TANZANIA WATER WELLS

Building a Machine Learning Algorithm to predict the condition of water wells in Tanzania

By Brian Kitainge Kisilu

BUSINESS UNDERSTANDING

The client is an NGO focused on locating water wells that need repair in order to provide access to clean drinking water in rural areas. The client has provided us with a dataset containing information on various water wells, including their location, construction date, and various physical measurements such as depth and water volume. The goal of this project is to build a classifier that can predict the condition of a water well based on this information, in order to help the client prioritize which wells to repair first.

OBJECTIVES

The research study was embodied based on the following objectives:

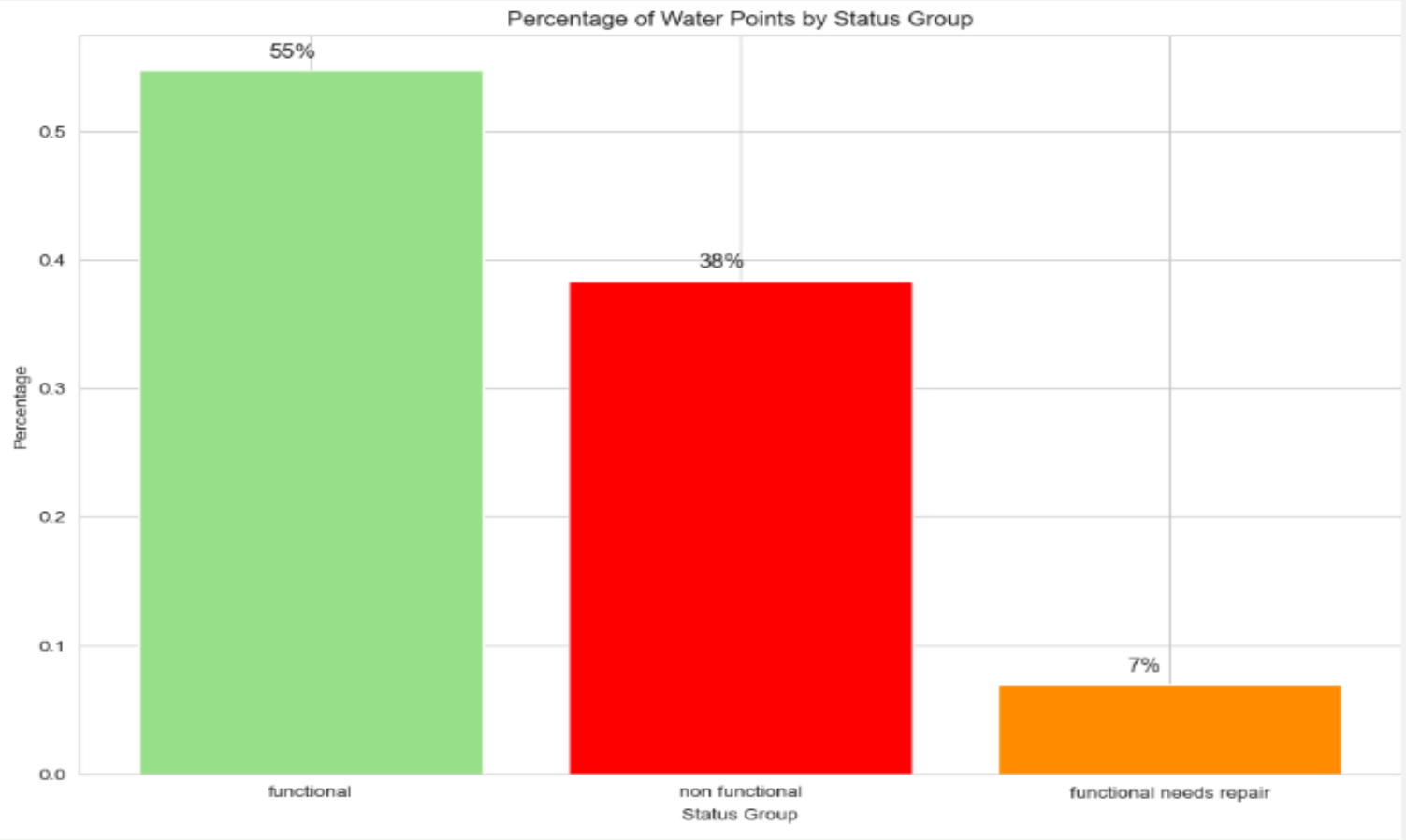
1. To predict the condition of a waterpoint pump based on the geographical location
2. To predict the condition of a waterpoint pump based on age
3. To find patterns in non-functional waterpoint to influence how new waterpoints are built
4. To identify effect of water quality on water pumps

DATA UNDERSTANDING

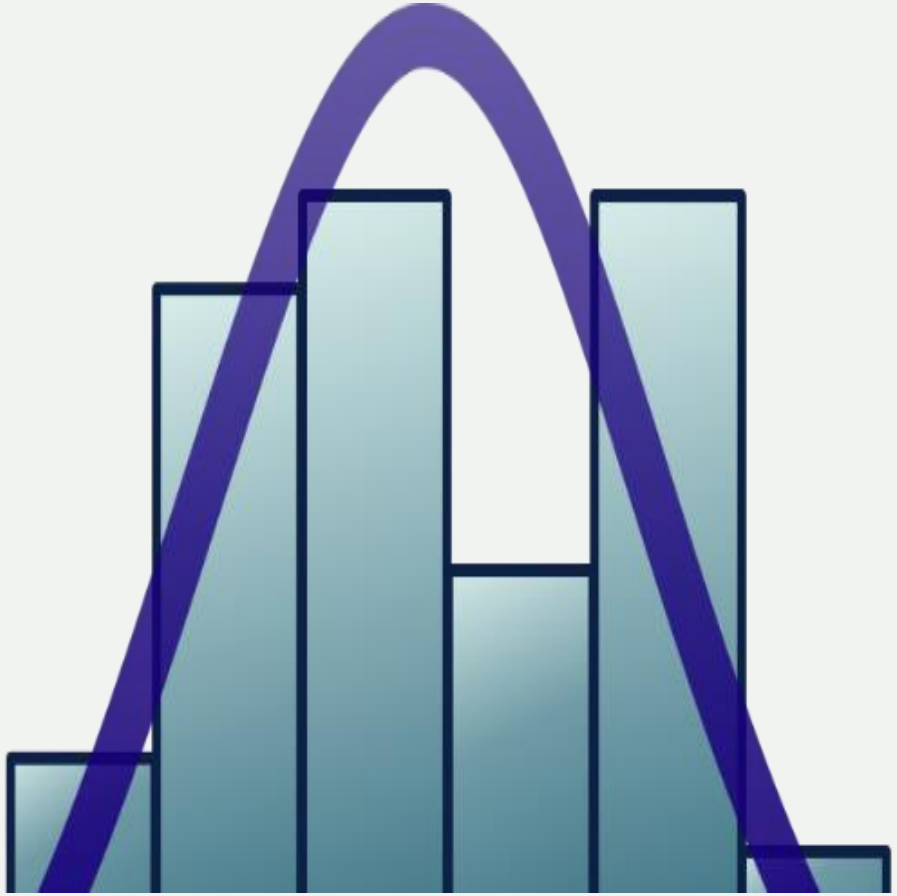
- The data was a record from Tanzanian public records; data set from the Driven Data competition (https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/), originally sourced from the Tanzanian Ministry of Water and supplied by Taarifa.
- The data had 59400 individual data points and 41 columns

WELL DISTRIBUTION

- The following graph shows the distribution of the wells according to functionality.



MODELING



For this study the following models were employed:

1. Decision Tree
2. Logistic Regression
3. K-Nearest Neighbors
4. Random Forest

EVALUATION

For this project the following evaluation metrics were used to rate the performance of the models used :

1. Accuracy
2. Precision
3. Recall
4. F1 Score

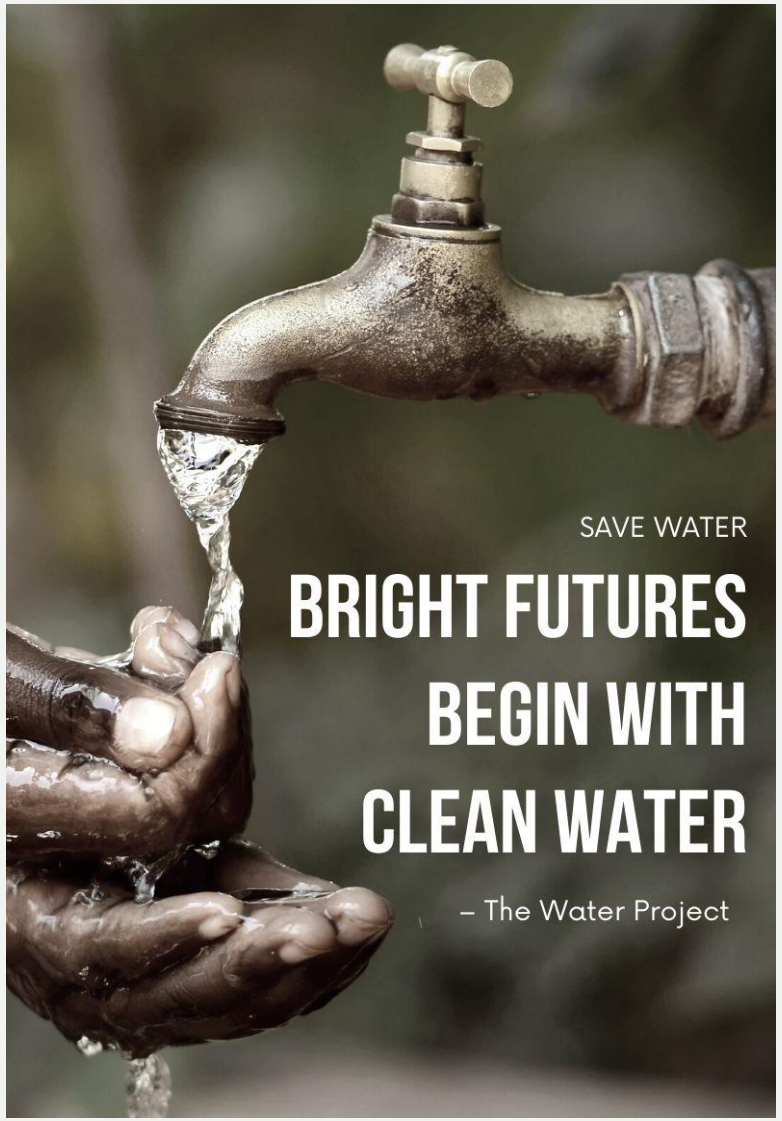
MODEL PERFORMANCE

The table on the right shows the metrics of performance observed for the four models used. It indicates the accuracy, precision, recall and F1 score values .

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Decision Tree	63%	64%	63%	59%
Logistic Regression	59%	56%	59%	53%
KNN	62%	60%	61%	60%
Random Forest	66%	66%	66%	62%

RECOMMENDATIONS

- The random forest model being the best performing model based on the performance metrics should be adopted for use in the project
- The Tanzanian government should prioritize drawing water from springs
- Wells with enough water should be closely monitored and maintained as they are frequently put to use by the population


SAVE WATER
BRIGHT FUTURES
BEGIN WITH
CLEAN WATER
– The Water Project

THE END