Capstone Project 2 Proposal

The question to answer is:

What is the operation status of a water point in Tanzania: functional, needs repair, non functional?

Slightly more than half of the population has access to clean water in Tanzania. The operation and maintenance costs are difficult to bear for local government authorities. Tanzania receives external support from several donor agencies.

The objective of this project is to predict the operation status of water points. This will help reducing operation and maintenance cost, while improving continuity of supply.

This is a DrivenData competition.
The dataset used comes from Taarifa and the Tanzanian Ministry of Water.

The dataset contains records of 59400 water points.
Each record has the following information about the water point:

- `amount_tsh` - Total static head (amount water available to waterpoint)
- `date_recorded` - The date the row was entered
- `funder` - Who funded the well
- `gps_height` - Altitude of the well
- `installer` - Organization that installed the well
- `longitude` - GPS coordinate
- `latitude` - GPS coordinate
- `wpt_name` - Name of the waterpoint if there is one
- `num_private` -

- `basin` - Geographic water basin
- `subvillage` - Geographic location
- `region` - Geographic location
- `region_code` - Geographic location (coded)
- `district_code` - Geographic location (coded)
- `lga` - Geographic location
- `ward` - Geographic location
- `population` - Population around the well
- `public_meeting` - True/False
- `recorded_by` - Group entering this row of data
- `scheme_management` - Who operates the waterpoint
- `scheme_name` - Who operates the waterpoint
- `permit` - If the waterpoint is permitted
- `construction_year` - Year the waterpoint was constructed
- `extraction_type` - The kind of extraction the waterpoint uses
- `extraction_type_group` - The kind of extraction the waterpoint uses
- `extraction_type_class` - The kind of extraction the waterpoint uses
- `management` - How the waterpoint is managed
- `management_group` - How the waterpoint is managed
- `payment` - What the water costs
- `payment_type` - What the water costs
- `water_quality` - The quality of the water
- `quality_group` - The quality of the water
- `quantity` - The quantity of water
- `quantity_group` - The quantity of water

- `source` - The source of the water
- `source_type` - The source of the water
- `source_class` - The source of the water
- `waterpoint_type` - The kind of waterpoint
- `waterpoint_type_group` - The kind of waterpoint

The training dataset is labelled in 3 categories: functional, functional needs repair, non functional.

This project consists of data exploration and multiclass classification.
I plan to use one vs. rest classification technique.

The deliverables will be Jupyter Notebook, a technical report and a presentation of the project.