Tanzania's Water Pumps

Analysis and Interpretation

by Leticia Drasler

What if you can't have access to clean water?

- In Tanzania most of the access to clean water is through water pumps.
- According to 2015 Tanzania Water Point Mapping Data a significant portion of these water points are non-functional.
- Could we predict which ones are likely to fail?
- Overall goal is to maintain water supply and forecast repair costs by region by having predictions of which pumps will fail.

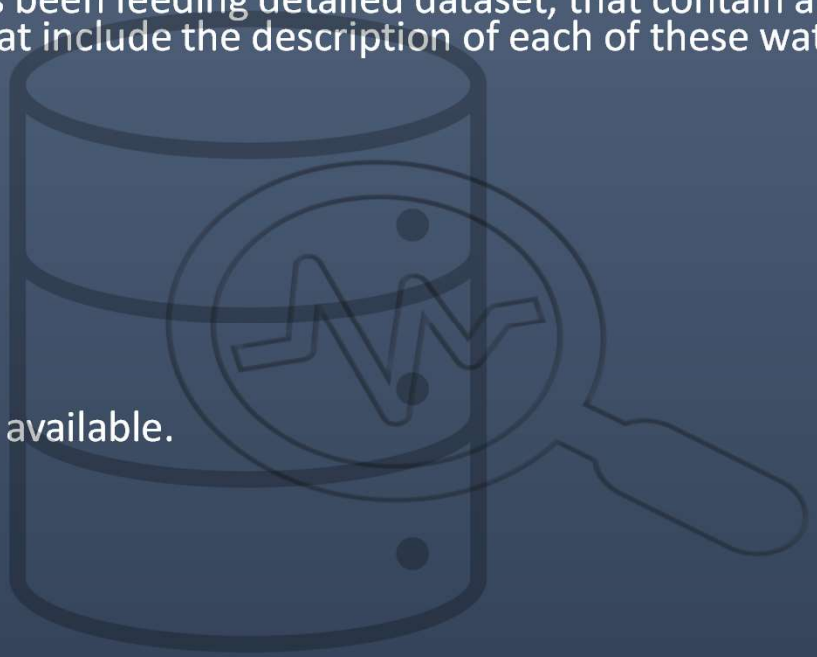
Data

Our goal through this project is to use Machine Learning models to make predictions on the functional status of water pumps with unknown status.

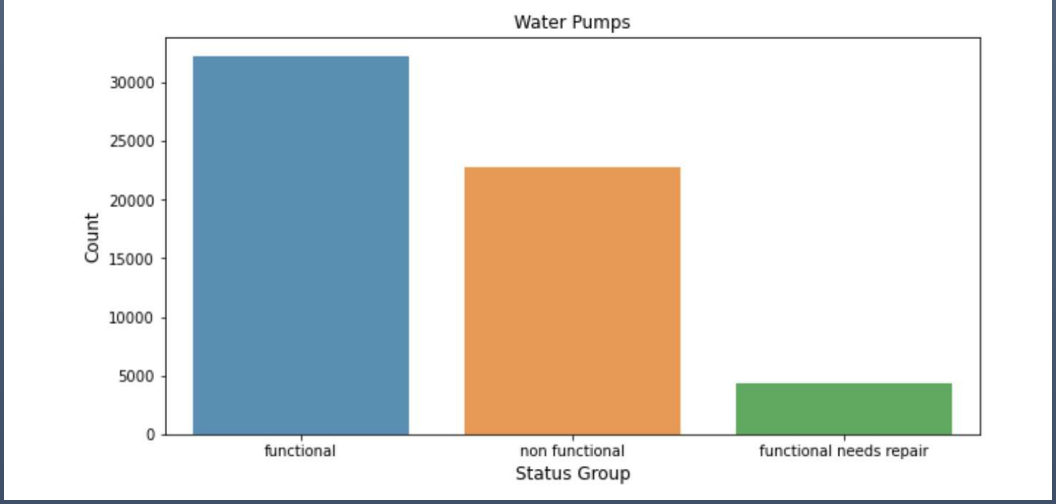
The Tanzania Ministry of Water has been feeding detailed dataset, that contain about 59 thousand values in 39 variables, that include the description of each of these water points, such as:

- Location,
- Funder,
- Installer,
- Etc.

The Data was provided by, and it is available.

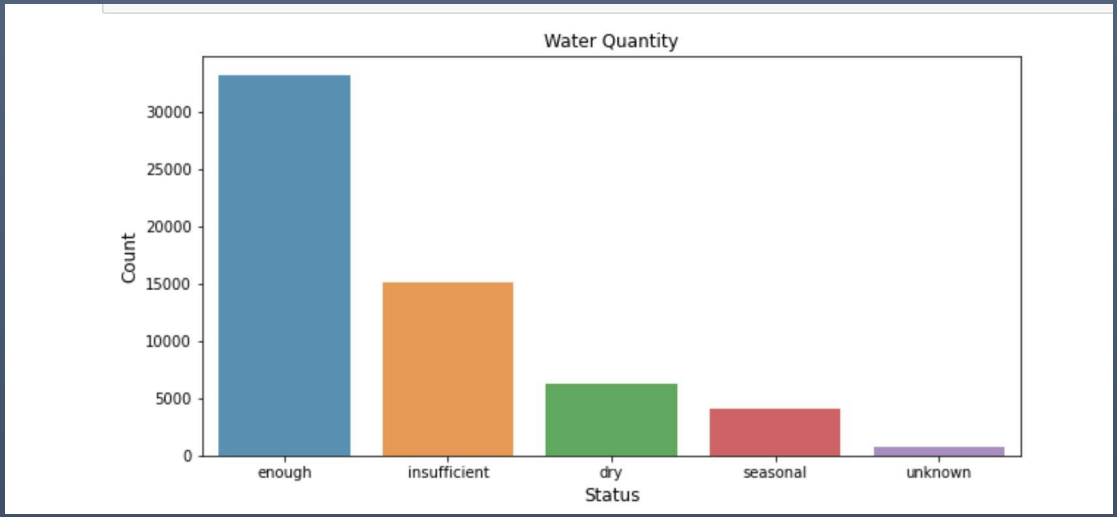


ANALYSIS



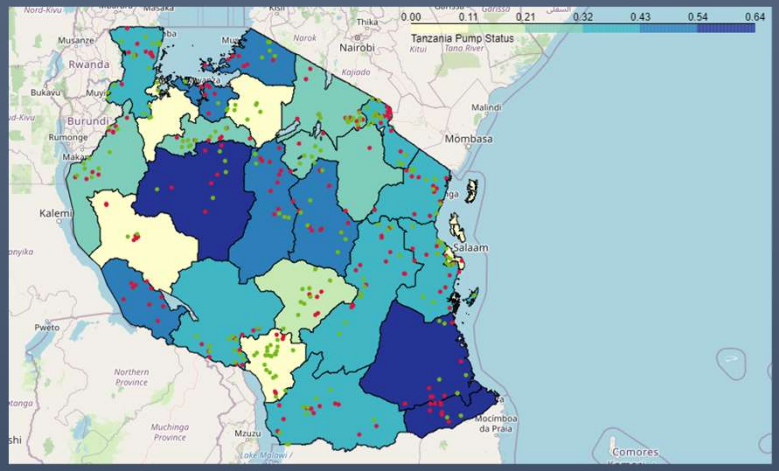
The data show us that almost half of the water pumps are non-functional

ANALYSIS

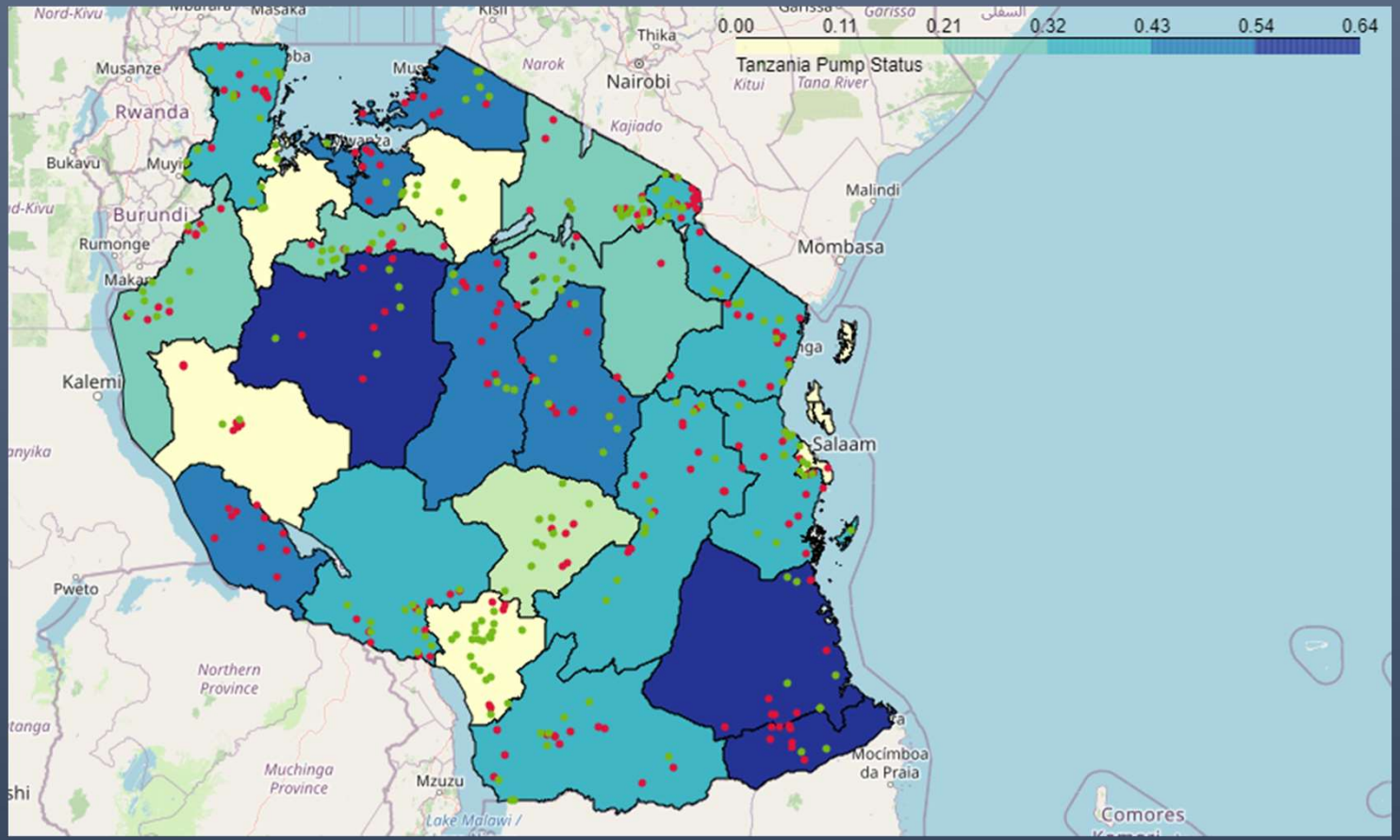


Water insufficient and dry regions also aggravates the functionally of the water pumps.

ANALYSIS



Plotting the regions for a better understanding of the functional status of water pumps


Tanzania Pump Status
0.00
0.11
0.21
0.32
0.43
0.54
0.64
Nairobi
Thika
Narok
Kajiado
Kitui
Tana River
Malindi
Mombasa
Rwanda
Musanze
Bukavu
Burundi
Rumonge
Kalemi
Salaam
Pweto
Northern Province
Muchinga Province
Mzuzu
Lake Malawi
Mocímboa da Praia
Comores

Modeling and Methods

For this analysis I applied multiple models to compare the results

Methods:

- Roc Curve
- Confusion Matrix

Models:

- DecisionTreeClassifier
- RandomForestClassifier
- BaggingClassifier
- Adaboost
- GradientBoost
- XGBoost

Results

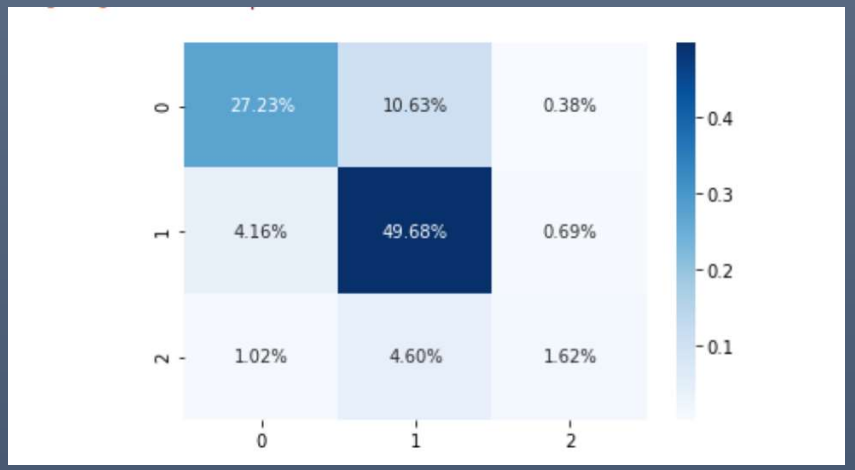
Final Model

XGBoost 0.785%

After all our model seems to work well

79% of the time we can well predict the functional status of unknown water pumps in Tanzania.

Results



Final Model

XGBoost 0.785%

Confusion Matrix illustrates

How well we apply each label to the

Test data.

We can see that we classify 10% of broken pumps as

Functional.

We also classify 4% of functional pumps as non-functional.

Conclusion

The analysis leads us to believe that the label status of the water pumps were best predicted using the XGBOOST algorithm. This algorithm achieved a 79% accuracy when classifying our test data. From these results, we feel confident that these data contain enough information to make meaningful recommendations regarding where to direct repair resources and how to best prevent interruptions in the water supply.

Future Work

- Data completeness and accuracy.
  - Based on our models, we can direct the agency in which parameters we feel to be important, and which are not.
- Data transformation. We will continue to look over our data and determine if other transformations improve our predictions.
- Algorithm – Inside of the machine learning community, there is continual development in the classifier algorithm. We will continue to explore these and apply them to our problem.

THANK YOU!

Leticia Drasler