Tanzania Well Failure Forecasting: Insights from Machine learning models.

Context.

1.Introduction.

2.Key objectives.

3.Key features.

4.Exploring Well Data.

5.Machine learning techniques.

6.Insights & Recommendations.

INTRODUCTION.

Access to clean water is essential, yet many water pumps in Tanzania fail due to aging infrastructure, poor maintenance, and environmental factors. These failures disrupt communities and increase repair costs.

This project analyzes key factors influencing pump functionality using data science and machine learning. By identifying patterns in pump failures, we aim to develop a predictive model to classify pumps based on their risk of failure.

The insights generated will help decision-makers optimize maintenance schedules, allocate resources efficiently, and extend pump lifespan, ultimately improving water access for communities.

KEY OBJECTIVES.

1. Identify key factors affecting water pump functionality.
2. Analyze geographic, technical, and managerial influences on failure rates.
3. Determine which pump types and water sources are most prone to failure.
4. Assess the impact of management on well maintenance.
5. Develop a predictive model to classify pump status.
6. Provide actionable recommendations based on data-driven insights.

Key features.

Identification & Status:

The **ID** and **Date Recorded** track pump data, while the **Status Group** (Functional, Non-functional, Needs Repair) serves as the target v**agement**:variable for predicting pump performance.

Installation & Management

Funder and **Installer** assess the impact of funding and expertise, while **Management** and **Payment Type** link ownership to pump sustainability.

Location & Environment:

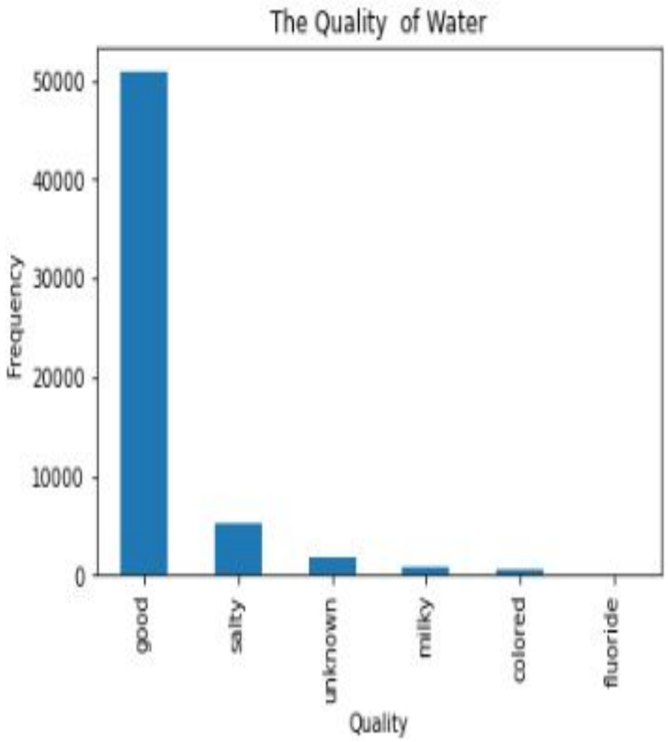
Region, **GPS**, and **Water Source** help identify geographical patterns in pump failures and environmental factors affecting performance.

Technical Aspects:

Technical features like **age**, **type**, and **maintenance history** provide insights into mechanical and operational factors influencing pump failure.

Data Source: DrivenData – Pump It Up

Exploring Well Data: Key Trends and Patterns.

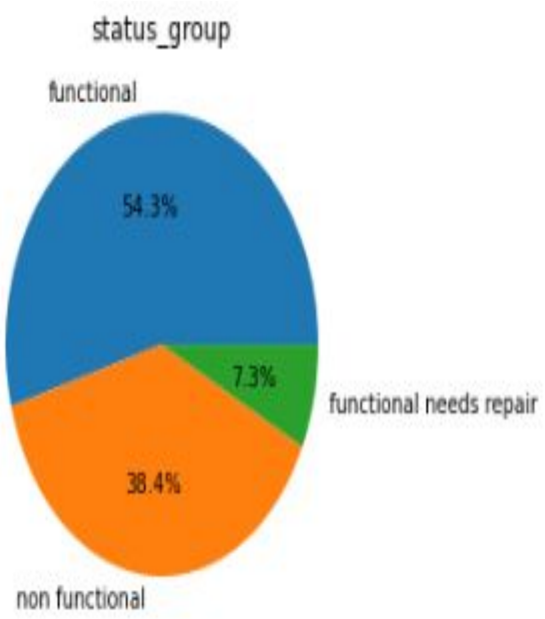


1. Pump age .

The pumps in the dataset range from **12 to 65 years**, with **14 years** being the most common age. **Older pumps**, particularly those closer to **65 years**, are more prone to **failure** due to **wear and tear**. While most pumps are relatively young, **aging infrastructure** increases failure risks, highlighting the need for **proactive maintenance**.

2.Water quality

Most wells provided **good water quality**, ensuring **safe usage**. However, a few wells showed signs of slight **salinity** or a **milky appearance**. These issues can affect water **usability** and may require additional **treatment** to meet safety standards. **Regular monitoring** is essential to address these concerns before they become larger problems.



3. Well status

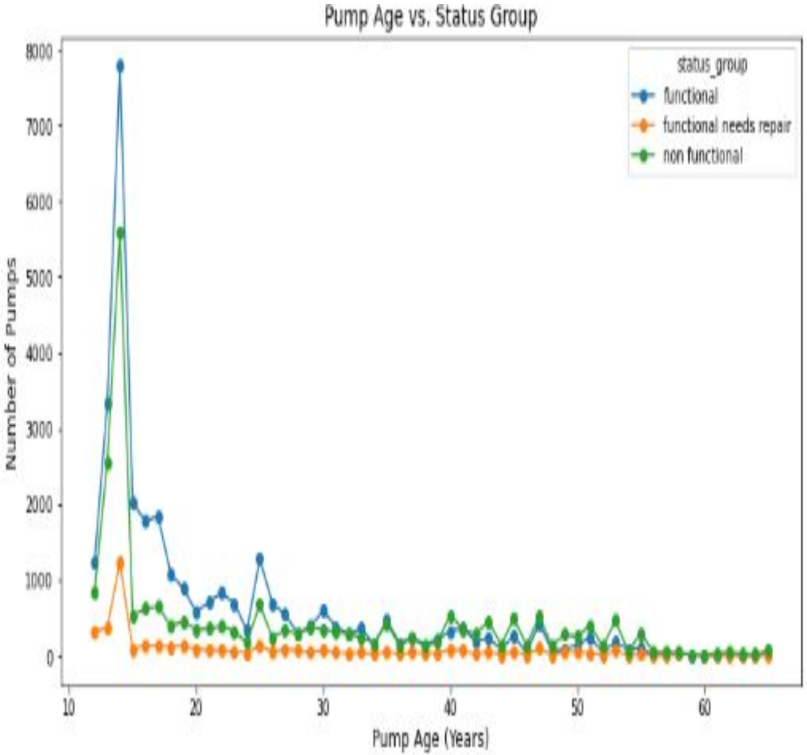
The majority of wells (**54.4%**) were **fully functional**, providing reliable water access. **7.3%** were **functional but needed repairs**, indicating potential maintenance issues. Meanwhile, **38.4%** were **non-functional**, highlighting significant challenges in water accessibility and infrastructure upkeep.

4.Well funders.

The **Government of Tanzania** is the leading funder of water wells, playing a major role in infrastructure development. Other key contributors include **DANIDA, HESAWA, RWSSP, World Bank, KKT, World Vision, UNICEF, TASAF, and District Councils**, highlighting the involvement of both international and local organizations in improving water access.

5..Water quantity.

Most wells provided **enough water** to meet community needs, while some were classified as **sufficient** but not abundant. A few wells were **dry**, others had **seasonal water availability**, and some had **unknown water quantity**, indicating gaps in data or inconsistent supply.



6.Permit issuance.

The **VMC scheme** had permits for most of its wells, ensuring regulatory compliance. However, a few wells operated **without permits**. This pattern was consistent across other schemes, highlighting both adherence to regulations and instances of unpermitted installations.

7.Management groups by population.

Most of the population relied on wells **managed by user groups**, ensuring community oversight. Others were managed by **commercial entities** and **parastatals**, playing a role in regulated water distribution and maintenance.

8. Pump age by performance.

Younger pumps are **performing well**, with fewer issues and consistent functionality. However, **older pumps** are more **prone to failure**, requiring frequent maintenance or replacement due to wear and aging infrastructure.

Machine Learning Techniques for Well Performance Prediction.

Model 1: Logistic Regression for Well Failure Prediction

Purpose:
The model predicts whether a well will fail (1) or not (0) based on available data.

Algorithm:
Logistic Regression is used, which outputs a probability. Wells are classified as failures (1) or non-failures (0) with a threshold of 0.5.

Evaluation.

Class 0 (no failure):
The model has a precision of 0.64, meaning it correctly predicts 64% of non-failing wells. The recall is 0.84, indicating it identifies 84% of actual non-failures. The F1-Score is 0.73, balancing precision and recall.

Class 1 (failure):
The precision is 1.00, meaning all predicted failures are correct. However, recall is 0.00, meaning it fails to identify any actual failures. The F1-Score is 0.00, reflecting poor performance in predicting failures.

Overall Accuracy:
The model's overall accuracy is 64%, showing the proportion of correct predictions across both classes.

Model 2: Decision Tree for Well Failure Prediction

Goal:

The objective is to improve the prediction of well failure using a Decision Tree model.

First Model (Gini Impurity):

The first model used Gini impurity for splitting the data. It achieved an accuracy of 73.64%, but the recall for pumps likely to fail was very low at 0.09, indicating poor identification of failing pumps.

Second Model (Entropy):

The second model used Entropy as the splitting criterion, which improved the results. The accuracy increased to 74.27%, and recall for pumps likely to fail improved to 0.17, showing better performance in detecting failures.

Model 3: Random Forest for Well Failure Prediction

Purpose:

- **Goal**: Improve predictions for well failure using a **Random Forest** model.

Results of the Random Forest Model:

- **Accuracy**: The accuracy score remained **consistent**(74.44%)with the previous models.
- **Recall**: The recall improved significantly to **0.51 / 51%**, making the model more effective at identifying wells likely to fail.
- **SMOTE Oversampling**: To address class imbalance, **SMOTE (Synthetic Minority Over-sampling Technique)** was applied, which helped in improving recall for the less frequent class (wells likely to fail).

Key insights from best performing model.

1. **Consistent Accuracy**: The Random Forest model achieved an accuracy of 74.44%, similar to the previous models, indicating stable performance.
2. **Improved Recall**: The recall for wells likely to fail increased significantly to 0.51%, making the model much more effective at identifying potential failures compared to earlier models.
3. **SMOTE Oversampling**: Applying SMOTE helped address class imbalance, improving recall for the less frequent class (wells likely to fail) and making the model better at predicting failures.

Key recommendations for Funders.

1. **Regular Maintenance**: Prioritize wells at high risk of failure for routine inspections and maintenance based on model predictions, reducing downtime.
2. **Early Detection**: Use the model for proactive detection of potential failures, allowing for timely interventions and minimizing unexpected breakdowns.
3. **Dynamic Threshold Adjustments**: Adjust the model's decision threshold to improve failure detection, ensuring more wells are inspected as needed.
4. **Predictive Maintenance**: Combine model predictions with predictive maintenance tools to better anticipate and address failures.
5. **Continuous Model Updates**: Regularly update and retrain the model with new data to maintain accurate predictions over time.

Conclusion.

By working closely with the government of Tanzania, its funders, and well installers, we can improve predictions, optimize maintenance, and ensure better well functionality. This collaboration will reduce failures and enhance the sustainability of the wells.

Thank you!

About Me.

Name: Linet Shammah Patriciah

Program: Data science

School: Moringa school

Email: linet.patriciah@student.moringaschool.com

Github : https://github.com/shammy-lp