PREDICTIVE MODELLING OF GROUNDWATER FLUORIDE CONTAMINATION

USING MACHINE LEARNING MODELS

GROUP MEMBERS

HARSHIT SHARMA 9919103016

YASHA JAFRI 9919103009

SIDDHARTH SINGH 9919103029

PRANAT JAIN 9919103017

MENTOR **DR. SHIKHA K MEHTA**

PROBLEM STATEMENT

India's Groundwater Overexploitation has led to a grave problem of - **Whether the Groundwater being consumed is safe or not?** There are various **contaminants like Arsenic, Fluoride, Hydrogen Sulfide, etc.** which have been responsible for the situation. Intake of these contaminants, even in small amounts, **can lead to Mass Health Crisis**.

Among these contaminants, Fluoride has an importance to Human Health. **WHO recommends fluoride levels of 0.5 - 1.0 mg/L**, with an exception of upto 1.5 mg/L. Measuring Fluoride levels in Groundwater in the whole country is a near-to-impossible task. We **can only use the surface parameters** for this purpose.

Our task is to **identify the surface parameters** responsible for the accumulation of Fluoride in Groundwater and **use these values** to **train a model** using different ML algorithms such that it can **predict Fluoride levels with high accuracy** for the values of parameters being considered.

STATE-OF-THE-ART

The technologies that we have used in our project are:-

- Python 3.8
- Python libraries such as:-
  - NumPy (1.21.4)
  - Pandas (1.3.4)
  - Scikit-Learn (1. 0.20)
  - Rasterlo (1.2)
- Microsoft Excel 2019
- Google Colab Notebooks (12 GB RAM)

LIMITATIONS

There is some limitations that we have to consider in our project:-

- According to the research papers that we had studied, there are nearly 15 factors on which fluoride concentration of an area depends out of which we are able to use only 6 factors this is because of the unavailability of correct resolution. (For example, we have used 30 arcs second format dataset in our project and some of the datasets are not available in this resolution).
- Another reason is the low computational power of Colab Notebooks , which cause RAM error when we tried to extract data from the dataset
- In some cases, the data we got from the internet was not for India so we were not able to include that in our project

OBJECTIVES & WORK DISTRIBUTION

- Learn the basics of Python & packages like NumPy, Pandas, etc.
- Learn the basics of Machine Learning & Predictive Modelling
- Understanding the GIS format, TIFF files and reading Data from them
- Implementation of related Machine Learning Models using the Dataset
- Improve the accuracy of current predictions

HARSHIT SHARMA - Design Ideas, Data Organization & Preparation, Model implementation, Training & Testing the model for Collected Data

SIDDHARTH SINGH - Dataset Collection, Dataset Organisation, Background Study and Research Paper Scanning, Model Implementation

YASHA JAFRI - Data organization, Background Study and Research Paper Scanning, Report & Presentation preparation

PRANAT JAIN Dataset Collection, Dataset preparation, Background studies, and research paper scanning, Report & Presentation preparation

PROPOSED DESIGN

Data points, compiled from **CGWB (Central Ground Water Board)** of India & few other sources, were **assigned to 1-sq.km resolution** (similar to the resolution of predictor variables available) against their fluoride concentrations for Modeling.

The Collected **Dataset is converted to low & high classes** of fluoride concentrations, where 0 is assigned to all concentrations <= 1.0 mg/L and 1 to all concentrations > 1.0 mg/L, **also maintaining a separate column of absolute values**, which then will be **divided into 80% Training Data and 20% Testing Data**, to train the model..

After collecting and organizing the data, **implement Machine Learning Models** on the data and make predictions. The ML Models to be implemented are **Random Forest and XGBoost**. These both models **provide Regressor as well as Classifier Models**.

Calculate the Prediction Accuracy of all these models and compare them with each other. The Scikit-learn module provides for built-in function to test the accuracy of these models.

IMPLEMENTATION

For Implementation purposes, the tasks we had to work upon were -

- **Understanding the GeoTIFF format used in GIS Data**. Most of the data-in-use followed the EPSG:4326 format. But, every dataset had slight variations like non-uniform coordinate distribution, or RAM problems. So, we had to lay off some predictor variables.
- **Implementing the Rasterio Python library to extract the files**, so that it can be operated upon to extract desired values. This library can be used to convert GeoTIFF files into ND-arrays, which can be easily operated upon. It is a new library we all had to work with. The documentation of this library is also not detailed. So, it was a challenge for us to figure our way out of the problems.
- **Working with ASCII files**. While collecting datasets, we also worked with ASCII files, which required a different approach for extracting the data values.

- **Working with Microsoft Excel files to filter out data collected to feed into the Model**. After all the data was collected, we needed to organise it for Model Implementation. MS Excel is a brilliant tool for the same. It provided various operations to work with Data Values.The Dataset of Coordinates for Observation Wells across India contained Geographical Coordinates in degrees, which could not be used directly for Implementation. These Coordinates had to be filtered and converted to decimal. Using Python for this process could have been troublesome, but Excel provides better features to deal with such data. Also, finding Null values, incorrect data, etc. is easier in Excel.

- **Exploring the Random Forest & XGBoost Algorithms for Regression and Classification** purposes & their implementation. Random Forest Algorithm is a multi-purpose algorithm with better accuracies than many models. XGBoost also has produced some great results when implemented. Thus, we used these 2 algorithms.

- **Predicting Accuracy and other Statistics**. While implementing these 2 algorithms, we checked the accuracy of the data for two variations of the models used - one with 250 estimator variations and the other with 500 estimator variations.

EXPERIMENTAL RESULTS & ANALYSIS

We had started with collecting 15 predictor variables, belonging to 9 different classes, but due to Resolution limitations or System Limitations, could only collect 6 different predictor variables of 5 different classes. But, this did not affect the accuracy of predictions much.

As explained earlier, we had used 2 variations of Regression & Classification Model, with 250 & 500 estimator values respectively. The results we achieved are described are as follows:

The Statistics of the 4 Classification Models have been depicted in the table below:

PREDICTION PARAMETER	RANDOM FOREST MODEL - 1	XGBOOST MODEL - 1
Accuracy	85.04259	84.55392

PREDICTION PARAMETER	RANDOM FOREST MODEL - 2	XGBOOST MODEL - 2
Accuracy	85.02278	84.73222

The Statistics of the 4 Regression Models have been depicted in the table below:

PREDICTION PARAMETER	RANDOM FOREST MODEL - 1	XGBOOST MODEL - 1
Mean Absolute Error	0.40358	0.41723
Mean Squared Error	1.48976	1.37457
Root Mean Squared Error	1.22056	1.17242

PREDICTION PARAMETER	RANDOM FOREST MODEL - 2	XGBOOST MODEL - 2
Mean Absolute Error	0.40260	0.41544
Mean Squared Error	1.47964	1.37553
Root Mean Squared Error	1.21641	1.17283

The results for the Regression & Classification models have detailed in the above tables. These Statistics reveal similar results have been achieved by both the Models. Hence, cross-verifying each other.

CONCLUSION

Although, there are **several predictor variables** that can enhance the accuracy of predictions in Fluoride Contamination, the **six predictors**, namely – **Actual Evapotranspiration**, **Potential Evapotranspiration**, **Soil pH**, **Slope** (0-60% in Majority area of the land), **Aridity** and **Precipitation** are enough to easily predict the levels of Fluoride.

It can also be concluded that **Random Forest Modelling and XGBoost Modelling**, both produce **nearly equal results**. An algorithm might perform better for Regression, while the another algorithm might perform better for Classification. These both **can be used interchangeably** without hesitation.

FUTURE SCOPE

Groundwater is excessively used in India and Fluoride contamination in Groundwater is a growing concern these days. It is **important to pinpoint these locations of high risk** and **provide remedial measures**, in order to free the population of the hazardous effects of diseases like Fluorosis. Also, **India has set a target to reach the National average of 1.0 mg/L** fluoride in Groundwater from the existing 1.5 mg/L mark. So, this project has a societal benefit added to it.

Our Project helps us to achieve both these goals. This project **can also be expanded to many other Groundwater contaminants** like Arsenic & Nitrate which in-turn can help to provide quality water to every household in India. This project **can be envisioned to improve the current accuracy of predictions**.

THANK YOU !