**Interim Report**

COMP4030 Data Modelling and Analysis

Yichen Lu 20215622 scyyl15@nottingham.ac.uk

Yuhao Zhang 20215559 scyyz12@nottingham.ac.uk

**Title of the paper**

Prediction of the status of waterpoints based on post-construction environmental and management factors

**The dataset and our research questions**

The dataset we choose is “Pump it Up: Data Mining the Water Table”, which describes water point data in Tanzania. The dataset contains 59,400 rows (records of water points) and 41 columns (1 unique id + 39 attributes + 1 label). The 3 possible values for the label (status groups) are: functional, non functional, and functional needs repair. The attributes that describe a water point can be grouped into 3 categories:

1. the condition of the water point itself, including the construction of the water point, e.g. the name of the water point (wpt\_name), the funder and installer of the water point;
2. the environmental factors of the water point, e.g. the geographic location of the water point (region, district\_code, …), the population around the water point, the water quality (quality\_group);
3. the management factors of the water point, e.g. the management scheme (scheme\_management, scheme\_name, management\_group), the cost of the water (payment, payment\_type).

The attributes of this dataset provide a variety of factors that affect the status of water points. Considering that the data describing the condition of the water points themselves (category 1) are not strongly correlated with each other, we decided to explore two other perspectives. Therefore, two of our sub-research questions are to determine the relationship between the water point status and environmental (category 2) and management factors (category 3), respectively. With the insights from these two sub-researches, we aim to accomplish the primary objective of our research, which is to predict the operating condition of waterpoints by taking into account both environmental and management factors.

**Our data wrangling and pre-processing approaches for the dataset**

1. [Irrelevant data - Deletion] Drop irrelevant attributes, e.g. id, funder.
2. [Missing data - Deletion] Drop attributes with too many missing values, e.g. num\_private, amount\_tsh, scheme\_name.
3. [Missing data - Imputation] There is a large amount of missing data in the construction\_year attribute, but our cluster analysis revealed that water points from the same installer have similar years of construction. We fill in the missing records using the average of the year of construction for the same installer. Likewise, we fill in the other missing values in water\_quality, payment\_type.
4. [Data Compression] Subtract the date\_recorded with the construction\_year to form a new attribute use\_duration, indicating how long the water point has been in use since it was built.
5. [Outliers - Deletion] Drop 1812 rows of data with invalid longitude and latitude values (0, -0.00000002), where waterpoints can not exist.
6. [Inconsistent data - Attribute Subset Selection] For geographical location of the water points, there are inconsistent attributes, e.g. 21 regions and 27 different region\_codes; there are excessive and repetitive attributes, such as subvillage, lga and ward. Thus, we decided to retain only region\_code and district\_code attributes for geographical location of water points.
7. [Binning] The data in attribute water\_quality is unbalanced because there are 50,818 soft water records out of 59,400 records. We define the water quality as two conditions, normal and polluted water, and give each condition a weight based on the water\_quality\_group.
8. [Data reduction] Upon inspection, the dataset did not have fully duplicate records, however, there is possible redundant data in 54,000 records. We plan to check whether the smaller representation of the data set will produce the same (or better) quality of analytical results.