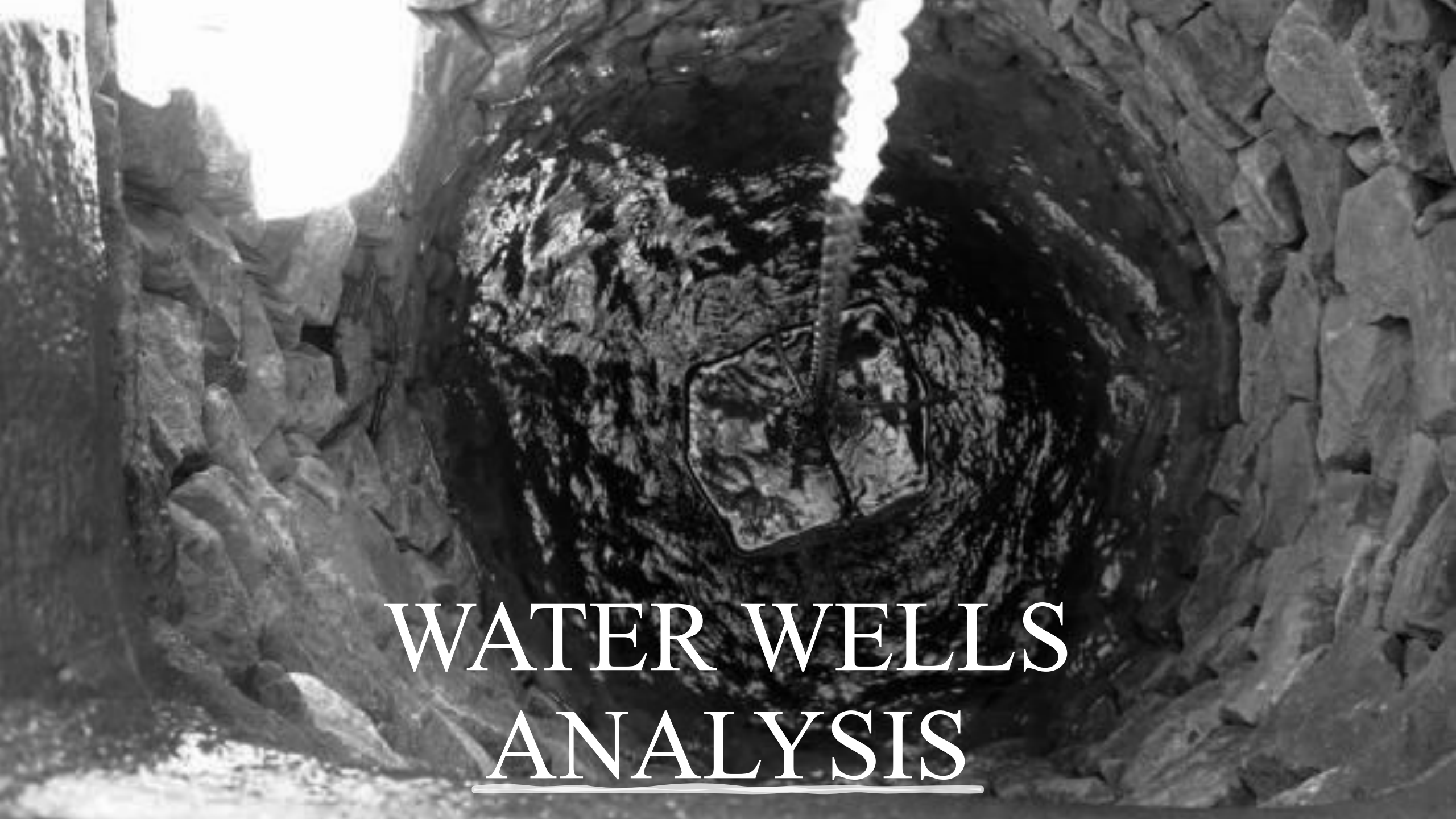


WATER WELLS ANALYSIS

Overview

This presentation is divided into:

- Business Understanding
  - Objectives
  - Business Questions
  - Success criteria
- Data Understanding
  - Exploratory Data Analysis
- Models Evaluation
- Conclusions
- Recommendations

Business Understanding

- In Tanzania, water wells become non-functional due to various factors such as:
  - quantity of water available in the wells
  - the source of the water feeding the well
  - the group managing the well
- Identifying non-functional wells will help prioritize maintenance efforts and improve water infrastructure planning.
- This analysis aims to develop a machine learning classifier that predicts the condition of water wells based on the availed data.
- The model will categorize wells as functional or non-functional.

Objectives

- ➔ Develop a classification model to predict whether a well is functional or non-functional using historical data.
- ➔ Implement and compare multiple algorithms to identify the most effective model for well functionality prediction.
- ➔ Improve prediction accuracy through feature selection, hyperparameter tuning, and handling class imbalances.

Business Questions

- ➔ What factors contribute most to well failures?
- ➔ Which regions have the highest concentration of non-functional wells?
- ➔ How does the construction year affect well failure rates?

Success Criteria

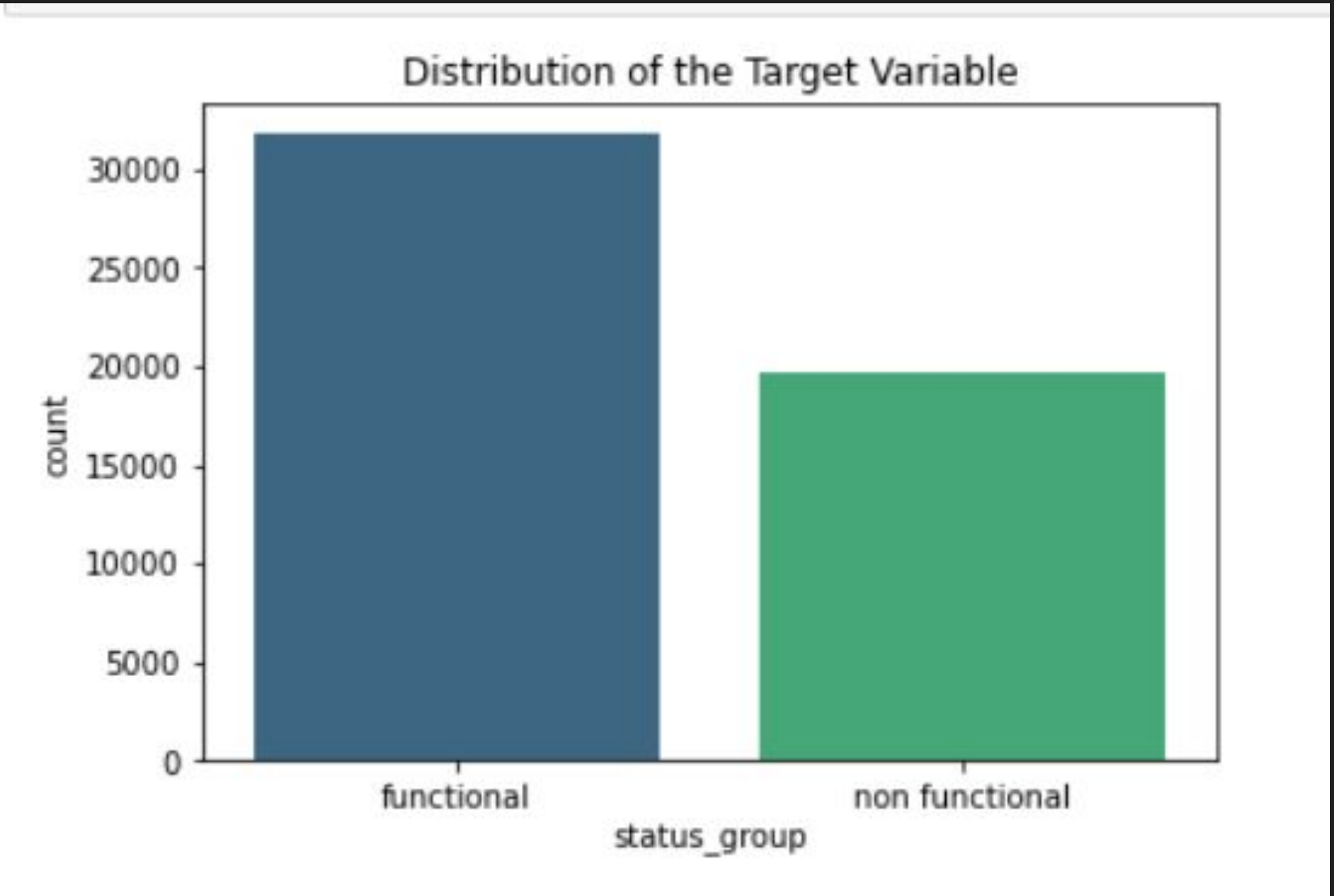
- A highly accurate and reliable model that effectively classifies water wells as functional or non-functional, ensuring strong performance in accuracy, recall, and F1-score.
- Identification of high-risk regions in Tanzania where wells require more attention and maintenance efforts.
- Understanding the impact of construction year on well functionality to support strategic planning and scheduling of maintenance based on well age.
- Identifying features that highly impact well functionality for strategic allocation of resources during future constructions.

Methods

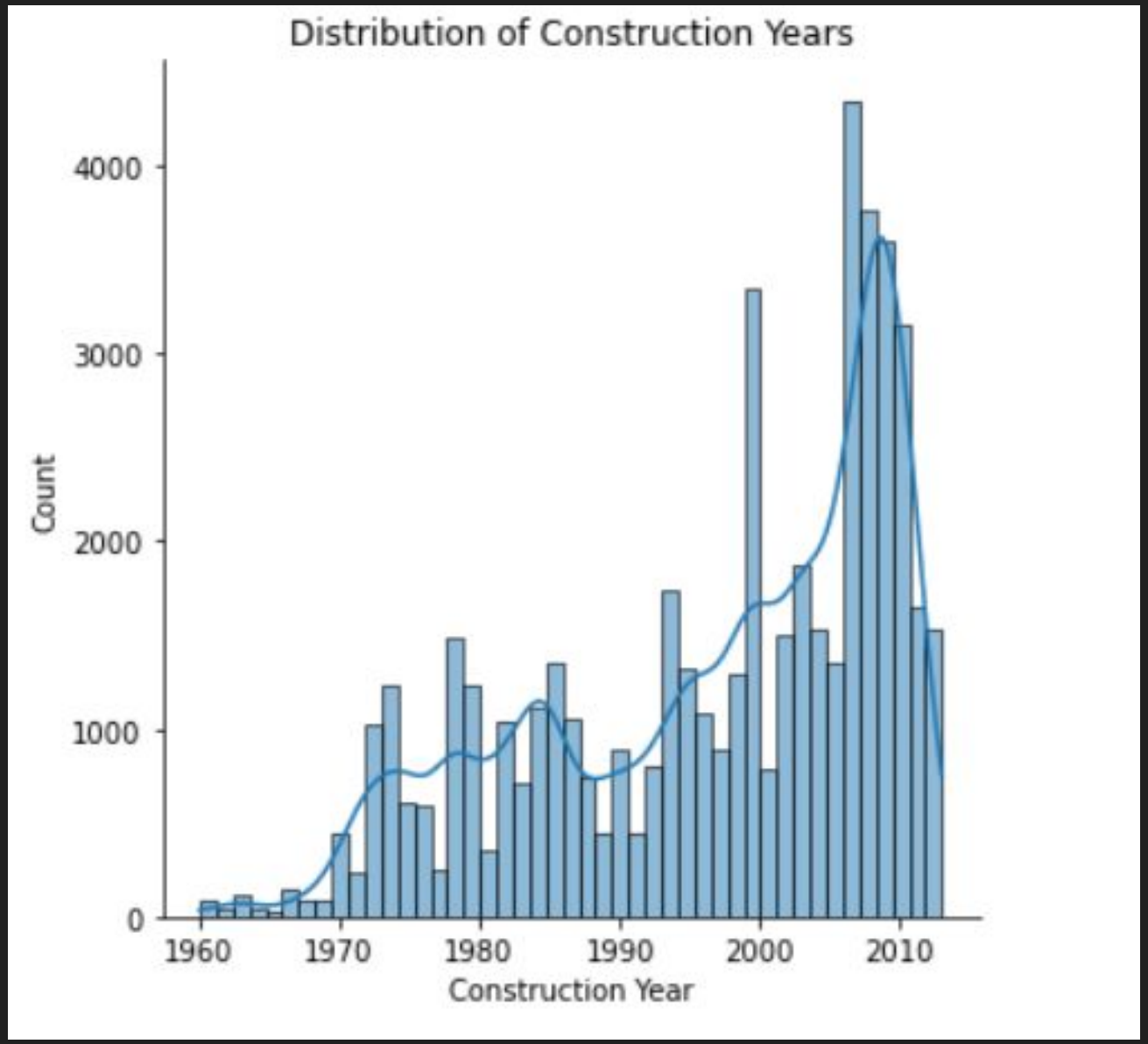
The methods used in the analysis are as follows:

- ➔ Data Collection
- ➔ Exploratory Data Analysis
- ➔ Statistical Analysis - correlations
- ➔ Linear regression
- ➔ Logistics regression
- ➔ Random Forest
- ➔ Decision Trees
- ➔ K-Nearest Neighbour

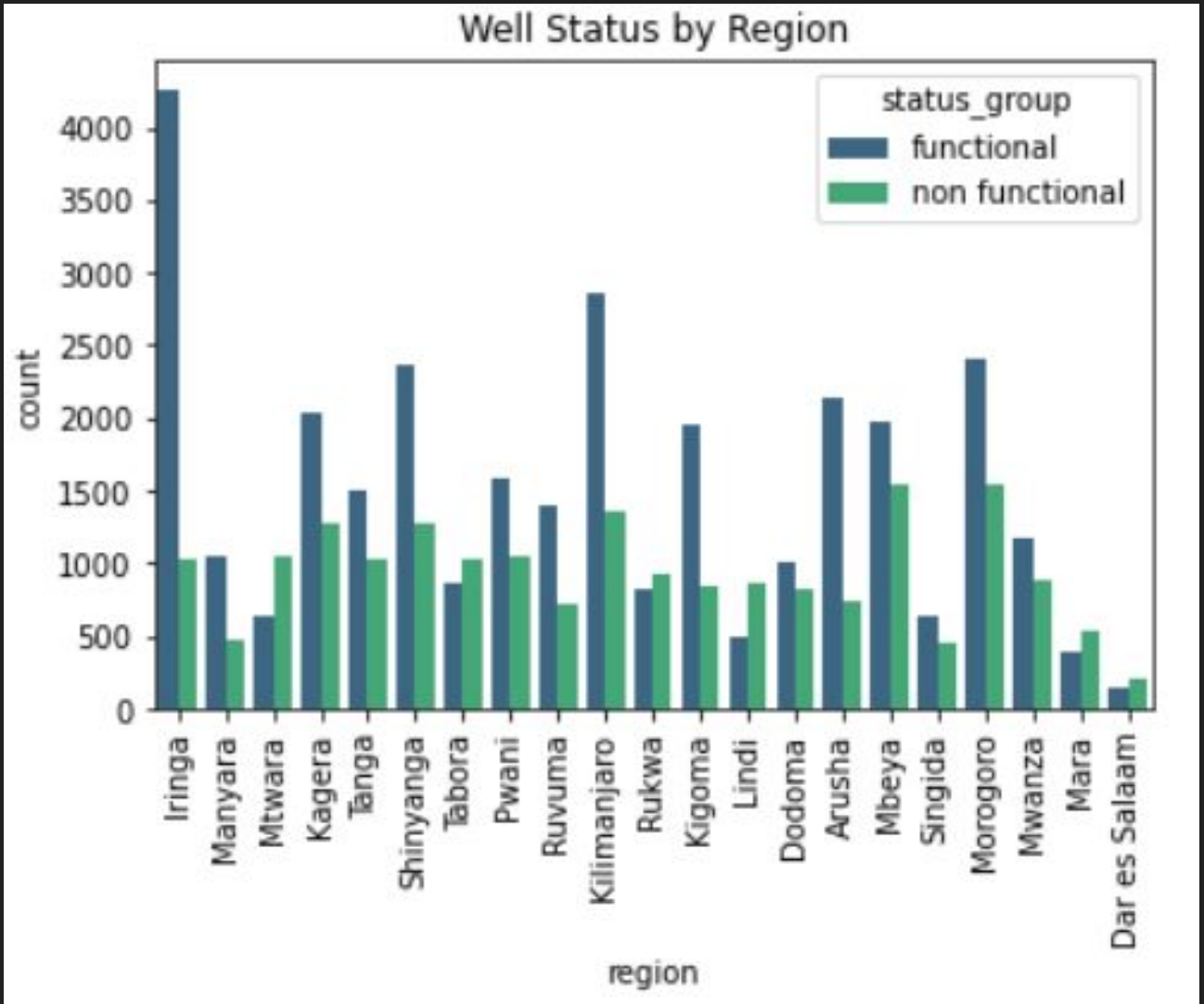
Data Understanding



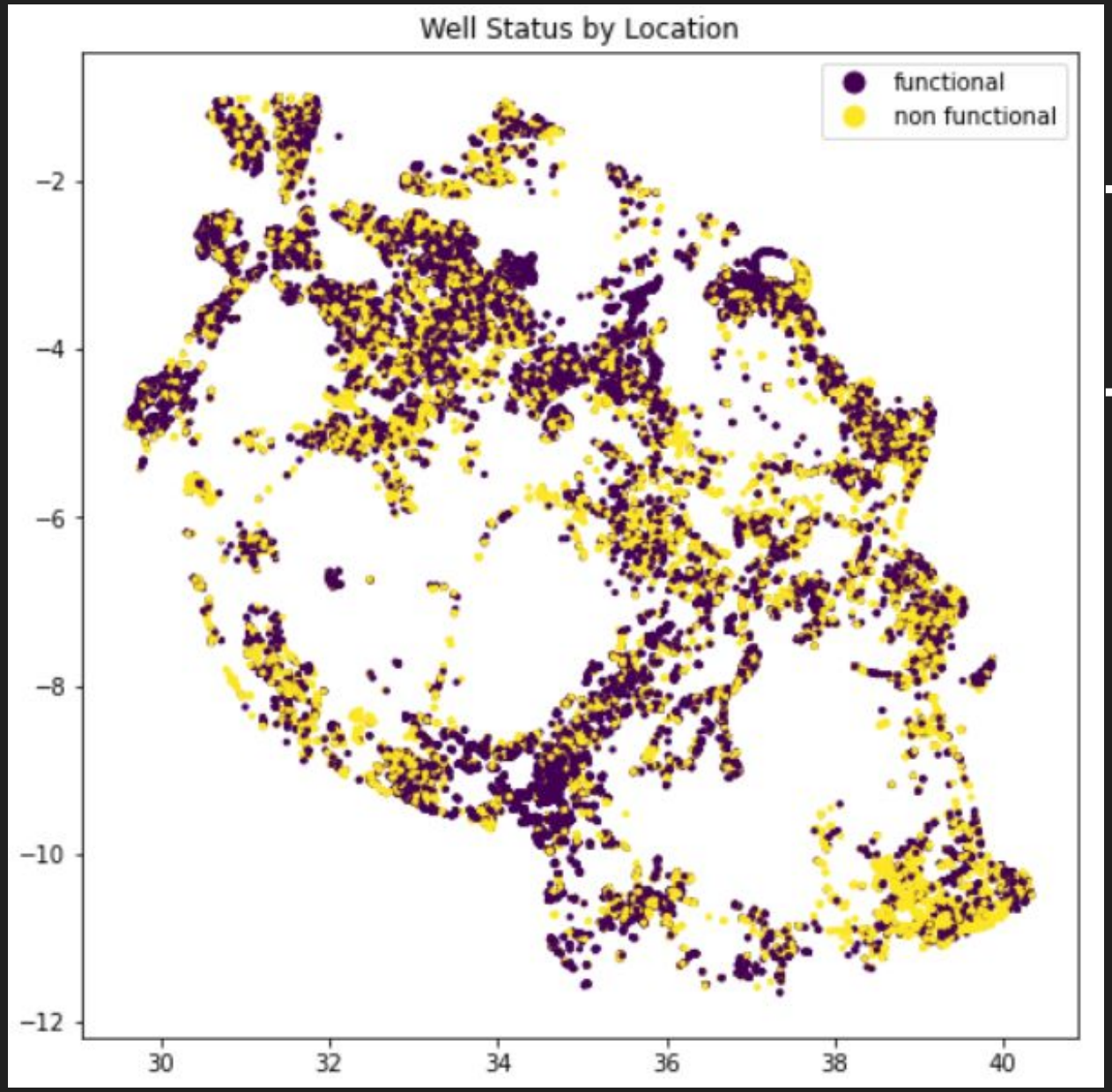
The majority of the wells are functional, with a distribution of approximately 3:2 between functional and non-functional wells.



This histogram illustrates the distribution of well construction years, revealing a significant rise in well development since the 1960s, with a peak in the late 2000s.



This plot shows the status of wells per region. We can see that Iringa has the most functional wells.

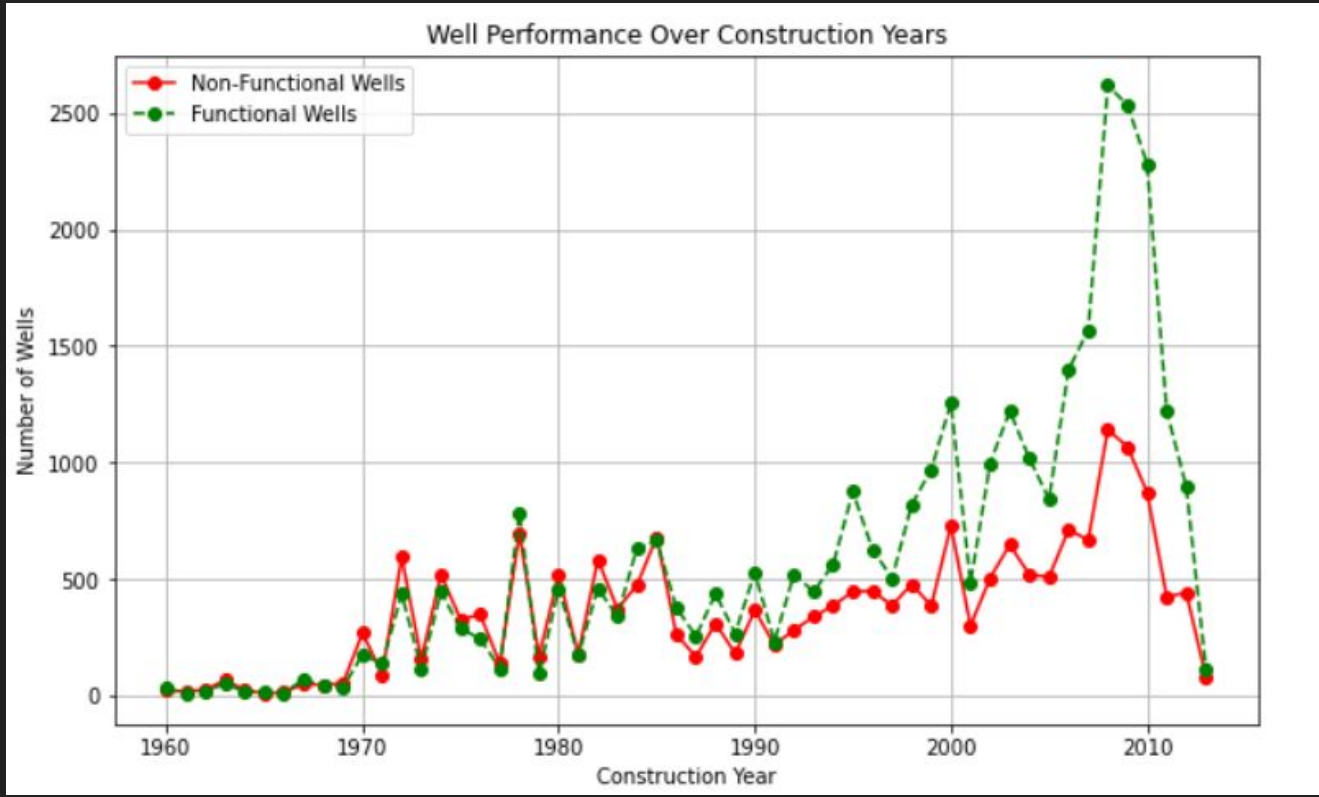


This plot shows well status by location. We can see:

- ➔ Non-functional wells are spread across the region but are more concentrated in the south eastern parts.
- ➔ Functional wells are more densely clustered in the central and western areas.

The higher concentration of non-functional wells in the south east of Tanzania could indicate challenges such as:

- ➔ poor management
- ➔ aged infrastructure
- ➔ unreliable sources of water feeding the wells

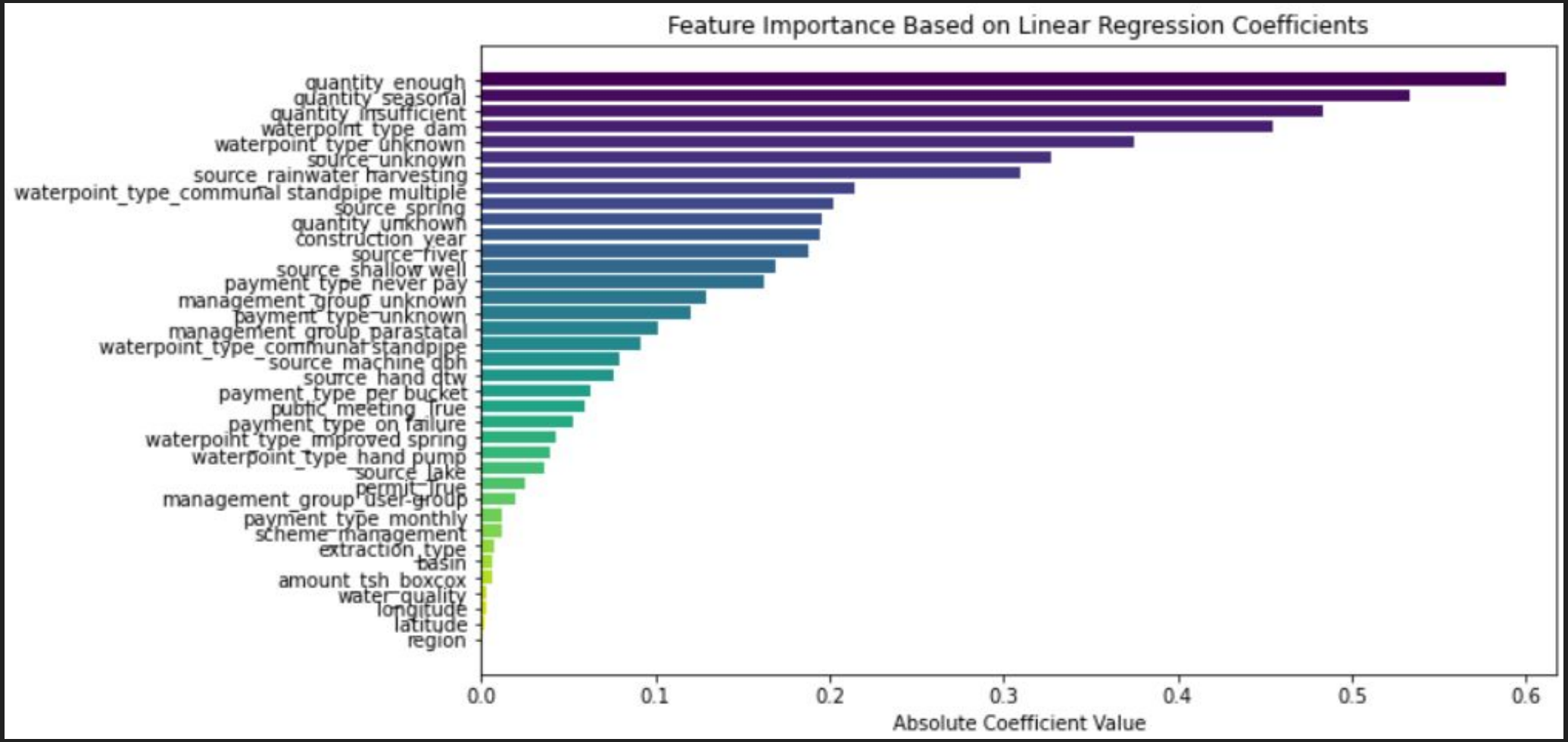


From this line plot, we can see:

➔ Older wells tend to have a higher proportion of non-functional status

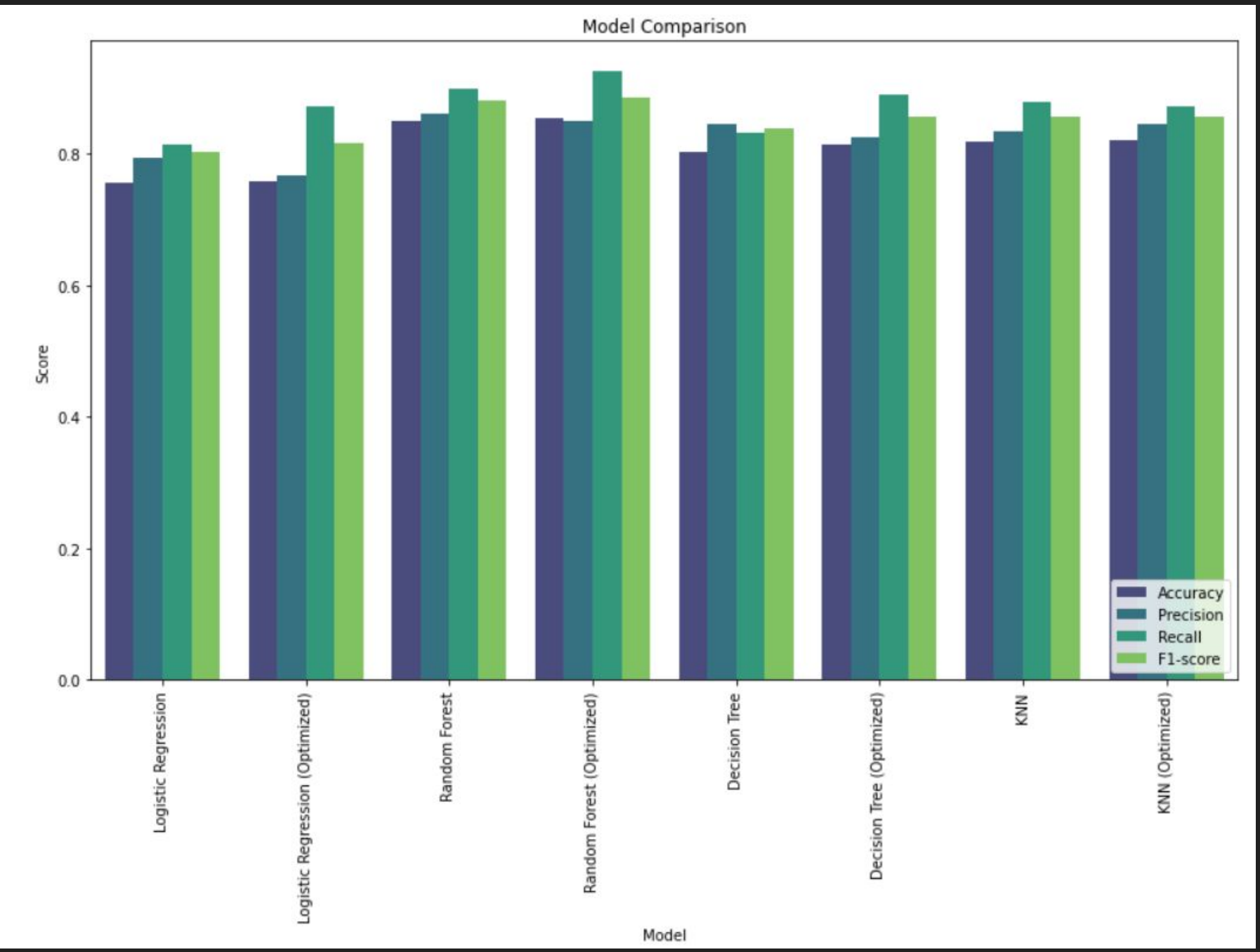
➔ More recent wells are more likely to be functional

However, the presence of non-functional wells in all time periods highlights that factors beyond age also play a role in well performance.



From the above, we can see that 'quantity', 'waterpoint_type', 'source', 'construction_year', 'payment_type' and 'management_group' are the leading indicators of the functionality of a well.

Models Evaluation



Random Forest is the best choice to maximize predictive performance and correctly classify as many wells as possible.

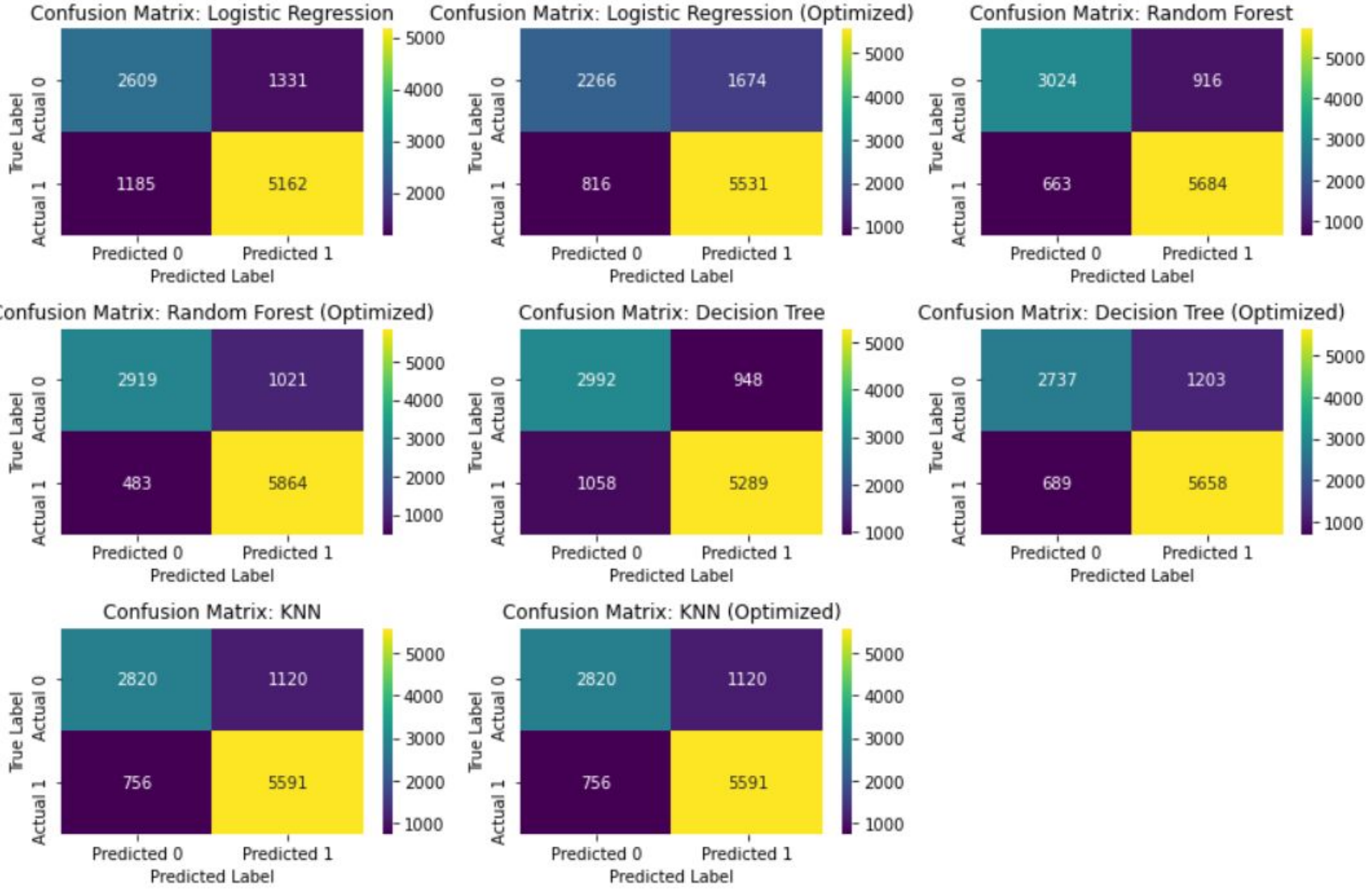
From the above plot, we can deduce that:

- ➔ Optimized models generally outperform their baseline counterparts. Tuning helped refine decision boundaries and improve model performance.
- ➔ Random Forest (Optimized) performs best overall. It has the highest Recall, which indicates that it effectively identifies wells that are functional or non-functional.
- ➔ Decision Tree and KNN perform similarly, but not as well as Random Forest. Their performance suggests that while they are effective classifiers, they may misclassify some wells.
- ➔ Logistic Regression has the lowest performance. This indicates that this model struggles with capturing complex patterns in the dataset, likely due to its linear nature.

Now let us take a look at the confusion matrices for the models:

It is important to note that:

- ➔ Misclassifying a non-functional well as functional could result in communities relying on a faulty water source.
- ➔ If a functional well is misclassified as non-functional, it may lead to unnecessary repairs or neglect.



Random Forest (Optimized) is the most reliable model, making it the best choice for minimizing misclassifications.

From the above plots, we can see that:

- ➔ Random Forest Optimised has the lowest false negatives (483) and false positives (1021) compared to other models. It correctly classifies most functional wells (5864 True Positives) and non-functional wells (2919 True Negatives).
- ➔ This suggests that it is the most reliable model for predicting well functionality.
- ➔ Decision Tree and KNN models perform well also but have higher false negatives.
- ➔ Logistic Regression struggles with classification.

Conclusions

Modeling Evaluation Summary

The Optimised Random Forest model stands out as the best performing model with:

- ➔ Accuracy-85.38%
- ➔ Recall-92.39%
- ➔ Precision-85.17%
- ➔ F1-score-88.63%.

The best parameters after tuning are:

- ➔ max_depth: 20
- ➔ min_samples_leaf: 1
- ➔ min_samples_split: 5
- ➔ n_estimators: 300

Business Answers

- The following features stand out as the leading predictors to the functionality of a well:
  - 'quantity',
  - 'waterpoint_type',
  - 'source',
  - 'construction_year',
  - 'payment_type'
  - 'management_group'
- There is a high cluster of non-functional wells in the south east of Tanzania. The government should consider allocating resources towards repair and maintenance in this area, to ensure the residents get good water supply.
- More recent wells are more likely to be functional than aged wells. The presence of non-functional wells in all time periods highlights that factors beyond age also play a role in well performance.

Recommendations

The above analysis could be improved in the following ways:

- ➔ Once the most influential factors affecting well functionality are identified, analyze their impact—do they contribute positively or negatively? Understanding these relationships will help the government of Tanzania make informed investment decisions to enhance well longevity.
- ➔ Instead of using construction year as a standalone feature, create a new variable, "well age", to better capture the patterns related to functionality. This could improve the model's ability to detect trends over time.
- ➔ The model could be further analysed as a ternary classification problem to include the 'functional but needs repair' as a category on its own. This will help the client be able to plan for maintenance early enough before the well condition deteriorates to the 'non-functional' category.

The end!

Any Questions?