Predicting Water Pump Functionality In Tanzania

School of Computer Science, University of Nottingham, UK

Vedashree Vittalapura Krishnamurthy
student, University of Nottingham
psxvv3@nottingham.ac.uk

Muhammed Razi Abdul Shukoor
student, University of Nottingham
psxma23@nottingham.ac.uk

Mohid Ali Gill
student, University of Nottingham
psxmg8@nottingham.ac.uk

***Abstract*—This study evaluated four machine learning algorithms - Random Forest, XGBoost, K-Nearest Neighbours, and Artificial Neural Networks - for predicting the functionality status of rural water pumps in Tanzania using data from the Tanzanian Ministry of Water. Extensive data preprocessing like imputation, discretization, feature selection, and oversampling were applied. The Random Forest algorithm performed best, achieving 74% accuracy and 66.3% balanced accuracy in classifying pumps as functional, non-functional, or functional requiring repairs. It excelled at capturing complex relationships between features like location, construction details, and management factors. XGBoost also showed strong results with 68.4% accuracy. The ensemble learning techniques outperformed other methods on this challenging prediction task involving highly categorical data. The findings can guide resource allocation and maintenance efforts for Tanzania's water infrastructure.**

I. Introduction

Providing clean water has been a critical challenge in rural areas all around the world, specifically in Tanzania effecting health, agriculture and economic stability. To supply water clean water pumps have been installed throughout the country but a portion of them either need repairs or are non functional. The ”Pump It Up:Data Mining the Water Table” competition, hosted by DrivenData aims to predict the operational status of water pumps throughout the country. .

The data provided is very rich containing records of approximately 60,000 water points. For each waterpoint a wide variety of features is provided such as the water point's geographical location (longitude, latitude), management organisation, construction year and the source of water among many others. These features assist in developing the models to accurately predict the water points condition i.e. water point is functional, need repairs or is non-functional. .

Advanced predictive models including Random Forest, Gradient Boosting and Neural Networks are used to predict the water pump functionality. Impact of over 40 features was analysed for how they impact the condition of water pumps which range from geographical data to technical specifications of the pumps. The aim is not only to improve the accuracy of the model but also to provide insights of what water pumps need repairs so that they can be repaired timely and prevented from being non functional.

II. Literature Review

The application of machine learning in predicting the water pump functionality in Tanzania is explored in this literature review and extends the study to explore other domains where similar methods have proven to be effective. .

A. Murugan, S.Anu H. Nair, and K.P. Sanal Kumar's research on ”Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers” [6] work on the skin cancer dataset used support vector machine (SVM), Random Forest and kNN classifiers to detect the cancer effectively. The study mainly focuses on how SVM classifier is particularly effective in medical field, and has potential to perform good in other fields.

”Comprehensive Review of Machine Learning for Water Quality Prediction” by Xiaohui Yan et al [7] examines the use of machine learning for predicting the water quality by reviewing over 170 researches previously held in the last five years. The study emphasises on the importance of data pre-processing, single indicator and double indicator predictions. The study helps us ensure that the prediction for condition of water pump based on comprehensive environmental data. .

”Research on Water Resource Modelling Based on Machine Learning Technologies” by Ze Liu and colleagues [8]again explores the machine learning models in various aspects of water resource modelling. Models are explored and discussed in the research which can guide us in selecting appropriate models and techniques for our study. .

The detailed literature review highlights the applications of machine learning techniques and algorithms which can deliver precise and accurate predictions for the water pump conditions.

III. Methodology

A systematic approach will be adopted to handle and model data effectively. Starting with Exploratory Data Analysis (EDA) to develop a thorough understanding of the dataset and determine the most appropriate data pre-processing techniques, Data will then undergo a comprehensive pre-processing stage focusing on cleaning, structuring, and transforming the data to prepare it for analysis and modelling.

A. Exploratory Data Analysis

The exploratory data analysis phase is the most crucial stage for gaining insights into the dataset and knowing its characteristics. In this study, we have performed the following EDA:

- Data Summary: A descriptive statistical information of the data were generated to understand the distribution of the data. The similar process was performed for the categorical variables, to understand the uniqueness and distribution of the categorical data.

- Data Visualisation: Data visualisation is used at every step of the analysis to help understand the patterns, relationships, and data distribution which is necessary for knowing the insights better.
- Target Variable Analysis: Target variable analysis is performed for the status_group to understand the class imbalance and distribution.

B. Data Pre-Processing

1) Data Cleaning: Data Cleaning is the first and foremost part of analyzing and modeling the data. Data with a high amount of noise and missing values will not be a great fit for modeling. Most of the machine learning algorithms fail to work with missing features, so we will need to take care of these before moving to the next phase. This is done by the following methods:

- Drooping the column/feature with high noise or missing values.
- Dropping the rows with high noise or missing values.
- Performing imputation techniques to fill in the data and remove noise from it.

This process ensures the data readability and usability in the model building. In this study, we have incorporated various data-cleaning techniques such as dropping the feature and performing imputation in various ways.

2) Data Transformation:

- Data Imputation: Data imputation of setting the missing values to some value (zero, the mean, the median, the mode of the data, etc). This can be easily accomplished by using the built-in imputers or using the replace() function to replace the missing values with the desired imputed values. In this study, we have used various imputation techniques such as replacing the missing values with the mean of the data in the column, replacing with mode, replacing with median, and replacing with mode by grouping the data by a particular feature to produce an unbiased imputation.
- Data Discretization: Data Binning is applied to continuous variables to convert them into categorical ones. Since the dataset predominantly consists of categorical features, binning helps to reduce data complexities by grouping continuous values into more manageable categories. This technique enhances model interpretability by transforming continuous variables into a clearer format, allowing patterns and relationships to be more easily identified. Additionally, binning aligns the continuous features with the existing categorical data, creating a consistent representation across the dataset.

3) Handiling Outliers: Data points that deviate a lot from other observations in the dataset are called outliers, if not handled properly they can affect the statistical analysis leading to misleading predictions. These outliers can be due to many reasons such as errors in the data collection, measurement errors, or natural variance in the population. Identifying outliers is a critical task, data visualization is done using scatter plots and box plots, or analytical techniques like Z-scores and Interquartile range can be used to detect outliers. Once the outliers are figured out we need to determine the reason for the outliers, if they are due to errors or genuine variations, and a final decision is made as to whether to keep, adjust, or drop them. Depending on the decision outliers can be removed, adjusted by capping or transforming the values, or kept as it is if they represent important aspects of the data. Dealing with the outliers properly helps in making more accurate models, especially in predictive analytics, where the model training and predictions can be skewed.

4) Feature Reduction: Feature reduction is used to reduce the Curse of Dimensionality. When a model is trained with high number of features, there are more chances of overfitting the model and leading the model to produce bad results. In this study, we have used several strategic ways to carefully perform the feature reduction without losing any valuable information.

5) Encoding: Encoding is the technique of converting categorical columns to numerical. In this study, we have used a one-hot encoding technique. This is done by creating dummies. For instance, if a categorical feature contains 3 unique values then three binary columns are created to represent the categorical feature.

6) Feature Selection: The Random Forest Classifier is a powerful machine learning algorithm used for feature selection due to its ability to rank the importance of features. It constructs multiple decision trees and aggregates their predictions to improve accuracy and control overfitting. For feature selection, the classifier calculates importance based on how frequently and significantly each feature contributes to data splitting across the ensemble of trees. The Random Forest was chosen for its robustness in handling a large number of features, resistance to overfitting, and effectiveness in identifying complex data relationships. This ranking helps prioritize impactful features while eliminating less significant ones, optimizing model performance and reducing computational complexity.

7) Oversampling: Oversampling is often employed to tackle the uneven distribution of data, specifically when one class has significantly fewer instances than another. Essentially, this technique involves boosting the number of cases in that underrepresented group until it evens out with its counterpart. This can be achieved by replicating original observations or synthesizing new ones through methods such as SMOTE, ADASYN, SMOTEEN, and more. By implementing oversampling into machine learning models, datasets are balanced, which promotes optimal performance and less partiality towards overburdened groups, amongst others. Overall results should then yield improvements to accuracy across all classifications involved within that system's workable capacity.

IV. Modelling

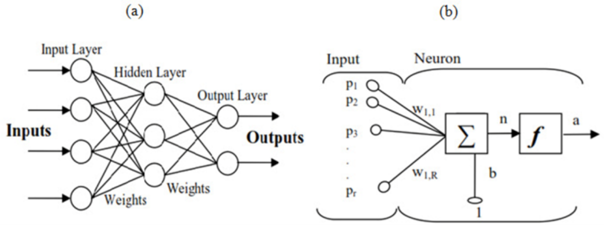
In this case study, we have used four classification algorithms to create predictive models for the water point classification.

1) Random Forest (RF): The Random Forest algorithm is a popular machine-learning technique used for both classification and regression. It is an ensemble learning technique that works by combining multiple decision trees to make the prediction. In our study, we use this to classify the water point pumps into functional, non-functional, and functional needs repair categories. RF contains all the hyperparameters of a DecisionTreeClassifier (this is to control how trees are grown) and all the hyperparameters of a BaggingClassifier to control the ensemble itself. RandomForestClassifier is chosen as a good fit model because RF is capable of handling high-dimensional data. It randomly selects a subset of features at each split which prevents the overfitting of the data. RF is capable of capturing complex non-linear relationships between features and the target variable (in our case 'status-group'). This aspect is beneficial in our current study where there are multiple non-linear relationships present between the features and the functional status of water pumps. RF is capable of handling imbalanced class distributions. It is majorly required in our study as we have a high imbalance in the number of functional, non-functional and functional but needs repair classes of water pumps. Additionally, a major advantage of RF is that it measures feature importance which can be a useful aspect for feature selection and understanding the importance of different variables in the classification task. This aspect can help us to identify the most influential features and improve the interpretability of the model.

2) eXtreme Gradient Descent (XGBoost): XGBoost is an optimized distributed gradient boosting library. It builds a series of decision trees based on the errors made by the previous tree. XGBoost was chosen as a potentially significant model for our classification task as it is designed to be highly efficient, flexible, and portable. The decision trees created by the XGBoost algorithm are not only used to minimize an objective function by accounting for the loss function, but they also protect the tree from overfitting by using a regularization process [2]. XGBoost is desired by data scientists as it has a high execution speed out of core computation. [3]

3) K-Nearest Neighbours (KNN): KNN is a supervised learning algorithm based on majority votes. The labels are predicted using the most frequent occurrence label of k-nearest neighbours present. The key challenge here is to find the optimal number of k where the prediction is more accurate. KNN is majorly chosen for classification problems as its simplicity and easy-to-implement nature. KNN is a non-parametric algorithm that can handle complex relationships between features and the target variable without making any underlying assumptions about data distribution. This nature of KNN is the best for non-linear classification like in our study. The optimal number of k for the algorithm can be found using the either Elbow method.

4) Artificial Neural Network (ANN): Artificial Neural Network (ANN) are simplified models of biological neurons. ANN consists of a highly interconnected network of simple processing neurons. Each neuron receives multiple input signals which are modified by connection weights. These weighted inputs are summed and passed to the non-linear activation function to generate the neuron's output signal. This is best explained by the diagram below: [4]



ANN is chosen widely for classification problems because of its ability to handle non-linearity and complex relationships between the data. It is capable of handling high-dimensional data.

A. Evaluation Metrics

Evaluation and Performance metrics are essential to assess the effectiveness and performance of the model for the given task. Appropriate evaluation of the model will help in making informed decisions. In this study, we have used the following evaluation and performance metrics in order to measure accuracy, robustness, and generalization capability

1) Classification Report: The classification report provides a comprehensive evaluation for each class in the classification problem. The metrics measured in this include precision, recall, F1-score, and support. Precision – Also known as Positive Predictive Value (PPV). It measures the proportion of true positives out of all the positive predictions. True positive is the positive instances that the model actually predicted to be positive. This can be calculated as below:

$$precision = \frac{TP}{TP + FP}$$

Recall - It is also known as sensitivity. It measures the proportion of true positives out of all the true values.

$$recall = \frac{TP}{TP + FN}$$

F1 score – F1 score is the harmonic mean on the precision and recall.

$$F1 - score = \frac{TP}{TP + \frac{FN+FP}{2}}$$

Support - The support is the number of records of each class in the data.

2) Confusion Matrix: The confusion matrix is a matrix that summarizes the performance of a classification model. It shows the number of true positives, true negatives, false positives, and false negatives.

3) Balanced Accuracy score: This is one of the important ways of evaluating the model. The balanced accuracy score is specifically instrumental when dealing with imbalanced datasets, where the number of instances in different classes is significantly different like in our data. It calculates the average recall obtained on each class. The balanced accuracy score provides a fair evaluation of the model's performance by considering the performance of the model on each class.

4) K-fold Cross-Validation Score: It is a model evaluating technique that evaluates the models by splitting the dataset into multiple subsets called folds. The model is trained on different combinations of these subsets and tested on the remaining subsets. The cross-validation score represents how accurate the model is on different folds of the dataset. It provides a more robust evaluation of the model's performance by considering its performance on various subsets of the data.

V. Results and Evaluation

A. Data Preprocessing Results

1) Data Cleaning: In this section data is thoroughly checked for missing values, duplicates, and outliers. The data did not contain any duplicate rows or columns as such. However, from the initial steps of reading and describing the data in this study, we found some data that were more of noise than useful information for modeling. These cases were handled appropriately by performing imputations or dropping the columns. One such instance from our data cleaning process is given as follows: The columns such as 'sheme_name', 'wpt_name', 'subvillage', 'date_recorded', 'num_private' etc were more noise than useful information for modeling. The column 'scheme_name' has 48.5% of the data missing. Such columns even after imputing will not provide the modelling with meaningful information but rather make it more biased and harder for the algorithm to learn. The columns which had less significant uniqueness were also dropped after analysis for ease of modeling. This process was aimed to ensure the integrity of the dataset and remove redundant information.

2) Data Transformation:

- Data Imputation Results: Data imputation is the technique of replacing the missing values with meaningful data such as the mean, median, or mode of the data. In this study, we have used different techniques of imputation such as replacing the missing values with the mean of the data, the mode of the data, and also the median of the data. Imputing all the missing values with mode when the number of missing values is high leads to imbalanced data. To tackle this, we performed a strategic way of imputing the data. One such example from this study is the imputation done for the column 'funder'. The 'funder' column had 3637 missing values. Imputing all the 3637 values with the single most occurring values would hamper the integrity of the data. Hence we came up with an approach to impute the values by the most occurring 'funder' for each region. This approach handled all the missing values by not losing the integrity of the data. Whereas in the case of the features 'scheme_management' and 'public_meeting', the missing values were saved to be imputed with the mode itself directly without grouping.
- Data Discretization Results: The construction year column in the dataset was binned into distinct intervals to convert the continuous feature into meaningful categories, such as decades. The data was categorized into the following groups: 1960s, 1970s, 1980s, 1990s, 2000s, 2010s, and an additional "unknown" category for years marked as 0 or missing. The distribution of data points (fig1) across these categories provides insights into the history of water point construction and data collection. The 1960s have the fewest instances, suggesting limited construction or incomplete records, while the 1970s and 1980s show increased activity. The 1990s and 2000s have the highest counts, indicating peak construction periods or more comprehensive data collection. The 2010s, though not as extensive as the previous decades, represent recent developments. However, the substantial "unknown" category, with over 20,000 instances, highlights the challenges of incomplete or inconsistent historical records for many water points. Moreover, the inequality in the number of data points for each decade could introduce bias and affect the model's performance. Considering these factors, the decision was made to drop the construction year column from further analysis to ensure a more balanced and reliable dataset for modeling purposes.

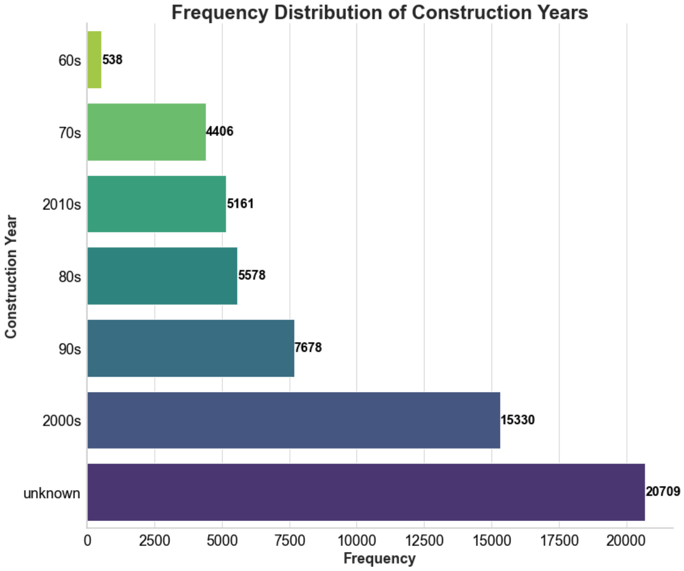


Fig. 1.

3) Handling Outliers Results: The outlier handling involved many processes and techniques Before handling the outliers figuring out the columns containing outliers need to be figured out, Z-score method is used with a threshold of 5 to flag the values that deviate from the mean by 5 standard deviations. This resulted in providing the columns which had values far away from the average and had to be handled. To visualize the outliers, violin plots are plotted for each identified column which help us to provide the distribution of the column and the extent the outlier affects it. The plots are as follow fig(2) The district code column had outliers but it was dropped as each code is unique and it can be the actual data that seems like an outlier. Longitude column has a few outliers showing 0 longitude which is not possible and considered to be null values and they are replaced with mean longitude values. Lastly, the columns with outliers i.e. amount tsh, longitude and population are handled with capping. The method involves calculating the 5th and 95th percentile of each column and then adjusting the values outside of this range to the nearest boundary value. Capping helps in moderating the extreme values while maintaining the integrity of the data. All these techniques ensured a balanced handling of the outliers.

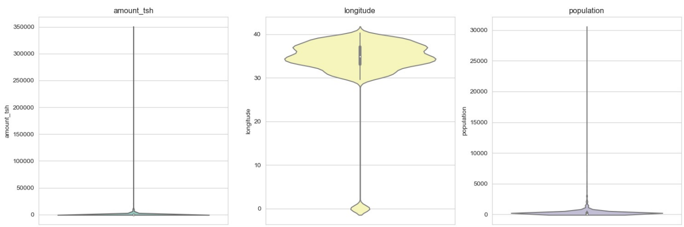


Fig. 2.

4) Feature Selection Results: To identify the most significant features from the random forest classifier, a threshold of 0.02 is applied to the feature importances. Features with importance values greater than 2% are considered important and selected for further analysis. The feature importance results fig(3), as shown in the graph below which reveal that the top two features, 'longitude' and 'latitude', have the highest importance values of 15.1% and 14.8%, respectively. These features collectively account for 29.9% of the total feature importance, indicating their significant influence on the target variable. The next important features are 'quantity_group_dry' (7.7%), 'gps_height' (7.5%), and 'population' (5.0%). The remaining features have lower importance values, with 'region_code' having the least importance of 2.4% among the top 11 features.

The fig 4 provides a correlation heatmap that visualizes the pairwise correlations between the features. The heatmap reveals some notable correlations, such as the strong positive correlation between 'quantity group dry' and 'waterpoint type group other' (0.39), and the moderate positive correlation between 'longitude' and 'latitude' (0.28). These correlations suggest potential interactions or dependencies between the features. The findings from the feature importance analysis and correlation heatmap provide valuable insights for further analysis and modeling. The top features identified, such as 'longitude', 'latitude', 'quantity group dry', 'gps height', and 'population', can be given priority in subsequent modeling steps. Additionally, the correlations observed between features can be taken into consideration when interpreting the results and making decisions based on the model.

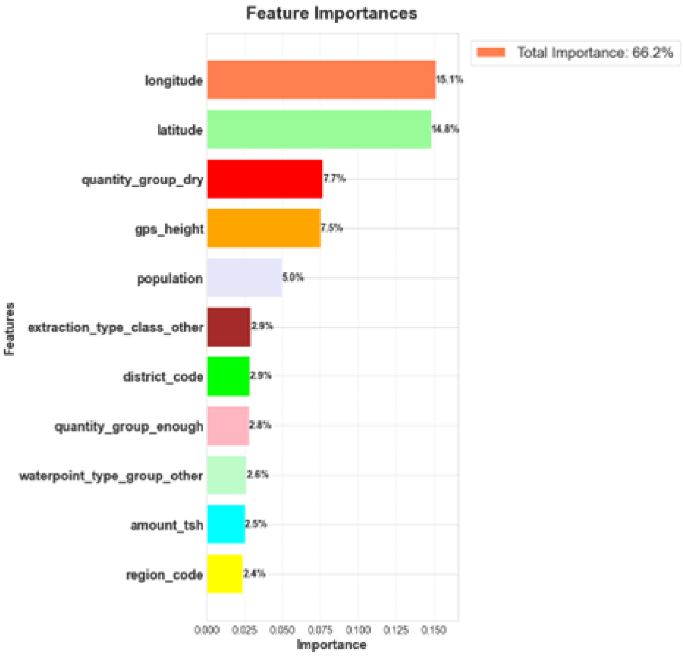


Fig. 3.

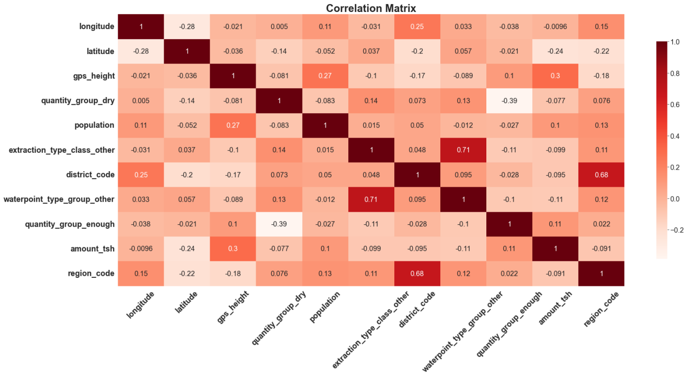


Fig. 4.

5) Oversampling Results: The class variable (Fig5) showed a notable imbalance, as the "Non-Functional" and "Functional Needs Repair" categories together made up just 45.7% of all data points. In contrast, a disproportionate amount of 54.3% was occupied by the "Functional" category which could result in significant obstacles for machine learning models since it may cause biased predictions with suboptimal accuracy levels overall.

In order to address this problem, we utilized the Adaptive Synthetic Sampling (ADASYN) approach. This oversampling technique aims to create a more equitable class distribution and its efficacy is shown in figure below. Initially, the data was imbalanced towards certain classes but after applying ADASYN, an approximately equal number of instances were

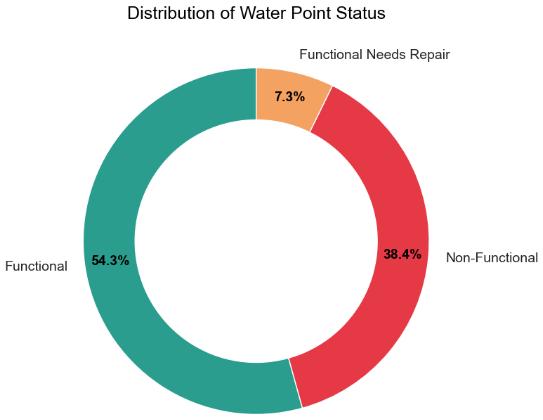


Fig. 5.

produced for each category resulting in a well-balanced set .

B. Modeling Results

1) Random Forest Results:: The first run of the Random Forest with default parameters. After training the model for the selected features from RF feature selection, it produces 73.9% classification accuracy and 66.4% balanced accuracy. Several studies have stated that the default parameters achieves a satisfactory results. However we try to optimise the outcomes by performing hyperparameter tuning using GridSearchCV function. GridSearchCV uses grid search and cross valication techniques to find the best combination of hyperparameters for the RandomForestClassifier. The hyperparameters to be optimised are defined in a dictionary named ‘param-grid’. By running the GridSearchCV to perform a 5 cross-validation , we derive the best hyperparameters as below: ’criterion’: ’gini’,’max-depth’: 50,’min-samples-split’: 2,’n-estimators’: 150 These hyperparameters is expected to give 79.16% classification accuracy. Once tested by fitting the test data in the best model obtained by GridSearchCV, we obtained the result of 74% classification and 66.4% balanced accuracy which is very similar to the default RandomForestClassifier.

2) XGBoost Algorithm Results: In this study, we model the XGBoost algorithm in two ways. One with default parameters and the other by performing hyperparameter tuning using RandomSearchCV. XGBoost aims at optimising a cost objective function composed of a loss function (d) and a regularization term (β)

$$\Omega\left(\theta\right)=\underbrace{\sum_{i=1}^{n}d\left(y_i,\hat{y_i}\right)}_{Loss}+\underbrace{\sum_{k=1}^{K}\beta\left(f_k\right)}_{regularization}, \qquad (1)$$

Fig. 6. This is a caption under the graphic.

For the start we run the XGBoost algorithm with the following parameters: min-child-weight=1, max-depth=10, learning-rate=0.05, gamma=0.1 colsample-bytree=0.4, booster=’gbtree’. min-child-weight: It determines the minimum sum of instance weight required in a leaf node. This parameter helped to control overfitting by adding regularization (β). max-depth : This sets the maximum depth of each decision tree. learning-rate : It determines the step size of each boosting iteration. gamma: It specifies the minimum loss reduction (d) required to make a further partition on a leaf node. colsample-bytree: It controls the subsampling of features while constructing each tree. booster: It specifies the type of booster used. From using the above parameters in the first run of the XGBoost we achieve the results of 68.1% model accuracy and 66.2% balances accuracy. For hyperparameter tuning on the algorithm, we use RandomisedSearchCV. RandomisedSearchCV is preferred when the hyperparameter search space is large. The usage is similar to that of GridSearchCV, but instead of trying out all possible combinations, it evaluates a fixed number of combinations, selecting a random value for each hyperparameter at every iteration. This process covers more number of values and it is faster than GridSearchCV. The hyperparameters to be optimised are defined in a dictionary named ‘param-grid’. By running the RandomisedSearchCV to perform a 5 cross-validation , we derive the best hyperparameters as follows: ’subsample’: 0.7,’n-estimators’: 110,’max-depth’: 9,’learning-rate’: 0.06,’gamma’: 0.1

The expected best accuracy for the model is 72.1%. After fitting the model on the scaled train data and testing it on the scaled test data we obtain a model accuracy of 68.4% and balanced accuracy of 66.6

3) KNN Algorithm Results: In this study, we used the KNN algorithm after finding the optimal number of the k using the elbow method. By using the Elbow method, and looking at the Distortion vs. Number of Neighbours graph below, we found that the optimal value of k is 2. Ideally, the distortion graph (fig7) should decrease as the number of neighbours increases. In our current study, we see that there is a constant increase in the distortion indicating that the higher value of k is not feasible. The least value of k we can consider is 2.

Post finding the optimal value of k, training the KNN algorithm with scaled training data obtained after feature selection and fitting the test data, we obtain a model accuracy of 66% and balanced accuracy as 63.1%.

4) Artificial Neural Network (ANN) Results: In this study, we define a sequential neural network using the keras library. The neural network defined consists of two hidden layers and an output layer with 3 neurons and softmax activation function. 3 neurons are because we have 3 classification categories. Both the hidden layers are built with 64 neurons each and relu as activation function. This model is trained for 50 epochs. Fitting the scaled train data and compiling this model using ‘adam’ optimiser and ‘sparse-categorical-crossentropy’ as loss function, we achieve 61.84% accuracy and 61.92% balanced

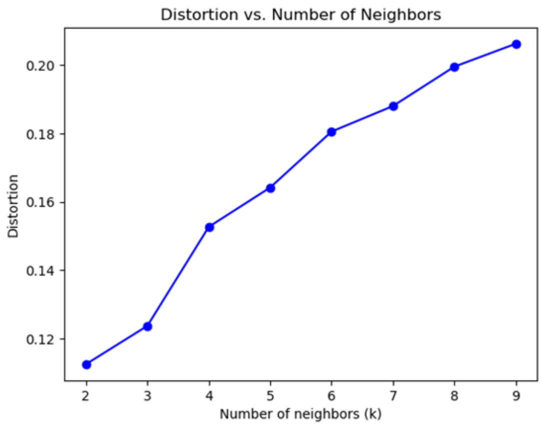


Fig. 7.

accuracy.

C. Interpretation of Results

In this study, 4 algorithms were tested for performing the task. Each algorithm was tested and assessed using the performance evaluation metrics after being trained on the pre-processed dataset. All four models provided decently good accuracies with the model (fig 8). However, Random Forest gave one of the best results of accuracy, and balanced accuracy of 74% and 66.3% respectively (Shown in Fig.no.). Although Random Forest model has the best results among our models, XGBoost, KNN and ANN are not bad for the classification task too. XGBoost achieved 68.4% accuracy and 66.6% balanced accuracy. The precision and recall measures are decently good for the model too. In spite of having imbalances in the dataset class passes, the models were able to achieve good results. KNN and ANN are not bad with the results either. (Fig). KNN has 66% accuracy and 63.1% balanced accuracy.(Fig.) ANN has 62% accuracy and balanced accuracy. (Fig.) A comparison of the accuracy result of the model is shown below:

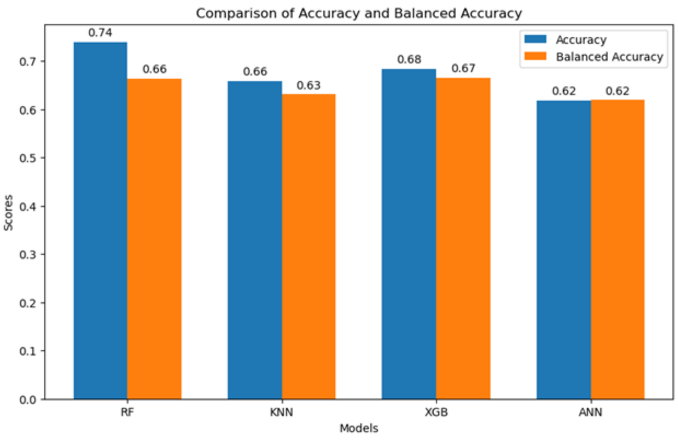
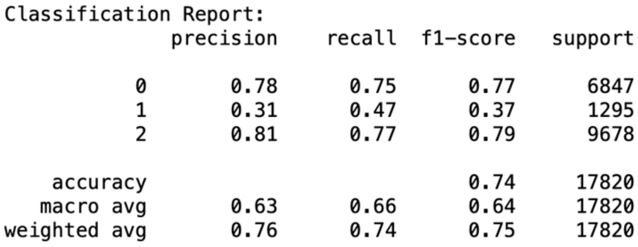


Fig. 8. Model Comparasion

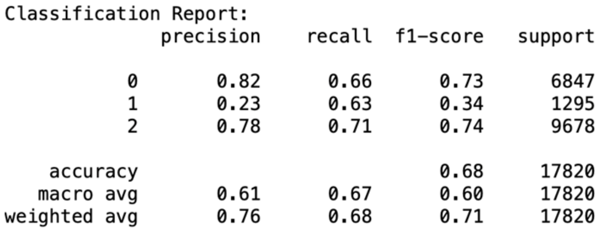


```
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.75      0.77      6847
           1       0.31      0.47      0.37      1295
           2       0.81      0.77      0.79      9678

    accuracy                           0.74     17820
   macro avg       0.63      0.66      0.64     17820
weighted avg       0.76      0.74      0.75     17820
```

Fig. 9. Random Forest Classification Report



```
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.66      0.73      6847
           1       0.23      0.63      0.34      1295
           2       0.78      0.71      0.74      9678

    accuracy                           0.68     17820
   macro avg       0.61      0.67      0.60     17820
weighted avg       0.76      0.68      0.71     17820
```

Fig. 10. XGB Classification Report

Overall, from the above results we see that all the above models have captured the complex relationships between the features and the target variable. This indicates that the models have reasonably good efficiency in predicting the working status of the water points and also forecast it for further if used correctly. Listing out the precision and recall of all the models built here for better comparison of the results: The weighted average precision for the models are as follows: • 76% for Random Forest • 76% for XGBoost • 72% for K-Nearest Neighbour • 72% for Artificial Neural Network The weighted average recall for the models are as follows: • 74% for Random Forest • 68% for XGBoost • 66% for K-Nearest Neighbour • 62% for Artificial Neural Network

All the above findings show that by using the chosen models, the functionality of the water point sites can be predicted and also forecasted for the water sites in Tanzania with high accuracy. Reliable results of prediction and forecast can be expected from these models. Random Forest and XG-Boost models demonstrate commendable use of the ensemble strategy in handling the categorical variables and capturing the non-linear relationships in the data. KNN and ANN also show decently good results and understanding of the data which can be used to predict and forecast the functionality. All the results and knowledge acquired by these models can be a great asset in assisting the resource allocation, maintenance and enhancing the decision-making processes for Tanzania's water infrastructure.

VI. Discussion

In this study, several algorithms and methodologies were used to solve the problem.

Firstly, for the pre-processing an elaborative exploratory Data Analysis was performed which provided multiple ways to handle the data. In one of the approaches 'construction-year' column was handled. No preprocessing was performed on the column as it had no missing values but eventually, it led to a lot of noise. The second approach here was to perform data wrangling method called 'binning'. In this approach, the data of 'construction-year' was grouped into decades ranging from 60s - 2010s which made data cleaner. Additionally, to this approach, columns with less unique data were discarded which reduced the multi-co-linearity in the dataset. For feature extraction Random Forest algorithm was used as it provides scores according to how important a feature is for the predicting by the model this resulted in 11 important features. 'longitude' and 'latitude' were the top two features comprising the highest importance of 30% in total.

Moving further, the data had imbalances in the output class variables. Data was divided 54.3%, 34.3% and 7.3% for functional, non-functional and need repairs respectively. To investigate, the first approach was to run the Random Forest algorithm without handling the imbalances. This resulted in 78% model accuracy and 62% Balance accuracy. Although a high accuracy was achieved, recall for need repair was considerably low. To improve this, Oversampling was performed using ADASYN to balance the data and a Random Forest model was trained again to compare. Achieved results are as shown in the (Fig.9). Precision and recall for each category have increased with a small sacrifice in accuracy.

Multiple different models are trained and their performance is measured. Random Forest's result was not disappointing so another ensemble learning algorithm called XGBoost was trained. Both options of training with default parameters and with hyperparameter tuning were explored, they resulted in similar results with a little better precision and recall in the hyperparameter tuned algorithm.

Thirdly, a K-Nearest Neighbour algorithm was trained. To determine the value of k, Elbow method is used. Results from kNN are decent but the previous models performed better.

Lastly, an Artificial Neural Network was implemented. ANN Performed with 62% accuracy and 63% balance accuracy. However, the recall values for all three classes here were significantly equal and good but there was a downfall in the precision of the class 'Functional Needs Repair'. It will affect finding the water points that are on the verge of failing

By analysing and comparing it can be concluded that the algorithms with an ensemble approach overpower the other used approaches. Among the ensemble approaches Random Forest performs better with or without oversampling and hyperparameter tuning achieving 74% accuracy and capable of achieving 78-80% accuracy in prediction. XGBoost achieves 68% accuracy. This comparative study underscores that Random Forest is the best-performing model for predicting water point functionality.

VII. Conclusion

This study paper presented is a comprehensive analysis and comparative study of different machine learning algorithms for solving the Pump it up: Data Mining the Water Table challenge. We studied four different machine learning algorithms in this. The algorithms studied are Random Forest, XGBoost, K-Nearest Neighbour, and Artificial Neural Networks. The study was aimed to accurately predict the working status of the water points in the Tanzanian region using the data from Taarifa and the Tanzanian Ministry of Water.

The KNN and ANN provided effective solutions and results in predicting the functional status. However, the ensemble technique algorithms, XGBoost and Random Forest outperformed the other two algorithms. Both XGBoost and Random Forest algorithms had competitive results. Random Forest algorithm gave the best results in both with and without data imbalances. The algorithms were capable of recording and handling the complex relationships between the features and the functional status. Random Forest's ability to handle highly categorical data out-powered the performance and demonstrated the ability to predict and forecast with the highest accuracy among all the models studied and tested in this paper.

Overall, the results of this study highlight the importance of ensemble learning algorithms in processing categorical features in challenging prediction and forecasting tasks. The results obtained from this study can be a significant contribution to the decision-making and maintenance process of the Tanzanian Ministry in efficiently allocating the resources to maintain and improve the water supply infrastructure.

Further research and study can be performed over this on exploring more ensemble methods and advanced neural network classifiers to see what results in better prediction and forecasting. Advanced feature selection techniques and data pre-processing techniques can also be discovered to improve the result and increase the robustness of the solution making it capable of handling real-time data.

References

[1] Cherif, I.L. and Kortebi, A. (2019) 'On using extreme gradient boosting (XGBoost) machine learning algorithm for Home Network Traffic Classification', 2019 Wireless Days (WD) [Preprint]. doi:10.1109/wd.2019.8734193.

[2] GARABAGHI, F.H., Benzer, S. and Benzer, R. (2021) Performance evaluation of machine learning models with ensemble learning approach in classification of water quality indices based on different subset of features [Preprint]. doi:10.21203/rs.3.rs-876980/v1.

[3] Nasir, N. et al. (2022) 'Water quality classification using machine learning algorithms', Journal of Water Process Engineering, 48, p. 102920. doi:10.1016/j.jwpe.2022.102920.

[4] Farokhzad, S. et al. (2012) 'Artificial neural network based classification of faults in centrifugal water pump', Journal of Vibroengineering, 14(4).

[5] Géron, A. (2017) Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques for Building Intelligent Systems. Beijing: O'Reilly.

[6] Murugan, A., Nair, S.A.H. and Kumar, K.P. (2019) 'Detection of skin cancer using SVM, Random Forest and Knn Classifiers', Journal of Medical Systems, 43(8). doi:10.1007/s10916-019-1400-8.

[7] Yan, X. et al. (2024) 'A comprehensive review of machine learning for water quality prediction over the past five years', Journal of Marine Science and Engineering, 12(1), p. 159. doi:10.3390/jmse12010159.

[8] Liu, Z. et al. (2024) 'Research on water resource modeling based on Machine Learning Technologies', Water, 16(3), p. 472. doi:10.3390/w16030472.