Data Mining Water Table

Debmalya Ray

Abstract— These Preservation of groundwater is very important for the development of agro-based countries. The majority of irrigation is dependent on it. There are several sources of groundwater systems including water pumps and bore-well that utilize groundwater to mitigate the immediate requirement. These sources are named as “water-points”. It is of vital importance to collect data about the water points from a particular region and also its operating conditions. Water infrastructure plays an important role in the development of agriculture and irrigation systems in various geographical regions. The data for this paper is taken from “Taarifa waterpoints dashboard, which aggregates data from the Tanzania Ministry of Water” (**Cited:** Caskey, Brandon, Jacob Gable, and William Lewis. "**Predicting Functionality of Tanzanian Waterpoints.**" Proceedings of Student Research and Creative Inquiry Day 6). Our focus is to do a comparative study of various machine learning techniques to find out the best model suitable to determine the status of the operating water points and willingness to help in building the best water infrastructure.

Index Terms— water table, classification problem, machine learning, water infrastructure, feature engineering, bagging and boosting algorithm.

1. Introduction

Water Scarcity can lead to serious problems for humanity and agriculture-based countries, especially in drought areas. The necessary measure can help us in bringing down the wastage of ground water which should be preserved for future and emergency use.

This global problem is not only confined to the counts of the water points in the various geographical region but also its operating condition. The malfunctioning water pumps can be identified from the data being collected and can be repaired with better effect in due course of time.

There are various factors based on which the operating conditions of the water points can be predicted. Some of the important factors include:

a) the altitude of the water points

b) geographical water basins

c) construction year. etc.

Using this data collected, we will use analytical techniques and machine learning algorithms with its evaluation metrics to predict the water point conditions.

The below diagram (Figure 1) represents the entire life cycle of our research project. The train data identified needs to be changed into its numerical format before feeding into the model. We emphasized the feature engineering techniques as well to make it fit for training. This includes the normalization of the data, checking multicollinearity and removal of outliers from the dataset.

Same techniques should be used for the test data as well. The selected model will be tested and made ready for the evaluation perspectives. A comparative study should be performed with a set of algorithms and evaluation metrics.

The entire life cycle is represented as below:

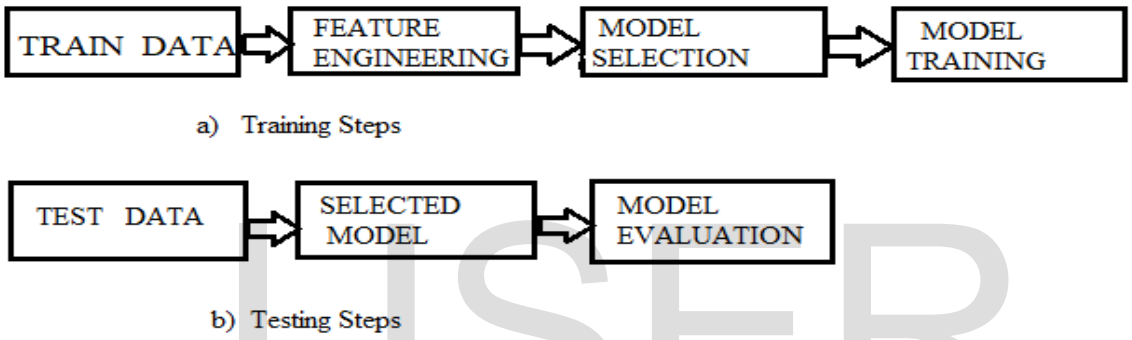


Figure 1

Results expected from the model created:

We did a comparative analysis with various ML models to their accuracy and find out the best model suitable for this problem statement. The algorithms used for this study are mentioned below (Table1):

Algorithms	Accuracy
Random Forest Classifier	Greater Than 50 %
Gradient Boosting Classifier	Greater Than 50 %
XG Boost Classifier	Greater Than 50 %

Table 1: Algorithms

This paper also concentrates on the tuning of the models to their best results. The initial models will help us in baselining its metrics. Once we achieve that, then various hyper-parameter tuning methods will be used to obtain our final result.

For each of the algorithms used, we define a parameter space. The tuning methods will search for the best parameter and then use it for our final model.

2. Problem Statement

To classify the operating conditions of water points in various geographical locations.

Background:

To mitigate the scarcity of water and solve the problem regarding the depletion of the water table, there are various solutions proposed previously. The problem is just not confined to depletion issues but also to understanding the operating conditions of these water points. Our study concentrates on finding these conditions which are broadly classified into three labels:

a) Functional
b) Non-Functional
c) Functional and need repairs

Literature Reviews:

Understanding The Problem	Citation
Preventing failures in water supply systems is of vital importance for the development of a population, especially when its economic engine is the agricultural sector. This paper performs a comparative analysis of three classification algorithms, random forest, support vector machines, and artificial neural networks, to predict failures in a water pumping system. The methodology employed considers the selection of a training dataset, data preprocessing, training, and	Herrera, Gerardo, and Paulina Morillo. "Benchmarking of Supervised Machine Learning Algorithms in the Early Failure Prediction of a Water Pumping System." *Communication, Smart Technologies and Innovation for Society*. Springer, Singapore, 2022. 535-546

evaluation of each model, and its subsequent performance comparison.	
Broken water pumps continue to impede efforts to deliver clean and economically-viable water to the global poor. The literature has demonstrated that customers' health benefits and willingness to pay for clean water are best realized when clean water infrastructure performs extremely well (>99% uptime)	Wilson, Daniel L., Jeremy R. Coyle, and Evan A. Thomas. "Ensemble machine learning and forecasting can achieve 99% uptime for rural handpumps." PLoS One 12.11 (2017): e0188808.
Long-term and accurate predictions of regional groundwater hydrology are important for maintaining environmental sustainability in arid agricultural areas that experience seasonal freezing and thawing where serious water-saving measurements are used.	Zhao, Tianxing, et al. "Machine-Learning Methods for Water Table Depth Prediction in Seasonal Freezing-Thawing Areas." *Groundwater* 58.3 (2020): 419-431.
In this paper, we describe the development and validation of two classification systems designed to identify the functionality and non-functionality of these electrical pumps, one an expert-informed conditional classifier and the other leveraging machine learning.	Thomas, Evan, et al. "A contribution to drought resilience in East Africa through groundwater pump monitoring informed by in-situ instrumentation, remote sensing and ensemble machine learning." *Science of The Total Environment* 780 (2021): 146486
This paper presents an open-source technology-based smart system to predict the irrigation requirements of a field using the sensing of ground parameters like soil moisture, soil temperature, and environmental conditions along with the weather forecast data from the Internet.	Goap, Amarendra, et al. "An IoT-based smart irrigation management system using Machine learning and open source technologies." Computers and electronics in agriculture 155 (2018): 41-49.

3. Aim and Objectives

i) To propose approximate counts of water points in various geographical locations.

ii) To find out the operating conditions of the water points in these locations.

iii) Effective and comparative use of feature engineering and machine learning techniques to find out the best solution for the problem statement.

iv) Feature selection techniques and

v) Determining the evaluation metrics or results applicable to the problem.

Based on the aim, we have created a set of objectives as follows:

- To define the dataframe suitable for the dataset
- To find out the best data cleansing and engineering techniques.
- To perform possible missing values imputation techniques.
- Suitable Encoding techniques should be used to convert the text data into numerical data.
- Scaling and Normalization techniques are suitable for the problem.
- Checking if the dataset is balanced on not.
- Using the best feature selection and model selection techniques.
- Analyze and compare various predictive models and evaluate their performances.
- To suggest the best hyper parameter and optimization techniques that can increase the model performance.

4. Significance of the Study

Our dataset contains a detailed summary of the water points data being collected. This source of the dataset is cited from: "**Predicting Functionality of Tanzanian Waterpoints." Cuskey, Brandon, Jacob Gable, and William Lewis.**

Many agro-based countries rely on local water pumps as their primary source of fresh water. It is of prime importance to make sure the downtime of the pumping system should be reduced to the minimum so that humanity does not face any issues.

Accomplishing a good result is very significant in the reduction of scarcity and building a solid foundation of water supply in rural and water-dependent areas.

5. Scope of the Study

The work will include:

i) Important feature variables: the altitude of the water points, geographical water basins and construction year. etc.
ii) Operating conditions of the water points

Challenges Involved:

All the important feature variables need to be collected before the research study can be performed. It requires a significant amount of effort in collecting data and structuring it for the research problem.

6. Research Methodology

Please find below the required methodology to be performed for achieving the aims and objectives:

1. Understanding the Problem: -
   i) Supervised Learning.
   ii) Target Column with its classes

Please find attached the water datasets:

dataset.csv

dataset2.csv

There are 41 columns. Some of the important feature variables are described below :

Description of the important variables:

Variable_Name	Variable_Description	Label
gps_height	The altitude of the well	Feature
wpt_name	Name of the waterpoint if there is one	Feature
basin	Geographic water basin	Feature
scheme_management	Who operates the water point	Feature
permit	If the waterpoint is permitted	Feature
extraction_type	Types of extraction	Feature
management	How the waterpoint is managed	Feature
water_quality	Quality of the water	Feature
quantity	Quantity of the water	Feature
source	The source of the water	Feature
source_type	Water Source Types	Feature
waterpoint_type	The kind of the waterpoint	Feature
status_group	Conditions of the waterpoint	**Target**

The dataset collected has been classified into two categories:

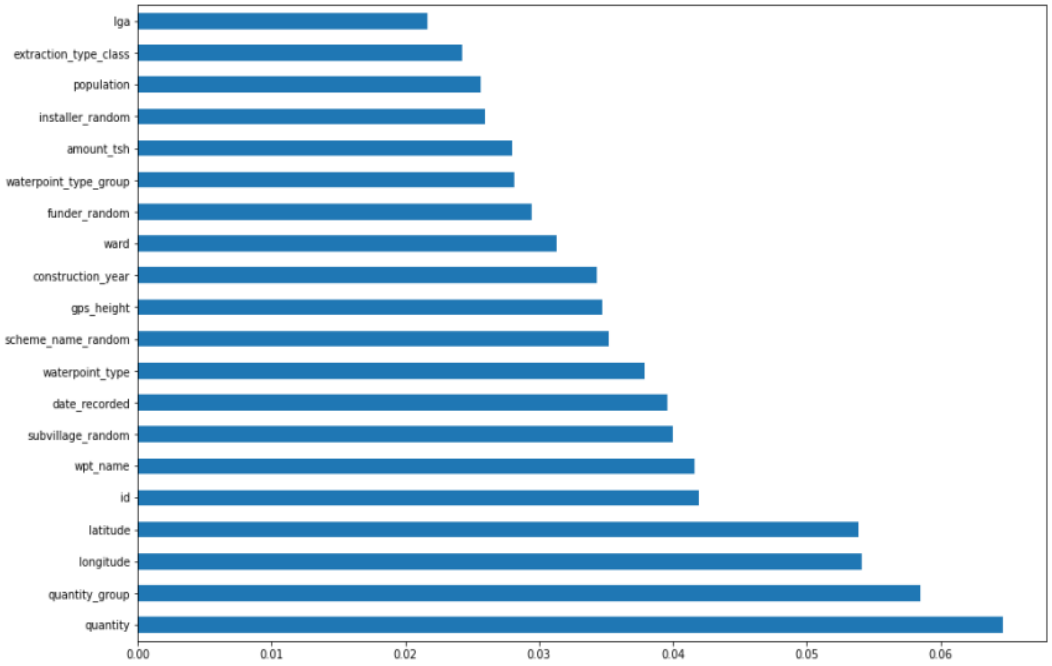
Train Dataset – More than 30,000 rows of data
Test Dataset – More than 13,000 rows of data

2. Feature Engineering:

These are the most important steps among all the methodologies used. The data collected needs to be analysed and prepared for the model. The steps mentioned below are very important in achieving the required results. All the required implications need to be taken care of step by step in a detailed manner. This includes,

i) Checking of Null Values
ii) Information about the dataframe
iii) Dividing the data into its train and test dataframes.
iv) Saving the dataframes as train/test files
v) Analysis of continuous and categorical variables
vi) Scaling and Normalization of variables
vii) Outliers Treatment
viii) Gaussian Transformation of variables
ix) Checking Multi-Collinearity and Feature Importances
x) Dropping the less important variables

The mentioned below diagram (Figure 2) represents the importance of feature variables concerning their target variable.



3. Data Modelling and the comparative study of its evaluation metrics:
   - i) Random Forest Classifier
   - ii) Gradient Boosting Classifier
   - i) XG Boost Classifier

4. Evaluation Metrics used are the same as that of a classification problem.
5. Hyper Parameter Tuning: -
   - i) Bayes Search CV
   - ii) Use of Optuna
6. Pie-Chart Representation of the predicted column with its categories.

Results Obtained:

The final result obtained from the dataset are mentioned below:

Predict	Count

functional	9206
non-functional	3748
functional needs repair	74

7. Required Resources

Programming Language used: Python

Software used: Jupyter Notebook, Anaconda Navigator

Hardware used:

Parameter	Minimum Configuration
CPU Frequency	2.30 GHZ
No. CPU Cores	2
RAM	12 GB(Upgradable to 26.75GB)

Please Note:

The above configuration is chosen based on the basic classical algorithms used in this problem statement.

Important Python Libraries used:

1. Data Visualization and Feature Engineering: Pandas, Numpy, Matplotlib
2. Data Modelling: Sklearn
3. Hyper Parameter Tuning: Optuna, skopt

8. Research Plan

Please find below Gantt Chart explaining the plan and timeline in Image 1:

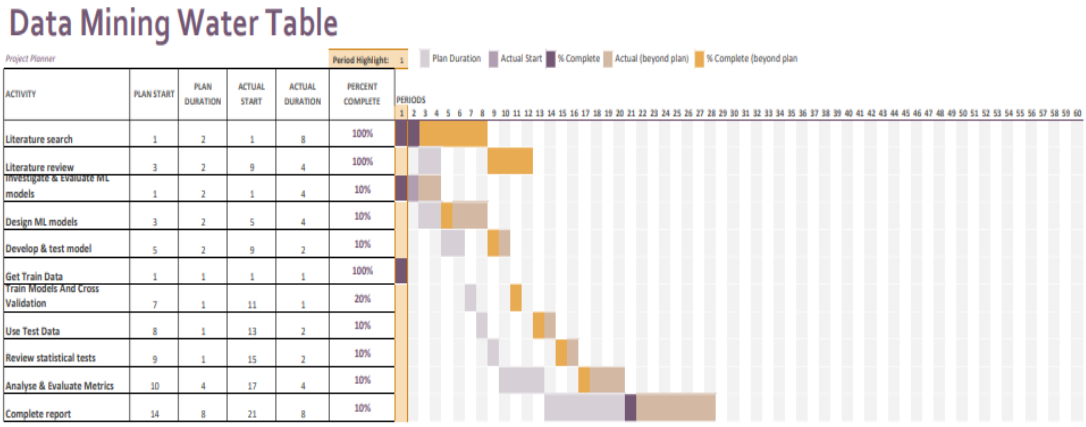


Image 1

9. References

- Herrera, Gerardo, and Paulina Morillo.
"Benchmarking of Supervised Machine Learning Algorithms in the Early Failure Prediction of a Water Pumping System."

- Wilson, Daniel L., Jeremy R. Coyle, and Evan A. Thomas.
"Ensemble machine learning and forecasting can achieve 99% uptime for rural handpumps."

- Zhao, Tianxing, et al.
"Machine-Learning Methods for Water Table Depth Prediction in Seasonal Freezing-Thawing Areas."

- Thomas, Evan, et al.
"A contribution to drought resilience in East Africa through groundwater pump monitoring informed by in-situ instrumentation, remote sensing and ensemble machine learning."

- Goap, Amarendra, et al.
"An IoT-based smart irrigation management system using Machine learning and open source technologies."