Predictive Analytics for Waterpoint Operational Status in Tanzania

Using Machine Learning to Improve Water Access and Maintenance

Introduction

- Tanzania relies heavily on waterpoints for clean water access.
- Many waterpoints are non-functional or in need of repair.
- Inefficiencies result in hardships for local communities.
- Predictive analytics can enable proactive maintenance and better resource allocation.

Project Objectives

Primary Objectives:

- Build a predictive model to determine waterpoint operational status.

Goals:

- Enable proactive maintenance scheduling.
- Improve water access for communities.
- Assist stakeholders in resource optimization.

Stakeholders

1. TANZANIA MINISTRY OF WATER: RESPONSIBLE FOR PLANNING AND RESOURCE ALLOCATION.

2. LOCAL COMMUNITIES: RELY ON WATERPOINTS FOR DAILY WATER NEEDS.

3. MAINTENANCE TEAMS: TASKED WITH WATERPOINT REPAIRS AND UPKEEP.

Data Overview

Dataset includes 59,400 waterpoints with 41 features.

Key variables: location, construction year, water quality.

Target: Functional and Non-Functional.

Workflow

1. Business Understanding

2. Data Understanding

3. Data Preparation

4. Modeling and Evaluation

Data cleaning

- **Missing Data:** Imputed missing values in key columns (funder, installer, public_meeting, etc.) using 'Unknown' for categories and median for numerical fields.
- **Outliers:** Removed outliers in longitude and latitude using the IQR method.
- **High Cardinality:** Columns like id and subvillage show diverse data with many unique values.
- **Geography:** Data covers a broad area, though some errors (e.g., negative GPS height) were noted.
- **Target Variable:** Most entries are *Functional*, with moderate *Needs Repair* and the rest *Non-Functional*.

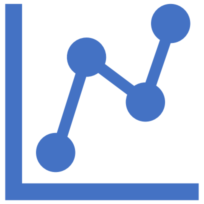
Model Comparison

GOAL: COMPARE THE PERFORMANCE OF THREE MODELS: LOGISTIC REGRESSION, DECISION TREE, AND RANDOM FOREST.

EVALUATE MODELS USING METRICS SUCH AS ACCURACY, AUC, AND INTERPRETABILITY.

PERFORM HYPERPARAMETER TUNING TO OPTIMIZE PERFORMANCE.

Model Explanations



Logistic Regression: Simple, linear model for binary classification. Strong predictive power, easy to interpret. Result: Performed well after tuning with good stability.

Decision Tree: Splits data into tree-like structures based on feature values. Result: Flexible but prone to overfitting; tuning mitigates this.

Random Forest: Ensemble method of multiple decision trees. Result: Best performance in terms of accuracy and AUC, stable and robust.

Model Comparison

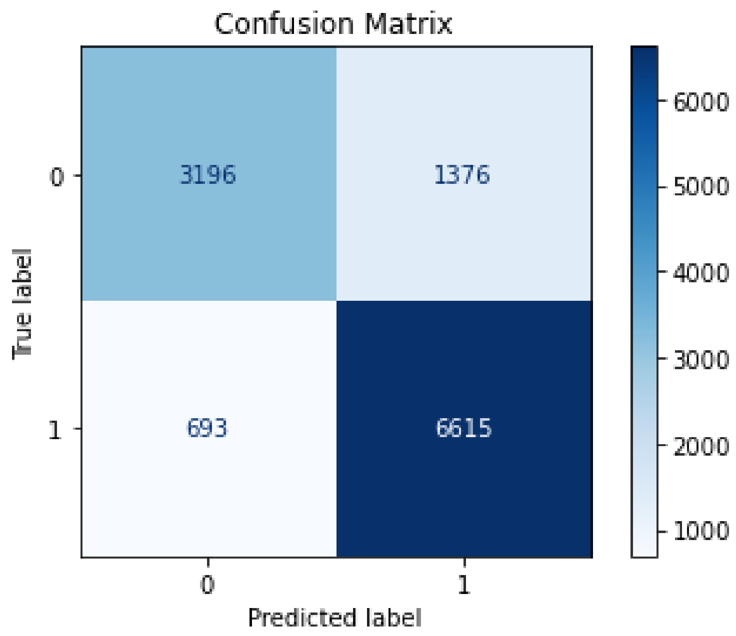
Metrics:

Accuracy: Random Forest led, followed by Logistic Regression and Decision Tree.

AUC: Random Forest outperformed both Logistic Regression and Decision Tree.

Interpretability: Logistic Regression > Decision Tree > Random Forest (due to complexity).

Final Recommendation: Random Forest



Why Random Forest?

Performance: Random Forest consistently outperformed other models in terms of **accuracy** and **AUC scores**, indicating strong predictive power and reliability across datasets.

Robustness: It handles feature interactions and complex datasets well, making it highly adaptable to real-world scenarios.

Scalability: Suitable for large, unbalanced datasets due to its ensemble nature.

Recommendations

- Focus on waterpoints identified as 'Needs Repair'.
- Regularly update the dataset for improved performance.
- Use GIS tools for better planning and resource allocation.

Thank You!

LinkedIn: www.linkedin.com/in/gideon-ochieng
GitHub: https://github.com/OchiengGideon/Phase-3-Project

Let's work together to ensure clean water access for all!

Contact us for further collaboration and insights.