Jonathan Dyer
Sriman Kolachalam

Team 02 – Proposal

Motivation

The motivation to study this problem comes from two articles that shed light on a critical issue the Tanzanian population faces today: access to clean and safe water. Both water.org (https://water.org/our-impact/tanzania/) and Drop4Drop (https://drop4drop.org/tanzanias-water-crisis/) have studied the Tanzanian water crisis, and report that nearly 25 million people in Tanzania—over 40% of all Tanzanians—lack access to safe water that underlies many basic necessities of life such as drinking water and adequate sanitation.

With this in mind, we look to study one potential area of the water problem in an effort to improve access to water. By analyzing the health of water pumps across Tanzania—a vital source of water for the country—we aim to learn scenarios in which water pumps may malfunction or not function at all, so that we can anticipate future cases of failure. In doing so, we can improve access to water and avoid costly or unnecessary trips by getting ahead of the problem and predicting those pumps that do and do not need maintenance.

Data

Our data comes from DrivenData. It contains a variety of measurements recorded for a given water pump (ex. amount of water available, the location of the water pump, population around the well, construction year, who manages the pump, etc). The full list of features can be found here: https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/

The training dataset includes 59,400 labeled rows, each representing a single water pump, with 40 features listed per row. There are an additional 14,850 unlabeled rows with all of the same features but no categorical label.

The goal of our work is to use the various features of a water pump to learn the possible scenarios under which water pumps may have issues or fail completely, so that in the future we can identify water pumps which may be likely to fail and deal with them before that happens.

Jonathan Dyer
Sriman Kolachalam

Learning Type + Problem Domain

In this project, we will use the data (described below) for a classification task. This is a supervised learning problem wherein we train the model using a variety of attributes describing water pumps in Tanzania, in hopes that we can predict which one of three categories those water pumps fall into: *functional*, *needs repairs*, and *nonfunctional.* We are given the true labels for most of the pumps, which is what enables us to use supervised learning techniques on this problem.

With regards to P(T, E+ΔE) > P(T,E), we have the following:

> **T:** The task is to *classify* any given water pump (based on its features) into a category in [functional, needs repairs, nonfunctional].
>
> **P:** Performance on this task will be measured by classification rate, defined as $\text{Classification Rate} = \frac{1}{N}\sum_{i=0}^{N} I(y_i = \hat{y_i})$, or in other words the percentage of rows where the predicted class $\hat{y}$ matches the true class y. We will seek to maximize this metric.
>
> **E:** Here the experience **E** includes the roughly 60k data points (pumps) for which we have attribute and label information.

As our model trains on more such information **E**, our hope is that the performance **P** on classification task **T** will improve. In other words, as we process additional water pump data, our model will correctly classify a greater percentage of the pumps.

This performance measure of classification rate is identical to another measure: *misclassification rate.* This is the number of incorrect classifications, which is what we are seeking to reduce through our learning process, and is represented simply as `(1 - Classification Rate)`. Overall this will inform us about how well our model is performing, which determines its utility and in turn its business value. If the model performs reliably, then it could lead to significant cost savings and timelier efforts in water pump maintenance, which could ultimately save lives and help improve the well-being of entire villages of people.