# Predictive Maintenance of Water Pumps in Tanzania

Enhancing Service Delivery Through Data Science,
by Leonard Koyio

# 1. Introduction

**Overview**: This project focuses on using data science to enhance the maintenance of water pumps in Tanzania.
**Importance**: Functional water pumps are crucial for reliable water supply, especially in rural areas.
**Goal**: Our aim is to predict the status of water pumps to ensure timely and efficient maintenance.

# 2. Business Understanding

- **Business Objective**:
  Assist the Tanzanian government in optimizing water pump maintenance by predicting pump functionality, ensuring consistent access to clean water for communities.

- **Business Problem**:
  Inefficient allocation of maintenance resources due to a lack of predictive insights, leading to over-servicing functional pumps and under-servicing those needing repairs, increasing costs and water supply disruptions.

# 2. Business Understanding

**Business Benefits**

- **Improved Resource Allocation**:
  Prioritize critical units, optimize resource use, and enhance service delivery by focusing efforts on pumps needing immediate attention.
- **Cost Savings**:Reduce repair costs and extend pump lifespan through preventative maintenance, leading to significant long-term savings.
- **Enhanced Water Access**:
  Ensure better access to clean water, vital for public health, by reducing non-functional pumps and improving water service reliability.

# 2. Business Understanding

- **Data-Driven Decision Making**:
  Foster a culture of data-driven strategies, improving efficiency in water pump maintenance and setting a precedent for other public services.

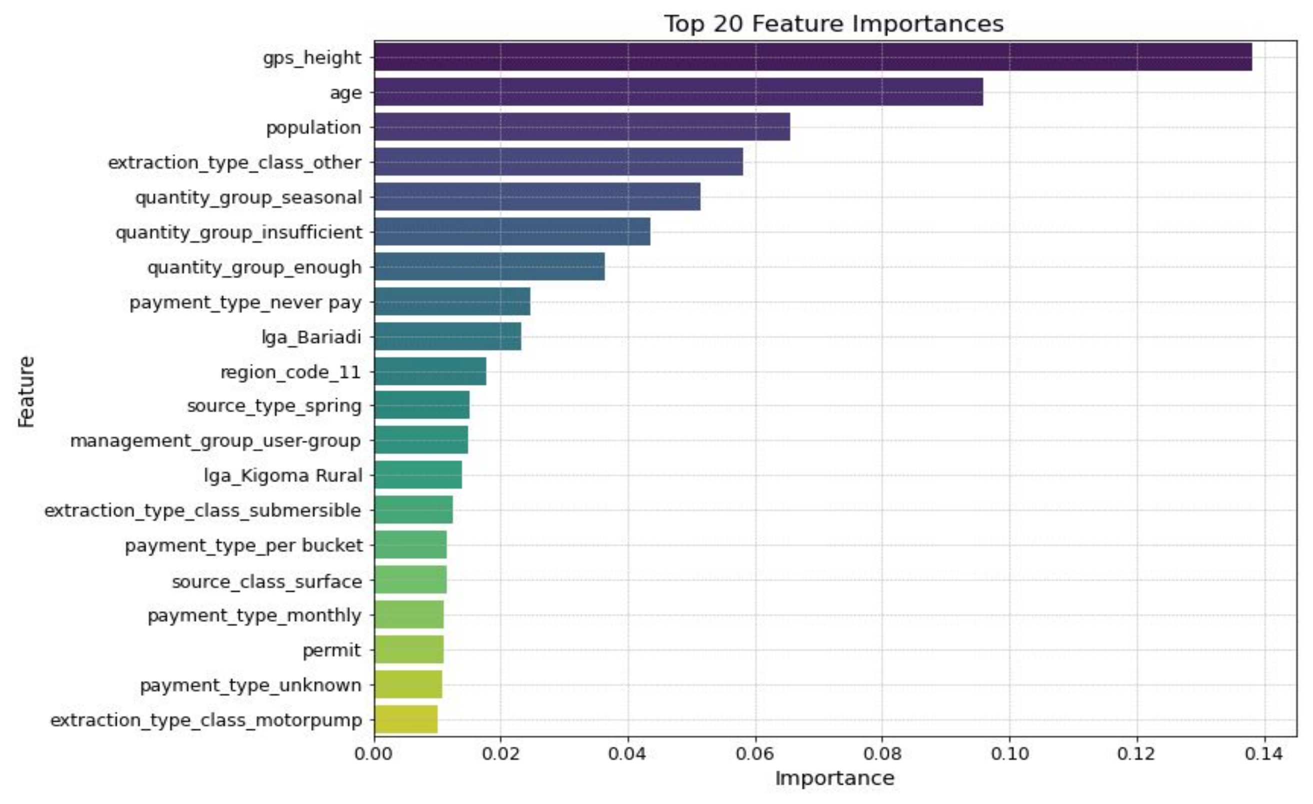
- **Improved Service Delivery**:
  Targeted and efficient maintenance approach reduces downtime and ensures quicker repairs, enhancing public trust and satisfaction with government services.

# 3. Data Understanding

**Dataset Description**: The dataset includes information on various features influencing the status of water pumps.
**Key Features**:

- **GPS Height**: Elevation of the pump location.
- **Age**: Age of the water pump.
- **Population**: Population served by the pump.
- **Extraction Type**: Type of extraction method used.
- **Quantity Group**: Water quantity availability (e.g., seasonal, insufficient, enough).

# Visualization : top 20 model features



# 4. Exploratory Data Analysis (EDA)

- **Data Cleaning**:
  - **Remove Redundant Columns**: Eliminated columns with no predictive value or columns with repeated data to streamline the dataset.
  - **Handle Missing Values**: Filled zeros and NaNs with suitable data, such as median or mean values, to maintain data integrity.
- **Data Transformation**:
  - **Min-Max Scaling**: Applied Min-Max Scaling to normalize the range of independent variables, ensuring all features contribute equally to the model.
  - **One-Hot Encoding**: Converted categorical variables into binary vectors to allow the model to interpret them effectively.
- **Feature Engineering**:
  - **Age of the Pump**: Calculated the age of each pump by subtracting the installation year from the current year, providing a crucial feature for predicting functionality.

# 5. Modeling

**Baseline Models**:

- Developed initial models using Logistic Regression and Decision Trees with default hyperparameters to establish a performance benchmark.
- Iteratively tuned models based on performance metrics.

**Class Imbalance Handling**:

- **Class Weights**: Adjusted weights to handle class imbalance within the model.
- **SMOTE**: Used Synthetic Minority Over-sampling Technique to balance class distribution by generating synthetic samples for the minority class.

# 5. Modeling

**Model Regularization**:

- **L1 Regularization**: Promotes sparsity by driving some coefficients to zero, improving model interpretability.
- **L2 Regularization**: Penalizes large coefficients to reduce overfitting, enhancing model generalizability.

**Decision Tree Tuning**:

- **Criteria**: Compared Gini and Entropy criteria to determine the best split quality measure.
- **Max Depth Optimization**: Tuned the maximum depth parameter to prevent overfitting and improve model performance.

# 6. Model Evaluation

**Settled Model**: Decision Tree trained on dataset without outliers.

- **Class Imbalance Handling**: SMOTE to balance class distribution.
- **Criterion**: Gini impurity criterion.
- **Max Depth**: None specified.

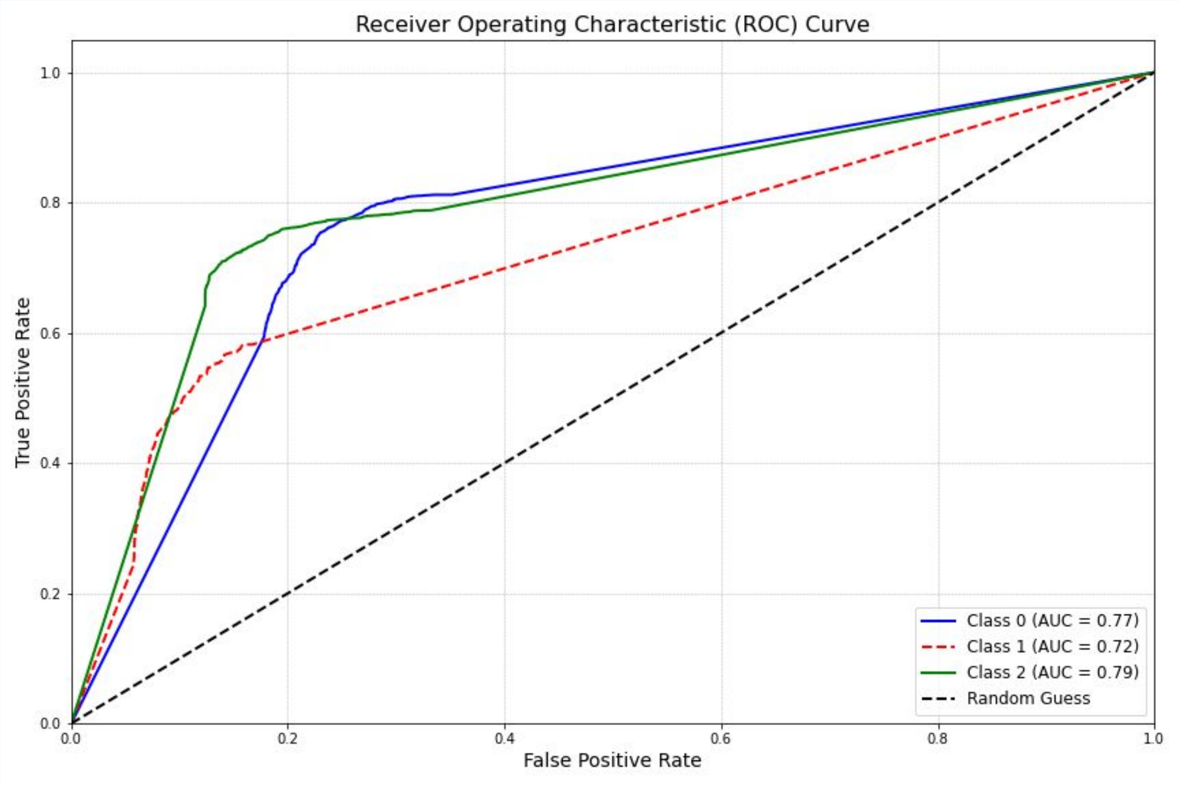
**Metrics**:

- **Weighted Recall**: 0.7128
  - Indicates that the model correctly identifies 71.28% of instances, considering class distribution, which is critical for ensuring non-functional and repair-needed pumps are detected.
- **Weighted Precision**: 0.7443
  - Shows that 74.43% of the positive predictions are correct, reflecting high accuracy in identifying pumps needing attention.

# 6. Model Evaluation

- **Weighted F1 Score**: 0.7253
  - Balances precision and recall, ensuring the model is effective in both identifying and accurately predicting pump statuses.
- **ROC AUC Score**: 0.7626
  - Measures overall performance; a score of 0.7626 indicates a good balance between sensitivity and specificity, which is suitable for our objective of predictive maintenance.

# 7. Model Visualizations : a. ROC Curve

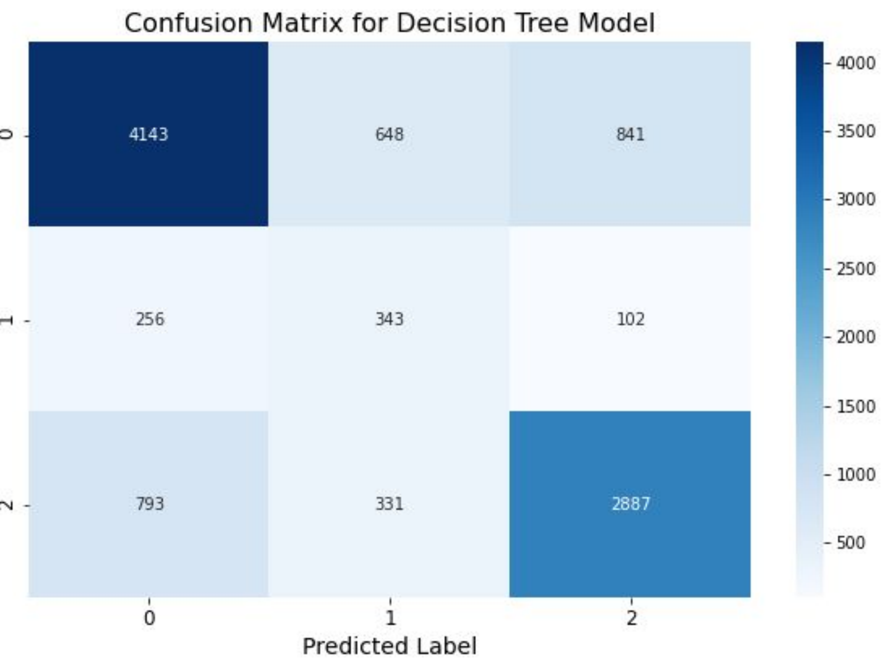



**Class 2 (Non-Functional):** High AUC of 0.79 ensures accurate identification of out-of-service pumps, crucial for timely repairs.

**Class 1 (Needs Repair):** AUC of 0.72 indicates good detection of pumps needing maintenance, preventing further issues.

**Class 0 (Functional):** AUC of 0.77 helps confirm operational pumps, reducing unnecessary checks and optimizing resource use.

# 7. Model Visualizations : b. Confusion Matrix




Key

0 - Functional

1- Functional Needs Repair

2- Non- Functional

The model is well-balanced, with no strong bias towards any class, ensuring fair consideration of all pump statuses.

# 8.Predictive Recommendations

**Model Utility**:

- **Most Useful**: In rural or remote areas where quick maintenance decisions are crucial. Helps prioritize repairs and maintenance.
- **Less Effective**: In rapidly changing conditions or where historical data is outdated.

**Suggestions for Improvement**:

- **Data Accuracy**: Ensure accurate and up-to-date data on pump conditions and maintenance.
- **Feature Engineering**: Add features related to usage patterns, maintenance history, and environmental conditions.

# 9. Project Impact

**Improved Maintenance Efficiency**: Enables timely repairs and consistent water supply.
**Cost-Effective Resource Allocation**: Optimizes budget use and reduces unnecessary expenses.
**Scalability and Adaptability**: Easily adapts to include more features or updated data.

# 10. Conclusion

**Project Success**:

- Developed a robust predictive model that categorizes water pumps as functional, needing repair, or non-functional using data from Taarifa and the Tanzanian Ministry of Water.
- Directly addresses the need for efficient water pump maintenance in Tanzania.

**Model Advantages**:

- **Accuracy**: The Decision Tree Classifier delivers reliable predictions, essential for timely maintenance decisions.
- **Efficiency**: Enhances maintenance scheduling and reduces downtime, ensuring consistent water supply.
- **Cost-Effectiveness**: Optimizes resource allocation and minimizes expenses by identifying priority pumps for repair.
- **Scalability**: Easily adaptable to incorporate new data or features, supporting ongoing infrastructure management.

**Stakeholder Value**:

- **For the Government**: A critical tool for optimizing water infrastructure management and aligning with national goals for improved water access.
- **For Citizens**: Ensures reliable access to clean water, contributing to enhanced community well-being.