TECHNICAL DATA REPORT:

Predictive Modeling for Water-Wells Maintenance in Tanzania

AUREL EPHY OCHIENG

Phase 3 project

DATA SCIENCE

Abstract

Access to clean and safe water remains a critical challenge in Tanzania, particularly in rural areas where well functionality is often compromised by limited infrastructure, harsh environmental conditions, and insufficient funding. This study leverages data from the "Pump It Up: Data Mining the Water Table" competition to address the problem of predicting well functionality using machine learning. The dataset contains information on well characteristics, location, management, and status (functional, non-functional, functional-needs repair). Through a structured methodology, we explored, cleaned, and preprocessed the data to enhance predictive modeling. Irrelevant and redundant columns, such as identifiers and geographic features already represented in coded form, were removed to reduce noise. Missing values were imputed using the most frequent value, and a preprocessing pipeline was developed to standardize numerical data and encode categorical features. Several Machine learning models, including Random Forest, XGBoost, and Stacking Classifiers, were evaluated using accuracy and ROC-AUC scores. The Stacking Classifier demonstrated the best performance, achieving an accuracy of 82.5% and a ROC-AUC score of 0. 9037.Visualizations highlighted key insights, such as the distribution of well statuses, water point types, and regional functionality trends, offering actionable insights for targeted well maintenance strategies. This predictive approach provides a scalable solution to optimize resource allocation, improve water access, and ensure sustainable well functionality in Tanzania.

Contents

List of figures

Predictive Modeling for Water-Wells Maintenance in Tanzania

Business understanding

Tanzania faces challenges in providing clean water to its population, particularly in rural areas where access to safe drinking water and well maintenance are significant issues. These challenges often result from a combination of limited infrastructure, funding constraints, and the harsh environmental conditions that affect the sustainability of water sources like wells. To address these issues, the goal is to identify wells requiring repair using predictive modeling. Accurate prediction helps allocate resources effectively and supports sustainable water access initiatives.

Dataset Understanding

The dataset from the "Pump It Up: Data Mining the Water Table" competition reflects real challenges in identifying wells needing repair in Tanzania, underscoring the importance of predictive modeling to address this issue. Key attributes:

1. Categorical Features: region, funder, installer, scheme_management.
2. Numerical Features: gps_height, population, amount_tsh (amount of water).
3. Target Variable: status_group (functional, non-functional, functional-needs repair).

Data Preparation

To streamline the dataset and reduce noise, the following columns have been dropped:

Irrelevant Features: These columns were removed as they do not directly affect well functionality:

1. id, wpt_name (identifiers).
2. scheme_name, funder, recorded_by, installer (names of operators, funders, or recorders).
3. date_recorded (date row was entered).

Geographic Redundancies: Geographic information is already captured in latitude, longitude, region_code, and district_code:region, subvillage, ward, lga, basin.

Feature Duplication: These columns were removed as their information is captured more accurately in other features:

1. quality_group (redundant with water_quality).
2. quantity_group (redundant with water_quantity).
3. source_type, source_class (redundant with source).

4. waterpoint_type_group (redundant with waterpoint_type).
5. extraction_type_group, extraction_type_class (redundant with extraction_type).
6. management_group (redundant with management).
7. payment_type (redundant with payment).

Recategorization of Numerical Features: region_code and district_code were recategorized as categorical features since they represent coded location descriptors.

Preprocessing Pipeline: A pipeline was implemented to streamline preprocessing, avoid data leakage, and standardize steps across all models.

Numerical Data:

1. No missing values were found in the numerical columns.
2. Applied StandardScaler to standardize and scale numerical data for consistency.

Categorical Data:

1. Missing values were imputed using SimpleImputer with the most frequent value in each column. A 'missing' indicator was added to denote imputed values.
2. Categorical variables were encoded using OneHotEncoder to transform them into binary variables for modeling.

Target Variable: The target variable, status_group, remained unaltered as it represents the functionality status of the wells.

Data Understanding through Visualization

Step 1: Data visualization results:

Through data understanding I was able to get results that helped in predictive data modelling as shown below:

Distribution of Well Statuses

The bar chart shows the distribution of well statuses, this provides a high-level overview of the well conditions in the dataset.

Functional: Operational wells.

Non-Functional: Wells no longer working.

Functional-Needs Repair: Wells requiring maintenance.

Water Point Type

A breakdown of water_point_type highlights the most common types of water points (e.g., communal standpipes, hand pumps) and their relationship to functionality.

Water Quantity

The quantity variable (e.g., "enough," "dry," "insufficient") reflects the availability of water and its connection to repair needs.

Extraction Type

An analysis of extraction_type (e.g., hand pumps, motor pumps) reveals which systems are more prone to failure.

Regional Distribution

Geographic variations in region_code and district_code show which areas report the highest percentage of non-functional wells, enabling targeted interventions

The above categories are represented in the figures below:

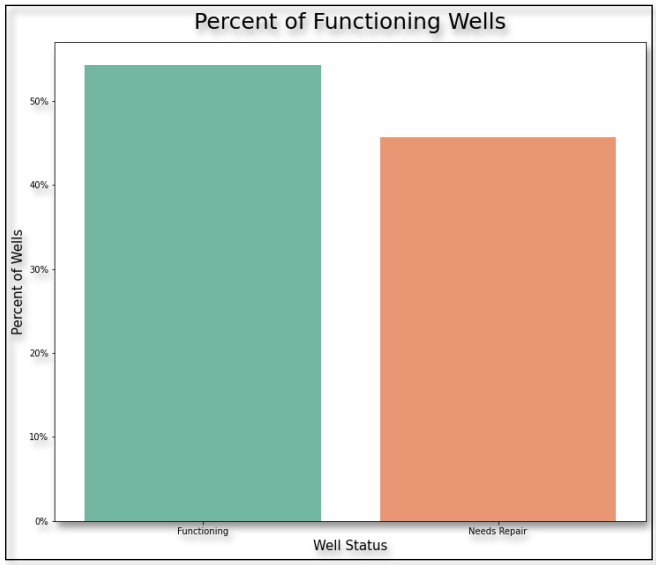

Figure 1: The bar chart shows the distribution of well statuses

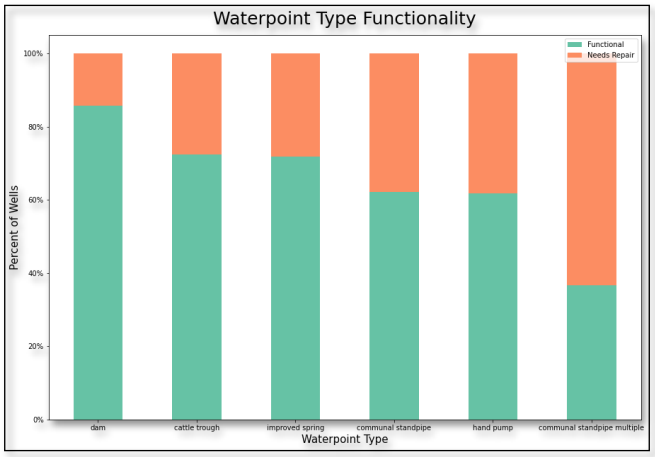

Figure 2:Bar chart showing functionality of water points

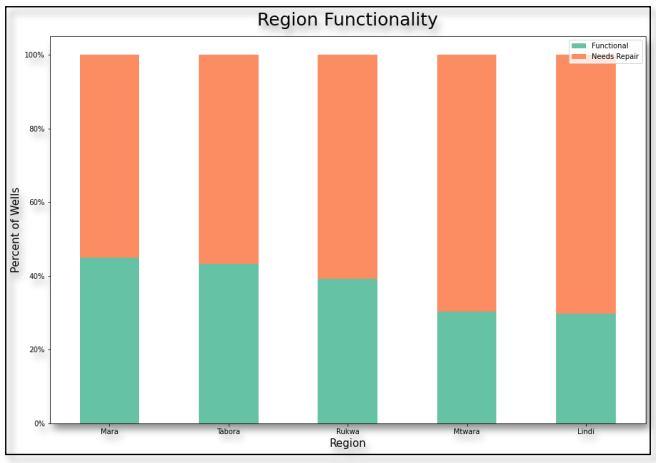

Figure 3: Bar chart showing regions with low functioning water wells

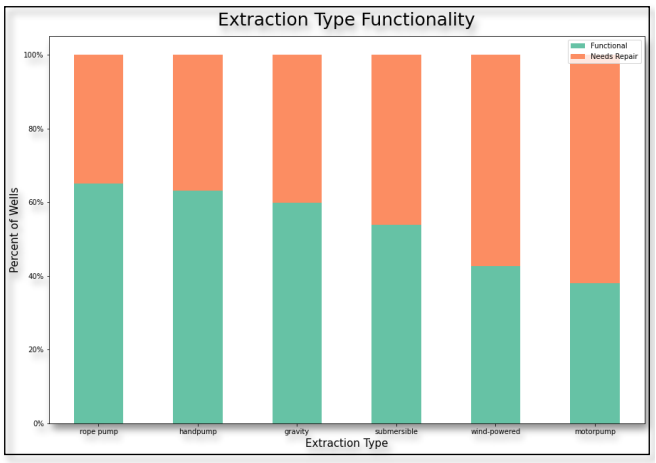

Figure 4: Bar chart showing extraction complexity and need of maintenance

Step 2: Model Selection and Training:

Experimented with a range of machine learning models, including Logistic Regression, Random Forest, SVC, and ensemble methods such as AdaBoost and Stacking Classifiers. I then split the data into training and testing sets to evaluate performance and avoid overfitting.

Step 3: Evaluation

Assessed each model using metrics such as accuracy and ROC-AUC score. Focused on the Stacking Classifier, which demonstrated the best overall performance.

Data modelling

In the analysis to predict Tanzanian water wells in need of repair, dozens of models were evaluated based on three main criteria: Accuracy, ROC-AUC score, and Fit and Run Time. Here's a detailed breakdown of how these factors influenced the selection of the final model:

Accuracy

Accuracy measures the proportion of correctly predicted wells (both those in need of repair and those not) out of the total predictions. While accuracy is a straightforward metric, it doesn't account for imbalances in true positives and true negatives, making ROC-AUC a complementary metric.

ROC-AUC Score

The ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) score evaluates the trade-off between true positive rates (sensitivity) and false positive rates. A higher score indicates better discrimination between wells needing repair and those not requiring attention. ROC-AUC was critical for this analysis as the cost of false negatives (missing wells that need repair) might outweigh that of false positives.

Fit and Run Time

The runtime is a vital consideration for large datasets or scenarios where predictions need to be updated frequently. It involves both the time required to train the model (fit time) and to make predictions (inference time).

The selection of a final model depends on balancing predictive performance (accuracy and ROC-AUC) with runtime efficiency: The above processes were displayed as shown in the figures below:

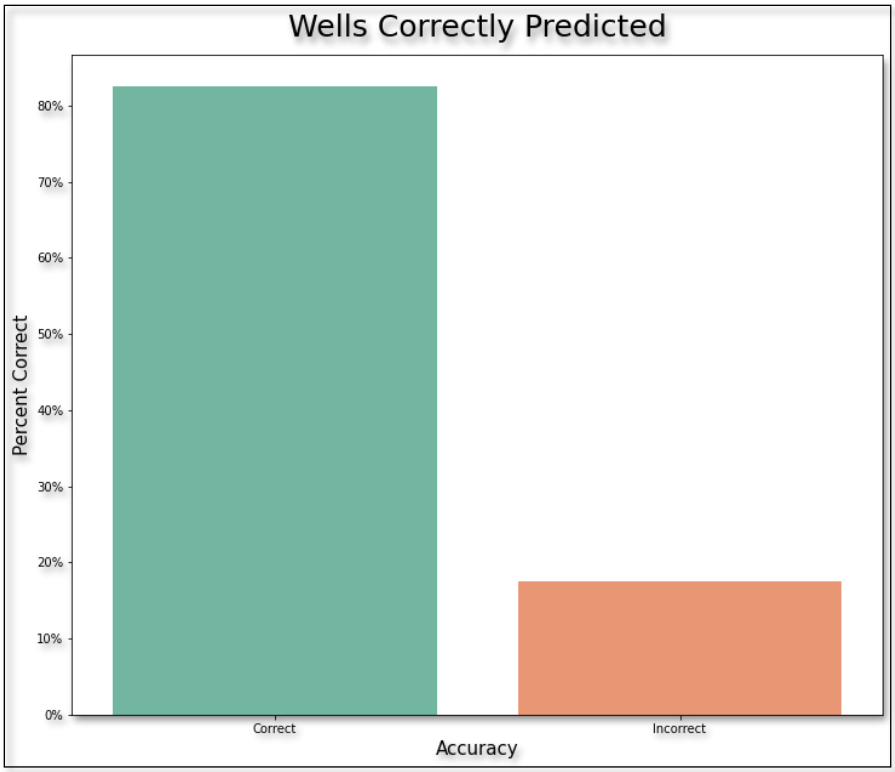

Figure 5: bar chart showing percentage of well correctly predicted

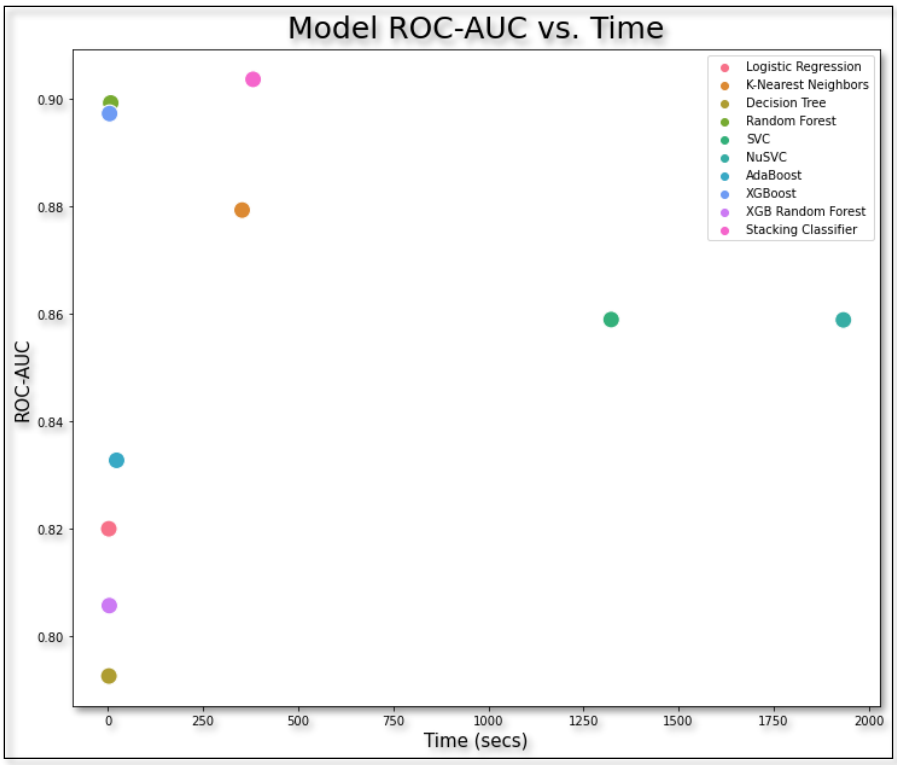

Figure 6: scatter plot correlation between ROC-AUC to time

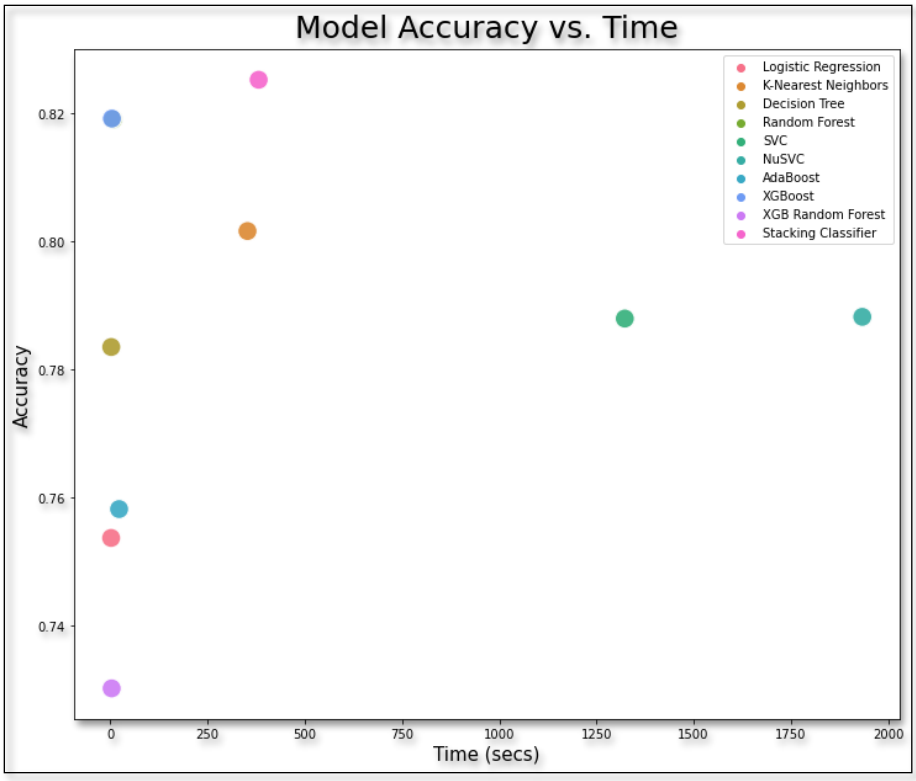

Figure 7:scatter plot correlation between model accuracy to time

Accuracy

The results table lets clearly shows the performance of the models run in this analysis. Overall, the Stacking Classifier has the best results with an accuracy of 82.5% and an ROC-AUC score of 90.3% on the test set. However, it took roughly 55 times longer to fit than the Random Forest Final Model. When looking at time-to-fit in addition to pure performance metrics, the Stacking Classifier took a little over 6 minutes to fit. It is recommended to use the Stacking Classifier for future predictions of wells needing repair in Tanzania.

Stacking Classifier is the best choice for performance-critical applications, as it achieved the highest accuracy and ROC-AUC score. Its higher computational cost is justified if predictions are made periodically or for smaller datasets.

Random Forest and XGBoost are more suitable for frequent or real-time predictions, offering excellent trade-offs between performance and runtime.

Simpler models like Logistic Regression are appropriate when speed and interpretability outweigh the need for high accuracy.

This multi-criteria evaluation ensures the final model aligns with the project's priorities, delivering reliable predictions for Tanzanian water wells in need of repair.

Table 1:Table showing accuracy-AUC score and runtime of the models

Models	Accuracy	ROC-AUC score	Run time
Logistic Regression	0.753670	0.820043	2.037007
K-Nearest Neighbors	0.801616	0.879372	352.673122
Decision Tree	0.783502	0.792631	2.322550
Random Forest	0.819057	0.899311	7.185174
SVC	0.787946	0.858988	1322.947074
NuSVC	0.788215	0.832784	1933.638887
AdaBoost	0.758182	0.832784	22.538390
XGBoost	0.819192	0.897334	4.091154
XGB Random Forest	0.730168	0.205740	3.413922
Stacking Classifier	0.825253	0.903708	281.458932

Model performance Comparison

This analysis underscores the trade-offs between model complexity, accuracy, and runtime. The decision to use the Stacking Classifier reflects prioritization of prediction quality over computation time. The Stacking Classifier can be used for future predictions of wells needing repair in Tanzania due to its top accuracy and ROC-AUC score, making it the best choice for performance-critical scenarios. If runtime is a significant constraint, the Random Forest model is an excellent alternative. It provides nearly equivalent performance at a fraction of the runtime.

Table 2: Table showing observation of the models

Models	**Observation**
Logistic Regression	Efficient and simple, though less accurate than ensemble methods
K-Nearest Neighbors	Decent accuracy but has a very long runtime for fitting compared to similar-performing models.
Decision Tree	Quick to fit but has lower accuracy and ROC-AUC compared to ensemble methods
Random Forest	A strong contender in terms of performance, with much faster training time compared to the Stacking Classifier.
SVC	Relatively strong performance metrics but prohibitively long runtimes
NuSVC	Relatively good
AdaBoost	Efficient runtime with moderate performance.
XGBoost	Comparable performance to Random Forest with slightly lower ROC-AUC and faster runtime.
XGB Random Forest	Poor ROC-AUC score makes this model unsuitable

Stacking Classifier	Best overall performance across accuracy and ROC-AUC but significantly higher computational cost compared to simpler models.

Conclusion

In conclusion, improving water well infrastructure in Tanzania requires a focused and strategic approach. By enhancing well technology, prioritizing year-round water availability, and investing in communal standpipe wells, the country can make significant strides in addressing water access challenges. Concentrating efforts in high-need regions such as Lindi and Mtwara ensures that resources are directed where they are most needed.

Deploying and refining predictive models will further augment traditional methods of identifying wells in need of repair, leading to more efficient use of resources. While the model may not be perfect, its integration promises a significant reduction in costs and manpower requirements, making well maintenance more sustainable.

Ultimately, these efforts will result in improved water access for Tanzanian citizens, enhancing health, well-being, and overall quality of life. With a commitment to continuous improvement and adaptation, Tanzania can ensure a brighter future for its communities through reliable and sustainable water supply systems.

Based on the evaluation of various regression and classification models, it is evident that machine learning algorithms can effectively predict the condition of waterpoints in Tanzania

Recommendations

Model Refinement and Optimization: Focus on reducing overfitting in models such as Decision Trees and Random Forests by applying techniques like hyperparameter tuning, cross-validation, and feature selection. This would enhance the robustness and predictive accuracy of both individual models and the Stacking Classifier.

Efficiency Improvements: Address the long run times of computationally expensive models like SVC and NuSVC by experimenting with reduced feature sets, kernel optimizations, or sampling techniques.

Enhanced Data Cleaning and Exploration: Allocate more time to expand initial exploratory data analysis (EDA). Investigate potential hidden patterns, interactions between variables, or anomalies that may have been overlooked during the first phase of data preparation.

Future Insights and Further Analysis: If additional resources are available, explore incorporating external data, such as climatic conditions, socioeconomic factors, or seasonal water availability, which could provide critical context to improve model predictions.

Stakeholder Collaboration and Application: Develop visual dashboards to communicate model outputs to stakeholders (e.g., government agencies, NGOs, and community organizations) for practical resource allocation. Provide recommendations on proactive maintenance strategies based on predictive insights, focusing on high-risk regions or water point types identified by the models.

Iterative Model Improvement: Use feedback from stakeholders to refine models and predictions over time, ensuring continuous improvement in identifying wells requiring repair.

References

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/24/

[1]Ahmed, Ashraf A., Sakina Sayed, Antoifi Abdoulhalik, Salissou Moutari, and Lukumon Oyedele. "Applications of Machine Learning to Water Resources Management: A Review of Present Status and Future Opportunities." *Journal of Cleaner Production* 441 (February 2024): 140715. https://doi.org/10.1016/j.jclepro.2024.140715.