Waterpoint Functionality Classification

Predicting Functional Status of Waterpoints in Tanzania

Charity Mwangangi

23nd July 2025

Introduction

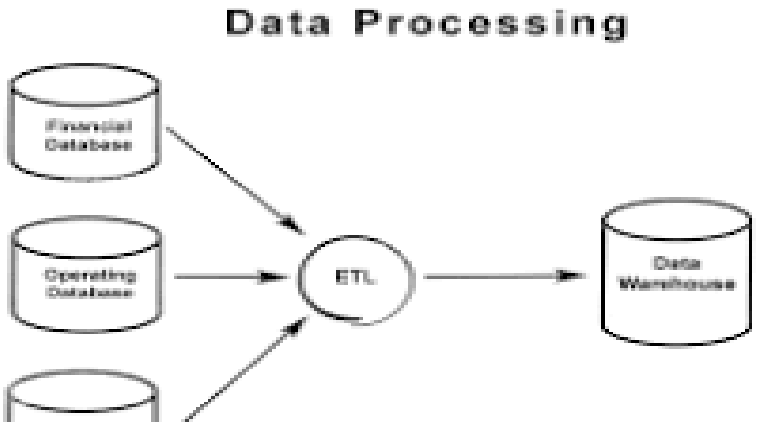
- Reliable water access is crucial for community health and development.
- Many waterpoints fail and need timely maintenance.
- Goal: Build machine learning models to classify waterpoints as functional or needs_attention.

Dataset Overview

- Waterpoint data with multiple features (e.g., location, age, water amount).
- Target classes: functional, needs_attention.
- Challenges: Imbalanced classes, Features on different scales.

Data Preprocessing

- Log-transformed skewed features like amount_tsh and population.
- Standardized numerical features using StandardScaler to normalize scales.
- Split into training and validation sets.



Modeling Approach

- Developed and compared three base models:
- 1. Logistic Regression
- 2. Decision Tree Classifier
- 3. Random Forest Classifier
- Evaluated based on accuracy, precision, recall, and F1-score.
- Tuned Random Forest with class weights for balance.

Logistic Regressior

- Accuracy: ~77.5%.
- Precision for 'needs_attention': 0.80.
- Recall for 'needs_attention': 0.68 (misses 32% failing wells).
- Strength: Good at identifying functional wells.

Decision Tree

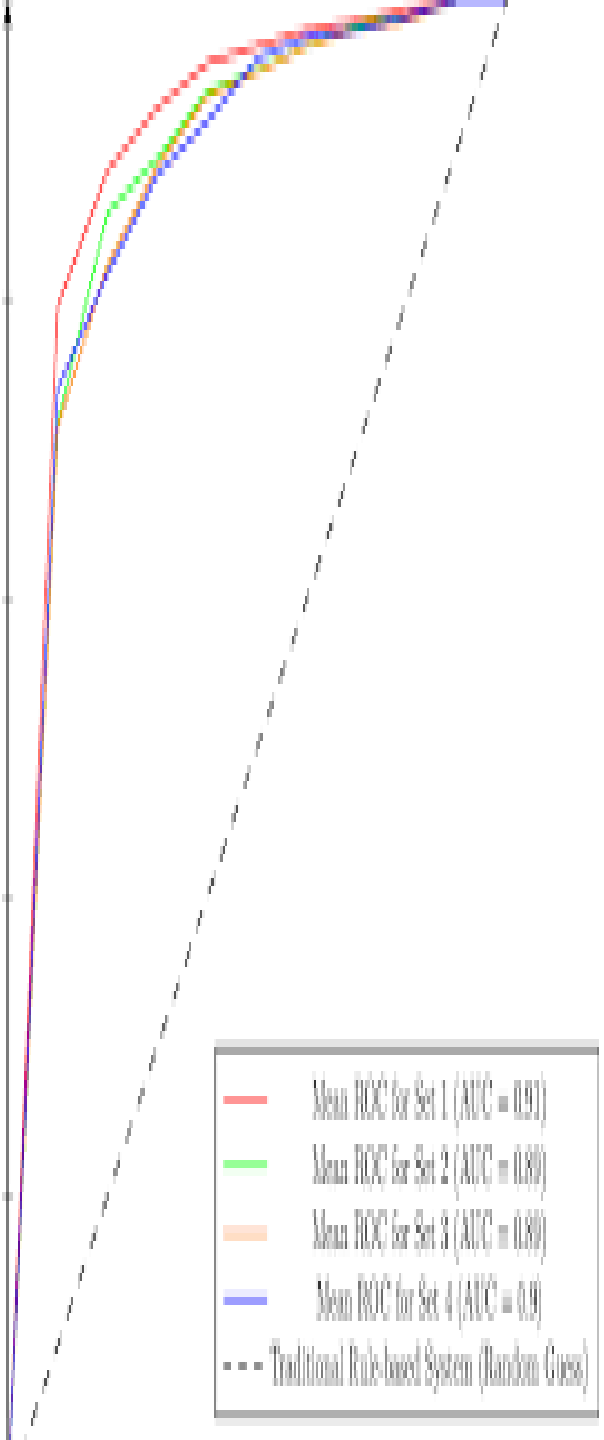
- Accuracy: ~77.0%.
- Recall for 'functional': 0.87.
- Recall for 'needs_attention': 0.64 (misses 36%).
- Slight overfitting to majority class observed.

Random Forest

- Accuracy: ~81.9%.
- Balanced precision and recall for both classes.
- Recall for 'needs_attention': 0.77 (improved detection).
- Most robust and reliable base model.

Hyperparameter Tuning

- Applied class_weight='balanced' to Random Forest.
- Final accuracy: ~82%.
- Balanced precision (0.82) and recall (0.77) for 'needs_attention'.
- Improved detection of failing wells.



ROC Curve & AUC

AUC score: 0.89

Indicates strong model discrimination ability.

Visualization: ROC curve shows true positive rate vs false positive rate.

Recommendations

- Deploy Random Forest model for practical waterpoint monitoring.
- Regularly retrain the model with new data to improve accuracy.
- Investigate false positives carefully to avoid unnecessary maintenance.
- Use model predictions to prioritize inspections and repairs.

Conclusion

- Machine learning can effectively classify waterpoint status.
- Random Forest provided best balance of accuracy and recall.
- Supports proactive maintenance and better resource allocation.
- Future work: Incorporate more features, explore ensemble methods, deploy in field.

Thank You!

- Questions?
- Contact: Charity Mwangangi.