairs effectively and to plan new constructions strategically.

## Insights and Patterns

The analysis revealed key factors contributing to well functionality, such as installation practices and geographical influences. Notably, certain regions exhibited higher incidences of well failure, which may inform targeted interventions.





# STAKEHOLDER RECOMMENDATIONS

## Potential benefits

- Prioritize the maintenance and repair of water wells based on the model's predictions to optimize resource allocation.
- Implement a system for continuous data collection and model retraining to adapt to changes over time.
- Consider targeted investigations in areas where the model has identified high rates of non-functionality or needs for repair, to understand and address underlying causes.
- Engage communities in reporting well statuses to supplement model predictions with real-time, on-the-ground insights.





## RECOMMENDATIONS

BASED ON OUR FINDINGS, WE RECOMMEND THE FOLLOWING ACTIONS:

**PROACTIVE MAINTENANCE:** LEVERAGE THE MODEL'S PREDICTIONS TO SCHEDULE TIMELY REPAIRS, PREVENTING WELL FAILURES.

**DATA-DRIVEN CONSTRUCTION:** UTILIZE INSIGHTS FROM NON-FUNCTIONAL WELLS TO IMPROVE NEW WELL DESIGNS AND LOCATIONS.

**COMMUNITY TRAINING:** IMPLEMENT COMMUNITY WORKSHOPS ON WELL MAINTENANCE, INFORMED BY COMMON ISSUES IDENTIFIED.

**CONTINUOUS MONITORING:** ESTABLISH A MONITORING SYSTEM TO TRACK THE PERFORMANCE OF WELLS POST-MAINTENANCE.

**POLICY AND AID:** USE DATA INSIGHTS TO SHAPE WATER RESOURCE POLICIES AND TO SUPPORT REQUESTS FOR INTERNATIONAL AID.

LET'S DIVE IN

# GITHUB LINK

[HTTPS://GITHUB.COM/PETEZDJ/PHASE3PROJECT](https://github.com/petezdj/phase3project)

Phase3Project Public

main 1 Branch 0 Tags


Go to file Add file Code

PeteZDJ Almost 95 99a3f31 · 42 minutes ago 12 Commits

images	Almost 90	1 hour ago
1.jpg	Almost 90	1 hour ago
SubmissionFormat.csv	Initial commit 1	3 days ago
TestSetValues.csv	Initial commit 1	3 days ago
TrainingSetLabels.csv	Initial commit 1	3 days ago
TrainingSetValues.csv	Initial commit 1	3 days ago
hand_pump_diagram.png	Sat Changes	2 days ago
index.ipynb	Almost 90	1 hour ago
readme.md	Almost 95	42 minutes ago
water-780x470.jpg	Sat Changes	2 days ago

README

## Phase-3-project



About

No description, website, or topics provided.

- Readme
- Activity
- 0 stars
- 1 watching
- 0 forks

Releases

No releases published  
[Create a new release](#)

Packages

No packages published  
[Publish your first package](#)

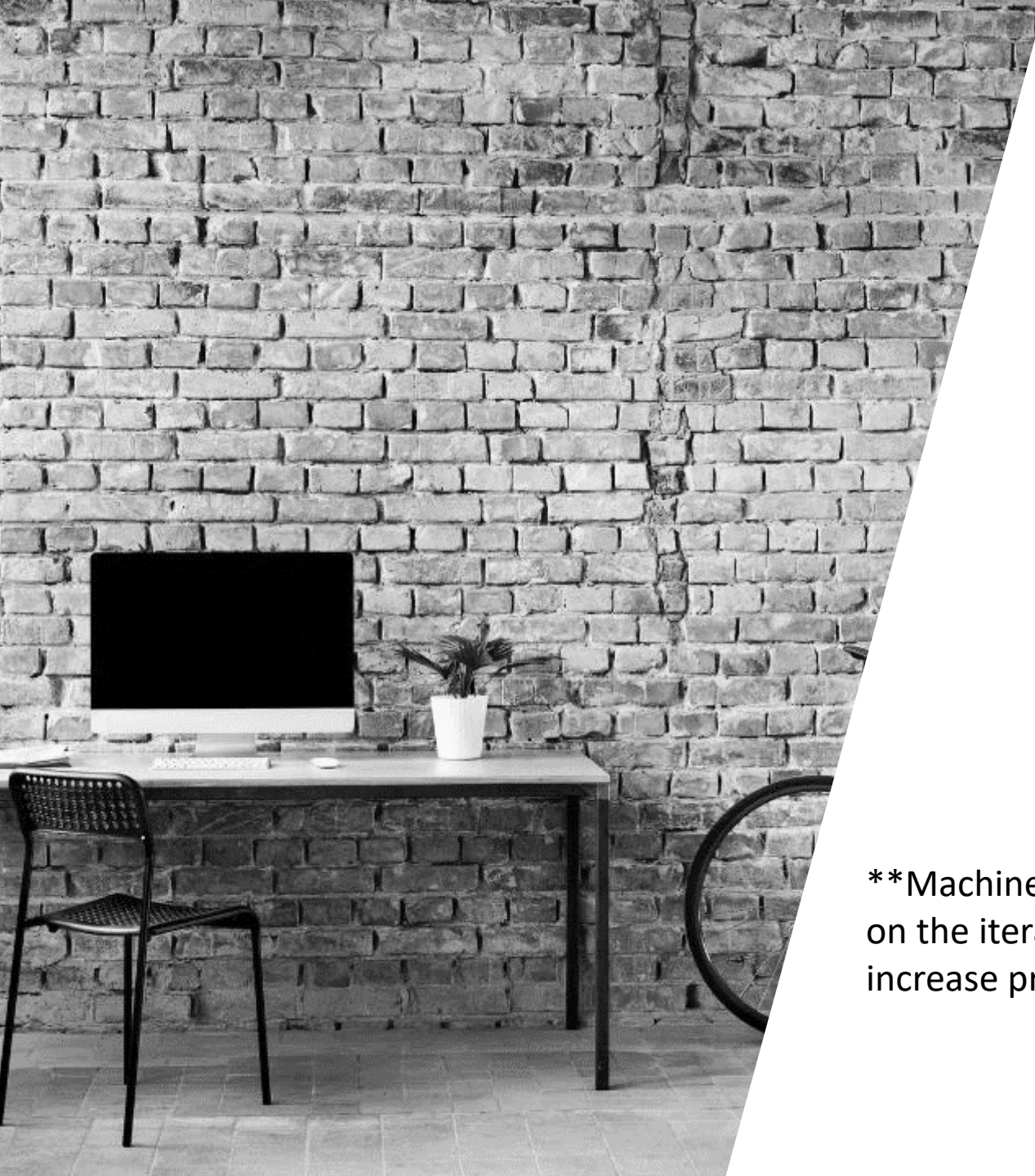
Languages

- Jupyter Notebook 100.0%

This initiative marks a pivotal step towards a data-informed approach in managing Tanzania's water resources. With continuous refinement and collaboration, the predictive model can become an indispensable tool for "Blue Life NGO" and the Government of Tanzania in their mission to provide clean water to all citizens.

CONCLUSION





# WHAT'S NEXT

## LOOKING AHEAD

**\*\*Machine Learning Model Refinement\*\***: Inclusion of expanded details on the iterative development and tuning of machine learning models to increase prediction accuracy and reliability.



# THANK YOU



<https://github.com/PeteZDj/Phase3Project>

We thank

