Tanzania's water pump functionality: the factors affecting their efficacy

Pointon J
MSc Data Science
University of Nottingham
United Kingdom
psxjp5@nottingham.ac.uk

Braganza J
MSc Computer Science
University of Nottingham
United Kingdom
psxjb9@nottingham.ac.uk

***Abstract*— This paper explores water pump functionality in Tanzania using Random Forest and Logistic Regression models. It is novel in that it employs nested cross-validation, with accuracy rates of 79.78% and 72.34% that are achieved for Random Forest and Logistic Regression, respectively. We computed the confidence interval for generalization error, finding it to be [78.96%, 80.60%] and [71.75%, 72.94%] and respectively. Most notably, we find that longitude, latitude, GPS height, and age emerging as the most important features for predicting functionality (using Random Forest). Furthermore, our analysis finds significant regional discrepancies in water pump functionality rates. These findings offer valuable insights for the Tanzanian government, enabling them to prioritize maintenance efforts and allocate resources efficiently to the regions most in need. The models we suggest can also assist in identifying and subsequently fixing water pumps.**

***Index Terms*— Tanzania water pumps, multinomial logistic regression, random forest**

I. Introduction

A. Context

Tanzanian authorities are bottling it when it comes to providing adequate water supply. The Ministry of Water has been rolling out its Water Sector Development Programme (WSDP) [1] since 2006. Since then, access to improved drinking water has only increased from 54% to 56% from 1990 to 2015, making "little or no progress" towards one of its Millennium Development Goals (MDGs) [2]. A key way that the Ministry of Water has been seeking to improve access to drinking water has been through the use of water pumps. However, these are often poorly maintained and malfunction after a few years [3].

Not only does clean water access affect nearly every one of the UN's Sustainable Development Goals (SDGs) [4], improved access to clean water and sanitation services has been found to enjoy 3.6 percentage points higher average annual growth rates [5]. However, the population expansion (3.2% annually in 2022 vs 2.7% in 2012 [6]) and extensive agricultural water use have caused increasing pressure on water resources in Tanzania.

B. Research question

It is therefore imperative that the regions with the worst access to water pumps are identified, and that the authorities are assisted in assessing which pumps are likely to be non-functional and/or in need of repairs, and can allocate resources accordingly. The Tanzanian authorities have recently been utilising technology to update and visualise the functional status of water points [7]. By using data aggregated from the Tanzanian Ministry of Water and Taarifa [8], we will seek to produce a model which can predict the functional status of water pumps in this dataset. We will compare and contrast the efforts of Random Forest and Multinomial Logistic regression in predicting water pump statuses.

1) What are the factors that are most influential in determining the working status of a water pump?
2) How does pump functionality vary by region and which regions have the least proportion of working pumps?
3) Does Random Forest classification or Multinomial (GLM) Regression best explain and predict the working status of a water pump, and what is the generalisation error of each approach?

C. Dataset

The dataset contains 59,400 records and 41 variables. The main variable of interest is the `status_group`, which outlines whether a pump is functional, non-functional, or functional but needs repairs; there is a bimodal distribution in Figure 1 with most pumps categorised as functional or non-functional, and only a few as partially functional.

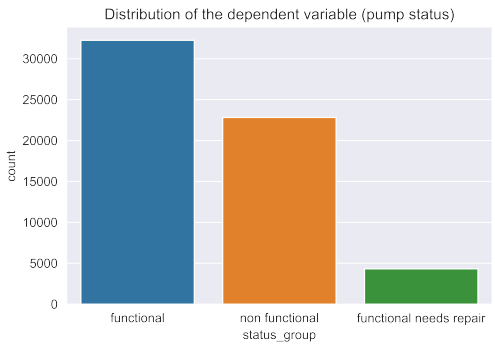


Fig. 1: Distribution of the working status of pumps in the training dataset.

The majority of variables are non-numeric; many key variables are categorical, Boolean, or strings, with only 10

numeric columns. Numeric columns are a mixture of 3 types: geographical, categorical encoding, and purely numerical. The gps_height only contains a few anomalies (< 0), and longitude/latitude values are as expected (see Figure 4). The region_code seems to use the value 99 as an encoding for missing values (and possibly the same for district_code with '80'). There is very little data regarding num_private, thus it is difficult to identify any anomalies with most values being 0. With regards to purely numerical data, amount_tsh and population are similar in distribution, being highly positively skewed and having a large variance: there are many 0 values and a portion of data with significantly large values. Finally, construction_year has some positive skew but relatively low variance, and has no anomalies, other than 0 indicating a missing value.

II. Literature

The literature is split into two strands: papers which seek to extract inference from the specific attributes (e.g. assessing the probability of certain types of pumps being functional), and papers which seek to take a purer machine learning approach in maximising the prediction accuracy. This paper is unique in that it seeks to achieve *both* using the same dataset, through comparing and contrasting two supervised learning approaches: Random Forest Classification and Multinomial Logistic Regression.

Papers which take the former approach over the latter have found that Sub-saharan Africa's pumps have an average non-functionality rate of 36%[1], with a huge variety: as low as 10% (Madagascar) and as high as 67% (DRC) [9]. Despite the work of significant foreign donors [10], Tanzania's water pumps are similarly poor quality [9], [11]. Additionally, the age of the pump has a significant impact on the likelihood of malfunction [11]. [9] found that 27% of Tanzania's pumps are non-functional within two years of construction. They also found that there was an interaction between the age of the pump and the factors affecting its likelihood of failure. Notably, that hydrological characteristics of the pump are more influential in the early years, but as the pump ages, the type of the pump and management type are much more influential.

By using similar data to this paper, [3] found that Tanzanian water pump functionality varied by pump type (Nira handpumps were more functional than Afridev and India Mark II handpumps), fee collection type (higher for monthly fees than in the event of pump malfunction), and the use of private operators rather than community-managed systems. The district the pump was situated in also had a statistically significant effect.

The second strand of literature focuses on the classification and prediction of water pumps. Most papers use either a Random Forest algorithm or Neural Network approach [12], [13], with some using Gradient Boosting [14]. Random Forest is a popular choice because of its relative resistance to overfitting using training data, ability to handle unbalanced classes, efficiency on large datasets, ability to handle thousands of input attributes, and success when using categorical data [15], [16].

There is one approach which seems to consistently achieve very high accuracy levels. [12] used a Random Forest algorithm to achieve an AUC of 0.91 and a classification rate of 0.8209 using the competition test dataset. Additionally, the author found that data cleaning only increased the AUC by 0.02 and pre-processing was hampered by limited domain knowledge. As a result, it seems like this machine learning technique is consistently successful in predicting water pump status using this dataset.

The different machine learning techniques have been compared against each other. [15] measured the classification rate (accuracy) of using a Decision Tree, Neural Network, Multinomial Logistic regression and found that Random Forest achieved the classification rate of 0.7706. The second most accurate technique was regression, but it only achieved an accuracy of 0.7298, which is significantly below the accuracy of Random Forest. Consequently, it is expected that the latter will provide the most accurate model for classification and prediction, while the former approach will provide more practical insights to policymakers and academics.

Finally, the prediction/classification strand of the literature has extracted the following attributes as contributing the most in this particular dataset: region, district, source, gps_height, quantity, extraction type, waterpoint_type, funder (particularly private/government) [12], [15]. Additionally, amount_tsh is often excluded because of the large amount of '0' values [12].

III. Methodology

A. Basic pre-processing

Using Python, the data is cleaned to maximise the value of statistical inference. Firstly, unlikely values are removed, such as record dates which were entered prior to the pump creation year, and seemed to be 10 years earlier all other data points. Due to a significant amount of incorrect and inconsistent values for amount_tsh and population (using domain knowledge and values using the 2012 Census [6]), these attributes are removed. Additionally, any '0' or '-' values are replaced with NaNs if applicable.

Missing values are typically attribute-based rather than record-based, meaning that we opted to take a column-based approach. There are many missing longitude/latitude/gps_height values, typically specific to a ward or subvillage. Therefore mean imputation is not possible and these are imputed using an external dataset (see code) using the centroids of each ward as the approximate location. Mean imputation, however, was used for construction year (and amount_tsh, which is subsequently dropped).

Several variables could not be trusted: region codes had a many-to-many mapping with the region names, and similarly with district codes; both attributes are subsequently

[1]Note that this is using data from UNICEF and other sources, and is a crude estimate due to the inconsistencies in defining functional pumps, and incomplete data

dropped. After this, we remove duplicates. based on all criteria except `id`, `amount_tsh`, `gps_height`, `population`, `construction_year`, `permit`, `funder`, `installer`, and `date_recorded`. There are many entries which had an identical entry with exactly the same data except the `amount_tsh` was NaN. Using this criterion, we found 105 duplicates, of which 68 are verified duplicates based on longitude/latitude data, and 37 are suspected duplicates based on the pump attributes. The reason that we didn't identify duplicates by using the record date is because all the other attributes were identical apart from the date, which may mean that there has been an error/typo which could artificially reduce the variance and bias the regression [17], [18]. After removing duplicates, there were 59,295 records.

B. Logistic Regression Pre-processing

Data-cleaning deviates once we begin processing categorical variables. Unlike Random Forest, Logistic regression requires many data points for each attribute used in order to converge and compute standard errors. With a large number of categorical features in the dataset, there are potentially over 100 columns. Consequently, logistic regression requires the use of attributes which has categories each containing more than 1,000 records.

The `funder` and `installer` attributes are grouped into just the top 3 categories. Furthermore, of the attributes that have similar columns, the following ones are chosen because of the large number of records in each category: `extraction_type_class`, `management_group`, `payment_type`, `water_quality`, and `quantity`. These attributes are then transformed into dummy variables.

Finally, an `age` column is created from the `construction_year` and `record_date` attributes, and the following attributes are standardised: `amount_tsh`, `gps_height`, `age`. The longitude and latitude values arre standardised by calculating the distance from a centroid of Tanzania (-6.3728, 34.8925).

Additionally, regression is much more sensitive to multi-collinearity, meaning that a simpler model is required [19]. An initial regression is fitted and the statistically insignificant variables at the 95% confidence level are dropped.

C. Random Forest Pre-Processing

Some of the basic pre-processing steps remain the same between logistic regression and random forest, however `amount_tsh` is initially included excluded from mean imputation, despite having many integer values of 0. An increasing proportion of 0's is seen as the pumps status worsens suggesting some importance. Certain categorical columns like `public_meeting` and `scheme_management` (excluded in logistic regression) have their NaNs replaced with a third category; 'other' and 'unknown' respectively. Columns like `amount_tsh`, `longitude`, and `latitude` are included as well.

An initial baseline run of Random Forest Classifier on the training data yields the confusion matrix in Figure 2 which clearly shows a class imbalance. Percentage weighting were 39%, 54%, 7% for 'non-functional', 'functional' and 'functional needs repair'.

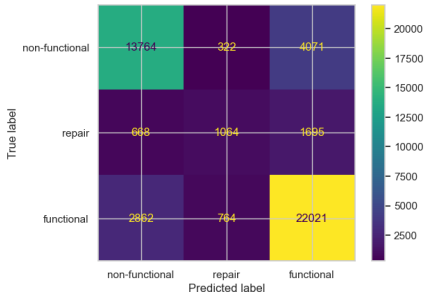


Fig. 2: Baseline Random Forest Prediction

Given Random Forest works well with high dimensional data, initially many columns are included in training the model. Those which are obviously unusable or contain too many categorical values with dubious labels like `id`, `recorded_by`, `funder_counts`, `installer_counts`, `source_class`, `wpt_name`, `scheme_name` are dropped.

Columns like `gps_height`, `longitude` and `latitude` are initially excluded from imputation previously implemented due to the external nature of the data sources and the possibility for distortion of the model. Standardisation is also as Random Forest does not require scaling as ensemble methods and decisions trees are not so sensitive to variance in the data. However standardisation is later implemented in order to run PCA analysis. The column `age` was initially excluded from replacing negative values with 0. Many imputation techniques are excluded in order to test the presence of a pattern within incorrect data inputs.

Duplicates are not dropped initially as they may contain significance in the few features of importance given the multitude of noise and poor quality of data collection. Initial implementation of pre-processing steps yields 59400 entries with 3 columns containing NaNs, `amount_tsh`, `latitude` and `longitude`. Both one hot encoding and label/ordinal encoding are compared in converting categorical data into numerical data with one hot encoding yielding better results.

D. Model Evaluation

We require the use of robust methods in order to evaluate model performance. Stratified KFold sampling is used to ensure each sample had the same distribution of pump functionalities. Additionally, we use nested cross validation in order to remove the methodological issue of fine-tuning hyperparameters to maximise the test dataset accuracy. This is a key unique aspect of this paper because of the lack of literature exploring the generalisation error for this dataset (see Section II).

There are two main metrics which are used to calculate model performance. The first is the outright accuracy. It is imperative that the model has a high overall classification

accuracy, because authorities will benefit from a model which has a low error rate. However, because of the class imbalance, a weighted F1-score is also used, because it is less affected by this.

A final, but related, method of assessing model performance is the use of One-vs-Rest Receiver Operating Characteristic (ROC) curves and the area under them (AUC). The ROC curve compares each of the possible classification thresholds a model can have; the AUC is a measure of performance calculated by summing the area under the ROC curve. This is important when there are multiple classes, and particularly when there is class imbalance.

IV. Results

A. Random Forest Exploratory Data Analysis

By looking at the percentage of pumps that are 'functional', 'non functional' and 'functional needs repair' for each category in a feature, we can see some relationships between certain features and worsening pump status. As seen in Figure 3 'quantity' is an interesting column showing overwhelming non functional status with 'dry' and 'unknown' categories and the most functional percentage of pumps seen in 'enough' with the greatest non functional plus functional in need of repair status percentage.

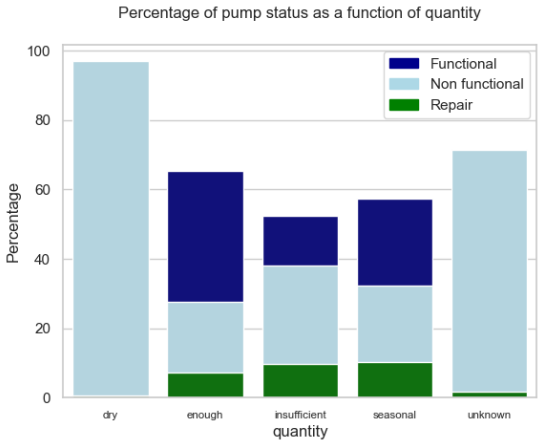


Fig. 3: Pump status functionality in relation to quantity

By plotting functional pumps in Figure 4 across a map of Tanzania we can see an uneven distribution among the country's regions, with wide variation in the proportion of functional pumps by region. Naturally there is some clustering where large cities are like dar es salaam (latitude: -6.776012, longitude: 39.178326) [20]. While there is no clear relationship between region and pump status, this map does open up the question of whether pump functionality may be related to other geographical sources like waterways. Another feature known as `source_type` provides some insights into the different water sources used. Boreholes, dam's and 'other' sources have a high percentage of non functional pumps. Strangely enough rainwater harvesting, rivers/lakes and springs had the highest repair status and also the highest working functional status as well.

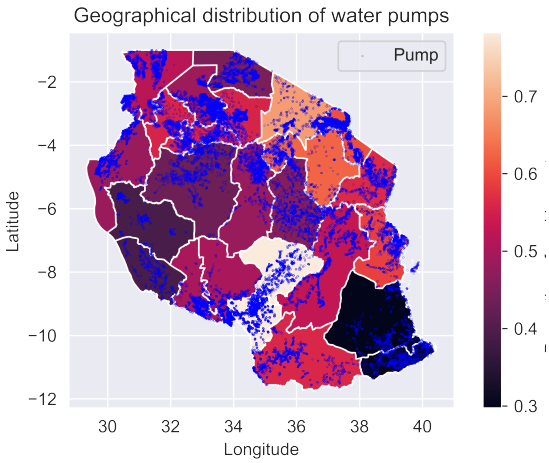


Fig. 4: Choropleth map showing the geolocation of pumps in Tanzania and proportion of functional pumps by region.

B. Logistic Regression Exploratory Data Analysis

We can also explore the breakdown of how different types of waterpumps are classified in terms of functional status. Figure 5 demonstrates how there are significant differences in the proportion of functional water pumps depending on the type of pump used. Communal standpipes have one of the highest malfunction rates, and they make up 58.4% of the sample. Similarly, handpumps have slightly higher rates of functionality. In contrast to this, Dams and cattle troughs have the highest functionality rates. However, it is important to note that there are only 7 dams in the dataset, and the top 3 have a combined share of around 1.5% of the dataset, and so the figures are less reliable. Finally, can deduce that almost 88% of Tanzanian water pumps are of the standpipe or hand pump type, and if the type is unknown, then it is highly likely to be non-functional.

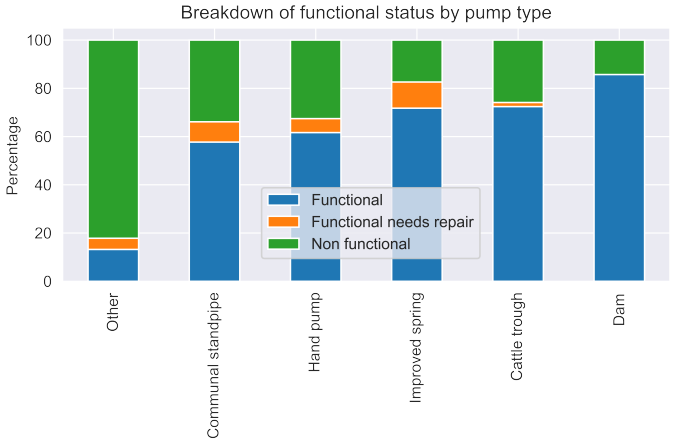


Fig. 5: Percentage breakdown of the functional status of each pump type (`waterpoint_type_group`)

Another factor affecting a water pump's functionality is its

age. As can be seen in Figure 6, there is a steady decline in the proportion of functional water pumps. Pumps that are up to 3 years old, typically having a functionality of 73.4%. However, this drops to 65.1% for pumps between 3 and 6 years old. This is a significant fall in the first few years of a pump's lifetime, meaning that the longevity of installations is poor. Furthermore, we can see that the decline in the likelihood of a pump being functional is linear, with pumps over 31 years old only having around 30% chance of being functional.

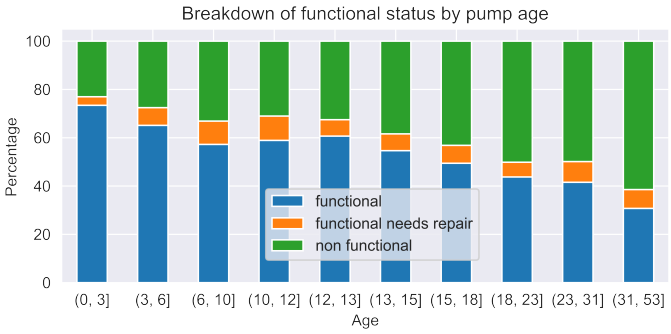


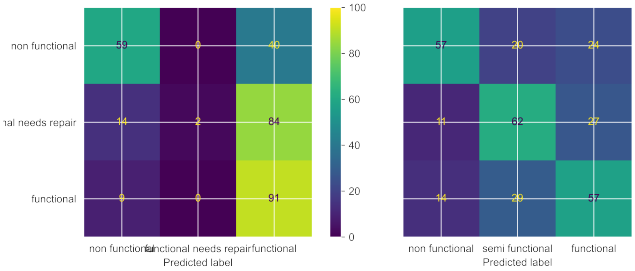
Fig. 6: Breakdown of the functional status by pump age

C. Logistic Regression

1) Imbalanced vs Balanced Sampling: We begin by outlining the comparison of controlling for imbalanced classes in the target variable (see Figure 1). The results are shown in Figure 7a: even though overall accuracy is high (due to 91% accuracy in the majority class), a large portion of pumps that are in need of repair are incorrectly predicted as being functional. There are two main techniques to combat this: under- and over-sampling.

Both techniques had a mixed impact on model performance. Under-sampling combined with stratified KFold sampling led to a vastly reduced F1 score (an average of 60.03% using the under-sampled dataset vs 69.65% using the full dataset). A significant (but lower) F1 score reduction of 5% was also observed for a MLP Neural Network model. Over-sampling also significantly impacted model performance, reducing the F1-score to 61.1%. However, as can be seen below in Figure 7, by increasing the weight of the minority class, the model has an increased prediction accuracy of the pumps that are in need of repair, but this is at the expense of the functional class, which falls from a true positive rate of 90% to 57%. Essentially controlling for class size means that each of the pump types have a reasonable true positive rate, but at the expense of the functional water pumps' true positive.

Neither technique was therefore used, because the large impact on the F1-score is not enough to justify the improved accuracy in predicting the minority class. There is no significant difference in societal impact of predicting a functional water pump as 'non-functional' or vise versa. Both lead to resource waste. Therefore, we are best to choose a technique which maximises the overall F1-score rather than the accuracy of a specific class. Consequently, our regression analysis will not use the over or under-sampled dataset, and instead use the original (imbalanced-class) dataset.



(a) Random sampling (b) Under sampling

Fig. 7: Confusion matrix for different sampling methods (in percentages)

2) Inference: Nested cross validation is used to assess the strength of L1-regularisation; stratified K-Fold sampling using 5 splits found that optimal strength was found to be $\alpha = 5.2$, as this hyperparemeter was limited by convergence tests (the Kuhn–Tucker conditions [21] must hold).

Using this optimal model, we can outline the regions which have the greatest and lowest likelihood of having a functional water pump. Figure 8 shows the results of the regression coefficients. Showing that, holding the other attributes constant, Iringa and Singida are the most likely to have functional water pumps, and Rukwa and Kilimanjaro are the least likely to.

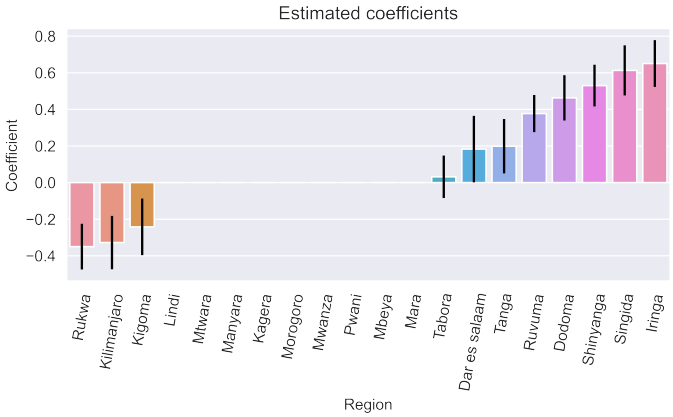


Fig. 8: Coefficients and confidence intervals for functional vs non-functional pumps, with Arusha being the baseline

The specific interpretation of these coefficients is nuanced; taking Iringa as an example, we can say that, since $e^{0.6507} = 1.92$, there is a 92% greater chance that the a given pump will be functional relative to non-functional compared to Arusha. Conversely, a negative coefficient (β) means that a particular region is $e^{\beta}\%$ less likely to have a functional water pump relative to a non-functional one. All coefficients are relative to the base case of Arusha, which is the dummy variable which was excluded to avoid the Dummy Variable Trap.

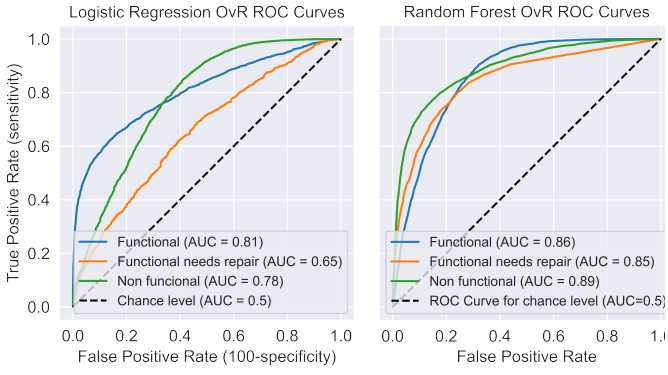
Metric		Logistic Regression	Random Forest
Generalisation error	Accuracy 95% CI	72.34% [71.75%, 72.94%]	79.78% [78.96%, 80.60%]
	Weighted F1-score 95% CI	69.21% [68.64%, 69.78%]	79.38% [78.55%, 80.21%]
Weighted AUC		78.11%	90.19%
Competition accuracy		72.30%	80.21%

TABLE I: Model performance comparison

The results in 8 show a different picture to the ones in 4. These coefficients are essentially 'unexplained' components in that they are the impact of the region on the likelihood of a pump being functional, once all other factors have been held constant (i.e. controlling for the type of pump and other factors). This can help to direct authorities to which regions require additional assistance and investigation into why water pump functionality is so poor.

3) Model performance: Nested cross validation can be used to estimate the generalisation error. A comparison of the models is shown in I. Logistic Regression achieved a 72.34% accuracy and 69.21% F1-score, which is an estimate of how well the model should perform on an unseen dataset in the real world. Indeed, this can be tested using the competition which the dataset was obtained from: an accuracy of 72.30% was achieved, which is remarkably similar and within the 95% confidence interval. This means the the model works with a consistent accuracy on unseen data, and is very unlikely to over-fitting the training dataset.

As can be seen in Figure 9a, there is a stark difference between the performance for each of the classes using logistic regression. Functional and non-functional pumps has the highest AUC, which indicates that the model is predicting their accuracy very well, with a high a low false positive rate for any given true positive rate. However, functional pumps in need of repair are poorly classified and have a low AUC value of 0.65.



(a) Logistic Regression (b) Random Forest

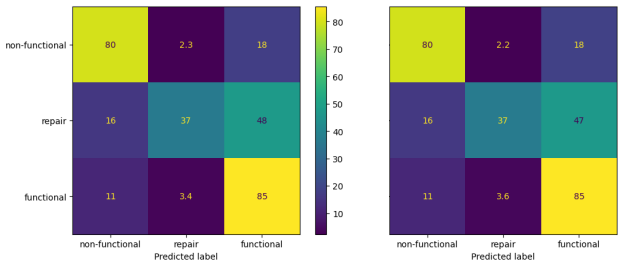
Fig. 9: One-vs-Rest ROC curves and AUC figures

D. Random Forest

As seen in I, the classification accuracy (CA) of Random Forest was 79.78% with a weighted F1-score of 79.38%. The recall score was 79.78% and the precision score was 79.17% with a 95% confidence interval for CA with a range of 1.64 and for F1-score 1.66. Nested cross validation is used to obtain the generalisation error for accuracy 79.78% +/- 0.042% and for F1 79.38% +/- 0.042%. The confidence intervals are quite wide suggesting a greater degree of uncertainty in the model. However with 15 splits implemented in stratified kfold sampling producing smaller sample size would have caused the CI interval to increase. As seen in in 9b the ROC curve with One-vs-Rest Multiclass method shows an overall similar performance between all three classes with functional being slightly higher at 0.89 compared to 0.86 for non functional and 0.85 for functional needs repair. The high AUC score (weighted) as seen in I indicates a good degree of separability between the classes with the ability to distinguish one class from the rest quite well.

As mentioned earlier, there is a significant class imbalance with functional needs repair around 7% of the total dataset. Random Forest Classifier contains a built in function called 'class_weight' [22] where you can adjust the weights. However in practice the 'class_weight' method slightly worsened the performance of the model to 79.50% accuracy and 79.12% F1 score from the baseline Random Forest model of 79.78% and 79.38% respectively.

Results of baseline Random Forest are shown in the confusion matrix. Despite a better accuracy overall than previous attempts, the functional needs repair class was mainly incorrectly classified as functional. Overall the percentages of correct classifications did not change from baseline to the weighted Random Forest model. A previous attempt has been made to implement oversampling and undersampling to account for significant class imbalance and the results appeared promising at first with a high CA score of 93% and much higher true positive rates. However when this model is tested and submitted to the competition with a score of 74.77%. It is also seen that increasing the correct classification of functional needs repair would decrease the prediction of the majority class as well.



(a) Baseline RF (b) Weighted RF

Fig. 10: Confusion matrix for different sampling methods (in percentages)

As seen in Figure 11, the most important features are longitude and latitude. The gini importance for these two features was 0.171 and 0.169 respectively. `gps_height` also plays an important role with a gini importance of 0.089. Longitude and latitude values did not undergo imputation but did undergo standardisation in the data wrangling process. Given that the top four most important features are related to geolocation, this suggests that location may be a reliable factor of predicting pump status in Tanzania.

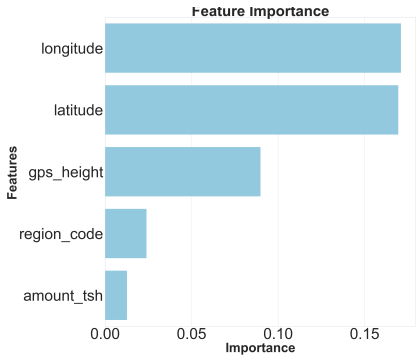


Fig. 11: Feature Importance of Baseline Random Forest

V. Discussion

A. Comparison of the 2 approaches

We will now compare and contrast the Logistic Regression results (Section IV-D) and Random Forest results (Section IV-C). The two approaches discussed provide starkly different outcomes. However, this is because two models were used with different aims and, as such, produce different outcomes. Logistic regression excels in producing interpretable inference from the parameter estimates, but doesn't achieve as high accuracy as Random Forest. However, this was known by the authors when selecting the models, because we sought to produce a holistic analysis with both accurate results and meaningful interpretation.

Logistic Regression provides greater insight into the specific impact of each variable on the likelihood that a pump is functional (or in needs of repairs) relative to being non-functional. It means that the model is more interpretable, but at the cost of lower accuracy. Conversely, Random Forest had a much higher accuracy (79.78% vs 72.34%, as shown in Table I), but at the expense of an interpretable result. Furthermore, Random Forest achieves a higher F1-score and a Weighted AUC value, which means that it has a higher true positive rate (sensitivity), when averaged over all the possible false positive rates.

Comparing the models on accuracy, F1-score, and Weighted AUC alone leads us to conclude that the Random Forest classifier produces much more reliable predictions and would therefore be the preferred algorithm, answering our 3rd research question (Section I-B). However, we have discussed how research question 2 can be answered using Logistic Regression (Figure 8), which provides valuable insights to the Tanzanian government as to which regions are in need of additional investment, and how to allocate resources.

Random Forest performed significantly better than Logistic Regression because it creates bootstrapped datasets to form multiple decision trees, where each one is a different combination of rows and attributes. These characteristics, combined with the use of majority voting, means that it is more robust to confounding variables and less sensitive to the variance in data. As such, we were able to include many more attributes than Logistic Regression (e.g. `amount_tsh`). Logistic Regression is much more sensitive to multi-collinearity, meaning that many attributes couldn't be included in the analysis.

While Logistic Regression has very high accuracy in correctly predicting functional pumps (over 90%, see Figure 7a), Random Forest only loses a small amount of the prediction accuracy for the majority class, yet excels at providing higher accuracy across each of the other classes. Therefore it is the preferred algorithm because it not only achieves a higher accuracy/F1-score, but is consistent in achieving it in each class. This is shown in the Confusion Matrices in Figure 7a and Figure 10a. Comparison of the accuracy/F1-score confidence intervals in Table I also show that, while the Random Forest confidence intervals are wider (almost 2% wide), they show that there is no overlap and as such we can say at the 95% confidence level that the results are statistically significant and Random Forest performs better than Logistic Regression.

We can say with a high degree of certainty that neither model seems to be overfitting, because the test data accuracy (competition accuracy in Table I) is within the confidence interval and similar or even higher than our optimised models' performance on training data.

B. Comparison vs literature

Our results corroborate the findings in the literature that Random Forest achieves higher accuracy than Logistic Regression. Furthermore, we find that the age of the pump has a significant impact on the functionality of the pump, providing further evidence that the longevity of the pumps seems to be a key issue behind poor functionality, and that any government investment would do well to focus on providing long-term solutions. In contrast to the literature, we found that longitude, latitude and `gps_height` are one of the most important features, which we attribute to the use of imputation using an external source in order to enrich the data and prevent record loss.

VI. Conclusions

A. Conclusion

Our novel work uses Regression and Random Forest classifiers found that we could achieve an accuracy of 72.34% and 79.78% respectively on the training dataset, and up to 80.21% on the competition test dataset. Therefore we conclude that Random Forest is the best algorithm to use on this dataset. The use of nested cross-validation to produce a generalisation error is one of our main contributions to the literature. In addition to this, we found that `longitude`, `latitude`,

gps_height and age were the most important attributes, that Iringa/Singida were the most likely to have functional pumps once other factors were taken into account (such as pump type), and Rukwa/Kilimanjaro were the most likely to have non-functional pumps (Section IV-C).

Logistic Regression was chosen because of its ability to provide interpretable insights about how each feature impacts the likelihood of a pump being functional, despite knowing that the literature has found it to be a poor classifier. However, if pure accuracy was the most important factor then we would have been best to choose a different algorithm (such as a neural network) which can be more resistant to overfitting and confounding variables.

There are some drawbacks with our approach. With the data cleaning, we opted to impute every column prior to deciding which features to extract for the analysis. As part of this, we used a computationally expensive method to fillna values in the competition test dataset with the training dataset values (for construction_year), by using the 'iterrows' method. In practice this will take a lot of time with a large dataset of previously unseen data that we wish to predict. It could also cause out-of-range values due to the test dataset having a different range to the training dataset. However, the next best alternative is to impute the test/training values as one whole/combined dataset but this means that the test data is influencing the training data, causing data leakage. We also note that implementations of this model on future data will require some manual work to enriching the data using our external dataset.

B. Future research

There are a number of research areas that we are unable to explore. We believe it would be fruitful for future work to look at the geographical distribution of pumps that the model fails to classify, to identify if the model struggles to predict pumps in specific regions. Furthermore, work could be done to adapt the model to predict *when* a pump will fail, which would make it a more applicable tool for managing agencies. Additionally, the use of nonlinear combinations of attributes could be explored, similar to [9], as these could improve classification accuracy further.

VII. Contributions

Joel Pointon and Joel Braganza both worked on the data cleaning process, regardless of which machine learning algorithm used it. Pointon completed the imputation of longitude/latitude values, and wrote the code for construction_year imputation. Both Pointon and Braganza created the chloropleth map.

Braganza specifically worked on the Random Forest classifier and wrote all content related to it, trialed PCA, and compared various combinations of attributes to find the best performance.

Pointon specifically worked on Logistic Regression and wrote all content related to it, constructed the code for comparing under- and over-sampling, and wrote the code for nested cross validation. Additionally, he wrote the introduction context (Section I-A and literature review (Section II).

References

[1] J. H. Aweso, "WATER SECTOR DEVELOPMENT PROGRAMME PHASE THREE (WSDP III) 2022/23 – 2025/26," https://www.maji.go.tz/uploads/publications/sw1664866566-WSDP%20III%20FINAL%20FINAL%202022%20(1).pdf, 2022, [Online; accessed 01/03/2023].

[2] WorldBank, "Progress on sanitation and drinking water – 2015 update and mdg assessment," https://data.unicef.org/wp-content/uploads/2015/12/Progress-on-Sanitation-and-Drinking-Water_234.pdf, 2015, [Online; accessed 01/03/2023].

[3] J. Bartram and R. Cronk, "Factors influencing water system functionality in nigeria and tanzania: A regression and bayesian network analysis," *Environmental Science & Technology*, vol. 51, 08 2017.

[4] R. Damania, Desbureaux, A.-S. Sébastien; Rodella, J. Russ, and E. Zaveri, "Quality unknown : The invisible water crisis," 2019.

[5] H. Tropp, "Making water a part of economic development: The economic benefits of improved water management and services," *Stockholm International Water Institute*, 2005, [Online; accessed 03/03/2023].

[6] Sensa, "2022 population and housing census - national bureau of statistics," 2022, [Online; accessed 01/03/2023].

[7] J. Katomero, Y. Georgiadou, J. Lungo, and R. Hoppe, "Tensions in rural water governance: The elusive functioning of rural water points in tanzania," *ISPRS International Journal of Geo-Information*, vol. 6, no. 9, 2017. [Online]. Available: https://www.mdpi.com/2220-9964/6/9/266

[8] P. Bull, I. Slavitt, and G. Lipstein, "Harnessing the power of the crowd to increase capacity for data science in the social sector," *CoRR*, vol. abs/1606.07781, 2016, accessed 09/02/2023. [Online]. Available: http://arxiv.org/abs/1606.07781

[9] B. Banks and S. Furey, "What's working, where, and for how long: A 2016 water point update," 2016.

[10] R. D. Carlitz, "Money flows, water trickles: Understanding patterns of decentralized water provision in tanzania," *World Development*, vol. 93, pp. 16–30, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305750X16303096

[11] G. Joseph, L. Andrés, G. Chellaraj, J. Grabinsky, Z. Sophie, C. Emi, A. Yi, and R. Hoo, "Why do so many water points fail in tanzania? an empirical analysis of contributing factors wps8729," 2019.

[12] J. Benoot, "Predicting the functional state of tanzanian water pumps," [Online; accessed 01/04/2023], available at https://libstore.ugent.be/fulltxt/RUG01/002/350/680/RUG01-002350680_2017_0001_AC.pdf.

[13] J. F. Flefil, M. A. Galanis, and V. Kozlow, "Pump it or leave it? a water resource evaluation in sub-saharan africa," 2018, [Online; accessed 01/04/2023], available at https://cs229.stanford.edu/proj2018/report/106.pdf.

[14] Darmatasia and A. M. Arymurthy, "Predicting the status of water pumps using data mining approach," in *2016 International Workshop on Big Data and Information Security (IWBIS)*, 2016, pp. 57–64.

[15] I. K. Chowdavarapu and V. Manikandan, "Data mining the water pumps: Determining the functionality of water pumps in tanzania using sas enterprise miner." SAS South Central User Group Forum, 2016.

[16] "Cfi's guide to random forest," *Corporate Finance Institute*, vol. [Online; accessed 01/04/2023], available at https://corporatefinanceinstitute.com/resources/data-science/random-forest/.

[17] F. Sarracino and M. Mikucka, "Estimation bias due to duplicated observations: a monte carlo simulation," *Munich Personal RePEc Archive*, vol. [Online; accessed 02/04/2023], available at https://mpra.ub.uni-muenchen.de/69064/1/MPRA_paper_69064.pdf.

[18] ——, "Bias and efficiency loss in regression estimates due to duplicated observations: A monte carlo simulation," vol. 11, pp. 17–44, 01 2017.

[19] E. Liddle and R. Fenner, "Water point failure in sub-saharan africa: The value of a systems thinking approach," *Waterlines*, vol. 36, pp. 140–166, 2017.

[20] LatLong, "Dar es salaam, tanzani," https://www.latlong.net/place/dar-es-salaam-tanzania-2451.html, 2023.

[21] H. W. Kuhn and A. W. Tucker, "Nonlinear programming. proceedings of the second berkeley symposium on mathematical statistics and probability. university of california press, berkeley-los angeles, calif. license: Cc by 3.0 igo.",," p. pp. 481–492, 1951.

[22] A. Cheng, "Machine learning: Step-by-step," https://towardsdatascience.com/machine-learning-step-by-step-6fbde95c455a, 2020.