Pump it Up: Data Mining the Water Table

Predicting the operating condition of a waterpoints across Tanzania

PROJECT OVERVIEW

Objective: to predict the operating condition of a waterpoint for each record

Importance: identify potential issues with existing water well projects, in order to promote access to clean, potable water across Tanzania.

BUSINESS AND DATA UNDERSTANDING

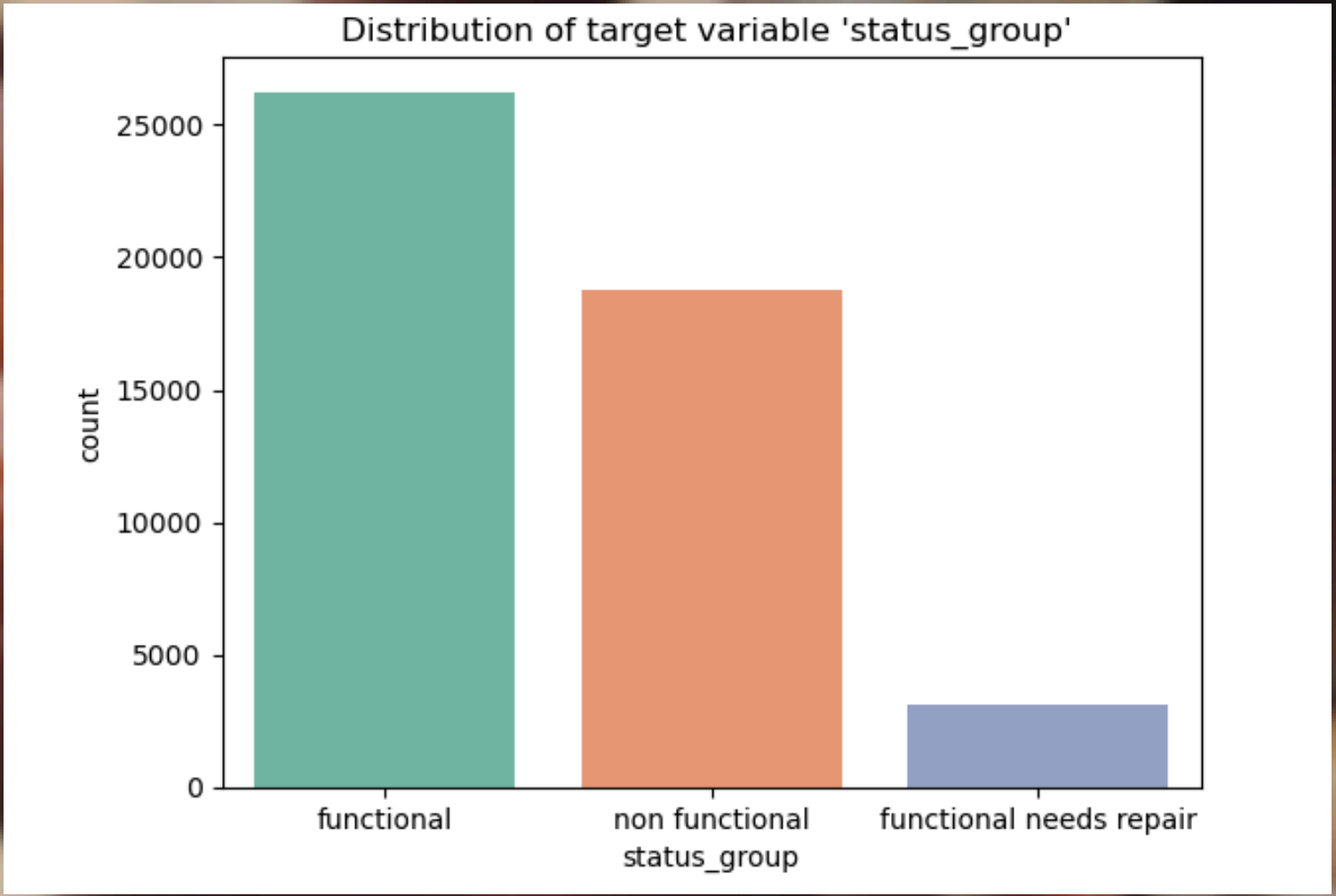
Business Context:

- Enhancement of access to clean water across Tanzania

Data Description:

- Provided by Taarifa Tanzania, downloaded from Driven Data
- Target Variable is status_group, which is a binary class indicating whether a well is functional, non-functional or needing repair

KEY FINDINGS

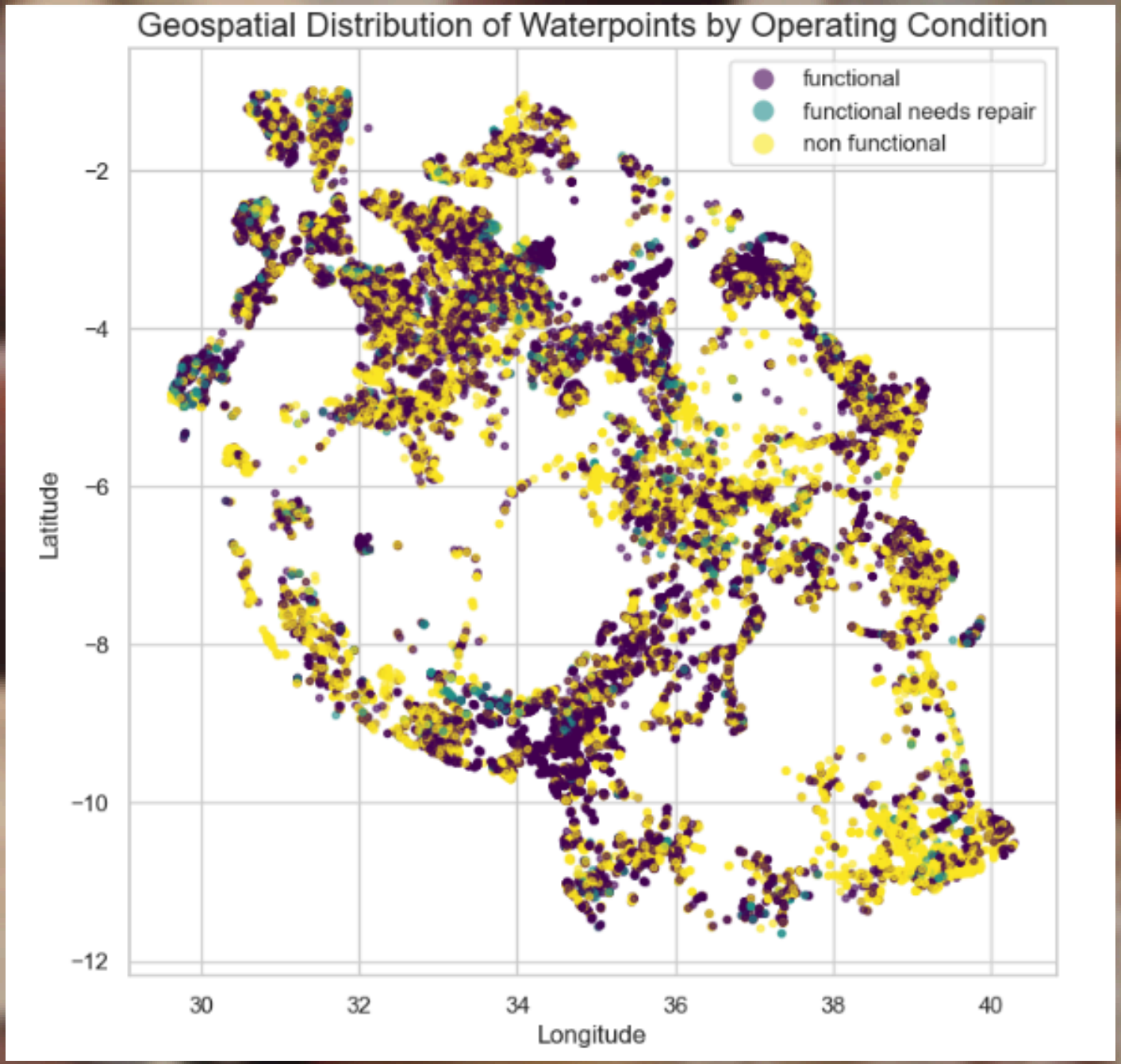


- 54.6% of the water points provide access to clean water
- 39.0% of the water points are not operational
- 6.4% of the water points need repair

KEY FINDINGS

Largest group of functional waterpoints uses gravity-fed systems (13709), but there is also a significant number of non-functional systems (6934).

KEY FINDINGS



Clusters of yellow and teal points indicate regions where water infrastructure might be failing or in need of urgent attention.

MODELING APPROACH

- Model Selection:
  - Logistic Regression
  - Decision Trees
  - Random Forest

MODEL PERFORMANCE AND EVALUATION

- Model Selection:
  - Logistic Regression
    - Train Accuracy - 64.54%
    - Test Accuracy - 64.02%
  - Decision Trees
    - Train Accuracy - 80.05%
    - Test Accuracy - 67.75%

MODEL PERFORMANCE AND EVALUATION

Random Forest

- Training Accuracy - 93.20%
- Testing Accuracy - 77.22%

- These results highlight the model's ability to learn from the training data effectively while still generalizing well to new, unseen data.

FEATURE IMPORTANCE

The geographical positioning, construction year and population around a water point greatly influence the operational status of a water point

CONCLUSION

The Random Forest Classifier was selected as the top-performing model for this classification task due to its ability to effectively balance complexity and generalization. The model was trained and evaluated on a dataset where the goal was to accurately predict the functional status of waterpoints.

RECOMMENDATIONS

Future Work:

- Refinement of the Model
- Alternative Evaluation Metrics
- Ensemble Methods
- Data Augmentation
- Periodic Retraining


THANK YOU