1. **What is the problem you want to solve?**

**Our client wants to predict which water pumps are functional, which need some repairs, and which do not work at all. Our goal is to 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. The client is hoping, by understanding of which water points will fail, to improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.**

1. **Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?**

Our client is the Tanzanian Ministry of Water. The ministry’s vision is to have a nation with clean and safe water for social and economic development. The ministry is working towards achieving its vision by ensuring that water resources are developed and managed sustainably through community and other stakeholders’ involvements in water resources protection for Tanzanians.

Freshwater is a basic natural resource, which sustains life and provides for various social and economic needs. In its natural state, water is an integral part of the environment whose quantity and quality determine how it can be used. Safe drinking water and good sanitation practices are basic considerations for human health. Despite its importance to our lives and development, water is unevenly distributed in time, space, quantity and with great variations in quality. Furthermore, water is a finite and a vulnerable resources.

Therefore, our client is going to use our results to have a better mechanism to monitor water pumps and take actions to **improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.**

1. **What data are you going to use for this? How will you acquire this data?**

The client has shared the data through [www.drivendata.org](http://www.drivendata.org). The shared data is described in the below table:

| **File** | **Description** |
| --- | --- |
| Training set values | The independent variables for the training set |
| Training set labels | The dependent variable (status\_group) for each of the rows in Training set values |
| Test set values | The independent variables that need predictions |
| Submission format | The format for submitting your predictions |

1. **In brief, outline your approach to solving this problem (knowing that this might change later).**

To solve the problem, I will follow the below framework:

1. **Define the problem and quantify the objective**

Reduce the time and resources required to maintain a water pump

1. **Understand the data**
2. **Decide on the machine learning technique**

In order to reduce the time and resources of a water pump, I need to predict the pumps that need immediate attention. This is a classification problem.

1. **Literature Review**

This would help in considering needed or additional variables.

1. **Data Cleaning**

* Missing values
* Duplicates
* Incorrect values

1. **Feature Engineering**

* Removing redundant features
* Transforming features

1. **Modelling**

* Logistic Regression
* Naïve Bayes, etc…

1. **Model Selection**

* Cross validation
* ROC
* Accuracy. Etc…

1. **Ensemble modelling**
2. **Improving best model**
3. **Model deployment (Dashboard, using Tableau)**
4. **What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.**

* Python code
* Report document
* Dashboard
* Presentation (using either Tableau or Power Point)