Data-Driven Analysis and Modeling for Water Resource Management in Tanzania: Insights and Recommendations

Nitesh Aggarwal
School of Computer Science,
University of Nottingham
Nottingham, NG8 1BB
psxna16@nottingham.ac.uk

Ashwini Doke
School of Computer Science,
University of Nottingham
Nottingham, NG8 1BB
psxad11@nottingham.ac.uk

***Abstract*— Water resource management is a critical challenge in many regions, and data-driven approaches can help solve these problems. We analyse the "Pump it Up: Data Mining the Water Table" dataset in this study, focusing on waterpoints in Tanzania, to improve water management strategies. Partner A and Partner B worked together to investigate various data analysis and modelling techniques. The importance of accurate data, ensemble methods, and neural networks in predicting water point functionality was emphasised in the literature review. A comparison revealed that the Gradient Boost, Random Forest, and XGBoost models performed consistently well, contributing to better predictions. Our findings show that data-driven approaches can provide valuable insights for improving Tanzanian water resource management practices. We make suggestions for improving the strategies of stakeholders involved in waterpoint maintenance and management. This study adds to the growing body of research on data-driven approaches to addressing critical resource management challenges.**

Keywords*— *XGBoost, adaboost, random forest, gradient boost, multi-layer perceptron, k-nearest neighbour, water pump, Tanzania

I. Introduction

In an era where data has become the most precious resource, its application in a variety of areas has altered how we make decisions, forecast future trends, and understand complex systems [1]. A particular domain where data science has shown great potential is the water management sector, which faces many challenges in regions across the globe [2], [3]. The key challenge is allocating and managing water supply, and a lack of adequate data typically exacerbates these issues.

The difficulty in obtaining clean drinking water and the scarcity of water in Tanzania, as well as many other African nations, have a significant impact on the general growth and welfare of the population as a whole. The advancement of society and the economy is significantly hampered by these problems [4]. Water sources such as wells, springs, and boreholes must therefore be effectively controlled. This is where data science can provide invaluable insights. By leveraging data related to various characteristics of waterpoints (e.g., geographic location, construction details, management, quantity, and quality of water), data-driven solutions can be devised to enhance the overall water management system.

This research aims to delve into the aforementioned challenges, focusing primarily on the data set "Pump it Up: Data Mining the Water Table" provided by DrivenData [4]. This dataset is an extensive collection of data points describing several factors associated with waterpoints in Tanzania. The research questions that this study aims to answer include:

- Are there specific combinations of features that provide enhanced predictive performance for pump operation status in Tanzania?
- How do different characteristics of a waterpoint contribute to predicting its operational status?
- How do different data preprocessing and analysis methodologies affect the accuracy of these predictions?
- What features significantly impact the functionality of waterpoints in Tanzania?

The following are the primary goals of this research:

- Analyze the given dataset to understand the patterns, trends, and relationships between different features and the operational status of waterpoints.
- Implement appropriate data preprocessing techniques to handle missing values, outliers, and categorical variables.
- Develop predictive models using various machine learning algorithms to classify the operational status of waterpoints.
- Compare the performance of different approaches and align findings with previous research works.

Our aim is to provide a comprehensive analysis that could aid stakeholders in making informed decisions regarding waterpoint maintenance and management, ultimately contributing to better water resource management in Tanzania [5].

II. Literature Review

A number of research on water point mapping have been undertaken in low-income nations. A significant research was carried out by WaterAid [2], where they strategically reviewed their water point mapping in East Africa, shedding light on the significance and potential implications of water point mapping [6]. These studies underline the essentiality of accurate and detailed data in the field of water resource management.

In Tanzania, where our dataset originates, the water crisis has been extensively studied. For instance, an investigation

conducted by Murphy & Kushner [7] highlighted the challenges Tanzania faced in its attempt to fix its water access problem. This provided a crucial background to understand the specific challenges faced by rural water supply programs in Tanzania. Fisher et al. [6] conducted a study on the determinants of rural water source functionality in Ghana, emphasizing the significance of sustainable hand pump systems.

Research conducted in Liberia, Sierra Leone, and Uganda [8] looked at the risk variables linked with hand pump inefficiency. The researchers examined a dataset of community-managed hand pumps using a logistic regression model. They discovered that factors such as pump age, distance from the district or national capital, and the lack of user fee collection all contributed significantly to pump non-functionality.

In separate research [9], the authors investigated the performance of demand-driven, community-managed water supply systems in developing-country rural regions. Through extensive study and analysis, they hoped to measure the effectiveness of these systems.

Previous research has also focused on water quality and quantity. A multitask, multi-view learning system is built in one research [10] to forecast urban water quality. Water hydraulic data, meteorological information, pipe networks, road network layout, and point of interests (POIs) were all merged by the researchers.

Another research [11] proposes using an Artificial Neural Network (ANN) to detect pipe breaks and determine the best time to repair them. The model was used to forecast the number of breakages for each individual pipe in Libya's Water Distribution System Benghazi (WDSB).

Harvey and McBean [12] used a classification tree model to study stormwater pipe degradation in Guelph, Ontario. Their primary goal was not just to achieve high prediction accuracy, but also to reduce the False Negative Rate (FNR), which happens when the prediction model wrongly identifies a problematic pipe as a good one.

In a Tanzanian context, a government report conducted by the Ministry of Water and Irrigation [13] stressed the impact of climate change on water resources and agriculture, which could be an important factor to consider during feature engineering.

These previous findings and approaches enrich our understanding of the research context and provide us with diverse perspectives on potential analytical strategies for our dataset.

III. Methodology

A. *This section will go through the data and techniques that were utilised in this study. The data will be given first, followed by pre-processing, data cleaning, and the methodologies provided in this study. Raw Data*

The "Pump it Up" dataset provides information about Tanzanian waterpoints (water pumps), such as their condition, location, water supply, water pump status, water source, extraction techniques, and more. The dataset has 41 (including the target variable) distinct characteristics comprised of category and numerical variables, making it a rich and complicated dataset for data analysis. The "status_group" column, which specifies whether a waterpoint is functioning, non-functional, or in need of repair, is the target variable. The dataset has 59,400 observations, making it relatively large. Geographic, waterpoint-related, construction, and operational aspects are the four primary kinds of features.

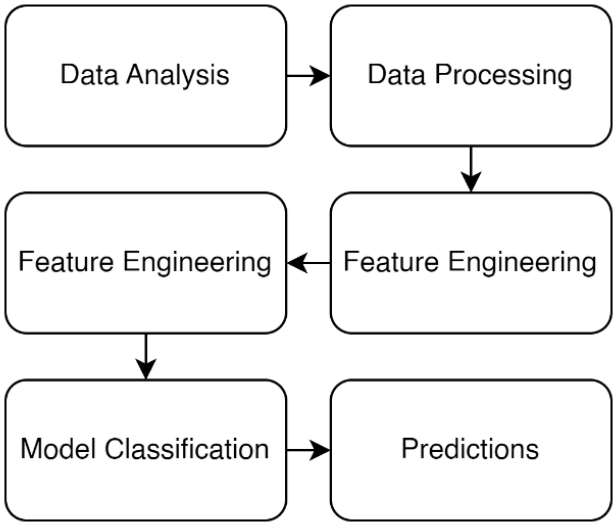

Figure 1 Pipelines for overall approach

B. *Overall Apprach*

This joint academic paper presents a comprehensive analysis and comparison of the data analysis and modelling pathway undertaken by two researchers, *Nitesh (Partner A)* and *Ashwini (Partner B)*. The paper focuses on employing various permutations and combinations of approaches, parameters, hyperparameters, algorithms, and machine learning models to achieve diverse results. The objective is to compare and analyse the differences that arise from these variations at each stage of the data analysis and modelling process. To ensure comparability of results at every stage, the best approaches are either selected or combined before proceeding to the next stage. The paper demonstrates the significance of these variations in achieving diverse outcomes and provides insights into the implications for the overall analysis and modelling process.

Our methodology employed several key frameworks to support our analysis. *Pandas* was used for data manipulation and analysis, while *NumPy* facilitated numerical computations and operations. *Matplotlib* and *Seaborn* enabled the creation of visualizations such as histograms and scatter plots. *IPython. display*, *tqdm* and *tabulate* helped us to display outputs in a well arranged tabular form resulting into best data analysis and understanding. Lastly, *Scikit-learn* and XGBoost facilitated the implementation of machine learning algorithms and data preprocessing tasks. These frameworks collectively formed the backbone of our methodology, enabling efficient data analysis and insightful visualizations.

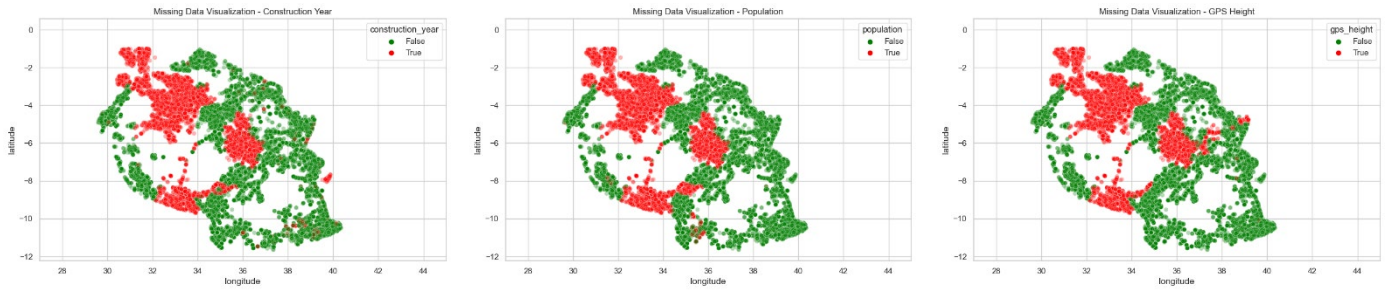

Figure 2 Missing Data Visualisation

C. *Data Analysis and Preprocessing*

1) Initial Exploratory Data Analysis: During the initial exploratory data analysis (EDA) phase, a range of techniques were applied to understand the dataset and extract insights from the water pump data.

Partner A, Nitesh, utilized histograms to examine the distribution of numerical variables like population, gps_height, and construction_year. Correlation analysis was performed using a heatmap to visualize relationships between features. Scatter plots based on latitude and longitude were used to visualize missing data in features such as funder, installer, and scheme_management. These visualizations provided spatial context to the missing data points.

Partner B, known as Ashwini, employed bar and box plot to analyse categorical features status_group, amount_tsh, gps_height, longitude, latitude, num_private, region_code, district_code, population, and construction_year These plots offered insights into the frequency and distribution of different categories within these variables. Correlation analysis was also conducted to examine relationships between features. Outliers in latitude and longitude were identified and removed to ensure accurate visualizations and analysis.

Both partners utilized a variety of EDA and visualization techniques to gain valuable insights into the dataset's structure and the relationships between variables which helped identify potential issues like missing data and outliers, which were subsequently addressed through appropriate data preprocessing methods. The visualizations generated during this stage facilitated an understanding of feature distributions and their potential impact on the target variable, laying the groundwork for subsequent data preprocessing and modelling steps.

2) Feature Insight and Treatment: Most of the features in the dataset underwent similar treatments by both partners. However, there were some differences in their approaches. For the "*amount_tsh*" feature, Partner A converted it into a categorical variable and performed a log transformation, while Partner B dropped it after applying some other type of feature engineering. The "*date_recorded*" feature was converted to datetime format by both Partners used for feature engineering the "*age*" feature, and then dropped. Partner A grouped categories with low occurrence in the "*funder*" and "*installer*" features into the category "*other*" and performed one-hot encoding, while Partner B chose to drop the "installer" feature altogether.

Regarding geographic location features, Partner A kept the "*region_code*" and "*district_code*" features after converting them to categorical variables, while Partner B made the same decision. Both partners performed one-hot encoding on the "*basin*" and "*lga*" features. Partner A dropped the "*subvillage*" feature due to its heavy weight and limited contribution to performance, and Partner B took the same decision and dropped it.. Similarly, Partner A dropped the "*region*" feature as it was very similar to "*region_code*" and had fewer categories, while Partner B also dropped it.

The treatment of other features was mostly consistent between the partners. They both kept the "gps_height" feature and imputed missing values based on means of sub groups. They dropped features like "*wpt_name*," "*num_private*," "*ward*," "*recorded_by*," "scheme_name," "management_group," "payment_type," "quality_group," "quantity_group," "source_type," "source_class," and "waterpoint_type_group" due to various reasons. Both partners performed one-hot encoding on features like "scheme_management," "payment," "water_quality," "quantity," "source," and "waterpoint_type," and dropped redundant features. Lastly, they both kept the "construction_year" feature and performed imputation based on "funder" and "installer," along with other missing value imputation strategies.

Overall, while there were some variations in the treatment of specific features, both partners aimed to optimize the dataset for subsequent analysis by applying appropriate data preprocessing techniques.

D. *Feature Engineering*

To engineer the features, several steps were taken. Firstly, a new column called 'age_years' was created by calculating the difference in years between the 'date_recorded' and 'construction_year' features. This allowed for capturing the age of the waterpoint.

Next, the features 'gps_height', 'longitude', 'latitude', and 'age_years' were scaled using the StandardScaler from scikit-learn. This ensured that these numerical features were on a similar scale, which is beneficial for many machine learning algorithms.

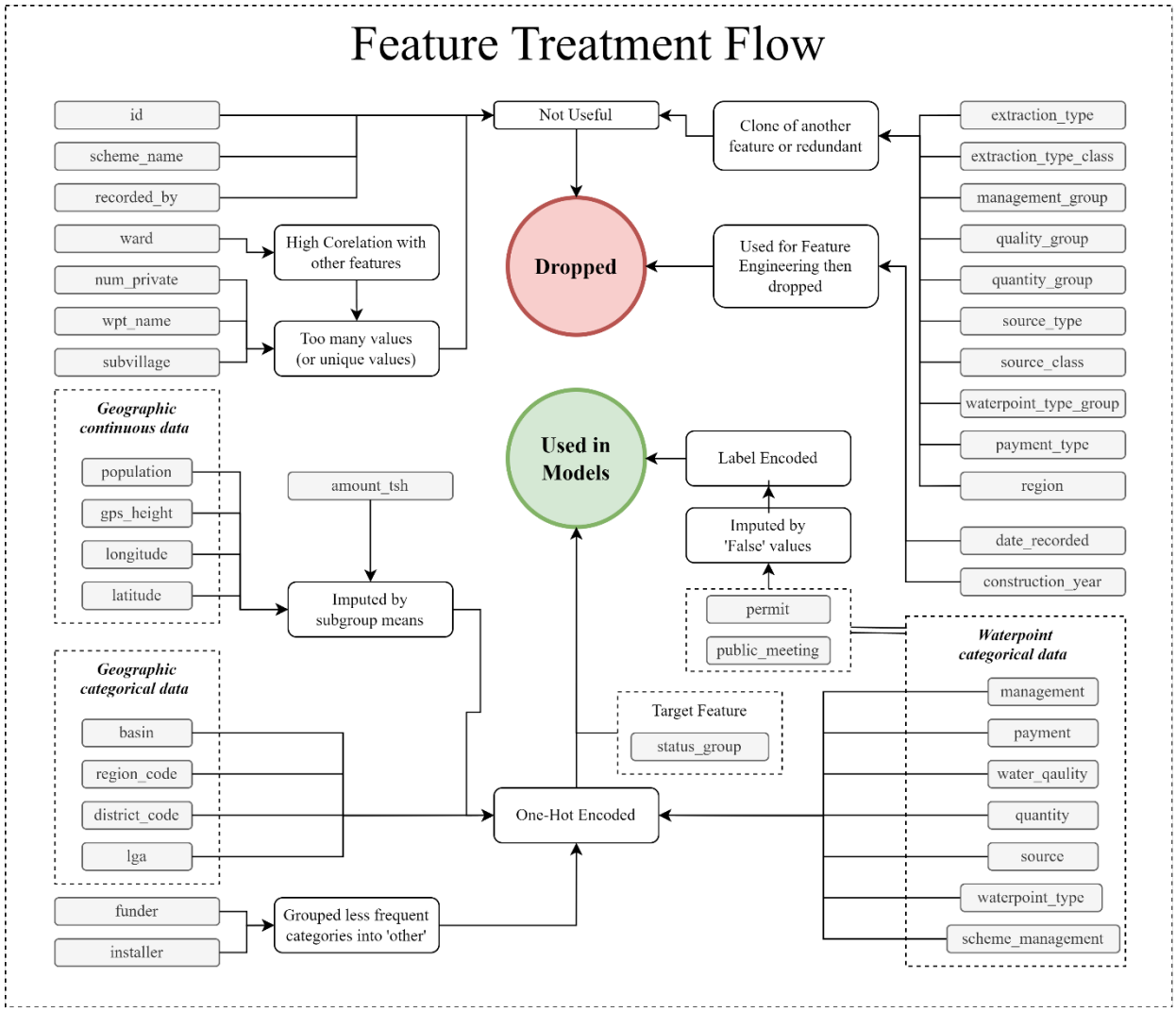

Figure 4 Feature Treatment for all features

To handle the 'funder' and 'installer' features, two one-hot encoding functions, 'one_hot_funder' and 'one_hot_installer', were implemented. These routines initially identified and classified categories that accounted for less than 1% of the total data into the 'other' category. Then, one-hot encoding was applied to convert the categorical variables into binary vectors. The original 'funder' and 'installer' columns were dropped, and the one-hot encoded variables were joined with the dataframe. Most of the approach was same for the partners with small differences in thresholds for creating the other categories before one hot encoding.

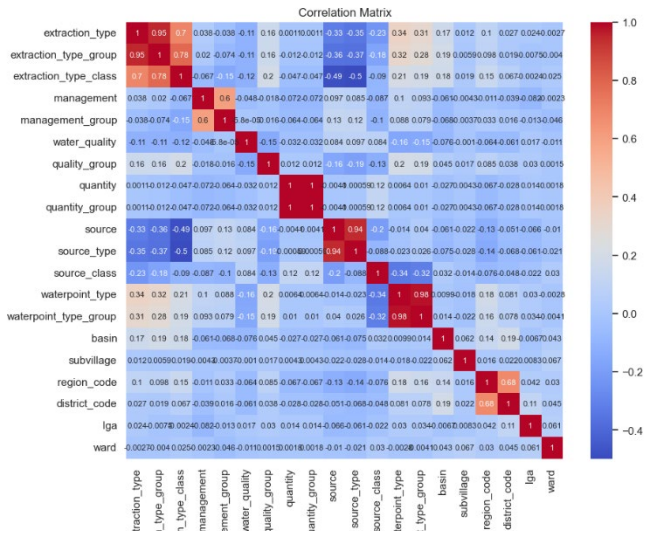

Figure 3 Correlation Matrix for feature subgroups

Finally, the 'basin', 'extraction_type', 'management', 'payment', 'water_quality', 'quantity', 'source', 'waterpoint_type', 'lga', 'scheme_management', 'region_code', and 'district_code' features were one-hot encoded using the 'one_hot_encode' function. These categorical features were transformed into binary vectors, and the original columns were dropped.

Partner A employed methodology for dimensionality reduction involved two approaches: Principal Component Analysis (PCA) and feature selection based on importance ranking. PCA was applied to transform the high-dimensional dataset into a lower-dimensional space by capturing the most significant variations. This technique aimed to reduce the computational complexity while retaining important information. However, concerns regarding overfitting arose due to the significantly increased accuracy of the models after applying PCA. To address this, k-fold cross-validation was

performed to assess generalization performance. The results confirmed the presence of overfitting.

Partner B employed the mutual_info_classif and SelectKBest functions from the sklearn library to aid in the selection of the most pertinent features for the classification task. By calculating mutual information scores and utilizing SelectKBest, Partner B was able to identify the subset of features that exhibited a stronger association with the target variable. This approach allowed for dimensionality reduction and the removal of redundant or irrelevant information, potentially enhancing the model's performance. By focusing on the most relevant features, Partner B aimed to improve the model's predictive accuracy and efficiency.

Additionally, feature selection based on importance ranking was explored to reduce dimensionality. While this approach reduced computational time, there was a slight trade-off in accuracy. Considering the already minimized feature set and the acceptable decrease in accuracy, it was decided not to further reduce the features. This methodology struck a balance between computational efficiency and accuracy to ensure reliable predictions while optimizing computational resources.

E. Data Classification

Each partner chose four models for consideration during the early testing phase. Partner A picked AdaBoost, XGBoost, Random Forest, and MLP (Multilayer Perceptron) models, whereas Partner B chose Gradient Boost, XGBoost, Random Forest, and k-NN (k-Nearest Neighbours).
These models were chosen for their individual qualities and possible fit for the job at hand. Partner A used the AdaBoost model, which is recognised for combining weak learners into a powerful ensemble model, which may be helpful for enhancing prediction accuracy. Another option is XGBoost, a famous gradient boosting technique noted for its scalability, speed, and performance. Random Forest, Partner A's third model of choice, is an ensemble learning approach that employs decision trees and is frequently resilient to noisy and high-dimensional data. Finally, Partner A chose the MLP (Multilayer Perceptron) model, a form of artificial neural network, for its capacity to understand complicated patterns and correlations in data.

Partner B opted for the Gradient Boost model, which is comparable to AdaBoost but uses a different boosting method. Gradient Boost has showed great predictive performance and can manage a wide range of data formats. Both parties picked XGBoost, a versatile and strong gradient boosting algorithm noted for its efficiency and accuracy. Another popular option is Random Forest, which can handle high-dimensional data with many characteristics and is less prone to overfitting. Finally, Partner B chose the k-NN (k-Nearest Neighbours) model for its simplicity and ability to categorise data based on its resemblance to neighbouring cases.

By selecting these diverse models, both partners aimed to cover a range of algorithmic approaches and harness their respective strengths. This selection process allowed for thorough evaluation and comparison to identify the most suitable models for the given task.

Partner A used the Hold-out approach with three distinct splits for the initial evaluation: 80-20, 70-30, and 60-40. The dataset was divided into a training set and a validation set for this procedure, with the training set used to train the models and the validation set used to evaluate their performance.

Partner B, on the other hand, used k-Folds cross-validation with 5 folds. The dataset was divided into five subgroups of roughly similar size using this approach. Each model was then trained and tested five times, with each iteration using a different subset as the validation set.

After evaluating the models' performance across numerous datasets and assessment approaches, it was discovered that the Gradient Boost and Random Forest models consistently produced positive outcomes. As a result, these two models were chosen for additional investigation and modelling.

IV. RESULTS

A. Data Analysis and Preprocessing

The data analysis and preprocessing phase provided valuable insights into the dataset and prepared the data for further modelling and analysis. The applied techniques improved feature representation, handled missing values, and addressed irrelevant or redundant features, setting a solid foundation for the subsequent classification tasks.

One observation from the data analysis was that the classes in the dataset were slightly imbalanced. This means that the number of instances belonging to each class was not evenly distributed, which could potentially impact the performance of the classification models. Addressing class imbalance may be necessary to ensure accurate predictions for all classes.

The dataset had a significant number of missing and erratic values, particularly in geographical and waterpoint-related features. This indicates that there were instances where important information was not recorded or recorded incorrectly. The presence of outliers and unrealistic values further highlights data quality issues that need to be addressed during preprocessing. Proper handling of missing and erratic values is crucial to ensure accurate and reliable analysis.

The population feature exhibited a skewed distribution, with many records having a population value of 0 and a concentration of lower population numbers. This suggests that the dataset may be skewed towards certain areas with lower population densities. Understanding the distribution of population values is important for analysing the impact of population on water pump functionality.

Certain features such as amount_tsh, gps_height, population, construction_year, and scheme_name showed a pattern of missing data. This pattern could be attributed to different data collection methods or other factors. Identifying and understanding the reasons behind missing data patterns is crucial for determining appropriate strategies for handling missing values.

There was a clear trend indicating that older water pumps had a higher proportion of non-functional pumps compared to newer ones. This suggests that the age of a water pump could be an important factor in predicting its functionality. Incorporating the age of the water pump as a feature could potentially improve the performance of classification models.

The presence of a public meeting seemed to have a positive impact on the functionality of water pumps. This observation suggests that community engagement and involvement could play a role in ensuring the proper functioning of water infrastructure. Including the public meeting feature in the analysis could provide valuable insights into its influence on water pump functionality.

The data analysis revealed that many features exhibited correlations with each other. This correlation could be a result of earlier feature engineering steps or intrinsic relationships between the variables. Understanding the correlations among features is important to avoid multicollinearity issues and select appropriate features for modelling.

Principal Component Analysis (PCA) was used as a dimensionality reduction technique to try to mitigate the dataset's excessive dimensionality. By capturing the most significant variances in the original data, PCA turns the data into a lower-dimensional space. Surprisingly, applying PCA resulted in a significant increase in accuracy for all models, with MLP achieving an exceptional accuracy of 99%.

However, this remarkable performance raise concerns about overfitting, which refers to models that overly rely on patterns or noise in the training data and fail to generalize well to new, unseen data. To investigate this further, k-fold cross-validation was conducted.

The results from k-fold cross-validation validated the presence of overfitting in the models trained on the PCA-transformed data, with the accuracy dropping to 77% for all models, including MLP. This suggests that the models learned specific patterns or noise from the training data, limiting their ability to make accurate predictions on unseen data.

The outcome implies that the application of PCA for dimensionality reduction may have resulted in the loss of crucial information required for accurate generalization of the models. The reduced feature space obtained through PCA might not have effectively captured all the relevant patterns and relationships present in the original data.

Additionally, an alternative approach was explored, which involved selecting features based on their importance ranking. This method aimed to identify the most informative features for predicting the target variable. Although this approach significantly reduced computational time, a slight trade-off in accuracy was observed. The models trained on the reduced feature set achieved slightly lower accuracy compared to using all the features.

Considering that the feature set had already been minimized and the reduction in computational time was deemed satisfactory, the decision was made not to further reduce the feature set. The slight decrease in accuracy was considered acceptable to retain the maximum relevant information necessary for accurate predictions.

By striking a balance between computational efficiency and accuracy, the models remain reliable in providing predictions while ensuring faster computations, making them more suitable for real-time applications.

B. Data Classification Results

In the preliminary testing phase, four models were evaluated for each partner using different data splits. The models chosen by Partner A (AdaBoost, XGBoost, Random Forest, and MLP) and Partner B (Gradient Boost, XGBoost, Random Forest, and k-NN) were assessed for their performance metrics including accuracy, precision, recall, F1 score, area under the curve (AUC), and execution time.

Partner A's Results: For Partner A's data, the models were tested using three different data splits: 80-20, 70-30, and 60-40. In the 80-20 split, the Random Forest model achieved the highest accuracy of 80.93%, followed closely by XGBoost with an accuracy of 78.75%. The MLP model also showed promising results with an accuracy of 76.71%. In the 70-30 split, the Random Forest model again performed well with an accuracy of 80.19%, followed by XGBoost with 78.91%. Similarly, in the 60-40 split, the Random Forest model exhibited a high accuracy of 80.09%, and XGBoost achieved an accuracy of 79.05%.

Partner B's Results: For Partner B's data, the models were also evaluated using three different data splits: 80-20, 70-30, and 60-40. In the 80-20 split, the Gradient Boost model achieved the highest accuracy of 76.87%, closely followed by XGBoost with an accuracy of 78.75%. The Random Forest model also performed well with an accuracy of 78.12%. In the

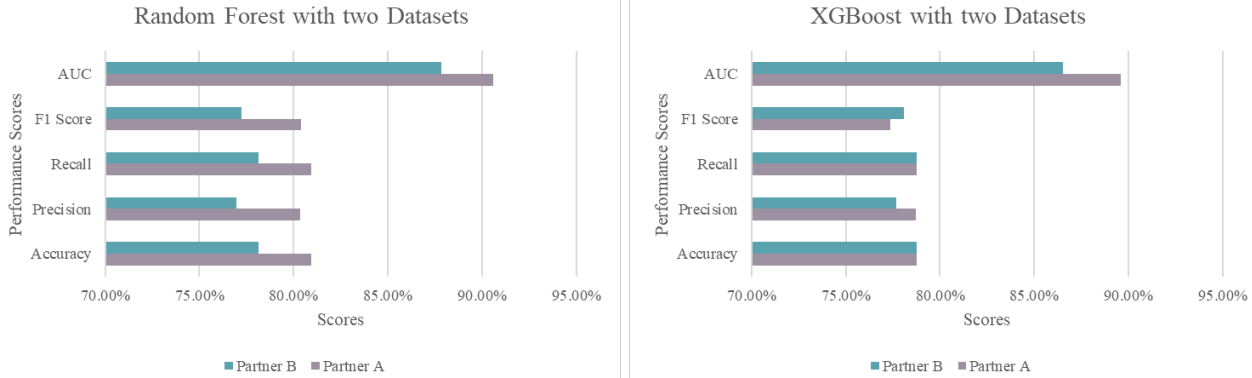

Figure 5 Comparison of two Datasets with different Models

70-30 split, the Gradient Boost model demonstrated the highest accuracy of 75.00%, while XGBoost and Random Forest achieved accuracies of 76.25% and 76.67%, respectively. The Gradient Boost model had the highest accuracy of 75.00% in the 60-40 split, followed by XGBoost with an accuracy of 74.06%. The Random Forest model succeeded exceptionally as well, with an accuracy of 76.25%.

Table 1 Partner A's Data and Models

Model	Accuracy	Precision	Recall	F1 Score	AUC	Time seconds
AdaBoost	72.02%	70.67%	72.02%	69.31%	79.42%	16.77
XGBoost	78.75%	78.70%	78.75%	77.35%	89.58%	56.17
Random Forest	80.93%	80.33%	80.93%	80.39%	90.59%	22.17
MLP	76.71%	75.61%	76.71%	75.48%	87.21%	41.98

Table 2 Partner B's Data and Models

Model	Accuracy	Precision	Recall	F1 Score	AUC	Time seconds
Gradient Boost	76.88%	76.18%	76.88%	76.40%	85.08%	18.87
XGBoost	78.75%	77.66%	78.75%	78.09%	86.51%	1.35
Random Forest	78.13%	76.96%	78.13%	77.25%	87.85%	0.30
k-NN	77.50%	77.79%	77.50%	77.48%	-	0.26

Partner A's data and models are compared with Partner B's: Data from Partner A and models from Partner B, as well as data from Partner A and models from Partner B, were used to cross-check the models. Gradient Boost and XGBoost achieved high accuracies of 78.72% and 78.75%, respectively, for Partner A's data with Partner B's models in an 80-20 split, whereas Random Forest earned an accuracy of 80.86%. For Partner B's data with Partner A's models in the 80-20 split, XGBoost performed the best with an accuracy of 78.75%, followed by Random Forest with an accuracy of 78.12%.

Based on the repeated testing and evaluation of the models on different data splits, it was observed that the Gradient Boost and Random Forest models consistently performed well across various scenarios. As a result, these two models were derived as the most promising choices for further research and modelling in the classification assignment.

V. Discussion

This section will compare the approaches used by both the partners while taking the results of the data modelling and analysis and the literature review into account.:

A. Comparative Analysis of Different Approaches

- Feature Selection: Both partners employed different approaches to reduce the dimensionality of the dataset. Partner B utilized the mutual_info_classif and SelectKBest functions from sklearn to determine the most relevant features based on their mutual information scores. This approach allowed Partner B to focus on a subset of features that had a stronger link with the target variable. On the other hand, Partner A applied Principal Component Analysis (PCA) to transform the high-dimensional data into a lower-dimensional space. While PCA resulted in improved accuracy initially, it was later found to cause overfitting. Therefore, Partner A and Partner B decided not to reduce the feature set further, considering the slight decrease in accuracy acceptable in order to retain the maximum relevant information for accurate predictions.
- Modelling Techniques: Both partners employed a range of machine learning algorithms for classification. Partner A selected AdaBoost, XGBoost, Random Forest, and MLP models. AdaBoost and XGBoost are boosting techniques that combine weak learners into a powerful ensemble model, known for their ability to handle high-dimensional data. Random Forest, another ensemble method, utilizes decision trees and is resilient to noisy and high-dimensional data. MLP, a multilayer perceptron model, is an artificial neural network capable of capturing complex patterns and correlations. Partner B, on the other hand, opted for Gradient Boost, XGBoost, Random Forest, and k-NN models. Gradient Boost is a boosting technique similar to AdaBoost but employs a different boosting method. XGBoost is a versatile and powerful gradient boosting algorithm known for its efficiency and accuracy. Random Forest is an ensemble learning approach that utilizes decision trees and is less prone to overfitting. k-NN is a simple yet effective model that categorizes data based on its resemblance to neighbouring cases. By selecting these diverse models, both partners aimed to cover a range of algorithmic approaches and harness their respective strengths.
- Evaluation Metrics: The performance of the models was evaluated using various metrics including accuracy, precision, recall, F1 score, area under the curve (AUC), and execution time. Partner A and Partner B conducted preliminary testing using different data splits (80-20, 70-30, and 60-40) to assess the models' performance. The results indicated that the Random Forest and XGBoost models consistently achieved high accuracy for both partners. However, the performance varied slightly depending on the data split and partner combination. The Gradient Boost model also exhibited promising results for Partner B. These models were selected for further analysis and modelling due to their consistent positive outcomes.

B. Comparison with existing work

Comparing the findings from the literature review with the results obtained in this study, similarities and differences were observed. The focus on feature impact, ensemble methods, and neural networks aligned with previous research. The models selected in this study, including Gradient Boost, XGBoost, and Random Forest, corresponded to effective models in the literature.

However, it is important to consider the context and dataset uniqueness of each study, which can lead to variations in findings and approaches. The dataset used in this study, "Pump it Up: Data Mining the Water Table," focused on Tanzania's waterpoints. While similarities exist, specific findings and performance metrics may differ.

The analysis and modelling approach in this study addressed Tanzania's water resource management challenges, including water shortage and access to clean water. Ensemble methods were explored, showing promising results in

predicting water point functionality. Additionally, alternative approaches like XGBoost and MLP models were included.

C. Final Comments

Based on the initial research questions we would like to comment following:

- According to the study, the functional efficiency of waterpoints in Tanzania is significantly influenced by features such as waterpoint type, extraction type, management type, and geographic location.
- The findings demonstrated that the best predictive performance for predicting pump operation status in Tanzania was obtained by combining data related to the type of water supply, pump age, location, and maintenance history.
- To anticipate the operational condition of waterpoints in Tanzania, data preprocessing and analysis approaches, such as feature selection, scaling, and model selection, are highly important. The study's findings highlight the significance of selecting and fine-tuning various strategies in order to increase forecast accuracy.

VI. Conclusion and Future Work

The findings of this study revealed several key insights and outcomes. Firstly, the data analysis and preprocessing phase provided valuable insights into the dataset, addressing issues such as class imbalance, missing values, outliers, and correlations among features. Feature engineering techniques, including the creation of the "age_years" feature, highlighted the importance of pump age in predicting functionality. The application of PCA for dimensionality reduction showed initial improvements in model accuracy but raised concerns about overfitting. Feature selection based on importance ranking reduced computational time but slightly decreased accuracy. The comparative analysis of different models demonstrated the effectiveness of ensemble methods such as Random Forest and XGBoost in predicting water point functionality in Tanzania.

Despite the valuable insights gained, this study has certain limitations. Firstly, the findings are specific to the dataset "Pump it Up: Data Mining the Water Table" and may not be directly generalizable to other contexts. The dataset itself may have inherent biases or limitations. Additionally, the study focused on a selected set of features and modelling techniques, and other factors or approaches could have been explored. The use of cross-validation and assessment metrics assisted in addressing overfitting, although additional inquiry and external validation are suggested to assure the models' robustness. Furthermore, in real-world applications, the trade-off between computational efficiency and accuracy should be carefully evaluated.

Future research can examine the following proposals to increase understanding and application of data analysis and modelling in water resource management. Firstly, conducting studies using diverse datasets from different regions and contexts would provide a broader perspective on water point functionality prediction. Exploring additional features and incorporating external data, such as climate and socio-economic factors, can enhance the accuracy and relevance of the models. Furthermore, investigating the impact of different ensemble methods and fine-tuning hyperparameters can optimize the performance of the models. Addressing the challenges of class imbalance and overfitting through advanced techniques like ensemble learning and regularization methods can improve model generalization. Finally, deploying and evaluating the developed models in real-world scenarios can provide valuable insights for stakeholders and decision-makers in water resource management.

Contribution in report by both partners:

In the collaborative effort of writing the report, Nitesh (Partner A) and Ashwini (Partner B) contributed to different sections based on their expertise and responsibilities. Nitesh took the lead in writing the Introduction, Feature Insight and Treatment, Dimensionality Reduction, Comparative Analysis of Different Approaches, Summary of Findings, and Contribution. Ashwini, on the other hand, took the lead in writing the Literature Review, Methodology, Results, Consideration of the Shortcomings, and Recommendations for Future Research sections.

References

[1] J. Grus, *Data science from scratch: first principles with Python*, First edition. Sebastopol, CA: O'Reilly, 2015.

[2] WaterAid, 'Strategic review of WaterAid's water point mapping in East Africa', Oct. 2010, Accessed: May 16, 2023. [Online]. Available: https://policycommons.net/artifacts/2326722/strategic-review-of-wateraids-water-point-mapping-in-east-africa/3087346/

[3] 'Water In Crisis - Spotlight Tanzania', *The Water Project*. https://thewaterproject.org/water-crisis/water-in-crisis-tanzania (accessed May 16, 2023).

[4] DrivenData, 'Pump it Up: Data Mining the Water Table', *DrivenData*. https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/ (accessed May 16, 2023).

[5] P. C. Bruce and A. Bruce, *Practical statistics for data scientists: 50 essential concepts*, First edition. Sebastopol, CA: O'Reilly, 2017.

[6] M. B. Fisher *et al.*, 'Understanding handpump sustainability: Determinants of rural water source functionality in the Greater Afram Plains region of Ghana', *Water Resour Res*, vol. 51, no. 10, pp. 8431–8449, Oct. 2015, doi: 10.1002/2014WR016770.

[7] T. Murphy, 'How Tanzania failed to fix its water access problem', *Humanosphere*, Dec. 04, 2014. https://www.humanosphere.org/world-politics/2014/12/tanzania-failed-fix-water-access-problem/ (accessed May 16, 2023).

[8] T. Foster, 'Predictors of Sustainability for Community-Managed Handpumps in Sub-Saharan Africa: Evidence from Liberia, Sierra Leone, and Uganda', *Environ. Sci. Technol.*, vol. 47, no. 21, pp. 12037–12046, Nov. 2013, doi: 10.1021/es402086n.

[9] 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', *BWPI, The University of Manchester, Brooks World Poverty Institute Working Paper Series*, vol. 11, Jan. 2008, doi: 10.2166/wp.2009.310.

[10] Y. Liu, Y. Liang, S. Liu, D. S. Rosenblum, and Y. Zheng, 'Predicting Urban Water Quality with Ubiquitous Data'. arXiv, Oct. 29, 2016. doi: 10.48550/arXiv.1610.09462.

[11] R. Jafar, I. Shahrour, and I. Juran, 'Application of Artificial Neural Networks (ANN) to model the failure of urban water mains', *Mathematical and Computer Modelling*, vol. 51, no. 9, pp. 1170–1180, May 2010, doi: 10.1016/j.mcm.2009.12.033.

[12] R. Harvey and E. McBean, 'Understanding Stormwater Pipe Deterioration Through Data Mining', *JWMM*, Feb. 2014, doi: 10.14796/JWMM.C374.

[13] 'Water Sector Status Report', MINISTRY OF WATER AND IRRIGATION, TANZANIA, Dar es Salaam, 2009. Accessed: May 18, 2023. [Online]. Available: https://www.maji.go.tz/uploads/publications/en1568462448-Water_Sector_Status_Report_2009_1_.pdf