

 CindyKN /  
Phase-3-Project




 Code  Issues  Pull requests  Actions  Projects  Wiki  Security  Insights  Set




☆ 0 stars    0 forks    1 watching    Activity

 Public repository


 main ▾



 Branches    Tags



CindyKN Final Documents ...

15 minutes ago 

[View code](#)

 README.md



## Tanzania Water Wells Conditions Project

### Project Overview/ Background

Tanzania, as a developing country with a population of over 57 million, faces significant challenges in providing clean and accessible water to its citizens. The country has established numerous water points; however, many of these water wells are in need of repair, and some have ceased functioning altogether. Access to clean and reliable water sources is vital for public health and economic development.

### Business Problem

The problem at hand pertains to a classification task focused on categorizing well conditions can be classified into three categories:

1. Functional - The well is in good working condition, currently in operational and good working condition, efficiently providing clean and reliable water to the local population.
2. Functional needs repair - The well is operational but requires maintenance or repair to ensure sustained functionality.
3. Non-functional - The well has stopped working, rendering them unproductive and incapable of providing water.

### Data Understanding

An in-depth analysis of a dataset sourced from the Taarifa (<http://taarifa.org/> (<http://taarifa.org/>)) open-source platform, the Tanzanian Water Ministry (<http://maji.go.tz/> (<http://maji.go.tz/>)), and publicly disseminated by Driven Data (<https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-watertable/page/23/> (<https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-watertable/page/23/>)). This dataset comprises extensive information pertaining to approximately 60,000 water pumps distributed throughout Tanzania. It encompasses a comprehensive set of 39 distinct attributes, encompassing details related to the pumps' geographical locations, management aspects, and technical specifications. Our analytical endeavor encompassed an exhaustive examination of 59,400 individual data points, incorporating a total of 39 attributes (columns) associated with water pumps situated in the Tanzanian context. Datasets have been compiled organized as follows:

1. Test Set Values: This portion comprises the independent variables for which predictions are required. It contains the data for which the model's predictions will be generated.
2. Training Set Labels: This component contains the dependent variable, namely "status\_group," which represents the target variable or the desired outcome for each corresponding row in the "Training Set Values." It essentially provides the ground truth for the model to learn from during training.
3. Training Set Values: This segment encompasses the independent variables associated with the training set. It contains the features and attributes that the machine learning model will utilize for training and predicting the "status\_group" variable.



## Modeling [↗](#)

We conducted tests on various models, and based on the accuracy score, we opted for the One-hot Coding classifier. Here's a breakdown of its performance:

- Precision: Precision gauges the proportion of accurate predictions relative to the total predicted instances for each class, with higher precision signifying fewer false positives. The model demonstrates good precision across all classes.
- Recall: Recall evaluates the proportion of accurate predictions relative to the total actual instances for each class, with higher recall suggesting fewer false negatives. The model exhibits reasonably good recall for all classes.
- F1-score: The F1-score, a balance between precision and recall, is the harmonic mean of these two metrics. The F1-scores for all classes are relatively high, indicating strong overall performance.

In summary, the model achieved an accuracy of 80.74%, meaning it accurately predicted the class for approximately 80.79% of instances in the validation set. The macro-averaged F1-score also stands at 80.79%, reflecting consistent performance across all classes. Additionally, the weighted average F1-score is 80.74%, taking into account the class imbalance in the dataset.



## Evaluation [↗](#)

Some conclusions made include:

Handling large and intricate datasets, such as the Tanzania wells dataset with 59,400 observations, can pose challenges when employing conventional data analysis methods like Excel. Machine Learning (ML) proves highly effective in such scenarios, as it can uncover intricate patterns and relationships that may elude simpler data analysis techniques. The Tanzania well dataset encompasses both categorical and numerical variables, rendering ML algorithms invaluable for revealing patterns and connections. For instance, ML algorithms can discern that wells situated in sparsely populated regions have limited use and require maintenance. This kind of insight plays a pivotal role in identifying wells in need of attention. Before applying ML algorithms, data cleaning assumes critical importance to ensure data quality, especially in the context of extensive datasets. In the case of the Tanzania well dataset, superfluous variables like latitude, longitude, and district codes were removed to streamline the model.

## Conclusion [↗](#)

Based on the information provided some recommendations include:

1. **Class Imbalance Handling:** Address the class imbalance issue by employing techniques such as oversampling or undersampling to rebalance the class distribution in the training data. This will help improve the model's ability to predict the minority classes, "functional needs repair" and "non-functional."
2. **Feature Engineering:** Invest more effort in feature engineering. Explore the creation of new features or transformations to enhance the model's predictive power. Additionally, consider incorporating external data sources that could provide valuable information for the problem at hand.
3. **Algorithm Selection:** Experiment with different classification algorithms beyond logistic regression. Algorithms like random forests, gradient boosting, or support vector machines may be more suitable for this specific dataset. Each algorithm has unique strengths and weaknesses, and alternative algorithms might yield better results.
4. **Hyperparameter Tuning:** Optimize the hyperparameters of the chosen algorithm(s) to enhance their performance. Utilize techniques such as grid search or random search to systematically explore various combinations of hyperparameters and identify the best configuration for the model. This will improve the model's generalization and prediction accuracy

## Releases

No releases published

[Create a new release](#)

## Packages

No packages published

[Publish your first package](#)

## Languages

● Jupyter Notebook 100.0%