A Data-Driven Approach to Waterpoint Functionality and Accessibility

Project Description

- This initiative aims to forecast the operational condition of waterpoints in Tanzania, categorizing them as either functional or non-functional, by utilizing machine learning models.
- By examining elements like physical traits, installation specifics, geographical settings, and usage behaviors, the objective is to pinpoint the primary factors influencing water point functionality.
- The findings obtained will aid stakeholders such as the Tanzanian Ministry of Water, NGOs, and local communities in effectively distributing resources, minimizing repair and maintenance expenses, and improving access to clean water.

Business Understanding

In the fast-evolving water infrastructure sector of today, effective water management is essential for guaranteeing access to safe and dependable water sources.

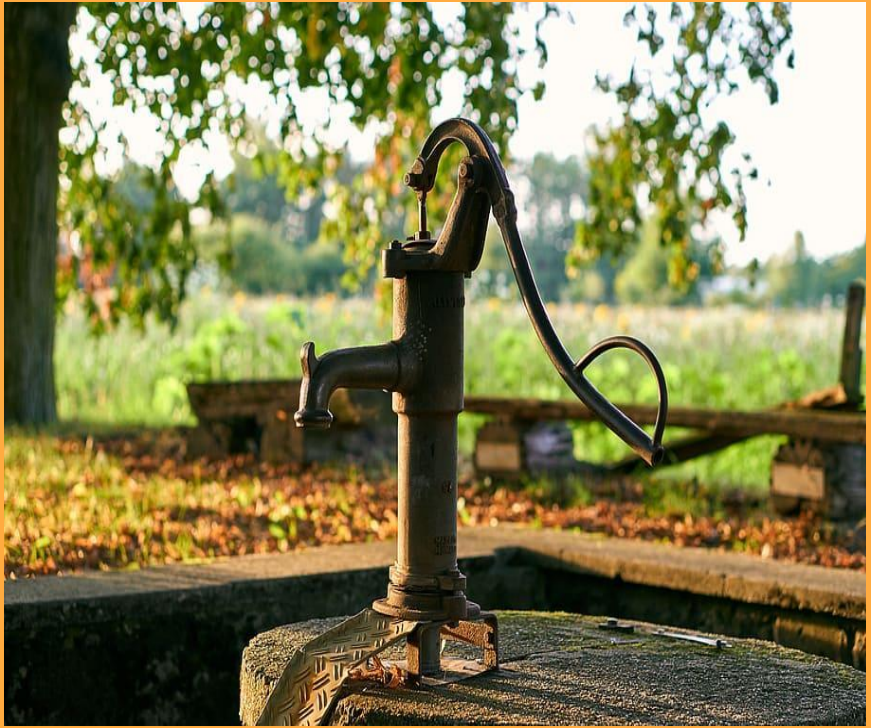
To tackle this issue, the Tanzanian Ministry of Water, together with NGOs and various stakeholders, is concentrating on forecasting the operational condition of waterpoints nationwide.

The project will examine data concerning water point functionality, encompassing elements like geographic location, installation specifics, usage behaviors, and physical characteristics.

Through the identification of essential determinants of waterpoint performance, the analysis seeks to reveal significant insights that will aid stakeholders in understanding which factors most influence waterpoint failures, how geographic and temporal changes affect functionality, and how to enhance resource distribution for repairs and maintenance.

Objective

1. Efficient Resource Allocation: Thus helping prioritize repairs and maintenance.
2. Improved Accessibility to Water:This will ensures functional water points for clean water, improving public health.
3. Cost Optimization: Prevents unnecessary repairs and optimizes resource allocation.
4. Sustainability: Identifies patterns to improve the durability of future waterpoints.

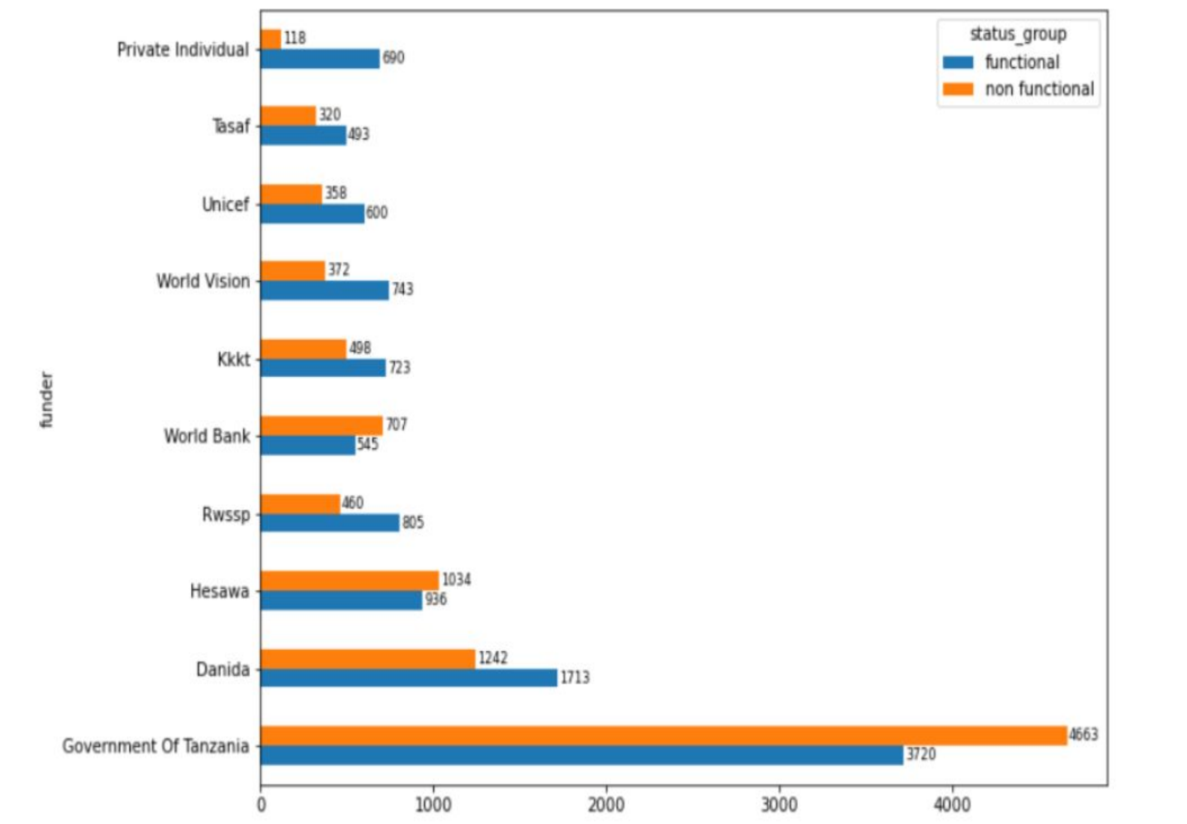


Data Preparation

Integrity of the data is always a very important step in any data analysis process. For this project the following techniques were employed to ensure accurate data:

- Check for missing values and imputing them
- Checking for duplicates
- Drop irrelevant columns
- Check for outliers and handling them in the data

Exploratory Data Analysis



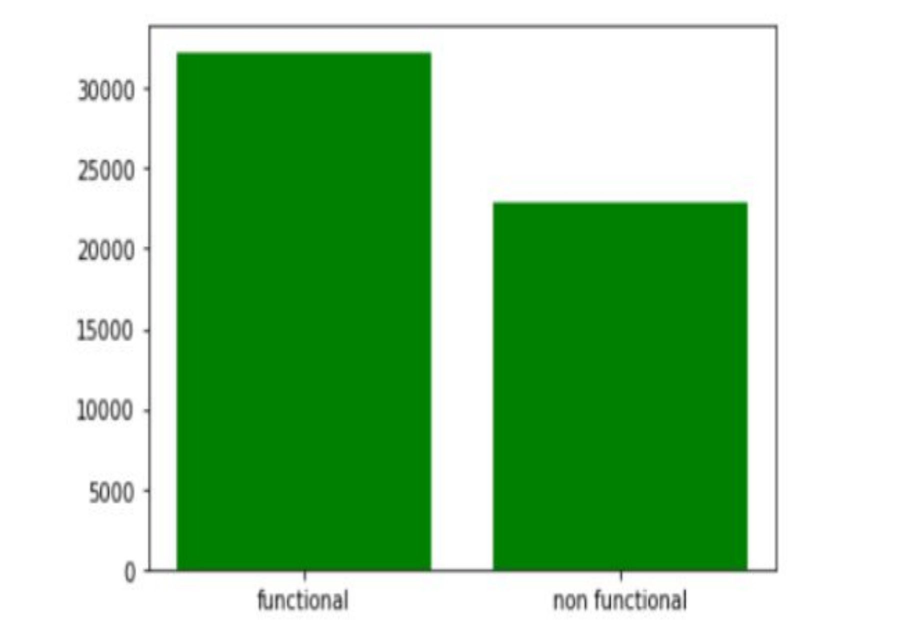
The chart displays the functionality status of water points across different funders.

- Funders like Government of Tanzania and HESAWA may require further analysis to understand why their non-functional water points are so high despite large investments.

- Private Individuals might benefit from technical or financial support to improve their success rate in maintaining functional water points.

- Focus on funders with a higher proportion of functional water points (e.g., DANIDA) could offer insights into best practices.

Modeling

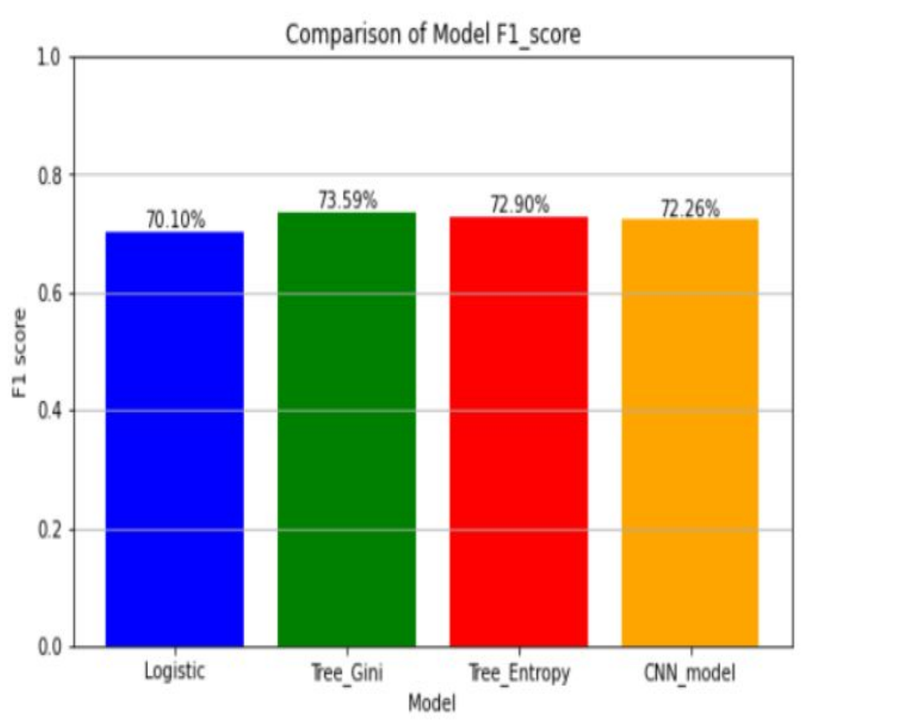


To address the imbalance in the data, we applied a technique called SMOTE-NC, designed to work effectively with datasets containing mostly categorical variables. We chose the F1 score as our primary metric to evaluate how well the model predicts each class, ensuring a balanced focus on both precision and recall. Additionally, we fine-tuned the model to achieve optimal performance while taking steps to prevent overfitting, ensuring reliable and accurate predictions.

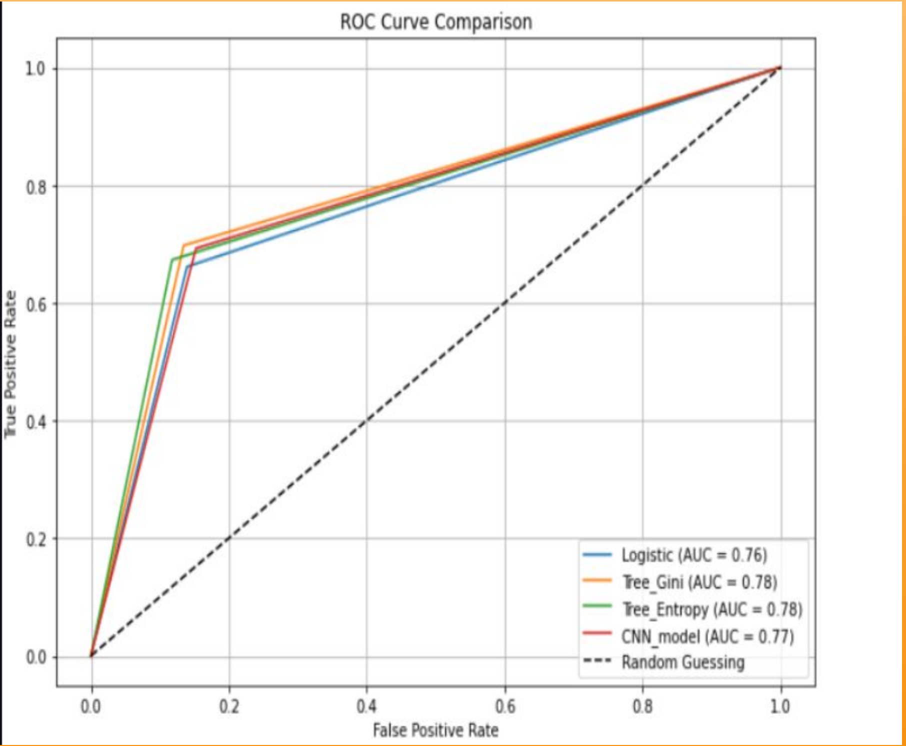
Model Comparison

The bar chart compares the F1 scores of four models used in the classification of water pump functionality: Logistic Regression, Decision Trees (Gini and Entropy), and a Convolutional Neural Network (CNN).

- The Decision Tree (Gini) model achieved the highest F1 score of 73.59%, slightly outperforming the Decision Tree (Entropy) model, which had an F1 score of 72.90%.
- The CNN model followed closely with an F1 score of 72.26%, demonstrating strong performance despite its complexity.
- The Logistic Regression model had the lowest F1 score at 70.10%, indicating it may not handle non-linear relationships as effectively as the other models.



Model Comparison



Models with higher AUC values are better at distinguishing between classes.

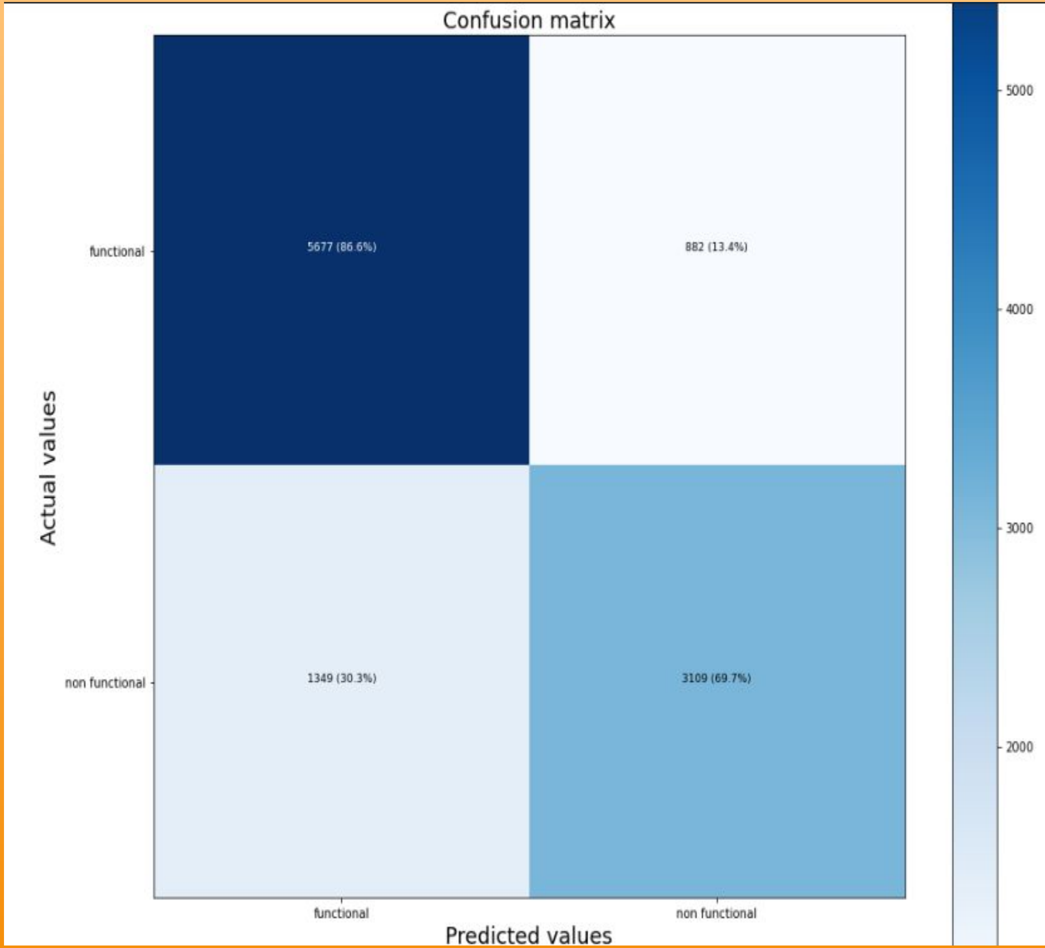
1. Tree_Gini and Tree_Entropy, with the highest AUC values (0.78), are the best-performing models.
2. Logistic regression is the weakest among the models but still acceptable for classification tasks.
3. The CNN model is competitive but slightly underperforms compared to the decision tree models.

Best Performing Model

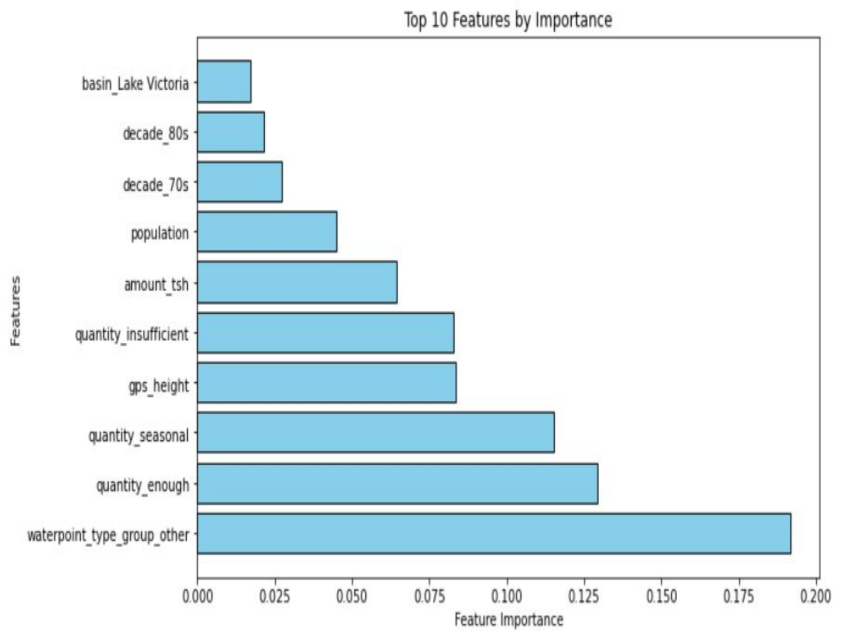
Interpretation:

Strengths: The model performs well in predicting functional items (high true negative rate, 86.6%).

Weaknesses: It struggles more with predicting non-functional items, as indicated by a lower recall (69.7%) and a significant false negative count (30.3%).



Important Features



- Water-related metrics quantity_enough, quantity_insufficient, quantity_seasonal, and amount_tsh) dominate as predictors, highlighting the importance of consistent water supply and flow.
- Geographic and demographic factors (gps_height, population, basin_Lake Victoria) suggest that environmental and community characteristics play a secondary role.
- The age of the infrastructure (decades) reflects the need for maintenance and modernization.

Thank You

QUESTIONS

Feel free to reach out with any questions.

📞 **TelePhone: 0715378370**
Gmail: savinskamau01@gmail.com

Linked Profile:
https://www.linkedin.com/in/nsavins/