

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
psytic8@nottingham.ac.uk

*Loo Yang Shen Jason
Computer Science
University of Nottingham
Nottingham, England
hfyy15@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 the most important features were ADD RESULTS, and the best classification model as 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 discussion 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>. This will also be discussed further in the Results section.

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.

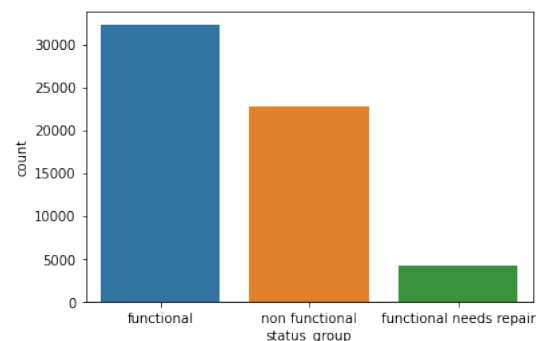


Fig. 1: Distribution of the target variable

II. LITERATURE REVIEW

In this section, we review the relevant literature on predicting machine failure, with a focus on studies that have used similar datasets to Pump It Up. Predicting machine failure is an important area of research, as it has the potential to reduce downtime and maintenance costs. Some machines may be part of a critical infrastructure, so preventing them from failure is of utmost importance.

One study by Pathak et al. (2023) [4] compared the performance of TabNet, a sequential attentive classification architecture designed for tabular data, and tree-based approaches such as XGBoost. They found that TabNet outperformed XGBoost, boasting an 83% accuracy compared to XGBoost's 78%. TabNet makes use of Transformers, a machine learning algorithm which uses self-attention to differentially weight the significance of each part of the input. A point of note is that TabNet does not require feature engineering to perform at these standards. While TabNet is an interesting solution, we have not used it in our report, as our primary goal was to showcase an end-to-end machine learning solution, and this includes feature engineering.

Jabeur et al [2] conducted a detailed review of different algorithms for predicting financial distress, including, SVMs, Neural Networks, RandomForest, XGBoost & CatBoost. They found that the ensemble methods worked the most effectively (RandomForest, XGBoost & CatBoost). Ensemble methods combine multiple models in order to improve accuracy.

Similarly, Mahabub [3] investigated the effectiveness of ensemble methods to detect fake news. They found that combining different classification methods into a Ensemble Voting Classifier produced the best results.

Finally, Celikmih et al [1] used classification models to predict the failure of aircraft equipment and employed the ReliefF feature selection algorithm. This algorithm estimates feature weights iteratively, according to their ability to make a distinction between neighboring models. They used this to select the best subset of features to feed into their classification model, resulting in the best performance.

III. METHODOLOGY

IV. RESULTS

A. Data Analysis

B. Data Preprocessing

C. Classification

V. DISCUSSION

VI. CONCLUSION

REFERENCES

- [1] Kadir Celikmih, Onur Inan, and Harun Uguz. Failure prediction of aircraft equipment using machine learning with a hybrid data preparation method. 2020.
- [2] Sami Ben Jabeur, Cheima Gharib, Salma Mefteh-Wali, and Wissal Ben Arfi. Catboost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166:120658, 2021.
- [3] Atik Mahabub. A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers. *SN Applied Sciences*, 2020.
- [4] Karan Pathak and L Shalini. Pump it up: Predict water pump status using attentive tabular learning, 2023.