[Interim Report] What factors are most important for determining the status of a well, and how accurately can we classify wells based on these features?

Loo Yang Shen Jason*
jason18501@gmail.com

Thomas Cotter*
thomascotter00@gmail.com

March 1, 2023

1 Introduction

This interim report is a short report that introduces the reader into our proposed research question, and the dataset we will be using to answer said question. We will also be discussing the data wrangling & pre-processing approaches for the dataset. We have decided that our research question will be:

What factors are most important for determining the status of a well, and how accurately can we classify wells based on these features?.

We choose this question because we are interested in the factors that determine the status of a well, and using ML to try to classify these wells into 1 of 3 classes: Functional, Non-Functional & Functional Needs Repair. From this question, we can think about some follow-up questions. These could include:

- Which features are most strongly correlated with well status, and how do they interact with each other?
- How does the accuracy of the classification model vary with different feature sets and classification algorithms?
- How do the results of the classification model compare with expert assessments of well status, and what insights can be gained from any discrepancies between the two?

We will be using the dataset from the Tanzanian Ministry of Water, which contains information on the status of wells in Tanzania. We will be using this dataset to answer our research question.

2 Introduction To Dataset

Here we should talk about the current set of features, and some basic statistical analysis: counts, means, medians etc - possibly some graphs if we have space.

3 Data Wrangling & Pre-Processing

Here we should talk about the steps we are ”going to” (we’ve already taken most of them) take in order to clean up the dataset for ML.