**Data-Driven Decision Support for Water Pump Maintenance in Tanzania: A Predictive Modelling and Analysis Study**

Adam Williams (*psyaw3@nottingham.ac.uk*), Filip Kornata (*psyfk2@nottingham.ac.uk*)

**1. Introduction**

**2. Dataset Description**

The provided dataset contains information on 59,400 water pumps in Tanzania. The dataset has 40 features containing a range of information on each pump, such as its geographic location, operator, type, surrounding population, construction year and water quality. For each pump, its current functionality is given in 3 classes: *Functional*, *Non-Functional*, and *Functional (Needing Repair)* – these classes are imbalanced, with 54.3% Functional (*N=32,259*), 38.4% Non-Functional (*N=22,824*), and 7.3% Needing Repair (*N=4,317*).

**3. Research Questions**

The goal of this project is to provide insights which could allow a decision maker to better allocate maintenance resources across Tanzania, and possibly lead to improved pump installation considerations into the future. We have formulated 4 research questions to provide these insights:

1. *What are the main factors associated with pump failure or malfunction, and how do these vary across different regions of Tanzania?*
2. *Which operators and/or management groups have the highest success rates in maintaining water pumps, and how do these rates vary based on factors which may make pump maintenance easier, such as water cost, pump type, or location remoteness?*
3. *What are the interactions between different features, such as water quantity and pump type, which could provide insights into the underlying causes of pump failure?*
4. *How does the age of a water pump relate to its functionality, and is there a point at which pumps become significantly more likely to break down or require replacement?*

**4. Data Pre-Processing**

Due to missing values and outliers in the dataset, data pre-processing steps are required.

There are no duplicate values in the dataset – no identical ID or latitude/longitude values were found.

Throughout the dataset, there are many missing values, both in discrete and continuous columns. These missing values can be dealt with in several ways:

1. *Some columns contain many missing values. These are unlikely to be useful in classification and can be dropped.*
2. *In columns with fewer missing values, data imputation can be used.*

A diagram describing missing values in each feature can be seen below.

**< diagram>**

Normalisation may also be a necessary step for use in our machine learning classifier, as features with a larger range of values could have a greater influence on the output than features with a smaller range. We have found that some features have numerous extreme outliers which would affect this process.

To overcome this, Min-Max normalisation could be used, with outlier values beyond 1.5x the

*d) Feature Engineering*

* Creating new features
  + Water pump age rather than year?
  + Water pump density/remoteness