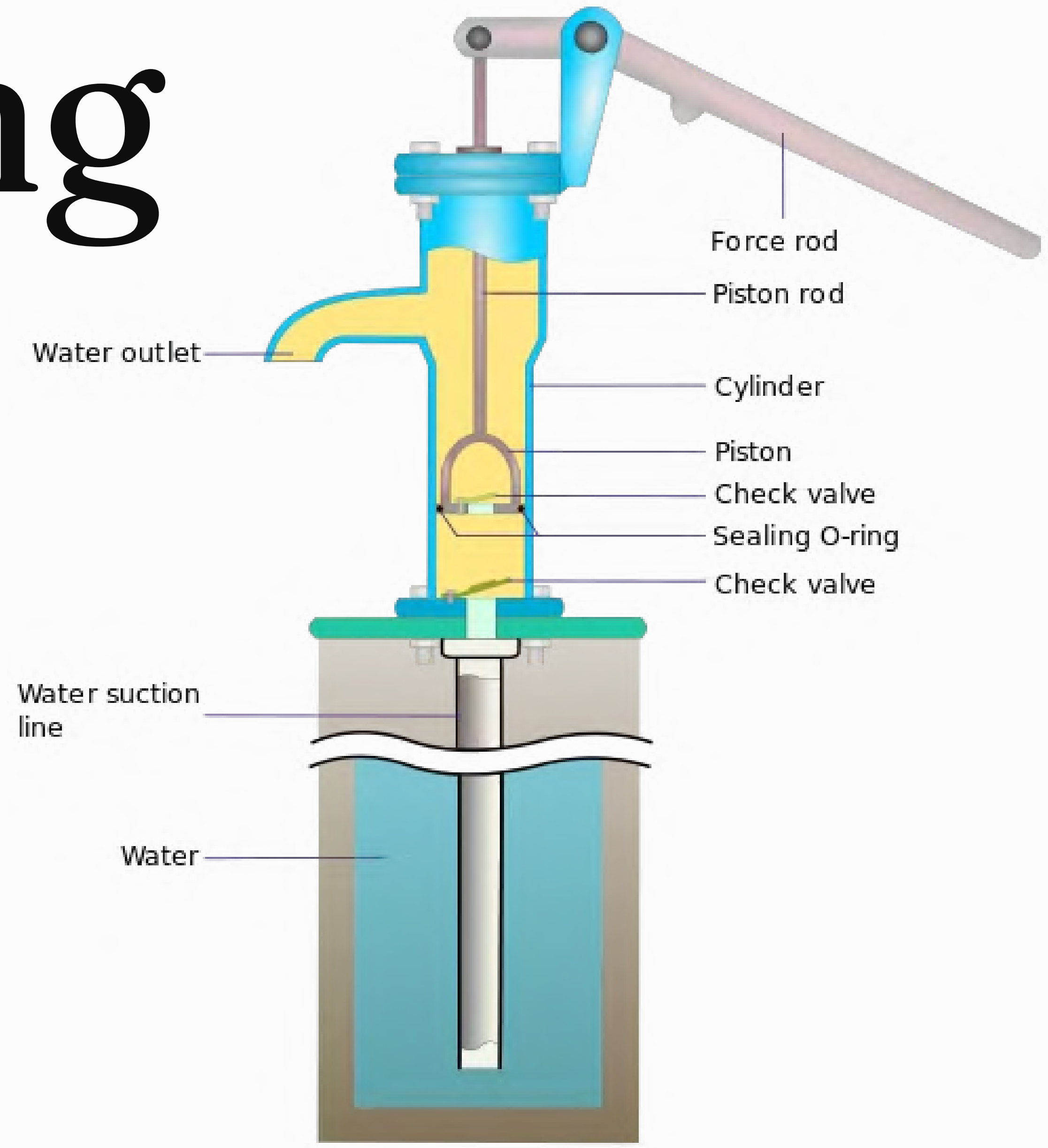
Team 37

AVISHKAR RAWAL, MANYA MITTAL, NIKITA JADHAV

16 MAY 2025

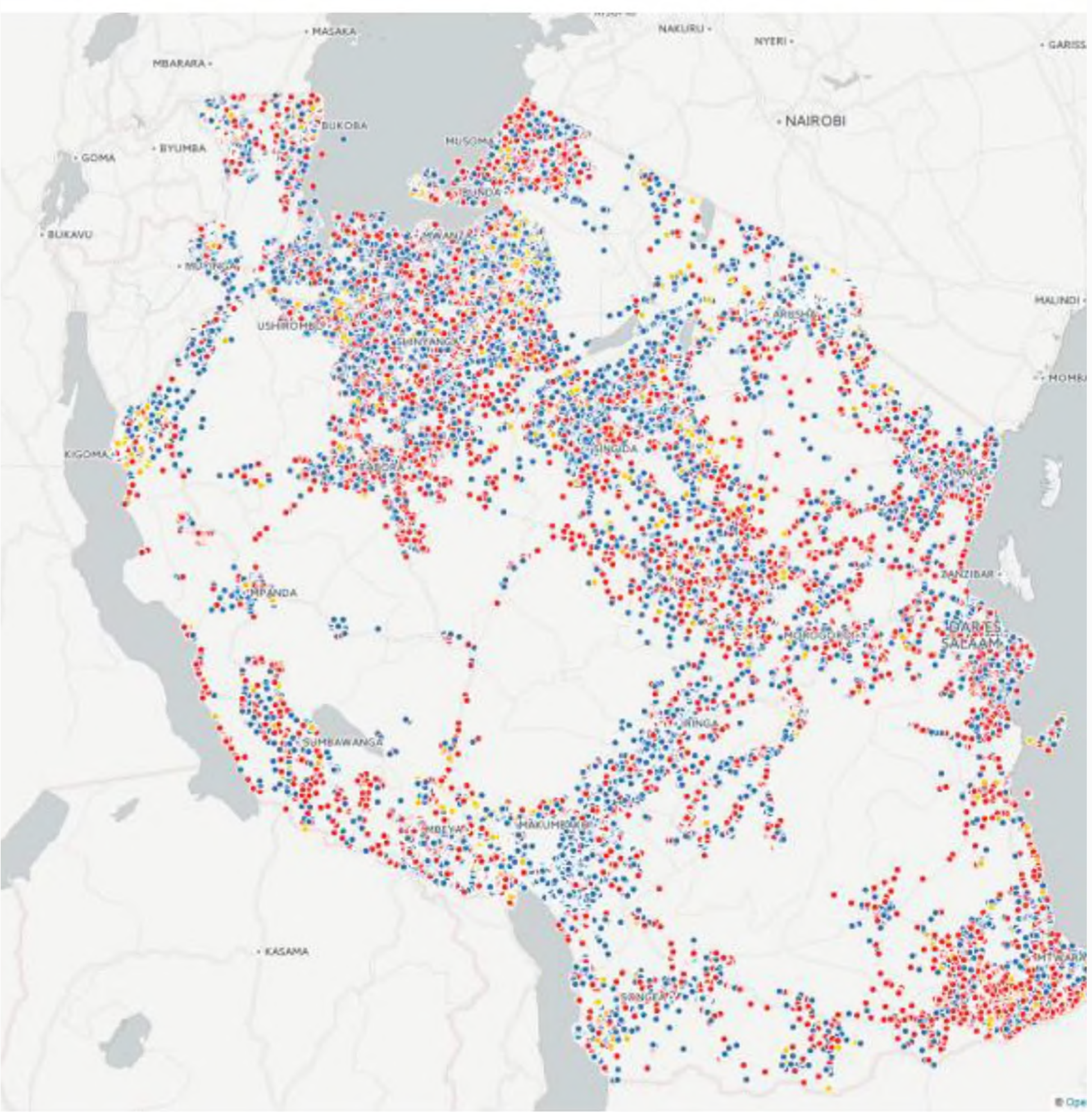
Pump it Up: Data Mining the Water Table

DRIVENDATA



The Problem

- **Access to clean water = Human Right**
- **SDG 6**: Ensure availability & sustainable management of water
- **Tanzania ranks 130/166** on the SDG index (World Bank Group, 2018)
- 21 million people, lack access to improved drinking water (Joseph et al., 2018)
- **29% of waterpoints fail shortly after installation (**Joseph et al., 2019)
- Current maintenance = **Reactive**, not **Preventive**
- **AI + ML → Predictive Maintenance**



THE CURRENT WATER DASHBOARD – FUNCTIONALITY STATUS OF ALL RURAL WATER POINTS—FUNCTIONAL (**BLUE**), NON FUNCTIONAL (**RED**), FUNCTIONAL NEEDS REPAIR (**YELLOW**)

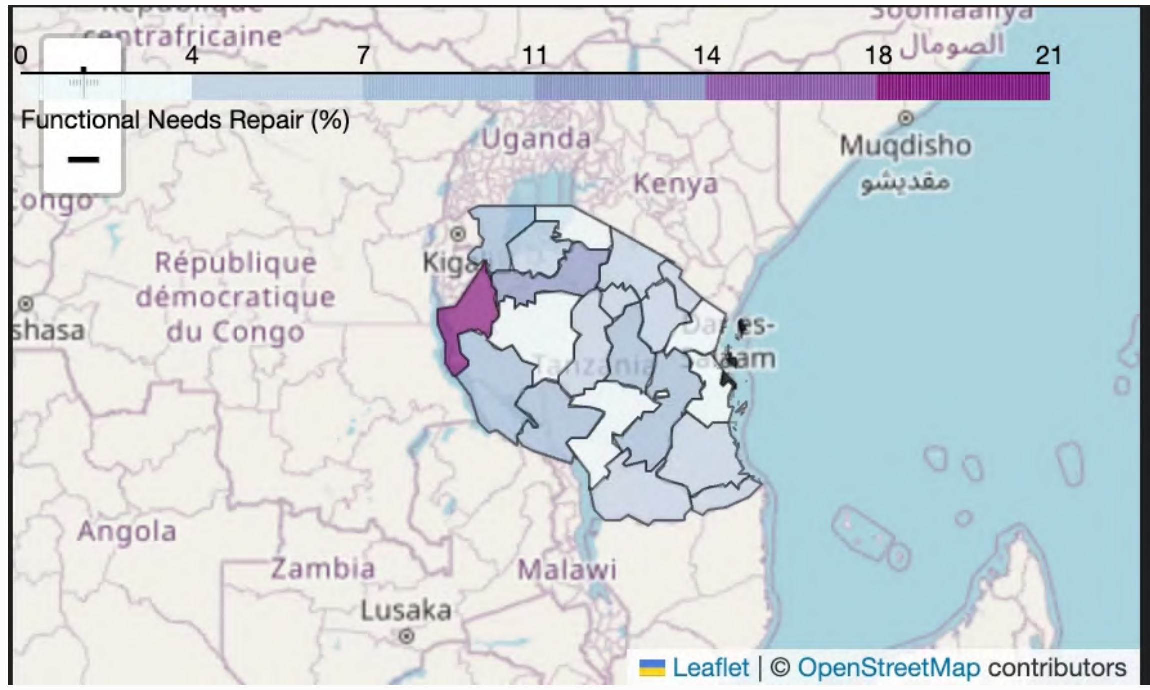
(KATOMERO ET AL., 2017)

THE PROBLEM STATEMENT

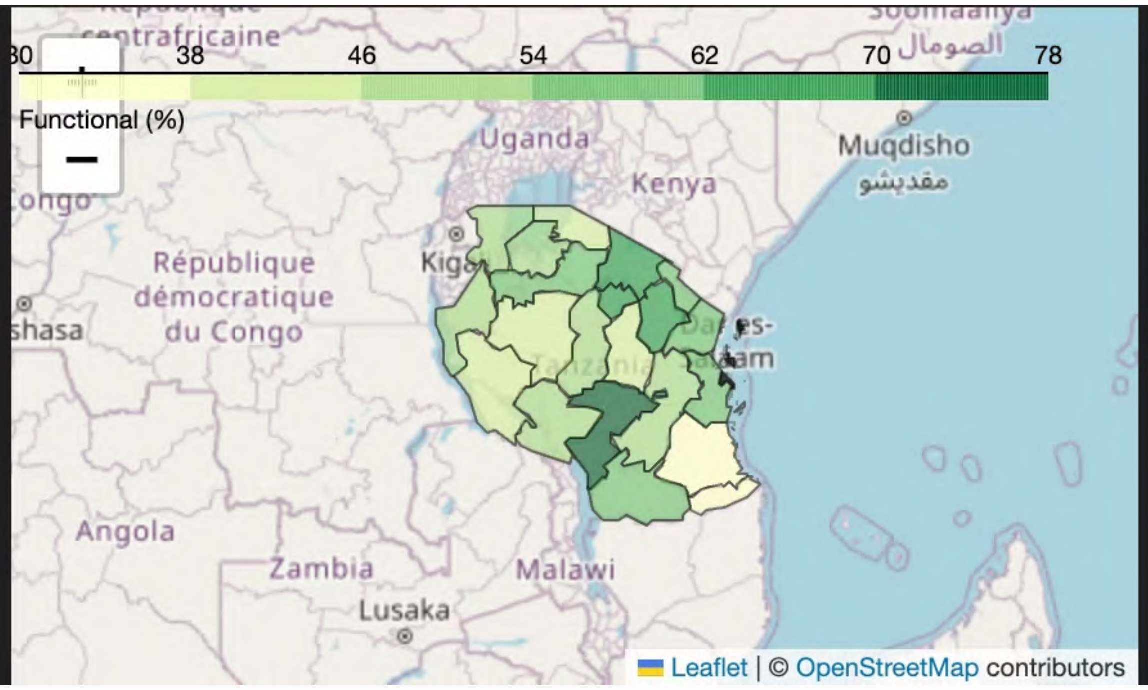
Can we predict the functionality of water points in rural Tanzania?

Why This Matters

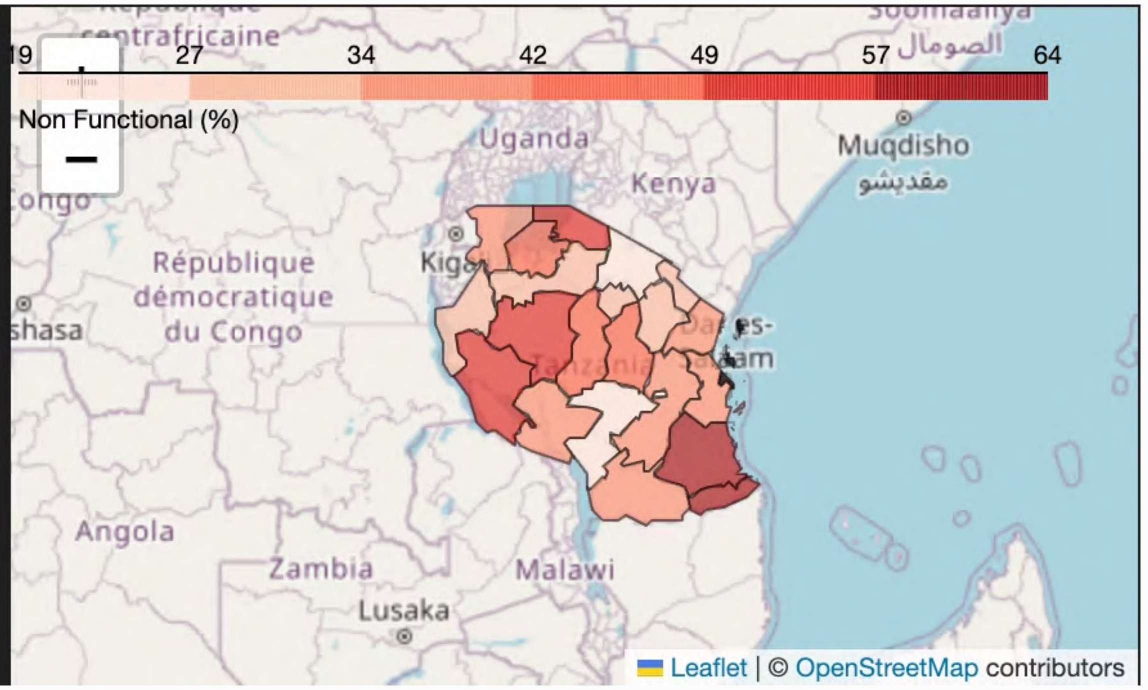
- **Tanzania faces critical water scarcity**, especially in rural regions.
- Thousands of water pumps exist, but **many are non-functional** due to poor maintenance and delayed repair.
- Manual monitoring is **expensive and inefficient**.



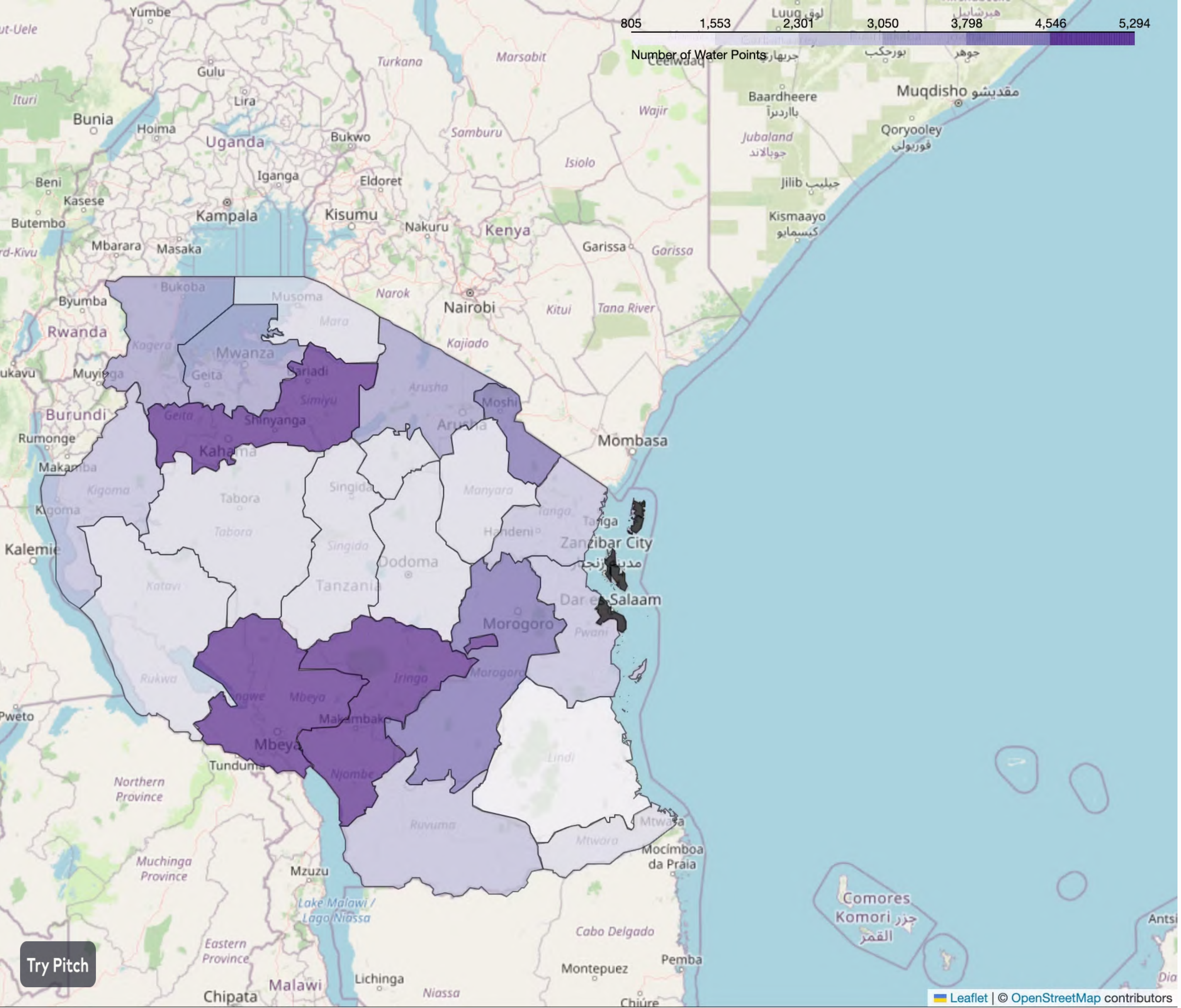
Functional but needs repair



Functional



Non Functional



Try Pitch

Potential Impact

- **Early identification** of failing or likely-to-fail waterpoints.
- Enables **timely intervention and maintenance**.
- Reduces **downtime** of essential water sources.
- Improves **access to safe drinking water** for rural communities.
- Supports **data-driven governance** and resource allocation.

Literature Review pt.1 - The historical overview

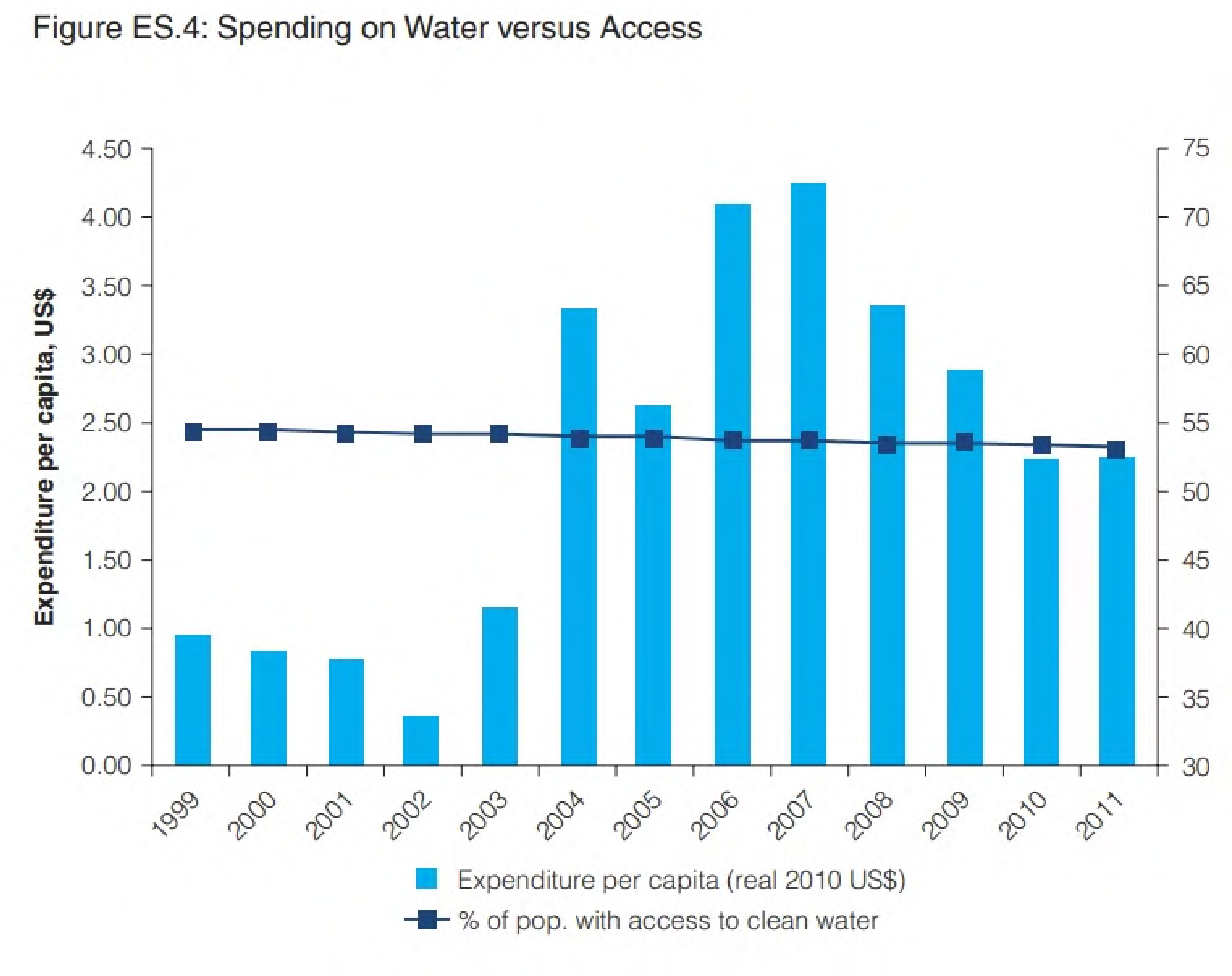
2002
NATIONAL WATER POLICY

2010-2013
WATER POINT MAPPING (WPM)

2005-2025
WATER SECTOR DEVELOPMENT PROGRAMME WSDP (20-YEAR PLAN)

2014-2017
SEMA PROJECT

1. Tanzanian government introduced the National Water Policy in 2002, emphasizing community ownership and management to enhance sustainability in rural water supply projects
2. This policy led to the establishment of Community Owned Water Supply Organizations (COWSOs), aiming to empower communities to select appropriate technologies, finance infrastructure, and maintain water points independently.
3. Despite this initiative, implementation has been limited. Only about 30% of Tanzania's 12,319 villages have active COWSOs, leaving the majority without structured management.
4. The Water Sector Development Programme (WSDP), intended to support these efforts, allocated approximately 80% of its funding to constructing new water points, with minimal investment in community training and maintenance of existing systems
5. This approach has led to a cycle of building and neglect, where new installations quickly fall into disrepair due to inadequate management and maintenance structures
6. The **SEMA (Sensors, Empowerment, and Accountability)** platform empowers COWSOs to report pump statuses via mobile tools, marking an early shift to **digital monitoring**.
(Source: Katomero et al., 2017)



(Joseph et al., 2018)

DrivenData

SORT OF LIKE A BREAKTHROUGH –
CROWDSOURCING THE PROBLEM

THE DATA

The data for this comeption comes from the Taarifa waterpoints dashboard, which aggregates data from the Tanzania Ministry of Water.

(DrivenData, n.d.)

59,400
DATAPOINTS

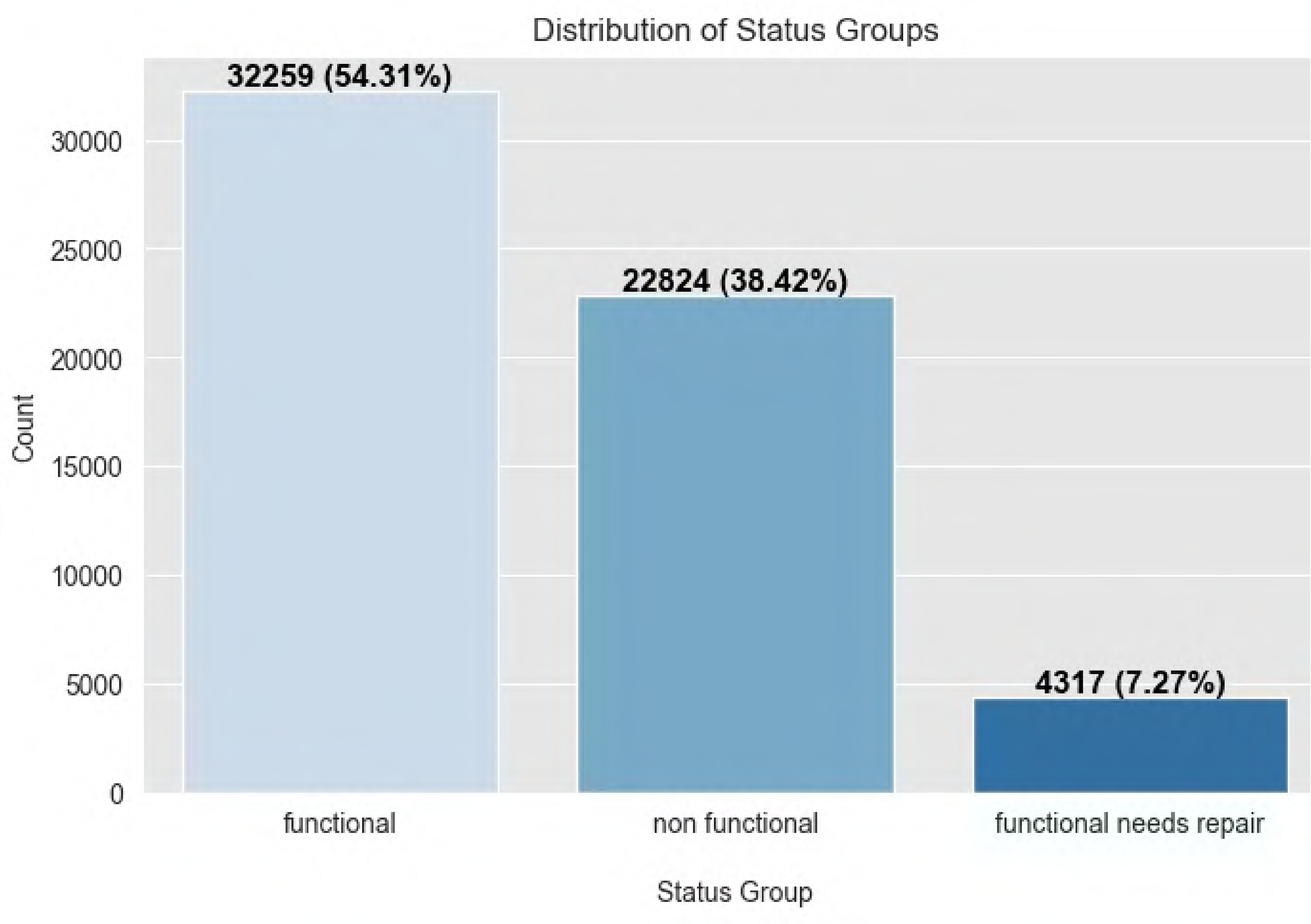
40 COLOUMNS
30 CATEGORICAL
10 CONTINUOUS

TARGET VARIABLE

STATUS_GROUP

{FUNCTIONAL,
NON-FUCNTIONAL,
FUNCTIONAL BUT
NEEDS REPAIR}

Location of the well	- latitude - longitute - region - district - Lga - ward
Building of the well	- Funder - Installer - whether there was a public planning meeting or not - not whether it was built with a permit or not - the year of construction
Operation of the well	- the management and its structure, - the price charged - the amount of water, - the quality of the water - the source of the water - how the water is extracted



Try Pitch

EDA

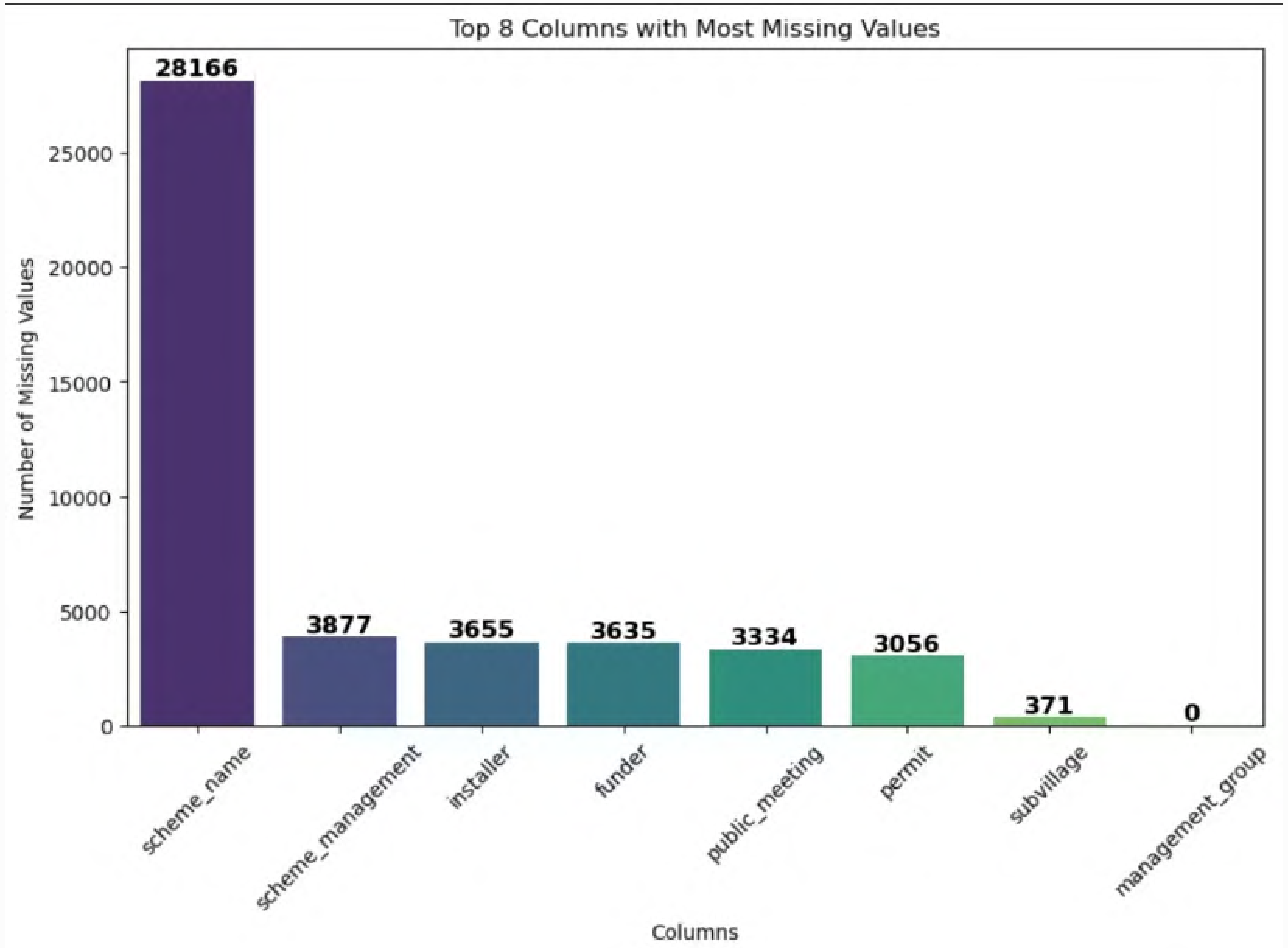
Some observations right off the bat

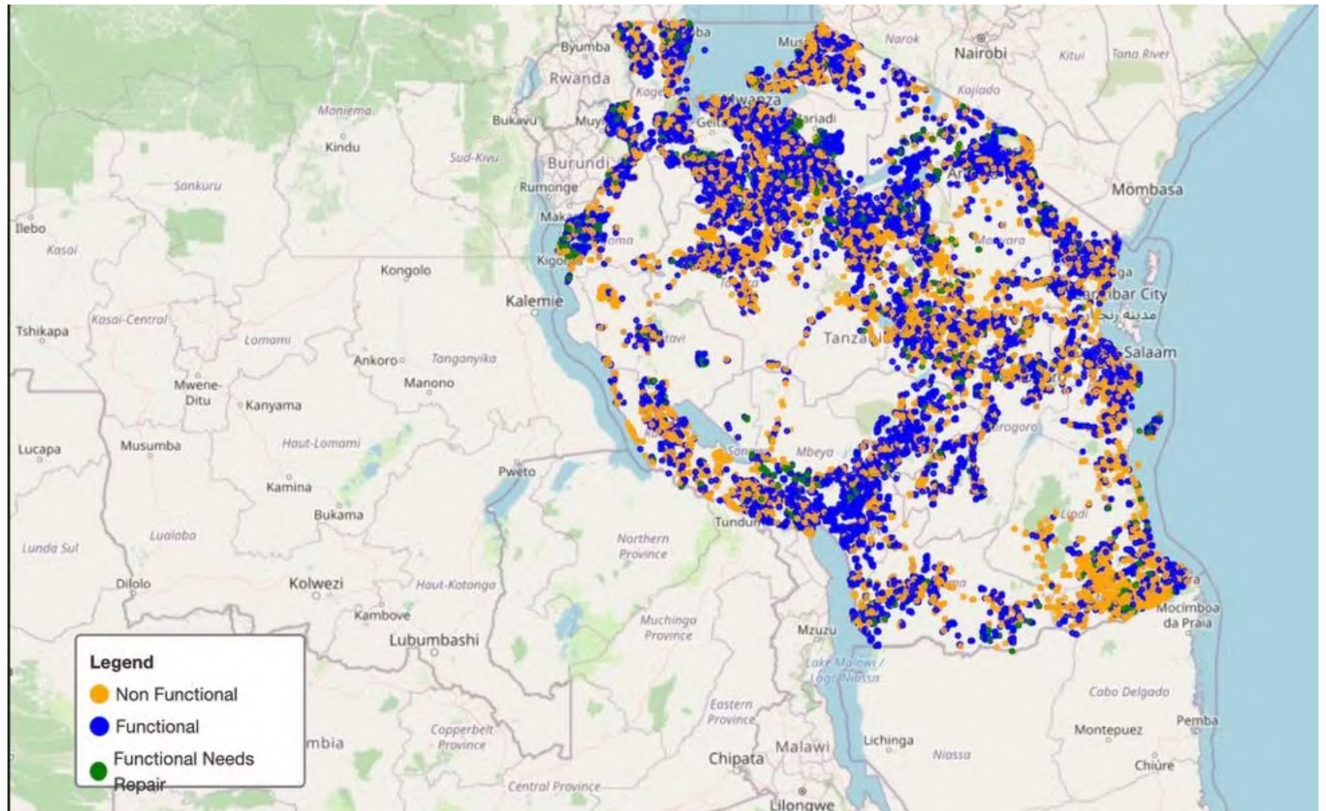
1. There is a class imbalance in the target variable - status_group (54% function, 38% non functional, 7% functional needs repair)
2. There are no missing values in the continuous columns
3. There are missing values in categorical columns
4. There are columns with duplicate or repetitive information

{quality, quality_group}, {quantity, quantity_group}, {source_type, source}, {waterpoint_type_group, waterpoint_type}, {payment, payment_type},

5. recorded_by column has one unique value - it is common for all rows
6. date_recorded coloumn shows that all the data is recorded in the span of just two years

```
quality Group: ['good' 'salty' 'milky' 'unknown' 'fluoride' 'colored']
Water_quality: ['soft' 'salty' 'milky' 'unknown' 'fluoride' 'coloured' 'salty abandoned'
 'fluoride abandoned']
Quality Group        Water_quality
--------------------------------------------------
good                 ['soft']
salty                ['salty', 'salty abandoned']
milky                ['milky']
unknown              ['unknown']
fluoride             ['fluoride', 'fluoride abandoned']
colored              ['coloured']
```





DATA PREPROCESSING

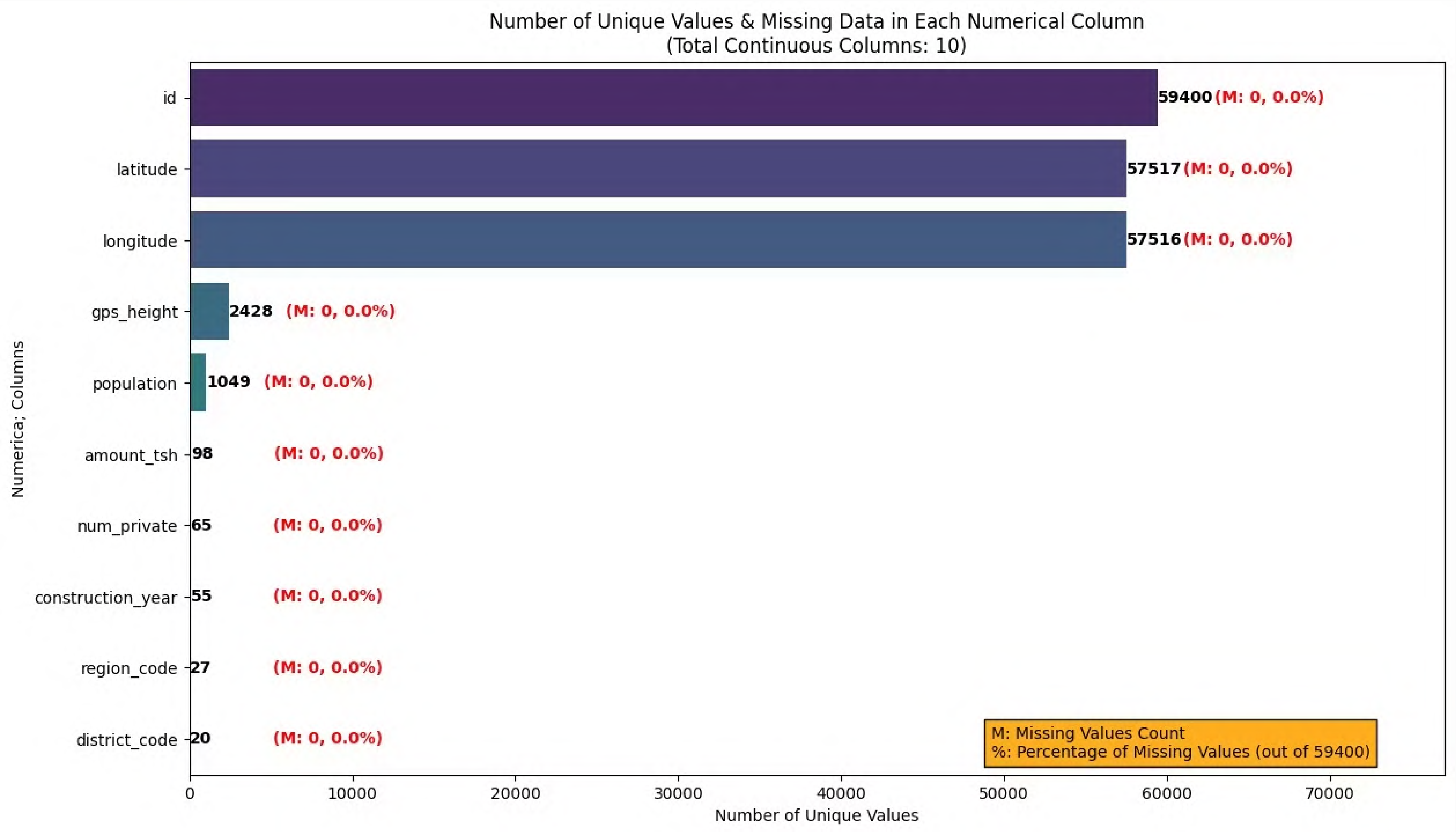
Addressing missing, incorrect, and uninformative data

Dealing with continuous variables

There are no missing values in the continuous variable

Problem solved !

NO...



A MEDIUM article showed that missing values for numerical columns were representation by zeros ((BrendaLoznik, n.d.)

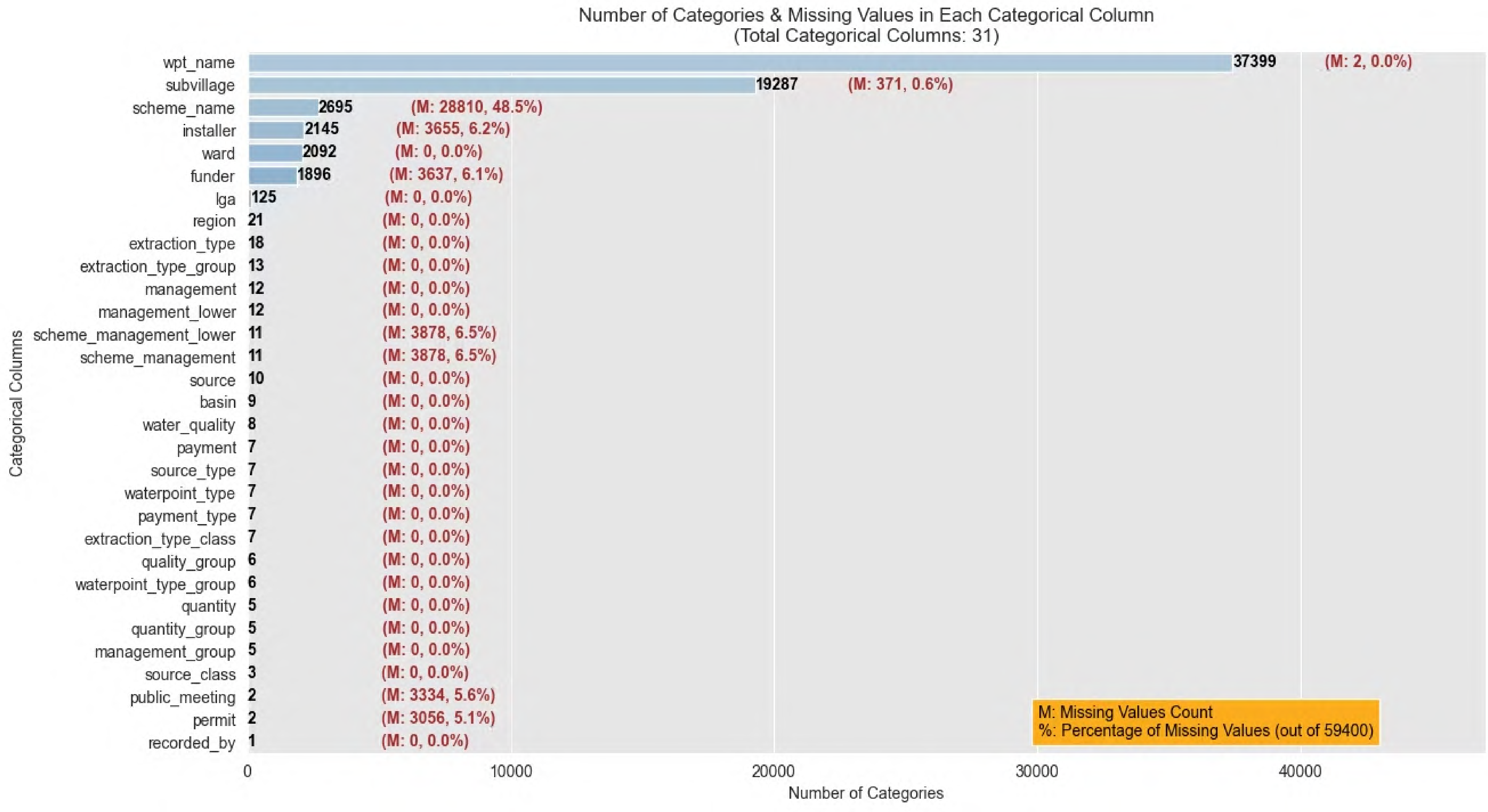
She imputed the values but she reported -

" I have performed three imputation strategies and assessed their performance using a random forest classifier. Despite the fact that the three imputation strategies had very different effects on the distribution of certain variables, their **resulting model performance is very comparable.** Enough variance was left for the model to learn patterns in the data. "

Dealing with categorical variables

Handling the missing values

The categorical column scheme_name was dropped. For other categorical columns, missing values were replaced with "other". Continuous columns had no missing values, but some had inval zeros — for example, construction_year had **43.9**% zeros, and longitude had **1,812** zeros, which are not valid for Tanzania. We used a **KNN imputer** to replace these zeros with more suitable values.



Encoding the categorical variables

Three primary encoding strategies were assessed for: Label Encoding, Frequency Encoding, and Label Encoding with Rare Category Grouping. Each encoding method was applied prior to training a baseline classification model. Accuracy using each encoding strategy was test on our final chosen classification model

Model / Encoding Method	Accuracy	Description of the method	Notable Strength
Label Encoding + RareCategory	0.8106	All categories except the top 20 are assigned a single numeric label	Strong general performance
Frequency Encoding	0.8113	encoded based on their frequencies.	Slightly improved consistency
Label Encoding	0.8125	assigns a unique integer to each category	Comparable to other encodings, but best

Dealing with class imbalance via SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a technique used to balance imbalanced datasets, especially in classification tasks.

```
🔹 Class distribution BEFORE SMOTE:
 functional: 25807
 non functional: 18259
 functional needs repair: 3454

SMOTE applied
🔹 Class distribution AFTER SMOTE:
 functional: 25807
 non functional: 25807
 functional needs repair: 25807

Trained on 77421 samples
Accuracy: 0.8088

Classification Report:
                         precision    recall  f1-score   support

             functional       0.82      0.88      0.85      6452
functional needs repair       0.49      0.34      0.40       863
         non functional       0.84      0.79      0.82      4565

               accuracy                           0.81     11880
              macro avg       0.72      0.67      0.69     11880
           weighted avg       0.80      0.81      0.80     11880
```

with Smote : 0.8080

```
 df[col].fillna("unknown", inplace=True)

🔹 Class distribution BEFORE SMOTE:
 functional: 25807
 non functional: 18259
 functional needs repair: 3454

SMOTE not applied. Using original training data.

Trained on 47520 samples
Accuracy: 0.8125

Classification Report:
                         precision    recall  f1-score   support

             functional       0.81      0.90      0.85      6452
functional needs repair       0.55      0.31      0.40       863
         non functional       0.85      0.78      0.82      4565

               accuracy                           0.81     11880
              macro avg       0.74      0.66      0.69     11880
           weighted avg       0.81      0.81      0.80     11880
```

without Smote : 0.8125

Making the class balanced might have created redundant data points, which might have caused the model to overfit.

FEATURE ENGINEERING

Introducing domain knowledge into the dataset

New features

- **Raininess Score:**
  Derived from seasonal indicators (month)
  helped capture rainfall patterns affecting water supply.

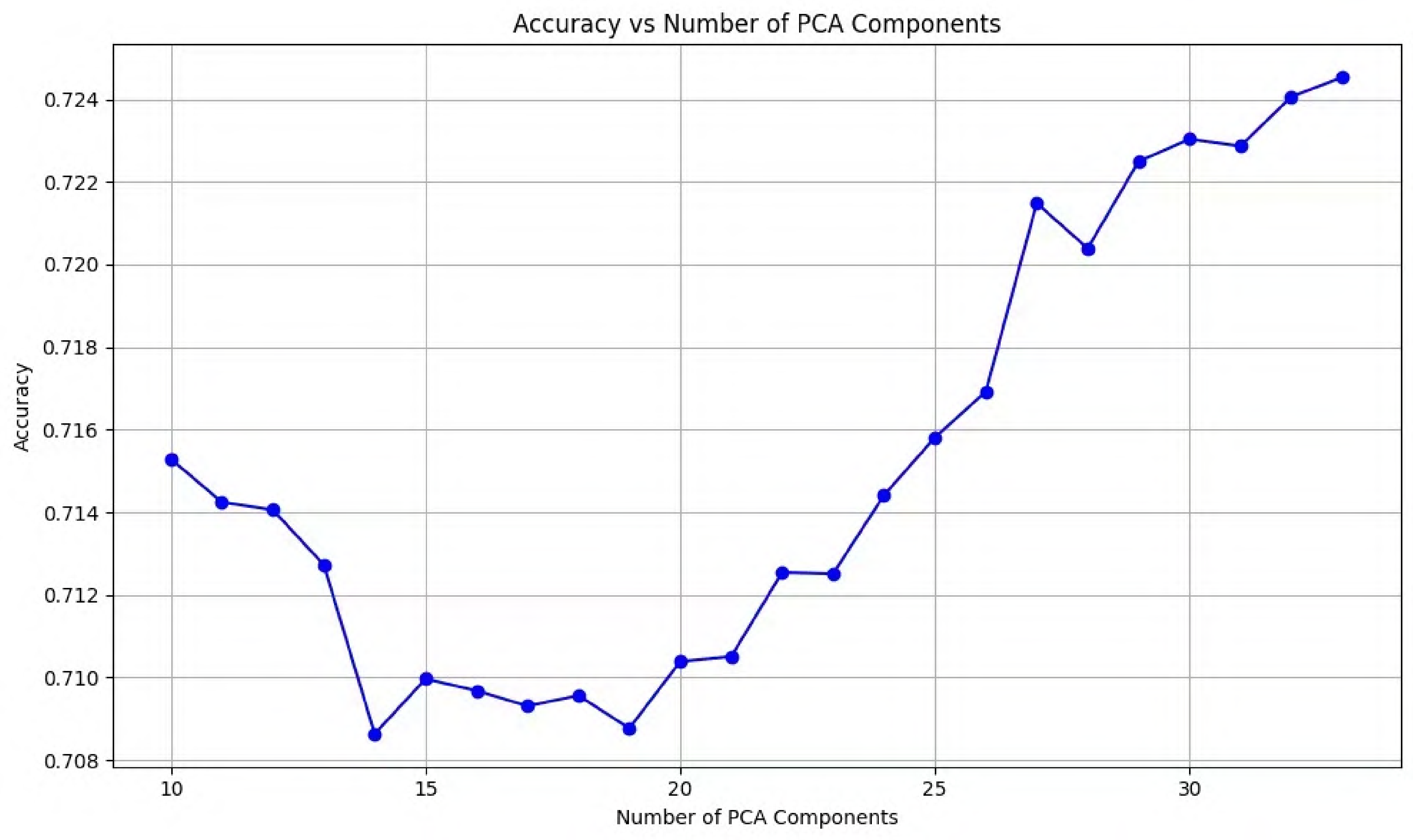
- **Pump Age:**
  Age = recorded_date - construction_year
  Helped identify age-related failures

- **Combined Regional Codes:**
  district_code & region_ode merged to capture location-specific behaviors.

- **Population:**
  Missing or zero values imputed using **KNN** based on similar categorical and numeric features

PCA FOR DIMENSIONALITY REDUCTION

- PCA is used to **combine highly correlated features**, it retains **maximum variance** with fewer dimensions.
- Our **final model was a Random Forest classifier**, which has an **intrinsic feature selection** capability
- **Model accuracy dropped** compared to original feature set.
- Dataset didn't have a very high number of features — PCA **not significantly beneficial**.



FEATURE SELECTION

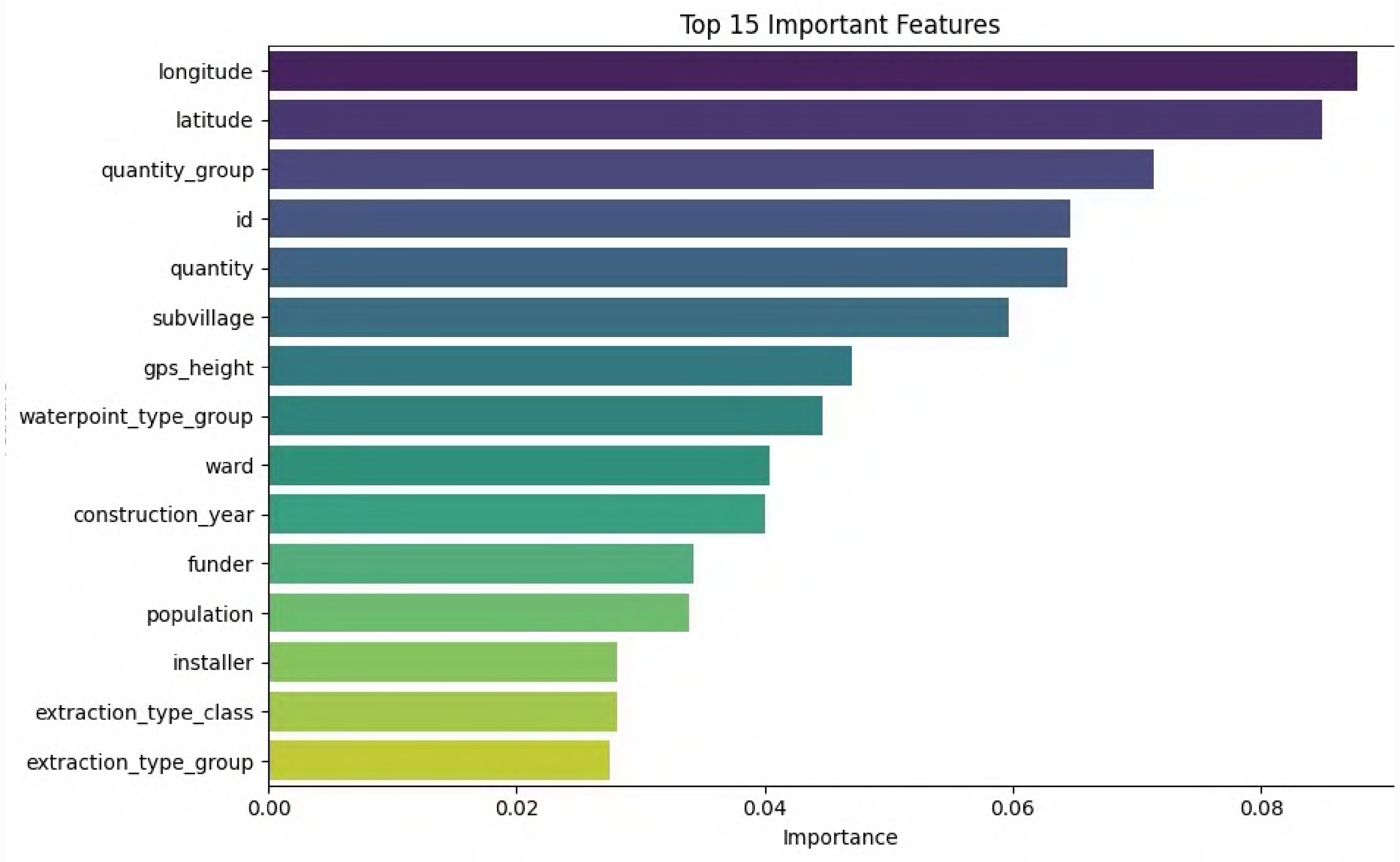
if you dont spark joy (or information), you're out!

Selected based on Importance

- Used **Random Forest's intrinsic feature importance**
- Checked **Logistic Regression weights** for interpretability

Minimal Redundancy Left After Data Cleaning

- Extensive preprocessing removed irrelevant and duplicate features
- After this step, very few features were left to drop
- So we simply used what remained and ensured it had signal



OUR BASELINE MODEL

why KNN

K-Nearest Neighbors (KNN) is a **lazy learning**, instance-based algorithm

KNN is straightforward to implement and easy to interpret

KNN naturally handles multi-class classification without requiring architectural adjustments

it was also the first supervised learning model we were introduced to

$$\hat{y} = \text{mode}(y_i \text{ for } i \in \text{K nearest neighbors})$$

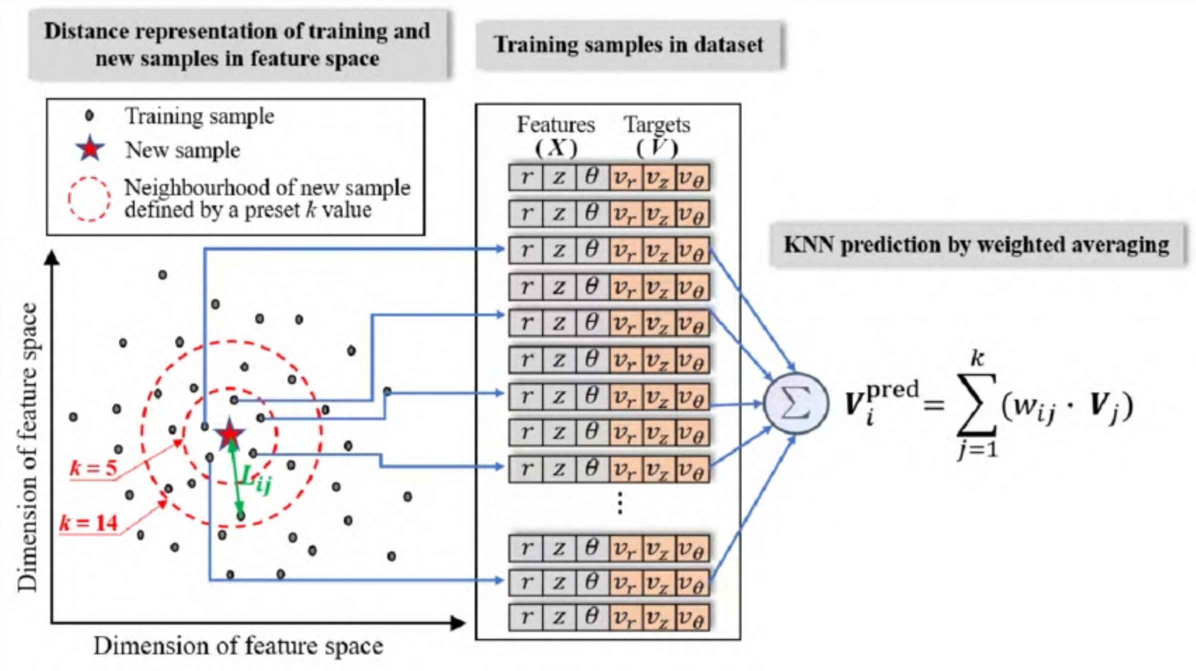
KNN - K NEAREST NEIGHBOURS

how KNN

1) For a new input sample, KNN calculates its distance (commonly Euclidean) to all other points in the training dataset.

2) It identifies the k closest training samples (neighbors) to the new input based on the computed distances.

3) For classification, it takes a majority vote of the labels of those k neighbors. The class with the most votes becomes the prediction.



eh KNN

classification accuracy = 0.5375
(using just continuous columns)

classification accuracy = 0.7185
(using both categorical and continuous columns

What next?

Find out what others did

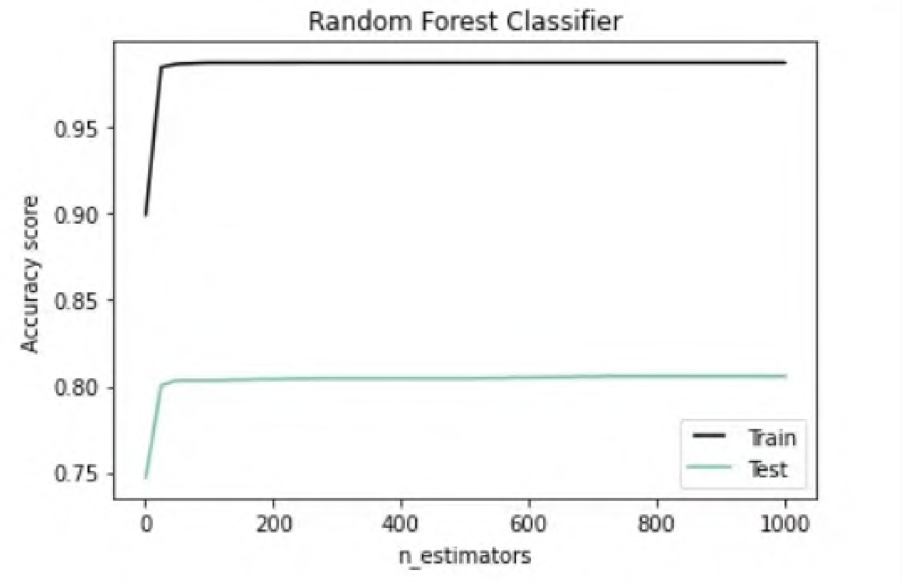
LITERATURE REVIEW PT.2

let's go through other participant's github/articles

- MODEL: RANDOM FOREST WITH WEIGHTED VOTING CLASSIFIER

BEST	CURRENT RANK	# COMPETITORS
0.8235	504	12803

- GridSearchCV for hyperparameter tuning
- best parameters (max_depth = 30, max_feature = log2, min_sample_spilt = 7, n_estimator = 200, random_state = 42
- no more than 200 trees required for tuning



Brenda Loznik - Data Scientist

- MODEL: H2O.RANDOM FOREST

BEST	CURRENT RANK	# COMPETITORS
0.8248	50	5235

- THEY BUILT A INTERACTIVE DASHBOARD TO MAKE THIS KNOWLEDGE ACCESSIBLE

Random Forest - R	Random Forest - h2o

• Faster to predict when model built
• Can only handle categorical variables with 53 unique levels
• No native multi-core processing
• Different way of splitting leaf nodes for categorical variables
	• Faster to build model
	  ○ Native multi-core processing
	  ○ Progress bar
	• Unlimited number of levels for categorical variables

Northwestern University MS capstone project - DaftPump

MOdel: 11 XGBOOST ensemble method

Accuracy: No data provided

A random guy who ranked 9th in 2017

model: h2O.Random Forest

accuracy: 0.8238

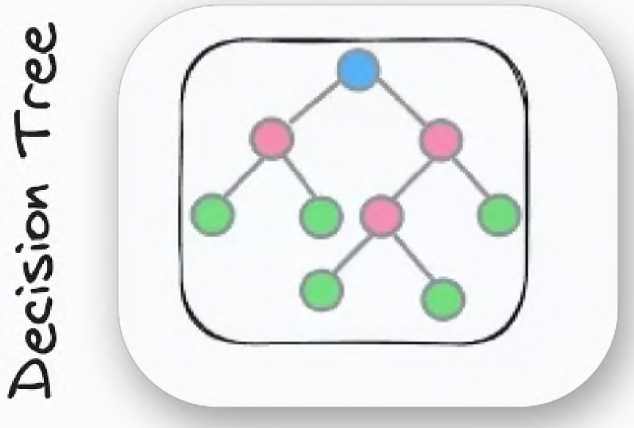
A random guy with good accuracy

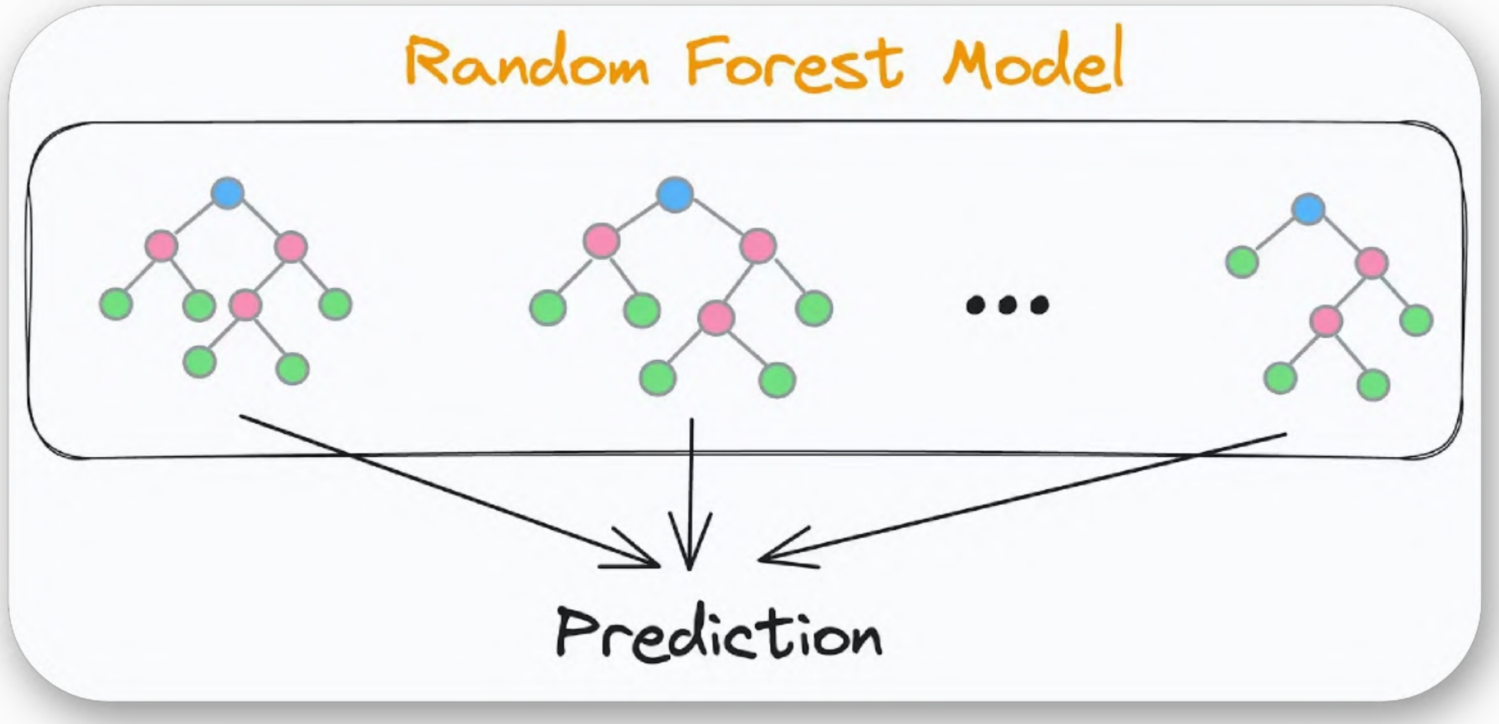
The winner

From Github Repos ~ A Word-of-Mouth...

RANDOM FOREST

RANDOM FOREST





img: https://blog.dailydoseofds.com/p/your-random-forest-is-underperforming

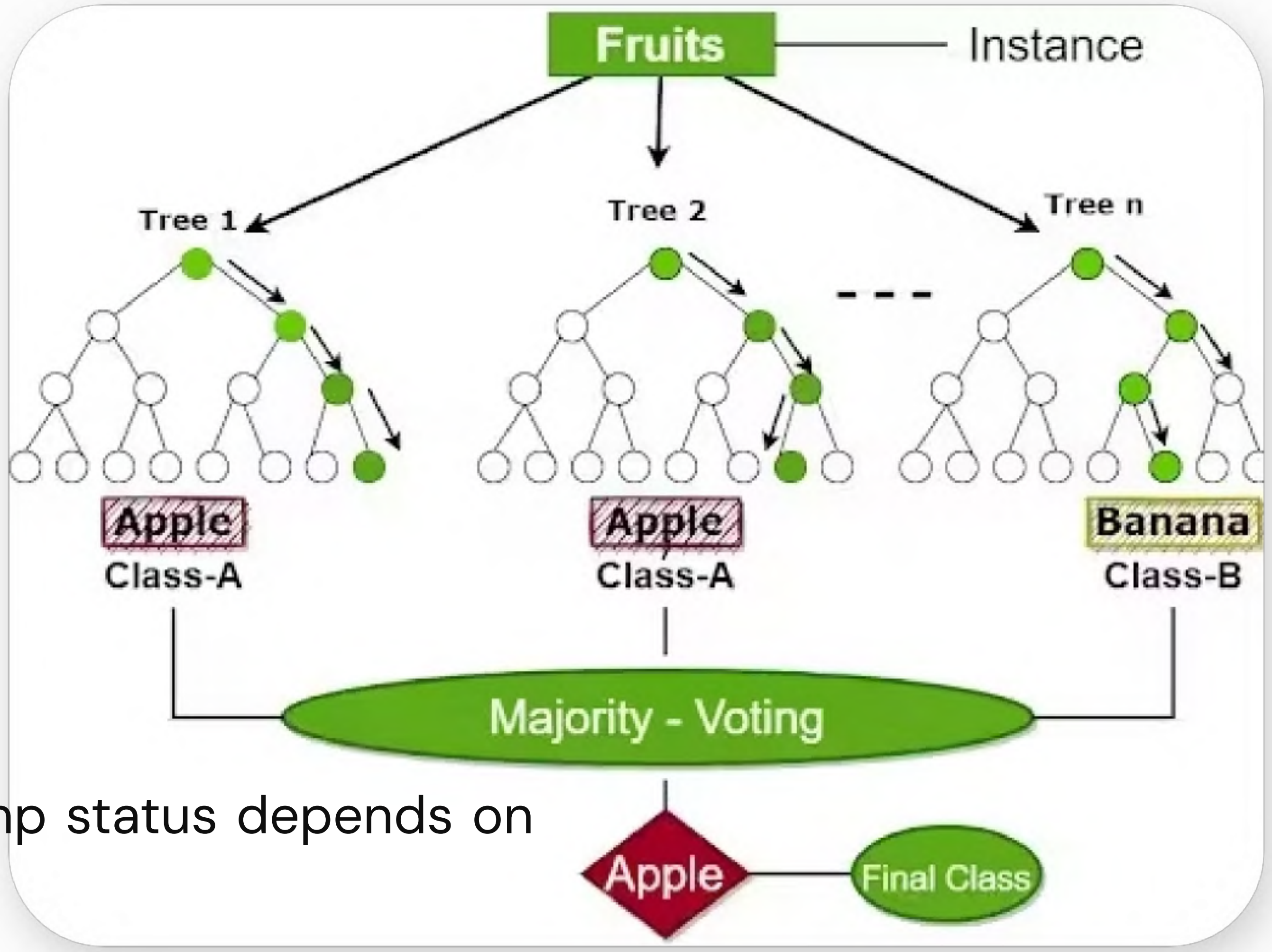
Random Forest is an **ensemble machine learning algorithm** that builds multiple **decision trees** and combines their outputs to make more accurate and stable predictions.

- **Builds many decision trees**
  Each tree is trained on a random bootstrap sample (random subset with replacement) of the dataset.

- **Random feature subsets per tree**
  Each tree splits nodes using a random subset of features, encouraging diversity among trees.

- **Combines predictions via hard voting (classification)**
  Each tree votes for a class; the class with the majority votes becomes the final prediction.

Why Use Random Forest?



img: https://www.turing.com/kb/random-forest-algorithm

Captures complex, non-linear patterns

Random Forest can uncover intricate relationships in the data, which is useful since pump status depends on many interacting factors.

Resistant to noise and overfitting

By averaging results from many trees trained on different data and feature subsets, it avoids overfitting-a common issue with single decision trees.

Handles both numeric and categorical features

It works well with the diverse types of data in "Pump It Up" (like location coordinates and management types), especially after encoding categories.

Built-in feature selection

Because each tree looks at different features, Random Forest naturally highlights the most important ones for predicting pump status(automatically prioritizes features that consistently improve split quality across trees,).

Other Models

CATBOOSTING

NUERAL NETWORKS

RF_MODEL WITH K FOLD TARGET ENCODING

XGBOOST

KNN

LOGISTIC REGRESSION

Catboost

```
Shrink model to first 294 iterations.
Model trained on 100000.0% of the data
Accuracy: 0.8011
Classification Report:
                        precision    recall  f1-score   support

            functional       0.79      0.91      0.84      6452
functional needs repair       0.61      0.24      0.35       863
        non functional       0.84      0.76      0.80      4565

              accuracy                           0.80     11880
             macro avg       0.75      0.64      0.66     11880
          weighted avg       0.80      0.80      0.79     11880
```

CatBoost is a gradient boosting algorithm, developed by **Yandex**.
IT works well with categorical features
It encodes categorical data automatically.
Builds an ensemble of decision trees using **gradient boosting**.
Each new tree focuses on correcting the errors made by the previous ones.

Accuracy:0.8011

XGboost

```
XGBoost trained using Optuna best params (80/20 split).
Accuracy on test: 0.8098
Classification Report:
                        precision    recall  f1-score   support

            functional       0.80      0.91      0.85      6452
functional needs repair       0.62      0.26      0.37       863
        non functional       0.85      0.77      0.81      4565

              accuracy                           0.81     11880
             macro avg       0.75      0.65      0.68     11880
          weighted avg       0.80      0.81      0.80     11880
```

XGBoost is an optimized and scalable implementation of the gradient boosting algorithm, designed for performance and speed.

Gradient Boosting is a method where models are added one after another, and each new model tries to fix the mistakes made by the one before it.

Builds an **ensemble of decision trees** using **gradient boosting**.
Uses **second-order derivatives** (like Newton's method) for better accuracy.
Incorporates **regularization (L1 & L2)** to reduce overfitting.
Supports **parallel processing**, making it **fast and efficient**.
Can handle **missing values internally**.

Accuracy:0.8098

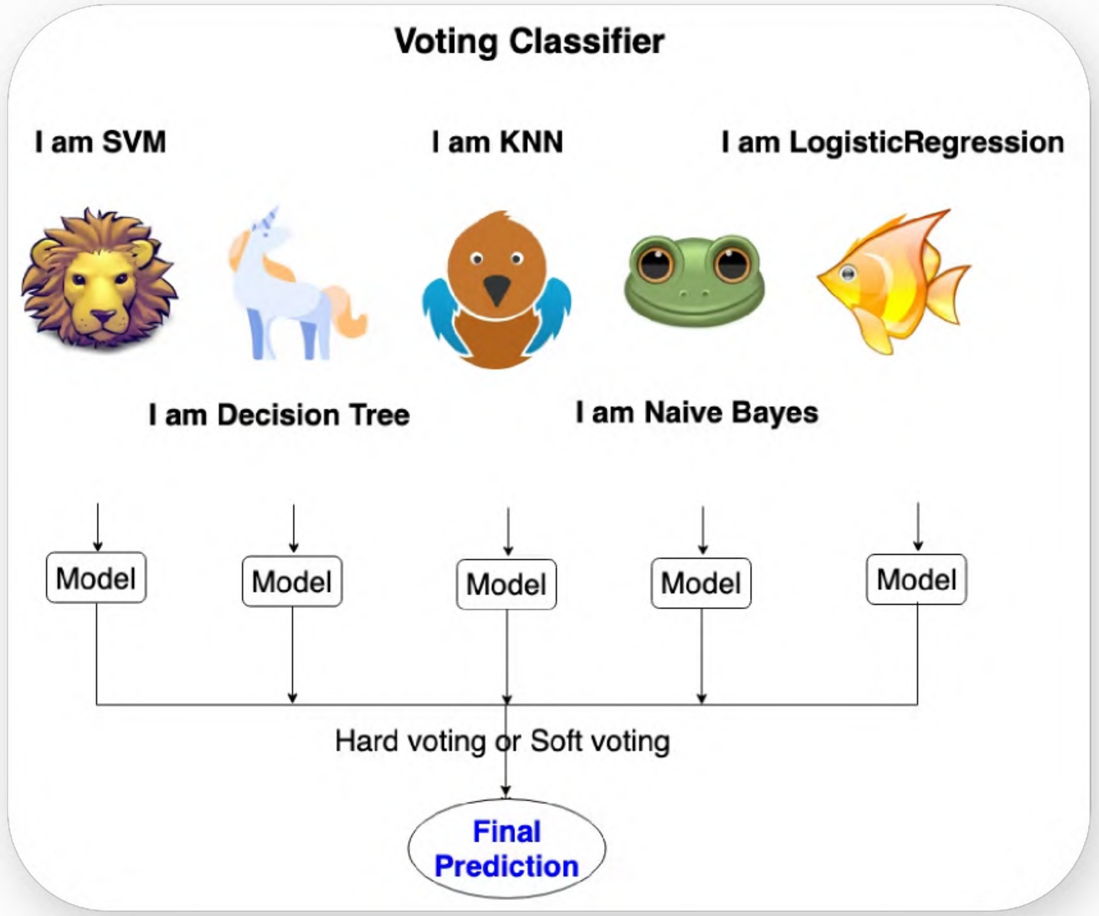
Soft Voting Ensemble

- Combines predictions from **Random Forest**, **XGBoost**, and **LightGBM**.
- Uses **soft voting**: averages predicted probabilities from all models.

```
🔍 Ensemble Soft Voting Classifier
Accuracy: 0.8074
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.90      0.85      6452
           1       0.85      0.77      0.81      4565
           2       0.56      0.31      0.40       863

    accuracy                           0.81     11880
   macro avg       0.74      0.66      0.69     11880
weighted avg       0.80      0.81      0.80     11880
```



Accuracy:0.8074

HYPERPARAMETER TUNING

PERFORMANCE METRICS

	Classification accuracy	leaderboard ranking
RF with 100% training set	0.8215	1517
RF with Frequency Encoding optimized with optuna	0.8182	2456
RF with Frequency Encoding	0.8143	3375
RF with K fold target encoding (k = 150)	0.7954	5451
KNN with Numerical Columns	0.5375	7187

RF: Random Forest

CHALLENGES ?

LAPTOP BAD

HOW DID WE TACKLE THEM ?

THE SAVIOUR

LIMITATIONS

- **Limited Data Availability**
  Manual survey data is static and often incomplete. It lacks the granularity and timeliness required for optimal predictive maintenance.

- **Absence of Real-Time Monitoring**
  Our model is trained on cross-sectional data that does not reflect real-time changes in pump conditions, water usage, or environmental stressors.

- **High Resource Cost of Advanced Alternatives**
  The referenced telemetry study (Thomas et al, 2021) used **sensors on handpumps** and **satellite data** (for 480 data points) to achieve high predictive accuracy and scalability. However, such solutions require substantial funding for optimal infrastructure often unavailable in rural regions like Tanzania.

- **AI/ML in Predictive Maintenance is Emerging**
  The application of machine learning to rural infrastructure maintenance is a **relatively recent field**. Data sparsity remains a challenge, particularly when real-time instrumentation is not feasible.

FINAL ACCURACY AND RANK

82.15%

top 8% !!! 🤯

CONCLUSION

- **Early intervention = faster repairs**
  → Prevents long-term pump failures.

- **Increased water accessibility**
  → Reduces water scarcity in rural communities.

- **Better planning for NGOs & governments**
  → Enables **targeted maintenance**.

Despite limitations, our model provides a **low-resource, scalable baseline** that can inform water point maintenance decisions and can be enhanced as better data becomes available.

Deployment ?

WHAT IS NEEDED

Can the solution be deployed at Plaksha to solve the problem you have chosen? If so, how?

WHAT we thought

- Planned to test the model on Indian data, including a case study on Mohali waterpoints.
- Goal: Evaluate accuracy and explore deployment feasibility in a new region.
- Challenge: Indian dataset had different features and collection methods compared to the Tanzanian data.
- Result: Direct model transfer was not feasible — a new model would need to be built for Indian conditions.

References

Katomero, J., Georgiadou, Y., Lungo, J., & Hoppe, R. (2017). Tensions in rural water governance: The elusive functioning of rural water points in Tanzania. *ISPRS International Journal of Geo-Information*, *6*(9), 266. https://doi.org/10.3390/ijgi6090266

Joseph, G., Andres, L. A., Chellaraj, G., Zabludovsky, J. G., Ayling, S. C. E., & Hoo, Y. R. (2019). Why do so many water points fail in Tanzania? An empirical analysis of contributing factors [Article]. *Policy Research Working Paper*, *WPS8729*, WPS8729. https://www.researchgate.net/publication/333827509

Joseph, G., Haque, S., Ayling, S., World Bank, & Swedish International Development Cooperation Agency. (2018). *Reaching for the SDGs: the untapped potential of Tanzania's water supply, sanitation, and hygiene sector* [WASH Poverty Diagnostic]. World Bank. https://www.worldbank.org

World Bank Group. (2018, March 21). Improving Water Supply and Sanitation Can Help Tanzania Achieve its Human Development Goals. *World Bank*. https://www.worldbank.org/en/news/press-release/2018/03/20/improving-water-supply-and-sanitation-can-help-tanzania-achieve-its-human-development-goals

DrivenData. (n.d.). *Pump it Up: Data Mining the Water Table*. DrivenData. https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/

BrendaLoznik. (n.d.). *waterpumps/2B. Dealing with missing data.ipynb at main · BrendaLoznik/waterpumps*. GitHub. https://github.com/BrendaLoznik/waterpumps/blob/main/2B.%20Dealing%20with%20missing%20data.ipynb

Thomas, E. A., Wilson, D., Kathuni, S., Libey, A., Chintalapati, P., & Coyle, J. (2021). A contribution to drought resilience in East Africa through groundwater pump monitoring informed by in-situ instrumentation, remote sensing and ensemble machine learning. *Science of the Total Environment, 780*, 146486. https://doi.org/10.1016/j.scitotenv.2021.146486

Pitch

Want to make a presentation like this one?

Start with a fully customizable template, create a beautiful deck in minutes, then easily share it with anyone.

Create a presentation (It’s free)