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

I. Introduction & Dataset Description

Communities across Tanzania are heavily reliant on water pumps for access to clean, potable water, so it is essential that pumps are continually maintained. Data from Taarifa and the Tanzanian Ministry of Water has been made available to facilitate the creation of a system which can predict water pump failure across the country, helping to improve maintenance operations and resource allocation. Our goal is to create a system which provides feature analysis and predictive modelling on this dataset to provide decision support for the ministry and local governments.

The data was collected using crowd-sourcing platform Taarifa, where members of communities can report water pump functionality. The provided dataset contains entries for 74,250 pumps across the country, split into training and testing datasets of *N=59,400* and *N=14,850* respectively. The dataset contains 39 features describing a range of information about each pump, including factors such as geographic location, operator, type, surrounding population, construction year and water quality. For each pump, its current functionality is given in a label split into 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*).

According to Steven's four-level scale, there are 31 nominal features, 5 ratio features, 2 interval features, no ordinal features, and one unclear feature (*num_private*). Many of the numerical values show central tendency towards 0, which is likely related to faulty data input, leading to inaccuracies in standard deviation. This will be further discussed addressed in Section III. The only relatively well-distributed numerical feature is *latitude* with approximate mean -5.7 and standard deviation 2.95. *Basin* and *source_type* are the most equally distributed nominal features in the dataset, while the rest of the features are dominated by one or more prominent values.

II. 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?

III. Data Pre-Processing

Data pre-processing steps are required for further analysis and modelling. There are many missing values throughout the dataset, including missing values in discrete columns and zero-values in continuous columns. These 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, different methods of data imputation can be used.

Diagrams describing missing values in each feature that contains them can be seen in Figure 1.

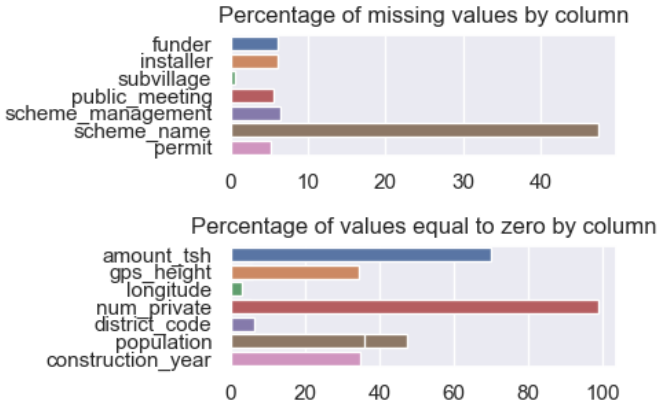

Figure 1. Diagram showing missing values in the dataset by feature.

Additionally, the nominal values will need to be standardized and filtered, as they contain different versions of the same categories, and some contain non-alphanumeric characters.

When dealing with features that have many extreme values, we will either replace or cap the outliers, or discretise the values. On the other hand, if a feature has only a few outliers, they will be removed, or the values will be standardised.

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. Some features will benefit from Min-Max normalization, and other non-linear ones such as *amount_tsh* will require a different solution. Dimensionality reduction methods will be used to extract meaningful insights from the dataset. We will use Principal Component Analysis and Linear Discriminant Analysis to analyse, create and remove features, where necessary.