TANZANIAN WATER WELLS PROJECT

- A MODEL THAT PREDICTS THE STATUS OF A WELL BASED ON TRAINING DATA

- **Presented by:** Charles Otieno Aloo

Overview

Project Goal: Predict the operational status of waterpoints across Tanzania to enhance access to clean, potable water.

Data Source: Taarifa waterpoints dataset from DrivenData.

Business Impact: Improve decision-making for water resource management and maintenance planning.

Business and Data Understanding

- **Business Problem:** Unreliable water pumps hinder access to clean water, impacting health and livelihoods.
- **Data Sources:** The dataset includes details on waterpoint characteristics, geographic location, water quality, and management.

Key Features:

- **Geographic data:** Longitude, latitude, region.
- **Waterpoint characteristics:** Construction year, management type, extraction type.
- **Outcome variable:** Waterpoint status - Functional, Functional needs repair, Non-functional.
- **Business Context:** Water access is critical in Tanzania. Timely maintenance of water pumps ensures uninterrupted water supply.

Data Description:

- 59,400+ waterpoints with attributes such as GPS coordinates, construction year, water quality, and management details.
- **Target Variable:** Operational status of the waterpoint (functional, functional needs repair, non-functional).
- **Features:** 40 features including numeric (e.g., `amount_tsh`, `gps_height`), categorical (e.g., `funder`, `source`), and date-based (`date_recorded`).

Data Exploration

Summary of Findings:

- Most waterpoints are functional, but a significant number require repairs.
- Geographic distribution of waterpoints reveals regional disparities in functionality.

Key Insights:

- Older waterpoints are more likely to be non-functional.
- Management type and extraction method play a crucial role in determining functionality.

Data Exploration

Summary of Findings:

- Most waterpoints are functional, but a significant number require repairs.
- Geographic distribution of waterpoints reveals regional disparities in functionality.

Key Insights:

- Older waterpoints are more likely to be non-functional.
- Management type and extraction method play a crucial role in determining functionality.

Modeling

Data Preprocessing:

- **Missing Data Handling:** Used most frequent value imputation for missing values.
- **Categorical Encoding:** Applied Label Encoding to categorical features.
- **Feature Scaling:** Standardized features using `StandardScaler`.

Model Selection:

- **Algorithm Used:** Random Forest Classifier, chosen for its robustness and ability to handle complex interactions.
- **Hyperparameter Tuning:** Employed GridSearchCV for fine-tuning parameters (e.g., `n_estimators`, `max_depth`).

Evaluation

Model Performance:

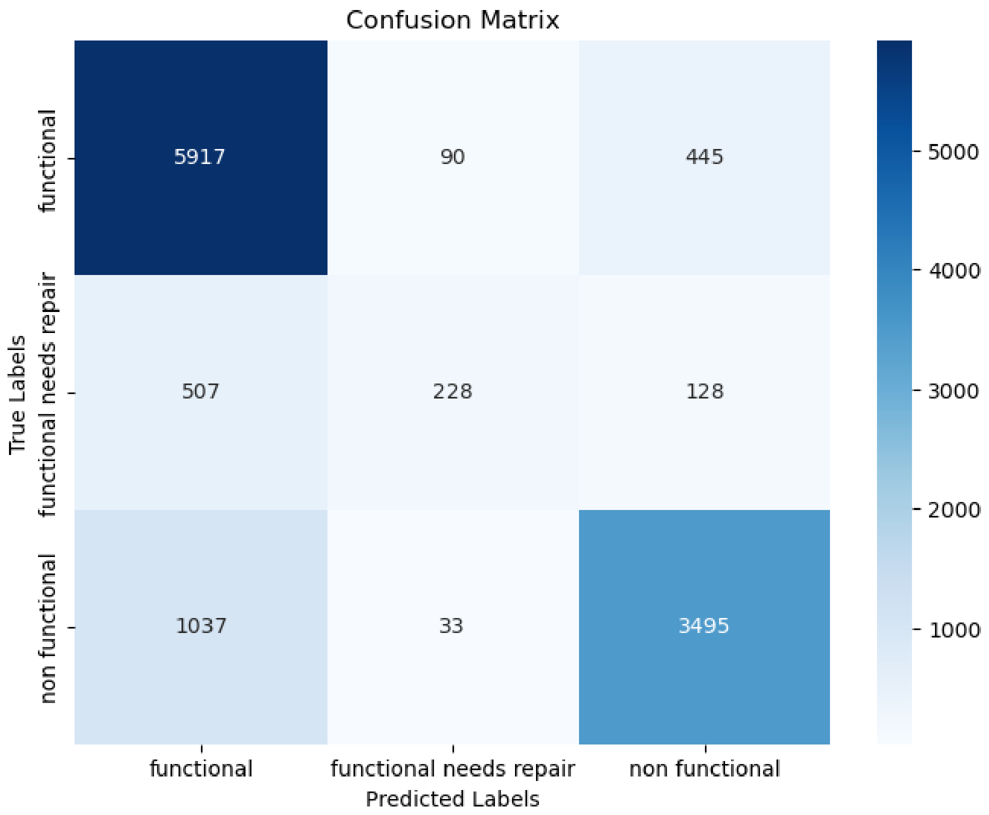
- **Accuracy:** 81.06% on validation set.
- **Confusion Matrix:** Visual representation showing true vs. predicted classifications.

Classification Report:

- **Precision:** High precision in predicting non-functional pumps.
- **Recall:** Good recall across all categories, with some room for improvement in "functional needs repair."
- **Best Model Parameters:** `n_estimators=300`, `max_depth=30`, `min_samples_split=5`.

Visuals:

- **Confusion Matrix**
- **Class Distribution**

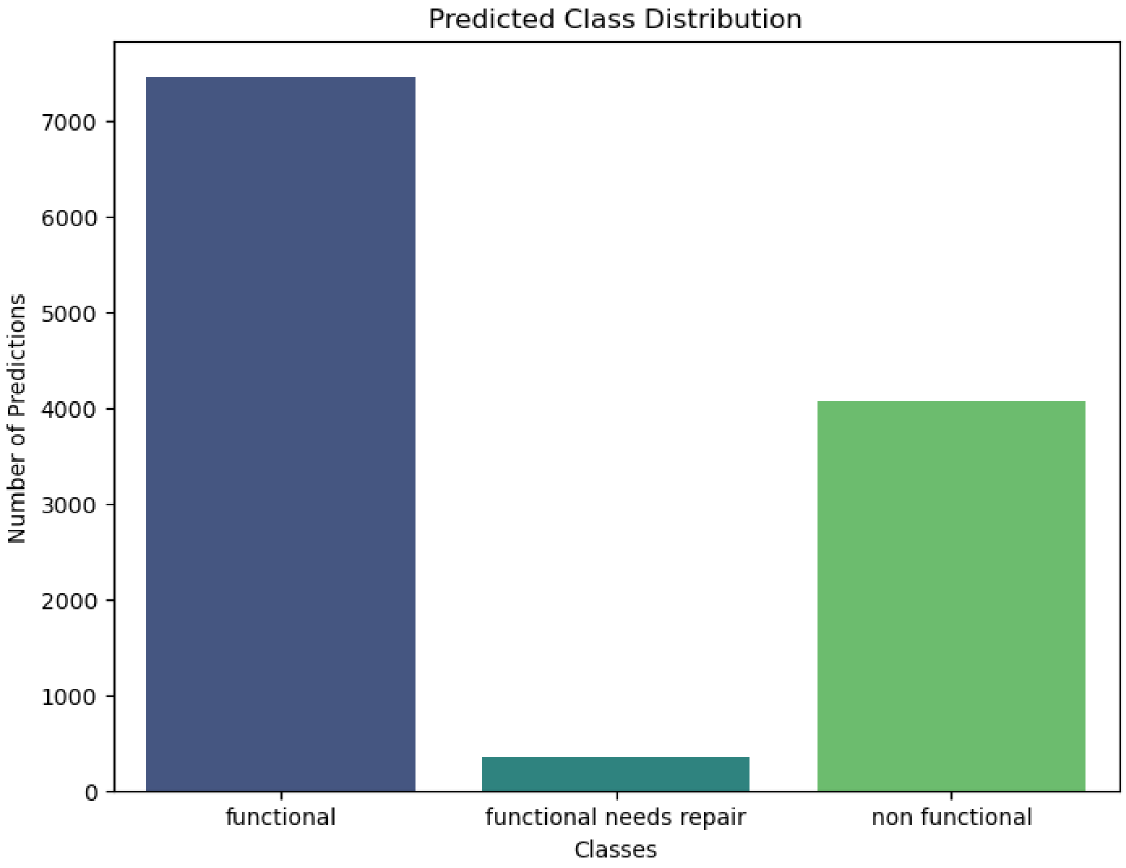


Confusion Matrix

- The model performs well in identifying `functional` and non `functional` waterpoints, as indicated by the higher precision and recall scores for these classes.
- The model struggles more with predicting `functional  needs repair` waterpoints, as shown by the lower precision and recall for this class.

Recommendations

- **Maintenance Scheduling:** Use predictions to prioritize maintenance for waterpoints flagged as "Functional needs repair."
- **Resource Allocation:** Focus on regions with a higher likelihood of non-functional waterpoints for resource allocation.
- **Further Research:** Investigate the impact of environmental factors on waterpoint functionality.
- **Operational Insights:** Focus maintenance efforts on regions with a higher concentration of non-functional waterpoints.
- **Model Deployment:** Integrate model predictions into resource management systems for real-time decision support.
- **Future Enhancements:** Incorporate additional data sources (e.g., weather patterns) to refine predictions.



Predicted Class Distribution

- Class 'functional': 7461 predictions
-
- Class 'functional needs repair': 351 predictions
-
- Class 'non functional': 4068 predictions