PREDICTING THE CONDITION OF TANZANIA WATER WELLS

A Machine Learning Approach

PRESENTED BY

Gloria Oseko

AGENDA

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling and Evaluation
5. External Validation
6. Conclusion and Recommendations

1. BUSINESS UNDERSTANDING

PROBLEM STATEMENT

Lack of clean and potable water is a major issue in communities across Tanzania. The Tanzania Ministry of Water has installed several water wells. Not all of these wells are functioning as intended, which results in a lack of access to clean water for communities.

The goal of this project is to use data from Taarifa and the Tanzania Ministry of Water to predict which pumps are functional, which need some repairs, and which don't work at all.

PROJECT GOAL

- The goal of this project is to build a predictive model that can accurately predict the condition of water wells in Tanzania based on the variables provided in the data.

- The aim is to improve maintenance operations and ensure that clean and portable water is available to communities across Tanzania.

OBJECTIVES

MAIN OBJECTIVE:

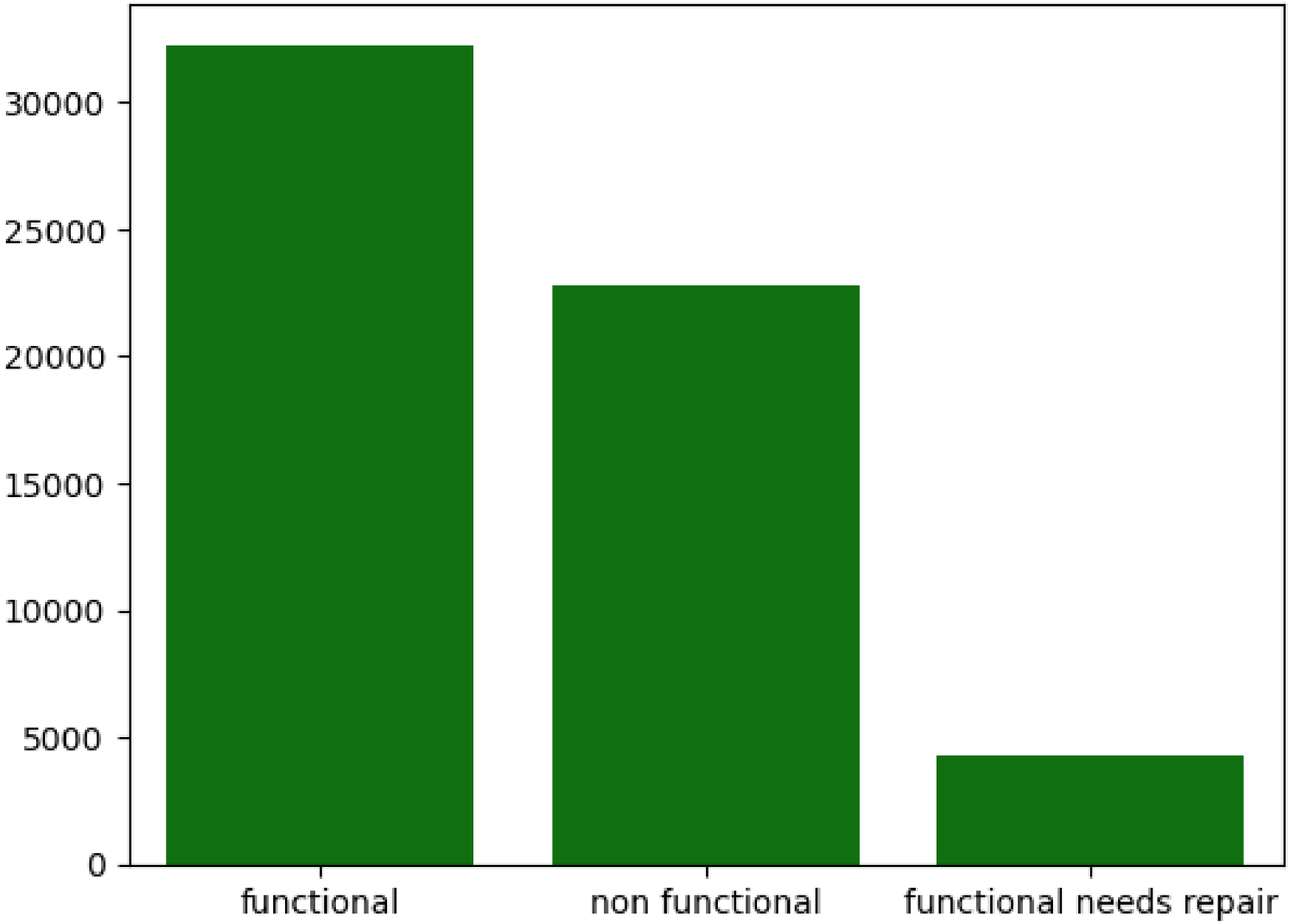
To predict the condition of water wells in Tanzania to ensure that clean and portable water is available to communities across Tanzania.

SPECIFIC OBJECTIVES:

1. To understand the problem statement and the goal of the project
2. To identify the variables that can impact the functionality of water wells
3. To determine the target variable (functional, need repairs, or non-functional)

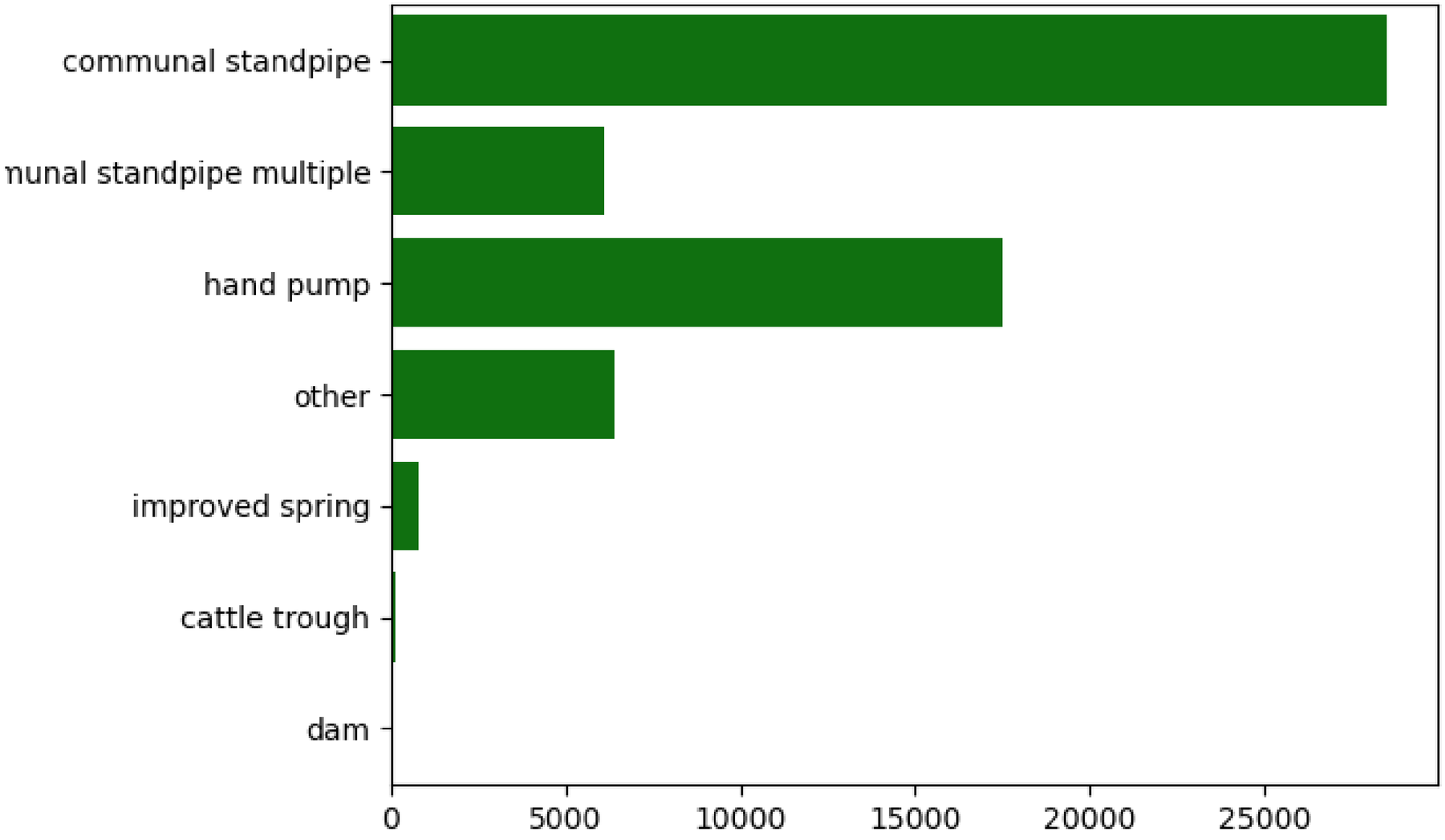
2. ANALYSIS

STATUS GROUP



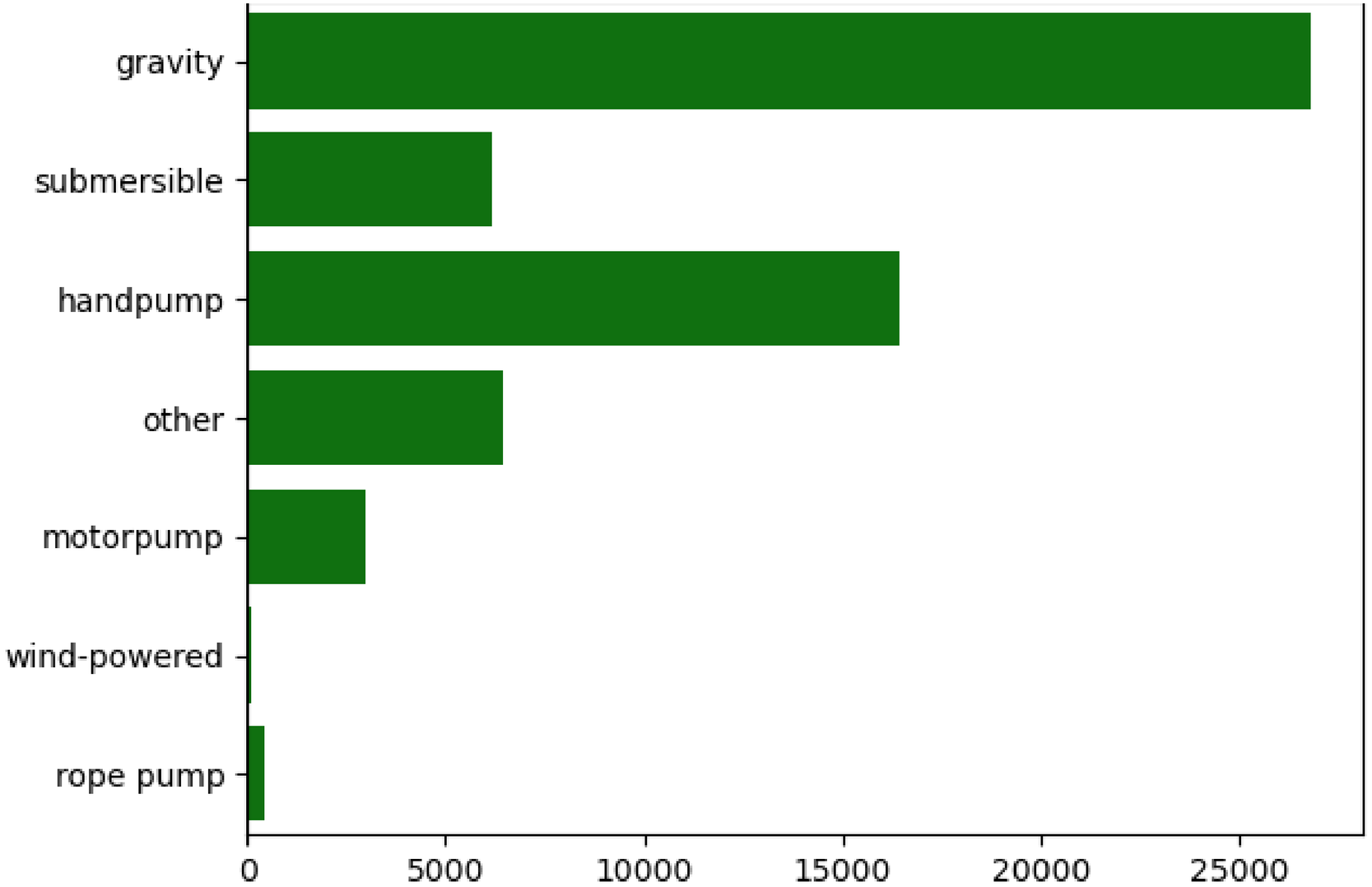
THE MAJORITY CLASS IS THE FUNCTIONAL CLASS WHILE THE MINORITY IS THE FUNCTIONAL NEEDS REPAIR CLASS

WATERPOINT TYPE



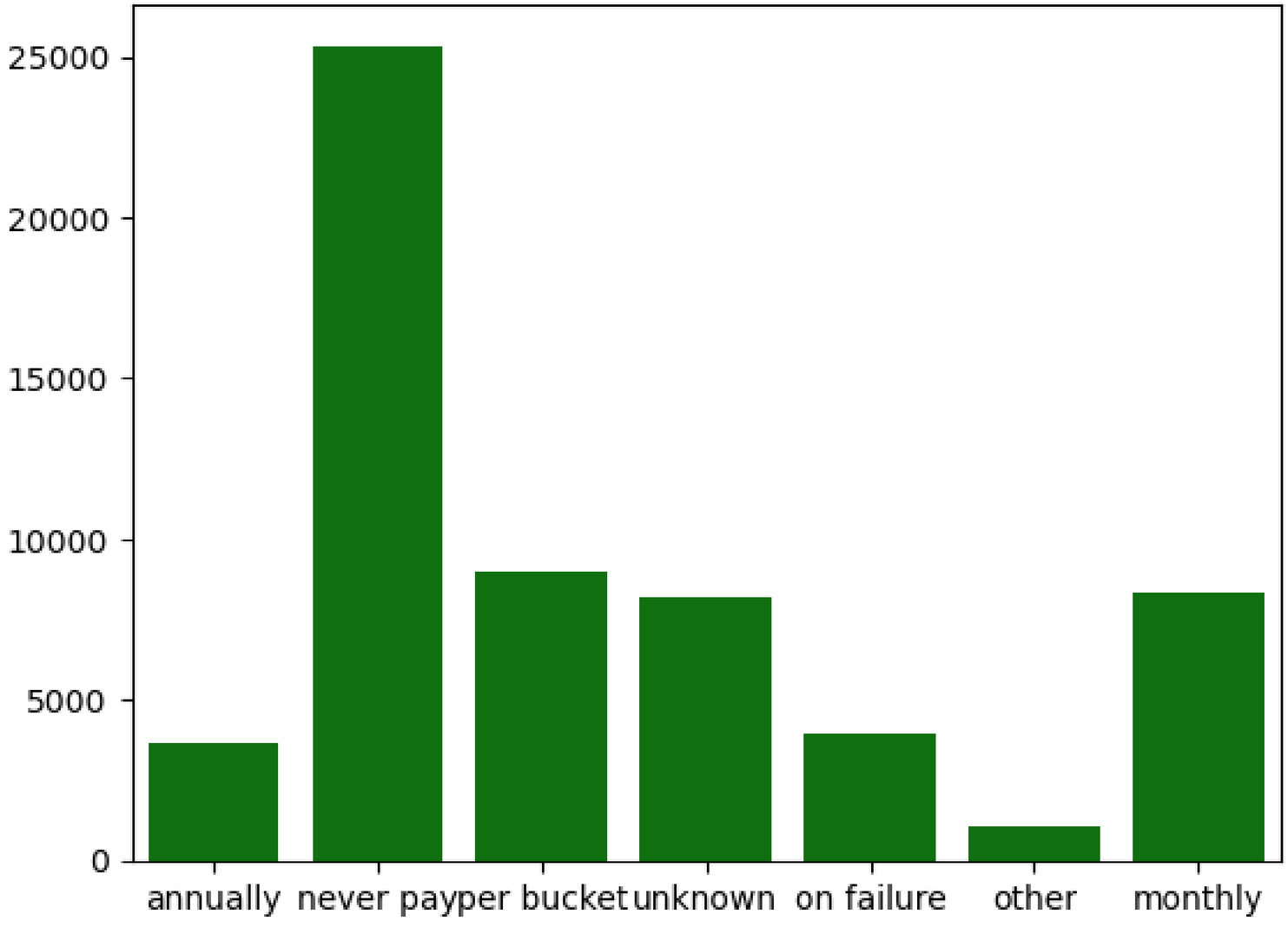
THE TYPE OF WATERPOINT FOR MOST WELLS IS THE COMMUNAL STANDPIPE FOLLOWED BY HAND PUMP

EXTRACTION TYPE



THE EXTRACTION TYPE FOR MOST WELLS IS THROUGH GRAVITY

PAYMENT TYPE



MOST WATERPOINTS ARE NEVER PAID FOR

3. DATA PREPARATION

- Handling Missing Values
- Feature Selection
- Encoding Categorical Variables
- Scaling
- Handling Class Imbalance

4. MODELLING AND EVALUATION

MODELING AND EVALUATION

- THE BEST MODEL (RANDOM FOREST) WAS SELECTED FROM FIVE CLASSIFIER MODELS
- THIS MODEL WAS TUNED TO IMPROVE ITS PERFOMANCE
- THE ACCURACY OF THE MODEL IS 0.8221, WHICH MEANS THAT THE MODEL CORRECTLY PREDICTS THE STATUS GROUP WITH AN ACCURACY OF 82.21%

5. EXTERNAL VALIDATION

EXTERNAL VALIDATION

SUBMISSION OF THE TEST PREDICTIONS MADE BY THE MODEL TO THE "PUMP IT UP: DATA MINING THE WATER TABLE" COMPETITION HOSTED BY DRIVENDATA:

- THE CLASSIFICATION RATE FOR THE SUBMISSION IS 0.8203 WHICH MEANS IT WORKS WELL ON BOTH SEEN AND UNSEEN DATA TO PREDICT THE WATER WELLS CONDITION

6. CONCLUSION AND RECOMMENDATIONS

CONCLUSION

THE MODEL COULD BE FURTHER IMPROVED BY INCORPORATING MORE DATA ESPECIALLY FOR THE FUNCTIONAL NEEDS REPAIR CLASS TO HANDLE IMBALANCE FOR THE CLASSES

RECOMMENDATIONS

- THE TANZANIA MINISTRY OF WATER SHOULD INVEST IN BETTER WATERPOINT TYPES SUCH COMMUNAL STANDPIPES AND HAND PUMPS
- THE TANZANIA MINISTRY OF WATER SHOULD ENSURE THAT THE EXTRACTION TYPE FOR THE WELLS IS MOSTLY THROUGH GRAVITY AND HANDPUMP
- THE TANZANIA MINISTRY OF WATER SHOULD ENSURE THAT THE GPS HEIGHT(ALTITUDE OF THE WELL) FOR MOST WATERPOINTS IS HIGH ENOUGH
- THE TANZANIA MINISTRY OF WATER SHOULD ALSO ENSURE THAT THE PEOPLE USING THE WATERPOINTS PAY EITHER MONTHLY, ANNUALY OR PER BUCKET TO ENSURE THAT THE WELLS ARE WELL MAINTAINED

Thankyou. Any Questions?