Phase 3 Submission: CSE 4/587 Data Intensive Computing

WATER QUALITY PREDICTOR

Submitted By:
Maitrai Sharma (50476049)
Mancy Saxena(50477980)

Phase 3 Submission: CSE 4/587 Data Intensive Computing

1. a. Code and Working Instructions to demo/use the finished product

The code consists of 2 parts in the /src folder

a. A ipynb notebook, which is fully documented and has detailed analysis. For our analysis, we have chosen the "Random Forest" for our problem statement.
b. Our product is a webpage that is implemented using Python's Flask framework can be found in the /src/app folder.

To use the demo :
Type the following commands on Windows machine to run flask
cd src/web app
.env\Scripts\Activate
cd app
On the terminal then type *flask run* to run the webapp

To run the application on Linux/Mac machine :
cd src/web_app
source .env/bin/activate
cd app
flask run

b. Model tuned and finalized:
Model Selection: Based on the problem statement mentioned above, the best performing was the Random Forest based on accuracy. However, the observations had more records for is_safe =0 values than is_safe =1, which was our target variable in determining if the water is safe or not. So we look at other evaluation metrics, such as precision, recall and f1-score

With the other metrics, the SVM (Support Vector Machine) model has the highest precision value and best f1 and recall values as well. So we are going with our SVM model

Evaluation Metrics :
The f1 score which is to assess the model's performance by considering both precision and recall is 0.97 for class is_safe=0 and 0.23 for class is_safe=1. Precision which is measure the accuracy of the positive predictions made by a model is 0.97 for class is_safe = 0 and 0.23 for class is_safe =1 . Lastly, recall which is to measure the ability of a model to correctly identify positive instances from the total actual positive instances in the dataset is 1 for class is_safe =0 and 0.14 for is_safe =1

c. Recommendations Related to the problem statement based on the analysis

The application can be used by laboratories to have an initial analysis of if the water is safe for usage. For comprehensive water analysis, laboratories can take up to several days to predict if water is fit for consumption. With our web app, the users can perform chemical tests, and based on the chemical tests they can predict if the water is contaminated or not.

There are the following features of the app:

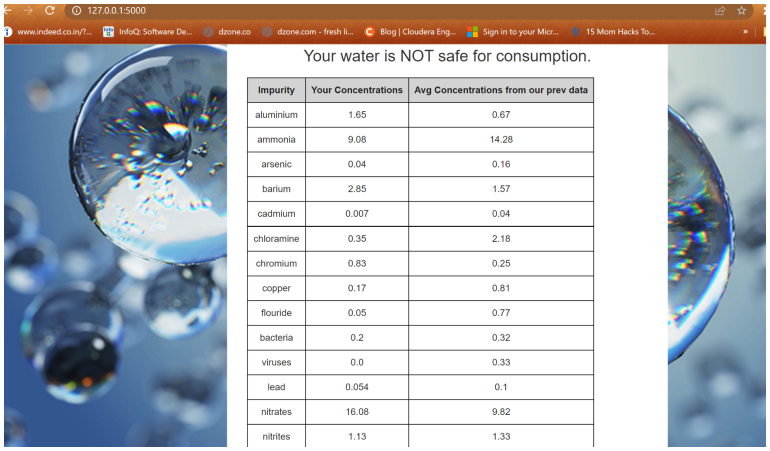
a. Quick Initial Analysis: The app can expedite the process of analyzing the water sample quicker, in seconds as opposed to traditional methods where laboratories may take up to several days to weeks to inform and predict if data is fit for consumption
b. Accessibility and Convenience: This tool allows naive users to access their house water quality and instantly learn about their household water quality / or can be used by farmers to check if water is fit for irrigation
c. Education and Learning: The easy-to-use portal with the prediction model can help people understand the various kinds of contamination present in the water and the range of contamination that is safe and unsafe for consumption

There are a few extensions possible with this project

a. Integration with IOT: IOT devices can be integrated to check contamination on a continuous/ periodic basis and alert if the contamination levels change
b. Mapping Water Quality geographically: To map the quality of water with the cities and check visually the contamination levels geographically
c. This can be extended to include more / lesser parameters and higher accuracy/ precision, f1 and recall values
d. Data Sharing Options: Sharing on various platforms can be done to provide the users to share their contamination status on social platforms such as Twitter etc and alert their local public water works department, effectively

Web App Screenshots:

Phase 3 Submission: CSE 4/587 Data Intensive Computing



Impurity	Your Concentrations	Avg Concentrations from our prev data
aluminium	1.65	0.67
ammonia	9.08	14.28
arsenic	0.04	0.16
barium	2.85	1.57
cadmium	0.007	0.04
chloramine	0.35	2.18
chromium	0.83	0.25
copper	0.17	0.81
flouride	0.05	0.77
bacteria	0.2	0.32
viruses	0.0	0.33
lead	0.054	0.1
nitrates	16.08	9.82
nitrites	1.13	1.33