Can We Pump It Up? An Exploratory Evaluation of the Water Pumps in Tanzania using Data Science

Joel Jacob Thomas
MSc Cyber Physical Systems
University of Nottingham
psxjj1@nottingham.ac.uk

Matthieu Blackler
MSc Computer Science
University of Nottingham
psxmb11@nottingham.ac.uk

Brendan Ezekiel Agnelo Vaz
MSc Data Science
University of Nottingham
psxbv1@nottingham.ac.uk

***Abstract*—Despite the fact that a water point mapping system has been an established part of the Tanzanian community for more than a decade, the persistent challenge in ensuring its functionality is evident through its water crisis. This paper investigates the application of techniques in data science to predict the operational status of water pumps in Tanzania. Utilising a dataset from the Taarifa waterpoints dashboard, we explore the performance of six machine learning models: Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Gaussian Naive Bayes (GNB), and Multilayer Perceptron (MLP). Following rigorous data pre-processing and model performance evaluation upon different metrics, we find that the Random Forest model is best suitable with 79% accuracy. Invaluable insights can be gathered from our addressal of research questions, through which we discover relationships between the factors affecting the water pumps. These insights could potentially guide proactive maintenance strategies, offering a scalable solution to improve water resource management in all developing countries. Future research is suggested to integrate more complex models and additional predictive variables to further refine the accuracy and applicability of the predictive models.**

***Keywords*—data science, machine learning, predictive analysis, water resource management**

I. Introduction

Health is wealth, reveals the paraphrase of the famous comment by the Roman poet Virgil. The notion emphasises that the wealth brought by our physical and mental well-being is inarguably equal or more in weightage and worth than the likes of money. Water is the most vital element that sustains this health of ours. The human body can last weeks without food, but only days without water. Many bodily processes depend on it, including supplying nutrition to cells, eliminating waste, protecting joints and organs, and regulating body temperature [15]. This drives the importance of availability of clean water. Water is not only used with a purpose of individuality but also for harmonious country's development such as the irrigation of agricultural land, public health provisions, household livelihood and so on. But not all regions in the global index are privileged to such essential provisions. One such region that is void of seamless access to water is Tanzania. In 2021, access to basic drinking water services was at 74 percent, household sanitation 72 percent, and access to handwashing 41.5 percent [16].

Since 2004, Tanzania has included water point mapping (WPM) into its approach to water management to trace the community water availability on a regional basis. The exercise produced maps of 55 of 132 rural water districts [17]. In 2010, the Tanzanian government implemented WPM as its primary instrument for tracking the availability of water in rural areas and created the Water Point Mapping System (WPMS), an online database. Every district in mainland Tanzania was mapped by the Ministry of Water between 2011 and 2013, and the data was uploaded to the WPMS. Following the creation of a new management system called Rural Service Delivery Management Information System (RSDMS) in 2019, the initiative aims had advanced to support monitoring for both operation and governance needs of rural water investments.

Despite the existence of established water points, touching a population of over 57 million people with access to clean water is challenging as the water pumps' functionality is still in question. Tanzania Ministry of Water collaborates with the Ministry of Education Science and Technology, Ministry of Health, the Rural Water Supply and Sanitation Agency (RUWASA), and the President's Office - Regional Administration and Local Government (PO-RALG), to resolve this crisis [16].

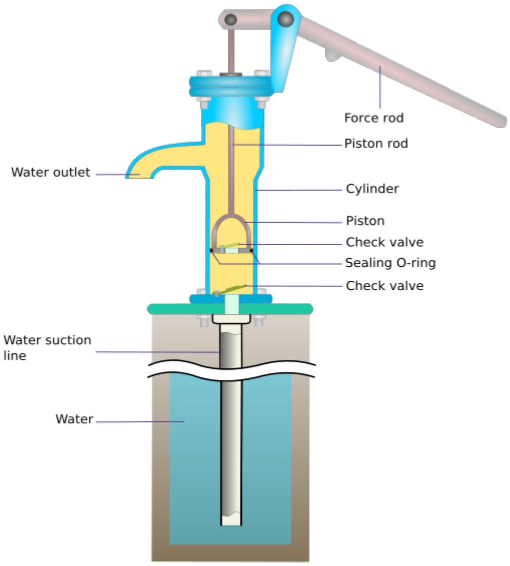

Fig. 1. Diagrammatic Representation of Water Pump [20]

This is where Taarifa comes into the picture. Taarifa is an open-source platform that facilitates the crowd-sourced reporting and resolution of infrastructure-related issues. The "Pump it up: Data Mining the Water Table" dataset comes from the Taarifa waterpoints dashboard [1], which aggregates data from the Tanzania Ministry of Water, providing detailed information on the status of each pump and various characteristics that may influence their operational functionality. This data is vital as it directly supports efforts in infrastructure management and planning, particularly in ensuring access to clean and reliable water sources for communities.

This dataset includes 38 features that can be used to predict the status of the pumps, categorised into three classes: functional, functional needs repair, and non-functional. It includes a range of variables related to the water pumps, such as:

- Type of pump: The technology used in the pump.
- Management: How the pump is managed and by whom.
- Geo-location: Exact coordinates of each water pump.
- Installation date: When the pump was installed, which can be critical for understanding wear and tear.
- Water quality: Descriptors of the water quality, such as good, salty, or milky.
- Quantity: The amount of water the pump yields, which could range from dry to enough for all uses.
- Funding: Details of financial support for the pump's installation and maintenance.

Mining and understanding the interplay between these factors is crucial in determining the functionality of the water pumps, as it allows for efficient allocation of maintenance resources, breakdown predictions, and look-ahead planning. The core aim of this study is to leverage machine learning techniques to predict the operational status of water pumps based on these factors. Therefore, we define the following research questions (RQ) to address as part of this objective to eviscerate the data:

RQ1: Can machine learning models accurately predict the functionality of water pumps in Tanzania?

We seek to validate the effectiveness of machine learning models in categorising water pumps into one of three: functional, functional but needs repairs, and non-functional. To ascertain the predictive accuracy of these models in a practical setting is the goal.

RQ2: What factors are most predictive of pump failures?

This question attempts to highlight areas where interventions could be made to improve pump dependability and service delivery by determining which factors in the dataset most closely correlate with pump failures.

RQ3: Are there any relationships between the features that can predict the functionality of the water pumps? In theory, if all water pumps were manufactured in the same way, and the environment had no effect on them, it would be impossible to predict the status of water pump from these variables. Thus, we seek to reveal the reliability of the role that factors, described in the features, have on the water pump functionality.

RQ4: How do various imputation techniques affect dataset skewing and distribution and which technique is most effective?

This question brings up the issue of dataset distribution in relation to imputation techniques; if a feature contains missing or invalid values, we will want to try and impute these values. However, preserving the distribution and pattern of values is crucial for the model training; as changing the pattern too severely will influence the model outcome and degrade performance.

II. Literature Review

Water pump functionality is a critical issue in many developing countries, not just Tanzania, where access to clean water is directly linked to the operational status of water pumps. Studies have shown that a considerable proportion of water pumps are either non-functional or need repairs at any given time [2]. The functionality of these pumps is influenced by several factors including the type of pump, management practices, and the physical and social environment in which they are installed [3]. For instance, the Modified India Mark II and Afridev pumps are commonly used in rural areas of sub-Saharan Africa and their performance varies widely based on local conditions and maintenance practices [2]. Here, machine learning offers powerful tools for classification and prediction tasks in large datasets. Its models are trained on various features including the type of pump, installation year, management group, and geographical information. Feature engineering plays a crucial role in improving model performance, where irrelevant or redundant features are pruned to enhance the model's predictive accuracy [4] [5].

Moreover, advanced techniques like 1D-CNN-LSTM networks have been explored for predicting the remaining useful life of water pumps, demonstrating the potential of deep learning methods in this field [6]. Data mining methods employed SAS® Enterprise Miner [7] to analyse the pump functionality, using models like random forest, decision trees, and neural networks, involving variables such as water quantity, pump type, and geographical coordinates. These methods might capture complex patterns in the data that simpler models miss, potentially improving prediction accuracy, which is what we are banking on.

Another study tested different algorithms like a voting ensemble classifier combining random forest, and neural networks was used for the best performance across different categories of hand pump sustainability, such as functionality, water quantity, and quality. The study aimed to enhance maintenance efficiency and support provision by predicting the sustainability of water resources but there seemed a lack in optimisation of each layer in the neural network [8].

With a focus on XGBoost and Recursive Feature Elimination, [9] optimised feature selection and achieve a high accuracy of 80.38%. This is an effective approach to handling structured data, but the paper fails to incorporate the class imbalance problem in the dataset and right methods of imputation were not followed. Additional validation methods could be implemented to test the model's generalisability beyond the initial dataset. Future research could integrate real-time data acquisition for dynamic prediction updates, enhancing predictive accuracy and operational responsiveness [9]. Factors influencing the water pump functionality were studied in [13] through multilevel logistic regression and Bayesian networks, where the data were collected via mobile surveys. But this establishes a possibility of potential bias from non-sampling errors like varied enumerator interpretations and reliance on cross-sectional data, which limits causal inferences.

Deploying an IoT-based system allowed remote monitoring and operation of water pumps [10]. The system demonstrates high accuracy (96.67%), specificity (95%), and sensitivity (100%) in controlling the pump based on environmental conditions like soil moisture, temperature, and humidity, instead of the direct factors related to the pump. This automation aims to enhance efficiency and decision-making in agricultural practices in general. The use of such a system provides automated data capture, which can lead to effective prediction, and can be a proactive solution to the water crisis [14].

III. Methodology

A. Data Cleaning

Handling Missing Values as part: Missing data was addressed by analysing the pattern of missingness.

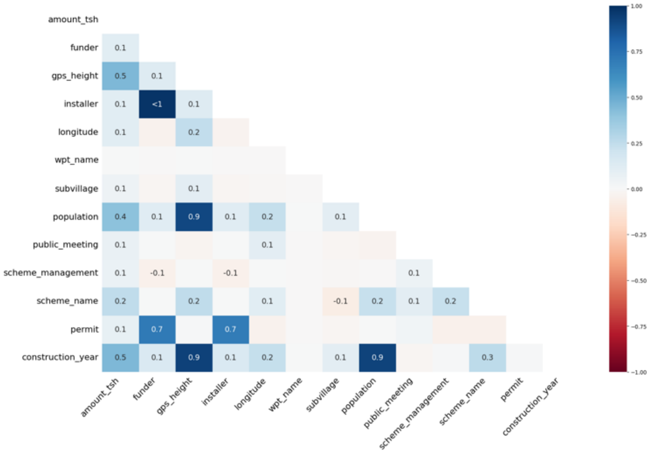

Fig. 2. Correlation in missing values

Upon the EDA, for columns with a significant percentage of missing values, imputation methods were employed on each of the split datasets separately, which helps prevent leaking of the imputation, thus overfitting the model. For numerical variables, mean imputation was trialled but did not produce desirable results as it skewed the data distribution; instead, we made use of MICE (Multivariate Imputation by Chained Equations) imputation which yielded better results [11]. When comparing an enhanced mean imputation of GPS coordinates, taking the mean of the area from less precise to more precise, mice imputation yielded closer results.

A select few redundant columns were removed, after observing that they had exact correlations with a different column. The few categorical variables that were missing were filled with either the mode or a "missing" category; depending on the impact this would have on the distribution of the data. Finally, we chose to remove certain columns which either did not provide any valuable information ("num_private") or that had a too great number of missing values ("amount_tsh" for example where 70% of values were missing).

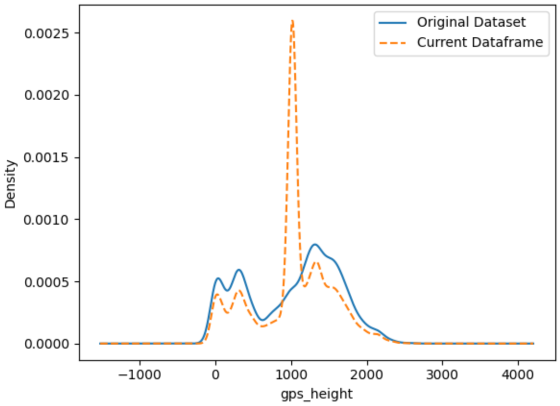

Fig. 3. Enhanced Mean imputation of GPS height

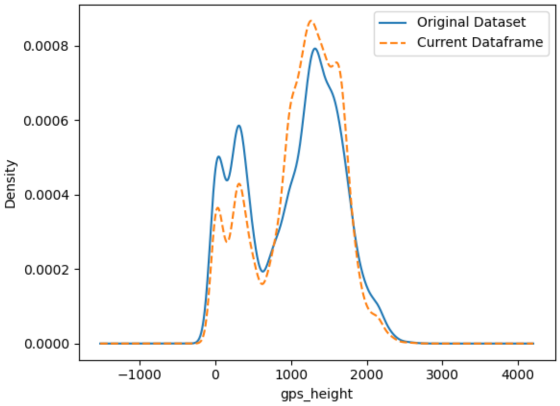

Fig. 4. MICE imputation of GPS Height

Data Type Corrections: The dataset required type conversions, such as converting date fields from strings to date/time objects to facilitate time-series analysis or age calculations of the water pumps.

B. Feature Engineering

Creating New Features: We created an age feature from the installation date, which is a significant predictor of failure likelihood; it is intuitively easier for a model to understand an age rather than a year.

Encoding Categorical Variables: At the modelling stage, non-numeric categories were transformed into numerical data using one-hot encoding for categorical features, as most machine models cannot read string of text. Due to the data preprocessing with one-hot encoding and our use of SVD, the features of our model are no longer comparable to the features of our original dataset; therefore, we cannot derive a relevant feature importance. Therefore, before any modelling, we run a simple version of a Random Forest Classifier Model without SVD and reconstruct the features, to gain an idea of the feature importance; even though these might not be directly applied to our final model.

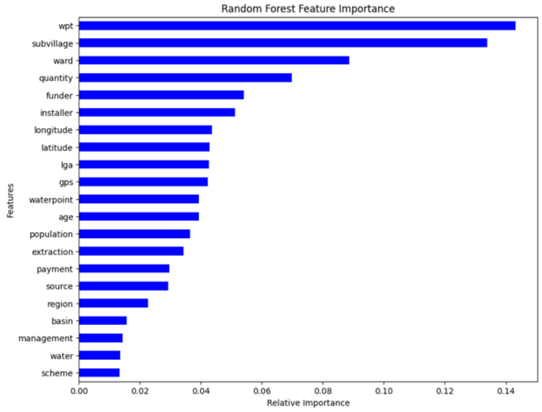

Fig. 5. Feature Importance by Random Forest

This showed that a few high cardinality features that are highly specific were having a high impact on our model (water point names, sub village and ward) and could lower the performance of it on a varying dataset; therefore, we decided not to use those columns in our model training.

Scaling and Normalisation: Features were scaled and normalised to treat all variables equally, especially important when using models sensitive to the scale of input data, like SVM or KNN.

C. Model Selection and Justification

The choice of machine learning models was dictated by the nature of the data and the specific challenges of the predictive task. We chose a wide range of models, as modern tools like Sklearn make it quite straightforward to trial varying models. Our dataset contained a wide array of categories whose relationship to the water pump's working order might not immediately be apparent, which is where a model such as decision trees could be useful due to it excelling in non-linear relationships in classifying tasks. Our task is finding a pattern in water pump functionality based on indirect patterns so decisions trees naturally feel suitable.

Furthermore, building upon this idea, we will also employ Random Forests as these have the same advantages as Decision Trees while being a more robust and modern implementation of that principle; they are an ensemble of decision trees. They aggregate the outcome of multiple trees, therefore reducing the risk of overfitting. They also aid in understanding feature importance, which we used to determine which features to keep, to remove unnecessarily high cardinality features.

We also included the supervised Support Vector Machines algorithm which is effective in such cases as high-dimensional datasets with many categories and attributes, thus sparse datapoints which some models struggle with, leading to overfitting.

Furthermore, with the non-linear and complex patterns present in our dataset, MLP was chosen because of its deep learning capabilities, allowing it to model intricate and unlikely relationships through multiple layers. Finally, K-Nearest Neighbours was chosen for its simplicity yet effectiveness. This algorithm predicts outcomes based on the assumption that similar instances are located near each other in the feature space.

IV. Results

A. Hyperparameter Tuning

A crucial component of improving machine learning models for increased performance is hyperparameter tuning. To optimise the parameters of the models we used for our study, a variety of hyperparameters must be methodically searched through and chosen.

To make the most of each model, we employed a Bayesian Optimisation algorithm using a library known as 'optuna' to smartly test each algorithm in an efficient and performant way; each 'study' was given a range of hyperparameter, and the algorithm smartly determines which hyperparameter will provide us with the best performance.

For instance, in the case of random forests, the first few trials of parameters only returned accuracies in the low 70s, while after a dozen trials, it was 80 percent. It must be noted, that the "ultimate" set of parameters for our dataset was not necessarily found, as each trial is computationally expensive, as so therefore, we are not to do more than 10 trials per model. The Support Vector Machine (SVM) yields an accuracy of 78%, the Decision Tree Classifier achieves an accuracy of 74% which is comparatively lesser. Multilayer Perceptron Classifier achieves an accuracy of 79%, with varying precision, recall, and f1-score for each class. And K-Nearest Neighbours (KNN) results with a 78% accuracy. However, the least accuracy score is witnessed in the Gaussian Bayes Naive model which has an accuracy of 66%. This thorough analysis emphasises how important hyperparameter tuning is for fine-tuning models so they can perform better in terms of prediction and greater generalisation.

B. Model Performance

With these "best performing" set of parameters determined, the performance of each selected model and its more optimal parameters was rigorously evaluated on the internal testing dataset using a series of metrics: accuracy, precision, recall, and F1-score. The results are presented below, for the best performing model; Random Forests, we have supplemented with visual aids such as confusion matrices and ROC curves to provide a comprehensive understanding of each model's performance. Precision and recall are considerable, defined by:

$$precision = \frac{\sum TruePositives}{\sum TruePositives + \sum FalsePositives}$$

$$recall = \frac{\sum TruePositives}{\sum TruePositives + \sum FalseNegatives}$$

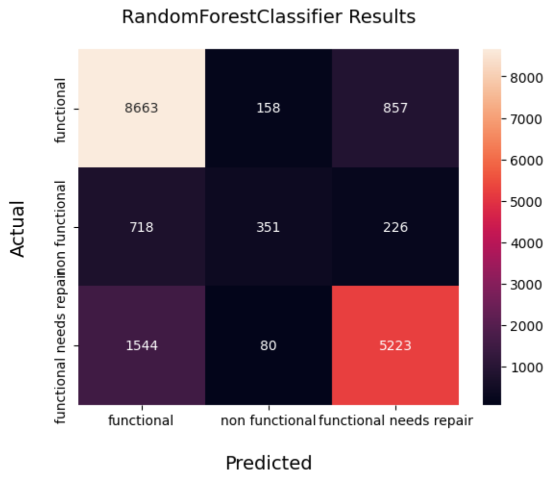

Fig. 6. Random Forest Confusion Matrix

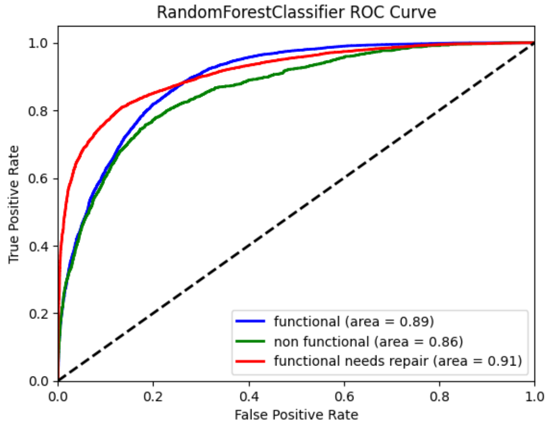

Fig. 7. Random Forest ROC Curve (One vs All)

Visual Aids:

Confusion Matrices: Each model's confusion matrix shows the number of correct and incorrect predictions for each class (functional, functional but needs repairs, and non-functional). These matrices help in visualizing the actual versus predicted classifications, highlighting areas where models may confuse one class for another.

ROC Curves: The Receiver Operating Characteristic (ROC) curves for each model illustrate the trade-offs between true positive rate (sensitivity) and false positive rate (1-specificity) across different thresholds. The area under the ROC curve (AUC) provides a single metric to compare the overall performance of the models. Due to the multi value nature of our model, and ROC traditionally being used for binary classification; we employed one-vs-rest [18].

C. Comparison of Models

We observed notable differences in how each model handled the classification of functional status among water pumps. The random forest classifier emerged as the highest performer, with a weight average score of F1 of 0.79. Its strength lies in balancing precision and recall consistently across classes, particularly for the 'functional' category with a score of 0.84; the ensemble approach is working when compared to the decision tree classifier which reported lower overall accuracy and F1-Scores, we can theorise that this may be due to overfitting. The latter, along with the KNN Classifier had difficulty in accurately classifying the 'functional needs repair' status which might be due to their sensitivity to the model's noise within this dataset class. The MLP Classifier and SVC did show strong performances with accuracy of 0.79 and 0.78; with the MLP and Random Forest exceeding all other models at identifying "functional" water pumps with a recall of 0.89. The SVM classifier was the most accurate at predicting the non-functionality of the water pumps with a precision of 0.85 but struggled with the 'functional needs repair' category, achieving the lowest recall at 0.15.

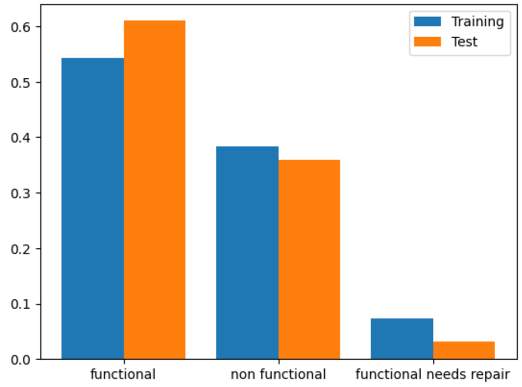

Fig. 8. Comparison between final predictions (test) and labelled data (train), using random forest

Lastly, the Gaussian Naïve Bayes Classifier was by far and large the weakest preforming classifier with an accuracy and weighted average F1-score of 0.66. This model was particularly weak in precision and recall for the "functional needs repair" category, highlighting potential limitations in handling datasets where the feature independence assumption does not hold.

TABLE I
COMPARISON OF MODEL PERFORMANCE

Model	Decision trees	Random Forests	SVM	KNN	MLP	GNB
Accuracy	0.75	0.80	0.78	0.77	0.78	0.66
Precision (wgt avg)	0.74	0.79	0.78	0.77	0.77	0.66
Recall	0.75	0.80	0.78	0.77	0.78	0.66

In conclusion, while all models provided valuable insights and modelling capabilities, Random Forests stand out in overall performance, making them particularly suitable for this application in predicting the operational status of water pumps in Tanzania. These findings suggest that while each model has its place depending on the scenario and requirements, ensemble approaches like Random Forests are more robust for complex, multi-class classification tasks such as this.

V. Discussion

A. Interpretation of Results

The results obtained from all the classifier models provide insightful answers to the primary research questions posited at the onset of this study. It has been proven that the accuracy in predicting operational status of water pumps can be extracted from machine learning methods. The Random Forest model demonstrated the highest accuracy at 79.991%, indicating a strong capability to classify the water pumps into functional, functional but needs repairs, and non-functional categories. This aligns with findings from the literature, where ensemble methods such as Random Forests are often more effective in handling complex classification problems with imbalanced datasets [12]. Upon submission to the competition page, our final predictions to the dataset made with our random forest model were compared with the answers; our model was given an accuracy score of 0.7991.

Decision Trees and SVMs showed lower accuracy but still provided valuable insights, especially in terms of model interpretability and handling high-dimensional spaces, respectively.

B. Comparison with Other Literature

TABLE II
Methods reflection with other literature

Author	Method	Accuracy
Jacqueline F. et al. (2018) [8]	Voting Ensemble Classifier	66.4%
Darmatasia et al. (2016) [9]	XGBoost	73.9%
Ryan C. et al. (2017) [13]	Bayesian networks and regression	71.8%

Table II gives a comparison of the models from other literature that use the same dataset. Our method of eviscerating the data and propelling a predictive analysis with random classifier proposes the best accuracy with 79%. This implies that with the right methodology of pre-processing and trialling, the classifier can help determine the operational status of the water pumps better than any other model. A good consideration of the imbalanced data classes is thus a mandate.

C. Critical Evaluation

Model Overfitting: Decision Trees were prone to overfitting, which might have led to less generalised predictions. Although Random Forests mitigate this, the risk of overfitting complex models to specific data characteristics remains a concern.

Data Quality and Completeness: The presence of missing values and potential inaccuracies in the dataset required assumptions and imputations that could affect the reliability of the predictions. While methods for handling missing data were employed, the inherent biases introduced through imputation could skew model outcomes, although this was mitigated as much as possible by following standard imputation practices such as running imputation tasks on the test and training data separately.

Computational Resources: SVMs, especially with non-linear kernels, required significant computational resources, which may not be feasible in all operational contexts, particularly in low-resource settings prevalent in many parts of Tanzania. This also means, we were unable to tune the models to their best parameters.

Lastly, the Gaussian Naïve Bayes Classifier was by far and large the weakest preforming classifier with an accuracy and weighted average F1-score of 0.66. This model was particularly weak in precision and recall for the ”functional needs repair” category, highlighting potential limitations in handling datasets where the feature independence assumption does not hold.

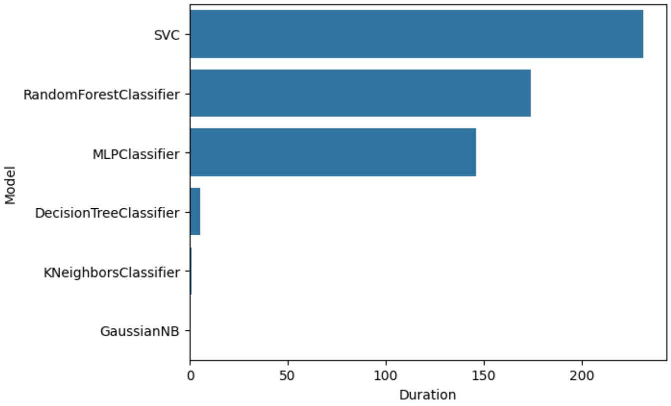

Fig. 9. Relative Model Runtime

D. Discrepancies between Outcomes

Performance Discrepancies: While literature suggested high efficacy of neural networks in similar tasks, the current study did not employ these due to their complexity and resource demands, highlighting a gap between potential and chosen methodologies.

From a layperson perspective, we can notice from Fig. 10 that the number of non-functional pumps is predicted to be higher, which is explainable by the real-world status of the crisis. This can also indicate that government records of water points may not be precise as there would be water pumps located in places which are humanly impossible to examine and predict status of. This proves that RQ1 has a strong standing as machine learning can escalate the ease of examination of practical assets like water pumps even in a geographically underserved spot.

Feature Importance Variability: Some features anticipated to be highly predictive, such as geographic location and installation year, did not show as strong an influence as expected. This could be due to the overriding impact of more immediate operational factors like pump type and maintenance history.

In summary, the discussion underscores the importance of selecting appropriate machine learning models based on

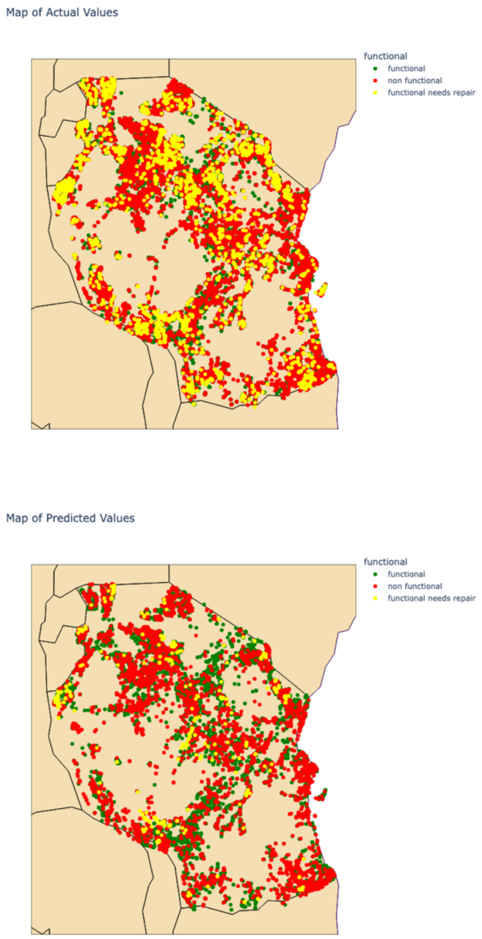

Fig. 10. Geographical Comparison of Predicted and Available statuses of water pumps across Tanzania

both their predictive performance and practical considerations such as model interpretability, computational demands, and adaptability to specific data conditions. The findings and methodologies applied in this study contribute valuable perspectives to ongoing efforts in using data science for effective public infrastructure management, particularly in resource-constrained environments.

Therefore, the research questions posited can be converged and elucidated in the following manner: Firstly, in such a dataset where a considerable portion of important values are missing, it is clear that the choice of imputation is extremely important; mean imputation for instance does provide very reliable results and changes the data distribution significantly while MICE imputation preserves it adequately. Machine learning models, specifically the Random Forest model, have indeed proven to be highly effective in predicting the functionality of water pumps in Tanzania, demonstrating an accuracy of at least 79%. This highlights the utility of machine learning in managing and maintaining critical infrastructure. Key factors influencing pump failures include the geographical location and age of the pump, manufacturing and management practices, and water quality. These insights enable targeted and effective maintenance strategies. Finally, the fact that our model is accurate shows that there is a pattern of unlikely causal relationships in water pump functionality which only these complex algorithms could ever perceive; and having proved this will be very useful in the future deployment of water pumps.

VI. Conclusion

To sum up, this research constitutes a noteworthy contribution in utilising machine learning methodologies to tackle the problem of water pump performance in Tanzania. Through the application of diverse models to forecast the operational state of water pumps, significant insights have been obtained, providing concrete advantages for the management of water resources in the area.

Among the models tested, the Random Forest model performed the best, with a high accuracy of 79%. Its capacity to classify water pumps into three categories—functioning, functional but in need of repairs, and non-functional—is demonstrated by its high degree of accuracy. Because the accuracy of the pump failure prediction is so high, resources may be deployed more effectively, focusing maintenance efforts where they are most needed. With any luck, this focused strategy will reduce downtime, maximise resource use, and eventually improve access to clean water in Tanzania communities.

VII. Future Work and Recommendations

To improve the precision and usefulness of predictive models for water pump functionality, new research directions must be suggested. The following are well-reasoned recommendations designed to enhance the report's conclusions and goals:

Enhanced Variable Integration:

Environmental Factors: By capturing the impact of external factors on pump performance, the models' prediction power can be enhanced by incorporating comprehensive environmental data, such as soil composition, precipitation patterns, and temperature changes.

Socio-Economic Data: By include socio-economic factors like income distribution, community demographics, and infrastructure investment, a more comprehensive understanding of water pump dynamics may be fostered by gaining insightful knowledge about usage patterns and maintenance needs.

Policy and Implementation Research:

Impact Assessment: To determine the practical effectiveness of predictive models and pinpoint areas for optimisation, it is imperative to carry out thorough evaluations of their application in policy formation and maintenance scheduling.

Cost-Benefit Analysis: Conducting thorough cost-benefit analyses to assess the financial effects of implementing predictive maintenance models can clarify the observable advantages in terms of lower costs, less downtime, and better water accessibility, strengthening the argument for spending money on data-driven infrastructure management systems.

Higher scale and Advanced Machine Learning Techniques:

Many shortcomings of this research were due to resource limitations. However, one could imagine a more complex testing pipelines being better at achieving higher accuracy; including testing different combinations of imputation techniques, experimenting with a wider array of hyperparameters. Techniques such as ensemble learning, such as gradient boosting and stacking, holds promise for enhancing predictive accuracy by leveraging the strengths of multiple algorithms and mitigating individual model biases.

Time Series Models: Delving into specialised time series modelling techniques enables the analysis of temporal trends in pump functionality, unveiling long-term degradation patterns and seasonal variations for more informed maintenance planning.

As a whole, the study's results highlight how machine learning can revolutionise water resource management initiatives. Our goal is to guarantee that every community in Tanzania and abroad has fair access to clean water by utilising data-driven strategies and embracing ongoing innovation.

Acknowledgment

We acknowledge that this is an original contribution based on the Taarifa dataset ”Pump it up: Data Mining the Water Table”. We thank Prof Praminda Caleb-Solly and Dr Jeremie Clos, and the entire team on COMP4030 Data Science and Machine Learning module of the University of Nottingham for their unconditional support towards the successful completion of this research study and beyond.

References

[1] DrivenData, Pump it up: Data Mining the Water Table. Available at: https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

[2] Fisher MB, Shields KF, Chan TU, Christenson E, Cronk RD, Leker H, Samani D, Apoya P, Lutz A, Bartram J. Understanding handpump sustainability: Determinants of rural water source functionality in the Greater Afram Plains region of Ghana. Water Resour Res. 2015 Oct;51(10):8431-8449. doi: 10.1002/2014WR016770

[3] S. Parry-Jones, R. Reed, and B. H. Skinner. Sustainable handpump projects in Africa. WEDC, Loughborough University, UK. https://wedcknowledge.lboro.ac.uk/docs/research/WEJW2/Literature-Review.pdf

[4] Nair, A. (2023) Predicting the functionality of water pumps with XGBoost, Medium. Available at: https://towardsdatascience.com/predicting-the-functionality-of-water-pumps-with-xgboost-8768b07ac7bb

[5] Green, S. (2019) Pump it up: Predict water pump condition using data science, Medium. Available at: https://towardsdatascience.com/pump-it-up-predict-water-pump-condition-using-data-science-2839d26638b8

[6] Jian Pan, Yujiang Li, Huandong Zhao. (2023) A prediction method of remaining useful life of water pump equipment based on 1D-CNN-LSTM. Journal of Physics: Conference Series 2450 (2023) 012006, IOP Publishing. doi:10.1088/1742-6596/2450/1/012006

[7] SAS® South Central User Group Forum (2016). Determining the functionality of water pumps in Tanzania. Available at: http://www.scsug.org/wp-content/uploads/2016/11/SS-Determining-the-functionality-of-Water-Pumps-in-Tanzania.pdf

[8] Flefil, J.F., Kozlow, V. and Galanis, M.A. (2018) Pump it or Leave it? A Water Resource Evaluation in Sub-Saharan Africa. Available at: http://www.scsug.org/wp-content/uploads/2016/11/SS-Determining-the-functionality-of-Water-Pumps-in-Tanzania.pdf

[9] Darmatasia and A. M. Arymurthy, ”Predicting the status of water pumps using data mining approach,” 2016 International Workshop on Big Data and Information Security (IWBIS), Jakarta, Indonesia, 2016, pp. 57-64, doi: 10.1109/IWBIS.2016.7872890

[10] View of monitoring and controlling water pumping system using IoT for agriculture purpose (2019). Available at: www.akademiabaru.com/submit/index.php/aram/article/view/1849/835

[11] Azur, M.J. et al. (2011) Multiple imputation by chained equations: What is it and how does it work?, International journal of methods in psychiatric research. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

[12] Using random forest to learn imbalanced data. Available at: https://www.researchgate.net/publication/254196943-Using-Random-Forest-to-Learn-Imbalanced-Data.

[13] Ryan Cronk and Jamie Bartram (2017). Factors Influencing Water System Functionality in Nigeria and Tanzania: A Regression and Bayesian Network Analysis. Environ. Sci. Technol. 2017, 51, 19, 11336–11345. https://doi.org/10.1021/acs.est.7b03287

[14] Dhandapani, V., Thomas, J.J., Sravanthi, Y.D. (2022). Integrated Smart Alert System for Industrial Applications using Transceiver Module Analysis. Sengodan, T., Murugappan, M., Misra, S. ICAECT 2021. Lecture Notes in Electrical Engineering, vol 881. Springer, Singapore. https://doi.org/10.1007/978

[15] Medicine, G.F.I.G. (2023) Why it's important for you to drink water and stay hydrated, good-food. Available at: https://health.ucdavis.edu/blog/good-food/why-its-important-for-you-to-drink-water-and-stay-hydrated/2022/07

[16] World Bank Group (2023) Expanded access to water supply, sanitation, and hygiene services in Tanzania, World Bank. Available at: https://www.worldbank.org/en/results/2023/11/20/expanded-access-to-water-supply-sanitation-and-hygiene-services-in-tanzania

[17] Water Point Data and governance in Tanzania (no date) WPdx. Available at: https://www.waterpointdata.org/2022/06/16/water-point-data-and-governance-in-tanzania/

[18] Hoo, Z.H., Candlish, J. and Teare, D. (2017) What is an ROC curve?, Emergency Medicine Journal. Available at: https://emj.bmj.com/content/34/6/357

[19] Yan, X. et al. (2024) A comprehensive review of machine learning for water quality prediction over the past five years, MDPI. Available at: https://www.mdpi.com/2077-1312/12/1/159

[20] Manco Capac, CC BY -SA 3.0 ¡https://creativecommons.org/licenses/by-sa/3.0¿, via Wikimedia Commons.

Contributions

Joel Jacob Thomas: Conceptualisation, Methodology, Software, Investigation, Writing. **Matthieu Blackler:** Data curation, Investigation, Software, Visualisation, Writing. **Brendan Ezekiel Agnelo Vaz:** Software, Formal analysis, Validation, Data curation, Writing.