MACHINE LEARNING APPROACH FOR THE PREDICTION OF THE STATUS OF TANZANIAN WELLS [COMP4030 CW2 - Data Science and Machine Learning]

*Thomas Cotter
Computer Science
University of Nottingham
Nottingham, England
psytc8@nottingham.ac.uk

*Loo Yang Shen Jason
Computer Science
University of Nottingham
Nottingham, England
hfyyl5@nottingham.ac.uk

***Abstract*—This paper details our approaches and results for the COMP4030 CW2. We have used a number of machine learning algorithms in order to predict the status ('Functional', 'Functional - Needs Repair' & 'Non-Functional') of water wells in Tanzania. This is important as not only does knowing the status of the well help keep the total percentage of functioning wells higher, but it also allows for more effective spending from the government as they no longer would have to send workers out to check the status of the well. Our results suggested that ADD RESULTS!**

***Index Terms*—Machine Learning, Data Science, Classification**

I. Introduction

This section will provide an introduction to the dataset and the research questions we have attempted to solve.

A. Research Questions

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:

- How does the accuracy of the classification model vary with different feature sets and classification algorithms?
- Could we use our results to ensure that wells are built and repaired so that fewer wells are non-functional?

Perhaps we could also include a section on the importance of this research or some more dicussion on the research questions?

B. Dataset

We will be using the dataset from the Tanzanian Ministry of Water, which contains information on the status of wells in Tanzania to answer our research question. This dataset has 59400 rows, with 40 different features. These 40 features could be broken down into three subgroups which hold information regarding: a) Geographic Location of the Wells. b) Management of the wells. and c) Water Condition of the wells. The dataset is originally split into 2 different files, one for labels and one for the actual data. These can be merged easily with pandas through left join on the "ID" column.

We have done a range of statistical analysis on the dataset, including the value counts and graphical representations of each feature. This can be seen here https://github.com/tomcotter7/dma-project/blob/main/Data%20Understanding%20%26%20Statistical%20Analysis.ipynb

Fig. 1 shows the distribution of the target variable, which is the status of the well. We can see from this that we might need to oversample the 'functional needs repair' class.

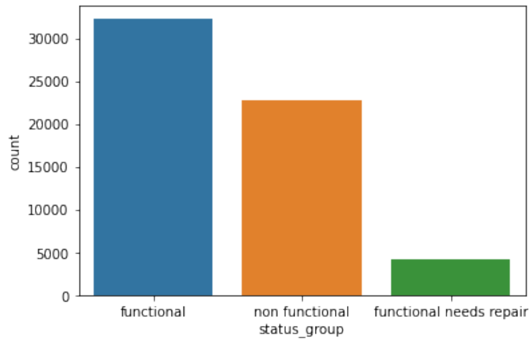

Fig. 1: Distribution of the target variable

As we have more space now, we could include some the results of the statistical analysis here?

II. Literature Review

III. Methodology

IV. Results

A. Data Analysis

B. Data Preprocessing

C. Classification

V. Discussion

VI. Conclusion

References

[1] OpenCV. [Online]. Available: https://opencv.org/. [Accessed: 02-Jan-2023]. **EXAMPLE**