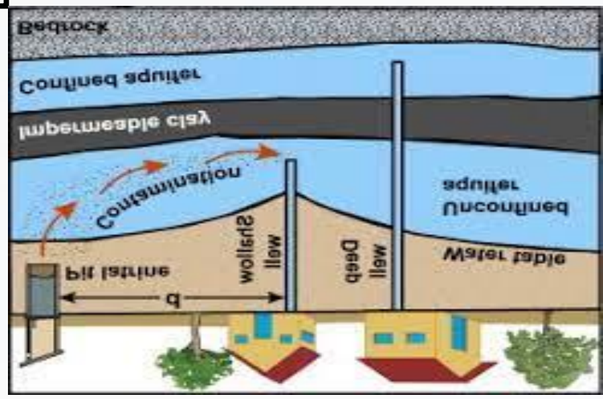
Water Works: Unearthing Insights from Tanzania's Aquifer Data



CONTENT

OVERVIEW

This project focuses on predicting the condition of water wells in Tanzania, aiming to aid stakeholders in effectively managing water infrastructure and ensuring access to clean water. By leveraging machine learning techniques, the project aims to provide insights into the factors influencing well conditions and offer actionable recommendations for maintenance and improvement strategies.

BUSINESS UNDERSTANDING

The stakeholders involved in this project include NGOs focused on water infrastructure management, governmental agencies responsible for water resource allocation, and communities relying on well water for daily needs. The dataset chosen contains valuable information about water points in Tanzania, including pump type, installation year, geographic location, and water quality indicators. This dataset was selected due to its relevance to the real-world challenges faced in providing clean water to the population.

PURPOSE

Forecasting Water Pump Conditions?

Anticipating Water Pump Failures?

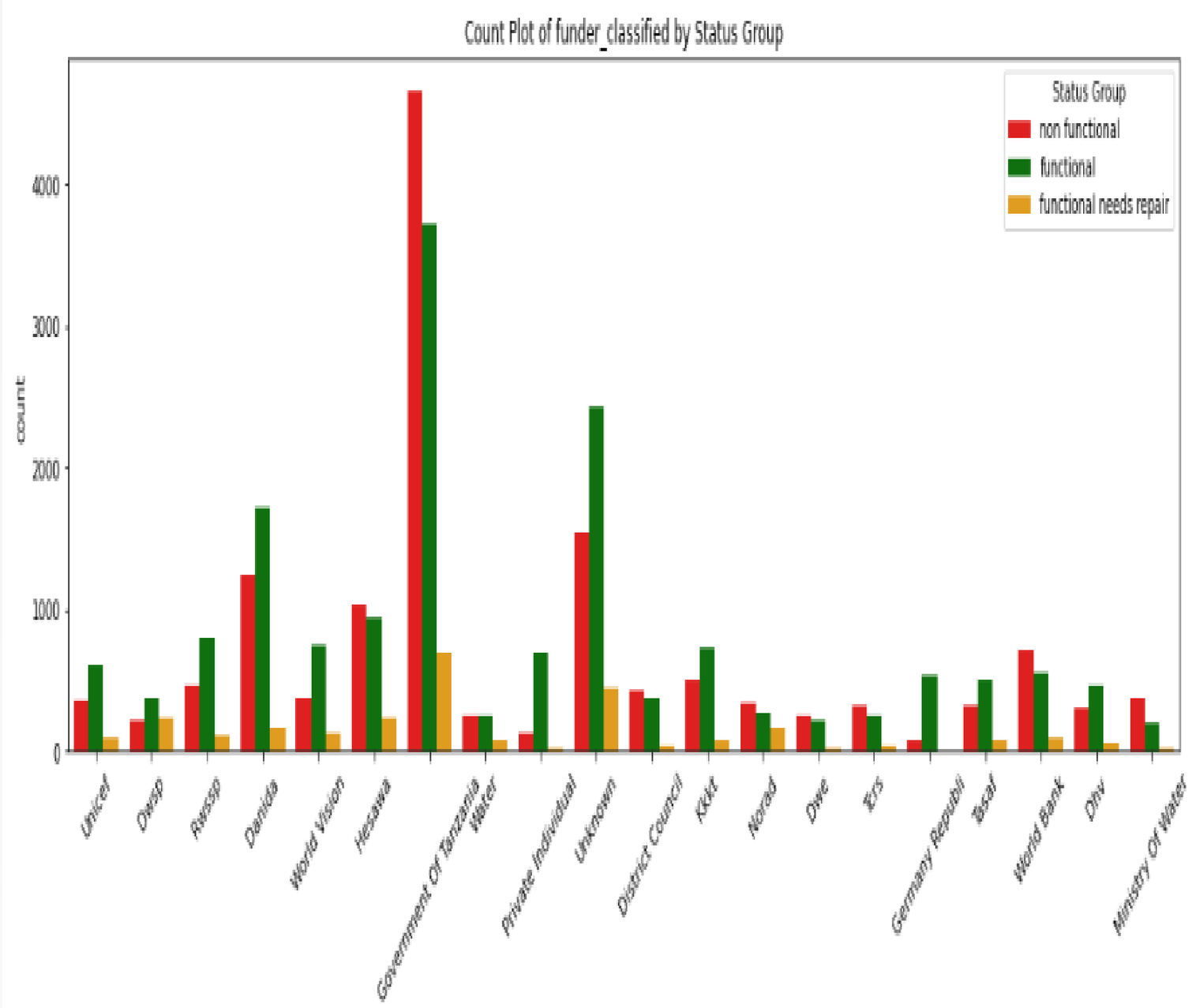
Predicting Water Pump Status?

DESCRIPTIVE ANALYSIS

In this section, we delve into key insights derived from our descriptive analysis of water well data. Through comprehensive visualizations and statistical summaries, we uncover crucial factors influencing water well conditions. Our exploration aims to provide valuable insights to guide strategic decisions for maintaining and improving water infrastructure in Tanzania.

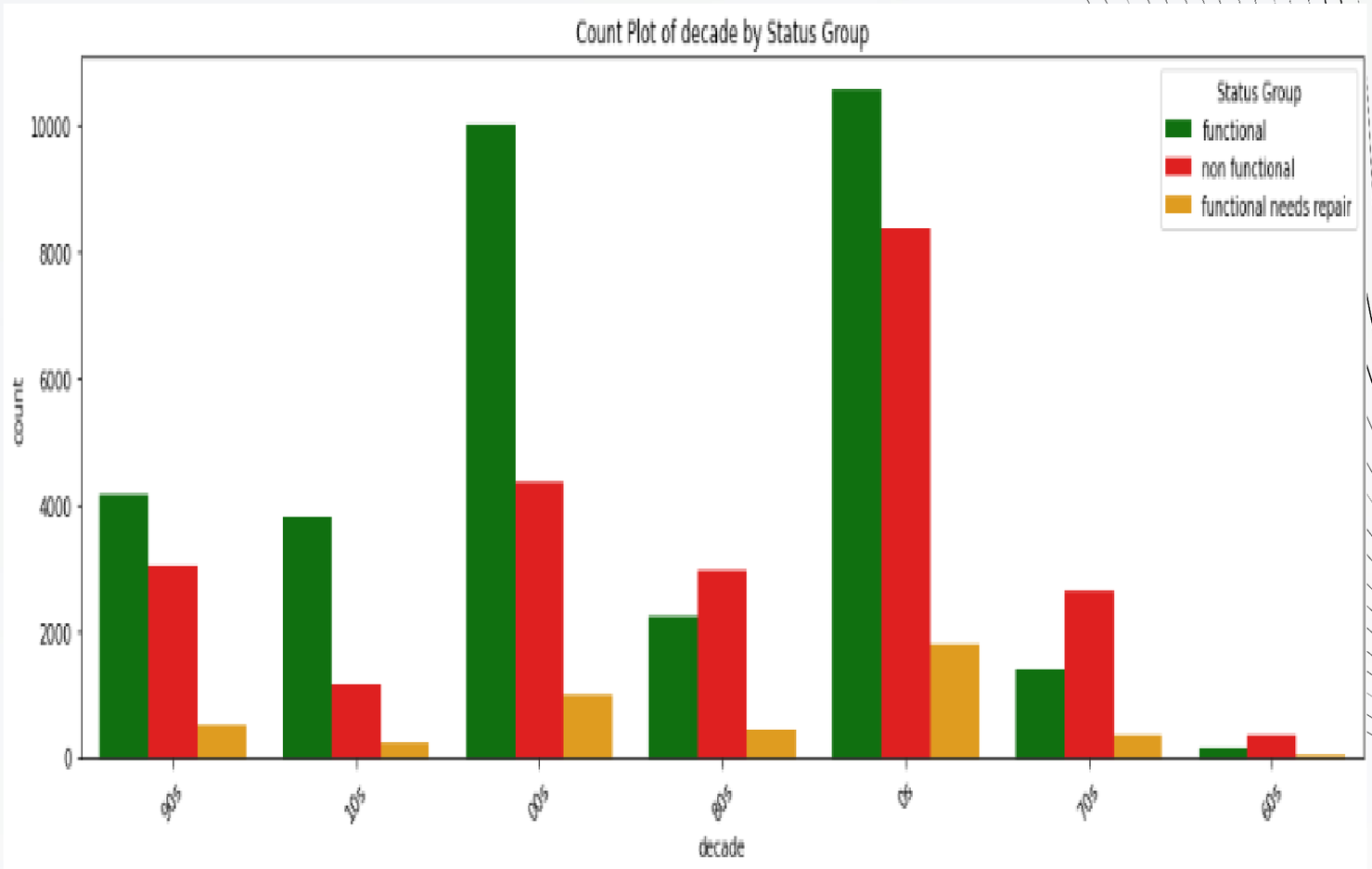
Funding

Government-funded wells often exhibit a higher likelihood of being non-functional, highlighting a need for improved oversight or maintenance practices in these projects



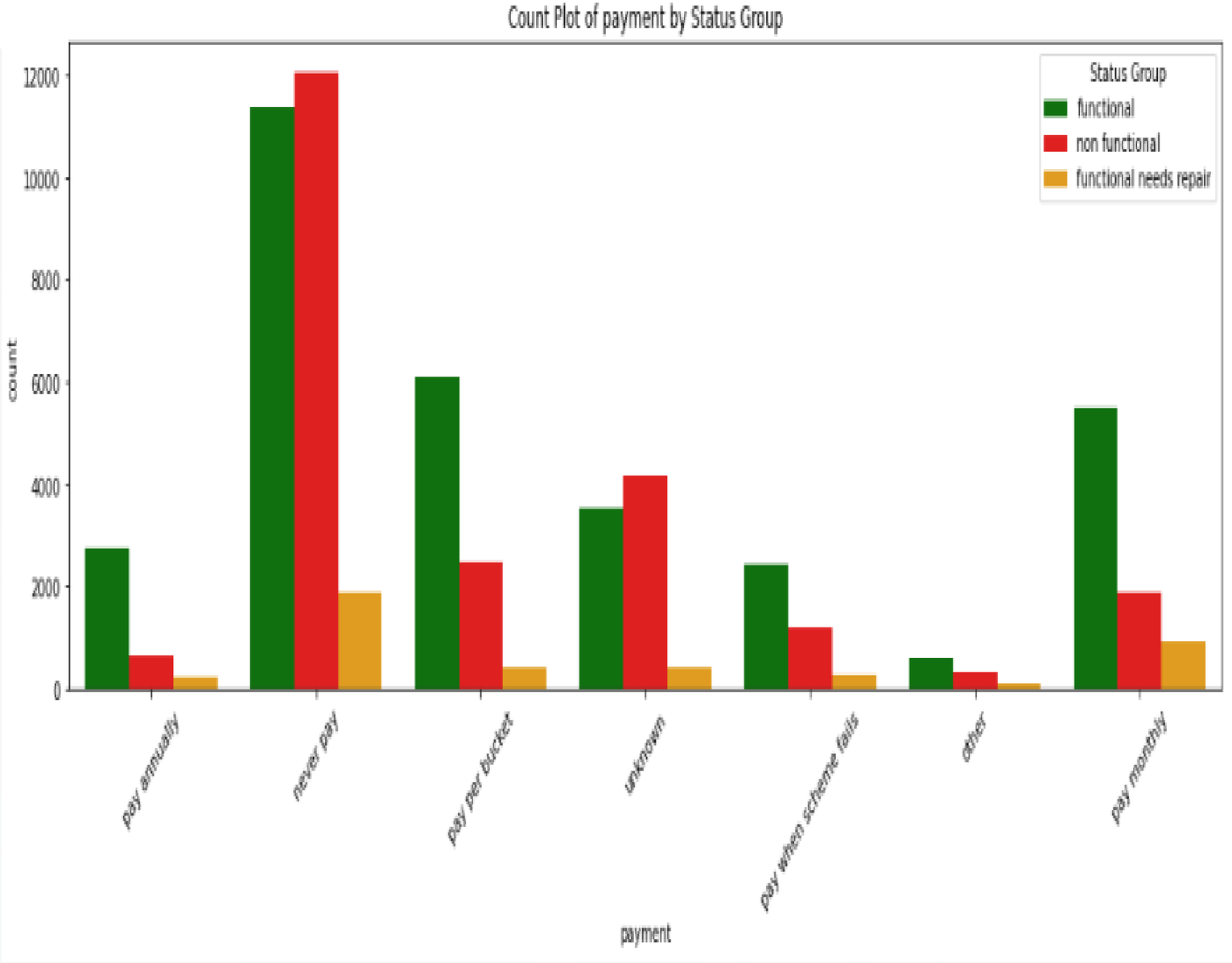
Year of Installation

Recent years have seen a higher proportion of functional wells compared to older ones, but there are still functional wells that require repair. This underscores the importance of timely maintenance to prevent functional wells from deteriorating into non-functional ones.



PAYMENT

The water wells that are paid for are majorly functional compared with water wells that are not paid for



MODELING

Data Preprocessing

- Handling missing values
- Encoding categorical variables (Target Encoder)
- Scaling numerical features (Robust Scaler)

Model Selection

A Random Forest Classifier was chosen as the predictive model due to its ability to handle complex relationships in the data and provide insights into feature importance

Hyperparameter Tuning

- Parameters tuned: n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap
- Best parameters identified through grid search and cross-validation.

EVALUATION

Model Performance Metrics

- Accuracy: Train Accuracy (0.9025), Test Accuracy (0.7870)
- Balanced Accuracy: Train Balanced Accuracy (0.9310), Test Balanced Accuracy (0.7161)

Model Selection

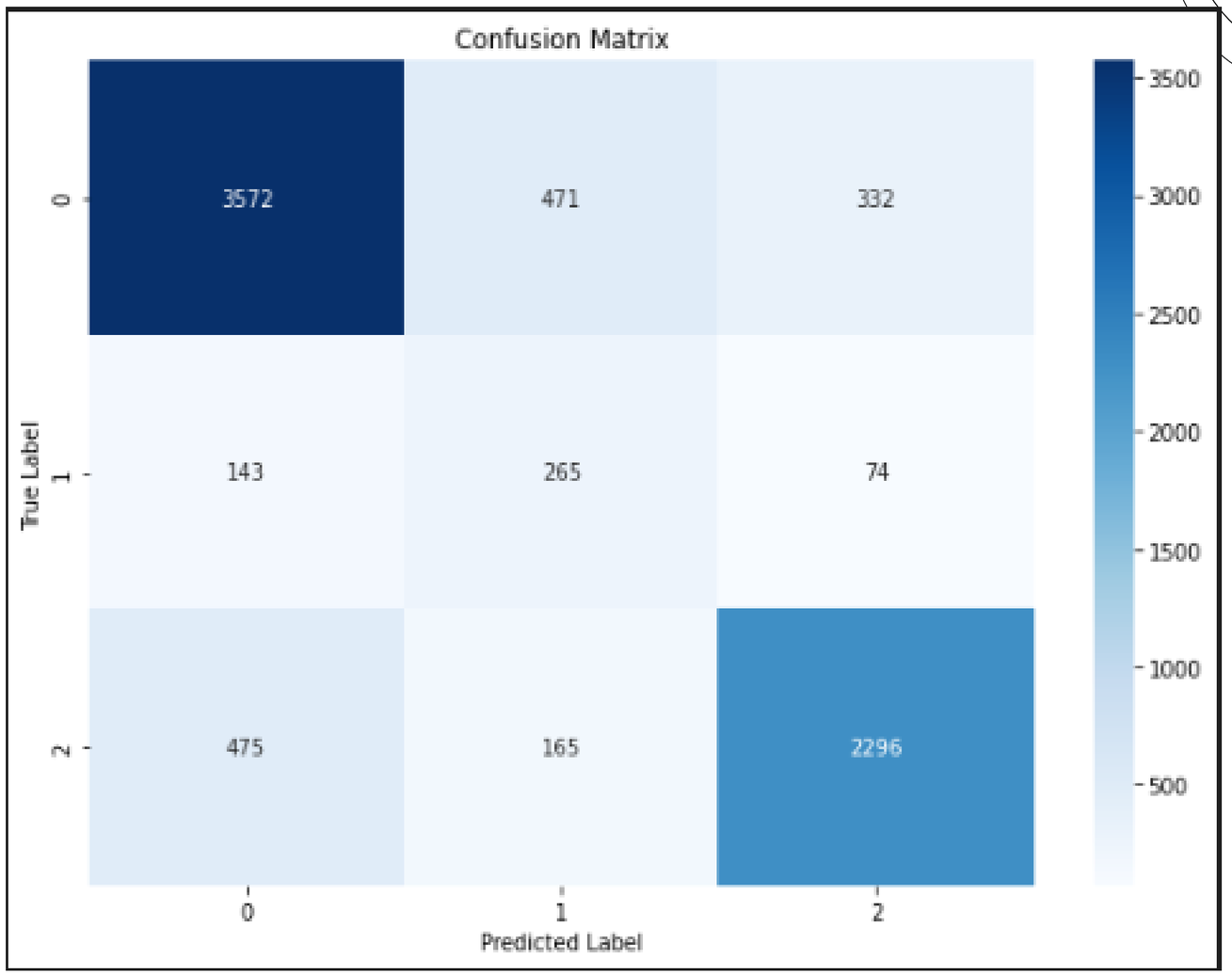
A Random Forest Classifier was chosen as the predictive model due to its ability to handle complex relationships in the data and provide insights into feature importance

Confusion Matrix

- Visualization of true vs predicted labels
- Interpretation of model performance across different classes

CONFUSION MATRIX DISPAY

- Test set balanced accuracy score of 0.7161.
- Room for improvement in handling class imbalance and overall prediction accuracy.



RECOMMENDATIONS

Targeted Maintenance Approach: Utilize insights from EDA and modeling to implement a targeted maintenance approach.

Optimize Resource Allocation: Leverage EDA findings and feature importance from the modeling phase to optimize resource allocation.

Continuous Monitoring and Feedback Loop: Establish a continuous monitoring system informed by EDA insights and model predictions

NEXT STEPS

Validation and Deployment of Model: Validate the predictive model using additional datasets or real-time data to ensure its accuracy and reliability. Once validated, deploy the model for ongoing monitoring and prediction of water well conditions.

Actionable Insights Implementation: Implement actionable insights derived from the EDA analysis, such as prioritizing maintenance in high-population areas, improving water quality monitoring, and investing in pumping infrastructure.

Optimize Resource Allocation: Leverage EDA findings and feature importance from the modeling phase to optimize resource allocation

CONCLUSION

In conclusion, this project demonstrates the potential of machine learning in addressing water infrastructure challenges in Tanzania. By predicting water well conditions and providing actionable insights, stakeholders can make informed decisions to improve clean water access and ensure sustainable water management practices. Further enhancements and collaborations with domain experts can lead to even greater impact and positive outcomes for communities relying on well water sources.

THANK YOU