Prediction of waterpoint status in Tanzania

Yuhao Zhang
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
scyyz12@nottingham.ac.uk

Yichen Lu
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
scyyl15@nottingham.ac.uk

***Abstract*—Due to the uncertainty of functionality in Tanzania's waterpoints, numerous regions face the challenge of ensuring a stable water supply. The issue poses a threat to the health and well-being of local people. This study aims to use multiple machine learning models to identify the waterpoint functionality in Tanzania. The dataset is obtained from Tanzania Ministry of Water. Data pre-processing is preformed on the dataset, while 2 sub-research questions regarding the geographical attributes and data time attributes are addressed for further criteria selection. The cleaned data is fed into 6 machine learning models for training and validation. The result shows that the Random Forest model outperforms other models in terms of accuracy, precision, recall, and F1-score.**

***Index Terms*—Data Science, Waterpoints, Pre-processing, KNN, RF, GBM, AdaBoost, SVM, MLP.**

I. Introduction

Water is the most essential natural resource on the earth, underpinning all aspects of human life, from our survival and health to our economic and industrial activities. Despite its vital importance, water scarcity remains a global challenge. According to the World Health Organisation (WHO), over 2 billion people worldwide live in water-stressed countries [1]. Tanzania, a country in East Africa where one third of the country is arid to semi-arid [2], serves as a prime example of a nation with water scarcity. Data from water.org reveals that out of Tanzania's population of 59 million people, 28% of the population lack access to safe water, and 73% lack access to safely managed household sanitation facilities [3]. The lack of clean water not only contributes to health issues such as water-borne illnesses like cholera and malaria, which account for over half of the diseases in the country, but also hinders economic and social development of the country.

In response to the water scarcity issue, various organisations and governments have initiated projects to construct waterpoints across Tanzania. These waterpoints aim to provide communities with a reliable source of clean water, thereby improving public health, promoting gender equality, and fostering economic growth. Based on the data from Tanzanian Ministry of Water [4], prior to 2013, there are 59,400 recorded waterpoints in the country. Despite the large number of waterpoints, the challenge persists, with 22,824 (38.42%) non-functional waterpoints and an additional 4,317 (7.27%) partially functional ones requiring repairs. A smart understanding of which waterpoints may experience failure and require repair can assist the government in enhancing maintenance operations and guaranteeing the provision of safe drinking water to communities throughout Tanzania.

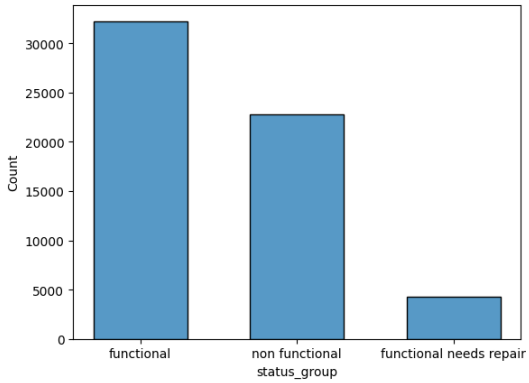

Fig. 1. Waterpoint status in Tanzania

The data from Tanzanian Ministry of Water consists of 40 attributes relating to the waterpoint itself (such as the name, the installer, and the funder of the waterpoint), the environmental factors (such as GPS coordinates and administrative location of the waterpoint, the water quality and quantity around the waterpoint), and the management factors (such as its management scheme and the cost of water). All the 59,400 records distribute the waterpoint status into three classes, functional, functional but needs repair, and non-functional. Distribution of the data can be seen in Figure 1.

In this research, we try to identify the status of waterpoints in Tanzania based on the identity, environmental, and management factors of the waterpoints.

II. Literature Review

In our investigation of predicting the status of waterpoints in Tanzania based on several aspects of factors, several prior studies provide valuable insights.

The research conducted by Bonsor et al. [5] suggests an overarching perspective on the impact of climate change on groundwater recharge in Africa. While not directly tied to post-construction factors, this study offers vital background information on the broader environmental factors that may influence water point functionality in Tanzania.

Although not based in Tanzania, the data-mining research by Dale et al. [6] scrutinised the impact of management factors

on the functionality of water supply projects. The study proffered insight into the performance of community-managed, demand-driven models in diverse geographic and cultural contexts. Their comprehensive assessment, incorporating technical, management, and water quality aspects, indicated a significant relationship between management factors and the functionality of waterpoints. Their work offers key insights into the management aspect of water point functionality.

Emboldened by advances in information communication technologies, the Ministry of Water has been developing computing, financial and administrative technologies to update and visualise the status of rural waterpoints. The research by Jesper et al. [7] pioneered a new digital research method to evaluate the ware resource in Tanzania. In a different vein, Michael et al. [8] proposed the use of water point mapping as an instrument for enhanced water governance in Tanzania. Their research underscores the potential of using geographical data for post-construction management of waterpoints, a perspective we aim to incorporate in our study.

Finally, the wealth of empirical data collected previously has made it possible for subsequent data mining and data analysis to be based on machine learning. Ryan et al. [9] utilised a regression and Bayesian network analysis to predict water point failures, incorporating a range of factors including system type, administrative unit, and management structure. For example, systems were more likely to be functional if they were used for both human and livestock consumption. Their study provides a model for predictive analysis in this field, particularly in its use of machine learning and statistical methods and looking for the strong dependencies between functionality and the principal factors.

These studies collectively offer an array of perspectives on the environmental and management factors affecting waterpoint functionality. Their methodologies, findings, and limitations will be essential in shaping our own investigation into the prediction of waterpoint status in Tanzania.

III. Methodology

This section will provide an explanation of the methodology employed in this research. Firstly, the description of the study area will be presented, followed by an explanation of the preprocessing and the proposed method utilised in this study.

A. Study Area

The study area is Tanzania. It is located between latitudes 1°S and 12°S and longitudes 29°E and 41°E. The country encompasses a vast area, stretching from the eastern coastline along the Indian Ocean to the western border with Lake Tanganyika and beyond. Regarding altitude, Tanzania features diverse elevations across its landscape. The country's lowest point is the shoreline along the Indian Ocean, which lies at M.S.L.; the highest point is the summit of Mount Kilimanjaro, which reaches an altitude of 5,895 meters above M.S.L.. Tanzania has approximately 64 million people. In terms of population distribution, Tanzania has a diverse demographic landscape, with higher population densities in coastal areas and several urban centres, and lower density in rural areas.

Figure 2 illustrates the current status of waterpoints in Tanzania, indicating a mixed situation across all regions. Operational and non-operational water points coexist in each region. In urban areas, the water supply situation is generally good, while in some areas, particularly those located in arid or semi-arid areas, there may be more non-functional waterpoints.

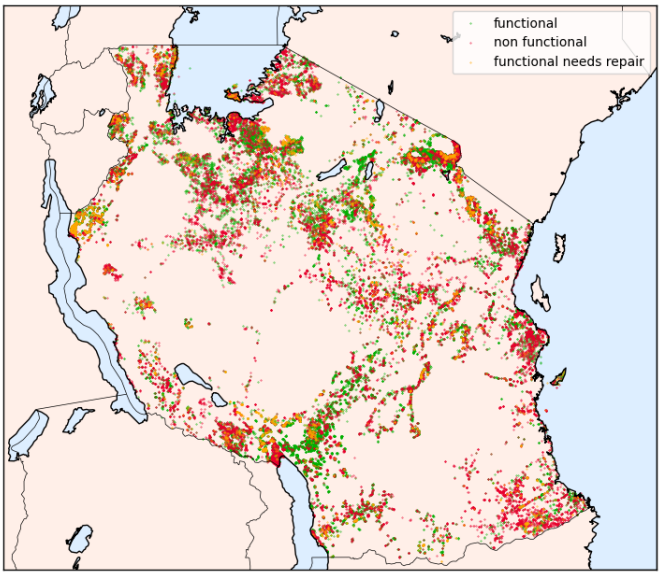

Fig. 2. Geographical distribution of waterpoints in Tanzania

B. Data Pre-processing

Upon initial inspection of the dataset, several issues were identified that could potentially degrade prediction accuracy, including missing values, duplicates, fractured attributes, and outliers. To address these issues, data pre-processing is performed on the training set.

Firstly, the dataset contains irrelevant and fractured attributes. For instance, the attribute *scheme_name* is absent in 48.5% (28,835) of the records, the attribute *wpt_name* is unrelated to the waterpoint status, and the attribute *subvillage* consists of 19,288 categories, resulting in excessively fine granularity. These attributes cannot be used for prediction, and are impossible to impute. Therefore, they are directly dropped.

In addition, the dataset exhibits inconsistencies in syntax and typos, particularly within the attributes *installer* and *ward*, where the same category may appear in various forms with differing spelling, case, spacing, and punctuation. A rule is applied on these attributes to unify the syntax of these attributes. Following the removal of irrelevant features and standardisation, all duplicated records are subsequently eliminated.

Missing data is present in both numerical and categorical attributes within the dataset. Numerical attributes with missing values are denoted as numerical zeroes (0), and they are

imputed using the mean of corresponding geographical groups. Categorical attributes, on the other hand, exhibit various forms of missing data, including 0, no value, *none*, *Not Known*, or *NaN*. To standardise these missing values, they are uniformly replaced with *unknown*.

Furthermore, outliers in the dataset, including the left-skewed *construction_year* and the right-skewed *population* numerical attributes, as well as categorical attributes like *installer* and *lga* that consist of categories with a single value, are also examined.

Finally, the dataset exhibits redundant attributes that measure the same feature of waterpoints and display highly similar data. For instance, as illustrated in the heat-map displayed in Figure 3, both *water_quality* and *quality_group* serve as indicators of the water quality of the waterpoints. Generally, these attributes demonstrate a one-to-one correspondence, with the exception of two instances where *salty* and *fluoride* show a one-to-two match. Due to the easily understandable relationship between the two categories, the attribute *quality_group* is manually excluded.

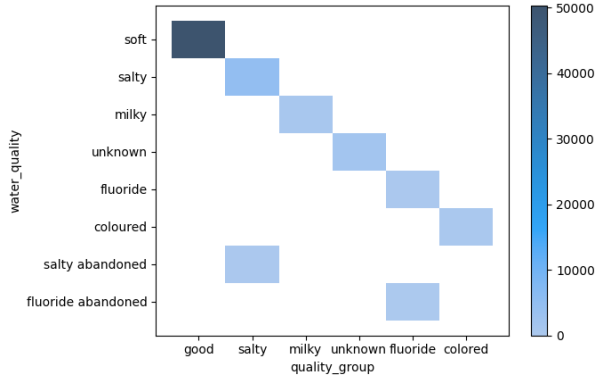

Fig. 3. Correlation between *water_quality* and *quality_group*

The same dimensionality reduction technique involving manual feature selection is also applied to attributes such as *management*, *payment*, *source*, and others.

The data pre-processing significantly reduces the dimensionality of attributes from 40 to 19, with the remaining attributes carefully examined and imputed. With respect to the records, only duplicate instances are eliminated, resulting in a total of 58,799 records remaining.

C. Criteria Selection

Researches have found that the waterpoint functionality is linked to population, climatological condition [10], water demand and water scarcity [11], and other environmental, geographical, geological and social conditions [12]. While this dataset provides attributes from different aspects, it is vital that these criteria are carefully selected and utilised for machine learning model training.

In this part, we are initially concerned with the multiple geographical attributes of the waterpoint in the dataset. It is also worth exploring the date time values with waterpoint functionality. We try to explore and answer these two sub-research questions, before we use machine learning models for prediction.

1. How to best utilise geographical attributes of the waterpoint to improve prediction accuracy

The dataset contains various geographical attributes for each waterpoint, namely *gps_height*, *longitude*, *latitude*, *basin*, *position_code*, and *ward*. Among these attributes, *gps_height*, *longitude*, *latitude* precisely determine the location of each waterpoint within a Coordinate System. On the other hand, *position_code* and *ward* offer insights into the administrative allocation of the waterpoints, while the attribute *basin* provides descriptive information about a nearby landform in relation to each waterpoint.

In our research, we aim to explore the correlations existing among all these attributes. However, it is important to note that certain attributes, such as *basin*, *position_code*, and *ward*, are categorical in nature. This categorical nature poses a challenge when attempting to calculate the correlation coefficient directly between these attributes. Furthermore, it is essential to consider the *longitude*, *latitude* attributes in conjunction as a pair, as they represent the spatial coordinates of the waterpoints and should be analysed together.

The correlations of categorical attributes and the *longitude* and *latitude* of waterpoints are inspected by visually mapping the waterpoints according to their respective groups. The attributes *basin* and *position_code* displayed clear clustering of waterpoints, with no overlap observed among different clusters, indicating a strong correlation. However, when plotting *ward* on the map, we found waterpoints in different wards messed up, and there are too many wards. Therefore, we regard the granularity of *ward* too fine and messy, and decide to drop the attribute.

Furthermore, spatial overlap is observed in the distribution of waterpoints based on their *position_code* and *basin* attributes. To further investigate the correlation between these attributes, a heatmap visualisation is used. The heatmap is depicted in Figure 4.

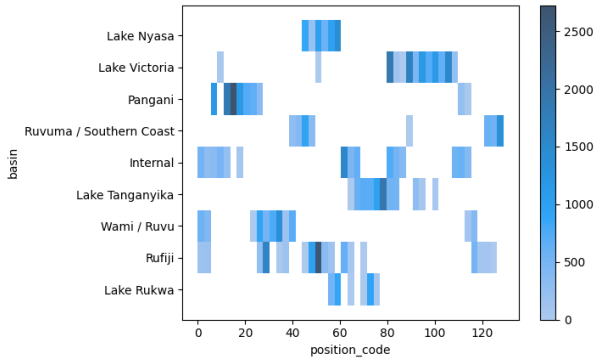

Fig. 4. Correlation between *basin* and (re-coded) *position_code*

Upon examining the plots, it becomes evident that the *position_code* and *basin* attributes offer distinct classifications

of the waterpoints. The waterpoints categorised under each *position_code* always fall within one to three different *basin* regions. Consequently, a weak to medium correlation can be observed between *position_code* and *basin* and therefore it is reasonable to utilise both attributes in prediction.

The correlation between *gps_height* and *basin* is also inspected based on the belief that the difference in height within the same basin should not be significant. However, the results show no correlations between these two attributes.

Based on the correlation analysis, we determined that all geographical attributes except *ward* hold significant value and should be incorporated into the prediction model for waterpoints. However, it is crucial to acknowledge that both *basin* and *position_code* are categorical attributes. Consequently, when applying One-Hot Encoding to these attributes, they greatly increases the number of attributes to over 150.

It is imperative to recognise that not all of these attributes bear equal importance or contribute equally to the predictive power of the model. To assess attribute importance, a Random Forest model is able to estimate attribute importance by calculating the average decrease in node impurity. Figure 5 gives an example of the 20 most important attributes selected by the model.

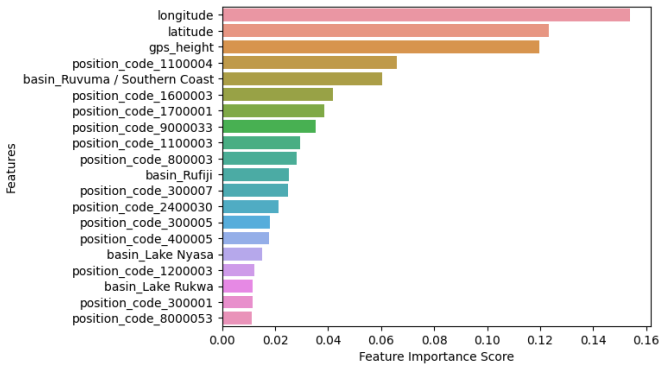

Fig. 5. Feature importance score estimated by a Random Forest

Consequently, a Recursive Feature Elimination (RFE) technique utilising Random Forest as the attribute importance estimator is employed to iteratively eliminate the least significant attributes until the top 10 important attributes are determined [13].

2. Is there any strong relationships between the date time and the waterpoint functionality?

The time data is a worth-exploring topic in the provided dataset. According to common sense, it is considered that the construction time and the data recording time are not directly related to waterpoint functionality, but only depend on how long they have been used. Therefore, our initial purpose is to create a new attribute called *usage_year* by subtracting *construction_year* from *date_recorded*. The waterpoint's age would in turn serve as a good feature in the modelling process.

On the other side, a research group led by Dr MacAllister [14] indicates that waterpoint had a complex relationship with the dry season. Moreover, the handpump cylinder was less than 10 meters below the water table at 38% of sites, which increases the risk of it running dry during intensive use. In this case, the waterpoints functionality is possibly related to the time when the records were created, as well.

Although the date of construction data showed no missing values, a total of $20,709$ data points registered as zero there, suggesting that those waterpoints have been around since the switch from B.C to A.D. We computed the correlation between *construction_year* and other attributes, and the attribute latitude displayed the highest correlation value of 0.039. However, it is insufficient for imputation purposes. Furthermore, dropping all the zero values in the *construction_year* can increase the correlation with the predicted label, *status_group*, to around 0.24. As a result, we made the decision to delete these attributes in our dataset.

D. Data Splitting

We split the dataset with an 70:30 splitting ratio. Table I shows the number of records for each class in training and test sets.

TABLE I
CLASS COUNT FOR TRAINING AND TEST SETS

Class	Training Set	Test Set	Total
functional (0)	22,365	9,547	31,912
functional needs repair (1)	2,955	1,288	4,243
non functional (2)	15,839	6,805	22,644

Upon analysis, it has been identified that the class distribution within the dataset is imbalanced. In light of this observation, it becomes imperative to address the issue of class imbalance in order to ensure fair and accurate model training and evaluation. To mitigate the impact of class imbalance, two common strategies are considered: oversampling and weighting.

Oversampling involves replicating instances from the minority class, thereby increasing its representation within the training set. Table II shows the number of records for each class after oversampling of the training set.

TABLE II
CLASS COUNT AFTER OVERSAMPLE OF THE TRAINING SET

Class	Training Set (after oversampling)
functional (0)	22,365
functional needs repair (1)	22,365
non functional (2)	22,365

On the other hand, weighting assigns higher importance or significance to instances from the minority class during the training process.

Both of these techniques aim to rectify the class imbalance and provide the model with a more balanced and representative learning experience. The balanced training set is fed into machine learning models when in need.

E. Proposed Method

Multiple machine learning models were deployed on the dataset to assess their performance. The models encompassed a non-parametric model KNN, decision tree-based models such as RF, GBM, and AdaBoost, linear models such as SVM, and a neural network MLP. These models were fine-tuned using optimal hyper-parameter settings to achieve their highest performance. The hyper-parameter configurations for these models are presented in Table III.

TABLE III
HYPER-PARAMETER SETTING OF MACHINE LEARNING MODELS

Model	Hyper-parameters
KNN	*n_neighbors*=7, *metric*='minkowski'
RF	*n_estimators*=200, *random_state*
GBM	*n_estimators*=1000, *learning_rate*=0.1
AdaBoost	*n_estimators*=1000, *algorithm*='SAMME.R', *learning_rate*=1.0
SVM	*kernel*='rbf', *C*=0.1, *gamma*=1
MLP	*activation*='tanh', *alpha*=0.0001, *hidden_layer_sizes*= (100,200,100), *learning_rate*='adaptive', *solver*='adam'

1. K-Nearest Neighbours

The K-Nearest Neighbours (KNN) classifier [15] is a widely used algorithm for classification tasks. It is a non-parametric approach that assigns labels to new instances based on the class labels of its K nearest neighbours in the feature space. The algorithm operates on the principle that similar instances tend to belong to the same class. To classify a new data point, the KNN algorithm computes the distances between the point and all other training instances, selecting the K nearest neighbours. The class label of the new point is then determined by a majority vote among the labels of its neighbours.

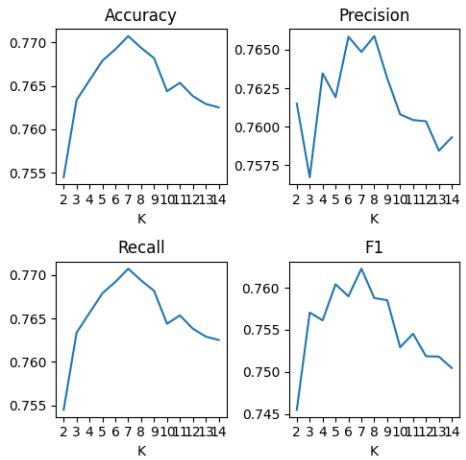

Fig. 6. !!!!!

To determine the optimal K value, various K values are tested and compared in terms of accuracy, precision, recall, and F1 Score. The results are presented in Figure 6. Among the tested values, $K = 7$ yields the highest accuracy, recall, and F1 Score, along with a relatively high precision. Therefore, it is selected as the hyper-parameter setting.

2. Random Forest

The random forest classifier [16] consists of a combination of tree classifiers. Each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote to classify an input vector based on the most popular class. We can define a parameter, *N*, as the number of trees to be grown. To classify a new dataset, each case of the datasets is evaluated by all *N* trees. The forest selects the class with the highest number of votes among the *N* trees for each case.

Furthermore, the random forest classifier utilises the Gini Index as an attribute selection measure, which quantifies the impurity of an attribute in relation to the classes. For a given training set *T*, selecting one case (pixel) at random and saying that it belongs to some class C_i, the Gini index can be written as:

$$\sum \sum_{j \neq i} \left(\frac{f(C_i, T)}{|T|} \right) \left(\frac{f(C_j, T)}{|T|} \right) \tag{1}$$

where $\frac{f(C_i,T)}{|T|}$ is the probability that the selected case belongs to class C_i.

Each tree is grown to the maximum depth on new training data using a combination of features, and these fully grown trees are not pruned. This characteristic represents a significant advantage of the random forest classifier over other decision tree methods.

3. Gradient Boosting Machine

Gradient Boosting Machine (GBM) [17] is an ensemble model used for regression and classification tasks. GBM combined weak learners such as decision trees to make prediction processes significant. Unlike RF, GBM uses boosting method, in which multiple weak learners are sequentially trained to correct the mistakes made by the previous learners. Error rate from first weak learner is transferred to second and then further ones to reduce it and get optimal training. The idea is to combine the predictions of these weak learners into a strong learner that performs better than any individual learner. We used GBM with 1000 boosting stages which means that 1000 decision trees will participate in the prediction procedure.

4. AdaBoost

Adaptive Boosting (AdaBoost) [18] is another ensemble learning algorithm that aims to improve the performance of weak classifiers (decision trees) by iteratively constructing a strong classifier. AdaBoost works by assigning weights to training instances, with higher weights given to misclassified instances. It trains weak classifiers on weighted versions of the training set and evaluates their performance based on weighted classification error. AdaBoost then adjusts the weights of the instances, emphasising the misclassified ones, and trains

subsequent weak classifiers to focus on the previously misclassified instances. The final strong classifier is formed by combining the weak classifiers using weighted majority voting. We used AdaBoost with 1000 boosting stages which means that 1000 weak classifiers will participate in the prediction procedure.

5. Support Vector Machine

Support Vector Machine (SVM) [16], based on statistical learning theory, determines decision boundaries which optimise class separation.

Dealing with linearly separable problems, SVM selects the boundary with the largest margin, as the distances to the hyperplane from the closest points of each class. Maximising the margin that is measured by few support vectors involves Quadratic Programming (QP).

For non-linearly separable cases, SVM will find a hyperplane that maximises the margin while minimising misclassification errors. User-defined parameter *C* controls the margin-misclassification trade-off. SVMs handle non-linearity by projecting data into a high-dimensional feature space using nonlinear mappings, reducing computation with kernel functions.

6. Multi-layer Perceptron

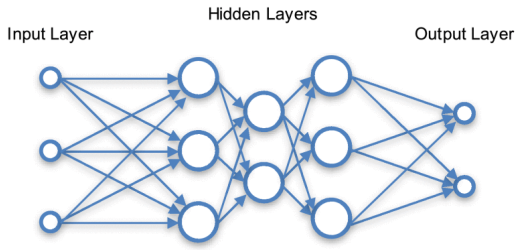

Fig. 7. Example of an MLP with 3 inputs, 3 hidden layers; the first with 3 neurons, the second with 2 neurons and the third with 3 neurons and an output layer with 2 neurons.

As Figure 7, a multi-layer perceptron (MLP) [19] is a fully connected class of feed-forward artificial neural network (ANN). It contains at least three layers of nodes, an input layer, a hidden layer, and an output layer. Each layer is fully connected to the next one, and except for the input nodes, each node uses a nonlinear activation function. The MLP is a feed-forward network, meaning information moves from input to output without looping back. MLPs are capable of learning complex patterns through a supervised learning technique called back propagation, which adjusts the weights of the connections to minimise the difference between actual and predicted outputs. This ability to model non-linear separability distinguishes MLPs from linear perceptrons. They are used in various applications such as speech recognition, image recognition, and machine translation, amongst others.

F. Performance Evaluation

This study employed four evaluation criteria to assess the performance of machine learning models, accompanied by the utilisation of a confusion matrix as a tool for evaluation. The primary evaluation criterion utilised was *accuracy*, which measures the proportion of correct predictions made by the model in relation to the total number of predictions, providing an indication of the model's correctness.

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions} \tag{2}$$

Furthermore, the confusion matrix consists of four essential components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). In case of this problem that being a non-binary classification problem that has 3 classes, these terms are generalised as:

- True Positive (TP): The number of instances correctly predicted as the target class
- False Positive (FP): The number of instances incorrectly predicted as the target class
- False Negative (FN): The number of instances incorrectly predicted as other classes than the target class

By leveraging these terms within the confusion matrix, it becomes possible to derive the evaluation parameters of *precision*, *recall*, and *F1 score*.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F1\text{-}score = \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

These parameters serve as additional indicators of the model's performance and effectiveness.

IV. Results

This section presents the outcomes of the machine learning models. All experiments were conducted on a machine equipped with an Intel(R) Xeon(R) CPU @ 2.20GHz. The performance evaluation of the models is based on metrics such as accuracy, precision, recall, and F1 Score. Tables IV to IX shows the results of all models.

TABLE IV
Results obtained using KNN

Accuracy	Class	Precision	Recall	F1-score
0.77	functional	0.77	0.88	0.82
	non functional	0.80	0.71	0.75
	functional needs repair	0.53	0.30	0.38
	macro avg	0.70	0.63	0.65
	weighted avg	0.76	0.77	0.76

TABLE V
RESULTS OBTAINED USING RF

Accuracy	Class	Precision	Recall	F1-score
0.80	functional	0.80	0.87	0.84
	non functional	0.83	0.77	0.80
	functional needs repair	0.50	0.34	0.41
	macro avg	0.71	0.66	0.68
	weighted avg	0.79	0.80	0.79

TABLE VI
RESULTS OBTAINED USING GBM

Accuracy	Class	Precision	Recall	F1-score
0.77	functional	0.76	0.90	0.82
	non functional	0.83	0.70	0.76
	functional needs repair	0.58	0.21	0.31
	macro avg	0.72	0.60	0.63
	weighted avg	0.77	0.77	0.76

TABLE VII
RESULTS OBTAINED USING ADABOOST

Accuracy	Class	Precision	Recall	F1-score
0.73	functional	0.71	0.88	0.79
	non functional	0.78	0.63	0.70
	functional needs repair	0.44	0.10	0.16
	macro avg	0.64	0.54	0.55
	weighted avg	0.72	0.73	0.71

TABLE VIII
RESULTS OBTAINED USING SVM

Accuracy	Class	Precision	Recall	F1-score
0.64	functional	0.86	0.50	0.63
	non functional	0.57	0.86	0.69
	functional needs repair	0.32	0.49	0.39
	macro avg	0.59	0.61	0.57
	weighted avg	0.71	0.64	0.64

TABLE IX
RESULTS OBTAINED USING MLP

Accuracy	Class	Precision	Recall	F1-score
0.75	functional	0.80	0.87	0.80
	non functional	0.79	0.66	0.72
	functional needs repair	0.48	0.27	0.34
	macro avg	0.67	0.60	0.62
	weighted avg	0.74	0.75	0.74

V. DISCUSSION

Given the importance of access to clean, safe water, accurately predicting the functionality of these is crucial. This study highlights the waterpoint functionality problems in Tanzania by analysing numerous attributes of waterpoints there that may cause a loss of their functionality. Based on the main purpose of predicting the operating condition for each record in the dataset, we determined how to use the critical attributes and choose classification models to reach our goal. Besides, the other two sub-questions are proposed in this study to explore the impact that geographical attributes and datetime have on the results.

During the present study, we have undertaken an investigation to predict the functionality of waterpoints in Tanzania by employing six different models. Each model was fine-tuned independently, with distinctive results. We compared the predictive performance of each model by incorporating the optimal parameters.

The Random Forest model exhibited superior performance in all evaluated aspects: precision, recall, and F1 score, while its accuracy reaching a value of 0.80. This indicates that the Random Forest model was able to accurately identify the functional status of the waterpoints, with a high degree of reliability. The F1 score, being the harmonic mean of precision and recall, suggests that the model's ability to balance both these aspects was exceptional.

However, the performance of SVM model is worst among all the models. The relatively low score of only 0.57 suggests that SVM is not the best choice for this task, possibly due to the complexity and characteristics of the data, or the need for multivariate predictions. Indeed, SVM can struggle with larger datasets, but it might not handle the noise in the data as effectively as tree-based models like Random Forest. Anyway, there is the need for multivariate predictions within three unbalanced groups of records, which would increase the error rate for a SVM model that is adept at binary predictions.

The remaining models delivered average performances, with scores ranging between 0.73 and 0.77. This range suggests that while these models were reasonably effective at predicting the functionality of the waterpoints, they did not perform as well as the Random Forest model. The disparity between their scores and that of the Random Forest model may be due to differences in the way these models handle feature interactions, noise, or outliers.

The superior performance of the Random Forest model in this context underscores its potential utility in supporting initiatives aimed at improving water access in Tanzania. However, it is significant to recognize that while our study has identified the Random Forest model as the most effective in this context, which may not always work perfectly for other scenarios or datasets. Generally speaking, the choice of model should depend on the specific characteristics of the problem and the collected dataset. Future work could explore the incorporation of other models and ensemble methods to improve prediction accuracy further.

VI. CONCLUSIONS

The study aimed to predict the operating condition of waterpoints in Tanzania using various machine learning models. Moreover, we raised two sub-research questions which each focus on analysing the relationship between the waterpoint status and geological attributes and date-time respectively. Data pre-processing was performed to address issues in the initial dataset, such as missing values, duplicates, fractured attributes, and outliers. Six different machine learning models were implemented, each optimally tuned for better results

with cross-validation, including K-Nearest Neighbours (KNN), Random Forest (RF), Gradient Boosting Machine (GBM), AdaBoost, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). The Random Forest model demonstrated the best performance across all metrics, with its precision, F1 score, and recall, which suggest that the Random Forest model holds promise for helping improve water access in Tanzania. Furthermore, other models showed average performances, with scores ranging between 0.73 and 0.77, indicating that the top ten important attributes have a strong correlation with the waterpoint functionality.

VII. Author contributions

YZ and YL contributed to conception, design, and statistical analysis of the study. YZ wrote the abstract, introduction, data pre-processing, sub research question 1, data splitting, performance evaluation. YL wrote the literature review, sub research question 2, and discussion sections, and conclusion sections. In terms of machine learning models, YZ is responsible for conducting experiments and result written up for KNN, GBM, and AdaBoost models, and YL is responsible for RF, SVM, and MLP models.

References

[1] "Drinking-water," World Health Organization, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/drinking-water. [Accessed: 06-May-2023].

[2] R. Shore, "Water in crisis - spotlight Tanzania," The Water Project. [Online]. Available: https://thewaterproject.org/water-crisis/water-in-crisis-tanzania. [Accessed: 06-May-2023].

[3] "Tanzania's water crisis - Tanzania's water in 2022," Water.org, 2022. [Online]. Available: https://water.org/our-impact/where-we-work/tanzania/. [Accessed: 06-May-2023].

[4] "Pump it Up: Data Mining the Water Table." [Online]. Available: https://www.drivendata.org/competitions/7/.

[5] H. Bonsor and A. MacDonald, "Groundwater and climate change in Africa: review of recharge studies," BRITISH GEOLOGICAL SURVEY, Jan. 2010, [Online]. Available: https://assets.publishing.service.gov.uk/media/57a08b3240f0b64974000a1a/60826_IR-10-075_recharge_review1.pdf [Accessed: 13-May-2023].

[6] D. Whittington et al., "How well is the demand-driven, community management model for rural water supply systems doing? Evidence from Bolivia, Peru and Ghana," Water Policy, vol. 11, no. 6, pp. 696–718, Dec. 2009, doi: 10.2166/wp.2009.310.

[7] J. Katomero, Y. Georgiadou, J. H. Lungo, and R. A. Hoppe, "Tensions in Rural Water Governance: The Elusive Functioning of Rural Water Points in Tanzania," ISPRS International Journal of Geo-information, vol. 6, no. 9, p. 266, Aug. 2017, doi: 10.3390/ijgi6090266.

[8] M. Fisher et al., "Understanding handpump sustainability: Determinants of rural water source functionality in the Greater Afram Plains region of Ghana," Water Resources Research, vol. 51, no. 10, pp. 8431–8449, Oct. 2015, doi: 10.1002/2014wr016770.

[9] R. Cronk and J. Bartram, "Factors Influencing Water System Functionality in Nigeria and Tanzania: A Regression and Bayesian Network Analysis," Environmental Science & Technology, vol. 51, no. 19, pp. 11336–11345, Sep. 2017, doi: 10.1021/acs.est.7b03287.

[10] M. J. Cordão, I. A. Rufino, P. Barros Ramalho Alves, and M. N. Barros Filho, "Water shortage risk mapping: A GIS-MCDA approach for a medium-sized city in the Brazilian semi-arid region," Urban Water Journal, vol. 17, no. 7, pp. 642–655, 2020. doi:10.1080/1573062x.2020.1804596

[11] K. Simukonda, R. Farmani, and D. Butler, "Intermittent water supply systems: Causal factors, problems and solution options," Urban Water Journal, vol. 15, no. 5, pp. 488–500, 2018. doi:10.1080/1573062x.2018.1483522

[12] A. Noori et al., "A reliable GIS-based FAHP-FTOPSIS model to prioritize urban water supply management scenarios: A case study in semi-arid climate," Sustainable Cities and Society, vol. 81, p. 103846, 2022. doi:10.1016/j.scs.2022.103846

[13] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, no. 1–2, pp. 273–324, 1997. doi:10.1016/s0004-3702(97)00043-x

[14] D. J. MacAllister et al., "Contribution of physical factors to handpump borehole functionality in Africa," Science of the Total Environment, vol. 851, p. 158343, Aug. 2022, doi: 10.1016/j.scitotenv.2022.158343.

[15] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, pp. 986–996, 2003. doi:10.1007/978-3-540-39964-3_62

[16] M. Pal, "Random forest classifier for remote sensing classification," International Journal of Remote Sensing, vol. 26, no. 1, pp. 217–222, Jan. 2005, doi: 10.1080/01431160412331269698.

[17] A. Natekin and A. Knoll, "Gradient Boosting Machines, a tutorial," Frontiers in Neurorobotics, vol. 7, 2013. doi:10.3389/fnbot.2013.00021

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997. doi:10.1006/jcss.1997.1504

[19] S. Haykin, Neural Networks: A Comprehensive Foundation. 1998. [Online]. Available: http://www.cis.hut.fi/Opinnot/T-61.3030/luennot2007/lect1.pdf