WATER WELLS CLASSIFIER MODELS

1.1. BUSINESS OVERVIEW

In Tanzania, a country of more than 57 million people, there's a big problem with providing clean water to everyone. Many people still don't have access to clean water, which means about 61% of Tanzanians can't get safe water easily. This puts a lot of people at risk of getting sick from diseases carried by dirty water. The government, along with different groups and organizations, has been working hard to improve the systems that give people clean water. But there are still problems like old or broken water wells that stop the progress. Not having clean water affects everyone's health, especially children and older people.

Because of dirty water sources, diseases like diarrhea, cholera, and typhoid are still common. These diseases could be prevented if people had clean water to drink. But because they don't, sadly, people get sick and some even die. This is a big ongoing problem that needs attention to make sure everyone in Tanzania has access to clean, safe water.

1.2. PROBLEM STATEMENT

The Ministry of Water and the World Health Organization (WHO) in Tanzania are teaming up to make sure clean water is available for everyone in communities. Even though Tanzania has been trying hard to provide more clean water in recent years, about 31,000 people still die each year because there isn't enough clean water and good sanitation. Shockingly, more than 10% of these deaths could have been prevented.

To tackle this problem, a plan is in motion to look closely at where water wells and pumps are and how well they're working all across Tanzania. This investigation found that some pumps work but need fixing, while others don't work at all. This really affects how easy it is for people to get clean water.

I'm leading a team of data experts in this effort. Our main job is to find clear patterns that can help us predict if a water pump will work or not. Our goal is to figure out what things make a pump work well or stop working. By doing this, we can give the Ministry of Water and WHO important information. They can use this info to know when pumps need fixing and where to spend money to help broken pumps. This plan aims to make clean water easier to get, save lives by preventing deaths caused by not having enough clean water and good sanitation, and support Tanzania's goals for sustainable development.

1.3. OBJECTIVES

1.3.1. MAIN OBJECTIVES

Our primary objective is to develop a robust machine learning model capable of accurately predicting water pump performance. This model will utilize various attributes, including pump type, installation date, and well conditions, to forecast how well a water pump is expected to function.

1.3.2. SPECIFIC OBJECTIVES

A) Conduct statistical analysis and utilize feature importance techniques to pinpoint the key factors that significantly affect the functionality of water pumps.

B) Employ a range of machine learning algorithms, such as Random Forest, Gradient Boosting, and Logistic Regression, among others, to construct a predictive model for water pump performance.

C) Generate data-driven recommendations aimed at enhancing the accessibility of clean water sources based on the insights gathered from our analysis and predictive modeling.

2.DATA UNDERSTANDING

Our dataset encapsulates crucial information about water points across Tanzania, covering geographical coordinates, water source details, management practices, and operational status. The primary focus lies in the 'status_group' variable, indicating whether water points are Functional, Need Repair, or Non-Functional.

Exploring this dataset reveals a diverse range of features and thousands of water points distributed across regions. However, challenges like missing values and categorical data require careful preprocessing. Addressing imbalanced classes within 'status_group' is pivotal, and methods like SMOTE will aid in rectifying this imbalance.

Understanding this dataset isn't merely data exploration; it's about unraveling insights to predict factors impacting water point functionality. Our analysis aims to inform strategic interventions and resource allocation, ultimately enhancing water accessibility for all Tanzanians

3.MODELING

We have employed several machine learning models for specific reasons:

Firstly, Logistic Regression was chosen due to its simplicity and interpretability, making it suitable for binary classification tasks where we needed clear insights into the relationships between variables.

Secondly, Decision Tree Classifier was utilized to construct a tree-like model based on feature splits, allowing us to visually comprehend the decision-making process and handle both categorical and numerical data effectively.

Thirdly, K-Nearest Neighbors (KNN) was used for its simplicity and intuitive approach in classification, relying on the classes of nearby data points to classify new instances.

Fourthly, Random Forest was employed as an ensemble learning method comprising multiple decision trees, which helped in addressing overfitting issues and improving predictive performance in classification tasks.

Lastly, Gradient Boosting was chosen to iteratively build strong predictive models by enhancing the performance of weaker models sequentially, resulting in highly accurate predictions for our classification tasks. Each model served a specific purpose, leveraging its unique characteristics to cater to different aspects of our analysis and predictive requirements.

3.1. MODEL PERFORMANCE

In evaluating various machine learning models, each exhibited distinct performance nuances. Logistic Regression, while achieving a 65.5% accuracy, struggled notably in categorizing instances of one class, revealing limitations in precision and recall for specific categories. On the other hand, Decision Tree Classifier surpassed Logistic Regression with a 71% accuracy, particularly excelling in precision and recall for a previously challenging class. K-Nearest Neighbors (KNN) displayed balanced performance at 67.2% accuracy, maintaining moderate precision and recall across classes. Random Forest, despite achieving 67.9% accuracy after tuning, faced challenges in maintaining balanced performance metrics. However, Gradient Boosting emerged as the top performer, boasting 70.6% accuracy after tuning and showcasing the best balance between precision and recall. Consequently, Gradient Boosting stands out as the most promising model among the evaluated ones, demonstrating superior potential for effective classification tasks in this specific dataset.

4.EVALUATION

The project's main focus was reducing False Negatives and improving the True Positive Rate, particularly concerning mortality rates associated with inadequate water supply. Our primary goal was to achieve a recall and accuracy score exceeding 70% for the model. Successfully meeting this objective through Gradient Boosting solidified its suitability for predicting water pump functionality. Furthermore, evaluating the model's efficiency revealed a root mean squared error close to zero, emphasizing its remarkable effectiveness.

5.CONCLUSION

The model's performance, although indicative of progress, requires further enhancement. Despite achieving a 70% accuracy rate, it might not be deemed commendable in certain high-stakes scenarios where precision is paramount. Continuous training with updated data could substantially refine the model's predictions and address the existing dataset imbalance. This process is fundamental for enhancing the model's robustness and predictive capabilities.

6.RECCOMMENDATION

1. The WHO should collaborate with the government to efficiently mobilize funds to support the initiative.

2. Once the WHO initiates the program, they should consider engaging DWE to handle water pump installations.
3. Features like amount_tsh (water pump pressure) and water quantity serve as crucial indicators of water pump functionality. Utilizing these features will aid in determining the functionality of a water pump.
4. Lake Victoria, despite being one of the largest water bodies in the region, has a high number of non-functional wells. An exploration by the WHO in this region could reveal the underlying causes and enable the formulation of effective solutions.

It's noteworthy that a significant portion of the population resides away from wells, and about two-thirds have access to water supply. Additionally, the country's urbanization rate is growing at a rate of 0.7% annually. This suggests that many individuals might have shifted to using piped water instead of well water. If the WHO expands its initiative to include access to water supply in urban areas alongside the repair and construction of water pumps, it would significantly contribute to achieving their Sustainable Development Goals by 2030.

7.NEXT STEPS

Future improvements could focus on enhancing the model's predictive capabilities by exploring additional relevant features, such as weather or water quality data, refining ensemble methods, and conducting more exhaustive hyperparameter tuning. Augmenting the dataset, incorporating domain-specific insights, and implementing real-time monitoring for updates are crucial for model adaptability. Addressing class imbalance issues and collaborating with domain experts could further refine the model, contributing to improved accuracy and applicability in predicting water pump functionality and mitigating associated mortality rates linked to water supply inadequacies.