Comparative Analysis of Machine Learning Techniques for Waterpoint Functionality Classification in Tanzania

1st Kunal Bafna
Department of Computer Science
University of Nottingham
Nottingham, UK
psxkb5@exmail.nottingham.ac.uk

2nd Shruti Sundaram
Department of Computer Science
University of Nottingham
Nottingham, UK
psxss34@nottingham.ac.uk

3rd Shefali Mahindrakar
Department of Computer Science
University of Nottingham
Nottingham, UK
psxsm34@nottingham.ac.uk

***Abstract*—In previous research on determining machine learning models for waterpoint functionality classification, Random Forest classifiers have typically been favored due to their high accuracy. However, this study aims to broaden the scope by comparing a variety of machine learning techniques - Voting Classifier, Boosting technique (GradientBoost and XGBoost), and Feature Selection technique (Recursive Feature Elimination) - to identify the most effective approach. Several preprocessing techniques were also applied in the data cleaning stage to ensure data integrity, including cleaning to handling outliers and fill missing values, feature scaling and standardization, data discretization for continuous variables , and encoding of categorical variables using One Hot Encoding technique. Among all the techniques evaluated, XGBoost and Voting classifier with base configuration as Random forest and XGBoost have almost same balanced accuracy but the sensitivity of Voting Classifier was slightly higher than XGBoost Classifier. This study illustrates the efficacy of integrating diverse machine learning methods for enhanced predictive accuracy in waterpoint functionality assessment.**

***Index Terms*—Gradient Boosting; Decision Tree Classifier, Recursive Feature Elimination(RFE), Voting Classifier, XGBoost Classifier, Ensemble Learning**

I. Introduction

The provision of clean and safe drinking water is the challenge across the world that has affected billions of people and has bad implications on public health.The water situation in Tanzania is graphically underscored by troubling statistics that show inactive progress and inequities in access to safe water sources. Over the last two decades, access to piped water has remained almost stable, with only 33.1% of the population having such access in 2010, compared to 33.5% in 1991/1992. Furthermore, metropolitan areas have witnessed a large loss in access, falling from 77.8% to 58.6%, and rural areas have only seen a slight improvement, rising from 19.2% to 24.1%. [1]

In Tanzania, approximately 58 million people out of its total population of 65 million people lacked access to basic drinking water services. This means that millions of Tanzanians still relied on unimproved water sources or had to travel long distances to access clean water. According to the Tanzania Demographic and Health Survey (TDHS) conducted in 2015-2016, approximately only 55% of rural households had access to improved drinking water sources, compared to 91% of urban households. [3]

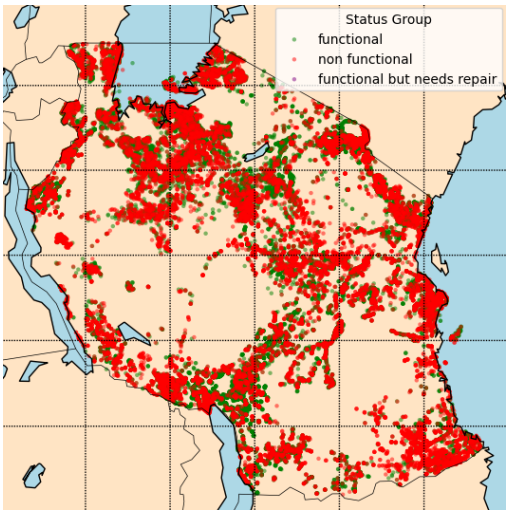

Fig. 1. Geographic Distribution of Waterpoints in Tanzania

Technology can play a crucial role in determining water availability and the functionality of water pumps in Tanzania. This research paper applies various machine learning techniques to a dataset of water points across Tanzania to predict the functionality status of these pumps. This involves exploring the relationships between different variables related to water point characteristics, such as type of pump, management organisations, and geographic locations, and their functionality status.

To tackle this issue, the Tanzanian Ministry of Water collaborated with the open-source platform 'Taarifa' to share data related to water infrastructure, particularly information about water pumps spread across Tanzania. Figure 1 shows the distribution of the water points as per the dataset produced by Taarifa. This paper makes use of the dataset to identify the functionality of water pumps, categorizing them into three

potential states: functional, non-functional, and functional but requiring repair. Before proceeding with any data analysis, thorough cleaning of the dataset was necessary due to the abundance of missing, redundant, and inconsistent data.

II. Literature Review

When reviewing related work, a vast majority had used Ensemble Learning techniques (mainly Random Forest Classifier) and Boosting approaches (mainly XGBoost) to forecast the functional status of water pumps. Although some investigations have explored the application of Logistic Regression, Support Vector Machines, and even Deep Learning techniques for predictive modeling, these methods weren't as accurate as the ones mentioned earlier. The majority of data cleaning and exploratory data analysis methods employed across the submitted studies exhibit a considerable degree of similarity.

The efficient prediction and management of water infrastructure integrity are critical for ensuring the reliability and sustainability of water resources, especially in regions prone to water scarcity. Several studies have been done using various data mining approaches to enhance predictive maintenance and operational strategies in this domain.Harvey and McBean (2014) employ a classification tree methodology to enhance pipe integrity forecasts as they examine stormwater pipe deterioration via a method of data mining.[2] Their research focuses on the development of a framework for decision-making that identifies significant deterioration criteria, such as construction year, diameter, and slope, in order to streamline maintenance operations. They achieve a 71% success rate in predicting pipe condition through analysing a portion of the stormwater system in Guelph, Ontario. This provides an affordable substitute for the labor-intensive and expensive full-scope CCTV inspections that are usually used. This method highlights the importance of data-driven approaches in public utility management by improving the prediction accuracy of assessments of municipal infrastructure and offering a scalable model for different urban systems. [7]

Joomi K introduced a new "*no funder data*" feature for missing categorical variables. Missing population values are initially kept as zeros, with potential prediction using methods like clustering or KNN later. A '*missing_construction_year*' feature is added for missing construction_year' values, and models are tested with and without median imputation. To address low-frequency or numerous levels, levels occurring 20 times or less are categorized as "other." The analysis explores two main modeling approaches: one using one-hot encoded data with and without dimensionality reduction, and another leveraging original data formats. Several machine learning models were evaluated, with Random Forest achieving notable performance. The best model achieved around 77% accuracy using GridSearch to fine-tune Random Forest parameters. Despite the model achieving a high F1-score for functional and non-functional labels, it struggles with the intermediate category. [4]

Wang et al. (2013) developed a novel data mining prediction system for urban water pipe failure, emphasising statistical models' practical application over more expensive physical models for pipe failure prediction.[4] The authors utilised a range of statistical techniques, such as single- and multivariate probabilistic models, deterministic models, and ranking-based models that were customised for their unique dataset of pipe conditions. They were able to manage the highly skewed data distribution related to pipe failures—where breaks are infrequent in comparison to non-break instances—by using this method. Their predictive algorithm showed that correct predictions could be generated even with the imbalanced data, indicating that skewness of this kind is not a formidable challenge in predictive analytics [10]

Bao Tram Duong prepared two datasets - one with SMOTE applied, and without. Various models ranging from Decision Tree, Logistic Regression, KNN, SVM, Random Forest, Gradient Boost, ADABoost and XGBoost were explored. In the end, the models were found to perform better on the imbalanced dataset. A feature importance list was first generated using Decision Tree, and many features were eliminated accordingly. The best model came up to be Imbalance Gradient Boost, which is without correcting class imbalance, with an accuracy of 81.3%. [2]

Based on our comprehensive review of previous research papers and analysis of the data characteristics and suitability, we have selected these approaches for further exploration and application:

- Voting Technique
- Boosting Technique
- Feature Selection Technique - Recursive Feature Elimination (RFE)

III. Methodology

A. Dataset Description

The dataset used in this research was sourced from Taarifa and the Tanzanian Ministry of Water, processed by DrivenData, an organisation known for hosting data science competitions with a social impact focus. The dataset consists of 59,400 entries spread across 40 different columns. These columns include a variety of attributes that provide detailed information about the location, management, and operation characteristics of water points, which are essential for analysing their distribution and functionality across different regions. The data is distributed across three instances such as *functional*, *non-functional*, and *functional but needs repair*.

B. Data Pre-Processing

The data preprocessing involves various stages in order to ensure the data is clean, relevant, and properly formatted for analysis.

1) Data Cleaning: This step involves analysis of the data to check if there are any outliers, missing values or redundant columns in the data.

2) Feature Scaling and Standardization: Feature Scaling is the preprocessing technique used to standardize the range of independent features in the dataset to a uniform

range. For the given dataset, it has been observed that different columns has different range values. Standardization technique was used to solve this problem within the dataset.

3) Data Discretization: To handle the continuous values in the dataset such as construction year columns, data discretization technique was used to convert the data into discrete intervals or bins.

4) Encoding Categorical Variables: To handle the categorical variables in the data, One Hot Encoding technique was employed to convert it into binary vectors.

The final pre-processed dataset contains 162 features. The dataset is then split into 70% training and 30% testing. This partitioning strategy was chosen to strike a balance between efficient model training and robust performance evaluation.

C. Proposed Methods

1) Voting Classification Technique

Voting Classifier is a ensemble learning that can be use for both regression and classification. It combines the prediction of multiple models and predicts the output based on most votes or highest average across all the models.The Voting Classifier model offers the potential to outperform individual models by mitigating bias and variance. This approach involves aggregating predictions from multiple classifiers, which is pivotal for handling the dataset with large number of categorical features. These ensemble models exhibit greater resilience to noise and outliers in the data, thereby reducing overfitting by balancing out individual model errors.
In order to implement the Voting Classification technique effectively, a comprehensive approach leveraging ensemble learning techniques is proposed. Building on the previous research, which demonstrated the effectiveness of ensemble methods in enhancing model stability and predictive accuracy, we are inspired to apply a similar approach to our water resource management problem. Smith et al. (2023) effectively used a weighted voting system integrating Decision Tree, Random Forest, and XGBoost—the DRX approach—for network intrusion detection in IoT environments. Their study achieved a impressive accuracy of 99.95% and an F1 score of 99.90% on the UNSW-NB15 dataset. These results significantly surpassed the performance of individual algorithms used in isolation and outperformed other contemporary models, which typically achieved accuracies around 98% [6].

Initially, three base classifiers are selected for ensemble learning: RandomForestClassifier, DecisionTreeClassifier, and XGBClassifier. This ensemble learning technique involves exploration of multiple configurations to understand the impact of different combinations of base classifiers on overall performance. Four distinct ensemble configurations were considered, varying in the inclusion of RandomForest, DecisionTree, and XGBoost classifiers. Hyperparameter optimization is critical for fine-tuning the performance of ensemble models. Hyperparameters for each ensemble configuration are optimised via grid search and cross-validation (cv=3). The hyperparameters studied are *n_estimators* and *max_depth* for RandomForestClassifier, *max_depth* and *min_samples_split* for DecisionTreeClassifier, *max_depth* and *n_estimators* for XGBClassifier. This iterative procedure allows for the determination of ideal hyperparameter values that maximise classification accuracy and generalisation performance.

Algorithm 1 Ensemble Voting Classifier with Hyperparameter Tuning

```
Input: Training data X_train, Y_train, Testing data X_test, Y_test Output: Best model accuracy, Precision, Recall, F1-Score
Initialize classifiers and hyperparameters
Define list of algorithms with configurations
for each configuration in algorithms do
    Create ensemble with soft voting
    Define parameter grid for GridSearchCV
    Perform grid search with cross-validation to get best estimator and best parameters
    Predict on testing data
    Calculate accuracy, balanced accuracy, and training accuracy
    Calculate precision, recall, and F1-score,confusion matrix
end for
```

2) Boosting Technique

Boosting works by building a series of weak learners in a sequential manner, where each subsequent model attempts to correct the errors of its predecessors. This approach focuses on the hardest-to-classify instances, which often include the minority classes in an imbalanced dataset.

A study by Tanha, J., Abdi, Y., Samadi, N. et al. (2020) focused on evaluating the effectiveness of different boosting techniques across multiple datasets to determine which algorithms perform best in terms of handling the challenges posed by class imbalance in multiclass settings. It was found that boosting methods like gradient boosting and XGBoost create models that generalize well on unseen data. They are notable for their ability to minimize overfitting and optimize loss functions, with XGBoost further improving efficiency and scalability. [9]

Based upon the findings of various research papers, we decided to use boosting approaches to predict pump functionality. The main advantage of XGBoost is that it is faster, accurate, efficient, flexible, and portable. In addition, XGBoost can identify the features that have more influence on the status of water pumps. Since Gradient boosting can use any differentiable loss function, it is considered adaptable for various types of data including categorical and continuous outputs in multiclass settings.

In our study, we proposed using Gradient Boost, AdaBoost and XGBoost. During the experimentation it was found that AdaBoost classifier was not giving good accuracy - most likely due to overfitting caused by non-separable data and class imbalance. Hence AdaBoost classifier was discarded, and only Gradient boosting and XGBoost classifiers were considered for the final models.

Algorithm 2 Boosting Classifier with Hyperparameter Tuning

Input: Training data $X_{\text{train}}, Y_{\text{train}}$, Testing data $X_{\text{test}}, Y_{\text{test}}$
Output: Best model accuracy, Precision, Recall, F1-Score

1: Create a base model
2: Define parameter grid for the Classifier
3: Perform GridSearchCV with cross-validation to get best estimator and best parameters
4: Fit GridSearchCV on training data
5: Predict on testing data using the optimized classifier
6: Calculate training accuracy, accuracy, and balanced accuracy
7: Print classification report including precision, recall, and F1-score for test predictions

An initial base model of both Gradient and XGboost model was created and trained on the training data. Predictions are made on both the training and test datasets using the trained model which helps evaluate the model's performance. A parameter grid is defined for hyperparameter tuning, and Grid Search Cross Validation was used to find the best combination of hyperparameters for both the model. The cross-validation strategy used here is k-fold cross-validation (cv=3) to find the and cross validation score. Test accuracy and balanced accuracy are calculated to assess the performance of the trained model.

The hyperparameters specified in the parameter grid for the Gradient Boosting Classifier include *max_depth* and hyperparameters used for tuning the XGBoost classifier are *booster*, *min_child_weight*, and *learning_rate*.

3) RFE Technique

In this methodology we are exploring two approaches: a Decision Tree Classifier and a Decision Tree Classifier with RFE. Here, the first prioritises simplicity while the other incorporates feature selection. By comparing both methods our aim is to identify the optimal balance between simplicity and feature selection efficiency.

In the research conducted by Awad and Fraihat (2023) Recursive Feature Elimination with Cross-Validation was used along with a Decision Tree model in order to optimise feature selection for IoT network intrusion detection systems which resulted in achieving high efficiency with a reduced feature set. Considering their approach,our method suggests a similar strategy for optimising feature selection. We use RFE CV with a Decision Tree classifier, initially focusing on 55 features from the dataset. The features have been selected based on initial observations and literature review which aims to balance model complexity and predictive performance effectively.

We also followed Lian et al. (2020)'s methodology by integrating a Decision Tree classifier with RFE to optimise feature selection, specifically focusing on reducing feature space dimensionality and selecting optimal features. [8]

Taking inspiration from Awad and Fraihat's approach to methodical feature reduction to 15 ideal features, we tested with the optimal subset size that maximizes classification accuracy and minimizes overfitting risks in order to improve the system. [5] Our method incorporates RFECV into a pipeline that includes a Decision Tree classifier. To further optimize the maximum depth hyperparameter and further fine-tune the model, we use a grid search technique. Additionally, we employed a 3-fold cross-validation scheme to ensure the robustness and reliability of our model performance estimations. The Decision Tree Classifier have hyperparameters like *max_depth*, *min_samples_split* and *min_samples_leaf* while the Decision Tree Classifier with RFE employs *max_depth*. These hyperparameters play a vital role in controlling the complexity, structure, and generalization capability of the decision tree models.

This dual approach enables a comprehensive assessment of the trade-offs between model complexity, feature selection efficiency, and classification performance, ultimately guiding the selection of the most suitable approach for the specific application at hand.

D. Performance Evaluation

In our research, we have used evaluation metrics to evaluate the performance of the model. These metrics provide the insights of the effectiveness of all the models. The evaluation is based on a confusion matrix, accuracy, precision, and recall. Accuracy is the proportion of correctly classified instances among the total instances. It is calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correctly classified Samples}}{\text{Total number of test instances}}$$

Precision, defined as the ratio of relevant records obtained to the total number of irrelevant and relevant records retrieved, assesses the model's ability to correctly identify relevant occurrences from the retrieved data set. Conversely, recall, calculated as the ratio of relevant records retrieved to the total number of relevant records in the dataset, evaluates the model's ability to capture all relevant instances within the dataset. Furthermore, balanced accuracy is applied to offer a more thorough evaluation of the model's performance. The arithmetic mean of the two metrics—sensitivity and specificity—is used to compute balanced accuracy.

IV. Results

A. *Data Analysis Results*

During the data analysis phase, several critical observations were identified, including issues related to data imbalance, missing data, and the presence of large zero values within certain columns. For instance, as shown in Fig. 2. , the distribution of the target variable is not even, indicating an imbalance within the dataset. Such imbalances can impact the accuracy of predictive models, particularly with regard to minority classes.

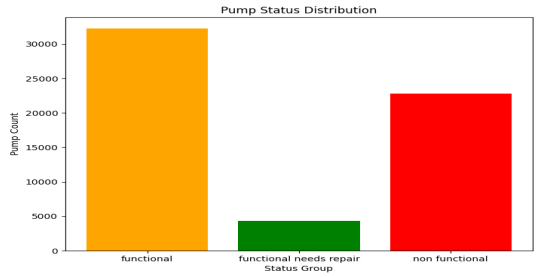

Fig. 2. Target Variable *status_group* distribution

There are some features that can impact the functioning of the water pumps, such as *Construction Year*. Fig. 3. shows the impact of *construction year* on *status group*.

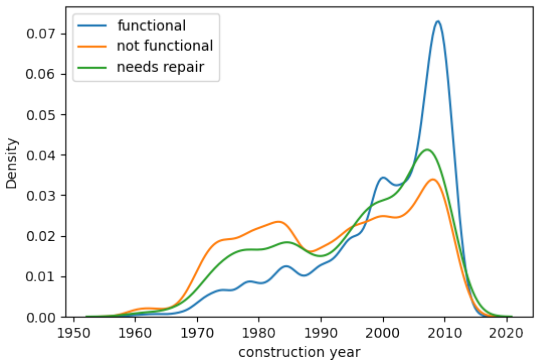

Fig. 3. contruction year vs status group

From Fig. 3, it can be observed the most of the pumps were constructed recently. Older pumps have high probability of failure and requires repair. This feature can be important to predict water pump failure.

B. *Data Preprocessing Results*

This section highlights the steps taken to prepare the data. It involves data cleaning , encoding categorical column, and data discretization.

1) Data Cleaning
This step involves handling missing values, detecting outlier in the dataset and check duplicates in the dataset. Missing values in numeric and categorical features were preprocessed to manage missing values and represent them in a numerical format using imputation technique with strategy as most frequent value. The dataset consists of many correlated columns such as payment type and payment. It can possibly introduce overfitting in the model and also reduces run time of the model. This column may potentially introduce noise and complexity in the model. In order to solve this problem we used cross-tabulation and heatmap visualisations as shown in Fig. 4. to get the insights into the frequency distribution of the categorical variables and detect overlapping patterns.

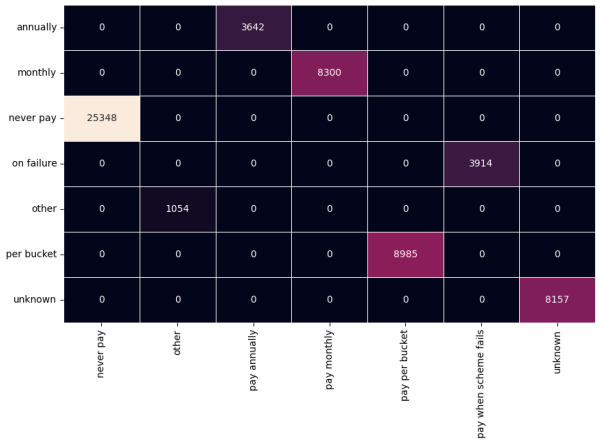

Fig. 4. Frequency Distribution of Payment and Payment Type

2) Encoding categorical Features
Categorical features in the dataset were appropriately one-hot encoded to create binary vectors. The technique was chosen since most of the columns in the dataset were nominal. In addition, certain columns in the dataset have a large number of categories, which could lead to sparsity issues, especially when using one-hot encoding. To order to address this issue, a conditional grouping strategy was implemented where less common categories were aggregated under a label called 'other'. This approach allowed us to manage the large number of categories effectively, reducing sparsity. For example, as shown in Fig. 5. , the *funder* column has various

categories appearing less than 500 times then it was added under a common category named as 'other'.

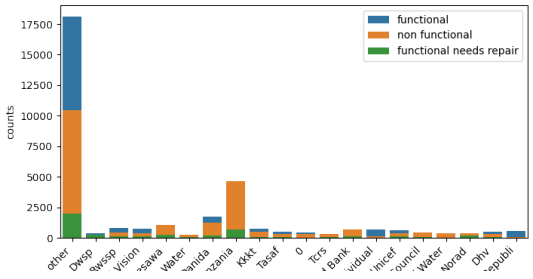

Fig. 5. Funder Categorization

3) Feature Scaling and Standardization
During the analysis of the numerical columns within the dataset, it was observed that certain features exhibit significant difference in scale range. This variation in scale could lead to feature domination during model training, thereby affecting the overall performance and generalization capability of the model. In order to handle this issue and ensure more balanced and effective model training, the features were normalized through scaling. Normalization standardizes the scale of features, reducing the impact of scale differences.

4) Data Discretization

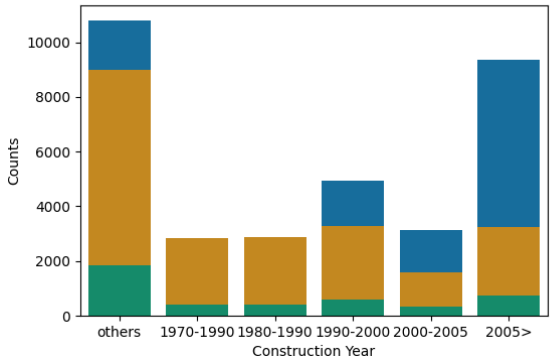

Fig. 6. Construction Year data discretization

In order to manage the continuous data within the dataset, a binning technique was used to convert the data into bins or intervals. The binning of the *construction_year* column was done was to categorize into distinct bins to carry out subsequent analysis. Fig. 6. is a bar chart representing the number of constructions across different time periods.

C. *Algorithm Results*

1) Voting Classification Technique

The performance of ensemble learning approach using the Voting Classification technique with different configuration of base estimators and optimised hyperparameters are shown in TABLE I.

TABLE I
VOTING CLASSIFIER BASE CLASSIFIER CONFIGURATIONS

Configurations	Base Classifier Included
Configuration1	RandomForest, Decision Tree, XGBoost
Configuration2	RandomForest, Decision Tree
Configuration3	Decision Tree, XGBoost
Configuration4	RandomForest, XGBoost

TABLE II
VOTING CLASSIFIER BASE CLASSIFIER CONFIGURATIONS RESULTS

Configurations	F1 Score	Precision	Recall
Configuration1	63	76	61
Configuration2	60	76	58
Configuration3	63	72	60
Configuration4	66	72	64

Configurations	Train Accuracy	Test Accuracy	Balanced
Configuration1	89	79	76
Configuration2	87	78	77
Configuration3	92	79	73
Configuration4	92	79	72

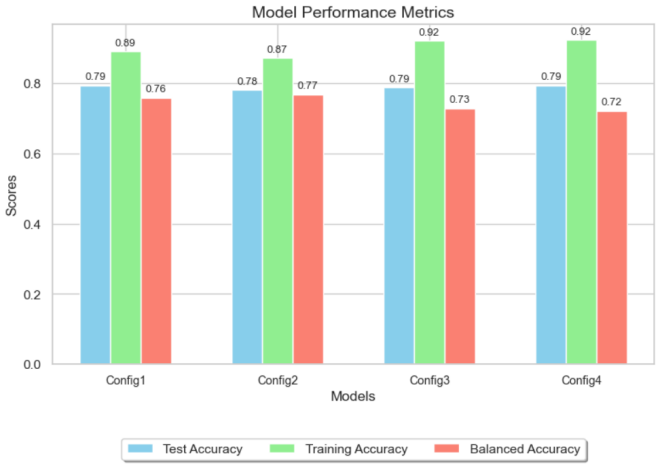

Fig. 7. Accuracy Graph of Voting Technique

In this technique, we have explored four different configurations of voting classifier. Configuration1, with high precision, is ideal for dealing with reducing unnecessary repairs and also perfect for situations where cost saving is essential. Configuration2 is ideal that has the ability to detect all repairs that are needed, which is important for environments where a pump failure could lead to a major problem. Configurations3 has not received good performance metrics , might be suitable for less critical

situation where consequences of not catching repair is minimal. Configuration4 model provides the most balanced performance that can help to catch failure without need of conducting unnecessary repairs.This configuration offers a robust solution by balancing precision and recall, making it ideal for environments that require reliable maintenance .The best accuracy was achieved with hyperparamters *estimator_RandomForest__max_depth*=20, *estimator__RandomForest__n_estimators*=20, *estimator__xgb_model__max_depth*= 10, *estimator__xgb_model__n_estimators*=80.

2) Boosting Technique
The performance of boosting approaches with optimised hyperparameters were calculated and compared in TABLE III. These comparisons were carried out to evaluate the effectiveness of the models in predicting the functionality of water pumps in Tanzania.

TABLE III
BOOSTING CLASSIFIER MODEL RESULTS

Configurations	F1 Score	Precision	Recall
Gradient Boosting	66	72	64
XGBoost	65	74	62

Algorithm	Train Accuracy	Test Accuracy	Balanced
Gradient Boosting	89	79	72
XGBoost	84	79	74

It is found that XGBoost classifier has a slight edge over the Gradient boosting classifier due to a higher F1 Score and Balanced Accuracy. A higher F1 Score is beneficial as it indicates both robustness and reliability in the model's predictive power, addressing both aspects of false positives and false negatives effectively. The best accuracy was achieved with the following hyperparamter combination: *booster* = '*gbtree*', *min_child_weight* = 3, *learning_rate*=0.5 .

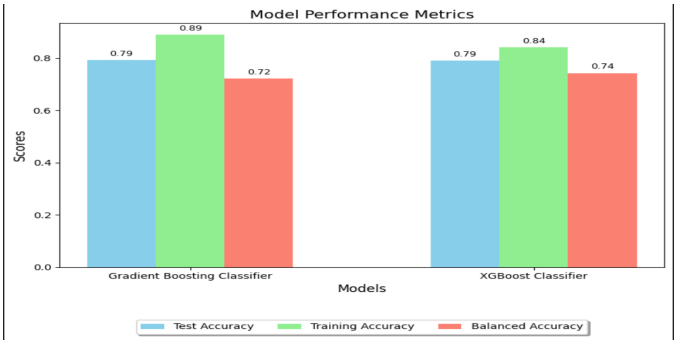

Fig. 8. Accuracy Graph of Boosting Technique

Overall comparisons of accuracies of Gradient Boosting and XGBoost classifiers as shown in Fig. 3. indicates that XGBoost has a marginally better training accuracy, which might suggest it has a slight edge in capturing complex patterns in the data or possibly overfitting the training data slightly more than the Gradient Boosting Classifier.

3) Recursive Feature Elimination Technique
The performance metrics of the Decision Tree with Recursive Feature Elimination (RFE) and the standard Decision Tree are detailed in TABLE IV. Analysis reveals the Decision Tree provides improved precision and a slightly enhanced F1 Score.The addition of RFE does not significantly enchance the Decision Tree's performance instead it slightly reduces the performance of model across the metrics. This could indicate that RFE is removing some features that are beneficial for the Decision Tree for the given dataset. The unchanges recall rate suggests that the model's ability to identify true positives (correctly identifying non-functional or functional-needs-repair) remains constant, regardless of feature selection.Additionally, RFE is generally used to reduce overfitting by eliminating less important features and simplifying the model. In this case, RFE doesn't seem to benefit Decision Tree, as indicated by reduced precison and balanced score.

In situations where minimizing false positives is critical, especially in our scenario where we look to avoid misclassification of a pump that is non-functional or needs repair as functional, Decision Tree is considered to be the better model.The hyperparameter that gave the best result is *estimator__max_depth*=20.

The comparison of the accuracies of the classifier with and without RFE is given in Fig. 9.

TABLE IV
RECURSIVE CLASSIFIER MODEL RESULTS

Algorithm	F1 Score	Precision	Recall
Decision Tree	63	66	61
Decision Tree with RFE	62	65	61

Algorithm	Train Accuracy	Test Accuracy	Balanced
Decision Tree	89	76	66
Decision Tree with RFE	89	75	65

V. DISCUSSIONS

There were various approaches implemented to solve the problem in each phase.
During the data preprocessing phase, several approaches were implemented to handle missing values and categorical features. Approach A1 involved using the mode imputation technique to replace missing values. Additionally, for categorical features with a large number of unique categories, values with a unique count below a certain threshold were replaced with

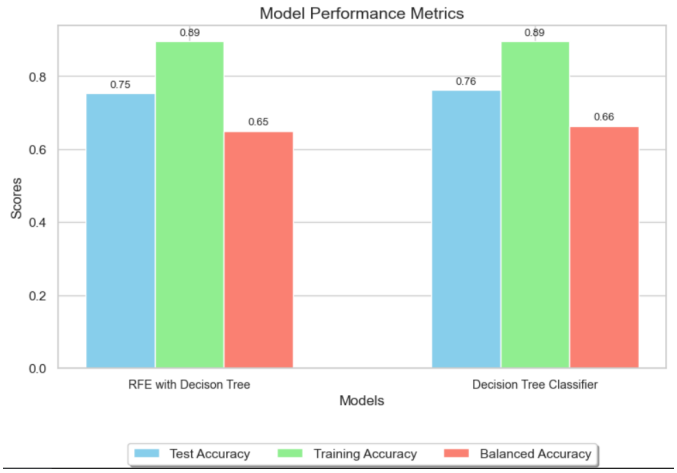

Fig. 9. Accuracy Graph of RFE Technique

the label '*other*'. Another approach, A2, focused on replacing missing values with the mode in numerical columns and using the label '*unknown*' for categorical columns. Approach A3 involved dropping rows containing NaN values entirely from the dataset. On analysis of all the approaches, it was found that Approach A1 yielded better accuracy and also helped reduce the dimensionality of the categorical data.

For the 'construction year' column, Approach A1 employed binning techniques to discretize continuous data into intervals, facilitating the categorization of construction years into meaningful groups based on the trends discovered for construction year column across the target variable. Meanwhile, Approach A2 implemented normalization on the '*construction_year*' data, in order to rescale it to a consistent range to minimize noise and improve data consistency. In contrast, Approach A3 maintained the '*construction_year*' column in its original form without applying any transformation or preprocessing steps.

In our research, various predictive models were studied. Gradient and XgBoost are chosen to handle complex patterns and anomalies like the varied factors influencing water pump functionality.This algorithm are effective at handling at handling imbalanced datasets. Gradient Boosting shows the strong overall performance in recall and balanced accuracy, making it effective for handling true positive. XGBoost, with higher precision, is preferable in scenarios where percentage of false positive is high. In constrast, models such Decision Tree and Decision Tree with Recursive Feature Elimination does not perform very well compared to other Ensemble and Boosting Techniques.

Voting Classifier was taken into consideration in order to design robust and generalized model. Both Voting Classifer with base ensemble configuration as random forest and XGBoost and XGBoost Classifier have achieved almost high balanced scores, indicating these models manage the trade-off between precision and recall effectively. They are robust across varies scenarios,providing a good balance of identifying true positive while minimizing false positives. XGboost and Voting Classifier base configuration(Configuration4) both have highest precision, indicating that both are preferable in scenarios where the cost of false positive is significant.

VI. Conclusion

This paper presents an extensive analysis of machine learning models applied to predicting the functionality of water pumps in Tanzania. Three main machine learning techniques were applied, namely Voting classifier technique, Boosting technique, and Feature Selection technique. The performance of Decision Trees was not up to standards. Although Voting Classifier model and XGBoost Algorithm have same balanced accuracy but the recall value of Voting Classifier is slightly higher than XGboost.The training and test accuracy of the Voting Classifier were comparatively close to those of other algorithms. It is a model with lower complexity, prioritizing generalization over achieving high accuracy on the training data.Based on the analysis of all the evaluation metrics,we decided to go forward with Voting Classifier for testing our model on test file.

In this study, class imbalance was not resolved. Future work may include reapplying and integrating our proposed method with techniques designed to solve class imbalances which include Oversampling (SMOTE/ADASYN), Undersampling (ENN), use of Modified Loss Functions etc.

Model stacking, i.e building a binary classification between *functional* vs *non-functional* and another binary classification between *functional* vs. *functional needs repair*, can be tried to improve accuracy.

Since water resources are dynamic, it may be possible to ensure continued performance by investigating the development of adaptive models that change over time in response to new data.

VII. Contributions

The Data Preprocessing, EDA and Report-writing were joint ventures, while the Voting classifier was implemented by Kunal, Gradient Boosting and XG Boost models were implemented by Shruti, and a Decision Tree model with and without RFE was implemented by Shefali.

References

[1] https://blogs.worldbank.org/en/africacan/tanzania-water-is-life-but-access-remains-a-problem. Accessed on May 5, 2024.
[2] https://medium.com/geekculture/data-science-vs-pump-it-up-competition-cccc8d58bb64/. Accessed on May 8, 2024.
[3] https://water.org/our-impact/where-we-work/. Accessed on May 5, 2024.
[4] Joomi https://joomik.github.io/waterpumps/. Accessed on May 8, 2024.
[5] Mohammed Awad and Salam Fraihat. Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. *Journal of Sensor and Actuator Networks*, 12(5):67, 2023.
[6] Ashfaq Hussain Farooqi, Shahzaib Akhtar, Hameedur Rahman, Touseef Sadiq, and Waseem Abbass. Enhancing network intrusion detection using an ensemble voting classifier for internet of things. *Sensors*, 24(1), 2024.
[7] Richard Harvey and Edward McBean. Understanding stormwater pipe deterioration through data mining. *Journal of Water Management Modeling*, 2014.

[8] Wenjuan Lian, Guoqing Nie, Bin Jia, Dandan Shi, Qi Fan, and Yongquan Liang. An intrusion detection method based on decision tree-recursive feature elimination in ensemble learning. *Mathematical Problems in Engineering*, 2020:1–15, 2020.

[9] Jafar Tanha, Yousef Abdi, Negin Samadi, Nazila Razzaghi, and Mohammad Asadpour. Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7:1–47, 2020.

[10] Rui Wang, Weishan Dong, Yu Wang, Ke Tang, and Xin Yao. Pipe failure prediction: A data mining method. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1208–1218. IEEE, 2013.