TANZANIAN WATER WELLS

Introduction

Recent data from the World Bank show that Tanzania has a population of about 60 million. According to Nsemwa (2022) many Tanzanians continue to struggle with insufficient or limited access to clean and safe water. Only 30.6% of Tanzanian households use recommended water treatment methods, and only 22.8% have adequate hand-washing facilities (Ministry of Health report, 2019). Poor sanitation is estimated to cause 432,000 diarrhea-related deaths per year and is a major contributor to several Neglected Tropical Diseases (NTDs) such as intestinal worms, schistosomiasis, and trachoma. Malnutrition is also made worse by poor sanitation (WHO, 2019).

Challenges

Tanzania, struggles with providing clean water to its population of over 57,000,000. There are several waterpoints already established in the country. Some of the waterpoints are Functional,Others need repair while others are completely non-functional

Problem Statement

An NGO is curious to know the state of the different waterpoints so that they may Help where needed

Data Understanding

The data has a 59,400 rows and 41 columns. Of those 41 columns, 10 were numeric columns and the remaining 31 were string columns; known as object in pandas. The observations are that the columns are very repetitive and some are irrelevant which led to dropping of most of them:

* extraction_type,extraction_type_class and extraction_type_group give same info

* management is the same as scheme_management
* payment and payment_type are the exact same
* water_quality and quality_group are the same ,
* Quantity and quantity_group are exact same
* source type and source are similar
* waterpoint_type and waterpoint_type_group are the same

* I'll also drop wpt_name,funder,permit,recorded_by i considered them not imporant and had missing values
scheme_name had alot of missing values

Data Cleaning

The data cleaning process wasnt very hectic since only 7 out of 40 columns had missing values. The data had a massive amount of outliers which I had to fix,especially in the height,amount and population column

Data Analysis

There are more functional water-points compared to the non_functional and those that required repair.

During the Analysis, here's what I concluded:

1)Gravity,nira/tanira are the most used extraction type

2)There are more functional wells than non functional. Though non-functional is a big number.The wells that need repair are much less

3)Most wells have good water compared to the other types

4)communal standpipe and hand pump are the most types of waterpoints

5)Most water sources are shallow wells

6)Functional waterpoints are mostly located at high altitudes

Modeling

I used Four models : LogisticRegression model,Kneighbors classifier model,Random Forest classifier model, and Gradient boosting classifier model and the results were:

LOGISTIC REGRESSION

First Model Training Accuracy: 61.62% Validation accuracy: 59.49% Iterated Model Training Accuracy: 61.62% Validation accuracy: 59.5% KNEAREST NEIGHBORS

First Model Train accuracy: 81.39481243840557 Test accuracy: 69.01366826745475 Iterated Model Train Accuracy: 60.659243882493556 Test Accuracy: 59.41632803841891

RANDOM FOREST CLASSIFIER First Model

Training Accuracy: 89.9% Validation accuracy: 72.21% Iterated Model Training Accuracy: 49.91% Validation accuracy: 70.92% GRADIENT BOOSTING CLASSIFIER

First Model Train accuracy: 70.43257637709421 Test accuracy: 65.68895456224602 Iterated Model Train accuracy: 89.75697720743956 Test accuracy: 73.06243073513114

Conclusion

GradientBoostingClassifier (iterated model)was my bet model. It has the highest test accuracy of 72.6%, indicating better performance on unseen data compared to the other models.The RandomForestClassifier first model is also a good one .This means that the Models are able to better predict the functionality of Tanzanian Water Wells better than the other models. Hence for my stakeholders i would advice the use of the model