TANZANIA
WATER WELLS

TABLE OF CONTENTS

- OVERVIEW
- BUSINESS UNDERSTANDING
- DATA UNDERSTANDING
- MODELLING
- RECOMMENDATIONS

OVERVIEW

IN TANZANIA

16 MILLION PEOPLE DO NOT HAVE ACCESS TO SAFE DRINKING WATER

40 MILLION PEOPLE LACK ACCESS TO IMPROVED SOURCES OF DRINKING WATER

60% HOUSEHOLDS IN TANZANIA CURRENTLY HAVE ACCESS TO A BASIC WATER-SUPPLY

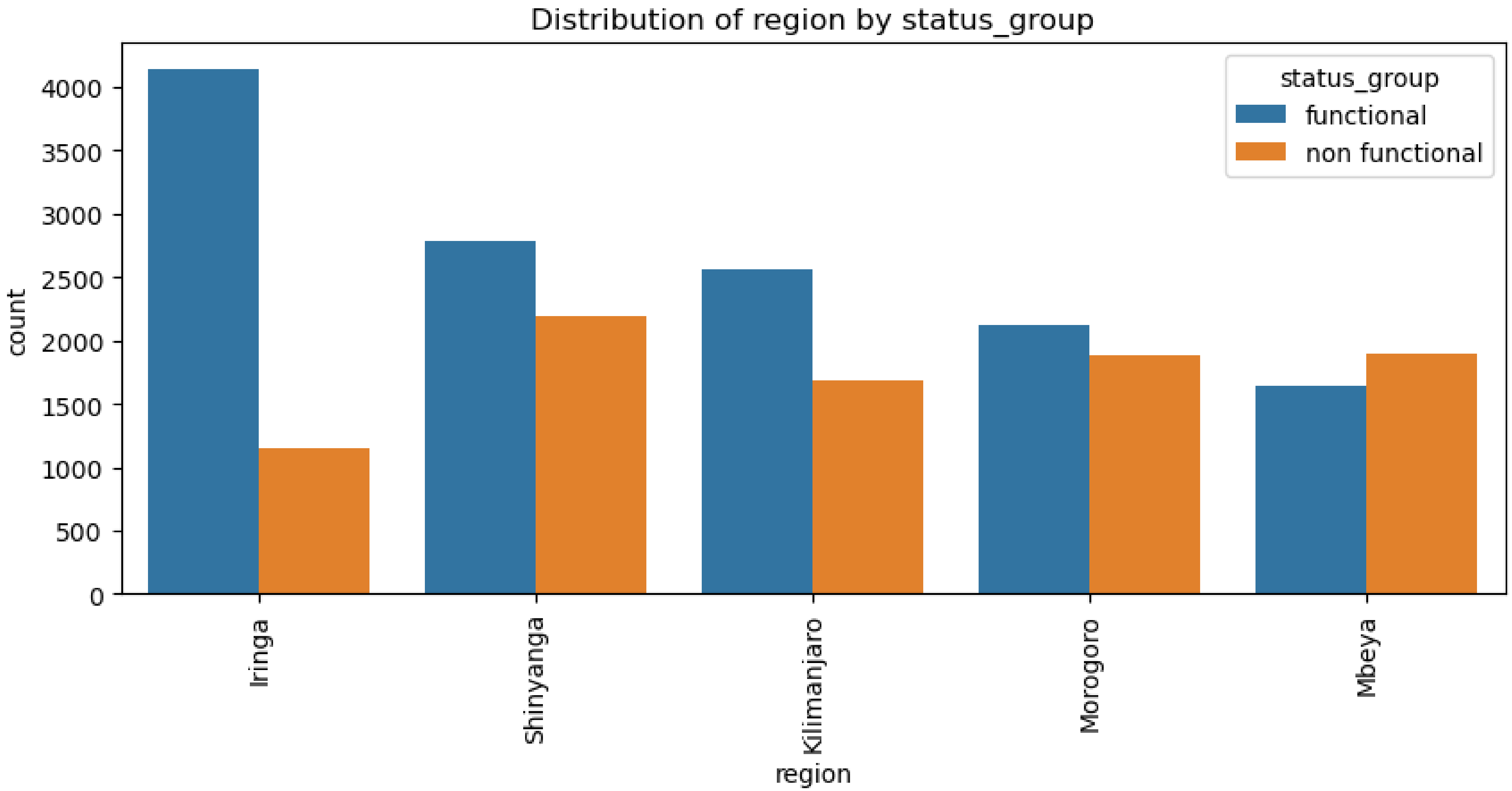
BUSINESS UNDERSTANDING

Despite efforts to improve water access in Tanzania, significant challenges persist in ensuring sustainable and reliable access to clean water for all communities. The lack of accurate predictive models for water wells hampers efficient planning and resource allocation, resulting in suboptimal drilling locations, unreliable well yields, and inadequate maintenance strategies. As a result, communities continue to face water scarcity, health risks from contaminated water sources, and economic hardships.

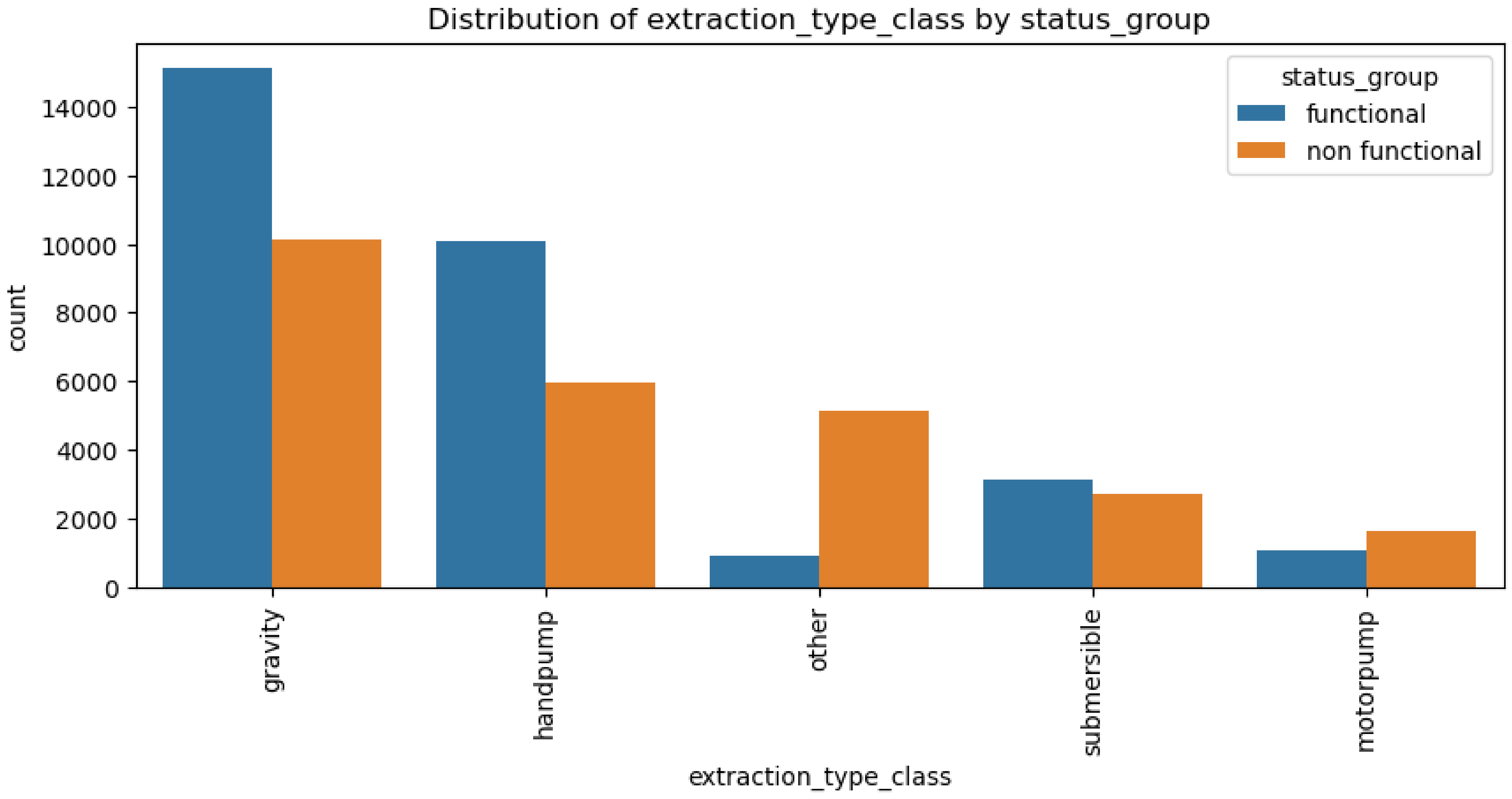
DATA UNDERSTANDING

In this project we shall use a dataset containing information about existing water wells in Tanzania sourced from an ongoing DrivenData competition.

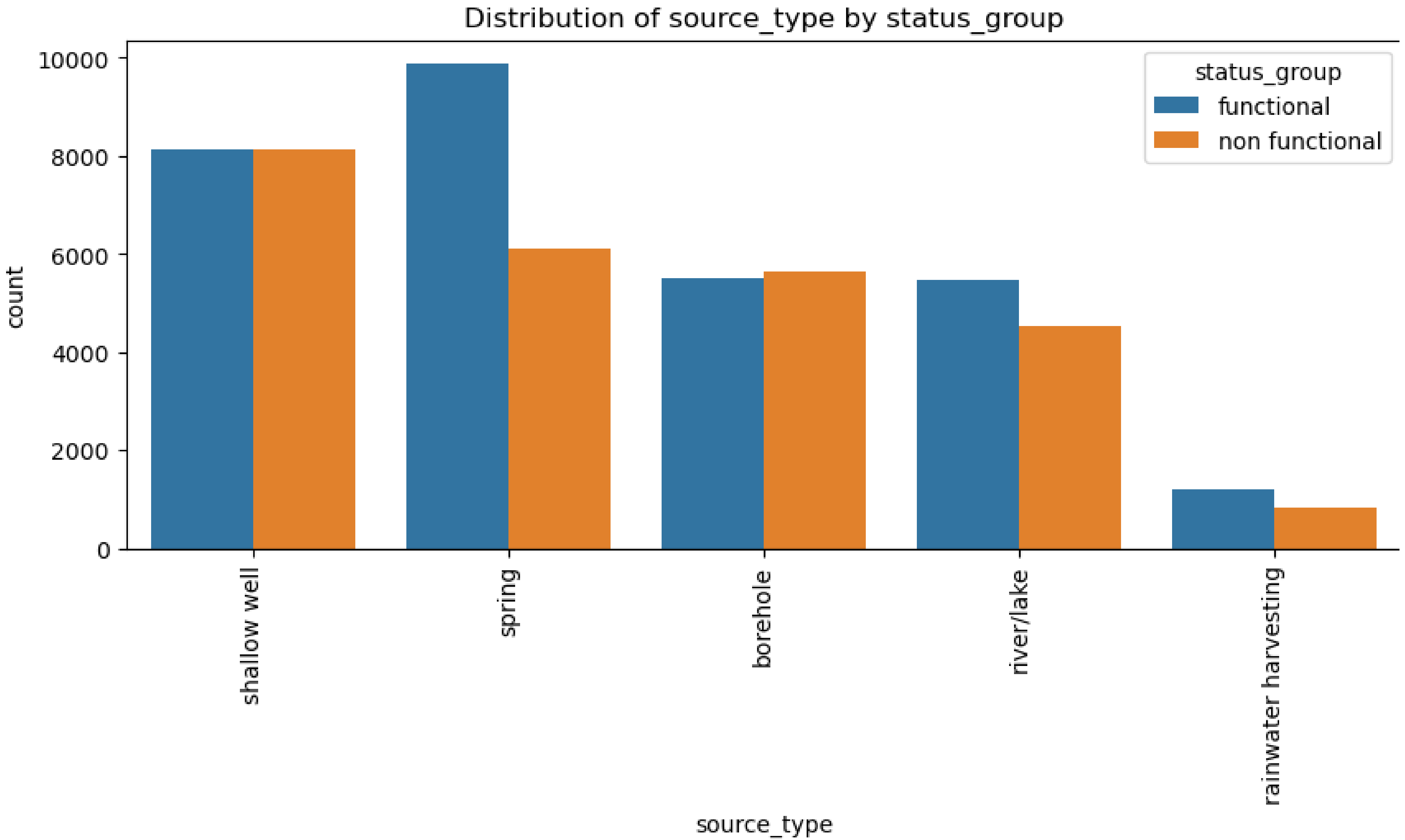
The dataset contains 59,400 records and 40 columns. Of these columns, we identified 31 to be categorical, and 9 as numerical. We were able to further group the columns into the general features being captured.



From the distribution above, Iringa region has the most functional wells compared to the rest of the regions followed by Shinyanga, Kilimanjaro, Morogoro and Mbeya.



Njombe region has the most functional wells followed by Moshi Rural, Arusha Rural, Bariadi and Kilosa.



Springs are the most common source of functional wells.

MODELLING

MODELS PERFORMED;

- DUMMY CLASSIFIER
- DECISION TREES
- SVM
- K- NEAREST- NEIGHBORS
- DECISION TREE WITH GRID SEARCH
- K- NEAREST NEIGHBORS GRID SEARCH
- LOGISTIC REGRESSION

Logistic Regression

The initial Logistic Regression model performed poorly, likely due to limitations in handling the complexity of the data. Logistic Regression, while a straightforward and interpretable model, often struggles with non-linear relationships and imbalanced data, which can result in low accuracy, especially when predicting more complex outcomes.

Given these challenges, it was essential to revisit and refine the model to better capture the underlying patterns in the data. This involved enhancing preprocessing techniques, engineering more informative features, tuning model parameters, and exploring ensemble methods. These steps aimed to improve the model's ability to generalize and make accurate predictions, ultimately leading to better performance in the specific task of classifying water pump statuses in Tanzania.

CONCLUSION

In conclusion, the Decision Tree with GridSearchCV and K-Nearest Neighbors (KNN) classifiers performed best, achieving accuracies of 0.78 and 0.77, respectively. The Support Vector Machine (SVM) classifier followed with 0.73 accuracy. Logistic Regression initially had 0.59 accuracy but significantly improved to 0.77 after several iterations. The objective of effectively classifying water pumps was met, with Decision Tree with GridSearchCV and KNN emerging as the most reliable models. Logistic Regression also showed notable improvement, making it a competitive option.

RECOMMENDATION

The In conclusion, the findings suggest several important considerations for stakeholders when planning the construction and maintenance of wells in Tanzania. The Dodoma region has a higher number of non-functional wells compared to functional ones, indicating a need for careful assessment and investigation into the causes and potential solutions before building new wells. Data analysis shows that wells with operating permits tend to remain functional longer than those without permits, highlighting the importance of regulatory compliance. Furthermore, wells that are not associated with any payment plan often become inoperable due to public misuse; therefore, implementing a reasonable payment plan could help maintain well functionality. To ensure sustainable water access, stakeholders should focus on regions with higher failure rates, verify that all wells have the necessary permits, and consider the establishment of payment plans to enhance the viability of the wells.