Predictive Analytics for Smart Water Management in Developing Regions

Gissella Bejarano[1], Mayank Jain[1], Arti Ramesh[1], Anand Seetharam[1], Aditya Mishra[2]
[1]Department of Computer Science, SUNY Binghamton, USA
[2]Department of Computer Science and Software Engineering, Seattle University, USA
{gbejara1, mjain8, artir, aseethar}@binghamton.edu, mishraa@seattleu.edu

***Abstract*—Water availability and management is an important problem plaguing many developing and under-developed countries. Many factors including geographic, political, management, and environmental factors affect the availability of water in these regions. In this paper, we develop an ensemble-learning based predictive-analytics framework for smart water management to predict: *i)* water pump operation status (e.g., functional, non functional), *ii)* water quality, and *iii)* quantity. In the predictive-analytics framework, we first perform feature engineering to select relevant features, use them to develop the XGBoost and Random Forest ensemble learning models, and then perform extensive feature analysis to identify the most predictive features, for each prediction problem mentioned above.**

We evaluate our framework on two publicly available smart water management datasets pertaining to Tanzania and Nigeria and show that our proposed models outperform several baseline approaches, including logistic regression, SVMs, and multi-layer perceptrons in terms of precision, recall and F1 score. We also demonstrate that our models are able to achieve a superior prediction performance for predicting water pump operation status for different water extraction methods. We conduct a detailed feature analysis to investigate the importance of the various feature groups (e.g., geographic, management) on the performance of the models for predicting water pump operation status, water quality and quantity. We then perform a fine-grained feature analysis to identify how individual features, not just feature groups, impact performance. We identify that among individual features, location (x, y, z coordinates) has the maximum impact on performance. Our analysis is helpful in understanding the types of data that should be collected in future for accurately predicting the different water problems.

I. Introduction

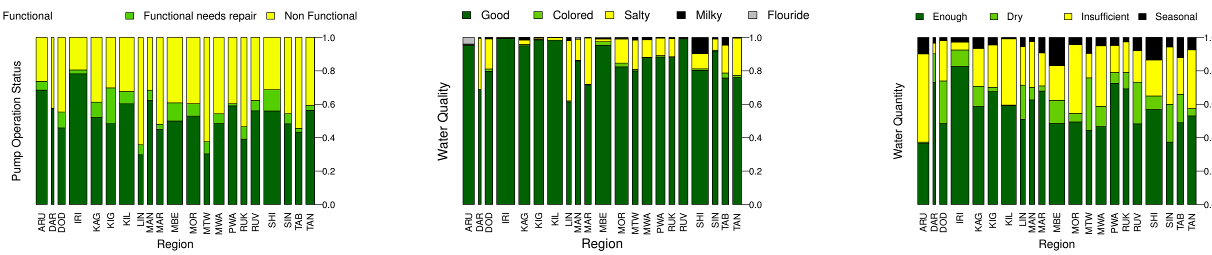
Water availability and management is an important problem plaguing many developing and under-developed countries. Many factors including geographic, political, management, and environmental factors affect the availability of water in these regions. In most developing regions, the primary means of extracting water is water pumps (e.g., manual, hand pump, solar pump). While significant initial capital is needed to install these pumps, a sustained long-term effort and investment is also required to maintain these pumps. For example, nearly 1.42 billion dollars have been donated to address the water access crisis in Tanzania [1]. Though significant strides have been undertaken in installing pumps across the country, many of these pumps are in a condition of decay, primarily due to the lack of adequate maintenance [1].

Therefore, in this paper, we develop a predictive-analytics framework to investigate important questions related to *water availability:* water pump operation status, water quality, and quantity. Specifically, we answer the following questions. *i)* Can we accurately predict water pump operation status, water quality, and quantity from the data? *ii)* How do individual or groups of features impact water pump operation status, quality, and quantity?

Answering these questions is essential for the well being and economic growth of communities in these regions. They provide valuable information that can be used by the authorities to identify and effectively allocate scarce resources (e.g., money, personnel) to places most in need. For example, they provide insight into where to install new pumps depending on the water availability and which pumps are in need of immediate maintenance. Additionally, our analysis helps identify the most important features that influence the operation of a pump and thus can be extremely beneficial and used as a reference by other regions/nations who plan to collect similar data in the future.

Specifically, our contributions are as follows:

1) We develop a predictive-analytics framework that incorporates the different features to predict different problems related to water availability: *i)* pump operation status, *ii)* quality, and *iii)* quantity. In the predictive-analytics framework, we first perform feature engineering to select relevant features, use them to develop the XGBoost and Random Forest ensemble learning models, and then perform extensive feature analysis to identify the most predictive features, for each of the above-mentioned prediction problem.
2) We evaluate our framework on two publicly available datasets related to the operation of pumps in Tanzania [2] and Nigeria [6], respectively and demonstrate that our models are able to accurately predict the different water-related problems and significantly outperform several baseline models.
3) We perform a fine-grained feature analysis to understand the contribution of different features in predicting the different water problems. We group similar features into groups and drop/include feature groups individually and measure the corresponding effect on prediction performance. Our feature analysis throws light on the contribution of each feature group, individually and in combination with other feature groups, thus helping in understanding the relationship between feature groups.



(a) Pump operation status (b) Water quality (c) Water quantity

Fig. 1. Tanzania: Comparison of pump operation status, water quality, and water quantity for different regions

From our analysis, we observe that geographic features and source-related features are the most important feature groups across the different prediction problems.

4) Finally, we rank the individual features in order of importance to understand their contribution toward prediction performance. We identify that among individual features, location (x, y, z coordinates) has the maximum impact on performance. This endeavor helps in understanding the importance of the different features/feature groups in the various water prediction problems.

II. Related Work

In this section, we outline existing work related to smart water management and contrast it with our work. Due to the lack of open datasets, there is limited research in this space. In [9], the authors study a relatively small dataset from Ghana's GAP region. They perform a Bayesian Network analysis and observe strong dependencies between pump functionality and features such as pump type and management. Jimenez *et al.* analyze the relationship between the functionality and the technology of the water points [13].

Another study focusing on Liberia, Sierra Leone, and Uganda [10] analyzes the risk factors associated with non-functionality of hand pumps. They apply a logistic regression model to a dataset of community-managed hand pumps and observe that age of the pump, distance from district/country capital, and absence of user fee collection all contribute significantly toward pump non-functionality. In [16], the authors investigate the performance of demand-driven, community-managed water supply systems in rural areas of developing countries through a large-scale study.

Water quality and quantity have also been investigated in prior work [3], [7]. The authors in [14] build a multi-task, multi-view learning framework to predict urban water quality by combining a number of data sources including water hydraulic data, weather, pipe networks, structure of road networks, and point of interests (POIs). In another work, the authors report results from water quality measurement studies carried out in the rural districts of Tanzania [7].

Our work is closest to [6], where the authors explore the same two datasets and investigate the factors influencing the water pump functionality using regression and Bayesian Network analysis. From their analysis, they identify that hand pumps of a particular make have higher functionality in comparison to others. They find strong correlation between management type and functionality. Our approach differs from this existing approach in that we adopt a predictive analytic approach. We design a framework to predict water pump operation status, water quality, and quantity and use that to identify the most predictive features for each of these problems. We analyze how accurately we are able to predict pump operation status across different extraction types. Our analysis helps in identifying the types of data that needs to be collected to address similar water problems in future. We note that similar machine learning techniques have been used to predict other environmental factors (e.g., air, river water quality, landslide) for enabling a smarter world [8], [15], [12].

III. Dataset

In this section, we describe the two water management datasets from Tanzania and Nigeria used in this work. The dataset for Tanzania and Nigeria have been made publicly available by Taarifa and the Tanzanian and Nigerian Ministry of Water, respectively [2], [6]. The Tanzania dataset was collected using hand-held sensors, paper reports, and feedback from people using cellular phones. The Tanzania dataset has $59,401$ instances and contains information such as the pump operation status, water quality, water quantity, pump location, source type, extraction technique, and population demographics in the region where the pump is installed. The Nigeria dataset has $132,542$ instances and has features similar to, but less in comparison to the Tanzania dataset.

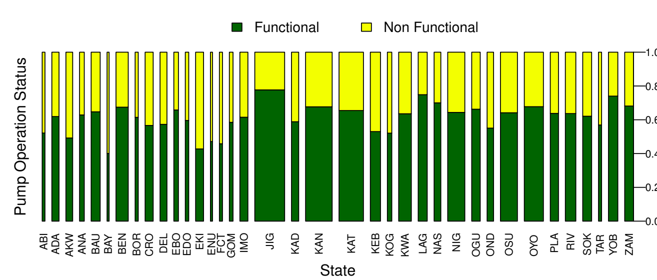


Fig. 2. Nigeria: Comparison of pump operation status for different states

The primary difference between the datasets is that the Nigeria dataset does not contain information regarding the water quality and quantity. Therefore, in this paper, we present significantly more results for Tanzania in comparison to Nigeria. In the Tanzania dataset, the pump operation status is described using three values namely *functional*, *functional needs repair* and *non-functional*, while in Nigeria the pump operation status is described using two values *functional* and *non-functional*. The water quality in the Tanzania dataset is described using the values *good*, *milky*, *salty*, *colored* and *fluoride* and the water quantity takes the values *dry*, *enough*, *insufficient* and *seasonal*.

Figures 1(a), 1(b), and 1(c) show the normalized distribution of the water pump operation status, the water quality, and quantity for the different regions for Tanzania and Figure 2 shows the normalized distribution of the water pump operation status for Nigeria. The width of the bars in the figures denote the number of instances that correspond to a particular region or state. We make multiple important observations from these figures - *i)* the total number of recorded data points varies with region/state, *ii)* there is a significant portion of pumps that are *non-functional* in almost all regions, with some regions such as LIN, MTW, RUK having greater than 50% non-functional pumps, and *iii)* there is an uneven distribution of the values for water quality and quantity. For example, if we consider water quality, a large fraction of instances have *good* value. But, for the same regions, we can observe that a considerable fraction of pumps do not have *enough* water quantity.

IV. Smart Water Management Prediction Framework

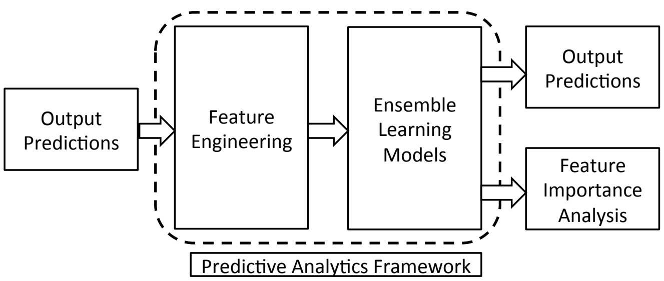


Fig. 3. Smart Water Management Prediction Framework

Having provided an overview of the datasets in the previous section, we next describe the smart water management framework that can accurately predict the pump operation status, water quality, and quantity. Figure 3 provides an overview of the different components in our smart water management framework. Our framework can potentially be extended to address water-related issues in other developing and underdeveloped regions.

A. Feature Engineering

In order to design effective models for the problems studied in this paper, we remove irrelevant features from the set of available features. For example, we remove features such as the name of the water point or village name as they are either unique or shared between few instances in the dataset. We use Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC) to determine the correlation between features and class variables. The correlation values are then used to eliminate features from the dataset that may not be useful in the prediction problem. Additionally, we also convert few features to relevant units. For example, we convert the latitude and longitude values in the datasets to x, y, and z coordinates. Similarly, we use the year in which the pump was manufactured to determine the age of the pump. Missing values in the dataset are replaced by a measure of central tendency (i.e., mean, median) or not applicable (NA) depending on the appropriateness.

B. Predictive Models

We leverage two ensemble learning models, namely Random Forest [4] and a recently developed ensemble model, XGBoost [5], to address the smart water management problem. Ensemble learning methods leverage multiple learning algorithms to obtain better prediction performance than what could be obtained from the respective individual learning algorithms in the ensemble and have been shown to be effective in a number of applications, particularly in problems that involve data that has class imbalance [11].

1) Random Forest: It constructs multiple decision trees based on bootstrapping and random attribute selection during the training phase. The algorithm uses them to predict the class during the test phase, and then outputs the result by carefully combining the results from the different trees [4]. Random Forest avoids overfitting by randomly selecting a set of attributes instead of taking all the available attributes into consideration for constructing the trees.

2) XGBoost: In contrast to Random Forest, XGBoost uses dependent but smaller decision trees. It uses a gradient boosting algorithm to improve the results of the previous trees to predict the next tree. The final output is decided on the basis of a voting algorithm that is applied on the results obtained from all the trees [5].

V. Experimental Results

In this section, we conduct experiments to demonstrate the efficacy of our framework. Specifically, we answer the following questions:

- How good are our models in predicting different attributes of water management: pump operation status, quality, and quantity in Tanzania, and pump operation status in Nigeria?
- What features/feature groups are most predictive of pump operation status, quality, and quantity?

In all our experiments, we use *5-fold* crossvalidation, where we divide the data into 5 partitions, iteratively train on four partitions and report the prediction performance on the fifth partition. We report standard performance metrics of precision, recall, and F1 score for all the models.

- *Precision* is defined as a ratio of the true positives to the sum of the true positives and false negatives.

- *Recall* is the ratio of the true positives to the true instances in the dataset (i.e., the sum of true positives and the false negatives).
- The *F1 score* is calculated as the harmonic mean of the precision and recall.

We compare our models with several classic machine learning approaches such as Support Vector Machines (SVM), Logistic Regression, Multilayer Perceptrons, and Naive Bayes. We report results for SVM, the model that performs the best on our dataset. Statistically significant differences evaluated at a rejection threshold of $p = 0.05$ are typed in **bold** in all the tables below. We measure statistical significance between XGBoost and Random Forest, wherever relevant, to show which of these models is a better fit for the prediction problem. For scores where we cannot establish statistical significance between XGBoost and Random Forest, we report statistical significance with SVM. We note that both our ensemble models perform statistically better than SVM across all prediction tasks and in all performance metrics.

A. Pump Operation Status Prediction

In this subsection, we report performance results for pump operation status for Tanzania and Nigeria. Table I gives the performance results for the pump operation status for Tanzania. We observe that Random Forest and XGBoost perform better than SVM across all performance metrics. Our models achieve a 78% performance improvement in F1 score over SVM for *non-functional*, 79% for *functional needs repair*, and 13% for *functional*, respectively. Looking closely at the results for individual class values, we observe from Table I that the performance of the proposed models is better for the *functional* and *non-functional* classes in comparison to the *functional needs repair* class for the Tanzania dataset. The main reason behind the lower performance for the *functional needs repair* class is the lack of enough instances pertaining to this class in our dataset (as shown in Figure 1(a)).

Similarly, from Table II, we observe that XGBoost and Random Forest perform better than SVM on the Nigeria dataset. We observe that the F1 score performance is higher for the *non-functional* class in the Nigeria dataset in comparison to the *functional* class for all the three models. We observe that our proposed models achieve a performance improvement of 39% in the *functional* class when compared to SVM.

In Tables III and IV, we evaluate how accurately are our models able to predict pump operation status across the different extraction methods. *Accuracy* is defined as percentage of instances predicted correctly by our models in the total number of instances. We observe that both our proposed models can accurately predict the pump operation status for the different water extraction methods. XGBoost outperforms Random Forest in the Tanzania dataset across all extraction types, while their performance is comparable for the Nigeria dataset.

Model	Class	F1 Score	Precision	Recall
SVM	Functional	0.75	0.62	0.94
	Functional needs repair	0.24	0.62	0.14
	Non Functional	0.46	0.79	0.32
XGBOOST	Functional	**0.85**	**0.81**	**0.91**
	Functional needs repair	0.42	**0.63**	0.31
	Non Functional	**0.82**	**0.85**	0.78
Random Forest	Functional	0.85	0.81	0.88
	Functional needs repair	**0.43**	0.54	**0.36**
	Non Functional	0.81	0.84	**0.79**

TABLE I
TANZANIA: F1 SCORES, PRECISION AND RECALL, AND FOR PREDICTING *pump operation status*

Model	Class	F1 Score	Precision	Recall
SVM	Functional	0.38	0.49	0.31
	Non Functional	0.74	0.68	0.82
XGBOOST	Functional	0.52	**0.57**	**0.47**
	Non Functional	**0.76**	**0.73**	**0.80**
Random Forest	Functional	**0.53**	0.50	0.32
	Non Functional	0.74	0.68	0.78

TABLE II
NIGERIA: F1 SCORES, PRECISION AND RECALL FOR PREDICTING *pump operation status*

Extraction Type	Accuracy	
	XGBOOST	Random Forest
Gravity	86.65%	79.21%
Hand pump	86.67%	72.88%
Motor pump	93.79%	90.86%
Rope pump	95.37%	85.64%
Submersible	93.31%	84.07%
Wind	84.31%	72.54%

TABLE III
TANZANIA: ACCURACY OF PREDICTIONS ACROSS DIFFERENT EXTRACTION TYPES

Extraction Type	Accuracy	
	XGBOOST	Random Forest
Animal	65.57%	63.92%
Diesel	65.06%	68.85%
Electric	64.47%	63.96%
Hand pump	64.70%	64.96%
Manual	65.03%	65.64%
Solar	63.70%	63.92%
Wind	86.66%	66.67%

TABLE IV
NIGERIA: ACCURACY OF PREDICTIONS ACROSS DIFFERENT EXTRACTION TYPES

B. Quality and Quantity Prediction

Recall that water quality and quantity measurements are only available for the Tanzania dataset. From Figure 1(b), we observe that majority of data instances correspond to *good* water quality, while for the remaining data instances the water quality is spread across *salty, fluoride, colored,* and *milky*. As there are less number of instances in *salty, fluoride, colored,* and *milky* classes, we group them into *bad* water quality class. In comparison, from Figure 1(c), we observe that there are sufficient data points in all classes for predicting quantity. Hence, in our quantity prediction models, we consider all the four different quantity classes.

Table V shows the performance results for water quality.

Model	Class	F1 Score	Precision	Recall
SVM	Good	0.94	0.90	0.99
	Bad	0.33	0.83	0.20
XGBOOST	Good	0.95	0.92	**0.99**
	Bad	0.49	**0.84**	0.35
Random Forest	Good	**0.96**	**0.95**	0.97
	Bad	**0.69**	0.78	**0.62**

TABLE V
TANZANIA:PRECISION, RECALL, AND F1 SCORES FOR PREDICTING *quality*

Model	Class	F1 Score	Precision	Recall
SVM	Dry	0.49	0.87	0.34
	Enough	0.77	0.64	0.97
	Insufficient	0.35	0.76	0.23
	Seasonal	0.46	0.83	0.31
XGBOOST	Dry	0.84	**0.90**	0.79
	Enough	**0.89**	0.85	**0.94**
	Insufficient	**0.77**	**0.82**	0.72
	Seasonal	**0.73**	**0.82**	**0.66**
Random Forest	Dry	**0.85**	0.86	**0.82**
	Enough	0.89	**0.87**	0.92
	Insufficient	0.77	0.80	**0.74**
	Seasonal	0.72	0.79	0.66

TABLE VI
TANZANIA: PRECISION, RECALL, AND F1 SCORES FOR PREDICTING *quantity*

We observe that predicting *bad* quality is a more challenging prediction problem as there are lesser number of instances for *bad* quality as opposed to number of instances for *good* quality. Here, our ensemble models achieve a significant improvement in the F1 score when compared to the SVM model, giving a performance improvement of 109% for the *bad* class, despite it having fewer number of instances.

Table VI shows the performance results for predicting water quantity. Here again, we observe that our ensemble models achieve a superior prediction performance in F1 score when compared to SVM, improving the prediction performance for *insufficient* and *dry* classes by 120% and 73%, respectively.

C. Fine-grained Feature Analysis

In this section, we perform a fine-grained feature analysis by: *i)* leaving groups of features out and including specific feature groups and examining the corresponding effect on prediction performance, and *ii)* ranking features based on their contribution to the prediction problem. We first conduct a fine-grained feature analysis to investigate the impact of the contribution of features groups on prediction performance and then extend this to identify the most important individual features affecting performance in each of the prediction problems. Our feature analysis is especially useful when extending the prediction models to new datasets/regions where only a subset of the features are available.

1) Feature Group Analysis: To conduct this analysis, we group similar features into features groups. For Tanzania, we identify three feature groups: *i)* geographic features (GF), *ii)* management features (MG), *iii)* source-related features (SR). The GF group includes features such as location (x, y, z coordinates) and the GPS height. The MG group includes features such as the installer of the pump and the entity responsible for maintaining the pump. The SR group contains features related to the water source including quality, quantity, and water extraction method. We conduct a similar grouping of features in the Nigeria dataset. The Nigeria dataset only contains geographic (GF) and source-related (SR) feature groups.

We investigate the predictive power of the various feature groups by adopting a two-pronged approach - *i)* dropping a particular feature group, *ii)* including only a particular feature group and examining the effect on the prediction performance. Dropping a particular feature group helps us understand how excluding it can adversely impact performance. In comparison, only including a particular feature group helps us appreciate the predictive power of the feature group when applied alone.

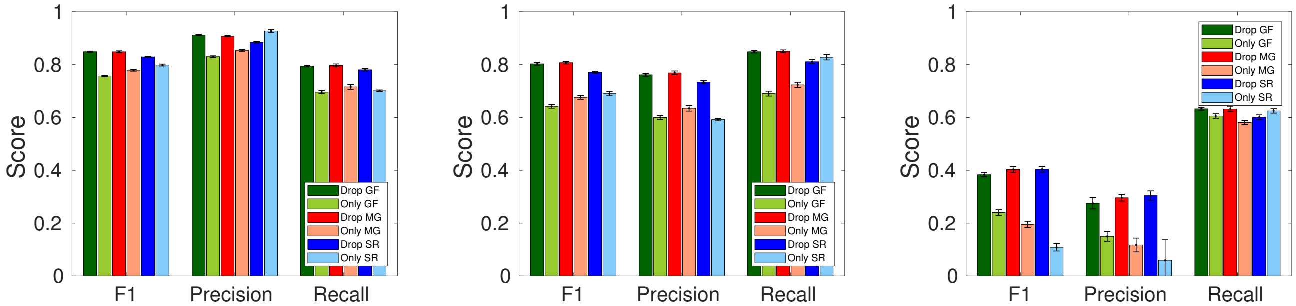
Figures 4(a), 4(b), and 4(c) capture the performance impact of leaving feature groups out and including only a single feature group for pump operation status for the three classes: *functional*, *functional needs repair*, and *non-functional*, respectively. We observe that dropping the SR feature group has a greater performance impact for the *functional* and *functional needs repair* classes in comparison to dropping the other two feature groups, while the GF feature group has the highest impact for the *non-functional* class. Similarly, including only the SR feature group achieves the highest prediction performance for *functional* and *functional needs repair* classes, while including only the GF feature gives the highest prediction performance for the *non-functional* class.

On average, we observe that the SR and GF individual feature groups have the highest impact on the *functional* and *functional needs repair* classes and the *non-functional* classes, respectively. From our analysis, we conclude that the geographical location of the pump (GF features) play an important role in predicting *non-functional* pumps. For distinguishing between pumps that are *functional* or *functional needs repair*, SR features are most helpful.

Figure 5 shows the leave one group out and include only a single group analysis for the pump operation status for the Nigeria dataset. We observe from the figure that the GF feature group plays a crucial role in predicting *functional* class for the Nigeria dataset. In comparison, it is hard to pick a winner for the *non-functional* group.

We also conducted the feature group analysis for water quality and quantity for the Tanzania dataset and observe similar results. We omit these results for the lack of space. Our extensive feature analysis helps in understanding the individual contribution of feature groups and their dependence with other feature groups. Our analysis and subsequent conclusions serves as a reference for collecting similar data in future and what information to focus on for already existing data, depending on what water-related problems are of interest.

2) Feature Importance Ranking: Having studied the impact of the different feature groups on performance, we investigate the contribution of individual features by examining how valuable a feature is in the construction of boosted decision trees in the XGBoost model. Feature importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by

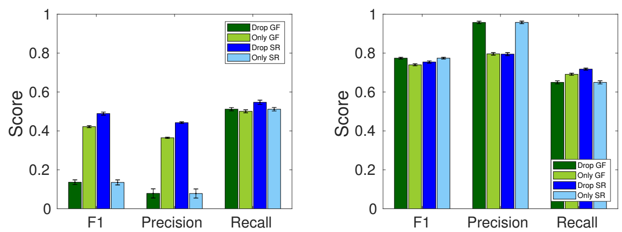


(a) Performance scores for *functional*

(b) Performance scores for *functional needs repair*

(c) Performance scores for *non-functional*

Fig. 4. Tanzania: Performance scores for *functional*, *functional needs repair* and *non-functional* when a group of features are dropped or only when a group of features is included.



(a) Performance scores for *functional*

(b) Performance scores for *non-functional*

Fig. 5. Nigeria: Performance scores for *functional* and *non-functional* when a group of features are dropped or only when a group of features is included.

the number of instances the node is responsible for. This importance measure is then averaged across all the decision trees in the model to calculate the feature importance of each feature. We observe that x coordinate (denoted by position_x), y coordinate (denoted by position_y), z coordinate (denoted by position_z), gps_height are the most predictive features across all the prediction problems. This helps us understand that the geographic location plays a key role in all the water management problems.

VI. Conclusion

In this paper, we developed prediction models for smart water management to predict the water pump operation status, water quality, and quantity in developing regions. Via experiments, we demonstrated that our models perform effectively across all the three water prediction problems and outperform the baseline SVM model in terms of precision, recall and F1 score. We also demonstrated that our models can accurately predict the pump operation status for the different water extraction methods. We then performed a detailed fine-grained feature analysis to understand the contribution of features/feature-groups in the different prediction problems and make important conclusions on which aspects of the data are helpful for the different prediction problems. Our analysis throws light on how to extend our models to other water management data which only have a subset of features. Our analysis also helps in understanding what kinds of data should be collected in future for accurately predicting the different water problems.

References

[1] Humanosphere report. http://www.humanosphere.org/world-politics/2014/12/tanzania-failed-fix-water-access-problem/.

[2] The taarifa platform. http://taarifa.org/.

[3] M. Ali and A. M. Qamar. Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in pakistan. In *Digital Information Management (ICDIM), 2013 Eighth International Conference on*, pages 108–113. IEEE, 2013.

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[6] R. Cronk and J. Bartram. Factors influencing water system functionality in nigeria and tanzania: a regression and bayesian network analysis. *Environmental Science & Technology*, 51(19):11336–11345, 2017.

[7] A. J. F. de Palencia and A. Pérez-Foguet. Quality and year-round availability of water delivered by improved water points in rural tanzania: effects on coverage. *Water Policy*, 14(3):509–523, 2012.

[8] S. Deleawe, J. Kusznir, B. Lamb, and D. J. Cook. Predicting air quality in smart environments. *Journal of ambient intelligence and smart environments*, 2(2):145–154, 2010.

[9] M. B. Fisher, K. F. Shields, T. U. Chan, E. Christenson, R. D. Cronk, H. Leker, D. Samani, P. Apoya, A. Lutz, and J. Bartram. Understanding handpump sustainability: Determinants of rural water source functionality in the greater afram plains region of ghana. *Water resources research*, 51(10):8431–8449, 2015.

[10] T. Foster. Predictors of sustainability for community-managed handpumps in sub-saharan africa: evidence from liberia, sierra leone, and uganda. *Environmental Science & Technology*, 47(21):12037–12046, 2013.

[11] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, July 2012.

[12] B. B. Gulyani, J. A. Mangai, and A. Fathima. An approach for predicting river water quality using data mining technique. In *Industrial Conference on Data Mining*, pages 233–243. Springer, 2015.

[13] A. Jiménez and A. Pérez-Foguet. The relationship between technology and functionality of rural water points: evidence from tanzania. *Water science and technology*, 63(5):948–955, 2011.

[14] Y. Liu, Y. Liang, S. Liu, D. S. Rosenblum, and Y. Zheng. Predicting urban water quality with ubiquitous data. *arXiv preprint arXiv:1610.09462*, 2016.

[15] D. Tien Bui, B. Pradhan, O. Lofman, and I. Revhaug. Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and naive bayes models. *Mathematical Problems in Engineering*, 2012, 2012.

[16] D. Whittington, J. Davis, L. Prokopy, K. Komives, R. Thorsten, H. Lukacs, A. Bakalian, and W. Wakeman. How well is the demand-driven, community management model for rural water supply systems doing? evidence from bolivia, peru and ghana. *Water Policy*, 11(6):696–718, 2009.