Capstone 1 proposal

Topic: Pump it Up: Data Mining the Water Table.

Task: Predict which water pumps are faulty? (predict which pumps are functional, which need some repairs, and which don't work at all?)

Using data from Taarifa and the Tanzanian Ministry of Water, Predict one of these three classes based on a number of variables about what kind of pump is operating, when it was installed, and how it is managed. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

Goal is to predict the operating condition of a waterpoint for each record in the dataset. provided the following set of information about the waterpoints:

Features

- 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 - no description
- 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

Labels

- **functional** - the waterpoint is operational and there are no repairs needed
- **functional needs repair** - the waterpoint is operational, but needs repairs
- **nonfunctional** - the waterpoint is not operational

Solution methods:

Planning to apply different classification algorithms like Naïve Bayes, Decision Tree, K Nearest Neighbors, Support Vector Machine and Random Forest on the data set. And will try to understand which algorithms can give best result out of which.