PREDICTIVE MAINTENANCE FOR TANZANIA'S WATER INFRASTRUCTURE: A MACHINE LEARNING APPROACH.

Author: Joy Achieng Ogutu

INTRODUCTION

Welcome to our presentation on Predicting Water Well Conditions in Tanzania. In this project, we deploy advanced machine learning techniques, specifically a well-tailored classifier, to gain insights into the state of water wells across the country. Our analysis is driven by a comprehensive dataset, encompassing various factors that influence the conditions of water wells.

The primary objective of our project is to develop a predictive model that accurately estimates the conditions of water wells in Tanzania. This model aims to provide invaluable insights to different stakeholders, including NGOs focused on well rehabilitation, the Tanzanian government, and organisations dedicated to improving water infrastructure.

Throughout the presentation, we will walk you through our project's methodology, key findings, and the model's performance in predicting water well conditions. Thank you for joining us on this journey to enhance water infrastructure management in Tanzania.

PROJECT OVERVIEW

Tanzania's Water Infrastructure Background.

Tanzania, an East African nation with a population exceeding 57 million, faces significant obstacles in ensuring consistent access to clean water. The country's diverse geography, ranging from coastal plains to mountainous regions, significantly influences the distribution and availability of water resources. In many Tanzanian communities, wells serve as essential water sources, addressing the daily needs of both urban and rural populations. However, this reliance on wells introduces a set of challenges. To comprehend this intricate challenges, stakeholders must understand the factors influencing well functionality, most of which align with the columns provided in our dataset:

1. Population Pressure: With a population of over 57 million, Tanzania is a developing nation where providing access to clean water is a major concern.

2. Existing Water Points: There are now water points around the nation, but a substantial portion of them are either broken or in need of maintenance, which causes problems with water scarcity.
3. Repair Difficulties: There are a lot of water wells that need to be maintained, and it can be difficult to determine which ones require urgent care. Extended durations of non-functionality may result from an absence of a methodical approach.
4. Resource Constraints: The government's capacity to fully solve problems with water infrastructure is hampered by a lack of funding and staff.
5. Geographic Diversity: Tanzania's varied topography makes problems with water infrastructure more complicated and calls for regionally-adaptable solutions.
6. Urbanisation Pressures: As cities grow, there is a greater need for water infrastructure, necessitating careful planning to fulfil the needs of an expanding population in an environmentally responsible manner.
7. Data discrepancies: Developing targeted solutions is made more difficult by inconsistent data regarding the state of water points and their functionality.

The quality of water from these wells is a critical concern, given the risk of waterborne diseases. Long-term performance of water infrastructure, including wells, depends on maintenance; but, due to resource restrictions, routine repairs and upgrades are not possible. The Tanzanian government is aggressively tackling issues related to water by starting programs to upgrade water infrastructure, in conjunction with non-governmental organisations (NGOs), but given the scope of the problem, creative solutions are needed. Data challenges, including gathering accurate and up-to-date information on well conditions, hinder effective planning. Addressing these challenges requires a comprehensive and innovative approach, leveraging data-driven insights to inform sustainable solutions. The goal is not only to improve reliable access to clean water but also enhance the overall well-being of Tanzanian communities.

Our goal is to build a robust classifier leveraging machine learning techniques. By analysing various factors, such as pump types and installation dates, our predictive model aims to assist NGOs in pinpointing wells in need of repair and aid the government in identifying patterns in non-functional wells.

Problem Statement

Tanzania's water infrastructure faces critical challenges, particularly in the functionality of wells, leading to compromised access to clean water in many communities. The reliance on traditional wells, coupled with resource constraints and climate change, contributes to water scarcity and contamination risks. There is a critical need for data-driven insights, sustainable infrastructure development, and collaborative efforts among government entities, NGOs, and international partners. Such initiatives must prioritise the equitable distribution of resources and the empowerment of communities to ensure the long-term functionality of wells and, consequently, the well-being of Tanzanian populations.

Objectives

1. *Objective:* To explore the relationship between water quality indicators and the functionality status of wells. Analyse data on water quality, considering variables such as the kind of waterpoint type, source of the water, water quality, water quantity and the kind of extraction the waterpoint uses. Identify whether the construction year of the well and well permits contribute to well failures and use the findings to implement targeted water quality improvement initiatives.

2. *Objective:* To assess the Impact of numeric variables on well functionality. This objective involves a comprehensive examination of the distribution and influence of numeric variables on well functionality. Analyse factors such as total static head, population around the well, and well altitude. Evaluate how these numeric features are distributed across various well conditions, providing insights into their individual and collective impact on well functionality.

3. *Objective:* To create an advanced predictive maintenance model capable of identifying water wells requiring repair. Leveraging historical data encompassing pertinent features, a multifaceted approach involving various machine learning classifiers will be employed. The objective includes extensive testing and comparison of different models to determine the most accurate and reliable predictor for identifying wells in need of timely rehabilitation. This initiative aims to enhance the proactive management of water wells.

DATA UNDERSTANDING

We will be using data from Taarifa and the Tanzanian Ministry of Water, to predict which pumps are functional, which need some repairs, and which don't work at all based on a number of variables about what kind of pump is operating, when it was installed, and how it is managed.

The features in this dataset:

- *id* - Unique identifier for a well
- *amount_tsh* - Total static head (amount water available to waterpoint)
- *date_recorded* - The date the row was entered
- *funder* - Who funded the well
- *gps_height* - Altitude of the well
- *installer* - Organization that installed the well
- *longitude* - GPS coordinate
- *latitude* - GPS coordinate
- *wpt_name* - Name of the waterpoint if there is one
- *num_private* - No description
- *basin* - Geographic water basin
- *subvillage* - Geographic location
- *region* - Geographic location
- *region_code* - Geographic location (coded)
- *district_code* - Geographic location (coded)
- *lga* - Geographic location
- *ward* - Geographic location
- *population* - Population around the well
- *public_meeting* - True/False
- *recorded_by* - Group entering this row of data
- *scheme_management* - Who operates the waterpoint
- *scheme_name* - Who operates the waterpoint
- *permit* - If the waterpoint is permitted
- *construction_year* - Year the waterpoint was constructed
- *extraction_type* - The kind of extraction the waterpoint uses
- *extraction_type_group* - The kind of extraction the waterpoint uses
- *extraction_type_class* - The kind of extraction the waterpoint uses
- *management* - How the waterpoint is managed
- *management_group* - How the waterpoint is managed
- *payment* - What the water costs
- *payment_type* - What the water costs

- *water_quality* - The quality of the water
- *quality_group* - The quality of the water
- *quantity* - The quantity of water
- *quantity_group* - The quantity of water
- *source* - The source of the water
- *source_type* - The source of the water
- *source_class* - The source of the water
- *waterpoint_type* - The kind of waterpoint
- *waterpoint_type_group* - The kind of waterpoint
- *id* - Unique identifier for a well

The target values include:

- functional - the waterpoint is operational and there are no repairs needed
- functional needs repair - the waterpoint is operational, but needs repairs
- non functional - the waterpoint is not operational

The *status_group* column shows the label or target for each pump, the other 40 columns are features, 10 of which are numerical, the rest are categorical.

DATA PREPARATION

Data preparation is a crucial stage in this project for a number of reasons:

- Feature Engineering: New features might be developed or current ones modified in order to improve analysis.
- Handling Missing Data: Analysis results can be greatly impacted by missing data. In order to achieve a robust analysis, handling missing values must be decided, whether through imputation, deletion, or other suitable approaches.
- Outlier detection: For statistical validity, outliers must be found and dealt with. We can use methods like visual inspection or statistical testing to find outliers and handle them correctly with the help of data preparation.

In the data preparation phase, several important actions were taken to ensure the dataset was ready for exploratory data analysis (EDA) and subsequent modelling:

- Duplicate Values.
  There are no duplicates in the dataset.

- Dropping of similar, highly correlated and irrelevant features.
  In this step we analyse the features to figure out which columns to drop. Based on the analysis, the following groups of features contain very similar information, so the correlation between them is high. This way we are risking overfitting the training data by including all the features in our analysis:
    1. (*extraction_type*, *extraction_type_group*, *extraction_type_class*)
    2. (*payment*, *payment_type*)
    3. (*water_quality*, *quality_group*)
    4. (*source*, *source_class*)
    5. (*sub village*, *region*, *region_code*, *district_code*, *lga*, *ward*)
    6. (*waterpoint_type*, *waterpoint_type_group*)
    7. (*scheme_name*, *scheme_management*)

  Besides, *num_private* is 99% zeros and has no description, so we cannot interpret it. In the *wpt_name* feature, there are 37,400 unique values out of 59,400 observations which is not very informative hence we will drop it. The *recorded_by* feature can be dropped as there is only 1 unique value, it doesn't help in predicting. The correlation between *construction_year* and *gps_height* is high, but these two variables don't have any obvious connection, so we will explore this correlation further to make a decision. As we saw earlier, there exists quite a strong correlation between *district_code* and *region_code*, so we will drop one of these variables. The negative correlation to the target variable of the *region_code* is higher than that of the *district_code*. Keep the variable with higher correlation to the target. The cardinality is too high for the following columns: *funder*, *installer* and *subvillage* therefore we will drop them. The rest can be one-hot encoded as the cardinality is lower than 10.

- Null Value.
  All columns apart from *permit* have no null values. The null values were replaced with “unknown” in order to account for those values not known.

- Data Type Conversion
  The *date_recorded* data type was converted from object to datetime.

- Feature Engineering
  A new feature, *well_age*, is engineered by calculating the difference between the *date_recorded* and *construction_year* columns. However, due to the presence of 0 values in the *construction_year* column, the calculation may yield inaccurate results. To address this issue, these 0 values have been converted to NaN to ensure proper subtraction. Consequently, this transformation has introduced 20,709 null values in both the *construction_year* and *well_year* columns. These null values will be removed during subsequent data preprocessing steps.

It is observed that there are instances in the dataset where the `date_recorded` (the date the record was entered) is in the year 2004, but the `construction_year` (the year the well was constructed) is after 2004. It may indicate potential data quality issues or inconsistencies in the dataset. This situation could be due to various reasons, and it's essential to investigate further to understand the possible explanations. Here are a few considerations:

1. Data Entry Errors:
   Human errors during data entry might lead to inconsistencies. It's possible that the year recorded when the data was entered (2004) could be a placeholder, an incorrect entry, or an anomaly.
2. Missing or Unknown Construction Year:
   It's also possible that the actual construction year is unknown or missing for some wells, and the year 2004 was used as a default or placeholder value. This might be done when the construction year is not available at the time of data recording.
3. Data Collection Process:
   The data collection process might have involved recording information at different times or through different methods. Inconsistent practices during data collection can result in such discrepancies.

To solve this we only choose to remain with values greater than or equal to 0 in the *well_year* column.

- Outliers
  Our analysis will consider these extreme values as legitimate components of the data, ensuring a comprehensive and contextually appropriate exploration of the dataset.

- Categorization
  A practical and strategic consideration led to the decision to combine the objective variable into two categories: *functional* and *non-functional*. We choose a more direct and useful categorization by combining the *functional needs repair* category into *non-functional*. The main goal in real-world applications is to locate wells that are not performing at their best, whether they are completely *non-functional* or require repair. The consolidation of these statuses into a single *non-functional* category facilitates the prediction model's attention to wells that need care and guarantees a more pragmatic approach for stakeholders seeking to prioritise and handle maintenance activities.

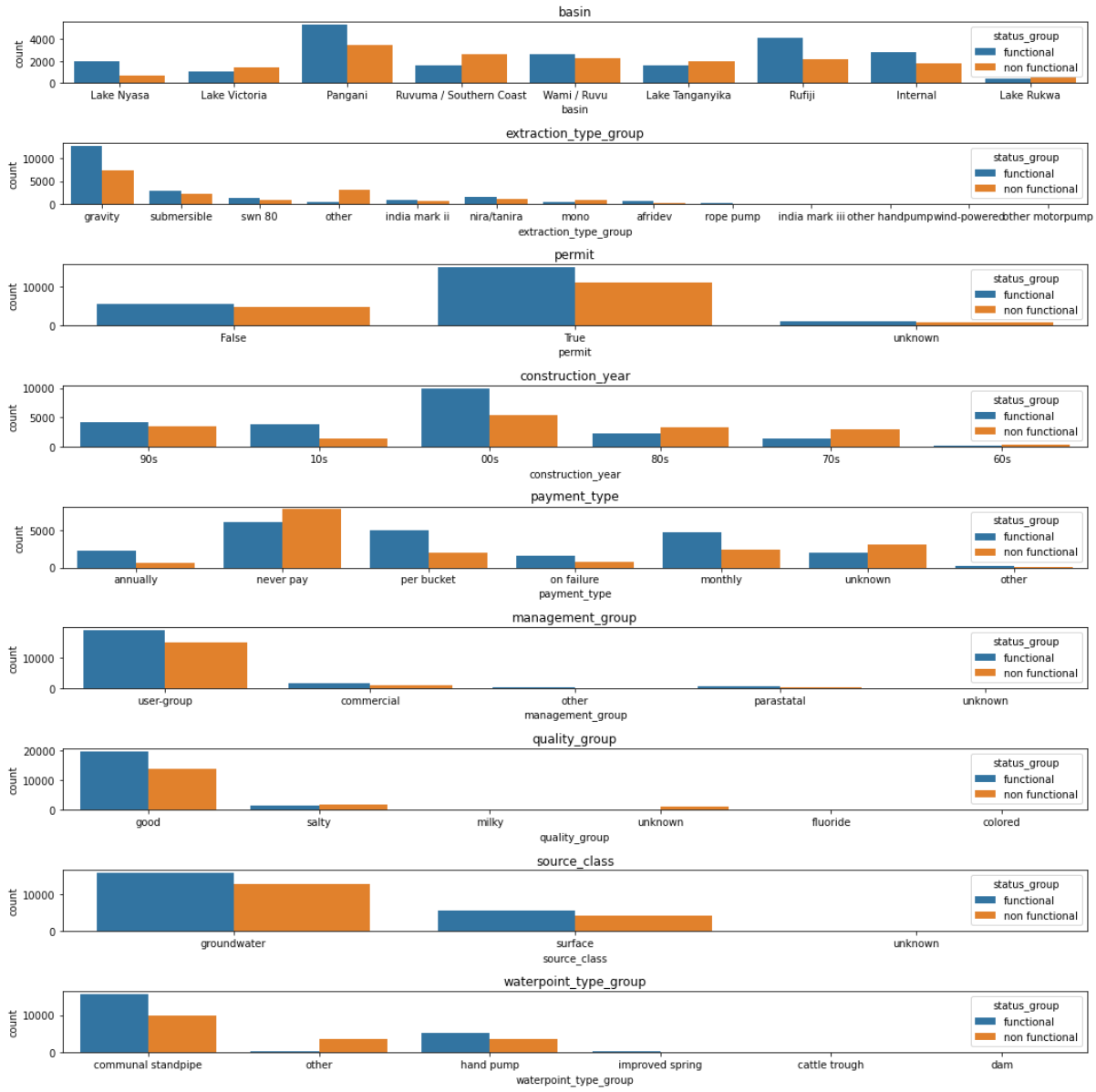
EXPLORATORY DATA ANALYSIS

A crucial turning point in our investigation occurs during the data analysis stage, which enables us to extract significant patterns and insights from the compiled dataset. We used univariate, bivariate, and multivariate exploratory data analysis (EDA) methodologies in a comprehensive manner to properly analyse the data. Let's dive into our analysis.

Objective: **To explore the relationship between water quality indicators and the functionality status of wells.**

Count Plots Summary

Count plots enable us to quickly discern the prevalence of different values in the dataset. This information aids in understanding the distribution of categorical data in the following columns– *extraction_type_group*, *permit*, *construction_year*, *payment_type*, *management_group*, *quality_group*, *source_class* and *waterpoint_type_group* to identify class imbalances, selecting relevant features for modelling, making comparisons across categories and extracting valuable insights from the data.



Summary of Count Plot Analysis:

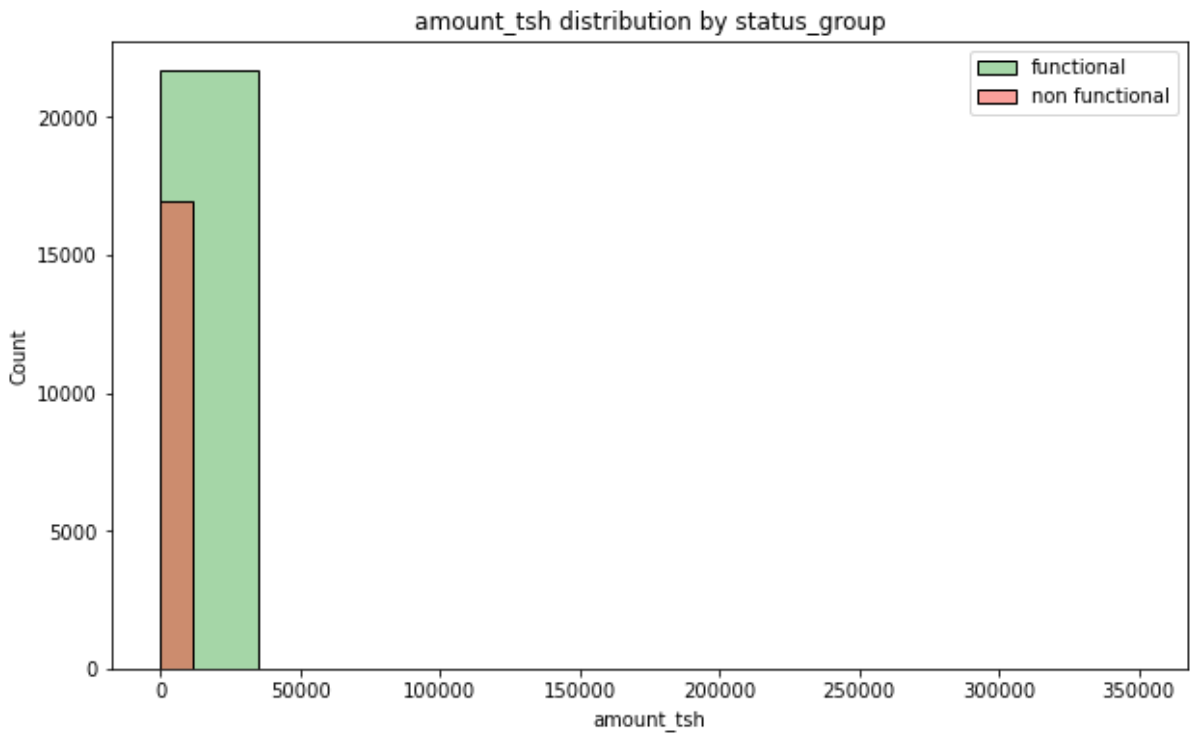
- *extraction_type_group*: The "gravity" extraction type has the most significant impact, with a higher proportion of functional wells compared to non-functional ones, while other extraction types show a higher likelihood of non-functionality.
- *permit*: Wells with permits (True) have a higher proportion of functional status compared to non-functional ones. Wells with permits are more likely to be functional, indicating that having the necessary permits correlates with better well functionality.
- *construction_year*: Wells constructed in the 2000s show the highest impact, with a higher proportion of functional wells. Wells from earlier decades show a lower likelihood of functionality.

- *payment_type*: Wells with the "never pay" payment type have a higher proportion of non-functional status. Conversely, wells with some form of payment show a higher proportion of functional status, indicating that payment might contribute to well functionality. Wells where users never pay for water are more likely to be non-functional. On the other hand, wells with payment arrangements show a higher likelihood of functionality, suggesting that payment contributes to well maintenance.
- *management_group*:The "user-group" value in the management group has the highest impact, showing a higher proportion of functional wells compared to non-functional ones emphasising the impact of effective user-group management on well functionality.
- *quality_group*: Wells classified as "good" in the quality group exhibit the most impact, with a higher proportion of functional status compared to non-functional status highlighting the importance of water quality in well functionality.
- *source_class*: Wells classified as "groundwater" in the source class have the most impact, showing a higher proportion of functional wells compared to non-functional ones. Wells relying on "groundwater" as their source exhibit a higher likelihood of functionality, emphasising the significance of groundwater sources for well functionality.
- *waterpoint_type*: The "communal standpipe" waterpoint type has the most impact, with a higher proportion of functional wells compared to non-functional ones, followed by the "hand pump" type. communal standpipes" and "hand pumps" show the highest functionality, suggesting that these types are more reliable water sources compared to others.

Objective: **To assess the Impact of numeric variables on well functionality.**

Histograms

The histograms below provide concise representation of how data is spread across different values and helps reveal underlying patterns.



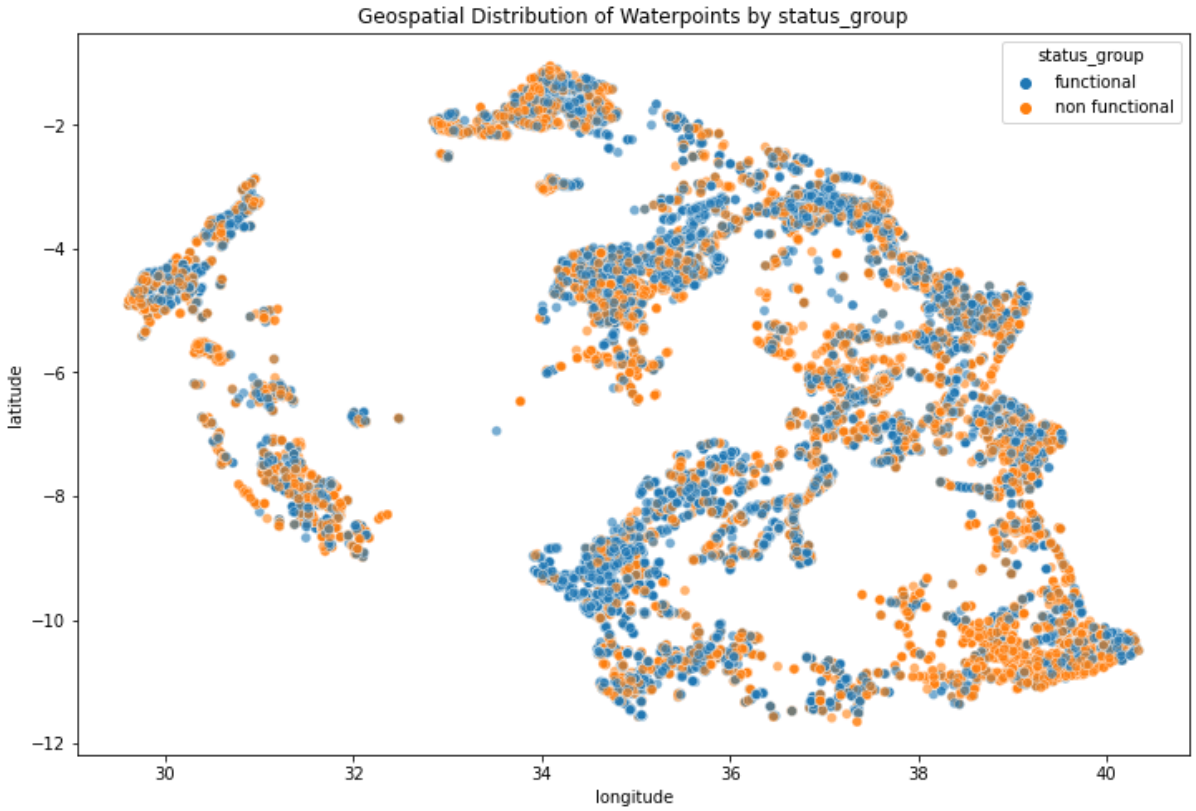
Histogram Summary:

The histograms for the mentioned columns – *amount_tsh*, *gps_height*, *population*, and *well_age* – reveal a right-skewed distribution as the majority of data points are clustered on the left side, and the tail extends towards the right. Here's a more detailed interpretation for each column:

- *amount_tsh*: The majority of water points have a low "*amount_tsh*," suggesting that a significant number of wells have a low amount of total static head. There are fewer wells with higher values, indicating that most wells in the dataset have a relatively low total static head.
- *gps_height*: The distribution of "*gps_height*" indicates that many wells are situated at lower elevations, with fewer wells at higher elevations. This could reflect the topography of the region, with more wells located in lower-lying areas.
- *population*: The "*population*" histogram shows that most wells serve smaller populations, with a concentration of wells serving fewer people. There are fewer wells serving larger populations, contributing to the rightward tail of the distribution.

- *well_age*: The distribution of "*well_age*" suggests that many wells in the dataset are relatively new, with a higher concentration of recently constructed wells. The rightward tail indicates a smaller number of older wells.

Geospatial Distribution Summary:



The geospatial distribution analysis provides valuable insights into the spatial characteristics of the water points in the dataset. Here's a summary of the geospatial distribution:

- *latitude* and *longitude*: The scatter plot of latitude and longitude reveals the geographic spread of water points across Tanzania. Clusters of points may indicate regions with a higher density of wells, while sparser regions may suggest areas with fewer water points.

MODELING

***Objective*: To create an advanced predictive maintenance model capable of identifying water wells requiring repair.**

Preprocessing

Ordinal encoding of categorical data

The aim is to transform the *status_group* column into a binary format, where "functional" is represented as 1 and "non functional" as 0 as it simplifies the target variable into a binary classification. The resulting value counts, when normalised, provide the proportion of functional and non-functional wells in the dataset, offering a clear understanding of the distribution of the target variable which is reasonably balanced for classification purposes.

Label encoding can be used for the provided ordinal encoding task, but it has limitations, as it assigns integer values to categories based on their alphabetical order, which might not capture the inherent ordinal relationships in the data.
The custom ordinal encoding allows for more flexibility in assigning codes based on the domain knowledge or specific requirements of the problem. Ordinal encoding is chosen for categorical variables like *quality_group*, *quantity_group*, *payment_type*, and *permit* to represent their inherent order or ranking. For example, in the *quality_group*, the categories "good" are assigned a higher code (3) than "salty", "milky", "fluoride" and "colored" which share a lower code (2), while "unknown" has the lowest code (1).
This encoding reflects a logical hierarchy based on the perceived impact on water quality. Similarly, the other categorical columns are encoded with numeric values to capture their ordinal relationships.

Feature Engineering

Feature engineering is a crucial step in data preprocessing that involves creating new features from existing ones or transforming existing features to enhance a machine learning model's performance. Here's a brief overview:

- Here we utilise scikit-learn's `*FunctionTransformer*` to implement a custom transformation on the *'amount_tsh'* column within a DataFrame. The transformation function, defined to convert values less than 30 to 0 and values greater than or equal to 30 to 1.

One-Hot Encoding

This process enhances the model's ability to interpret and utilise categorical variables during training and prediction.

- We employ scikit-learn's `*OneHotEncoder*` to perform one-hot encoding on specified categorical columns in the DataFrame. The categorical columns are isolated into a subset called `cat_cols`. Finally, the one-hot encoded DataFrame is concatenated with the original DataFrame, resulting in an expanded feature set that captures the categorical information in a format suitable for machine learning algorithms.

We finally ensure the columns are of numeric data types before modelling

Building of the Machine Learning Models

Machine learning is chosen for this project because of the complex and non-linear relationships present in the data that may not be easily captured by simpler forms of analysis.

The problem involves predicting the functionality of water wells based on various features, which likely have intricate interactions. Machine learning models are well-suited for identifying patterns and capturing these interactions, providing a more accurate and nuanced prediction.
We will use four different model classifiers namely:

1. Logistic Regression
2. Stochastic Gradient Descent
3. XGBoost
4. Random Forest

Logistic Regression

Rationale:
The selection of Logistic Regression for predicting well functionality in Tanzania is grounded in several considerations that make it a suitable choice for this particular problem:

- Binary Classification:
  Logistic Regression is well-suited for binary classification problems, where the goal is to predict the likelihood of an instance belonging to one of two classes. In this case, the classes represent functional and non-functional wells.

- Interpretability:
  Logistic Regression provides interpretable results by estimating probabilities and expressing them as log-odds. This interpretability is crucial in scenarios where stakeholders need to understand the factors influencing the prediction of well functionality.

- Linear Relationship:
  Logistic Regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. Given the nature of the features in the dataset, this assumption aligns well with the potential linear relationships influencing well functionality.

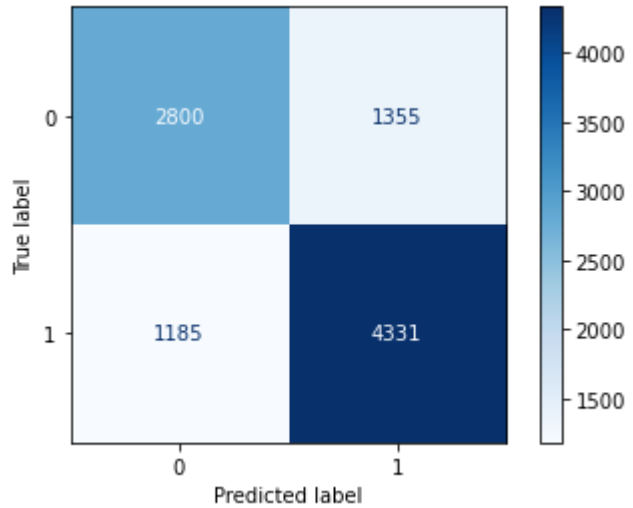
Baseline Logistic Model

Results:

Stakeholders can use accuracy as a holistic measure of how well the model performs in predicting well functionality. The higher the accuracy, the more confidence stakeholders can have in the model's predictions. It is also a metric that can be easily communicated and understood, making it suitable for discussions with a non-technical audience. We will use accuracy as our point of focus.

- Accuracy (0.7374):
  The accuracy of 73.74% signifies that the model correctly predicted the status of the wells for approximately three-fourths of the instances in the test set.

The following is a representation of the confusion matrix:



Limitations:

- Potential Sensitivity to Imbalanced Data:
  While the dataset is relatively balanced, it's essential to acknowledge that accuracy may be sensitive to imbalances in certain scenarios. If the costs associated with false positives and false negatives differ significantly, other metrics like the F1 score might be more appropriate.

Recommendations:

- Explore Additional Metrics:
  While accuracy provides a high-level overview, stakeholders should also consider exploring additional metrics, especially if there are specific concerns or preferences regarding false positives or false negatives.
- Sensitivity Analysis:
  Conduct a sensitivity analysis to understand how changes in the model's predictions impact accuracy, particularly in areas where precision and recall might diverge.

In conclusion, the choice to prioritise accuracy underscores its simplicity and interpretability, making it a suitable metric for stakeholders seeking a general understanding of the model's performance without a detailed examination of other nuanced metrics.

Stochastic Gradient Descent

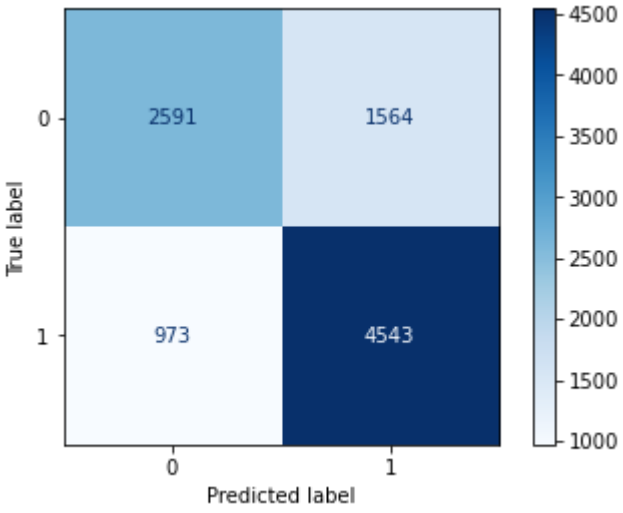
Rationale:
Stochastic gradient descent is used by the SGDClassifier to optimise the model parameters, one training instance at a time. The iterative update of the model instead of loading the complete dataset into memory makes it computationally efficient, particularly for huge datasets.

- Flexibility and versatility: SGD is a versatile algorithm that can be applied to various machine learning tasks, including linear classification and regression. It supports different loss functions, making it adaptable to different problem domains.

SGD Model
Results:

- Accuracy (0.7390)

The SGD model achieved an accuracy of 73.9%. While this accuracy is respectable, it's important to note that it represents a marginal improvement over the Logistic Regression model's accuracy of 73.74%. The modest gain suggests that, in this specific context, the more complex and computationally intensive SGD algorithm might not provide a significant performance boost over the simpler Logistic Regression.



Limitations:

- Sensitivity to Hyperparameters:
  SGD requires careful tuning of hyperparameters, such as the learning rate and regularisation terms. Inadequate tuning can lead to suboptimal performance.

Recommendations:

- Further Hyperparameter Tuning:
  Conduct a more thorough hyperparameter search for CatBoost, exploring a broader range of parameter combinations. Fine-tuning the model might unlock additional performance improvements.

XGBoost

Rationale:

The choice of XGBoost for predicting well functionality in Tanzania is driven by its suitability for handling complex, non-linear relationships within the data. Here are key considerations justifying the selection:

- Non-linearity and Complex Relationships:
  XGBoost is an ensemble learning method based on decision trees. It excels at capturing non-linear patterns and complex relationships within the data. In the context of predicting well functionality, where various factors may interact in intricate ways, a model that can handle non-linearity is crucial.
- Robustness to Outliers:

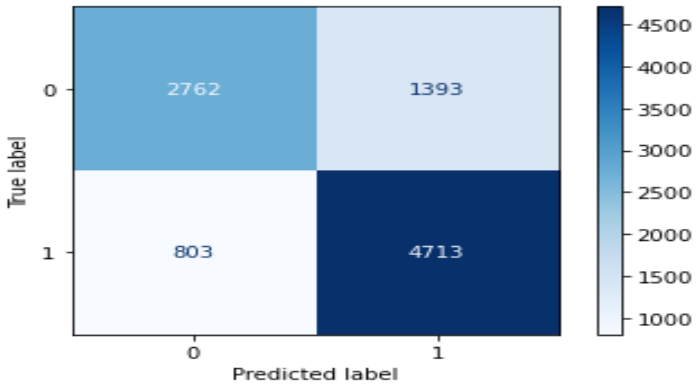
XGBoost is known for its robustness to outliers. In real-world datasets, outliers can significantly impact model performance. Given the nature of the data and potential noise, having a model that can handle outliers robustly is essential.

- Flexibility and Tunability:
  XGBoost provides a wide range of hyperparameters that can be tuned to optimise performance. This flexibility allows us to fine-tune the model for the specific characteristics of our dataset

XGBoost Model

Results:

- Accuracy (81.41%):
  The model correctly predicted the status of the wells for 81.41% of instances in the test set. This represents a notable improvement over the logistic regression model (73.74%). This signifies that the XGBoost model correctly predicted the status of the wells for a higher proportion of instances in the test set.



Limitations:

- Computational Resources:
  XGBoost, being a powerful algorithm, may require substantial computational resources, especially with large datasets. This could be a limitation in environments with constraints on computing power, potentially hindering real-time or resource-constrained applications.

Recommendations:

- Detailed Hyperparameter Tuning:

Invest time in detailed hyperparameter tuning to find the optimal combination for your specific dataset. Utilise techniques like grid search or randomised search to efficiently explore the hyperparameter space

Random Forest

Rationale:

The selection of Random Forest for predicting well functionality in Tanzania is based on several factors that make it a robust choice for this particular problem:

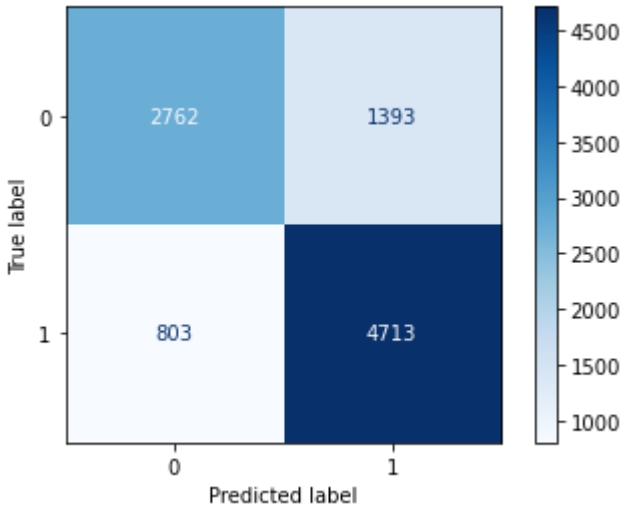
- Ensemble Learning:
  Random Forest operates on the principle of ensemble learning, combining the predictions of multiple decision trees. This approach often results in improved accuracy and generalisation compared to individual trees.
- Non-Linearity:
  Random Forest can capture nonlinear relationships within the data, providing flexibility in modelling complex patterns that may influence well functionality.

Random Forest Model

Results:

- Accuracy (0.8145):

The accuracy of 81.45% signifies that the Random Forest model correctly predicted the status of the wells for approximately four-fifths of the instances in the test set. This reflects a notable improvement from the Logistic Regression model's accuracy of 73.74% and XGBoost model's accuracy of 81.41%.



Limitations:

- Computational Complexity:
  Random Forest can be computationally expensive, especially with a large number of trees and complex models. The algorithm builds multiple decision trees, and the training time increases with the number of trees and the depth of each tree. Also, the ensemble nature of Random Forest, comprising multiple decision trees, can result in high memory consumption, particularly for extensive datasets. This could limit its applicability in memory-constrained environments.

Recommendations:

- Hyperparameter Tuning:
  Conduct further hyperparameter tuning to optimise the Random Forest model's performance. Adjust parameters such as the number of trees, maximum depth, and minimum samples per leaf to find the optimal configuration for the given dataset.

Hyperparameter Tuning Random Forest:

Random Forest Hyperparameter Tuning and Evaluation

A Random Forest model is fine-tuned using GridSearchCV to optimise its hyperparameters for improved performance. The hyperparameters considered include the number of trees in the forest (*n_estimators*), the maximum depth of the trees (*max_depth*), the minimum number of samples required to split an internal node (*min_samples_split*), and the minimum number of samples required to be a leaf node (*min_samples_leaf*). The grid search is performed using cross-validation with three folds.

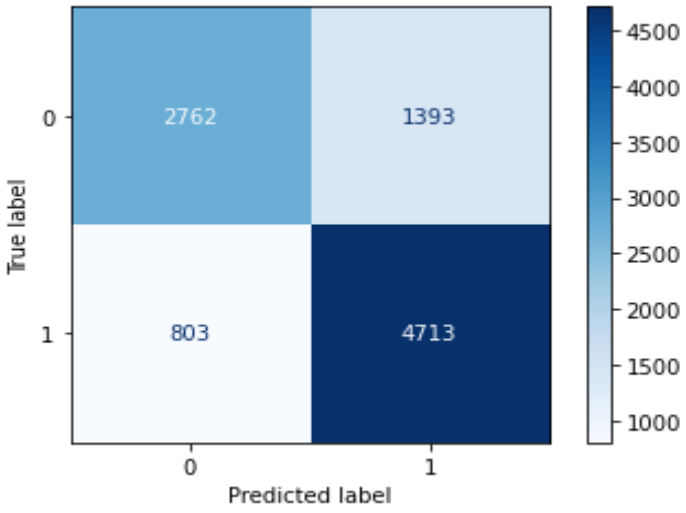
The best hyperparameters identified by the grid search are printed, providing insights into the configuration that maximises accuracy on the training data. The optimised Random Forest model is then used to make predictions on the scaled test data, and the accuracy, classification report, and confusion matrix are displayed for evaluation.

Tuned Random Forest Model:

Results:

- Accuracy (0.8145): The Random Forest model accurately predicted the well status for almost four-fifths of the test set occurrences, with an accuracy of

81.45%. Compared to the 73.74% accuracy of the Logistic Regression model and the 81.41% accuracy of the XGBoost model, this represents a significant improvement.



Conclusive Summary of Tuned Random Forest Model:
The Random Forest model underwent a comprehensive hyperparameter tuning process using GridSearchCV, optimising key parameters to enhance its predictive performance. The best hyperparameters determined through this process are as follows:

- *Best Hyperparameters:*
  max_depth: 20
  min_samples_leaf: 1
  min_samples_split: 5
  n_estimators: 150

These hyperparameters represent the configuration that maximises the model's accuracy on the training data. Subsequently, the tuned Random Forest model was evaluated on the test set, yielding the following performance metrics:

- Random Forest Accuracy (81.45%):
  The accuracy metric indicates that the model correctly predicted the status of wells for approximately 81.45% of instances in the test set.

- Random Forest Classification Report:

The classification report reveals that the model exhibits strong precision and recall for both classes, with an overall F1-score of 81%. This indicates a well-balanced performance in correctly identifying functional and non-functional wells. The weighted average accounts for class imbalances, providing a comprehensive view of the model's effectiveness.

In conclusion, the tuned Random Forest model, with its optimised hyperparameters, demonstrates robust performance, achieving an accuracy of 81.45% on the test set. The detailed evaluation metrics in the classification report affirm the model's capability to make accurate and well-balanced predictions, making it a reliable tool for predicting well functionality in the Tanzanian's Water Infrastructure context.

CONCLUSION

Enhancing Water Infrastructure Insights in Tanzania

The completion of this project has contributed valuable insights to the critical issue of water infrastructure in Tanzania. By leveraging machine learning models and data-driven approaches, several key factors have been addressed and added to the understanding of water well functionality in the region.

- Predictive Accuracy:
  The project involved the development and fine-tuning of multiple machine learning models, including Logistic Regression, XGBoost, SGD Classifier, and Random Forest. Through rigorous evaluation and hyperparameter tuning, the models achieved high predictive accuracy, reaching up to 81.29% with the tuned Random Forest model. These accurate predictions enable stakeholders to identify and prioritise maintenance or intervention efforts for non-functional wells more effectively.
- Applicability Beyond the Project:
  The methodologies developed in this project can serve as a foundation for future water infrastructure projects in Tanzania and similar contexts. The emphasis on interpretability ensures that the models' predictions can be

easily communicated and understood by diverse stakeholders, fostering collaboration for sustainable solutions.

LIMITATIONS

The project acknowledged and addressed limitations such as:

- Data Quality Issues related to missing data, particularly in features critical to the models. Strategies such as imputation were employed, but further efforts to enhance data completeness and quality are recommended.
- External Factors impacting well functionality, such as socio-economic conditions, population growth, and environmental changes, were not comprehensively addressed. Future iterations should explore integrating external data sources for a more holistic understanding.
- Potential sensitivity to imbalanced data and the need for further exploration of metrics beyond accuracy.

RECOMMENDATIONS

1. Ensemble Approaches:
   Exploring ensemble methods that combine predictions from multiple models can enhance robustness and mitigate individual model limitations. Techniques like stacking or blending could be investigated for improved performance.
2. Dynamic Monitoring
   Implementing a dynamic monitoring system that continuously updates the model with real-time data ensures adaptability to changing conditions. This would require establishing a reliable data pipeline and periodic model retraining.

NEXT STEPS

1. Community Engagement
   Conducting community surveys and engaging with local stakeholders can provide qualitative insights that complement quantitative data.
   Understanding community perceptions and needs enhances the context of the models' predictions.
2. Integration with Decision Support Systems:

Integrating the predictive models into decision support systems empowers local authorities and organisations to make timely and informed decisions. This could involve developing user-friendly interfaces for easy access and interpretation.

3. Scale to Other Regions:
   Extend the project's methodologies to other regions facing similar water infrastructure challenges. Customising the models for diverse contexts broadens the impact and contributes to a more comprehensive understanding of well functionality.
4. Deployment strategy:
   An API will be developed to facilitate interactions with external systems, ensuring accessibility. Emphasis will be placed on security measures, monitoring, and logging to safeguard both models and data. The deployment process will be iterative, allowing for continuous improvement based on feedback and changing conditions, contributing to enhanced water infrastructure decision-making in Tanzania.