# Department of Electrical and Computer Engineering
# North South University



## Directed Research

## Detecting Arsenic of Ground Water

**Submitted By:**

| Name | ID |
|------|-----|
| Paplu Dash | 1611420043 |
| Tanbin Akter Mitaly | 1712438642 |

**Faculty Advisor:**
**Dr. Ahsanur Rahman**
**Assistant Professor**
**ECE Department**
**Fall,2020**

## Declaration

This is to declare that no part of this report or the project has been previously submitted elsewhere for the fulfilment of any other degree or program. Proper acknowledgement has been provided for any material that has been taken from previously published sources in the bibliography section of this report.

......................................................

Paplu Dash

ECE Department

North South University, Bangladesh

......................................................

Tanbin Akter Mitaly

ECE Department

North South University, Bangladesh

# Approval

The Directed Research Project entitled "Detecting Arsenic of ground water" by Tanbin Akter Mitaly (1712438642) ,Paplu Dash (1611420043), has been accepted as satisfactory and approved for partial fulfilment of the requirement of BS in CSE degree program on June 2019.

**Supervisor's Signature**

**Dr. Ahsanur Rahman**

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

**Department Chair's Signature**

**Dr. Rezaul Bari**

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

# Acknowledgement

First of all, we would like to express our profound gratitude to our honourable course instructor,

**Dr. Ahsanur Rahman** for his constant and meticulous supervision, valuable suggestions, his patience and encouragement to complete the thesis work.

We would also like to thank the ECE department of North South University for providing us with the opportunity to have an industrial level design experience as part of our curriculum for the undergraduate program.

# Introduction:

Arsenic contamination of ground water in Bangladesh is one of the main concerned issue. Tubewell water is the main source of drinking water in the rural areas of Bangladesh and Arsenic contamination is found in it. It has become a health hazard due to drinking the Arsenic contaminated water regularly. Then Government tried to control it by painting the tubewell with red and green to mark safe and unsafe. Then planting deep tube well is another solution. To know which areas of a region is highly to have Arsenic contamination in ground water with the known values of other geometric factor to determine is a needed solution for to cope with the problem.

# Related Work:

The work of Amini *et al.* worked with 2000 data of arsenic concentration that has been collected from around the world and they used digital map for the characteristic of soil, geology, climate and elevation that used for to produce a probability maps. They used regression and neuro-fuzzy .where they have gotten 77% by using ANFIS, 63% by using linear regression.

# Methodology:

First of all we have some missing values in our dataset, to resolve it we used interpolation, regression and datawig. Then we compared with the accuracy by using the data from interpolation and datawig. We tried in two way one classification another is regression. Classification gives us specific class where regression give continues data. First we needed to pre-process data as our arsenic column had continues value that's why we converted it to two class safe ad unsafe. Then used several scaling but the most accuracy we have gotten from is random forest classifier and it does not have any impact of scaling. We used 10 fold of cross validation for to train all the model.

# Result:

| Model | Accuracy |
|---|---|
| BayesNet | 85.6 |
| NaiveBayes | 90 |
| Logistic | 81.9 |
| Multilayer Perceptron | 91.9 |
| Simple Logistic | 90 |
| IBK | 88.1 |
| Kstar | 90.6 |
| Adaboost | 88.75 |
| Bagging | 89.375 |
| Decision Table | 87.5 |
| HoeffdingTree | 93.125 |
| LMT | 90 |
| RandomForest | 93.125 |
| RandomForest,Max_depth=4 | 94.375 |
| RepTree | 86.875 |

Table 1: Classification Model accuracy with missing value computed using datawig

The first table is showing the accuracy for the data's that was missing calculated using datawig. We have gotten the maximum from random forest classifier with 100 random decision tress that is our ensemble parameter with 42 random state that gave us the optimum result. For the other model like Multilayer Perceptron and other are less than 94 but still gotten a high accuracy.

| Model | Accuracy |
|---|---|
| BayesNet | 86.125 |
| NaiveBayes | 90.625 |
| Logistic | 85 |
| Multilayer Perceptron | 93.75 |
| Simple Logistic | 91.125 |
| IBK | 88.1 |
| Kstar | 91.125 |
| Adaboost | 88.125 |
| Bagging | 89.75 |
| Decision Table | 88.125 |
| HoeffdingTree | 92.5 |
| LMT | 91.25 |
| RandomForest,Max_depth=0 | 95 |
| RandomForest,Max_depth=4 | 94.375 |
| RepTree | 86.875 |

Table 2: Classification Model accuracy with missing value computed using interpolation and regression

For the table 2 the accuracy we have gotten is from using the missing value by using interpolation and regression to compute the missing values. The difference in accuracy does not have that much change compared to the data that we have got from imputing with datawig.

| Model | Accuracy |
|---|---|
| RepTree | 80 |
| RandomTree | 74 |
| RandomForest | 90 |
| M5p | 89.5 |
| DecisionStamp | 78 |
| Decision Table | 75 |
| M5 Rules | 84 |
| Additive Regression | 81 |
| Bagging | 85 |
| Linear regeression | 87 |
| Gaussian process | 86 |

Table 3: Regression Model accuracy with missing value computed using datawig

The third table shows us the accuracy for the regression model. The missing value computed using datawig. The most accuracy we have got from is Random Forest Regression model where we have use 100 random decision tree to compute the random forest. With cv fold 10

| Model | Accuracy |
|---|---|
| RepTree | 80.7 |
| RandomTree | 77.9 |
| RandomForest | 89.5 |
| M5p | 88.12 |
| DecisionStamp | 78.33 |
| Decision Table | 76 |
| M5 Rules | 81.94 |
| Additive Regression | 84 |
| Bagging | 84.92 |
| Linear regeression | 84.92 |
| Gaussian Process | 87.4 |

Table 4: Regression Model accuracy with missing value computed using interpolation and regression

Table 4 is showing the accuracy of our regression models with dataset missing value computed using interpolation and regression. The most accuracy we got from the random forest regression that is 89.5

## Conclusion:

Having safe drink water is one the primary need for citizen of any country. Bangladesh, India and 48 more countries are suffering from the issues of arsenic contaminated water. To having a knowledge geometrically which regions in our countries' ground water are mainly contaminated with arsenic and to classify them into safe and unsafe our work have done.

## Bibliography:

1. Amini, M., Abbaspour, K.C., Berg, M., Winkel, L., Hug, J.S., Hoehn, E., Yang, H., Johnson, C.A., 2008. Statistical modeling of global geogenic arsenic contamination in groundwater. Environ. Sci. Technol. 42, 3669–3675.