TANZANIA WATER WELL CLASSIFICATION PROJECT

Project Done By: Leonard Mwangi Gachimu
October 23, 2023

PROJECT OVERVIEW

To build a Machine Learning classifier algorithm that can predict the condition of a water well (functional, functional-but-needs-repair, and non-functional), using data such as the kind of pump, when it was installed, the installer, the region, and so on.

GROUND WATER SITUATION IN TANZANIA

> Only 61% of households in Tanzania currently have access to a basic water-supply.

> Ground water development has concentrated

> mainly on shallow wells for domestic purposes.

> The average water project in Tanzania about USD 6,000 to USD 8,000

> Up to 90% of pumps and other equipment used for water extraction fail due to a lack of repair and maintenance.

BUSINESS OBJECTIVES

1. To build a Machine Learning classifier that will predict the condition of a water well, using data such as the kind of pump, when it was installed, the installer, the region, etc.
2. To help the Government of Tanzania find patterns in water well pumps.
3. To find the most important factors that influence the condition of a pump.
4. To find out the geographical distribution of the three classes of water pumps.

STUDY QUESTIONS

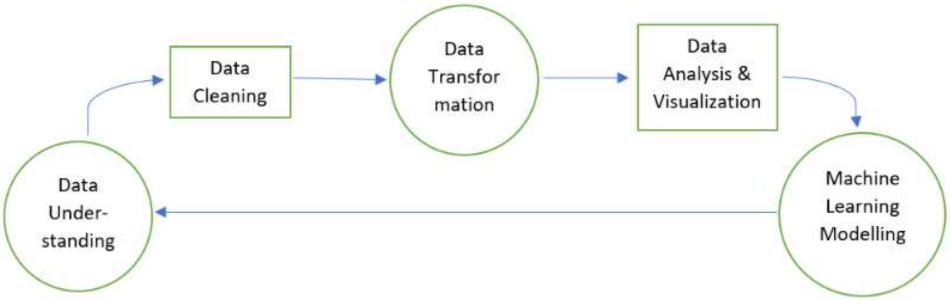
1. Is it possible to correctly predict whether a pump is functional, functional-but-needs-repair, or non-functional given data such as the kind of pump, when it was installed, the installer, the region, and so on?
2. Which are the 20 most important factors that influence pump condition?
3. What is the geographical distribution of pump condition?
4. Is there a relationship between a pump's total static head and its reliability?
5. What is the relationship between the installation year and a pump's functionality?

6. What is the relationship between population and a pump's functionality?

DATA DESCRIPTION

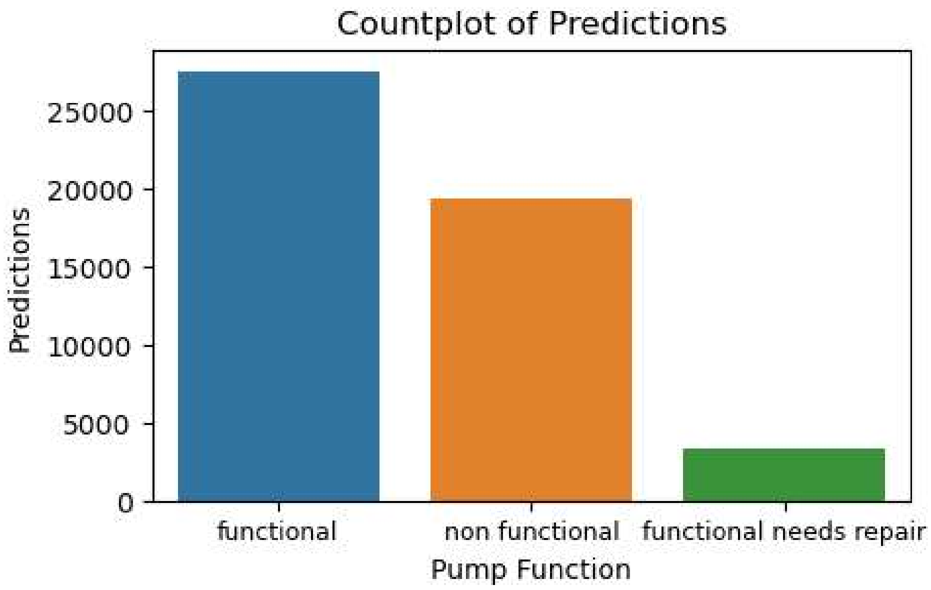
The training dataset contains data about 59,400 water points across Tanzania and 39 features about the waterpoints.

METHODOLOGY



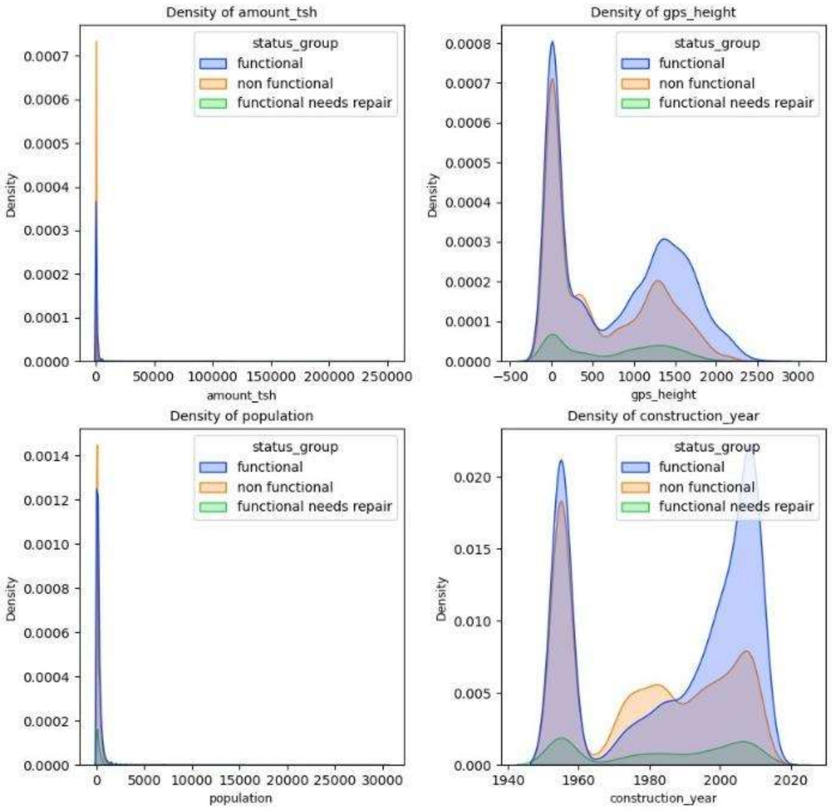
FINDINGS

1. Count Plot of Predictions



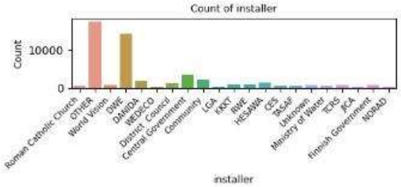
The target classes are not balanced.

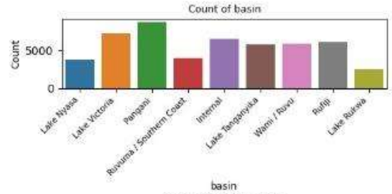
2. Relationship Between Pump Functionality and Continuous Variables

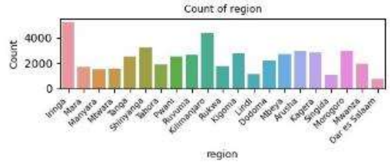


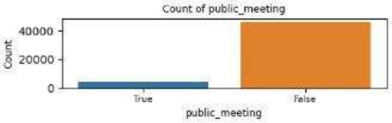
- Non-functional pumps are the most frequent in water points with zero static head.
- Functional pumps are the most frequent in waterpoints at high GPS heights **(1,000 to 2,500)**.
- Non-functional pumps are most frequent in areas with zero population.
- Newer pumps **(1990-2020)** have a higher likelihood of being functional than older pumps.

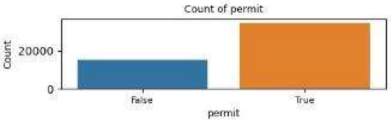
3. Relationship Between Pump Functionality and Categorical Variables

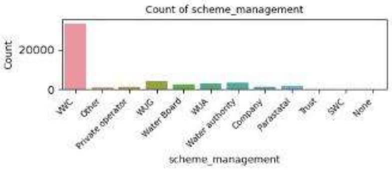


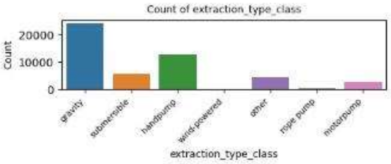


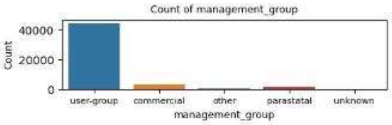


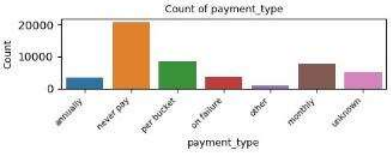


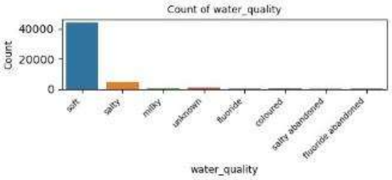


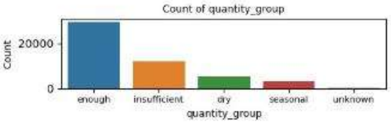


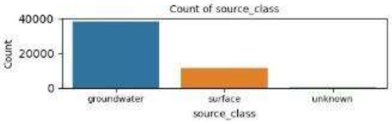


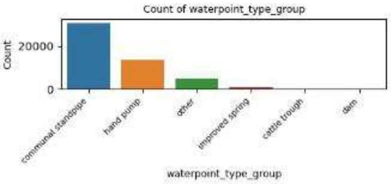








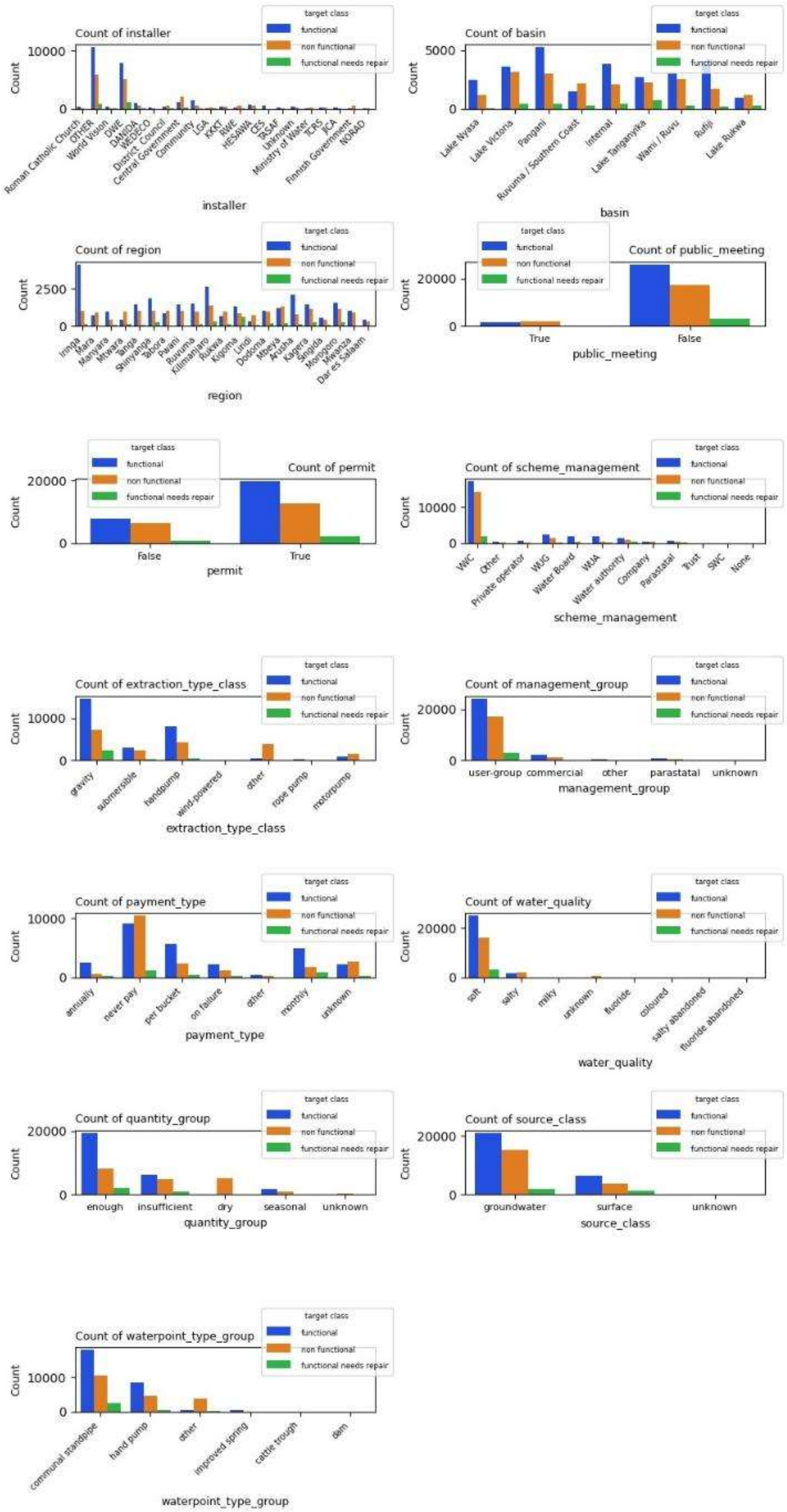




- Some classes of categories are more popular than others. For example,

- Iringa and Kilimanjaro regions have the highest number of pumps.

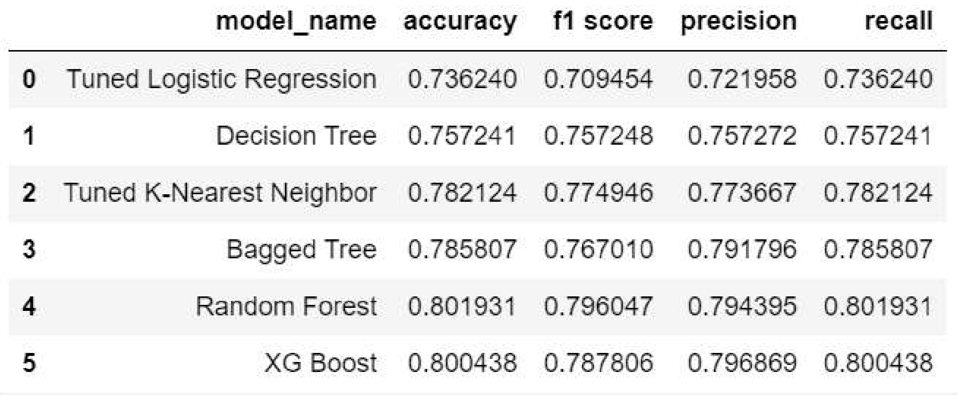
- The never-pay payment scheme is most popular and over **40,000 out of 59,400 wells** have soft water quality.



- We notice that the functional pumps are more frequent than functional-needs-repair and non-functional pumps.

- A notable deviation from this trend is the never-pay class of the payment-type category, where non-functional pumps are more than the other classes of pumps.

4. Predictive Modelling Results



	model_name	accuracy	f1 score	precision	recall
0	Tuned Logistic Regression	0.736240	0.709454	0.721958	0.736240
1	Decision Tree	0.757241	0.757248	0.757272	0.757241
2	Tuned K-Nearest Neighbor	0.782124	0.774946	0.773667	0.782124
3	Bagged Tree	0.785807	0.767010	0.791796	0.785807
4	Random Forest	0.801931	0.796047	0.794395	0.801931
5	XG Boost	0.800438	0.787806	0.796869	0.800438

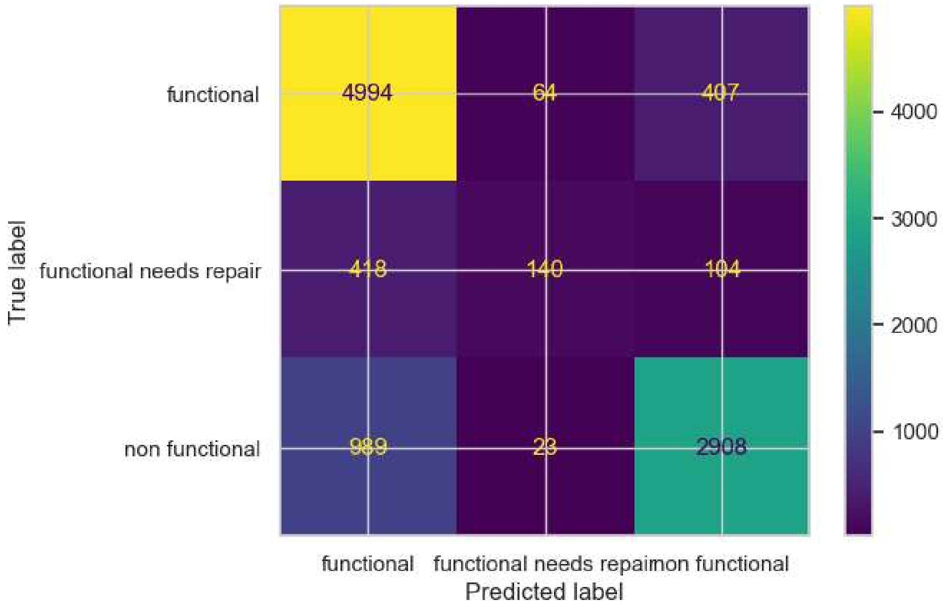
- The XGBoost model is the best for predicting whether a pump is functional, functional-needs-repair or non-functional given the different data features provided. It has an **accuracy score of 0.800 (80%)**, an **F1-score of 0.**788, a **precision of 0.**797, and a **recall of 0.800**.

- Even though the Random Forest model has slightly higher scores, the train scores reveal that it is overfitting the data while the XGBoost model is a good fit (not underfitting or overfitting).

Classification Report

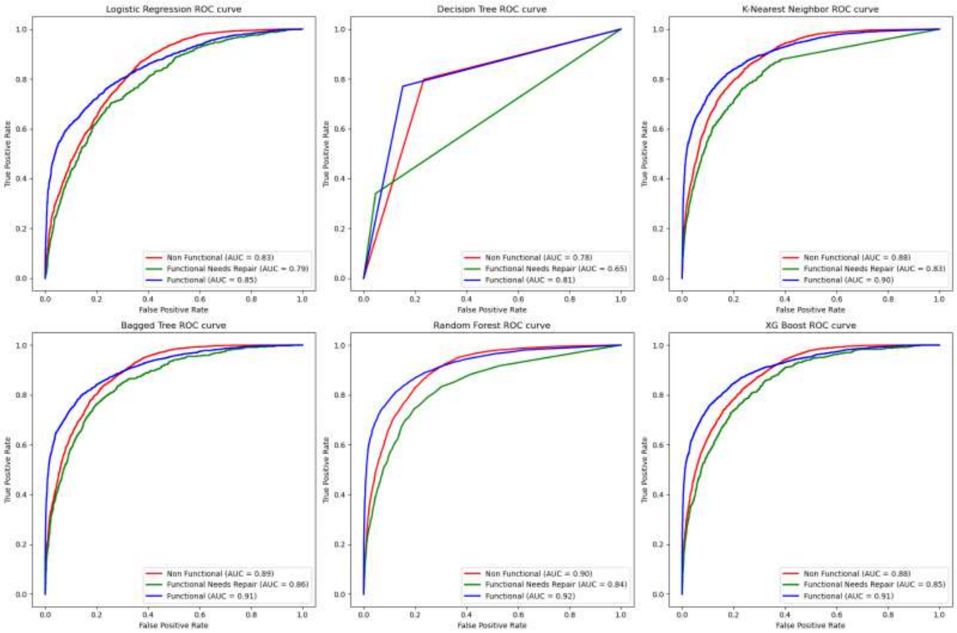
```
                         precision    recall  f1-score   support
            functional       0.78      0.91      0.84      5465
functional needs repair      0.62      0.21      0.31       662
        non functional       0.85      0.74      0.79      3920

              accuracy                           0.80     10047
             macro avg       0.75      0.62      0.65     10047
          weighted avg       0.80      0.80      0.79     10047
```

Confusion Matrix



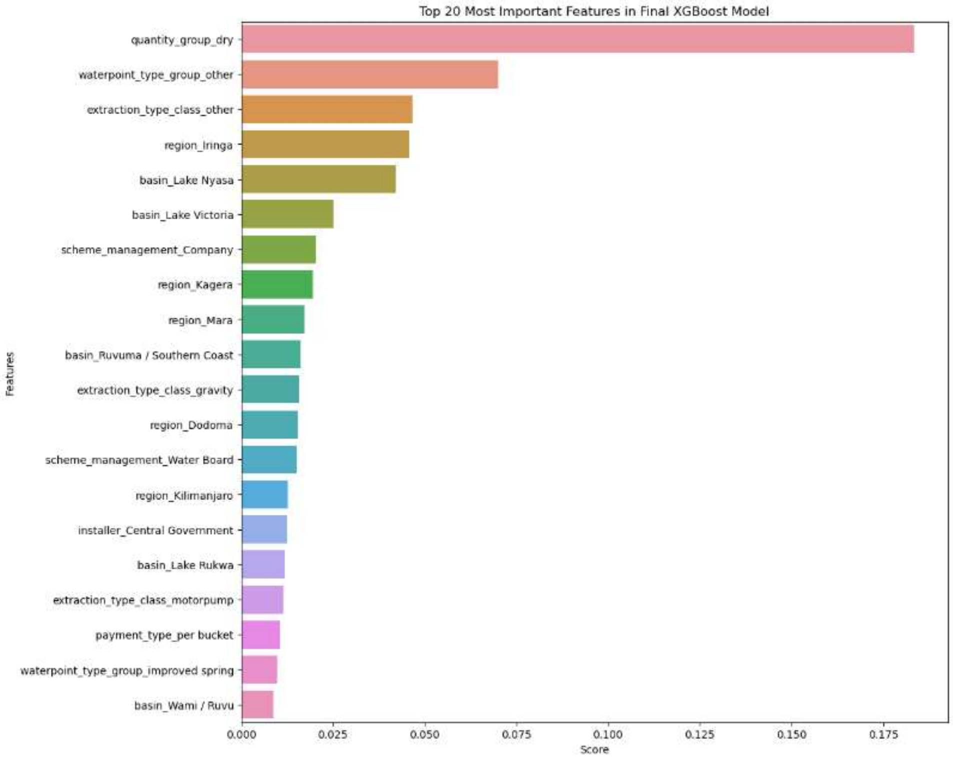
The XGBoost model has high prediction scores for both functional and non-functional pump conditions but low scores for the functional-needs-repair condition. I attribute this to the low count of the functional-needs-repair condition, which makes up only **4.42%** of the dataset.

5. ROC-AUC Analysis



The XGBoost model is the best in terms of performance metrics and goodness of fit, and ROC curves show that it has the best combined AUC scores for the three classes of outcome.

6. Feature Importances



Some of the top features influencing a prediction include:
i.) quantity-group (the quantity of water)
ii.) The water point type
iii.) The extraction type class
iv.) The basin
v.) scheme management
vi.) The installer
vii.) payment type

CONCLUSION

1. Waterpoints with zero static head have the highest density of non-functional pumps but I could not determine the reason for this fact.

2. Pumps having total static head (tsh) above approx. 125,000 are all functional. This could be an indicator that the higher the tsh, the higher the likelihood of a pump lasting longer.
3. Zero population areas have the highest density of non-functional pumps. However, there is no information about whether the wells have been abandoned or the population has relocated.
4. The density of functional pumps is higher among the newest pumps while non-functional pumps are higher among the older pumps, from around **1965 to 1990.**
5. From the geographical map of pumps, we observe the following:
   a.) There is a higher density of functional pumps in the Northern regions of Mwanza and Shinyanga, as well as the Southern region of Njombe.
   b.) There is a higher density of functional-needs-repair pumps in the Northern regions of Bukoba and Arusha, as well as the Western region of Kigoma.
   c.) There is a higher density of non-functional pumps in the Central region of Dodoma and the South West region of Mtwara.
6. Some classes of categorical variables are more frequent than others. For example, Iringa and Kilimanjaro regions have the highest number of pumps. The never-pay payment scheme is the most popular and over **40,000 out of 59,400 wells have soft water quality**.
7. Functional pumps are the most frequent among almost all classes of predictor features but a notable deviation from this trend is the **never-pay class** of the payment-type category, where non-functional pumps are the most frequent.
8. From the predictive section, I conclude that it's possible to correctly **predict at least 80% accuracy**, the condition of a pump given the data features from the Ministry of Water in Tanzania.

   The XGBoost model is the best for predicting whether a pump is functional, functional-needs-repair or non-functional given the different data features provided. It has an **accuracy score of 0.800 (80%)**, an **F1-score of 0.788**, a **precision of 0.797**, and a **recall of 0.800**.

RECOMMENDATIONS TO THE GOVERNMENT OF TANZANIA

1. It should apply my final machine learning model in predicting the condition of water points across Tanzania.
2. It needs to give more attention to the Northern regions of Bukoba and Arusha where there is the highest density of functional-needs-repair pumps, as well as the regions of Dodoma and Mtwara, where there is the highest density of non-functional pumps.
3. It needs to find out more about why there are more non-functional pumps among records with zero static head and in areas recorded as having zero population.
4. It will need to implement and operationalize a payment scheme for the water points, having observed that the sites where people never pay for water had the highest frequency of non-functional pumps.

STUDY LIMITATIONS

1. There is no information about why about half of the dataset has zero population, GPS height, total static head, and construction year values.
2. There is a high number of pumps with zero static head, the majority of which are non-functional. It is not possible to conclude whether the pumps are actually faulty or they are in good condition but the wells have run dry.
3. Tuning the Machine Learning models was computationally intensive and some of them took a couple of hours to execute. Therefore, I could not tune combinations of all possible parameters in due time.
4. The target classes were highly imbalanced and oversampling the minority class did not improve prediction scores.

RECOMMENDATIONS FOR FUTURE RESEARCH AND MACHINE LEARNING MODELLING

1. Finding out more information about the high count of zero values in the population, GPS height, total static head, and construction year columns may lead to better analysis, prediction, and recommendations.
2. Finding out if there is more data that can balance the target classes. The current classes are imbalanced with the most frequent class comprising 37.2% of the data while the least class comprises only 4.42%. This affected the prediction score of the least class compared to the other classes. Availability of more data that can balance the classes would realize much better prediction scores.
3. The use of a more robust computer or cloud server would enable tuning to be done for all parameters of the different classifier models.