PUMP IT UP:

MINING THE WATER TABLE

I. ABSTRACT

This report presents the findings of a comprehensive data mining analysis conducted on the "Pump it Up: Data Mining the Water Table" focusing on the water point data provided by the Tanzanian Ministry of Water. The primary objective of this project was to identify key factors influencing the functionality of water pumps in Tanzania. Through the application of various data mining techniques, including data preprocessing, exploratory data analysis, feature selection, and model development, valuable insights were derived from the dataset. The results revealed significant relationships between specific features, and the functionality status of the water pumps, providing valuable guidance for decision-making in water resource management. The findings from this study contribute to a deeper understanding of the challenges surrounding water pump functionality in Tanzania and offer actionable recommendations for enhancing the maintenance and sustainability of water supply systems.

II. INTRODUCTION

This project is centered around a competition organized by Driven Data® that focused on water pumps in Tanzania, a large country with limited access to clean water. The competition gathered data from the Tanzania Ministry of Water using an open-source platform called Taarifa [2].

The collected data includes various features related to water pumps, such as geographic locations, organizations involved in their construction and management, regional and local government information, extraction and payment methods, and quantities [3]. The goal of the competition is to determine the functionality of water pumps, classifying them as functional, non-functional, or functional but in need of repair.

Objectives The primary objective of this study was to conduct a comprehensive data mining analysis of the water point data available on the data driven website. By applying various data preprocessing, feature engineering, and model development techniques, we aimed to build a predictive model that accurately determines the functionality status of water pumps in Tanzania. Additionally, we sought to identify the most influential features associated with pump functionality and provide actionable recommendations for optimizing maintenance and sustainability efforts.

This document examines the factors associated with water pump functionality. Predictive models were developed to provide an accurate solution, enabling the Ministry to make informed decisions regarding water pumps across different regions in Tanzania. The metric used to evaluate the models' performance is "Accuracy," which measures the precision of the predictions. Data balancing techniques were applied to address any imbalances in the target feature. Finally, several conclusions were drawn from the study.

The project's results were submitted for scoring based on the predictions made. The achieved result was impressive, with a score of 0.8215, slightly lower than the first-place score of 0.8294. Considering the minimal difference between these results, the outcome is considered highly successful.

III. PROPOSED MODEL

- **Model Development and Evaluation**

In this section, we describe the development and evaluation of various machine learning models for predicting the functionality status of water pumps in Tanzania. To ensure

robustness and achieve the best accuracy in the competition, we employed an iterative process where each model was trained multiple times, using different combinations of features and parameters. The following models were selected for experimentation:

- **Random Forest Classifier:** The Random Forest Classifier is an ensemble learning method that combines multiple decision trees to make predictions. We trained the model using different feature subsets and parameter configurations, optimizing for the best accuracy in the competition. The Random Forest Classifier achieved a top competition accuracy of 0.8215.

- **XgBoost Classifier:** The XgBoost Classifier is a gradient boosting algorithm known for its efficiency and accuracy. We experimented with various feature sets and hyperparameter values to maximize the model's performance. The XgBoost Classifier achieved a competition accuracy of 0.8196.

- **CatBoost Classifier:** The CatBoost Classifier is another powerful gradient boosting algorithm that handles categorical features effectively. Multiple iterations were performed with different feature engineering techniques and hyperparameter tuning to enhance the model's accuracy. The CatBoost Classifier achieved a competition accuracy of 0.8125.

- **Bagging Classifier:** The Bagging Classifier is an ensemble method that combines multiple models trained on different subsets of the dataset. We conducted several iterations, varying the feature selection and model combination approaches to improve performance. The Bagging Classifier achieved a competition accuracy of 0.8161.

- **Mixed Voting Classifier:** The Mixed Voting Classifier combines the predictions of multiple models using a voting mechanism. We experimented with different combinations of base models and weighting strategies to optimize accuracy. The Mixed Voting Classifier achieved a competition accuracy of 0.8200.

- **Weighted Voting Classifier:** The Weighted Voting Classifier is a variant of the Mixed Voting Classifier that assigns different weights to each model's predictions. We iteratively adjusted the weights and fine-tuned the feature sets to improve accuracy. The Weighted Voting Classifier achieved a competition accuracy of 0.8202.

- **AdaBoost Classifier:** The AdaBoost Classifier is an ensemble method that combines multiple weak classifiers to create a strong classifier. We performed multiple iterations, varying the choice of weak classifiers and adjusting hyperparameters to achieve the best performance. The AdaBoost Classifier achieved a competition accuracy of 0.8201.

- **Gradient Boosting Classifier:** The Gradient Boosting Classifier is a powerful algorithm that builds an ensemble of weak models in a sequential manner. We conducted several rounds of experimentation, modifying the boosting parameters and exploring different feature combinations to enhance accuracy. The Gradient Boosting Classifier achieved a competition accuracy of 0.7921.

- **Support Vector Machine (SVM):** SVM is a popular supervised learning algorithm used for classification tasks. We conducted multiple experiments, adjusting the kernel function, regularization parameters, and feature engineering techniques to maximize accuracy. The SVM model achieved a competition accuracy of 0.8181.

- **Adaboost X Random Forest:** This model combines the Adaboost algorithm with Random Forest to improve overall performance. We performed extensive iterations, adjusting the boosting parameters and exploring different subsets of features to optimize accuracy. The Adaboost X Random Forest achieved a competition accuracy of 0.8197.

Based on the competition accuracies achieved, the Random Forest Classifier achieved the highest accuracy of 0.8215, closely followed by the Weighted Voting Classifier with an accuracy of 0.8202. The other models also demonstrated competitive performance, ranging from 0.8125 to 0.8201 in terms of accuracy.

Evaluation Metrics: To assess the performance of the models, we utilized competition accuracy as the evaluation metric. The competition accuracy represents the accuracy achieved during the model evaluation phase on the competition dataset.

It is important to note that the iterative approach employed for each model, involving feature selection, hyperparameter tuning, and model combination, helped maximize the accuracy achieved in the competition. The selection of different feature subsets and parameter configurations aimed to optimize model performance and reduce the risk of overfitting.

	Model	Best Competition accuracy
0	Random Forest Classifier	0.8215
1	XgBoost Classifier	0.8196
2	CatBoost Classifier	0.8125
3	Bagging Classifier	0.8161
4	Mixed Voting Classifier	0.8200
5	Weighted Voting Classifier	0.8202
6	AdaBoost Classifier	0.8201
7	Gradient Boosting Classifier	0.7921
8	SVM	0.8181
9	adaboost X Randomforest	0.8197

IV. Experimental Work

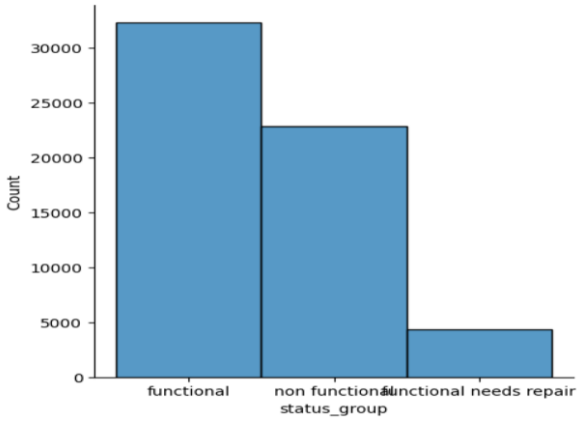
- Description of the Dataset

The dataset was obtained through the utilization of an open-source web application called "Taarifa." For the training set, the data consists of 59,400 rows and 40 columns, excluding the label which is provided in a separate file. As for the testing data, it comprises 14,850 rows. Detailed information about each column can be found in [1]. The target variable is named "status_group" and it can take one of three values: functional, non-functional, or functional but in need of repair.

- **Exploratory Data Analysis (EDA)**

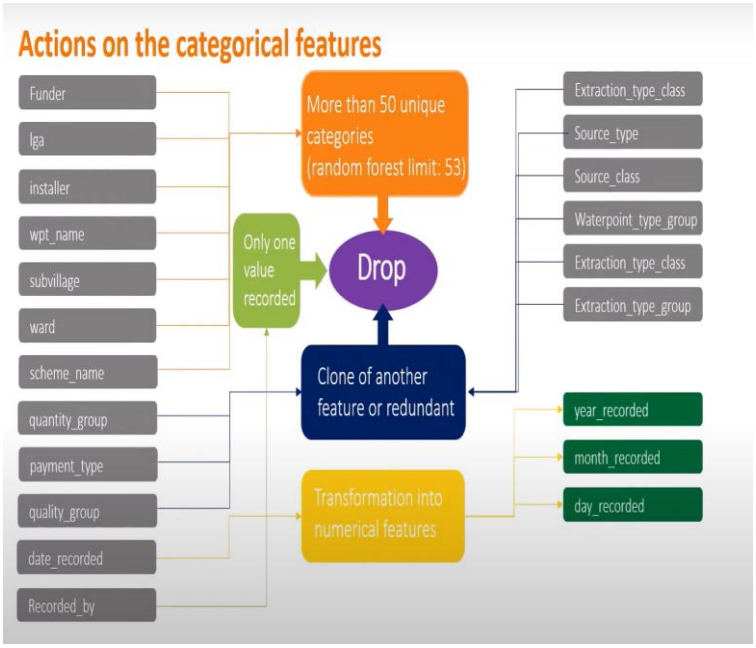
To conduct the exploratory data analysis, the context of the problem and the characteristics of the variables were thoroughly examined. This analysis was conducted within the project's directory, specifically in the "DM.ipynb" file located in the notebooks directory. The target variable in this analysis exhibits three potential outcomes:

1) - Functional
2) - Non-functional
3) - Functional but requiring repair.



- ***Feature Selection and Engineering***

In this section, we discuss the feature selection and engineering process undertaken to optimize the model's performance. The original dataset comprised 40 variables, and to enhance the model's efficiency and reduce redundancy, we narrowed it down to 23 variables. The variables excluded from the analysis included duplicates, features with only one recorded value, and those with over 50 unique categories.



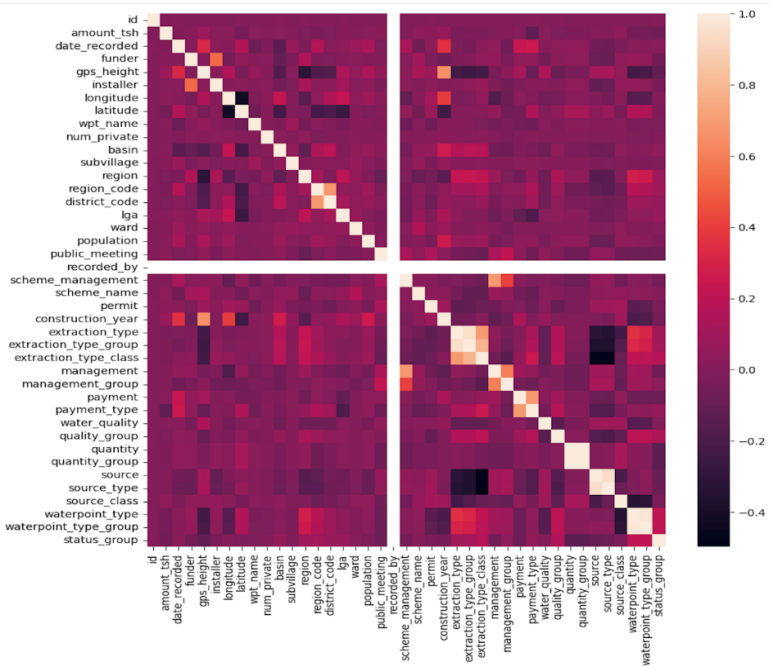
After performing the feature selection, we proceeded with feature engineering to address missing values. For the remaining features, we replaced the zero values with the mean value. This approach was determined to be the most effective after comparing it with alternative methods such as using the mode or median values.

To refine the model further, we initially constructed a model using all available variables. Subsequently, we evaluated the impact on performance by removing duplicate variables and observed how it influenced the model's accuracy. Additionally, the feature importance function was employed to assess the influence of each variable on the model.

The is an example of how I transfer the categorical features into numerical labels:

train.funder = enc.fit_transform(train.funder.values.reshape(-1, 1))

- This line of code applies label encoding to the "funder" variable, which represents the organization or individual who funded the water pump. By transforming the categorical funder values into numerical labels, the variable can be used as an input feature for the machine learning model.



While substantial progress has been made in optimizing the model, there may still be potential for improvement. For example, the installer variable was excluded from the analysis, but it might possess predictive power. It could be beneficial to reduce the number of factors by grouping certain installers together, which could potentially enhance the model's accuracy and predictive capabilities.

V. RELATED WORK

After reviewing the available literature [4][5][6], it was observed that previous approaches to the problem mostly employed different machine learning algorithms. However, by utilizing the random forest model in the context of CS230: Deep Learning, Winter 2018 at Stanford University, CA, groups were able to achieve accuracy levels exceeding 82 percent. Many of these groups also utilized grid search to facilitate the process of tuning hyperparameters.

The notable advantage of this model lies in its relatively straightforward implementation. Using scikit-learn, one can easily set up the model, with most of the work revolving around data cleaning. Additionally, implementing grid search appears to be more straightforward for this basic machine learning algorithm compared to the approach of neural networks. We spent a considerable amount of time constructing the model, mainly focused on finding the optimal combination.

However, a potential weakness of this approach is its lack of complexity compared to a neural network. We speculated that a neural network with multiple layers and nodes could potentially achieve higher accuracy. Furthermore, since all published solutions to this problem seem to converge on using the random forest model.

In previous attempts by other researchers, it was universally acknowledged that cleaning and parsing the given data was the most time-consuming and challenging aspect of the project. Various strategies and ideas were discussed, some of which we implemented in our data preprocessing, while others were deemed impractical. Additionally, we devised some original methods of processing the data, which we believed could positively impact our results.

VI. CONCLUSION & FUTURE WORK

For future research, we have identified three main paths to pursue. Firstly, we propose identifying important features by utilizing data visualization techniques and further preprocessing. This approach aims to enhance the accuracy of our analysis by understanding the significance of individual features. By doing so, we can achieve better data representation, faster model training, and improved accuracy.

Secondly, instead of performing a three-class classification, we suggest estimating the decay rate and lifetime of pumps. To accomplish this, we plan to collaborate with Taarifa to obtain construction year data for approximately 20,000 pumps that are currently missing this information in the dataset. Analyzing the relationship between pump decay and other features will allow us to identify trends and patterns.

Lastly, we aim to enhance our success metrics. While our current work focuses on assessing the functionality of pumps, we believe it would be valuable to develop a success metric that calculates the number of lives improved per pump. This approach would enable us to determine which pumps are critical for ensuring water access for Tanzanians.

VII. REFERENCES

[1] Tanzania Ministry of Water & Data Driven. (n.d.). *Pump it Up: Data Mining the Water Table*. Pump It Up: Data Mining the Water Table. Retrieved 2015, from https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/

[2] Tanzania Ministry of Water & Data Driven. (n.d.). *Pump it Up: Data Mining the Water Table*. Retrieved 2015, from https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/24/

[3] Tanzania Ministry of Water & Data Driven. (n.d.). *Pump it Up: Data Mining the Water Table*. Retrieved 2015, from https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/

[4] Addison, Joseph. "Pump it up - Data Mining the Water Table." Yet Another Compsci Guy, 12 Jul.2017

https://jitpaul.blog/2017/07/12/pump-it-up/.

[5] Kim, Joomi. "Predicting Non-Functional Water Pumps in Tanzania." Joomi K Blog, 3 Feb. 2017, joomik.github.io/waterpumps/.

[6] Kremonic, Zlatan. "Predicting Status of Tanzanian Water Pumps." Zlatan Kremonic - My Mathematical Mind, 23 Jan. 2017, zlatankr.github.io/posts/2017/01/23/pump-it-up.