WATER WELLS CONDITION PREDICTION IN TANZANIA

Data Science Project

Predicting the status of wells for better maintenance prioritization

PROBLEM OVERVIEW

The objective of this analysis is to predict the condition of water wells in Tanzania. The goal is to classify each well as:

- Functional

- Functional Needs Repair

- Non-Functional (Requires urgent repair)

This classification is critical for prioritizing maintenance interventions and improving access to safe water.

DATASET OVERVIEW

The dataset contains information about:

- Pump types	- Installation year	- Geographical information	- Maintenance records

The target variable is `status_group`, which contains the categories:

- Functional	- Functional Needs Repair	- Non-Functional

MODEL SELECTION PROCESS

We tested several models to predict well conditions:

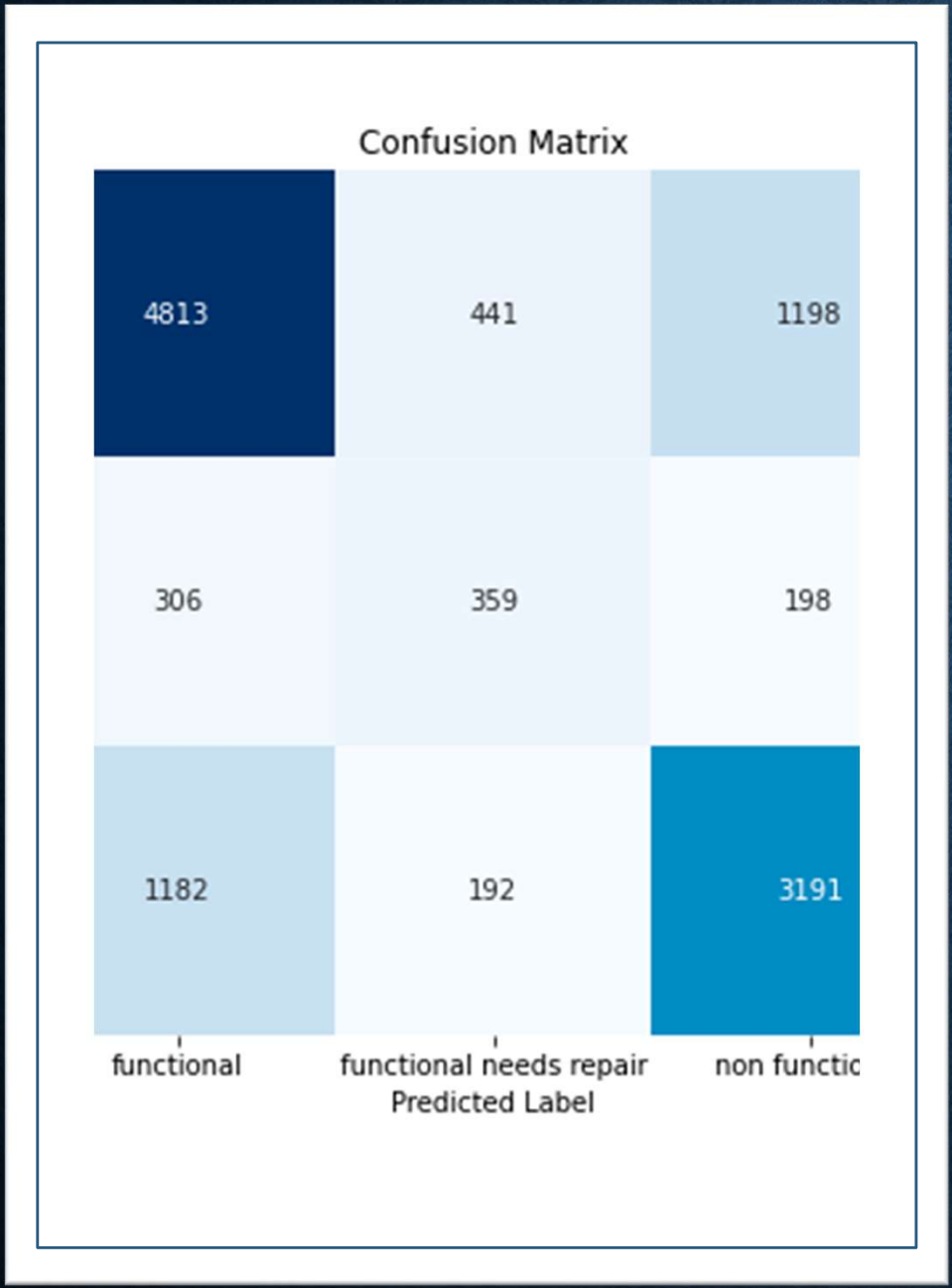
Logistic Regression (baseline model)	**Random Forest** (tuned model)

The Random Forest model outperformed the Logistic Regression model in terms of **accuracy** and **macro-F1 score.**

MODEL PERFORMANCE COMPARISON

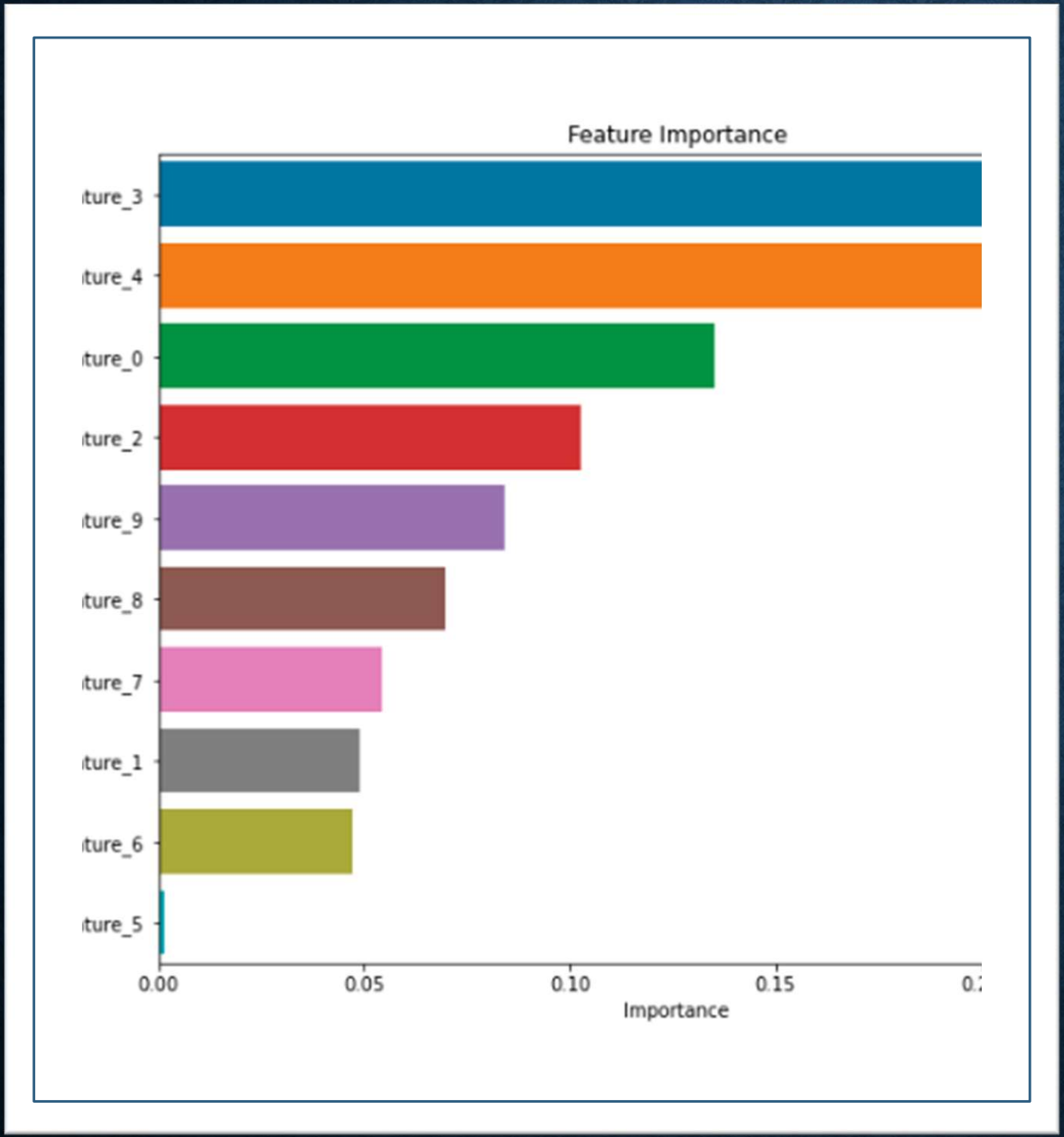
Random Forest achieved better performance across key metrics:

- **Accuracy** : 70%
- **Macro-F1** : 0.613
- **Recall** for Non-Functional Class : 70%



CONFUSION MATRIX

The confusion matrix visualizes how well the model is classifying each category (Functional, Functional Needs Repair, Non-Functional).

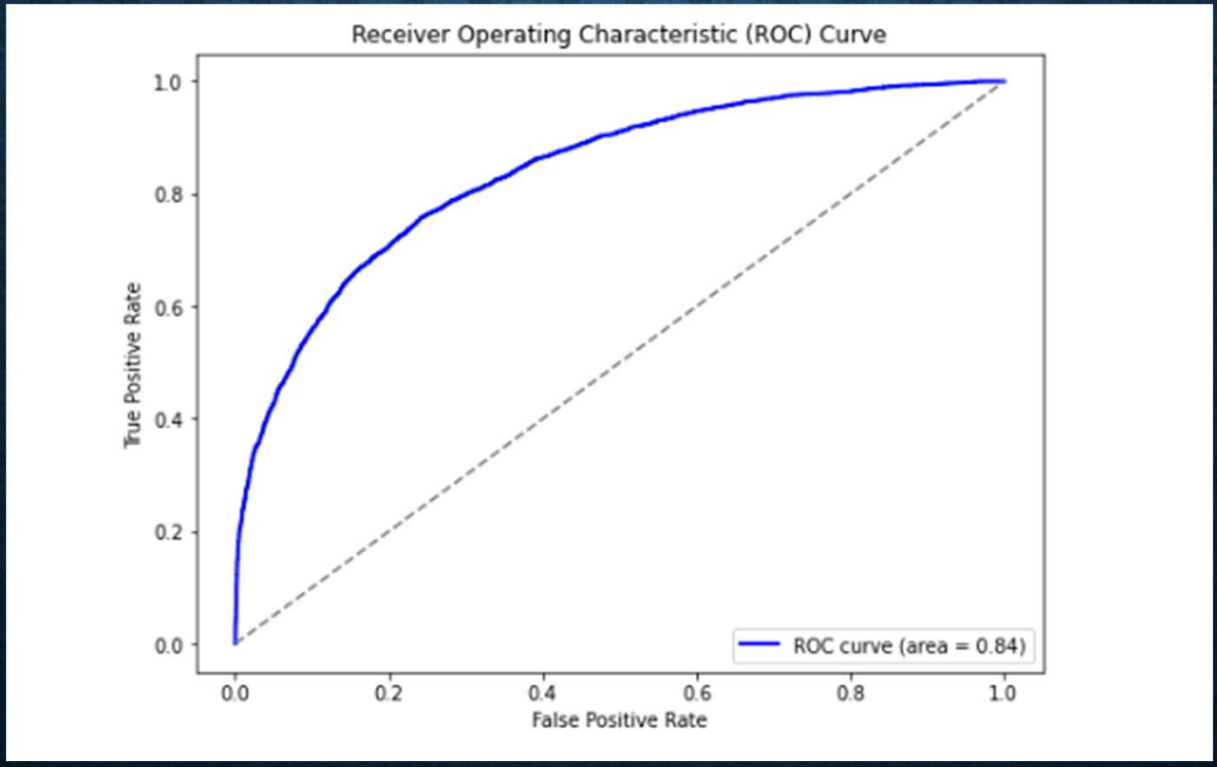


FEATURE IMPORTANCE

Feature importance shows which features the model considers most important when making decisions.

ROC CURVE

The ROC Curve helps evaluate the model's ability to distinguish between the classes, especially for the non-functional class.



CONCLUSION

The Random Forest model is the best for predicting the condition of wells, especially identifying non-functional wells.

This model will help NGOs and local authorities prioritize maintenance tasks, ensuring broken wells are repaired quickly.

Next steps:

- Deploy the model for real-time predictions.
- Continuously update the model with new data.