Detection of Non-functional Bore wells Using Machine Learning Algorithms

T Janani Priya¹, K P Sabari Priya², Jennifer Priscilla Vincent³, K V Uma⁴ and S Pudumalar⁵

¹, ², ³ Student, Department of Information Technology, Thiagarajar College of Engineering, Madurai, India
⁴, ⁵ Assistant Professor, Department of Information Technology, Thiagarajar College of Engineering, Madurai, India

Email: tjananipriya2000@gmail.com, kp.sabaripriya@gmail.com, jennifervincent2001@gmail.com, kvuit@tce.edu, spmit@tce.edu

Abstract. India is the largest user of ground water with approximately 27.5 million bore wells drawing over 230 cubic kilolitre per year. Bore wells are vertical drilled wells, bored into an underground water-bearing layer in the earth’s surface, to extract water for various purposes. Less rainfall, water scarcity, depletion of underground water has spearheaded the number of bore wells dug in a year. Excessive number of bore wells has led to squandering of groundwater at higher rates than the rate of water replenished and caused depletion of the groundwater levels. When the water gets dried, the motor is removed and the outer surface is not properly covered or sealed. The diameter is large enough for the child to fall inside. The inside of the bore well now non-functioning or left unused might have collapsed. According to a report, since 2009 more than 40 children fell into bore well, sadly seventy percentage of the rescue operations fail. The current paper incorporates the analysis of datasets of bore well acquired from Kaggle. A predictive model has been built by employing machine learning algorithms like Random forest, Extra-trees classifier, logistic regression ‘to predict the non-functional bore wells present which is to be reported and to be taken action in the immediate future’. The performance of the models have been evaluated by performance metrics, among which Random forest shows excellent performance in predicting non-functional killer bore wells.

1. Introduction

1.1. Bore wells in India

In India the main source for irrigation is ground water. Farmers and people in rural regions of India depend on bore wells due to the unavailability of other water resources and lack of water supply. Most of the government schemes aims to supply water through bore wells. Due to ground water level depletion, most of these bore wells become abandoned and are left open without a proper closure.

A survey by the Ministry of Water Resources had discovered that 85 % of rural, 50 % of urban drinking and industrial needs, and 55 % of irrigation needs were satisfied through bore wells. The increasing population and urbanization has led to the deeper drilling of these vertical wells into the underground which exploits the groundwater at higher rates than its replenishment. To prevent the depletion of groundwater level, the government has come up with laws and statutory authority to regulate and monitor groundwater utilization.

However, these bore wells later fail to yield water. The unused bore well might be partially sealed and left as it stands leaving vegetation to take over the spot. These abandoned bore wells are forgotten but their diameter is large enough for a child to fall inside. However, the time taken to realize that the
lost child could have fallen into the bore well is pretty delayed. The inside of the bore well now non-functioning or left unused might have collapsed.

The percentage of bore well accidents occurred in different states of India is shown in figure 1. These recent tragic incidents has now forced us to look after the matter seriously. According to National Disaster Response Force, in the consecutive years from 2006, still more than 33 deaths occurred and 92% of them are under the age of 10. The incident of losing lives trapped in bore well was known widespread in 2006 where a 5-years old child named Prince was rescued after a tough rescue combat which lasted 49 hours. Subsequently, there were number of incidents in the various parts of the country. The last reported incident occurred in October 2019 in which a 3-years-old child Sujith Wilson died despite of the rescue operation which lasted for about 80 hours which again reminded the seriousness of the issue.

![Figure 1: State-wise bore well accidents.](image)

### 1.2. Machine Learning

Machine Learning (ML) is the technology which learns from the dataset. It provides ability to a system to learn and improve without explicit programming. Machine Learning takes data as input and generates rules/patterns which is used for getting inference from the data. ML can be categories into 3 types such as supervised, unsupervised and Reinforcement learning.

Supervised ML algorithms learn the relationship between the target variable (feature) and the predictors (independent feature). It contains target feature (labelled data). From the dataset, using various models, it accurately predicts the future observations. Supervised model can be of classification or regression type. In classification, the target feature is the categorical value whereas in regression the target value is of continuous value.

In the unsupervised Machine learning algorithm, it learns the structure of the dataset. It does not contain target feature. Here instead of predicting the target feature, it will discover the patterns/structure of the data using trained model. Models using for unsupervised learning are k-means clustering, Hierarchical clustering etc.

In reinforcement learning the machine makes trial and error to find solution. It learns form the past experience and tries to capture the best knowledge.

### 2. Objective

- To develop a predictive model using machine learning algorithms like Logistic regression, Random forest, Extra-trees classifier and evaluate their performance.
- To predict the non-functional bore wells that require immediate action.
3. Literature survey
Neda Bihamta Toosi, Ali Reza Soffianian, Sima Fakheran, Saeid Pourmanafi, Christian Ginzler, Lars T. Waser (2019) used four classification algorithms: SVM Radial, RDA, RF, SVM Linear to classify land cover and tested the classification accuracies using K-folds cross-validation and independent validation.

Manal Blnkhonain, Liping Zhao (2019) reviewed machine learning algorithms for identification and classification of non-functional requirements. They used 24 ML based approaches, 16 different ML algorithms and evaluated using 7 different performance metrics.

Chiranth Hegde, Harry MillWater, Ken Gray (2019), classified drilling stick slip index using four algorithms, logistic regression, SVM, random forest and GMMs and evaluated the algorithms using F1 score, accuracy, AUC-ROC.

Arun D Kulkarni and Barrett Lowe (2016) used Random Forest Algorithm for Land Cover Analysis. In their paper “Random Forest Algorithm for Land Cover Classification”, they compare classification accuracy of Random Forest with other commonly used algorithms such as maximum likelihood, minimum distance, decision tree, neural network and support vector machine classifiers.

Nouf Rahimi, Fathy Elasa and Lamiaa Elrefaei (2020) introduce a new ensemble ML technique for classifying FR statements to improve their accuracy and availability. This technique combines different ML models (Naïve Bayes, Support Vector Machine, Decision Tree, Logistic Regression and Support Vector Classification) and uses enhanced accuracy. The accuracy of classifying FR was 99.45% and the required time was 0.7 s.

L Mary Gladence, M Karthi and V Maria Anu (2015) – compare Naïve Bayes and Logistic Regression by using some standard dataset results as input and also show how they differ from each other in terms of performance factor. They also show various Bayes classifiers like Bayes Network, Naïve Bayes, Naïve Bayes Multinomial Text and Naïve Bayes Updateable are working and how they differ with each other based on given data and these results are effectively compared with Logistic Regression.

4. Proposed methodology

4.1. System requirements and tools
- Processor: Intel (R) Core(TM) 2 Duo CPU E7500
- Installed Memory: 8.00 GB
- System type: 64-bit Operating System
- Tools used: Python3, Jupyter notebook

4.2. Dataset description

The dataset is taken from Kaggle.com. This site provides various data including health care, education, infrastructure and commerce etc. We have collected all the datasets related to bore wells. We have collected two datasets which can be matched by the Bore well ID. The first dataset has 40 attributes including id, construction year, extraction type, region code, ward, Population etc. The second dataset contains bore well id and the functionality (functional / non-functional) of the bore well. The dataset acquired is a detailed dataset of bore wells in African Villages. Real-time data in can also be collected and processed using real-time computing and deliver near-instantaneous predictions.
4.3. System design

![System design diagram](image)

**Figure 2. System design**

4.4. Training the data

The data set is pre-processed to convert raw data to complete and consistent data for efficient application of algorithms. The missing values are handled. StandardScaler is fitted to the data frame to transform the data with a mean as 0 and standard deviation as 1 to make it internally consistent. Irrelevant features are removed by applying feature selection using Sequential Forward Selection (SFS). For 59400 records, we have divided the dataset into 80% of training set and 20% of testing data. Python packages are used to build the predictive model.

4.5. Classification model

4.5.1. Logistic regression. It is one of the supervised Machine Learning models which predicts classification task. It ranges in the form of probabilities which lies between 0 and 1. Output value is generated by applying log of x-value using logistic function. The aim of logistic regression is to find the best fitting model to describe the relationship between the characteristic of dependent variable and a set of independent variables.

*H(z) = \frac{1}{1+e^{-z}}*

4.5.2. Extra-trees classifier. It is an ensemble-based algorithm which constructs multiple decision trees and correlates to obtain the classification result. Each decision tree is constructed from the original sample and also allows to bootstrap replicas. The cut points of a decision tree is chosen randomly in extra trees classifier.

4.5.3. Random forest. It is an ensemble-based algorithm for classification and regression. This algorithm measures the relative significance of each feature in the process of prediction and creates many decision trees contributing to a robust forest. The output class is determined by the mode of the classes.

Comparative analysis of all the above machine learning algorithms are carried out. The algorithms are analysed in terms of accuracy, precision, recall and f1 score. The one with the best figures are carried forward to predict the non-functional bore wells that require immediate action.
5. Results and discussions

5.1. Data visualization

Data visualization is the process of graphically representing the given data to understand complex patterns in data sets.

**Figure 3.** The correlation matrix of the various attributes of the dataset. The attributes region_code and district_code are strongly correlated same as construction_year and population. Thus any one of the attribute in the correlated pair can be dropped to improve accuracy and reduce training time.

**Figure 4.** The visualization of the attribute source type shows the class spring is the major source type for 28.6% bore wells.
6

Figure 5. The visualization of attribute extraction type class shows a maximum of 17% of the bore wells constructed have gravity and a minimum of less than 1% of bore wells have rope pump as their extraction type.

5.2. Performance metrics
The performance of the algorithms is evaluated using accuracy, precision, recall and f1 score.

- Classification accuracy is the ratio of correct to the total number of predictions made.
- Precision is the correct predictions among classified positive instances.
- Recall is the correct predictions of positive instances.
- F1 score ranges from 0 to 1 with 0 as the worst and 1 as its best case representing the weighted harmonic mean of precision and recall.

On comparing the performance of algorithms using various performance metrics, Random Forest shows greater efficiency in classification of bore wells. The limitation of Extra-trees classifier is that the features selected and the split points are chosen at random rather than choosing the optimum value. Hence decreasing the overall efficiency of the model compared to random forest. The logistic regression does not work well with larger datasets with higher dimensions as it may lead to over fitting of the model. It is difficult for a logistic regression algorithm to apprehend complex patterns in a data and requires the variables to be related linearly. Thus Random Forest model is carried forward to predict the non-functional bore wells that require action to be sealed immediately.

6. Conclusion
Bore wells are currently the most common source of water in India. The two major problems it has contributed is the depletion of underground water levels and children losing their lives in the deadly holes. Various statistical analyses of water quantity, extraction type, source of water etc., have been explored and represented. The government should concentrate on imposing strict measures to take immediate action to seal the dangerous non-functional bore wells as projected in the analysis made.
Predicting the non-functional bore wells is a challenging task since it involves a lot of physical and geographical factors. Machine learning algorithms like logistic regression, Random forest and Extra-tree classifiers are applied and tested. For the classification, the Random Forest model shows excellent accuracy in predicting non-functional bore wells. Developing a predictive model for non-functional bore wells will enable to identify risk at early and further encourage the government to take prompt measures to properly seal those wells. Our future work is to predict the functionality of the bore wells by collecting real-time large dataset and applying optimization techniques.

References

[1] Manal Binkhonain, Liping Zhao (2019), A review of machine learning algorithms for identification and classification of non-functional requirements.
[2] Breiman L(2001), Random Forests. Machine Learning.
[3] Arun D Kulkarni and Barrett Love(2016)- “Random Forest Algorithm for Land Cover Classification”
[4] Nouf Rahimi , Fathy Eassa and Lamiaa Elrefaei (2020)-“An Ensemble Machine Learning Technique for Functional Requirement Classification”
[5] L Mary Gladence , M Karthi and V Maria Anu(2015)- “A Statistical Comparison of Logistic Regression and Different Bayes Classification Methods for Machine Learning”
[6] Chiranth Hegde,Harry MillWater, Ken Gray (2019), Classification of drilling stick slip severity using machine learning.
[7] Strobl, C., Boulesteix, AL., Zeileis, A. et al. Bias(2007) in random forest variable importance measures: Illustrations, sources and a solution.
[8] Lei Zhang, Falk Huettmann, Xudong Zhang, Shirong Hui, Pengsen Sun, Zhen Yu, Chunrong Mi(2019), The use of classification and regression algorithms using the random forests method with presence-only data to model species’ distribution.
[9] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP(2003), Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.
[10] Amit, Y., Blanchard, G., & Wilder, K. (1999), Multiple randomized classifiers: MRCL Technical Report, Department of Statistics, University of Chicago.
[11] Neda Biahamta Toosi, Ali Reza Soffianian,Sima Fakheran, Saeid Pourmanafi,Christian Ginzler,Lars T.Waser (2019), Comparing different classification algorithms for monitoring mangrove cover changes in southern Iran.
[12] Mareike Ließ, Bruno Glaser, Bernd Huwe (2012), Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models
[13] Nisha Arora, Pankaj Deep Kaur (2020), A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment.
[14] Qi Y, Bar-Joseph Z, Klein-Seetharaman J(2006), Evaluation of Different Biological Data and Computational Classification Methods for Use in Protein Interaction Prediction.
[15] Caruana, R., and Niculescu-Mizil, A. (2006). “An empirical comparison of supervised learning algorithms,”
[16] C. Chen, A. Liaw, and L. Breiman (2004), Using random forests to learn imbalanced data.
[17] Dietterich, T. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization
[18] Fu B, Wang Y, Campbell A (2017) Comparison of object-based and pixel-based Random Forest algorithm for wetland vegetation mapping using high spatial resolution GF-1 and SAR data.
[19] Manish Raj, P.Chakraborty, G.C.Nandi (2014), Rescue Robotics in Bore well Environment.