Tanzania Water Wells Conditions Project

Project Overview

- Tanzania, a developing country with a population of over 57 million, grapples with the substantial challenge of ensuring clean and accessible water for its people. While the country has set up numerous water points, a significant portion of these wells requires repair, and some are no longer operational. Access to clean and dependable water sources is essential for public health and economic progress.

Business Problem

The problem involves classifying well conditions into three categories:

1. Functional: Wells in good working condition, providing clean and reliable water.

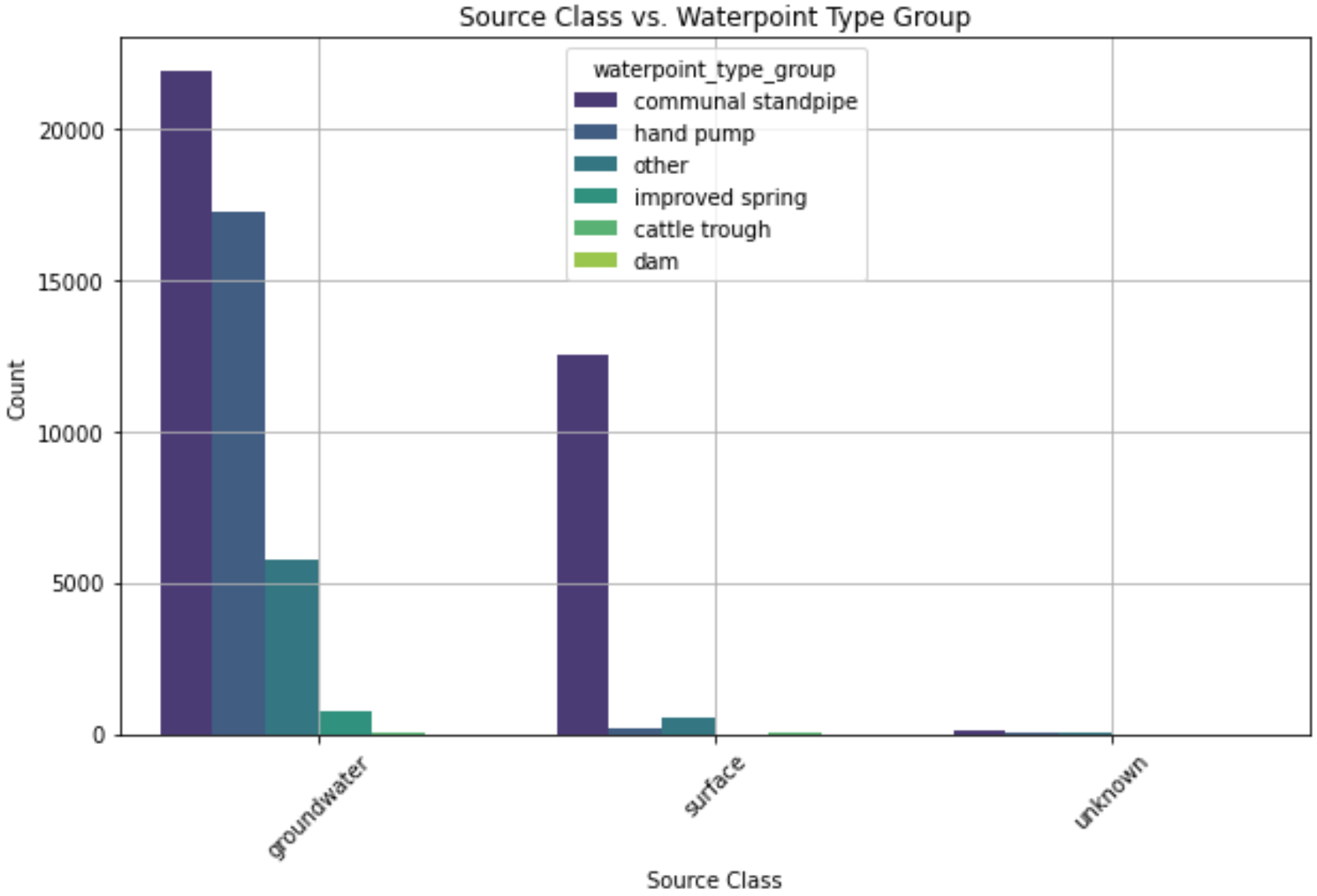
2. Functional Needs Repair: Operational wells needing maintenance for sustained functionality.

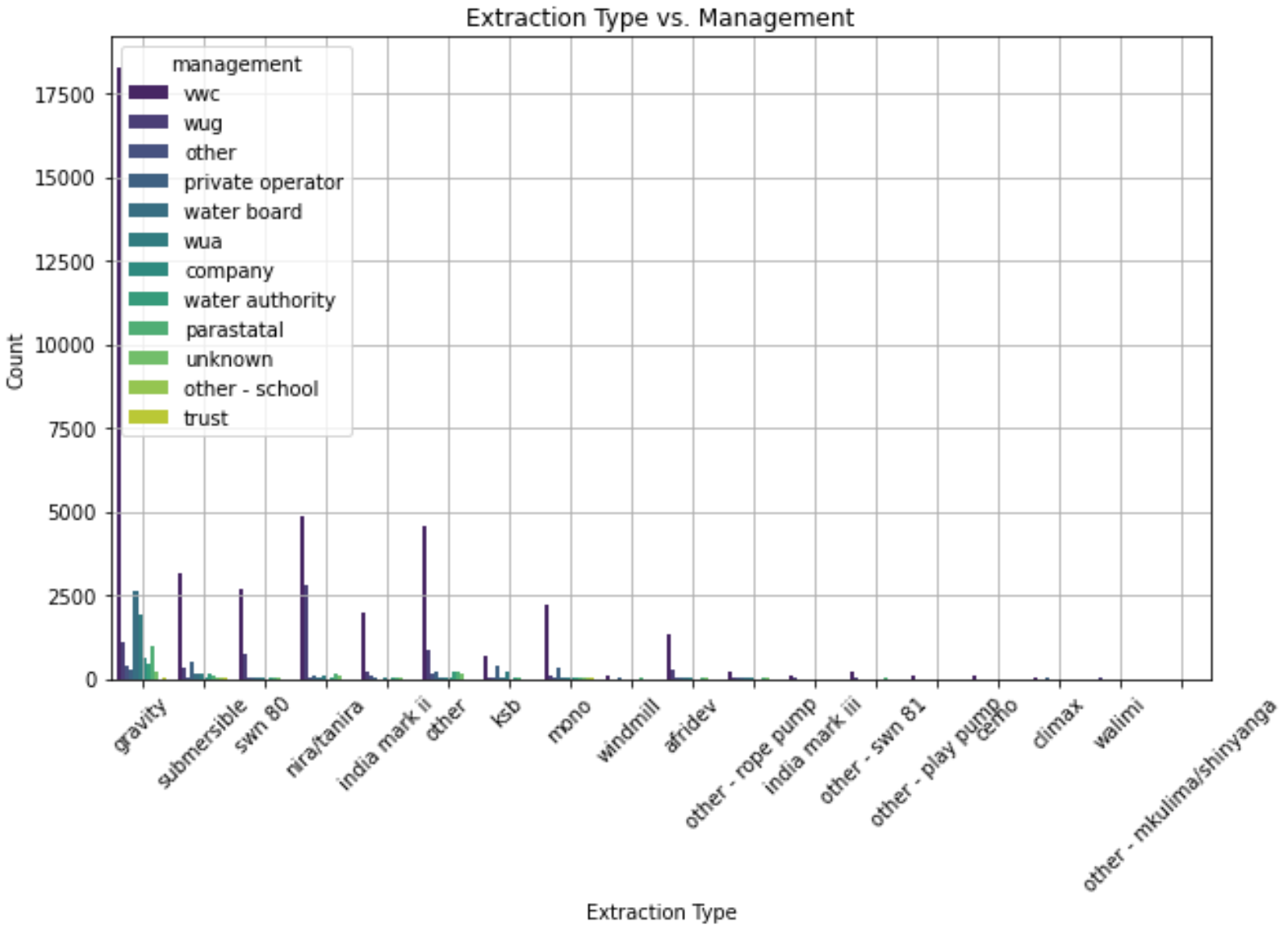
3. Non-functional:Wells that have stopped working and can't provide water.

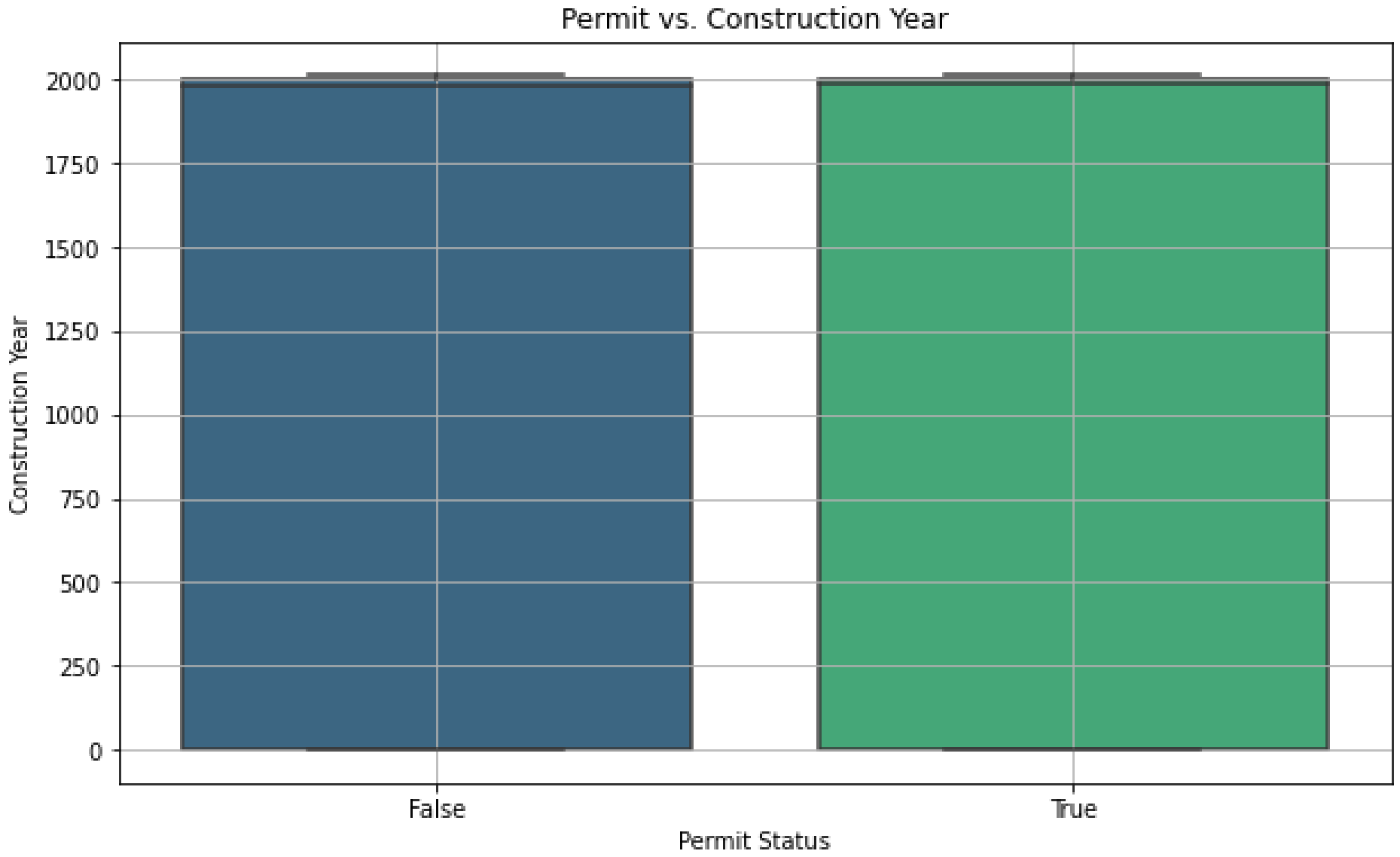
Data Understanding

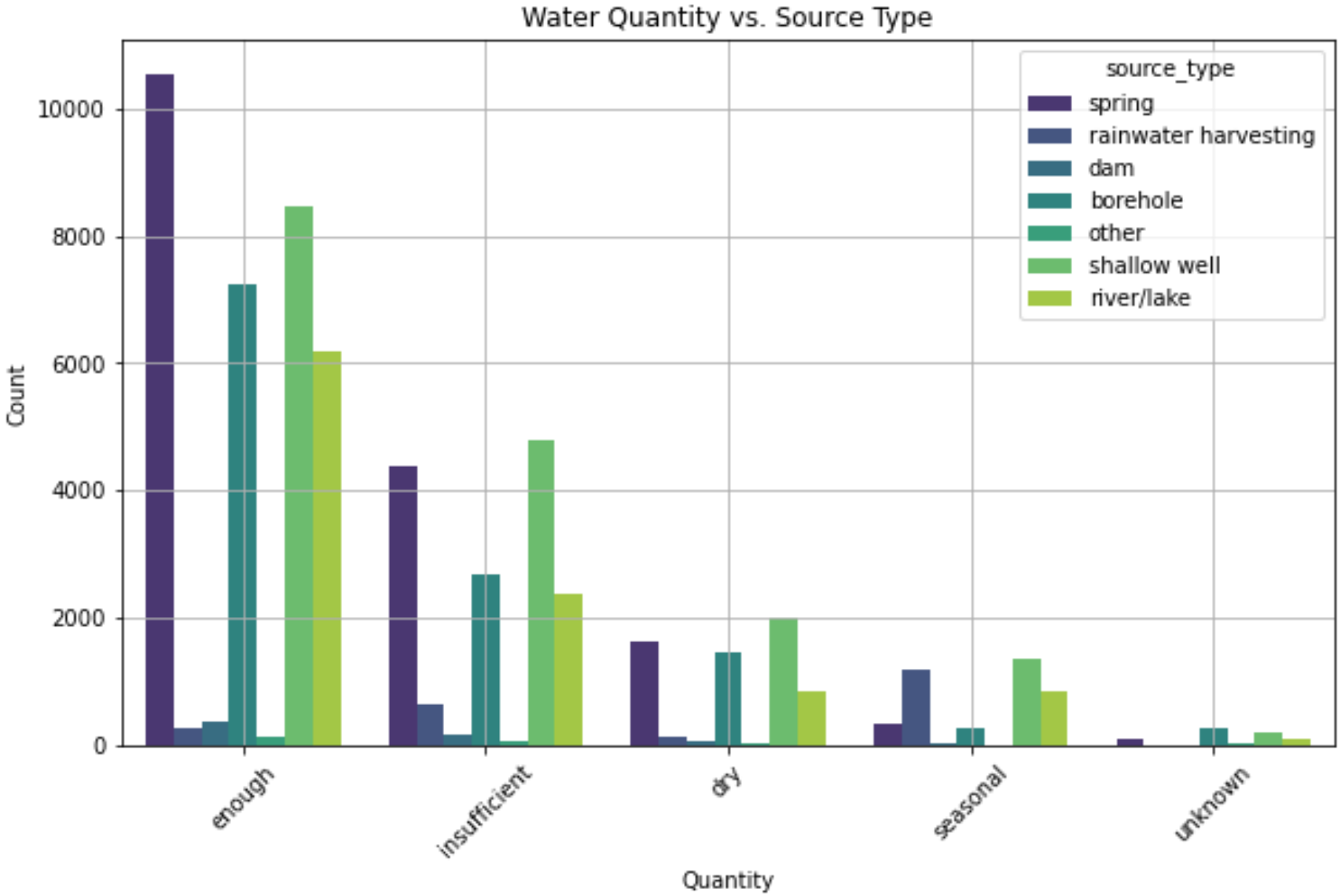
- The analysis involved a dataset from multiple sources, including Taarifa, the Tanzanian Water Ministry, and Driven Data. This dataset contains information on about 60,000 water pumps in Tanzania, with 39 attributes covering geographical, management, and technical details.
- The dataset was split into three parts:

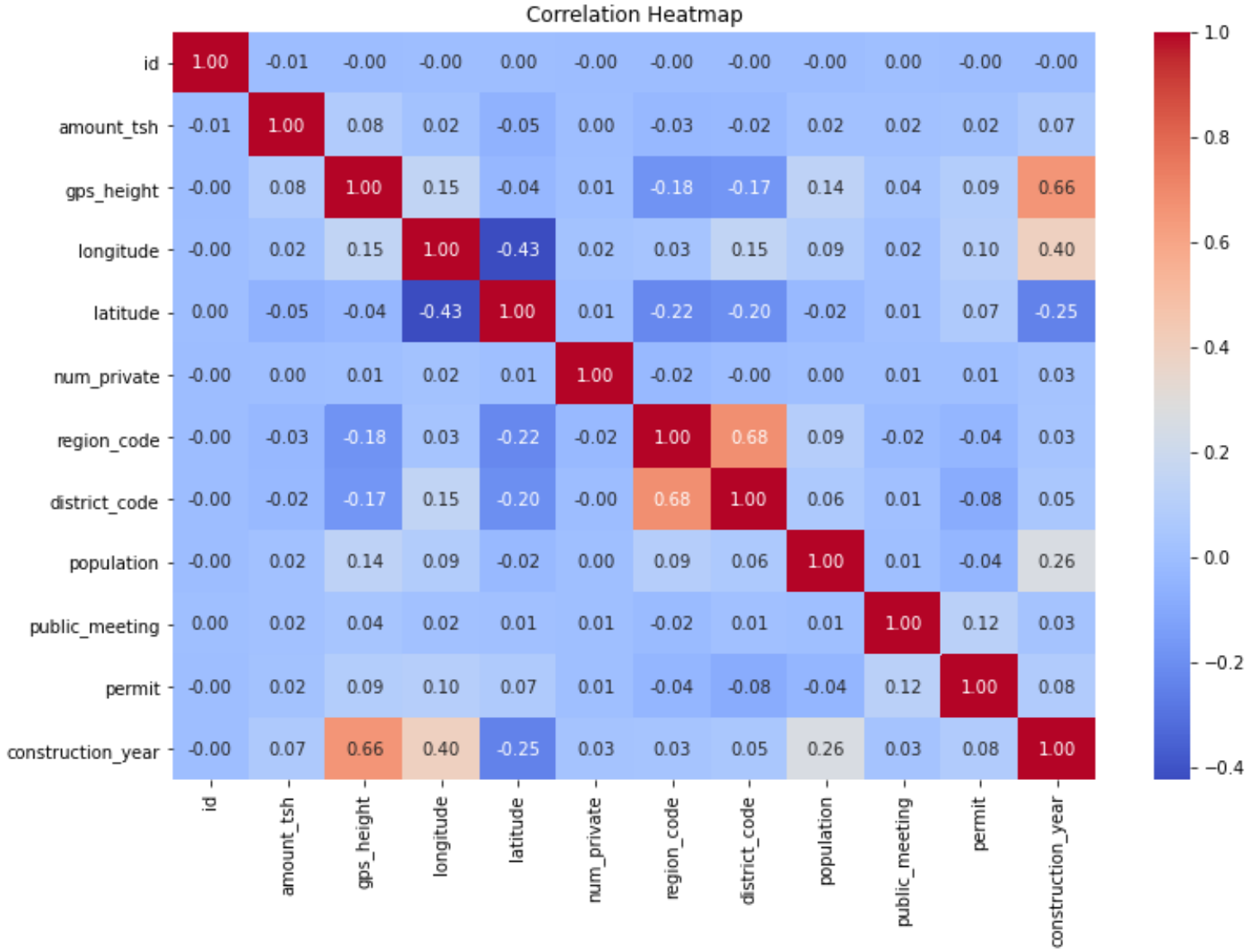
- 1. Test Set Values: These are the independent variables for making predictions.
- 2. Training Set Labels: This contains the target variable "status_group" for training.
- 3. Training Set Values: These are the features used to train the machine learning model.


Source Class vs. Waterpoint Type Group
waterpoint_type_group
communal standpipe
hand pump
other
improved spring
cattle trough
dam
20000
15000
10000
5000
0
Count
groundwater
surface
unknown
Source Class


Extraction Type vs. Management
management
vwc
wug
other
private operator
water board
wua
company
water authority
parastatal
unknown
other - school
trust
Count
17500
15000
12500
10000
7500
5000
2500
0
gravity
submersible
swn 80
nira/tanira
india mark ii
other
ksb
mono
windmill
afridev
other - rope pump
india mark iii
other - swn 81
other - play pump
cemo
climax
walimi
other - mkulima/shinyanga
Extraction Type


Permit vs. Construction Year
Construction Year
2000
1750
1500
1250
1000
750
500
250
0
False
True
Permit Status


Water Quantity vs. Source Type
source_type
spring
rainwater harvesting
dam
borehole
other
shallow well
river/lake
Count
10000
8000
6000
4000
2000
0
enough
insufficient
dry
seasonal
unknown
Quantity


Correlation Heatmap
id 1.00 -0.01 -0.00 -0.00 0.00 -0.00 -0.00 -0.00 -0.00 0.00 -0.00 -0.00
amount_tsh -0.01 1.00 0.08 0.02 -0.05 0.00 -0.03 -0.02 0.02 0.02 0.02 0.07
gps_height -0.00 0.08 1.00 0.15 -0.04 0.01 -0.18 -0.17 0.14 0.04 0.09 0.66
longitude -0.00 0.02 0.15 1.00 -0.43 0.02 0.03 0.15 0.09 0.02 0.10 0.40
latitude 0.00 -0.05 -0.04 -0.43 1.00 0.01 -0.22 -0.20 -0.02 0.01 0.07 -0.25
num_private -0.00 0.00 0.01 0.02 0.01 1.00 -0.02 -0.00 0.00 0.01 0.01 0.03
region_code -0.00 -0.03 -0.18 0.03 -0.22 -0.02 1.00 0.68 0.09 -0.02 -0.04 0.03
district_code -0.00 -0.02 -0.17 0.15 -0.20 -0.00 0.68 1.00 0.06 0.01 -0.08 0.05
population -0.00 0.02 0.14 0.09 -0.02 0.00 0.09 0.06 1.00 0.01 -0.04 0.26
public_meeting 0.00 0.02 0.04 0.02 0.01 0.01 -0.02 0.01 0.01 1.00 0.12 0.03
permit -0.00 0.02 0.09 0.10 0.07 0.01 -0.04 -0.08 -0.04 0.12 1.00 0.08
construction_year -0.00 0.07 0.66 0.40 -0.25 0.03 0.03 0.05 0.26 0.03 0.08 1.00
id
amount_tsh
gps_height
longitude
latitude
num_private
region_code
district_code
population
public_meeting
permit
construction_year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Modeling

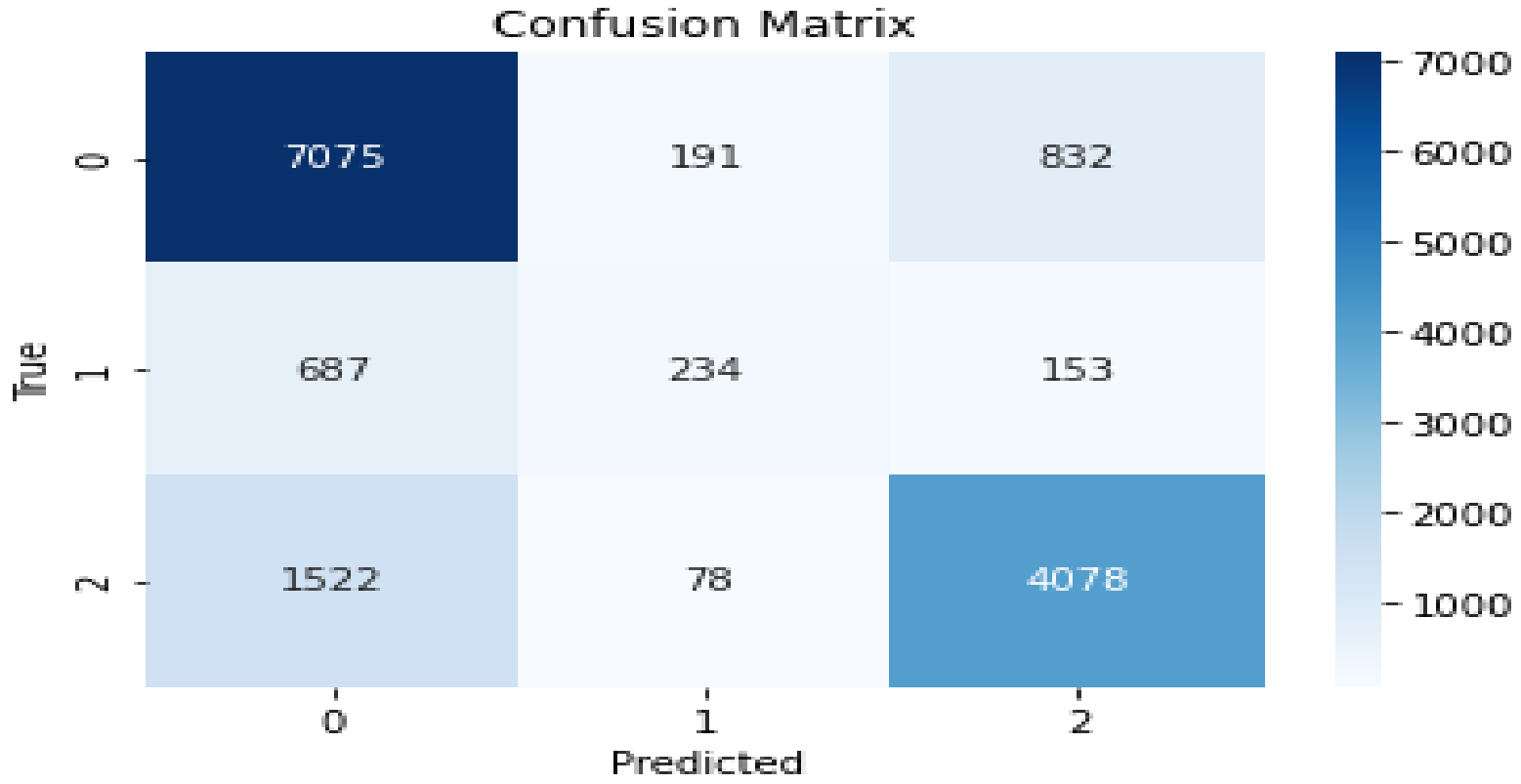
AFTER TESTING VARIOUS MODELS, WE SELECTED THE ONE-HOT CODING CLASSIFIER BASED ON ITS ACCURACY SCORE. HERE'S A PERFORMANCE BREAKDOWN:

PRECISION: MEASURES ACCURATE PREDICTIONS, WITH HIGHER PRECISION INDICATING FEWER FALSE POSITIVES. THE MODEL DEMONSTRATES GOOD PRECISION FOR ALL CLASSES.

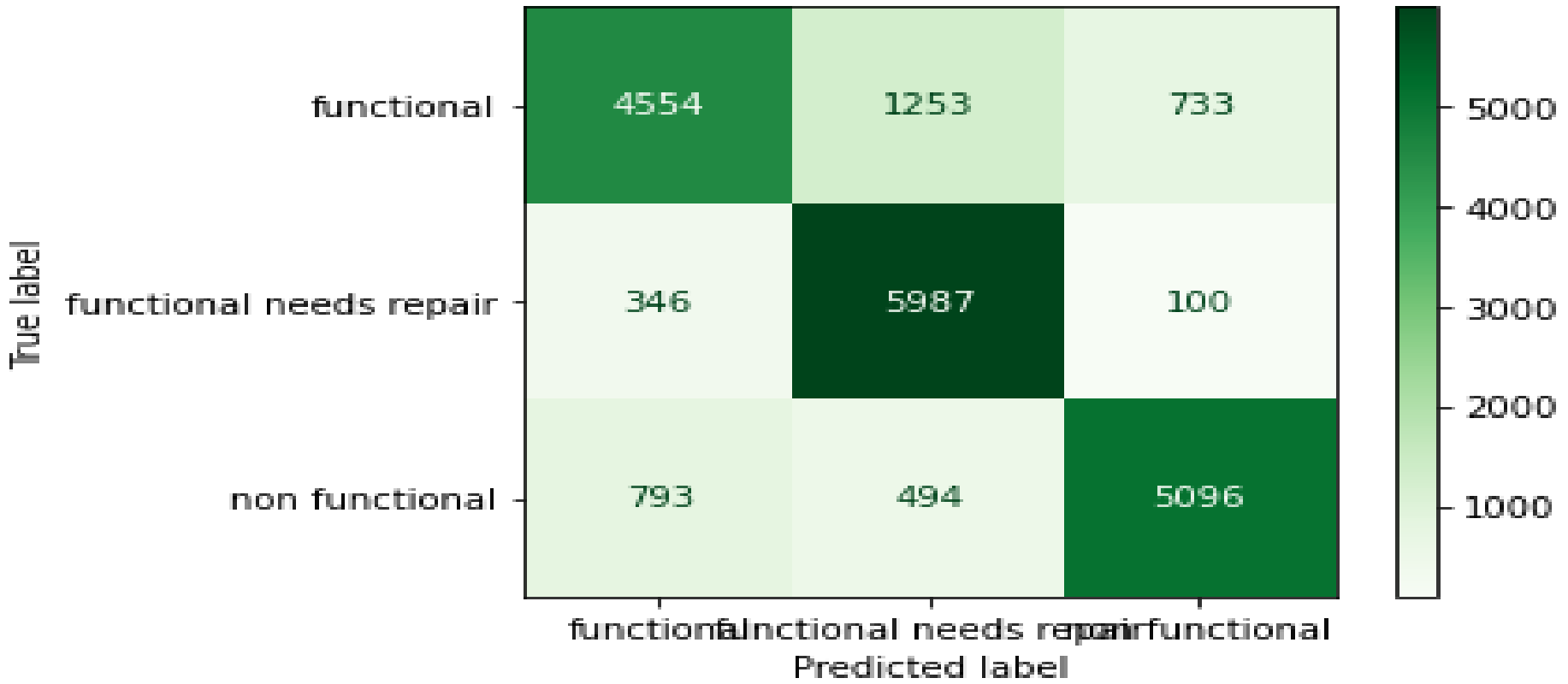
RECALL: MEASURES ACCURATE PREDICTIONS, WITH HIGHER RECALL INDICATING FEWER FALSE NEGATIVES. THE MODEL EXHIBITS REASONABLY GOOD RECALL FOR ALL CLASSES.

- F1-SCORE: BALANCING PRECISION AND RECALL, THE F1-SCORES ARE RELATIVELY HIGH FOR ALL CLASSES, INDICATING STRONG OVERALL PERFORMANCE.

IN SUMMARY, THE MODEL ACHIEVED AN 80.74% ACCURACY, CORRECTLY PREDICTING THE CLASS FOR ABOUT 80.79% OF INSTANCES IN THE VALIDATION SET. THE MACRO-AVERAGED F1-SCORE IS ALSO 80.79%, REFLECTING CONSISTENT PERFORMANCE ACROSS ALL CLASSES. THE WEIGHTED AVERAGE F1-SCORE IS 80.74%, CONSIDERING THE DATASET'S CLASS IMBALANCE.

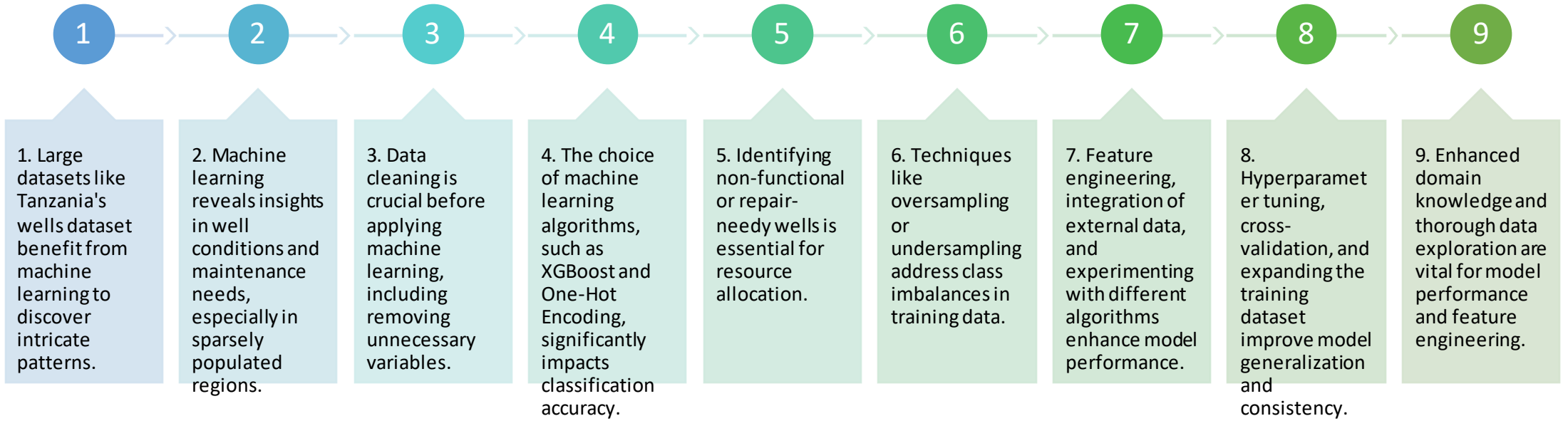


First Confusion Matrix



Second Confusion Matrix

Evaluation


1
2
3
4
5
6
7
8
9
1. Large datasets like Tanzania's wells dataset benefit from machine learning to discover intricate patterns.
2. Machine learning reveals insights in well conditions and maintenance needs, especially in sparsely populated regions.
3. Data cleaning is crucial before applying machine learning, including removing unnecessary variables.
4. The choice of machine learning algorithms, such as XGBoost and One-Hot Encoding, significantly impacts classification accuracy.
5. Identifying non-functional or repair-needy wells is essential for resource allocation.
6. Techniques like oversampling or undersampling address class imbalances in training data.
7. Feature engineering, integration of external data, and experimenting with different algorithms enhance model performance.
8. Hyperparameter tuning, cross-validation, and expanding the training dataset improve model generalization and consistency.
9. Enhanced domain knowledge and thorough data exploration are vital for model performance and feature engineering.

Recommendations

Improving Tanzania Wells Model

- Class Imbalance Handling: Use oversampling or under sampling to address class imbalance and predict "functional needs repair" and "non-functional" wells more accurately.
- Feature Engineering: Create new features and explore external data sources to enhance predictive power.
- Algorithm Selection: Experiment with algorithms like random forests, gradient boosting, or support vector machines for better results.
- Hyperparameter Tuning: Optimize hyperparameters systematically for improved generalization and prediction accuracy.
- Cross-Validation: Implement cross-validation for robust model evaluation and reduced overfitting.
- Data Collection: Collect more data to expand the training set and improve generalization.
- Domain Knowledge & Exploration: Gain deep domain knowledge and explore data thoroughly to identify quality issues, missing values, and outliers.

Result: Improved model accuracy, particularly in identifying non-functional or repair-needy wells, aiding resource allocation for maintenance and repairs.

Next Steps

- Class Imbalance Handling: Implement oversampling or under sampling to address class imbalance and enhance predictions for "functional needs repair" and "non-functional" wells.
- Feature Engineering: Invest in creating new features and exploring external data sources to boost predictive power.
- Algorithm Selection: Experiment with algorithms like random forests, gradient boosting, or support vector machines for improved model performance.
- Hyperparameter Tuning: Optimize hyperparameters systematically for better generalization and prediction accuracy.
- Cross-Validation: Implement cross-validation for robust model evaluation and to prevent overfitting.
- Data Collection: If possible, collect additional data to expand the training set and improve model generalization.
- Domain Knowledge & Exploration: Deepen domain understanding and conduct thorough data exploration to identify quality issues and feature-target relationships.

Implementing these steps will result in an enhanced model, particularly in identifying non-functional or repair-needy water pumps, aiding resource allocation for maintenance and repairs.

Thank You