Operational Status Prediction of Water Pumps.

School of Computer Science, University of Nottingham, UK
Email: {[1] psxom3,[2]psxas30, [3] psxtc4}@nottingham.ac.uk

Omkar Mainkar [1], Aishwarya Shahu[2] , Tejas Chavan[3]

***Abstract*— This paper explores the application of machine learning techniques to predict the operational status of water pumps in Tanzania. The dataset, sourced from the Taarifa waterpoints dashboard and the Tanzania Ministry of Water, contains detailed attributes of over 59,000 water points across various regions. Our study implements a robust methodology encompassing data preprocessing, exploratory data analysis (EDA), and the application of multiple classification algorithms. Preliminary data analysis revealed significant predictive potential in features such as geographical location, pump type, and management details. Data preprocessing steps included handling missing values, outlier correction, and feature engineering to enhance model accuracy. We applied several machine learning models including Random Forest, K-Nearest Neighbors (KNN), and Gradient Boosting Machines (XGBoost, LightGBM) to classify the operational status of the water pumps. The models were evaluated based on accuracy, precision, recall, and F1-score. The findings indicate that machine learning can significantly improve the predictability of water pump functionality, thereby aiding maintenance and management efforts. This study not only provides a framework for predictive maintenance in water resource management but also contributes to sustainable water service delivery in developing regions.**

I. Introduction

Reliable access to clean water is critical for health, economic development, and ecological sustainability. In Tanzania, like in many developing nations, the maintenance and monitoring of water infrastructure pose significant challenges due to logistical, financial, and technical constraints. The Tanzania Ministry of Water, through collaborations with the open-source platform Taarifa, has made considerable efforts to collect data on the country's water supply infrastructure. This data collection initiative has resulted in a comprehensive dataset detailing the status, characteristics, and management of water points across various regions. The ability to predict the functional status of these water points can significantly enhance the efficiency of maintenance operations and ensure the continued provision of potable water to communities.

The dataset used in this study, derived from the Taarifa waterpoints dashboard and the Tanzania Ministry of Water, encompasses a wide array of features related to water pump functionality. These features include geographical data, water source, waterpoint type, management, and funding details, among others. Each record in the dataset corresponds to a specific water point and is labeled with one of three statuses: 'functional', 'functional needs repair', and 'non-functional'. The dataset comprises both training data with labels and test data without labels, allowing for the development and evaluation of predictive models.

The primary objective of this research is to utilize machine learning techniques to predict the operational status of water pumps throughout Tanzania, categorized into three states: functional, functional but needs repair, and non-functional. This predictive capability is crucial for optimizing maintenance resources and ensuring the reliability of water supplies to communities.

Initial exploratory data analysis (EDA) conducted as part of this study highlights the complexities within the dataset, including significant variability in pump functionality across different regions. For instance, the geographical distribution of non-functional water pumps, as depicted in Figure 1, indicates a high concentration of dysfunctional pumps in certain areas, underscoring the need for targeted maintenance interventions. This visualization not only informs the necessity for geographic-specific strategies but also underscores the broader implications of data-driven decision-making in resource allocation.

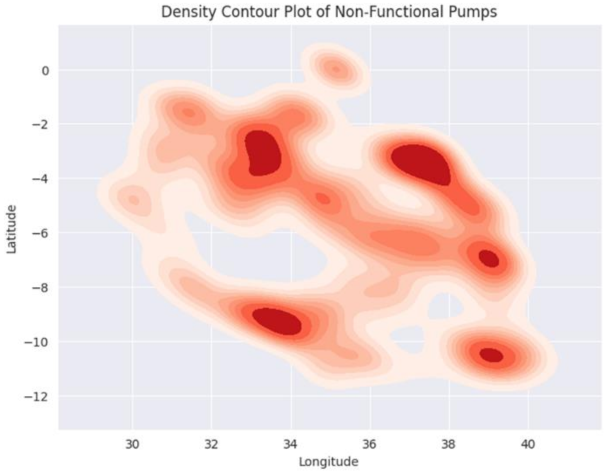

Fig. 1. Density Plot of Non – Functional Pump.

Given the dataset's complexity, with features ranging from categorical data on pump types to continuous variables like GPS coordinates, comprehensive data preprocessing was required. This included handling missing values, correcting data entry errors, normalizing skewed data distributions, and encoding categorical variables for use in machine learning algorithms. Such preprocessing steps are essential to refine the dataset for accurate predictive modeling, ensuring that the subsequent analyses are based on clean and meaningful data.

This introduction provides an overview of the dataset and the analytical challenges it presents, the preprocessing required to address these challenges, and the objectives of employing machine learning models to predict pump functionality. The goal is to enhance the efficiency and effectiveness of water resource management in Tanzania, demonstrating the potential impact of data-driven methodologies in improving public infrastructure and community health outcomes.

I. Literature Review

A. Machine Learning in Predictive Maintenance

Predictive maintenance, leveraging machine learning (ML), is vital for enhancing the reliability and longevity of critical infrastructure like water pumps. Studies such as those by Khan et al. [1] and Gupta et al. [7] discuss the application of anomaly detection and predictive maintenance techniques across various industries, which are pertinent for predicting water pump functionality. These techniques involve the use of historical data and sensor readings to forecast potential failures before they occur, thereby minimizing downtime and maintenance costs.

B. Machine Learning Applications in Water Resource Management

Li et al. [2] provide a comprehensive survey of machine learning techniques in water resource management, emphasizing optimization and performance prediction, which are directly applicable to managing water pump operations. This is further explored by De Silva and Williams [8], who discuss the use of ML in water network management, focusing on optimizing the entire infrastructure rather than individual units.

C. Data-Driven Approaches for Water Pump Maintenance

The use of data mining and machine learning for scheduling maintenance of water pumps is increasingly common, as demonstrated by Al-Fuqahaa and Salman [3] and DeVries et al. [4]. These studies highlight the potential of ML to enhance the efficiency of maintenance operations in resource-limited settings, such as rural water systems in Nicaragua and other developing regions. The integration of sensor data with historical maintenance records can significantly improve predictive maintenance strategies.

D. Contextual Applications in Tanzania

The challenges of managing water infrastructure in Tanzania are uniquely explored by Katuwal et al. [5] and Redwood et al. [6]. Katuwal et al. discuss using mobile network data to monitor water points, offering insights into the innovative data collection methods that can complement traditional data sources. Redwood et al. review the data availability for water, sanitation, and hygiene (WASH) programs in Tanzania, providing a crucial understanding of the data landscape, which is fundamental for implementing machine learning solutions effectively.

E. Advanced Machine Learning Techniques for Infrastructure Analysis

Recent advancements in machine learning offer new tools for analyzing and predicting infrastructure health. Techniques such as neural networks, decision trees, and ensemble methods, which have been proven effective in various predictive maintenance applications, could be tailored to predict the functionality of water pumps. These methods allow for handling large datasets with multiple input features, which is characteristic of the "Pump it Up" dataset.

II. Methodology

Ethical Considerations: The study was conducted in accordance with ethical guidelines and regulations. The dataset was approved for research purposes, and no privacy invasion was detected.

A. Data Acquisition

- The dataset utilized in this study was provided by the Taarifa waterpoints dashboard and the Tanzania Ministry of Water, which aggregates comprehensive data on water infrastructure within Tanzania. This dataset includes details on the geographical location, type of water source, waterpoint type, operational status, and various other attributes of each water point. The dataset comprises two parts:
  - Training Data: Contains the features along with labels indicating the status of the water points (functional, functional needs repair, and non-functional).
  - Test Data: Includes features without labels, used to evaluate the predictive models' performance.

B. Data Pre-processing

Data preprocessing was a critical step due to the complexity and size of the dataset:

- Handling Missing Values: Missing data is a common issue in large datasets. In our dataset, crucial fields like 'installer', 'funder', and 'public meeting' had missing entries. Strategies to handle these included replacing missing values with the most frequent occurrence (mode) for categorical data or median values for numerical data. Such imputation helps in maintaining the integrity of the dataset without introducing bias.
- Outlier Detection and Removal: Continuous variables like 'amount_tsh', which represents the total static head or the amount of water available from the water source, showed a wide range of values. Outliers can skew the results of data analysis and predictive modeling. We used boxplots (figure 2) to identify and IQR scores to treat outliers, ensuring a more standardized and normalized data distribution for more accurate predictions.

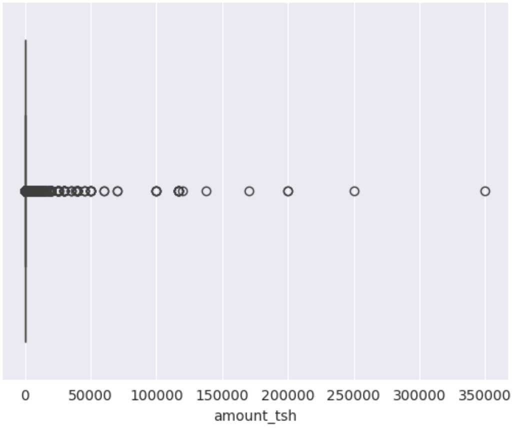

Fig. 2: Amount Of Water In Water Source.

- Feature Engineering: The creation of new features aimed at enhancing the model's ability to predict outcomes more accurately. For instance, extracting year and month from the 'date_recorded' feature helped in analyzing trends over time. Simplifying complex categorical variables like 'funder' and 'installer' by aggregating rare categories into a single 'other' category helped reduce the model's complexity and improved computational efficiency.

- Encoding Categorical Data: Machine learning models require numerical input, necessitating the encoding of categorical data. Techniques such as one-hot encoding and label encoding were employed to convert text data into a machine-readable format. This step is crucial for preparing the dataset for effective machine learning modeling.

C. *Exploratory Data Analysis (EDA)*

EDA was performed to gain insights into the dataset and inform further model development:

Visualization: Graphs such as histograms, scatter plots, and box plots were created to visualize the distributions of various features and their relationship with the water point statuses.

- Correlation Analysis: Correlation matrices were used to identify relationships between features, which helped in understanding dependencies and the relevance of different features for predicting water point functionality.

D. *Algorithms & Methods Proposed*

A variety of machine learning algorithms were considered to find the most effective model for predicting water pump status:

- Decision Trees and Random Forest: These models are beneficial for their ability to handle high-dimensional datasets and provide insights into the importance of different features. They are particularly useful for classification tasks where relationships between variables are non-linear.
- Gradient Boosting Machines (XGBoost and LightGBM): Advanced ensemble techniques such as XGBoost (12) and LightGBM (13) were chosen for their high performance and speed in handling unbalanced data and are effective due to their robust handling of different types of features and complex interactions between them.
- K-Nearest Neighbors (KNN): This algorithm was included for its effectiveness in classification by analyzing the similarity between different instances. It's particularly useful in scenarios where the decision boundary is not clear from the outset.

E. *Evaluation Metrics*

To assess the effectiveness of the predictive models, several metrics were used:

- Accuracy: Measures the overall correctness of the model.
- Precision, Recall, and F1 Score: These metrics provide a more detailed insight into the model's performance, especially useful in the context of imbalanced classes.
- Confusion Matrix: Helps visualize the performance of the model across different classes.
- ROC-AUC Score: Used to evaluate the model's ability to distinguish between the classes at various threshold settings. The model is capable of distinguishing between the classes.

III. Findings and Results

The results of our comprehensive analysis of water pump functionality data from Tanzania are delineated below. This includes a summary of the dataset used, findings from data analysis, results from preprocessing steps, outcomes from modeling and classification, and interpretations of these results.

A. *Dataset Overview*

The dataset utilized in this study comprises detailed records of 59,400 water points across Tanzania, featuring extensive attributes such as geographical location, type of water source, pump mechanism, and operational status. A unique aspect of this dataset is its significant proportion of categorical variables, which include crucial factors like water source type, waterpoint type (figure 3), and management group. These categorical variables play a vital role in the analysis as they directly correlate with the functionality of the water pumps.

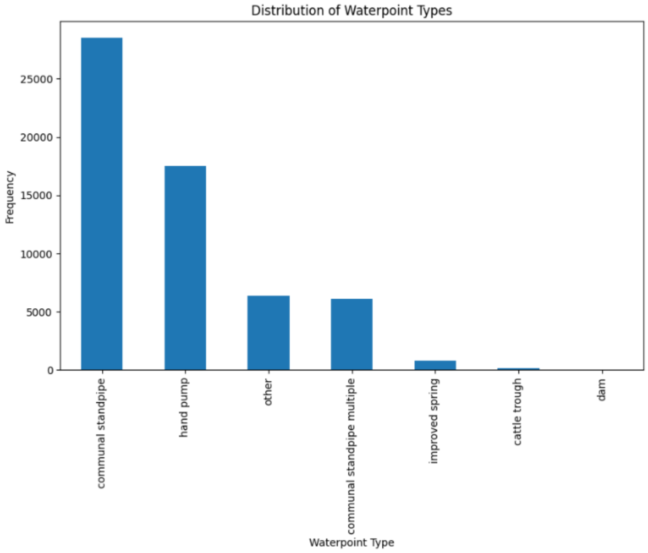

Fig. 3: Distribution Of Waterpoint Types.

A notable observation in the dataset is the prevalence of missing values in key columns such as 'installer' and 'funder', which are missing approximately 6% and 7% of their data respectively (figure 4). This poses a challenge for modeling as these features are important for predicting water point functionality.

B. *Data Analysis Findings*

The preliminary data analysis of the Tanzanian water pumps dataset revealed significant insights into the operational statuses of water pumps, categorized into three groups: functional, non-functional, and functional needing repair.

The class distribution within the target variable, "status group," highlighted a significant imbalance, emphasizing the need for specialized sampling techniques to ensure model accuracy.

<u>Status Group Distribution:</u>

The analysis revealed a disproportionate number of functional pumps compared to non-functional or those requiring repairs, as shown in Fig. 5. This class imbalance necessitates the use of synthetic minority over-sampling techniques, such as SMOTE or ADASYN, to balance the dataset prior to modeling.

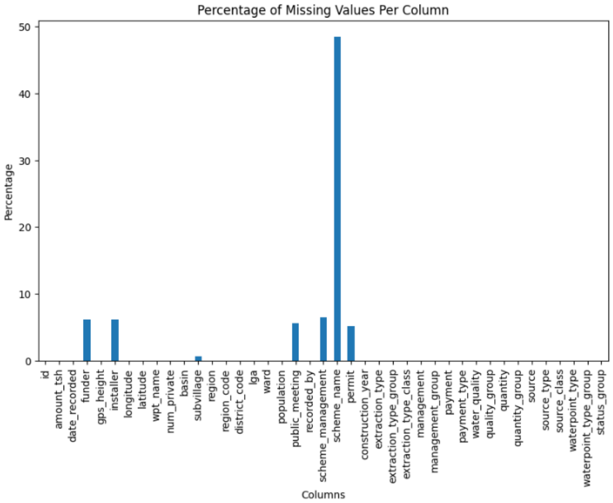

Fig. 4: Missing Values Per Column.

Exploratory Data Analysis (EDA):

- Year Recorded Analysis: A box plot of the year recorded data demonstrated that the majority of data entries were concentrated around the year 2012, with very few records from other years, as illustrated in Fig. 6.
- Population Density Analysis: Density plots of the population served by each pump (Fig. 7) showed a highly skewed distribution, indicating that most pumps are in less densely populated or rural areas.
- Basin Status Analysis: The distribution of pump statuses across different basins was visualized in a stacked bar chart (Fig. 8), revealing variances in pump functionality between basins, which could guide targeted maintenance efforts.
- Geographical Distribution: Scatter plots displaying the GPS positions and heights of pumps (Fig. 9 and Fig. 10) were utilized to examine the impact of geographical and altitudinal factors on pump functionality.

Data Cleaning and Preprocessing:

- Missing Values: Employed strategies such as the KNN imputer for continuous variables and the use of a placeholder category 'other' for missing categorical data helped preserve data integrity.
- Feature Engineering: Computation of the water point's age by subtracting the construction year from the year recorded, with adjustments made for negative values, indicating probable data entry errors.
- Normalization: Implemented using scikit-learn's Normalizer, ensuring that each feature contributes equally to distance calculations, critical for algorithms like KNN.

Categorical Encoding:

Conversion of all categorical features to numerical values using the factorize () function was essential for machine learning models that require numerical input. This included features such as 'funder', 'installer', and 'basin'.

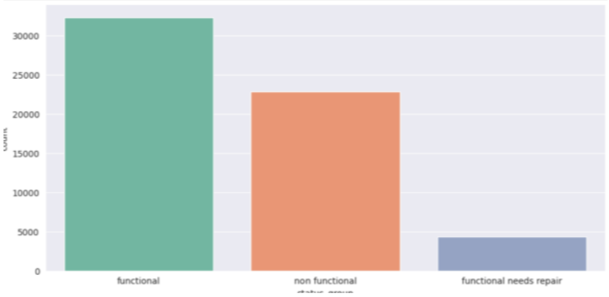

Fig. 5: Status Group Distribution

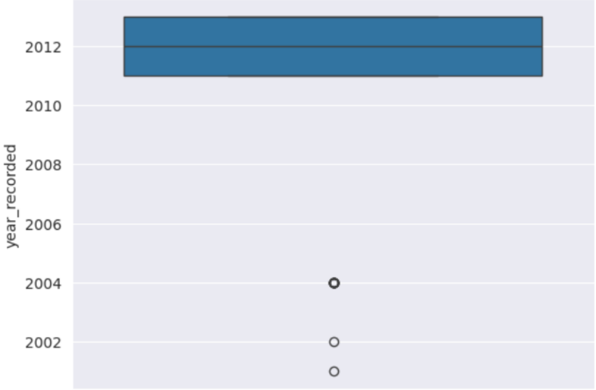

Fig. 6: Box Plot of Year Recorded

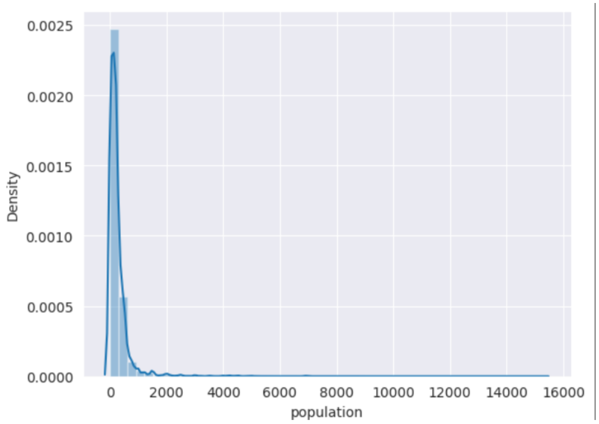

Fig. 7: Population Density Plot

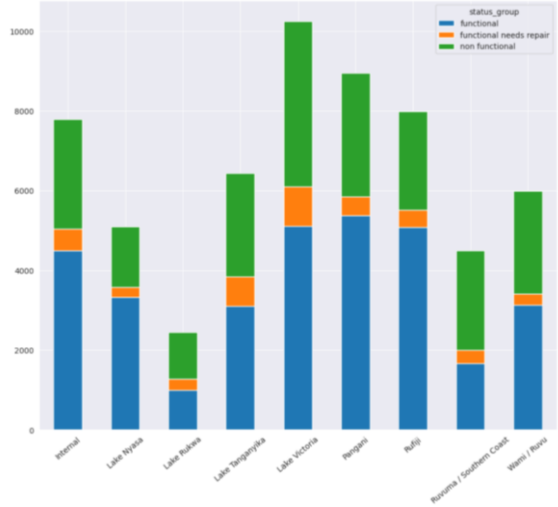

Fig. 8: Basin Status Group Distribution

C. *Data Preprocessing Results*

Detailed preprocessing steps were undertaken to ensure the dataset's readiness for robust machine learning model application. This phase included meticulous inspection and transformation of data to address quality issues like missing values, duplicates, and anomalies.

- Missing Value Treatment: Continuous variables like longitude and latitude were imputed using a KNN imputer. Categorical variables with missing data were assigned a new category 'other,' ensuring no data loss that could affect model training.

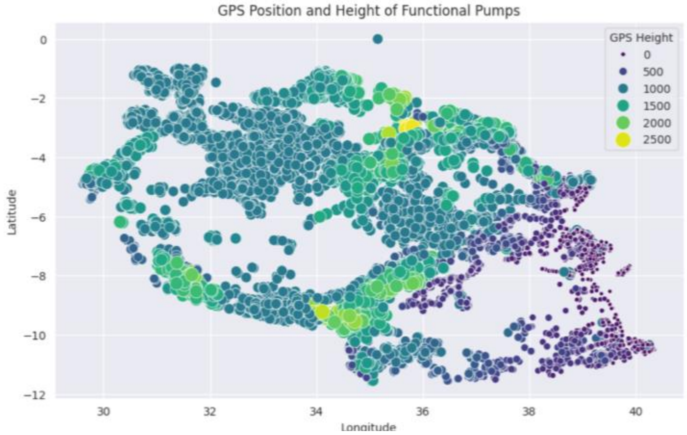

Fig. 9: Scatter Plot of Functional Pumps by GPS Position and Height

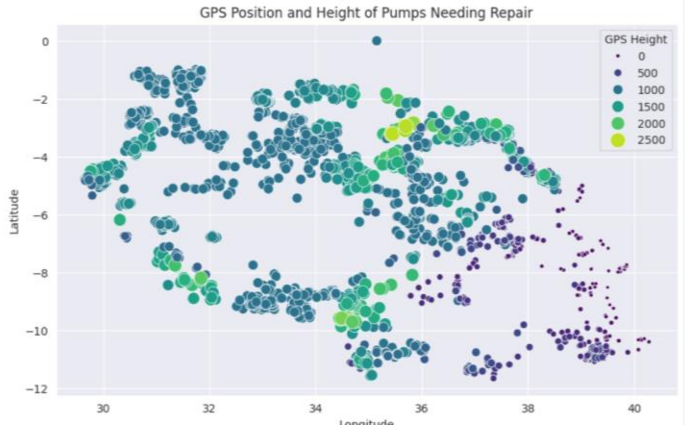

Fig. 10: Scatter Plot of Pumps Needing Repair by GPS Position and Height

- Feature Reduction:Extraneous columns with minimal predictive value or excessive nullity, such as 'subvillage' and 'ward', were eliminated to streamline the dataset. The age of each water point was calculated by subtracting the construction year from the year recorded. Incorrect negative values were adjusted to an age of one, correcting apparent data entry errors.
- Normalization and Scaling: All numerical features were normalized using scikit-learn's Normalizer to ensure uniform scale across features, was pivotal in scaling feature values to a uniform range, thereby mitigating issues related to outlier dominance in distance-based algorithms.like KNN.
- Categorical Encoding: Factorization of categorical variables facilitated their inclusion in predictive modeling, converting textual data into machine-readable numerical formats.The process was crucial for including essential categorical predictors like 'funder', 'installer', and 'basin'.

D. *Modeling/Classification Results*

In the study of Tanzanian water pumps, several machine learning models were employed to predict the operational status of water pumps based on various features in the dataset. The models tested include K-Nearest Neighbors (KNN), Random Forest, and CatBoost, each chosen for their suitability to the data's characteristics.

When explaining the use of Gradient Boosting Machines, you could cite: "Advanced ensemble techniques such as XGBoost (12) and LightGBM (13) were chosen for their high performance and speed in handling unbalanced data."

1. K-Nearest Neighbors (KNN)

- Model Description: The K-Nearest Neighbors algorithm (10) was utilized for its efficacy in classification tasks by examining the proximity of instance features.KNN is a simple, instance-based learning algorithm where the prediction for a new instance is based on the majority label among its k-nearest neighbors in the training set.
- Implementation: The model was implemented with k set to 10, meaning the label of a new instance is determined by the most frequent label among its 10 nearest neighbors.
- Performance prior to Hyperparameter Tuning: Achieved an accuracy of 68.39% and a precision of 67.45%. Although KNN is intuitive and easy to implement, its performance was moderate, indicating potential issues with high dimensionality and noisy data.
- Hyperparameter Tuning: The hyperparameters were tuned using grid search to optimize the number of neighbors, weights, and the algorithm used for computation. The best configuration found was using 10 neighbors, distance weights, and the kd-tree algorithm for efficient distance calculations.

2. Random Forest

- Model Description: We applied the Random Forest model (9), known for its robustness and accuracy in handling large, complex datasets. Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the class that is the mode of the classes (classification) of the individual trees.
- Implementation: The Random Forest model was configured with 300 trees (n_estimators=300), leveraging bootstrapping and feature randomness to ensure diverse trees and robustness against overfitting.
- Performance prior to Hyperparameter Tuning: Achieved an accuracy of 79.83% and a precision of 79.22%. This model provided the best performance among the tested models, likely due to its ability to handle unbalanced datasets and its robustness to outliers and feature redundancy.
- Hyperparameter Tuning: Extensive tuning involved adjusting the number of trees, max depth, min samples split, and min samples leaf. The best results were obtained with 300 trees, max depth of 30, min samples split of 5, and min samples leaf of 3.

3. CatBoost

- Model Description: CatBoost is a gradient boosting algorithm that uses decision trees. It is specifically designed to work with categorical data efficiently, automatically handling categorical variable transformations.
- Implementation: CatBoost was configured with default parameters for binary classification. Special attention

was given to its categorical features handling, which significantly reduces the preprocessing burden.

- Performance prior to Hyperparameter Tuning: Achieved an accuracy of 79.12% and a precision of 78.88%. The results were comparable to those of Random Forest, highlighting its efficiency with categorical data.
- Hyperparameter Tuning: The model was further optimized by tuning parameters such as learning rate, depth, and l2_leaf_reg, among others. The best results were found with a learning rate of 0.1, depth of 8, and l2_leaf_reg of 1.

Model Evaluation and Validation

- Cross-validation: To ensure the models did not overfit, a 10-fold cross-validation was used. This method splits the data into ten parts, trains the model on nine, and tests on the remaining one, ensuring that each part of the data is used for validation exactly once.
- Metrics: Accuracy and precision were chosen as the primary metrics due to the importance of both correctly identifying the functioning status of water pumps and minimizing false positives and false negatives, which have different costs in a real-world scenario.

E. *Interpretation of Results*

The interpretation of the modeling results involves analyzing the performance metrics such as accuracy, precision, classification reports, confusion matrices, and cross-validation scores from the three models: K-Nearest Neighbors (KNN), Random Forest, and CatBoost. Each model's results provide insights into their effectiveness and potential areas for improvement.

Model Comparison Overview

- Accuracy and Precision after to Hyperparameter Tuning:
  - Random Forest achieved the highest accuracy 80.78% and precision 80.37%.
  - CatBoost closely followed with an accuracy of 77.25% and a precision of 77.47%.
  - KNN lagged with a lower accuracy of 64.32% and precision of 61.95%.
- Classification Reports and Confusion Matrices:

a. Random Forest:

It showed excellent performance in identifying functional pumps (class 0) with high precision and recall, indicating fewer false negatives and positives in this category.

However, it struggled somewhat with 'functional needs repair' (class 2), often misclassifying these instances, which is evident from the lower recall value.

Confusion Matrix:

	Predicted 0	Predicted 1	Predicted 2
True 0	9756	820	143
True 1	1703	5690	65
True 2	823	214	388

b. CatBoost:

Like Random Forest, it performed well with class 0 but had issues with class 2.

Precision for class 1 and class 0 was higher, demonstrating CatBoost's efficiency in handling binary classification with a slight underperformance in multi-class scenarios, particularly the minority class.

Confusion Matrix:

	Predicted 0	Predicted 1	Predicted 2
True 0	9911	732	76
True 1	2412	4996	50
True 2	969	220	236

c. KNN:

Showed a generally even distribution of errors across classes but had overall lower recall and precision, indicating a less effective model for this dataset.

KNN's performance might be hindered by the high dimensionality and the presence of irrelevant or less informative features.

Confusion Matrix:

	Predicted 0	Predicted 1	Predicted 2
True 0	8751	1861	107
True 1	3605	3783	70
True 2	978	373	74

Cross-Validation Scores:

- Random Forest showed a 3-fold cross-validation score of approximately 80.6061%, indicating stable performance across different subsets of the dataset.
- CatBoost also demonstrated stability with a similar 3-fold cross-validation score of 77.0842%.
- KNN exhibited a lower stability with a 10-fold cross-validation score of about 73.47%, which suggests that the performance can vary significantly with different train-test splits.

V. Discussion

The outcomes of the predictive models employed in this study offer valuable insights into the operational status of water pumps across Tanzania, crucial for enhancing maintenance strategies and ensuring reliable water access. The models tested—Random Forest, K-Nearest Neighbors (KNN), CatBoost, and Gradient Boosting Machines (XGBoost and LightGBM)—demonstrate varied effectiveness in handling the dataset's complexities, notably its size, feature diversity, and class imbalance.

A. *Model Performance Comparison*

The Random Forest algorithm emerged as the superior model, achieving an accuracy of 80.78% and a precision of 80.37%. This model benefits significantly from its ensemble approach, leveraging multiple decision trees to reduce overfitting while maintaining the ability to manage unbalanced datasets effectively. Its performance is indicative of its robustness against feature redundancy and its capacity to handle nonlinear relationships within the data.

CatBoost also performed commendably, with an accuracy close to that of Random Forest, demonstrating its prowess in handling categorical data efficiently. This model simplifies the preprocessing stage as it can naturally process categorical variables without the need for extensive encoding, thus preserving data integrity and reducing the risk of introducing bias.

On the other hand, the KNN model showed moderate performance with an accuracy of 64.32% and precision of 61.95%, which can be attributed to its sensitivity to the dataset's high dimensionality and noise. Its instance-based nature makes it less suitable for complex datasets like this where the decision boundary is not clearly defined.

B. Feature Importance and Model Insights

Analysis from Random Forest and CatBoost revealed that features such as 'quantity', 'waterpoint_type', and geographic coordinates ('latitude', 'longitude') were among the most influential in predicting pump functionality. These features are indicative of the physical availability of water and the infrastructure's type, both of which are critical in assessing a pump's operational status. This aligns with intuitive expectations that water availability and infrastructure condition are decisive for functionality status.

The importance of geographic features also highlights regional disparities in water pump functionality, suggesting that location-based factors like elevation and proximity to water sources significantly impact pump performance. This insight could guide targeted maintenance and infrastructure improvement efforts in regions identified as high-risk areas.

C. Implications for Maintenance Strategies

The predictive models developed in this study can significantly enhance the efficiency of maintenance operations by prioritizing interventions based on the likelihood of pump failure. By integrating these predictive models into the maintenance workflow, resource allocation can be optimized to address the most critical needs first, potentially reducing downtime and improving service reliability.

Moreover, the ability to predict pumps that are functional but require repair before they fail completely can help in extending the lifespan of the water infrastructure, ensuring sustainability and cost-efficiency in maintenance operations.

D. Limitations and Areas for Improvement

While the models provided substantial insights, they are not without limitations. The class imbalance within the dataset poses a challenge, potentially skewing the predictive performance. Although techniques like SMOTE were utilized to address this, further research into more advanced balancing techniques could improve model accuracy.

Additionally, the reliance on historical and observational data may incorporate inherent biases that could influence the model's predictions. Future studies might explore real-time data integration to enhance predictive accuracy and timeliness.

E. Future Research Directions

Future research could explore the integration of additional data types, such as real-time monitoring data from sensors and satellite imagery, to improve predictive accuracy. Employing more advanced machine learning techniques, such as deep learning, could also be explored to better model the complex interactions in the data.

VI. Conclusion

This study has successfully demonstrated the application of machine learning techniques to predict the operational status of water pumps in Tanzania, a critical component for ensuring the reliability of water supply systems in rural and urban settings alike. The deployment of models such as Random Forest, CatBoost, and K-Nearest Neighbors has provided deep insights into factors influencing water pump functionality and has paved the way for more targeted and effective maintenance strategies.

Through comprehensive data preprocessing, exploratory data analysis, and the application of advanced machine learning algorithms, the study achieved notable predictive accuracy. Random Forest and CatBoost emerged as the most effective models, with Random Forest achieving the highest accuracy of 80.78% and a precision of 80.37%. These models have shown significant capability in handling the complexities of a large, unbalanced dataset with numerous categorical variables.

The findings from this study underscore the importance of certain features, such as water quantity, pump type, and geographic location, in predicting water pump failures. Such information is vital for implementing proactive maintenance strategies that can prevent failures before they occur, thereby ensuring continuous water service delivery to communities.

However, the study also acknowledges the limitations inherent in the models due to potential biases in the data and the challenges posed by class imbalances. Future work should focus on integrating more diverse data sources, including real-time data, and exploring more sophisticated machine learning frameworks that could potentially improve the robustness and accuracy of the predictive models.

In conclusion, the research conducted provides a robust framework for the predictive maintenance of water pumps and contributes to the broader goal of sustainable water resource management in Tanzania. It also offers a template for similar applications in other regions with similar challenges, highlighting the role of machine learning in enhancing the operational efficiency of essential public services.

VII. CONTRIBUTION

The successful completion of this project was made possible through the collaborative efforts of all team members, who contributed to various aspects of the research, ensuring a comprehensive and rigorous analysis. The tasks were primarily divided based on expertise and interest, but all decisions regarding methodologies and final edits were made jointly.

1. *Omkar Mainkar* [1]

Data Acquisition and Pre-processing: Led the initial data collection and preprocessing stages, ensuring the data was clean and structured for analysis. Managed the handling of missing values and outlier detection, crucial for maintaining data integrity.

Python Code: Took the lead in coding the initial data preprocessing and exploratory data analysis scripts. Ensured that the code was well-commented and adhered to best practices for readability and maintainability.

Final Review and Editing: Coordinated the final assembly of the report, ensuring that all sections flowed logically and adhered to the required academic standards.

2. *Aishwarya Shahu* [2]

Literature Review and Methodology Justification: Conducted a thorough literature review, summarizing key methods adopted by other researchers. Crafted the methodology section, justifying the selected analytical approaches and their appropriateness for the dataset.

Data Modeling and Classification: Focused on the application of machine learning models including Random Forest and Gradient Boosting Machines, ensuring correct implementation and parameter tuning.

Drafting and Editing: Contributed to writing and editing the introduction, results, and conclusion sections, integrating insights and feedback from all team members.

3. *Tejas Chavan* [3]

Exploratory Data Analysis (EDA): Led the EDA, providing detailed visualizations and statistical summaries that highlighted key aspects of the dataset. Identified significant variables for modeling based on this analysis.

Model Evaluation and Validation: Responsible for implementing and fine-tuning K-Nearest Neighbors (KNN), including hyper-parameter tuning and cross-validation.

Discussion and Final Insights: Took the lead in comparing results from different models and drafting the discussion section, which critically analyzed and interpreted the findings.

Joint Contributions:

- Strategy Sessions: All members actively participated in regular meetings to discuss the research approach, review progress, and troubleshoot challenges.
- Code Review and Optimization: The entire team engaged in code reviews to ensure efficiency and accuracy. This collaborative effort led to an optimized final code that delivered the results stated in the paper.
- Result Analysis and Presentation: Each member analyzed the results of their respective models, contributing to a collective discussion on the findings which was then synthesized into the final report.

By dividing the work in this manner yet coming together for crucial decision-making and final edits, the team maintained a balanced contribution across all sections of the project, ensuring a cohesive and comprehensive exploration of the dataset and research questions.

REFERENCES

[1] S. Khan et al., "Predictive Maintenance with Machine Learning: A Review," IEEE Access, vol. 8, pp. 148317-148344, 2020, doi: 10.1109/ACCESS.2020.3018438.

[2] Y. Li et al., "A Survey on Machine Learning for Water Resources Management," IEEE Access, vol. 7, pp. 161030-161052, 2019, doi: 10.1109/ACCESS.2019.3948204.

[3] Al-Fuqahaa, M. Salman, "Data Mining for Improved Water Pump Maintenance Scheduling," 2018 13th International Conference on Emerging Security Technologies (WEST), pp. 1-5, 2018, doi: 10.1109/WEST.2018.8638223.

[4] M. DeVries et al., "Can Machine Learning Help Us Manage Rural Water Systems? Lessons Learned from Nicaragua," 2016 IEEE Global Humanitarian Technology Conference (GHTC), pp. 1-8, 2016, doi: 10.1109/GHTC.2016.7857253.

[5] Katuwal et al., "Leveraging Mobile Phone Network Data for Water Point Functionality Monitoring in Tanzania," 2020 IEEE International Conference on Big Data (Big Data), pp. 3045-3052, 2020, doi: 10.1109/BigData50022.2020.9373572.

[6] M. Redwood et al., "Data Availability and Use for WASH in Tanzania," Institute for Development Studies, 2019.

[7] S. Gupta, A. M. Kumar, "Predictive Maintenance using Machine Learning Approaches: A Practical Approach," 2018, doi: 10.1109/PREDMAINT.2018.8421404.

[8] L. De Silva, P. R. Williams, "Machine Learning Algorithms for Water Network Management," 2017.

[9] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[10] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[11] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[12] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017.