Failing Water Points in Tanzania: An Analysis into the Factors Affecting Human-Powered Pumps

Abinaya Maruthalingam
School of Computer Science
Univeristy of Nottingham
Nottingham, United Kingdom
psyam23@nottingham.ac.uk

Kangle Yuan
School of Computer Science
Univeristy of Nottingham
Nottingham, United Kingdom
psyky2@nottingham.ac.uk

***Abstract*—Water scarcity is a global threat that affects more than half of the world. In this paper, we develop a predictive model to aid in water management in Tanzania by predicting the functional status of water points. We specifically look at points that utilise human-powered pumps and the affecting factors: i) population, and ii) management. Multiple approaches are taken throughout the process. We first preprocess the data to make it usable. We select relevant features using Random Forest, correlation formulas, or by assuming using common sense. We use these to develop our models, K-Nearest Neighbours, Support Vector Machines, and Random Forest. In predicting whether human-powered pumps are more likely to fail in regions with high population as opposed to other regions, the Support Vector Machine performed the best overall with an accuracy of 71.2%. In predicting whether locally managed water points are more likely to fail than state managed ones, Random Forest performed the best.**

Keywords—Machine Learning, Data Mining, K-Nearest-Neighbours, Support Vector Machines, Random Forest

I. Introduction

Water scarcity is a global threat that is predicted to affect two-thirds of the world's population by 2025. Tanzania is a country in East Africa that, even with their respectable economic progress, has suffered deeply from water shortages. Despite significant investments towards waterpoints from both the Tanzanian government and the international community, approximately 30% of the population still lack access to safe water [1]. Tanzania has multiple water pumps around the country to supply water. However, many of these are unreliable. Monitoring and maintaining these water pumps are key in providing clean, and reliable access to water for the local areas.

In this paper, a data mining approach is proposed to identify the status of the waterpoints. It compares multiple approaches at various stages. Data is pre-processed differently and different models are explored to comprehend the effects of data wrangling on predictive models, and which models are most effective. It uses data from Taarifa's waterpoint dashboard, which aggregates data from the Tanzanian Ministry of Water. In particular, this research paper explores whether we can predict the likelihood of human-powered, specifically hand and rope, water points failing due to external factors. This is investigated through two aspects: population, and management type of the water point.

We are looking to answer:

1. Can we predict whether human-powered pumps are more likely to fail in regions with a high population than those with a low population?
2. Can we predict whether locally managed human-powered water points are more likely to fail than state managed ones?

Identifying and understanding the key factors that influence the functionality of water points can help the government and management better maintain existing pumps and ensure the availability of water across Tanzania.

II. Related Work

This section provides an overview of key methods related to water management that have been adopted by other researchers.

Bejarano et al. [2] develop a predictive model for smart water management in Tanzania and Nigeria. They predict the operating status of water points, along with the water quantity and quality. Feature engineering is performed to remove irrelevant features using Pearson's correlation coefficient and Spearman's rank correlation coefficient. Ensemble learning algorithms Random Forest and XGBoost are used to model the data. This paper is particularly relevant due to its predictive approach. It is most useful, however, for setting up new water points as it evaluates factors involving water - a geographic factor we have no influence over.

Darmatasia & Arymurthy [3] use data mining to predict the status and future status of water points in Tanzania. Their dataset is the same as this paper's. Recursive Feature Elimination is proposed to select important features. XGBoost is the chosen implementation for the model over Random Forest, Gradient Boosting Machine, and SVM which were also evaluated, as it had the highest accuracy at 80.38%. Results from [3] can also be used to identify influencing features, such as coordinate location and gps height, indicating that spatial analysis is important when considering building a new water point.

Cronk & Bartram [4] explore influencing factors by using regression and Bayesian Network analysis. They found a strong correlation between management type and functionality, similar to what we are investigating. Our approach differs, however, as we are adopting a predictive approach.

III. Methodology

Question 1 – Population

A. Pre-Processing

Abinaya's Approach

The training set values and labels are merged into a single dataset using the water point's ID to allow for easier processing. The dataset is reduced to focus solely on "handpump" and "rope pump" extraction as that is the focus of the paper. The dataset is simplified by removing redundant or irrelevant columns, such as *payment_type*, reducing the risk of overfitting, and to focus on the project's main goals.

Regarding duplicates, as this was considered after the removal of columns, more points flagged as duplicates. Instead of removing duplicates, they were kept as at it now ran the risk of deduplicating genuinely unique water points. Missing values, other than *population,* which is addressed later, were left as they had negligible impact.

Label encoding is done for relevant categorical values, including *basin* and *water_quality*. The newly encoded numerical values are used for training the predictive model.

Kangle's Approach

The training set values and labels are merged into one data frame based on *id*, reducing duplicates and the dataset size to 59363*40 for ease of preprocessing. Missing values are handled by removing columns with over 50% of missing values, dropping rows with more than 1% missing values in certain columns, and omitting columns where missing values can be substituted with valid non-missing values. These measures preserve original data features, refining the dataset size to 51670*38.

Duplication and overlapping information are addressed by calculating the overlap percentage between columns with similar values, keeping only one representative column from each set, thus refining the dataset to 51670*24, thereby optimizing the dataset size while preserving essential features. She also resolves discrepancies in *region_code* by reordering the regions alphabetically and reassigning them numerically.

The percentage of 0 values is evaluated, eliminating columns with more than 50% zeroes to further optimize the dataset to 51670*20, ensuring a cleaner dataset for later analysis and modelling. The functionality categories are simplified, merging 'functional' and 'functional needs repair' into 'functional'. *Functionality_status* is encoded for easy statistical analysis.

Finally, the dataset is then limited to 'handpump' and 'rope pump' entries, shrinking it to 14994*20 entries for focussed analysis.

B. Binning

We want to bin the population into categories of "Very Low", "Low", "Medium", and "High". Abinaya's approach bins based on region population and is separated into three bins (omitting "Very Low"), and Kangle's approach on subvillage population and uses all four bins.

Abinaya's Approach

Before we can bin, the region population is calculated. This is done by summing up *population* for each region which we call *region_population_value*. As there are many zero values, imputation is done to calculate the population by using the mean *population* for waterpoints in the same *region* and *subvillage*. It is likely that population was not initially recorded rather than being true zero values, as waterpoints in Dodoma are resulting in 0 which does not seem correct as it is the capital of Tanzania. Regions with 0 population are categorised as "Not Recorded". Region population is distributed into three bins of equal interval: Low, Medium, and High.

Kangle's Approach

Zeroes are substituted in the population with the *subvillage* mean or the overall median if all *subvillage* values are zero, thus enhancing data validity. Outliers in population data are adjusted and binned into four categories called *population_group*.

C. Exploratory Data Analysis

EDA is begun by converting all object data types into numeric formats, specifically an integer, bool, or float. This encoding allows for a more thorough and accurate correlation analysis using the Pearson Correlation Coefficient and P-value, as these measures only work with numerical values.

Pearson Correlation Coefficient and P-value is calculated for each column relative to the water point's working state. These metrics provide an initial insight of which factors are more likely to be reasonable predictors of the water point's functional status.

To make these statistical results more readable, data visualization techniques are employed, specifically drawing regression plots to represent the correlation analysis outcomes. These plots allow for a visual interpretation of the correlations and their strength, enhancing the accessibility and understanding of the data analysis.

D. Classification

There are three possible classes that represent the status of water points for our dataset: "functional", "non-functional", and "functional but needs repair". Abinaya works with two or three of the classes, whereas Kangle strictly works with two, combining "functional but needs repair" into "functional". We propose experimenting with multiple classifiers.

Abinaya's Approach

K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) are used to classify the models. KNN is chosen for its suitability for multiclass classification as the dataset uses 3 classes. The Minkowski metric is chosen over Euclidean distance as it is more robust to outliers and works with data that are not linearly separable.

SVM is also run with multiple kernels to determine the optimal fit. Since only 13 features are utilised, SVM becomes an efficient choice due to its efficacy with large datasets and non-linear data, without being computationally expensive. It is important to note however that SVM works with two classes only, positive class ("Functional") and negative class ("Non Functional"). The model is limited in this sense as it is not able to distinguish the third class.

Kangle's Approach

Random Forest (RF) classifier is used to identify key features. This model, enhanced by 5-fold cross-validation, handles all 13 variables and mitigates overfitting. It provides comprehensive data by determining functional water point percentages.

For the SVM classifier, the top four features and population are used to increase precision and efficiency. Different kernels account for potential non-linearity, enabling selection of the best fitting function. Like RF, it also yields functional water point percentages to answer our question.

Question 2 – Management

A. Pre-processing

Using the cleaned dataset with size 51670*20, the percentage of 0s in each column is calculated for quality control. The same data imputation method is adopted as the first question, imputing all 0s with the mean population of its *subvillage* or the median if a *subvillage* has all population values as 0. This procedure is performed before binning to secure better population imputations using a larger sample set.

Zeroes in the *gps_height* and *construction_year* are treated as errors and rows containing these are removed. This reduces the dataset's size and minimizes noise, leading to a cleaner, more accurate dataset.

B. Binning and EDA

The dataset is refined by retaining only the rows with 'vwc', 'water authority', and 'water board' under management. 'vwc' is binned as 'local', and both 'water authority' and 'water board' are binned as 'state', replacing these strings with numeric values 1 and 2, respectively. This method can enhance the readability of data distribution. By encoding all object values into numeric values, the statistical correlation analysis could be easily employed to determine the important features, supporting the subsequent model training. However, after initial EDA analysis using Pearson Correlation Coefficient and P-value, the final feature extraction is based on the importance of features produced by Random Forest classifier.

C. Classification

The RF model is used to extract the top four attributes and *management_level_num* as features, contributing to the classifier's predictive performance. The five-fold cross-validation is crucial to validating the model's performance across different subsets of the dataset, reducing the risk of overfitting and promoting model generalizability. Thethe functional water point percentage for each management group from *y_pred* is extracted, calculating the overall mean to answer our question.

In addition, the KNN model with a k-value ranging from 2 to 6 is used, assisting in identifying the model with the optimal result. She retrieves the functional water point percentage for each management group from *y_pred*, similar to the Random Forest model, and visualises it. This visualization provides a clear, intuitive representation of the findings, aiding in drawing a conclusion for our question.

IV. Results

Question 1 – Population

We observe the results from our pre-processing and training of the data.

A. Pre-processing and Binning

Abi's Results

Once we binned the region population, it is clear that there's an imbalance in our data, with a higher number of waterpoints in regions with low population than in medium or high population.

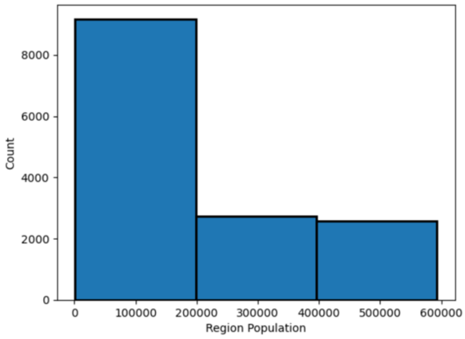

Fig. 1. Number of Water Points in Each Region Population Bin (from left to right: Low, Medium, High

This imbalance is apparent when we categorise the data into population bins in Figure 1. For instance, regions with low population (0 to 200,000) have three times as many water points as the other regions. This makes it more difficult to predict whether the region's population influences the status of the water pump as there is not as much data for Medium and High areas.

Kangle's Results

The data in Figure 2 reveals the spread of water point functionality across different population groups.

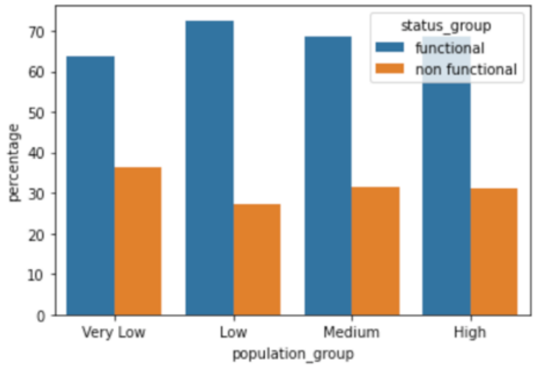

Fig. 2. Status of Water Points across Population Groups

The functionality rate is highest in 'Low' population group at approximately 72.57%. 'Very Low' population group has the lowest functional rate at 63.57%, indicating lack of maintenance in these areas. 'Medium' and 'High' population groups display similar functionality rates, 68.53% and 68.73% respectively, suggesting higher population regions might come with routine maintenance scheme. However, the 'non-functional' status across all population groups further emphasises the need for improved maintenance and management.

B. EDA

Used in Kangle's Approach

It can be concluded from Figure 3 that 13 out of 15 variables (including *population*, *region_code*, *public_meeting*, *permit*, *construction_year*, *management*, *funder*, *installer*, *payment_type*, *quality_group*, *source_class*, *waterpoint_type_group*) show weak correlation to the functional state value, as the first two regression plots are obviously not linearly correlated. However, all correlations are relatively weak, suggesting complex relationships between these variables and the *status_group_num*. This information is crucial as it guides the initial decision-making on which features to include in the classification model, ensuring a focused and effective approach to model building.

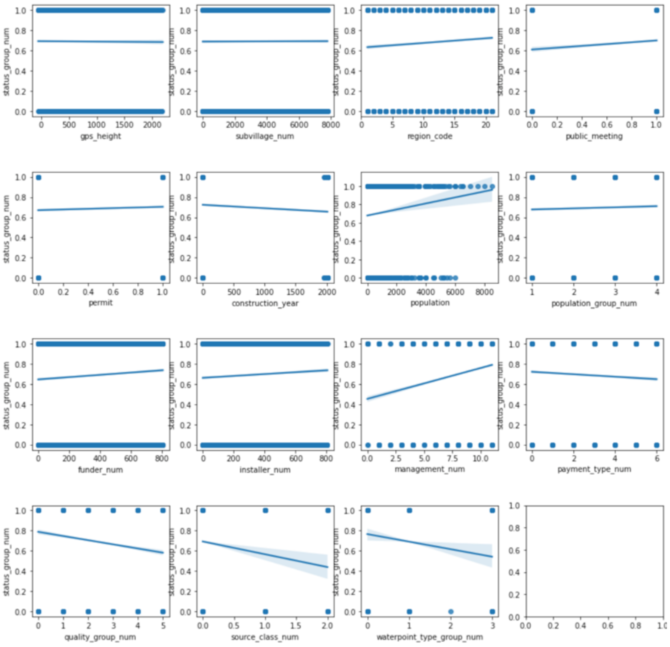

Fig. 3. Correlation between Status of Water Point and Other Features

C. *Classification*

Accuracy, Precision, and Recall are used to evaluate the performance of a model. Accuracy is the percentage of data points that the model correctly classifies. Precision is the percentage of data points classified as positive that are indeed positive. Recall is the percentage of positive data points that the model correctly classifies.

Abi's Results

The performance of the KNN and SVM models were evaluated using accuracy, precision, and recall.

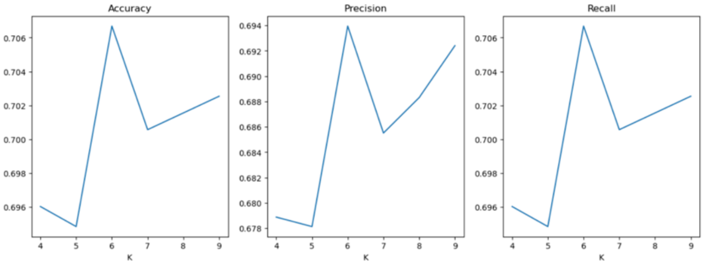

Fig. 4. KNN Accuracy, Precision, and Recall Scores

For KNN, k-values from 2 to 10, 50 to 54, and 1000 to 1004 were evaluated using the Minkowski metric. Among these, a 'k' value of 6 demonstrated the best overall with an accuracy and recall of 70.6% and 69.4% precision (see Fig. 4).

The confusion matrix in Fig. 5 shows that there are quite a significant number of misclassified instances. 68% of functional water points, 72% of non-functional waterpoints, and 50% of functional but needs repair water points were correctly classified.

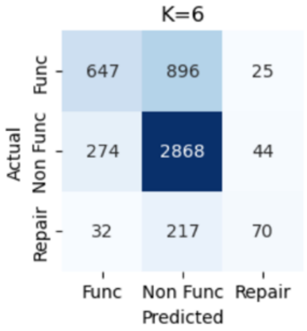

Fig. 5. KNN Confusion Matrix for k=6

On the other hand, SVM performed slightly better. The SVM Radial Basis Function (RBF) – the default kernel – had the best accuracy at 71.2% (see Table 1). Additional SVM kernels, including linear, poly, and sigmoid, were also evaluated, each presenting unique performance metrics. The table below shows all the values. We can see that the sigmoid kernel performed by far the worst, at only 50.6% accuracy and recall, and 49% precision.

	RBF	**Linear**	**Poly**	**Sigmoid**
Accuracy	71.2	62.8	68.6	50.6
Precision	72.7	76.6	69.6	49.0
Recall	71.2	62.8	68.6	50.6

Table 1. SVM Various Kernel Accuracy, Precision, and Recall Scores / %

Kangle's Results

The Random Forest model was evaluated using 5-fold cross-validation. Overall, the classifier has mean accuracy, precision, and recall of about 70%. Moreover, "Very Low" and "Low" population groups had the highest functional water points.

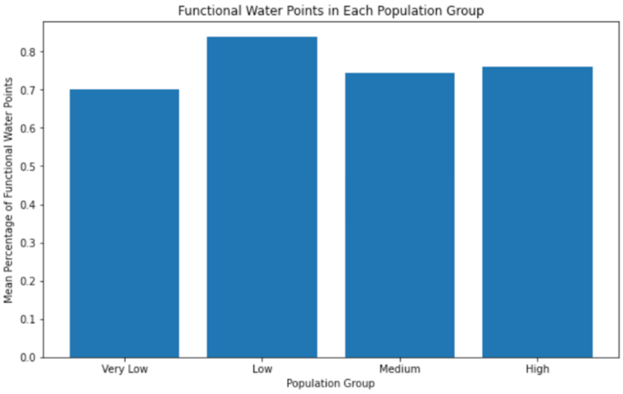

Fig. 6. Percentage of Functional Water Points across Population Groups after RF

The Support Vector Machines (SVM) model iterates through four kernels (linear, poly, sigmoid, and rbf), and the linear, poly, and rbf kernels have same accuracy, precision, and recall at 0.693, with 100% functional water points across all population bins. The sigmoid kernel, however, achieved lower metrics, indicating a poorer fit, but the reason might need to be further investigated.

Question 2 – Management

A. *Pre-Processing*

The data reveals the spread of water point functionality across different management groups. Based on the graph in Figure 7, around 60% of the local-managed water points are functional, while roughly 40% are non-functional. In contrast, state-managed water points perform significantly better with approximately 79% being functional and only about 21% non-functional. This suggests that water points under state management schemes are generally more reliable than those under local management schemes.

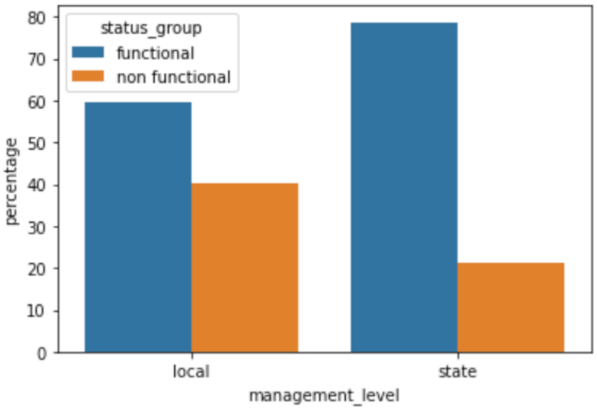

Fig. 7. Status of Water Point across Management Groups

B. *EDA*

After the correlation analysis, it is obvious that only 13 out of 16 has correlation, where variables *subvillage*, *population* and *funder_num* show no significant correlation.

Variable	Pearson Correlation Coefficient	P-value
gps_height	-0.2169	0
region_code	0.1725	0
population	0.0009	0.8844
construction_year	-0.2626	0
management_level_num	-0.1260	0
funder_num	-0.0067	0.2738
installer_num	0.0146	0.0168
basin_num	0.1545	0
subvillage_num	0.009274572	0.1294
extraction_type_class_num	0.271857496	0
payment_type_num	0.0686	0
quality_group_num	0.1629	0

Table 2. Pearson Correlation Coefficient and P-Value of numeric values

As seen in Table 2, the *gps_height*, *region_code*, and *construction_year* show significant correlations with values -0.22, 0.17, and -0.26 respectively, and all with P-values close to 0. The *installer_num* and *basin_num* have a weak positive correlation while *management_level_num* has a weak negative one. Finally, *extraction_type_class_num*, *payment_type_num*, and *quality_group_num* show a slight but significant correlation.

C. *Classification*

The use of two different models, RF and KNN, allows for a more robust analysis by comparing the results of both models. The feature importance produced by random forest classifier indicates *gps_height*, *construction_year*, and *installer_num* as the top three significant factors impacting the model, with 36.6%, 16.8%, and 9% importance respectively, and *management_level_num* was least important, at 0.8%. However, to answer our question, management should always be extracted.

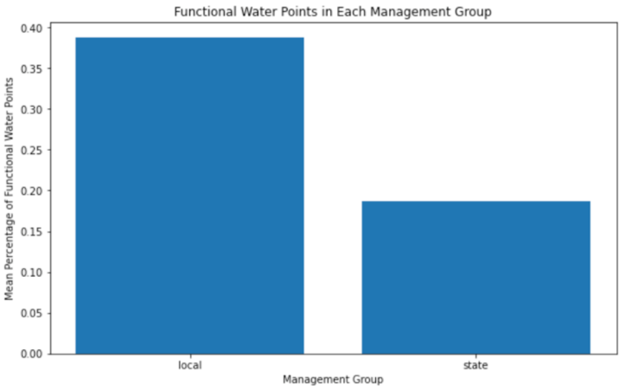

Fig. 8. Percentage of Water Point across Management Groups after RF

During Random Forest Model, over 5 folds, the accuracy, precision, and recall values were consistently around 0.77, indicating a reliable model with minor fluctuations. The average functional water points were notably higher in the local management group at 38.7% compared to the state at 18.7%, which is opposite to the values from pre-processing.

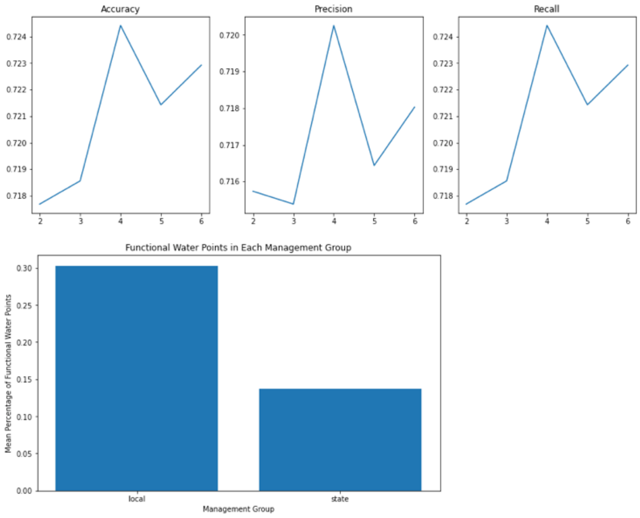

Fig. 9. KNN Accuracy, Precision, and Recall Score (top), and Percentage of Functional Water Points across Management Groups after KNN (bottom)

For KNN, k-values from 2 to 6 were evaluated using the Minkowski metric, and k-value at 4 demonstrated the best overall with an precision, accuracy and recall around 72% (see Figure 9). Overall, the functional water points percentage under local management is more than the ones under state management, which is also opposite to the results from pre–processing.

V. Discussion

We compare and critique the different approaches taken in pre-processing and when training our models.

A. *Findings*

Our findings indicate some discrepancies between pre-processing results and those derived from our classifiers, suggesting potential areas for improvement in our analysis.

In our first question, predictions from the Random Forest model correlate with pre-processing results, indicating that human-powered water points are more likely to fail in lower-populated areas. The SVM model, however, does not concur. This discrepancy may be due to imbalances in the classes of the testing data split.

In our second question, the pre-processing suggests that water points managed by the state are typically more reliable than those under local schemes. However, the predictions generated by the KNN and Random Forest models contradict this, possibly due to data imbalance or unreliable data sources.

Interestingly, neither question one nor two are population and management significant predictors of a water point's working state, according to statistical correlation analysis. In our second question, management ranks as the least reliable predictor. This is surprising, given conventional wisdom would expect these factors to have at least moderate relevance.

Several reasons might contribute to these unexpected findings, including inappropriate pre-processing, flaws in model training, or insufficient data. Future work should focus on addressing these potential issues to improve the predictive power of our models.

B. *Small-scale Comparison on Kangle's Work*

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

During pre-processing, columns where more than half of the data was missing were removed. This may have been a mistake as the dataset was not restricted to human-powered water pumps until much later, thus potentially removing incorrect columns. The strategy of removing zeroes should have been potentially treated with more caution as there are features where it may have been genuine.

In terms of feature selection, there could have been a more in-depth analysis of influencing features rather than blindly assuming features such as *longitude* and *latitude* are irrelevant to the classifier.

It's also worth noting that Pearson's correlation coefficient is only appropriate on numerical data, not encoded categorical data as it has been used here. This means that the correlation analysis may not be entirely valid.

Comparing Random Forest and SVM

Random Forest has a higher accuracy, precision and recall than the SVM classifier, thus better at identifying the status of water pumps. This makes sense as a random forest classifier is an ensemble learning algorithm that uses multiple decision trees to make predictions, compared to SVM, which is a single decision tree. Random Forest consistently performed better than all the SVM models in terms of accuracy, with a rate of 72.21% compared to the highest SVM accuracy which is 69.3%. However, we must keep in mind that there is a class imbalance problem with this dataset - there is a significantly higher number of functional water points compared to non-functional water points, and thus this difference may not be all that significant.

Random Forest also outperformed SVM in precision, with a score of 70.49% compared to 69.3%, suggesting that it's more reliable to positive predictions. Recall, on the other hand, is perfect for all SVM kernels bar Sigmoid.

Comparatively, Random Forest has a score of 72.2%. SVM's perfect recall score, however, is suspicious. Perfect recall rates are unusual, but in this context, it may be due to SVM only using four features and population whereas Random Forest uses 13. This could mean that the hyperparameters are perfectly tuned and the data is well-curated. It is more likely though that the data set is biassed or overfitted.

C. *Small-scale Comparison on Abinaya's Work*

Similarly, to Kangle, feature selection should have been done more carefully. Duplicates should also have been handled at the beginning when they could have been easily identified as now, they may affect the training data.

Comparing KNN and SVM

KNN works on all three classes, whereas SVM works on two classes: functional and non-functional, as its positive and negative classes respectively. "Functional but needs repairing" is difficult to define and measure as the water point is working but may not run for a long time before breaking.

SVM produced more accurate results. KNN is also sensitive to noisy data which wasn't specifically accounted for in pre-processing in the resulting features. The RBF kernel (radial basis function) works with non-linear data. Attributes such as *gps_height* aren't linearly related to the functionality of the pump.

RBF has an accuracy recall of 71.2% and precision is 72.7%. A high precision score is important as it's necessary to avoid false negatives to minimise the number of unnecessary maintenance visits, thus saving resources. In this case, the linear kernel has the best precision score with 76.6%. However, the accuracy and recall rates are much lower at 62.8%. A high recall is also important. A high recall means that faulty pumps are effectively detected - minimising the number of false negatives, preventing water supply disruptions. Both are key, however recall is arguably more important as the initial goal of water pumps is to provide access to water for everyone across Tanzania. A high precision, however, may be more important to government officials as they would be more willing to allocate resources. In our context where are we are attempting to predict the status of a water point, all three metrics are important. RBF's precision is only 4% lower than Linear's, whereas Linear's accuracy and recall rates are less than RBF's by nearly 8%. Due to this, RBF is a better performing kernel overall.

Every SVM kernel has the same score for accuracy and recall. This suggests that the model performs equally well in correctly identifying positive instances and correctly classifying instances. This is good as it means that the model is not biassed towards identifying one class over the other. This could also mean, however, that there is something incorrect with the data as we would expect some bias due to the class imbalance problem.

To answer question 1, based on the data, water points in areas of low population are more likely to fail. However, this may not be entirely reflective of the situation. Due to the imbalance in region population - there being three times as many water points in regions of low population than in medium or high - this skews the analysis. Further analysis needs to be done to determine this, such as feature grained feature analysis as have done. This would be useful as we can examine the effect of excluding and including certain features. We can use this to rank the importance of the features on the prediction.

D. *Overall Comparison*

Both approaches require more finetuning in its pre-processing stage. As mentioned, feature selection needs to be done more carefully. This can be approached in multiple ways, including using the Correlation Matrix or Recursive Feature Elimination as [3] does. Population ideally should have been handled better to better reflect the real-world. Missing data, our zero values in this case, could have been substituted as in [3]. This was, however, difficult due to a lack of access to similar resources.

Comparing Abinaya and Kangle's SVMs

Abinaya's best SVM outperforms Kangle's. This can be due to the way the data was pre-processed and the number of features used when training the model. Recall is not considered due to perfect recall being possibly incorrect in Kangle's results. Sigmoid for both approaches performed the worst. It is clear that this kernel is unsuitable for our predictive model. This makes sense as it is imost appropriate for datasets where the data is linearly separable, which ours is not. Kangle's Sigmoid performing better may be due to using fewer attributes, making the data more linearly separable. This is in line with the values obtained for the linear kernel, another SVM that works with linear data. Kangle's linear performs better than Abi's in accuracy. However, Kangle's polynomial performs better in accuracy, but not in precision (only just). This is contradictory, however the difference in precision can be considered negligible. This could, again, be due to the fact that Kangle's data is more linearly separable. RBF however performs better for Abi than Kangle for both accuracy and precision which is contradictory. This could be due to overfitting in Abi's SVM, which is a risk when there are more features.

Comparing Abinaya's SVM and Kangle's Random Forest

Random Forest had accuracy, precision and recall scores of 70%. SVM RBF had a similar score, albeit slightly higher. Both classifiers use 13 features. However, it is possible that the features used in the SVM are more relevant and informative to the prediction than in Random Forest. SVM, typically, are more robust to noisy data than Random Forest. Even though RF's data was pre-processed, it is possible that not all of it was handled.

VI. Conclusion & Future Considerations

In this paper, we developed prediction models for water management in Tanzania. The aim was to explore whether we can predict whether water pumps will fail in areas of high population and whether locally managed ones are more likely to fail than locally managed ones. Multiple predictive models were explored, including Random Forest, SVM and KNN classifiers. The best method based on the evaluation metrics is SVM using the radial basis function kernel, however Random Forest was a close second. To answer our first research question, the results were not as distinguishable as anticipated. Kangle concluded from Random Forest that water points are more likely to fail in regions of low population than in high. Abinaya was unable to draw a definite conclusion as, while the data also shows Low, the dataset is biassed as there are many more water points in regions of low population than any other. Even when viewed as a percentage, it is unclear. This is possibly due to the way population and population binning is handled as Kangle did not have this issue.

To further improve our prediction model and analysis, a number of steps can be taken. When reducing the dataset, feature selection should be done to identify relevant features. Random Forest can be used to draw important features, or Correlation Matrix and Recursive Feature Elimination as in [3] can be explored. A fine-grained feature analysis as in [2] could be done to rank feature importance across the prediction problems. Population is another limitation in our work. Ideally, we would have values that more accurately reflect the state of Tanzania rather than missing values as it is important in understanding which populations are mostly affected by water point failures. As further exploration of our second factor, management, was ceased, this would be an area to further investigate with better feature selection.

The research can be also extended to evaluate other extraction types, such as motor powered pumps, and compare how human-powered water pumps and motorised water pumps perform in various contexts, such as for different water qualities or quantities.

References

[1] "Water scarcity (no date) WWF. World Wildlife Fund. Available at: https://www.worldwildlife.org/threats/water-scarcity (Accessed: March 6, 2023).

[2] G. Bejarano, M. Jain, A. Ramesh, A. Seetharam, and A. Mishra, "Predictive analytics for smart water management in developing regions," 2018 IEEE International Conference on Smart Computing (SMARTCOMP), 2018. doi:10.1109/smartcomp.2018.00047

[3] Darmatasia and A. M. Arymurthy, "Predicting the status of water pumps using data mining approach," 2016 International Workshop on Big Data and Information Security (IWBIS), 2016. doi:10.1109/iwbis.2016.7872890

[4] R. Cronk and J. Bartram, "Factors influencing water system functionality in Nigeria and Tanzania: A regression and bayesian network analysis," Environmental Science &amp; Technology, vol. 51, no. 19, pp. 11336–11345, 2017. doi:10.1021/acs.est.7b03287

Contributions

Abinaya

- Formatted and edited final report
- Abstract
- Introduction
- Literature Review

Question 1

- Methodology - Everything but content of Kangle's approach
- Results – Everything but content of Kangle's approach
- Discussion – Everything but Finding
- Conclusion
- References

Kangle

Question 1

- Methodology - Kangle Section, condensed Abi's content
- Results - Kangle Section, condensed Abi's content
- Discussion - Finding

Question 1 and 2

- EDA

Question 2

- Everything